Statistik

Corona - Notversion

M. Oettinger

15. April 2020

Altersverteilung

In der nachfolgenden Tabelle sind als einfaches Beispiel das Geschlecht, die Körpergröße und das Alter der Teilnehmer eines Statistik-Kurses in Ravensburg aufgeführt.

w	180	18
w	168	28
w	167	23
w	176	20
w	168	20
w	162	19
w	166	20
m	183	21
m	175	29
w	168	21
w	172	19
w	164	21
w	165	20
w	177	21

Nullstellen

Die relevante Information des Merkmals Alter kürzer durch eine sog. *Urliste*, einen Vektor der einzelnen Daten, dargestellt:

$$\{x_i\} = (18, 28, 23, 20, 20, 19, 20, 21, 29, 21, 19, 21, 20, 21, 19,$$

$$21, 19, 20, 22, 20, 21, 24, 19, 22, 19, 25, 40)$$

Dabei wird noch nicht auf Information verzichtet, sofern die Reihenfolge der Personen der Reihenfolge der Daten in der Urliste entspricht. Ebensogut könnte jedoch ein Vektor von $n = 32$ Zahlen benutzt werden, der die Alterswerte bereits in geordneter Form enthält:

$$(18, 19, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, 21, 21,$$

$$21, 21, 21, 21, 22, 22, 23, 24, 25, 28, 29, 40)$$

Häufigkeiten

Tabelle der *absoluten* Häufigkeiten h_i bzw. *relativen* Häufigkeiten $f_i = h_i/n$ für den i-ten in der Stichprobe auftretenden Wert:

Alter x_i	Häufigkeit h_i	$h_i \cdot x_i$	rel. Häufigkeit f_i
18	1	18	0.037
19	6	114	0.222
20	6	120	0.222
21	6	126	0.222
22	2	44	0.074
23	1	23	0.037
24	1	24	0.037
25	1	25	0.037
28	1	28	0.037
29	1	29	0.037
40	1	40	0.037
Summe	591		1,00

Tabelle: absolute Häufigkeiten h_i und relative Häufigkeiten f_i zum Alter.

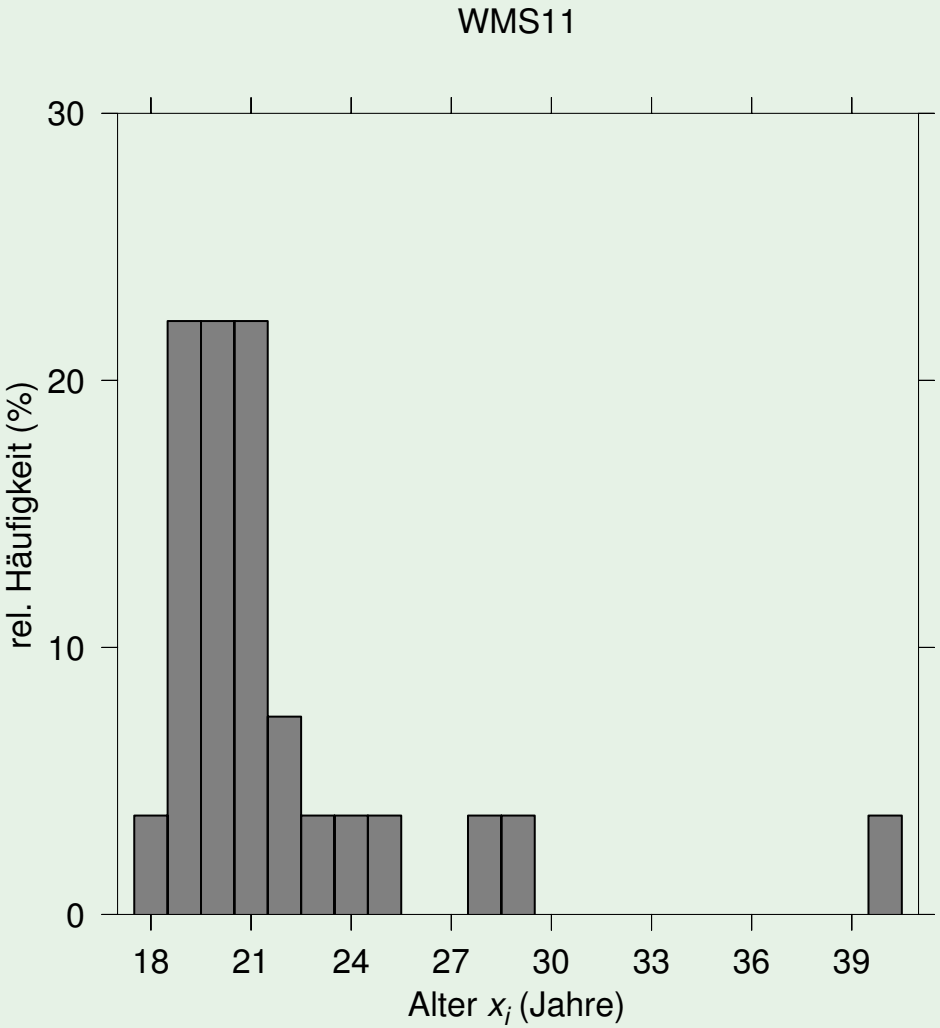
Häufigkeiten

In einer Stichprobe vom Umfang n vorhandene Information eines kardinalen Merkmals X kann allgemein natürlich in Form einer Urliste $(x_1, x_2, x_3, ...x_n)$ der bereits geordneten Werte angegeben werden. Treten einzelne Merkmalswerte mehrfach in einer Stichprobe auf, so dass nur $m < n$ der Merkmalswerte verschieden sind, werden sie in einem Vektor $(x_1, x_2, x_3, ...x_m)$ zusammengefasst, der dann lediglich die verschiedenen auftretenden Merkmalswerte enthält. Um die erhobene Information wiederzugeben, muss zusätzlich ein Vektor der auftretenden absoluten oder relativen Häufigkeiten $(h_1, h_2, h_3, ...h_m)$ bzw. $(f_1, f_2, f_3, ...f_m)$ angegeben werden (absolute bzw. relative Häufigkeitsverteilung).

X	x_1	x_2	x_3	...	x_m	Summe
$h_i = h(X = x_i)$	h_1	h_2	h_3	...	h_m	n
$f_i = f(X = x_i)$	f_1	f_2	f_3	...	f_m	1

Beispiel: Balkendiagramm der Altersverteilung im Kurs.

Als Balkendiagramm ergibt sich folgendes Bild:



statistische Analyse

Statistische Parameter (auch statistische Maßzahlen) sind charakteristische Werte, die eine Menge von Beobachtungen *einfach* beschreiben. Zweck ist die Verdichtung von Daten einer Stichprobe in einzelne, möglichst einfache Parameter. Dabei wird Information vernichtet, aber Übersichtlichkeit gewonnen! Für eine Menge von Beobachtungen lassen sich viele solcher Maßzahlen angeben, wir behandeln einige der am häufigsten benutzten.

Lagemaße: geben für eine Stichprobe repräsentative, typische Werte an (beispielsweise einen Durchschnittswert)

Streuungsmaße: geben an wie dicht (oder wie weit entfernt) einzelne Merkmalswerte bei einem Mittelwert liegen

Schiefemaße: liefern Information über die Symmetrie oder Asymmetrie einer Verteilung von Daten

Lagemaße

Lagemaße sind Werte, die für eine gegebene Stichprobe einen einzelnen, für die vorliegenden Daten repräsentativen Wert angeben, beispielsweise einen Mittelwert.
Sie müssen dabei nicht selbst Werte aus dem Bestand des vorliegenden Datenmaterials sein. So spricht beispielsweise bei einer Erhebung von Lebensaltern in ganzen Jahren nichts gegen einen Mittelwert, der als Wert zwischen zwei vollen Jahren angegeben wird.

Lagemaße

Das arithmetische Mittel ist der am weitesten verbreitete Mittelwert. Das arithmetische Mittel kann nur für kardinale Merkmale berechnet werden, wird aber auch für ordinale Merkmale verwendet (unsinnig!)

Definition: arithmetisches Mittel (AM)

Das AM ist die Summe der Merkmalswerte geteilt durch ihre Anzahl:

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

Einschub: Summe

Definition: Summe

Für $n, m \in \mathbb{Z}, m <= n$ und $x_i \in \mathbb{R}$ für $i \in [m; n]$ ist

$$\sum_{i=m}^{n} x_i := x_m + x_{m+1} + \cdots + x_{n-1} + x_n$$

die Summe aller x_i mit $i \in [m; n]$. m heißt untere, n obere Grenze.

Für $x_1 = 1, x_2 = 3, x_3 = 4$ und $x_4 = 6$:

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 1 + 3 + 4 + 6 = 14$$

Lagemaße

Herleitung des AM: Die Summe aller Merkmalswerte ist

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

Als typischen Wert $\bar{x}$ für sie Daten wählen wir den Wert, der n-mal summiert dieselbe Summe S ergibt:

$$\underbrace{\bar{x} + \bar{x} + \bar{x} + \bar{x} + \cdots + \bar{x}}_{n \cdot x} = \sum_{i=1}^{n} \bar{x} = S = \sum_{i=1}^{n} x_i$$

$$\Leftrightarrow n \cdot \bar{x} = \sum_{i=1}^{n} x_i \Leftrightarrow \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Dieser Wert ist das *arithmetische Mittel*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ mit } x_i = (x_1, x_2, x_3, \ldots x_n) \tag{1}$$

Eigenschaften des arithmetischen Mittels

Schwerpunkteigenschaft:

$$n\bar{x} = x_1 + x_2 + x_3 + ... + x_n \Longleftrightarrow x_1 + x_2 + x_3 + ... + x_n - n\bar{x} = 0$$

umsortieren der Summanden liefert

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + ... + (x_n - \bar{x}) + = 0$$

Abweichungen der Einzelwerte vom arithmetischen Mittel heben sich in der Summe auf.

Alternative Berechnung

Kommen einzelne Merkmalswerte mehrfach vor und gibt es in Wirklichkeit nur $m < n$ verschiedene Merkmalswerte $(x_1, x_2, ...x_m)$, die mit den absoluten Häufigkeiten $(h_1, h_2, ...h_m)$ auftreten, so ist das arithmetische Mittel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{m} h_i x_i = \frac{h_1 x_1 + h_2 x_2 + h_3 x_3 + ...h_m x_m}{n} \tag{2}$$

Alternative Berechnung

Manchmal sind relative Häufigkeiten f_i praktischer:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} h_i x_i = \frac{h_1}{n}x_1 + \frac{h_2}{n}x_2 + \ldots \frac{h_m}{n}x_m = \sum_{i=1}^{m}\frac{h_i}{n}x_i = \sum_{i=1}^{m} f_i x_i \qquad (3)$$

Die einzelnen Faktoren f_i können als Faktoren aufgefasst werden, mit denen die jeweiligen Merkmalswerte gewichtet werden. Im allgemeinen sind diese Gewichtungsfaktoren natürlich nicht identisch, man spricht vom *gewogenen* arithmetischen Mittel. Das arithmetische Mittel, das für n verschiedene Einzelwerte $(x_1, x_2, ...x_n)$ gebildet wird, kann ebenfalls als ein gewogenes Mittel betrachtet werden, allerdings mit n identischen Gewichten $1/n$:

Lineare Transformation

Lässt sich ein Merkmal Y über eine allgemeine lineare Transformation

$$Y = a + bX$$

durch ein kardinales Merkmal X ausdrücken, so ergibt sich das arithmetische Mittel $\bar{y}$ des Merkmals Y durch

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

Jeder Wert x_i ergibt durch eine lineare Transformation einen Wert $y_i = a + bx_i$. Das arithmetische Mittel ist also

$$\begin{aligned} \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_i (a + bx_i) = \frac{1}{n}\sum_i a + \frac{1}{n}\sum_i bx_i \\ = \frac{n}{n}a + b\frac{1}{n}\sum_i x_i = a + b\bar{x} \end{aligned} \tag{4}$$

Lineare Transformation

Beispiel: Umrechnung zwischen Fahrenheit und Celsius

Die Temperatur T_F in Grad Fahrenheit ergibt sich aus der Temperatur T_C in Grad Celsius nach der Vorschrift

$$T_F = \frac{9}{5} T_C + 32$$

Die Temperaturen $x_1 = 10, x_2 = 20, x_3 = 30$ Grad Celsius können damit in die Werte $y_1 = 50, y_2 = 68, y_3 = 86$ Grad Fahrenheit umgerechnet werden.

Lineare Transformation

Beispiel: Umrechnung zwischen Fahrenheit und Celsius

Für die arithmetischen Mittel ergeben sich die Werte

$$\bar{x} = 20^\circ\text{Celsius und } \bar{y} = \frac{50 + 68 + 86}{3} = 68^\circ\text{Fahrenheit}$$

Genausogut kann der Mittelwert $\bar{y}$ aber über die lineare Transformation bestimmt werden:

$$\bar{y} = \frac{9}{5}\bar{x} + 32 = \frac{9}{5} \cdot 20 + 32 = 68$$

arithmetisches Mittel

Beispiel: Altersverteilung im Kurs.

die Einzelwerte des Alters der Teilnehmer in Jahren:

i	Alter (x_i)
1	19
2	21
3	19
4	18
5	20
6	19
7	20
8	21
9	21
10	25
11	20
12	20
13	20

Beispiel: Altersverteilung im Kurs.

Zur Berechnung des arithmetischen Mittels $\bar{x}$ wird die Summe der einzelnen Lebensalter durch den Umfang n der Stichprobe (die Zahl der erfassten Personen) geteilt:

$$\bar{x} = \frac{685}{32} = 21,42$$

Beim Vergleich mit der grafischen Darstellung der erhobenen Daten in Abb. 6 wird deutlich, dass der Mittelwert hier nur eingeschränkt sinnvoll eingesetzt werden kann. Der Mittelwert erscheint im Vergleich mit der Grafik als zu groß - er wird durch die Ausreißer auf der rechten Seite zu höherem Alter hin verschoben.

Median

Statt des arithmetischen Mittels ist bei Merkmalsausprägungen mit Rangfolge oft das Konzept des *Medians* (auch Zentralwert) sinnvoll. Der Median kann bei ordinalen oder kardinalen Merkmalen benutzt werden, bei nominalen Merkmalen ist keine Rangfolge festgelegt.

Der Median entspricht in einer Stichprobe der Merkmalsausprägung $\bar{x}_Z$, die das vorhandene Datenmaterial in zwei (gleichgroße) Hälften aufteilt. Er ist also auch ein mittlerer Wert (eine typischer Wert für die vorliegenden Daten).

Median

Definition: Beschreibende Definition des Medians

Der Median $\bar{x}_Z$ ist derjenige Merkmalswert eines kardinalen Merkmals X, den mindestens 50% aller Merkmalswerte einer Stichprobe vom Umfang n unterschreiten oder höchstenfalls erreichen **und** den mindestens 50% aller Merkmalswerte überschreiten oder zumindest erreichen.

Median

Beispiel: Berechnung des Medians.

Als Beispiel ein Vektor $(x_1, x_2, ...x_11)$ von 11 *geordneten* Einzelwerten: beim 6. Wert einer geordneten Reihe von 11 Werten sind stets 6 Werte kleiner oder gleich dem Wert x_6 und ebenso 6 von 11 Werten größer oder gleich x_6:

$$\bar{x}_Z = x_6.$$

Die beschreibende Definition des Medians legt ihn aber leider nicht immer eindeutig fest.

Median

Beispiel: Berechnung des Medians.

Wären statt der 11 Werte im Vektor 12 Werte $(x_1, x_2, ...x_12)$ vorhanden, kämen nach der obigen Definition zwei Werte in Frage - der 6. und der 7. Wert. Als Median wird in diesem Fall das arithmetische Mittel der beiden Werte festgelegt:

$$\bar{x}_Z = \frac{x_6 + x_7}{2}$$

Median

Beispiel: Berechnung des Medians.

Als Beispiel könnte der folgende Vektor von 12 Zahlen vorgelegen haben:

$$(0, 1, 1, 2, 3, 3, 4, 4, 9, 9, 10, 32)$$

Der Median lautet in diesem Fall

$$\bar{x}_Z = \frac{x_6 + x_7}{2} = \frac{3 + 4}{2} = 3,5$$

Dieser Wert teilt die Menge von 12 Zahlen in zwei gleichgroße Hälften, Teilmengen gleichen Umfangs, auf:

$$(0, 1, 1, 2, 3, 3) \text{ und } (4, 4, 9, 9, 10, 32)$$

Lageparameter müssen nicht in der Stichprobe enthalten sein.

Median

im Gegensatz zum arithmetischen Mittel werden beim Median nicht alle Werte einer Stichprobe in dessen Berechnung einbezogen. Insbesondere spielen der kleinste und der größte Datenwert für die Berechnung des Medians i.A. keine Rolle. Diese Eigenschaft macht der Median robust gegenüber positiven und negativen Ausreißern bzw. dem Auftreten von extremen Werten.

Berechnung des Medians

Definition: Berechnung des Medians

Bezeichnet $(x_1, x_2, x_3, ...x_n)$ einen Vektor von geordneten Merkmalswerten eines kardinalen Merkmals, so ist der Median $\bar{x}_Z$ in eindeutiger Weise definiert durch

$$\bar{x}_Z = \begin{cases} x_i & \text{mit } i = (n+1)/2 \text{ für ungerade } n, \\ \dfrac{x_i + x_{i+1}}{2} & \text{wobei } i = n/2 \text{ für gerade } n. \end{cases} \tag{5}$$

der Modus

Bei nominalen Merkmalen kann keiner der bisher festgelegten Mittelwerte berechnet werden. Für diesen Fall existiert das Konzept des Modus (der modalen Klasse).

Definition: Definition des Modus und der modalen Klasse

Der Modus einer Stichprobe ist die am häufigsten auftretende Merkmalsausprägung. Liegen statt Einzelwerten klassierte Daten eines Merkmals vor, wird die Klasse mit der größten Häufigkeit modale Klasse genannt.

geometrisches Mittel

Beispiel:

Ein Sparbrief der Spaßkasse Nirgendwo verspricht bei Anlage einer Summe von $K_0 = 10.000$ im 1. Jahr einen Zins von $q_1 = 6\%$, im 2. Jahr von $q_2 = 7\%$ und im 3. Jahr von $q_3 = 8\%$. Nach drei Jahren erfolgt die Rückzahlung. Das arithmetische Mittel ist
$\bar{x} = 1/3 \cdot (6\% + 7\% + 8\%) = 7\%$.

Der hypothetische Kapitalbetrag nach Ende des ersten Jahres lautet

$$K_1 = K_0 + q_1 \cdot K_0 = (1 + q_1)K_0 = 10.600$$

Der durch das Ausklammern von K_0 entstehende Ausdruck $(1 + q_1)$ wird auch als *Kapitalwachstumsfaktor* bezeichnet.

geometrisches Mittel

Beispiel:

Die geometrische Folge der Kapitalbeträge K_1, K_2 und K_3 errechnet sich wie folgt:

$$\begin{aligned} K_1 &= (1+q_1)K_0 = 10.600 \\ K_2 &= (1+q_2)K_1 = 1,07 \cdot 10.600 = 11.342 \\ K_3 &= (1+q_3)K_2 = (1+q3)K_2 = 1,08 \cdot K_2 \\ &= (1+q_3)(1+q_2)K_1 = 1,08 \cdot 1,07 \cdot K_1 \\ &= (1+q_3)(1+q_2)(1+q_1)K_0 = 1,08 \cdot 1,07 \cdot 1,06 \cdot K_0 \end{aligned}$$

Wie lässt sich hier ein mittlerer Zinssatz ermitteln?

geometrisches Mittel

Der Mittlere Zuwachs $\bar{q}$ ist derjenige, der für den Kapitaleinsatz K_0 nach drei Jahren denselben Endbetrag K_3 ergibt:

$$K_0(1+\bar{q})(1+\bar{q})(1+\bar{q}) = (1+\bar{q})^3 K_0 = K_3$$

Nach Einsetzen von $K_3 = (1+q_3)(1+q_2)(1+q_1)K_0$ und Kürzen von K_0 erhält man

$$(1+\bar{q})^3 = (1+q_3)(1+q_2)(1+q_1)$$

$$\Longleftrightarrow \bar{q} = \sqrt[3]{(1+q_3)(1+q_2)(1+q_1)} - 1 \tag{6}$$

Mit Zahlenwerten ergibt sich für unser Beispiel das geometrische Mittel

$$\bar{q} = \sqrt[3]{1,08 \cdot 1,07 \cdot 1,06} - 1 = 0,06997 = 6,997\%.$$

Definition: geometrisches Mittel

Für n einzelne Werte $(x_1, x_2, ...x_n)$ kann die Formel (**??**) verallgemeinert werden: das geometrische Mittel x_G ist

$$\bar{x}_G := \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}. \tag{7}$$

Treten einzelne Merkmalswerte mehrfach auf und sind nur $m \leq n$ Werte $(x_1, x_2, ...x_m)$ verschieden, die mit den Häufigkeiten $(h_1, h_2, ..., h_m)$ auftreten, so ist das geometrische Mittel

$$\bar{x}_G = \sqrt[n]{(x_1)^{h_1} \cdot (x_2)^{h_2} \cdot ...(x_m)^{h_m}} = (x_1)^{h_1/n} \cdot (x_2)^{h_2/n} \cdot ...(x_m)^{h_m/n} \tag{8}$$

harmonisches Mittel

Beispiel: Berechnung der mittleren Geschwindigkeit.

Herr Andy Theke legt mit seinem PKW eine Strecke s von 300km - die ersten 100km mit einer Durchschnittsgeschwindigkeit von $v_1 = 120\text{km/h}$, auf dem zweiten Teilstück von ebenfall 100km Länge erreicht er noch $v_2 = 100\text{km/h}$ Durchschnittsgeschwindigkeit, die letzten 100km mit einem Schnitt von $v_3 = 80\text{km/h}$ zurück.
Wie hoch war die durchschnittliche Geschwindigkeit auf der gesamten Strecke? Das arithmetische Mittel ist:

$$\bar{v} = \frac{v_1 + v_2 + v_3}{3} = \frac{120 + 100 + 80}{3}\ \text{km/h}\ = 100\ \text{km/h}$$

harmonisches Mittel

Beispiel: Berechnung der mittleren Geschwindigkeit.

Rechnet man jedoch nach der intuitiven Formel

Durchschnittliche Geschwindigkeit = Gesamtweg geteilt durch Gesamtzeit

die gesamte Wegstrecke $s = 300\text{km}$ durch die gesamte Fahrzeit

$$T = \frac{100\text{km}}{120\text{km/h}} + \frac{100\text{km}}{100\text{km/h}} + \frac{100\text{km}}{80\text{km/h}} = 5/6\text{h} + 1\text{h} + 5/4\text{h} = 37/12\text{h}$$

so ergibt sich die korrekte Durchschnittsgeschwindigkeit von

$$\bar{v}_H = \frac{s}{T} = \frac{300\text{km}}{37/12\text{h}} = 97,30\text{km/h}$$

harmonisches Mittel

Mit Hilfe der allgemeinen Bezeichnungen v_1, v_2, v_3 für die Geschwindigkeiten lautet die analoge Formel zur Berechnung der Durchschnittsgeschwindigkeit

$$\bar{v}_H = \frac{1}{\frac{1}{3}\left(\frac{1}{v_1} + \frac{1}{v_2} + \frac{1}{v_3}\right)} \tag{9}$$

Durch Verallgemeinerung der Formel (9) erhält man die Definition des *harmonischen Mittels*.

Definition: Definition des harmonischen Mittels

Für n Einzelwerte $(x_1, x_2, x_3, ...x_n)$ ist das harmonische

$$\bar{x}_H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}\right)} \tag{10}$$

Sind aus den n Einzelwerten lediglich $m < n$ Merkmalswerte $(x_1, x_2, ...x_n)$ mit den Häufigkeiten $(h_1, h_2, ...h_m)$ verschieden, gilt

$$\bar{x}_H = \frac{1}{\frac{1}{n}\left(\frac{h_1}{x_1} + \frac{h_2}{x_2} + ... + \frac{h_m}{x_m}\right)} \tag{11}$$

Zusammenfassung Lageparameter

- der Modus $\bar{X}_M$: Merkmalswert mit der größten Häufigkeit
- arithmetisches Mittel

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

  Vorteil: eingängig
- Median (Zentralwert) $\bar{x_Z}$: teilt die Stichprobe in gleichgroße Teilmengen.
  Vorteil: robust gegen Ausreißer
- harmonisches Mittel $\bar{x_H}$: zur Mittelwertbildung von Quotienten
- geometrisches Mittel $\bar{x_G}$: zur Mittelung bei multiplikativen Zusammenhängen (Wachstumsprozesse)

Zusammenfassung Lageparameter

Oft sind in der Statistik mehrere Lageparameter korrekt - es gibt tatsächlich 'alternative facts' !

Für einfache Stichproben mit additiver Verknüpfung x_i können sich AM und Median deutlich unterscheiden, sie sind aber beide korrekt. Dadurch sind oft auch unterschiedliche Darstellungen eines Sachverhalts möglich

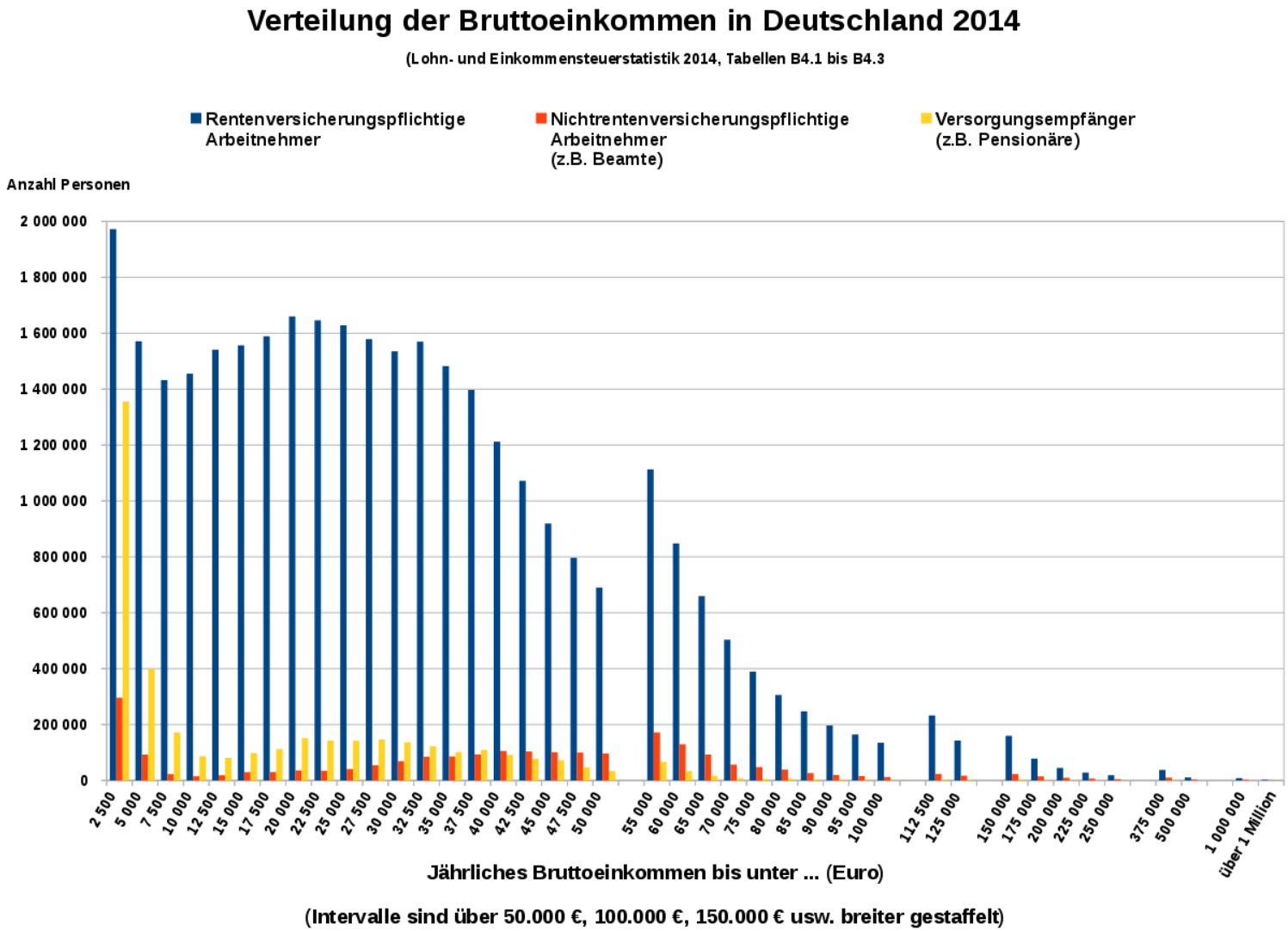
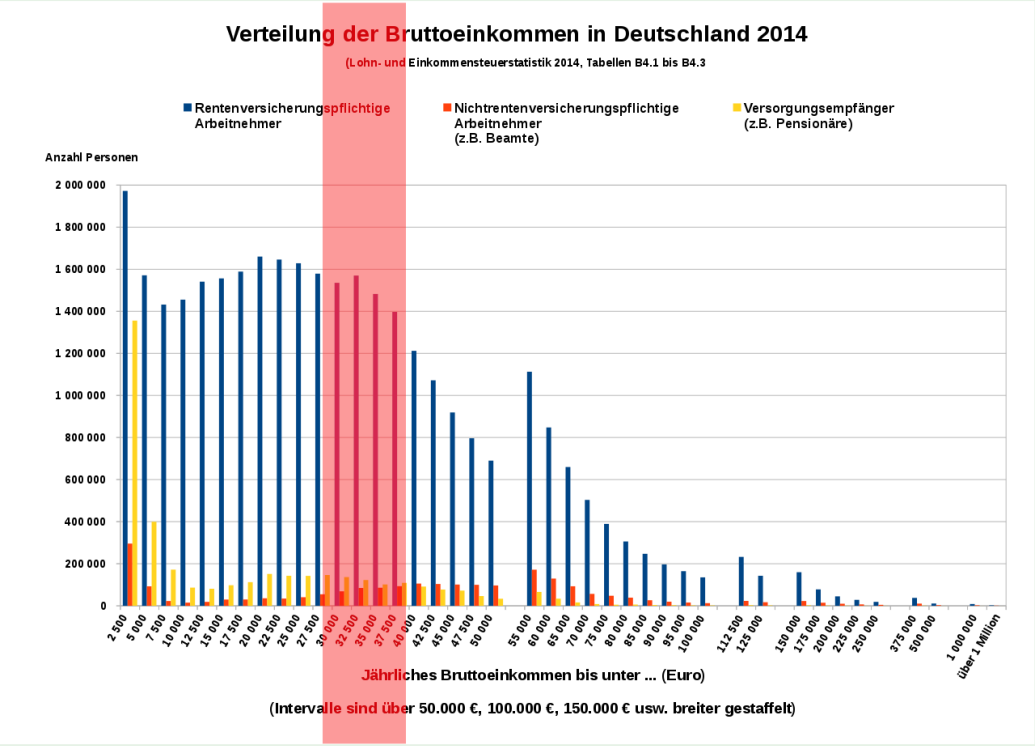


Abbildung: Einkommen deutscher Haushalte

Zusammenfassung Lageparameter

Beispiel: Einkommensverteilung



Für die Einkommensverteilung ist das AM $\bar{x} = 39\,077€$, der Median $\bar{x}_Z = 27\,850€$[1]. Beides sind typische Werte für diese Stichprobe!

[1] Statistisches Bundesamt, Fachserie 14 Reihe 7.1, 2014

klassierte Daten

Für diskrete Merkmale sind erhobene Daten nur an den Stellen $x = x_i$ empirisch gehaltvoll ('Ehepaare haben in D im Mittel 1,72 Kinder').

Liegt aber ein stetiges Merkmal vor, so ist für x auch jeder Wert dazwischen möglich. Oft ist es sinnvoll, schon bei der Erhebung der Daten Beobachtungswerte Intervallen zuzuordnen, den sog. Klassen. Die Zahl und die Größe dieser Klassen wird vom Untersuchungsziel und den Möglichkeiten der Datenerhebung bestimmt.

klassierte Daten

Als Beispiel die Verteilung der Körpergröße eines Statistik-Kurses mit $n = 27$ erfassten Teilnehmern. Die Werte des Merkmals *Größe* werden in $k = 6$ Größenklassen[2] eingeteilt.

Größenklasse	absolute Häufigkeit h_k	relative Häufigkeit f_k	Breite Δx_k	Dichte f_k^*
[1,50 ; 1,65]	4	0.148	0.15	0.988
]1,65 ; 1,70]	7	0.259	0.05	5.185
]1,70 ; 1,75]	3	0.111	0.05	2.222
]1,75 ; 1,80]	6	0.222	0.05	4.444
]1,80 ; 1,85]	6	0.222	0.05	4.444
]1,85 ; 2,05]	1	0.037	0.2	0.185

Tabelle: die Größenverteilung in klassierter Form. Die Dichte f_k^* ist der Quotient f_k/Δ_k.

[2] $]a; b]$: die Klasse erstreckt sich von a bis b, wobei a nicht enthalten ist.

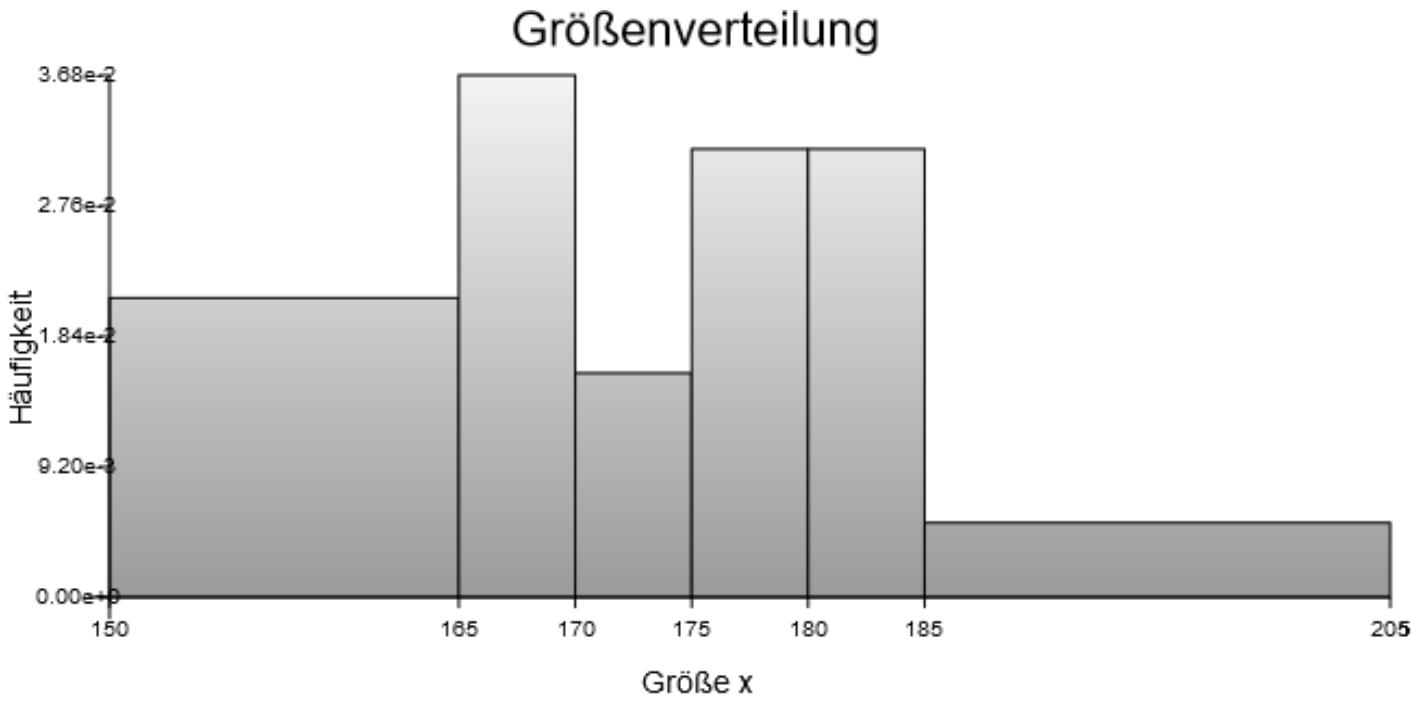
klassierte Daten

Im Gegensatz zur vorherigen Tabelle der Verteilung des Lebensalters ist hier bereits Information in Form einzelner Körpergrößen vernichtet worden.

Während bei einem geringen Stichprobenumfang (hier $n = 27$ Werte) diese Reduktion der Daten nicht nötig gewesen wäre, ist sie bei bei größer angelegten Stichproben unumgänglich: man stelle sich alleine die Verteilung der Einkommen deutscher Haushalte ohne die Reduktion durch klassierte Angaben vor!
Auch klassierte Daten werden oft grafisch dargestellt.

klassierte Daten

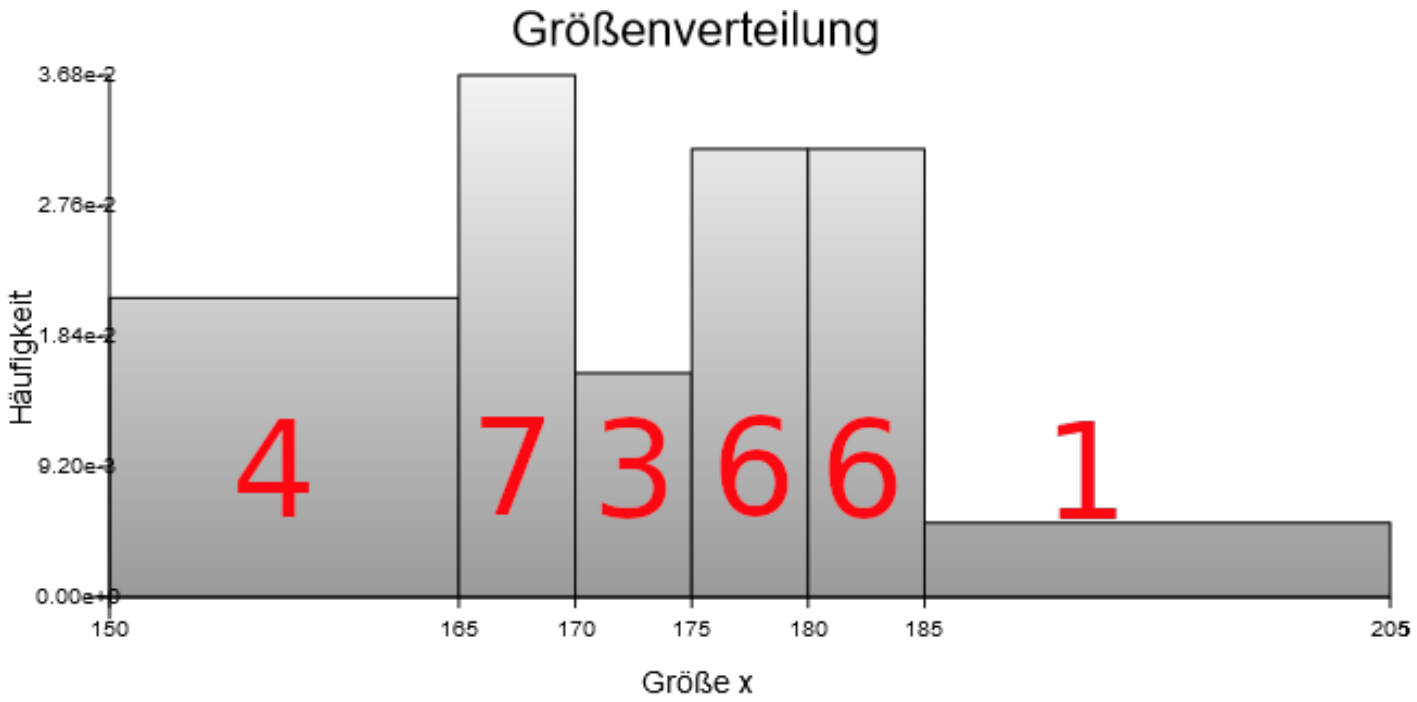
In einer einfachen Balkengrafik sieht die Verteilung der Größe so aus:



Das ist nicht gut!

klassierte Daten

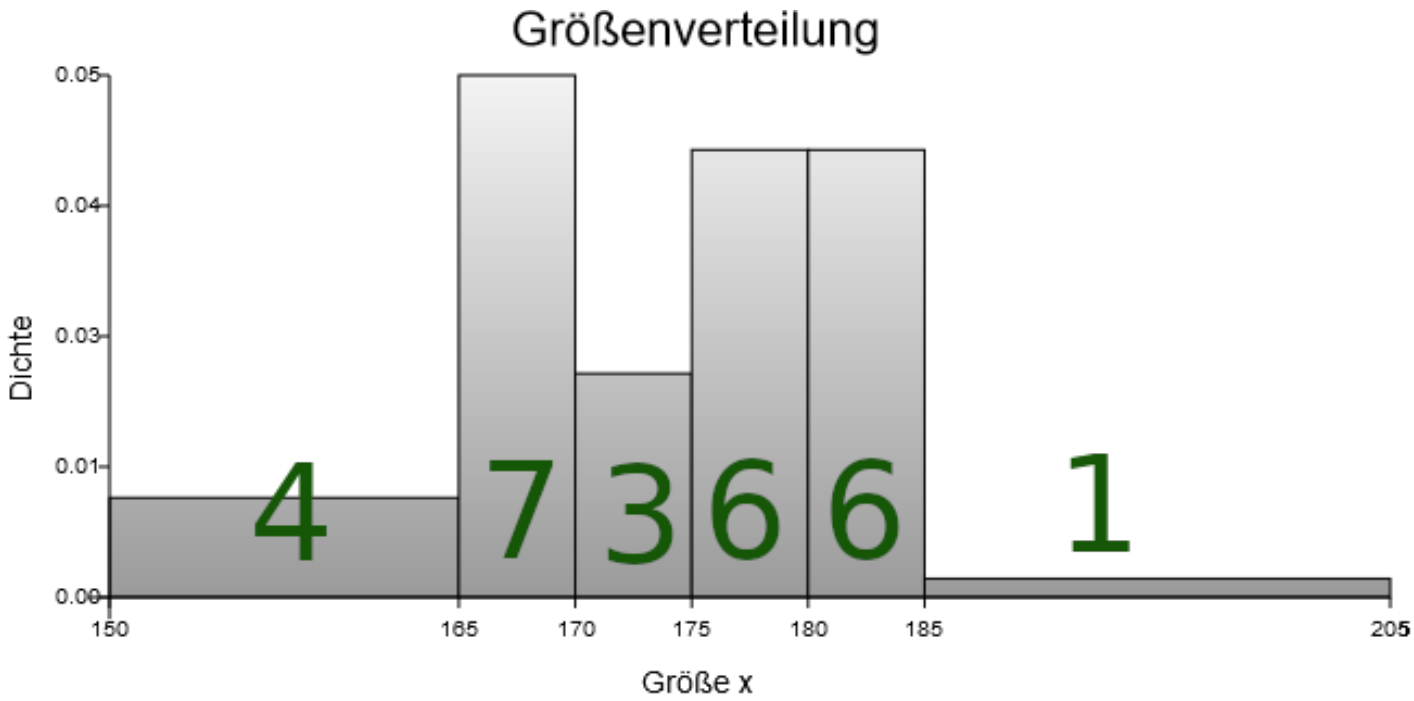
In einer einfachen Balkengrafik sieht die Verteilung der Größe so aus:



Das ist nicht gut!

klassierte Daten

Besser:



In einem Histogramm wird die Zahl in der Fläche der Balken versteckt.
Dazu trägt man die Häufigkeitsdichte über die Klassengrenzen auf.

klassierte Daten

Definition: Häufigkeitsdichte

Die Klassenbreite einer Klasse $[x_i^u; x_i^o]$ ist $\Delta_i = x_i^o - x_i^u$. Damit kann die absolute Häufigkeitsdichte $h_i^* = \frac{h_i}{\Delta_i}$ bzw. relative Häufigkeitsdichte $f_i^* = \frac{f_i}{\Delta x_i}$ bestimmt werden.

Der Mensch interpretiert die Fläche der Balken - das Histogramm bildet die klassierte Häufigkeitsverteilung deutlich besser ab.