Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik	**Grundbegriffe der Statistik**	

In der Statistik haben wir es mit Stichproben zu tun, die aus einer Grundgesamtheit (alle Einwohner eines Landes, alle Äpfel aus einer Lieferung ...) entnommen werden. Die Elemente der Stichprobe werden auf ein bestimmtes Merkmal untersucht, das in verschiedenen Ausprägungen auftreten kann.

n: Umfang der Stichprobe
$x_1, x_2, \ldots, x_n$: gemessene Werte (Ausprägungen des untersuchten Merkmals)
$H_1, H_2, \ldots$: absolute Häufigkeit
$h_1, h_2, \ldots$:relative Häufigkeit ($h_i = H_i/n$)

Je nach Art eines Merkmals unterscheidet man verschiedene Skalenniveaus:

Nominalskala: verschiedene Eigenschaften, keine vorgegebene Reihenfolge (z.B. Geschlecht, Wohnort)
Ordinalskala: die Werte können geordnet werden, man kann aber keine Abstände zwischen ihnen angeben (z.B. Rangplätze, Schulnoten)
Intervallskala: der Abstand zwischen zwei Werten lässt sich messen, der Nullpunkt ist willkürlich festgelegt (z.B. Jahreszahlen, Temperatur in °C)
Verhältnisskala: es gibt einen natürlichen Nullpunkt, man kann also sowohl die Differenz als auch das Verhältnis zweier Werte angeben (z.B. Alter, Einkommen). Solche Daten liefern die meiste Information.

Die Häufigkeiten stellt man gern in einem **Histogramm** dar.
Bei großen Datenmengen teilt man die Werte in **Klassen** ein (z.B. Größe 150 - 160 cm, 160 - 170 cm ...)

Mathematik – Intensivkurs: Statistik		
Thema	**Gesetze und Regeln**	**Musterbeispiele**
Darstellungen	**Strichliste** Die Strichliste ist die Urform vieler Graphiken, wird aber praktisch nur noch bei manueller Daten-Erfassung (Inventur, Bier-Deckel usw.) eingesetzt. Dabei wird für jede Merkmalsausprägung die Häufigkeit ihres Vorkommens durch eine entsprechende Anzahl von Strichen symbolisiert. Bisher wurde jeder fünfte Strich quer durch die vier vorhergehenden gemacht, also 5-er-Päckchen gebildet, moderner sind 10-er-Päckchen, bei denen zuerst die vier Seiten eines Quadrates, dann die beiden Diagonalen und schließlich Punkte auf den vier Ecken gezeichnet werden; es sollen so weniger Fehler beim Auszählen passieren; wer's glaubt... **Säulendiagramm – Histogramm** Ein Säulen-(Balken)Diagramm erlaubt den direkten visuellen Größenvergleich der bei den Merkmalsausprägungen dargestellten Werte. Dabei werden auf einer Achse die Merkmalsausprägungen und auf der anderen die darzustellenden Variablen-Werte aufgetragen. Ist die Merkmalsachse horizontal, spricht man von einem Säulen-Diagramm, ist sie vertikal, von einem Balken-Diagramm. Letzteres ist besonders dann vorteilhaft, wenn die Beschriftungen der Merkmalsausprägungen lange Worte sind, die so gut lesbar bleiben. Prinzipiell wird bei einem Säulen-(Balken-)Diagramm jeder Variablenwert durch ein Rechteck dargestellt. In einem Säulen-Diagramm ist die Höhe der Rechtecke zu den Variablenwerten proportional und die Breite immer gleich, bei einem Balken-Diagramm ist es genau umgekehrt. Bei sehr vielen Merkmalsausprägungen (wie immer bei stetigen Variablen) bildet man besser sog. Klassen, die eine bestimmte Anzahl von Merkmalsausprägungen zusammenfassen. Dabei ist auf eine eindeutige Zuordnung von Merkmalsausprägungen zu Klassen unbedingt zu achten. Liegen diskrete (einzelne) Merkmalsausprägungen vor oder wurden gleichbreite Klassen gebildet, dann sind sowohl die Höhe(Breite) als auch die Fläche der Säulen(Balken) zu den Variablen-Werten proportional. So kann die zu kommunizierende Information vom menschlichen visuellen System sehr einfach erfaßt werden. Ein Säulen(Balken)-Diagramm betont grundsätzlich die Unterschiede zwischen den Variablen-Werten verschiedener Merkmalsausprägungen. Bei Daten mit zumindest Ordinal-Niveau liegt die Reihenfolge der Merkmalsausprägungen meist fest, ansonsten kann man eine beliebige wählen, um bestimmte Aspekte durch die Art der Darstellung zu betonen. Sind die dargestellten Werte die Häufigkeiten, mit denen die Merkmalsausprägungen (Klassen) in den Daten vorkommen, so nennt man dieses spezielle Diagramm ein Histogramm.	

Mathematik – Intensivkurs: Statistik		
Thema	**Gesetze und Regeln**	**Musterbeispiele**
Darstellungen	**Kreissektordiagramm** Bei einem Kreis-Sektoren-Diagramm wird jede Merkmalsausprägung durch ein Kreis-Segment (Torten-Stück) abgebildet, dessen Größe (Winkel) sich nach der relativen Häufigkeit der jeweiligen Merkmalsausprägung richtet. Wahrnehmungsforscher glauben herausgefunden zu haben, daß viele Menschen solche Diagramme wie eine Uhr lesen, also oben anfangen und im Uhrzeigersinn fortfahren. Bei Daten auf Ordinal-Niveau ist die Reihenfolge der Sektoren meist festgelegt; haben die Daten aber lediglich Nominal-Niveau, so kann mit der dann frei wählbaren Anordnung der Sektoren eine für den Zweck der Darstellung optimale Reihenfolge wählen! **Liniendiagramm** Bei einer Linien-Graphik werden die Werte der dargestellten Größe durch eine Linie verbunden, daher der Name. Die Linie betont die Kontinuität in einem Trend; eine Linien-Graphik sollte daher auch nur entsprechend eingesetzt werden, nämlich vorzugsweise zur Darstellung desselben Merkmals bspw. zu verschiedenen Zeiten. Auch jede andere stetige Variable kann zur Unterscheidung dienen, also als sog. unabhängige Variable eingesetzt werden. Wenn die unabhängige Variable schon nicht stetig ist, dann sollte sie wenigstens im dargestellten Ausschnitt sehr viele kleine Schritte umfassen, also quasi-stetig sein, ansonsten wird ein falscher Eindruck erweckt und ein Säulen(Balken)-Diagramm ist ggf. vorzuziehen. Das in einer Linien-Graphik dargestellte Merkmal, die sog. abhängige Variable, sollte eine quantitative Skala aufweisen und ebenfalls zumindest quasi-stetig sein; ansonsten ist wiederum ein Säulen(Balken)-Diagramm vorzuziehen.	
 Treff Lernen CHRISTINA STRAUCH		

Thema	Gesetze und Regeln	Musterbeispiele

Beschreibende Statistik

Lagemaße

Modalwert (Modus) M

Der **Modalwert** (auch **Modus**) **M** ist die häufigste Merkmalsausprägung. Es ist also derjenige Wert, welcher die größte Häufigkeit aufweist. Der Modus (Modalwert oder auch Dichtemittel) einer Verteilung ist jener Wert, der in einer Datenmenge am häufigsten vorkommt. Der Modalwert ist nur für große Datensätze von Interesse, da er für kleine Stichproben stark von Zufälligkeiten in den Daten abhängt.

- Der Modus ist für beliebig skalierte Daten definiert. Für nominalskalierte Daten ist er der einzige Lageparameter.
- Im Gegensatz zu den anderen Lageparametern kann der Modus mehrdeutig sein bzw. es kann mehrere „Modi" geben. Gibt es z.B. zwei Modi (doppelgipflige Verteilung), spricht man von
- einer **bimodalen** Verteilung (warum ist z.B. die Geschwindigkeitsverteilung auf Autobahnen oft nahezu bimodal?)
- Bei klassierten kardinalskalierten Daten liegt der Modus, dann auch **Dichtemittel** genannt, innerhalb der Klasse i mit der höchsten Häufigkeitsdichte f^*_i .

Beispiel
7,12,8,9,2,5,9
M = 9

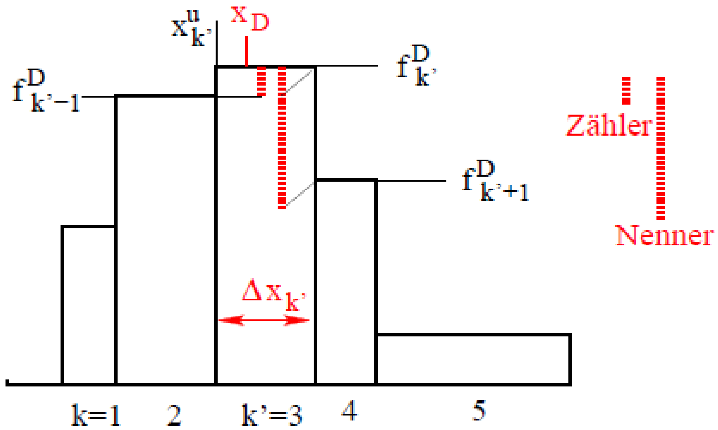
Modalwert bei klassierten Daten

Wie bei den Quantilen gibt es für klassierte Daten eine „Feinberechnung", bei der Parabeln durch jeweils drei Punkte der Punktmenge $\{ (x_k; f^D_k) \}$ gelegt werden:

$$x_D = x^u_{k'} + \frac{f^D_{k'} - f^D_{k'-1}}{(f^D_{k'} - f^D_{k'-1})(f^D_{k'} - f^D_{k'+1})} \Delta x_{k'}$$

mit k' der Klasse mit der höchsten Häufigkeitsdichte f*.

k'	die am häufigsten besetzte Klasse
$x^u_{k'}$	untere Klassengrenze der am häufigsten besetzten Klasse
$\Delta x_{k'}$	Klassenbreite
$f^D_{k'}$	Häufigkeit der k-ten Klasse
f^D_{k-1}, f^D_{k+1}	Häufigkeit der benachbarten Klassen



Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele

Beschreibende Statistik

Median $x_{0,5}$

Wenn wir die Verteilungskurve (oder das Histogramm) einer Stichprobe betrachten, so liegt der Median an der Stelle auf der x-Achse, an der die Fläche unter der Kurve (oder die Fläche des Histogramms) exakt in zwei Teile geteilt wird. Die relative Position des Modus, des Medians und des Mittelwerts liefert einen Hinweis auf die Schiefe einer Verteilung (siehe rechts).

Der Median wird wie folgt berechnet:

- Alle Werte werden in aufsteigender Reihenfolge geordnet.
- Wenn die Anzahl der Werte ungerade ist, wird die mittlere Zahl verwendet.
- Wenn die Anzahl der Werte gerade ist, wird der Durchschnitt der zwei mittleren Zahlen ausgewählt.

Die Summe der absoluten Abweichungen eines Stichprobenwerts von ihrem Median sind geringer als die absoluten Abweichungen eines beliebigen anderen Werts. Unter bestimmten Umständen kann der Median ein stabileres Lagemaß als der Mittelwert sein. Der Median ist weniger anfällig für Ausreisser (Extremwerte) als der Mittelwert. Die Medianstatistik wird daher oft in Zusammenhang mit der robusten Statistik genannt.

> Der Median teilt die Stichprobe so, dass **mindestens** 50% der Daten kleiner oder gleich diesem Wert und **mindestens** 50% der Daten größer oder gleich diesem Wert sind.

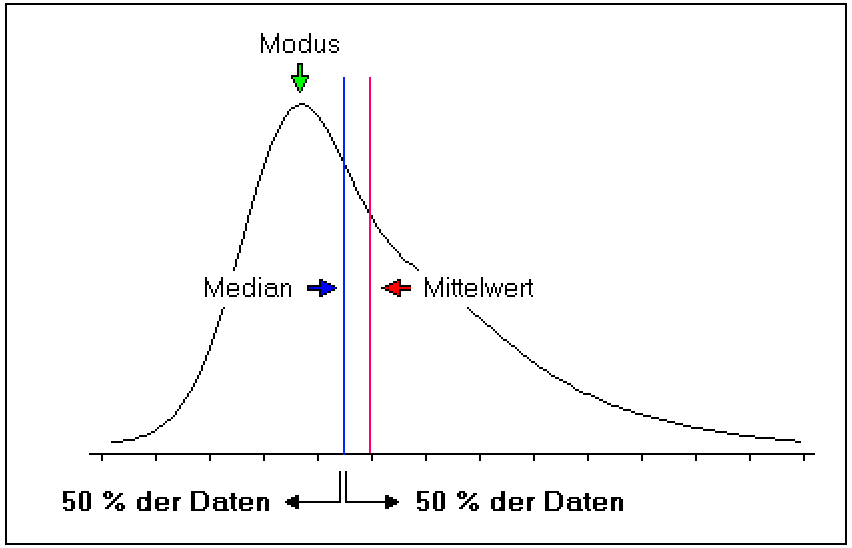
Der Median kennzeichnet den Wert einer Verteilung, bis zu dem 50% aller Daten vorliegen. Der Median ist eng verbunden mit den später beschriebenen Quantilen. Für diese Quantile gilt ein Prozentwert für den Anteil der betrachteten Daten. Diesen Anteil kann man beim Median mit p = 0,5 ansetzen, da die Hälfte der Daten erfaßt sein soll.

In Anlehnung an Quantile wird der Median auch als $x_{0,5}$ bezeichnet, als Quantil mit 50%

Eigenschaften des Medians:

- Er ist im Gegensatz zum arithmetischen Mittel unempfindlich gegenüber Fehlern
- Er nutzt im Gegensatz zum arithmetischen Mittel nicht die ganze Information der Häufigkeitsverteilung aus
- Der Median minimiert die Betragssumme der Abweichungen:

$$\sum_{i=1}^{n} |x_i - x_{0,5}|$$



Aus der Definition links folgt:
Wenn die Anzahl der Daten ungerade ist, gibt es genau ein mittleres Element. Bei einer geraden Anzahl ist die Mitte der beiden mittleren Werte zu bestimmen

$$x_{0,5} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{(n ungerade)} \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right) & \text{(n gerade)} \end{cases}$$

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik	**Median einer Urliste**	

Gegeben sei eine univariate statistische Erhebung mit

- einer Grundgesamtheit mit dem Erhebungsumfang n,
- einem quantitativen Merkmal X und
- der durch die Erhebung gewonnen Urliste mit den Messwerten x1,..., xn , die sich in einer eindeutigen Reihenfolge, z.B. $x_1 \leq . . .. \leq x_n$ nach steigender Größe anordnen lassen

Dann bestimmt sich den Median (oder Zentralwert) $\tilde{x}$ der gewonnenen Daten als diejenige Zahl, für die die eine Hälfte der Messwerte kleiner oder gleich und die andere Hälfte größer oder gleich ist, d.h. durch die Bedingung

$$\underbrace{x_1 \leq . . . . \leq}_{\text{die Hälfte der ´kleineren´ Messwerte}} \ \tilde{x} \ \underbrace{\leq . . . . \leq x_n}_{\text{die Hälfte der ´größeren´ Messwerte}}$$

Wenn n • p eine ganze Zahl ergibt (gerade Anzahl von Daten):

Daten: 2,1; 3,4; 7,3; 8,9; 9,3; 10,1; 11,2; 11,9 → n = 8

1. Bestimmung des unteren und oberen Datenpunktes:

   oberer Datenpunkt:
   n • 0,5 = 8 • 0,5 =4 → Stelle 4 → Wert der Stelle 4: 8,9

   unterer Datenpunkt:
   n • 0,5+1 = 8 • 0,5 +1= 5 → Stelle 5 → Wert der Stelle 5: 9,3

2. Berechnung des Medians:
   (Wert des unteren Datenpunktes + Wert des oberen Datenpunktes)/2
   (8,9 + 9,3)/2 = 9,1 → Median: 9,1

Gerade Zahl von Werten
7,12,8,10,2,5,9,12
Sortiert nach Größe: 2, 5, 7, 8, 9, 10, 12, 12
$\tilde{x}$ = 8, 5

Erklärung: Beim Median muss die Hälfte der Werte links vom Median liegen und die Hälfte der Wert rechts. Bei einer geraden Anzahl von Werten in einer Liste kann es einen solchen Punkt, der auch noch zur Liste gehört nicht geben. Deshalb muss man den Mittelwert der beiden mittleren Daten bilden

Wenn n • p keine ganze Zahl ergibt (ungerade Anzahl von Daten):

Daten: 2,1; 3,4; 7,3; 8,9; 9,3; 10,1; 11,2 → n = 7

n • 0,5 = 7 • 0,5 = 3,5 → aufrunden!! → Stelle 4 → Wert der Stelle 4: 8,9
→ Median: 8,9

Es **muß** aufgerundet werden, da die Definition verlangt: .. es muß 50% der Daten vorliegen. Beim abgerundeten Wert wären es noch keine 50% der Daten.

Ungerade Zahl von Werten
7,12,8,10,2,5,9
Sortiert nach Größe: 2, 5, 7, 8, 9, 10, 12
$\tilde{x}$ = 8

Bei einer ungeraden Anzahl gibt es genau einen Wert, für den die Hälfte der Daten kleiner oder gleich ist und die Hälfte der Daten größer oder gleich.

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele

Beschreibende Statistik

Median aus absoluten Häufigkeiten

Gegeben sei eine univariate statistische Erhebung mit
- einer Grundgesamtheit mit dem Erhebungsumfang n,
- einem quantitativen Merkmal X und
- den Absoluten Häufigkeiten $H(a_1),\dots, H(a_m)$ der einzelnen Merkmalsausprägungen.

Dann bestimmt sich den Median (oder Zentralwert) $\tilde{x}$ der gewonnenen Daten als diejenige Zahl, für die die eine Hälfte der Messwerte kleiner oder gleich und die andere Hälfte größer oder gleich ist, d.h. durch die Bedingung

$$\underbrace{x_1 \leq \dots\dots\dots \leq}_{\text{die Hälfte der ´kleineren´ Messwerte}} \tilde{x} \underbrace{\leq \dots\dots\dots \leq x_n}_{\text{die Hälfte der ´größeren´ Messwerte}}$$

Beispiel: Gegeben sind die Absoluten Häufigkeiten

a_i	1,3	1,4	1,5	1,6	1,7	1,8	1,9
$H(a_i)$	1	4	5	3	4	2	1

Berechne den Median $\tilde{x}$ der gewonnenen Daten.

Da 10 Messwerte kleiner oder gleich 1,5 und 10 Messwerte größer oder gleich 1,6 sind, ergibt sich $\tilde{x} = \frac{1{,}5+1{,}6}{2} = \frac{3{,}1}{2} = 1{,}55$.

Median aus relativen Häufigkeiten

Gegeben sei eine univariate statistische Erhebung mit
- einer Grundgesamtheit mit dem Erhebungsumfang n,
- einem quantitativen Merkmal X und
- den Relativen Häufigkeiten $h(a_1),\dots, h(a_m)$ der einzelnen Merkmalsausprägungen.

Dann bestimmt sich den Median (oder Zentralwert) $\tilde{x}$ der gewonnenen Daten als diejenige Zahl, für die die eine Hälfte der Messwerte kleiner oder gleich und die andere Hälfte größer oder gleich ist, d.h. durch die Bedingung

$$\underbrace{x_1 \leq \dots\dots\dots \leq}_{\text{die Hälfte der ´kleineren´ Messwerte}} \tilde{x} \underbrace{\leq \dots\dots\dots \leq x_n}_{\text{die Hälfte der ´größeren´ Messwerte}}$$

Beispiel: Gegeben sind die Relativen Häufigkeiten

a_i	1,3	1,4	1,5	1,6	1,7	1,8	1,9
$h(a_i)$	5%	20%	25%	15%	20%	10%	5%

Berechne den Median $\tilde{x}$ der gewonnenen Daten.

Da 50% der Messwerte kleiner oder gleich 1,5 und 50% der Messwerte größer oder gleich 1,6 sind, ergibt sich $\tilde{x} = \frac{1{,}5+1{,}6}{2} = \frac{3{,}1}{2} = 1{,}55$.

Mathematik – Intensivkurs: Statistik

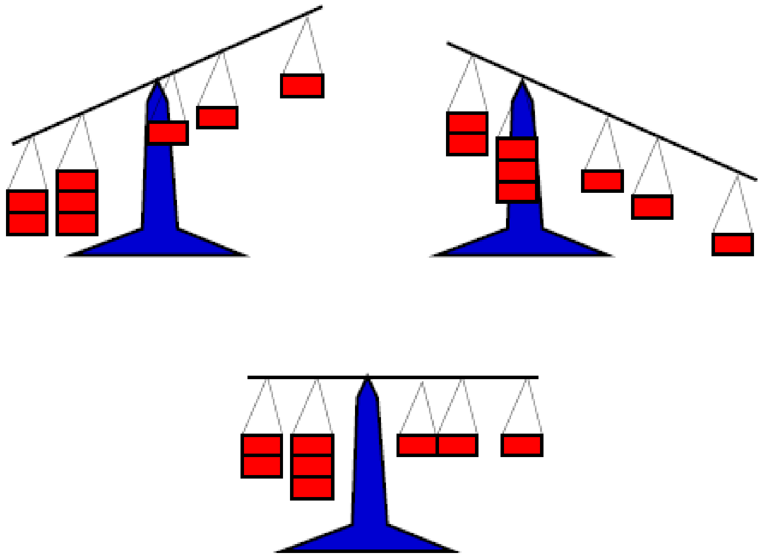
Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik	**Arithmetisches Mittel $\bar{x}$** Das arithmetische Mittel (Mittelwert) liegt in der Mitte aller vorhandenen Werte. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *(Der Bruch vor dem Summenausdruck hat im Nenner die Anzahl der Werte stehen)* Der Mittelwert ist geeignet Verteilungen mit einer Spitze und möglichst symmetrischen Verhaltens vor und nach der Spitze zu beschreiben. Der Mitelwert ist empfindlich gegenüber Ausreißern. Einzelne abnorme Wert beeinflussen den Mittelwert sehr stark.	Mittelwert aus einfacher Liste 7.5; 7.7; 8.1; 8.8; 8.9; 9.3; 9.5 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7} \cdot 59{,}8 = 8{,}543$
	Arithmetisches Mittel für gruppierte Daten $\bar{x}$ $\bar{x} = \frac{\sum_{i=1}^{m} x_i \cdot h_i}{\sum_{i=1}^{m} h_i} = \sum_{i=1}^{n} x_i \cdot \pi_i$ x_i Note h_i Häufigkeit π_i Relative Häufigkeit	Berechnung eines Notendurchschnitts für eine Gruppe von Schülern: Note (x_i): 1 2 3 4 5 Häufigkeit (h_i) 2 4 8 2 1 $\Sigma = 17$ Rel. Häufigk. (π_i) 2/17 4/17 8/17 2/17 1/17 $\Sigma = 1$ $x_i \times h_i$ 2 8 24 8 5 $\Sigma = 47$ $\bar{x} = 47/17 = 2,8$
	Arithmetisches Mittel für klassifizierte Daten $\bar{x}$ Sind die Daten in K Klassen mit den Klassenmitten m_k und den Häufigkeit h_k klassifiziert, so gilt: $\bar{x} = \frac{\sum_{k=1}^{K} m_k \cdot h_k}{\sum_{k=1}^{K} h_k} = \sum_{k=1}^{K} m_k \cdot \pi_k$ Das berechnete arithmetische Mittel ist nur ein Näherungswert, da durch die Klassifizierung Information verloren geht.	

Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik	**Gewichtetes arithmetisches Mittel $\bar{x}_g$**	

Betrachtet man die einzelnen zu mittelnden Größen als „Gewichte" vom Gewicht gi, die an die Stellen xi einer Waage mit verschiebbaren Lastarm gehängt werden, so git das arithmetische Mittel den Punkt an, bei dem die Waage im Gleichgewicht ist!

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

(Der Bruch vor dem Summenausdruck hat im Nenner die Summe aller Gewichte stehen.)



Das Bild entspricht etwa folgender Häufigkeitstabelle:

x_i	0	1	3	4	6
h_i	2	3	1	1	1

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele

Beschreibende Statistik

Geometrisches Mittel $\overline{w}^g$

Wenn sich innerhalb eines bestimmten Zeitintervalls ein Wachstumsprozess mit variierendem Wachstum abspielt, dann ist das geometrische Mittel die zentrale Tendenz für den Wachstumsfaktor:

$$\overline{w}^g = \sqrt[t]{\prod_{i=1}^{n} w_i}$$

$\overline{w}_i$ ist dabei der Quotient zwischen aufeinanderfolgenden Werten (Wachstums-faktor) der wachsenden Größe und t ist das Zeitintervall (z.B. in sec, h, a), über welches das Wachstum verfolgt wird. $\overline{w}^g$ ist die mittlere Wachstumsrate innerhalb des Zeitintervalls t.

Das geometrische Mittel aus einer Häufigkeitstabelle mit den Häufigkeiten h_i wird bestimmt aus:

$$\overline{w}^g = \sqrt[t]{\prod_{i=1}^{n} w_i^{h_i}}$$

Musterbeispiel: Durchschnittliches Bevölkerungswachstum

Es sei bekannt, dass eine Gemeinde zum Zeitpunkt t 100.000 Einwohner hat und die Einwohnerzahl f`unf Jahre später auf 124.000 Einwohner gewachsen ist (w_1 = 1, 24). Weitere 5 Jahre später ist die Bevölkerung auf 136.000 Einwohner gewachsen (w_2 = 1, 0968). Daraus ergibt sich für das durchschnittliche Wachstum in den vergangenen 10 Jahren

$\overline{w}^g = \sqrt[10]{1{,}24 \times 1{,}0968} = 1{,}031226$.

Die mittlere Wachstumsrate beträgt damit 0,031226. D.h. die Bevöl-kerung ist in den letzten 10 Jahren pro Jahr um 3,12 % gewachsen.

Harmonisches Mittel $\overline{v}^h$

- Das harmonische Mittel entspricht dem arithmetischen Mittel der Kehrwerte.
- Man kann zeigen: $\overline{x}_H$ ist immer kleiner oder gleich $\overline{x}$.
- Die Entscheidung, wann das harmonische Mittel sinnvoll ist, setzt Kenntnis des Sachverhalts voraus!
- Kandidaten sind wegen der obigen Beziehung v.a. Verhältniszahlen, bei denen der Zähler variiert, z.B.
  - Ermittlung von Durchschnittsgeschwindigkeiten
  - Mehrfacher Einkauf ein- und desselben Gutes mit schwankendem Stückpreis X(t) für jeweils einen festen Geldbetrag

Bei Größen, die Verhältnisse darstellen, wie z.B. Geschwindigkeiten, berechnet sich der Mittelwert wie folgt als harmonisches Mittel:

$$\overline{v}^h = \frac{n}{\sum_{i=1}^{n} \frac{1}{v_i}}$$

Musterbeispiel: Mittelwert von Geschwindigkeiten

Ein Skilift bringt Skifahrer mit einer Geschwindigkeit von 5 km/h zum Berggipfel. Bei der Abfahrt beträgt die Geschwindigkeit 15 km/h. Wie groß ist die Durchschnittsgeschwindigkeit bei Auf- und Abfahrt zusammen?

$$\overline{v}^h = \frac{2}{\frac{1}{5} + \frac{1}{15}} = \frac{30}{4} = 7{,}25$$

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele

Beschreibende Statistik
 	
	Betrachtung der verschiedenen Mittelwerte
	• Das arithmetische Mittel $\bar{x}$ ist der Standard-Mittelwert"
	• Das harmonische Mittel x_H wird bei Mittelwerten von Verhältniszahlen verwendet, wenn die im Zähler stehende Größe zur Gewichtung herangezogen werden soll:
	z.B. Geschwindigkeit=Strecke/Zeit, Leistung=Arbeit/Zeit, Kfz-Dichte=Zahl der Kfz/Zahl der Einwohner etc
	• Das geometrische Mittel x_G wird zur Mittelung von Wachstumsprozessen verwendet, bei der es auf die prozentuale Änderung ankommt.
	• Der Median ist das „Mittel“ der Wahl, wenn Extremwerte und/oder offene Randklassen keine Rolle spielen sollen, sowie bei ordinalskalierten Daten
	• Der Modus ist sinnvoll bei mehrgipfligen (multimodalen) Verteilungen, bei stark asymmetrischen Verteilungen sowie bei nominalskalierten Daten
	• Grundsätzlich gilt:
	  ◆ Immer: $x_H \le x_G \le \bar{x}$
	  ◆ Bei symmetrischen unimodalen Verteilungen: $\bar{x} = x_{0,5} = x_D$
	  ◆ Bei rechtsschiefen (= linkssteilen) Verteilungen: $x_D < x_{0,5} < \bar{x}$, bei rechtssteilen Verteilungen umgekehrt.
		Für nominal skalierte Daten ist das einzige Lage-Maß der Modalwert. Bei ordinal skalierten Daten kommt, als zweites Quartil, der Median hinzu. Bei Daten auf zumindest Intervall-Niveau kann ein Vergleich des ersten (bzw. nullten) und dritten (bzw. vierten) Quartils mit dem Median aufdecken, ob die Daten deutlich unsymmetrisch um das Lage-Maß herum verteilt sind oder nicht.
		Haben die Daten zumindest Intervall-Niveau, so lassen sich drei verschiedene Lagemaße ermitteln: der Modalwert, der Median und der (arithm.) Mittelwert. In der Praxis weichen diese drei Werte für ein und dieselbe Verteilung mehr oder weniger voneinander ab. Es stellt sich dann die Frage, welcher der drei Werte ist in einem solchen Fall als Maß für die zentrale Tendenz am verläßlichsten? Allgemein kann man dies nicht beantworten!
		
		Besonders "schlimm" zu handhaben sind deutlich mehrgipflige Verteilungen; bei diesen kann es passieren, daß der Mittelwert in einer Region liegt, die (fast) keinerlei Daten enthält! Bei diesen sog. multi-modalen Verteilungen sollte man generell kein Lagemaß angeben! Manchmal helfen die verschiedenen Arten der Mittelung von arithmetischem Mittelwert, Median und Modalwert vielleicht, eine Entscheidung zu treffen: der arithmetische Mittelwert minimiert die Abweichungen auf einer quadratischen Skala, der Median auf einer linearen, der Modalwert schließlich nur deren Anzahl. Soll also ein Lage-Maß angegeben werden, das allzu extreme Abweichungen verhindert, so ist der arithmetische Mittelwert sinnvoll, soll die Summe aller Abweichungen möglichst klein bleiben, dann ist der Median das Lage-Maß der Wahl, soll einfach die Anzahl der Werte am kleinsten sein, die sich vom Lage-Maß unterscheiden, ist der Modalwert richtig. Weil beim arithmetischen Mittelwert die Abweichungen vor der Mittelung quadriert werden, ist die maximale Abweichung bei ihm kleiner als beim Median; dies heißt aber auch, daß ein Ausreißer, also ein weit von allen anderen entfernt liegender Wert, den arithmetischen Mittelwert viel mehr aus seine Seite "zieht" als den Median.
		Im Alltag hat der arithmetische Mittelwert eine dominante Rolle erreicht. Dies liegt wohl daran, daß man ihn schon dann berechnen kann, wenn man nur die Summe der Variablenwerte und ihre Anzahl kennt, aber kein einziges Datum aus der Verteilung! Das sollte man dann aber auch bei der Einstufung der Verläßlichkeit solcher Zahlen berücksichtigen...

Mathematik – Intensivkurs: Statistik		
Thema	**Gesetze und Regeln**	**Musterbeispiele**
Beschreibende Statistik	**Streuungsmaße** Streuungsmaße drücken aus, wie stark die einzelnen Daten um das Zentrum herum verstreut liegen. Sind fast alle Daten eng um das Zentrum konzentriert, ist die Streuung (Variabilität, Variation) klein. Liegen dagegen einige Daten weit vom Zentrum weg ist, die Streuung groß. **Minimum, Maximum, Spannweite** Am einfachsten läßt sich die Streuung einer Verteilung durch den Wertebereich (Minimum bis Maximum, Bereich (range)) charakterisieren. Die Differenz zwischen Maximum und Minimum ist die Spannweite. **Spannweite aus Urlisten** Gegeben sei eine univariate statistische Erhebung mit • einer Grundgesamtheit mit dem Erhebungsumfang n, • einem quantitativen Merkmal X und • der durch die Erhebung gewonnen Urliste mit den Messwerten x1,..., xn , die sich in einer eindeutigen Reihenfolge, z.B. $x_1 \leq . . .. \leq x_n$ nach steigender Größe anordnen lassen Dann berechnet sich die Spannweitedurch die Differenz des größten vom kleinsten Meßwert: $\mathbf{r = x_n - x_1}$ Für die Spannweite spielt nur die geordnete Urliste eine Rolle. Die Häufigkeit des Auftretens einzelner Werte hat keine Bedeutung.	Beispiel: Gegeben ist die Urliste 1,3|1,8|1,6|1,7|1,7|1,6|1,7|1,5|1,5|1,5|1,4|1,4|1,5|1,5|1,6|1,8|1,4|1,9|1,7|1,4 Daraus ergibt sich eine geordnete Urliste 1,3|1,4|1,4|1,4|1,4|1,5|1,5|1,5|1,5|1,5|1,6|1,6|1,6|1,7|1,7|1,7|1,7|1,8|1,8|1,9 und daraus die Spannweite r = 1,9 – 1,3 = 0,6

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele

Thema: **Beschreibende Statistik**

Quantile

Quantile (Fraktile) sind bestimmte Datenwerte, die eine Verteilung in Bereiche mit einer gleichen Zahl von Werten aufteilen.
Das q-Quantil x_q gibt den Wert der Merkmalsausprägung X an, bei dem ein Anteil q der Merkmalsausprägungen unterhalb von x liegen.
Wenn f der Bruch ist, der dem Quantil (Fraktil) zugeordnete ist, so gilt für das entsprechende Quantil $Q_f = X[i\,]$, wobei gilt:

$$f = \frac{i-1}{n-1} \quad \text{oder} \quad i = f \cdot (n-1) + 1$$

Resultiert für i eine nichtganze Zahl, so muss für die Berechnung des entsprechenden Quantils interpoliert werden:

$$Q_f = X\,[\text{floor}(i)] + (X\,[\text{ceiling}(i)] - X\,[\text{floor}\,(i)]\,)\,(i - \text{floor}\,(i\,))$$

floor ist die größte ganze Zahl, die nicht größer als i ist (=abgerundet)
ceiling ist die kleinste ganze Zahl, die nicht kleiner als i ist (=aufgerundet)

Eines der am Häufigsten benutzten Quantile ist das Quartil, bei dem die geordneten Daten in vier gleiche Bereiche eingeteilt werden.

Quartile: Einteilung in 4 gleich große Bereiche

$$\underbrace{1, 3, 4, 4, 9, 10}_{\text{1.Quartil}},\ \underbrace{17, 20, 23, 24, 24}_{\text{2.Quartil}},\ \underbrace{30, 33, 34, 35, 41}_{\text{3.Quartil}},\ \underbrace{45, 50, 55, 56, 56}_{\text{4.Quartil}}$$

n = 21	f	Wert Nr. (i)	Kurzbez.	Wert
1. Quartil	1/4	6	$Q_{1/4} = Q_1$	10
2. Quartil	2/4 = $\tilde{x}$	11	$Q_{2/4} = Q_2$	24
3. Quartil	3/4	16	$Q_{3/4} = Q_3$	41
4. Quartil	4/4	21	$Q_{4/4} = Q_4$	56

Für jedes Quantil kann man auch eine Prozentzahl p angeben, wie es beim Median der Fall ist. Es sind stets zwei Fälle zu unterscheiden: Die Multiplikation mit der Anzahl ergibt eine Ganze Zahl oder ergibt keine ganze Zahl. Im zweiten Fall ist der Positionswert immer aufzurunden. (s.Ausführungen zum Median)
Insbesondere spricht man von

- **Quartilen**: $x_{0,25}$=erstes Quartil, Median=zweites Quartil, $x_{0,75}$=drittes Quartil.
- **Dezilen**: $x_{0,10}$ = erstes Dezil, $x_{0,20}$ = zweites Dezil, etc.
- **Perzentilen**: $x_{0,01}$ = erstes Perzentil, etc.

Man kann
- das **untere Quartil** als Median der unteren Datenhälfte ansehen und
- das **obere Quartil** als Median der oberen Datenhälfte.

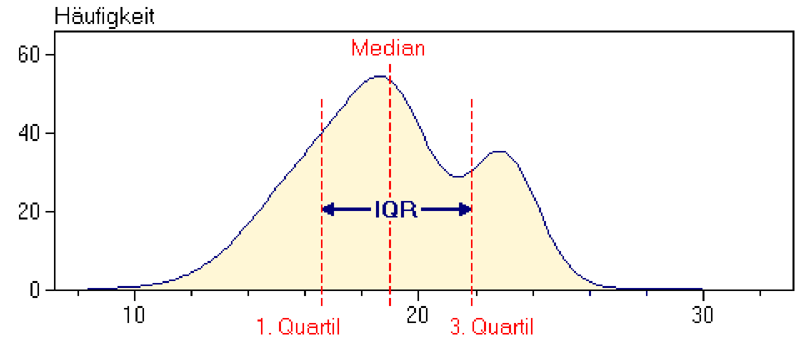
Berechnung des 1. Quartils: 1, 3, 4, 7, 7, 8, 9, 10, 10, 11

$i = f \cdot (n - 1) + 1 = ¼ \cdot 9 + 1 = 3,25$

Q1 = X [floor(3,25] + (X[ceiling(3,25)] − X[floor(3,25)])•(3,25 − floor(3,25))

Q1 = X[3] + (X[4] − X[3]) • (3,25 − 3)

Q1 = 4 + (7 − 4) • 0,25 = 4, 75



Interquartilabstand

Ebenso wie mit Mittelwert und Standardabweichung können Verteilungen auch mittels Median und einer Reihe von Quantilen rund um den Median beschrieben werden. Besonders der Interquartilsabstand IQR wird zur Beschreibung der Streuung von Daten verwendet. Der Interquartilsabstand ist als der Abstand zwischen dem ersten und dem dritten Quartil definiert. Es ist zu beachten, dass der IQR genau 50 % der Daten innerhalb der Verteilung enthält.

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik		

Mittlere Absolute Abweichung

Mittlere Absolute Abweichung einer Urliste

Gegeben sei eine univariate statistische Erhebung mit
- einer Grundgesamtheit mit dem Erhebungsumfang n,
- einem quantitativen Merkmal X und
- der durch die Erhebung gewonnen Urliste mit den Messwerten x1,..., xn , die sich in einer eindeutigen Reihenfolge, z.B. $x_1 \leq. . . .. \leq x_n$ nach steigender Größe anordnen lassen
- einem Mittelwert, meist dem Meridian $x_{0,5}$

Dann berechnet sich die mittlere Absolute Abweichung d vom Mittelwert:

Die Summe der Einzelabstände vom Mittelwert ist für jeden Datensatz gleich Null.

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

Als Maß für die Variation eines Datensatzes kann man die mittlere absolute Abweichung d verwenden

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

Die mittlere absolute (durchschnittliche) Abweichung d ist gleich dem arithmetischen Mittelwert aller Abstände der Daten vom Lage-Maß (meist Median).

Der Begriff absolut, in der Formel durch die vertikalen Striche symbolisiert, meint, daß auch die ja eigentlich negativen Differenzen bei Werten kleiner als das Lage-Maß positiv gezählt werden.

Musterbeispiel:

Gegeben ist die Urliste

x_i	1,3	1,8	1,6	1,7	1,7	1,6	1,7	1,5	1,5	1,5	1,4	1,4	1,5	1,5	1,6	1,8	1,4	1,9	1,7	1,4

mit dem Median $\tilde{x} = 1{,}55$.

Berechne die Mittlere Absolute Abweichung A_{abs} der Messwerte vom Median.

Es ergibt sich

$$A_{abs} = \frac{|1{,}3 - 1{,}55| + ... + |1{,}4 - 1{,}55|}{20} = \frac{0{,}25 + ... + 0{,}15}{20} = 0{,}135$$

Mittlere Absolute Abweichung absoluter Häufigkeiten

Gegeben sei eine univariate statistische Erhebung mit
- einer Grundgesamtheit mit dem Erhebungsumfang n,
- einem quantitativen Merkmal X und
- den Absoluten Häufigkeiten $H(a_1),..., H(a_m)$ der einzelnen Merkmalsausprägungen.
- einem Mittelwert, meist dem Meridian $\tilde{x}$

Dann berechnet sich die mittlere Absolute Abweichung d vom Mittelwert:

$$\frac{\sum_{i=1}^{n} H(a_i)\,|x_i - \bar{x}|}{n} = d$$

Musterbeispiel:

Gegeben sind die Absoluten Häufigkeiten

a_i	1,3	1,4	1,5	1,6	1,7	1,8	1,9
$H(a_i)$	1	4	5	3	4	2	1

mit dem Median $\tilde{x} = 1{,}55$.

Berechne die Mittlere Absolute Abweichung A_{abs} der Messwerte vom Median.

Es ergibt sich

$$A_{abs} = \frac{1 \cdot |1{,}3 - 1{,}55| + ... + 1 \cdot |1{,}4 - 1{,}55|}{20} = \frac{1 \cdot 0{,}25 + ... + 1 \cdot 0{,}15}{20} = 0{,}135$$

Mathematik – Intensivkurs: Statistik		
Thema	**Gesetze und Regeln**	**Musterbeispiele**
	Mittlere Absolute Abweichung relativer Häufigkeiten	

Gegeben sei eine univariate statistische Erhebung mit

- einer Grundgesamtheit mit dem Erhebungsumfang n,
- einem quantitativen Merkmal X und
- den Relativen Häufigkeiten $h(a_1),\ldots, h(a_m)$ der einzelnen Merkmalsausprägungen.
- einem Mittelwert, meist dem Meridian $\tilde{x}$

Dann berechnet sich die mittlere Absolute Abweichung d vom Mittelwert:

$$\sum_{i=1}^{n} h(a_i)\,|x_i - \bar{x}| = d$$

ACHTUNG! Da es sich hier um relative Häufigkeiten handelt wird nicht mehr durch die Anzahl geteilt, die eventuell gar nicht mehr bekannt ist.

Treff Lernen
CHRISTINA STRAUCH

Thema	Gesetze und Regeln	Musterbeispiele

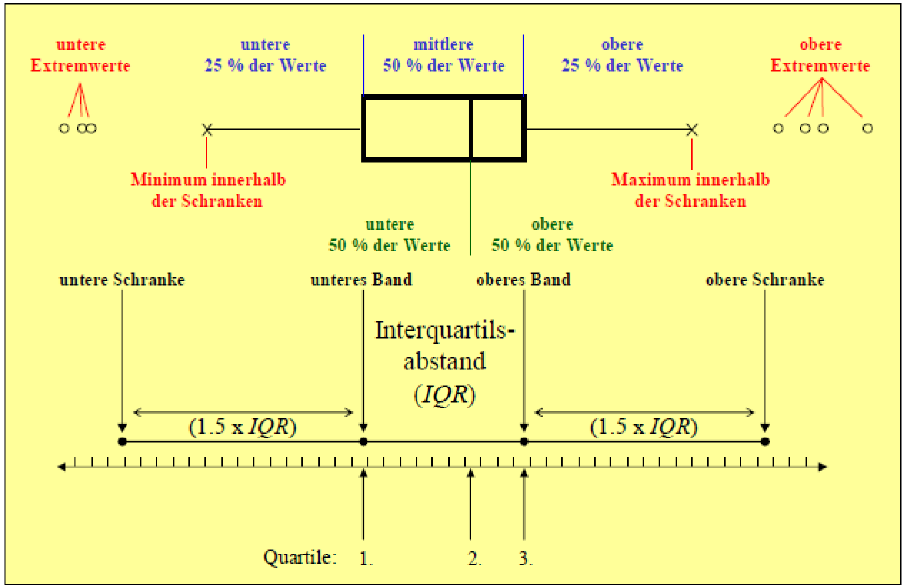
Beschreibende Statistik

Boxplot

Eine kompakte Darstellung der Situation einschließlich der Streuung wird durch den **Boxplot** oder auch **Box & Whisker-Plot** ermöglicht. Die Lage und Breite der „Box" ist dabei durch das erste und dritte Quartil gegeben, die „Whiskers" (Barthaare) erstrecken sich bis zu den Extremwerten.

Tukey's 5 Zahlen:
Minimum, 1. Quartil, Median, 3. Quartil, Maximum

Interquartil Abstand (IQR): 3. Quartil - 1. Quartil



Quartile und Boxplot

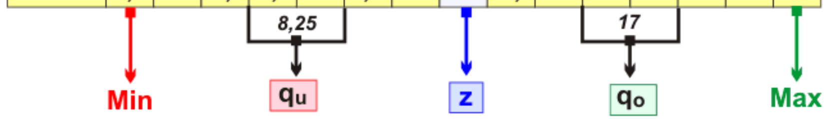
dwu-Unterrichtsmaterialien.de
mdz010fL © 2007

Es ist üblich, die Daten einer geordneten Liste (Rangliste) in Viertel-Bereiche aufzuteilen. Das hierzu passende Diagramm nennt man den Boxplot.

Beispiel einer geordneten Liste (Rangliste):

Bei einem Test hatten die 15 Teilnehmer folgende Ergebnisse:

Rang	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Punkte	2,5	4	5,5	7,5	9	9,5	10	12	13,5	16	16	18	22	23	25

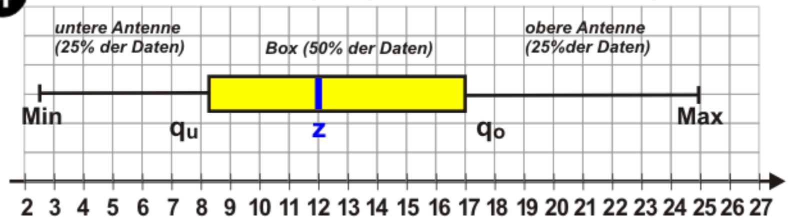


1 Der Zentralwert z (Median) ist bei 15 Teilnehmern der Wert von Rang 8, weil gleichviele Werte oberhalb und unterhalb davon liegen (z = 12).
Ergeben sich zwei mittlere Werte, dann ist es der Mittelwert aus den beiden.

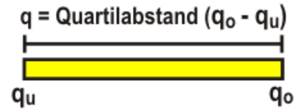
2 Der untere Quartilwert q_u wird ebenso im Bereich zwischen Min und z bestimmt.
Da es hier zwei mittlere Werte gibt, nimmt man den Durchschnittswert.
Hier gilt für q_u (7,5 + 9):2 = 8,25
(Beim vereinfachten Verfahren nimmt man den Wert mit dem höheren Rang.)

3 Der obere Quartilwert q_o wird ebenso im Bereich zwischen z und Max bestimmt.
Da es hier zwei mittlere Werte gibt, nimmt man den Durchschnittswert.
Hier gilt für q_o (16 + 18):2 = 17
(Beim vereinfachten Verfahren nimmt man den Wert mit dem höheren Rang.)

4 Mit den 5 Kennwerten Min, q_u, z, q_o und Max lässt sich der Boxplot zeichnen.



Den Bereich zwischen Minimum und Maximum nennt man die Spannweite w



Den Bereich zwischen dem unteren Quartil und dem oberen Quartil nennt man den Quartilabstand q

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele

Thema: Beschreibende Statistik

Maßzahlen für die Form der Verteilung

Für multimodale Häufigkeitsverteilungen sind dies vor allem Lage (d.h. Modus x_D) und Höhe aller "Gipfel". Für unimodale Verteilungen gibt es zwei Kategorien von Kennzahlen:

- Asymmetriemaße, z.B. "Schiefe" S,
- Wölbungsmaße, z.B. "Exzess" (Kurtosis) E.

Zur Definition von Schiefe und Exzess werden die zentralen Momente benötigt:

Das N-te zentrale Moment von n kardinalskalierten Merkmalswerten ist gegeben durch:

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

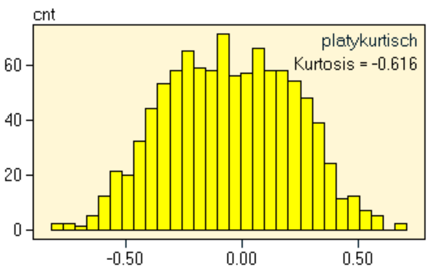
Kurtosis (Wölbung)

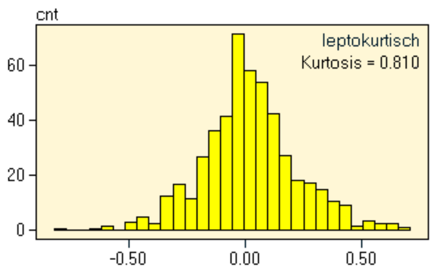
Die Kurtosis (auch Exzess oder Wölbung) ist ein Maß für die relative "Flachheit" einer Verteilung (im Vergleich zur Normalverteilung, die eine Kurtosis von null aufweist). Eine positive Kurtosis zeigt eine spitz zulaufende Verteilung (eine so genannte leptokurtische Verteilung), wohingegen eine negative Kurtosis eine flache Verteilung (platykurtische Verteilung) anzeigt. Die Kurtosis einer Stichprobe ist durch folgende Formel bestimmt:

$$g_2 = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4\right) - 3 = \frac{M_4}{\sigma^4} - 3$$

Diese Formel für die Kurtosis der Stichprobe ist ein nicht erwartungstreuer Schätzer der Kurtosis der Grundgesamtheit. Um die Kurtosis für die Population zu schätzen verwendet man daher folgende Formel:

$$G_2 = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4\right) - \frac{3(n-1)^2}{(n-2)(n-3)}$$





Musterbeispiele

- Kennziffern der Verteilungsform sind Schiefe (skewness) und Wölbung (kurtosis).
- Die Maße der Schiefe und der Wölbung ändern sich bei Lineartransformationen der Werte nicht.
- Ist das Maß der Schiefe positiv, wird die Verteilung als rechtsschief (oder linkssteil) bezeichnet. Dabei gilt: Modus < Median < arithmetisches Mittel.
- Ist das Maß der Schiefe negativ, wird die Verteilung als linksschief (oder rechtssteil) bezeichnet. Dabei gilt: arithmetisches Mittel < Median < Modus.
- Die Flächenanteile unter der Kurve der Standardormalverteilung $\mu = 0.0$, $\sigma = 1.0$) sind:
  - $[-1.00\ \sigma, +1.00\ \sigma] = 68.3\%$
  - $[-1.96\ \sigma, +1.96\ \sigma] = 95.0\%$
  - $[-2.58\ \sigma, +2.58\ \sigma] = 99.0\%$
- Verteilungskurven normalverteilter Werte mit Standardabweichungen $\neq 1.0$ schneiden die Kurve der Standardnormalverteilung auf jeder Seite nur an einer Stelle. Durch z-Standardisierung der Werte lässt sich jede Normalverteilung in die Standardnormalverteilung überführen.
- Kurtosis und Varianz sind unabhängig von einander.
- Ist das Maß der Kurtosis $\neq 0$, schneidet die Verteilungskurve die Kurve der Standardnormalverteilung auf jeder Seite (links und rechts) an zwei Stellen.
- Ist das Maß der Kurtosis positiv, verläuft die Verteilungskurve steiler als die Kurve der Standardnormalverteilung, dabei hat die Verteilungskurve „schwächere Schultern": Es liegt mehr „Masse" im Zentrum und an den Enden der Verteilung.
- Ist das Maß der Kurtosis negativ, verläuft die Verteilungskurve flacher als die Kurve der Standardnormalverteilung, dabei hat die Verteilungskurve „stärkere Schultern": Es liegt weniger „Masse" im Zentrum und an den Enden der Verteilung.
- Normalverteilte Werte haben eine kurtosis von 0.0, gleichverteilte eine kurtosis von -1.2; zweigipflig verteilte eine kurtosis < -1.2.
- Bei nicht normalverteilten Werten (Schiefe $\neq 0$ und/oder Kurtosis $\neq 0$) ist die gängige Interpretation der Flächenanteile anhand der Standardabweichung und des arithmetischen Mittels als Maß der zentralen Tendenz möglicherweise unangemessen!

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik	**Schiefe**	Linksseitige Asymmetrie

Beschreibende Statistik

Schiefe

Beim Vorhandensein einer echten Schiefheit fallen arithemetisches Mittel $\bar{x}$ und Modus M (Dichtemittel) auseinander.
Die Schiefheit tritt um so weniger in Erscheinung, je größer die Streuung des Datensatzes ist. Dementsprechend gilt für die Schiefe nach Pearson:

$$S = \frac{\bar{x} - M}{s}$$

Eine Verteilung wird rechtsschief (bzw. linkssteil) genannt, wenn der Hauptanteil der Verteilung auf der linken Seite konzentriert ist. Für linksschiefe (bzw. rechtssteile) Verteilungen gilt dasselbe für die rechte Seite der Verteilung. Der Grad der Schiefe wird durch das dritte Moment der Verteilung bestimmt:

$$g_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3 = \frac{M_3}{\sigma^3}$$

Diese Formel für die Schiefe der Stichprobe ist ein nicht erwartungstreuer Schätzer der Schiefe der Grundgesamtheit. Um die Schiefe für die Population zu schätzen verwendet man daher folgende Formel:

$$G_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

Um zu testen, ob die berechnete Schiefe tatsächlich auf eine schiefe Verteilung schließen lässt, berechnet man folgende Testgröße:

$$Z_n = g_1\sqrt{\frac{n}{6}}$$

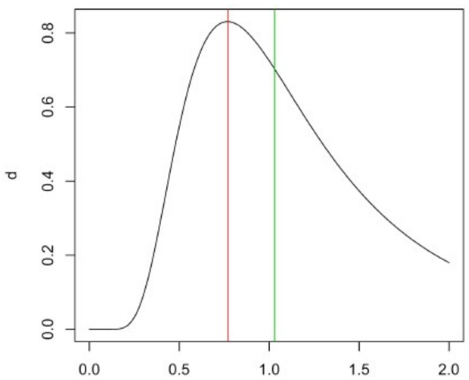
Übersteigt diese das (1-α/2)-Quantil der Standardnormalverteilung, so kann man die Annahme der Symmetrie auf einem Signifikanzniveau von verwerfen. Für diesen Test sollte n größer als 100 sein.

Hinweis: In manchen Statistikprogrammen wird die oben berechnete Testgröße Z_n als "standardisierte Schiefe" bezeichnet. Als Faustregel gilt, dass mit 95%iger Wahrscheinlichkeit eine schiefe Verteilung vorliegt, wenn die standardisierte Schiefe kleiner als -2 bzw. größer als +2 ist.

Bitte beachten Sie, dass die Schiefe manchmal durch unterschiedliche Formeln definiert wird, was zu veschiedenen Schiefemaßen führt.

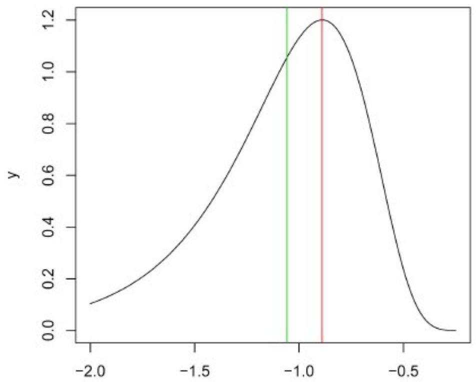
Musterbeispiele

Linksseitige Asymmetrie



$S = \frac{\bar{x}-M}{s} > 0$
Linkssteil, rechtsschief

Rechtsseitige Asymmetrie



$S = \frac{\bar{x}-M}{s} < 0$
Linksschief, rechtssteil

Treff Lernen
CHRISTINA STRAUCH

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele

Beschreibende Statistik

Varianz

Zusätzlich zu den Lagemaßen, die man für die Beschreibung der Position der Verteilung einer Variablen braucht, muss man die Ausdehnung der Verteilung kennen und auch ihre Form.
Die Ausdehnung einer Verteilung kann durch verschiedene Parameter beschrieben werden, von denen die Varianz die gebräuchlichste ist. Die Varianz v ist die Summe der quadrierten Abweichungen vom Mittelwert dividiert durch die Zahl der Proben minus 1.

Die Varianz (σ^2) bezieht sich auf den Mittelwert der Verteilung, indem sie die durchschnittliche quadrierte Abweichung vom Mittelwert berechnet.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = M_2$$

Die Summe $\sum_{i=1}^{n} (x_i - \bar{x})^2$ wird als Summe der Abstandsquadrate bezeichnet.

Bei genauerer Betrachtung dieser Formel ergeben sich folgende Fragen:

- Warum wird die Quadratsumme und nicht etwa die Summe der absoluten Abweichungen vom Mittelwert zur Berechnung verwendet? Die Antwort: Die mathematische Analyse ist einfacher, wenn die Quadratsumme verwendet wird.
- Warum wird die Summe durch n-1 dividiert; wäre es nicht logischer, einfach n zu verwenden? Die Antwort ist nicht schwer, bedarf aber der Einführung des Konzepts der Freiheitsgrade.
- Was ist mit σ^2 in der Formel? Die Antwort: Der Parameter σ, der die Quadratwurzel der Varianz ist, wird Standardabweichung genannt.

Varianz σ^2 für gruppierte Daten

$$\boldsymbol{\sigma^2} = \frac{\sum_{i=1}^{m} h_i \cdot (x_i - \bar{x})^2}{\sum_{i=1}^{m} h_i} = \sum_{i=1}^{n} \pi_i \cdot (x_i - \bar{x})^2$$

x_i Note
h_i Häufigkeit
π_i Relative Häufigkeit

Berechnung die Notenvarianz für eine Gruppe von Schülern:

Note (x_i):	1	2	3	4	5	
Häufigkeit (h_i)	2	4	8	2	1	Σ = 17
Rel. Häufigk. (π_i)	2/17	4/17	8/17	2/17	1/17	Σ = 1
$(x_i - x)^2 \cdot h_i$	6,228	2,339	0,443	3,052	4,997	Σ = 17,05

σ^2 = 17,059 / 17 = 1, 003
Standardabweichung σ = 1,00

Mathematik – Intensivkurs: Statistik

Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik		

Varianz σ² für klassifizierte Daten

Sind die Daten in K Klassen mit den Klassenmitten m_k und den Häufigkeit h_k klassifiziert, so gilt

$$\sigma^2 = \frac{\sum_{k=1}^{K} (m_k - \bar{x})^2 \cdot h_k}{\sum_{k=1}^{K} h_k} = \sum_{k=1}^{K} (m_k - \bar{x})^2 \cdot \pi_k$$

Das berechnete Varianz ist nur ein Näherungswert, da durch die Klassifizierung Information verloren geht.

Varianz und Standardabweichung einer Stichprobe

Da wir es in der Statistik meistens mit Stichproben zu tun haben, ist mit „Varianz“ und „Standardabweichung“ meistens die auf die Stichprobe bezogene empirische Varianz (s^2) bzw. Standardabweichung (s) gemeint:

$$\mathbf{s^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Variationskoeffizient

Spezifiziert man nur die Standardabweichung, kann man wenig über eine Verteilung aussagen, da diese u.a. von den absoluten Maßzahlen der Messungen (z.B. Messungen in mm oder km) abhängt. Es macht natürlich einen erheblichen Unterschied, ob eine Standardabweichung von 5 bei einem Mittelwert von = 100 oder einem Mittelwert von = 3 auftritt. Am einfachsten kann man dieses Problem durch den Bezug der Standardabweichung auf den Mittelwert lösen. Wir definieren darum den Variationskoeffizienten durch

$$c_v = \frac{\sigma}{\bar{x}}$$

das nun ein relatives Maß der Standardabweichung darstellt. Wie auch immer, man sollte sich immer im Klaren sein, dass der Variationskoeffizient mehr oder weniger unbrauchbar wird, wenn der Mittelwert gegen Null geht. Man sollte daher den Variationskoeffizienten keineswegs z.B. für den Vergleich von Nachweisgrenzen verwenden

Thema	Gesetze und Regeln	Musterbeispiele
Beschreibende Statistik	**Standardabweichung**	

Um die Varianz in ein Maß zu transformieren, das die gleiche Einheit wie die Ausgangsdaten besitzt, berechnet man die Quadratwurzel:

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

σ wird als Standardabweichung bezeichnet.

Es gibt manchmal einige Verwirrungen bezüglich der Standardabweichung und ihrer Interpretation. Man sollte vorsichtig zwischen der formalen Definition der Standardabweichung und deren Interpretation unterscheiden. Die Standardabweichung kann als Zahlenwert immer berechnet werden, vorausgesetzt, es sind genügend Proben verfügbar. Im Gegenteil dazu kann die Interpretation der Standardabweichung als ein Maß der Breite nur dann vollständig verwendet werden, wenn die Art der Verteilung bekannt ist. Das Theorem von Tschebyscheff gibt jedoch einige Richtlinien für jede beliebige (!) Verteilung.

Bitte beachten Sie bei der Bezeichnung für die Varianz und die Standardabweichung:

- Sie wird mit s^2 (bzw s) bezeichnet, wenn sie aus einer Stichprobe berechnet wurde.
- Wenn sie aus einer Grundgesamtheit berechnet wurde, wird die Standardabweichung durch den griechischen Buchstaben σ (Sigma) dargestellt.