VORKURS

STATISTIK

Glaube nur der Statistik,
die Du selbst gefälscht hast.

Die Notlüge,
die gemeine Lüge
und
die Statistik.

Ablaufdefinitionen:

- Praktische Statistik
  In diesem ersten Schritt werden lediglich Daten gesammelt bzw. Informationen erhoben.
  Es folgen keine Interpretation und Prognosen

- Deskriptive Statistik:
  Die im ersten Schritt erhaltenen Daten werden aufbereitet, gefiltert (gestrafft) und dargestellt.
  Sie endet mit einer Interpretation ohne Prognosen

- Induktive Statistik:
  Die Resultate der ersten beiden Schritte werden mittels mathematischer Methoden und Algorithmen der Wahrscheinlichkeitsrechnungen analysiert.
  Als Ergebnis werden Prognosen für die Zukunft erzeugt

Begriffsdefinitionen I:

- Statistische Einheit:
  Die statistische Einheit stellt ein einzelnes, elementares Objekt dar, das Träger der Informationen bzgl. der Untersuchung ist.

- Identifikationskriterien je Einheit:
  Jede statistische Einheit wird aufgrund des Untersuchungsziels durch **sachliche**, **räumliche** und **zeitliche** Kriterien identifiziert.

- Statistische Masse (Gesamtheit):
  Eine statistische Masse (endlich/unendlich) ist eine Menge von statistischen Einheiten, die mit übereinstimmenden Identifikationskriterien ausgestattet und anhand dieser abgrenzbar sind.

Beispiel: Kommunalwahlen in Hessen

- Masse: Alle wahlberechtigten Personen in Hessen
- Einheit: einzelne wahlberechtigte Bürger
- sachlich: Parteiauswahl
- räumlich: Gebiet der Stadt / Gemeinde
- zeitlich: Wahltag (27.03.2019)

Aufgabe: Studiendauer in Deutschland
Es soll die durchschnittliche Studiendauer von Studierenden an deutschen Hochschulen bis zum Abschluss Bachelor ermittelt werden.

- Masse: Studenten an deutschen Hochschulen (ohne Abschluss)
- Einheit: einzelne(r) Student(in)
- sachlich: Bachelor Absolvent einer Hochschule
- räumlich: Deutschland
- zeitlich: Zeitraum für den die Studiendauer ermittelt wird.

Begriffsdefinitionen II:

- Bestandsmasse:

  Die Bestandsmasse besteht aus Elementen, die auch nebeneinander vorkommen können und für die eine gewisse Lebensdauer (Zeitstrecke) existiert.
  Die Erfassung erfolgt zu einem definierten Zeitpunkt und erzeugt den sogenannten **Bestand**.

  Beispiele:
  Einwohner von Fulda am 27.03.2019.
  Position eines Lagers am 31.12.2019
  Kassenbestand eines Warenhauses am 01.01.2020

Begriffsdefinitionen III:

- Bewegungs-/Ereignismasse:

  Die Ereignismasse besteht aus Elementen, die ausschließlich zu Einem bestimmten Zeitpunkt existieren
  Das Element, das zu einem bestimmten Zeitpunkt die Bestandsmasse verändert, heißt **Ereignis**.

  Beispiele:
  Geburten in Deutschland im Jahr 2019.
  Baufertigstellungen im September 2019.
  Bei einer Bank im Monat Januar 2020 eingegangene Schecks.

  Anfangsbestand + Zugang – Abgang = Endbestand

Merkmale:

- Merkmale:
  In einer statistischen Untersuchung interessierende Eigenschaften nennt man Merkmale
  (statistische Einheit = **Merkmalsträger**).
  Beispiele:
  Merkmalsträger Student
  Merkmale Alter, Schulabschluss, Studienfach...

- Merkmalsausprägung:
  Die möglichen Werte oder Kategorien, die ein Merkmal annehmen kann, heißen Merkmalsausprägungen und werden in verschiedenen Skalen gemessen und eingeteilt.
  Beispiele:
  Merkmal Geschlecht
  Ausprägung männlich, weiblich

- Häufbares Merkmal:
  Ein Merkmal heißt häufbar, wenn an derselben Einheit mehrere Ausprägungen des betreffenden Merkmals zum gleichen Zeitpunkt vorkommen können.
  Beispiele:
  Unfallursache: überhöhte Geschwindigkeit und Trunkenheit

- Merkmalsklassen:
  Sollen bei einer Erhebung/Aufbereitung nicht alle möglichen Ausprägungen erfasst werden, können benachbarte Werte zu einer Klasse zusammengefasst werden.
  Beispiele:
  Einkommensklasse: von 1 bis 2500, von 2501 bis 5000...

Messskalen I:

Die innerhalb einer Untersuchung / Erhebung erzeugten Daten / Werte werden in verschiedenen Skalen gemessen und eingeteilt:

- Nominalskala:
  Die Ausprägungen haben keine natürliche Reihenfolge, sondern existieren gleichberechtigt nebeneinander.
  Religion, Geschlecht, Farbe und Autokennzeichen

- Ordinalskala:
  Es besteht eine natürliche Rangordnung zwischen den Werten, d.h. man kann eine größer als – Beziehung aufstellen.
  Allerdings sind die Abstände nicht quantifizierbar.
  Examensnoten, Güteklasse (Lebensmittel), Rangplätze (Fußballliga)

- Intervallskala:
  Neben der Rangordnung können zusätzlich auch die Abstände zwischen den Merkmalen angegeben werden, wobei der NULL-Punkt frei wählbar ist.
  Temperaturmessung in °C, Kalenderzeitrechnung

Messskalen II:

- Verhältnisskala:
  Zusätzlich zu der Intervallskala existiert ein absoluter NULL-Punkt, jedoch keine natürliche Einheit (Quotient sinnvoll)
  Körpergröße, Alter, Einkommen

- Absolutskala:
  Eine metrische Skala mit einem natürlichen NULL-Punkt und einer natürlichen Einheit

- metrisch: Alle Skalen, denen ein Maßsystem zugrunde liegt.
  (Intervall-/Verhältnisskala)

- diskret: Es werden nur bestimmte Werte angenommen (metrisch).
  Geldeinkommen, Zahl der Studenten

- stetig: Es kann jeder beliebige Wert erreicht werden.
  Füllgewicht, Lebensalter

Aufgabe 1 - Merkmalsausprägung:
Geben Sie zu folgenden Merkmalen mind. 3 Ausprägungen an.

a) Haarfarbe	b) Einkommen	c) Klausurnote
d) Körpergewicht	e) Studienfach	f) Kinderzahl

Aufgabe 2 -Messskalen:
Auf welcher Skala können folgende Merkmale gemessen werden

a) Militärdienstgrad	b) Semesterzahl	c) Alter
d) Temperatur	e)Klausurpunkte	f) Geschlecht
g) Breitengrad	h) Studienfach	i) Körpergröße
j) Fahrpreise	k) Nationalität	l) Handelsklasse

Aufgabe 3 - Häufbarkeit:
Welche der folgenden Merkmale sind häufbar?

a) Körpergröße	b) Schulbildung	c) Alter
d) Kinderanzahl	e) ausgeübter Beruf	f) Augenfarbe

Empirische Verteilung I

- Urliste:

  Werden die bei einer statistischen Untersuchung erhobenen, erzeugten Beobachtungswerte nacheinander aufgeschrieben, so erhält man eine **statistische Reihe (Urliste)**.

  Diese kann geordnet oder ungeordnet dargestellt werden.

  Beispiel:

  Eine Untersuchung vom Kaufverhalten einer bestimmten Zeitung ergab folgende geordnete Reihe (Dauer 30 Tage):

  0,0,0,1,1,1,1,1,1,2,2,2,2,3,3,3,3,3,3,3,3,3,3,4,4,5,5,5,5,5

Empirische Verteilung II

- Absolute Häufigkeit:
  Die Anzahl der Beobachtungswerte mit der Merkmalsausprägung x_j heißt absolute Häufigkeit und wird mit $h(x_j)$ bezeichnet.
  Die Summe aller absoluter Häufigkeiten ergibt die Gesamtzahl n der Beobachtung $\sum_{j=1}^{k} x_j = n$

- **Beispiel:**

h(0)=3	h(1)=6	h(2)=4
h(3)=10	h(4)=2	h(5)=5

Gesamtzahl n = 3 + 6 + 4 + 10 + 2 + 5 = 30

Empirische Verteilung III

- ✓ Relative Häufigkeit:
  Der relative (prozentuale) Anteil der absoluten Häufigkeit $h(x_j)$ einer Merkmalsausprägung x_j an der Gesamtzahl n der Beobachtungswerte heißt relative Häufigkeit und wird mit $f(x_j)$ bezeichnet.
  Es gilt $f(x_j) = \frac{h(x_j)}{n}$, mit $0 \leq f(x_j) \leq 1$ und $\sum_{j=1}^{k} f(x_j) = 1$

  Beispiel:

  $f(0) = \frac{3}{30} = \frac{1}{10}$ $\quad f(1) = \frac{6}{30} = \frac{1}{5}$ $\quad f(2) = \frac{4}{30} = \frac{2}{15}$

  $f(3) = \frac{10}{30} = \frac{1}{3}$ $\quad f(4) = \frac{2}{30} = \frac{1}{15}$ $\quad f(5) = \frac{5}{30} = \frac{1}{6}$

Darstellungen I

- Tabelle:
  Eine Tabelle dient der systematischen und übersichtlichen Zusammenstellung von Daten.
  Wichtig: Leichte Lesbarkeit und unmissverständliche Bezeichnungen

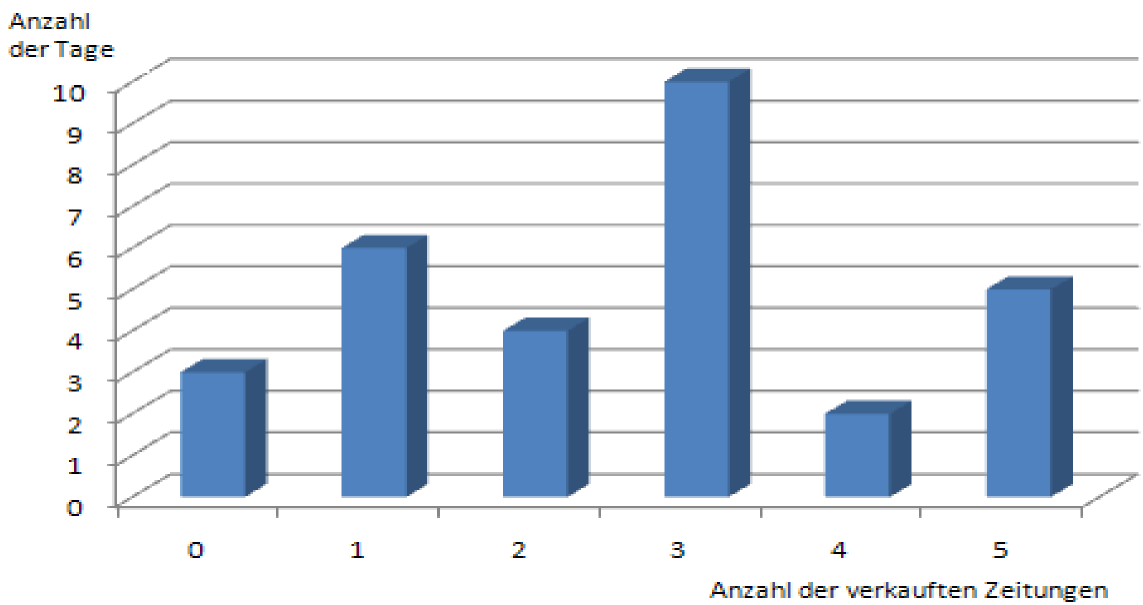
Beispiel:

Verkaufte Zeitungen	0	1	2	3	4	5	∑
Anzahl der Tage	3	6	4	10	2	5	30
Anteil der Tage in %	10%	20%	13%	33%	7%	17%	100%

Darstellungen II

- Stab- / Säulendiagramm :
  Die Häufigkeiten der Merkmale werden durch Längen von Strecken dargestellt (höhenproportionale Darstellung).
  Die Stäbe bzw. Säulen können noch unterteilt werden und haben meist Zwischenräume (siehe Wahlergebnisse).
  Es ist für nominal und ordinal messbare Merkmale geeignet.

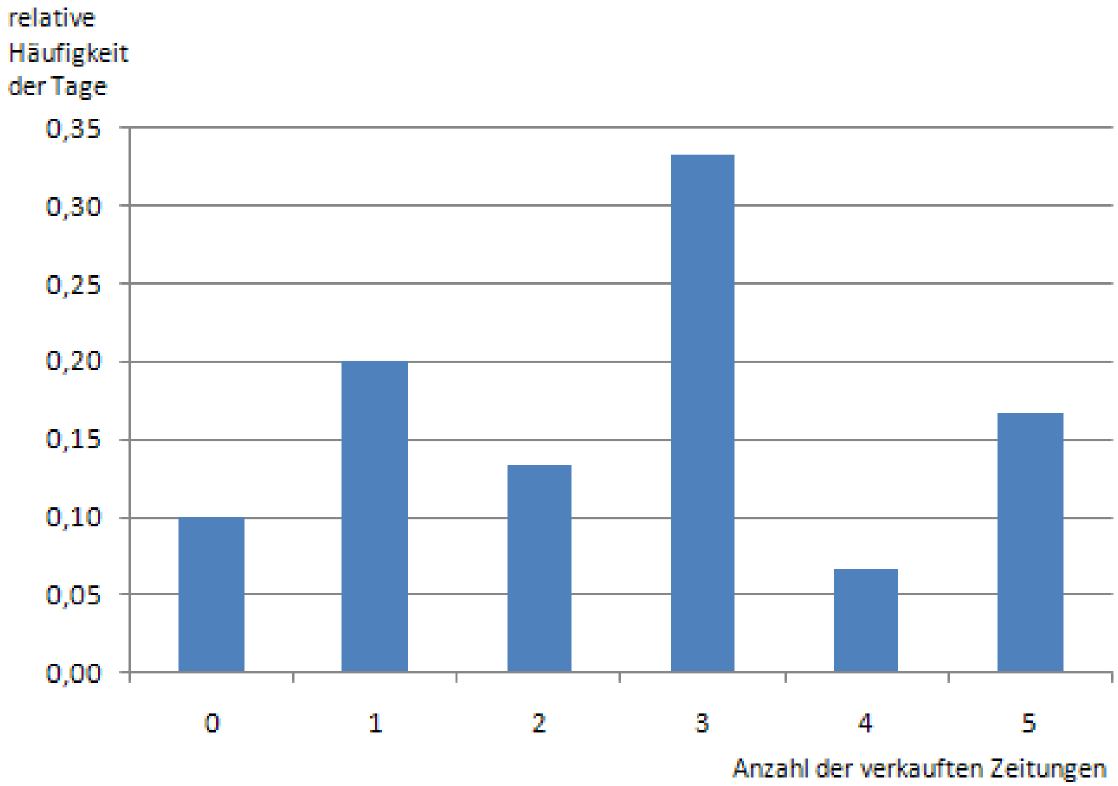
Beispiel:



Darstellungen III

- <u>Histogramm:</u>
  Ein Histogramm wird mittels eines Säulendiagramms dargestellt, wobei der Flächeninhalt der Säulen exakt der relativen Häufigkeit entspricht, so dass die Summe aller Flächen stets 1 ergibt.

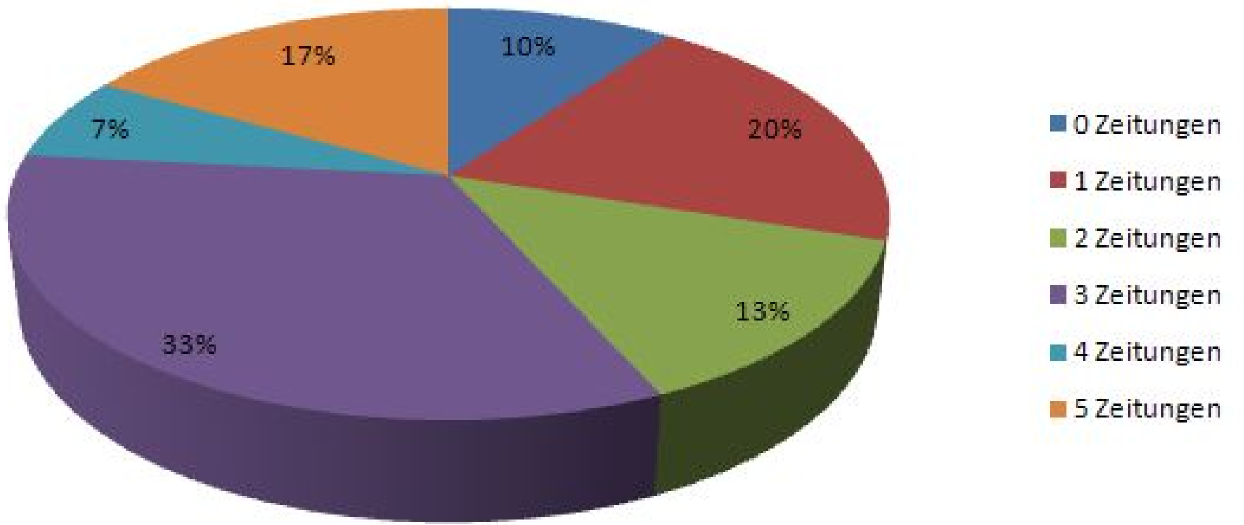
Beispiel:



Darstellungen IV

- <u>Kreisdiagramm:</u>
  Die grafische Darstellung von relativen Häufigkeiten durch die sektorale Aufteilung einer Kreisfläche heißt Kreisdiagramm.
  Wichtig: Bei zu großen Merkmalsausprägungen nicht einsetzbar.

Beispiel:



Summenhäufigkeit (kleiner gleich) I

✓ Definition:

Durch fortlaufende Summierung (Kumulierung) lassen sich aus den absoluten Häufigkeiten die **absoluten** Summenhäufigkeiten H_i wie folgt bilden:

$$H(x_i) = h(x_1) + h(x_2) + \cdots + h(x_i) = \sum_{j=i}^{i} h(x_i)$$

H_i gibt die Anzahl der Elemente an, die einen Merkmalswert besitzen, der höchstens x_i beträgt. In entsprechender Weise lassen sich die **relativen** Summenhäufigkeiten F_i berechnen:

$$F(x_i) = f(x_1) + f(x_2) + \cdots + f(x_i) = \sum_{j=i}^{i} f(x_i)$$

oder

$$F(x_i) = \frac{H(x_i)}{N}$$

Summenhäufigkeit II

Mit Hilfe der relativen Summenhäufigkeit lässt sich die Summenhäufigkeitsfunktion (**empirische Verteilungsfunktion**) $F(x)$ definieren. $F(x)$ gibt den Anteil der Elemente mit einem Merkmalswert kleiner oder gleich x an.

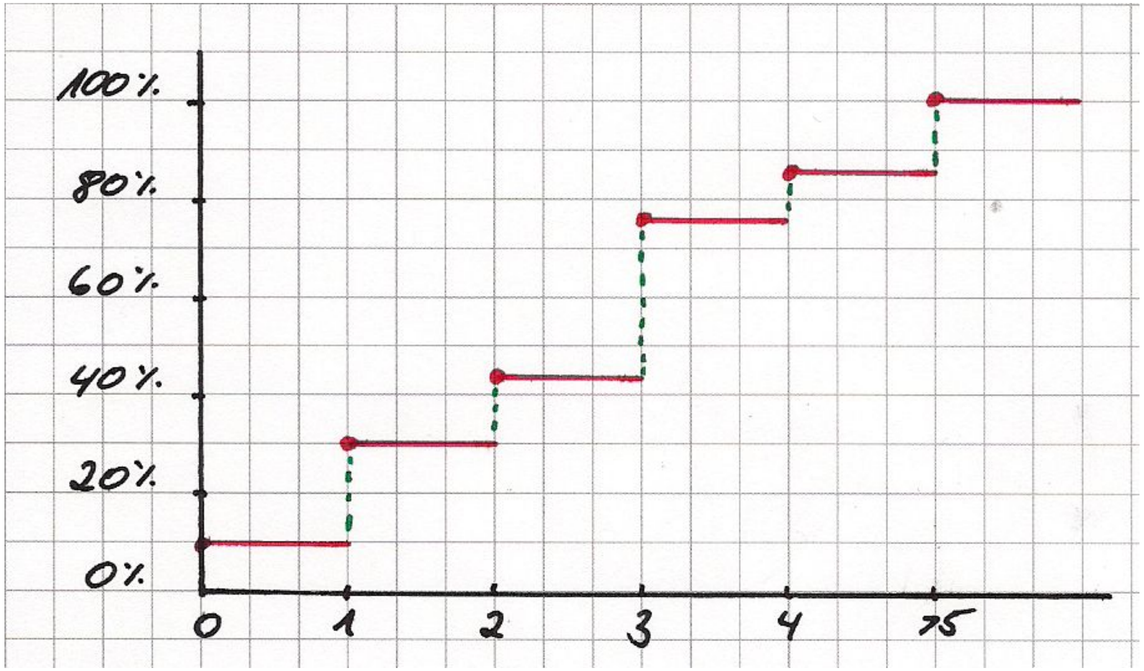
Beispiel:

Verkaufte Zeitungen	**0**	**1**	**2**	**3**	**4**	**>5**
Anzahl der Tage	3	6	4	10	2	5
Anteil der Tage in %	10%	20%	13%	33%	7%	17%
absolute Summenhäufigkeit	3	9	13	23	25	30
relative Summenhäufigkeit	10%	30%	43%	77%	83%	100%

Darstellung Summenhäufigkeit I

<u>Treppenfunktion:</u>
Ähnlich wie bei der Gauß-Klammer-Funktion werden die zugehörigen relativen, kumulierten Häufigkeiten eingetragen, wobei die Funktion an jedem Merkmalswert auf die Summenhäufigkeit springt.

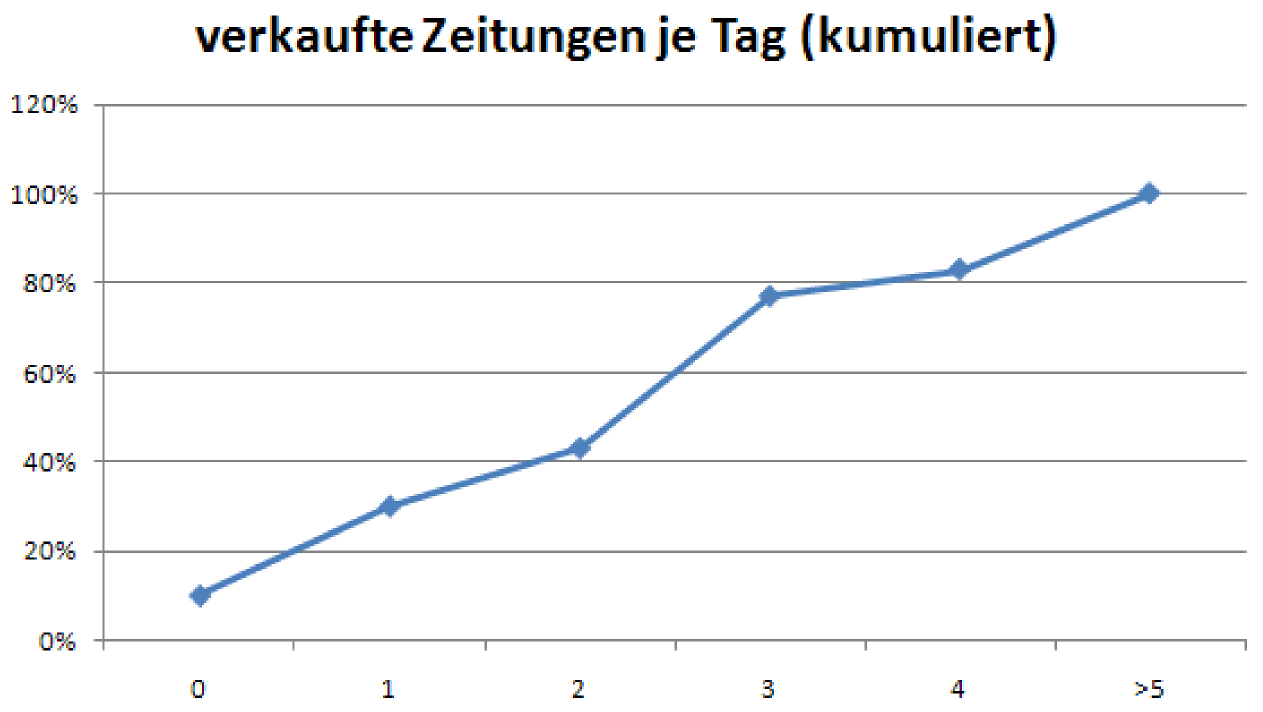
Beispiel:



Darstellung Summenhäufigkeit II

<u>Liniendiagramm:</u>
Es werden die jeweiligen Punkte zu den Merkmalswerten in ein Koordinatensystem eingetragen und diese Punkte anschließend mittels einer Geraden verbunden.

Beispiel:



Resthäufigkeit (größer gleich)

Definition:

Die einer Merkmalsausprägung oder Klasse zugeordnete Häufigkeit der Werte, die diese Ausprägung / Grenze überschreiten, heißt Resthäufigkeit und stellt somit die Umkehrung / das Gegenteil der Summenhäufigkeit dar.

absolute Resthäufigkeit:

$$H_R(x_i) = \sum_{x_k > x_i} h(x_k) = N - H(x_i)$$

relative Resthäufigkeit:

$$F_R(x_i) = \sum_{x_k > x_i} f(x_k) = 1 - F(x_i)$$

Beispiel:

Verkaufte Zeitungen	0	1	2	3	4	5
Anzahl der Tage	3	6	4	10	2	5
Anteil der Tage in %	10%	20%	13%	33%	7%	17%
absolute Summenhäufigkeit (bis)	3	9	13	23	25	30
relative Summenhäufigkeit	10%	30%	43%	77%	83%	100%
absolute Resthäufigkeit (über)	27	21	17	7	5	0
relative Resthäufigkeit	90%	70%	57%	23%	17%	0%

Aufgaben:

1) An einer Prüfung, bei der maximal 10 Punkte erreicht werden können, nahmen 50 Personen teil.

   Es wurde folgende Ergebnisreihe erzielt:

   0,1,1,1,2,2,2,2,3,3,4,4,4,4,4,5,5,5,5,5,5,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,7,7,8,8,8,8,9,9,9,9,9,10,10.

a) Berechnen Sie die absolute und relative Häufigkeit.
b) Bestimmen Sie die absolute und relative Summenhäufigkeiten.
c) Bestimmen Sie die absolute und relative Resthäufigkeiten.
d) Zeichnen Sie Ihre Ergebnisse in ein Schaubild Ihrer Wahl.

Statistische Maßzahlen

Bisher wurden die auftretenden Merkmalsausprägungen verdichtet, klassifiziert und zusammengefasst. Daraus entstanden verschiedene Häufigkeitsformen, Klassenbildungen und Darstellungsformen. Mittels der statistischen Maßzahlen werden diese Informationen noch weiter komprimiert und in knapper Form interpretiert.

- ✓ Mittelwert:

  Das arithmetische Mittel sollte stets bei metrisch skalierten Merkmalen gebildet werden und entspricht dem Durchschnitt einer Untersuchung:

$$\mu = \frac{1}{N}(a_1 + a_2 + \cdots + a_N) = \frac{1}{N}\sum_{i=1}^{N} a_i$$

Oder, sofern ein Beobachtungswert mehrfach vorkommt, wird der Mittelwert mittels der absoluten Häufigkeiten gebildet:

$$\mu = \frac{1}{N}(x_1 h_1 + x_2 h_2 + \cdots + x_k h_k) = \frac{1}{N}\sum_{i=1}^{k} x_i h_i$$

Beispiel:

Verkaufte Zeitungen	0	1	2	3	4	5
Anzahl der Tage (h_i)	3	6	4	10	2	5
Anzahl der Zeitungen ($x_i h_i$)	0	6	8	30	8	25

Summe der Tage: $N = 30$

Summe der Zeitungen: $X = 77$

Arithmetischer Mittelwert: $\mu = \frac{X}{N} = \frac{77}{30} \cong 2{,}5$

Demzufolge wurden innerhalb der letzten 30 Tage im Durchschnitt 2,5 Zeitungen je Tag verkauft.

✓ <u>Modus / Modalwert:</u>

Dieser Wert gibt an, welcher Wert am häufigsten vorkommt und wird u.a. auch dichtester Wert genannt.
Zusätzlich kann er auch auf nominalen bzw. ordinalen Skalen bestimmt werden:

$$Modus = \overline{X_D} = \max\left(h(x_i)\right)$$

Beispiel:
In unserem Beispiel ergibt sich als Modus 3 Zeitungen je Tag.

✓ <u>Median / Zentralwert :</u>

Der Median ist die Merkmalsausprägung desjenigen Elements, das in der der Größe nach geordneten Beobachtungsreihe in der Mitte steht.
Daraus ergibt sich automatisch, dass mindestens eine ordinale Skala vorhanden sein muss.

<u>ungerade Anzahl N:</u> $\overline{X_Z} = x_{\left[\frac{n+1}{2}\right]}$

5,5,6,6,6,7,8,8,9,9,10 $\overline{X_Z} = x_{\left[\frac{11+1}{2}\right]} = x_6 = 7$

<u>gerade Anzahl N:</u> $\overline{X_Z} = \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right)$

1,1,2,2,3,3,3,4,4,5,6,7 $\overline{X_Z} = \frac{1}{2}\left(x_{\left[\frac{12}{2}\right]} + x_{\left[\frac{12}{2}+1\right]}\right) = \frac{1}{2}(3+3) = 3$

Liegen die Daten in Form einer Häufigkeitsverteilung vor, dann ist der Median die Ausprägung, bei der die Summenhäufigkeit den Wert 0,5 überschreitet.

Beispiel:
In unserem Beispiel ergibt sich als Median 3 Zeitungen, da an der Stelle 3 der Wert 0,5 (mit 77%) zum ersten Mal überschritten wird.

- ✓ <u>Mittelwert (geometrisch):</u>

  Hat man es mit zeitlich aufeinanderfolgenden Wachstumsraten, Zuwächsen oder ähnlichen zu tun, ist das arithmetische Mittel nicht der sachlich korrekte Durchschnittswert, sondern das geometrische Mittel:

$$\overline{X_G} = \sqrt[N]{{x_1}^{h(x_1)} \cdot {x_2}^{h(x_2)} \cdot \ldots \cdot {x_k}^{h(x_k)}} = \sqrt[N]{\prod_{i=1}^{k} {x_i}^{h(x_i)}}$$

Beispiel:

Im Jahr 1 einer Untersuchung macht ein Unternehmen 1 Mio Umsatz. Im darauffolgenden Jahr 1,8 Mio und im 3. Jahr 1.98 Mio.

<u>Gesamtzuwachsrate</u>: $1{,}8 * 1{,}1 = 1{,}98$ demzufolge 98%

<u>Geometrische Mittel</u>: $\overline{X_G} = \sqrt[2]{1{,}8 * 1{,}1} = \sqrt[2]{1{,}98} \cong 1{,}4$

✓ <u>Mittelwert (harmonisch):</u>

Beziehen sich die Merkmalsausprägungen auf verschiedene Bezugssysteme wie z.B. die Durchschnittsgeschwindigkeit auf unterschiedlich lange Teilstrecken, liefert das harmonische Mittel den exakten Wert:

$$\overline{X_H} = \frac{N}{\sum_{i=1}^{k} \frac{h(x_i)}{x_i}}$$

Beispiel:

Teilstrecke Nr.	1	2	3	4
Länge der Teilstrecke [km]	30	10	40	20
Geschwindigkeit [km/h]	40	50	80	100

$$\overline{X_H} = \frac{30 + 10 + 40 + 20}{\frac{30}{40} + \frac{10}{50} + \frac{40}{80} + \frac{20}{100}} = \frac{100}{1{,}65} = 60{,}6$$

Die Durchschnittsgeschwindigkeit beträgt demzufolge 60,6 km/h.

Aufgabe I:

1. Ein Angestellter legt den Weg zu seiner Arbeitsstätte an 5 Tagen in 12, 10, 16, 12 und 17 Minuten zurück.

   a) Berechnen Sie das geometrische Mittel.
   b) Bestimmen Sie Modus und Median.
   c) Wie groß ist das 30% und das 80%-Quantil?

2. Ein Unternehmen erzielte in den Jahren von 2003-2007 folgende Jahresumsätze:

Jahr	2003	2004	2005	2006	2007
Umsatz (in MIO)	2,0	2,4	2,9	2,7	3,1

   a) Berechnen Sie den jährlichen prozentualen Zuwachs.
   b) Wie groß ist der durchschnittliche, relative Jahreszuwachs?

Aufgabe II:

3. Ein Formel-1-Fahrer notierte die Anzahl seiner Startplatzierungen aus den letzten 11 Rennen:

   1 5 2 3 1 4 5 2 1 3 4

   a) Berechnen Sie den Modus, Median und Durchschnittswert.
   b) Bestimmen Sie den 25% als auch 75%-Quantil.

4. Zeigen Sie, dass folgende Gleichung gilt:

$$\sum_{i=1}^{k} (x_i - \mu) \cdot f_i = 0$$

✓ <u>Quantile (Definition):</u>

Durch den Zentralwert (Median) wird eine Beobachtungsreihe in zwei gleich große Teile zerlegt.

Wird nun eine solche Kette in k-Teile auf gespaltet, so spricht man von k-Quantilen (Quartile, Dezentile, Perzentile[%]).

Definition:

Die Merkmalsausprägungen $\overline{X_{p/k}}$ $(p = 1, \ldots, k-1)$ eines mind. ordinal messbaren Merkmals, die die geordnete Reihe der Beobachtungswerte in k gleiche Teile zerlegen, heißen k-Quantile.

Es gilt:

$$\sum_{x_i < x_{p/k}} f(x_i) \leq \frac{p}{k}$$

und

$$\sum_{x_i > x_{p/k}} f(x_i) \leq 1 - \frac{p}{k}, p = 1, \ldots, k-1$$

$\overline{x_{p/k}}$ heißt p-tes k-Quantil.

✓ <u>Quantile (Algorithmus):</u>

Ein Quantil beschreibt bis zu welchem Wert ein bestimmter Prozentsatz einer geordneten Reihe erreicht wird bzw. dieser überschritten wird.

Bestimmung des k-Quantils Q_k:

- Sortierung der Urliste (aufsteigend)
- Bestimmung der Position $i = k \cdot n$
- Ist die Position ganzzahlig:
  $i \in \mathbb{Z} \rightarrow Q_k = \frac{1}{2} \cdot (x_i + x_{i+1})$
- Ist die Position nicht ganzzahlig:
  $i \notin \mathbb{Z} \rightarrow$ Aufrunden von i auf die nächste ganze Zahl

- <u>Quantile (Beispiel):</u>

Betrachtet wird die Beobachtungsreihe der folgenden Urliste $2; 5; 8; 4; 10$.
Unabhängig von dem gesuchten Quantil, musss diese Liste im ersten Schritt sortiert werden, so dass wir $2; 4; 5; 8; 10$ erhalten.

- Gesucht wird das 0,2 Quantil oder auch das 20. Perzentil.
  Bestimmung der Position: $i = k \cdot n = 0{,}2 \cdot 5 = 1$
  Bestimmung des Quantils: $1 \in \mathbb{Z} \rightarrow Q_{0,2} = \frac{1}{2} \cdot (x_1 + x_2)$
  $Q_{0,2} = \frac{1}{2} \cdot (2 + 4) = 3$

Also sind mindestens 20% der Werte kleiner oder gleich 3 und mindestens 80% der Werte größer oder gleich als 3.

- Gesucht wird das 0,7 Quantil oder auch das 70. Perzentil.
  Bestimmung der Position: $i = k \cdot n = 0{,}7 \cdot 5 = 3{,}5$
  Bestimmung des Quantils: $3{,}5 \notin \mathbb{Z} \rightarrow$ Aufrunden von 3,5
  $Q_{0,7} = x_4 = 8$

Also sind mindestens 70% der Werte kleiner oder gleich 8 und mindestens 30% der Werte größer oder gleich der 8.

✓ <u>Quartile:</u>

Wird eine Messreihe in die jeweiligen 25% Teilsgemente zerlegt, so nennt man diese Quartile. Dabei entspricht das 2. Quartil dem Median (50% = Mitte).

Will man nun den Bereich wissen, in dem sich 50% aller Messwerte aufhalten und der in der Mitte der Untersuchung liegt, so bestimmt man die Differenz zwischen dem 3. und 1. Quartil. Dieser Abstand wird auch als Interquartilabstand bezeichnet.

$$IQA = Q_{0,75} - Q_{0,25}$$

Beispiel: 5,6,6,7,7,7,8,9,9,11,11,13,13,14,14,14,16,17,19,28

1. Quartil: $\mathrm{i} = 0{,}25 \cdot 20 = 5 \rightarrow Q_{0,25} = \frac{1}{2} \cdot (7 + 7) = 7$

2. Quartil: $\mathrm{i} = 0{,}5 \cdot 20 = 10 \rightarrow Q_{0,5} = \frac{1}{2} \cdot (11 + 11) = 11$

3. Quartil: $\mathrm{i} = 0{,}75 \cdot 20 = 15 \rightarrow Q_{0,75} = \frac{1}{2} \cdot (14 + 14) = 14$

$$IQA = Q_{0,75} - Q_{0,25} = 14 - 7 = 7$$

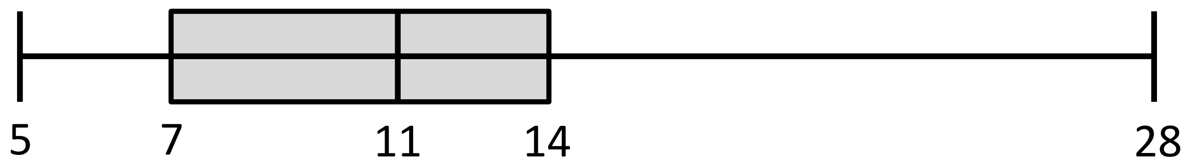
✓ <u>Box-Plot:</u>

Um die wesentlichen bzw. charakteristischen Zahlen einer Messreihe und deren Verteilung darzustellen, nutzt man die Grafik Box-Plot.

In Excel heißt diese Art der grafischen Darstellung Kastengrafik.

Es werden dabei 5 Werte näher betrachtet.	Beispiel
➢ Kleinster Wert	5
➢ 1. Quartil	7
➢ 2. Quartil (Median)	11
➢ 3. Quartil	14
➢ Größter Wert	28

Box-Plot:



Statistische Streuungsmaße I

Die zuvor beschriebenen statistischen Maßzahlen reichen oft zur Charakterisierung einer eindimensionalen Verteilung nicht aus.
Aus diesem Grund ist es wichtig den Abstand der Beobachtungswerte vom Mittelwert zusätzlich zu betrachten (Streuungsverhalten).

- ✓ Spannweite (Range):
  Ein einfaches Streuungsmaß, da es den lediglich aus der Differenz zwischen der größten und kleinsten Merkmalsausprägung besteht.
  Die Spannweite kann auch bei ordinalen Daten angewandt werden.

$$R = x_{max} - x_{min}$$

Nachteil:
Die Spannweite hängt nur von 2 Werten ab und ist aufgrund von „Ausreißern“ wenig aussagekräftig.

Beispiel:
Bei dem Kaufverhalten von Zeitungen ergibt sich eine Spannweite von 5 Zeitungen (5-0).

Statistische Streuungsmaße II

- ✓ <u>Varianz:</u>
  Um dem Problem der positiven und negativen Zahlenwerte aus den Differenzen der Ausprägungen zum arithmetischen Mittel zu begegnen, werden daraus die Abweichungsquadrate gebildet und aus diesen Werten der Durchschnitt berechnet.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

Im Fall der absoluten Häufigkeiten errechnet sich die Varianz aus:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 \cdot h(x_i) = \sum_{i=1}^{k}(x_i - \mu)^2 \cdot f(x_i)$$

bzw.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} {x_i}^2 \cdot h(x_i) - \mu^2 = \sum_{i=1}^{k} {x_i}^2 \cdot f(x_i) - \mu^2$$

Statistische Streuungsmaße III

- ✓ Standardabweichung:
  Die positive Quadratwurzel aus der Varianz bezeichnet man als Standardabweichung .
  Die Dimension der Standardabweichung entspricht der Einheit der Beobachtungswerte.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- ✓ Variationskoeffizient (VC):
  Damit Streuungsmaße verschiedener Untersuchungen verglichen werden können, bezieht man die Standardabweichung auf das arithmetische Mittel und erhält das relative Streuungsmaß VC.

$$VC = \frac{\sigma}{\mu} \quad \text{bzw.} \quad VC = \frac{\sigma}{\mu} \cdot 100\ \%$$

Statistische Streuungsmaße IV

✓ <u>Mittlere absolute Abweichung (MAD):</u>
Bezieht man die Berechnung des Streuungsmaß nicht auf das arithmetischen Mittels μ, sondern hat $\overline{X_Z}$ (Median) als Bezugswert gewählt, so spricht man von der mittleren <u>absoluten</u> Abweichung.

$$d = \frac{1}{N}\sum_{i=1}^{N}|x_i - \overline{X_Z}|$$

bzw.

$$d = \frac{1}{N}\sum_{i=1}^{k}|x_i - \overline{X_Z}| \cdot h(x_i) = \sum_{i=1}^{k}|x_i - \overline{X_Z}| \cdot f(x_i)$$

Beispiel I:

Zeitungsverkauf bei einem Durchschnitt von 2,5 und einem Median von 2 Zeitungen

Verkaufte Zeitungen	x_i	0	1	2	3	4	5	**Summe**
absolute Häufigkeit	$h(x_i)$	3	6	4	10	2	5	**30**
Varianz I	$x_i^2 \cdot h(x_i)$	**0**	**6**	**16**	**90**	**32**	**125**	**269**
relative Häufigkeit	$f(x_i)$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{2}{15}$	$\frac{1}{3}$	$\frac{1}{15}$	$\frac{1}{6}$	$\frac{30}{30}$
Varianz II	$x_i^2 \cdot f(x_i)$	$\frac{0}{10}$	$\frac{1}{5}$	$\frac{8}{15}$	$\frac{9}{3}$	$\frac{16}{15}$	$\frac{25}{6}$	$\frac{269}{30}$
Abweichung (Betrag)	$\lvert x_i - \overline{X_Z} \rvert$	2	1	0	1	2	3	
MAD (absolut)	$\lvert x_i - \overline{X_Z} \rvert \cdot h(x_i)$	6	6	0	10	4	15	**41**
MAD (relativ)	$\lvert x_i - \overline{X_Z} \rvert \cdot f(x_i)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{0}{15}$	$\frac{1}{3}$	$\frac{2}{15}$	$\frac{1}{2}$	$\frac{41}{30}$

Beispiel II:

Varianz I: $\frac{1}{30}(0+6+16+90+32+125)-2{,}5^2$

$\frac{269}{30}-6{,}25\approx 2{,}72$

Varianz II: $\left(\frac{0}{10}+\frac{1}{5}+\frac{8}{15}+\frac{9}{3}+\frac{16}{15}+\frac{25}{6}\right)-2{,}5^2$

$\frac{269}{30}-6{,}25\approx 2{,}72$

Standardabweichung: $\sigma=\sqrt{2{,}72}\approx 1{,}5$ Stück

Variationskoeffizient: $VC=\frac{1{,}5}{2{,}5}=0{,}6=60\%$

MAD (absolut): $\frac{1}{30}(6+6+0+10+4+15)=\frac{41}{30}=1{,}37$

MAD (relativ): $\left(\frac{1}{5}+\frac{1}{5}+\frac{0}{15}+\frac{1}{3}+\frac{2}{15}+\frac{1}{2}\right)$

$\left(\frac{6}{30}+\frac{6}{30}+\frac{0}{30}+\frac{10}{30}+\frac{4}{30}+\frac{15}{30}\right)=\frac{41}{30}=1{,}37$

Aufgaben:

1. Für ein Waschpulver eines bestimmten Herstellers wurden in 10 Geschäften Fuldas folgende Preise für ein 1-kg-Paket ermittelt:
   1,40 – 1,60 – 1,70 – 1,50 – 1,40 – 1,80 – 1,70 – 1,60 – 1,50 – 1,80

   a) Berechnen den Durchschnittspreis und Zentralwert der Preise.
   b) Bestimmen Sie die Varianz und Standardabweichung.
   c) Geben Sie die mittlere absolute Abweichung an.

2. Berechnen Sie zur Aufgabe 2 (Seite 17) die Spannweite, Varianz, Standardabweichung, mittlere absolute Abweichung und den Variationskoeffizienten.

3. Die folgende Tabelle gibt die Körpergröße in cm von 5 Kindern an:

Name	Hugo	Fred	Peter	Paul	Maria
Größe	120	130	125	130	135

   a) Berechnen das arithmetische Mittel und den Median.
   b) Bestimmen Sie die Varianz und Standardabweichung.
   c) Geben Sie die mittlere absolute Abweichung an.
   d) Wie groß ist der Variationskoeffizient?