Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Deskriptive Statistik Kapitel 3

Prof Dr. Andrea Wirth

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Gliederung

1. **Einführung und Aufgaben**
   Data Science und Statistik**,** deskriptive Statistik, Wahrscheinlichkeitsrechnung, induktive Statistik, Vorgehensweisen und Beispiele

2. **Grundbegriffe, statistische Untersuchung und Darstellung der Daten**
   Datenquellen, Merkmale, Grundgesamtheit, Stichprobe, Messskalen, Vorgehensweise bei statistischen Untersuchungen Planung, Datenerhebung und -techniken, Datenaufbereitung, tabellarische und grafische Darstellung, Interpretation

3. **Eindimensionale Häufigkeitsverteilungen**

   3.1 Terminologie und grafische Darstellungen
   3.2 Lageparameter
   3.3 Streuungsparameter
   3.4 Schiefe und Konzentration

3.1 Eindimensionale Häufigkeitsverteilung

Werden die Merkmalsträger hinsichtlich eines einzigen Merkmals (Dimension) untersucht, ist das Ergebnis der Erhebung und Aufbereitung eine ***eindimensionale Häufigkeitsverteilung***. Sie beschreibt, wie sich die Merkmalsträger auf die Merkmalswerte des einen Merkmals verteilen (häufen).

Bei einer Erhebung wird an n Merkmalsträgern ein Merkmal X beobachtet, d.h. an jeder Einheit wird die Ausprägung dieses Merkmals festgestellt. Sind a_1, a_2, a_3,...a_n die möglichen Ausprägungen des Merkmals X, so wird der i-ten Untersuchungseinheit (i=1,....,n) seine Ausprägung a_i als Merkmalswert zugeordnet. Die insgesamt n (ggf. auch gleichem) Merkmalswerte a_1,...,a_n heißen auch ***Urliste***. D.h. werden die Beobachtungswerte so notiert, wie sie sich bei einer statistischen Erhebung nacheinander ergeben, so spricht man von einer Urliste.

Beispiel: Ein Zeitungsverkäufer notiert 200 Tage lang täglich die Anzahl der verkauften Exemplare einer bestimmten Zeitung: die Ergebnisse sind in der folg. Urliste ausschnittsweise dargestellt:

Laufende Nummer des Beobachtungstags	Anzahl der verkauften Zeitungen a_i
1	3
2	1
3	0
4	2
199	2
200	5

- Die Urliste ist die Masse der statistischen Daten in ursprünglicher willkürlicher Reihenfolge
- Vorteil: vollständige und richtige Daten
- Nachteil: unübersichtlich

Quelle: in Anlehnung an Bleymüller/Gehlert/Gülicher: 1998

Erinnerung Kapitel 2: Datenaufbereitung und -darstellung

3. ***Datenaufbereitung und -darstellung*:** Verdichtung und Ordnung hin zu Tabellen und Schaubildern.
   - Zu Beginn der Datenaufbereitung oder auch schon während der Erhebungsphase müssen die Daten ***geprüft bzw. kontrolliert*** werden. Die Kontrolle erstreckt sich auf die Vollständigkeit der Erfassung und der Beantwortung sowie auf die Glaubwürdigkeit bzw. Plausibilität der erfassten Daten.
   - Nach der Erhebung liegen die Daten bzw. Merkmalswerte (Urwerte, Urdaten) zunächst in Form einer sogenannten ***Urliste (statistische Reihe)*** vor. In der Urliste sind die Merkmalswerte und eventuell auch die zugehörigen Merkmalsträger nacheinander aufgereiht.
   - In der ***Strichliste*** werden alle in der Urliste enthaltenen Merkmalswerte aufgelistet. Die Anordnung der Merkmalswerte ist vom Skalenniveau abhängig.
   - Zur Erstellung der Häufigkeitstabelle werden in der Strichliste die Striche ausgezählt und dem jeweiligen Merkmalswert als Häufigkeit zugeordnet. Die ***Häufigkeitstabelle*** gibt also die Häufigkeitsverteilung eines Merkmals wieder, d.h. man kann aus ihr ersehen, wie sich die Merkmalsträger auf die verschiedenen Merkmalswerte verteilen.

Quelle: Bourier: Beschreibende Statistik

3.1 Eindimensionale Häufigkeitsverteilung

Die ***einfache Häufigkeit*** gibt an, wie häufig ein Merkmalswert x_i aufgetreten ist. Die einfache Häufigkeit kann ***absolut*** oder ***relativ*** ausgedrückt werden.

Eine erste Aufbereitung der durch die Urliste gegebenen Daten besteht in der Auszählung, mittels einer Strichliste. Hat das Merkmal k Merkmalsausprägungen $x_1,...,x_k$ so ist h_i (i=1,...,k) die Anzahl der Elemente welche die Merkmalsausprägung x_i besitzen. Man bezeichnet h_i als die absolute Häufigkeit der Ausprägung x_i. Dividiert man die absoluten Häufigkeiten h_i durch die Gesamtzahl der Elemente n, so erhält man relative Häufigkeit f_i

h_i = ***absolute einfache Häufigkeit*** (i.d.R. kurz: absolute Häufigkeit)
d.h. Anzahl der Merkmalsträger mit dem Merkmalswert x_i (i = 1, .., k)

f_i = ***relative einfache Häufigkeit*** (i.d.R. kurz: relative Häufigkeit)
d.h. Anteil der Merkmalsträger mit dem Merkmalswert x_i (i = 1, .., k)

n = Gesamtzahl der Merkmalsträger

k = Anzahl verschiedener Merkmalswerte

$$f_i = \frac{h_i}{n} \quad (i = 1,...,k)$$

$$0 \le h_i \le n \quad (i=1,...,k) \text{ und } h_1 + h_2 + ... + h_k = \sum_{i=1}^{k} h_i = n$$

$$0 \le f_i \le 1 \quad (i=1,...,k) \text{ und } \sum_{i=1}^{k} f_i = 1$$

Quelle: Bourier: Beschreibende Statistik 2018

3.1 Eindimensionale Häufigkeitsverteilung

Bei quantitativen Merkmalen ist es sinnvoll, die Häufigkeitstabelle mit ihren ***absoluten bzw. relativen Summenhäufigkeiten*** zu ergänzen. Diese sind wie folgt definiert:

$$H_i = h_1 + h_2 + \ldots + h_i = \sum_{j=1}^{i} h_j \qquad (i=1,\ldots,k)$$ ***absolute Summenhäufigkeit***

$$F_i = f_1 + f_2 + \ldots + f_i = \sum_{j=1}^{i} f_j = \frac{H_i}{n} \qquad (i=1,\ldots,k)$$ ***relative Summenhäufigkeit***

Beispiel: Im o.g. Beispiel sind die Beobachtungstage die statistischen Einheiten, das untersuchte Merkmal ist die Anzahl der an einem Tag verkauften Exemplare einer bestimmten Zeitungen. Die Ermittlung der Häufigkeiten erfolgt über eine Strichliste aus der sich wiederum die Häufigkeits-verteilung ergibt.

Anzahl verkaufter Zeitungen x_i	Anzahl Tage h_i mit verkauften Zeitungen	Anteil Tage f_i	Prozentanteil Tage f_i	Anzahl der Tage H_i an denen höchstens x_i Zeitungen verkauft wurden	Anteil der Tage F_i an denen höchstens x_i Zeitungen verkauft wurden
0	21				
1	46				
2	54				
3	40				
4	24				
5	10				
6	5				
	200				

Ermitteln Sie bitte die fehlenden Werte und illustrieren die relative Summenhäufigkeit.

Quelle: in Anlehnung an Bleymüller/Gehlert/Gülicher: 1998

3.1 Eindimensionale Häufigkeitsverteilung

Beispiel: Im o.g. Beispiel sind die Beobachtungstage die statistischen Einheiten, das untersuchte Merkmal ist die Anzahl der an einem Tag verkauften Exemplare einer bestimmten Zeitungen. Die Ermittlung der Häufigkeiten erfolgt über eine Strichliste aus der sich wiederum die Häufigkeits-verteilung ergibt.

Anzahl verkaufter Zeitungen x_i	Anzahl Tage h_i mit verkauften Zeitungen	Anteil Tage f_i	Prozentanteil Tage f_i	Anzahl der Tage H_i an denen höchstens x_i Zeitungen verkauft wurden	Anteil der Tage F_i an denen höchstens x_i Zeitungen verkauft wurden
0	21				
1	46				
2	54				
3	40				
4	24				
5	10				
6	5				
	200				

Quelle: in Anlehnung an Bleymüller/Gehlert/Gülicher: 1998

3.1 Eindimensionale Häufigkeitsverteilung

Bei quantitativen Merkmalen ist es sinnvoll, die Häufigkeitstabelle mit ihren ***absoluten bzw. relativen Summenhäufigkeiten*** zu ergänzen. Diese sind wie folgt definiert:

$$H_i = h_1 + h_2 + \ldots + h_i = \sum_{j=1}^{i} h_j \qquad (i=1,\ldots,k)$$ ***absolute Summenhäufigkeit***

$$F_i = f_1 + f_2 + \ldots + f_i = \sum_{j=1}^{i} f_j = \frac{H_i}{n} \qquad (i=1,\ldots,k)$$ ***relative Summenhäufigkeit***

Beispiel: Im o.g. Beispiel sind die Beobachtungstage die statistischen Einheiten, das untersuchte Merkmal ist die Anzahl der an einem Tag verkauften Exemplare einer bestimmten Zeitungen. Die Ermittlung der Häufigkeiten erfolgt über eine Strichliste aus der sich wiederum die Häufigkeits-verteilung ergibt.

Anzahl verkaufter Zeitungen x_i	Anzahl Tage h_i mit verkauften Zeitungen	Anteil Tage f_i	Prozentanteil Tage f_i	Anzahl der Tage H_i an denen höchstens x_i Zeitungen verkauft wurden	Anteil der Tage F_i an denen höchstens x_i Zeitungen verkauft wurden
0	21	0,105	10,5	21	0,105
1	46	0,23	23	67	0,335
2	54	0,27	27	121	0,605
3	40	0,2	20	161	0,805
4	24	0,12	12	185	0,925
5	10	0,05	5	195	0,975
6	5	0,025	2,5	200	1,000
	200	1	100		

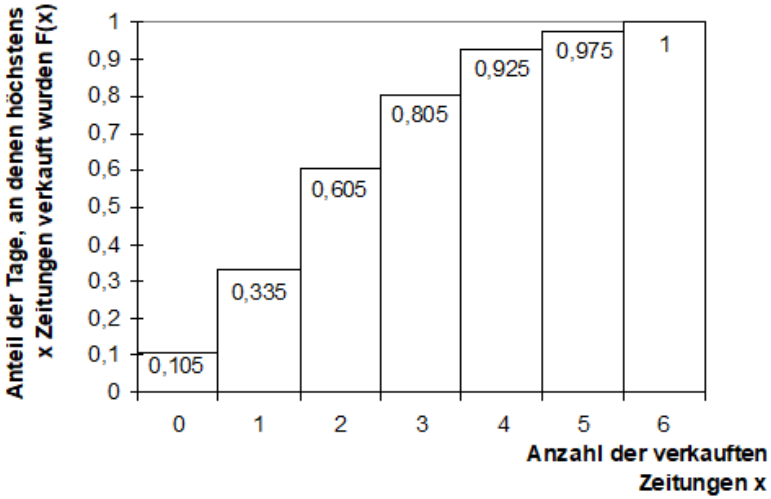
Quelle: in Anlehnung an Bleymüller/Gehlert/Gülicher: 1998

3.1 Eindimensionale Häufigkeitsverteilung

Mittels der relativen Summenhäufigkeiten kann die ***Summenhäufigkeitsfunktion*** *(****empirische Verteilungsfunktion****)* F(x) definiert werden. F(x) gibt den Anteil der statistischen Elemente mit einem Merkmalswert kleiner oder gleich x an. Es handelt sich dabei um eine Treppenfunktion mit Sprungstellen:

$$F(x) = \begin{cases} 0 & \text{für } x < x_i \\ F_i & \text{für } x_i \le x < x_{i+1} \quad (i=1,\ldots,k-1) \\ 1 & \text{für } x \ge x_k \end{cases}$$

Beispiel: Im o.g. Beispiel ergibt sich die folg. Summenhäufigkeitsfunktion



Quelle: in Anlehnung an Bleymüller/Gehlert/Gülicher: 1998

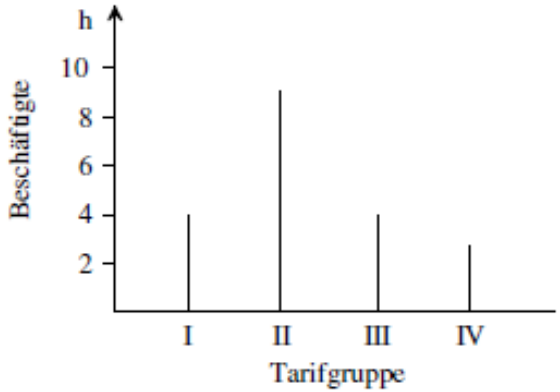
3.1 Grafische Darstellungen

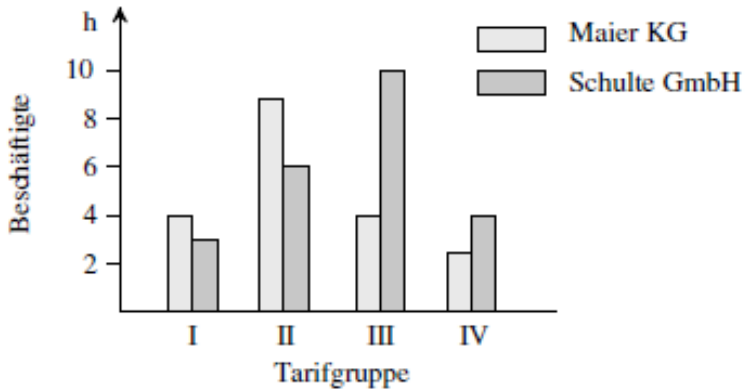
Das ***Stabdiagramm*** ist geeignet für die Darstellung von Häufigkeitsverteilungen qualitativer Merkmale und diskreter, nicht-klassifizierter Merkmale.

Auf der Abszisse eines rechtwinkeligen Koordinatensystems werden die Merkmalswerte abgetragen. Ihre Anordnung hat entsprechend der natürlichen Rangordnung zu erfolgen, bei nominalskalierten Merkmalen ist die Anordnung beliebig. Auf der Ordinate werden die einfachen absoluten und/oder relativen Häufigkeiten h_i bzw. f_i abgetragen. Über den Merkmalswerten werden Stäbe (Linien) senkrecht errichtet, deren Höhe der jeweiligen Häufigkeit entspricht ist. Das Stabdiagramm ist daher als ***höhenproportional.***

Beispiel: Tarifliche Eingruppierung der 20 Beschäftigten der Maier KG und Erweiterung um die Schulte GmbH

Tarifgruppe	h_i
I	4
II	9
III	4
IV	3





Hinweis: Eine Unterbrechung der Häufigkeitsskala sollte vermieden werden, da dies einen Verstoß gegen die Höhenproportionalität darstellt. Wird ein Teil der Skala ausgelassen werden, so ist dies deutlich zu vermitteln. Dies geschieht i.d.R. dadurch, dass die Unterbrechung durch eine gezackte Linie wiedergegeben wird.

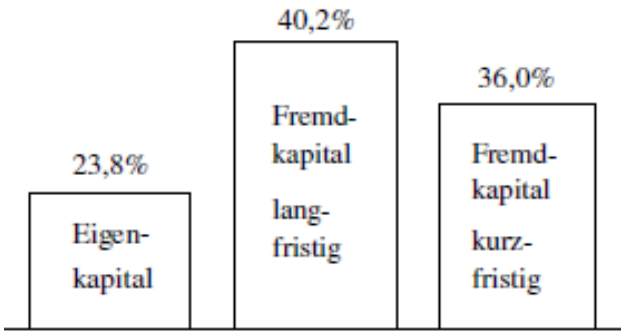
Quelle: Bourier: Beschreibende Statistik 2018

3.1 Grafische Darstellungen

Das ***Rechteck- oder Fächendiagramm*** ist geeignet für die Darstellung von Häufigkeitsverteilungen qualitativer Merkmale und diskreter, nicht-klassifizierter Merkmale.

Jedem Merkmalswert wird ein Rechteck zugeordnet. Die Rechtecke werden in gleichem Abstand nebeneinander auf einer Linie angeordnet. Grundlinie und Seitenhöhe sind so festzulegen, dass die Fläche des Rechteckes proportional zur Häufigkeit ist. Das Rechteckdiagramm ist also eine ***flächenproportionale*** Darstellung. Konstruktion und Interpretation fallen leichter, wenn die Grundlinie für alle Rechtecke identisch ist, da das Diagramm dann ***zugleich höhenproportional*** ist. Die Seitenhöhe entspricht in diesem Fall direkt der Häufigkeit. In oder unter den Rechtecken können die Merkmalswerte, über den Rechtecken zusätzlich deren Häufigkeiten angegeben werden. Sind die Grundlinien für alle Rechtecke gleich lang, dann können die Rechtecke auch zu einem Turm aufgestapelt werden.

Beispiel: Die Kapitalstruktur der Maier KG



Ist diese Darstellung sinnvoll?

Quelle: Bourier: Beschreibende Statistik 2018

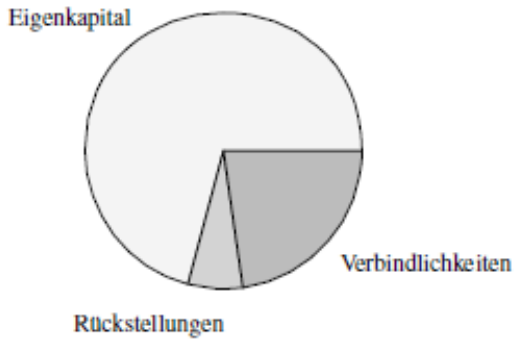
3.1 Grafische Darstellungen

Das ***Kreisdiagramm*** ist geeignet für die Darstellung von Häufigkeitsverteilungen qualitativer Merkmale und diskreter, nicht-klassifizierter Merkmale. Es ist insbesondere zum Aufzeigen der inneren Struktur einer Gesamtheit geeignet.

Der Kreis ist derart in Kreissektoren zu untergliedern, dass die Flächen der Kreissektoren den Häufigkeiten proportional sind. Das Kreisdiagramm ist also eine ***flächenproportionale*** Darstellung. Die Flächenproportionalität wird hergestellt, indem der Kreiswinkel von 360° den Häufigkeiten entsprechend auf die Merkmalswerte aufgeteilt wird. Entfallen auf einen Merkmalswert 20% der Gesamtheit, dann entfallen auf ihn auch 20% des Kreiswinkels, also 72°. Der Winkel des Kreissektors ist damit festgelegt.

Beispiel: Passivseite der Beständebilanz der Medicus-Klinik AG

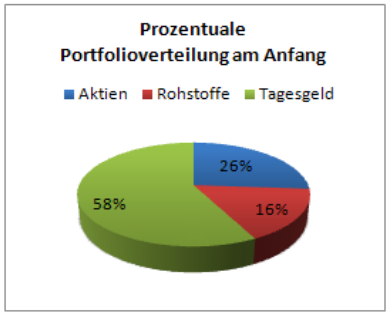
Passiva	Mio € (h_i)	α_i (in °)
Eigenkapital	43,3	255
Rückstellungen	3,9	23
Verbindlichkeiten	13,9	82
Gesamtkapital	61,1	360



Hinweis: Durch ein Nebeneinanderreihen mehrerer Kreisdiagramme können die inneren Strukturen der Gesamtheiten anschaulich verglichen werden. Unterschiede in den Gesamthäufigkeiten n können dabei durch eine entsprechend unterschiedlich große Gestaltung der Kreise (Flächen) wiedergegeben werden..

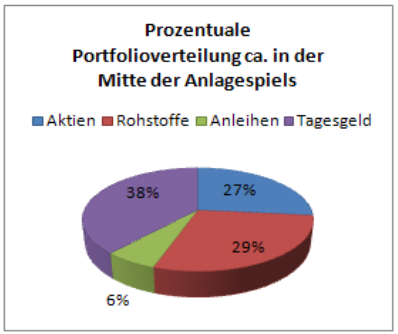
Quelle: Bourier: Beschreibende Statistik 2018

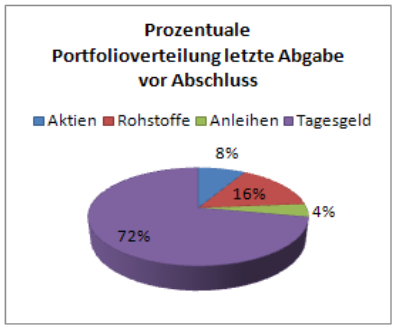
3.1 Gut oder schlecht?



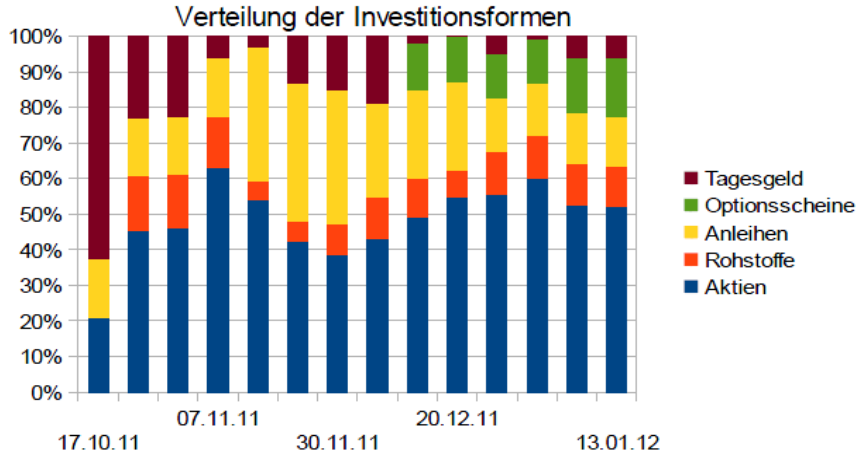
Zu Beginn des Planspiels hatte ich nur 2 Anlagearten. Aktien und Rohstoffe. Der größte Teil des Kapitals lag noch auf dem Tagesgeldkonto.

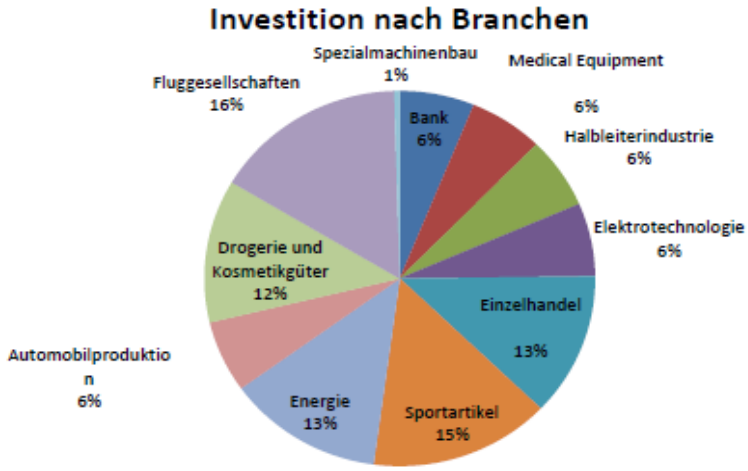
Ca. zur Mitte des Planspiels verteilten sich die Anlagen fast zu gleichen Teilen in Aktien und Rohstoffe. Dies ist darin begründet, dass ich durch die Eurokrise nur noch sehr zögerlich in neue Aktien investiert habe und mir eine Investition in Rohstoffe als sicherer erschien. Dazu kam die sichere Investition in die BRD Anleihe.





Vor dem Abschluss des Planspiels, hatte ich bereits die meisten Anlagen wieder verkauft und nur noch ein geringer Restbestand befand sich in meinem Portfolio. Ich hab keine neuen Investitionen getätigt, da das Planspiel vor dem Abschluss stand.

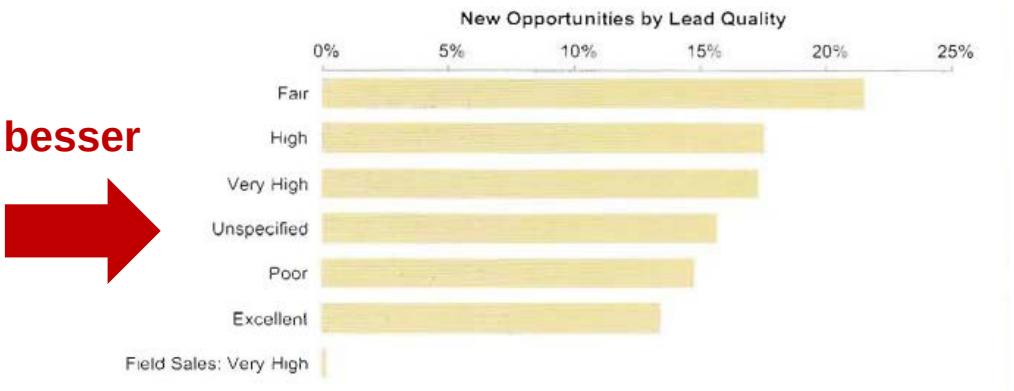


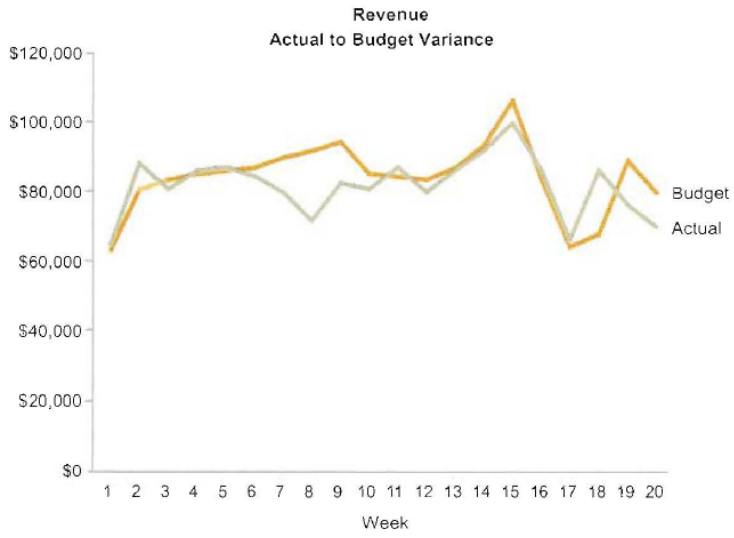


Quelle: Abgaben Finanzmanagement

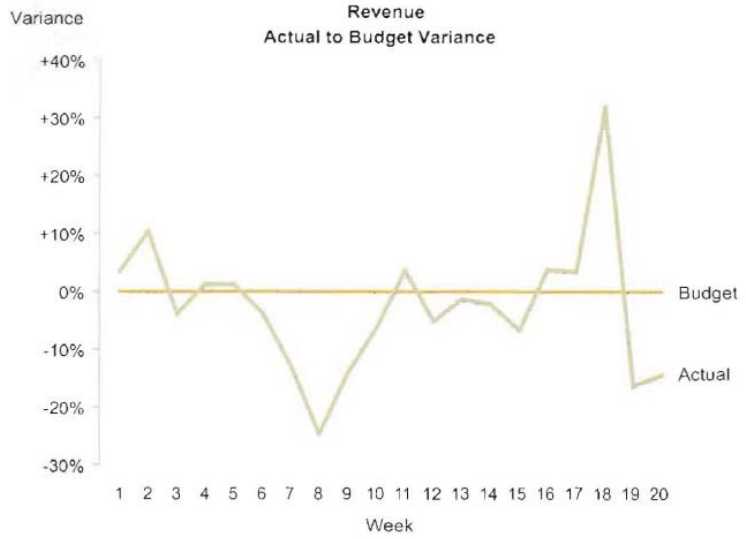
3.1 Gut oder schlecht?

besser





besser



Quelle: Vorlesung Unternehmenssteuerung

3.1 Klassifizierte Häufigkeitsverteilung

Bei diskreten Merkmalen mit vielen Merkmalsausprägungen oder bei stetigen Merkmalen wird empfohlen die Merkmalsausprägungen zu ***Klassen*** zusammenzufassen. Teilt man die gesamte Menge der statistischen Daten in Teilmengen, so bilden die Daten jeder Teilmenge eine Klasse. Diese Einteilung ist nur bei quantitativen Merkmalen sinnvoll. Ein stetiges Merkmal mit begrenzter Messgenauigkeit ergibt so ein diskretes Merkmal mit vielen möglichen Ausprägungen.

Die i-te Klasse wird durch die ***untere*** x_i^u und die ***obere Klassengrenze*** x_i^o begrenzt

für die gilt $x_i^o = x_{i+1}^u$ (i=1,...k-1)

Als ***Klassenbreite*** bezeichnet man Δx, für die gilt: $\Delta x = x_i^o - x_i^u$ (i=1,...,k)

Als ***Klassenmitte*** bezeichnet man x_i^*, für die gilt: $x_i^* = \frac{1}{2}\left(x_i^o + x_i^u\right) = x_i^u + \frac{1}{2}\left(x_i^o - x_i^u\right)$ (i=1,...,k)

Hinweis: Es ist unbedingt notwendig, das Intervall zu betrachten, in dem alle statistischen Daten liegen. Die erste und die letzte Klasse sind häufig offen, man bezeichnet dies ***als offene Randklassen***. Teilintervalle sind nach Möglichkeit gleich lang zu wählen. Die Daten eines Teilintervalls bilden nach Definition eine Klasse, wobei die obere und die untere Grenze eines Teilintervalls gleichzeitig die obere und untere Klassengrenze darstellen. Die Klassenmitte sollte ein repräsentativer Wert der Klasse für weitere Berechnungen und eine möglichst einfache Zahl sein. Die absolute und die relative Häufigkeit sollten zur Kontrolle pro Klasse berechnet werden. Die Tatsache wie viele Klassen gebildet werden, ist vom Problem abhängig und häufig willkürlich. Bei der statistischen Auswertung von Messergebnissen ist es üblich, dass die Klassen disjunkt definiert werden.

Quelle: in Anlehnung an Bleymüller/Gehlert/Gülicher: 1998

3.1 Klassifizierte Häufigkeitsverteilung

Beispiel: Rechnungsbeträge von 140 Kunden

Grundgesamtheit: 140 Kunden
Merkmal X: Rechnungsbetrag (€)
Merkmalswert x_i: 0,25, 1,18, ..., 116,00, 119,80.

j	Rechnungsbetrag (€) von ...	bis unter ...	h_j	H_j	f_j	F_j
1	0	20	10	10	0,07	0,07
2	20	40	20	30	0,14	0,21
3	40	60	60	90	0,43	0,64
4	60	80	35	125	0,25	0,89
5	80	100	10	135	0,07	0,96
6	100	120	5	140	0,04	1,00
			140		1,00	

j = Laufindex für die Klasse (Klassenindex), $j = 1, ..., v$

x_j^u = Untergrenze der Klasse j

x_j^o = Obergrenze der Klasse j

h_j = absolute einfache Klassenhäufigkeit (kurz: absolute Klassenhäufigkeit)
Anzahl der Merkmalsträger mit einem Merkmalswert x_i, der in die j-te Klasse fällt, d.h.

$$x_j^u \leq x_i < x_j^o$$

h_2 = 20, d.h. 20 Kunden haben eine Rechnung über einen Betrag von 20 € bis unter 40 €

H_j = absolute kumulierte Klassenhäufigkeit
Anzahl der Merkmalsträger mit einem Merkmalswert x_i, der kleiner als die Obergrenze der j-ten Klasse ist, d.h.

$$x_i < x_j^o$$

H_2 = 30, d.h. 30 Kunden haben eine Rechnung über einen Betrag von weniger als 40 €

Quelle: Bourier: Beschreibende Statistik 2018

3.1 Klassifizierte Häufigkeitsverteilung

Beispiel: Ergebnisse einer Klausur mit einer maximalen Punktezahl von 75 Punkten

Punkte	Häufigkeit	Punkte	Häufigkeit	Punkte	Häufigkeit	Punkte	Häufigkeit
12	1	27	1	42	5	56	2
14	2	28	1	46	3	57	1
15	1	30	1	47	1	58	2
18	1	34	1	48	2	59	1
20	1	35	1	49	3	61	1
22	1	37	2	52	1	62	1
24	1	39	2	54	1	65	1
25	1	41	1	55	1	Σ	45

Note	Klassengrenzen	Klassenmitte	Absolute Häufigkeit h_i	Relative Häufigkeit f_i	Absolute Summen-häufigkeit H_i	Relative Summen häufigkeit F_i
6	$0 \leq x \leq 19$	9,5	5	0,111	5	0,111
5	$19 < x \leq 30$	24,5	7	0,156	12	0,267
4	$30 < x \leq 42$	36	12	0,267	24	0,533
3	$43 < x \leq 54$	48	11	0,244	35	0,778
2	$54 < x \leq 65$	59,5	10	0,222	45	1,000
1	$65 < x \leq 75$	67,5	0	0,000	45	1,000
			45	1		

- Vorteil: sehr übersichtlich.
- Nachteil: Informationsverlust durch Zusammenfassung der Daten; alle Daten einer Klasse werden jetzt als gleichwertig behandelt.
- Im Beispiel ist aus quantitativem Ergebnis (Punktezahlen) durch Klassenbildung nahezu ein qualitatives Ergebnis (Prädikate) geworden.
- Vorsicht bei der Bildung von Notendurchschnitten.

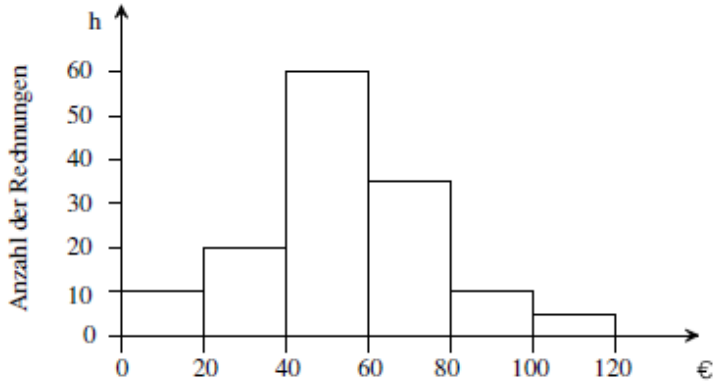
Quelle: in Anlehnung an Bleymüller/Gehlert/Gülicher: 1998

3.1 Grafische Darstellungen

Das ***Histogramm*** ist geeignet zur graphischen Darstellung klassifizierter Häufigkeitsverteilungen.

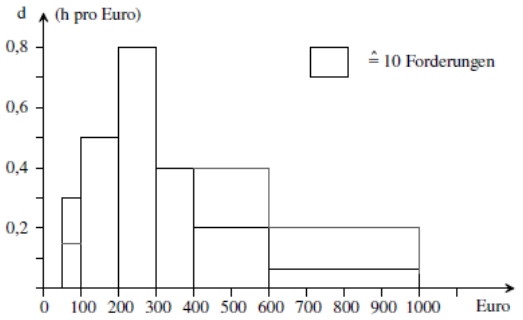
Auf der Abszisse eines rechtwinkeligen Koordinatensystems werden die Merkmalswerte bzw. die Klassen abgetragen. Bei offenen Randklassen ist für die offene Grenze ein plausibel erscheinender Wert anzusetzen. Über den Klassen werden Rechtecke errichtet, wobei die Flächen der Rechtecke den jeweiligen Klassenhäufigkeiten proportional sind. Da die Grundlinie des Rechteckes durch die Klassenbreite festgelegt ist, ist die ***Flächenproportionalität*** über die Höhe des Rechteckes herzustellen. Bei der Bestimmung der Rechteckhöhe ist es sinnvoll, zwischen konstanter und unterschiedlicher Klassenbreite zu differenzieren.

Beispiel: konstante Klassenbreiten:
Rechnungsbeträge



Beispiel: unterschiedliche Klassenbreiten über die Häufigkeitsdichte: Forderungsbestand

Forderung (€)		h_j	d_j
von ...	bis unter ...		
50	100	15	0,30
100	200	50	0,50
200	300	80	0,80
300	400	40	0,40
400	600	40	0,20
600	1.000	20	0,05



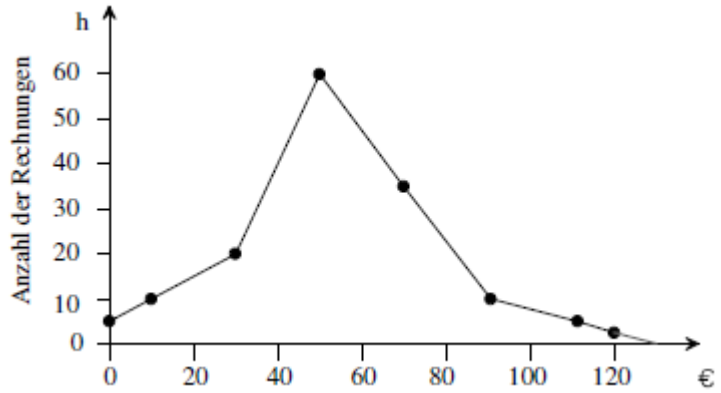
Quelle: Bourier: Beschreibende Statistik 2018

3.1 Grafische Darstellungen

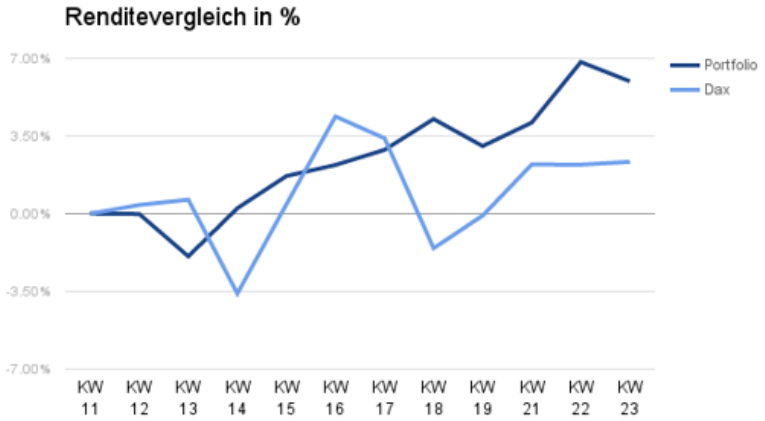
Der ***Polygonzug*** ist geeignet zur graphischen Darstellung klassifizierter Häufigkeitsverteilungen, insbesondere wenn es um den Vergleich mit anderen Häufigkeitsverteilungen geht.

Auf der Abszisse eines rechtwinkeligen Koordinatensystems werden die Merkmalswerte bzw. die Klassen abgetragen und auf der Ordinate die Häufigkeiten bzw. die Häufigkeitsdichten. Auch hier ist wieder zwischen konstanter und unterschiedlicher Klassenbreite zu differenzieren. Der Polygonzug ist sehr gut zur graphischen Darstellung von Vergleichen mit anderen Gesamtheiten geeignet. Dazu ist in das Koordinatensystem der Polygonzug einer zweiten Gesamtheit oder weiterer Gesamtheiten einzutragen.

Beispiel: konstante Klassenbreiten:
Rechnungsbeträge



Beispiel: Renditevergleich

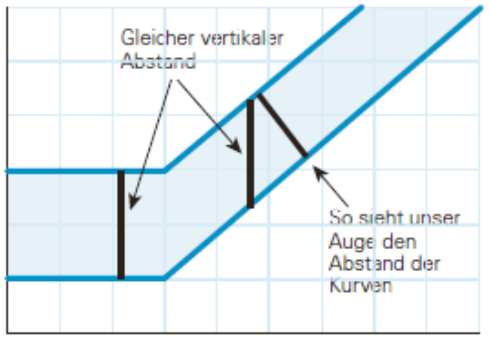


Quelle: Bourier: Beschreibende Statistik 2018

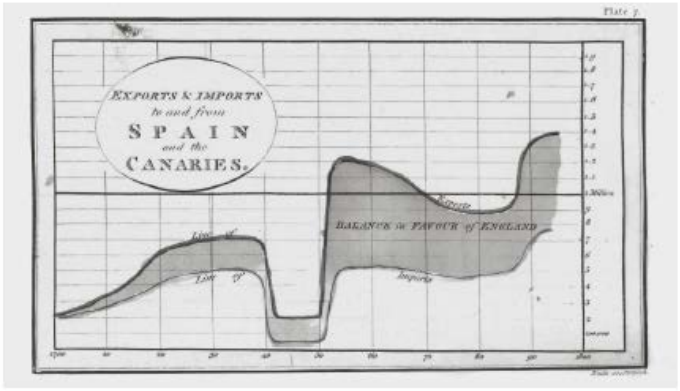
3.1 Grafische Darstellungen

Grundsätzlich bieten sich bei der Erhebung und der tabellarischen Darstellung insbesondere aber bei der grafischen Darstellung zahlreiche Manipulationsmöglichkeiten :

- Auswahl der Darstellungsart
- Beschriftung des Schaubildes
- Formatauswahl
- Maßstabswahl (z.B. nichtlinear sondern logarithmisch)
- Gruppen- und Klassenbildung
- Flächentreue
- Bei Säulendiagrammen z.B. lange Balken absägen, d.h. kein Nullpunkt.
- oder optische Täuschungen



Unser Auge interpretiert Abstände nicht senkrecht, sondern orthogonal



Eine der ersten Datengrafiken der Welt

Leider gibt es recht wenige Möglichkeiten sich vor derartigen Manipulationen zu schützen und man sollte durch genaue Betrachtung und Analyse auch die grafische Darstellung prüfen.

Grafiken: Krämer: So lügt man mit Statistik

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Gliederung

1. **Einführung und Aufgaben**
   Data Science und Statistik**,** deskriptive Statistik, Wahrscheinlichkeitsrechnung, induktive Statistik, Vorgehensweisen und Beispiele

2. **Grundbegriffe, statistische Untersuchung und Darstellung der Daten**
   Datenquellen, Merkmale, Grundgesamtheit, Stichprobe, Messskalen, Vorgehensweise bei statistischen Untersuchungen Planung, Datenerhebung und -techniken, Datenaufbereitung, tabellarische und grafische Darstellung, Interpretation

3. **Eindimensionale Häufigkeitsverteilungen**

   3.1 Terminologie und grafische Darstellungen
   3.2 Lageparameter
   3.3 Streuungsparameter
   3.4 Schiefe und Konzentration

3.2 Nachdenkliches

»Hunde beißen am liebsten Männer, Katzen bevorzugen ältere Frauen und Pferde Mädchen. Das fand Eilif Dahl von der norwegischen Ärztevereinigung heraus.«

Die Nachrichtenagentur Associated Press

Ein Einzelhändler kauft eine Ware für 100 Euro ein und schlägt sie für 200 Euro wieder los. Wie viel Prozent macht seine Handelsspanne aus?

»Eine Unverschämtheit!«, sagen wir als Kunde. »Ein Aufschlag von glatt 100 Prozent!«

»So schlimm ist das nun auch wieder nicht«, sagt der Händler. »50 Prozent Verdienst sind wirklich nicht zu viel.«

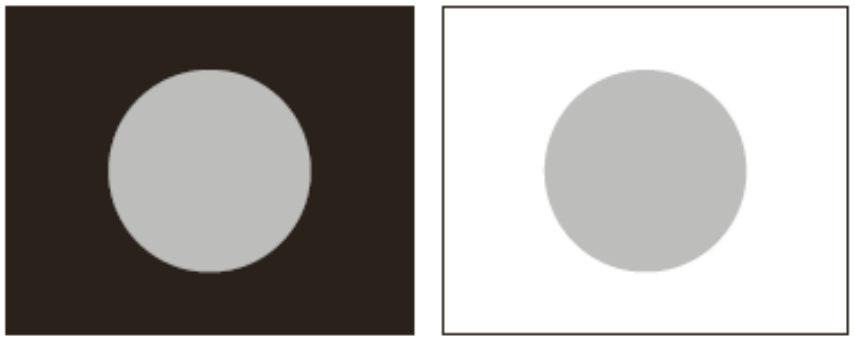
Offenbar haben beide recht. 100 Euro sind 100 Prozent von 100 Euro und 50 Prozent von 200 Euro. Trotzdem ist der Eindruck je nach der Basis durchaus ein anderer.

Quelle: Krämer: So lügt man mit Statistik

Wer ist Teil von was?

Diese Bedeutung der Basis bei Vergleichen zeigt sich am besten, wenn man dabei, ob aus Versehen oder in Betrugsabsicht, total danebengreift. So philosophierte eine amerikanische Zeitung einmal über Gewaltverbrechen in den USA. Die meisten geschehen zu Hause, in Küche, Wohn- und Schlafzimmer. Die Zeitung schloss daraus: Man schläft nachts sicherer im Central Park.

Ein und dieselbe Sache sieht also sehr verschieden aus, je nachdem, womit man sie vergleicht. Betrachten wir etwa im folgenden Diagramm die Farbe Grau. Vor schwarzem Hintergrund erscheint sie eher weiß, vor weißem Hintergrund dagegen eher schwarz:



Die gleiche Farbe Grau erscheint einmal dunkel, einmal hell

3.2 Nachdenkliches

Jeder weiß, was ein Durchschnitt ist. Wenn Bauer A drei und Bauer B fünf Kühe hat, so hat im Durchschnitt jeder vier. Nichts ist einfacher als das.

Leider tut es der guten Dinge dabei oft zu viel, wie folgende Version eines uralten Statistikerwitzes zeigt (Copyright Franz Josef Strauß): »Zwei Männer sitzen im Wirtshaus, der eine verdrückt eine Kalbshaxe, der andere trinkt zwei Maß Bier. Statistisch gesehen ist das für jeden eine Maß Bier und eine halbe Haxe, aber der eine hat sich überfressen und der andere ist besoffen.«

Das arithmetische Mittel verkleistert oft eine große Ungleichheit – es schweigt sich zur Streuung um den Mittelwert völlig aus. Wenn es in unserem Dorf zehn Bauern gibt, von denen einer 40 Kühe hat und alle anderen haben nichts, so hat im Mittel jeder vier. Für die neun Habenichtse ist das aber nur ein schwacher Trost. Offenbar macht es einen Unterschied, ob sich die Werte dicht um das Mittel sammeln oder ob sie in alle Winde streuen, aber diesen Unterschied sieht man dem Mittelwert nicht an.

Ein Mittelwert ohne Streuung ist also nur die Hälfte wert. Zu einem seriösen Durchschnitt gehört in aller Regel auch ein Maß für die Abweichung davon. Wie dieses aussieht, sei uns hier egal – wichtig ist: Bei nackten Mittelwerten ist immer Vorsicht angezeigt.

Quelle: Krämer: So lügt man mit Statistik

3.2 Statistische Maßzahlen

Für viele statistische Fragestellungen möchte man, eine Häufigkeitsverteilung durch einige wenige informative Größen oder ***Maßzahlen*** ersetzen. Die Charakterisierung der Häufigkeitsverteilung durch statistische Maßzahlen ist wie die Klasseneinteilung eine ***Datenreduktion***, um das Datenmaterial überschaubarer und aussagekräftiger zu machen. Man unterscheidet Lageparameter, Streuungsparameter, Momente und Konzentrationsmaße.

Eine radikalere Komprimierung der Ausgangsdaten als beim Übergang auf einen Häufigkeitsverteilung ist dann gegeben, wenn die ***gesamte Urliste durch eine einzige Zahl***, einen sogenannten ***Lageparameter*** charakterisiert wird. Dieser soll „möglichst gut“ beschreiben, wo das gesamte Datenmaterial auf der Merkmalsachse lokalisiert ist

Lageparameter sollen folgenden Forderungen genügen:

- Der Lageparameter repräsentiert die Verteilung möglichst gut
- Der Lageparameter ermöglicht den Vergleich verschiedener Verteilungen des gleichen Merkmals bzgl. verschiedener statistischer Massen
- Der Lageparameter ermöglicht die Beurteilung von Einzelwerten/Gruppen von Einzelwerten

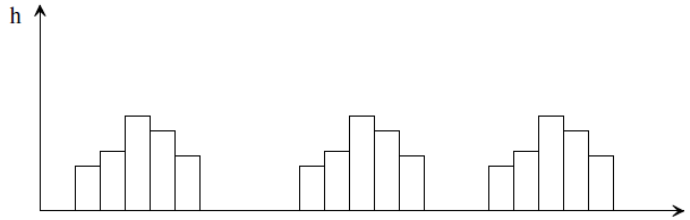


Abb. 3.1: Häufigkeitsverteilungen mit unterschiedlicher Lage

Was fallen Ihnen für Lageparameter ein?

Quelle: Bamberg/Bauer: Statistik 2017

3.2 Der Modus

Der ***Modus/Modalwert*** ist derjenige Merkmalswert, der am häufigsten beobachtet wurde.

Da für die Bestimmung des Modus allein die Häufigkeiten der Merkmalswerte maßgebend sind, werden an die Skalierung der Merkmale keine Voraussetzungen gestellt. D.h. der Modus ist prinzipiell für jede Verteilung bestimmbar.

Beispiel: Die größte Religionsgemeinschaft in der Bundesrepublik Deutschland ist die evangelische Kirche.

Beispiel: Für die Beschäftigten der Maier KG und der Schulte GmbH wurden jeweils die in der vergangenen Woche geleisteten Überstunden erfasst.

Maier KG

Überstunde x_i	h_i
0	3
1	5
2	4
3	4
4	4

Schulte GmbH

Überstunde x_i	h_i
0	3
1	10
2	4
3	3
4	2
12	1

Bestimmen Sie den Modus und diskutieren, ob dies sinnvoll ist

Quelle: Bourier: Beschreibende Statistik 2018

3.2 Der Median

Der ***Median/Zentralwert*** $\bar{x}_Z$ ist durch die Eigenschaft definiert, dass mindestens 50% aller Merkmalswerte kleiner oder gleich als $\bar{x}_Z$ sind und mindestens 50% aller Merkmalswerte größer oder gleich $\bar{x}_Z$. Es ist derjenige Merkmalswert, dessen Merkmalsträger in der Rangordnung aller Merkmalsträger genau die mittlere Position einnimmt.

Zur Bestimmung des Medians müssen die Merkmalswerte bzw. die Merkmalsträger in eine Rangordnung gebracht werden. Der Median kann daher nur dann bestimmt werden, wenn das Merkmal mindestens ordinalskaliert ist. Es empfiehlt sich zwischen gerader und ungerader Anzahl an Merkmalsträgern zu unterscheiden.

$$\bar{x}_Z = \begin{cases} a_{\frac{n+1}{2}} & \text{falls n ungerade} \\ \dfrac{a_{\frac{n}{2}} + a_{\frac{n}{2}+1}}{2} & \text{falls n gerade} \end{cases}$$

Beispiel: Legt ein Angestellter den Weg zwischen Wohnung und Arbeitsstätte an 5 Tagen in 12, 10, 16, 12, und 17 bzw. geordnet in 10, 12, 12, 16, 17 Minuten zurück, so beträgt der Median

Beispiel: Die Ermittlung des Intelligenzquotienten bei 9 Kindern habe – bereits geordnet - die Merkmalsausprägungen 88, 92, 92, 95, 98, 102, 103, 105, 110. Da n ungerade ist, ergibt sich der Median

Beispiel: Die Ermittlung des Intelligenzquotienten bei 9 Kindern habe – bereits geordnet - die Merkmalsausprägungen 86, 92, 95, 96, 98, 100, 102, 105. Da n gerade ist, ergibt sich der Median

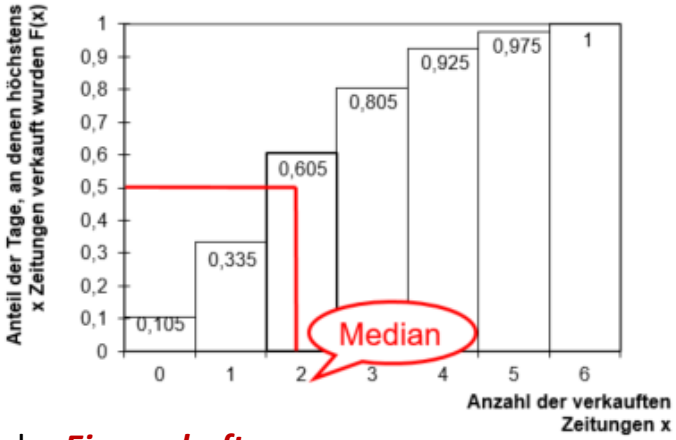
Quelle: Bourier: Beschreibende Statistik 2018, Beispiele aus meinem alten Skript

3.2 Der Median

Eigenschaften und Eignung: Der Median ist unbeeinflusst von Ausreißern, da er allein von der Anzahl der Merkmalsträger abhängig ist. Der Median ist ein geeigneter Mittelwert für schiefe Verteilungen. Der Median kann über die relativen Summenhäufigkeiten auch grafisch bestimmt werden.

»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

Beispiel: Im o.g. Beispiel des Zeitungsverkäufers bedeutet dies einen Wert von $\bar{x}_Z = 2$ verkauften Zeitungen



Der Median besitzt 3 charakteristische ***Eigenschaften***:

- Mindestens 50% aller Merkmalswerte sind kleiner oder gleich dem Median $\bar{x}_Z$ und mindestens 50% aller Merkmalswerte sind größer oder gleich dem Median $\bar{x}_Z$
- Liegen die Merkmalswerte als Zahlen vor, so hat die Summe der absoluten Abweichungen von einer beliebigen Zahl c d.h. der Ausdruck $\sum_{i=1}^{n} |a_i - c|$ ein Mimimum, wenn man die Abweichungen bzgl. des Mediums bildet
- Der Median ist unempfindlich gegenüber Ausreißern, d.h. extreme Merkmalswerte haben keinen wesentlichen Einfluss auf den Median.

Quelle: Bourier: Beschreibende Statistik 2018, Bleymüller/Gehlert/Gülicher[1998]

3.2 Der Median

Aus der ***klassifizierten Häufigkeitsverteilung*** kann der Median nicht mehr exakt abgelesen werden. Er lässt sich nur näherungsweise bestimmen. Für die Feinberechnung wird angenommen, dass in der Medianklasse i eine Gleichverteilung vorliegt und der Median über lineare Interpolation ermittelt wird

$$\bar{x}_z = x_i^u + \frac{0,5 - F_i^u}{F_i^o - F_i^u}(x_i^o - x_i^u) \overset{\text{alternativ}}{=} x_i^u + \frac{\frac{n}{2} - H_{i-1}}{h_i}(x_i^o - x_i^u)$$

Beispiel: Verdeutlichung am Beispiel von klassifizierten Forderungen

Forderung (€)		h_j	H_j
von ...	bis unter ...		
50	100	15	15
100	200	50	65
200	300	80	145
300	400	40	185
400	600	40	225
600	1000	20	245

Quelle: Bourier: Beschreibende Statistik 2018

3.2 Quantile als Erweiterung des Median

Ein ***Quantil*** $\bar{x}_{p/q}$ mit $q \geq 2$, $p = 1,2,...,q-1$, ist ein Merkmalswert, durch den die Gesamtheit in zwei Teile zerlegt wird. So wie der Median die Gesamtheit in zwei Hälften zerlegt, zerlegen die ***Quartile*** die Gesamtheit in vier Viertel, die ***Dezile*** (Dezentile) in zehn Zehntel, die ***Perzentile*** in 100 Hundertstel etc.

Das Quantil $\bar{x}_{p/q}$ hat die Eigenschaft, dass höchstens n*p/q statistische Einheiten echt kleiner und höchstens n*(1-p/q) statistische Einheiten echt größere Merkmalswerte haben. D. das Quartile teilen die aufsteigend geordnete Reihe von Merkmalswerten im Verhältnis p/q zu (1-p/q)

Das 1. Quartil (auch: 25%-Quantil mit p=1 und q=4) z.B. zerlegt die Gesamtheit derart, dass 25% kleiner und 75% größer als das 1. Quartil sind. Bei den Dezilen und Perzentilen interessieren i.d.R. nur die am Rand liegenden Werte wie z.B. das 5. Perzentil (p=5 und q=100), das die Gesamtheit in die Teile 5% : 95% zerlegt. Auf diese Weise werden weitere Informationen über die Lage und die Struktur der Verteilung gewonnen.

Beispiel: Bestimmen Sie das 3. Quartil der o.g. klassifizierten Forderungen

Forderung (€)		h_j	H_j
von ...	bis unter ...		
50	100	15	15
100	200	50	65
200	300	80	145
300	400	40	185
400	600	40	225
600	1000	20	245

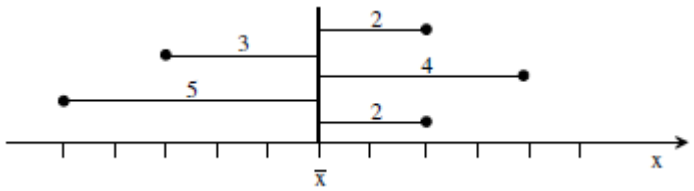
Quelle: Bourier: Beschreibende Statistik 2018

3.2 Das arithmetische Mittel

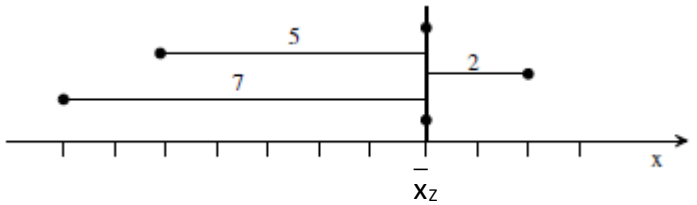
Das ***arithmetische Mittel*** (***Durchschnittswert***, ***gewogenes arithmetisches Mittel***) $\bar{x}$ ist der bekannteste Lageparameter und der Wert, der sich bei gleichmäßiger Verteilung der Summe aller beobachteten Merkmalswerte auf die Merkmalsträger ergibt. Er ist folgendermaßen definiert:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n}\sum_{i=1}^{k} h_i x_i = \sum_{i=1}^{k} f_i x_i$$ wobei k wie üblich die Anzahl der Klassen ist

Die Addition von Merkmalswerten ergibt nur dann einen Sinn, wenn die Abstände zwischen den Werten messbar sind. Die Bestimmung des arithmetischen Mittels ist daher nur zulässig, wenn das Merkmal mindestens intervallskaliert ist.



Entfernung von 5 Merkmalswerten zu ihrem arithmetischen Mittel



Entfernung von 5 Merkmalswerten zum Median

Es gibt keinen anderen Wert als den Median, bei dem die Summe der Entfernungen (falls messbar) aller Merkmalswerte kleiner ist bzw. zu dem die Merkmalswerte - in der Summe gesehen - näher liegen.

Quelle: Bourier: Beschreibende Statistik 2018, mein altes Skript

3.2 Das arithmetische Mittel

Beispiel: Legt ein Angestellter den Weg zwischen Wohnung und Arbeitsstätte an 5 Tagen in 12, 10, 16, 12, und 17 Minuten zurück, dann beträgt die durchschnittliche Zeit, die er für den Weg benötigt

Beispiel: Für die 23 Beschäftigten der Schulte GmbH ist die durchschnittliche Überstundenzahl für die letzte Woche zu berechnen

Überstunde x_i	0	1	2	3	4	12
h_i	3	10	4	3	2	1

Beispiel: Bestimmen Sie auf Basis der absoluten und der relative Häufigkeiten für das o.g. Beispiel des Zeitungsverkäufers das arithmetische Mittel

Anzahl verkaufter Zeitungen x_i	Anzahl Tage h_i mit verkauften Zeitungen	Anteil Tage f_i	Prozentanteil Tage f_i	Absolute Summenhäufigkeit H_i	Relative Summenhäufigkeit F_i
0	21	0,105	10,5	21	0,105
1	46	0,23	23	67	0,335
2	54	0,27	27	121	0,605
3	40	0,2	20	161	0,805
4	24	0,12	12	185	0,925
5	10	0,05	5	195	0,975
6	5	0,025	2,5	200	1,000

Quelle: Bourier: Beschreibende Statistik 2018, altes Skript, Bleymüller/Gehlert/Gülicher[1998]

3.2 Das arithmetische Mittel

Eigenschaften und Eignung: Das arithmetische Mittel ist der Mittelwert, der in der Praxis am häufigsten zum Einsatz kommt. Die Vorstellung, was wäre, wenn alle Merkmalsträger gleich gestellt wären, ist dafür ausschlaggebend. Die Anwendung erfolgt mitunter zu unkritisch. So kann die Abhängigkeit des arithmetischen Mittels von sämtlichen Merkmalswerten nachteilig sein, wenn die Verteilung Ausreißer besitzt oder eine schiefe Verteilung vorliegt.

Das arithmetische Mittel besitzt folgende charakteristische ***Eigenschaften***:

- Die Summe der Abweichungen der Merkmalswerte vom arithmetischen Mittel ist Null: $\sum_{i=1}^{n}(a_i - \bar{x}) = 0$
- Werden die Merkmalswerte einer linearen Transformation unterworfen, $a_i^* = m * a_i + t$ (t, m beliebig $\neq 0$) so gilt, dass das arithmetische Mittel $\bar{x}^*$ der gleichen Transformation unterliegt.
- Die Summe der absoluten Abweichungen von einer beliebigen Zahl c d.h. der Ausdruck $\sum_{i=1}^{n}(a_i - c)^2$ hat ein Minimum, wenn man die Abweichungen bzgl. des arithmetischen Mittels bildet.
  Dies heißt ***Minimaleigenschaft***.
- Das arithmetische Mittel einer Grundgesamtheit, die sich aus zwei Teilgesamtheiten n_1 und n_2 zusammensetzt kann folgendermaßen bestimmt werden:
$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$
- Der arithmetische Mittel ist empfindlich gegenüber Ausreißern, d.h. extreme Merkmalswerte können die Aussagekraft erheblich einschränken.

Quelle: Bourier: Beschreibende Statistik 2018, Bleymüller/Gehlert/Gülicher[1998]

3.2 Das arithmetische Mittel

Für ***klassifizierte Häufigkeitsverteilungen*** kann das arithmetische Mittel nur näherungsweise berechnet werden. Es werden in der Formel dazu die Merkmalswerte x_i gegen die Klassenmitten x_i^* ausgetauscht. Die Klassenmitte wird als Repräsentant für die Merkmalswerte in der Klasse angesehen. D.h. es wird für jede Klasse eine Gleichverteilung oder eine um die Klassenmitte symmetrische Verteilung unterstellt.

Beispiel: erneut am Beispiel der Forderungen

Forderung (€)		h_j	x_j'	$x_j' \cdot h_j$
von ...	bis unter ...			
50	100	15	75	1.125
100	200	50	150	7.500
200	300	80	250	20.000
300	400	40	350	14.000
400	600	40	500	20.000
600	1000	20	800	16.000
		245		78.625

Quelle: Bourier: Beschreibende Statistik 2018

3.2 Beispiele zum Nachdenken

Beispiel: In einer Gruppe von 10 Personen beziehen 9 Personen ein Jahreseinkommen von 20.000 € und eine Person ein Jahreseinkommen 200.000 €. Eignet sich das arithmetische Mittel?

Beispiel: Eine Autovermietung berechnet für ihre Wagen eine feste Tagesgebühr von 40 € und einen Kilometersatz von 0,4 €/km; ferner sei bekannt, dass die Wagen im Durchschnitt täglich 350 km zurücklegen. Damit ergeben sich die durchschnittliche täglichen Einnahmen zu

$$\bar{x} = 40 + 0{,}4 * 350 = 180 €.$$

 Quelle: Bourier: Beschreibende Statistik 2018, altes Skript, Bleymüller/Gehlert/Gülicher[1998]

3.2 Beispiele zum Nachdenken

Beispiel: Ein Student fährt täglich 60km zur Hochschule. Morgens hat er nur eine Durchschnittsgeschwindigkeit von 60km/h erreicht. Wie ist die Durchschnittsgeschwindigkeit für Hin- und Rückfahrt, wenn er abends einen Schnitt von 100km/h fährt?

Beispiel: Ein Student tankt immer für 50 € Der Preis pro Liter war 0,88; 0,98 und 1,08. Welchen Durchschnittspreis hat er bezahlt?

Quelle: Bourier: Beschreibende Statistik 2018, altes Skript, Bleymüller/Gehlert/Gülicher[1998]

3.2 Das harmonische Mittel

Das ***harmonische Mittel*** ist derjenige Wert, zu dem die in der Häufigkeitsverteilung vor ihm liegenden Merkmalswerte in der Summe gesehen relativ gleich weit entfernt sind wie die nach ihm liegenden Merkmalswerte:

$$\bar{x}_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{a_i}} = \frac{n}{\sum_{i=1}^{k}\frac{h_i}{x_i}}$$

mit k = Anzahl der Merkmalsausprägungen/ Klassen

Voraussetzungen: Zur Berechnung der relativen Entfernungen müssen Quotienten aus Merkmalswerten gebildet werden. Das Merkmal muss daher verhältnisskaliert sein. Die Merkmalswerte müssen alle positiv oder alle negativ sein.

Beispiel von vorne: Ein Student fährt täglich 60km zur Hochschule. Morgens hat er nur eine Durchschnittsgeschwindigkeit von 60km/h erreicht. Wie ist die Durchschnittsgeschwindigkeit für Hin- und Rückfahrt, wenn er abends einen Schnitt von 100km/h fährt?

Überprüfung der Formel am Beispiel

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{k}\frac{h_i}{x_i}} = \frac{120}{\frac{60}{60}+\frac{100}{60}} = 75$$

Eignung: Das harmonische Mittel ist zur Berechnung des Durchschnitts einzusetzen, wenn das Merkmal aus einem Quotienten hervorgeht und wenn der Zähler des Quotienten und die Häufigkeit auf dieselbe Dimension (z.B. Stundenkilometer; Eigenkapitalquote) bezogen sind. Für klassifizierte Häufigkeitsverteilungen kann das harmonische Mittel – analog dem arithmetischen Mittel - nur näherungsweise berechnet werden. Dazu sind in der o.g Formel die Merkmalswerte x_i durch die Klassenmitten x_i^* zu ersetzen.

Quelle: Bourier: Beschreibende Statistik 2018, mein altes Skript

3.2 Noch ein Beispiel zum Nachdenken

Beispiel: Wenn Aktie A im ersten Jahr eine Rendite von 50% erwirtschaftet und im zweiten Jahr eine Rendite von -33%, haben Sie eine durchschnittliche Gesamtrendite von 17% bzw. ohne Zinseszinseffekte eine jährliche Rendite von 8,5% erwirtschaftet?

t = -2	t = -1	t = 0
100 €	150 €	100 €

+ 50% - 33%

Lösung: Die Rendite von 50% hat eine ***andere Bezugsbasis*** wie die negative Rendite von 33%. Die 50% beziehen sich auf einen Wert von 100 €, die -33% hingegen auf einen Wert von $150€ = 100 \cdot (1 + 0{,}5)$. Korrekt ist die Ermittlung der Rendite durch ***Multiplikation mit den Zuwachsfaktoren***:

t = -2	t = -1	t = 0
100 €	150 €	100 €

+ 50% - 33%

Quelle: Vorlesung Finanzmanagement, Wirth

3.2 Das geometrische Mittel

Das ***geometrische Mittel*** ist der Wert, der mehrere aufeinanderfolgende Vervielfachungen einer Größe als durchschnittliche Vervielfachung wiedergibt.

$\overline{x_G} = \overline{z_G} - 1$ mit z_i Zuwachsfaktoren $z_i = 1 + x_i$ und $\overline{z_G} = \sqrt[n]{z_1^{p_1} \cdot z_2^{p_2} \cdot \ldots \cdot z_n^{p_m}}$

Voraussetzungen: Die den Merkmalswerten zugrunde liegenden Größen müssen wegen der Division verhältnisskaliert sein.

Beispiel: Für den Zeitraum 2012 bis 2017 sei die Gewinnentwicklung der Software KG angegeben. Die Geschäftsleitung interessiert sich für den durchschnittlichen prozentualen Gewinnanstieg pro Jahr (Vervielfachung) im Betrachtungszeitraum.

Jahr	Gewinn (€)	Wachstumsfaktor x_i	Wachstumsrate (%)
2012	120.000		
2013	138.000		
2014	165.600		
2015	157.320		
2016	188.784		
2017	235.980		

Quelle: Bsp. Bourier: Beschreibende Statistik 2018

3.2 Das geometrische Mittel

Schrittfolge zur Bestimmung des geometrischen Mittels:

- Schritt 1: Berechnung der n Wachstumsfaktoren aus den Ausgangswerten
- Schritt 2: Berechnung des Produktes der Wachstumsfaktoren
- Schritt 3: Ziehen der n-ten Wurzel aus dem Produkt

Eignung: Das geometrische Mittel ist die einzige Möglichkeit, die durchschnittliche prozentuale (relative) Entwicklung einer Größe im Zeitablauf exakt zu beschreiben. Darin liegt die Bedeutung des geometrischen Mittels. Das geometrische Mittel ist zwingend anzuwenden, wenn die *durchschnittliche prozentuale (relative)* Entwicklung einer Größe (Gewinn, Kapital, Aktienkurs, Sozialprodukt, Bevölkerung, Preis etc.) zu bestimmen ist. Da die zu mittelnden Wachstumsfaktoren nicht additiv, sondern multiplikativ verbunden sind, ist der Einsatz des arithmetischen Mittels nicht zulässig.

Lesen und überlegen Sie Sie genau !

Das Problem hier ist der Wechsel der Basis. Selbst Experten fallen zuweilen darauf herein. In einem Mathematiklehrbuch habe ich einmal gelesen: Wenn eine Kuh 25% mehr Milch produziert, dann braucht der Bauer 25% weniger Kühe für die gleiche Menge Milch.

Pustekuchen! Der Bauer braucht 20% weniger Kühe für die gleiche Menge Milch. Wenn er vorher zehn Kühe hat, jede gibt 10 Liter Milch, dann hat er 100 Liter Milch. 25% mehr macht 12,5 Liter Milch pro Kuh. Also braucht er für 100 Liter jetzt 8 Kühe, das sind 20% und nicht 25% weniger als zuvor.

Quelle: Bourier: Beschreibende Statistik 2018, Krämer: So lügt man mit Statistik, 2018

3.2 Zusammenfassung Lageparameter

Es wird empfohlen für folgende Fragestellungen die folgenden Lageparameter zu verwenden:

Skalierung	Nominal	Ordinal	Intervall	Verhältnis
Zu verwendende Lageparameter	Modalwert	Median	Arithmetisches Mittel bei Wachstumsfaktoren geometrisches Mittel	Arithmetisches Mittel bei Wachstumsfaktoren geometrisches Mittel

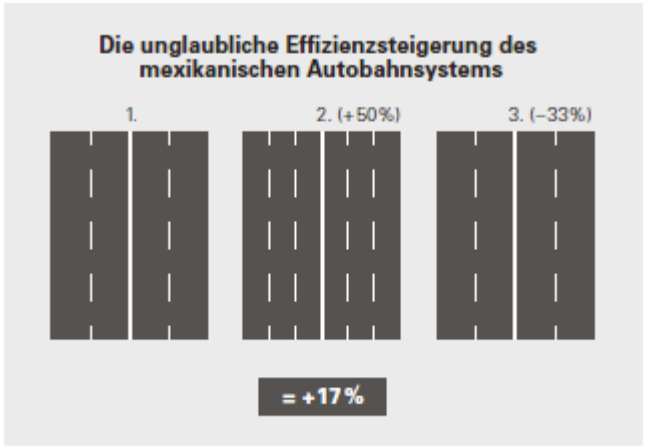
Einige relativierende und ergänzende Bemerkungen sollten jedoch hinzugefügt werden:

- Im Gegensatz zum Modus stimmen das arithmetische Mittel sowie der Median u. U. mit keinem der Beobachtungswerte überein. Der altbekannte Vorwurf, es gebe den „***mittleren Bundesbürger***“ oder den „mittleren Haushalt“ überhaupt nicht, kann in gewissen Fällen berechtigt sein.
- In der Tat kann der Modus auch bei ordinal- oder kardinalskalierten Merkmalen den sinnvolleren Lageparameter (z.B. bei Schuhgrößen oder Konfektionsgrößen) darstellen.
- Im Unterschied zum arithmetischen Mittel, das auf ***Ausreißer*** (oder Fehler) in den Daten sehr empfindlich reagiert, erweist sich der Median gegenüber Ausreißern als robust.
- Bei klassierten Daten ist der exakte Mittelwert nicht zu ermitteln. Ersatzweise wird das gewogene Mittel der Klassenmitten verwendet, wobei die relativen Klassenhäufigkeiten als Gewichte fungieren.

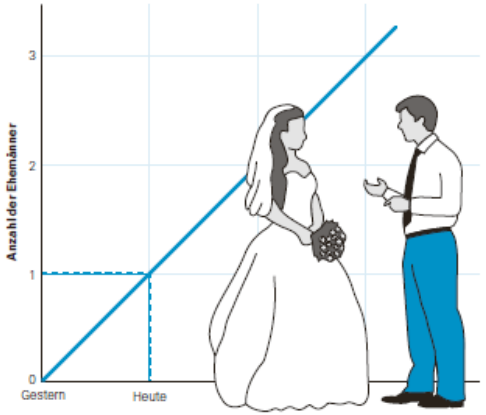
Quelle: Bamberg: Statistik 2017

3.2 Zusammenfassung der Lageparameter

»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«



Wachstum trotz Konstanz



»Sehen Sie mal, Ende des nächsten Monats werden Sie mehr als zwei Dutzend Ehemänner haben. Besser, Sie denken schon mal über einen Rabatt für Hochzeitstorten nach.«

Quelle: Krämer: So lügt man mit Statistik, 2018

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Gliederung

1. **Einführung und Aufgaben**
   Data Science und Statistik**,** deskriptive Statistik, Wahrscheinlichkeitsrechnung, induktive Statistik, Vorgehensweisen und Beispiele

2. **Grundbegriffe, statistische Untersuchung und Darstellung der Daten**
   Datenquellen, Merkmale, Grundgesamtheit, Stichprobe, Messskalen, Vorgehensweise bei statistischen Untersuchungen Planung, Datenerhebung und -techniken, Datenaufbereitung, tabellarische und grafische Darstellung, Interpretation

3. **Eindimensionale Häufigkeitsverteilungen**

   3.1 Terminologie und grafische Darstellungen
   3.2 Lageparameter
   3.3 Streuungsparameter
   3.4 Schiefe und Konzentration

3.3 Wiederholung: Statistische Maßzahlen

Für viele statistische Fragestellungen möchte man, eine Häufigkeitsverteilung durch einige wenige informative Größen oder ***Maßzahlen*** ersetzen. Die Charakterisierung der Häufigkeitsverteilung durch statistische Maßzahlen ist wie die Klasseneinteilung eine ***Datenreduktion***, um das Datenmaterial überschaubarer und aussagekräftiger zu machen. Man unterscheidet Lageparameter, Streuungsparameter, Momente und Konzentrationsmaße.

Eine radikalere Komprimierung der Ausgangsdaten als beim Übergang auf einen Häufigkeitsverteilung ist dann gegeben, wenn die ***gesamte Urliste durch eine einzige Zahl***, einen sogenannten ***Lageparameter*** charakterisiert wird. Dieser soll „möglichst gut“ beschreiben, wo das gesamte Datenmaterial auf der Merkmalsachse lokalisiert ist

Lageparameter sollen folgenden Forderungen genügen:

- Der Lageparameter repräsentiert die Verteilung möglichst gut
- Der Lageparameter ermöglicht den Vergleich verschiedener Verteilungen des gleichen Merkmals bzgl. verschiedener statistischer Massen
- Der Lageparameter ermöglicht die Beurteilung von Einzelwerten/Gruppen von Einzelwerten

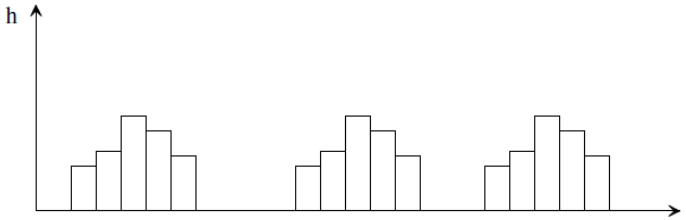


Abb. 3.1: Häufigkeitsverteilungen mit unterschiedlicher Lage

Quelle: Bamberg/Bauer: Statistik 2017

3.3 Streuungsmaße

Das arithmetische Mittel verkleistert oft eine große Ungleichheit – es schweigt sich zur Streuung um den Mittelwert völlig aus. Wenn es in unserem Dorf zehn Bauern gibt, von denen einer 40 Kühe hat und alle anderen haben nichts, so hat im Mittel jeder vier. Für die neun Habenichtse ist das aber nur ein schwacher Trost. Offenbar macht es einen Unterschied, ob sich die Werte dicht um das Mittel sammeln oder ob sie in alle Winde streuen, aber diesen Unterschied sieht man dem Mittelwert nicht an.

»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

Ein Mittelwert ohne Streuung ist also nur die Hälfte wert. Zu einem seriösen Durchschnitt gehört in aller Regel auch ein Maß für die Abweichung davon. Wie dieses aussieht, sei uns hier egal – wichtig ist: Bei nackten Mittelwerten ist immer Vorsicht angezeigt.

Die ***Streuung der Merkmalswerte*** ist neben den Mittelwerten die zweite wesentliche Eigenschaft einer Häufigkeitsverteilung, die die Lageparameter flankieren. So ist es ein wesentlicher Unterschied, ob die Merkmalswerte in einem engen Bereich oder in einem sehr breiten Bereich um den Mittelwert herum liegen bzw. wie weit sie um das Zentrum der Verteilung streuen.

Quelle: Bourier: Beschreibende Statistik 2018, Krämer: So lügt man mit Statistik, 2018

3.3 Die Spannweite

Die ***Spannweite*** ist die Differenz aus dem größten und dem kleinsten beobachteten Merkmalswert. Sie gibt also die Länge des Bereiches an, über den sich die Merkmalswerte verteilen.

$$w = \max_i a_i - \min_i a_i$$

Die Spannweite setzt voraus, dass das Merkmal mind. Intervallskaliert ist. Bei der Spannweite werden nur die beiden extremen Werte berücksichtigt, was zu insbesondere bei Ausreißern zu starken Verzerrrungen führen kann. Die Spannweite hat deshalb eine geringe Aussagekraft und wird in der Regel nur als erste grobe Abschätzung verwendet.

Beispiel: Überstunden der Beschäftigten der Schulte GmbH

Überstunde	0	1	2	3	4	12
Beschäftigte	3	10	4	3	2	1

Die Spannweite vermittelt, sofern keine Ausreißer vorliegen, eine grobe Vorstellung von der Streuung. Sie ist geeignet, wenn allein die Länge des Streubereiches interessiert. Dies ist insbesondere der Fall, wenn die äußersten Werte der Häufigkeitsverteilung von Bedeutung sind. In der praktischen Anwendung wird die Spannweite oft unter Nennung des kleinsten und größten Merkmalswertes angegeben. Man denke z.B. an die Angaben "höchst/tiefst" bei Börsenkursen oder "minimal/maximal" bei Temperaturangaben. Damit beschreibt die Spannweite zugleich die Lage der Häufigkeitsverteilung. Bei der klassifizierten Häufigkeitsverteilung werden als kleinster Wert die Untergrenze der ersten Klasse und als größter Wert die Obergrenze der letzten Klasse verwendet.

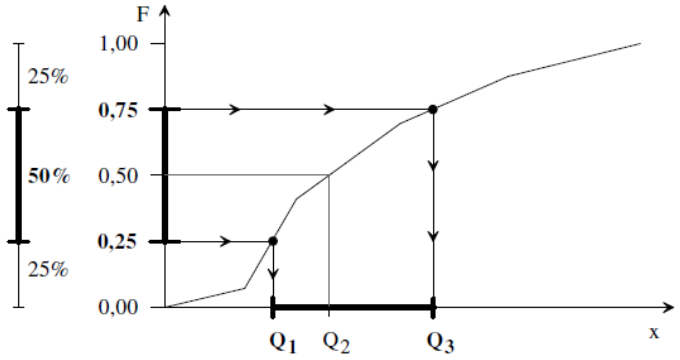
Quelle: Bourier: Beschreibende Statistik 2018

3.3 Der zentrale Quartilsabstand

Der ***zentrale Quartilsabstand*** ist die Entfernung zwischen den beiden Merkmalswerten, welche die in der Rangordnung zentral gelegenen 50% der Merkmalsträger eingrenzen.

$$z = Q_3 - Q_1 \quad \text{mit } Q_1 = a_{\left[\frac{1*n}{4}\right]} \quad \text{und} \quad Q_3 = a_{\left[\frac{3*n}{4}\right]}$$

Der zentrale Quartilsabstand ist ein anschauliches Streuungsmaß. Wie bei der Spannweite wird über den Streubereich informiert, nicht aber wie die Merkmalswerte in diesem Bereich streuen. Im Unterschied zur Spannweite tritt das Ausreißer-Problem hier nicht auf, da die unteren und oberen 25% der Häufigkeitsverteilung abgeschnitten werden.



Beispiel: Fehlzeiten der Beschäftigten der Maier KG

Fehltage	0	2	5	6	7	11	12	14
h_i	4	2	2	2	4	3	2	1
H_i	4	6	8	10	14	17	19	20

Der zentrale Quartilsabstand ist aufgrund seiner Konstruktion als Streuungsmaß geeignet, wenn der Kernbereich - hier 50% - einer Häufigkeitsverteilung interessiert. So ist es z.B. bei der Verteilung des Einkommens oder des Vermögens von Interesse, in welchem Bereich die mittleren 50% der Haushalte streuen. Bei klassifizierten Häufigkeitsverteilungen wird der zentrale Quartilsabstand mit der obigen Formel berechnet, die näherungsweise Berechnung der Quartilswerte wurde weiter vorne aufgezeigt.

Quelle: Bourier: Beschreibende Statistik 2018

3.3 Die mittlere absolute Abweichung

Die ***mittlere absolute Abweichung*** ist die durchschnittliche Entfernung aller beobachteten Merkmalswerte von einem Lageparameter m (arithmetisches Mittel oder Median).

$$d = \frac{1}{n}\sum_{i=1}^{n}|a_i - m| = \frac{1}{n}\sum_{i=1}^{k} h_i |x_i - m| = \sum_{i=1}^{k} f_i |x_i - m|$$

Beispiel: Für die Merkmalswerte 3, 7, 8, 9, 13 beträgt der Median 8 und die mittlere absolute Abweichung

Beispiel: Fehlzeiten der Beschäftigten der Schulte GmbH. Bestimmen Sie die mittlere absolute Abweichung vom arithmetischen Mittel, welches bei 2,04 liegt.

x_i	h_i
0	3
1	10
2	4
3	3
4	2
12	1
	23

Quelle: Bourier: Beschreibende Statistik 2018

3.3 Die mittlere absolute Abweichung

Bei ***klassifizierten Häufigkeitsverteilungen*** repräsentiert die Klassenmitte die gesamte Klasse und es ergibt sich als Näherungswert für die mittlere absolute Abweichung

$$d = \frac{1}{n}\sum_{i=1}^{k} h_i^* \left|x_i^* - m\right| = \sum_{i=1}^{k} f_i^* \left|x_i^* - m\right|$$

Beispiel: Klassifizierte Forderungen, wobei das arithmetische Mittel bei 320,92€ liegt

Forderungen €			
von	bis unter	x_i^*	h_i
50	100	75	15
100	200	150	50
200	300	250	80
300	400	350	40
400	600	500	40
600	1000	800	20
			245

Da die Abstände zwischen den Merkmalswerten und ihrem Mittelwert zu berechnen sind, muss das Merkmal mindestens intervallskaliert sein. Die mittlere Abweichung entspricht der allgemeinen Vorstellung von Streuung und ist zur Messung der Streuung sehr gut geeignet, falls nicht Ausreißer zu einer Verzerrung führen. Sie ist für die beschreibende Statistik deutlich geeignet als die Varianz/Standardabweichung, die aufgrund ihrer Bedeutung in der schließenden Statistik die mittlere Abweichung im praktischen Einsatz leider verdrängen.

Quelle: Bourier: Beschreibende Statistik 2018

3.3 Die Varianz und die Standardabweichung

Die ***Varianz σ²*** ist definiert als das arithmetische Mittel der quadratischen Abweichungen der Merkmalswerte vom arithmetischen Mittel und die ***Standardabweichung σ*** als positive Wurzel daraus.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} h_i (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$$

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

Bei ***klassifizierten Häufigkeitsverteilungen*** repräsentiert die Klassenmitte x_i^* die gesamte Klasse und es ergibt sich als Näherungswert für die Varianz

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} h_i^* (x_i^* - \bar{x})^2 = \sum_{i=1}^{k} f_i^* (x_i^* - \bar{x})^2$$

Beispiel: Für das bekannte Beispiel des Arbeitsweges, bei der ein Angestellter den Weg zwischen Wohnung und Arbeitsstätte an 5 Tagen in 12, 10, 16, 12, und 17 Minuten zurücklegt, ergibt sich

Quelle: Bourier: Beschreibende Statistik 2018

3.3 Die Varianz und die Standardabweichung

Zur Rechenvereinfachung lässt sich die Varianz σ^2 folgendermaßen umformen

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

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} h_i x_i^2 - \bar{x}^2 = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2$$

Beispiel: Überstunden der Beschäftigten der Schulte GmbH

x_i	h_i
0	3
1	10
2	4
3	3
4	2
12	1
	23

Alternativ:

Quelle: Bourier: Beschreibende Statistik 2018

3.3 Die Varianz und die Standardabweichung

Zur Rechenvereinfachung bei ***klassifizierten Häufigkeitsverteilungen***:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} h_i^* x_i^{*2} - \bar{x}^2 = \sum_{i=1}^{k} f_i^* x_i^{*2} - \bar{x}^2$$

Beispiel: Die Berechnung ist als ***Selbstlerneinheit*** am Beispiel der klassifizierten Forderungen durchzuführen

Forderungen €			
von	bis unter	x_i^*	h_i
50	100	75	15
100	200	150	50
200	300	250	80
300	400	350	40
400	600	500	40
600	1000	800	20
			245

Quelle: Bourier: Beschreibende Statistik 2018

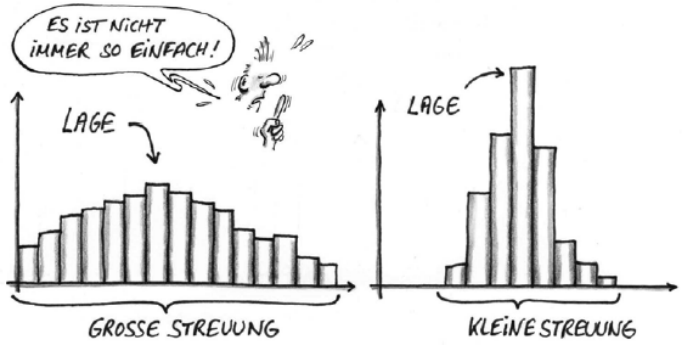
3.3 Die Varianz und die Standardabweichung

Die Varianz besitzt die folgenden wesentlichen ***Eigenschaften***:

- Werden die Merkmalswerte einer ***linearen Transformation*** $a_i^* = m * a_i + t$ unterworfen (t beliebig, m beliebig $\neq 0$) so gilt, für die Varianz $\sigma^{*2} = m^2\sigma^2$. Insbesondere gilt für den wichtigen Spezialfall m=1 , $\sigma^{*2} = \sigma^2$ d.h. die Varianz bzw. auch die Standardabweichung bleiben unverändert, wenn man die Merkmalswerte um einen bestimmten Wert vergrößert oder verkleinert.
- Ähnlich wie beim arithmetischen Mittel einer Grundgesamtheit, die sich aus zwei Teilgesamtheiten mit den Umfängen n_1 und n_2 zusammensetzt beträgt die Varianz dieser Grundgesamtheit

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}$$

- Das arithmetische Mittel der quadratischen Abweichung der Merkmalswerte von einem bestimmten Wert wird minimal, wenn man die Abweichungen bzgl. des arithmetischen Mittels bildet. Dies bezeichnet man als ***Minimumseigenschaft***, d.h. die mittlere quadratische Abweichung von einem beliebigen Wert M kann nie kleiner sein als die mittlere quadratische Abweichung bezogen auf das arithmetische Mittel



Quelle: Bleymüller/Gehlert/Gülicher, 2015, https://datatab.de/tutorial/standardabweichung-varianz-spannweite

3.3 Der Variationskoeffizient

Die behandelten Streuungsmaße messen die Streuung, ohne die Lage (Niveau) der Häufigkeitsverteilung zu berücksichtigen. So wird eine Abweichung von 5€ bei einem Preisniveau von 50€ als genauso hoch angesehen wie bei einem Preisniveau von 10.000€. Die absolute Abweichung ist in beiden Fällen mit 5€ identisch. Betrachtet man jedoch die Abweichung im Verhältnis zum Preis, dann ist die Abweichung im zweiten Fall deutlich geringer. Diese relative Betrachtungsweise liegt dem Variationskoeffizienten VK zugrunde.

Der Variationskoeffizient misst nicht die absolute, sondern die ***relative Streuung***, d.h. er setzt die Streuung in Relation zur Lage der Häufigkeitsverteilung. Als Parameter haben sich das arithmetische Mittel und die Standardabweichung durchgesetzt.

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

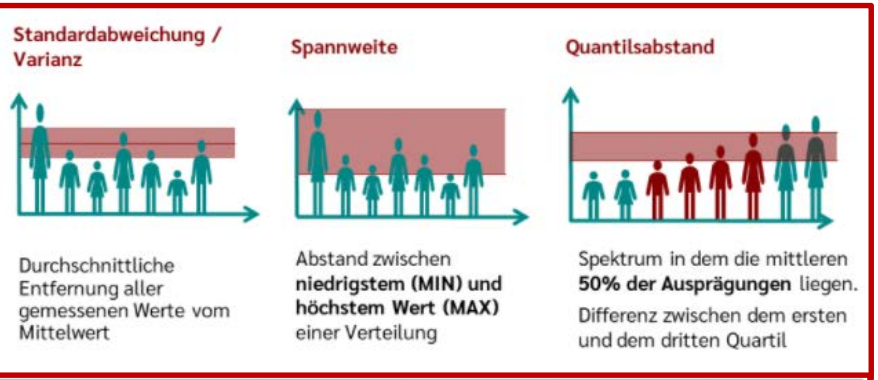
Beispiel: Die Arbeiter eines Unternehmens verdienen durchschnittlich 12,2 €/Std. und die Arbeiterinnen durchschnittlich nur 8,40 €/Std. Die Standardabweichungen betragen bei den Männern 1,80 €/Std. und bei den Frauen 1,30 €/Std. Daraus könnte man schließen, dass die Löhne der Männer stärker streuen als bei den Frauen. Die Bewertung der Standardabweichungen ist aber von den Lohnniveaus abhängig. Dieser Effekt lässt sich durch den Variationskoeffizienten ausschalten, der folgendermaßen ergibt:

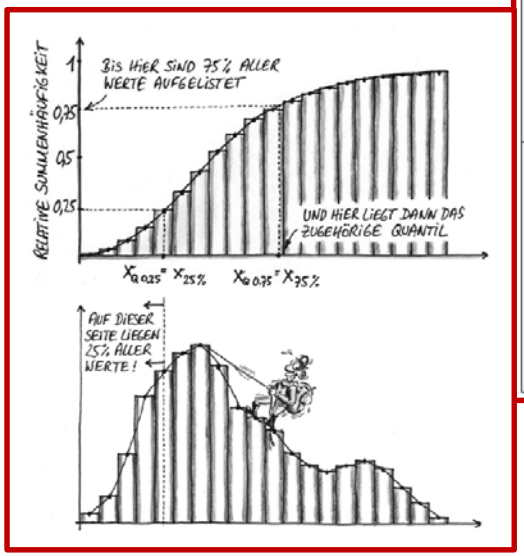
Quelle: Bourier: Beschreibende Statistik 2018

3.3 Die Streuungsmaße

Streuungsmaße

Standardabweichung, Varianz und Spannweite gehören zu den **Streuungsmaßen** in der **deskriptiven Statistik**. Sie werden auch Maße der Dispersion genannt und dienen dazu die Streuung von Werten einer Stichprobe rundum einen Lageparameter zu beschreiben. Vereinfacht ausgedrückt, sind **Streuungsparameter** ein Maß dafür, wie sehr eine Stichprobe um einen Mittelwert schwanken.





Skalierung		Nominal	Ordinal	Intervall	Verhältnis
Lageparameter	Modalwert	X	X	X	X
	Median		X	X	X
	Arithmetisches Mittel			X	X
	Geometrisches Mittel				X
Streuungs-parameter	Spannweite	X	X	X	X
	Mittlere absolute Ab-weichung		X	X	X
	Varianz Standardabweichung			X	X
	Variationskoeffizient				X

Nach unserm Exkurs in die Wahrscheinlichkeits-rechnung geht es mit der Schiefe/Wölbung weiter

Quelle: Bleymüller/Gehlert/Gülicher, 2015, https://datatab.de/tutorial/standardabweichung-varianz-spannweite

3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

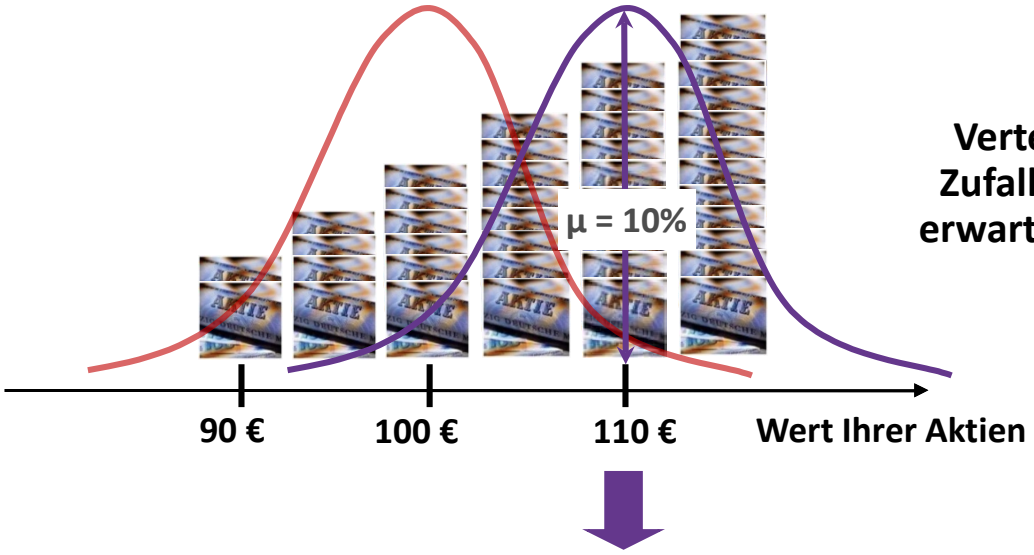
- **Wie ermittelt sich die Rendite?**
  Gewinn und Rendite werden direkt aus den Aktienkursveränderungen ermittelt:

$$\text{Rendite} = \frac{\text{Gewinn}}{\text{eingesetztes Kapital}} = \frac{110-100}{100} = \frac{10}{100} = 10\%$$

☺ **Chance**

deutlich schlimmer: $\text{Rendite} = \frac{90-100}{100} = \frac{-10}{100} = -10\%$

☹ **Risiko**



Verteilung der Zufallsvariablen erwartete Rendite

- **Erwartungswert der Rendite μ = 10%**
  ist die wahrscheinlichkeitsgewichtete Rendite, die Ihre Aktien nach einer Zeit haben. Sie tritt ***nicht mit Sicherheit***, sondern mit einer (Standard-)Abweichung auf.
- **Ziel einer Kapitalanlage: *Maximierung der Rendite***

Quelle: Wirth: Vorlesung Risikomanagement

3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

Erwartungswert

Der Erwartungswert einer Zufallsvariablen X ist jener Wert, der sich bei oftmaligem Wiederholen des Experiments als ***Mittelwert der Ergebnisse*** ergibt. Hat eine Zufallsvariable X die Werte $x_1;x_2;.....x_n$ und p_i bezeichne die Wahrscheinlichkeit mit der der Wert eintritt, so heißt E(x) Erwartungswert von X:

$$E(X)=x_1 \cdot P(X=x_1) + x_2 \cdot P(X=x_2) +...+ x_n \cdot P(X=x_n)$$

Während sich der Mittelwert – eine Größe aus der beschreibenden Statistik – auf die Vergangenheit bezieht, also auf Werte, die in der Stichprobe tatsächlich aufgetreten sind, beschreibt der ***Erwartungswert*** eine Größe, die sich auf die ***Zukunft*** bezieht.

Erwartungswert – arithmetisches Mittel

Häufig wird die erwartete Rendite E(X) aus den Vergangenheitsdaten und damit aus dem arithmetischen Mittel approximiert, das den ***statistischen Durchschnittswert*** beschreibt.

$$E(X) = \bar{x} = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{n} p_j x_j$$

mit m<n
Merkmalswerte $x_1,..., x_i$
Einzelwahrscheinlichkeiten $p_1,...,p_n$

Der Erwartungswert der Rendite wird mit μ bezeichnet

Quelle: Wirth: Vorlesung Risikomanagement

3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

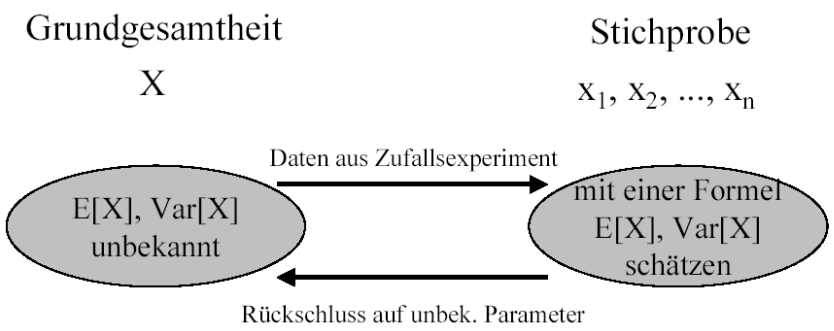
Stochastische Modelle

dienen zur Beschreibung von Finanzmärkten bzw. zur Bewertung der dort gehandelten Finanztitel

- Diese Finanztitel sind risikobehaftet, d.h. dass ihre zukünftige Entwicklung nicht präzise vorhergesagt werden kann
- Die Werte der Finanzinstrumente sind als Zufallsvariable aufzufassen und entsprechend zu modellieren
- Die Modellierung basiert auf der Überlegung Wertentwicklungen als stochastische Prozesse zu beschreiben

Abgrenzung

Während die ***Statistik*** als methodischer Werkzeugkasten dient, verwendet man die ***Stochastik*** um die Realität modellhaft abzubilden. Hierzu bedient man sich Dichtefunktionen, die eine Wahrscheinlichkeitsverteilung einer Größe modellieren und beschreiben. Auch hier benötigt man die ***Lage- und die Streuungsparameter*** zur genauen Spezifikation.



$x_1, x_2, ..., x_n$ sind Realisationen einer Zufallsvariablen
Folglich sind auch die Schätzungen für Erwartungswert und Varianz (aus Vergangenheitsdaten) Realisationen von Zufallsvariablen

Quelle: Wirth: Vorlesung Risikomanagement

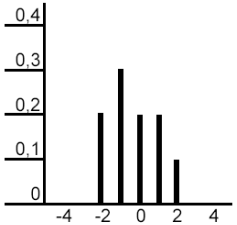
3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

- **Dichte- und Verteilungsfunktion**

Dichtefunktion

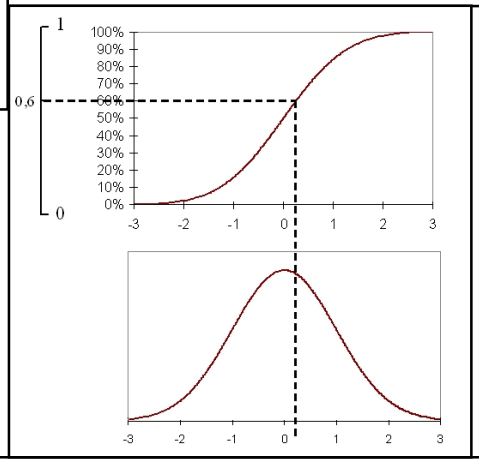
- ➔ mit welcher Wahrscheinlichkeit ein bestimmter Wert P(X) der Zufallsvariable X eintritt
- ➔ legt fest, wie die gesamte Wahrscheinlichkeitsmasse 1 auf der Zahlengeraden zu verteilen ist

Wahrscheinlichkeiten addieren sich im diskreten Fall zu eins



diskrete Dichte: 1. Angabe der Trägermenge (x_1,x_2, ...); 2. Bestimmung der Einzelwahrscheinlichkeiten, deren Summe gleich 1 ergibt

stetige Dichte: ein Bereich/Intervall $[a,b]$ erhält eine positive Wahrscheinlichkeit; diese bestimmt sich aus der Fläche unter der Dichtekurve in den Grenzen a und b. Die Gesamtfläche unter der Dichtekurve ergibt den Wert 1



Verteilungsfunktion

- ➔ Summe der Einzelwahrscheinlichkeiten (Integral der Dichtefunktion)
  - liefert die Wahrscheinlichkeit, dass eine Zufallsvariable X maximal einen bestimmten Wert a annimmt [$F_X(a) = P(-\infty < X \leq a)$]
  - wird umgekehrt eine bestimmte Wahrscheinlichkeit p vorgegeben, so liefert die Umkehrfunktion der Verteilungsfunktion $F_X^{-1}(p)$ [=p-Quantil] den Wert, den die Zufallsvariable X mit der vorgegebenen Wahrscheinlichkeit maximal annimmt

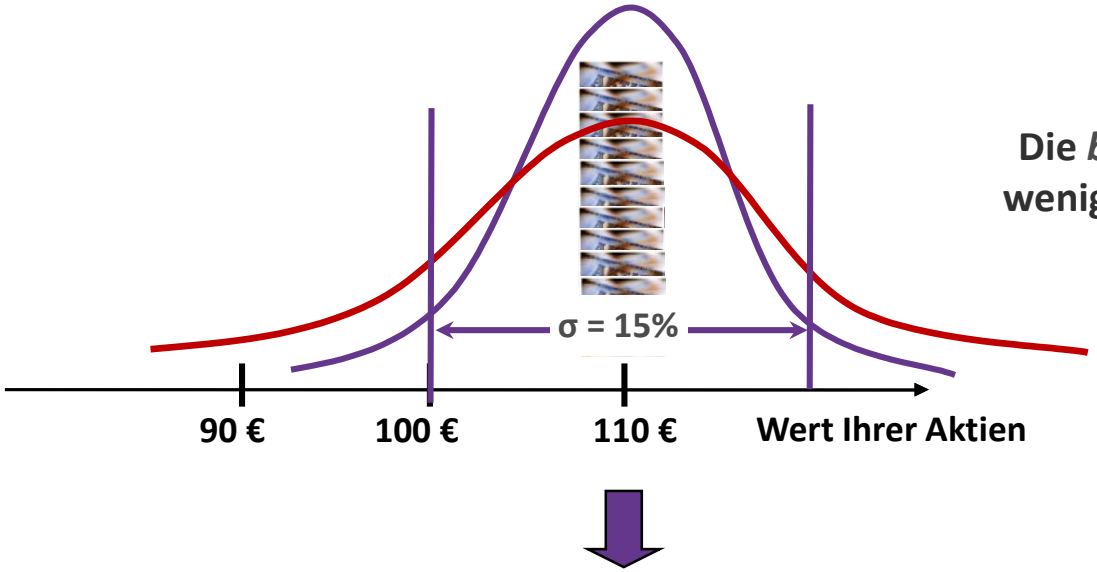
Quelle: Wirth: Vorlesung Risikomanagement

3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

- **Wo liegt jetzt hier Ihr Risiko?**

  Als ***Risiko*** empfinden Sie einen Wert kleiner als 100 €, z.B. eine Rendite = −10% ☹

- **Risiko und seine Interpretation**
  - Das ***Risiko*** ein Ziel nicht zu erreichen: ***downside risk*****.**
  - Dem gegenüber steht die ***Chance***, dass man besser ist als das Ziel, das man sich gesetzt hat: ***upside risk***.



Die *blaue Kurve* hat weniger Risiko als die *rote Kurve*

- **Risiko auf Basis der Standardabweichung σ = 15%**

  häufig verwendetes Risikomaß für die Streuung der Werte um den Mittelwert.

Quelle: Wirth: Vorlesung Risikomanagement

3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

Risikomaße

Es haben sich unterschiedliche Risikomaße als zweckmäßig erwiesen: die Varianz, die Standardabweichung oder auch der Korrelationskoeffizienten bzw. der Beta-Faktor…

Streuungsparameter

Das Maß zur Messung der Streuung, d.h. die ***Verteilung*** von Werten ***um den Mittelwert,*** ist die ***Varianz***, die erwartete quadrierte Abweichung vom Erwartungswert. Je weiter die Realisationen vom Erwartungswert entfernt, desto größer die Varianz. Deren Quadrat-wurzel bezeichnet man als ***Standardabweichung***, ebenfalls ein ***Maß für die Streuung***. Für den diskreten Fall gilt:

$$\text{Standardabweichung} = \sigma(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Streuungsparameter – Abhängigkeit und Unabhängigkeit

Für den stetigen Fall gilt:

Varianz = $\text{Var}(X) = E\left[(X - E(X))^2\right]$ es gilt: $\text{Var}(aX+b) = a^2\text{Var}(X)$ mit a,b reelle Konstanten

$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(y)$ falls X und Y ***unabhängig***

$\text{Var}(X+Y) = \text{Var}(X) + 2\text{Cov}(X,Y) + \text{Var}(Y)$ falls X,Y ***abhängig***

Standardabweichung = $\sigma(X) = \sqrt{\text{Var}(X)}$ es gilt: $\sigma(aX+b) = a\sigma(X)$ mit a,b reelle Konstanten

$\sigma(X+Y) = \sigma(X) + \sigma(y)$ falls X und Y ***unabhängig***

$\sigma(X+Y) = \sqrt{\sigma^2(X) + 2\rho_{XY}\sigma(X)\sigma(Y) + \sigma^2(Y)}$

falls X,Y ***abhängig***

Quelle: Wirth: Vorlesung Risikomanagement

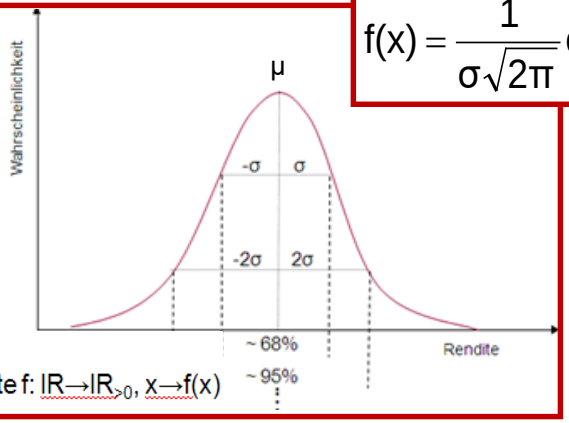
3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

Die Normalverteilung

Die Normalverteilung (***Gaußsche Glockenkurve***) unterstellt eine ***symmetrische Verteilungsform*** numerischer Daten. Ihr Kurvenverlauf ist symmetrisch, Median und Mittelwert sind identisch. Für die Normalverteilung gilt, dass sich die Werte der Zufallsvariablen in der Mitte der Verteilung konzentrieren und mit größerem Abstand zur Mitte immer seltener auftreten.

n	innen	außen
1	0,68268	0,31732
2	0,9545	0,0455
3	0,9973	0,0027

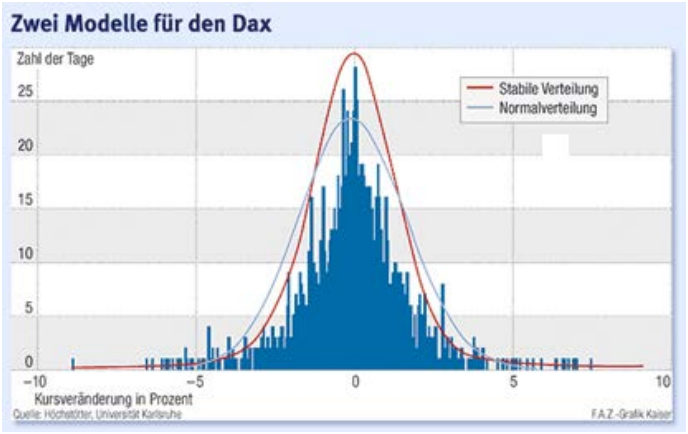
Anschaulich bedeutet dies, dass in ca. 2/3 aller Fälle sich z.B. die Rendite im Intervall ± eine Standardabweichung befindet und in ca. 1/3 aller Fälle außerhalb.



Stetige Zufallsvariable X

Wahrscheinlichkeitsdichte f: $\mathbb{R} \to \mathbb{R}_{>0}$, $x \to f(x)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Kritik an der Normalverteilung

Viele Modelle basieren auf der Normalverteilungsannahme, die allerdings in der Praxis durchaus bei Extremereignissen Schwierigkeiten macht. Wäre der DAX normalverteilt, so dürfte bei einer Standardabweichung von 1,6% eine negative Abweichung vom Mittelwert alle 120.000 Jahre (bei 250 Börsentagen) auftreten. In der Realität führen 5 Standardabweichungen zu einem daraus resultierenden Verlust von ca. 8% und der tritt alle 3-10 Jahre auf: Anschlag Gorbatschow-Putsch, Anschlag auf das World Trade Center, Finanzkrisen, Corona,.... ***Dennoch: für Schätzungen und Schlussfolgerungen im Normalfall passt das alles*** ☺

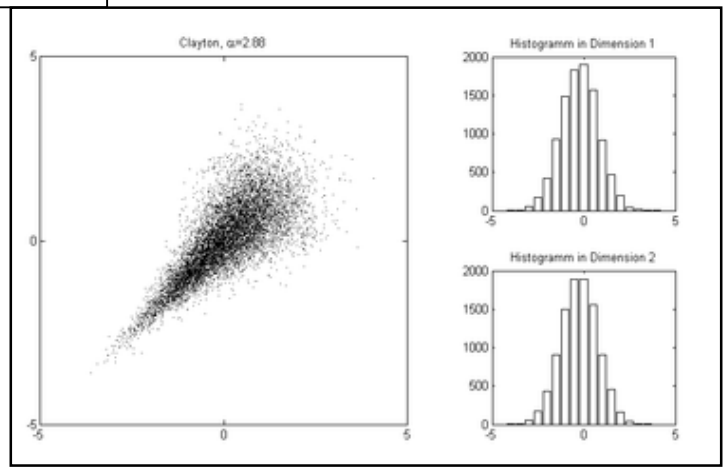
Quelle: Siegert: Risikomanagement

3.3 Exkurs: Lageparameter und Wahrscheinlichkeiten

Multivariate Zufallsvariablen

Multivariate Zufallsvariablen sind gekennzeichnet durch ihre Erwartungswerte μ, die Varianzen σ_i der einzelnen Komponenten, sowie durch die Kovarianzen $\sigma_{i,j} = Cov(X_i, X_j)$ zwischen den einzelnen Elementen. Multivariate Verteilungen modellieren mehrere voneinander abhängige oder unabhängige Risikofaktoren.

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \quad \text{und} \quad \sum = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{2n} & \sigma_{12} & \cdots & \sigma_n^2 \end{pmatrix}$$



Sie sehen, es bleiben spannende Themen im Data Science, damit geht es in der Wahrscheinlichkeitsrechnung und in der induktive Statistik weiter

Quelle: Siegert: Risikomanagement

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Gliederung

1. **Einführung und Aufgaben**
   Data Science und Statistik**,** deskriptive Statistik, Wahrscheinlichkeitsrechnung, induktive Statistik, Vorgehensweisen und Beispiele

2. **Grundbegriffe, statistische Untersuchung und Darstellung der Daten**
   Datenquellen, Merkmale, Grundgesamtheit, Stichprobe, Messskalen, Vorgehensweise bei statistischen Untersuchungen Planung, Datenerhebung und -techniken, Datenaufbereitung, tabellarische und grafische Darstellung, Interpretation

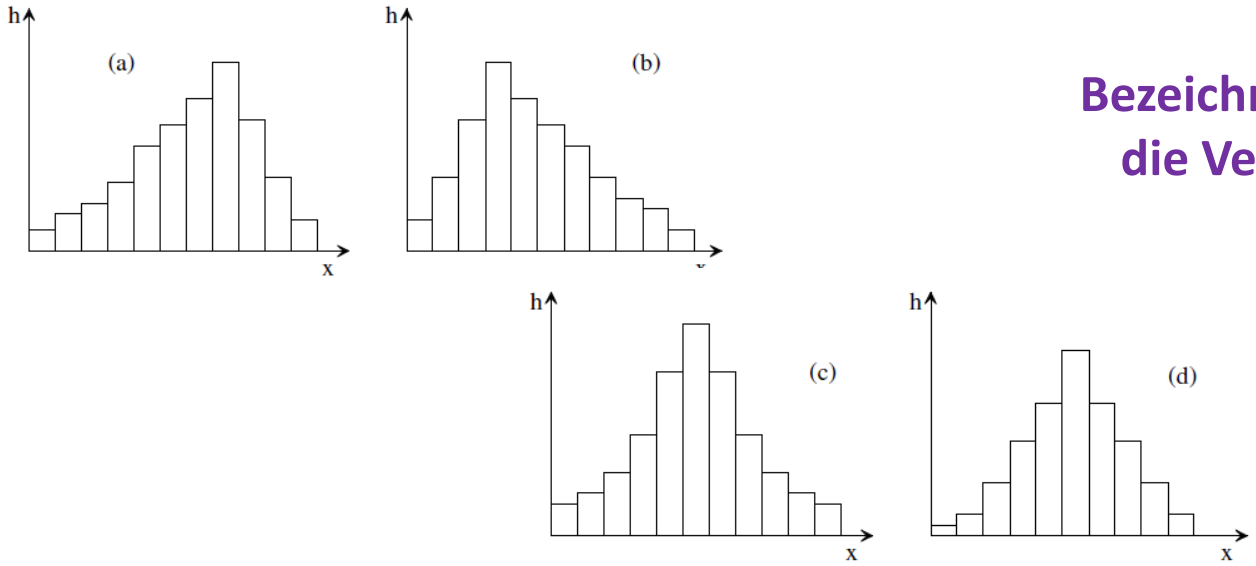
3. **Eindimensionale Häufigkeitsverteilungen**

   3.1 Terminologie und grafische Darstellungen
   3.2 Lageparameter
   3.3 Streuungsparameter
   3.4 Schiefe und Konzentration

3.4 Schiefe und Wölbung

Neben der Lage und Streuung sind die ***Schiefe*** und die ***Wölbung*** weitere wesentliche Eigenschaften einer Häufigkeitsverteilung. Sie sind durch die Betrachtung der graphischen oder tabellarischen Darstellung der Verteilung die beiden Eigenschaften i.d.R. gut zu erkennen.

Häufigkeitsverteilungen können ***symmetrisch*** oder ***asymmetrisch***, d.h. ***schief*** verlaufen. Im Falle der Schiefe ist zwischen ***rechtsschiefen*** (linkssteilen) und ***linksschiefen*** (rechtssteilen) Häufigkeitsverteilungen zu unterscheiden. Linksschiefe Verteilungen weisen bis zum Modus ein langsames (schiefes) Ansteigen und nach dem Modus ein schnelles (steiles) Abfallen der Häufigkeiten auf; bei rechtsschiefen Verteilungen ist dies umgekehrt.



Bezeichnen Sie bitte die Verteilungen

Quelle: Bourierr: Statistik 2018

3.4 Momente und Schiefe

Die ***Momente*** stellen die Verallgemeinerung des arithmetischen Mittels und der Varianz dar. Man unterscheidet zwischen ***Momenten um den Nullpunkt*** und ***Momenten um das arithmetische Mittel***.

$$m_r(0) = \frac{1}{n}\sum_{i=1}^{n} a_i^r = \frac{1}{n}\sum_{i=1}^{k} h_i x_i^r = \sum_{i=1}^{k} f_i x_i^r$$ wobei k die Anzahl Merkmalsausprägungen bzw. der Klassen ist

$$m_r(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} (a_i - \bar{x})^r = \frac{1}{n}\sum_{i=1}^{k} h_i (x_i - \bar{x})^r = \sum_{i=1}^{k} f_i (x_i - \bar{x})^r$$

Auch hier können wie bereits bei den anderen Lageparametern, die r-ten Momente um den Nullpunkt bzw. um das arithmetische Mittel über die Klassenmitte näherungsweise berechnet werden. Liegen z.B. links vom Modus mehr Merkmalsträger als rechts vom Modus, dann ist die Verteilung linksschief (linkslastig). Oder ist die Entfernung vom Median zum 1. Dezil weiter als zum 9. Dezil, dann ist die Verteilung linksschief.

Das 3te Moment um das arithmetische Mittel wird als Maßzahl für die ***Schiefe*** verwendet und gibt an, wie stark eine Häufigkeitsverteilung von der Symmetrie abweicht. Ferner gibt das Vorzeichen an, ob es sich um eine ***linksschiefe*** (negatives Vorzeichen) oder um eine ***rechtsschiefe*** Häufigkeitsverteilung handelt.

$m_3(\bar{x}) > 0$ rechtsschiefe Verteilung

$m_3(\bar{x}) = 0$ symmetrische Verteilung

$m_3(\bar{x}) < 0$ linksschiefe Verteilung

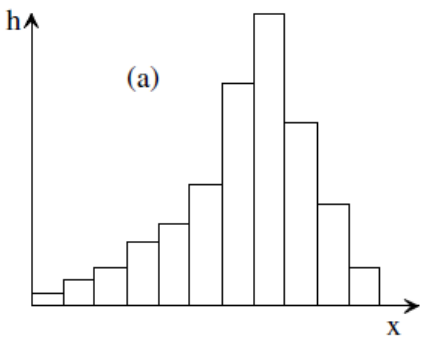
Quelle: Bourierr: Statistik 2018

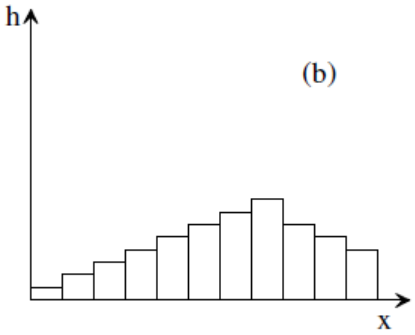
3.4 Momente und Schiefe

Die ***Wölbung*** (Exzess, Kurtosis) beschreibt die Steilheit, die Aufwölbung einer Häufigkeitsverteilung. Eine Verteilung kann z.B. ***steil*** oder ***flach*** aufgewölbt sein. Zur Messung der existieren verschiedene Wölbungsmaße, stellvertretend sei hier der ***Wölbungskoeffizient*** (***Kurtosis***) wiedergegeben.

$$w_4(\bar{x}) = \frac{\frac{1}{n}\sum_{i=1}^{k} h_i (x_i - \bar{x})^4}{\sigma^4} = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^4}{\sigma^4}$$

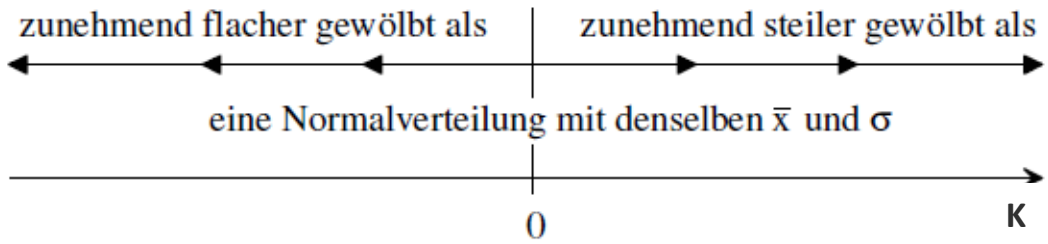
$$K = \frac{\frac{1}{n}\sum_{i=1}^{k} h_i (x_i - \bar{x})^4}{\sigma^4} = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^4}{\sigma^4} - 3$$





Interpretation der Kurtosis: mit der Kurtosis wird die Verteilung mit der Normalverteilung verglichen, deren Wölbung bei einem Wert von 3 liegt.

Interpretation der Kurtosis



3.4 Konzentration

Die Merkmalswertsumme kann z.B. gleichmäßig auf die Merkmalsträger verteilt sein oder sich auf nur wenige Merkmalsträger konzentrieren. Gegenstand der ***Konzentrationsmessung*** ist es, das Ausmaß der Konzentration zu beschreiben. Dies ist z.B. von großem Interesse bei der Verteilung des Einkommens (***Merkmalswertsumme***) auf die Haushalte (Merkmalsträger) oder bei der Verteilung der Marktanteile (Merkmalswertsumme) auf die Unternehmen (Merkmalsträger). Die Messung der Konzentration kann dabei relativ oder absolut erfolgen. Voraussetzung für die Messung ist, dass das Merkmal ***extensiv*** ist, d.h. die Addition der Merkmalswerte ist sinnvoll (z.B. Einkommen, Umsatz).

Unter einem ***Konzentrationsmerkmal*** versteht man ein Merkmal, bei dem die Bildung der Merkmalssumme möglich und sinnvoll ist. Wir gehen im folgenden von n Merkmalswerten $a_1<a_2...<a_q$, mit ihren absoluten Häufigkeiten $H_1,...,H_q$ und ihren relativen Häufigkeiten $h_1,...h_q$ aus. $a_1<a_2...<a_q$ können auch die Klassenmitten sein. Die Fragestellung der ***relativen Konzentrationsmessung*** lautet:

Welcher Anteil der Merkmalswertsumme entfällt
auf welchen Anteil der Merkmalsträger?

Es werden also zwei relative kumulierte Häufigkeiten gegenübergestellt.

Quelle: z.T. Bourier: Statistik 2018

3.4 Konzentration

Die relative Konzentrationsmessung wird am Beispiel einer klassifizierten Häufigkeitsverteilung erklärt. Die Ausführungen können auf nichtklassifizierte Verteilungen übertragen werden, indem die Klassenmitten durch die Merkmalswerte x_i ersetzt werden

Beispiel: 5000 Lagerpositionen: Welcher Anteil des gesamten Lagerwertes entfällt auf welchen Anteil der Lagerpositionen?

Lagerwert in Tsd. €			
von	bis unter	x_i*	h_i
1	5	3	2.000
5	15	10	1.200
15	25	20	800
25	50	37,5	700
50	100	75	200
100	395	247,5	100
			5.000

Ermittlung in Schritten:

1. Rangordnung der Merkmale x_i bzw. x_i* , hier sowieso gegeben
2. Ermittlung des gesamten Lagerwertes
3. Berechnung beider relativer Summenhäufigkeiten(relativer Anteil an den Lagerpositionen F_i und relativer Anteil am gesamten Lagervolumen S_i)
4. Treffen von Konzentrationsaussagen durch Gegenüberstellung und Verknüpfung von F_i und S_i

Quelle: z.T. Bourier: Statistik 2018

3.4 Konzentration

Die relative Konzentrationsmessung wird am Beispiel einer klassifizierten Häufigkeitsverteilung erklärt. Die Ausführungen können auf nichtklassifizierte Verteilungen übertragen werden, indem die Klassenmitten durch die Merkmalswerte x_i ersetzt werden

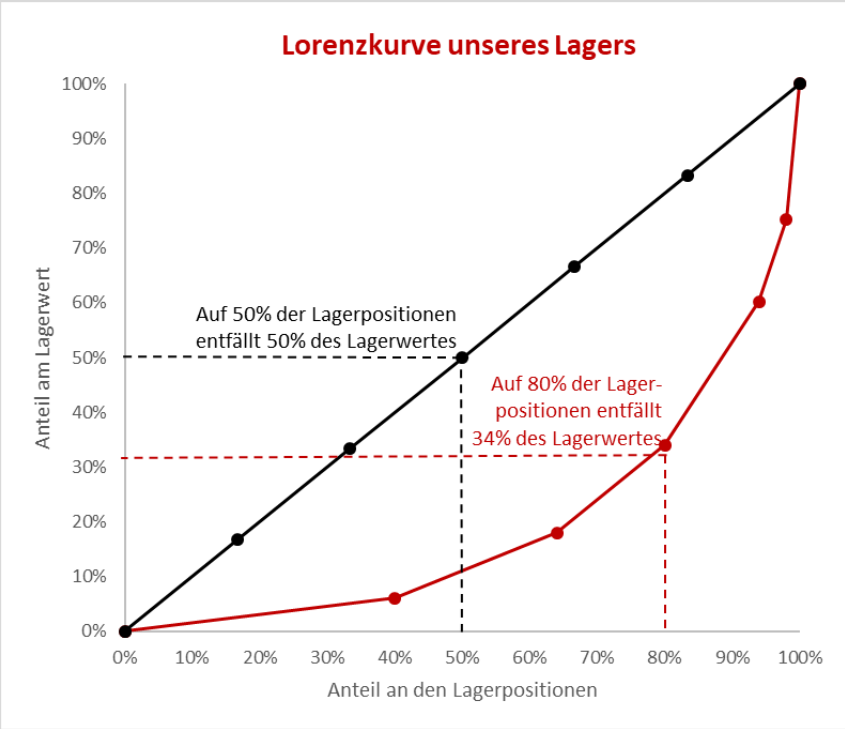
Beispiel: 5000 Lagerpositionen: Welcher Anteil des gesamten Lagerwertes entfällt auf welchen Anteil der Lagerpositionen?

Lagerwert in Tsd. €					1. rel. Summenhäufigkeit			2. rel. Summenhäufigkeit	
von	bis unter	x_i*	h_i	H_i	f_i	F_i	x_i* h_i	s_i	S_i
1	5	3	2.000	2.000	40%	40%	6.000	6%	6%
5	15	10	1.200	3.200	24%	64%	12.000	12%	18%
15	25	20	800	4.000	16%	80%	16.000	16%	34%
25	50	37,5	700	4.700	14%	94%	26.250	26%	60%
50	100	75	200	4.900	4%	98%	15.000	15%	75%
100	395	247,5	100	5.000	2%	100%	24.750	25%	100%
			5.000	5.000	100%		100.000	100%	

Quelle: z.T. Bourier: Statistik 2018

3.4 Konzentration und Lorenzkurve

Die Ergebnisse der relativen Konzentrationsmessung und damit das Ausmaß der Konzentration werden mit Hilfe der ***Lorenzkurve*** (Lorenz, Max; 1876 - 1959) oder ***Konzentrationskurve*** graphisch veranschaulicht. Die Koordinatenpunkte der Lorenzkurve ermöglichen punktuelle Aussagen zur Konzentration wie z.B., dass auf 85% der Lagerpositionen 43% des gesamten Lagerwertes entfallen.



Aussagen zur Lorenzkurve

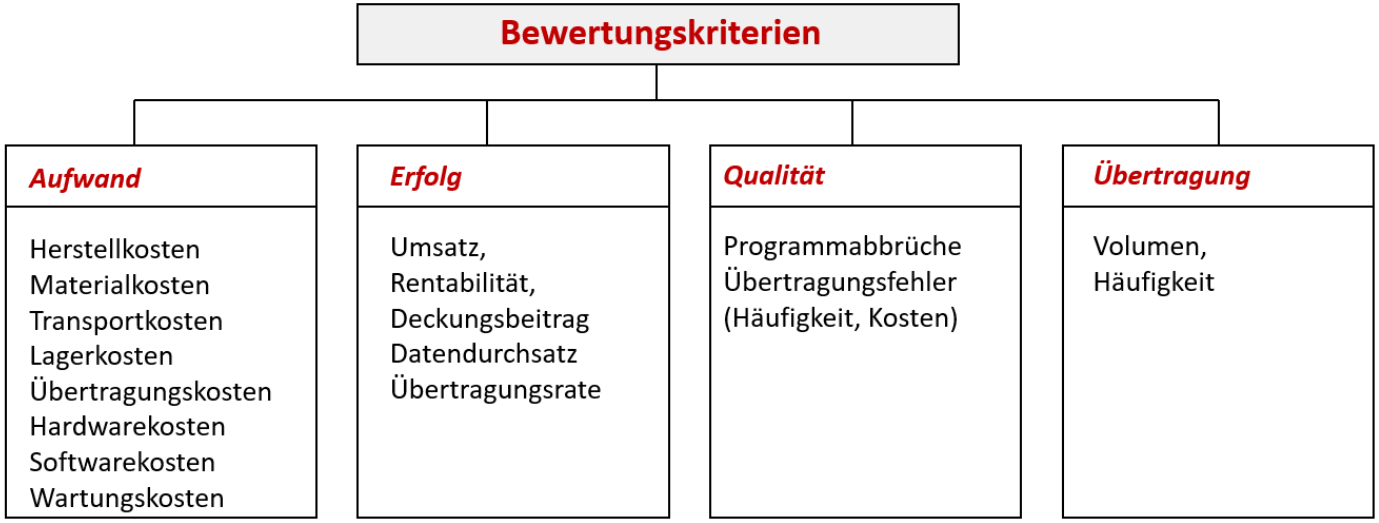
- Haben alle Merkmalsträger denselben Merkmalswert, dann liegt ***keine Konzentration*** vor. Auf 10% der Merkmalsträger entfallen 10% der Merkmalswertsumme, auf 50% entfallen 50% etc. In diesem Fall der Gleichheit ist die Lorenzkurve identisch mit der Diagonalen, die als ***Gleichheitsgerade*** bezeichnet wird. Die Fläche zwischen Lorenzkurve und Diagonale ist gleich Null.
- Bei ***maximaler Konzentration*** vereinigt ein einziger Merkmalsträger die gesamt Merkmalswertsumme auf sich, während auf die anderen n-1 Merkmalsträger nichts entfällt.
- ***Je näher die Lorenzkurve zur Diagonalen liegt, desto geringer ist die Konzentration. Je entfernter die Lorenzkurve zur Diagonalen liegt, desto größer ist die Konzentration***.

Quelle: z.T. Bourier: Statistik 2018

3.4 Ähnliche Anwendungen – die ABC-Analyse

Aufgabe der ***ABC-Analyse*** ist das ***Ermitteln der wirtschaftlichen Bedeutung*** verschiedener Gegenstände über eine Rangordnung und Zuordnung zu den Wertgruppen A, B und C. Sie soll:

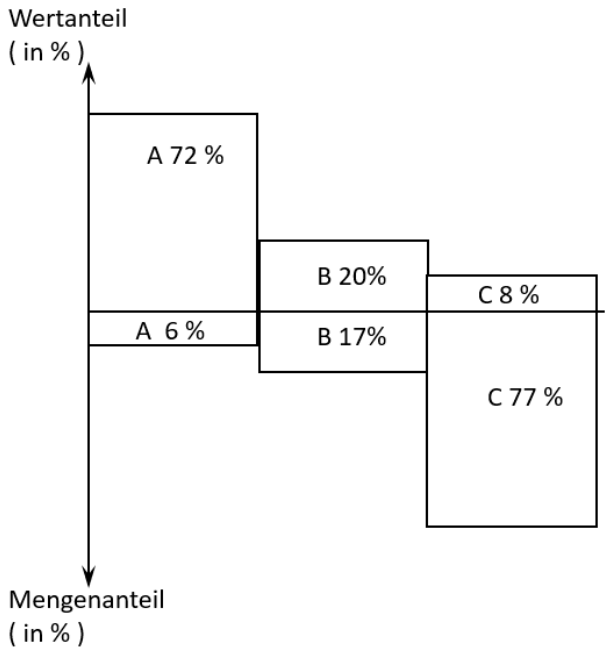
- das Wesentliche vom Unwesentlichen unterscheiden
- die Aktivitäten schwerpunktmäßig auf den Bereich hoher wirtschaftlicher Bedeutung lenken und gleichzeitig
- den Aufwand für die übrigen Gebiete durch Vereinfachungsmaßnahmen senken,
- die Effizienz von Management-Maßnahmen durch die Möglichkeit eines gezielten Einsatzes erhöhen.

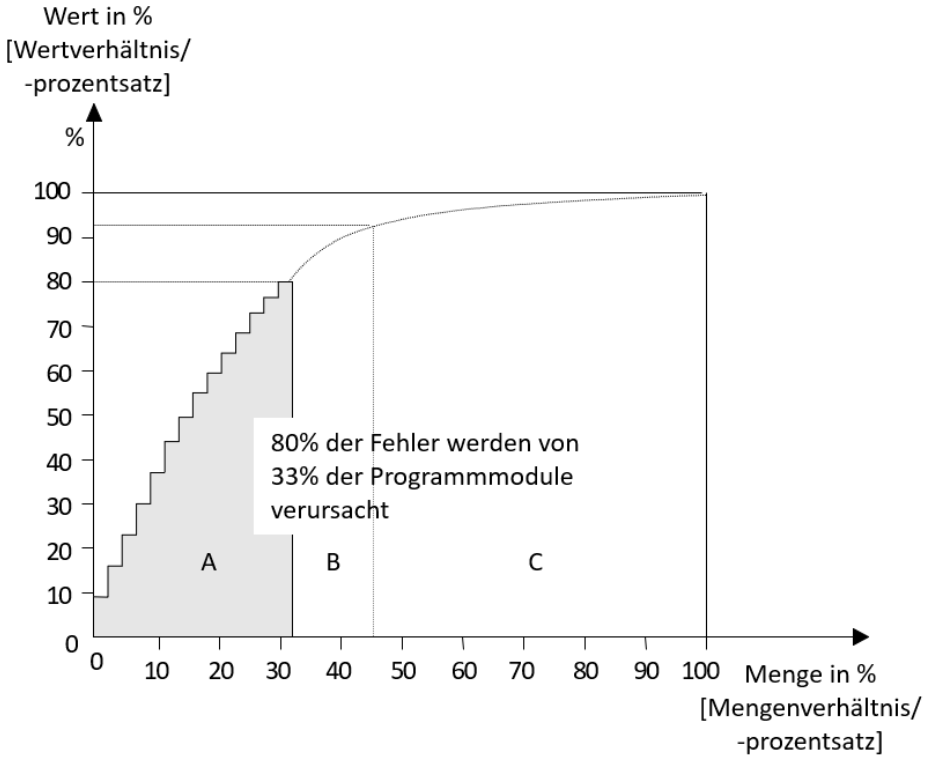


3.4 Ähnliche Anwendungen – die ABC-Analyse

Bedeutung der Klassen

- ***A-Objekte*** - Wenige Objekte haben einen großen Wertanteil.
- ***B-Objekte*** - Zahlreiche Objekte haben einen relativ kleinen Wertanteil.
- ***C-Objekte*** - Die größte Menge der Objekte hat einen sehr kleinen Wertanteil.





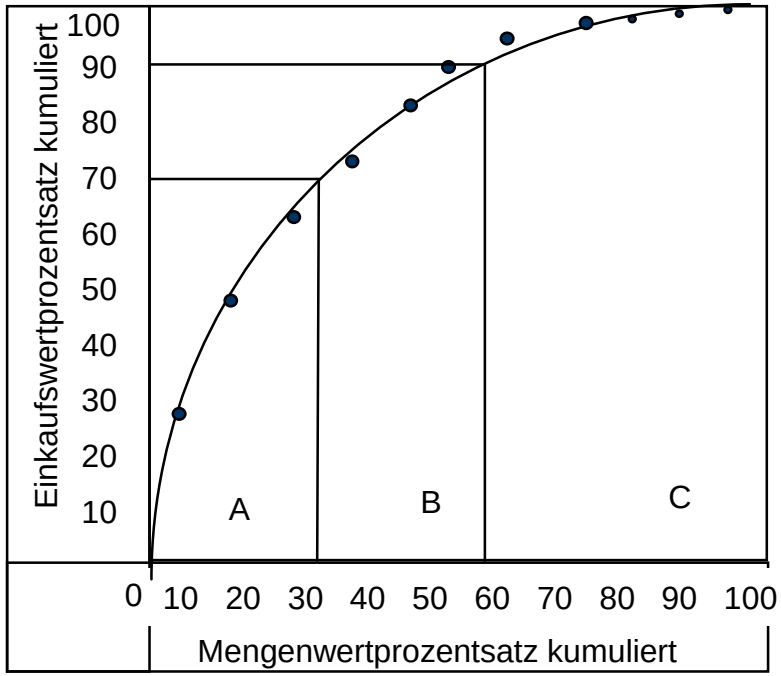
Quelle: Graf

3.4 Ähnliche Anwendungen – die ABC-Analyse

Beispiel einer ABC- Analyse

Material-art	Preis des Materials €/St	Menge St. / Jahr	Einkaufs-wert € / Jahr	Einkaufs-wertpro-zentsatz	Rangfolge
Scheibe	0,02	120 000	2 400,--	2,18	9
Spann-band	0,07	22 000	1 540,--	1,40	10
Klappe	0,09	15 000	1 350,--	1,22	11
Dichtung	0,15	30 000	4 500,--	4,08	6
Schelle	0,20	6 000	1 200,--	1,08	12
Halter	1,60	2 000	3 200,--	2,90	8
Gleitring	3,00	3 500	10 500,--	9,52	5
Zahnrad	20,00	750	15 000,--	13,60	3
Welle	48,00	500	24 000,--	21,76	2
Lager-bock	98,00	300	29 400,--	26,66	1
Kupplung	600,00	22	13 200,--	11,97	4
Antrieb	2 000,00	2	4 000,--	3,63	7
			110 290,--	100,00	

Rangfolge	Material-art	Einkaufs-wertpro-zentsatz	Einkaufs-wertpro-zentsatz kummuliert	Mengen-prozent-satz	Mengen-prozent-satz kummuliert	ABC Klassifi-zierung
1	Lagerbock	26,66	26,66	8,33	8,33	
2	Welle	21,76	48,42	8,33	16,67	A-Teile
3	Zahnrad	13,60	62,02	8,33	25,00	
4	Kupplung	11,97	73,99	8,33	33,33	
5	Gleitring	9,52	83,51	8,33	41,67	B-Teile
6	Dichtung	4,08	87,59	8,33	50,00	
7	Antrieb	3,63	91,22	8,33	58,33	
8	Halter	2,90	94,12	8,33	66,67	
9	Scheibe	2,18	96,30	8,33	75,00	C-Teile
10	Spannrad	1,40	97,70	8,33	83,33	
11	Kappe	1,22	98,92	8,33	91,67	
12	Schelle	1,08	100,00	8,33	100,00	



Siehe auch hier die Analogie zur Pareto-Regel

Quelle: Graf

3.4 Konzentrationsmessung über den Gini-Koeffizient

Konzentrationsmaße beschreiben das Konzentrationsausmaß durch einen einzigen Wert. Dadurch können z.B. Konzentrationsprozesse im Zeitablauf einfacher beurteilt und/oder Vergleiche mit anderen Gesamtheiten leichter durchgeführt werden. Das bekannteste Konzentrationsmaß ist der ***Gini-Koeffizient GK***, der folg Erkenntnis ausnutzt:

Je größer die Fläche F, desto größer die Konzentration;
je kleiner die Fläche F, desto kleiner die Konzentration.

Die Fläche zwischen der Geraden (0,0) und (1,1) und der Lorenzkurve bezeichnet man als **Konzentrationsfläche**. Je größer die Konzentrationsfläche ist, desto größer ist auch die Konzentration. Bei völlig gleicher Verteilung der Merkmalssumme ist die Konzentrationsfläche gleich 0, bei maximaler Konzentration gleich ½. Über mathematische Umformungen lässt sich aus der Division der Fläche in Relation zur Gesamtfläche der Gini-Koeffizient gewinnen, der im Intervall von [0, 1) liegt.

Der Quotient aus der Konzentrationsfläche und der Fläche bei maximaler Konzentration heißt ***Gini-Koeffizient***.

$$G = 1 - \sum_{i=1}^{k} f_i \left(S_i + S_{i-1} \right) \text{ mit } S_0 = 0 \text{ und } 0 \le G < 1 \text{ bzw } 0 \le G \le \frac{n}{n-1}$$

Interpretation des Gini-Koeffizienten G: Je näher der Gini-Koeffizient gegen Null geht, desto geringer ist die Konzentration; je näher der Gini-Koeffizient gegen 1 geht, desto größer ist die Konzentration.

3.4 Konzentrationsmessung über den Gini-Koeffizient

Die relative Konzentrationsmessung wird am bereits besprochenen Beispiel einer klassifizierten Häufigkeitsverteilung erklärt. Die Ausführungen können auf nichtklassifizierte Verteilungen übertragen werden, indem die Klassenmitten durch die Merkmalswerte x_i ersetzt werden

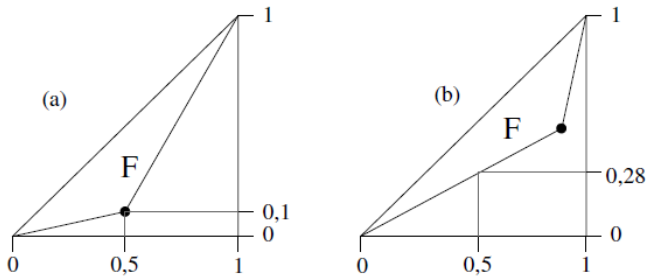
Beispiel: 5000 Lagerpositionen: Welcher Anteil des gesamten Lagerwertes entfällt auf welchen Anteil der Lagerpositionen?

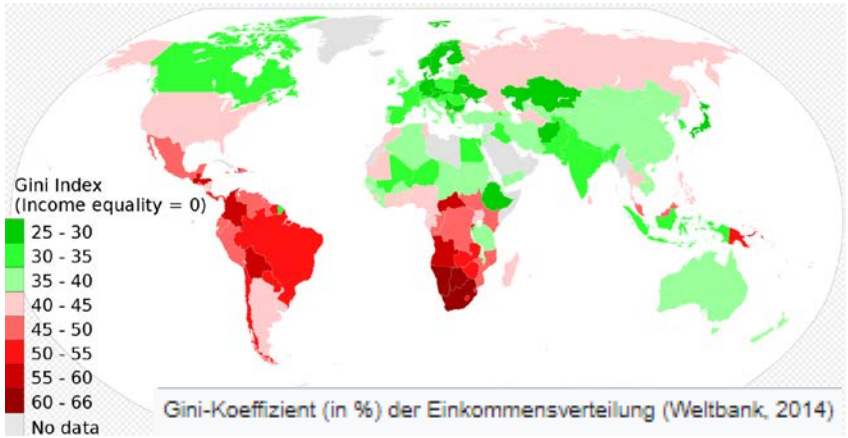
Lagerwert in Tsd. €					1. rel. Summenhäufigkeit			2. rel. Summenhäufigkeit	
von	bis unter	x_i*	h_i	H_i	f_i	F_i	x_i* h_i	s_i	S_i
1	5	3	2.000	2.000	40%	40%	6.000	6%	6%
5	15	10	1.200	3.200	24%	64%	12.000	12%	18%
15	25	20	800	4.000	16%	80%	16.000	16%	34%
25	50	37,5	700	4.700	14%	94%	26.250	26%	60%
50	100	75	200	4.900	4%	98%	15.000	15%	75%
100	395	247,5	100	5.000	2%	100%	24.750	25%	100%
			5.000	5.000	100%		100.000	100%	

Quelle: z.T. Bourier: Statistik 2018

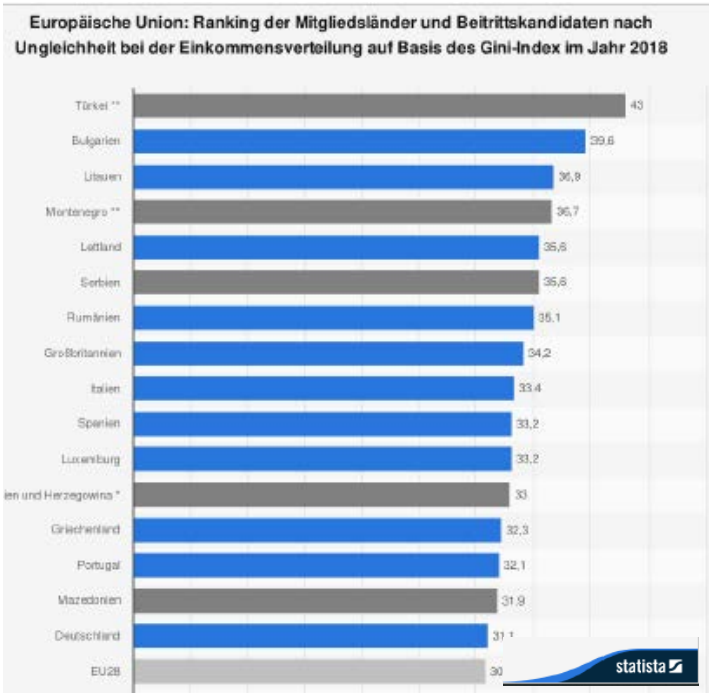
3.4 Konzentrationsmessung über den Gini-Koeffizient

Deutlich unterschiedliche Verteilungen der Merkmalswertsumme auf die Merkmalsträger können zu demselben oder fast demselben Gini-Koeffizienten führen. Hierin liegt ein Nachteil des Gini-Koeffizienten. In den nachfolgend gegenübergestellten Lorenzkurven ist die Merkmalswertsumme unterschiedlich verteilt. So entfallen etwa auf 50% der Merkmalsträger im Fall a) 10% und im Fall b) zirka 28% der Merkmalswertsumme. Dennoch besitzen beide Verteilungen mit 0,30 denselben Gini-Koeffizienten. Deswegen empfiehlt sich die begleitende Betrachtung der Lorenzkurve.





Gini-Koeffizient (in %) der Einkommensverteilung (Weltbank, 2014)



Quelle: z.T. Bourierr: Statistik 2018, Statista und Wikipedia

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Gliederung

1. **Einführung und Aufgaben**
   Data Science und Statistik**,** deskriptive Statistik, Wahrscheinlichkeitsrechnung, induktive Statistik, Vorgehensweisen und Beispiele

2. **Grundbegriffe, statistische Untersuchung und Darstellung der Daten**
   Datenquellen, Merkmale, Grundgesamtheit, Stichprobe, Messskalen, Vorgehensweise bei statistischen Untersuchungen Planung, Datenerhebung und -techniken, Datenaufbereitung, tabellarische und grafische Darstellung, Interpretation

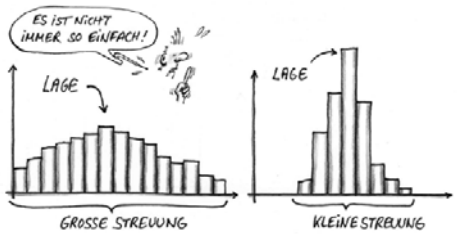
3. **Eindimensionale Häufigkeitsverteilungen**

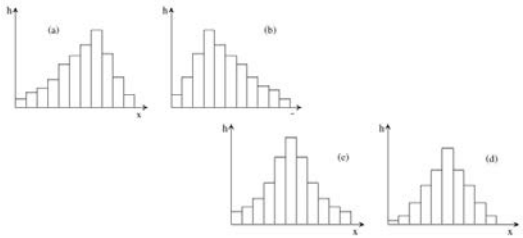
   3.1 Terminologie und grafische Darstellungen
   3.2 Lageparameter
   3.3 Streuungsparameter
   3.4 Schiefe und Konzentration

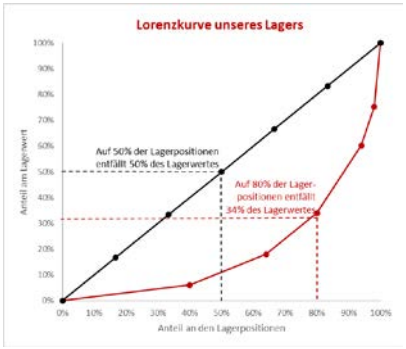
Was sollten Sie in diesem Kapitel mitgenommen haben?

- Tieferes Verständnis der eindimensionaler Häufigkeitsverteilungen und deren grafischer Darstellung
- Verstehen, wie Lage- und Streuungsparameter eine eindimensionale Häufigkeitsverteilung charakterisieren und deren Anwendbarkeit einschätzen können.
- Verstehen, welcher Mittelwert (Modus, Median, Quantile sowie arithmetisches, harmonisches und geometrisches Mittel) sich für welche Fragestellung eignet sowie die Anwendung und Berechnung für eine vorgegebene Fragestellung.

»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«







- Ein Verständnis für die Streuungsparameter (Spannweite, Quantilsabstände, mittlere absolute Abweichung, Varianz, Standardabweichung und Variationskoeffizient) eindimensionaler Häufigkeitsverteilungen sowie die Anwendung und Berechnung der unterschiedlichen Streuungsparameter.
- Die (A)Symmetrie eindimensionaler Häufigkeitsverteilungen grafisch und methodisch erkennen sowie Anwendung und Berechnung von Schiefe, Momente und Wölbung.
- Einsatzfelder der Konzentration aufzeigen können sowie Anwendung und Ermittlung von Lorenzkurve und Gini-Koeffizient.

Quelle: Grafik: Fernuni Hagen, Statistik