

Was versteht man in der Umgangssprache unter einer "Statistik"?
Eine „Statistik“ ist eine systematische Zusammenstellung von Zahlen und Daten zur Beschreibung von Zuständen, Entwicklungen und Phänomenen.
Beispiele: Häufigkeitsverteilungen, Zeitreihenvergleich und -analyse, Zusammenhangs- und Abhängigkeitsanalysen, statistische Kennzahlen zur Beschreibung von Verteilungen, Zusammenhängen.
Worin unterscheiden sich „beschreibende“ und „schließende Statistik“
Im Rahmen der „**Beschreibende Statistik**“ sammelt man Daten bei **allen** Untersuchungseinheiten, über die man Informationen erhalten will. Die **beschreibende Statistik** hat zum Ziel, empirische Daten durch Tabellen, Kennzahlen (auch: Maßzahlen oder Parameter) und Grafiken übersichtlich darzustellen und so ordnen. Dies ist vor allem bei umfangreichem Datenmaterial sinnvoll, da dieses nicht leicht überblickt werden kann. In der **schließenden Statistik** wählt man aus der Grundgesamtheit, über die man Informationen haben will, eine **Teilmenge = Stichprobe** aus. (Grund für diese Vorgehensweise ist meistens die Größe der Grundgesamtheit). Die **schließende Statistik** wird zu einem wesentlichen Teil zum Beweis oder zur Widerlegung von vorher aufgestellten Behauptungen, den Hypothesen, die sich auf definierte Grundgesamtheit bezieht, eingesetzt. Nur bei den Einheiten der Stichprobe erhebt man Daten, die man dann mit statistischen Methoden auswertet. Von den Stichprobenergebnissen versucht man, auf die Eigenschaften der Grundgesamtheit zu schließen.
Was ist die wichtigste (in der Praxis aber sehr oft nicht gegebene) Voraussetzung für eine "Random-Auswahl"?
Siehe c): Liste/Datel aller Einheiten der Grundgesamtheit

Bei einer Teilerhebung muss man Entscheidungen fällen über den Stichprobenumfang und das Auswahlverfahren. Welche Entscheidung ist wichtiger?
Die Entscheidung über das Auswahlverfahren ist wesentlich wichtiger. Eine Stichprobe „schlecht“ ausgewählt kann noch so groß sein, sie liefert keine brauchbaren Erkenntnisse über die Grundgesamtheit. So ist unmittelbar einleuchtend, dass man bei der Grundgesamtheit „Einwohner in Deutschland“ (rund 82 Mio.) und der daraus gezogenen Stichprobe „Studierende in Deutschland“ (rund 2 Mio.) zwar einen sehr großen Stichprobenumfang hat, dass man aber nicht von der Stichprobe auf die Grundgesamtheit schließen kann. Diese Stichprobe ist „verzerrt“ oder wie man oft auch sagt „nicht repräsentativ“.
c) Überlegen und Beschreiben Sie den Unterschied zwischen einer „willkürlichen Auswahl“ und einer „zufälligen Auswahl“.
Bei einer **„willkürlichen Auswahl“** (= Auswahl aufs Geratewohl oder „convenience sample“) gibt es keinen Auswahlplan. Die Interviewer sind frei in der Auswahl ihrer Interviewpartner. Daher suchen sie sich die Personen aus, die für sie am bequemsten zu erreichen sind. Das führt meist zu einer verzerrten Stichprobe. Bei einer **„zufälligen Auswahl“** ist die Auswahl zufalls gesteuert, d.h., jede Einheit der Grundgesamtheit (über die man Informationen erhalten will) muss mit gleicher Wahrscheinlichkeit in die Stichprobe gelangen können. Dies setzt voraus, dass eine Liste/Datel aller Einheiten der Grundgesamtheit vorliegt. Die Interviewer sind nicht frei in der Auswahl ihrer Interviewpartner. Sie bekommen feste Zielpersonen vorgegeben. Nur bei der Zufallsauswahl (=Random-Auswahl) lässt sich mit Hilfe der Wahrscheinlichkeitsrechnung der Stichprobenfehler berechnen.

Beschreiben Sie die "Quota-Auswahl".
Die **„Quota-Auswahl“** ist ein „bewusstes Auswahlverfahren“. Über für die Interviewer verbindliche Quotenpläne wird erzwungen, dass für ausgewählte Quotierungsmerkmale (meist Geschlecht, Alter, Beruf) in der Stichprobe eine gleiche Verteilung wie in der Grundgesamtheit erreicht wird. Voraussetzung ist, dass man die Verteilung der Quotierungsmerkmale in der Grundgesamtheit kennt. Amtliche Statistiken liefern hier in vielen Fällen die gewünschten Verteilungen.
Welche Art von Stichprobe ergibt sich bei einer so genannten TED-Umfrage im Fernsehen? (Bei der TED-Umfrage werden Fernsehzuschauer aufgefordert Fragen zu aktuellen Themen zu beantworten. Jeder der Antwortmöglichkeiten ist eine Telefonnummer zugeordnet, die dann – je nach persönlicher Meinung - gewählt werden soll.)
Bei der TED-Umfrage im Fernsehen liegt eine **willkürliche Auswahl** (der Bewohner eines Landes) vor. Es gibt keinerlei Auswahlplan, jeder kann, wenn er Lust hat sich an der Umfrage beteiligen. Es gibt Personen, die grundsätzlich an solchen Umfrage niemals teilnehmen würden, andere versuchen mehrmals ihre Meinung zu äußern.
Geben Sie für die folgenden Merkmale jeweils an, welcher Merkmalstyp vorliegt und auf welcher der 5 Skalen das Merkmal gemessen wird.
Geschwindigkeit e. Fahrzeugs = quantitatives Merkmal, Verhältnisskala
Rechtsform e. UN = qualitatives Merkmal, Nominalskala
Preis e. Produktes = quantitatives Merkmal, Verhältnisskala
Umsatzklasse des UN = Rangmerkmal, Rangskala
Wohnort = qualitatives Merkmal, Nominalskala
Mitarbeiterzahl = quantitatives Merkmal, Absolutskala
Kundenzufriedenheit = Rangmerkmal, Rangskala / Ratingskala
Umsatz eines UN = quantitatives Merkmal, Verhältnisskala
Fachsemesterzahl = quantitatives Merkmal, Absolutskala

Steuerklasse = qualitatives Merkmal, Nominalskala
Einkommensklasse = Rangmerkmal, Rangskala
Geburtsjahr = quantitatives Merkmal, Intervallskala
Nehmen Sie zum Schaubild (Abbildung) kritisch die Stellung. Überlegen Sie was darf bei einer Tabelle und einem Diagramm auf keinen Fall fehlen?

1995 - 2000 hat sich der Umsatz des Produktes ABC rasant entwickelt ...

a) Bei der Abbildung fehlt die Maßeinheit, die Quellenangabe,
b) Aussage in Überschrift übertrieben. Jährlicher Anstieg nur etwa 1 %.
Anstieg des Umsatzes über die Jahre bei ca 5,5 %. Skalierung der Ordinate Y achse täuscht, einen höheren Anstieg vor. Grund: Umsatz-Achse beginnt bei 700 nicht bei 0.

Im Folgenden ist die Zielsetzung verschiedener statistischer (= empirischer) Untersuchungen beschrieben. Geben Sie jeweils an, wer oder was die Untersuchungseinheiten sind oder sein könnten und wie Sie die Grundgesamtheit unter sachlichen, räumlichen und zeitlichen Aspekten abgrenzen würden. Bitte beachten Sie, dass in den Aufgaben nicht zu allen Punkten Hinweise vorhanden sind. Hier geht es um die Operationalisierung (OP) der Aufgabe.
A) Hersteller von Schokolade möchte Infos über Verbrauchsgewohnheiten von Jugendlichen in Süddeutschland haben:
1. Potentielle Untersuchungseinheit: Menschen.
2. Jugendliche: Möglich 10-18 Jahre. (Definition nicht eindeutig)
3. Räumliche Abgrenzung: Süddeutschland (OP, wo genau Bundesland?)
4. Zeitliche Abgrenzung: keine Infos, wahrscheinlich Jahresende
5. Glühbirnenfabrik untersucht bei Qualitätskontrolle Brenndauer und Funktionstüchtigkeit von Glühbirnen
1. Potentielle Untersuchungseinheit: Glühbirnen
2. Sachliche Abgrenzung: Neu produzierte Glühbirnen in dem UN.
3. Räumliche Abgrenzung: No Infos, Beschränkung auf Hallen/Maschine
4. Zeitliche Abgrenzung: No Infos, normal zeitlicher Intervall
5. Teilerhebung: Brenndauer, durch Stichprobe
6. Vollerhebung: Funktionstüchtigkeit
UN in Versandhandelsbranche stellt fest, dass aus neuen Bundesländern viele Beschwerden kommen. Welche Produkte besonders betroffen?
Potentielle Untersuchungseinheiten: Reklamationen.
Sachliche Abgrenzung: schriftliche Reklamationen, die eingegangen sind.
Räumliche Abgrenzung: aus den „neuen Bundesländern“ (OP eindeutig).
Zeitliche Abgrenzung: 1. – 31. 12. 2006

Hausbrauerei Tauffenbach Bochum veranstaltet im Sommer jeden Sonntagvormittag einen „Frühschoppen mit Jazz-Musik“. Betreiber will wissen, ob Gäste zufrieden mit Musik & Angebot Getränken / Speisen
1. Potentielle Untersuchungseinheiten: Menschen.
2. Sachliche Abgrenzung: Gäste der Hausbrauerei Tauffenbach.
3. Räumliche Abgrenzung: z.B. die Räumlichkeiten der Hausbrauerei.
4. Zeitliche Abgrenzung: Sonntagvormittage im Sommer, an denen der „Frühschoppen mit Jazz“ stattfindet.
Eine Einzelhandelskette will die räumliche Anordnung des Warensortiments „optimieren“. Dazu will das UN untersuchen lassen, ob und in welchem Umfang es bei den Einkäufen ihrer Kunden so genannte „Verbundeffekte“ gibt, d.h., bestimmte Produkte aus dem Sortiment häufig zusammengekauft werden (z.B. Kaffee, Filtertüten und Gebäck).
1. Potentielle Untersuchungseinheiten: Kassensbons / Warenkörbe
2. Sachliche Abgrenzung: Kassensbons der Einzelhandelskette
3. Räumliche Abgrenzung: Keine Infos, Geschäfte einz. Regionen / Filialen
4. Zeitliche Abgrenzung: no infos, Sinnv: Beschränkt. Auf best. Jahreszeit
Überlegen Sie bitte, was bei der Einteilung der Umsatzklassen in Tabelle 1 und bei dem Diagramm in Abbildung 1 (Modul 1 Folien 13 und 14), bis unter („Abkürzung „b.u.“) bedeutet. Warum schreibt man statt „bis unter“ nicht einfach „bis“, z.B. „100 bis 200 Tsd. €?“
In mathematischer Schreibweise bedeutet die Umsatzklasse „100 bis unter 200 Tsd. €“, dass alle Großhändler mit einem Umsatz x, mit 100 >= x < 200 Tsd. €, in diese Klasse fallen. D.h. wenn der Umsatz genau 200 Tsd. € beträgt, dann fällt er in die Umsatzklasse, wenn der Umsatz aber genau 200 Tsd. € beträgt, dann fällt er nicht mehr in diese Klasse, sondern in die nächste Klasse. Liegt der Umsatz allerdings ganz knapp unter 200 Tsd. €, z.B. bei 199 999,99 €, dann liegt er in der Klasse „100 bis unter 200 Tsd. €“.

Wenn man wie folgt klassieren würde: „100 bis 200 Tsd. €“, „200 bis 300 Tsd. €“ usw., dann könnte ein Umsatz von genau 200 Tsd. € nicht eindeutig einer Umsatzklasse zugeordnet werden. Eine eindeutige Zuordnung zu genau einer Klasse ist aber bei der Klassierung von Daten unbedingt notwendig.
Welche zwei Probleme hat man bei der Klassierung von Daten?
Übersichtlichkeit – Informationsverlust für klassierte Daten können keine exakten statistischen Kennzahlen (z.B. Mittelwerte) berechnet werden. Näherungswerte können nur unter bestimmten Annahmen (z.B. Gleichverteilung in den Klassen) berechnet werden.
Was versteht man unter „offenen Randklassen“?
Bei „**offenen Randklassen**“ ist entweder die Klassenuntergrenze nicht angegeben (untere offene Randklasse, z.B. „b.u. 100 kg“) oder die Klassenobergrenze (obere offene Randklasse, z.B. „200 kg und schwerer“). Bei offenen Randklassen kann die Klassenmitte und die Klassenbreite nicht berechnet werden.
Bestimmen Sie für die Klasse „150 b.u. 180 cm“ die Klassenbreite und die Klassenmitte.
Klassenbreite = 180 – 150 = 30 cm
Klassenmitte = (150 + 180)/2 = 165 cm.
Von 2006 bis 2007 ist der Umsatz eines Unternehmens um 50% gesunken. Im darauffolgenden Jahr 2008 war das Unternehmen erfolgreicher: der Umsatz stieg von 2007 bis 2008 um 70%. Welche der folgenden Aussagen ist richtig? (kurze Begründung durch Rechnung) Der Umsatz ist im Zeitraum von 2006 bis 2008
a) um 10 % gestiegen - Falsch
b) um 15 % gestiegen - falsch
c) um 15 % gesunken – richtig Bsp: 100/2 = 50, 50/100*70 = 35. 50 + 35 = 85. 100 + x = 85, x = - 15

Für 200 Unternehmen liegt für das Jahr 2006 die folgende Umsatzverteilung vor. Vervollständigen Sie die klassierte Häufigkeitsverteilung.
a) Bestimmen Sie den Modus, den Median und das arithmetische Mittel
b) Bestimmen Sie die Streuungsparameter Spannweite w, Varianz s2, Standardabweichung s, Variationskoeffizient v.

Lösung der Aufgabe 6

Klasse Nr. i	Umsatzklasse (Mio €)	Anzahl Unternehmen hi	Anteil fi (%)	Hi	Fi (%)	Klassenmitte mi
1	0 b.u. 1	60	30%	60	30%	0,5
2	1 b.u. 2	80	40%	140	70%	1,5
3	2 b.u. 5	40	20%	180	90%	3,5
4	5 b.u. 10	10	5%	190	95%	7,5
5	10 b.u. 20	10	5%	200	100%	15
Σ		200	100%			

a) $\bar{x} = 1,5$ Mio € (Mitte der Klasse 2 mit größtem hi)
Einfallsklasse k = 2
(s. Bsp. 2 Folie 15 Modul 4) $\bar{x} = 1 + (2 - 1) \cdot 0,5 = 1,5$ (Mio €)
 $\bar{x} = \sum mi \cdot fi = 1 \cdot 0,3 + 1,5 \cdot 0,4 + 3,5 \cdot 0,2 + 7,5 \cdot 0,05 + 15 \cdot 0,05 = 1,5$ (Mio €)
 $s^2 = \sum mi^2 \cdot fi - \bar{x}^2 = 1^2 \cdot 0,3 + 1,5^2 \cdot 0,4 + 3,5^2 \cdot 0,2 + 7,5^2 \cdot 0,05 + 15^2 \cdot 0,05 - 1,5^2 = 1,28$
b) $w = 20 - 0 = 20$

Aufgabe 3
Im Rahmen einer Marktforschungsstudie wurden n = 12 Personen u.a. gefragt nach den drei Merkmalen Geschlecht G (w = weiblich, m = männlich), Alter A (Alter in Jahren) und Markenpräferenz M (A = Produkt A, B = Produkt B). Die Erhebung ergab die folgenden 12 Befragungsergebnisse (Beobachtungswertkombinationen): (w, 37, A), (m, 65, A), (w, 26, A), (m, 37, B), (w, 21, B), (m, 29, A), (w, 52, B), (m, 43, A), (w, 48, A), (m, 58, B), (w, 24, A), (m, 58, B).
Lesepbeispiel: Die 1. Person ist weiblich, 37 Jahre alt und bevorzugt Produkt A.
a) Erstellen Sie die zwei folgenden zweidimensionalen Kreuztabellen: Geschlecht x Markenpräferenz klassiertes Alter x Markenpräferenz (2 Altersklassen: 1. Klasse: bis unter 40 Jahre, 2. Klasse: 40 Jahre und älter)
b) Versuchen Sie eine dreidimensionale Kreuztabelle zu erstellen für die drei Merkmale: Geschlecht x klassiertes Alter x Markenpräferenz
c) Welche der drei Merkmale kann man als unabhängige bzw. abhängige Merkmale betrachten?
d) Sie erheben bei einer Grundgesamtheit Daten für 3 Merkmale. Was liefert mehr Information: die dreidimensionale Häufigkeitsverteilung oder alle drei möglichen ein- und zweidimensionalen Häufigkeitsverteilungen zusammen?

M / A	bis unter 40 Jahre alt	40 Jahre und älter	Σ
Produkt A	4	3	7 58,3 %
Produkt B	2	3	5 41,7 %
Σ	6 50%	6 50%	12 100%

b)

M / G A	weiblich		männlich		Σ
	bis unter 40 Jahre alt	40 Jahre und älter	bis unter 40 Jahre alt	40 Jahre und älter	
Produkt A	3	1	1	2	7 58,3 %
Produkt B	1	1	1	2	5 41,7 %
Σ	4 33,3%	2 16,7%	2 16,7%	4 33,3%	12 100 %

c) Geschlecht G und Alter A sind **unabhängige** Merkmale und Markenwahl ist das **abhängige Merkmal**. Die Antwort ist hier eindeutig. Denn Geschlecht und Alter können nicht vom Merkmal Markenwahl abhängige Merkmale sein; die Markenwahl kann das Geschlecht bzw. das Alter **nicht** beeinflussen. Es ist aber sinnvoll zu untersuchen, ob die Markenwahl bei Frauen und Männern unterschiedlich oder gleich ist bzw. ob die Markenwahl bei einer Produktgruppe altersabhängig ist. Denn es ist möglich, dass Frauen und Männer ein unterschiedliches Kaufverhalten haben. Genauso ist möglich bzw. sogar wahrscheinlich, dass jüngere Konsumenten andere Marken wählen als ältere Konsumenten.
d) Die dreidimensionale Häufigkeitsverteilung liefert mehr Information als alle möglichen ein- und zweidimensionalen Häufigkeitsverteilungen zusammen. Nur aus ihr kann man erfahren, welchen Zusammenhang es zwischen **allen drei Merkmalen** gibt.
Ein Merkmal X mit m Merkmalsausprägungen (x1, ..., xm) wird bei n Untersuchungseinheiten gemessen. Beantworten Sie für die entsprechende Häufigkeitsverteilung die folgenden Fragen:
a) Wie groß ist $\sum h(x_{mi}=1)$ Lösung: $\sum h(x_{mi}=1) = n$
b) Wie groß ist $\sum f(x_{mi}=1)$ Lösung: $\sum f(x_{mi}=1) = 1$ bzw. 100%
c) Welchen Wert hat H(xm), F(xm)? Lös. H(xm)=n, F(xm)=1 bzw. 100%

Vervollständigen Sie die folgende Häufigkeitstabelle (relative Häufigkeiten sind ganz Zahlen!):

Lösung der Aufgabe 5

i	Merkmalsausprägung xi	h(xi)	f(xi) (%)	H(xi)	F(xi) (%)
1	3	240	15%	240	15%
2	4	400	25%	640	40%
3	5	320	20%	960	60%
4	6	640	40%	1.600	100%
Σ		1.600	100%	-	-

Vervollständigen Sie die folgende klassierte Häufigkeitstabelle (relative Häufigkeiten mit einer Nachkommastelle!)

Lösung der Aufgabe 6

Klasse Nr. i	Umsatzklasse (Mio €)	Anzahl Filialen hi	Anteil fi (%)	Hi	Fi (%)
1	0 b.u. 40	50	16,7%	50	16,7%
2	40 b.u. 100	150	50,0%	200	66,7%
3	100 b.u. 200	50	16,7%	250	83,4%
4	200 b.u. 400	50	16,7%	300	100,1%
Σ		300	100,1%	-	-

Aufgabe 9 (Klausuraufgabe WS18/19 mit 9 Punkten):
Um die Entwicklung der Telefonkosten der letzten 6 Monate des vergangenen Jahres zu analysieren, wird Claudia von ihrem Vater beauftragt, die mittleren Telefonkosten sowie deren Streuung zu berechnen. Die Telefonkosten (in €) sind in der folgenden Tabelle aufgeführt. Berechnung mit 2 Nachkommastellen!

Monat	Juli	August	September	Oktober	November	Dezember
Kosten (€)	31,44	30,18	31,04	33,60	38,16	132,40

a) (5 Punkte) Berechnen Sie das arithmetische Mittel sowie die Streuung der Telefonkosten.
 $\bar{x} = 16 + (31,44 + 30,18 + 31,04 + 33,6 + 38,16 + 132,4) = 296,826 = 49,47$
 $s^2 = 16 + (31,442 + 30,182 + 31,042 + 33,62 + 38,162 + 132,42) - 49,472 = 22977,69 - 2447,28 = 1382,34$
 $s = \sqrt{s^2} = \sqrt{1382,34} = 37,18$

b) (4 Punkte) Claudia, die im Dezember häufig bei teuren Hotlines angerufen hat, ist entsetzt über den hohen Mittelwert und befürchtet Taschengeldentzug durch ihren Vater. Helfen Sie Claudia aus der Patsche, indem Sie ein alternatives Lageparameter, zu Claudias Gunsten, vorschlagen. Begründen Sie Ihren Vorschlag kurz (maximal drei Sätze) und berechnen Sie den Wert Ihres vorgeschlagenen Lagemaßes.
b) (4 Punkte) „Bessere“ Vorschlag: Median – weniger empfindlich gegenüber Ausreißern. n gerade, Reihe erst sortieren!
 $\bar{x} = 12(31,44 + 33,60) = 32,52$

Im Rahmen einer Bürgerbefragung wurden n = 1.200 Personen u.a. gefragt nach ihrem Alter (A) und ihrer Meinung zur Ausweitung der Fußgängerzone in der Innenstadt (F). Das Merkmal Alter wurde wie folgt klassiert: „bis unter 40 Jahre“ und „40 Jahre und älter“. Auf die Frage nach der Ausweitung der Fußgängerzone konnte nur mit „ja“ oder „nein“ geantwortet werden. Die Befragungsergebnisse wurden ausgezählt. Es ergaben sich die in der folgenden Kreuztabelle aufgeführten absoluten Häufigkeiten für die Merkmalsausprägungskombi.

a) Tragen Sie in die obige Kreuztabelle ein (absolut und relativ in % mit Nachkommastelle):

(1) die beiden Randverteilungen
(2) die relativen Spaltenhäufigkeiten in %
(3) die relativen Zeilenhäufigkeiten in %
(4) die relativen Häufigkeiten der Merkmalsausprägungskombi. in %

Lösung der Aufgabe 7

Auswertung der Fußgängerzone	F / A → S	A		Σ
		bis unter 40 Jahre alt	40 Jahre und älter	
ja		396 (2) 60,0% (3) 66,0% (4) 33,0%	204 (2) 37,8% (3) 34,0% (4) 17,0%	600 50,0%
	nein	264 (2) 40,0% (3) 44,0% (4) 22,0%	196 (2) 62,2% (3) 56,0% (4) 28,0%	460 50,0%
Σ		660 55,0%	540 45,0%	1.200 100,0%

b) Beantworten Sie mit Hilfe der obigen Kreuztabelle die folgenden Fragen:

(1) Wie viel % der Befragten, die die Ausweitung der Fußgängerzone befürworten, sind 40 Jahre und älter? **34,0 %**.

(2) Wie viel % der jüngeren Befragten (bis unter 40 Jahre), befürworten die Ausweitung der Fußgängerzone? **60,0 %**.

(3) Wie viel % der Befragten sind bis unter 40 Jahre alt und befürworten die Ausweitung der Fußgängerzone? **33,0 %**.

(4) Wie viel % der Befragten, die 40 Jahre und älter sind, sind gegen die Ausweitung der Fußgängerzone? **62,2 %**.

a) Sie lesen in einer Studie über die Altersverteilung in einer Gruppe, dass $\bar{X} = 32$ Jahre und $s = 40$ Jahre ist. Welche Schlüsse können Sie daraus über die Altersverteilung ziehen?

b) Sie lesen in einer Studie über die Einkommensverteilung einer Berufsgruppe: $\bar{X} = 30.000$ €, $s = 40.000$ €, $Q_1 = 25.000$ €, $Q_3 = 45.000$ €. Welche Informationen erhalten Sie aus diesen 4 statistischen Kennzahlen über die Einkommensverteilung? Erhalten Sie auch Informationen über die Streuung der Verteilung? **LÖSUNG:**

a) Wegen $\bar{X} = 32 < 40 = s$, liegt hier eine rechtsschiefe Altersverteilung vor. Es gibt Ausreißer im oberen Altersbereich, die das arithmetische Mittel nach oben ziehen, den Median aber nicht beeinflussen.

b) Die Einkommensverteilung ist rechtsschief, da $\bar{X} < s$. Es gibt Ausreißer im oberen Einkommensbereich. 25% der Personen verdienen weniger und 75% mehr als 25.000 €. 75% verdienen weniger und 25% mehr als 45.000 €. Ein Einkommen von 30.000 € wird von 50% über- und von den anderen 50% unterschritten. Der Interquartilsabstand IQR = $Q_3 - Q_1 = 20.000$ € gibt die Spannweite bei den mittleren 50% an und informiert somit über die Streuung des "mittleren" Teils der Einkommensverteilung.

Aufgabe 7 (Klausuraufgabe WS17/18 mit 18 Punkten):
In der folgenden Tabelle ist die Verteilung der männlichen Teilnehmer bei einer Umfrage auf Altersklassen dargestellt. Dabei wurde zwischen Personen mit und ohne Migrationshintergrund unterschieden.

a) Berechnen Sie approximativ die Altersquartile (Q1, Q2, Q3) für die beiden Gruppen. Alle Quartile auf ganze Zahlen runden. TIPP: vervollständigen Sie zuerst die Häufigkeitstabelle.

b) Zeichnen Sie die Boxplots für beide Gruppen in einem Diagramm. Bitte denken Sie an die „Lesbarkeit“ des Boxplots (Achsenbeschriftung und Legende nicht vergessen!)

c) Wie groß ist der Anteil der Personen im Alter zwischen 15 und 75 Jahre für jede Gruppe? (gemeint ist das Intervall [15;75])

Lösung der Aufgabe 7

Klassen-Nr.	Alter (von – b.u.) in Jahren	mit Migrationshintergrund f(x) in %	ohne Migrationshintergrund f(x) in %	mit Migrationshintergrund kumm. F(x) in %	ohne Migrationshintergrund kumm. F(x) in %
1	b.u. 15	22	12	22	12
2	15 – 35	31	23	53	35
3	35 – 55	29	33	82	68
4	55 – 75	15	25	97	93
5	75 und älter	3	7	100	100

a) Median = Q2 → Median ist auch Quartil (s. Folie 38 Modul 4).
Berechnung von Q1, Q2, Q3 wie Median für klassierte Daten! (s. Folie 13, 15, 16 Modul 4)

$\bar{X} = \bar{x}_k - 1 + (\bar{x}_k - \bar{x}_{k-1}) \cdot 0,5 - \frac{F_k - 1}{f_k} \cdot 1$ mit Migrationshintergrund: für Q1 (25%) $k=2$, $Q_1M=15+(35-15) \cdot (0,25-0,22)/0,31=17$
für Q2 (50%) $k=2$, $Q_2M=15+(35-15) \cdot (0,5-0,22)/0,31=33$
für Q3 (75%) $k=3$, $Q_3M=35+(55-35) \cdot (0,75-0,53)/0,29=50$
ohne Migrationshintergrund: für Q1 (25%) $k=2$, $Q_1oM=15+(35-15) \cdot (0,25-0,12)/0,23=26$
für Q2 (50%) $k=3$, $Q_2oM=35+(55-35) \cdot (0,5-0,35)/0,33=44$
für Q3 (75%) $k=4$, $Q_3oM=55+(75-55) \cdot (0,75-0,68)/0,25=61$

b)

Boxplot:

c) Mit $M = 31 + 29 + 15 = 75\%$ (ODER: $97 - 22 = 75\%$) → Anteile Personen aus den Klassen 2, 3, 4 addieren ODER subtrahieren Anteil Personen der Klasse 1 von kumuliertem Anteil Personen der Klasse 4.
Ohne $M = 23 + 33 + 25 = 81\%$ (ODER: $93 - 12 = 81\%$)

Im Rahmen einer Marktforschungsstudie wurden n = 2.000 Personen u.a. gefragt nach den drei Merkmalen Geschlecht G, Alter A (klassiert in 2 Klassen: "b.u. 40 Jahre" und "40 Jahre und älter") und Markenpräferenz M (A = Produkt A, B = Produkt B). Die Befragungsergebnisse sind in der folgenden dreidimensionalen Häufigkeitstabelle zusammengefasst:

M / G A	weiblich		männlich		Σ
	bis unter 40 Jahre alt	40 Jahre und älter	bis unter 40 Jahre alt	40 Jahre und älter	
Produkt A	400	150	200	450	1200
Produkt B	100	450	200	50	800
Σ	500	600	400	500	2000

Beantworten Sie auf der Basis der obigen Tabelle die folgenden Fragen (mit 1 Nachkommastelle):

a) Wieviel % der jüngeren Frauen (bis unter 40 J) bevorzugen Produkt B?
Lösung $100 / 500 = 20,0 \%$

b) Wieviel % der älter. Männer (40 Jahre & älter) bevorzugen Produkt A?
 $450 / 500 = 90,0 \%$

c) Wie viel % der Frauen bevorzugen Produkt A?
 $550 / 1100 = 50,0 \%$

d) Wie viel % der A-Käufer sind Frauen?
 $550 / 1200 = 45,8 \%$

e) Wie viel % der Befragten sind Frauen, unter 40 Jahre & A-Käuferinnen?
 $400 / 2000 = 20,0 \%$

f) Wie viel % der Frauen, die 40 Jahre & älter sind, bevorzugen Produkt A?
 $150 / 600 = 25,0 \%$

g) Wie viel % der Befragten sind Männer?
 $900 / 2000 = 45 \%$

h) Welchen Wert hat $f(B | \text{b.u. 40 J.})$?
 $300 / 900 = 33,3 \%$

i) Wie viel Prozent der A-Käufer, die 40 Jahre und älter sind, sind Frauen?
 $1500 / 600 = 25,0 \%$

j) Wie viel % der Befragten sind A-Käufer, 40 Jahre und älter & Männer?
 $450 / 2000 = 22,5 \%$

Für die folgenden Umsätze (in Mio. €) ermittelt: 35, 150, 190, 20, 8, 74, 44, 89, 25, 12, 17, 5, 22, 10, 13, 150, 47, 65, 49, 55. Erstellen Sie für diese statistische Reihe in der folgenden Tabelle eine klassierte Häufigkeitsverteilung für die gegebenen Klassen (mit absoluten und relativen in % mit 1 Nachkommastelle) Häufigkeiten und Summenhäufigkeiten).

Lösung der Aufgabe 3

Klasse Nr. i	Umsatzklasse (Mio €)	Anzahl Filialen h _i	Anteil f _i (%)	H _i	F _i (%)
1	0 b.u. 20	6	30,0%	6	30,0%
2	20 b.u. 50	7	35,0%	13	65,0%
3	50 b.u. 100	4	20,0%	17	85,0%
4	100 b.u. 200	3	15,0%	20	100,0%
Σ		20	100,0%	-	-

Auf die Frage „Wie viel Stück des Produktes ABC haben Sie im letzten Monat gekauft?“ gab es bei der Hauptuntersuchung unterschiedliche Antworten zur Zahl der gekauften Stücke von ABC. Die statistische Reihe wurde zusammengefasst in der folgenden Häufigkeitsverteilung:

a) Bestimmen Sie den Modus, den Median und das arithmetische Mittel.

b) Bestimmen Sie die Streuungsparameter Spannweite w, Varianz s², Standardabweichung s, den Variationskoeffizient v

c) Welche Aussagen können Sie auf der Basis der Werte der Lageparameter über die Form der Verteilung machen?

Lösung der Aufgabe 5

i	gekauft Stückzahl von ABC x _i	Anzahl der Nennungen h(x _i)	Anteil f(x _i) (%)	H(x _i)	F(x _i) (%)	(x _i - \bar{x}) ²	(x _i - \bar{x}) ² · h(x _i)
1	0	100	20%	100	20%	1,44	144
2	1	300	60%	400	80%	0,04	12
3	2	50	10%	450	90%	0,64	32
4	3	20	4%	470	94%	3,24	64,8
5	4	20	4%	490	98%	7,84	156,8
6	5	5	1%	495	99%	14,44	72,2
7	7	5	1%	500	100%	33,64	168,2
Σ		500	100%				650

a) $\bar{X} = 1$
 $\bar{X} = 100 + 1 \cdot 300 + 2 \cdot 50 + 3 \cdot 20 + 4 \cdot 20 + 5 \cdot 5 + 7 \cdot 5 = 5500 = 600500 = 1,2$

b) $w = x_{\max} - x_{\min} = 7 - 0 = 7$

1. Variante (Formel (2), s. Folie 20 Modul 5)
 $s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{500} [(0+1-0,6+4+0,1+9+0,04+16+0,04+25+0,01+49+0,01) - 1,22 \cdot 0,6 + 0,4 \cdot 0,36 + 0,64 \cdot 0,25 + 0,49 \cdot 0,1 + 1,44 \cdot 0,1] = 1,3$

2. Variante (Formel (1), s. Folie 20 Modul 5)
 $s^2 = 1/n \sum (x_i - \bar{x})^2 \cdot h(x_i) = 650500 / i = 1,3$
 $s = \sqrt{s^2} = 1,3 = 1,4$ $v = s/\bar{x} = 1,4 / 1,2 = 0,95$

c) Die Verteilung ist rechtsschief wegen $\bar{X} = 1 < 1,2 = \bar{x}$

Die Tabelle zeigt die Zahl der Eheschließungen bzw. die Zahl der Ehescheidungen je 10.000 Ehen in Deutschland

Jahr	Eheschließungen	Ehescheidungen je 10.000 Ehen
2001	389.000	198,2
2000	418.550	194
1999	430.674	187,7
1998	417.420	191,4
1997	422.776	181,2
1996	427.297	161
1995	430.534	153,8
1994	440.244	150
1993	442.605	135,6
1992	452.428	104,8
1991	454.291	104

a) Bestimme den Modalwert, den Zentralwert und das arithmetische Mittel sowohl von den Eheschließungen als auch von den Ehescheidungen. Was fällt dir an den Ergebnissen auf?

b) Bestimme die Varianz und die Standardabweichung sowohl von den Eheschließungen als auch von den Ehescheidungen.

c) Welche Veranschaulichungsmöglichkeiten für solch einen tabellarischen Zusammenhang hast du bereits kennen gelernt? Wähle zwei davon aus und realisiere sie! Welche Visualisierungsform ist in diesem Fall besonders geeignet bzw. ungeeignet und warum?

a) Worüber informiert ...

a. ... die Standardabweichung? Informiert über die (absolute) Streuung einer Verteilung. Die Standardabweichung hat die gleiche Maßeinheit wie die Beobachtungswerte und die Mittelwerte.

b. ... das Quartil Q3? Informiert darüber, welcher Wert von 75% der Beobachtungswerte einer Verteilung unterschritten wird und von 25% der Beobachtungswerte überschritten wird.

c. ... der Variationskoeffizient? Informiert über die relative Streuung einer Verteilung.

d. ... der Interquartilsabstand? Informiert über die Spannweite der mittleren 50% der Beobachtungswerte einer Verteilung.

e. ... das 5%-Quantil? Informiert über den Wert, der von 5% der Beobachtungswerte einer Verteilung unterschritten wird und von 95% der Beobachtungswerte überschritten wird.

f. ... der Median? Informiert über den Wert, der von der Hälfte der Beobachtungswerte überschritten und von der anderen Hälfte unterschritten wird.

g. ... der Modus? Informiert darüber, welche Merkmalsausprägung(en) in der Häufigkeitsverteilung am häufigsten vorkommen.

h. ... die Spannweite? Informiert über den maximalen Abstand der Merkmalsausprägungen, die in der Häufigkeitsverteilung vorkommen. Sie ergibt sich aus der Differenz zwischen dem größten und kleinsten Merkmalswert der Verteilung.

i. ... das Quartil Q1? Informiert darüber, welcher Wert von 25% der Beobachtungswerte einer Verteilung unterschritten wird und von 75% der Beobachtungswerte überschritten wird.

b) Was haben die statistischen Parameter Varianz, Standardabweichung und Variationskoeffizient gemeinsam? Wodurch unterscheiden sich die drei Parameter?

Lösung: Varianz, Standardabweichung und Variationskoeffizient sind Streuungsparameter. Mit ihnen lässt sich die Streuung einer Häufigkeitsverteilung charakterisieren. Varianz und Standardabweichung informieren über die absolute Streuung, der Variationskoeffizient über die relative Streuung.

In einer Statistik über die Einkommen (Jahreseinkommen) von leitenden Angestellten im Rechnungswesen lesen Sie:

1. Quartil Q1: 75.000 €, 3. Quartil Q3: 150.000 €, Median \bar{X} : 100.000 €, Mittelwert \bar{x} : 140.000 €. Welche der folgenden Aussagen über die Einkommensverteilung sind richtig? (Zutreffendes ankreuzen!)

Lösung der Aufgabe 2

	RICHTIG	FALSCH
(1) Die Einkommensverteilung ist linkschief.		X
(2) 25% der Befragten verdienen weniger als 75.000 €.	X	
(3) 50% der Befragten verdienen zwischen 75.000 und 150.000 €.	X	
(4) 75% der Befragten verdienen mehr als 150.000 €.		X
(5) Die Einkommensverteilung ist symmetrisch.		
(6) 50% der Befragten verdienen weniger als 140.000 €.		X

Erstellen Sie auf der Basis der folgenden Angaben eine eindimensionale, unklassierte Häufigkeitsverteilung:

In einer Stadt haben vier Taxi-Unternehmen jeweils drei Wagen, ein Taxi-Unternehmen hat 26 Wagen. Die übrigen Taxi-Unternehmen in der Stadt sind kleiner und haben weniger Wagen: acht Taxi-Unternehmen haben jeweils nur einen Wagen, sieben haben jeweils zwei Wagen. Fragen auf der Basis der erstellten Tabelle beantworten

Fragen (mit 1 Nachkommastelle):

a) Wie viele Taxi-Unternehmen gibt es in dieser Stadt?
20 Taxi Unternehmen in dieser Stadt.

b) Wie viele Wagen bieten in dieser Stadt ihre Leistungen an?
60 Wagen stehen in der Stadt zur Verfügung. $1^*8+2^*7+3^*4+26^*1=60$

c) Bestimmen / berechnen Sie für die obige Verteilung

a. den Modus
 $\bar{x}_D = 1$ (da Anz. Unternehmen mit 1 Wagen am höchsten ist)

b. den Median
 $\bar{x}_Z = 2$ (s. Folie 11 Modul 4) Reihe: 1 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 26

c. das arithmetische Mittel
 $\bar{x} = 1 \cdot 8 + 2 \cdot 7 + 3 \cdot 4 + 26 \cdot 120 = 6020 = 3$

d) Machen Sie Aussagen über die Schiefe der Verteilung
rechtsschiefe Verteilung

e) Berechnen Sie die Standardabweichung
 $s^2 = 1/20 \cdot [(1-3)^2 \cdot 8 + (2-3)^2 \cdot 7 + (3-3)^2 \cdot 4 + (26-3)^2 \cdot 1] = 120 \cdot (32 + 7 + 0 + 529) = 56820 = 28,4$ $s = \sqrt{s^2} = \sqrt{28,4} = 5,33$

f) Wie viel % der Taxi-Unternehmen haben mehr als einen Wagen
100 - 60% = 40%

g) Wie viel % der Taxi-Unternehmen haben weniger als 3 Wagen
40 + 35 = 75%

Lösung der Aufgabe 3

i	Anzahl Wagen x _i	Anzahl Taxi-Unternehmen h(x _i)	f(x _i) (%)	H(x _i)	F(x _i) (%)
1	1	8	40,0%	8	40,0%
2	2	7	35,0%	15	75,0%
3	3	4	20,0%	19	95,0%
4	26	1	5,0%	20	100,0%
Σ		20	100,0%	-	-

Lösung der Aufgabe 8

a) Der Modalwert der Eheschließungen in Deutschland und Ehescheidungen je Ehe existiert nicht, denn jeder Wert kommt nur ein einziges Mal vor. Kein Wert der Datenreihe ist ein Modalwert!

Zentralwert = Median → ist der 6. Element der sortierten Reihe für Eheschließungen $\bar{X} = 430.534$ für Ehescheidungen $\bar{X} = 1.610.000$ oder 161 je 10.000 Ehen

\bar{X} Eheschließungen = 389.000 + ... + 454.29111 = 4.725.81911 = 429.619,909 = 429,620
 \bar{X} Ehescheidungen = 198,2 + ... + 10411 = 1.76211 = 160,154545 = 160 je 10.000 Ehen

Das arithmetische Mittel entspricht ungefähr dem Zentralwert der Datenmenge sowohl bei den Eheschließungen als auch bei den Ehescheidungen, d.h. dass die Entwicklung der Zahlen gleichmäßig ist und keine Ausreißer aufweist.

b) Berechnung der Varianz über die absolute Häufigkeit:
 s^2 Eheschließungen = $111[(1-1)^2 \cdot 8 + (2-1)^2 \cdot 7 + (3-1)^2 \cdot 4 + (26-1)^2 \cdot 1] = 111[(389.000 - 429.620)^2 + \dots + (45.291 - 429.620)^2] = 307.808.635$
 s^2 Ehescheidungen = $s^2 = 307.808.635 = 17.545$
 s^2 Eheschließungen = $111[(1-1)^2 \cdot 8 + (2-1)^2 \cdot 7 + (3-1)^2 \cdot 4 + (26-1)^2 \cdot 1] = 111[(198,2 - 160)^2 + \dots + (104 - 160)^2] = 1.066,89$ je 10.000 Ehen
 s^2 Ehescheidungen = $s^2 = 1.066,89 = 32,66$ je 10.000 Ehen

c) Darstellungsformen: Graph, Säulendiagramm, Stabdiagramm, Balkendiagramm, Kreisdiagramm. Besonders geeignete Darstellungsformen: Graph, Balkendiagramm, Säulendiagramm und Stabdiagramm; ungeeignete Darstellungsform ist Kreisdiagramm.

