

2 Häufigkeitsverteilungen

Lernziele

In diesem Kapitel geht es um „beschreibende Statistik“. Nach erfolgreicher Bearbeitung sind Sie in der Lage, eine zunächst unübersichtliche Menge beobachteter Daten so aufzubereiten, dass die Daten an Aussagekraft gewinnen. Sie können dazu unter verschiedenen Typen von Tabellen und Grafiken die jeweils geeigneten auswählen und diese Aufbereitungsform dann für Ihre Daten nutzen. Es geht dabei zunächst um qualitative Daten, also solche, die entweder nur nominale Skalen oder doch höchstens ordinale zulassen. Danach behandeln wir die Aufbereitung quantitativer Daten.

Praxisbeispiel

Das umfangreiche Internetangebot des Statistischen Bundesamtes enthält u. a. die Tabelle 2.1.

Tabelle 2.1 **Bevölkerungsstand in Deutschland 2015**

Bevölkerungs- stand	30.09.2015	31.12.2015
	1 000	
Insgesamt	81 770,9	82 175,7
männlich	40 238,5	40 514,1
weiblich	41 532,4	41 661,6
Deutsche	73 536,3	73 523,8
männlich	35 912,2	35 910,0
weiblich	37 624,2	37 613,7
Nichtdeutsche	8 234,6	8 652,0
männlich	4 326,4	4 604,1
weiblich	3 908,3	4 047,8

Quelle: Statistisches Bundesamt (2016a)

Es handelt sich um eine zweidimensionale Häufigkeitstabelle. Die beobachteten Merkmale sind „Bevölkerungsgruppe“ und „Stichtag“, die erste Spalte und die erste Zeile enthalten die jeweiligen Merkmalsausprägungen, in den Spalten 2 und 3 finden Sie die absoluten Häufigkeiten, mit denen die Kombinationen der Merkmalsausprägungen beobachtet wurden. Durch das Einfügen von Summenzeilen wurde hier die Tabelle für den Leser aussagefähiger gemacht, eigentlich hätten offenbar vier Zeilen für die elementaren Merkmalsausprägungen „deutsch männlich“, „deutsch weiblich“, „nichtdeutsch männlich“ und „nichtdeutsch weiblich“ ausgereicht.

Anstelle der tabellarischen Form können Informationen aber auch grafisch dargestellt werden. Zum Thema Zu- und Abwanderung finden wir z. B. die Grafik Abbildung 2.1.

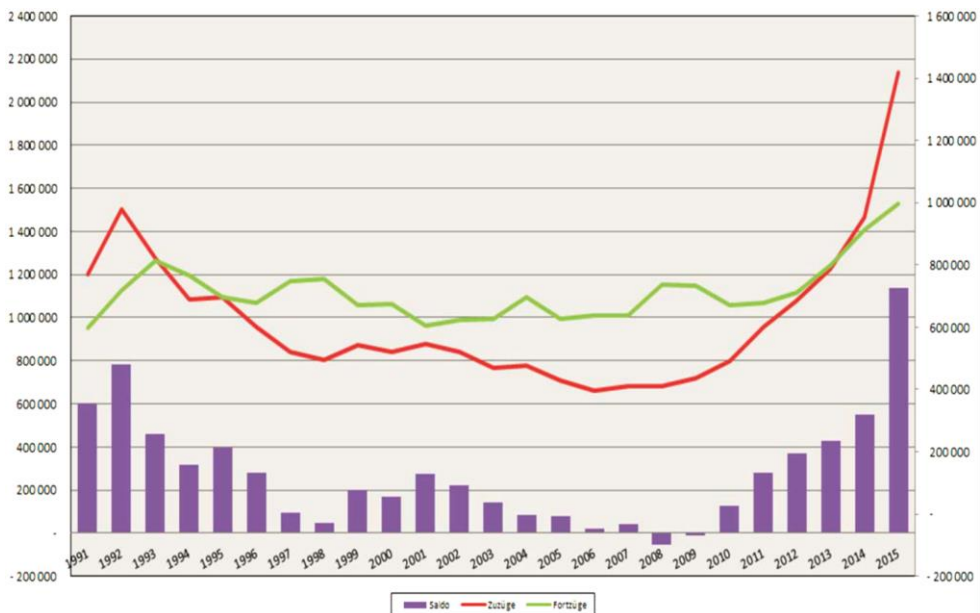


Abbildung 2.1 Zu- und Abwanderungen in Deutschland 1991-2015

Quelle: Statistisches Bundesamt (2016d)

Hier sind die Darstellungsformen Säulendiagramm und Liniendiagramm in einer Grafik zusammengefasst. Auf Liniendiagramme greift man im Allgemeinen nur dann zurück, wenn – wie hier – zeitliche Abläufe darzustellen sind. Dies wird für uns im vorliegenden Buch keine besondere Rolle spielen.

Eine andere Darstellungsform verwendet das Statistische Bundesamt in der Übersicht zu den Einkommensquellen deutscher Haushalte, siehe Abbildung 2.2. Dieser Ring ist ein modifiziertes Kreisdiagramm, bei dem die Sektoren nicht bis zum Mittelpunkt durchgezeichnet sind. Statt von Kreisdiagrammen sprechen süßigkeitsliebende Statistiker auch von Kuchen-

oder Tortendiagrammen. Abbildung 2.2 zeigt dann eher einen Frankfurter – Wiesbadener? – Kranz.

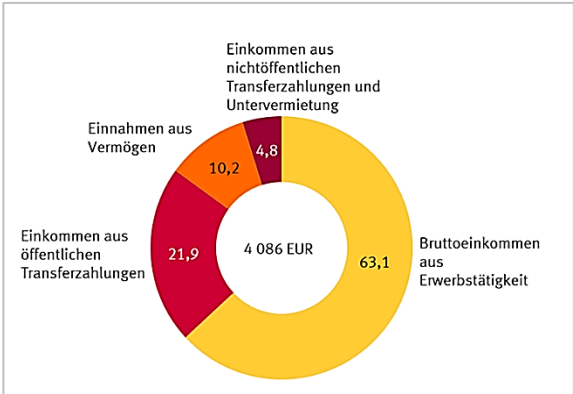


Abbildung 2.2 Struktur des Haushaltsbruttoeinkommens privater Haushalte 2013, Anteile in Prozent
Quelle: Statistisches Bundesamt (2015b)

Wann man welche Diagramme verwendet, und wie sie konstruiert werden, lernen Sie in diesem Kapitel.

2.1 Aufbereitung qualitativer Daten in Tabellen

Der Manager des Mittelklasse-Hotels „Gute Nacht“ hat beschlossen, mehr über die Meinung seiner Gäste zu erfahren und deshalb beim Auschecken gebeten, den Eindruck über die Qualität der Unterbringung nach folgenden Kategorien zu beurteilen: ausgezeichnet, gut, ordentlich, verbesserungswürdig, unakzeptabel. Es handelt sich somit um eine Ordinalskala. Mit einer ungeraden Anzahl von Merkmalsausprägungen einer Ordinalskala wird die Möglichkeit zu neutralen Aussagen eröffnet. Dies ist in der Regel zu empfehlen, da man einen Befragten nie zwingen soll, sich zwischen Alternativen zu entscheiden, wenn er tatsächlich unentschieden ist. Es wurden bei zwanzig Gästen folgende Aussagen gesammelt:

Tabelle 2.2 Hotelbewertungen „Gute Nacht“ – erfasste Daten

verbesserungswürdig	ordentlich	verbesserungswürdig	gut
gut	ordentlich	gut	gut
gut	verbesserungswürdig	ordentlich	ausgezeichnet
unakzeptabel	gut	unakzeptabel	ordentlich
gut	ordentlich	gut	gut

Offensichtlich bedarf es hier der ordnenden Hand, damit aus den Rohdaten leicht überschaubare „Information“ wird. In diesem Fall wird man einfach zählen, wie oft die einzelnen Antworten gegeben wurden, und die Ergebnisse auflisten. Die Merkmalsausprägungen werden

selbstverständlich in der Reihenfolge der Ordinalskala aufgelistet. Die gezählte *Häufigkeit* als die reine Anzahl (*absolute Häufigkeit*) finden Sie in Spalte 2 der Tabelle 2.3.

Außerdem kann man die Häufigkeit als Anteil der Merkmalsausprägung an der Gesamtzahl (*relative Häufigkeit*) oder als Prozentwert (*prozentuale Häufigkeit*) angeben (Spalten 3 und 4).

$$\text{relative Häufigkeit} = \frac{\text{absolute Häufigkeit}}{\text{Gesamtzahl der Beobachtungen}}$$

$$\text{prozentuale Häufigkeit} = \text{relative Häufigkeit} \cdot 100 \text{ (in Prozent)}$$

Tabelle 2.3 Häufigkeitsverteilung der Hotelbewertungen „Gute Nacht“

Bewertung	Absolute Häufigkeit (h_i)	Relative Häufigkeit ($h_i/\text{Gesamtzahl}$)	Prozentuale Häufigkeit ($100 \cdot h_i/\text{Gesamtzahl}$)
ausgezeichnet	1	0,05	5
gut	9	0,45	45
ordentlich	5	0,25	25
verbesserungswürdig	3	0,15	15
unakzeptabel	2	0,1	10

Die Zuordnung von Häufigkeiten (in den jeweiligen Tabellenspalten) zu den Merkmalsausprägungen nennen wir die *absolute*, *relative* bzw. *prozentuale Häufigkeitsverteilung* des beobachteten Merkmals.

Beobachtungen eines qualitativen Merkmals, das sich ausschließlich mit Hilfe einer Nominalskala erfassen lässt (z. B. die Nationalität der Hotelgäste), lassen sich auf dieselbe Weise tabellarisch darstellen. Allerdings haben in diesem Fall die Eintragungen keine natürliche Reihenfolge mehr. Um nicht reine Willkür walten zu lassen, lässt sich eine (allerdings problemfremde) Reihenfolge etwa durch die alphabetische Ordnung der Nationalitätsbezeichnungen wählen. Stärker am Problem orientiert wäre eine Ordnung nach der beobachteten Häufigkeit der Merkmalsausprägungen.

Liegt ein mindestens ordinal skaliertes Merkmal vor, ist es in manchen Fällen hilfreich, kumulierte Häufigkeiten wie folgt zu bilden:

Sind die k Merkmalsausprägungen der Reihe nach geordnet und ist h_i die Häufigkeit der i -ten Merkmalsausprägung bei dieser Anordnung ($i = 1, \dots, k$), dann ist

$$H_i = \sum_{j=1}^i h_j$$

die kumulierte Häufigkeit. Für die Hotelbewertungen sieht das wie folgt aus:

**Tabelle 2.4 Kumulierte Häufigkeitsverteilung der Hotelbewertungen
„Gute Nacht“**

Bewertung	Kumulierte absolute Häufigkeit (h _i)	Kumulierte relative Häufigkeit (h _i / Gesamtzahl)	Kumulierte prozentuale Häufigkeit (100 · h _i / Gesamtzahl)
ausgezeichnet	1	0,05	5
gut oder besser	10	0,5	50
ordentlich oder besser	15	0,75	75
verbesserungswürdig oder besser	18	0,9	90
insgesamt	20	1	100

Kumulierte Häufigkeitsverteilungen kommen für Merkmale mit nur nominalen Skalen in der Regel nicht in Betracht, da sie ja die Summation von Werten in einer definierten – bei Nominalskalen nicht gegebenen – Reihenfolge erfordern. In geeigneten Fällen kann auch bei einer nominal skalierten Variablen eine kumulierte Verteilung erstellt werden, indem man die Anordnung der Merkmalsausprägungen nach der Häufigkeit der Beobachtungen vornimmt. Damit werden dann Aussagen der Form „80 % der Beobachtungen entfallen auf die drei häufigsten Merkmalsausprägungen“ möglich.

Überwiegend verwendet man aber kumulierte Häufigkeitsverteilungen ohnehin bei quantitativen Merkmalen.

Bisher haben wir je Beobachtungseinheit (im Beispiel: Hotelgast) *ein* Merkmal (Zufriedenheit) beobachtet. Beobachten wir für unsere Beobachtungseinheiten jeweils zwei Merkmale, dann ist eine *Kreuztabelle* die passende Darstellungsform.

In dem Fünf-Sterne-Hotel „Bellevue“ haben 42 Gäste Schulnoten von 1 (für ausgezeichnet) bis 5 (für unakzeptabel) gewählt, außerdem gaben sie ihre Nationalität an: siehe Tabelle 2.5.

Daraus ergibt sich die Kreuztabelle mit den absoluten Häufigkeiten: siehe Tabelle 2.6.

Die Häufigkeitsauszählung geschieht in der gleichen Weise wie bei einer eindimensionalen Häufigkeitsverteilung. Neu sind in Tabelle 2.6 die letzte Spalte und letzte Zeile. Hier werden

jeweils die Zeilensummen bzw. Spaltensummen dargestellt. Die letzte Spalte stellt somit beispielsweise die eindimensionale Häufigkeitsverteilung der Zufriedenheitsvariablen dar. Es waren also zum Beispiel fünf Gäste überhaupt nicht zufrieden. Die Gäste aus den USA waren insgesamt wenig zufrieden.

Tabelle 2.5 Hotelbewertungen „Bellevue“ nach Bewertungsergebnis und Nationalität – Rohdaten

4; DE	3; DE	4; IT	5; US	3; DE	2; UK
2; FR	3; US	2; DE	3; IT	2; IT	5; DE
2; UK	4; UK	3; DE	2; DE	4; US	2; DE
5; US	2; BE	5; UK	3; US	1; DE	1; DE
2; DE	3; IT	2; FR	4; UK	5; DE	3; DE
1; SE	3; DE	2; DE	1; SE	3; IT	3; UK
2; FR	2; SE	3; DE	2; SE	2; FR	2; SE

Tabelle 2.6 Hotelbewertungen „Bellevue“ nach Bewertungsergebnis und Nationalität – aufbereitete Daten

Bewertung ↓ Nationalität →	BE	DE	FR	IT	SE	UK	US	Σ
1	0	2	0	0	2	0	0	4
2	1	5	4	1	3	2	0	16
3	0	6	0	3	0	1	2	12
4	0	1	0	1	0	2	1	5
5	0	2	0	0	0	1	2	5
Σ	1	16	4	5	5	6	5	42

An dieser Stelle soll nicht darüber spekuliert werden, welche Schlüsse der Hotelmanager hieraus nun ziehen wird, sicher ist jedenfalls, dass man im Gegensatz zur ursprünglichen Datenmenge mit Hilfe der Kreuztabelle überhaupt etwas „sieht“.

Lernkontrolle zu 2.1

1. In einer Umfrage wollen Sie die Befragten veranlassen, zu einem Problem Stellung zu nehmen, dem man nach Ihrer Überzeugung auch neutral gegenüberstehen kann. Die Antwort soll auf einer Skala von „stimme voll zu“ bis „lehne vollständig ab“ gegeben werden.
 - a. Sie wählen eine ungerade Anzahl von Antwortalternativen.
 - b. Sie wählen eine gerade Anzahl von Alternativen.
 - c. Sie wählen genau zehn Alternativen.
 - d. Auf die Anzahl der Alternativen kommt es nicht an.
2. Von dreißig Bewerbern haben sich sechs für eine Aufgabe qualifiziert. 0,2 ist ...
 - a. ... die absolute Häufigkeit der Qualifikation.
 - b. ... die relative Häufigkeit.
 - c. ... die prozentuale Häufigkeit.
 - d. ... keines von allem.
3. In einer Tabelle, in der Daten zu 150 Werkstücken gesammelt sind, ist als relative Häufigkeit der Eigenschaft „zu rau“ 0,02 angegeben. Das bedeutet:
 - a. 2 Werkstücke sind zu rau.
 - b. 2 % der Werkstücke sind zu rau.
 - c. 3 % der Werkstücke sind zu rau.
 - d. Drei Werkstücke sind zu rau.
 - e. Hieraus lässt sich die absolute Zahl zu raueren Werkstücke nicht ermitteln.
4. In einer Urne liegen zwei gelbe, drei rote und fünf blaue Kugeln.
 - a. Die relative Häufigkeit der roten Kugeln ist 30 %.
 - b. Die kumulierte relative Häufigkeit der blauen Kugeln ist 1.
 - c. Die relative Häufigkeit der blauen Kugeln ist 0,5.
 - d. Der Begriff der relativen Häufigkeit ist in diesem Beispiel unsinnig.
 - e. Der Begriff der kumulierten Häufigkeit ist in diesem Beispiel unsinnig.
5. Zur tabellarischen Darstellung von zwei Merkmalen je Beobachtung verwendet man ...
 - a. ... eine Kreuzungstabelle.
 - b. ... eine Überkreuztabelle.
 - c. ... eine Kreuztabelle.
 - d. ... eine Kreuzweistabelle.



2.2 Grafische Aufbereitung qualitativer Daten

Mit der tabellarischen Aufbereitung der Daten ist bereits ein wichtiger Schritt gemacht. Aber letztlich sagt ein Bild nicht nur mehr als tausend Worte, sondern eine Grafik kann häufig auch sehr viel einprägsamer und aussagekräftiger sein als die Tabelle, deren Zahlen sie wiedergibt.

Für die grafische Darstellung qualitativer Daten kommen standardmäßig Säulen- bzw. Balkendiagramme und Kreisdiagramme (Kuchen-, Tortendiagramme) in Frage.

Für ein *Säulen-* oder *Balkendiagramm* vermerkt man auf einer Koordinatenachse die Merkmalsausprägungen. Im Fall einer Ordinalskala erfolgt dies natürlich in der durch die Skala vorgegebenen Reihenfolge. Die andere Achse teilt man so ein, wie es der gewählten Häufigkeitsdarstellung (absolut, relativ oder prozentual) entspricht. Über (neben) jeder Merkmalsausprägung gibt eine Säule (ein Balken) in geeigneter Länge die Häufigkeit der Ausprägung an. Üblicherweise verwendet man für die Merkmalsausprägungen die horizontale Achse, dann ist der Begriff der „Säule“ über der jeweiligen Ausprägung anschaulich. Insbesondere bei einer großen Anzahl von Merkmalsausprägungen und/oder langen Bezeichnungen kann es praktischer sein, die Ausprägungen an der senkrechten Achse anzuschreiben und waagerechte „Balken“ für die Häufigkeiten zu verwenden.

Da die einzelnen Ausprägungen qualitativer Merkmale isoliert nebeneinanderstehen, verwenden wir Säulen/Balken, die sich gegenseitig nicht berühren. Kaum der Erwähnung bedarf, dass die Säulen/Balken alle dieselbe Breite haben, sodass ihre Höhe bzw. Länge ebenso wie ihr Flächeninhalt das Verhältnis der Merkmalsausprägungen optisch verdeutlicht. Für das Beispiel der Hotelbewertungen ist in Abbildung 2.3 das Säulendiagramm für die absoluten Häufigkeiten gezeichnet:

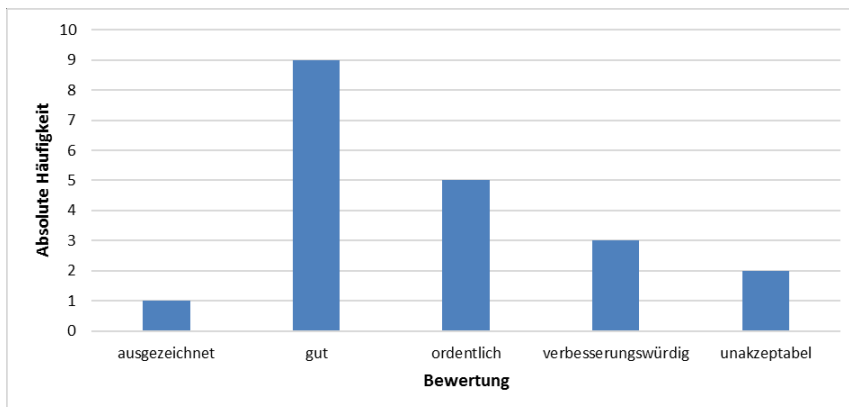


Abbildung 2.3 Säulendiagramm der Bewertungen für das Hotel „Gute Nacht“

Computerübung 2.1

Zur Erstellung dieses Diagramms in Excel gibt man die Merkmalsausprägungen und ihre Häufigkeiten in Spalten des Arbeitsblatts ein und markiert diesen Bereich. (Zur Erinnerung für Schreibunlustige, die sich etwas Arbeitserleichterung wünschen: In der Ihnen jetzt schon wohlbekannten zentralen Linkliste finden Sie auch einen Link zu einer Excel-Datei mit allen für die Computerübungen benötigten Daten. Alternativ verwenden Sie den QR-Code in der Box „Computerübungen“ in Kapitel 1.) Auf der Registerkarte „Einfügen“ wählt man dann in der Gruppe „Diagramme“ den passenden Typ – hier also das Säulendiagramm – und ist praktisch fertig. Das Programm geht davon aus und berücksichtigt automatisch, dass in der ersten Spalte die Merkmalsausprägungen und in der zweiten die Häufigkeiten stehen, *sofern die erste Spalte Text enthält*. Die Beschriftungen (Titel, x-Achse, y-Achse) gibt man ein, indem man bei markiertem Diagramm unter „Diagrammtools/Entwurf“ ganz links „Diagrammelement hinzufügen“ wählt.

Versuchen Sie dasselbe auch mit R. Der Befehl, ein Diagramm zu erzeugen, heißt `barplot`. Zuvor müssen Sie aber einen Vektor mit den absoluten Häufigkeiten befüllen. Das geschieht durch den Befehl `HotBew <- c(1, 9, 5, 3, 2)`. (Zur Erinnerung: Auch hier hilft bei Bedarf der Inhalt einer Datei weiter, zu der man durch den im Kapitel 1.5 angegebenen QR-Code oder durch die zentrale Linkliste gelangt.)

Das *Kreisdiagramm* ist eine andere Möglichkeit der Visualisierung qualitativer Daten. Die 360° des Kreises werden im Verhältnis der ermittelten Häufigkeiten aufgeteilt. Einer Merkmalsausprägung mit einer relativen Häufigkeit r wird ein Sektor von $r \cdot 360^\circ$ zugewiesen:

$$\text{Winkel} = \text{Relative Häufigkeit} \cdot 360.$$

Hier die Torte der Hotelbewertungen, guten Appetit!

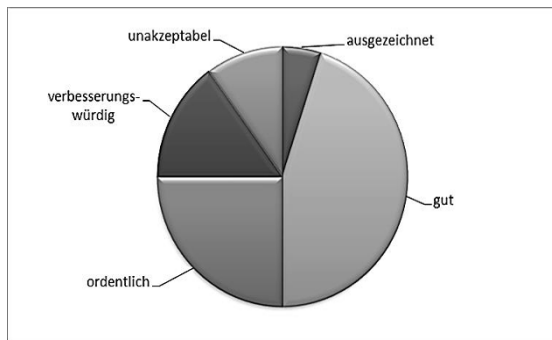


Abbildung 2.4
Kreisdiagramm der Bewertungen
für das Hotel „Gute Nacht“

Computerübung 2.2

Mit Excel ist dies wieder eine Angelegenheit von wenigen Klicks. Nach Markieren des Datenbereichs und Ansteuern der Registerkarte „Einfügen“ wählt man lediglich in der Gruppe „Diagramme“ jetzt das Kreisdiagramm aus. Zur Befriedigung ästhetischer Bedürfnisse haben wir eine besonders tortenähnliche Variante gewählt.

In R heißt der zuständige Zeichenbefehl sehr intuitiv `pie`.

Die in einer Kreuztabelle enthaltenen Daten erfordern bei der Visualisierung durch ein Säulendiagramm eine *dreidimensionale Darstellung*. Dabei wird allerdings das Ziel der leichteren Erfassbarkeit gegenüber der zahlenmäßigen Darstellung nicht immer erreicht. Dies auch – aber nicht nur – wegen der möglichen gegenseitigen Verdeckung der Säulen. Aus Tabelle 2.6 gewinnen wir Abbildung 2.5.

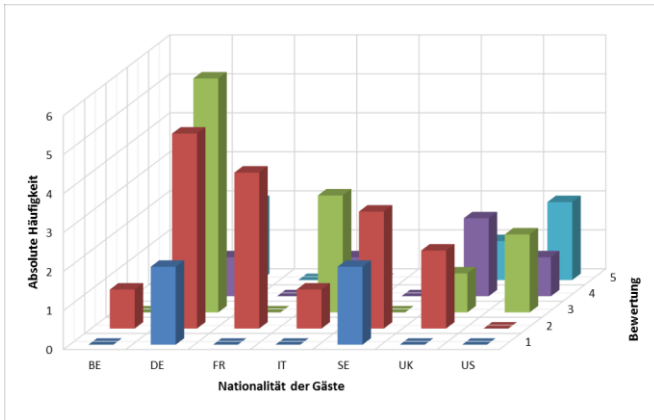


Abbildung 2.5 Dreidimensionales Säulendiagramm der Gästenumfrage des Hotels „Bellevue“ nach Bewertungsergebnis und Nationalität

Optisch aussagefähiger ist häufig ein *Blasendiagramm*, in dem die Zahlen der Kreuztabelle durch Kreise ersetzt sind, deren Flächeninhalt der jeweiligen Häufigkeit entspricht.

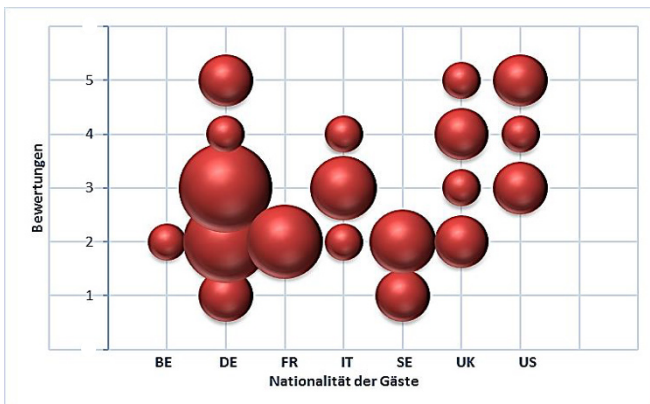


Abbildung 2.6 Blasendiagramm der Gästenumfrage des Hotels „Bellevue“ nach Bewertungsergebnis und Nationalität

Computerübung 2.3

Während die Erzeugung des dreidimensionalen Säulendiagramms aus Abbildung 2.5 keine Besonderheiten mit sich bringt (man erfasst die Kreuztabelle im Arbeitsblatt, markiert sie und steuert den passenden Diagrammtyp an), bedarf es für das Blasendiagramm einer besonderen Vorbereitung. Excel braucht für jede Blase ein Datentripel (x-Wert, y-Wert, Blasengröße). Dabei müssen auch die x- und y-Werte numerisch sein. Während (zufällig) die Bewertungsnoten schon numerisch sind, ist das bei den Nationalitätencodes nicht der Fall. Diese müssen also umcodiert werden, z. B. 1 für BE, 2 für DE usw. Dann muss für jedes Feld der Kreuztabelle eine Zeile einer dreispaltigen Tabelle im Arbeitsblatt geschrieben werden: numerischer Nationalitätencode, numerischer Bewertungscode, Häufigkeit. Wenn diese Tabelle markiert ist, geht es weiter wie üblich, Blasendiagramme findet man unter „Einfügen/Empfohlene Diagramme/Alle Diagramme/Punkt (X Y)“. Unschön ist natürlich die sachfremde numerische Beschriftung der Achsen, zumindest der Nationalitätenachse. Hier kann man sich helfen, indem man die Achsenwertbeschriftung löscht und durch einen mit etwas Probieren angepassten Achsentitel ersetzt.

In R kann man sich mit der Zeichenfunktion `symbol` helfen, die allerdings keine sehr schönen Blasen erzeugt. Einen speziellen Diagrammtyp für Blasendiagramme enthalten sonst nur Spezialpakete für R. Wer darauf besonderen Wert legt, muss ein wenig recherchieren.

Lernkontrolle zu 2.2

1. In einem Säulendiagramm ...
 - a. ... haben alle Säulen dieselbe Höhe.
 - b. ... haben alle Säulen dieselbe Breite.
 - c. ... grenzen die Säulen unmittelbar aneinander.
 - d. ... sind die Säulen der Höhe nach sortiert.
 - e. ... sind die Säulen durch Zwischenräume getrennt.
 - f. ... sind die Säulen im Fall einer Ordinalskala in der Reihenfolge der Merkmalsausprägungen sortiert.
2. In einem Kreisdiagramm ...
 - a. ... sind die Flächen der Sektoren proportional zur Häufigkeit der Merkmalsausprägungen.
 - b. ... sind die Winkel der Sektoren proportional zur Häufigkeit der Merkmalsausprägungen.
 - c. ... sind die Häufigkeiten durch unterschiedliche Radien der Sektoren dargestellt.
 - d. ... ist die Fläche des Kreises proportional zur Anzahl der Beobachtungen.
3. Ein Blasendiagramm ...
 - a. ... wird für Beobachtungen mit einem Merkmal verwendet.
 - b. ... wird für Beobachtungen mit zwei Merkmalen verwendet.
 - c. ... wird für Beobachtungen mit drei oder mehr Merkmalen verwendet.
 - d. ... gibt die Häufigkeit der Ausprägung der Merkmale durch die Höhe des Blasenmittelpunkts an.



Lösung LK 2.2

Videos 2.1 und 2.2

Die Präsentation qualitativer Daten im Rahmen der beschreibenden Statistik können Sie sich noch einmal durch zwei Videos erläutern lassen – die QR-Codes dazu finden Sie hier rechts. Als Werkzeug wird Microsoft Excel benutzt.

Sie brauchen etwa 14 Minuten (Video 2.1) bzw. 10 Minuten (Video 2.2).



2.3 Aufbereitung quantitativer Daten in Tabellen

Den jetzt zu behandelnden Daten liegt mindestens eine Intervallskala oder sogar eine Verhältnisskala zu Grunde. Wir sprechen von einem *diskreten Merkmal*, falls die Merkmalsausprägungen nur Werte aus einer (endlichen oder unendlichen) Folge von Zahlen annehmen können. Ein *stetiges* oder *kontinuierliches Merkmal* liegt vor, falls der Wertebereich der Merkmalsausprägungen ein Kontinuum von Werten ist (ein endliches oder unendliches Intervall). Die Personenzahl in einer Familie ist ein diskretes Merkmal, die einer Familie zur Verfügung stehende Wohnfläche ein stetiges Merkmal.

Begrifflich ist die Unterscheidung zwischen diskret und stetig scharf, bei ganz genauem Hinsehen sind die Übergänge in der Realität allerdings doch ein wenig fließend. Ein diskretes Merkmal mit 10.000 möglichen Ausprägungen unterscheidet sich nicht viel von einem stetigen Merkmal, dessen Ausprägungen theoretisch „alle“ Werte zwischen 10 cm und 20 cm annehmen können, wenn unsere Messgenauigkeit auf 1/100stel Millimeter begrenzt ist. Solche spitzfindigen Abgrenzungsfragen sollen aber für uns im Folgenden keine Rolle spielen.

Weitere Beispiele für diskrete Merkmale: Anzahl der Gäste eines Restaurants an einem Tag (ermittelt an allen Tagen des Jahres 2016), Anzahl der bei einem Basketballspiel geworfenen Körbe, Anzahl der Fahrzeuge, die in einem Zeitintervall von zehn Minuten eine Kreuzung passieren.

Weitere Beispiele für stetige Merkmale: Temperatur um 12:00 Uhr mittags in Bonn (ermittelt an allen Tagen des Jahres 2016), Gewicht der Patienten einer Arztpraxis, Zeit für einen Boxenstopp beim Formel-1-Rennen.

Bei qualitativen Merkmalen ist es eher die Regel, dass in der Gesamtheit der Beobachtungseinheiten die Ausprägungen mehrfach vorkommen. Die Darstellungsmöglichkeiten dafür, wie häufig dies jeweils geschieht, wurden im vorigen Abschnitt erörtert. Bei quantitativen Merkmalen ist das mehrfache Auftreten ein und derselben Merkmalsausprägung meist selten. Dies wird höchstens bei diskreten Merkmalen in Kombination mit vielen Beobachtungen und/oder wenigen Merkmalsausprägungen vorkommen.

Die sehr viel häufigeren Fälle, in denen eine einzelne Merkmalsausprägung – wenn überhaupt – nur einmal, im Ausnahmefall ein paar wenige Male, vorkommt, erfordern vor der

Tabellierung eine *Klasseneinteilung der Merkmalsausprägungen*. Sonst wird die Tabelle kaum etwas anderes als die Auflistung der Rohdaten sein.

Im Allgemeinen wird diejenige, die eine gegebene Menge von Beobachtungen aufbereitet, selbst sehr schnell sehen, ob eine gewählte Einteilung zum Erkenntnisgewinn oder zur Verwirrung führt. Es gibt allerdings bei der Klasseneinteilung einige Grundregeln, die beachtet werden sollten.

- Wählen Sie Klassen, die alle dieselbe Breite haben. Sinnvolle Ausnahmen können die erste und die letzte Klasse sein.
- Als Klassengrenzen wählen Sie „runde“ Zahlen.
- Finden Sie einen vernünftigen Kompromiss zwischen Klassenzahl und Klassenbelegung (Klassenzahl in der Regel zwischen 5 und 20). Je größer die Zahl der Beobachtungen ist, umso größer kann und soll die Anzahl der Klassen sein.

Zur Bestimmung der Klassenbreite wird zunächst

$$(\text{Maximalwert} - \text{Minimalwert}) / \text{Klassenzahl}$$

berechnet und das Ergebnis dann mathematisch gerundet, erforderlichenfalls noch so angepasst, dass das Ziel „runde Klassengrenzen“ erreicht wird.

Beispiel

Als Klasseneinteilung für die Körpergröße Erwachsener (nennen wir sie x) könnte sinnvoll sein (Angaben in cm):

$$x \leq 150 \mid 150 < x \leq 160 \mid 160 < x \leq 170 \mid 170 < x \leq 180 \mid 180 < x \leq 190 \mid 190 < x \leq 200 \mid x > 200$$

Weniger sinnvoll ist diese Einteilung, da das Ziel „runde“ Klassengrenzen verfehlt ist:

$$x \leq 152 \mid 152 < x \leq 162 \mid 162 < x \leq 172 \mid 172 < x \leq 182 \mid 182 < x \leq 192 \mid 192 < x \leq 202 \mid x > 202$$

(Vielleicht war der Statistiker 2,01 m lang und wollte nicht zur Randklasse gehören?)

Da die in diesem Abschnitt betrachteten quantitativen Daten immer auch ordinal – also sortierbar – sind, kann man immer und wird man in passenden Fällen häufig die *kumulierten Werte* tabellieren.

Bei den Körpergrößen ergibt sich für die kumulierte Verteilung die Einteilung:

$$x \leq 150 \mid x \leq 160 \mid x \leq 170 \mid x \leq 180 \mid x \leq 190 \mid x \leq 200 \mid x \leq \text{höchster } x\text{-Wert}$$

Beispiel

Ein Computerreparaturbetrieb will sich einen Überblick über die Ersatzteilkosten pro Reparatur verschaffen. Zu diesem Zweck wurden die Ersatzteilkosten (in €) von insgesamt fünfzig Reparaturen aufgezeichnet.

Tabelle 2.7 Ersatzteilkosten (in €) – erfasste Daten

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Der Ladenbesitzer will die Daten tabellarisch aufbereiten, um Informationen über die Kostenstruktur zu erhalten. Da es sich bei den Ersatzteilkosten um eine quantitative Variable handelt, müssen zuerst Klassen gebildet werden. Da die Anzahl der Beobachtungen nicht sonderlich groß ist, entscheiden wir uns für sechs Klassen. Laut Formel lässt sich die Klassenbreite folgendermaßen berechnen:

$$\text{Klassenbreite} = (\text{Maximalwert} - \text{Minimalwert}) / \text{Klassenzahl} = (109 - 52) / 6 = 9,5 \approx 10$$

Als nächstes werden die Klassen bestimmt. Zweckmäßigerweise startet man mit einer runden Zahl knapp unterhalb des kleinsten Wertes. Somit ergeben sich bei einer Klassenbreite von 10 diese Klassen:

$$50 < x \leq 60 \mid 60 < x \leq 70 \mid 70 < x \leq 80 \mid 80 < x \leq 90 \mid 90 < x \leq 100 \mid 100 < x \leq 110$$

Nun wird ausgezählt, wie viele Rechnungen in die jeweilige Klasse fallen. Damit ergibt sich die folgende Häufigkeitstabelle

Tabelle 2.8 Ersatzteilkosten (in €) – aufbereitete Daten

Reparaturkosten	Absolute Häufigkeit	Relative Häufigkeit	Prozentuale Häufigkeit
$50 < x \leq 60$	2	0,04	4
$60 < x \leq 70$	13	0,26	26
$70 < x \leq 80$	16	0,32	32
$80 < x \leq 90$	7	0,14	14
$90 < x \leq 100$	7	0,14	14
$100 < x \leq 110$	5	0,1	10
Gesamt	50	1	100

Da wir jetzt quantitative Merkmale in die Betrachtung einbeziehen, sind bei Beobachtungsreihen mit *zwei Variablen* außer den in Abschnitt 2.1 betrachteten *Kreuztabellen* für zwei qua-

litative Merkmale jetzt auch die Kombinationen quantitativ/qualitativ und quantitativ/quantitativ möglich. Im Grundsatz ergibt sich daraus nichts Neues, wenn man – wie soeben erläutert – die quantitativen Beobachtungen in Klassen gruppiert.

Beispiel

Interessiert man sich für die Körpergröße x in Abhängigkeit vom Geschlecht, so ist ein Merkmal quantitativ und das andere qualitativ. Man nimmt also eine Klasseneinteilung für das quantitative Merkmal vor, z. B. die oben vorgestellte. Jede Kombination

(Ausprägung des qualitativen Merkmals; Klasse des quantitativen Merkmals)

besetzt dann ein Tabellenfeld. Die Tabelle hat diese Struktur:

Kreuztabelle von Körpergrößen und Geschlecht

	$x \leq 150$	$150 < x \leq 160$	$160 < x \leq 170$	$170 < x \leq 180$	$180 < x \leq 190$	$190 < x \leq 200$	$x > 200$
weiblich							
männlich							

Lernkontrolle zu 2.3

- Die Anzahl der Ehen von berühmten Filmschauspielern ...
 - ... ist ein diskretes Merkmal.
 - ... ändert sich kontinuierlich und ist daher ein kontinuierliches Merkmal.
 - ... ist kein stetiges Merkmal.
 - ... ist kein diskretes Merkmal, weil Diskretion in diesem Bereich unbekannt ist.
- Mehrfaches Vorkommen derselben Merkmalsausprägung ist eher typisch ...
 - ... für qualitative Merkmale.
 - ... für quantitative Merkmale.
 - ... für verhältnisskalierte Merkmale.
 - ... für unbekannte Merkmale.
- Bei der tabellarischen Darstellung quantitativer Merkmale ...
 - ... nimmt man selten eine Klasseneinteilung vor.
 - ... nimmt man Klasseneinteilungen vor, weil solche Merkmale fast immer auf natürliche Weise in Klassen zerfallen.

- c. ... nimmt man Klasseneinteilungen vor, um zu vermeiden, dass sehr oft nur die Häufigkeit 1 einzutragen ist.
- d. ... sollte man immer in mindestens zwanzig Klassen unterteilen.
- e. ... richtet sich die Breite jeder einzelnen Klasse nach der Häufigkeit ihrer Belegung.
- f. ... kann man die Tabellierung kumulierter Häufigkeiten in Betracht ziehen.



Lösung LK 2.3

2.4 Grafische Aufbereitung quantitativer Daten

Wie für die tabellarische Aufbereitung ergibt sich auch hier gegenüber dem Fall qualitativer Merkmale nichts grundsätzlich Neues. Im Wesentlichen wird es darauf ankommen, zunächst eine Gruppierung zu Klassen vorzunehmen und dann die für quantitative Daten passenden Diagrammformen anzuwenden.

Anstelle der oben besprochenen Säulendiagramme/Balkendiagramme wählen wir eine leicht modifizierte Darstellungsform. Die Säulen zu qualitativen Merkmalen hatten wir durch Zwischenräume getrennt und wollten damit auch optisch zum Ausdruck bringen, dass die Merkmale in keiner Beziehung oder doch höchstens in einer Ordnungsbeziehung zueinander standen. Die Klassen quantitativer Merkmale sind hingegen aneinander anschließende Intervalle reeller Zahlen oder Mengen diskreter Werte aus solchen Intervallen. Es ist demzufolge intuitiv, die „Säulen“ über diesen Intervallen die ganze Intervallbreite annehmen zu lassen. Das hat zur Folge, dass die Säulen ohne Zwischenräume aneinandergrenzen. Wir nennen diese Diagrammform *Histogramm*.

In einem Histogramm ist der Flächeninhalt immer proportional zur Häufigkeit. Die Säulenhöhe stellt die sogenannte *Häufigkeitsdichte* dar, das ist die Häufigkeit je Maßeinheit der waagerechten Achse. Fallen z. B. in unserer Beobachtungsreihe die Körperlängen von 120 der gemessenen Personen in die Klasse $170 < x \leq 180$, dann haben wir es in diesem Größenbereich mit einer Häufigkeitsdichte von 120 Personen auf 10 cm zu tun.

Wollte man ganz präzise sein, müsste man also die entsprechende Achse eines Histogramms auch mit einem Begriff beschriften, der die *Häufigkeitsdichte* bezeichnet, z. B. „Anzahl Personen je 10 cm-Intervall der Körpergröße“ und nicht nur „Anzahl der Personen“. In wissenschaftlichen Arbeiten kann eine solche begriffliche Präzision erwünscht oder notwendig sein, in einer Tageszeitung wirkt sie vermutlich bestenfalls albern, schlimmstenfalls unverständlich.

Folgt man der in Abschnitt 2.3 ausgesprochenen Empfehlung konstanter Klassen-/Intervallbreite und trägt in geeignetem Maßstab die Klassenhäufigkeit auf der senkrechten Achse auf, dann gibt nicht nur der Flächeninhalt, sondern auch die Höhe der jeweiligen Säule einen optischen Eindruck von dieser Häufigkeit.

Da die Säulenbreite immer das gesamte zugehörige Merkmalsintervall überdecken soll, ist klar, dass zur Darstellung im Histogramm „offene“ Randklassen ungeeignet sind. Man könnte zwar durch eine geeignete Darstellung noch grafisch andeuten, dass eine Klasse einseitig unbegrenzt sein soll, für die Säulenhöhe ergibt sich aber aus der Forderung, dass der Flächeninhalt proportional zur Klassenhäufigkeit ist, kein sinnvoller Wert mehr. Man wird daher zum Zweck der grafischen Veranschaulichung anstreben, ohne die offenen Randklassen auszukommen, d. h. alle vorkommenden Werte durch geschlossene Klassen abzudecken.

Wir illustrieren die Ausführungen am Beispiel von Körpergrößen.

Körpergröße	Absolute Häufigkeit
≤ 150	3
$150 < x \leq 160$	50
$160 < x \leq 170$	80
$170 < x \leq 180$	120
$180 < x \leq 190$	90
$190 < x \leq 200$	30
$x > 200$	5

Tabelle 2.9
Absolute Häufigkeitsverteilung von Körpergrößen

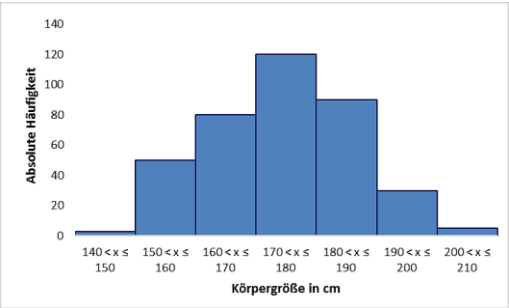


Abbildung 2.7
Histogramm absolute Häufigkeitsverteilung von Körpergrößen

Kommt bei den Rohdaten z. B. kein Wert unter 140 und keiner über 210 vor, dann benötigen wir die offenen Randklassen nicht wirklich und können sie jeweils durch eine geschlossene Klasse am unteren und oberen Rand ersetzen. Allgemein sollte zum Zweck der grafischen Veranschaulichung die unterste und oberste Klasse geschlossen sein und so gewählt werden, dass alle Werte durch die Klassen abgedeckt werden.

Bei quantitativen Merkmalen ist es häufig sehr aufschlussreich, sich mit den kumulierten Häufigkeiten zu beschäftigen. Die grafische Darstellung sieht so aus, dass man den Wert „Summe aller Häufigkeiten bis einschließlich dieser Klasse“ dem Maximalwert jeder Klasse zuordnet, falls es sich um eine stetige Variable handelt. Diese Punkte werden dann miteinander verbunden. Bei einer diskreten Variablen wird die kumulierte Häufigkeit dem Durchschnitt zwischen dem Maximalwert einer Klasse und dem Minimalwert der nächsten Klasse

zugeordnet. Haben beispielsweise 30 % der Fünfjährigen bis zu 10 neue Zähne und die nächste Klasse fängt bei 11 Zähnen an, so wird der Datenpunkt bei 10,5 Zähnen und einer prozentualen Häufigkeit 30 % eingetragen. Eine solche Kurve von kumulierten Werten nennt man *Polygonzug*.

Das Beispiel der Körpergrößen ergibt die folgende kumulierte Tabelle und die zugehörige Grafik.

Körpergröße	Kumulierte absolute Häufigkeit
≤ 150	3
≤ 160	53
≤ 170	133
≤ 180	253
≤ 190	343
≤ 200	373
≤ 300	378

Tabelle 2.10
Kumulierte absolute Häufigkeitsverteilung von Körpergrößen

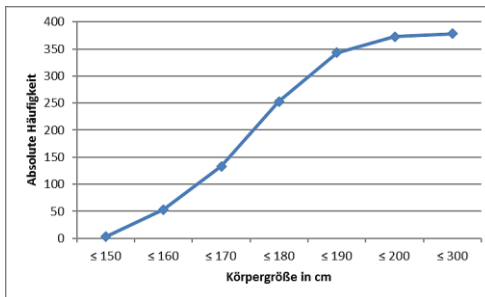


Abbildung 2.8
Polygonzug kumulierte absolute Häufigkeitsverteilung von Körpergrößen

Es sind also z. B. 253 Personen der Stichprobe 1,80 m groß oder kleiner.

Für Kreisdiagramme ergibt sich bei quantitativen Daten nach Unterteilung der Werteskala in Klassen und Zuordnung der Beobachtungen zu den Klassen grundsätzlich nichts Neues gegenüber dem Fall qualitativer Daten, sofern alle Klassen gleich breit gewählt wurden. Andernfalls ist beabsichtigten oder unbeabsichtigten Fehlinterpretationen Tür und Tor geöffnet. Verwenden Sie Kreisdiagramme bei quantitativen Daten höchstens ausnahmsweise.

Wir wenden uns schließlich den Beobachtungen von Merkmalspaaren (x, y) zu. Wir nehmen jetzt an, dass beide Merkmale quantitativ sind. Dann sind in der Regel für jedes Merkmal viele Werte und entsprechend sehr viele Wertepaare möglich. Von Ausnahmefällen abgesehen wird jedes beobachtete Wertepaar einen eigenen Punkt in der x - y -Ebene darstellen. Die

geeignete Visualisierung ist ein *Streudiagramm* oder *Streuungsdiagramm*. (In der Literatur finden sich beide Bezeichnungen.) Daraus können häufig gute erste Hinweise für die Beziehung der Merkmale zueinander gewonnen werden.

Beispiel

Zusätzlich zur Körpergröße ist auch das Gewicht jeder Person in einer Stichprobe erfasst worden.

Tabelle 2.11 Körpergröße in cm und Körpergewicht in kg – Rohdaten

Fall	Körpergröße	Körpergewicht	Fall	Körpergröße	Körpergewicht	Fall	Körpergröße	Körpergewicht
1	172	68	8	167	70	15	179	83
2	191	85	9	180	75	16	169	68
3	155	65	10	189	77	17	178	82
4	175	80	11	199	102	18	179	70
5	177	75	12	167	65	19	184	88
6	188	83	13	181	83	20	163	66
7	165	65	14	187	95	21	201	120

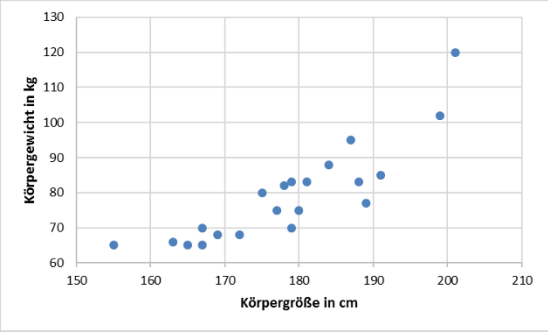


Abbildung 2.9 Streudiagramm Körpergröße und Körpergewicht

Das Streudiagramm legt die Vermutung nahe, dass größere Menschen tendenziell mehr wiegen. Aha.

Computerübung 2.4

Sie werden mit der Diagrammerzeugung in Excel inzwischen so gut vertraut sein, dass Sie die Abbildungen 2.7 bis 2.9 ohne nähere Erläuterungen nachvollziehen können. In Excel heißen die entsprechenden Diagrammtypen Histogramm, Linien- und Punktdiagramm. Dabei sollten Sie auch etwas mit den verschiedenen Einstellungsmöglichkeiten spielen, um ein Gefühl für das Werkzeug zu erhalten. Erstellen Sie zusätzlich ein Histogramm aus den Größenwerten der Tabelle 2.11. Sie können dazu in Excel zwei Wege gehen: Entweder legen Sie eine Tabelle der gewünschten Klassengrenzen an, verwenden die Funktion **HÄUFIGKEIT** (Beachten Sie: Das ist eine Matrixfunktion – oder auch Arrayfunktion – und muss mit **STEUERUNG + UMSCHALTEN + EINGABE** abgeschlossen werden.) und lassen ein Säulendiagramm zeichnen, bei dem Sie im Nachhinein die Säulenabstände auf 0 festsetzen (Doppelklick auf Diagramm/Reihenoptionen/Abstandsbreite = 0), oder Sie gehen von den Rohdaten aus und wählen die Grafik „Histogramm“, in der Sie die Anzahl der Säulen festlegen können.

In R verwendet die Leserin die Befehle `hist`, `barplot` und `plot`.

Über die zentrale Linkliste oder die Links aus der Box „Computerübungen“ von Kapitel 1 erreichen Eilige die Dateien mit bereits eingetragenen Daten und für R erforderlichenfalls auch die Lösungen.

Videos 2.3 und 2.4

Diese beiden Videos zeigen die tabellarische und grafische Aufbereitung quantitativer Daten. Wir nutzen dabei Microsoft Excel.

Sie brauchen etwa 15 Minuten bzw. 10 Minuten, um sich die Videos anzusehen.



Lernkontrolle zu 2.4

1. Hat man quantitative Daten beobachtet und will diese nun grafisch darstellen, ...
 - a. ... lässt man im Unterschied zur Darstellung qualitativer Daten keinen Zwischenraum zwischen den Säulen der Klassenhäufigkeiten.
 - b. ... werden die Merkmalsklassen normalerweise auf der x-Achse eingetragen.
 - c. ... dann beschreibt bei konstanter Klassenbreite die Säulenhöhe die Häufigkeit.
 - d. ... nennt man die Art der Darstellung ein Histogramm.
2. Bei quantitativen Merkmalen ...
 - a. ... kann man immer kumulierte Häufigkeitsdiagramme zeichnen.
 - b. ... sind kumulierte Häufigkeitsdiagramme nie sinnvoll.
 - c. ... sollten kumulierte Häufigkeitsdiagramme grafisch durch ein Balkendiagramm dargestellt werden.
 - d. ... sollten kumulierte Häufigkeitsdiagramme grafisch durch ein Histogramm dargestellt werden.

3. Bei quantitativen Merkmalen ...
- a. ... sind Streudiagramme ein Mittel zur Veranschaulichung der Beobachtungen eines einzelnen Merkmals.
 - b. ... verwendet man Streudiagramme bei der Beobachtung von zwei Merkmalen.
 - c. ... erfordern Streudiagramme zwingend eine vorausgehende Klasseneinteilung.
 - d. ... können Streudiagramme Einsichten bezüglich der Beziehung zweier beobachteter Merkmale vermitteln.



Lösung LK 2.4

Zusammenfassung

Sie wissen jetzt, dass es Unterschiede in der Aufbereitung qualitativer bzw. quantitativer Daten gibt. Sie wissen außerdem, dass die Aufbereitung von Beobachtungen mit einem Merkmal andere Hilfsmittel erfordert als es bei zwei Merkmalen der Fall ist. Ihnen ist bekannt, dass die Darstellung kumulierter Häufigkeiten mindestens eine Ordinalskala voraussetzt. Sie wissen, dass bei der Aufbereitung quantitativer Merkmale regelmäßige Klasseneinteilungen erforderlich sind und kennen Faustregeln für eine solche Einteilung. Die Unterschiede zwischen absoluter, relativer und prozentualer Häufigkeit sind Ihnen klar. Im Einzelnen haben Sie eindimensionale Häufigkeitstabellen und Kreuztabellen, Säulen und Balkendiagramme, Kreisdiagramme, Blasendiagramme, Histogramme, Polygonzüge und Streudiagramme kennengelernt.

Übungsaufgaben zu Kapitel 2

Auch hier haben Sie wieder Gelegenheit, die gewonnene Sicherheit auf dem behandelten Gebiet noch einmal zu überprüfen und zu festigen. Der QR-Code führt Sie zu einer Reihe von Aufgaben, die Ihnen so oder ähnlich auch in Klausuren begegnen könnten. Am Ende der Aufgaben-Datei finden Sie die Lösungen oder Lösungsvorschläge. Sofern sich die Bearbeitung mit Hilfe eines Kalkulationsprogramms anbietet, enthalten die Lösungen auch einen Link zu einer Excel-Tabelle.



Statistik für Wirtschaftswissenschaftler

Ein Lehr- und Übungsbuch für das Bachelor-Studium

Schuster, Th.; Liesen, A.

2017, XII, 261 S. 80 Abb., Softcover

ISBN: 978-3-662-49835-4