# Teil I: Beschreibende Statistik

# 1 Grundbegriffe

**In diesem Kapitel…**

- Grundgesamtheit, Stichprobe und Repräsentativität
- Merkmalsträger, Merkmale, Merkmalswerte und Merkmalsausprägungen
- Merkmalsarten Qualitativ und Quantitativ
- Skalierung von Merkmalen
- Urliste und Häufigkeitsverteilung
- Gruppieren, Kumulieren, Klassieren und Symmetrie
- Grafische Darstellungen

Um Statistik erfolgreich zu lernen, müssen Sie einige Begriffe kennen. Ich werde Ihnen diese nach und nach vorstellen und Ihnen auch sagen, was es mit diesen Fachbegriffen auf sich hat. Dabei führe ich Sie auch in die statistische Nomenklatur ein, einfach gesagt: Ich sage Ihnen, welches Zeichen in einer statistischen Formel oder Gleichung was bedeutet. Am Ende des Kapitels sollten Sie die Aufgabenstellung einer Statistikfrage richtig lesen, verstehen und lösen können.

## Grundgesamtheit, Stichprobe und Repräsentativität

### Grundgesamtheit

Bei einer statistischen Untersuchung müssen wir zuerst einmal die Grundgesamtheit definieren, für die Aussagen getroffen werden sollen. Als Symbol für die *Anzahl der Objekte* in einer Grundgesamtheit, z. B die Anzahl aller registrierten Fahrzeuge in Deutschland, nehmen wir den großen lateinischen Buchstaben *„N“* (es gibt N = 61,5 Millionen registrierte Fahrzeuge in Deutschland). Weitere Beispiele für Grundgesamtheiten sind: Die Menge aller Per-

sonen wohnhaft in Deutschland oder die Menge aller Facebook-Nutzer auf der Welt.
In der Praxis sind die Grundgesamtheiten meist nicht so einfach zu bestimmen, Sie müssen sie zu Anfang eindeutig definieren.

### Stichprobe

Eine Stichprobe ist eine ausgewählte Teilmenge der Grundgesamtheit, auch *Teilgesamtheit* genannt. Diese Stichprobe sollte die Grundgesamtheit repräsentieren, wir sagen auch *repräsentativ* sein. Als Symbol für die Anzahl der Elemente in einer Stichprobe nehmen wir den kleinen lateinischen Buchstaben *„n“*.

### Repräsentativität

Repräsentativität werden Sie in der Realität höchstens annähernd erreichen können. Was wir in der Regel damit meinen, ist die Auswahl einer Teilgesamtheit so vorzunehmen, dass aus dem Ergebnis der Teilerhebung möglichst exakt auf die Verhältnisse der Grundgesamtheit geschlossen werden kann.

#### Zusammenfassung Grundgesamtheit, Stichprobe und Repräsentativität

Die wichtigste Eigenschaft einer Stichprobe ist, dass sie die Grundgesamtheit bezüglich relevanter Untersuchungskriterien repräsentiert.

## Merkmalsträger, Merkmale, Merkmalswerte und Merkmalsausprägungen

### Merkmalsträger

Merkmalsträger bezeichnen, wie der Name schon sagt, die Träger von Merkmalen. Aus einer Grundgesamtheit mit N Merkmalsträgern, z. B. die Menge aller Kunden einer Großbank (N =1,2 Millionen), werden bei einer statistischen Stichprobenuntersuchung alle Merkmalsträger n mit einem Vermögen von über 5 Millionen Euro ausgewählt (n = 25.000). Typische Merkmalsträger sind Käufer, Bankkunden, Fernseher, Autos, chemische Verbindungen, Patienten, usw.

## Merkmale

An Merkmalsträgern werden ein oder mehrere Merkmale, also Eigenschaften untersucht. Merkmale werden auch *Variablen* (Veränderliche) genannt und mit lateinischen Großbuchstaben bezeichnet, gerne zum Beispiel mit X oder Y. Zum Beispiel kann das Alter so bezeichnet werden: X = Alter. Weitere Merkmale können zum Beispiel sein: Kreditwürdigkeit, Preis, PS, Farbe, Schweregrad einer Krankheit.

## Merkmalswerte und Merkmalsausprägungen

Wir bezeichnen die Merkmalswerte eines Merkmals X mit einem lateinischen Kleinbuchstaben x. Die Merkmalswerte sind schlicht und einfach die Daten, mit denen Sie weitere Untersuchungen anstellen möchten. Ein Merkmalswert von 38 Jahren kann zum Beispiel so ausgedrückt werden: x = 38 (Jahre). Wenn Sie mehrere Merkmalswerte von n Merkmalsträgern an einem Merkmal beobachten, zum Beispiel das Alter X in Jahren von n Personen, nummerieren Statistiker gerne mit einem sogenannten Index *i* durch. Das i fängt bei 1 an und geht bis zu der Anzahl der Merkmalsträger n in der Untersuchung. Wir schreiben also: Merkmalswerte $x_i$ (mit i = 1, ..., n). Zum Beispiel könnten wir folgende n = 100 Merkmalswerte für das Merkmal Alter der Reihe nach erhoben haben: $x_1 = 38$, $x_2 = 22$, $x_3 = 41$, $x_4 = 23$, $x_5 = 63$, ..., $x_{n=100} = 33$.
Merkmalsausprägungen sind die theoretisch möglichen Werte eines Merkmals und werden zur Unterscheidung von den tatsächlich vorgekommenen Merkmalswerten $x_i$ mit dem Index j (mit j = 1, ..., m) abgekürzt, geschrieben $x_j$. Im vorherigen Beispiel über das Alter X in Jahren von n = 100 Personen könnte die Merkmalsausprägung x = 10 theoretisch möglich gewesen sein, praktisch ist sie aber bei den 100 Personen nicht vorgekommen. Hingegen könnte es sein, dass die Merkmalsausprägung x = 33 mehrfach vorgekommen ist. Wenn jeder Merkmalswert genau ein einziges Mal vorkommt, gilt m = n, jedoch gilt immer $m \leq n$. Ansonsten gibt m an, wie viele unterschiedliche Merkmalsausprägungen vorkommen.

### Zusammenfassung Merkmalsträger Merkmale, Merkmalswerte und Merkmalsausprägungen

Beispiele für diese Bezeichnungen finden Sie in dieser Tabelle:

| Grundgesamtheit | Merkmalsträger | Merkmal | Merkmals-ausprägungen |
|---|---|---|---|
| Studenten in Deutschland | Einzelner Student | Familienstand | ledig, verpartnert, verheiratet |
| Kunden einer Bank | Einzelner Kunde | Kreditwürdigkeit | uneingeschränkte -, bedingte -, fehlende Kreditwürdigkeit |

**Tabelle 1.1** Zusammenfassung Grundgesamtheit, Merkmalsträger, Merkmal und Merkmalsausprägungen

## Merkmalsarten Qualitativ und Quantitativ

Merkmalswerte kommen grob zusammengefasst in zwei unterschiedlichen Arten vor, entweder als Wörter, zum Beispiel „männlich“ oder „gefällt mir gut“, oder als Zahlen, zum Beispiel „30 (Grad Celsius)“ oder „46 (Jahre)“. Daher gibt es zwei Merkmalsarten, wir nennen sie qualitative und quantitative Merkmale.

Unter **qualitativen Merkmalen** verstehen wir Merkmale, die eine Beschreibung (Qualität) darstellen. Meist haben solche Merkmale Ausprägungen, die mit Wörtern umschrieben sind. In der englischsprachigen Literatur wird der Begriff „qualitative“ auch gerne mit **„kategorial“** gleichgesetzt.

Unter **quantitativen Merkmalen** verstehen wir Merkmale, die ein Ausmaß darstellen. Meist haben solche Merkmale Ausprägungen, die in Zahlen gemessen sind.

**Tipp**

**Unterschied Qualitativ und Quantitativ**

Eine Kollegin der Nelson Mandela Metropolitain University in Pt. Elizabeth, Südafrika, hat während ihrer Lehrzeit an der Fachhochschule Münster im Marketing-Masterstudiengang ihren Studenten den Unterschied sehr anschaulich erklärt: „There is one differentiation between qua**L**itative and qua**N**titative. It is the letter „L“ and „N“ which stand for the difference between both. „L“ stands for Letters and „N“ for Numbers.“ Leider funktioniert die Erklärung in der Übersetzung nicht, aber ich hoffe, sie hilft Ihnen trotzdem.

## Quantitative Merkmalswerte: Diskrete und Stetige

Zweckmäßig unterscheidet wir zwei Typen von quantitativen Merkmalen: **Diskrete und Stetige Merkmale**

- Diskrete Merkmale können nur bestimmte Werte annehmen, die streng voneinander getrennt sind, sodass keine Zwischenwerte möglich sind. Ein typisches Beispiel ist die Anzahl von Büchern im Regal, die Anzahl von Kindern, usw.
- Stetige Merkmale können jeden Wert auf einem Intervall als Ausprägung annehmen. Beispiele dafür sind Gewichte, Zeiten, Prozentangaben, usw.

Es gibt häufiger den Fall, dass ein Merkmal diskret mit einer extrem hohen Anzahl möglicher Ausprägungen vorliegt. In einem solchen Fall spricht dann von **quasi-stetig**. Zum Beispiel ist die Einwohnerzahl in Deutschland ein diskretes Merkmal. Zweckmäßig betrachten Sie es trotzdem als (quasi-)stetig, da Sie es so besser analysieren können. Dabei können Ergebnisse vorkommen, die etwas erklärungsbedürftig sind, zum Beispiel dass die Fruchtbarkeitsrate in Deutschland bei 1,38 Kindern liegt. Eine Frau kann natürlich nicht 1,38 Kinder gebären, das ist aber der Preis den wir für die quasi-stetige Betrachtung bezahlen, um zu einem schnell erfassbaren Ergebnis zu kommen.
Manchmal fällt es schwer, ein Merkmal in eine der beiden Kategorien einzuteilen, weil es gute Gründe gibt, es der einen oder der anderen zuzuordnen.

**Tipp**

**Unterschied Diskret und Stetig**
Wenn wir es uns ganz einfach machen möchten, dann wäre vielleicht folgende Erklärung hilfreich. Stellen Sie sich vor, Sie zeichneten die Merkmalswerte eines stetigen und eines diskreten Merkmals in ein XY-Koordinatensystem. Dann würden Sie beim stetigen Merkmal eine Funktion zeichnen und den Stift dabei nicht absetzen, Sie zeichnen eine Linie über ein vorgegebenes Intervall durch. Anders verhält es sich bei einem diskreten Merkmal. Hier würden Sie Ihren Stift immer wieder absetzen, um bei einem neuen X-Wert neu anzufangen und die „Lücken" zwischen den diskreten Ausprägungen nicht durchzeichnen.

## Skalierung von Merkmalen

Merkmale können entweder **nominal** oder **ordinal** oder **metrisch** skaliert sein. Nominal skalierte Merkmale sind am niedrigsten skaliert, metrisch skalierte Merkmale am höchsten. Wir meinen damit, dass wir mit am niedrigsten skalierten Merkmalen die wenigsten Rechenoperationen durchführen können. Ordinal skalierte Merkmale liegen dazwischen.

- Nominal skalierte Merkmale besitzen keine natürliche Rangfolge, wir können sie nicht der Größe nach anordnen. Zum Beispiel ist „männlich“ nicht größer oder kleiner als „weiblich“. Nominal skalierte Merkmale mit genau zwei Ausprägungen nennen wir **binär** oder **dichotom**. Bei mehr als zwei Ausprägungen nennen wir sie **multinomial**.
- Ordinal skalierte Merkmale können wir der Größe nach anordnen, aber wir können keine sinnvollen Additionen oder Differenzen bilden. Zum Beispiel ist eine Schulnote von „Eins“ besser als eine „Zwei“, doch die Differenz ergibt oft inhaltlich keinen Sinn, weil die Abstände zwischen den Merkmalswerten nicht immer gleich groß sind.
- Metrisch skalierte Merkmale können wir der Größe nach anordnen und wir können auch sinnvolle Additionen oder Differenzen bilden, weil die Abstände der Merkmalswerte gleich groß (äquidistant) sind. Für Metrisch wird auch gerne der Begriff **kardinal** verwendet.

### Intervall- oder verhältnisskaliert

Innerhalb der metrisch skalierten Merkmale unterscheiden wir noch einmal danach, ob das Merkmal keinen natürlichen Nullpunkt hat, also intervallskaliert ist oder doch einen natürlichen Nullpunkt hat, also verhältnisskaliert ist. Typischerweise sind zum Beispiel die meisten ökonomischen Größen, die in Geldeinheiten gemessen werden, verhältnisskaliert.

BEISPIEL

**Unterscheidung intervall- und verhältnisskaliert**
Sie können zum Beispiel mit den üblichen Temperaturangaben in Celsius keine sinnvollen Verhältnisse bilden, weil eine durchschnittliche Temperatur im Juni in Münster von 15 Grad Celsius und in Neapel von 30 Grad Celsius nicht bedeutet, dass es in Neapel doppelt so warm ist wie in Münster. Die Celsius Skala hat keinen natürlichen Nullpunkt, daher ergibt hier eine Verhältnisbildung keinen Sinn. Anhand der Kelvin-Skala (eine andere Temperaturskala) wären Verhältnisvergleiche sinn-

voll, diese Skala hat einen natürlichen Nullpunkt (er liegt bei –273 Grad Celsius). Dann könnten wir die Temperaturen in Münster (288 Kelvin) mit denen in Neapel (303 Kelvin) ins Verhältnis setzen und die Aussage treffen, dass es in Neapel um etwa 5 % $\left(\frac{303}{288}\right)$ wärmer ist als in Münster.

### Ordinal oder Intervall skaliert

In der Praxis gibt es Merkmale, die sich nicht immer genau einem Skalenniveau zuordnen lassen. So könnte es vorkommen, dass sich bei einem Merkmal nicht sicher belegen lässt, dass es intervallskaliert ist, Sie sich aber sicher sind, dass es mehr als ordinal skaliert ist. Ein Beispiel dafür ist die *Likert Skala*, die gerne in Fragebögen zur Erfassung von Einstellungen gewählt wird (zum Beispiel 1 = absolute Zustimmung, 2 = Zustimmung, 3 = neutral, 4 = Ablehnung, 5 = absolute Ablehnung), die eigentlich ordinal skaliert sind, in der Praxis aber aus rechnerischen Gründen meist als metrisch skaliert ausgewertet werden. Hier könnten wir eine Interpretation auf einer Intervallskala versuchen, diese Annahme aber bei der anschließenden Interpretation berücksichtigen und entsprechend vorsichtig vorgehen.

#### Zusammenfassung Merkmalsarten und Skalierung der Merkmale

Je nach Merkmalsart und Skalierung der Merkmale können unterschiedliche Rechenoperationen an den Merkmalen durchgeführt werden. Nachfolgende Tabelle fasst die unterschiedlichen Rechenoperationen zusammen.

| Merkmalsart | Skalierung | Welche math. Operationen sind möglich | Beispiele |
|---|---|---|---|
| Qualitatives Merkmal | Nominal | aufzählen | Beruf, Geschlecht |
| | Ordinal | aufzählen, ordnen | Ratings, Schulnoten |
| Quantitatives Merkmal (diskret und stetig) | Metrisch (Intervall) | aufzählen, ordnen, Differenzen bilden | Zeitskala, Temperatur (in Celsius) |
| | Metrisch (Verhältnis) | aufzählen, ordnen, Differenzen bilden, Quotienten bilden | Kosten, Alter, Einkommen, Größe, Temperatur (in Kelvin) |

**Tabelle 1.2** Merkmalsarten und Skalierungen

## Urliste und Häufigkeitsverteilung

In den vorangegangenen Abschnitten habe ich Ihnen die Merkmalsarten und die unterschiedlichen Skalierungen von Merkmalen aufgezeigt. Jetzt möchte ich Ihnen erklären, wie Merkmalswerte, also die Daten, übersichtlicher dargestellt werden können.

### Urliste

In der Regel liegen Daten in Form einer Urliste vor. Das ist eine unsortierte Abfolge von Merkmalswerten $x_i$, zum Beispiel dem Merkmal Feinstaubplakette für n = 10 Autos, wie Sie aus angefügter Tabelle ersehen können.

| Modelle | PS | Hubraum | Verbrauch l/100km | Feinstaub-plakette | Herstellungs-land |
|---|---|---|---|---|---|
| BMW 325 | 250 | 2.400 | 14 | 4 | Deutschland |
| VW Touran | 135 | 1.800 | 7 | 4 | Deutschland |
| Toyota Corolla | 135 | 1.500 | 6 | 5 | Japan |
| Volvo XC 90 | 180 | 2.200 | 10 | 4 | Schweden |
| VW Golf | 80 | 1.300 | 6 | 4 | Deutschland |
| Toyota Auris | 80 | 1.500 | 4 | 5 | Japan |
| Toyota Land Cruiser | 175 | 2.400 | 12 | 5 | Japan |
| Mercedes C 180 | 175 | 2.000 | 8 | 3 | Deutschland |
| Volvo XC 60 | 175 | 2.000 | 6 | 5 | Schweden |
| Saab 900 | 135 | 1.800 | 9 | 2 | Schweden |

**Tabelle 1.3** Autodatensatz mit n = 10 Autos und 6 Merkmalen

Diese Merkmalswerte werden zur Weiterverarbeitung üblicherweise der Größe nach aufsteigend sortiert. Sortieren ergibt für nominal skalierte Merkmale natürlich keinen Sinn, für die anderen Skalierungen schon. Hier führt die Sortierung des Merkmals Feinstaubplakette zu der Zahlenreihe: 2, 3, 4, 4, 4, 4, 5, 5, 5, 5.

## Häufigkeitsverteilung

Eine Häufigkeitsverteilung $h(x_j)$ der jeweils auftretenden unterschiedlichen Merkmalsausprägungen $x_j$ erfolgt dann durch Zuordnung der Häufigkeiten der Merkmalswerte $x_i$ zu den Merkmalsausprägungen $x_j$. Einfacher ausgedrückt: wie häufig wurde ein Merkmalswert, zum Beispiel $x = 4$ genannt? Für die absoluten Häufigkeiten gilt dann:

$$h(x_j) = h_j = \text{Anzahl der Merkmalswerte mit der Ausprägung } x_i$$

$$\sum_{j=1}^{m} h_j = n$$

Zum Summenzeichen $\Sigma$: Wir lesen: „Summe aller $h_j$ von $j = 1$ bis $m$“.

**Tipp**

**Das Summenzeichen $\Sigma$ als moderne App**

Das Summenzeichen $\Sigma$ ist der griechische Großbuchstabe Sigma. Viele Studenten lassen sich durch dieses Summenzeichen verwirren, obschon dieses Summenzeichen etwas Ähnliches ist wie eine App (Applikation) auf dem Smartphone. Eine App ist eine Möglichkeit, über eine Abkürzung zu dem zu gelangen, was hinter dieser Applikation steckt und die Applikation dann ausführt, zum Beispiel ein Spiel oder eine Suche. So verhält es sich mit dem Summenzeichen.

Das Summenzeichen wurde von Mathematikern entwickelt und dient als Kurzform für das Aufsummieren vieler Zahlen nach einem Muster, das nach dem Summenzeichen steht. Würden wir (gedanklich) also auf das Summenzeichen klicken, öffnete sich die App und eine Anzahl Summanden würde addiert.

Tabellarisch sieht die absolute Häufigkeitsverteilung für das Merkmal Feinstaubplakette so aus.

| Merkmalsausprägungen $x_j$ | Absolute Häufigkeiten $h_j$ |
|---|---|
| $x_1 = 2$ | $h(x_1) = 1$ |
| $x_2 = 3$ | $h(x_2) = 1$ |
| $x_3 = 4$ | $h(x_3) = 4$ |
| $x_4 = 5$ | $h(x_4) = 4$ |
| $m = 4$ | $n = 10$ |

**Tabelle 1.4** Absolute Häufigkeitsverteilung Feinstaubplakette

$$\sum_{j=1}^{m=4} h_j = 1 + 1 + 4 + 4 = 10$$

Neben den absoluten Häufigkeiten folgt im nächsten Abschnitt die Darstellung der *relativen Häufigkeiten*.

### Relative Häufigkeitsverteilung

Absolute Häufigkeiten sind teilweise nur bedingt aussagekräftig, gerade wenn wir viele Merkmalswerte vorliegen haben, also größere Datensätze. Dann ist es oftmals besser, wenn wir die absoluten Zahlen in Relation (ins Verhältnis) zur gesamten Anzahl der Merkmalswerte setzen. So erhalten wir die relative Häufigkeit:

$$f(x_j) = f_j = \frac{\text{absolute Häufigkeit einer Merkmalsausprägung } h_j}{\text{Anzahl aller Merkmalswerte } n}$$

$$\sum_{j=1}^{m} f_j = \frac{h_1}{n} + \frac{h_2}{n} + \ldots + \frac{h_m}{n} = 1$$

Im Beispiel für das Merkmal Feinstaubplakette mit n = 10 gilt dann:

$$\sum_{j=1}^{m=4} f_j = \frac{1}{10} + \frac{1}{10} + \frac{4}{10} + \frac{4}{10} = 10\,\% + 10\,\% + 40\,\% + 40\,\% = 100\,\% = 1$$

Die zugehörige Tabelle für die absolute und relative Häufigkeitsverteilung sieht dann so aus:

| Merkmalsausprägungen $x_j$ | Absolute Häufigkeiten $h_j$ | Relative Häufigkeiten $f_j$ |
|---|---|---|
| 2 | 1 | 0,1 = 10 % |
| 3 | 1 | 0,1 = 10 % |
| 4 | 4 | 0,4 = 40 % |
| 5 | 4 | 0,4 = 40 % |
| | $\sum h_j = 10$ | $\sum f_j = 1$ |

**Tabelle 1.5** Absolute und relative Häufigkeitsverteilung Feinstaubplakette

### Zusammenfassung Urliste und Häufigkeitsverteilung

In einer Urliste liegen die Daten ungeordnet vor. Um diese Daten besser weiterverarbeiten zu können, ordnen wir diese Daten gerne mit Hilfe einer absoluten/relativen Häufigkeitsverteilung.

## Gruppieren, Kumulieren, Klassieren und Symmetrie

Nachdem ich Ihnen gezeigt habe, wie Sie Daten zur besseren Weiterverarbeitung aufbereiten können, möchte ich Ihnen jetzt verschiedene Möglichkeiten näher bringen, wie Sie Daten rechnerisch weiter komprimieren können, um sie noch übersichtlicher darzustellen.
Je nach Merkmalstyp und Skalierung können die Ausprägungen eines Merkmals zur weiteren Analyse unterschiedlich aufbereitet und dargestellt werden. Das heißt, nicht alle Möglichkeiten sind für alle Merkmalstypen geeignet oder erlaubt. Hier sehen Sie eine Übersicht und anschließend im nachfolgenden Text die dazugehörigen Erklärungen für die jeweiligen Analysemöglichkeiten:

| | | | Gruppieren | Kumulieren | Symmetrie | Klassieren |
|---|---|---|---|---|---|---|
| Qualitativ | Nominal | | x | | | |
| | Ordinal | | x | x | | |
| Quantitativ | Metrisch | Diskret | x | x | x | |
| | | Stetig | x | x | x | x |

**Tabelle 1.6** Gruppieren, Kumulieren, Symmetrie und Klassieren nach Merkmalstyp

### Gruppieren

Eine **gruppierte Häufigkeitsverteilung** erhalten wir, wenn wir alle Merkmalswerte, die zu einer bestimmten Merkmalsausprägung passen, zu einer Gruppe zusammenfassen. Das ist ein intuitives Vorgehen, zum Beispiel die Einteilung in „abgasfreundliche Autos“ (Feinstaubplakette von 1–3) und „abgasunfreundliche Autos“ (Feinstaubplakette von 4–5), um das Zahlenmaterial übersichtlicher zu gestalten. Dieses Gruppieren lässt sich über absolute oder relative Häufigkeiten erstellen. Das Gruppieren ist bei allen Merkmalstypen möglich, besonders sinnvoll aber bei qualitativen und diskreten Merkmalen. Statt gruppieren sagen wir manchmal auch *kategorisieren*.
Ein Beispiel mit den n = 10 Autos aus dem vorherigen Kapitel soll die Gruppierung anhand des ordinal skalierten Merkmals Feinstaubplakette verdeutlichen.

BEISPIEL

**Gruppieren**
Das Merkmal Feinstaubplakette könnte ich in zwei sinnvolle Gruppen aufteilen. Die Gruppe 1 ist die „abgasfreundliche Gruppe“ mit Werten ≤3, während die Gruppe 2 die „abgasunfreundliche“ mit Werten ≥4 ist.

| Modell | Feinstaubplakette | Gruppierung Feinstaubplakette |
|---|---|---|
| BMW 325 | 4 | Gruppe 2 |
| VW Touran | 4 | Gruppe 2 |
| Toyota Corolla | 5 | Gruppe 2 |
| Volvo XC 90 | 4 | Gruppe 2 |
| VW Golf | 4 | Gruppe 2 |
| Toyota Auris | 5 | Gruppe 2 |
| Toyota Land Cruiser | 5 | Gruppe 2 |
| Mercedes C 180 | 3 | Gruppe 1 |
| Volvo XC 60 | 5 | Gruppe 2 |
| Saab 900 | 2 | Gruppe 1 |

**Tabelle 1.7** Beispiel Gruppieren mit Autodatensatz

## Kumulieren

Um kumulierte Häufigkeiten bilden zu können, müssen Merkmalswerte zuerst einmal der Größe nach aufsteigend geordnet werden. Daher müssen die Merkmale mindestens ordinal skaliert sein. Die **kumulierten Häufigkeiten** (absolut oder relativ) geben an, welche Anzahl bzw. welcher Anteil der Merkmalswerte kleiner oder gleich einer Merkmalsausprägung $x_j$ ist.

BEISPIEL

**Kumulieren**

Hier sehen Sie am Beispiel des Autodatensatzes mit dem Merkmal Feinstaubplakette die Berechnung von kumulierten Häufigkeiten.

| Feinstaubplakette $x_j$ | Absolute Häufigkeiten $h_j$ | Kumulierte absolute Häufigkeiten $H_j$ | Relative Häufigkeiten $f_j$ | Kumulierte relative Häufigkeit $F_j$ |
|---|---|---|---|---|
| 2 | 1 | 1 | 10 % | 10 % |
| 3 | 1 | 2 | 10 % | 20 % |
| 4 | 4 | 6 | 40 % | 60 % |
| 5 | 4 | 10 | 40 % | 100 % |
| Gesamt | 10 | | 100 % | |

**Tabelle 1.8** Kumulierte Häufigkeitsverteilung für das Merkmal Feinstaubplakette

Hier gibt es jetzt beispielhaft folgende Interpretationsmöglichkeiten:

- 20 % aller Autos aus dem Datensatz haben eine Feinstaubplakette mit der Schadstoffgruppe 3 und weniger.
- 40 % aller Autos aus dem Datensatz haben eine Feinstaubplakette mit der Schadstoffgruppe größer als 4.
- 90 % aller Autos aus dem Datensatz haben eine Feinstaubplakette mit der Schadstoffgruppe größer als 2.

Kumulierte Häufigkeiten sind also nichts anderes als aufsummierte Häufigkeiten und bieten eine andere Form der Aussage. Die kumulierte Häufigkeitsverteilung wird auch *Summenhäufigkeitsfunktion* oder *empirische Verteilungsfunktion* genannt. Empirisch sind Verteilungsfunktionen, wenn sie aus tatsächlich vorliegenden Daten berechnet werden. Damit soll der Unterschied zum entsprechenden, später im Kapitel 6 erscheinenden, Begriff für *Zufallsvariablen* angedeutet werden.

## Klassieren

Die Klassierung dient der übersichtlichen Darstellung eines stetigen Merkmals mit vielen Merkmalswerten. Die Klassenbildung erfolgt durch die Erzeugung von Klassen mit deren Klassenbreiten und Klassenmitten. Eine Klassierung ist nichts anderes als eine Einteilung aller Merkmalswerte in sinnvolle Klassen. Bei der Klasseneinteilung ist Folgendes zu beachten:

1. Es dürfen sich keine Zwischenräume oder Überschneidungen bei der Klassenbildung ergeben.
2. Die Klassen müssen den gesamten Wertebereich des Merkmals voll abdecken.
3. Die Klassen müssen nicht die gleiche Breite besitzen. In manchen Fällen ist das aber wünschenswert.
4. Die Anzahl der Klassen variiert je nach Anwendungsfall. In der Praxis arbeiten wir gerne mit fünf bis zehn Klassen.
5. Die Klassenbildung sollte robust sein, d. h. eine Änderung der Klasseneinteilung darf keine wesentliche Änderung der Analyseergebnisse nach sich ziehen.

$$\text{Klassen} \qquad K_k = [c_0; c_1), [c_1; c_2), \ldots, [c_{k-1}; c_k]. \quad (k = 1, 2, 3 \ldots)$$

$$\text{Klassenmitten} \qquad KM_k = \frac{c_{k-1} + c_k}{2}$$

$$\text{Klassenbreiten} \qquad KB_k = c_k - c_{k-1}$$

mit k = Anzahl der Klassen.

**Tipp**

**Geschlossene und offene Klammern**
Die geschlossene Klammer „[" bedeutet, dieser Wert ist noch enthalten, die offene Klammer „)" bedeutet, dieser Wert ist nicht mehr enthalten. Zum Beispiel bedeutet: $K_k = [c_0; c_1), [c_1; c_2)$ dass $c_1$ in der ersten Klasse nicht enthalten ist, aber dafür in der zweiten Klasse.

Ein Beispiel aus dem schon erwähnten Autodatensatz, dieses Mal mit dem Untersuchungsmerkmal PS, soll die Klasseneinteilung verdeutlichen.

BEISPIEL

**Klassieren**

Das stetige, quantitative Merkmal PS aus dem Autodatensatz soll in drei Klassen (k = 3) eingeteilt werden.

| k | $K_k$ | $KM_k$ | $KB_k$ | $n_i$ | $f(x_i)$ | $F(x_i)$ |
|---|---|---|---|---|---|---|
| 1 | [0 ; 100) | 50 | 100 | 2 | 20 % | 20 % |
| 2 | [100 ; 200) | 150 | 100 | 7 | 70 % | 90 % |
| 3 | [200 ; 300] | 250 | 100 | 1 | 10 % | 100 % |
| Gesamt | | | 300 | 10 | 100 % | |

**Tabelle 1.9** Klasseneinteilung für das Merkmal PS in drei Klassen

Eine mögliche Interpretation lautet hier, dass 90 % der untersuchten Autos weniger als 200 PS haben.
Jetzt können wir die drei kreierten Gruppen griffig benennen, zum Beispiel als

- 1. Klasse sind „Einsteigerklasse-Autos“
- 2. Klasse sind „Mittelklasse-Autos“
- 3. Klasse sind „Oberklasse-Autos“.

Dieses Beispiel fällt unter den allgemeineren Begriff *Klassifikation*. Damit meinen wir, dass wir ein ursprünglich höher skaliertes Merkmal (meistens ein metrisch skaliertes Merkmal) über eine Klassenbildung in ein niedriger skaliertes (meistens ein ordinales Merkmal) überführen. Der Hauptgrund dafür ist, dass wir dieses Merkmal übersichtlicher und plakativer darstellen möchten.

## Symmetrie

Die Symmetrieeigenschaften einer Häufigkeitsverteilung lassen sich nur sinnvoll bei sogenannten *ein-gipfligen (unimodalen)* Häufigkeitsverteilungen erklären, diese haben nur ein relatives Maximum. Bei sogenannten *mehr-gipfligen (multimodalen)* Häufigkeitsverteilungen ergeben sie keinen Sinn, weil diese mehr als ein relatives Maximum haben.

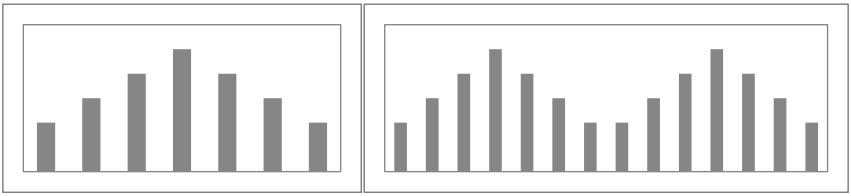

**Abbildung 1.1** Ein-gipflige (links) und zwei-gipflige (rechts) Verteilungen

**Linkssteil, rechtssteil, symmetrisch**
Eine Häufigkeitsverteilung heißt linksteil bzw. rechtsschief, wenn sich die Merkmalswerte am linken Rand der Verteilung häufen. Eine Häufigkeitsverteilung heißt rechtssteil bzw. linksschief, wenn sich die Merkmalswerte am rechten Rand der Verteilung häufen. Eine Häufigkeitsverteilung heißt **symmetrisch**, wenn sich die Merkmalswerte in der Mitte der Verteilung häufen und zu beiden Seiten spiegeln lassen.

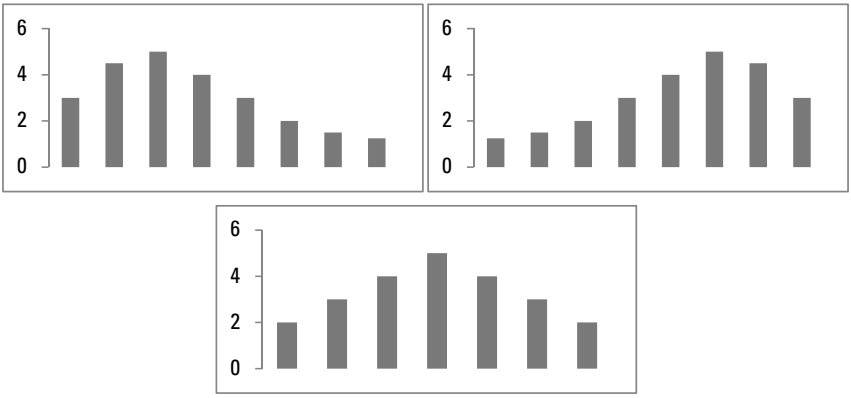


**Abbildung 1.2** Eine beispielhafte linkssteile (oben links), rechtssteile (oben rechts) und symmetrische Häufigkeitsverteilung (unten)

### Zusammenfassung Gruppieren, Kumulieren, Klassieren und Symmetrie

Um Daten für die weitere Verarbeitung zu komprimieren, gibt es je nach Merkmalsart (qualitativ und quantitativ) und Skalierung der Merkmale (nominal, ordinal und metrisch) verschiedene Verfahren zur Gruppen- oder Klassenbildung. Mit Hilfe des Kumulierens und der Symmetrieeigenschaften lassen sich die Daten noch besser beschreiben bzw. die Verteilung der Daten besser charakterisieren.

## Grafische Darstellungen

Neben der Möglichkeit, Daten über Tabellen übersichtlicher darzustellen, gibt es natürlich auch viele grafische Möglichkeiten, dies zu tun. Auf diese Möglichkeiten gehe ich in diesem Abschnitt ein. Grafische Darstellungen benutzen wir, um Daten komprimiert übersichtlich darzustellen. Daher ist es wichtig zu wissen, welche Darstellungsform Sie bei welchen Merkmalen am besten benutzen, zum Beispiel in Microsoft Excel, um einen höchstmöglichen Nutzen zu erhalten. Diese Übersicht gibt eine Anleitung, wie Sie sinnvoll vorgehen können.

| Skalierung | Stab/Säulen/ Balken/Kreis | Stamm-Blatt (Stem-Leaf plot) | Histogramm | Empirische Verteilungs-funktion |
|---|---|---|---|---|
| Nominal | x | | | |
| Ordinal | x | | | |
| Metrisch (diskret und n klein) | x | x | | x |
| Metrisch (diskret und n groß) | | | x | x |
| Metrisch (stetig und n klein) | x | x | | x |
| Metrisch (stetig und n groß) | | | x | x |

**Tabelle 1.10** Grafische Darstellungsarten nach Skalierungsart

Im Einzelfall müssen Sie aber selbst anhand der Datenmenge, Anzahl der Merkmale und anderer Gegebenheiten entscheiden. Ich werde Ihnen nachfolgend diese grafischen Darstellungsarten anhand von Merkmalen aus dem Autodatensatz erklären.

1. **Stab-, Säulen-, Balken- und Kreisdiagramme** sind die bekanntesten Darstellungsformen kategorialer (nominal und ordinal) oder diskreter Merkmale, wenn die Anzahl $m$ der verschiedenen Merkmalsausprägungen klein ist.
   - Bei einem **Stabdiagramm** werden auf der horizontalen Achse die Merkmalsausprägungen abgetragen und auf der vertikalen die absoluten (oder relativen) Häufigkeiten in Form eines Stabes.

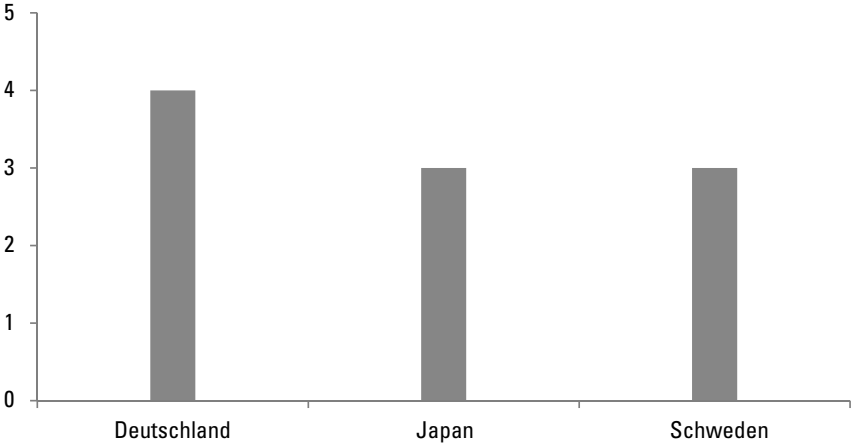


**Abbildung 1.3** Ein Stabdiagramm für das Merkmal Herstellungsland

   - Ein **Säulendiagramm** ist das Gleiche, nur werden die Stäbe durch Rechtecke ersetzt, die mittig über die Merkmalsausprägungen gezeichnet werden und nicht aneinander stoßen sollten.

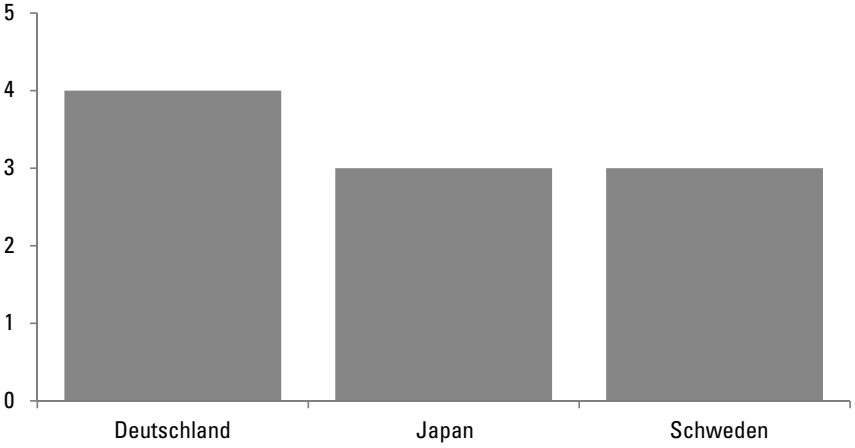


**Abbildung 1.4** Ein Säulendiagramm für das Merkmal Herstellungsland

- Das **Balkendiagramm** ist eine weitere Form des Säulendiagramms, nur werden hier die horizontale und vertikale Achse vertauscht.

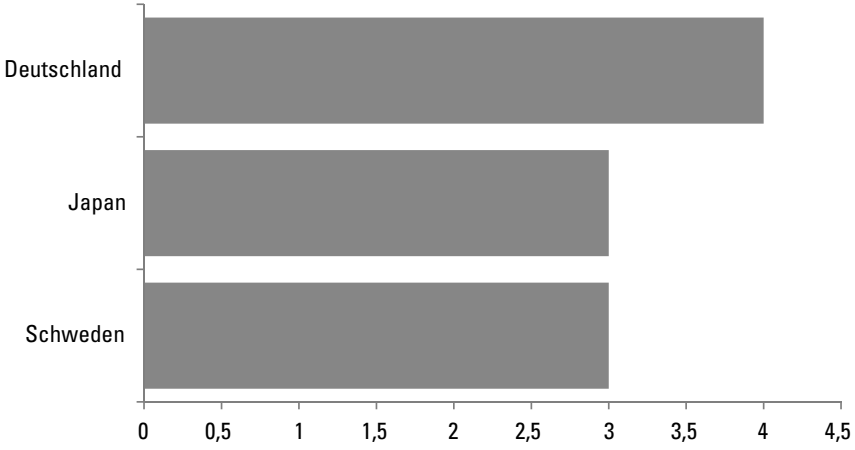


**Abbildung 1.5** Ein Balkendiagramm für das Merkmal Herstellungsland

- Bei einem **Kreisdiagramm** verhalten sich die zugehörigen Flächen einer Merkmalsausprägung proportional zur absoluten (oder relativen) Häufigkeit dieser Merkmalsausprägung. Das nennen wir das *Prinzip der Flächentreue*.

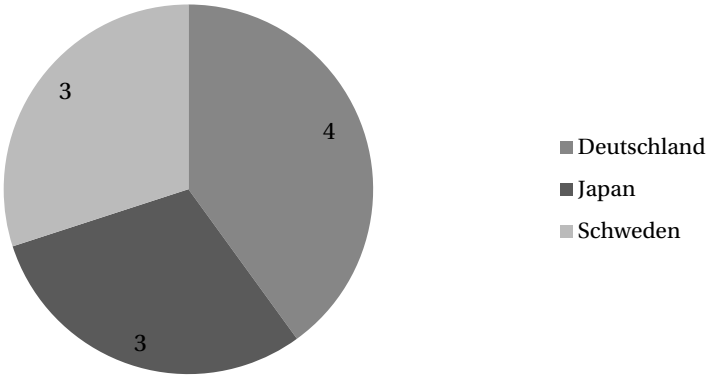


**Abbildung 1.6** Ein Kreisdiagramm für das Merkmal Herstellungsland

2. **Stamm-Blatt-Diagramme** (Stem-leaf plot) sind halbgrafische Darstellungsformen für metrische Merkmale, die für kleine Datenumfänge auch per Hand ausführbar sind. Das Diagramm besteht aus zwei Spalten. In der ersten Spalte steht der Stamm, in der zweiten Spalte die Blätter. Je nachdem wie Sie sich für den Stamm entscheiden, kann ein und derselbe Datensatz zu unterschiedlichen Diagrammen führen. In Bus- und Straßenbahnfahrplänen zum Beispiel werden die Abfahrtszeiten in Form eines Stamm-Blatt-Diagramms dargestellt.
   Anhand des Autodatensatzes möchte ich Ihnen gerne die Einfachheit dieser grafischen Darstellung präsentieren. Dabei beziehe ich mich auf das Merkmal PS. Die Zehnerzahl wird als Stamm genutzt und die Zahl rechts davon (die Einerzahl) als Blatt.

| Stamm | Blatt |
|---|---|
| 8 | 0 0 |
| 13 | 5 5 5 |
| 17 | 5 5 5 |
| 18 | 0 |
| 25 | 0 |

**Abbildung 1.7** Ein Stamm-Blatt Diagramm für das Merkmal PS

Es gibt demnach zwei Autos mit 80 PS, drei Autos mit 135 PS, drei Autos mit 175 PS, eins mit 180 PS und eins mit 250 PS.

3. **Histogramme** werden dann benutzt, wenn die Datenmengen für die bisher genannten Darstellungsarten zu groß werden. Es werden dann immer die Daten *gruppiert* oder *klassiert* (siehe vorheriger Abschnitt) und als Histogramm dargestellt. Auch hier sind die dargestellten Flächen direkt proportional zu den absoluten/relativen Häufigkeiten (*Flächentreue*).

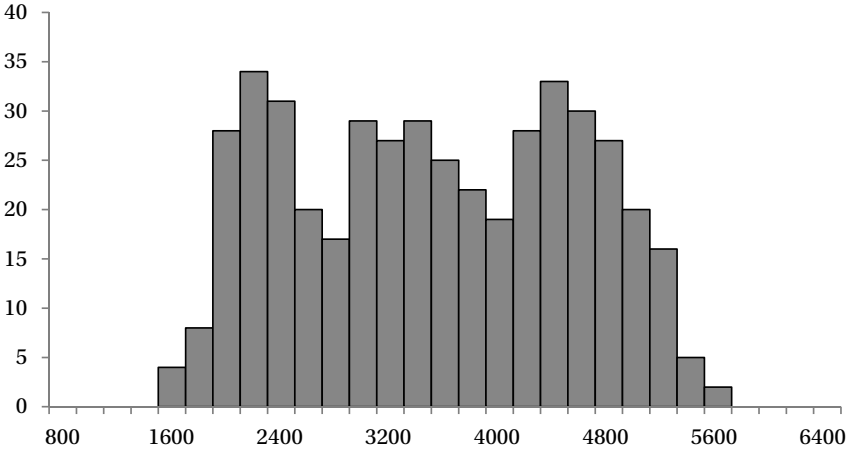


**Abbildung 1.8** Ein Histogramm für das Merkmal Gewicht

4. **Empirische Verteilungsfunktionen** beantworten Fragestellungen nach dem Anteil, der über oder unter einer bestimmten Merkmalsausprägung liegt. Also zum Beispiel: wie viel Prozent der Autos haben höchstens 180 PS? Um diese Frage zu beantworten, sortieren wir die Autos der Größe nach und summieren die Anzahl der Autos, die bis zu 180 PS haben. In der Regel verwenden wir relative statt absolute Häufigkeiten.

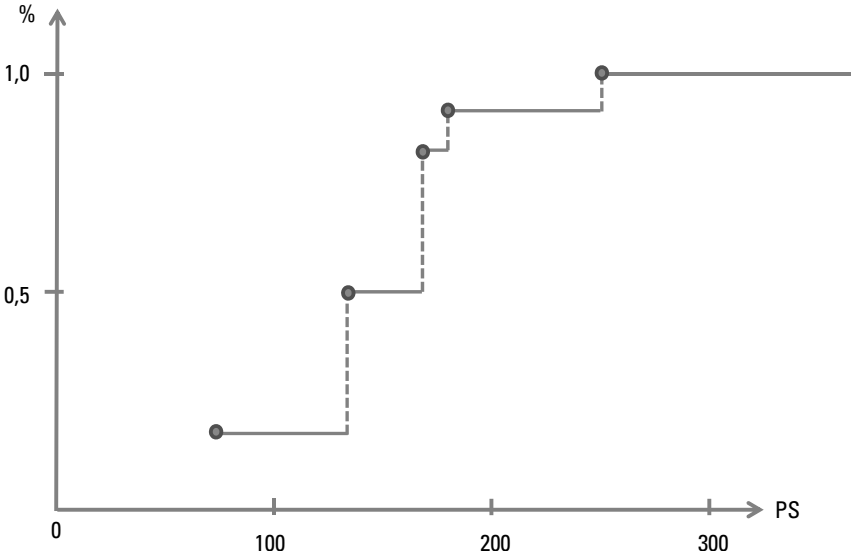


**Abbildung 1.9** Eine empirische Verteilungsfunktion für das Merkmal PS

Der Punkt bei 135 PS zum Beispiel deutet an, dass dieser Wert zur (x, y)-Koordinate (135, 50 %) gehört und das bedeutet, dass 50 % der Autos (5 von 10) höchstens 135 PS haben.

### Zusammenfassung grafische Darstellungen

Neben den tabellarischen Verfahren sind die grafischen Darstellungen ein wichtiges Visualisierungsmittel, um Information anhand von zusammengefassten Daten zu transportieren.

**AUF EINEN BLICK**

- Die wichtigste Eigenschaft einer Stichprobe ist, dass sie die Grundgesamtheit repräsentiert.
- Merkmale können wir hinsichtlich ihrer Merkmalsart als quantitatives oder qualitatives Merkmal beschreiben.
- Merkmale können wir auch hinsichtlich ihrer Skalierungsart in nominal, ordinal oder metrisch skaliert unterscheiden.

- Je nach Merkmalsart und Skalierung können die Ergebnisse zur weiteren Analyse unterschiedlich aufbereitet werden.
- Neben der Möglichkeit, Daten über Tabellen übersichtlicher darzustellen, gibt es je nach Merkmalsart und Skalierung auch grafische Möglichkeiten dies zu tun.

## Übungsaufgaben

### Übung 1.1

Fassen Sie die beiden großen Teilbereiche der Statistik „Beschreibende Statistik“ und „Schließende Statistik“ mit nur jeweils einem Begriff treffend plakativ zusammen.

### Übung 1.2

Eine Gruppe von Studenten in einer Lehrveranstaltung „Angewandte Statistik“ muss eine Hausarbeit zu folgenden Themen schreiben und sich dabei entscheiden, ob sie die Grundgesamtheit oder eine Stichprobe untersucht. Begründen Sie, für welche Erhebungsform Sie sich entscheiden:

1. Untersuchung der Auswirkungen der durchschnittlichen Computerspieldauer pro Woche auf die Schulnoten von Jugendlichen.
2. Qualitätsprüfung hochwertiger Rotweine der Marke Brunello aus der Provinz Montalcino in der Südtoskana.
3. Mechanische Überprüfung der Wirkungsweise von Herzschrittmachern in der Produktion.

### Übung 1.3

Im Kapitel 1 wird in einem Beispiel der Autodatensatz mit n = 10 Merkmalswerten und 6 Merkmalen vorgestellt. Ordnen Sie jedes Merkmal nach Merkmalsart und Skalierungsart ein.

### Übung 1.4

Gegeben sei das Autodatensatzbeispiel aus der vorherigen Übung 1.3. Lösen Sie folgende Aufgaben:

1. Erzeugen Sie eine tabellarische kumulative absolute und relative Häufigkeitsverteilung für das Merkmal Hubraum.
2. Wie viele Autos (absolut und prozentual) haben mehr als 1800 $cm^3$ Hubraum?
3. Wie viele Autos (absolut und prozentual) haben höchstens einen Hubraum von 1500 $cm^3$?
4. Wie viele Autos (absolut und prozentual) haben genau einen Hubraum von 2000 $cm^3$?

### Übung 1.5

Warum unterscheiden wir überhaupt zwischen unterschiedlichen Skalierungen bei Merkmalen? Erklären Sie das an einem Beispiel.

### Übung 1.6

Gegeben seien zwei Merkmalswertereihen von zwei Merkmalen X und Y. Berechnen Sie für diese Wertereihen:

a) $\sum_{i=1}^{8} x_i$, b) $\sum_{i=1}^{8} y_i$, c) $\sum_{i=1}^{8} x_i^2$, d) $\sum_{i=1}^{8} y_i^3$, e) $\sum_{i=1}^{8} x_i \sum_{i=1}^{8} y_i$, f) $\sum_{i=1}^{8} (x_i y_i)$