2

Die Darstellung empirisch erhobener Daten in Tabellen und Diagrammen

In einem ersten Schritt wollen wir Daten **organisieren**, **strukturieren**, **zusammenfassen** und möglichst anschaulich **darstellen** und übersichtlich **präsentieren**. Ausgangspunkt ist dabei die Erfahrung, dass es für die meisten nicht so leicht ist, sich in einem «Zahlenhaufen» zurecht zu finden, aber einen guten Eindruck von der Verteilung der Daten aus einer grafischen Präsentation in **Diagrammen** ablesen können, zumindest aber die Aufbereitung in **Tabellen** erwarten.

2.1 Klassenbildung

Qualitative Daten repräsentieren eine bestimmte Kategorie eines Merkmals; sie gehören zu einer bestimmten «Klasse». Manchmal werden auch quantitative Daten in mehreren Teilbereiche zusammengefasst und *klassifiziert*:

Klassenbildung bedeutet, den Wertebereich der (möglichen) Merkmalswerte in Teilbereiche (*Klassen*) aufzuteilen, und jede Realisierung (also jede Messung, Beobachtung, ...) einer Klasse zuzuordnen. Die Klassen müssen in ihrer Gesamtheit den Wertebereich vollständig (d.h. auch *lückenlos*) überdecken und außerdem einander ausschließen, d.h. es kann nicht sein, dass ein Messwert in mehrere Klassen hineinpassen würde.

Ergebnis der Klassierung der Daten ist letztlich, dass man nicht mehr jeden einzelnen Merkmalswert und die Häufigkeit seines Auftretens angibt, sondern nur noch für jede Klasse die Gesamtanzahl der in ihr enthaltenen Merkmalswerte.

Je weniger Klassen man bildet, desto übersichtlicher und «einfacher» wird die Stichprobe zwar, es geht aber auch ein mehr oder weniger großes Stück an konkreter Information verloren. Umgekehrt: Je größer die Anzahl der Klassen ist, desto unübersichtlicher bleibt die Stichprobe. Üblich sind, je nach Stichprobengröße, in etwa 5 bis 20 Klassen.

Bei quantitativen numerischen Daten sollten die *Klassengrenzen* «runde» und «einfache» Zahlenwerte sein. Die erste und letzte Klasse werden oft als *offene* Klassen geführt, d.h. von $-\infty$ oder 0 (untere Grenze der ersten Klasse) bzw. $+\infty$ (obere Klasse der letzten Klasse) begrenzt. Die *Klassenbreiten* (= obere minus untere Klassengrenze) werden so gewählt, dass sie möglichst gleich lang und die Klassenhäufigkeiten (Anzahl der Messwerte pro Klasse) nicht extrem unterschiedlich sind. (Die Forderung nach gleich großen Klassenbreiten ist nicht zwingend, in den meisten Anwendungsfällen aber üblich).

Bezeichnen wir den größten beobachteten Wert mit x_{max} und den kleinsten mit x_{min}, so ergibt sich die Klassenbreite d bei einer Einteilung in m Klassen zu:

$$d \approx \frac{x_{\max} - x_{\min}}{m} \tag{2.1}$$

wobei bei offenen Klassen x_{min} und x_{max} in den beiden offenen Klassen liegen sollten (also x_{min} in der ersten und x_{max} in der letzten Klasse). Für die Anzahl m der Klassen kann man als Faustregel[1] heranziehen:

$$m \approx \begin{cases} 1 + 3.322 \cdot {}_{10}\log n & \text{für } n \leq 100 \\ 3.322 \cdot {}_{10}\log n & \text{für } n > 100 \end{cases} \tag{2.2}$$

Jedenfalls sollte gelten:

$$2^m \geq n \tag{2.3}$$

Neben der weiter oben erhobene Forderung nach «runden» Klassengrenzen sollten auch die Klassenbreiten d «runde» Zahlen sein (z.B. 2, 5, 10 oder Vielfache von 5 oder 10).

Damit Messwerte nicht genau auf einer Klassengrenze zu liegen kommen, werden üblicherweise die unteren Klassengrenzen in die jeweilige Klasse eingeschlossen, die oberen hingegen ausgeschlossen und zur nächsten Klasse hinzugezählt. D.h. ein Wert, der genau auf einer Klassengrenze liegt, wird immer zur größeren Klasse gezählt. Oder man legt das gleich explizit fest. Wenn du zum Beispiel das Merkmal «Einwohner:innenzahl» von Städten in Klassen einteilen willst, kannst du als Intervallgrenzen wählen (Tab.2.1):

[1] Woran wir sehen: Nicht jede Faustregel kann leicht im Kopf gerechnet werden. . . .
${}_{10}\log n$ bedeutet übrigens: Logarithmus von n zur Basis 10.

Klassennummer i	explizite Angabe der Klassengrenzen (von-bis)		mathematisch elegantere Angabe der Grenzen
1	0	1999	$x < 2000$
2	2000	4999	$2000 \leq x < 5000$
3	5000	19 999	$5000 \leq x < 20\,000$
4	20 000	99 999	$20\,000 \leq x < 100\,000$
5	100 000	$+\infty$	$100\,000 \leq x$

Tabelle 2.1: Zwei Möglichkeiten zur Angabe von Klassengrenzen (Beispiel: Einwohner:innenzahlen)

Auch bei qualitativen Daten sollte die Anzahl der Klassen überschaubar gehalten und gegebenenfalls übergeordnete Klassen aus mehreren Kategorien gebildet werden. Oft werden auch Elemente, die in der Stichprobe nur selten vorkommen, in einer einzigen Klasse zusammengefasst, die den Namen **«andere»** oder **«sonst»** oder ähnliches trägt.

2.2 Darstellung der Daten in Häufigkeitstabellen

Es mag vielleicht ein wenig verwundern, dass ein Kapitel über *Visualisierung* mit einer *Tabelle* beginnt. Tatsächlich haben Tabellen auch Vorteile gegenüber «bildlichen» Darstellungen: Sie sind meist intuitiv erfassbar und können beliebige Datentypen enthalten, auch (beinahe) beliebig viele Zufallsvariablen (sofern die Tabelle noch lesbar bleibt). Nachteilig ist, dass es letztlich Text bleibt und das visuelle System des Menschen visuelle Informationen und Muster viel schneller erfassen und verarbeiten kann als Text, den man erst lesen muss.

Eine **Häufigkeitstabelle** beinhaltet zumindest zwei Spalten: In der linken Spalte stehen alle *möglichen* Merkmalswerte – entweder in Form von Intervallen (Klassen) oder als explizite Angabe, bei numerischen Daten meist in aufsteigender Form geordnet vom kleinsten zum größten Wert. In der rechten Spalte steht oft die Anzahl, wie oft der jeweilige Datenwert in der Stichprobe vorkommt. Letzteres nennen wir auch die **Häufigkeit**. Häufigkeitszahlen können auch den Wert Null haben – offensichtlich dann, wenn dieser Wert in der konkreten Stichprobe nicht vorkommt.

Wir nennen diese tabellarische Beschreibung auch die **Häufigkeitsverteilung** des Merkmals bzw. der Zufallsvariable.

Die Häufigkeitsverteilung kann außer mit Häufigkeitszahlen auch mit der relativen Häufigkeit (meist in Form von Prozentangaben) beschrieben werden und

darüber hinaus mit der zusätzliche Angabe von absoluten oder relativen Häufigkeitssummen ergänzt werden:

Absolute Häufigkeit und kumulierte Häufigkeit

Die **absolute Häufigkeit** f_i ist die Anzahl der Elemente der Stichprobe, die gleich einem vorgegebenen Wert sind oder in eine bestimmte Klasse i von Werten gehören. Es muss gelten:

$$\sum_{i=1}^{m} f_i = n \tag{2.4}$$

das heißt: Die Summe aller absoluten Häufigkeiten muss die Gesamtanzahl aller Werte – also den Umfang der Stichprobe oder Grundgesamtheit – ergeben.

Sowohl in *MS Excel* als auch in *Libre Office Calc* gibt es verschiedene Möglichkeiten, eine Häufigkeitstabelle zu erstellen, je nachdem, ob die Daten klassifiziert werden sollen oder nicht:
Will man Klassen bilden und zählen, wieviele Elemente in der jeweiligen Klasse vorhanden sind, verwendet man (in Excel und in Calc) die Funktion `=HÄUFIGKEIT(Daten; Klassen)`. Bei unklassifizierten Daten verwendet man zum Erstellen einer Häufigkeitstabelle `=ZÄHLENWENN(Bereich wo die Daten stehen; Kriterium welche Daten daraus mitgezählt werden sollen)`

Die **absolute Häufigkeitssumme** (auch: *kumulierte Häufigkeit*) F_i ist die Anzahl der Beobachtungswerte, die einen vorgegebenen Wert (nämlich die Klassengrenze der $i-$ten Klasse) nicht überschreiten.

Wir erhalten die kumulierten Häufigkeiten, indem wir in der Tabelle neben den absoluten Häufigkeiten eine neue Spalte einfügen und dort in jeder Zeile alle bisherigen absoluten Häufigkeiten (also die f_i) zusammenzählen (siehe Tab. 2.2).

Relative Häufigkeit und relative Häufigkeitssumme

Die **relative Häufigkeit** h_i ist die absolute Häufigkeit dividiert durch den Stichprobenumfang:

$$h_i = \frac{f_i}{n} \tag{2.5}$$

Es muss gelten:

$$\sum_{i=1}^{m} h_i = 1 \tag{2.6}$$

das heißt die Summe aller relativen Häufigkeiten muss 1 ergeben.

Sehr oft geben wir relative Häufigkeiten auch prozentuell an, indem wir jedes h_i mit 100 multiplizieren und das Prozentzeichen % hinzufügen. Die Summe aller relativen Häufigkeiten ist dann 100%.

Die **relative Häufigkeitssumme** (auch: *kumulierte relative Häufigkeit*) H_i ist die jeweilige absolute Häufigkeitssumme dividiert durch den Stichprobenumfang.

Wir erhalten die kumulierten relativen Häufigkeiten, indem wir in der Tabelle neben den relativen Häufigkeiten eine neue Spalte einfügen und dort in jeder Zeile alle bisherigen relativen Häufigkeiten (also die h_i) zusammenzählen (siehe Tab. 2.2).

Beispiel 1 *Die Darstellung der Häufigkeitsverteilung einer Stichprobe in einer Häufigkeitstabelle:*

i	Klassengrenzen	f	F	h	H
1	$160 \leq x < 165$	1	1	0.042	4%
2	$165 \leq x < 170$	3	4	0.125	17%
3	$170 \leq x < 175$	4	8	0.167	33%
4	$175 \leq x < 180$	8	16	0.333	67%
5	$180 \leq x < 185$	3	19	0.125	79%
6	$185 \leq x < 190$	4	23	0.167	96%
7	$190 \leq x < 195$	0	23	0	96%
8	$195 \leq x$	1	24	0.042	100%
Summe		24		1	

Tabelle 2.2: Häufigkeitstabelle zu erhobenen Daten über die Körpergröße (in *cm*)

Wir können aus Tabelle 2.2 zum Beispiel herauslesen:
33% sind kleiner als 175 *cm. Oder:*
8 Personen der Stichprobe sind größer oder gleich 175 *cm aber kleiner als* 180 *cm.*
Wir wissen aber zum Beispiel nicht, wie viele Personen 181 *cm sind; die sind in der Klasse* $180 \leq x < 185$ *«versteckt» und es könnte sein, dass kein einziger* 181 *cm groß ist, oder aber 1, 2 oder 3 Personen.*

Beachte: statt $[160 \leq x < 165]$ *schreiben wir manchmal auch* $[160 - 164]$*, statt* $[165 \leq x < 170]$ $[165 - 169]$ *etc. Zwischen* 164 *aus der ersten und* 165 *aus der*

zweiten Klasse könnte dann aber eine Lücke entstehen. Diese Lücke darf nicht größer sein als die kleinstmögliche Differenz zwischen zwei Datenwerten. Wenn wir also nur auf cm-Genauigkeit messen, passt die verkürzte Angabe $[160 - 164]$*. Wenn wir aber mitunter auch einen mm-genauen Wert erhalten könnten, geht das nicht mehr, weil ja ein Wert* $164,7$ *nirgends in der Tabelle eingetragen werden kann. Dann müsste man* $[160,0 - 164,9]$ *und* $[165,0 - 169,9]$ *etc. schreiben.*

Noch einige Hinweise

...zur Klassifizierung von quantitativen Daten: Wenn es nur wenige mögliche Merkmalsausprägungen gibt (ca. 10-15), ist im Allgemeinen keine Klassifizierung vorzunehmen, sondern es werden gleich die Häufigkeiten der einzelnen Merkmalsausprägungen angegeben.

...zu Häufigkeitssummen: Aufsummiert werden nur die (absoluten oder relativen) Häufigkeiten f oder h, nicht aber die Merkmalswerte x_i. Und: Eine Summenbildung macht nur Sinn, wenn die Daten zumindest ordinalskaliert sind. Bei Nominaldaten ist eine Summenspalte nicht sehr aussagekräftig.

...zu Prozentangaben: Manchmal können Merkmalsträger auch mehr als eine Ausprägung eines Merkmals haben. Auf die Frage nach dem Geburtsort kann üblicherweise nur eine Antwort gegeben werden, aber auf die Frage *«Welche Komponistinnen und Komponisten der klassischen Musik kannst du namentlich benennen?»* kann auch mehr als eine Antwort kommen[2]. In der Statistik dokumentiert man das dann mit *«Mehrfachnennungen möglich»*, und immer, wenn Mehrfachnennungen möglich sind, kann die Summe über 100% liegen; die Angabe von Häufigkeitssummen sind dann nicht mehr sinnvoll. Das schließt auch einige Diagramme wie ein Kreisdiagramm (siehe 2.3, Seite 40) aus.

...ganz allgemein zu Tabellen: Die Tabellen, die wir tatsächlich publizieren und z.B. in wissenschaftlichen Arbeiten oder beruflichen Präsentationen verwenden, müssen nicht unbedingt alle hier angegebenen Spalten und Parameter (f, F, h, H) enthalten. Zu viele Zahlen verwirren eher. Gib nur diejenigen an, die zum Verständnis und zur Nachvollziehbarkeit deiner Aussagen notwendig sind.

[2] Laut Wikipedia ist der Vorrat, aus dem du schöpfen kannst, wirklich sehr, sehr groß: Siehe de.wikipedia.org/wiki/Liste_von_Komponisten_klassischer_Musik

. . . zum Aufbau von Tabellen: Sie sollten auch dann lesbar sein, wenn jemand «nur» die Tabellen anschaut und sich den restlichen Text nicht durchliest. Daher sollten sie einen **Titel** haben (oft steht der auch als «Caption» unter der Tabelle) und jede Spalte sollte eine Kopfzeile enthalten, in dem angegeben wird, welche Elemente in dieser Spalte zu finden sind. Entweder in der Spaltenüberschrift oder in der Beschreibung im Titel sollte auch angegeben sein, welche Einheiten verwendet werden (siehe auch Seite 25).

Sehr umfangreiche Tabellen sind oftmals für diejenigen, für die wir die Daten aufbereiten, zu kompliziert zu lesen und sie verlieren leicht den Überblick. Insbesondere für einen mit visuellen Präsentationsmedien unterstützten Vortrag aber auch für Texte und Berichte ist es in den meisten Fällen hilfreich, die Häufigkeitsverteilung graphisch aufzubereiten und (in schriftlichen Dokumenten, fast nie aber auf Präsentationsfolien) die Tabellen ergänzend hinzuzufügen.

2.3 Graphische Visualisierungen

Die grafische Darstellung der Daten ist meistens sehr hilfreich, um einen guten Eindruck von ihrer Verteilung zu erhalten und um zum Beispiel Häufigkeiten oder Muster in den Daten «auf einen Blick» zu erfassen. Für viele Menschen ist eine Grafik viel einprägsamer als eine Tabelle oder Liste voller Zahlen. Grafiken erlauben auch einen optischen – und damit meist schnelleren – Vergleich zwischen einzelnen Werten. Andererseits stellen Grafiken alleine (ohne die zugrundeliegenden Zahlen) immer auch einen gewissen Informationsverlust dar, weil die ursprünglich beobachteten Werte eventuell nicht mehr erkennbar sind.

Je nachdem, welche Daten wir präsentieren und Information wir dabei herausstreichen wollen, können wir verschiedene Diagrammtypen[3] verwenden.

Säulen- und Balkendiagramm

Säulen- und Balkendiagramme werden verwendet, wenn man die Verteilung von *diskreten* Daten darstellen will. Man sieht gut, in welcher Größenordnung Unterschiede zwischen den einzelnen Klassen bestehen und auch die Rangordnung innerhalb der Daten gut visualisieren.

In einem **Säulendiagramm** (auch: *Stabdiagramm*) werden die *Häufigkeiten* des Auftretens eines bestimmten Merkmalswertes (oder der Elemente einer Klasse)

[3] Unsere Beispiele sind nur eine kleine Auswahl aus unzähligen möglichen Darstellungsformaten und Diagrammtypen.

dargestellt, indem über den jeweils auf der Abszisse eingetragenen Bezeichnungen der Merkmalsträger schmale Rechtecke parallel zur Ordinate eingezeichnet werden, deren Länge proportional zum tatsächlichen Merkmalswert ist. Die Breite der Säulen hat hingegen keine Bedeutung und kann frei (aber gleich breit) gewählt werden – nach Möglichkeit aber so, dass alle vorkommenden Werte auch sinnvoll untergebracht werden können. Für ein Beispiel siehe Abb.2.1.

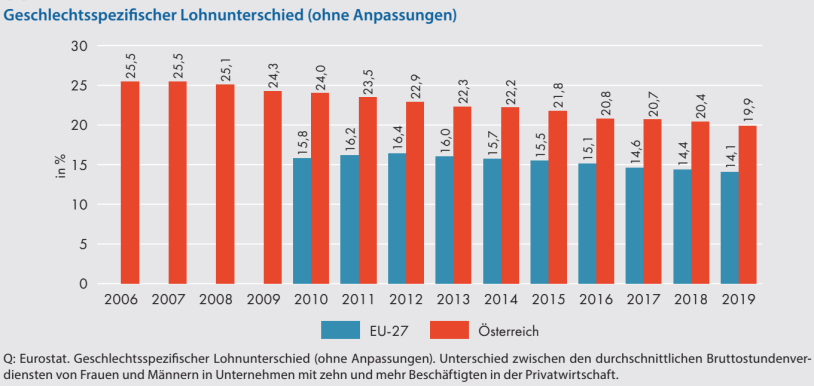


Abb. 2.1: Gender Pay Gap: Differenz zwischen den durchschnittlichen Bruttostundenverdiensten der männlichen und der weiblichen Beschäftigten in Prozent der durchschnittlichen Bruttostundenverdienste der männlichen Beschäftigten. (Quelle: Statistik Austria. 2021. *Wie geht's Österreich?* p.56)

Manchmal wird das Koordinatensystem, in dem das das Säulendiagramm eingebettet ist, auch um 90 Grad gedreht (Merkmalsträger werden auf der senkrechten Achse eingetragen, Merkmalswerte auf der waagerechten) und dann **Balkendiagramm** genannt. Balkendiagramme sind gegenüber Säulendiagramme insbesondere dann im Vorteil, wenn man mehr Klassen darstellen will, als sich nebeneinander ausgehen und auch wenn die Klassenbezeichnungen länger sind und Beschriftungen nur gegen die Leserichtung um 90 Grad gedreht möglich wären. (Abb.2.2).

Wenn in einem Balken- oder Säulendiagramm negative Werte dargestellt werden müssen, werden die negativen Daten (Balken oder Säulen) immer links bzw. unterhalb der Nulllinie platziert (siehe Abb.2.3). Das gilt auch, wenn gegebenenfalls die gesamte Datenreihe nur aus negativen Werten besteht.

Im Säulen- oder Balkendiagramm lassen sich auch zwei oder mehrere Datensätze darstellen, was oft einen anschaulichen Vergleich zwischen den Zufallsvariablen erlaubt. Dabei ist darauf zu achten, dass ein Vergleich zweier oder mehrerer Datensätze auf Basis der absoluten Häufigkeiten nur dann sinnvoll ist, wenn die Datensätze vom gleichen Umfang sind. Bei unterschiedlichem Umfang werden besser die relativen Häufigkeiten repräsentiert.

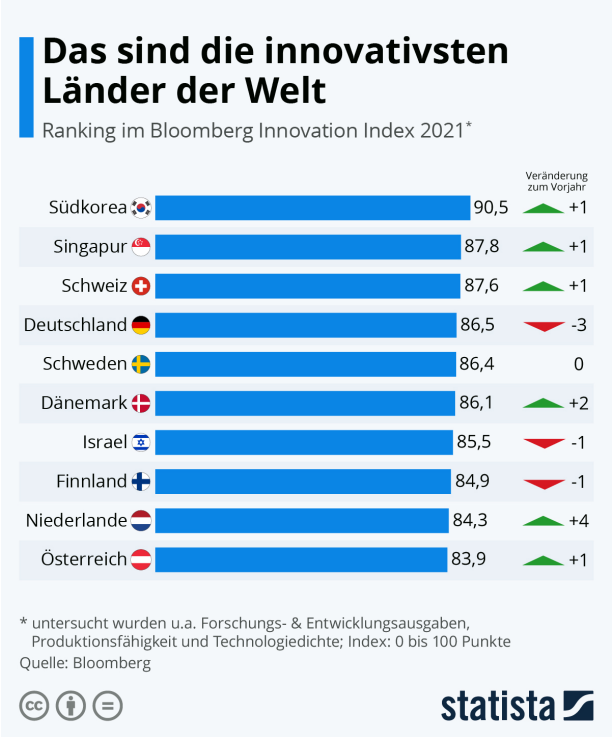


Abb. 2.2: Die innovativsten Länder der Welt (Quelle: de.statista.com/infografik/20548/)

Man sollte mit den Mehrfachsäulendiagrammen aber auch nicht mit der Anzahl der unterschiedlichen Säulen (Kategorien) übertreiben. Fünf oder mehr Säulen können vermutlich nicht mehr verglichen und interpretiert werden.

Und noch ein wichtiger Hinweis: Grundsätzlich sollten sowohl die x- als auch die y-Achse immer bei Null beginnen. Andernfalls werden die Differenzen zwischen den einzelnen Werten (oder Klassen) größer dargestellt, als sie in Wirklichkeit sind. Mitunter können einzelne Säulen überhaupt verschwinden[4] (siehe Abb.2.4).

In der englischsprachigen Literatur werden Säulendiagramme oft als «Histogram» bezeichnet. Im Deutschen sind wir da etwas präziser, und verwenden «Säulendiagramme» für diskrete Daten, «Histogramme» hingegen für stetige.

[4]Vermutlich würde es dir auffallen, wenn ganze Säulen verschwinden – sofern das Diagramm händisch erstellt wird. Im Fall der automatisierten Datenauswertung aber könnte das durchaus passieren, daher sollte man eine Verschiebung der Basislinie wirklich gut überlegen.

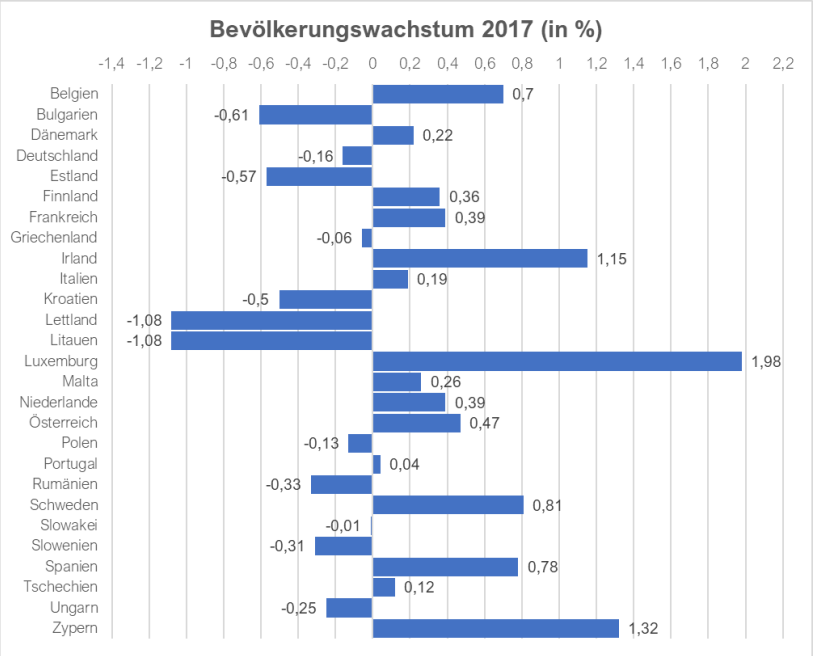


Abb. 2.3: Bevölkerungswachstumsraten der EU-Mitgliedsstaaten: Sie gibt die durchschnittliche jährliche prozentuale Veränderung der Bevölkerung an, die sich aus einem Überschuss oder Defizit zwischen Geburten und Todesfällen und der Differenz der Migrantinnen und Migranten ergibt, die in ein Land ein- oder aus einem Land ausreisen. (Datenquelle: CIA World Factbook, abgerufen: 23.7.2018)

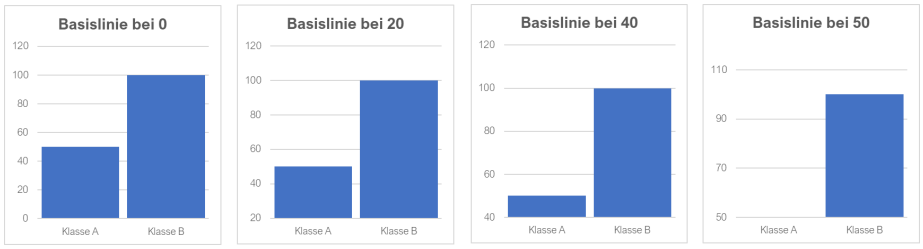


Abb. 2.4: Beginnt die y-Achse nicht bei 0, werden die Differenzen zwischen einzelnen Klassen überdeutlich dargestellt, was manchmal auch dazu führen kann, dass sie überhaupt verschwinden

Histogramm

In einem **Histogramm**[5] werden die Häufigkeiten *stetiger* Daten dargestellt. Stetige Zufallsvariable können ja jeden Zahlenwert aus $\mathbb{R}$ annehmen und so sehen wir die x-Achse des Diagramms als Zahlengerade[6], auf der jede denkbare Zahl abgebildet werden kann und auf der wir auch Intervalle bilden können, die eine

[5] Das Wort hängt vermutlich mit dem griechischen *ἱστός* (*histos*) = Mast(baum) zusammen. Die Bezeichnung wurde um 1895 von *Karl Pearson* eingeführt.

[6] auch unter *Zahlenstrahl* bekannt.

bestimmte Klasse von Merkmalswerten beinhaltet. Abb.2.5 zeigt zum Beispiel das Histogramm zur Tabelle 2.2.

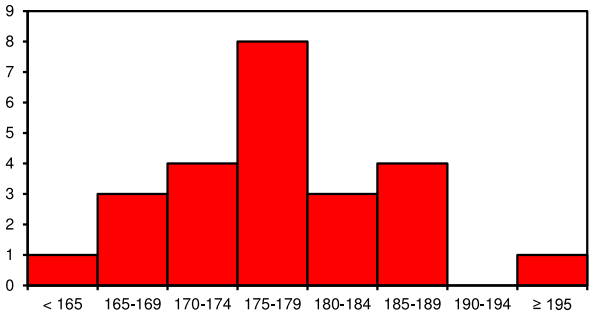


Abb. 2.5: Beispiel für ein Histogramm. Dargestellt sind die Daten aus Tab.2.2.

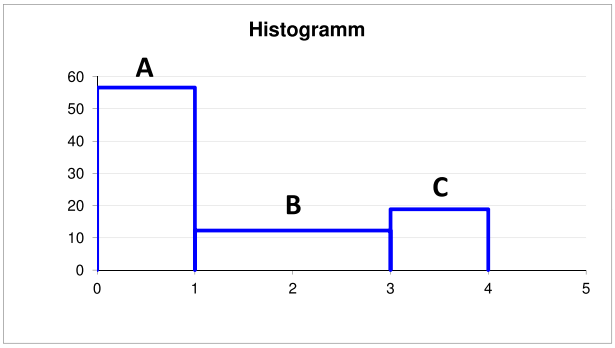
Auf der Abszisse werden im Histogramm die Klassengrenzen aufgetragen und über den Klassenintervallen Rechtecke errichtet, deren **Flächen** (!) proportional zu den Häufigkeiten sind. Beschriftet werden auf der Abszisse entweder die Klassengrenzen, die Klassenindizes oder die Klassenmitten (= obere minus untere Klassengrenze dividiert durch Zwei). Auf der Ordinate (y-Achse) wird die **Häufigkeitsdichte** angegeben, das ist der Quotient

$$\text{Häufigkeitsdichte} = \frac{\text{Häufigkeit}}{\text{Klassenbreite}} \tag{2.7}$$

Auf den ersten Blick schauen Histogramme genau aus wie die auf S.35 beschriebenen Säulendiagramme. Sie unterscheiden sich aber in wichtigen Punkten:

- Zwischen zwischen den Rechtecken (den Säulen) des Histogramms sind keine *Abstände*. Sie repräsentieren ja stetige, also kontinuierliche Daten, und ein Abstand würde der Stetigkeit entgegenstehen.
- Die Verwendung von Säulendiagrammen ist nur für den Fall *gleich breiter* Klassen angeraten. Bei unterschiedlichen Klassenbreiten benötigen wir ein Histogramm.
- Wie schon oben erwähnt: nicht die Höhe sondern die *Fläche* ist das Maß für die Häufigkeit. Nur im Fall gleicher Klassenbreiten spielt dieser Unterschied keine Rolle, dann könnten auf der y-Achse auch direkt die Häufigkeiten aufgetragen werden – streng genommen handelt es sich dann aber nicht mehr um ein Histogramm, sondern um ein Säulendiagramm, denn:
- Im Histogramm wird auf der y-Achse die Häufigkeitsdichte aufgetragen, im Säulendiagramm die (absolute oder relative) Häufigkeit.

Beispiel 2 *Das folgende Histogramm zeigt die Häufigkeitsverteilung einer Zufallsgröße, die in die drei Klassen A, B, und C eingeteilt wurde:*



In welcher Klasse / in welchen Klassen befinden sich die wenigsten Elemente?

Lösung:

In einem Histogramm ist nicht die Höhe der rechteckigen Säulen ausschlaggebend für die Anzahl der in der jeweiligen Klasse enthaltenen Elemente, sondern die Fläche.

In der Klasse A befinden sich demnach knapp unter 60 Elemente, in der Klasse B 24 (Breite = 2 × Höhe = 12) und in der Klasse C 20 (Breite = 1 × Höhe = 20) Elemente.

Somit befinden sich in der Klasse C die wenigsten Elemente (obwohl die Säule selbst höher ist als jene der Klasse B).

Kreisdiagramme

Beim **Kreisdiagramm** (auch: *Tortendiagramm*) wird jeder Ausprägung des Merkmals ein Kreissektor zugewiesen. Auch hier geht es also um die Fläche: Die Fläche jedes Sektors spiegelt die *relative Häufigkeit* seines Auftretens wider. Die Sektorgrenzen können berechnet werden, indem die relativen Häufigkeiten jeweils mit 360° multipliziert werden. Damit erhält jeder Merkmalswert ein «Tortenstück», dessen Größe der relativen Häufigkeit entspricht. Die einzelnen Kreissektoren erhalten zur besseren Lesbarkeit meist unterschiedliche Färbungen oder Grafikmuster. Abb.2.6 zeigt ein Beispiel für ein Kreisdiagramm:

Kreisdiagramme eignen sich besonders auch für nominalskalierte, kategorische Werte. Man erhält mit ihnen einen guten Gesamtüberblick über die Daten. Insgesamt sollten aber nicht mehr als 7 bis 9 Segmente (Klassen, Kategorien) vorliegen, damit es noch lesbar ist. Außerdem ist ein direkter Vergleich zweier Merkmale schwierig, wenn die betroffenen «Tortenstücke» nicht zufällig benachbart sind. Und selbst dann kann es sein, dass der Unterschied so gering ist, dass man

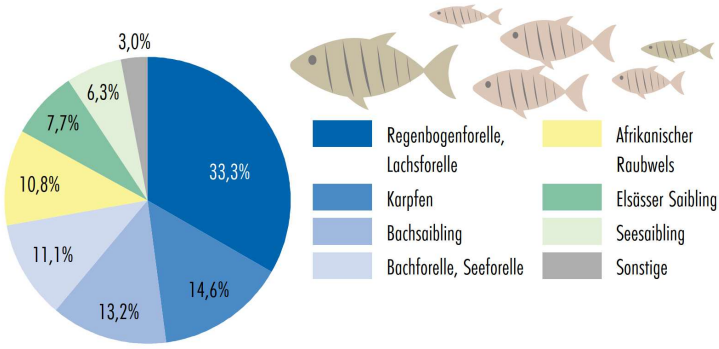


Abb. 2.6: Kreisdiagramme zur Produktion von Speisefischen in Österreich 2019 (Quelle: Statistik Austria: Österreichischer Zahlenspiegel Jänner 2021, p.7)

das aus der Größe der Tortenstücke alleine (also ohne Datenbeschriftung) nicht erkennen kann.

Tortendiagramme können übrigens auch mit einem «Loch» in der Mitte dargestellt werden (und ähneln dann eher einem *Donut* als einer Torte). Sie werden dann als *Ringdiagramme* bezeichnet. Siehe Abb.2.7.

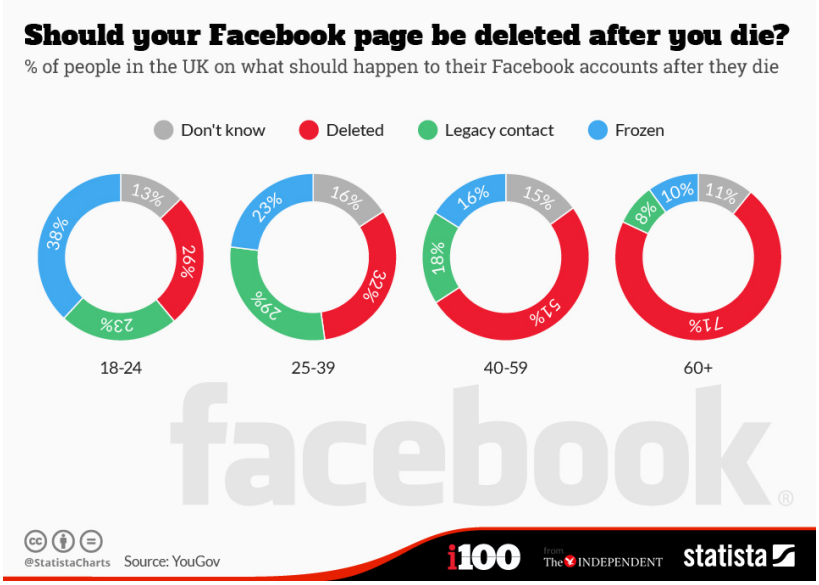


Abb. 2.7: Beispiel für ein «Donut-Diagramm»
(Quelle: http://www.statista.com/chart/3403/should-your-facebook-page-be-deleted-after-you-die/)

In klassischen Tortendiagrammen wie jedem der Abb.2.6 können keine negativen Werte visualisiert werden – dazu müssten Kreissegmente mit einer «negative Flächen» dargestellt werden. Donut-Diagramme können da Abhilfe schaffen, indem negative Werte «nach innen» in das Loch in der Mitte ausgerichtet werden.

Streifendiagramm

Ein **Streifendiagramm** (auch: **gestapeltes Säulendiagramm** oder **gestapeltes Balkendiagramm**) ist von der Aussage einem Tortendiagramm sehr ähnlich; es handelt sich aber nicht um einen Kreis, sondern um Balken oder Säulen, wo mehrere Merkmalswerte je Variable neben- oder übereinandergestapelt werden (siehe Abb.2.8). Im Gegensatz zu einem Tortendiagramm können hier nicht nur Relativ- sondern auch Absolutwerte dargestellt werden.

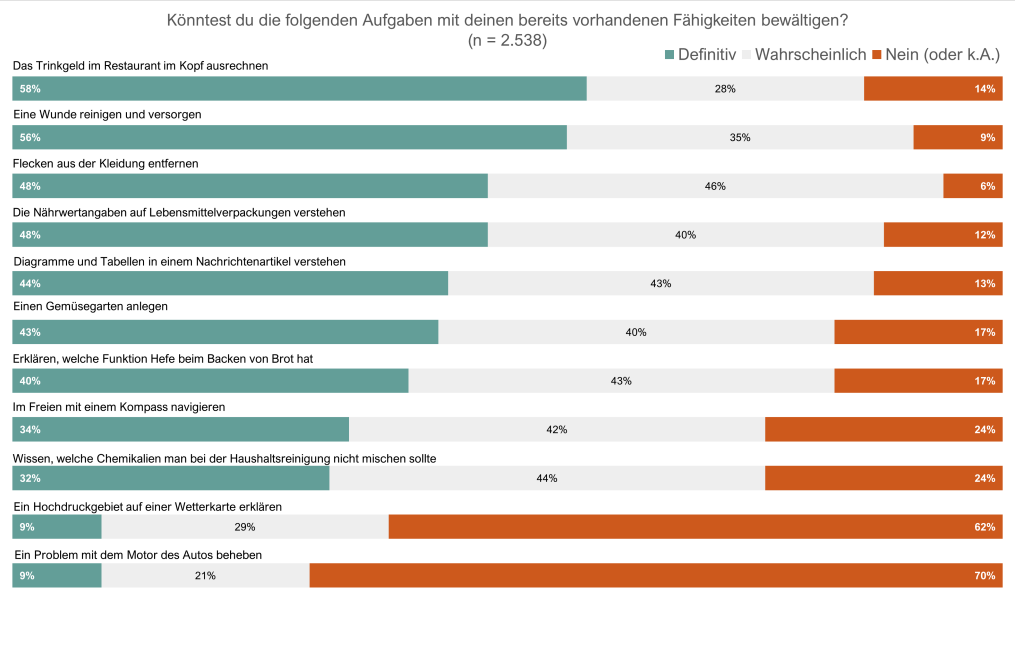


Abb. 2.8: Was US-Amerikaner:innen (Personen ab 18 Jahren, die in den Vereinigten Staaten leben) über ihre praktischen Fähigkeiten denken. (n = 2.538, Konfidenzniveau: 95%, Fehlerspanne: $\pm 1,5$ Prozentpunkte) (Datenquelle: https://www.pewresearch.org/wp-content/uploads/sites/20/2025/10/SR_25.10.10_science-skills_topline.pdf)

Statt einfacher rechteckiger Balken können in einem Diagramm übrigens auch Piktogramme verwendet werden. Das eignet sich insbesondere für die Darstellung der Häufigkeit von nominalskalierten Daten, siehe zum Beispiel Abb.2.9.

Linien- und Flächendiagramm

Liniendiagramme eignen sich vor allem dann, wenn mehrere Datenreihen verglichen werden sollen und wenn wir Trends (zeitliche Änderungen) darstellen wollen. Ein Beispiel ist in Abb.2.10 zu sehen.

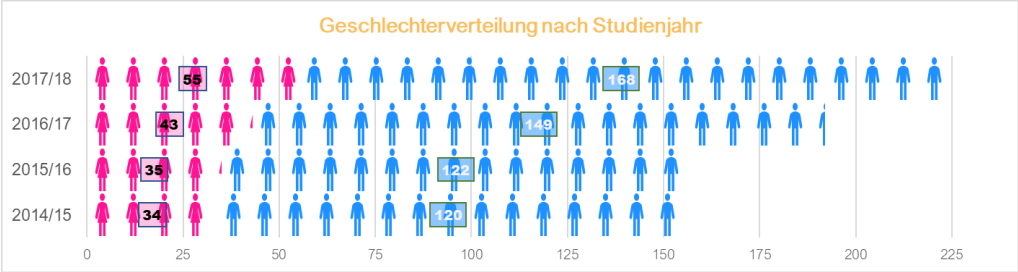


Abb. 2.9: WIBA-Studierende 2015-18 nach Geschlecht

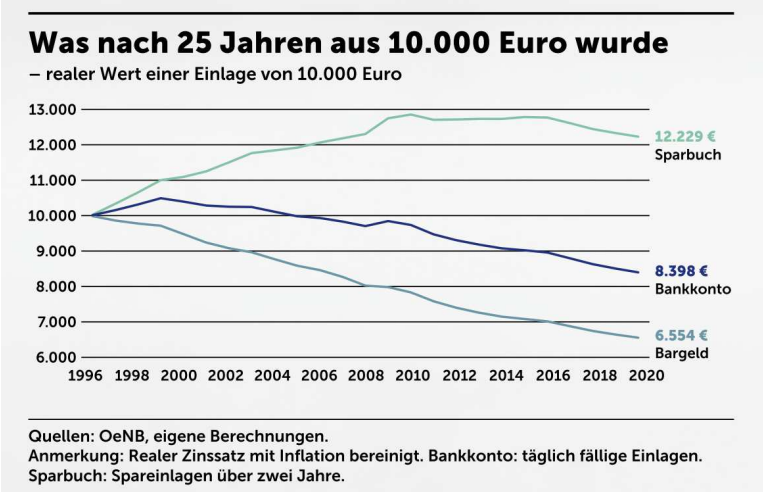


Abb. 2.10: Was über die vergangenen zweieinhalb Jahrzehnte aus 10.000 Euro wurde. (Quelle: Agenda Austria: Balken, Torten, Kurven Zweitausendeinundzwanzig. 2022. p.105)

Zur Erstellung des Liniendiagramms werden zunächst ähnlich wie beim Säulendiagramm über jedem Punkt der x-Achse die Datenwerte eingetragen, allerdings nicht in Form einer Säule, sondern nur in Form eines Punktes. Diese Punkte werden dann mit einer Linie verbunden. Die Punkte können anschließend angezeigt oder auch wieder weggelassen werden – Informationsträger ist ja jetzt die «Daten-Linie».

In Liniendiagrammen können relativ große Datenmengen auf relativ kleinem Raum abgebildet werden. Beachte aber, dass nach ca. 5-7 Linien die Übersichtlichkeit verloren geht. Zur Unterscheidung der einzelnen Linien verwendet man am besten unterschiedliche Farben (Achtung auf Personen mit Farbenfehlsichtigkeiten!), falls mit einem Schwarz-Weiß-Ausdruck zu rechnen ist, dann verschiedene Graustufen. «Gestrichelte» oder «gepunktete» schwarze Linien werden hingegen heutigen ästhetischen Ansprüchen nicht mehr ganz gerecht.

Ein artverwandtes Diagramm zum Liniendiagramm ist das **Flächendiagramm**. Dabei wird die Fläche zwischen der Linie des Liniendiagramms und der x-Achse

noch «ausgemalt». Beispiel: Abb.2.11.

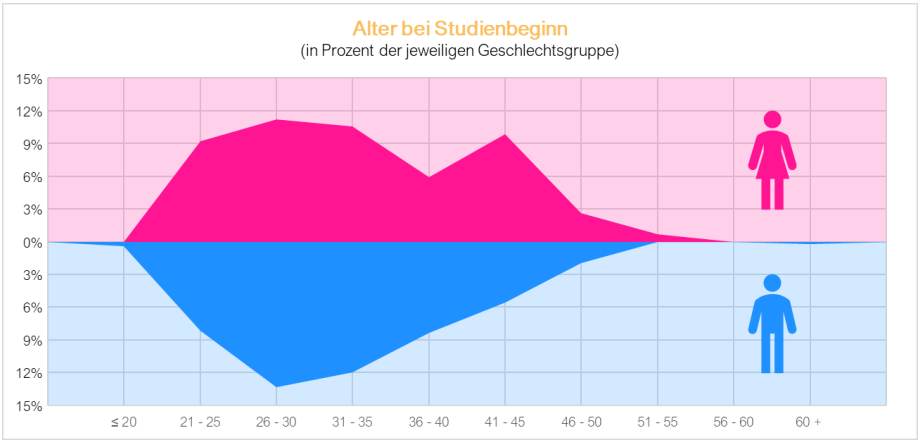


Abb. 2.11: Altersverteilung der WIBA-Studierenden bei Studieneintritt in Prozent der jeweiligen Geschlechtsgruppe

Linien- und Flächendiagramme sind nur schlecht geeignet, wenn Daten klassifiziert wurden.

Punktdiagramm

Streng genommen können wir Liniendiagramme nur für stetige Daten verwenden, weil wir ja meistens gar nicht kontinuierlich gemessen haben. Vermutlich wurde zum Beispiel der Realwert des Euros nicht jede Minute eruiert, wie Abb.2.10 suggeriert, sondern nur an bestimmten Stichtagen (z.B. am Beginn jeden Jahres), und diese Messpunkte einfach durch Linien verbunden. Will man «auf Nummer sicher» gehen und wirklich nur die Daten darstellen, für die gemessene Werte vorliegen, verwendet man ein **Punktdiagramm**, wie in Abb.2.12 dargestellt.

Die hier genannten Diagrammtypen stellen nur einen kleinen Ausschnitt der Möglichkeiten dar. Es gibt unzählige «Unterarten» oder Mischformen, wie das Beispiel der Abb.2.13.

Eine interessante Übersicht über alle möglichen Diagramme findest du zum Beispiel auf der Seite www.data-to-viz.com.

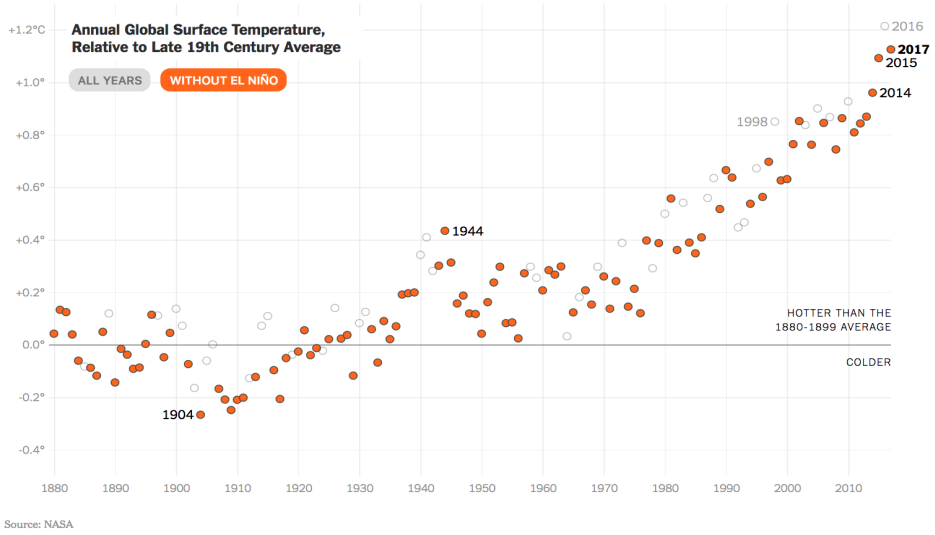


Abb. 2.12: Die durchschnittliche Temperatur an der Erdoberfläche von 1880 bis 2017 in Relation zur Durchschnittstemperatur der letzten 20 Jahre des 19. Jahrhunderts. (Bildquelle: The New York Times. 18.1.2018. http://www.nytimes.com/interactive/2018/01/18/climate/hottest-year-2017.html)

2.4 Hinweise für die Erstellung von Tabellen und Diagrammen

Computerprogramme wie MS Excel machen es ziemlich einfach, Grafiken zu erstellen und Daten zu visualisieren. Allerdings haben sie auch die Tendenz, mit (vermeintlichen) optischen Raffinessen zu übertreiben und die eigentliche Informationsdarstellung dem «Styling» unterzuordnen oder diese sogar zu verfälschen oder Betrachter:innen zumindest zu verwirren. Hier daher einige Tipps:

- Verschiedene Statistikprogramme bieten die oben genannten Diagramme und Histogramme auch in einer dreidimensionalen Ausprägung an. Dies kann eventuell dann Verwendung finden, wenn wir die statistische Verteilung zweier Merkmale zugleich darstellen wollen. In der Regel muss man aber darauf achten, dass durch den 3D-Effekt die Informationen auch verzerrt dargestellt werden können. Weniger ist oft mehr.
- Bei Histogrammen: Achtung auf unterschiedliche Klassenbreiten!
- Für alle Diagramme in einem Koordinatensystem: Die Ordinate (y-Achse) sollte ungefähr $2/3$ bis $3/4$ der Länge der Abszisse (x-Achse) haben. Sie sollte im Ursprung des Koordinatensystems mit dem Wert 0 beginnen, da es sonst zu irreführenden Maßstabsverzerrungen kommen kann.
- Vermeide unnötige «Dekorationen» oder Illustrationen im Hintergrund einer grafischen Darstellung sowie «Zierrahmen».

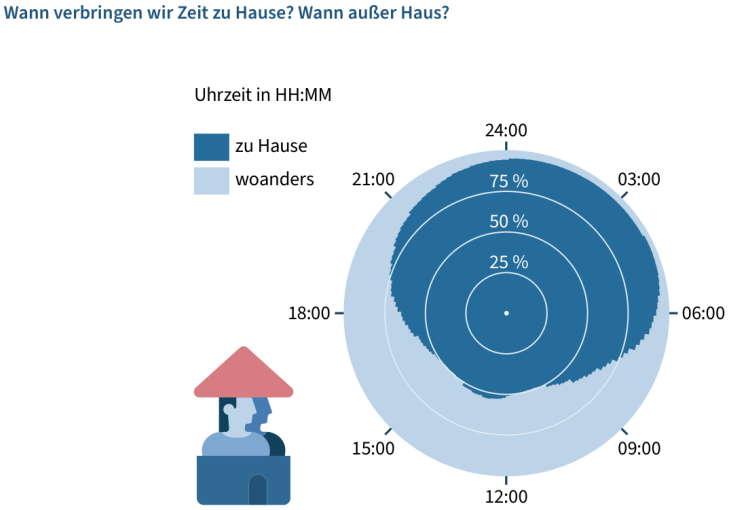


Abb. 2.13: Wann verbringen wir wo unsere Zeit?
(Quelle und Grafik: STATISTIK AUSTRIA, Zeitverwendungserhebung)

- Vergiss nicht: Jedes Diagramm braucht einen Titel, eine Beschriftung der Achsen und eine Beschriftung (möglichst direkt an Ort und Stelle – eine Legende ist nur die zweitbeste Lösung). Und einen Quellenverweis auf die Herkunft der Daten.
- Denke auch an eine ansprechende Schriftart und Farbgebung. Beachte dabei aber auch die Möglichkeit der Farbenfehlsichtigkeit mancher Menschen[7].
- Verwende nach Möglichkeit keine gedrehten oder schrägstehenden Beschriftungen (also keinen Text in anderer als horizontaler Ausrichtung), weder aus «künstlerischen» Gründen, noch um Platz zu sparen. Falls du dann zum Beispiel Bezeichnungen in einem Säulendiagramm nicht mehr unterbringen kannst, verwende – wenn nichts anderes dagegenspricht – Balkendiagramme.
- Fettschrift dient dazu, etwas hervorzuheben, weil es entweder eine Überschrift oder eine Kernaussage ist, oder aber, um es auf einem eingefärbten Hintergrund lesbarer zu machen. Wenn du *alles* fett schreibst, geht dieser Effekt verloren.
- Du musst nicht alle Farben verwenden, die dein Monitor oder Drucker schafft. Bedenke auch: Eventuell läuft deine Präsentation über einen kaputten Beamer oder deine Datei wird auf einem Schwarz-Weiß-Drucker gedruckt.
- Verwende bei Säulen- und Balkendiagrammen möglichst keine «Schatten» hinter den Säulen, außer du zeichnest von Hand auf ein Flipchart (falls das heute noch vorkommt). Schatten beinhalten keine besondere Information.

[7] siehe www.datylon.com/blog/data-visualization-for-colorblind-readers

- Bei Tabellen ist es nicht ratsam, alle möglichen Rasterlinien (Zellrahmen) einzuzeichnen; besser z.B. nur nach jeder dritten oder fünften Zeile eine Linie zeichnen (und nach der ersten Zeile mit den Spaltenüberschriften).
- Ganze Zahlen sollten in Tabellen nicht linksbündig, sondern rechtsbündig gesetzt werden; Dezimalzahlen weder links- noch rechtsbündig, sondern am Komma ausgerichtet.

Und das Wichtigste: Verwende immer *aktuelle* Daten aus zuverlässigen Quellen.
Und überleg immer *zuerst*, welche *Frage* du eigentlich mit deiner Visualisierung beantworten möchtest. Dann wähl den Diagrammtyp aus.
Nicht umgekehrt.