Zur Datenerhebung:

Vor einer Datenerhebung muss man sich über die Identifikation des zu messenden Merkmals einigen. Wenn man den Prozentsatz der blonden Menschen in einer Stichprobe ermitteln will, muss vorher klar sein, wo die Grenze zwischen blond- und braunhaarigen Menschen verlaufen soll. Bei einer Verkehrszählung muss vorher klar sein, ob etwa Fahrräder oder Fußgänger ebenfalls zu zählen sind.

Häufig wird man die erhobenen Daten in Klassen einteilen wollen. Hier ist viel Spielraum für *Manipulationen*:

Beispiel: Gehaltsstatistik eines Betriebes

Verdienst in €	Anzahl der Mitarbeiter		Verdienst in €	Anzahl der Mitarbeiter		Verdienst in €	Anzahl der Mitarbeiter
1400	3		1000 – 1500	3		1200 – 1700	5
1600	2		1500 – 2000	4		1700 – 2200	4
1800	2		2000 - 2500	5		2200 - 2700	3
2100	2						
2400	3						

Die Spalten in der Mitte vermitteln einen ganz anderen Eindruck als die Spalten rechts.

Zur Repräsentativität einer Stichprobe:

Hier geht es nur darum, grobe Fehler bei der Datenerhebung zu vermeiden, wie etwa eine Wahlumfrage vormittags in der Fußgängerzone durchzuführen (weil dadurch fast die gesamte arbeitende Bevölkerung ausgeblendet ist).

Zur Standardabweichung:

Hat man die Messdaten $x_1, x_2, ..., x_n$, so ist $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ das arithmetische Mittel und

$\sigma := \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$ die empirische Standardabweichung. Diese wird mitunter (u.a. auf Taschenrechnern) mit dem Nenner $n-1$ statt des Nenners n angegeben. Wann verwendet man welchen Nenner?
Beschreibt man Daten, so ist der Nenner n sinnvoll. Will man die Standardabweichung einer Gesamtpopulation aus der Standardabweichung einer Stichprobe *schätzen*, so muss man den Nenner $n-1$ verwenden. Dies Schätzproblem tritt im Schulunterricht nicht auf.

<u>Wege zur empirischen Standardabweichung</u>:

Die empirische Standardabweichung ist ein Streumaß. In früheren Jahrgängen haben die Lernenden die Spannweite als einfaches (und wenig aussagekräftiges) Streumaß kennengelernt. So haben die Messdaten 1, 3, 4, 7, 8, 9 und die Messdaten 1, 4, 4, 5, 5, 9 dieselbe Spannweite, aber die zweite Serie weist weniger Streuung auf.

Eine erste Idee, wie man die Streuung von Messdaten besser beschreiben kann, besteht darin, die Summe $\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)}{n}$ zu betrachten. Wegen $\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)}{n}=\frac{\sum_{i=1}^{n}x_i}{n}-\overline{x}=\overline{x}-\overline{x}=0$ ist das noch keine gute Idee. Lernende schlagen dann oft $\frac{\sum_{i=1}^{n}\left|x_i-\overline{x}\right|}{n}$ vor. Diese „mittlere absolute Abstandssumme" ist in der beschreibenden Statistik mittlerweile auch gebräuchlich (wird allerdings i.a. auf den Median bezogen). Nun möchte man in der (nicht mehr schulnahen) weiterführenden Statistik das Streuungsmaß auch ableiten können, und da sind Beträge unpraktisch. Aus diesem Grunde hat sich als populäres Streuungsmaß die „mittlere quadratische Abstandssumme" $\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n}$ durchgesetzt, die *empirische Varianz* genannt wird.

Sind die Messdaten z. B. Körpergrößen in cm, so hat die empirische Varianz die Einheit Quadratzentimeter. Das ist unpraktisch. Aus diesem Grund verwendet man die Quadratwurzel aus der empirischen Varianz, und das ist die oben angegebene empirische Standardabweichung.

Bei Verwendung der Binomialverteilung und der Normalverteilung später in der Qualifikationsphase wird der empirischen Standardabweichung die theoretische Standardabweichung zugeordnet, und man hat dann Aussagen wie „Im Abstand von maximal einer Standardabweichung vom Erwartungswert liegen etwa 68,3 % aller Daten".

<u>Darstellungen von Häufigkeitsverteilungen in Säulendiagrammen mit GeoGebra:</u>

Die x-Werte der Häufigkeitsverteilung kommen in eine Liste, etwa

X=Folge[k, k, 1, 12]

(Listen sind in GeoGebra stets Folgen), die y- Werte in eine andere Liste, etwa

Y={1, 2, 8, 7, 1, 4, 6, 6, 1, 4, 7, 3}.

Mit

Balkendiagramm[X, Y]

erhält man das Säulendiagramm (es gibt in GeoGebra keinen Befehl „Säulendiagramm").