Statistik:

Vorbemerkung 1: Viele Programme (z.B. PRAAT) benutzen die ‘englische’ Darstellungsweise für Zahlen mit einem Dezimal**punkt** während deutsche Programme ein Dezimal**komma** erwarten. Manche Programme/Betriebssystemversionen konvertieren das automatisch, bei manchen kann man es einstellen, häufig hilft es aber nur, die Daten in ein Textfile zu kopieren (bzw. mit einem Texteditor wie Notepad++ (Windows) oder TextWrangler (Mac) aufzumachen), dort mit einem generellen Ersetzen alle Punkte durch Kommata zu ersetzen, und dass dann wieder in das ‘deutsche’ Programm zu kopieren (Achtung: nicht im ‘Word’, ‘OpenOffice’, ‘Pages’, o.ä. Format speichern, sondern als ‘Text’). Zahlenwerte werden von Programmen (z.B. Excel) häufig rechtsbündig dargestellt, Text (*Strings* = Zeichenketten) linksbündig. Also: “1,000” ist im Englischen “eintausend”, “1.000” ist im deutschen Excel ein Text (die Buchstabenfolge “Eins-Punkt-Null-Null-Null”).

Vorbemerkung 2: Der Wert ‘0’ ist etwas anderes als ein ‘fehlender Wert / *missing value*’ (ein ‘*missing value*’ wird in vielen Programmen mit einem Punkt, im Statiskprogramm ‘R’ als “NA” und in PRAAT als “-- undefined --” dargestellt. Wenn man vier Messungen mit den Werten $\{3, 0, 4, 5\}$ hat und den Mittelwert berechnet, gibt das (3+0+4+5)/4 = 12/4 = 3. Hat man vier Messungen gemacht und die zweite ist fehlgeschlagen $\{3, ., 4, 5\}$ dann ist der Mittelwert (3+4+5)/3 = 4.

Maße zur Bestimmung eines ‘Mittelwertes’ (*mean*):

(arithmetischer) Mittelwert (Durchschnittswert, *average*): Summe aller Messwerte geteilt durch die Anzahl der Werte.

Mathematisches Zeichen: $\bar{x}$, mathematische Schreibweise: $\frac{\sum_{i=1}^{n} x_i}{n}$, bzw. $\frac{\sum x_i}{n}$

(‘Summiere alle Messwerte x_i und teile die Summe durch die Anzahl n der Messwerte’). Der (arithmetische) Mittelwert wird sehr häufig als Maßzahl verwendet, um die ‘durchschnittliche’ Größe von Messwerten zusammenfassend zu beschreiben. Manchmal kann er aber nicht berechnet werden (‘das Geschlecht der Einwohner von Deutschland ist im Mittel ??’), dürfte eigentlich nicht berechnet werden (z.B. bei Ordinalzahlen, s.u.), oder kann von Extremwerten (Ausreißern) verfälscht werden (z.b. hat die Messreihe $\{3, 5, 4, 30, 3\}$ den Mittelwert “9”).

Median ($\hat{x}$): Ordne alle Werte der Größe nach und nehme den Wert, der in der Mitte steht (bei einer geraden Anzahl von Werten, ‘nehme den Mittelwert der beiden Werte in der Mitte’). Die Hälfte aller Werte sind größer, bzw. kleiner als der Median. Er ist nicht so stark von Ausreißern abhängig (der Median von $\{3, 5, 4, 30, 3\}$ ist “4”), aber steht in keinem ‘rechnerischen’ Zusammenhang mit allen Werten. Der Vergleich von Median und arithmetischem Mittelwert ist eine einfache Kontrolle, ob es wohl Ausreißer oder ‘Schieflagen’ (s. Histogramm) gibt. Bei Nominaldaten (s.u.) kann man keinen Median berechnen.

Modal: Der am häufigsten auftretende Wert (z.B. “Christian” ist der häufigste männliche Vorname).

Streuungsmaße (wie stark die Messwerte von ihrem Mittelwert abweichen):

Summe der Abweichungsquadrate (SAQ, '*sum of squares*', SS): Summe aller quadrierten Abweichungen vom Mittelwert. Mathematische Schreibweise: $\sum\left(x_i-\bar{x}\right)^2$ ('Bilde die Differenz zwischen Messwert und dem arithmetischen Mittelwert, quadriere sie, und summiere alle diese quadrierten Differenzen'). Wird vor allem in der Prüfstatistik verwendet.

Varianz ('*variance*'): Summe der Abweichungsquadrate geteilt durch die Anzahl der Messwerte minus 1: $\frac{\sum\left(x_i-\bar{x}\right)^2}{n-1}$. Gewissermaßen das 'mittlere Abweichungsquadrat'.

Standardabweichung (St.Abw., '*standard deviation*', sdev, SD): Die Wurzel aus der Varianz: $\sqrt{\frac{\sum\left(x_i-\bar{x}\right)^2}{n-1}}$. Gewissermaßen die 'mittlere Abweichung'.

Quartile (25%, 75% Quartile, 'Viertelschritte'): Bei der Anordnung der Werte der Größe nach derjenige Wert, bei dem 25% aller Werte größer (oder kleiner) sind. 50% aller Messwerte liegen zwischen dem 25% und 75% Quartil.

Quantile, bzw. **Perzentile** (10%, 20%, 80%, 90%, aber auch 5%, 2.5%, 1% etc. Quantil / Perzentil): Bei der Anordnung der Werte der Größe nach diejenigen Werte, bei denen die entsprechende prozentuale Anzahl der Werte größer, bzw. kleiner sind; z.B. sind 99% aller Werte größer als das 1% Perzentil (bzw. 1% der Werte sind kleiner). Z.B. liegen 90% aller Messwerte zwischen dem 5% und dem 95% Perzentil. (Häufig wird von 'Quantilen' gesprochen, die eigentlich nur 10%-Schritte sind, wenn es um kleinere Schritte geht, also z.B. '5% Quantil'.)

Minimum, Maximum: Kleinster, größter Wert.

Bereich ('*range*'): Wertebereich, in dem alle (bzw. Perzentil-Anteile) aller Daten liegen.

Skalen (Datentypen):

Nominaldaten: Dinge, die einen Namen haben (Hans; Ilse; a; i; weiblich...)
Ordinaldaten: Dinge die man anordnen kann (der Größe nach; 1.; 2.; 3... im Wettbewerb)
Intervalldaten: Werte, bei denen gleiche Abstände das gleiche Bedeuten (0; 123; 10,76; -25,3)
Verhältnisdaten: Wie Intervalldaten, aber es gibt einen absoluten Nullpunkt und man kann Verhältnisse angeben (z.B. macht es auf einer Celsiusskala keinen Sinn zu sagen, dass 10°C 'doppelt so warm ist' wie 5°C; auf einer Kelvinskala kann man das sagen)

Z.B. hat man bei einem Abfahrtslauf die Teilnehmerinnen „Isolde", „Brunhilde" und „Kunigunde" (Nominaldaten), die auf dem 1., 2. und 3. Platz landen (Ordinaldaten), und um 10 Uhr 30, 11 Uhr 15, und 17 Uhr 31 losgelaufen sind (Intervalldaten) und 100, 112 und 113 Sekunden gebraucht haben (Verhältnisdaten).

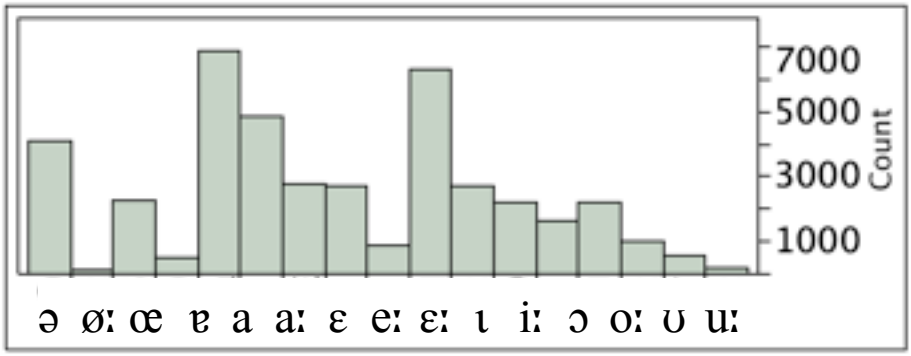
Nominaldaten:
Bei den Nominaldaten kann man nur feststellen, wie häufig Werte (Namen) auftreten; der häufigste Wert ist der **Modalwert** (z.B. der häufigste Familienname in Deutschland ist *Müller*).

Ordinaldaten:
Bei Ordinaldaten kann man zusätzlich feststellen, welcher Wert ‚in der Mitte steht'; dies ist der **Median**. Streuungsmaße sind **Wertebereich, Quantile, Quartile**.

Intervalldaten:
Hier kann man addieren und subtrahieren (und damit auch den **arithmetischen Mittelwert** und die **Standardabweichung, Varianz** und **SAQ** berechnen.

Verhältnisdaten:
Alles was man bei Intervalldaten machen darf und noch mehr, weil man auch multiplizieren und dividieren darf (z.B. kann man sagen 'doppelt so groß').

Verteilung ('*distribution*') und **Histogramm** (graphische Darstellung der Häufigkeitsverteilung):



z.B. Häufigkeiten der Vokaldauern in einem deutschen Text-Korpus (Verhältnisdaten):

Quantile:

100.0%	maximum	840.75	ms	Mittelwert:	76.68 ms
99.5%		332.41	ms	St.Abw.	50.62 ms
97.5%		214.96	ms	N:	42 540
90.0%		128.75	ms		
75.0%	quartile	89.06	ms		
50.0%	median	64.56	ms		
25.0%	quartile	46.81	ms		
10.0%		34.75	ms		
2.5%		24.28	ms		
0.5%		17.06	ms		
0.0%	minimum	0.063	ms