Mogelzettel: Deskriptive Statistik

Betrachtet wird eine **Grundgesamtheit** quantitativer Daten $\{x_1, x_2, \ldots x_N\}$, die *alle* möglichen Daten[1] enthält. Dieser unhandliche Datensatz wird durch "typische" Zahlen charakterisiert.

Arithmetischer **Mittelwert:** $\boxed{\langle x\rangle := \overline{x} := \frac{1}{N}\sum_{i=1}^{N} x_i}$

Entsprechend kann der Mittelwert abgeleiteter Größen $f(x_i)$ (z.B. Kugelvolumina für gegebene Radien) definiert werden:

$$\langle f(x)\rangle := \overline{f(x)} := \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Klasseneinteilung

Bei einer großen Anzahl von Werten wird häufig mit einer Klasseneinteilung und Histogrammen gearbeitet. Dies erfolgt oft bereits bei der Datenerhebung.
Klassen: disjunkte, aufeinanderfolgende Intervalle, dh. $[x_i^u, x_i^o)$ oder $(x_i^u, x_i^o]$ $\quad i = 1, \ldots, k$
Klassenbreite: $\Delta x_i = x_i^o - x_i^u \;= x_{i+1}^u - x_i^u$
Als Repräsentant der Klasse wird oft die **Klassenmitte** (Intervallmitte) $\quad x_i = \frac{x_i^o + x_i^u}{2}$ benutzt. Ist n_i die Anzahl der Werte in der i-ten Klasse, so definiert man:

$$N = \sum_{i=1}^{k} n_i \;, \qquad \overline{x} := \frac{1}{N}\sum_{i=1}^{k} n_i x_i \;, \qquad \overline{f(x)} := \frac{1}{N}\sum_{i=1}^{k} n_i f(x_i)$$

was eigentlich nur bei symmetrischen Verteilungen korrekte Resultate liefert.

Alternativen zum arithmetischen Mittelwert:

Geometrisches Mittel: $\boxed{\langle x\rangle_{\mathrm{g}} = \sqrt[n]{x_1 x_2 \cdots x_n}}$ (wichtig für Wachstumsraten)

quadratisches Mittel: $\langle x\rangle_{\mathrm{rms}} = \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}}$ (z.B. Effektivwerte in der Elektrik)

harmonisches Mittel: $\langle x\rangle_{\mathrm{h}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$

Im Alltag sind auch noch der **Median** (Wert in der Mitte) und der **Modus** (häufigster Wert) gebräuchlich.

[1] also keine Stichprobe!

Neben dem Mittelwert ist auch die Breite der Verteilung wichtig, was auf die Varianz und Standardabweichung führt:

Varianz: $$\boxed{V(x) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \ldots = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \boxed{\overline{x^2} - \overline{x}^2}$$

Standardabweichung: $\boxed{\sigma = \sqrt{V(x)}}$ (hat dieselbe Einheit wie x)

Details sind abhängig von der Verteilung, aber als Faustformel:
σ ist eine typische Abweichung vom Mittelwert, Abweichungen um 2σ kommen vor, aber Abweichungen von mehr als 3σ sind verdächtig!

Alternative Breitenmaße:

In der Physik ist die **volle Halbwertsbreite (FWHM)**[2] gebräuchlich, da sie einfach zu bestimmen ist und für viele Verteilungen eine unmittelbare Berechnung von σ erlaubt (für Gaußverteilungen gilt: FWHM$\approx 2.35\sigma$).

Die zum Median gehörenden Breitenmaße sind die Quartile und Dezile; z.B. liegen beim unteren Quartil 25% der Daten darunter und 75% darüber, beim oberen Dezil liegen 90% darunter und 10% darüber.
Seltener: **Spannweite** $\max(x_i) - \min(x_i)$ und **mittlere absolute Abweichung** $\frac{1}{n}\sum_i |x_i - \overline{x}|$

höhere Momente:
m-tes (zentrales) Moment: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^m$ "Momentenentwicklung"

Höhere Momente werden in der Physik eher selten benutzt; Statistiker benutzen auch renormierte Versionen wie *Schiefe* und *Kurtosis*, mit von Autor zu Autor teilweise leicht abweichenden Definitionen.

Paare von Daten - Korrelation

Datensätze: $(x_1, y_1), ..., (x_n, y_n)$ (z.B. Größe und Gewicht)
Information über Zusammenhang zwischen den Größen liefert die **Kovarianz**:

$$\boxed{\mathrm{cov}(x,y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})} = \cdots = \overline{xy} - \overline{x}\,\overline{y}$$

> 0 falls tendenziell große x mit großen y auftreten
< 0 falls große x mit kleinen y auftreten (oder umgekehrt)
$= 0$ falls kein Zusammenhang besteht

Korrelationskoeffizient: $-1 \leq \boxed{\rho = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}} \leq 1$

Eine Korrelation sagt nicht notwendigerweise etwas über kausale Zusammenhänge aus!
(Zahl der Störche / Geburtenrate)

Eine Verallgemeinerung auf k Variablen führt auf die $k \times k$ Kovarianzmatrix V_{ij}.

[2] Full Width at Half Maximum