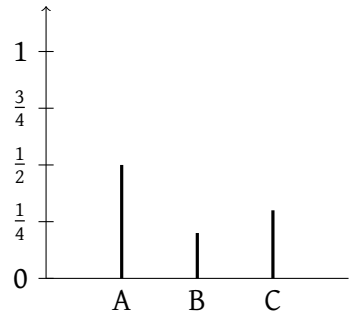
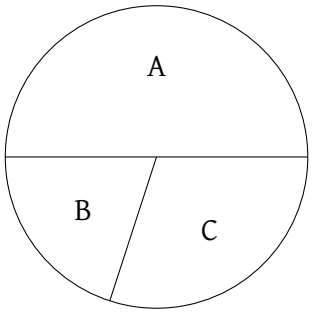
§ 28 Einführung in die Statistik

28.1 Definition: Seien X und Y endliche Mengen, und sei $f : X \to Y$ eine Abbildung. Die Abbildung $Y \to \mathbb{N}_0, y \mapsto |\{x \in X \mid f(x) = y\}|$ wird *(absolute) Häufigkeitsverteilung* von f genannt und mit h_f bezeichnet. Als *relative Häufigkeitsverteilung* r_f von f definiert man die Abbildung $Y \to \mathbb{R}_0^+, y \mapsto \frac{1}{|X|}\mathrm{h}_f$.◇

28.2 Bemerkung: Absolute und relative Häufigkeitsverteilungen lassen sich graphisch darstellen. Ist eine relative Häufigkeitsverteilung r_f gegeben, so wird etwa in einem sogenannten Stabdiagramm mit Skalierungsfaktor a für jedes Element y des Definitionsbereichs von r_f ein Stab der Höhe $a \cdot \mathrm{r}_f(y)$ eingezeichnet, während bei einem sogenannten Kreisdiagramm ein Kreissektor eingezeichnet wird, dessen Winkel gleich $2\pi \cdot \mathrm{r}_f(y)$ ist. ◇

Beispiel: Die relative Häufigkeitsverteilung $\mathrm{r}_f : \{\mathrm{A}, \mathrm{B}, \mathrm{C}\} \to \mathbb{R}_0^+$ mit $\mathrm{r}_f(\mathrm{A}) = \frac{1}{2}$ und $\mathrm{r}_f(\mathrm{B}) = \frac{1}{5}$ sowie $\mathrm{r}_f(\mathrm{C}) = \frac{3}{10}$ läßt wie folgt in einem Stab- beziehungsweise Kreisdiagramm darstellen.





◇

28.3 Definition: Sei $n \in \mathbb{N}$, und sei $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Dann bezeichnet man den Wert $\frac{1}{n}(x_1 + \cdots + x_n)$ als *arithmetisches Mittel* der Datenreihe x und schreibt dafür auch $\bar{x}$. ◇

Beispiel: Das arithmetische Mittel von $(1, 0, -3, 6, 16)$ ist $\frac{1}{5}(1 + 0 - 3 + 6 + 16) = 4$. ◇

28.4 Definition: Sei $n \in \mathbb{N}$, und sei $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Ein *Median* der Datenreihe x ist ein Wert $x^* \in \mathbb{R}$ mit $|\{i \in \{1, \ldots, n\} \mid x_i < x^*\}| \leqslant \frac{n}{2}$ und $|\{i \in \{1, \ldots, n\} \mid x_i > x^*\}| \leqslant \frac{n}{2}$. ◇

Beispiel: Gegeben sei die Datenreihe $(1, 0, -3, 6, 16)$. Für $x^* = 1$ enthält sowohl die Menge der Daten kleiner x^* als auch die Menge der Daten größer x^* höchstens die Hälfte aller Daten (nämlich jeweils genau 2: man hat $0, -3 < 1$ und $6, 16 > 1$). Andere Werte von x^* erfüllen diese Eigenschaft nicht, also ist 1 der einzige Median der Datenreihe. ◇

Beispiel: Gegeben sei die Datenreihe $(1, 0, -3, 6)$. Für $x^* = \frac{1}{2}$ enthält sowohl die Menge der Daten kleiner x^* als auch die Menge der Daten größer x^* höchstens die Hälfte der Daten (nämlich jeweils genau 2: man hat $0, -3 < \frac{1}{2}$ und $1, 6 > \frac{1}{2}$). Diese Eigenschaft wird von allen $x^* \in [0, 1]$ erfüllt, also sind alle Werte aus $[0, 1]$ Mediane der Datenreihe. ◇

28.5 Bemerkung: Die Mediane einer Datenreihe $(x_1, \ldots, x_n)$ lassen sich am besten bestimmen, wenn man die Datenreihe aufsteigend sortiert. Aus der sortierten Datenreihe $(x_{(1)}, \ldots, x_{(n)})$ liest man für n ungerade den Wert $x_{(\frac{n+1}{2})}$ als einzigen Median ab, während für gerades n genau die Werte aus dem Intervall $[x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}]$ die Mediane der Datenreihe sind. ◇

28.6 Definition: Sei $n \in \mathbb{N}$ mit $n \geqslant 2$, und sei $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Die *Stichprobenvarianz* von x wird definiert als $\frac{1}{n-1}((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)$, und die Quadratwurzel aus diesem Wert wird *Stichprobenstandardabweichung* der Datenreihe x genannt und mit s_x bezeichnet. ◇

Beispiel: Die Datenreihe $(1, 0, -3, 6)$ hat $\frac{1}{4}(1 + 0 - 3 + 6) = 1$ als arithmetisches Mittel. Die Stichprobenvarianz berechnet sich damit zu $\frac{1}{3}((1-1)^2+(0-1)^2+(-3-1)^2+(6-1)^2) = \frac{1}{3}\cdot 42 = 14$, und die Stichprobenstandardabweichung ist damit $\sqrt{14} \approx 3{,}74$. ◇

28.7 Bemerkung: Kann das Ergebnis eines Zufallsexperiments durch eine Normalverteilung mit Parametern μ und σ modelliert werden, so kann man, falls $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe des Zufallsexperiments darstellt, den Wert $\bar{x}$ als Näherung von μ und den Wert s_x als Näherung von σ verwenden. ◇

28.8 Definition: Sei $n \in \mathbb{N}$ und $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Datenreihe und $p \in (0, 1)$ sowie $\sigma \in \mathbb{R}^+$. Weiter sei Φ die Verteilungsfunktion der Standardnormalverteilung. Wählt man dann $z \in \mathbb{R}^+$ so, daß $2\Phi(z) - 1 = p$ gilt, dann nennt man $[\bar{x} - \frac{z\sigma}{\sqrt{n}}, \bar{x} + \frac{z\sigma}{\sqrt{n}}]$ das *(symmetrische) Konfidenzintervall* der Datenreihe x zum *Konfidenzniveau* p für die Schätzung des Parameters μ bei Normalverteilung mit Parametern μ und σ. ◇

28.9 Bemerkung: Für die Interpretation des Konfidenzintervalls betrachte ein gegebenes Zufallsexperiment, das durch eine Normalverteilung mit Parametern μ und σ modelliert werden kann. Führt man das Zufallsexperiment n-mal durch und berechnet das Konfidenzintervall der erhaltenen Datenreihe zum Konfidenzniveau p wie in 28.8, so ist der Parameter μ mit Wahrscheinlichkeit p in diesem Konfidenzintervall enthalten. Einige für Konfidenzintervalle wichtige Werte der Verteilungsfunktion der Standardnormalverteilung kann man der folgenden Tabelle entnehmen.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,0000	0,5000	1,1503	0,8750	1,9600	0,9750	2,5758	0,9950
0,2500	0,5987	1,2500	0,8946	2,0000	0,9772	2,7500	0,9970
0,5000	0,6915	1,2816	0,9000	2,2414	0,9875	2,8070	0,9975
0,6745	0,7500	1,5000	0,9332	2,2500	0,9878	3,0000	0,9987
0,7500	0,7734	1,6449	0,9500	2,3263	0,9900	3,0902	0,9990
1,0000	0,8413	1,7500	0,9599	2,5000	0,9938	3,2500	0,9994

◇

Beispiel: Die Datenreihe $x = (106, 98, 103, 100, 102, 103)$ habe man durch sechsmalige Messung einer Größe erhalten. Zu berechnen sei das Konfidenzintervall der Datenreihe x zum Konfidenzniveau 95 % für die Schätzung des Parameters μ bei Normalverteilung mit Parametern μ und σ, wobei σ gleich der Stichprobenstandardabweichung von x ist. Man berechnet zunächst $\bar{x} = 102$

und $s_x \approx 2{,}757$. Nun ist für die Verteilungsfunktion Φ der Standardnormalverteilung das $z \in \mathbb{R}^+$ zu bestimmen, für das $2\Phi(z) - 1 = 0{,}95$ gilt. Man hat

$$2\Phi(z) - 1 = 0{,}95 \quad \Leftrightarrow \quad 2\Phi(z) = 1{,}95 \quad \Leftrightarrow \quad \Phi(z) = 0{,}975.$$

Nach obiger Wertetabelle ist also $z \approx 1{,}9600$ zu wählen. Mit $\sigma = s_x$ berechnet man nun $\frac{z\sigma}{\sqrt{6}} \approx 2{,}206$, also erhält man als Konfidenzintervall zum Konfidenzniveau $95\,\%$ näherungsweise das Intervall $[99{,}8, 104{,}2]$. ◇