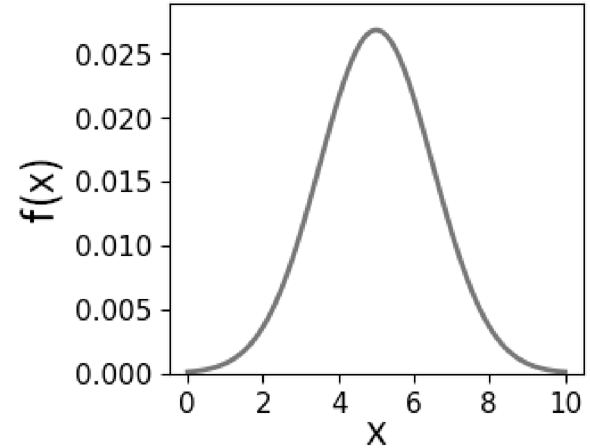
M24 Statistik 1: Sommersemester 2024

Vorlesung 09: Wahrscheinlichkeitsdichte

Prof. Matthias Guggenmos

Health and Medical University Potsdam

Theoretische Häufigkeitsverteilungen



In der letzten Vorlesung haben wir erarbeitet, dass sich die Stichprobenverteilung durch eine Normalverteilung der Form

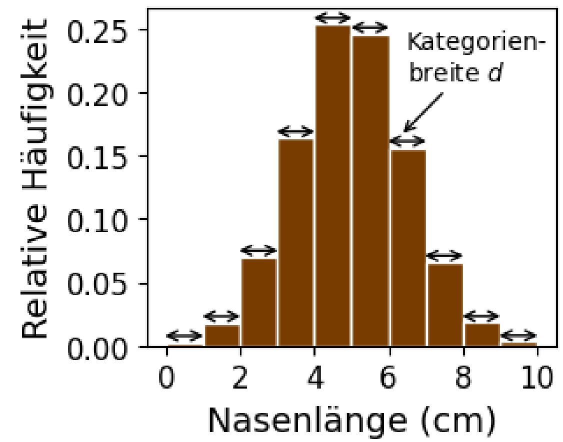
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

beschreiben lässt (wobei die Variable x im Fall der Stichprobenverteilung der Stichprobenkennwert $\hat{\theta}$ ist).

- Eine theoretische Häufigkeitsverteilung wie die Normalverteilung gibt für *jeden beliebigen Wert* x des Merkmals X eine Häufigkeit $f(x)$ an.
- Eine wichtige Frage haben wir bislang jedoch nicht beantwortet: was für eine Art von Häufigkeit ist $f(x)$ an? Wie ist also die y-Achse im Diagramm rechts oben zu interpretieren?

Rückblick: das Histogramm

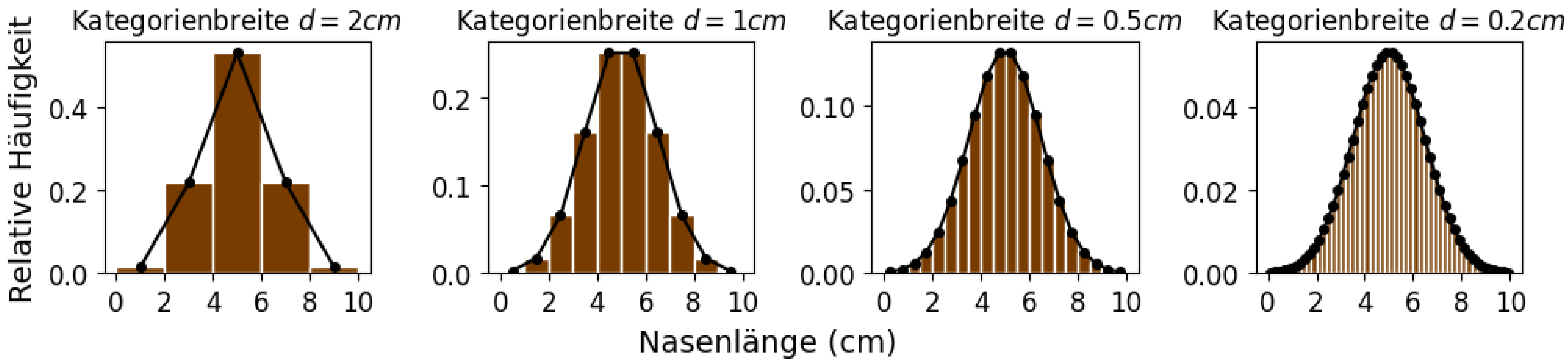
- Um die Bedeutung von $f(x)$ zu verstehen, gehen wir zunächst zurück zum **Histogramm**.
- Histogramme stellen die Häufigkeit der Merkmalsvariable X in einer Stichprobe oder Population dar. Wird die **relative Häufigkeit** aufgetragen, so ordnet das Histogramm jedem Intervall auf der x-Achse eine relative Häufigkeit dar, die z.B. in Prozent angegeben werden kann.



- Die Breite des Intervalls – die **Kategorienbreite *d*** – ist dabei ein Kompromiss zweier Faktoren:
  1. **Auflösung:** je schmaler das Intervall, desto feiner wird das Merkmal X unterteilt.
  2. **Fallzahl:** je breiter das Intervall, desto höher die Zahl der Fälle im Intervall, desto präziser die Schätzung des Häufigkeitswertes im Intervall.
- Die Gesamtsumme aller Säulen im Histogramm mit relativer Häufigkeit ist immer 1 (oder 100%).

Rückblick: das Histogramm

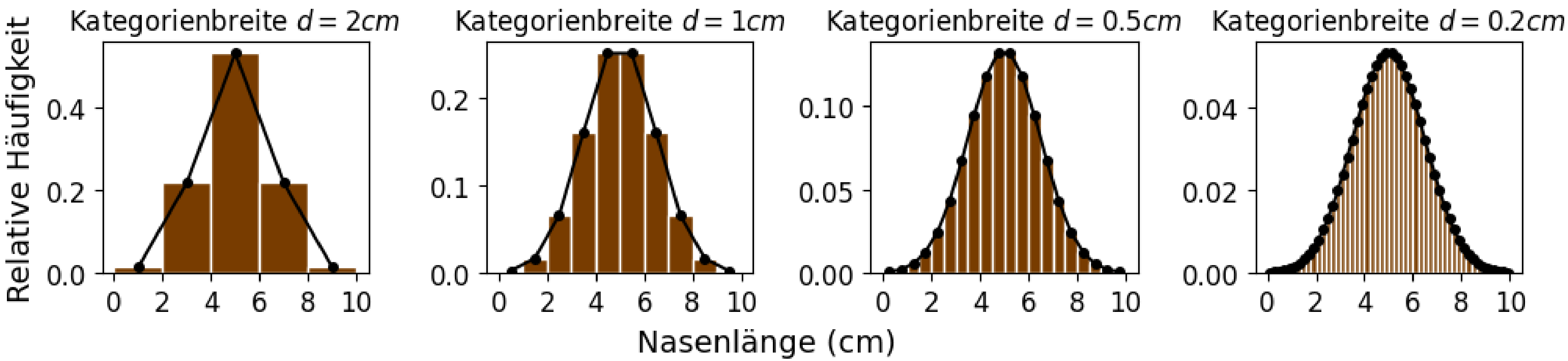
- Umgekehrt gilt: je kleiner die Kategorienbreite, desto mehr Säulen gibt es, desto kleiner die relativen Häufigkeitswerte jeder einzelnen Säule.



- Beispiel:
  - Im ersten Plot gilt $d = 2cm$. Die relative Häufigkeit, dass ein Merkmal etwa im Intervall $[6cm; 8cm]$ liegt ist ca. 0.22 (oder 22%).
  - Im zweiten Plot gilt $d = 1cm$. Nun wird etwa das Intervall $[6cm; 8cm]$ aufgeteilt in zwei relative Häufigkeiten für die Bereiche $[6cm; 7cm]$ und $[7cm; 8cm]$, die jeweils beide geringer sind (ca. 0.16 und 0.06).

Rückblick: das Histogramm

- Nun scheint ein Brückenschlag naheliegend: ist die theoretische Häufigkeitsverteilung $f(x)$, die Funktionswerte für beliebige x-Werte ausgibt, gleich einem Histogramm, bei dem die Kategorienbreite gegen Null geht?



- Im Prinzip ja, allerdings gibt es noch ein Problem: geht die Kategorienbreite d gegen 0, gehen die relativen Häufigkeitswerte des Histogramms ebenfalls gegen Null!
- Würde die theoretische Häufigkeitsverteilung $f(x)$ also relative Häufigkeiten angeben, so wäre $f(x)$ für jedes x Null. Das ist natürlich sinnlos.
- **Theoretische Häufigkeitsverteilungen $f(x)$ geben aus diesem Grund keine relative Häufigkeit an**.
- Bleibt die Frage: was stattdessen?

Von der Wahrscheinlichkeit zur Wahrscheinlichkeitsdichte

Zunächst ein Hinweis zur Nomenklatur:

Definition	Wir verwenden den Begriff **relative Häufigkeiten** bei empirischen Daten und meinen damit den Anteil einer Merkmalsausprägung relativ zu allen Datenpunkten. Beispiel: in einer Stichprobe von 100 Würfelversuchen lag die relative Häufigkeit von Zahlen größer 3 bei 0.48 oder 48%.

Definition	Wir verwenden den Begriff **Wahrscheinlichkeit**, wenn die theoretische Häufigkeitsverteilung eines Merkmals bekannt ist, und meinen damit den Anteil einer Merkmalsausprägung laut Theorie. Beispiel: bei einem perfekten Würfel ist die Wahrscheinlichkeit einer Zahl größer 3 exakt 0.5.

- Im Kontext von theoretischen Häufigkeitsverteilungen können wir daher von **Wahrscheinlichkeiten** sprechen.
- Klar ist auch: durch die Nomenklaturänderung *relative Häufigkeit* → *Wahrscheinlichkeit* ist noch nichts gewonnen.

Von der Wahrscheinlichkeit zur Wahrscheinlichkeitsdichte

Der entscheidende Trick theoretischer Häufigkeitsverteilungen ist der **Übergang von Wahrscheinlichkeiten zu Wahrscheinlichkeitsdichten**.

Ist das Merkmal X eine kontinuierliche Variable (z.B. Nasenlänge in cm), so geben theoretische Häufigkeitsverteilungen $f(x)$ eine **Wahrscheinlichkeitsdichte** an.

Wie kann man sich "Wahrscheinlichkeitsdichte" vorstellen?

- Wir kennen das Konzept der "Dichte" bei Stoffen: z.B. ist die Dichte von Eis ist ca. $0{,}918\ {}^{g}/_{cm^3}$, d.h. dass sich in einem $1cm^3$ Würfel ein knappes Gramm Eis befindet.
- Eine Dichte ist also immer eine bestimmte Masse *pro* Maßeinheit.

Wir können daher Wahrscheinlichkeitsdichte wie folgt definieren:

Definition
$$\text{Wahrscheinlichkeitsdichte} = \text{Wahrscheinlichkeits(masse)}\ \textit{pro}\ \text{Maßeinheit}$$

Die Maßeinheit ist somit durch die Skala unseres Merkmals gegeben: Wahrscheinlichkeit *pro* Zentimeter Nasenlänge, Wahrscheinlichkeit *pro* IQ-Punkt, Wahrscheinlichkeit *pro* Fragebogenpunkt.

Wahrscheinlichkeitsdichte

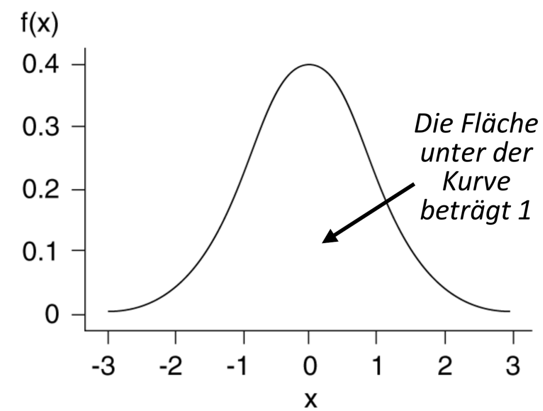
- Theoretische Häufigkeitsverteilungen $f(x)$ für kontinulierliche Merkmale X werden auch als **Wahrscheinlichkeitsdichtefunktion** bezeichnet (engl. *probability density function*).
- Wie bei Histogrammen mit relativen Häufigkeiten ist die gesamte Wahrscheinlichkeits*masse* von Wahrscheinlichkeitsdichtefunktionen $f(x)$ immer 1. Sie sind **normalisiert**.

Beispiel Normalverteilung:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Der Normalisierungsfaktor $\frac{1}{\sigma\sqrt{2\pi}}$ sorgt in diesem Fall dafür, dass die Fläche unter der Normalverteilung gleich 1 ist:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

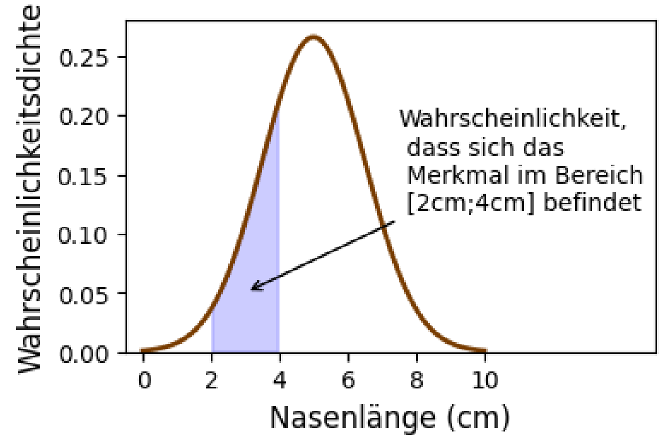


Von der Wahrscheinlichkeitsdichte zurück zur Wahrscheinlichkeit

- Um aus einer Wahrscheinlichkeits*dichte* eine Wahrscheinlichkeit zu erhalten, muss die Dichte über einen bestimmten **Wertebereich** $[x_0; x_1]$ des Merkmals summiert (integriert) werden.
- Mathematisch beschreiben wir diese Operation als ein Integral:

$$P(x_0 < x < x_1) = \int_{x_0}^{x_1} f(x)dx$$

- P ist die Wahrscheinlichkeit, dass das Merkmal einen Wert zwischen x_0 (Untergrenze) und x_1 (Obergrenze) aufweist.
- Das Integral setzt die *Wahrscheinlichkeitsdichte* $f(x)$ mit der *Wahrscheinlichkeit* $P(x_0 < x < x_1)$ in Verbindung.

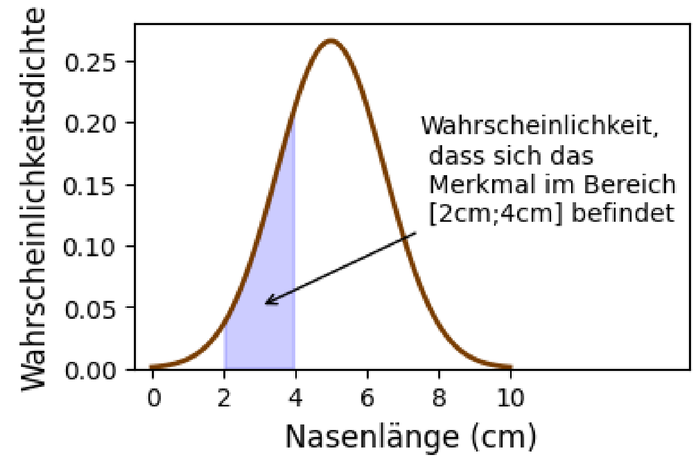


Berechnung einer Wahrscheinlichkeit P auf Basis einer Wahrscheinlichkeitsdichtefunktion $f(x)$ (hier der Normalverteilung).

Wahrscheinlichkeitsdichte: Beispiel 1

Nehmen wir an, dass Nasenlängen in der Population *normalverteilt* sind, mit Mittelwert $\mu = 5$ und Standardabweichung $\sigma = 1{,}5$.

Frage: wie hoch ist die Wahrscheinlichkeit, dass eine zufällig gezogene Nase aus der Population eine Länge zwischen $2cm$ und $4cm$ hat?



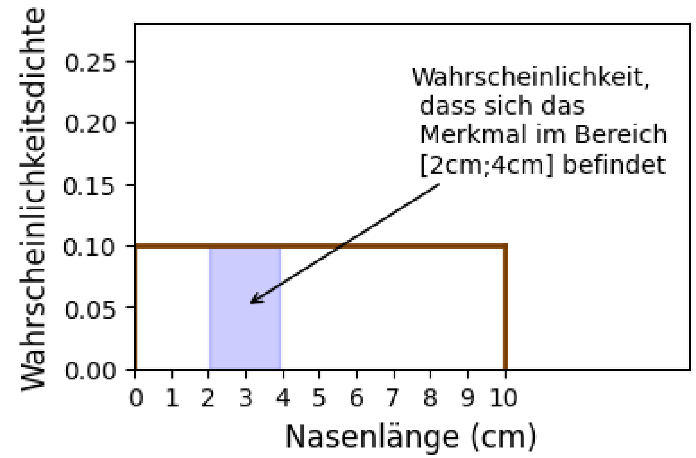
$$P(2 \leq x \leq 4) = \int_2^4 f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_2^4 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx =$$

$$= \frac{1}{1{,}5\sqrt{2\pi}} \int_2^4 \exp\left(-\frac{(x-5)^2}{2 \cdot 1{,}5^2}\right) dx \overset{(Computer!)}{\approx} 0{,}23$$

Wahrscheinlichkeitsdichte: Beispiel 2

Nehmen wir nun an, dass Nasenlängen in der Population *uniform zwischen 0 und 10 cm verteilt* sind.

Gleiche Frage: wie hoch ist die Wahrscheinlichkeit, dass eine zufällig gezogene Nase aus der Population eine Länge zwischen $2cm$ und $4cm$ hat?



Wir wissen: die Fläche unter der Verteilung muss 1 sein. Daher muss die Wahrscheinlichkeitsdichte für jeden Wert zwischen $0cm$ und $10cm$ gleich $0.1cm^{-1}$ betragen ($10cm \cdot 0.1cm^{-1} = 1$).

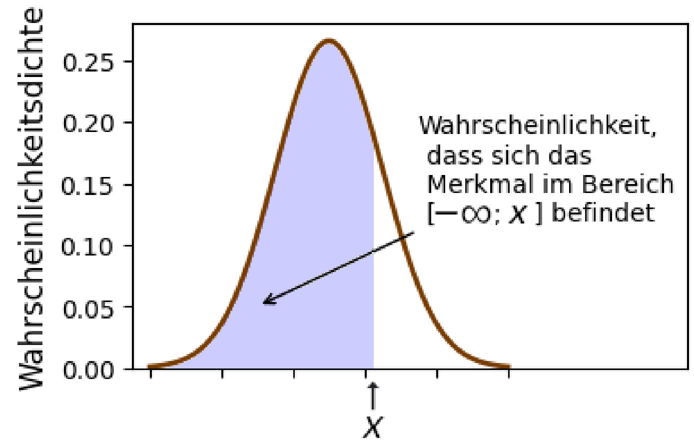
Die Berechnung des Flächeninhalts im Intervall $[2cm; 4cm]$ geht in diesem Fall ohne Integration, denn er entspricht der Fläche eines Rechteckes mit Breite $2cm$ und Höhe $0.1cm^{-1}$. Es gilt:

$$Wahrscheinlichkeit = Intervallbreite \cdot Wahrscheinlichkeitsdichte =$$
$$= 2cm \cdot 0.1cm^{-1} = 0.2$$

Verteilungsfunktion

Die Integration einer Wahrscheinlichkeitsdichte *bis zu einem bestimmten Wert* x ist ein sehr häufiger Fall im Umgang mit Wahrscheinlichkeitsdichten. Daher definieren wir dafür eine eigene Funktion, die **Verteilungsfunktion** $F(x)$:

$$F(x) = \int_{-\infty}^{x} f(x')dx'$$



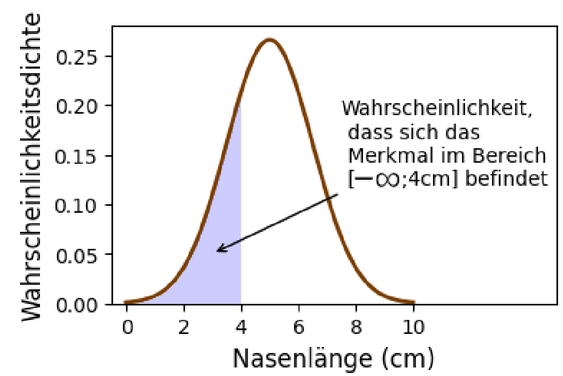
Die Verteilungsfunktion F gibt uns den Flächeninhalt der Dichtefunktion f "links von x" an.

Beispiel

Nehmen wir wieder die normalverteilte Nasenlängen-Population an mit Mittelwert $\mu = 5$ und Standardabweichung $\sigma = 1{,}5$. Die Wahrscheinlichkeit, dass eine zufällig gezogene Nase eine Länge kleiner $4cm$ hat, ist gegeben durch den Wert $F(4)$ der Verteilungsfunktion dieser Normalverteilung:

$$F(4) = \int_{-\infty}^{4} f(x')dx' =$$

$$= \frac{1}{1{,}5\sqrt{2\pi}} \int_{-\infty}^{4} \exp\left(-\frac{(x'-5)^2}{2 \cdot 1{,}5^2}\right) dx' \overset{(Computer!)}{\approx} 0{,}25$$



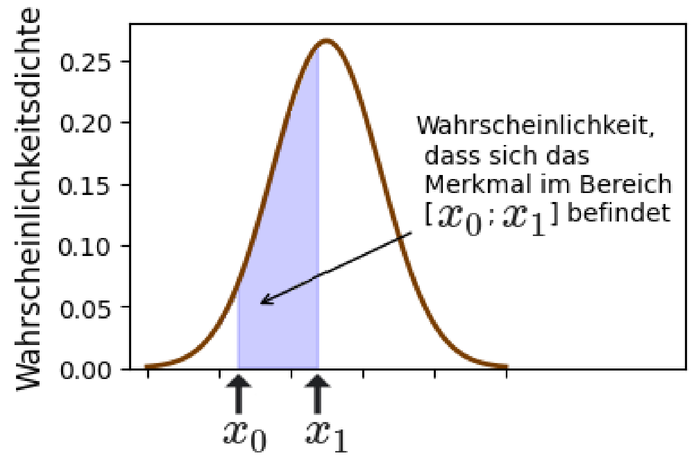
Verteilungsfunktion

Mithilfe der Verteilungsfunktion, lässt sich nun das Integral

$$P(x_0 < x < x_1) = \int_{x_0}^{x_1} f(x)dx$$

mit dem wir die Fläche zwischen einer Untergrenz x_0 und Obergrenze x_1 berechnen, auch folgendermaßen aufstellen:

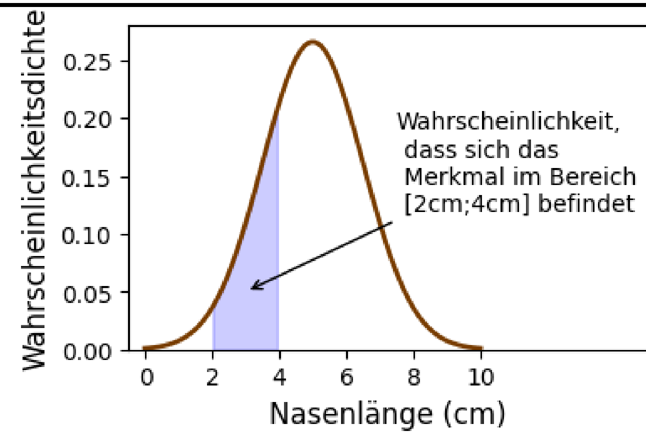
$$P(x_0 < x < x_1) = F(x_1) - F(x_0)$$



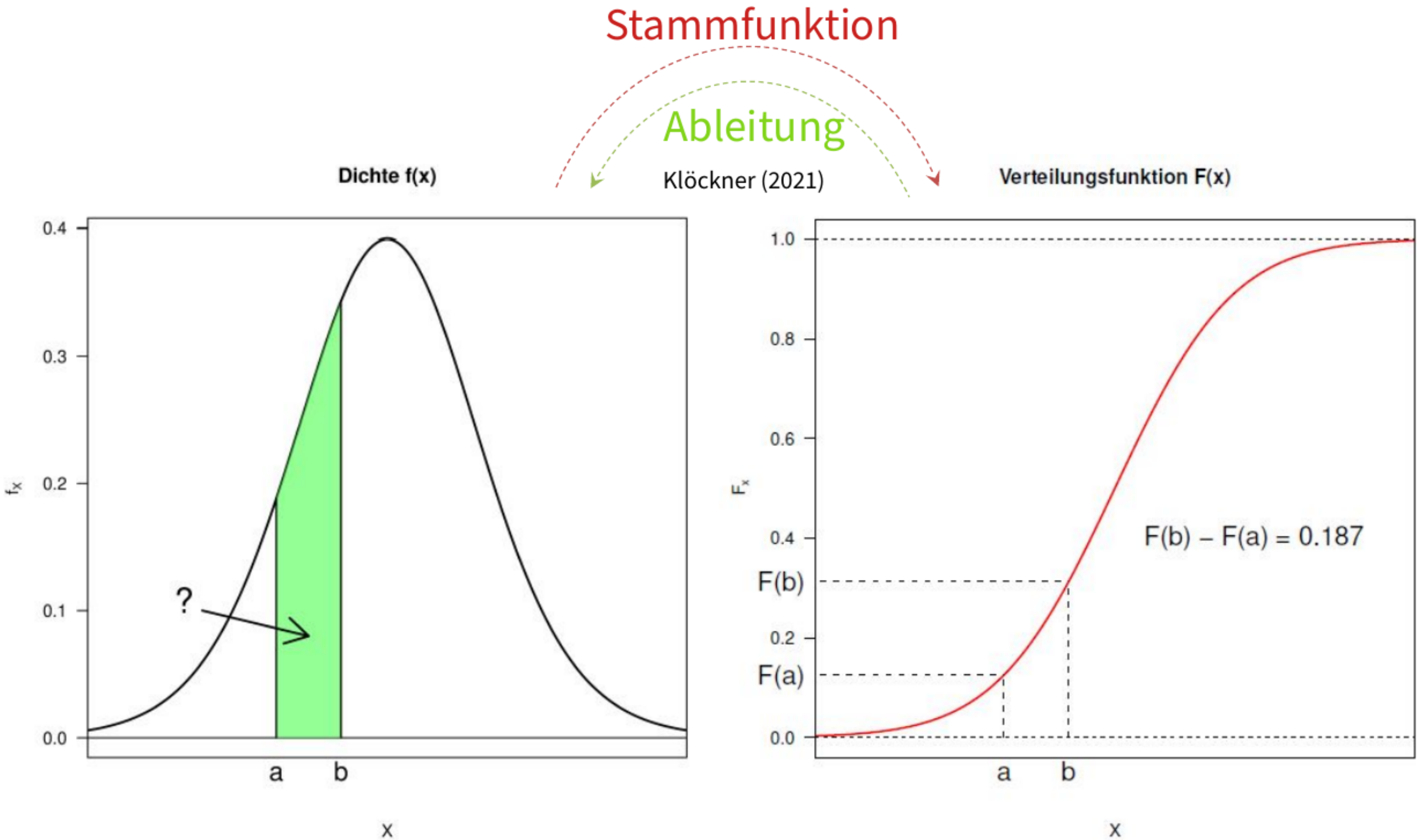
Beispiel

Die eingezeichnete Fläche aus unserem vorherigen Beispiel lässt sich berechnen als:

$$P(2 < x < 4) = F(4) - F(2) \overset{(Computer!)}{\approx} 0{,}23$$



Verteilungsfunktion und Stammfunktion

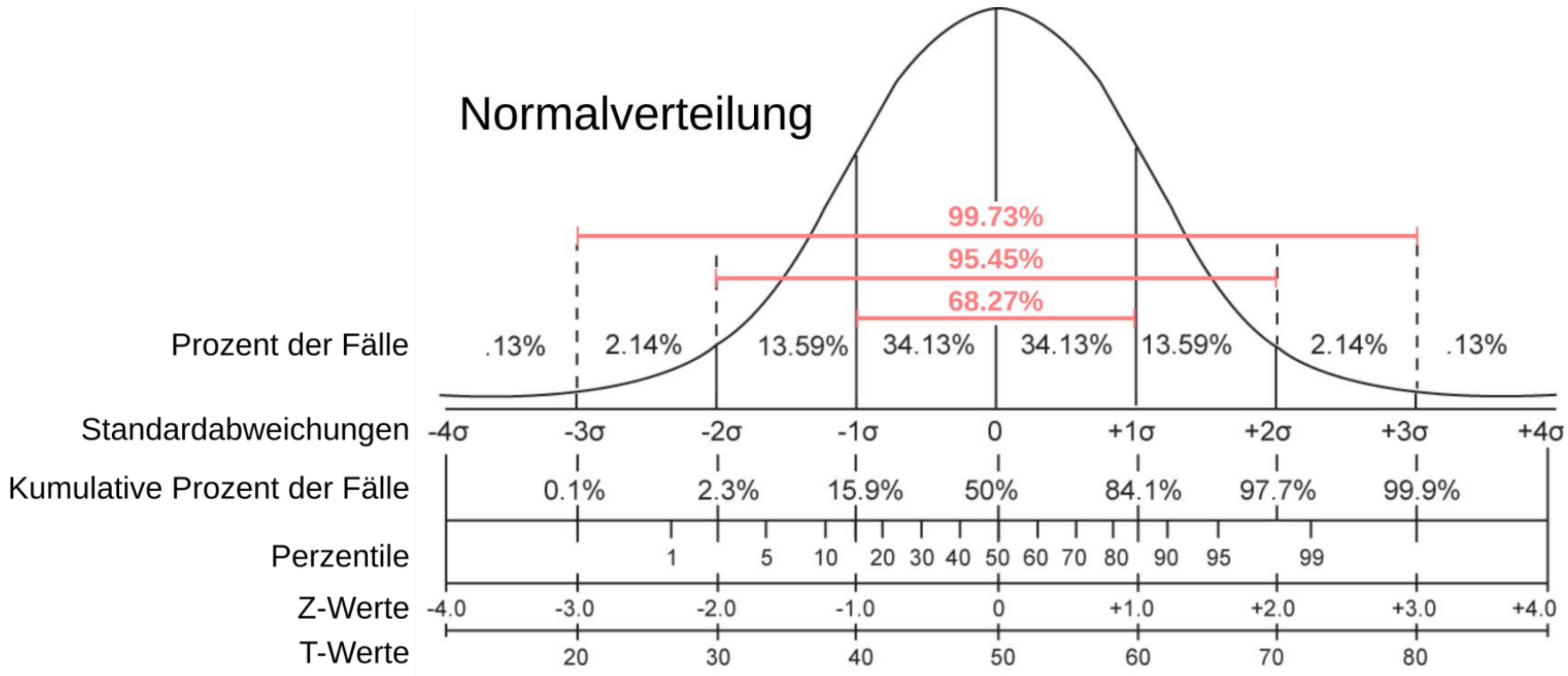


- $F(x)$ ist eine Stammfunktion von $f(x)$ wenn gilt: $\frac{dF}{dx} = f(x)$ bzw. $F(x) = \int_a^x f(x')dx'$.
- Die Verteilungsfunktion $F(x)$ entspricht der Stammfunktion $\int_a^x f(x')dx'$ mit $a = -\infty$.

68-95-99.7-Prozentregel

Mithilfe der Verteilungsfunktion lassen sich charakteristische Flächeninhalte der Normalverteilung berechnen. Als Faustregel ergibt sich die **68-95-99.7-Prozentregel**:

- Der Bereich Mittelwert $\pm$ eine Standardabweichung ($\mu \pm 1\sigma$) umfasst **68%** der Daten
- Der Bereich Mittelwert $\pm$ zwei Standardabweichungen ($\mu \pm 2\sigma$) umfasst **95%** der Daten
- Der Bereich Mittelwert $\pm$ drei Standardabweichungen ($\mu \pm 3\sigma$) umfasst **99.7%** der Daten



Vorschau

Im nächsten Schritt kehren wir zurück zur **theoretischen Stichprobenverteilung**. Die Erkenntnisse zur Wahrscheinlichkeitsdichte und Verteilungsfunktion lassen sich auf die theoretische Stichprobenverteilung übertragen und eröffnen so zwei wesentliche Methoden der Inferenzstatistik:

- **Hypothesentestung** bzw. **Signifikanztestung** (u.a. auch Idee des p-Wertes)
- **Konfidenzintervalle** (Verallgemeinerung des Standardfehlers)