Numerische Integration und Normalverteilung

Carl-Ferdinand Oliver[*], Tom Springer [†], Jonas Münz [‡]
Max Valentin Orlemann [§], Alexander Kögel [¶]
TES23

1. November 2024

Inhaltsverzeichnis

[*] es23035@lehre.dhbw-stuttgart.de
[†] es23026@lehre.dhbw-stuttgart.de
[‡] es23039@lehre.dhbw-stuttgart.de
[§] es23023@lehre.dhbw-stuttgart.de
[¶] es23031@lehre.dhbw-stuttgart.de

1 Statistische Grundlagen

1.1 Normalverteilung

Die Gaußsche Normalverteilung beschreibt die Verteilung einer stetigen Zufallsvariablen und ist durch zwei Parameter, den Mittelwert μ und die Standardabweichung σ, charakterisiert. Sie wird aufgrund ihrer Form auch als *Glockenkurve* (s. Abbildung 1) bezeichnet und ist symmetrisch um den Mittelwert.

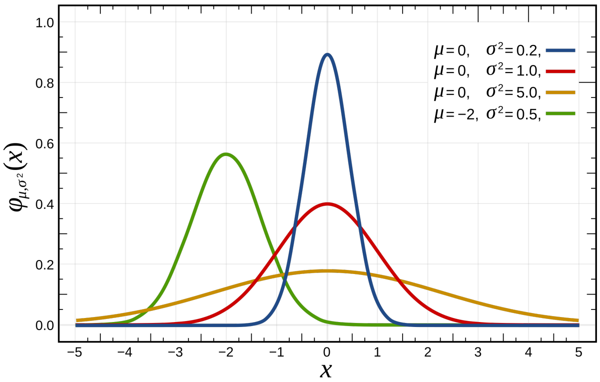

Abbildung 1: Auswahl verschiedener normalverteilten Funktionen [1]

Wichtige Eigenschaften

- Die Dichtefunktion der Normalverteilung lautet:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot u^2\right) \quad (-\infty < u < \infty)$$

- Die Verteilungsfunktion ist gegeben durch:

$$\Phi(u) = P(U \leq u) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} \cdot t^2\right) dt$$

- Die Normalverteilung ist symmetrisch zum Mittelwert μ, hat ein Maximum bei $x = \mu$ und Wendepunkte bei $x = \mu \pm \sigma$.

- Sie ist normiert, d.h., das Integral über die gesamte Dichtefunktion ergibt 1:

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

Berechnung der Wahrscheinlichkeiten

Für die Berechnung von Wahrscheinlichkeiten, die der Gaußschen Normalverteilung folgen, werden in der Praxis oft Tabellen oder eben numerische Verfahren verwendet, da die Integrale analytisch nicht lösbar sind (deswegen machen wir ja den ganzen Spaß mit der Trapezregel). Die Wahrscheinlichkeit für einen Bereich $a \leq X \leq b$ kann mit der Verteilungsfunktion berechnet werden:

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Wahrscheinlichkeiten in Abhängigkeit der Standardabweichung

Da die Normalverteilung symmetrisch um den Mittelwert μ ist, lassen sich bestimmte Wahrscheinlichkeitsbereiche in Abhängigkeit von der Standardabweichung σ um μ angeben:

- Etwa 68,27% aller Werte einer normalverteilten Zufallsvariablen liegen im Intervall $[\mu - \sigma, \mu + \sigma]$, also innerhalb einer Standardabweichung um den Mittelwert.
- Etwa 95,45% der Werte befinden sich im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$, also innerhalb von zwei Standardabweichungen.
- Rund 99,73% der Werte liegen im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$, also innerhalb von drei Standardabweichungen.

Diese Bereiche werden auch als *Empirische Regel* oder *68-95-99,7-Regel* bezeichnet und sind besonders nützlich zur Einschätzung, wie wahrscheinlich es ist, dass eine Zufallsvariable in einem bestimmten Bereich um den Mittelwert liegt. Die Wahrscheinlichkeit, dass ein Wert weiter als 3σ vom Mittelwert entfernt liegt, ist sehr gering und beträgt nur ca. 0,27%.
In der Praxis wird die Normalverteilung zur Modellierung vieler realer Phänomene verwendet, wie z.B. Fehlerverteilungen, Körpergrößen oder Filteralgorithmen (z.B. Kalman Filter/EKF), da viele natürliche Prozesse einem normalverteilten Muster folgen. [2] [1]

1.2 Standardnormalverteilung

Eine Normalverteilung lässt sich durch die Standardisierung mit $z = \frac{x-\mu}{\sigma}$ auf die sogenannte Standardnormalverteilung $\mathcal{N}(0,1)$ zurückführen, die Mittelwert 0 und Standardabweichung 1 hat:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{z^2}{2}\right)$$

Die Standardnormalverteilung ist ein nützliches Werkzeug in der Statistik, da sie es ermöglicht, beliebige Normalverteilungen durch eine Transformation auf diese Standardform zu bringen. Dies wird durch die Berechnung des *z-Scores* erreicht:

$$z = \frac{X-\mu}{\sigma}$$

Der z-Wert gibt an, wie viele Standardabweichungen ein Wert X vom Mittelwert entfernt ist. Diese Transformation erleichtert die Berechnung von Wahrscheinlichkeiten und statistischen Tests, da die Eigenschaften der Standardnormalverteilung bekannt sind und in Tabellen dokumentiert werden. Die Standardnormalverteilung ist also quasi eine Normalverteilung wo der Hochpunnkt bei $x = 0$ liegt

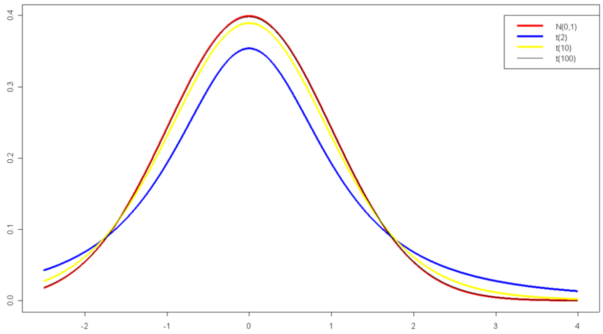

Abbildung 2: Dichten von t-verteilten Zufallsgrößen [3]

1.3 t-Student-Verteilung

Die standardisierte Schätzfunktion des Stichprobenmittelwerts normalverteilter Daten ist nicht mehr normalverteilt, sondern t-verteilt, wenn die zur Standardisierung des Mittelwerts benötigte Varianz des Merkmals unbekannt ist und mit der Stichprobenvarianz geschätzt werden muss. Seine t-Verteilung ermöglicht die Berechnung der Verteilung der Differenz zwischen dem Mittelwert der Stichprobe und dem wahren Mittelwert der Grundgesamtheit.
Die t-Werte hängen vom Signifikanzniveau sowie von der Stichprobengröße n ab und bestimmen das Vertrauensintervall sowie die Aussagekraft der Schätzung des Mittelwertes. Die t-Verteilung wird mit wachsendem n schmaler und konvergiert für $n \to \infty$ zur Standardnormalverteilung. Hypothesentests, bei denen die t-Verteilung Verwendung findet, bezeichnet man als t-Tests. [3]
Die t-Student-Verteilung wird verwendet, wenn die Stichprobengröße klein ist ($n < 30$) und die Varianz der Grundgesamtheit unbekannt ist. Sie ähnelt der Normalverteilung (s. Abbildung 2), hat jedoch „schwerere" Ränder. Dies bedeutet, dass extremere Werte wahrscheinlicher sind als bei der Normalverteilung. Die t-Student-Verteilung hat einen Parameter, die Freiheitsgrade df, der die Form der Verteilung bestimmt. Je größer die Freiheitsgrade, desto mehr nähert sich die t-Verteilung der Normalverteilung an. Bei einer großen Stichprobengröße (ca. $df > 30$) ist der Unterschied zur Normalverteilung vernachlässigbar. In der Statistik wird die t-Verteilung häufig für Hypothesentests verwendet. Sie spielt auch bei der Berechnung von Konfidenzintervallen eine wichtige Rolle, wenn die Varianz der Grundgesamtheit nicht bekannt ist.

1.4 χ^2-Verteilung

Die **χ^2-Verteilung** entsteht als Summe der quadrierten z-Werte aus einer Standardnormalverteilung. Sie wird häufig in der Statistik verwendet, um Varianzen zu schätzen und in Hypothesentests wie dem *Chi²-Anpassungstest* oder dem *Chi²-Test zur Unabhängigkeit*. Die Chi²-Verteilung hat einen Parameter, die Freiheitsgrade df, der die Form der Verteilung bestimmt. Sie ist asymmetrisch und nur für positive Werte definiert. Für $df \geq 30$ nähert sich die Chi²-Verteilung zunehmend der Normalverteilung an. Ein häufiges Anwendungsgebiet der Chi²-Verteilung ist die Analyse von Kontingenztafeln, bei denen überprüft wird, ob zwei kategorische Variablen unabhängig voneinander sind.

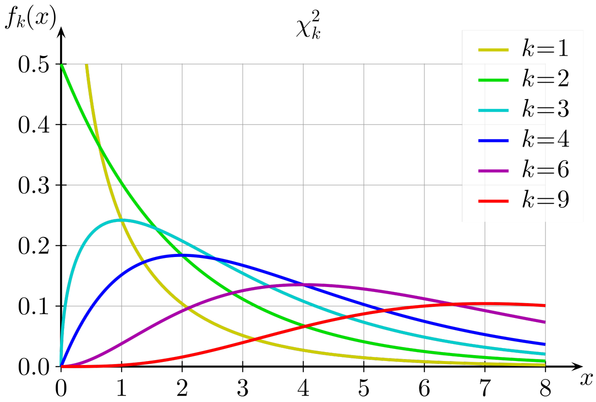

Abbildung 3: Weitere Einzelheiten Dichten der Chi-Quadrat-Verteilung mit unterschiedlicher Anzahl an Freiheitsgraden k [4]

2 Trapezregel

Diese brauche wir zur numerischen Integration, beipielweise um Warscheinlichkeit in einer Normalverteilung zu berechnen. Diese Integrale können, wie in Kapitel 1 besprochen, nicht analytisch gelöst werden und müssen damit genähert werden. Die Trapezregel ist eine einfache Methode hierzu.

2.1 Herleitung

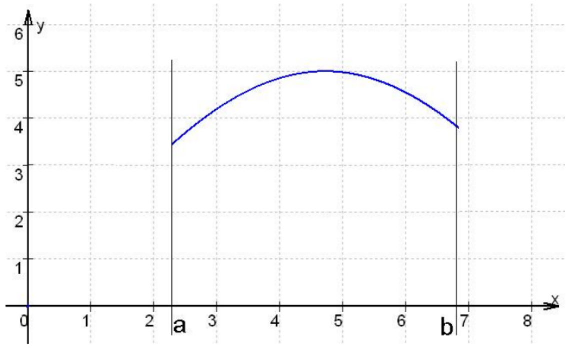

Abbildung 4: Keplersche Fassregel [5]

Wir betrachten eine Funktion f, deren Schaubild im gewünschten Intervall $I = [a, b]$ in Abbildung 5 gezeigt ist.

A. Die zwei Sehnentrapeze

Zur Flächenberechnung zeichnen wir zwei Sehnentrapeze in das gegebene Schaubild, wie in Abbildung 5 zu sehen ist. Für die Fläche eines Trapezes mit den Grundseiten a und b und der Höhe h gilt:

$$A_{\text{Trapez}} = \frac{a+b}{2} \cdot h = m \cdot h$$

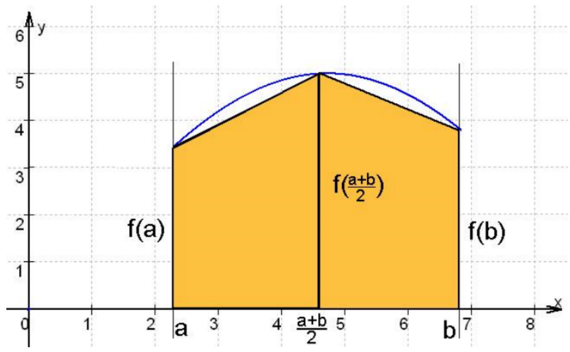

Abbildung 5: Sehnentrapeze für die anstehende Rechnung

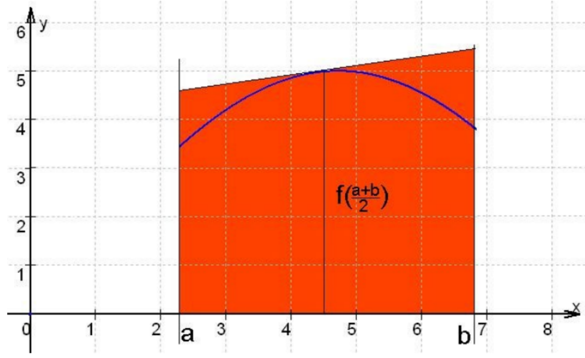

Abbildung 6: Tangententrapez [5]

Übertragen auf die beiden Sehnentrapeze aus Abbildung 5 ergibt sich die Gesamtfläche S als:

$$S = \frac{b-a}{2} \cdot \left(\frac{f(a) + f\left(\frac{a+b}{2}\right)}{2} + \frac{f(b) + f\left(\frac{a+b}{2}\right)}{2} \right)$$

$$\Leftrightarrow S = \frac{b-a}{2} \cdot \left(\frac{f(a)}{2} + f\left(\frac{a+b}{2}\right) + \frac{f(b)}{2} \right)$$

B. Das Tangententrapez

Nun legen wir ein weiteres Tangententrapez in das Schaubild, wie in Abbildung 6 dargestellt. Nach der Trapezformel ergibt sich:

$$T = f\left(\frac{a+b}{2}\right) \cdot (b-a)$$

Da wir **doppelt so viele Sehnentrapeze wie Tangententrapeze** haben, gewichten wir die Flächen entsprechend. Die Gesamtfläche $I[f]$ nähert sich durch:

$$I[f] \approx A = \frac{1}{3} \cdot (2S + T)$$

$$\Leftrightarrow A = \frac{b-a}{6} \cdot \left(f(a) + 4 \cdot f\left(\frac{a+b}{2}\right) + f(b) \right)$$

Die Keplersche Fassregel

Ist die Funktion f auf dem Intervall $I = [a, b]$ stetig, so gilt die Keplersche Fassregel:

$$I[f] \approx A = \frac{b-a}{6} \cdot \left(f(a) + 4 \cdot f\left(\frac{a+b}{2}\right) + f(b) \right)$$

Dabei ist zu beachten, dass die Keplersche Fassregel nur dann gute Näherungswerte liefert, wenn sich die Funktion im betrachteten Intervall durch eine Parabel annähern lässt (z.B. eine Normalverteilung). Daher ist es ratsam, das Intervall $[a, b]$ in **viele kleine, gleich große Teilintervalle** zu unterteilen und die Keplersche Fassregel auf jedes Teilintervall anzuwenden. Dabei sollte für eine bessere Genauigkeit das Intervall in möglichst viele, kleine Intervalle unterteilt werden. [5]

2.2 Fehlerabschätzung

Für die Fehlerberechnung muss zunächst überhaupt definiert werden, was dieser Fehler ist:

Definition des Fehlers

Der Fehler wird definiert als die Breite des Konfidenzintervalls:

$$\text{Breite} = 2 \cdot c \cdot \frac{\sigma}{\sqrt{n}}$$

Hierbei ist c eine Konstante und wird mit der Trapezregel berechnet.

Berechnung des Standardfehlers

Der Standardfehler (SE) des Mittelwerts wird berechnet als:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Daraus folgt die Fehlerabschätzung:

$$\text{Fehlerabschätzung} = c \cdot SE$$

Interpretation

Ein kleiner Fehler zeigt an, dass das Konfidenzintervall eine präzise Schätzung des wahren Mittelwerts bietet. Ein größerer Fehler deutet auf mögliche Unsicherheiten in den Daten oder eine unzureichende Stichprobengröße hin. Beispielsweise können so auch die Genauigkeit von Sensordaten oder ähnlichen eingeschätzt werden.

2.3 Vertrauensniveau γ

Das Vertrauensniveau γ ist ein Wert, mit dem Angegeben wird wie sehr den Messwerte den wahren Wert des Parameters enthält (sprich wie genau die Sensoren sind). Es wird in der Regel als Dezimalzahl zwischen 0 und 1 dargestellt, wobei typische Werte 0.95 (95%) und 0.99 (99%) sind.

- **Einfluss auf die Breite des Intervalls:** Ein höheres Vertrauensniveau führt zu einem breiteren Konfidenzintervall. Das bedeutet, dass wir mit größerer Sicherheit sagen können, dass der wahre Wert in diesem Intervall liegt. Ein niedrigeres Vertrauensniveau hingegen führt zu einem schmaleren Intervall, was die Schätzung präziser macht, aber auch das Risiko erhöht, dass der wahre Wert außerhalb liegt.
- **Berechnung:** Das Vertrauensniveau wird verwendet, um kritische Werte zu bestimmen, die notwendig sind, um das Konfidenzintervall zu berechnen.

3 Umsetzung im Programm

Die Umsetzung des Programms erfolgt in Python mithilfe von numpy. Die Berechnung des Mittelwertes, der Grenzen des Konfidenzintervalls, der Varianz σ^2 und der Konstante c erfolgt in der Funktion `konfidenz(dateiname, gamma)`.

3.1 Berechnung des Mittelwerts

Der Mittelwert (μ) der Daten wird berechnet, indem die Summe aller Datenpunkte durch die Anzahl der Punkte (n) geteilt wird:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} A_i$$

3.2 Berechnung der Varianz

Die Varianz (σ^2) wird berechnet, um zu messen, wie die Daten um den Mittelwert streuen. Der Code verwendet folgende Formel für die erwartungstreue Varianz:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(A_i - \mu)^2$$

3.3 Berechnung der Konstante c durch numerische Integration

Der Code verwendet die Trapezregel, um die Fläche unter der Standardnormalverteilung zu berechnen. Die Formel wird in Unterkapitel 2.1 hergeleitet.

3.3.1 Ziel der Berechnung

Die Fläche unter der Standardnormalverteilung wird so lange berechnet, bis sie einen bestimmten Zielwert erreicht, der durch $(\gamma + 1)/2$ definiert ist. Dies entspricht dem gewünschten Konfidenzniveau (hier gewählt als 99% $\vee$ 0,99).

3.3.2 Funktion der Standardnormalverteilung

Die Funktion `standardnormalverteilung(x)` berechnet den Wert der Standardnormalverteilung für einen gegebenen Wert x:

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

3.4 Berechnung des Konfidenzintervalls

Nach der Berechnung der Konstante c wird das Konfidenzintervall mit der folgenden Formel berechnet:

$$\text{Oben} = \mu + c \cdot \sqrt{\frac{\sigma^2}{n}}$$

$$\text{Unten} = \mu - c \cdot \sqrt{\frac{\sigma^2}{n}}$$

Damit können wir die beiden Grenzen des Konfidenzintervalls herausfinden, zwischen denen mit einer bestimmten Wahrscheinlichkeit (von uns gewählt als 99%) der wahre Mittelwert der Grundgesamtheit liegt.

Literatur

[1] Wikipedia. *Normalverteilung — Wikipedia, die freie Enzyklopädie*. [Online; Stand 30. Oktober 2024]. 2024. URL: https://de.wikipedia.org/w/index.php?title=Normalverteilung&oldid=248003193.

[2] J. Gehrke. *MATHEMATIK 4 – Mathematische Anwendungen; Statistik Grundlagen*. ZDM und SKT an der DHBW Stuttgart, Abgerufen am 28.10.2024. 2018.

[3] Wikipedia. *Studentsche t-Verteilung — Wikipedia, die freie Enzyklopädie*. [Online; Stand 29. Oktober 2024]. 2024. URL: https://de.wikipedia.org/w/index.php?title=Studentsche_t-Verteilung&oldid=249125602.

[4] Wikipedia. *Chi-Quadrat-Verteilung — Wikipedia, die freie Enzyklopädie*. [Online; Stand 29. Oktober 2024]. 2024. URL: https://de.wikipedia.org/w/index.php?title=Chi-Quadrat-Verteilung&oldid=249446964.

[5] J. Gehrke. *Skript zur Mathematik 4 Mathematische Anwendungen*. Stand: 2022, ZDM und SKT an der DHBW Stuttgart, Abgerufen am 28.10.2024. 2022.