Prof. Dr. Klaus Eckhardt

Standardisierung der Normalverteilung

veröffentlicht im Internet unter aufgabomat.de

Gliederung:

1 Vorwort

Auf die Einführung einer mathematischen Gleichung in einer Lehrveranstaltung folgt erfahrungsgemäß schnell die Frage nach einem Rechenbeispiel. Wie hilfreich Rechenbeispiele bei Gelegenheit auch sein mögen, kann dieser Wunsch doch in eine völlig falsche Richtung führen.

Im vorliegenden Skript beispielsweise werden mehrere Grundgleichungen der Statistik vorgestellt. Dabei ist die Rede von wenig vertrauten Begriffen wie Wahrscheinlichkeitsdichte, treten Integrale auf und Exponentialfunktionen mit kompliziert aussehenden Exponenten. Mit diesen Gleichungen Berechnungen durchzuführen, wäre tatsächlich nicht einfach. Es ist an dieser Stelle aber überhaupt nicht notwendig, sich mit solchen Berechnungen eingehender zu befassen. Dass es in der Mathematik darum gehe, irgendetwas zu berechnen, ist ein weit verbreiteter Irrtum. In der Mathematik geht es vielmehr darum, Strukturen und Zusammenhänge zu erkennen.

Nehmen Sie daher die in den Abschnitten 2 und 3 vorgestellten Gleichungen als genau das und nur das, was sie sind: die kompakte Darstellung von Zusammenhängen in der Sprache der Mathematik. Sie müssen mit diesen Zusammenhängen nichts berechnen, Sie müssen diese Zusammenhänge aber kennen, denn diese sind fundamental für die Statistik. Auf der Grundlage eines Verständnisses von Zusammenhängen können Sie Ihr Wissen dann erweitern. Das gilt hier und das gilt auch sonst. In Abschnitt 4 wird abschließend auch etwas zu berechnen sein.

2 Wahrscheinlichkeitsdichte- und Verteilungsfunktion

Eine stetige Zufallsvariable X ist dann vollständig beschrieben, wenn ihre **Wahrscheinlichkeitsdichtefunktion** f(x) bekannt ist[1]. Die beiden zentralen statistischen Kenngröße zur Beschreibung von Zufallsvariablen, der Erwartungswert E(X) und die Varianz Var(X), lassen sich in diesem Fall ermitteln als

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

[1] Sie beispielsweise Skript „Wahrscheinlichkeitsdichte“ unter der Internet-Adresse aufgabomat.de.

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} f(x)\,[x - E(X)]^2\,dx\,. \qquad (2)$$

Mit der Varianz Var(X) ist dann zugleich auch die Standardabweichung $\sqrt{\mathrm{Var}(X)}$ bekannt. Ferner kann die Wahrscheinlichkeit P, dass die Variable einen Wert in einem beliebigen Intervall [a; b] annimmt, berechnet werden als

$$P(a \le X \le b) = \int_{a}^{b} f(x)\,dx\,. \qquad (3)$$

Ein bestimmtes Integral, wie es auf der rechten Seite dieser Gleichung steht, entspricht einer Fläche unter einer Funktionskurve. Daher lässt sich auch sagen: Wahrscheinlichkeiten, mit denen Zufallsvariablen Werte in vorgegebenen Intervallen annehmen, entsprechen Flächen unter der zugehörigen Wahrscheinlichkeitsdichtefunktion (Abbildung 1).

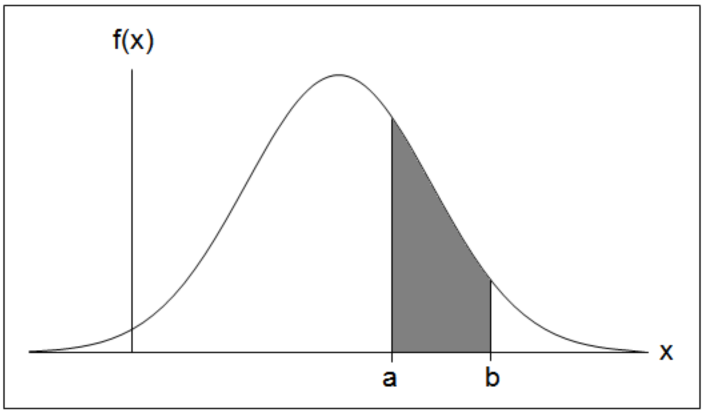

Abbildung 1: $P(a \le X \le b)$.

Einfacher ist die Berechnung der Wahrscheinlichkeit $P(a \le X \le b)$ mithilfe der zugehörigen **Verteilungsfunktion**. Die Werte $F(x_p)$ der Verteilungsfunktion sind definiert als

$$F(x_p) = \int_{-\infty}^{x_p} f(x)\,dx\,. \qquad (4)$$

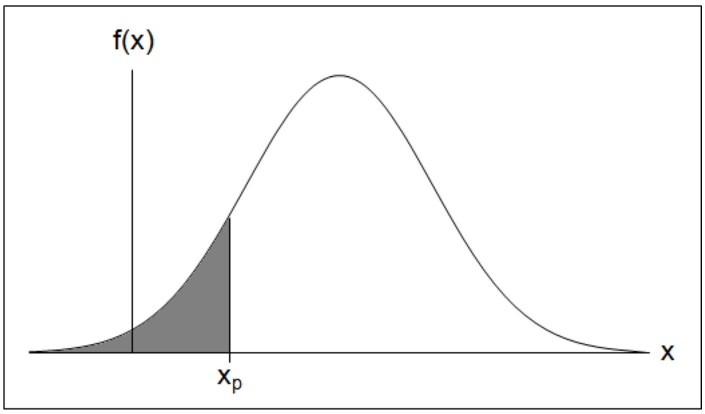

Abbildung 2: $P(X \le x_p)$.

Sie entsprechend der Wahrscheinlichkeit, dass die Variable den Wert x_p unterschreitet (Abbildung 2):

$$F(x_p) = P(X \le x_p). \tag{5}$$

Mithilfe der Verteilungsfunktion ergibt sich die Wahrscheinlichkeit $P(a \le X \le b)$ als

$$P(a \le X \le b) = F(b) - F(a). \tag{6}$$

Die Richtigkeit dieser Aussage lässt sich, wie in Abbildung 3 dargestellt, grafisch begründen.

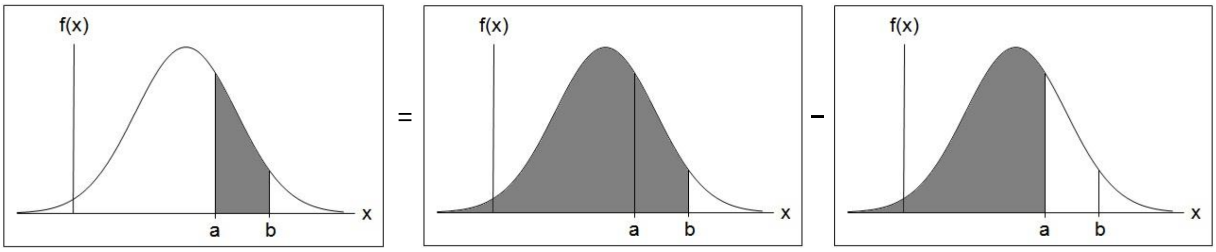

Abbildung 3: $P(a \le X \le b) = F(b) - F(a)$.

3 Die Normalverteilung

Auf viele Variablen trifft der **Zentrale Grenzwertsatz** der Wahrscheinlichkeitsrechnung. Er besagt: Setzt sich eine Zufallsvariable additiv aus einer großen Zahl beliebig verteilter, stochastisch unabhängiger Zufallsvariablen zusammen, so ist sie selbst näherungsweise normalverteilt.

Als **normalverteilt** werden solche Variablen bezeichnet, deren Wahrscheinlichkeitsdichtefunktion die Form

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{7}$$

hat. Darin steht „exp“ für die Exponentialfunktion mit der Basis e ($\exp(x) = e^x$).

Gleichung 7 enthält zwei mit μ und σ bezeichnete Größen, so genannte Parameter, welche die Form der Wahrscheinlichkeitsdichtefunktion bestimmen und damit ihre Anpassung an die jeweils betrachtete Variable erlauben:

- Der **Mittelwert** μ gibt an, wo das Maximum der Wahrscheinlichkeitsdichtefunktion liegt.
- Die **Standardabweichung** σ bestimmt die Breite der Verteilung. Sie ist gleich dem Abstand des Mittelwerts von den beiden Wendepunkten auf den Flanken der Funktionskurve (Abbildung 4).

Zwischen den statistischen Kenngröße Erwartungswert und Varianz und den beiden Parametern der Normalverteilung bestehen einfache Beziehungen. Es ist

$$E(X) = \mu \tag{8}$$

$$\sqrt{\mathrm{Var}(X)} = \sigma. \tag{9}$$

Daher die Namensgleichheit des Parameters σ und der statistischen Kenngröße Standardabweichung.

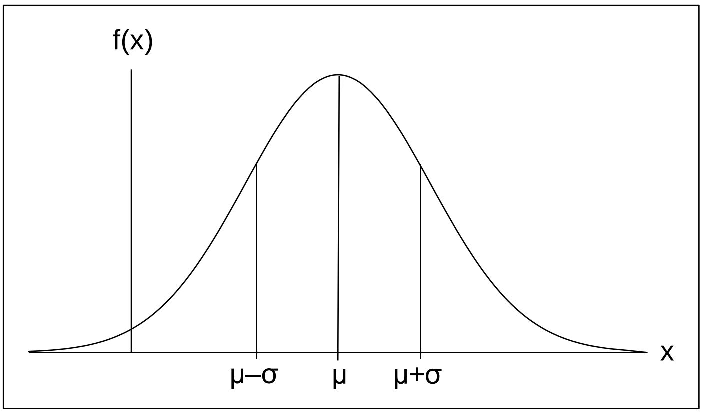

Abbildung 4: Die Wahrscheinlichkeitsdichtefunktion der Normalverteilung.

4 Standardisierung bzw. Z-Transformation

Von **Standardisierung** spricht man, wenn eine Zufallsvariable so transformiert (umgewandelt) wird, dass die resultierende Variable den Erwartungswert $E(X) = 0$ und die Varianz $Var(X) = 1$ bzw. die Standardabweichung $\sqrt{Var(X)} = 1$ hat.

Die Normalverteilung mit dem Erwartungswert bzw. Mittelwert $\mu = 0$ und der Standardabweichung $\sigma = 1$ wird **Standardnormalverteilung** genannt. Bei der Standardisierung der Normalverteilung geht es also darum, aus einer beliebig normalverteilten Variablen eine standardnormalverteilte Variable zu machen. Da eine solche Variable üblicherweise symbolisch mit Z bezeichnet wird, spricht man auch von der **Z-Transformation** der Variable.

Die Verteilungsfunktion normalverteilter Variablen ist nach den Gleichungen 4 und 7

$$F(x_p) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x_p} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \,. \tag{10}$$

Die Verteilungsfunktion speziell einer standardnormalverteilten Variable Z ist dann

$$\begin{aligned} F(z_p) &= \frac{1}{\sqrt{2\pi}\cdot 1} \int_{-\infty}^{z_p} \exp\left[-\frac{1}{2}\left(\frac{z-0}{1}\right)^2\right] dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_p} \exp\left(-\frac{1}{2} z^2\right) dz \,. \end{aligned} \tag{11}$$

Die Werte der Verteilungsfunktion nach dieser Gleichung zu berechnen, erfordert eine numerische Integration. Der Einfachheit halber bietet sich die Verwendung bereits berechneter, tabellierter Werte an. In Tabelle 1 sind Werte der Verteilungsfunktion der Standardnormalverteilung von F(0,00) bis F(2,99) aufgeführt. Das Argument der Funktion setzt sich aus der Information in der ersten Spalte und obersten Zeile der Tabelle zusammen. Suchen Sie aus der ersten Spalte die Vorkomma- und erste Nachkommastelle heraus und aus der obersten Zeile die zweite Nachkommastelle. Am Schnittpunkt der entsprechenden Zeile und Spalte lässt sich dann der zugehörige Wert der Verteilungsfunktion ablesen.

	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1,0	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
2,0	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995
2,6	0,995	0,995	0,996	0,996	0,996	0,996	0,996	0,996	0,996	0,996
2,7	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
2,8	0,997	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
2,9	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,999	0,999	0,999

Tabelle 1: Werte der Verteilungsfunktion der Standardnormalverteilung.

Beispiel: Gesucht ist F(0,12). Vorkomma- und erste Nachkommastelle finden Sie in diesem Fall in der dritten Zeile der Tabelle, dort, wo 0,1 steht. Die zweite Nachkommastelle steht in der vierten Spalte, in deren Kopf ,02 zu lesen ist. Am Schnittpunkt der dritten Zeile und vierten Spalte können Sie F(0,12) = 0,548 ablesen, d. h. eine standardnormalverteilte Variable nimmt mit der Wahrscheinlichkeit 0,548 einen Wert kleiner gleich 0,12 an.

Nun noch einmal zurück zu Gleichung 10, zur Verteilungsfunktion beliebig normalverteilter Variablen. Im Exponenten der e-Funktion steht der Ausdruck $(x - \mu)/\sigma$. Wenn von einer Zufallsvariable eine Konstante (hier: μ) subtrahiert und die Differenz durch eine Konstante (hier: σ) dividiert wird, entsteht wieder eine Zufallsvariable. Diese sei im Folgenden mit U bezeichnet:

$$U = \frac{X - \mu}{\sigma} .$$

Wie sieht die Verteilungsfunktion für die Variable U aus? Wenn in Gleichung 10 die Integrationsvariable X durch die Variable U ersetzt werden soll, so muss „dx“ ersetzt werden durch einen Ausdruck, der stattdessen „du“ enthält. In welcher Weise dies zu geschehen hat, sieht man, wenn man die Ableitung dU/dX bildet:

$$\frac{dU}{dX} = \frac{1}{\sigma} \quad \Rightarrow \quad dX = \sigma\, dU.$$

Damit ergibt sich

$$\frac{1}{\sqrt{2\pi}\,\sigma}\int \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\int \exp\left(-\frac{1}{2}u^2\right)\sigma\, du$$

und, nach Kürzen von σ, für die Verteilungsfunktion der Variable U

$$F(u_p) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_p} \exp\left(-\frac{1}{2}u^2\right)du\,.$$

Vergleichen Sie diese Gleichung mit Gleichung 11, mit der Verteilungsfunktion standardnormalverteilter Variablen. Ob eine Variable mit Z oder mit U bezeichnet wird, ist ohne Belang. Entscheidend ist die mathematische Form der Gleichungen. Und diese stimmt überein. Die Variable U = (X – μ)/σ wird genauso beschrieben wie eine standardnormalverteilte Variable, d. h. (X – μ)/σ *ist* standardnormalverteilt. Da standardnormalverteilte Variablen üblicherweise mit Z bezeichnet werden, kann zu dieser Bezeichnung zurückgekehrt und festgehalten werden:

Durch die **Z-Transformation**

$$Z = \frac{X-\mu}{\sigma} \tag{12}$$

wird aus jeder wie auch immer normalverteilten Zufallsvariable X mit dem Mittelwert μ und der Standardabweichung σ eine standardnormalverteilte Zufallsvariable.

Dies lässt sich nutzen, um allein mithilfe von Tabelle 1 Wahrscheinlichkeiten für beliebig normalverteilte Variablen zu berechnen.

Beispiel: Der Intelligenzquotient (IQ) ist eine normalverteilte Größe mit dem Mittelwert μ = 100 und der Standardabweichung σ = 15. Als hochbegabt werden Menschen bezeichnet, die einen IQ größer 130 besitzen. Wie hoch ist die Wahrscheinlichkeit, dieser Gruppe anzugehören?

Die Wahrscheinlichkeit, dass die Zufallsvariable IQ irgendeinen beliebigen Wert annimmt, ist 1, und damit auch die Wahrscheinlichkeit, dass sie einen Wert kleiner, gleich oder größer 130 annimmt:

$$P(IQ \le 130) + P(IQ > 130) = 1$$
$$P(IQ > 130) = 1 - P(IQ \le 130).$$

Auf der rechten Seite der Gleichung wird eine Z-Transformation der Variable IQ und ihres Wertes 130 vorgenommen:

$$
\begin{aligned}
P(IQ > 130) &= 1 - P\left(\frac{IQ-\mu}{\sigma} \leq \frac{130-100}{15}\right) \\
&= 1 - P(Z \leq 2{,}00).
\end{aligned}
$$

Die Wahrscheinlichkeit $P(Z \leq 2{,}00)$, dass eine standardnormalverteilte Variable den Wert 2,00 unterschreitet, ist der Wert der Verteilungsfunktion der Standardnormalverteilung an der Stelle 2,00. Dieser lässt sich aus der Tabelle 1 ablesen.

$$
\begin{aligned}
P(IQ > 130) &= 1 - F(2{,}00) \\
&= 1 - 0{,}977 \\
&= 0{,}023
\end{aligned}
$$

Rund 2,3 % der Bevölkerung gelten nach dem oben genannten Kriterium als hochbegabt.

Ist es überflüssig, sich heutzutage, da meist ein Computer für die Datenanalyse zur Verfügung steht, mit solchen Berechnungen zu befassen? Nicht überflüssig ist jedenfalls das Wissen um die Z-Transformation. Sie begegnet Ihnen beispielsweise wieder, wenn Sie sich eingehender mit der Intervallschätzung für den Erwartungswert normalverteilter Variablen beschäftigen, siehe Skript „Intervallschätzung für den Erwartungswert“, Abschnitt 2, unter aufgabomat.de.