Vorlesung Mathematik und Statistik

Sven König

Teilgebiet Statistik

Teil IV

Kontinuierliche Wahrscheinlichkeitsverteilung

Kontinuierliche Wahrscheinlichkeitsverteilung

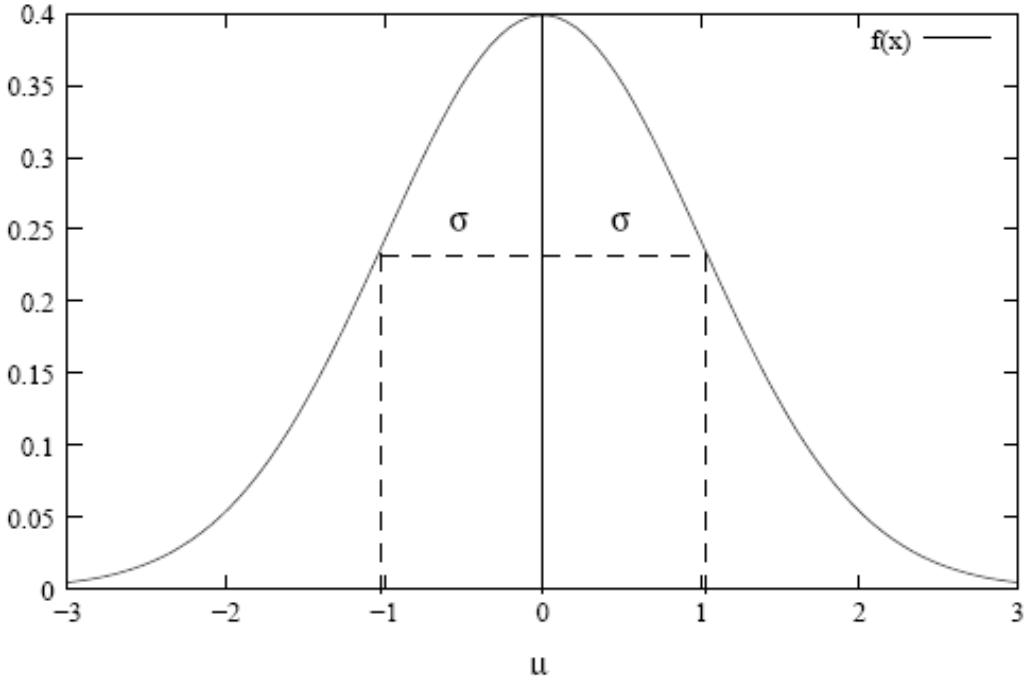
Die bisherigen Zufallsvariablen hatten alle eine **diskrete** (diskontinuierliche) Verteilung. Die Anzahl weiblicher Tiere bei ET war 0, 1, 2, oder 3; nicht aber 1.23 oder 2.41. Der Wertebereich hier ist N = Menge der natürlichen Zahlen

Die Mehrzahl aller wichtigen Zufallsvariablen kann aber alle möglichen reellen Zahlen in einem bestimmten Bereich einnehmen. Sie heißen **kontinuierliche** Variablen. Beispiele sind:

- Erntemengen, Milchmenge, Gewichte, Zunahmen, Futterverwertung,.....

Wichtigste kontinuierliche Verteilung = Normalverteilung

Mit Standardnormalverteilung bezeichnet man eine Normalverteilung mit dem Mittelwert 0 und der Varianz 1.



Eigenschaften der Standardnormalverteilung:

- **Eingipfelig**

- **Symmetrisch**
- **Glockenförmig**
- **Wertebereich von –unendlich bis +unendlich**
- **Vollständig bestimmt durch Mittelwert und Standardabweichung**
- **Mehr als 99% aller Werte liegt zwischen –3 und +3**

Wahrscheinlichkeitsfunktion (Kurvengleichung):

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

Bei kontinuierlichen Verteilungen bezeichnet man die Wahrscheinlichkeitsfunktion oft auch als *Dichte* der Verteilung. Die *Verteilungsfunktion* ist, wie wir bereits wissen, das Integral der Wahrscheinlichkeitsfunktion im Bereich von $-\infty$ bis zum Wert z.

$$F(z) = \Pr(Z \leq z) = \int_{-\infty}^{z} f(x)dx$$

Leider läßt sich dieses Integral nicht in geschlossener Form darstellen, d.h. es exisitiert keine Formel, die direkt den Wert von $F(z)$ liefert. Aus diesem Grunde sind die Flächen unter der Kurve für die Standardnormalverteilung tabelliert (s. Anhangstabelle I).

Wir haben gesehen, daß die Wahrscheinlichkeit $F(z) = \Pr(Z \leq z)$. Die Frage ist, wie hoch ist die Wahrscheinlichkeit $\Pr(Z < z)$? Es gilt generell für kontinuierliche Verteilungsfunktionen: $\Pr(Z < z) = \Pr(Z \leq z)$. Dies rührt daher, daß die Wahrscheinlichkeit, daß Z *exakt* den Wert z annimmt gleich 0 ist, wie folgende Überlegung zeigt: Die Wahrscheinlichkeit, daß eine Realisierung von Z im Intervall a, b liegt ist:

$$\Pr(a \leq z \leq b) = F(b) - F(a)$$

In der Statistik benötigt man ständig bestimmte Flächenanteile unter der Standardnormalkurve (sog. Quantile). Diese kann man aus der Tabelle im Anhang

entnehmen. Die Tabelle enthält die Werte der Verteilungsfunktion, also das Integral von $-\infty$ bis zum gesuchten Abszissenwert. Für $z = 1.28$ liefert die Tabelle z.B. den Wert .900, d.h. daß eine Realisierung einer standardnormalverteilten Zufallsvariablen mit 90-prozentiger Wahrscheinlichkeit kleiner ist als 1.28.

Wir erinnern uns, daß die Wahrscheinlichkeiten die relativen Häufigkeiten von bestimmten Werten im Stichprobenraum angeben. Man argumentiert deshalb auch oft in der Form "90% aller Werte im Stichprobenraum sind kleiner als 1.28", was identisch ist mit der Aussage "die Wahrscheinlichkeit, daß eine Realisierung z der Zufallsvariablen Z kleiner ist als 1.28 beträgt .9".

Oftmals werden auch Flächen zwischen zwei Grenzen benötigt. So interessiert zum Beispiel, wie hoch die Wahrscheinlichkeit ist, daß ein Wert in den Bereich $[-1, 1]$ fällt. Offensichtlich handelt es sich hierbei um die Wahrscheinlichkeit:

$$\Pr(-1 \leq x \leq 1) = \Pr(x \leq 1) - \Pr(x \leq -1) = F(1) - F(-1)$$

Die Tabelle liefert hierfür die Werte .841 - .159 = .682. Es ist also mit einer Wahrscheinlichkeit von 68.2% zu erwarten, daß eine Realisierung der Zufallsvariablen im Bereich von $[-1, 1]$ liegt. Einige weitere interessante Intervalle gibt die folgende Tabelle:

Flächenanteile für einige Bereiche der Standardnormalverteilung ($\mu = 0, \sigma = 1$)

von	bis	$\Pr(von \leq x \leq bis)$
-1.000	1.000	.682
-1.645	1.645	.900
-1.960	1.960	.950
-2.000	2.000	.954
-2.600	2.600	.990
-3.000	3.000	.997

Aufgrund der symmetrischen Natur der Normalverteilung werden die Werte für negative z nicht tabelliert. Es gilt hier die einfache Beziehung:

$$F(-z) = 1 - F(z)$$

Standardisierung beliebiger Normalverteilungen

So nützlich die Standardnormalverteilung auch ist, praktische Merkmale sind selten so verteilt, daß der Mittelwert 0 und die Standardabweichung 1 ist ($N(0,1)$). Eine Tabellierung der Normalverteilungen für andere Werte als $N(0,1)$ wäre ein mühsames Unterfangen. Glücklicherweise kann man jede normalverteilte Variable in eine Standardnormalvariable transformieren.

Nehmen wir an, die niedersächsischen Herdbuchkühe hätten eine mittlere Leistung von 6000 kg und eine Standardabweichung von 1000 kg ($N(6000,1000)$). Mit welcher Wahrscheinlichkeit hat eine zufällig gezogene Kuh dann eine Leistung von mehr als 8000 kg? Der Anteil Kühe mit einer Leistung über 8000 kg ist gleich 1 minus dem Anteil Kühe mit weniger als 8000 kg:

$$\Pr(X > 8000) = 1 - \Pr(X \leq 8000)$$

Wir führen bei dieser Gelegenheit eine neue Schreibweise ein, die der Klarheit beim Argumentieren mit verschiedenen Wahrscheinlichkeits- oder Verteilungsfunktionen dient. Um klarzumachen, daß unser Wert sich auf eine Verteilung mit den Parametern $\mu = 6000$ und $\sigma = 1000$ bezieht schreiben wir:

$$\Pr(X \leq 8000 \mid 6000, 1000)$$

sprich "Wahrscheinlichkeit X kleiner gleich 8000, gegeben μ gleich 6000 und σ gleich 1000". Der Wert 8000 liegt 2000 kg über μ. In Einheiten der Standardabweichung ausgedrückt erhalten wir:

$$\frac{2000}{1000} = 2$$

Also liegt unser Grenzwert genau zwei Standardabweichungen über dem Mittelwert. Nun gilt aber ganz allgemein:

> Eine Normalverteilung ist durch ihren Mittelwert und ihre Standardabweichung vollständig bestimmt. Das bedeutet, daß bei jeder Normalverteilung die Fläche unter der Kurve zwischen $-\infty$ und 1σ ($2\sigma, 3\sigma$ usw.) gleich ist.

Daher müssen wir nun lediglich in der Normalverteilungstabelle nachsehen, wie hoch die Wahrscheinlichkeit ist, daß ein Wert aus einer Standardnormalverteilung weniger als 2 Standardabweichungen über dem Mittel liegt. Die Tabelle ergibt den Wert .977. Unsere gesuchte Wahrscheinlichkeit ist also: $\Pr(X > 8000 \mid 6000, 1000) = 1 - .977 = .023$. In Worten: Die Wahrscheinlichkeit, daß eine zufällig aus der niedersächsischen Herdbuchpopulation entnommene Kuh eine Leistung über 8000 kg aufweist, beträgt nur 2.3%.

Was haben wir gemacht? Wir haben im mathematischen Sinne eine Transformation unserer Zufallsvariablen X in eine Standardnormalvariable Z vorgenommen. Hierzu wurde zunächst der Mittelwert von X von x abgezogen und anschließend durch die Standardabweichung von X dividiert:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

Dieser Vorgang wird als z-Transformation oder oft auch als Standardisierung bezeichnet. Man kann dies auch in Form der Verteilungsfunktionen darstellen:

$$F(z \mid 0, 1) = F\left(\frac{x - \mu}{\sigma} \mid \mu, \sigma\right)$$

Schätzen von Mittelwerten

Die Grundgesamtheit

Wie wir bereits wissen, ist die Grundgesamtheit die Gruppe von Objekten (Betriebe, Tiere usw.), aus denen die Stichprobe gezogen wird. Die Grundgesamtheit kann sehr unterschiedliche Größe haben. Bei Wahlprognosen besteht die Grundgesamtheit aus allen deutschen Wahlberechtigten, bei einer Untersuchung über den Schlachtkörperwert der Tiere in der Genreserve des Bunten Bentheimer Schweines dagegen aus einigen hundert Tieren insgesamt. Die Grundgesamtheit stellt auch die Grenze der Aussagefähigkeit der Stichprobe dar. Alle Aussagen aus der Stichprobe beziehen sich auf die Grundgesamtheit und auf nichts anderes (eine Meinungsumfrage, die ausschließlich in Altersheimen durchgeführt wurde, gibt die Meinung von Altersheiminsassen wieder und nicht die der gesamten Bevölkerung).

Wir beginnen mit einer handlichen Grundgesamtheit, der Körpergröße von 100 Studenten, die in einem bestimmten Semester das Landwirtschaftsstudium aufgenommen haben. Der Einfachheit halber haben wir das kontinuierliche Merkmal Körpergröße in ein diskretes Merkmal verwandelt, das folgende Verteilung aufweist:

Größe (X)	Frequenz	$p(x)$	$x \cdot p(x)$	$(x-\mu)^2 \cdot p(x)$
165	1	.01	1.65	2.25
170	6	.06	10.20	6.00
175	24	.24	42.00	6.00
180	38	.38	68.40	0.00
185	24	.24	44.40	6.00
190	6	.06	11.40	6.00
195	1	.01	1.95	2.25
	N=100	1.00	$\mu = 180.00$	$\sigma^2 = 28.5$
				$\sigma = 5.34$

Wenn wir nun einen Studenten aus der Population zufällig herausziehen, wie hoch ist die Wahrscheinlichkeit, daß er 1.75m groß ist? Offensichtlich beträgt diese

Wahrscheinlichkeit 24%. Und die Wahrscheinlichkeit, daß er 1.90m groß ist beträgt 6%. Für eine einzelne Beobachtung ist also die Wahrscheinlichkeit eines bestimmten Wertes gleich der Häufigkeit dieses Wertes in der Grundgesamtheit. Wenn wir den Studenten "wieder zurücklegen" und ein weiteres Mal ziehen, dann gelten für die neue Beobachtung dieselben Wahrscheinlichkeiten. Dies ist eine wichtige Erkenntnis:

> Jede einzelne Beobachtung in einer Stichprobe besitzt die Wahrscheinlichkeitsverteilung der Grundgesamtheit. Mit anderen Worten: Das Populationsmittel (μ) und die Populationsvarianz[1] (σ^2) sind die Parameter von *einzelnen Beobachtungen*.

Die Stichprobe und ihre Parameter

Wir ziehen aus der oben beschriebenen Grundgesamtheit eine Stichprobe der Größe $n = 5$. Dabei erhalten wir die folgenden Werte:

Beob.	Wert
1	170
2	195
3	175
4	180
5	170

Uns interessiert zunächst der Mittelwert der Stichprobe. Bisher haben wir nur Populationsmittelwerte berechnet, aber die Berechnung des Stichprobenmittelwerts erfordert keine neuen Prinzipien. Da Stichproben in der Regel relativ klein sind, macht man sich nicht die Mühe, die Häufigkeitsverteilung zu berechnen. Vielmehr summiert man einfach alle Beobachtungen auf und dividiert durch die Anzahl der Beobachtungen n. Bei dieser Vorgehensweise erhält jede Beobachtung das gleiche Gewicht, da man ja a priori die Parameter der Population nicht kennt und daher nicht weiß, das z.B. der Wert 180 sehr viel wahrscheinlicher ist als der Wert 195. Somit erhalten wir:

$$\text{Stichprobenmittel} = \bar{x} = (170 + 195 + 175 + 180 + 170)/5 = 178$$

Zur Erinnerung halten wir die allgemeine Formel für den Stichprobenmittelwer fest:

$$\bar{x} = \frac{1}{n}[x_1 + x_2 + \ldots + x_n]$$

Um auch in Formeln immer eindeutig zwischen dem Populationsmittel und den Stichprobenmittel unterscheiden zu können, verwenden wir für das Stichproben mittel die Bezeichnung $\bar{x}$. Durch das zufällige Ziehen der Stichprobe ist es wahr scheinlich, daß die ZS für die zugrundeliegende Grundgesamtheit repräsentati ist. Wir beobachten, daß der Stichprobenmittelwert nicht sehr weit vom wahre Mittel entfernt liegt. Der eine sehr große Wert (195), der in unserer Stichprob enthalten ist wird durch mehrere mittlere bis kleine Werte wieder ausgeglichen Wir erwarten, daß dies in der Mehrzahl der Fälle so sein wird.

Aufgrund der Durchschnittsbildung hat das Stichprobenmittel $\bar{x}$ einen weniger extremen Wert (es variiert weniger) als die einzelnen Beobachtungen.

$$\text{Stichprobenvarianz} = s^2 = \frac{1}{(n-1)}\sum(x - \bar{x})^2$$

Standardfehler des Stichprobenmittels

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

Die Standardabweichung des Stichprobenmittelwertes wird gemeinhin als *Standardfehler* bezeichnet [1] und gibt an, wie stark die einzelnen Stichprobenmittelwerte um ihren Erwartungswert (das Populationsmittel) schwanken. Beachten Sie, daß in der Formel zur Berechnung des SE die *Populationsvarianz* auftaucht, diese muß strenggenommen zur Berechnung des Standardfehlers bekannt sein.

Die letzte Formel gibt uns auch Aufschluß über die Bedeutung der Stichprobengröße für die Güte einer Stichprobe. Je größer die Stichprobe wird, umso geringer werden die Abweichungen der beobachteten Stichprobenmittelwerte vom Populationsmittelwert, d.h. desto geringer wird die Wahrscheinlichkeit, ein Stichprobenmittel zu erhalten, das "völlig daneben" liegt. Um diese Wahrscheinlichkeit zu kennen, müssen wir aber die *Form* der Wahrscheinlichkeitsverteilung der Stichprobenmittelwerte (uniform, Binomial, Normal) kennen. Dies ist das Thema des nächsten Abschnitts.

In *Zufallsstichproben* der Größe n streuen die Stichprobenmittelwerte ($\overline{x}$) um das wahre Populationsmittel (μ) mit einem Standardfehler von $\sigma/\sqrt{n}$, wobei σ die Standardabweichung der Grundgesamtheit ist.
Je größer n wird, desto mehr konzentrieren sich die Stichprobenmittelwerte um das Populationsmittel. Dabei nähert sich die Verteilung der Stichprobenmittelwerte immer mehr einer Normalverteilung an.

Ein Beispiel mag die Anwendung dieser Zusammenhänge verdeutlichen. Die Wahrscheinlichkeit, daß ein zufällig aus der Grundgesamtheit von 100 Studenten gezogener Student größer als 185 cm ist, beträgt .07 (s.o.). Wie hoch ist dagegen die Wahrscheinlichkeit, daß der Mittelwert einer ZS mit der Größe $n = 9$ größer als 185 ist?

Die Verteilung der Grundgesamtheit ist eine angenäherte Normalverteilung (symmetrisch, eingipfelig, wenn auch diskret). Bei einer Stichprobengröße von 9 können wir erwarten, daß die Stichprobenmittel tatsächlich normalverteilt sind. Der Erwartungswert ist das Populationsmittel ($\mu = 180$). Die Populationsstandardabweichung ist gleich 5.34. Der Standardfehler ist demnach $5.34/\sqrt{9} = 1.78$. Gesucht ist die Wahrscheinlichkeit $\Pr(\overline{x} > 185) = 1 - \Pr(\overline{x} < 185)$. Zur Bestimmung dieser Wahrscheinlichkeit müssen wir eine z–Transformation durchführen:

$$z = \frac{\overline{x} - \mu}{SE} = \frac{185 - 180}{1.78} = 2.808$$

Die Normalverteilungstabelle liefert uns für $z = 2.81$ einen Wert von .998. Damit ergibt sich eine Wahrscheinlichkeit von .002 für einen Stichprobenmittelwert größer als 185. Nur einer von 500 Stichprobenmittelwerten wird also über 185 cm liegen.