

---

## 2.1 Motivation

Dieses Kapitel gibt eine kurze Übersicht über die im weiteren Verlauf des Buches benötigten mathematischen und statistischen Grundlagen. Zu Beginn wird der Begriff der Wahrscheinlichkeitsverteilung oder auch Wahrscheinlichkeitsdichte eingeführt und die wichtigsten Eigenschaften werden definiert, sowie einige wichtige Verteilungen vorgestellt. Im weiteren Verlauf wird erläutert, was eine Korrelation zwischen verschiedenen Größen bedeutet und wie sie mathematisch behandelt werden kann. Danach folgt eine kurze Einführung in die Bayessche Statistik und eine Übersicht über den Umgang mit fehlerbehafteten Größen.

Es sei hervorgehoben, dass diese Einführung nur eine kurze Übersicht bieten kann, für eine tiefergehende Behandlung sei auf die entsprechende Fachliteratur, z. B. [BL98] verwiesen.

---

## 2.2 Wahrscheinlichkeitsverteilung

### 2.2.1 Definition der Wahrscheinlichkeitsdichte

Verschiedene Größen können aufgrund statistischer Prozesse verschiedene Werte annehmen. Diese Größen werden Zufallsvariablen genannt. Man unterscheidet diskrete Zufallsvariablen (z. B. Münzwurf, Roulette) und kontinuierliche Zufalls-

variablen. Letztere sind typischerweise das Ergebnis einer Messung oder eines Experiments.

Die Größe  $f(x)$  einer kontinuierlichen Zufallsvariablen  $x$  wird als Wahrscheinlichkeitsdichte (engl. probability density function, PDF) bezeichnet. Sie ist positiv semidefinit und auf 1 normiert:

$$f(x) \geq 0 \quad \forall x \quad \text{und} \quad \int_{-\infty}^{\infty} f(x) dx = 1 \quad (2.1)$$

Die Wahrscheinlichkeitsdichte selbst ist *keine* Wahrscheinlichkeit, sondern die Größe  $f(x)\Delta x$  geht gegen die Wahrscheinlichkeit, dass die Zufallsvariable  $x$  zwischen  $x$  und  $x + \Delta x$  liegt für  $\Delta x \rightarrow 0$ . Daraus folgt auch, dass die Wahrscheinlichkeit gleich null ist, dass  $x$  (bei kontinuierlichem  $f(x)$ ) einen präzisen vorher festgelegten Wert (d. h.  $\Delta x = 0$ ) hat.

Die Größe

$$F(x) = \int_{-\infty}^x f(x') dx' \quad (2.2)$$

mit  $F(-\infty) = 0$  und  $F(+\infty) = 1$  wird kumulative *Verteilungsfunktion* oder Wahrscheinlichkeitsverteilung (engl. cumulative distribution function - CDF) genannt. Dies ist in Abb. 2.1 am Beispiel der Gaußschen Normalverteilung illustriert.

### 2.2.2 Erwartungswerte

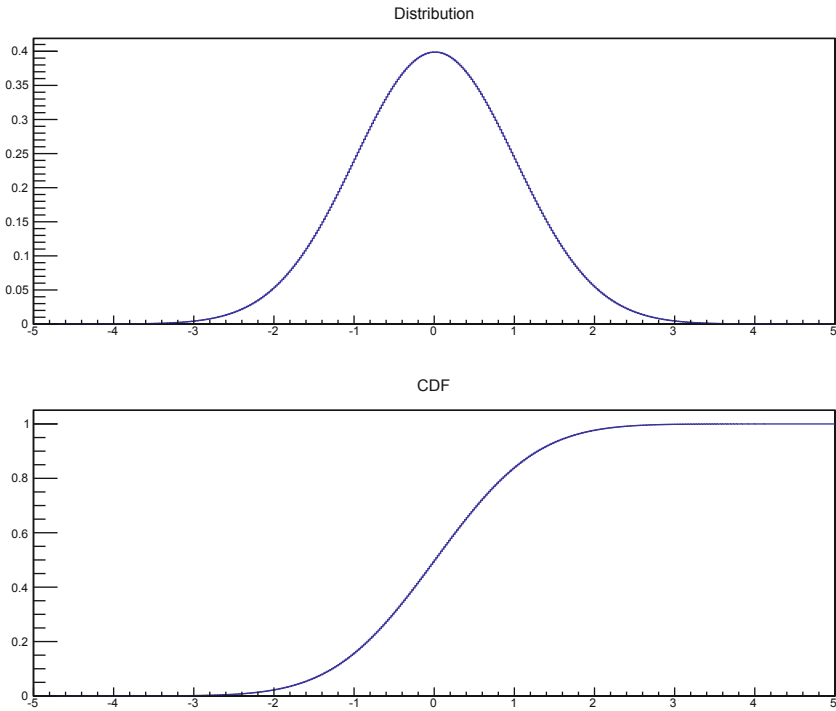
Eine wichtige Größe, um Wahrscheinlichkeitsverteilungen zu charakterisieren, ist der Erwartungswert bezüglich einer Funktion  $g(x)$ , der definiert wird als:

$$E[g] = \int_{-\infty}^{\infty} g(x) f(x) dx \quad (2.3)$$

Die Erwartungswerte von  $g(x) = x^n$  werden  $n$ -te algebraische Momente  $\mu_n$ , die von  $g(x) = (x - \langle x \rangle)^n$   $n$ -te zentrale Momente  $\mu'_n$  genannt. Eine Wahrscheinlichkeitsdichte ist eindeutig definiert durch alle ihre Momente. Ein Beweis ist in [BL98] zu finden.

Der *Mittelwert* als Erwartungswert der Funktion  $g(x) = x$  ist ein Spezialfall davon:

$$\langle x \rangle = E[x] = \int_{-\infty}^{\infty} x f(x) dx \quad (2.4)$$



**Abb. 2.1** Gaußverteilung und kumulative Wahrscheinlichkeitsverteilung

Das zweite zentrale Moment  $\mu'_2$  ist ein Maß für die Breite einer Wahrscheinlichkeitsdichte und wird als *Varianz* bezeichnet:

$$V[x] = E[(x - \langle x \rangle)^2] = \int_{-\infty}^{\infty} (x - \langle x \rangle)^2 f(x) dx \quad (2.5)$$

Es gilt:

$$V[x] = E[x^2] - \langle x \rangle^2$$

$$V[ax] = a^2 V[x]$$

Die Varianz ist positiv und die *Standardabweichung* wird definiert als

$$\sigma = \sqrt{V[x]} \quad \text{oder} \quad V[x] = \sigma^2 \quad (2.6)$$

### 2.2.3 Quantile

Quantile unterteilen die Wahrscheinlichkeitsdichte in zwei Bereiche und werden durch eine Zahl  $p$  mit  $0 < p < 1$  charakterisiert. Links vom Quantil  $Q_p$  liegen  $p * 100\%$  aller Werte, rechts davon  $(1 - p) * 100\%$ . Sie sind also definiert als

$$\int_{-\infty}^{Q_p} f(x)dx = p \quad (2.7)$$

Ein wichtiges Quantil ist der *Median*, der die Wahrscheinlichkeitsverteilung in zwei gleiche Teile unterteilt:

$$Q_{0,5} := \int_{-\infty}^{x_{0,5}} f(x)dx = 0,5$$

Gegenüber dem Mittelwert hat der Median den Vorteil, dass er robust ist und nicht so sensitiv auf die Ausläufer der Wahrscheinlichkeitsdichte reagiert wie der Mittelwert.

Darüber hinaus werden die Quantile  $Q_{0,84}$  und  $Q_{0,16}$  gerne betrachtet, da sie bei der Gaußverteilung den Positionen entsprechen, die vom Mittelwert eine Standardabweichung entfernt sind:  $\mu \pm \sigma$ , sowie die Quantile  $Q_{0,25}$  und  $Q_{0,75}$ .

### 2.2.4 Konfidenzintervall

Das Konfidenzintervall gibt an, ob die Daten mit einer bestimmten Hypothese verträglich sind. Meist wird nur die Größe der Abweichung betrachtet, so lässt sich das beidseitige Konfidenzintervall definieren als:

$$P(t_- \leq t \leq t_+) = \int_{t_-}^{t_+} f(x)dx \quad (2.8)$$

wobei  $f(x)$  die Wahrscheinlichkeitsdichte der Variable  $x$  ist. Die folgenden Werte werden oft verwendet:

$$\begin{aligned} P &= 68\% \quad (1\sigma) \\ &= 95\% \quad (1,96\sigma) \quad \text{oder} \quad 95,4\% \quad (2\sigma) \\ &= 99\% \quad (2,58\sigma) \end{aligned}$$

wobei die Werte in Klammern für eine Gaußsche Wahrscheinlichkeitsverteilung und ein beidseitig symmetrisches Konfidenzintervall gelten. In der Experimentalphysik wird verlangt, dass ein neuer Effekt mindestens eine statistische Signifikanz von  $3\sigma$  aufweist, um von einer Evidenz für diesen neuen Effekt sprechen zu können. Von einer Entdeckung spricht man erst, wenn eine Signifikanz von  $5\sigma$  erreicht ist. In anderen wissenschaftlichen Zweigen ist es hingegen üblich, bereits ab einem Konfidenzintervall von 95 %, also  $1,96\sigma$ , von einer gesicherten Erkenntnis zu sprechen. Dies bedeutet im Umkehrschluss, dass hier jede zwanzigste „Entdeckung“ nur auf zufälligen Schwankungen beruht, also keinen wissenschaftlichen Hintergrund hat und entsprechend auch keinen Erkenntnisgewinn liefern kann, siehe auch Kap. 2.4.

### 2.2.5 Wichtige Wahrscheinlichkeitsdichten

In diesem Abschnitt werden einige wichtige Wahrscheinlichkeitsverteilungen vorgestellt, die in praktischen Anwendungen eingesetzt werden. Hierbei ist zwischen diskreten und kontinuierlichen Wahrscheinlichkeitsverteilungen zu unterscheiden. Bei diskreten Verteilungen kann die Zufallsvariable nur bestimmte (diskrete) Werte annehmen, bei kontinuierlichen Verteilungen ist jeder Wert des Zahlenraums möglich.

#### Binomial-Verteilung

Bei der Binomial-Verteilung handelt es sich um eine diskrete Wahrscheinlichkeitsverteilung. Falls die Wahrscheinlichkeit für das Auftreten eines Ereignisses  $p$  ( $p \in [0, 1]$ ) ist, dann ist die Wahrscheinlichkeit, dass es bei  $n$  Versuchen *genau*  $r$  mal auftritt

$$P(r|n, p) = \binom{n}{r} p^r (1-p)^{n-r} \quad r = 0, 1, 2, \dots, n \quad (2.9)$$

Für Mittelwert und Varianz gilt:

$$\langle r \rangle = \sum_{r=0}^n r P(r) = np \quad (2.10)$$

$$V[r] = \sigma^2 = (r - \langle r \rangle)^2 P(r) = np(1-p) \quad (2.11)$$

Oft wird  $P(r|n, p)$  mit  $P(p|n, r)$  verwechselt, also die Wahrscheinlichkeit für die Häufigkeit des Auftretens des Ereignisses in  $n$  Versuchen mit der Wahrscheinlichkeit, dass das Ereignis auftritt ( $p$ ).

Darüber hinaus ist zu sehen, dass aus der Wahrscheinlichkeit, das Ereignis nicht (also genau Null mal) zu beobachten nicht folgt, dass die Wahrscheinlichkeit für das Ereignis Null ist ( $p > 0$ ) oder die Varianz Null ist ( $\sigma^2 > 0$ ).

### Poisson-Verteilung

Eine wichtige Verteilung ist die ebenfalls diskrete *Poisson-Verteilung*, die definiert ist als:

$$P(r|\mu) = \frac{\mu^r e^{-\mu}}{r!} \quad (2.12)$$

Sie gibt die Wahrscheinlichkeit an, bei  $n$  Versuchen im Grenzfalle  $n \rightarrow \infty, p \rightarrow 0$  mit  $n \cdot p = \mu$  genau  $r$  Ereignisse zu erhalten. Diese Verteilung hat nur einen freien Parameter, ihren Mittelwert  $\mu$ , und bildet diskrete Ereignisse ab. Sie tritt oft in Fällen auf, in denen Dinge oder Ereignisse gezählt werden. Sie ist mit der Exponentialverteilung verknüpft (siehe Abschn. 2.2.6) und beschreibt die Verteilung zufälliger Ereignisse (z. B. von Blitzen bei einem Gewitter, radioaktive Zerfälle in einem bestimmten Zeitintervall, ...). Die Varianz von  $r$  beträgt ebenfalls  $\mu$ . Das bedeutet, dass bei Poisson-Prozessen die Standardabweichung  $\sigma = \sqrt{\mu}$  mit der Wurzel des Erwartungswertes ansteigt, und der relative Fehler  $\sigma/\mu$  mit  $\frac{1}{\sqrt{\mu}}$  abfällt. Dieses Verhalten ist typisch für alle statistischen Prozesse.

### Gamma-Funktion

Die Gamma-Funktion  $\Gamma(x)$  ist definiert durch die Erweiterung der Fakultät auf nicht ganzzahlige Werte:

$$\Gamma(x+1) = x! \quad (2.13)$$

wobei

$$x! = \int_0^\infty y^x e^{-y} dy$$

Es gilt:  $\Gamma(x+1) = x\Gamma(x)$ .

### Gamma-Poisson-Verteilung

Die ebenfalls diskrete *Gamma-Poisson-Verteilung* geht zurück auf [GY20] und hat ähnliche Eigenschaften wie die Poisson-Verteilung, ist aber durch einen weiteren Parameter charakterisiert. Daher kann sie empirisch dann eingesetzt werden, wenn diskrete Ereignisse abgebildet werden sollen, die Daten aber nur durch eine breitere Verteilung erfolgreich beschrieben werden können. Dies gilt insbesondere dann, wenn der Mittelwert  $\mu$ , der die Poisson-Verteilung als einzigen Parameter charakterisiert, selbst Schwankungen unterworfen ist,

bzw. als Zufallsvariable interpretiert werden muss. Wird angenommen, dass der Mittelwert  $\mu$  einer Gamma-Verteilung mit Formparameter  $r$  und Rate  $p/(1-p)$  folgt, so ergibt sich für die Gamma-Poisson-Verteilung:

$$P(k|r, p) = \int_0^\infty f_{\text{Poisson}(\mu)}(k) \cdot f_{\Gamma(r, \frac{p}{1-p})}(\mu) d\mu \quad (2.14)$$

$$= \frac{\Gamma(r+k)}{k! \Gamma(r)} p^k (1-p)^r \quad (2.15)$$

### Gaußverteilung

Die kontinuierliche Gaußverteilung (auch Gaußsche Normalverteilung oder Normalverteilung) ist die wichtigste Verteilung und ist definiert als

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.16)$$

Sie wird durch 2 Parameter charakterisiert,  $\mu$  und  $\sigma$ , wobei

$$\mu = E[x]$$

$$\sigma = \sqrt{V[x]}$$

Die Gaußverteilung kann hergeleitet werden als Grenzfall der Binomialverteilung für große Werte von  $n$  und  $r$ , sowie als Grenzfall der Poisson-Verteilung für große Werte von  $\mu$ .

Es gilt:

$$|x - \mu| \geq 1\sigma \quad x \text{ außerhalb von } \pm 1\sigma : 31,74\%$$

$$|x - \mu| \geq 2\sigma \quad x \text{ außerhalb von } \pm 2\sigma : 4,55\%$$

$$|x - \mu| \geq 3\sigma \quad x \text{ außerhalb von } \pm 3\sigma : 0,27\%$$

Das bedeutet also, dass im Mittel ca. 32 % der Fälle *außerhalb* des  $1\sigma$  Bereichs liegen müssen.

Ihre besondere Bedeutung stammt aus dem *zentralen Grenzwertsatz*: Die Wahrscheinlichkeitsdichte der Summe  $\sum_{i=1}^n x_i$  einer Stichprobe aus  $n$  unabhängigen Zufallsvariablen  $x_i$  mit einer beliebigen Wahrscheinlichkeitsdichte mit Mittelwert  $\bar{x}$  und Varianz  $\sigma^2$  geht in der Grenze  $n \rightarrow \infty$  gegen eine Gaußverteilung mit Mittelwert  $\bar{y} = n\bar{x}$  und Varianz  $V[y] = n\sigma^2$ . Diese Eigenschaft kann auch dazu benutzt werden, aus einer beliebigen normierten Verteilung eine Gaußverteilung zu erzeugen.

### Log-Normalverteilung

Die kontinuierliche Log-Normalverteilung ist mit der Gaußverteilung verwandt und wird ebenfalls durch zwei Parameter,  $\mu$  und  $\sigma$ , charakterisiert. Sie ist definiert als

$$P(x|\mu, \sigma) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \quad x > 0 \quad (2.17)$$

Es gilt:

$$E[x] = e^{\mu + \frac{1}{2}\sigma^2}$$

$$V[x] = (e^{\sigma^2} - 1) (E[x])^2$$

Diese Verteilung hat besondere Bedeutung in vielen Prozessen, z. B. in der Biologie und Wirtschaft. Sie liegt unter anderem der Black-Scholes Formel [BS73] zur Modellierung der Preise (europäischer) Optionen am Finanzmarkt zugrunde.

### Weibull-Verteilung

Die Wahrscheinlichkeitsdichte der kontinuierlichen Weibull-Verteilung [Wei39] ist definiert als:

$$P(x|\lambda, k) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2.18)$$

wobei  $k > 0$  der Formparameter (auch Weibull-Modul genannt) und  $\lambda > 0$  der Skalenparameter ist. Es gilt:

$$E[x] = \lambda \Gamma\left(1 + \frac{1}{k}\right)$$

$$V[x] = \lambda^2 \left[ \Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2 \right]$$

$$Q_{0.5} = \lambda (\ln(2))^{1/k}$$

Die Weibull-Verteilung ist mit verschiedenen anderen Verteilungen verwandt, für  $k = 1$  erhält man die Exponentialverteilung, für  $k = 2$  die Rayleigh-Verteilung. Diese Verteilung entstammt ursprünglich aus den Materialwissenschaften und kann beispielsweise zur Beschreibung von Lebensdauern von Bauelementen oder Werkstoffen verwendet werden, beschreibt aber auch die Ausfallrate oder Lebensdauer allgemeiner Systeme. Der Formparameter  $k$  charakterisiert den Prozess:



- $k < 1$  : Die Ausfallrate nimmt mit der Zeit ab, d. h. die meisten Komponenten fallen zu einem frühen Zeitpunkt aus.
- $k = 1$  : In diesem Fall geht die Weibull-Verteilung in die Exponentialverteilung über. Die Ausfallrate ist dabei zeitlich konstant und durch den Zufall bestimmt.
- $k > 1$  : Die Ausfallrate nimmt mit der Zeit zu, z. B. wenn ein Ermüdungsprozess das System beeinflusst.

Wie oben erwähnt, ist die Exponentialverteilung mit der Poisson-Verteilung assoziiert und beschreibt also das Auftreten zufällig verteilter Ereignisse. Da für  $k = 1$  die Weibull-Verteilung in die Exponentialverteilung übergeht und sie so quasi als „generalisierte Exponentialverteilung“ betrachtet werden kann, erlaubt eine darauf aufbauende Betrachtung von diskreten Ereignissen eine theoretisch fundiertere Analyse Poisson-artiger Prozesse, als beispielsweise *ad-hoc* eine Gamma-Poisson-Verteilung anzunehmen (siehe z. B. [MABF08]).

### 2.2.6 Beziehung zwischen einem Zählexperiment und der Zeit zwischen Ereignissen

Bei vielen Fragestellungen wird die Häufigkeit des Auftretens diskreter Ereignisse betrachtet. Beispiele hierfür sind die Anzahl der Blitze bei einem Gewitter, der Verkauf eines bestimmten Artikels, der Ausfall von technischen Anlagen, Werkstoffen oder Geräten, etc. Die gleichen Ereignisse lassen sich auch unter einem anderen Gesichtspunkt betrachten: Wieviel Zeit vergeht zwischen dem Eintreffen von zwei Ereignissen?

Beide Betrachtungsweisen sind miteinander verknüpft, da sie die gleichen Ereignisse beschreiben, wenn auch auf eine andere Art. Das bedeutet auch, dass die entsprechenden Wahrscheinlichkeitsverteilungen bei der jeweiligen Betrachtungsweise miteinander verknüpft sind: Die Zeit, die zwischen zwei Ereignissen vergeht, steht wie folgt mit der Verteilung der Anzahl der Ereignisse in Beziehung: Es sei  $Y_n$  die Zeit, die von Beginn einer Messung bis zum Auftreten des  $n$ -ten Ereignisses vergeht und  $X(t)$  die Anzahl der Ereignisse, die bis zur Zeit  $t$  aufgetreten sind. Dann gilt für die Beziehung zwischen der Zeit, die zwischen den Ereignissen vergeht und der Anzahl:

$$Y_n \leq t \Leftrightarrow X(t) \geq n \quad (2.19)$$

Oder: Die bis zum  $n$ -ten Ereignis vergangene Zeit ist kleiner oder gleich  $t$  dann und nur dann, wenn die Anzahl der eingetretenen Ereignisse größer oder gleich

$n$  ist. Für die Verteilung der Anzahl, bezeichnet als  $C_n(t)$ , gilt dann:

$$\begin{aligned} C_n(t) &= P(X(t) = n) \\ &= P(X(t) \geq n) - P(X(t) \geq n+1) \\ &= P(Y_n \leq t) - P(Y_{n+1} \leq t) \end{aligned}$$

Sei  $F_n(t)$  die kumulierte Wahrscheinlichkeitsverteilung (CDF) von  $Y_n(t)$ , dann lässt sich das schreiben als

$$C_n(t) = F_n(t) - F_{n+1}(t) \quad (2.20)$$

Für weitere Details siehe z. B. [MABF08, KS96].

Für die Beziehung zwischen Exponentialverteilung und Poisson-Verteilung kann man dies wie folgt sehen: Bei einem Poisson-Prozess treten Ereignisse mit einer mittleren Rate  $\lambda$  auf. Sei  $L$  die Zeit, die bis zu einem (ersten) Ereignis vergeht. Dann gilt für die Wahrscheinlichkeit, dass die Zeit bis zum Auftreten des Ereignisses länger als die Zeit  $t$  ist:

$$P(L > t) = P(\text{Kein Ereignis in Zeit } t) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t} \quad (2.21)$$

also

$$P(L \leq t) = 1 - e^{-\lambda t} \quad (2.22)$$

Dies ist die kumulative Wahrscheinlichkeitsverteilung, die zugehörige Dichte ergibt sich dann durch die Ableitung

$$f(x) = \lambda e^{-\lambda t} \quad \text{für } t > 0, \quad (2.23)$$

also die Exponentialverteilung.

## 2.3 Korrelierte Variablen

### Korrelationskoeffizient

Bei statistischen Testreihen trifft man häufig auf die Situation, dass zwei oder mehr Größen stark oder weniger stark miteinander zusammenhängen. Zunächst

Prognosen bewerten

Statistische Grundlagen und praktische Tipps

Feindt, M.; Kerzel, U.

2015, XII, 86 S., Softcover

ISBN: 978-3-662-44682-9