

2.5 Normalverteilung

Die in der Statistik am häufigsten benutzte Verteilung ist die Gauss- oder Normalverteilung. Wir haben bereits gesehen, dass diese Verteilung aus den Binomial- und Poisson-Verteilungen im Grenzfall großer Zahlen (n bzw. λ) folgt. Wir werden weiter unten den ‘zentralen Grenzwertsatz’ besprechen, der solche Grenzübergänge noch allgemeiner behandelt.

Eine Normalverteilung ergibt sich, wenn viele kleine Änderungen ϵ_i aufsummiert werden. Anschaulich kann man sich das zum Beispiel anhand des Galton-Brettes (Abb. 2.1) klar machen: Die Kugel entscheidet n -mal, ob Sie links oder rechts um einen Nagel fällt entsprechen einem Versatz um $\epsilon_i = \pm\Delta\epsilon$. Die Verteilung der Auftrefforte unter dem Brett $x = \sum_{i=1}^n \epsilon_i$ nähert sich einer Normalverteilung im Grenzfall großer n .

Die Normalverteilung $N(\mu, \sigma)$ ist durch die beiden Parameter Mittelwert μ und Varianz σ^2 gegeben:

$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (2.40)$$

Normierung: Die Normierung wird durch den Faktor $(\sqrt{2\pi}\sigma)^{-1}$ sichergestellt, was sich mit folgendem bestimmten Integral ergibt:

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} \quad (2.41)$$

Mittelwert: Der Mittelwert ergibt sich aus:

$$\langle x \rangle = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx \quad (2.42)$$

Zur Berechnung des Integrals setzt man $x = (x - \mu) + \mu$ und erhält damit die beiden Integrale:

$$\langle x \rangle = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu) \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx}_{=0} + \mu \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx}_{=1} = \mu \quad (2.43)$$

Varianz: Die Varianz ergibt sich mit Hilfe eines weiteren bestimmten Integrals:

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2} dx = \frac{1}{2a} \sqrt{\frac{\pi}{a}} \quad (2.44)$$

Damit erhält man:

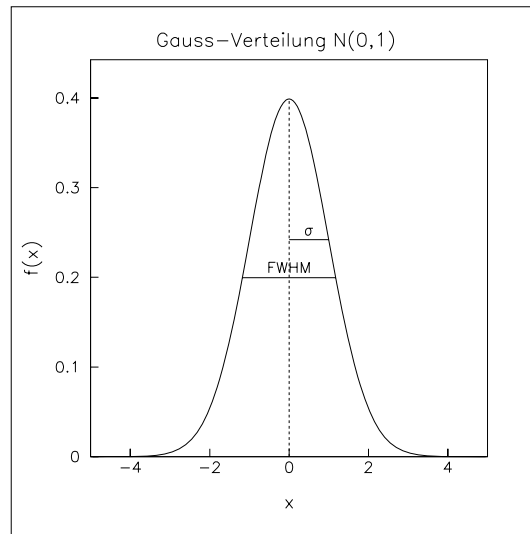
$$\langle (x - \mu)^2 \rangle = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^2 \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx = \sigma^2. \quad (2.45)$$

Standardisierte Normalverteilung: Durch die Transformation

$$x \rightarrow \frac{x - \mu}{\sigma} \quad (2.46)$$

erhält man eine Normalverteilung $N(0, 1)$ mit Mittelwert 0 und Varianz 1:

$$f(x) = f(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (2.47)$$

Abbildung 2.4: Standardisierte Normalverteilung $N(0, 1)$.

Eine standardisierte Normalverteilung ist in Abb. 2.4 gezeigt. Neben dem Mittelwert und der Standardabweichung σ ist auch die **volle Breite auf halber Höhe** des Maximums (FWHM = full width at half maximum) gezeigt. Diese Größe ist relativ einfach (mit Lineal und Bleistift) aus einer gemessenen Verteilung zu bestimmen. Für eine Gauss-Verteilung gibt es eine feste Beziehung zwischen FWHM und σ :

$$\frac{f(0)}{2} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(FWHM/2)^2}{2\sigma^2}\right) \implies FWHM = 2\sigma\sqrt{2\ln 2} \approx 2.355 \cdot \sigma \quad (2.48)$$

Verteilungsfunktion: Die Verteilungsfunktion der Normalverteilung ist nicht analytisch zu berechnen. Zahlenwerte findet man in Tabellen, in der Regel für die standardisierte Normalverteilung $N(0, 1)$ als Funktion von x . Den Übergang zu Verteilungen $N(\mu, \sigma)$ findet man durch Skalieren von x mit σ und Verschieben um μ :

$$x = \frac{x' - \mu}{\sigma} \quad (2.49)$$

Statt der Verteilungsfunktion findet man auch die sogenannte Fehlerfunktion (error function) (oder ‘Gauss’sches Fehlerintegral’) $erf(x)$ tabelliert:

$$\begin{aligned} erf(x) &= \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2} d\xi \\ \implies F(x) &= \frac{1}{2} \left[1 + erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \right] \end{aligned} \quad (2.50)$$

Vertrauensintervalle: Die Verteilungsfunktion benötigt man häufig zur Bestimmung der Wahrscheinlichkeit, dass ein Ereignis innerhalb bestimmter Grenzen für x liegt. Für die Beurteilung von Messergebnissen mit normalverteilten Fehlern benutzt man zum Beispiel die Wahrscheinlichkeit, in einem zentralen ‘Vertrauensintervall’ von $\pm n\sigma$ um den Mittelwert zu liegen (Abb. 2.5a, Tab. 2.1a):

$$p(\pm n\sigma) = F(\mu + n\sigma) - F(\mu - n\sigma) = erf\left(\frac{n\sigma}{\sqrt{2}\sigma}\right), \quad (2.51)$$

Tabelle 2.1: Wahrscheinlichkeiten innerhalb von $\pm n\sigma$ -Bereichen einer Normalverteilung.

a)	n	$p(\pm n\sigma)$	b)	$p(\pm n\sigma)$	n
	1	0.6827		0.900	1.645
	2	0.9545		0.950	1.960
	3	0.9973		0.990	2.576
	4	$1 - 6.3 \cdot 10^{-5}$		0.999	3.290

Häufig gibt man auch die Wahrscheinlichkeit, das ‘Vertrauensniveau’ (confidence level, c. l.), vor und fragt nach den entsprechenden Grenzen (Tab. 2.1b).

Innerhalb von 2 Standardabweichungen, $\pm 1\sigma$, um den Mittelwert liegen also 68.27% aller Ereignisse. Häufig werden Fehler so definiert, dass 68.27% innerhalb der Fehlergrenzen liegen, auch wenn die zugrundeliegende Verteilung nicht die Normalverteilung ist (‘Standardfehler’). Bei asymmetrischen Verteilungen können die Fehler auch asymmetrisch um den Mittelwert definiert werden, zum Beispiel so, dass jeweils 16% oberhalb und unterhalb des Fehlerbereichs liegen.

Welches Vertrauensniveau man für eine Aussage verlangt, hängt von der Problemstellung ab. Während man standardmäßig bei Messergebnissen das 1σ -Niveau angibt, verlangt man zur Festlegung von Toleranzgrenzen für Risiken, die das Leben von Menschen gefährden, viel höhere Vertrauensniveaus. Ob man nun 90% oder 99,9% oder 99,9999% verlangt, hängt unter anderem von der ‘a priori’ Wahrscheinlichkeit für das Risiko, also zum Beispiel die Größe der gefährdeten Gruppe, ab (‘Bayesischer Ansatz’). Wenn ein Fahrstuhl zum Beispiel im Mittel 1 Million mal während seiner Lebensdauer benutzt wird, sollte die Wahrscheinlichkeit für das Reißen des Seils kleiner als 10^{-6} sein.

Ausschlussgrenzen: Häufig möchte man ein bestimmtes Vertrauensniveau angeben, dass bei einem gegebenen Messwert x^{mess} der wahre Wert x^{wahr} oberhalb oder unterhalb einer Grenze liegt.

Beispiel: Um in der Elementarteilchenphysik die Entdeckung eines neuen Teilchens zu etablieren, wird ein Vertrauensniveau von mindestens 5 Standardabweichungen verlangt, weil jeder Physiker, der mal 1000 Histogramme mit je etwa 100 Bins angeschaut hat, eine gute Chance hat, wenigstens einen 4σ -Effekt zu beobachten. Ist dagegen ein Teilchen vorhergesagt und man findet oberhalb eines Untergrundes kein Signal, gibt man in der Regel untere Grenzen für die Häufigkeit der Erzeugung des Teilchens mit 90% oder 95% Vertrauensniveau an.

Will man zum Beispiel mit 95% Vertrauensniveau (95% c. l.) bei gegebenem Messwert x^{mess} eine obere Grenze für x^{wahr} angeben, stellt man die Frage: Was ist der Wert x_{95}^o , für den die Wahrscheinlichkeit, einen Messwert x^{mess} oder kleiner zu erhalten, 5% beträgt. Die Grenze x_{95}^o wird also als Mittelwert einer Gauss-Verteilung (mit bekannter, gemessener oder geschätzter Standardabweichung) gesucht, deren Integral von $-\infty$ bis x^{mess} 5% beträgt (Abb. 2.5b). Wegen der Symmetrie der Gauss-Verteilung kann man aber auch von einer entsprechenden Gaussverteilung um den gemessenen Wert ausgehen und x_{95}^o als denjenigen Wert bestimmen, für den das Integral über $x > x_{95}^o$ die geforderten 5% bzw. das Komplement 95% ergibt:

$$F(x_{95}^o) = 0.95 \quad (2.52)$$

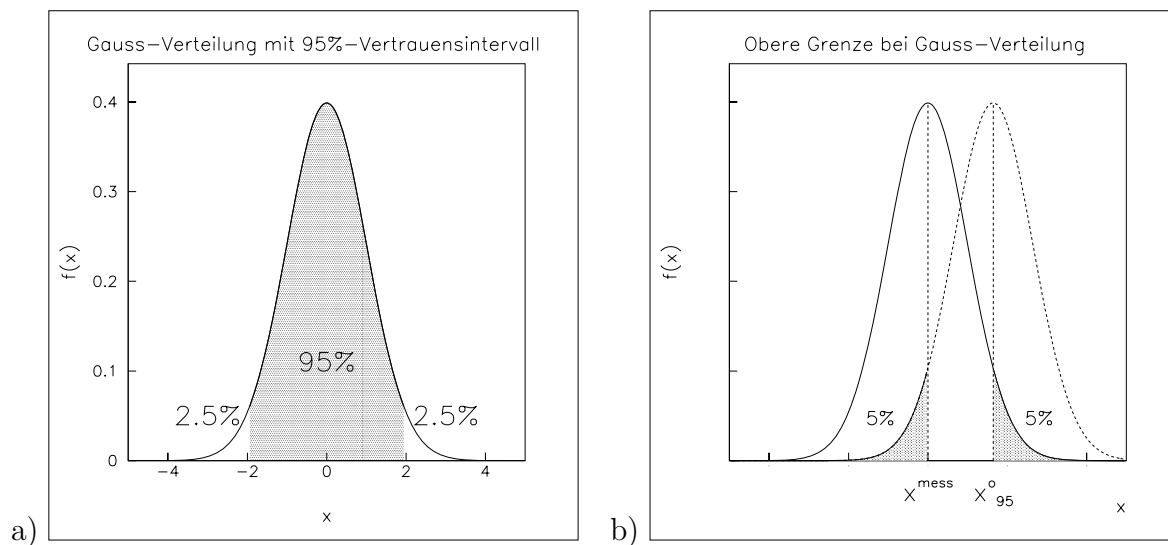


Abbildung 2.5: a) Fläche unter einer Gauss-Kurve, die einem Vertrauensintervall von 95% entspricht. b) Bestimmung einer oberen Grenze bei normalverteilten Fehlern, hier mit einem Vertrauensniveau von 95%. Links ist die Verteilung um den Messwert, rechts die Verteilung um den Wert der oberen Grenze. Die schattierten Bereiche entsprechen jeweils 5% Wahrscheinlichkeit. Siehe weitere Erläuterungen im Text.

Entsprechend ergibt sich für eine untere Grenze mit 95% Vertrauensniveau:

$$F(x_{95}^u) = 0.5 \quad (2.53)$$

Man schreibt dann zum Beispiel:

$$x < x_{95}^u, \quad 95\% \text{ c. l.} \quad (2.54)$$

Bei angenommenen gauss-verteilten Fehlern sind also die die Grenzen einfach aus der Verteilungsfunktion zu bestimmen. Im allgemeinen Fall muss man aber auf die oben angegebene Definition zurückgreifen. Zum Beispiel kommt es häufig vor, dass man auf der Suche nach einem Ereignis nichts findet, also ein Nullergebnis hat. Wenn es sich um ein Zählratenexperiment handelt, ergibt sich bekanntlich für eine Poisson-Verteilung eine endliche Wahrscheinlichkeit auch bei einem nicht-verschwindenden Mittelwert ($\lambda \neq 0$) ein Nullergebnis zu erhalten. Man kann dann nur eine obere Grenze für den wahren Wert von λ geben. Entsprechend der oben angegebene Definition fragt man für ein gefordertes Vertrauensniveau ϵ : für welchen Mittelwert λ_ϵ^o ist die Wahrscheinlichkeit die Zählrate 0 (oder kleiner) zu erhalten gerade $1 - \epsilon$:

$$p(n, \lambda) = p(0, \lambda_\epsilon^o) = \frac{(\lambda_\epsilon^o)^0}{0!} e^{-\lambda_\epsilon^o} = e^{-\lambda_\epsilon^o} \stackrel{!}{=} 1 - \epsilon \quad (2.55)$$

$$\implies \lambda_\epsilon^o = -\ln(1 - \epsilon) \quad (2.56)$$

Die Grenzen für 90 und 95% Vertrauensniveau sind bei 0 beobachteten Ereignissen:

$$\begin{aligned} \lambda_{90}^o &= 2.30 \\ \lambda_{95}^o &= 3.00 \end{aligned} \quad (2.57)$$

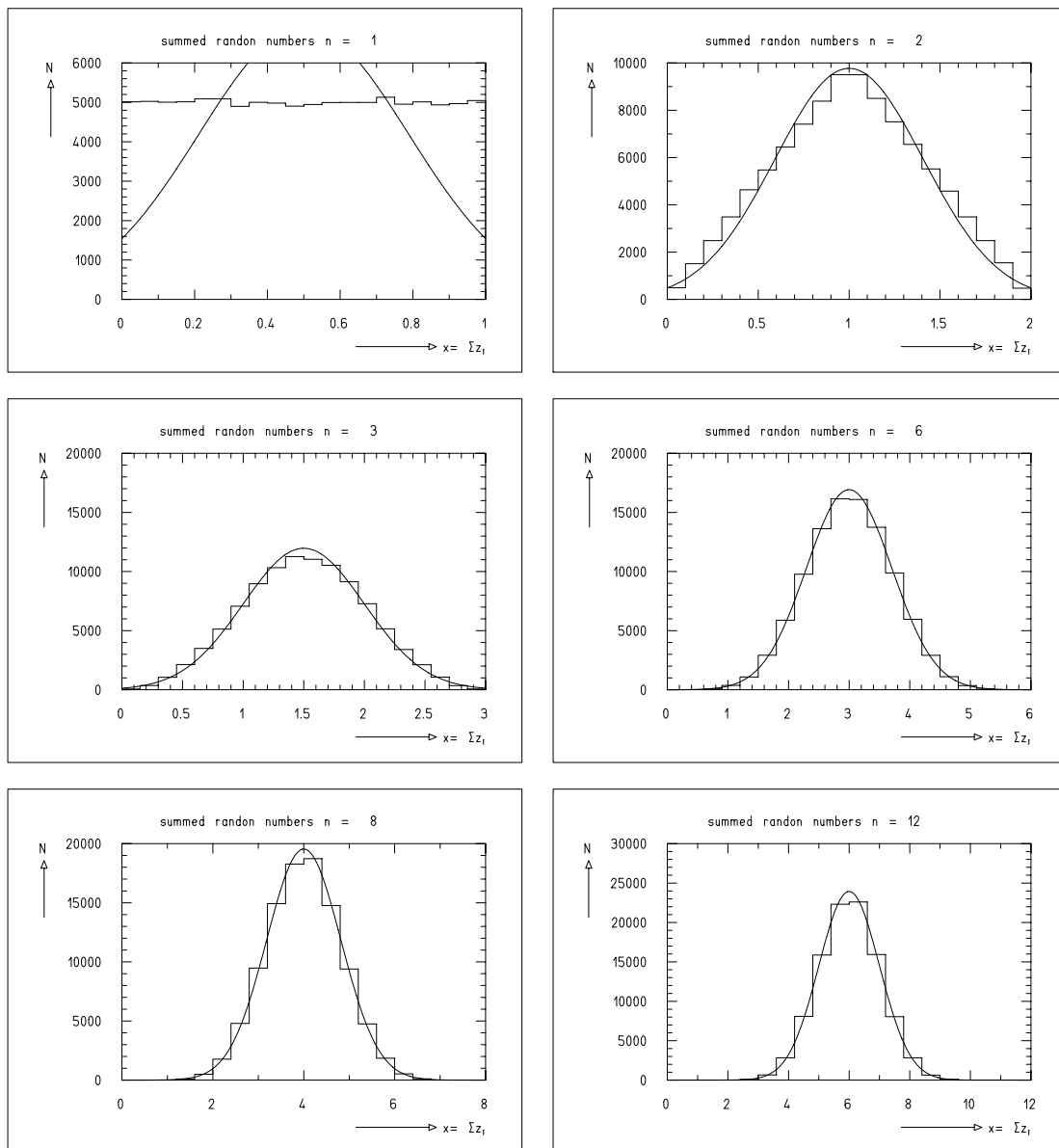


Abbildung 2.6: Beispiele von Verteilungen der Summen von n zwischen 0 und 1 gleichverteilten Zufallszahlen. Die Verteilungen werden mit Gauss-Verteilungen mit Mittelwert $\mu = n/2$ und Varianz $\sigma^2 = n/12$ verglichen.

2.6 Zentraler Grenzwertsatz

Die Gauss-Verteilung hat unter allen Verteilungen eine besondere Bedeutung, weil sie für viele Verteilungen ein Grenzfall für große Zahlen darstellt. Wir hatten das bereits für die Binomial- und die Poisson-Verteilung gesehen, die beide im Grenzfall großer Mittelwerte in die Gauss-Verteilung übergehen.

Die Gauss-Verteilung kann interpretiert werden als Verteilung von Abweichungen um einen Mittelwert, die sich als Überlagerung vieler kleiner Störungen ergeben. Tatsächlich findet man, dass die Summe von n beliebigen Zufallsvariablen für große n einer Gauss-Verteilung zustrebt. In Übungsaufgabe 8 wurde das für die Summe von gleichverteilten Zufallszahlen gezeigt, wobei sich zeigte, dass die Verteilung der Summe von 12 solchen Zufallszahlen bereits sehr gut eine Gauss-Verteilung approximiert (Abb. 2.6).

Diese Eigenschaft der Gauss-Verteilung wird mathematisch im Zentralen Grenzwertsatz formuliert: Gegeben seien n unabhängige Variablen x_i , $i = 1, \dots, n$, die jeweils einer Verteilung mit Mittelwert μ_i und Varianz σ_i entnommen sind (die Verteilungen sind ansonsten beliebig). Dann hat die Verteilung der Summe

$$X = \sum_{i=1}^n x_i \quad (2.58)$$

folgende Eigenschaften:

(i) Erwartungswert:

$$\langle X \rangle = \sum_{i=1}^n \mu_i; \quad (2.59)$$

(ii) Varianz:

$$\sigma_X^2 = \sum_{i=1}^n \sigma_i^2; \quad (2.60)$$

(iii) die Verteilung nähert sich einer Gauss-Verteilung für

$$n \rightarrow \infty. \quad (2.61)$$

Zum Beweis von (2.59) und (2.60) benutzt man die Linearität der Erwartungswertbildung: der Erwartungswert einer Summe unabhängiger Zufallszahlen ist die Summe der Erwartungswerte. Für den Erwartungswert von X ergibt sich:

$$\langle X \rangle = \left\langle \sum_i x_i \right\rangle = \sum_i \langle x_i \rangle = \sum_i \mu_i. \quad (2.62)$$

Entsprechend ergibt sich für die Varianz:

$$\begin{aligned} \sigma_X^2 &= \langle (X - \langle X \rangle)^2 \rangle = \left\langle \left(\sum_i x_i - \sum_i \mu_i \right)^2 \right\rangle = \left\langle \left(\sum_i (x_i - \mu_i) \right)^2 \right\rangle \\ &= \sum_i \langle (x_i - \mu_i)^2 \rangle + \sum_i \sum_{j \neq i} \underbrace{\langle (x_i - \mu_i)(x_j - \mu_j) \rangle}_{=0, \text{ wenn } i, j \text{ unabhängig}} = \sum_i \sigma_i^2 \end{aligned} \quad (2.63)$$

Der Beweis der wichtigen Aussage (2.61) ist schwieriger und kann in Statistikbüchern nachgelesen werden, zum Beispiel [1, 2]. Abbildung 2.6 zeigt die Summe gleichverteilter Variablen, die sich der Gauss-Verteilung mit wachsender Anzahl Variabler annähert.