Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun
http://blog.ruediger-braun.net

Heinrich-Heine-Universität Düsseldorf

17. Januar 2014

Verteilungsfunktionen

Verteilungsfunktion

- X sei eine Zufallsvariable. Die Funktion

$$F(x) = P(X \leq x)$$

  ist die *Verteilungsfunktion* von X
- Die Verteilungsfunktion $F(x)$ gibt an, mit welcher Wahrscheinlichkeit ein **kleinerer** Wert als x angenommen wird (x eingeschlossen)
- Die Verteilungsfunktion wächst also monoton

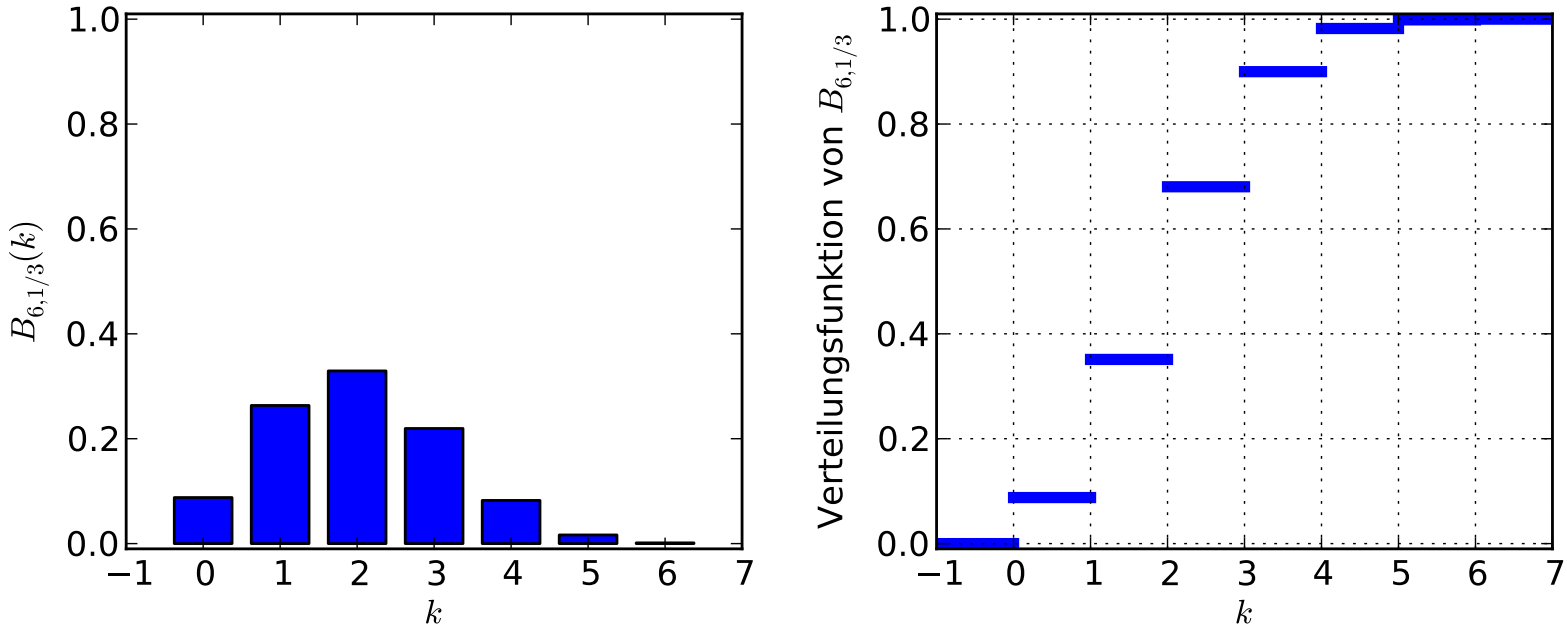
Beispiel: Verteilungsfunktion von $B_{6,1/3}$

F sei die Verteilungsfunktion von $B_{6,\,1/3}$, also

$$F(r) = \sum_{k \leq r} B_{6,\,1/3}(k)$$

k	$B_{6,\,1/3}(k)$	$F(k)$
0	0.0878	0.0878
1	0.2634	0.3512
2	0.3292	0.6804
3	0.2195	0.8999
4	0.0823	0.9822
5	0.0165	0.9986
6	0.0014	1.0000

Beispiel: Verteilungsfunktion von $B_{6,1/3}$



Stabdiagramm und Graph der Verteilungsfunktion von $B_{6,1/3}$

Kumulierte Tabellen

- Kumulierte Tabellen der Binomialverteilung zeigen die Verteilungsfunktion
- Beispiel: 47 Versuche mit Erfolgswahrscheinlichkeit $p = 0.88$ im Einzelfall wurden gemacht. Mit welcher Wahrscheinlichkeit gelingen weniger als 44, aber mehr als 39, beide Grenzen eingeschlossen?

$$\begin{aligned} P(39 \leq X \leq 44) &= P(X \leq 44) - P(X \leq 38) \\ &= F(44) - F(38) \\ &= 0.93236 - 0.10408 \\ &= 0.82828 \end{aligned}$$

- Beachte: Weil X nur ganze Zahlen annimmt, ist

$$P(X < 39) = P(X \leq 38)$$

Tabelle der Werte $\sum_{k=0}^{r} B_{n,p}(k)$ für $n = 47$

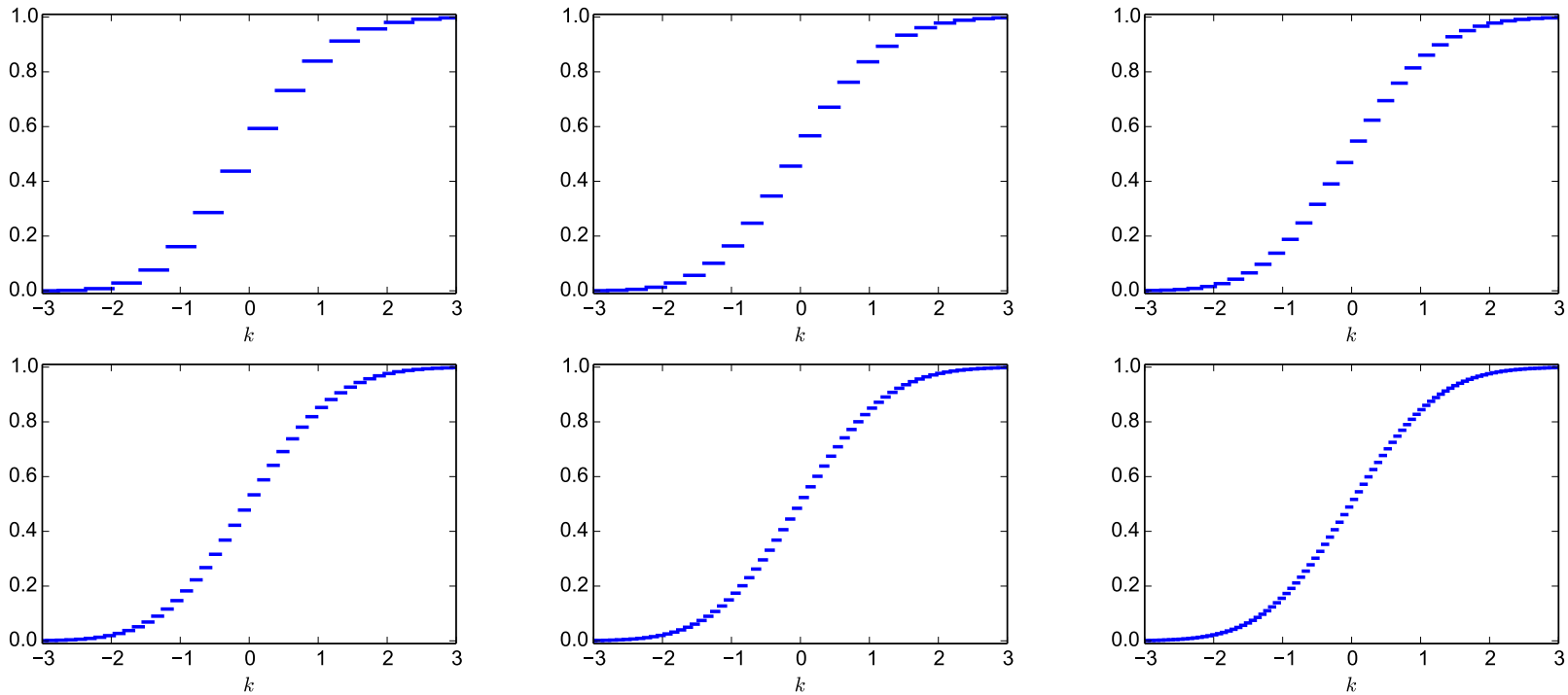
r	p	0.85	0.86	0.87	0.88	0.89
27	0.	00001				
28		00002	00001			
29		00008	00003	00001		
30		00029	00012	00005	00002	00001
31		00093	00043	00018	00007	00002
32		00274	00137	00063	00026	00010
33		00742	00398	00199	00091	00038
34		01832	01060	00573	00286	00130
35		04128	02571	01503	00817	00408
36		08463	05663	03578	02115	01156
37		15768	11311	07707	04946	02957
38		26660	20441	14978	10408	06792
39		40904	33384	26208	19651	13952
40		57047	49285	41238	33208	25538
41		72665	65962	58411	50182	41543
42		85309	80597	74830	67964	60042
43		93639	91050	87606	83128	77447
44		97931	96887	95379	93236	90248
45		99552	99278	98847	98178	97153
46		99952	99917	99856	99754	99582

Stetige Zufallsvariable, Normalverteilung

Die standardisierte Verteilung

- Die Zufallsvariable X besitze den Erwartungswert $E(X) = \mu$ und die Varianz $Var(X) = \sigma^2$
- Setze
$$Y = \frac{X - \mu}{\sigma}$$
- Dann $E(Y) = 0$ und $Var(Y) = 1$
- Y ist die *standardisierte Zufallsvariable* zu X

Standardisierte Binomialverteilungen



Verteilungsfunktionen der Standardisierungen von $B_{n,p}$ für $p = 0.2$ und $n = 40, 80, 160, 320, 640$ und 1280

Zentraler Grenzwertsatz

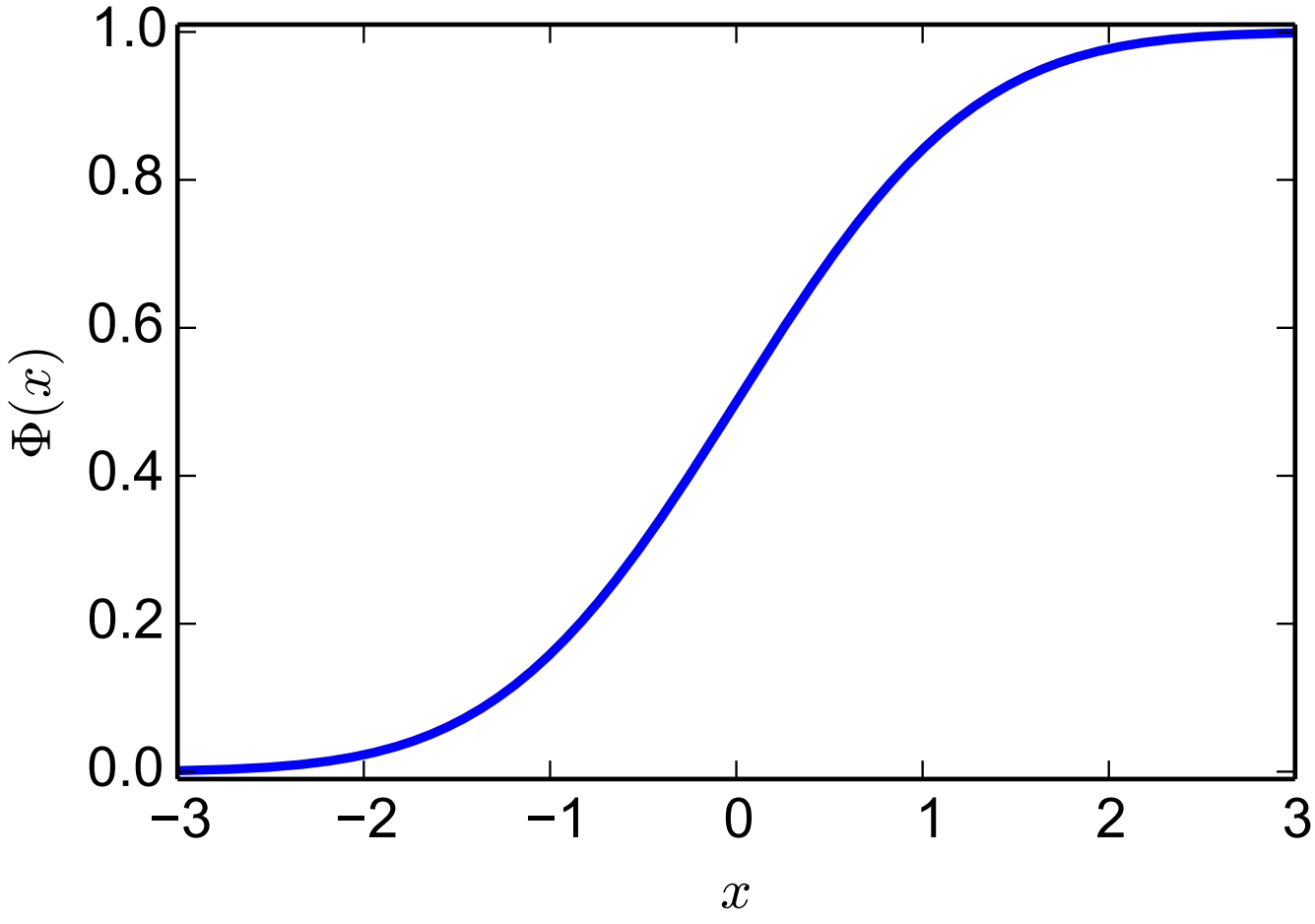
Für jeden Parameterwert p konvergiert der Grenzprozess auf der letzten Folie gegen dieselbe Funktion.

Diese Funktion heißt

Verteilungsfunktion der Standard-Normalverteilung

Sie wird mit Φ bezeichnet.

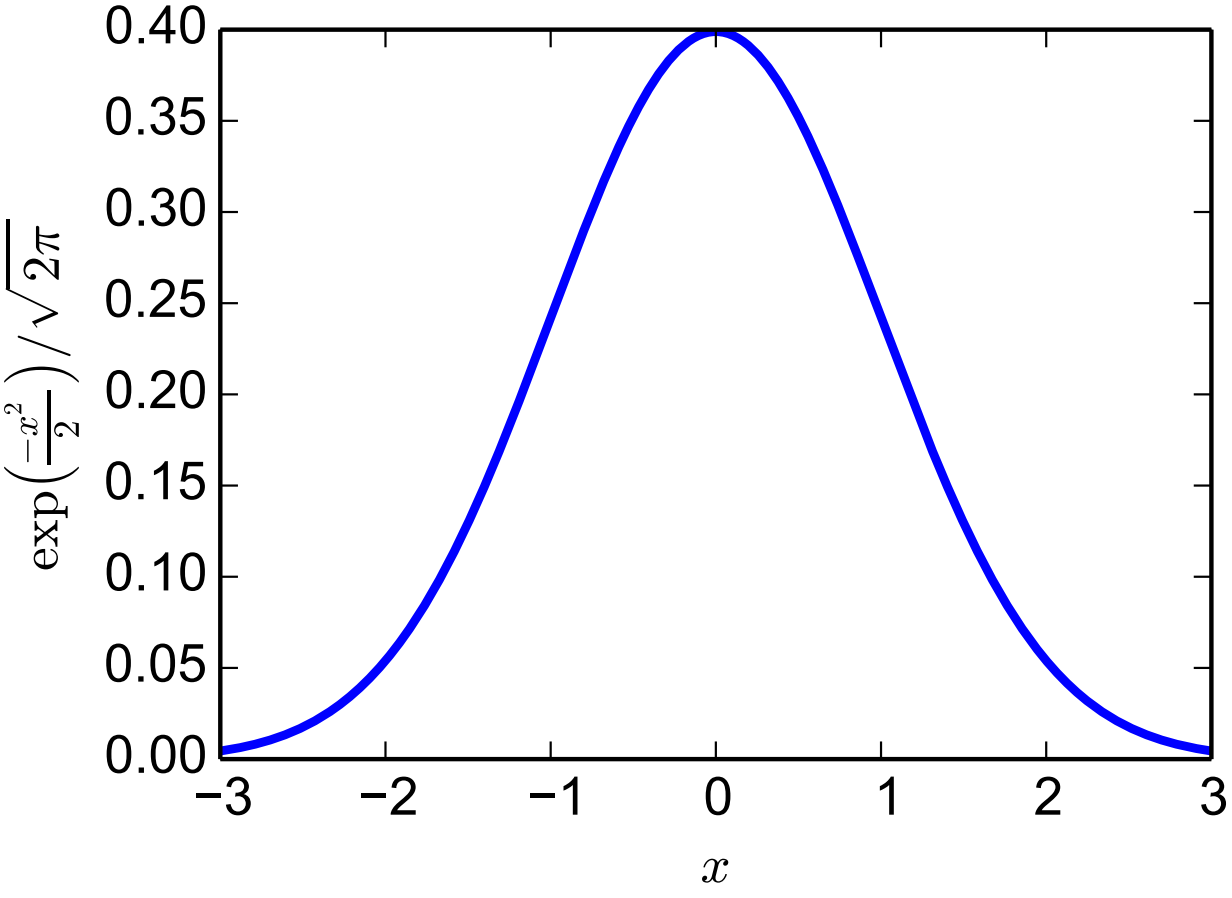
Verteilungsfunktion der Standard-Normalverteilung



Gaußsche Glockenkurve

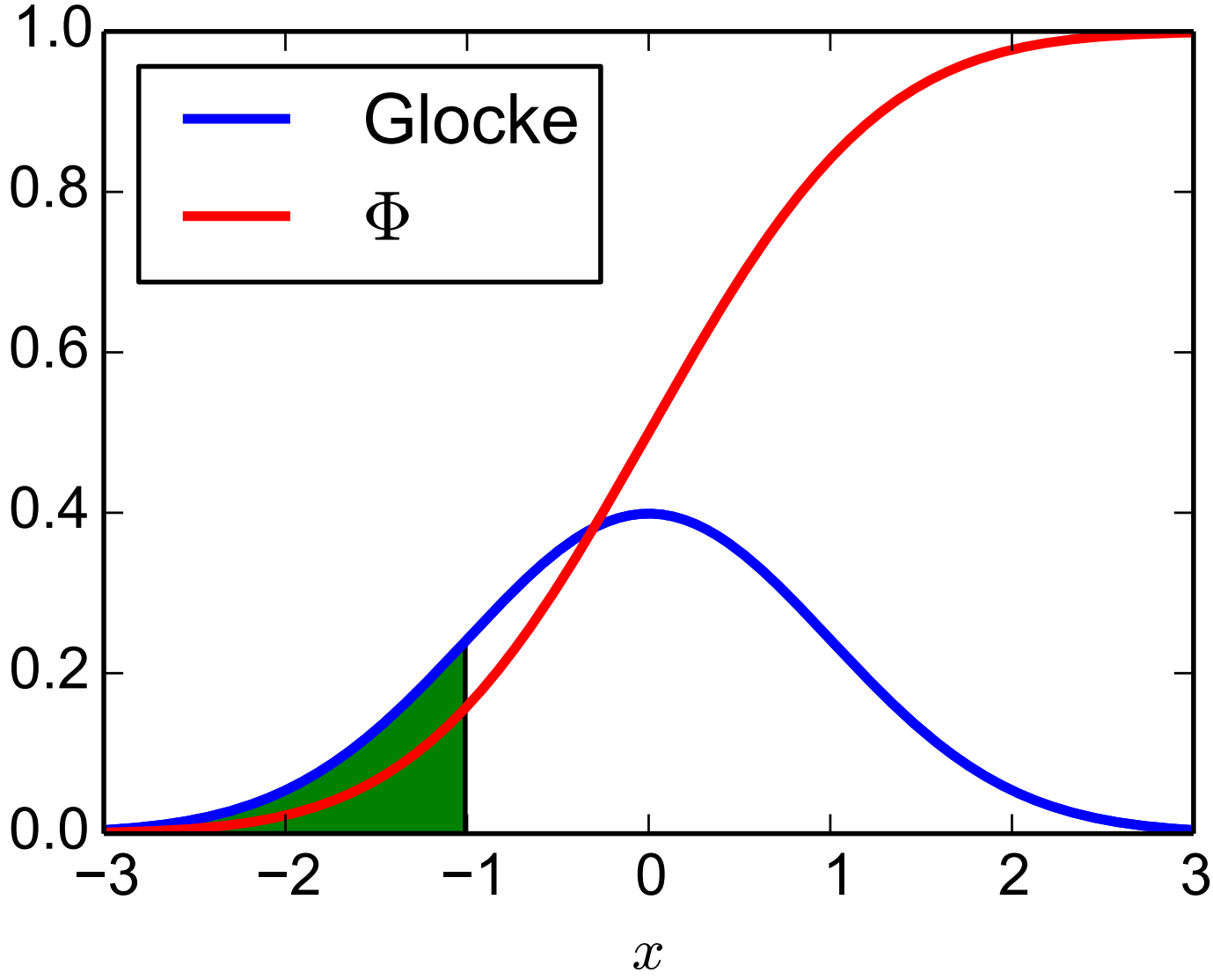
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

heißt *Gaußsche Glockenkurve*



Standard-Normalverteilung als Fläche

$\Phi(x)$ ist die Fläche links von x unter der Gaußschen Glockenkurve



Standard-Normalverteilung als Integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt$$

Für dieses Integral existiert keine geschlossene Form in Termen klassischer Funktionen

Stetige Zufallsvariable

Eine Zufallsvariable X, deren Verteilungsfunktion von der Gestalt

$$P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

ist, heißt *stetig*. Dann bezeichnet man f als die Dichte von X.

Die Gaußsche Glockenkurve ist die Dichte der Standard-Normalverteilung.

$$P(x < X \leq y) = \int_{x}^{y} f(t)dt$$

$$P(X = x) = 0$$

$$P(x \leq X \leq y) = \int_{x}^{y} f(t)dt$$

Für **stetige** Zufallsvariable gilt also $P(x < X \leq y) = P(x \leq X \leq y)$. Bei diskreten ist das nicht so.

Verteilungsfunktion der Standard-Normalverteilung

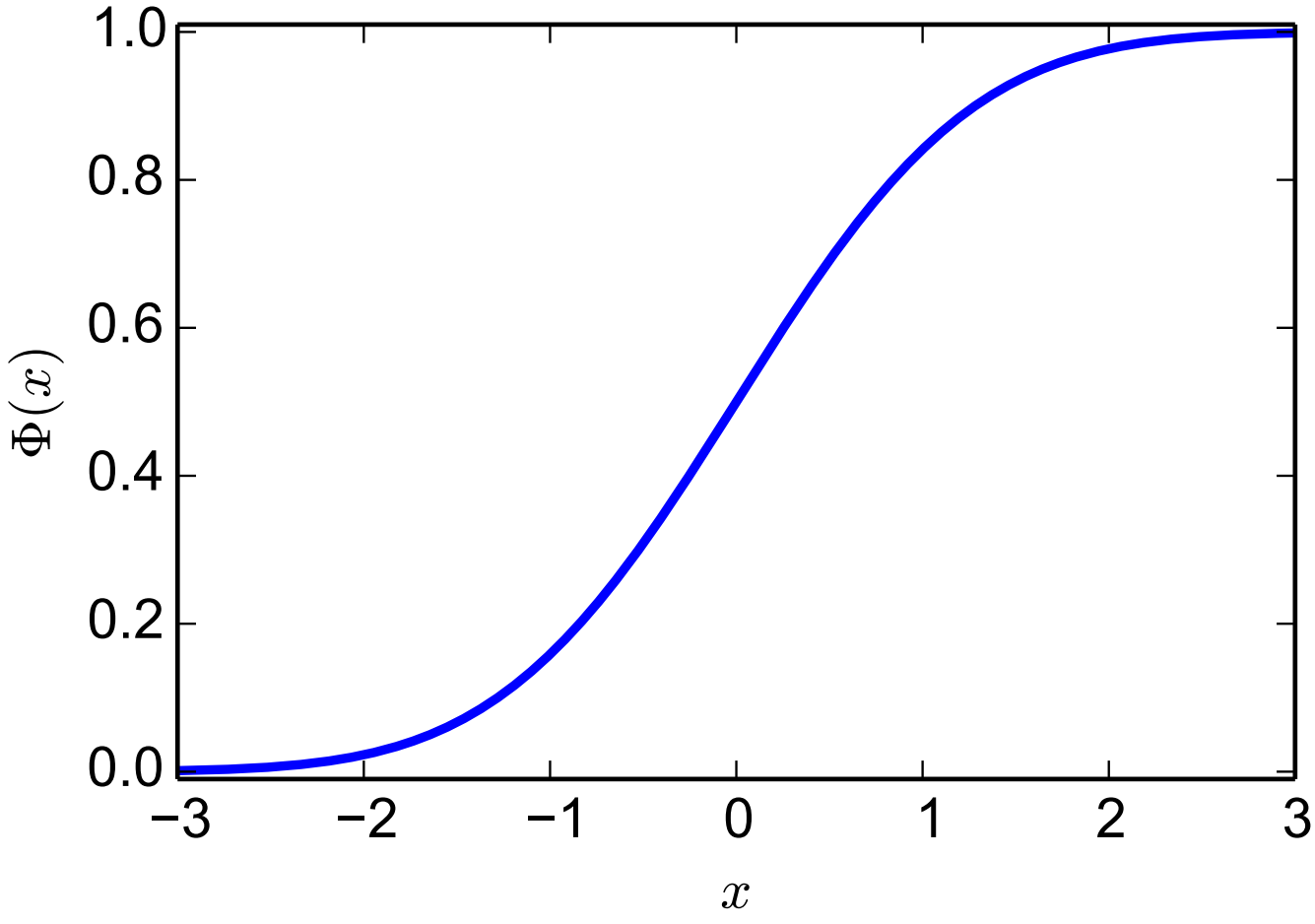


Tabelle der Standard-Normalverteilung, linke Seite

u	0.00	0.01	0.02	0.03	0.04
0.0	0,500000	0,503989	0,507978	0,511966	0,515953
0.1	,539828	,543795	,547758	,551717	,555670
0.2	,579260	,583166	,587064	,590954	,594835
0.3	,617911	,621720	,625516	,629300	,633072
0.4	,655422	,659097	,662757	,666402	,670031
0.5	,691462	,694974	,698468	,701944	,705401
0.6	,725747	,729069	,732371	,735653	,738914
0.7	,758036	,761148	,764238	,767305	,770350
0.8	,788145	,791030	,793892	,796731	,799546
0.9	,815940	,818589	,821214	,823814	,826391
1.0	,841345	,843752	,846136	,848495	,850830
1.1	,864334	,866500	,868643	,870762	,872857
1.2	,884930	,886861	,888768	,890651	,892512
1.3	,903200	,904902	,906582	,908241	,909877
1.4	0,919243	0,920730	0,922196	0,923641	0,925066

Tabelle der Standard-Normalverteilung, rechte Seite

u	0.05	0.06	0.07	0.08	0.09
0.0	0,519939	0,523922	0,527903	0,531881	0,535856
0.1	,559618	,563559	,567495	,571424	,575345
0.2	,598706	,602568	,606420	,610261	,614092
0.3	,636831	,640576	,644309	,648027	,651732
0.4	,673645	,677242	,680822	,684386	,687933
0.5	,708840	,712260	,715661	,719043	,722405
0.6	,742154	,745373	,748571	,751748	,754903
0.7	,773373	,776373	,779350	,782305	,785236
0.8	,802337	,805105	,807850	,810570	,813267
0.9	,828944	,831472	,833977	,836457	,838913
1.0	,853141	,855428	,857690	,859929	,862143
1.1	,874928	,876976	,879000	,881000	,882977
1.2	,894350	,896165	,897958	,899727	,901475
1.3	,911492	,913085	,914657	,916207	,917736
1.4	0,926471	0,927855	0,929219	0,930563	0,931888

Lesehinweise

- Auf dem Weblog gibt es die komplette Tabelle im Format A4 unter http://www.math.uni-duesseldorf.de/~braun/bio1314/tabNorm.pdf
- Beispiel $\Phi(0.31) = 0.621720$
- Wegen der Punktsymmetrie, also wegen

$$\Phi(-x) = 1 - \Phi(x)$$

kann man die Tabelle auch für negative Argumente verwenden. Also beispielsweise
$\Phi(-1.34) = 1 - \Phi(1.34) = 1 - 0.909877 = 0.090123$

- Durch Aufsuchen des Funktionswertes erhält man die Umkehrfunktion: Gesucht u mit $\Phi(u) = 0.79$. Man liest ab: $u = 0.81$

Tabelle der Standard-Normalverteilung, linke Seite

u	0.00	0.01	0.02	0.03	0.04
0.0	0,500000	0,503989	0,507978	0,511966	0,515953
0.1	,539828	,543795	,547758	,551717	,555670
0.2	,579260	,583166	,587064	,590954	,594835
0.3	,617911	,621720	,625516	,629300	,633072
0.4	,655422	,659097	,662757	,666402	,670031
0.5	,691462	,694974	,698468	,701944	,705401
0.6	,725747	,729069	,732371	,735653	,738914
0.7	,758036	,761148	,764238	,767305	,770350
0.8	,788145	,791030	,793892	,796731	,799546
0.9	,815940	,818589	,821214	,823814	,826391
1.0	,841345	,843752	,846136	,848495	,850830
1.1	,864334	,866500	,868643	,870762	,872857
1.2	,884930	,886861	,888768	,890651	,892512
1.3	,903200	,904902	,906582	,908241	,909877
1.4	0,919243	0,920730	0,922196	0,923641	0,925066

Beispiel

- Normalverteilungen werden beispielsweise zur Modellierung von Messfehlern benutzt
- Beispiel: Die Wirkstoffkonzentration in einem Heilmittel soll 4g/l betragen. Herstellungsabhängig beträgt die Streuung 100mg/l
- Mit welcher Wahrscheinlichkeit liegt die Konzentration über 4.12g/l?
- Y sei die Konzentration in g/l
- Dann $Y = 4 + 0.1 \cdot X$, wobei X standard-normalverteilt ist
- Gesucht $P(Y > 4.12)$
- Das ist $P(X > 1.2) = 1 - P(X \leq 1.2) = 1 - \Phi(1.2) = 1 - 0.884930 = 0.115070$

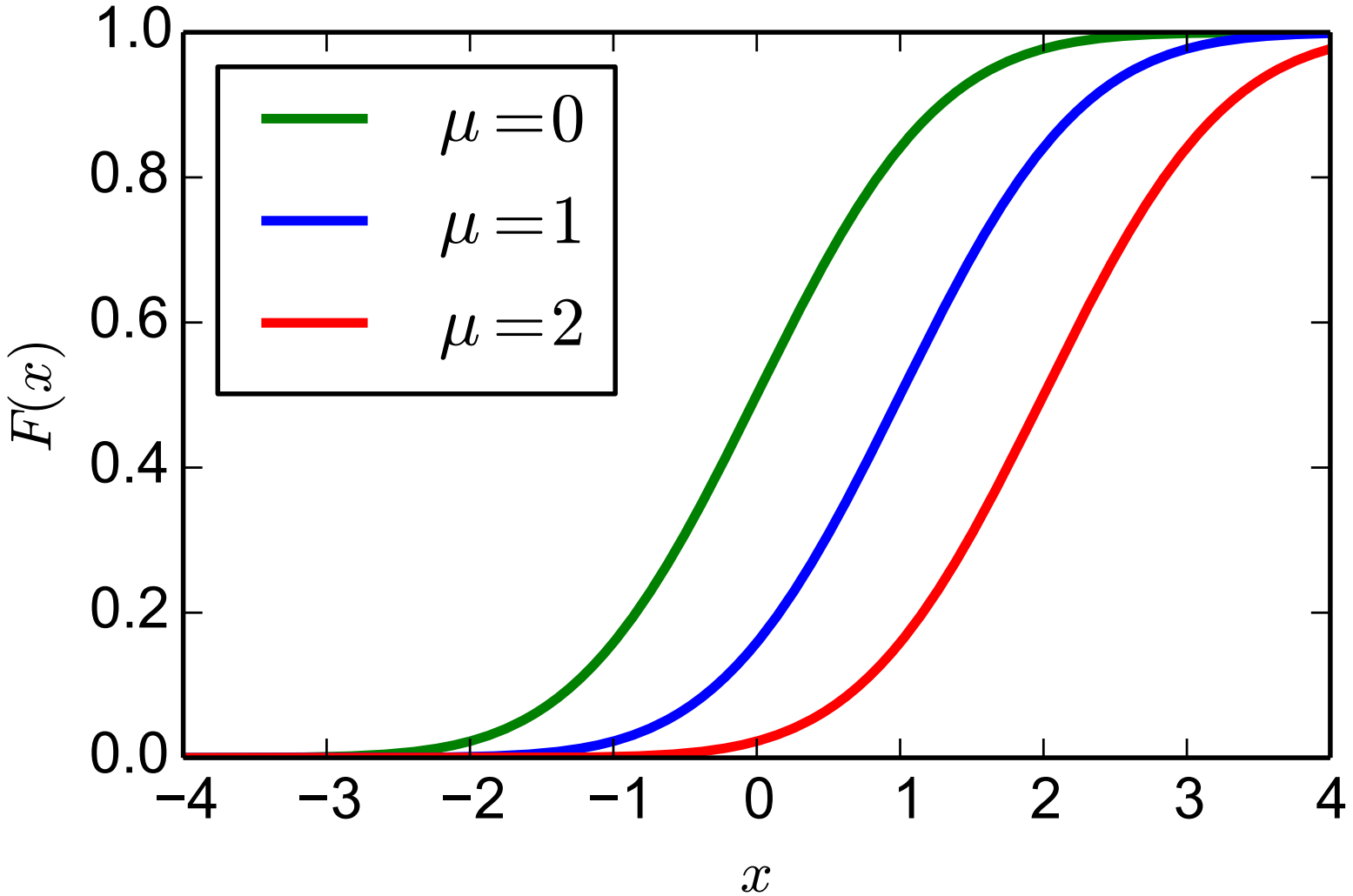
Normalverteilungen

- Die Zufallsvariable X heißt *normalverteilt* zum Erwartungswert μ und der Varianz σ^2, wenn

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

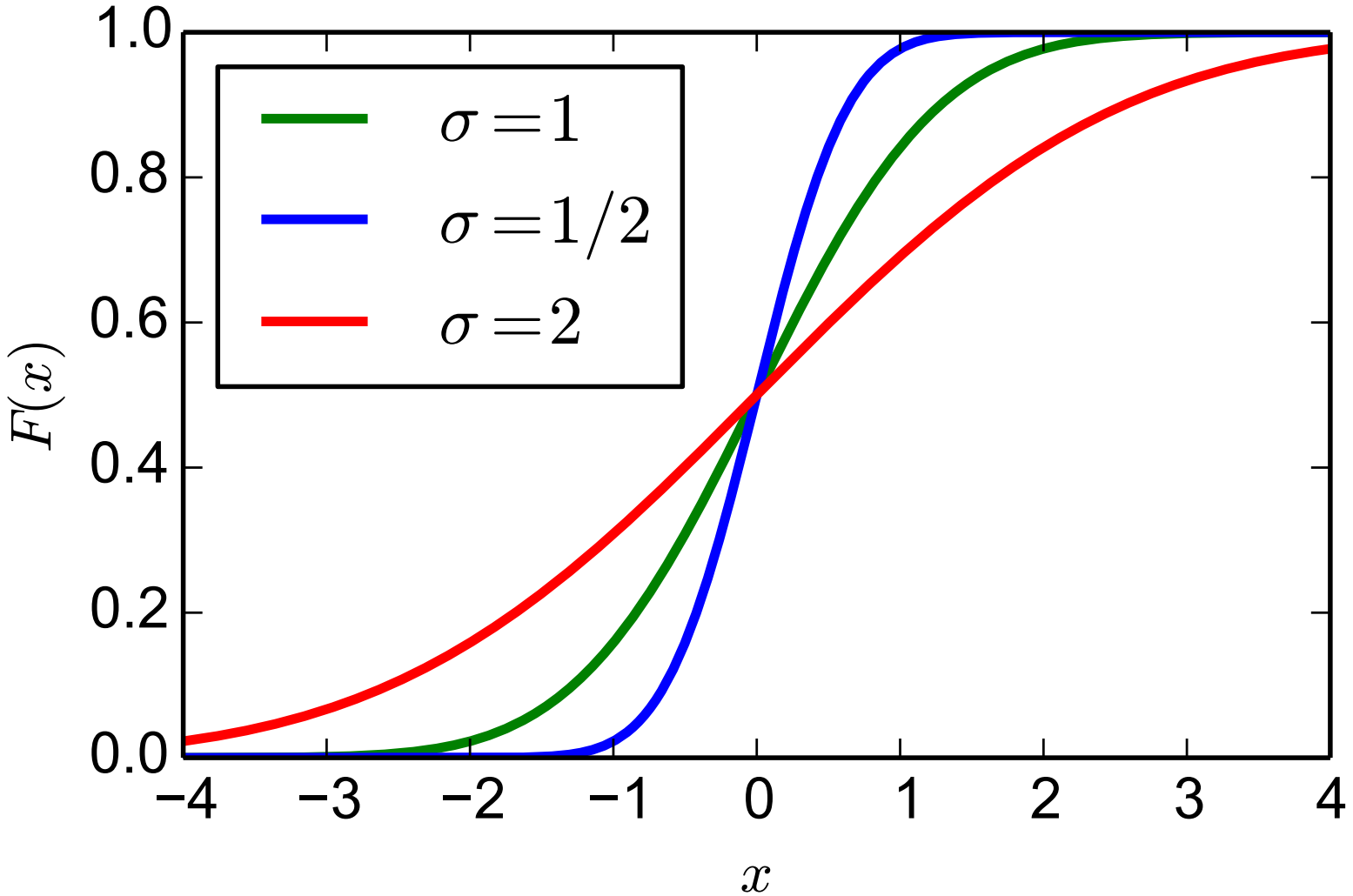
  standard-normalverteilt ist. Man sagt dann, X sei $N(\mu, \sigma^2)$-verteilt
- $X = \mu + \sigma \cdot Y$
- Y ist die Standardisierung von X

Normalverteilungen für verschiedene Erwartungswerte



Verteilungsfunktionen für $N(\mu, 1)$-verteilte Zufallsvariable

Normalverteilungen für verschiedene Streuungen



Verteilungsfunktionen für $N(0, \sigma^2)$-verteilte Zufallsvariable

Umrechnung auf Standardnormalverteilung

Die Zufallsvariable X sei $N(\mu, \sigma^2)$-verteilt. Dann gelten für $a < b$

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

$$P(a < X) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

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

Beispiel: natürliche Variabilitäten

- Roggenpflanzen erreichen eine mittlere Höhe von $0.98m$. Dabei streut die Höhe um $19cm$. Welcher Prozentsatz aller Pflanzen erreicht mindestens $1.10m$ Höhe?
- $X =$ Höhe der Pflanze
- Wir rechnen in Metern. Dann $E(X) = 0.98$ und $\sigma = 0.19$
- Wir suchen

$$P(1.1 < X) = 1 - \Phi\left(\frac{1.1 - 0.98}{0.19}\right) = 1 - \Phi(0.6316)$$
$$= 1 - 0.735653 = 0.2644$$

- ca 26% der Pflanzen sind mindestens $1.10m$ hoch

Kritische Betrachtung des Modells

- Das Modell erlaubt auch den unsinnigen Fall, dass Roggenpflanzen eine negative Höhe aufweisen
- Mit welcher Wahrscheinlichkeit geschieht das?

$$P(X < 0) = \Phi\left(\frac{-0.98}{0.19}\right) = \Phi(-5.158) = 1 - \Phi(5.158)$$
$$\leq 1 - \Phi(5) = 1 - (1 - 2.867 \cdot 10^{-7}) = 2.867 \cdot 10^{-7}$$

- Das Modell sagt für weniger als eine unter 3 Millionen Pflanzen eine negative Höhe voraus
- Damit können wir leben

Beispiel: IQ-Tests

- IQ-Tests sind so skaliert, dass die Werte in der Population normalverteilt mit Erwartungswert $\mu = 100$ und Streuung $\sigma = 15$ sind
- Welcher Anteil der Bevölkerung hat einen IQ über 130?
- X messe den IQ
- X ist $N(100, 225)$-verteilt.
- Also

$$P(130 < X) = 1 - \Phi\left(\frac{130 - 100}{15}\right)$$
$$= 1 - \Phi(2) = 1 - 0.977250 = 0.02275$$

- Ungefähr 2.3% der Population weist einen IQ von mindestens 130 auf

Die 3σ-Regel

- $X_1, \ldots, X_n$ unabhängig, alle mit derselben Verteilung
- $\mu = E(X_1) = \cdots = E(X_n)$ und
  $\sigma^2 = Var(X_1) = \cdots = Var(X_n)$
- Arithmetisches Mittel der X_j

$$Y = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

- n ausreichend groß, dann näherungsweise

$$P\left(\left|Y - \mu\right| \leq \frac{\sigma}{\sqrt{n}}\right) \geq \frac{2}{3}$$

$$P\left(\left|Y - \mu\right| \leq \frac{2\sigma}{\sqrt{n}}\right) \geq 0.95$$

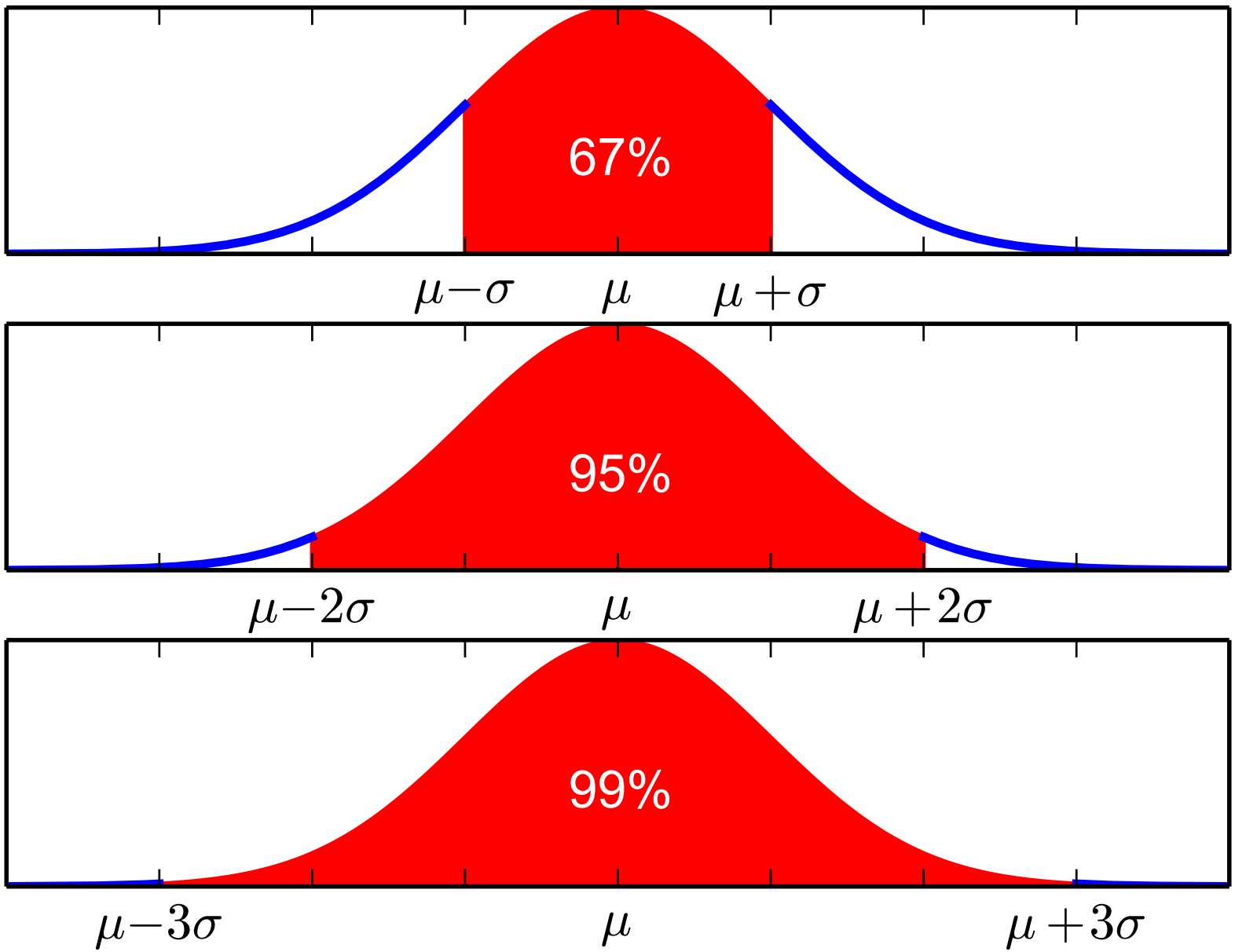
$$P\left(\left|Y - \mu\right| \leq \frac{3\sigma}{\sqrt{n}}\right) \geq 0.99$$

- Die 3σ-Regel ist nur eine Faustregel.

Erläuterung der 3σ-Regel

- $E(Y) = \mu$
- $Var(Y) = n\sigma^2$
- Also ist $n(Y - \mu)$ die Standardisierung von Y
- Für große n sieht die Standardisierung aus wie Φ

3σ-Regel, Zeichnung



3σ-Regel, Beispiel

- Der Wirkstoffgehalt von Düngetabletten wird gemessen (in mg). Die Streuung betrage $8mg$
- Wie viele Messungen schreibt die 3σ-Regel vor, um mit 95% Wahrscheinlichkeit einen Messfehler von weniger als $4mg$ zu haben?
- Wir brauchen

$$4 = \frac{2\sigma}{\sqrt{n}}$$

- Also

$$4 = \frac{2 \cdot 8}{\sqrt{n}}$$

- Das bedeutet $n = 16$

Schätzung eines Konfidenzintervalls mittels der 3σ-Regel

- $X_1, \ldots, X_n$ seien normalverteilt gemäß $N(\mu, \sigma^2)$ für bekanntes σ und unbekanntes μ
- Für μ soll ein Konfidenzintervall zum Konfidenzniveau 95% geschätzt werden.
- Das ist ein Intervall, welches den wahren Wert mit 95%-tiger Wahrscheinlichkeit enthält
- Den Mittelpunkt des Konfidenzintervalls bildet das arithmetische Mittel

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

Schätzung eines Konfidenzintervalls mittels der 3σ-Regel, Fortsetzung

- 3σ-Regel

$$P\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

- Beachte die Gleichwertigkeit der Ungleichungen

$$\overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}} \quad \text{und} \quad \mu \geq \overline{X} - \frac{2\sigma}{\sqrt{n}}$$

- Damit sieht die 3σ-Regel so aus

$$P\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

- Also ist folgendes Intervall ein Konfidenzintervall zum Konfidenzniveau 95%

$$\left[\overline{X} - \frac{2\sigma}{\sqrt{n}}, \quad \overline{X} + \frac{2\sigma}{\sqrt{n}}\right]$$

Beispiel: Roggenpflanzen

- Gesunde Roggenpflanzen einer bestimmten Art sind im Mittel 102.5 *cm* lang, wobei die Länge um 7 *cm* streut. Die Länge sei normalverteilt
- Durch Umwelteinflüsse änderte sich die mittlere Halmlänge, die Streuung aber nicht
- Die folgenden Längen wurden gemessen

96.62	94.91	85.05	101.61	109.55
93.05	97.86	96.66	95.08	98.87

- Arithmetisches Mittel der Daten

$$\overline{x} = 96.93$$

Roggenpflanzen, Fortsetzung

- $\overline{x} = 96.93$
- $\frac{2\sigma}{\sqrt{n}} = \frac{14}{\sqrt{10}} = 4.43$
- $96.93 - 4.43 = 92.50$ und $96.93 + 4.43 = 101.36$
- Also ergibt sich das folgende Konfidenzintervall zum Konfidenzniveau 95%

$$[92.50,\ 101.36]$$

- Die Länge gesunder Pfanzen beträgt im Mittel $102.5cm$. Wir können also mit 95%-tiger Sicherheit feststellen, dass die beobachteten Pflanzen nicht gesund sind.

Definition

- Es sei Θ eine Menge von Parameterwerten. Zu jedem Parameterwert $\theta \in \Theta$ gebe es eine Verteilung P_θ
- Von der Zufallsvariablen X sei bekannt, dass ihre Verteilung gleich einem der P_θ ist. Für dieses θ soll ein Konfidenzintervall geschätzt werden
- Ein Intervall $[G_u, G_o]$ mit der Eigenschaft

$$P_\theta(G_u \leq \theta \leq G_o) = 1 - \alpha$$

heißt *Konfidenzintervall* für den Parameter θ zum Konfidenzniveau $1 - \alpha$

- Übliche Konfidenzniveaus sind 90%, 95% und 99%
- G_u und G_o sind Zufallsvariable. Man bezeichnet sie als untere und obere *Vertrauensgrenze*

Quantile

- Φ die Verteilungsfunktion der Standardnormalverteilung
- Die Zahl q_α mit $\Phi(q_\alpha) = \alpha$ heißt α-*Quantil* der Standardnormalverteilung. Die wichtigsten Quantile der Standardnormalverteilung sind tabelliert

$\Phi(u)$	70%	80%	90%	95%	97.5%	99%	99.5%
u	0.524	0.842	1.282	1.645	1.960	2.326	2.576

- Z. B. ist das 0.975-Quantil gleich 1.960, also näherungsweise gleich 2
- Umrechnungsformel

$$q_\alpha = -q_{1-\alpha}$$

- Beispiel

$$q_{0.05} = -q_{0.95} = -1.645$$

Schätzung eines Konfidenzintervalls für den Erwartungswert bei bekannter Varianz mittels Quantilen

- $X_1, \ldots, X_n$ seien normalverteilt gemäß $N(\mu, \sigma^2)$ für bekanntes σ und unbekanntes μ
- Für μ soll ein Konfidenzintervall zum Konfidenzniveau $1 - \alpha$ geschätzt werden
- Benötigt wird das $1 - \frac{\alpha}{2}$-Quantil $q_{1-\alpha/2}$ der Standardnormalverteilung
- und das arithmetische Mittel $\overline{X}$ der Daten
- Das Konfidenzintervall ist

$$\left[\overline{X} - \frac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}}, \overline{X} + \frac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}}\right]$$

Beispiel: Roggenpflanzen

- Gesunde Roggenpflanzen einer bestimmten Art sind im Mittel 102.5 *cm* lang, wobei die Länge um 7 *cm* streut. Die Länge sei normalverteilt
- Durch Umwelteinflüsse änderte sich die mittlere Halmlänge, die Streuung aber nicht
- Die folgenden Längen wurden gemessen

96.62	94.91	85.05	101.61	109.55
93.05	97.86	96.66	95.08	98.87

- Arithmetisches Mittel der Daten

$$\overline{x} = 96.93$$

- Gesucht: Konfidenzintervall zum Konfidenzniveau 90%

Roggenpflanzen, Fortsetzung

- $1 - \alpha = 0.90$, also $\alpha = 0.10$, also $1 - \alpha/2 = 0.95$
- $q_{0.95} = 1.645$
- Konfidenzintervall

$$\left[\overline{X} - \frac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}}, \overline{X} + \frac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}}\right]$$
$$= \left[96.93 - \frac{7 \cdot 1.645}{\sqrt{10}}, 96.93 + \frac{7 \cdot 1.645}{\sqrt{10}}\right]$$
$$= [93.29, 100.6]$$

- Zum Vergleich: Konfidenzintervall zum Konfidenzniveau 95%

$$[92.50, 101.36]$$