Statistik und Wahrscheinlichkeitsrechnung

– Wintersemester 2019/20 –

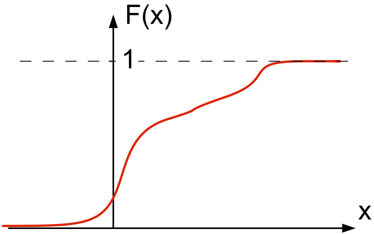
Kapitel 05: Spezielle Verteilungen

Prof. Dr. Adrian Ulges

Angewandte Informatik (B.Sc.) /
Informatik - Technische Systeme (B.Sc.) /
Wirtschaftsinformatik (B.Sc.)

Fachbereich DCSM
Hochschule RheinMain

Von Verteilungen zu Verteilungstypen Bild: [6]



- Verteilungsfunktionen F verraten uns wie Zufallsvariablen verteilt sind. *Aber: Wie ermitteln wir F?*
- Annahme: F gehört zu einem bestimmten Verteilungstyp.
- Im folgenden Kapitel werden wir uns mit einigen wichtigen Verteilungstypen auseinandersetzen, jeder mit anderen *Anwendungen, Parametern und Eigenschaften*.

Outline

1. Verteilungen für diskrete Zufallsvariablen
   - Die Bernoulli-Verteilung und Binomialverteilung
   - Die hypergeometrische Verteilung
   - Die Poisson-Verteilung

2. Verteilungen für stetige Zufallsvariablen
   - Die Uniformverteilung
   - Die Exponentialverteilung
   - Die Gauß'sche Normalverteilung

Outline

Die Bernoulli-Verteilung Bild: [8]

Definition (Bernoulli-Variable)

Wir bezeichnen einen Zufallsvariable mit nur zwei möglichen Realisierungen (üblicher Weise 0 und 1) als **Bernoulli-Variable**.

Anmerkungen

- Wir bezeichnen das Ereignis $X = 1$ als "Erfolg" und $X = 0$ als "Misserfolg".

Beispiele

- Münzwurf
- Defekt von Bauteilen
- ...

Die Bernoulli-Verteilung: Eigenschaften

Parameter

- Die Erfolgswahrscheinlichkeit $p := P(X = 1)$.

Erwartungswert und Varianz

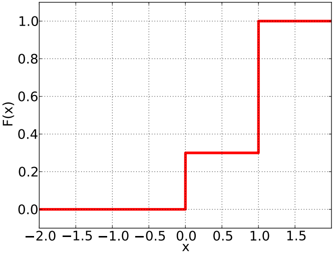
$$= (1-p)\cdot 0 + p \cdot 1 = p$$

$$E(X) = p = \sum_i p_i x_i \qquad \text{// Beweis: Übung}$$

$$Var(X) = p \cdot (1 - p)$$

Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{falls } x < 0 \\ 1 - p & \text{falls } 0 \le x < 1 \\ 1 & \text{falls } 1 \le x \end{cases}$$



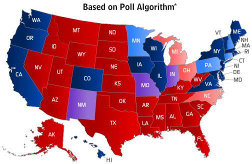
Die Binomial-Verteilung Bilder: [2] [4]

- Aus der Bernoulli-Verteilung lässt sich eine der wichtigsten Standard-Verteilungen herleiten, die Binomial-Verteilung.
- Wir wiederholen ein Bernoulli-Experiment n-mal und erhalten n unabhängige Bernoulli-Variablen $X_1, \dots, X_n$.
- Die Verteilung der Bernoulli-Variablen sei *stationär*, d.h. sie ändert sich zwischen den Wiederholungen nicht:

$$P(X_1 = 1) = P(X_2 = 1) = \dots = P(X_n = 1) = p$$

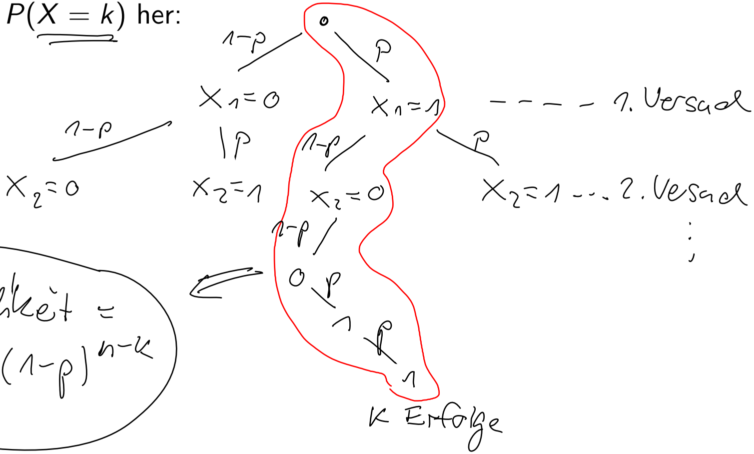
- Wir definieren eine neue Zufallsvariable X, die die Anzahl der erfolgreichen Versuche misst:

$$X := X_1 + X_2 + \dots + X_n$$



Die Binomial-Verteilung: Herleitung

Wir leiten $P(\underline{X = k})$ her:



Wahrschkeit = $p^k \cdot (1-p)^{n-k}$

1. Wahrscheinlichkeit einer konkreten Variation, z.B. $\underline{(1,0,0,1,0)}$:
   (Pfad)

Die Binomial-Verteilung: Herleitung

2. Wieviele Variationen mit k erfolgreichen Versuchen gibt es?

(Pferde)

$$\overbrace{\underbrace{\otimes\ \otimes\ \otimes}_{k\ \text{Erfolge}}\ \bigcirc\ \ldots\ \bigcirc}^{n\ \text{Kugeln / Versuche}}$$

$$\boxed{\underbrace{\frac{n!}{k! \times (n-k)!}}_{} = \binom{n}{k}}$$

Die Binomial-Verteilung

Definition (Binomial-Verteilung)

Eine diskrete Zufallsvariable X mit der Verteilung

$$P(X = k) = \begin{cases} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} & \textit{falls } 0 \leq k \leq n \\ 0 & \textit{sonst} \end{cases}$$

nennen wir binomial-verteilt.

Die Binomial-Verteilung: Do-it-yourself

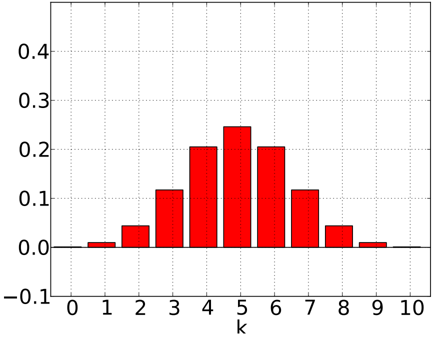
Berechnen Sie die Binomialverteilung für $n = 3$ und $p = \frac{3}{4}$.
Skizzieren Sie die zugehörige Verteilung als Säulendiagramm.

k	0	1	2	3
$\binom{n}{k}$				
$P(X = k)$				

Die Binomial-Verteilung: Beispiel ($n = 10, p = 0.5$)

$$P(X = k) = \binom{10}{k} \cdot 0.5^k \cdot 0.5^{n-k}$$

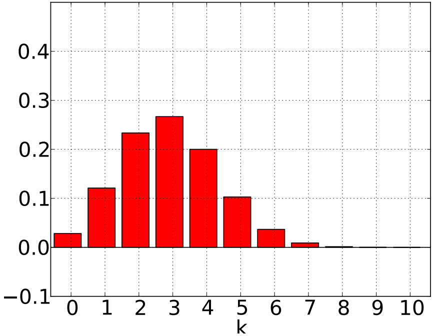
k	0	1	2	3	4	5	6	7	8	9	10
$\binom{n}{k}$	1	10	45	120	210	252	210	120	45	10	1
P(X=k)	0	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0



Die Binomial-Verteilung: Beispiel ($n = 10, p = 0.3$)

$$\boxed{P(X=k) = \binom{10}{k} \cdot 0.3^k \cdot 0.7^{n-k}}$$

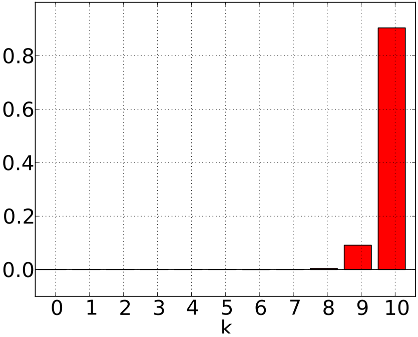
k	0	1	2	3	4	5	6	7	8	9	10
$\binom{n}{k}$	1	10	45	120	210	252	210	120	45	10	1
P(X=k)	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0	0	0



Die Binomial-Verteilung: Beispiel ($n = 10, p = 0.99$)

$$\boxed{P(X = k) = \binom{10}{k} \cdot 0.99^k \cdot 0.01^{n-k}}$$

k	0	1	2	3	4	5	6	7	8	9	10
$\binom{n}{k}$	1	10	45	120	210	252	210	120	45	10	1
P(X=k)	0	0	0	0	0	0	0	0	0.004	0.091	0.904



Anwendungsbeispiel: Service Level Agreement

Die tägliche Ausfallwahrscheinlichkeit eines Servers betrage $p = 1\%$. Berechnen Sie die Wahrscheinlichkeit, dasss der Server in einem Monat häufiger als 2 mal ausfällt.

Wir definieren X als die Anzahl der Ausfälle in einem Monat.

$$
\begin{aligned}
P(X > 2) &= 1 - P(X \leq 2) \\
&= 1 - \Big(P(X = 0) + P(X = 1) + P(X = 2)\Big) \\
&= 1 - \underbrace{\binom{30}{0} \cdot 0.01^0 \cdot 0.99^{30}}_{=P(X=0)} - \underbrace{\binom{30}{1} \cdot 0.01^1 \cdot 0.99^{29}}_{P(X=1)} \\
&\qquad - \underbrace{\binom{30}{2} \cdot 0.01^2 \cdot 0.99^{28}}_{P(X=2)} \\
&\approx 0.33\%
\end{aligned}
$$

Die Binomial-Verteilung: Eigenschaften

Parameter

- n: Anzahl der Wiederholungen
- p: Erfolgswahrscheinlichkeit eines Versuchs.

Notation und Eigenschaften

- Ist eine Zufallsvariable X binomialverteilt, schreiben wir auch "$X \sim B(n, p)$".
- Nützliche Rekursionsformel (für $p < 1$):

$$P(X = k+1) = P(X=k) \cdot \left(\frac{p}{1-p}\right) \cdot \left(\frac{n-k}{k+1}\right) \quad \text{für } k = 0, \ldots, n-1$$

Die Binomial-Verteilung: Eigenschaften

Erwartungswert $B(\cancel{n}, n=10, p=0{,}2) \Rightarrow E(X)=$

$$E(X) = E(X_1 + X_2 + \ldots + X_n) \quad \text{// Rechenregeln}$$
$$= E(X_1) + E(X_2) + \ldots + E(X_n) \quad \text{// Bernoulli-Variablen}$$
$$= p + p + \ldots + p$$
$$= n \cdot p$$

Varianz

$$Var(X) = Var(X_1 + X_2 + \ldots + X_n)$$

weil unabhängig

$$= Var(X_1) + Var(X_2) + \ldots + Var(X_n)$$
$$= p\cdot(1-p) + p\cdot(1-p) + \ldots + p\cdot(1-p)$$
$$= n \cdot p \cdot (1-p)$$

Binomialverteilung: Beispiel mit Karten Bild: [5]

- Wir ziehen 5 Karten aus einem Stapel von 32 Karten mit Zurücklegen.
- Was ist die Wahrscheinlichkeit, 3 Kreuz-Karten zu ziehen?
- X := Anzahl der (gezogenen) Kreuzkarten.
- $X \sim B(n{=}5, p{=}\frac{1}{4})$.

$$\boxed{\begin{aligned} P(X{=}3) &= \binom{5}{3} \cdot p^3 \cdot (1-p)^2 \\ &= \frac{5 \cdot 4 \cdot 3}{1 \cdot 2 \cdot 3} \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^2 \\ &\approx 8.7\% \end{aligned}}$$

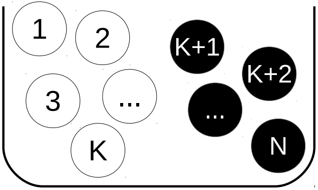
Binomialverteilung: Beispiel mit Karten II Bild: [5]

- Wir wiederholen das Experiment, ziehen aber ohne Zurücklegen!
- Können wir immer noch die Binomialverteilung verwenden?
  - Nein! Die Wahrscheinlichkeit eine Kreuz-Karte zu ziehen ändert sich mit jedem Versuch!
  - Wir können keine konstante Erfolgswahrscheinlichkeit p annehmen!
- Für diese Situationen verwenden wir die **hypergeometrische Verteilung** (im Folgenden).

Outline

Hypergeometrische Vertlg.: Herleitung (Urnenmodell)

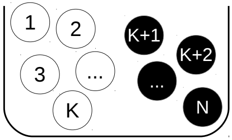
Wir leiten die gesuchte Verteilung mit Hilfe des Urnenmodells her.



- Analogie: Lotto-Spielen *(wieviele der "Erfolgs"-Kugeln werden gezogen?).*
- N Kugeln insgesamt,
  K weiße ("Erfolg"),
  $N-K$ schwarze ("Misserfolg").
- Wir ziehen n Kugeln ohne Zurücklegen und ohne Berücksichtigung der Reihenfolge.
- Gesucht: $P(X = k)$
  ("von den n gezogenen Kugeln sind k Stück weiß").

$$P(X = k) = \frac{\text{Anzahl der Kombinationen mit } k \text{ weißen und } n-k \text{ schwarzen Kugeln}}{\text{Anzahl aller möglichen Kombinationen}}$$

Die hypergeometrische Verteilung: Herleitung

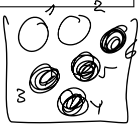


- N = Anzahl aller Kugeln
- K = Anzahl aller weißen Kugeln
- $N - K$ = Anzahl aller schwarzen Kugeln

- n = Anzahl der gezogenen Kugeln
- k = Anzahl der gezogenen weißen Kugeln
- $n - k$ = Anzahl der gezogenen schwarzen Kugeln

Möglichkeiten mit $n=3$ Kugeln und $k=1$ weißer Kugel?

$N=6$



$n=3$

Nenner

- Wir betrachten zuerst den Nenner (alle möglichen Kombinationen): Standardformel für Ziehen *ohne Zurücklegen* und *ohne Reihenfolge*:

$$C(N; n) = \binom{N}{n}$$

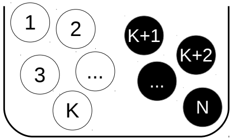
$\{1,2,3\}, \{1,2,4\}, \ldots \{4,5,6\}$

$\binom{6}{3}$

$\binom{2}{1}\cdot\binom{4}{2}$

$$\boxed{\begin{matrix}1\\2\end{matrix}} \times \begin{pmatrix}3 & 4\\3 & 5\\3 & 6\\4 & 5\\4 & 6\\5 & 6\end{pmatrix} = 2\cdot 6 = 12$$

Die hypergeometrische Verteilung: Herleitung



- N = Anzahl aller Kugeln
- K = Anzahl aller weißen Kugeln
- $N - K$ = Anzahl aller schwarzen Kugeln

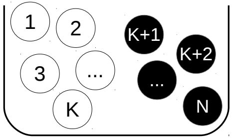
- n = Anzahl der gezogenen Kugeln
- k = Anzahl der gezogenen weißen Kugeln
- $n - k$ = Anzahl der gezogenen schwarzen Kugeln

Zähler

- Wieviele Kombinationen mit k (aus K) weißen Kugeln und $n - k$ (aus $N - K$) schwarzen Kugeln gibt es?
- Für die weißen Kugeln gibt es $C(K; k)$ Möglichkeiten (ohne Zurücklegen und Reihenfolge):

$$C(K; k) = \binom{K}{k}$$

Die hypergeometrische Verteilung: Herleitung



- N = Anzahl aller Kugeln
- K = Anzahl aller weißen Kugeln
- $N - K$ = Anzahl aller schwarzen Kugeln

- n = Anzahl der gezogenen Kugeln
- k = Anzahl der gezogenen weißen Kugeln
- $n - k$ = Anzahl der gezogenen schwarzen Kugeln

Zähler

- Für die schwarzen Kugeln gibt es $C(N-K; n-k)$ Möglichkeiten, d.h.

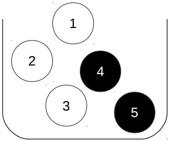
$$C(N-K; n-k) = \binom{N-K}{n-k}$$

- Insgesamt lautet die Anzahl der Möglichkeiten im Zähler:

$$C(K; k) \cdot C(N-K; n-k) = \binom{K}{k} \cdot \binom{N-K}{n-k}$$

Die hypergeometrische Verteilung: Beispiel

- Wir ziehen aus der abgebildeten Urne 3 Kugeln
- Gesuchtes Ereignis: 2 weiße Kugeln



Weiße Kombinationen (2 aus 3) $\binom{3}{2}$ x Schwarze Kombinationen (1 aus 2) $\binom{2}{1}$ = Kombinationen Mit 2 weißen *und* 1 schwarzen $\binom{3}{2} \cdot \binom{2}{1}$

1 2, 1 3, 2 3 x 4, 5 = 1 2 4, 1 2 5, 1 3 4, 1 3 5, 2 3 4, 2 3 5

$$P(X=2) = \frac{\text{1 2 4, 1 2 5, 1 3 4, 1 3 5, 2 3 4, 2 3 5}}{\text{1 2 3, 1 2 4, 1 2 5, 1 3 4, 1 3 5, 1 4 5, 2 3 4, 2 3 5, 2 4 5, 3 4 5}} = \frac{6}{10}$$

Alle Kombinationen (3 aus 5) $\binom{5}{3}$

Die hypergeometrische Verteilung: Herleitung

Definition (Hypergeometrische Verteilung)

Eine diskrete Zufallsvariable X mit der Verteilung

$$P(X = k) = \begin{cases} \dfrac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} & \text{falls } 0 \leq k \leq n \text{ und } k \leq K \\ 0 & \text{sonst} \end{cases}$$

nennen wir hypergeometrisch verteilt.

Parameter

- $N > 0$: Gesamtzahl der Kugeln
- $K \geq 0$: Gesamtzahl der weißen Kugeln
- $n > 0$: Anzahl der gezogenen Kugeln

Hypergeometrische vs. Binomialverteilung

- In der hypergeometrischen Verteilung ändert sich mit jedem Zug die Erfolgswahrscheinlichkeit p.
- Wir vergleichen die Binomialverteilung (mit Zurücklegen) mit der hypergeometrischen Verteilung (ohne Zurücklegen):

	Binomialverteilung	hypergeometrische Verteilung
1. Zug	K/N	K/N
2. Zug	K/N	K/(N-1) oder (K-1)/(N-1)
3. Zug	K/N	K/(N-2) oder (K-1)/(N-2) oder (K-2)/(N-2)
...	...	...

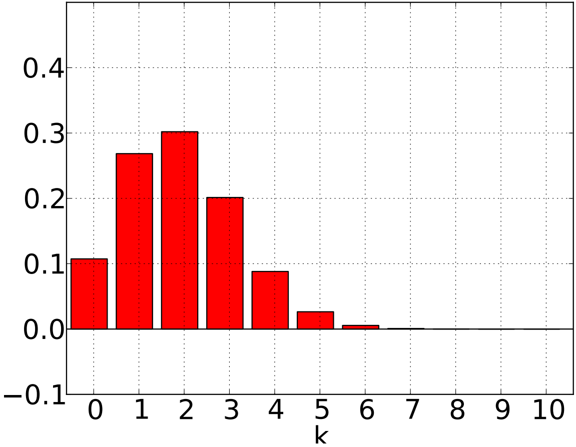
- Für $N >> n$ sind die Unterschiede gering, d.h.

$$p_1 \approx p_2 \approx ... \approx p_n \approx p.$$

- In diesem Fall können wir die hypergeometrische Verteilung durch die Binomialverteilung annähern!
- Faustregel: Die Approximation ist sinnvoll, wenn $n < 0.05 \cdot N$.

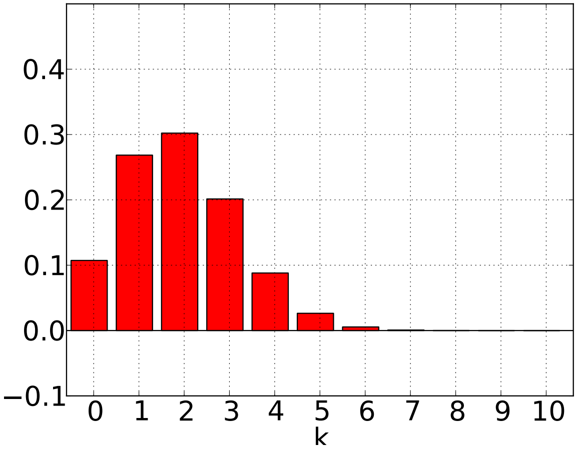
Hypergeometrische vs. Binomialverteilung

- Die Binomialverteilung mit $n = 10$, $p = 0.2$



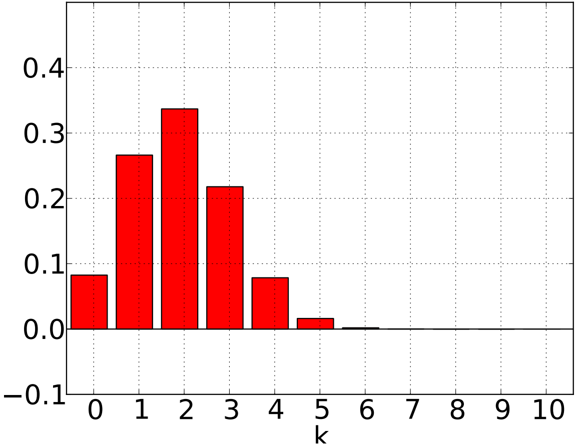
Hypergeometrische vs. Binomialverteilung

- Wir reduzieren schrittweise die Kugeln in der Urne...



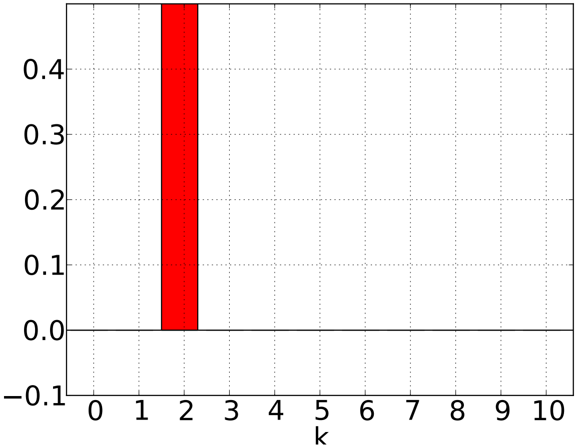
Hypergeometrische vs. Binomialverteilung

- $N = 50, K = 10$

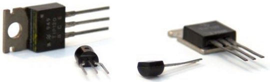


Hypergeometrische vs. Binomialverteilung

- $N = 10, K = 2$



Hypergeometrisch: Do-it-yourself Bild: [3]



Eine Lieferung aus 100 Transistoren enthält 5 defekte. In einer Kontrolle werden 4 Transistoren geprüft.

$N = 100,\ n = 4,\ K = 5$

- Wie lauten N, n, und K ?
- Berechnen Sie die Wahrscheinlichkeit, dass genau ein geprüfter Transistor defekt ist.
- Wie lautet die Wahrscheinlichkeit, wenn wir sie mit der Binomialverteilung approximieren?

$$P(X=1) = \frac{\binom{5}{1}\cdot\binom{95}{3}}{\binom{100}{4}} \qquad 1+3 = 4$$

$$= \frac{5 \cdot \frac{95 \cdot 94 \cdot 93}{\cancel{1} \cdot \cancel{2} \cdot \cancel{3}}}{\frac{100 \cdot 99 \cdot 98 \cdot 97}{\cancel{1} \cdot \cancel{2} \cdot \cancel{3} \cdot 4}} \approx 17{,}65\,\%$$

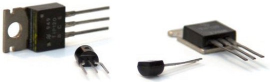
Binomial: $n = 4$, $p = 0{,}05$

$$P(X=1) = \binom{4}{1} \cdot 0{,}05^1 \cdot 0{,}95^3$$

$$\approx 17{,}15\,\%$$

$\approx$ ü

Hypergeometrisch: Do-it-yourself



Outline

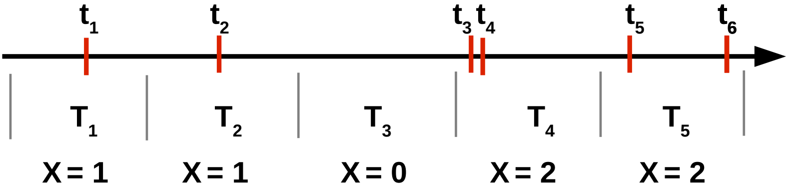
1. Verteilungen für diskrete Zufallsvariablen
   - Die Bernoulli-Verteilung und Binomialverteilung
   - Die hypergeometrische Verteilung
   - Die Poisson-Verteilung

2. Verteilungen für stetige Zufallsvariablen
   - Die Uniformverteilung
   - Die Exponentialverteilung
   - Die Gauß'sche Normalverteilung

Die Poisson-Verteilung: Anwendungsszenarien Bild: [1]

- Wir betrachten gleichartige Ereignisse, die unabhängig voneinander zu unregelmäßigen Zeitpunkten auftreten:
  - Kunden die ein Kaufhaus betreten
  - Jobs die in eine Warteschlange eingefügt werden
  - Autounfälle
  - Server-Ausfälle
  - Radioaktiver Zerfall von Atomen
  - ...
- Wir beobachten einen festen Zeitraum und möchten wissen, wieviele Ereignisse zu erwarten sind.

Die Poisson-Verteilung: Formalisierung



Es sei X die Anzahl der Ereignisse über den gegebenen Zeitraum.

Definition (Die Poisson-Verteilung)

Gegeben $\lambda \in \mathbb{R}$, nennen wir eine diskrete Zufallsvariable X mit

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad \text{für } k = 0, 1, 2, \ldots$$

Poisson-verteilt.

Die Poisson-Verteilung: Eigenschaften

Anmerkungen

- Die Formel für die Poisson-Verteilung lässt sich mit Hilfe der Binomialverteilung herleiten *(hier nicht)*.

Parameter

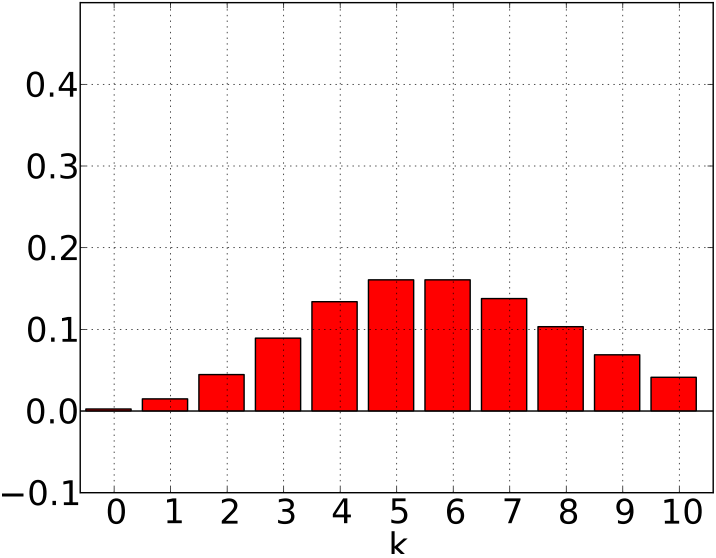
- $\lambda > 0$ (= die erwartete Anzahl an Ereignissen
  über den Beobachtungszeitraum)

Erwartungswert und Varianz

Ist X Poisson-verteilt (mit Parameter $\lambda > 0$), gilt:

$$E(X) = Var(X) = \lambda.$$

Die Poisson-Verteilung: Beispiel ($\lambda = 6$)



Do-Poisson-Yourself

Ein Fußballfeld hat eine Fläche von 0.75 ha. Pro Quadratkilometer schlägt im Schnitt alle zwei Jahre ein Blitz ein.
Was ist die Wahrscheinlichkeit, dass binnen eines Monats der Blitz häufiger als einmal auf dem Fußballfeld einschlägt?

λ? ⇒ wieviele Blitze pro Fußballfeld pro Monat im Durchschnitt?

0,5 Blitze pro Jahr pro km²

$0{,}5/12$ " " Monat " "

$\underbrace{0{,}5/12 \cdot \frac{0{,}75}{100}}_{\lambda = 0{,}0003125}$ " " " " Fußballfeld.

$$P(X>1) = 1 - \big(P(X=0) + P(X=1)\big)$$
$$= 1 - \left(\frac{\lambda^0}{0!} \cdot e^{-\lambda} + \frac{\lambda^1}{1!} \cdot e^{-\lambda}\right)$$
$$\approx 4{,}88 \cdot 10^{-8}$$

Do-Poisson-Yourself

Outline

Outline

Die Uniformverteilung

Die 'einfachste' stetige Zufallsvariable: Alle Werte innerhalb eines Intervalls $[a, b]$ sind gleich wahrscheinlich.

Definition (Uniformverteilung)

Es seien $a, b \in \mathbb{R}$ mit $a < b$. Dann nennen wir eine stetige Zufallsvariable X mit der Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{falls } x \in [a, b] \\ 0 & \text{sonst} \end{cases}$$

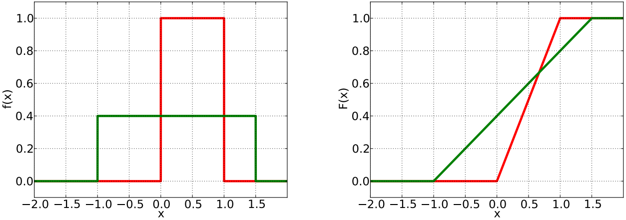
und der Verteilungsfunktion

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{falls } x < a \\ \frac{x-a}{b-a} & \text{falls } x \in [a, b] \\ 1 & \text{falls } x > b \end{cases}$$

uniformverteilt.

Die Uniformverteilung

Illustration



Parameter

- Intervallgrenzen a, b (mit $a < b$)

Anwendungsbeispiele

- Regentropfen die auf ein Einheitsquadrat fallen.
- Zeitpunkt von HTTP Requests an einen Server
  (alle Sekunden in einer Minute sind gleich wahrscheinlich).
- Die Wartezeit auf die nächste U-Bahn.

Die Uniformverteilung: Eigenschaften

Erwartungswert und Varianz

$$E(X) = \frac{a+b}{2}$$
$$Var(X) = \frac{1}{12}(b-a)^2$$

Outline

Die Exponentialverteilung

Definition (Die Exponentialverteilung)

Eine stetige Zufallsvariable X mit der Dichtefunktion

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \textit{falls } x \geq 0 \\ 0 & \textit{sonst} \end{cases}$$

(mit $\lambda > 0$) und der zugehörigen Verteilungsfunktion

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & \textit{falls } x \geq 0 \\ 0 & \textit{sonst} \end{cases}$$

nennen wir exponentialverteilt.

Anmerkungen

- Bei Zuverlässigkeitsuntersuchungen wird die Funktionsdauer eines Bauteils/Systems meist als eine exponentialverteilte Zufallsvariable modelliert *(siehe unser Beispiel "Festplatte")*.

Die Exponentialverteilung: Eigenschaften

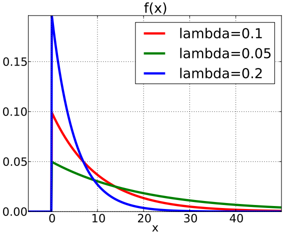
Parameter

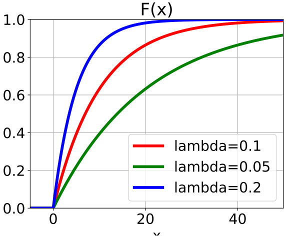
- Einziger Parameter: $\lambda > 0$

Erwartungswert und Varianz

- Erwartungswert: $E(X) = \frac{1}{\lambda}$ // haben wir schon in Kapitel 4 berechnet
- Varianz: $Var(X) = \frac{1}{\lambda^2}$

Dichte- und Verteilungsfunktion

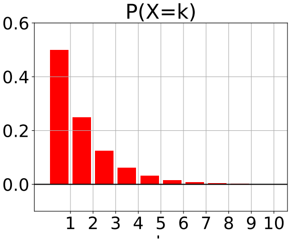


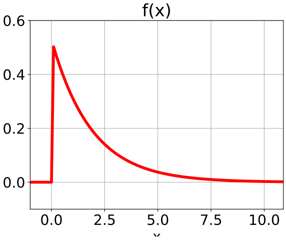


Die Exponentialverteilung: Veranschaulichung

Annäherung durch diskrete Zufallsvariable

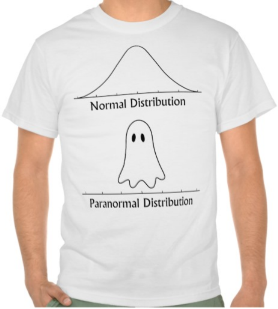
- Pro Zeiteinheit (z.B. jeden Tag) ein Bernoulli-Experiment: Bauteil fällt mit Wahrscheinlichkeit $1 - p$ aus (Abbruch), ansonsten weiterer Betrieb.
- $X :=$ Anzahl der Tage bis zum Ausfall
- $P(X = k) = p^{k-1} \cdot (1 - p) \rightarrow$ fällt exponentiell mit k ab!
- Im Kontinuierlichen wird hieraus $f(x) \propto e^{-\lambda x}$.





Outline

Die Normalverteilung: Bedeutung



Die Normalverteilung ist die Verteilung

- Körpergröße, Gewicht (je Geschlecht)
- Reisezeit (auf fester Strecke)
- Größe von Schrauben
- Gewicht von Kartoffeln, Schokotafeln, ...
- Messfehler, Pixelwerte in Bildern
- IQ
- ...

Warum ist die Normalverteilung so weit verbreitet?

- Eine mathematische Begründung liefert der Zentrale Grenzwertsatz.
- Umgangssprachlich: "Mitteln wir viele Zufallsvariablen, so konvergiert die Verteilung des Mittelwerts gegen eine Normalverteilung."

Die Gauß'sche Normalverteilung

Definition (Gauß'sche Normalverteilung)

Eine stetige Zufallsvariable X mit der Dichtefunktion

$$\varphi(x;\mu,\sigma^2) := f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(wobei $\mu \in \mathbb{R}$ und $\sigma^2 > 0$) und der Verteilungsfunktion

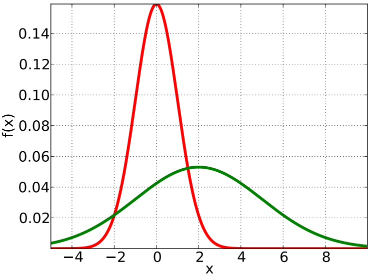
$$\mathcal{N}(x;\mu,\sigma^2) := F(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

nennen wir normalverteilt.

Anmerkungen

- μ entspricht dem Erwartungswert: $E(X) = \mu$
- σ^2 entspricht der Varianz: $Var(X) = \sigma^2$
- Für $\mu = 0$ und $\sigma^2 = 1$ nennen wir X standardnormalverteilt. Statt $\mathcal{N}(x;0,1)$ schreiben wir einfach nur $\mathcal{N}(x)$.

Die Gauß'sche Normalverteilung: Eigenschaften



- Die Gestalt des Graphen von φ erinnert an eine Glocke. Deshalb sprechen wir auch von der *Gauß'schen Glockenkurve* oder Gauß-Glocke (engl. *"bell curve"*).
- Die Dichtefunktion φ ist symmetrisch zu $x = \mu$.
- φ besitzt das einzige lokale Maximum bei $x = \mu$.
- Die einzigen beiden Wendepunkte von φ liegen bei $\mu \pm \sigma$.
- Das **Beispiel** zeigt: $\varphi(x; 2, 3^2)$ und $\varphi(x; 0, 1^2)$ *(die Standardnormalverteilung)*.

Die Gauß'sche Normalverteilung: Berechnung

Wie berechnen wir die Wahrscheinlichkeit $\mathbf{P(a \leq X \leq b)}$
(z.B. dass die Körpergröße zwischen 170 und 190 cm liegt)?

Unser bekannter Ansatz

$$
\begin{aligned}
P(a \leq X \leq b) &= F(b) - F(a) \\
&= \int_a^b \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}}}_{\varphi(t;\mu,\sigma^2)} \, dt
\end{aligned}
$$

Schlüsselproblem: Nicht-Integrierbarkeit der Normalverteilung

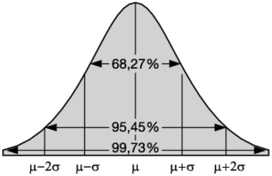
- Wir können für die Stammfunktion von φ keine geschlossene Form bestimmen!
- **Übung**: *Versuchen Sie die Stammfunktion von e^{-t^2} zu finden.*

Die Gauß'sche Normalverteilung: Berechnung

- Wir approximieren die gesuchte Wahrscheinlichkeit $P(a \leq X \leq b)$ mittels numerischer Integration.
- Techniken: Rechteckregel, Trapezregel, Simpson-Regel, Monte-Carlo-Integration, ...
- Die resultierenden Werte speichern wir in Tabellen und lesen sie bei Bedarf ab.

"Faustregel" Bild: [9]

- Es gibt Faustregeln für die Bereiche $[\mu \pm k \cdot \sigma]$
- *Beispiel:* $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68.27\%$



Die Gauß'sche Normalverteilung: Berechnung Bild: [7]

Genereller Ansatz: Standardisierung + Wertetabelle

Wir **standardisieren** die Zufallsvariable X
und lesen $\mathcal{N}$ aus einer Tabelle ab. *Wird im Folgenden erklärt.*

u	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5639	0,5675	0,5714	0,5754
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7258	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7518	0,7549
0,7	0,7580	0,7612	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7996	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8398
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545

Standardisierung normalverteilter Variablen

Satz (Standardisierung normalverteilter Variablen)

Es sei $X \sim \mathcal{N}(x; \mu, \sigma^2)$ eine normalverteilte Zufallsvariable mit Erwartungswert μ und Varianz σ^2. Dann ist

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

eine standardnormalverteilte Zufallsvariable.

Anmerkung

- Mit Hilfe dieses Satzes verwandeln wir X in eine standardnormalverteilte Variable Y. Wir bezeichnen diesen Prozess als **Standardisierung**.
- Durch die Standardisierung brauchen wir keine Tabelle für jedes μ und σ: Eine einzige Tabelle für die Standardnormalverteilung reicht!

Standardisierung normalverteilter Variablen

Beweis

Standardisierung normalverteilter Variablen

Standardisierung normalverteilter Variablen

Wie wir $P(a \leq X \leq b)$ berechnen

1. Standardisiere X, d.h. definiere $Y \sim \mathcal{N}(0,1)$ als

$$Y := (X - \mu)/\sigma$$

2. Berechne die Grenzen a', b':

$$a' := (a - \mu)/\sigma$$
$$b' := (b - \mu)/\sigma$$

3. Es gilt: $P(a \leq X \leq b) = P(a' \leq Y \leq b') = \mathcal{N}(b') - \mathcal{N}(a')$
4. Lese $\mathcal{N}(a')$ und $\mathcal{N}(b')$ aus Tabellen der Standardnormalverteilung ab *(interpoliere ggfs. zwischen den diskreten Werten der Tabelle, um die Genauigkeit zu erhöhen)*.

Beispiel: IQ

- Der IQ lässt sich als in der Gesamtbevölkerung normalverteilte Zufallsvariable X mit $\mu = 100$ und $\sigma = 15$ beschreiben.
- Bob hat einen IQ von 120, Charly einen IQ von 90.
- Wieviele Prozent der Bevölkerung besitzen einen IQ zwischen dem Charlys und Bobs?
- Wieviele Prozent besitzen einen höheren IQ als Bob?

Beispiel: IQ Bild: [7]

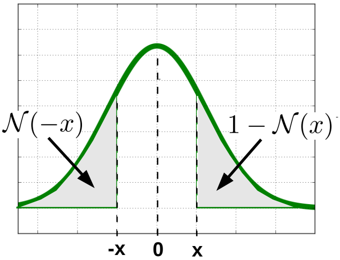
u	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5639	0,5675	0,5714	0,5754
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7258	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7518	0,7549
0,7	0,7580	0,7612	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7996	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8398
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545

Beispiel: IQ

Standardisierung: Anmerkungen

Symmetrie

- Warum fehlen in der Tabelle negative Werte ?
- Aufgrund der Symmetrie von φ gilt: $\mathcal{N}(-x) = 1 - \mathcal{N}(x)$. Wir können negative Werte also auch aus der Tabelle gewinnen.



Quantile

- Oft ist eine Wahrscheinlichkeit α gegeben, und wir suchen das α-Quantil, d.h. jenes x_α für das gilt $P(X \leq x_\alpha) = \alpha$
- Diese Werte sind ebenfalls in vorgefertigten Tabellen verfügbar.

α	x_α
0.001	-3.090
0.005	-2.576
0.01	-2.326
0.025	-1.960
0.05	-1.645
0.1	-1.282
0.9	1.282
0.95	1.645
0.975	1.960
0.99	2.326
0.995	2.576
0.999	3.090

Standardnormalverteilung: Do-it-yourself

Bei der Massenproduktion von Gewindeschrauben ist der gemessene Durchmesser (in mm) eine Zufallsvariable $X \sim \mathcal{N}(x; \mu = 10, \sigma^2 = 4)$.

- Eine Abweichung vom Sollwert ($= 10$ mm) von 3 mm kann gerade noch toleriert werden. Welcher Ausschuss-Anteil ist zu erwarten?
- Wie müsste die Standardabweichung lauten, damit der Ausschuss-Anteil nur 2% beträgt?

Standardnormalverteilung: Do-it-yourself

Standardnormalverteilung: Do-it-yourself

Praxisbeispiel: "Six Sigma"***

- Produktionsprozesse werden immer weiter perfektioniert – hierbei spielen statistische Verfahren eine große Rolle.
- Ziel ist in der Regel die Produktion "möglichst gleicher" Teile, d.h. die **Standardabweichung** der Artikel einer Produktionslinie soll möglichst gering sein.
- Unter dem Namen **"Six Sigma"** wurde eine Methode des Qualitätsmanagements bekannt, um die Standardabweichung σ soweit zu senken, dass nur noch Werte außerhalb eines **Toleranzbereiches von 6 Standardabweichungen** (6σ, daher der Name) als Ausschuss betrachtet werden müssen.
- Da in der Praxis die Lage des Mittelwerts driften kann, wird das Qualitätslimit in der Praxis von 6σ auf 4.5σ gesenkt.
- Wir suchen also: $P(|X-\mu| > 4.5\sigma)$

$$P(|X-\mu| > 4.5\sigma) = 1 - \Big(\mathcal{N}(4.5) - \mathcal{N}(-4.5)\Big) = 0.0000068$$

- Das bedeutet: Ziel sind höchstens **6.8 Defekte** unter **1 Million** produzierter Artikel.

References

[1] Dylan Meconis: Parent Line.
https://flic.kr/p/5R2khv (changed to black and white, CC license, retrieved: Dec 2016).

[2] Fairphone: Fairphone quality control check.
https://flic.kr/p/iQLoHE (retrieved: Dec 2016).

[3] Martin Bircher: Transistors.
https://flic.kr/p/94gcGL (retrieved: Dec 2016).

[4] Nationwide Poll Results for 2008 Presidential Election.
https://commons.wikimedia.org/wiki/File:Nationwide_Poll_Results_for_2008_Presidential_Election.svg (retrieved: Oct 2016).

[5] Olli Henze: 4 Asse.
http://www.ohenze.de/ (retrieved: Nov 2016).

[6] Sean Fallon: Statistical Distribution Plushies: Hug The Math.
http://nerdapproved.com/approved-products/statistical-distribution-plushies-hug-the-math/ (retrieved: Dec 2016).

[7] L. Papula.
Mathematik für Ingenieure und Naturwissenschaftler, volume 3.
Friedrich Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 4 edition, 2001.

[8] unknown.
Jakob Bernoulli.
https://commons.wikimedia.org/wiki/File:Jakob_Bernoulli.jpeg (retrieved: Dec 2016).

[9] Eggert Winter, editor.
Gabler Wirtschaftslexikon.
Gabler, 18. edition, 2013.