Stochastik (AIN)

Prof. Dr. Barbara Staehle

HTWG Konstanz
Fakultaet für Informatik

SS 2021

Teil III

Kontinuierliche Wahrscheinlichkeitstheorie

Teil III Kontinuierliche Wahrscheinlichkeitstheorie

Abschnitt 1

Zufallsvariablen

Diskrete und stetige Verteilungen I

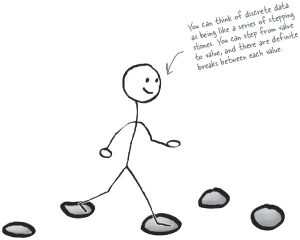


Bild: Modell einer diskreten Verteilung [Griffith, 2014]

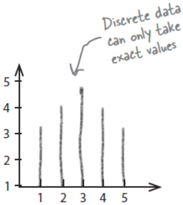


Bild: Histogramm einer diskreten Verteilung [Griffith, 2014]

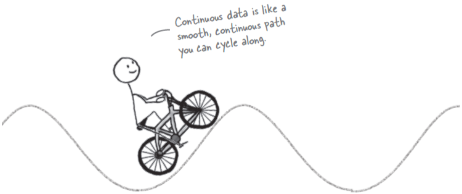


Bild: Modell einer stetigen (kontinuierlichen) Verteilung [Griffith, 2014]

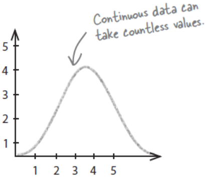


Bild: Histogramm einer stetigen (kontinuierlichen) Verteilung [Griffith, 2014]

Diskrete und stetige Verteilungen II

- Durch diskrete ZV lassen sich die Zufallsexperimente modellieren, die endlich viele oder abzählbar unendlich viele Werte annehmen können (Glücksspiele, Funktionalität von Bauteilen, Übertragungserfolg, Paketwiederholungen, ...).
- Für andere Experimente benötigt man aber die Variablen mit einem Kontinuum an möglichen Werten (Lebensdauer von Bauteilen, Übertragungszeit, Serverantwortzeiten, Messfehler, ...), dies kann nur durch stetige ZV modelliert werden.
- Eine Variable X heißt stetig oder kontinuierlich oder reell, falls zu beliebigen Werten $a < b$ auch jeder Zwischenwert aus dem reellen Intervall $[a, b]$ möglich ist.
- Im Gegensatz zu diskreten ZV können stetige ZV überabzählbar viele Werte annehmen.

Stetige Zufallsvariablen I

Erinnerung: Eine diskrete ZV X

- kann nur abzählbar viele Werte annehmen,
- hat eine Verteilung, die jedem Wert eine Wahrscheinlichkeit zuordnet und die Summe über alle Wahrscheinlichkeiten der Verteilung ist gleich 1.

Eine stetige Zufallsvariable X

- kann überabzählbar unendlich viele Werte annehmen,
- hat eine (Wahrscheinlichkeits-) Dichtefunktion (PDF), f mit:
  - f ist auf ganz $\mathbb{R}$ definiert ist,
  - für alle $x \in \mathbb{R}$ gilt $f(x) \geq 0$,
  - die Gesamtfläche unter dem Graphen von f ist gleich 1:

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1,$$

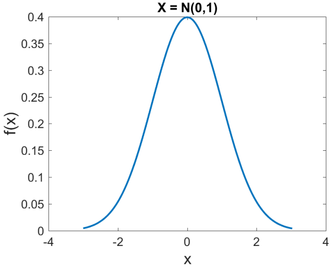


Bild: PDF der Standardnormalverteilung

Bemerkung: PDF ist kurz für "probability density function".

Stetige Zufallsvariablen II

Die (Wahrscheinlichkeits-)Verteilungsfunktion (CDF) einer stetigen ZV X erhält man als Integral der Dichtefunktion:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

- $F(x) = P(X \leq x)$ ist eine stetige und (mindestens stückweise) differenzierbare Funktion mit
  - $\lim_{x \to -\infty} F(x) = 0$
  - $\lim_{x \to \infty} F(x) = 1$
- Umgekehrt ist die Dichtefunktion die Ableitung der Verteilungsfunktion: $f(x) = F'(x)$.

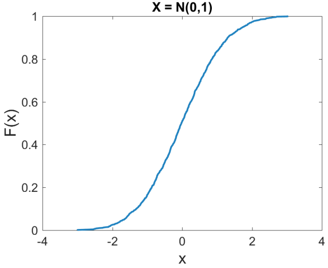


Bild: CDF einer Normalverteilung mit Mittelwert 0, Standardabweichung 1 (Standardnormalverteilung)

Bemerkung: CDF ist kurz für "cumulative distribution function".

Stetige Zufallsvariablen III

Wichtig: Für jede stetige ZV X gilt:

- Die Wahrscheinlichkeit, dass X genau einen Wert annimmt, ist 0:

$$\forall_{x \in \mathbb{R}} \; P(X = x) = 0.$$

- Wahrscheinlichkeit $\neq$ Dichte:

$$\forall_{x \in \mathbb{R}} \; P(X = x) \neq f(x)$$

- die Wahrscheinlichkeit, dass X einen Wert im Intervall $[a; b]$ (oder $(a; b]$ oder $[a; b)$ oder $(a; b)$) annimmt, ist

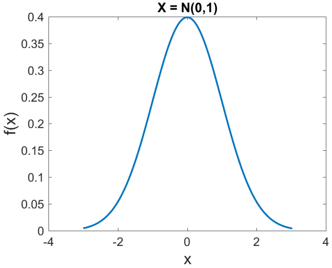


Bild: PDF der Standardnormalverteilung

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = F(b) - F(a) = \int_a^b f(t)\,\mathrm{d}t.$$

Bemerkung: Die Art der Intervallgrenzen ist unwichtig, da $P(X = x) = 0, \quad$ für alle $x \in \mathbb{R}$.

Beispiel: Exponentialverteilung I

Modelliert man ein Kommunikationssystem, z.B. ein Mobilfunknetz, nimmt man oft an, dass die Zwischenankunftszeit A, d.h. die Zeit die zwischen zwei Anforderungen (z.B. Anrufe, Datenpakete, SMS, ...) liegt, durch eine Verteilungsfunktion der Form

$$A(t) = \begin{cases} 1 - e^{-\lambda t} & \text{falls } t \geq 0, \\ 0 & \text{falls } t < 0. \end{cases}$$

beschrieben wird. Dabei ist λ eine Konstante, welche die Ankunftsrate, oder die Auslastung des Netzes beschreibt. Je größer λ, desto kleiner ist die mittlere Zeit zwischen zwei Anforderungen.

- Geben Sie die Dichtefunktion $a(t)$ an

$$a(t) = A'(t) = \begin{cases} \lambda e^{-\lambda t} & \text{falls } t \geq 0, \\ 0 & \text{falls } t < 0. \end{cases}$$

Beispiel: Exponentialverteilung II

- Für das Netz des Providers P_3 in Konstanz ist $\lambda = 0.5$ sec^{-1}. Wie groß ist die Wahrscheinlichkeit, dass zwischen zwei Ankünften
  - höchsten 1 Sekunde $P(A \leq 1) = A(1) = 1 - e^{-0.5} = 0.393 = 39.3\%$
  - zwischen 1 und 2 Sekunden $P(1 < A < 2) = A(2) - A(1) = 1 - e^{-1} - (1 - e^{-0.5}) = e^{-0.5} - e^{-1} = 0.239 = 23.9\%$
  - mehr als 2 Sekunden liegen? $P(A > 2) = 1 - P(A \leq 2) = 1 - A(2) = 1 - (1 - e^{-1}) = e^{-1} = 0.368 = 36.8\%$

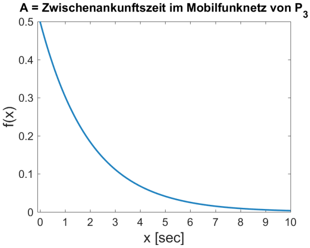


Bild: PDF der Zwischenankunftszeit

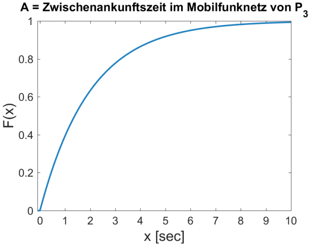
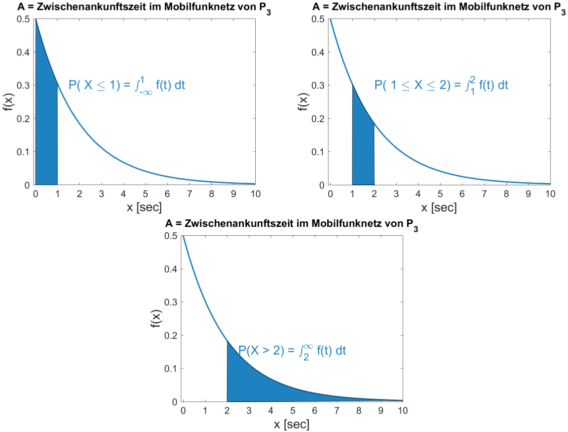


Bild: CDF der Zwischenankunftszeit

Beispiel: Exponentialverteilung III

Graphische Darstellung der berechneten Wahrscheinlichkeiten:



Beispiel: Gleichverteilung I

Angenommen, ein Bus fährt pünktlich alle 10 Minuten und ein Fahrgast hat als Wartezeit die ZV X, die kontinuierlich alle Werte von 0 bis 10 annehmen kann, wobei jede Wartezeit gleich wahrscheinlich ist.
Die zugehörige Wahrscheinlichkeitsdichte ist daher

$$f(x) = \begin{cases} k & \text{falls } 0 < x < 10, \\ 0 & \text{sonst }, \end{cases}$$

mit einer Konstanten $k \in \mathbb{R}$.

- Bestimmen Sie k. k muss so gewählt werden, dass die Fläche unter dem Graphen von f gleich 1 ist:

$$1 = \int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = \int_0^{10} k\,\mathrm{d}t = [kt]_0^{10} = 10k \quad \Rightarrow k = 0.1.$$

- Geben Sie die Verteilungsfunktion F an. Siehe nächste Folie.

Beispiel: Gleichverteilung II

- Geben Sie die Verteilungsfunktion F an.
  - Für $x \leq 0$ gilt $F(x) = 0$.
  - Für $0 < x < 10$ gilt: $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} 0.1\,\mathrm{d}t = 0.1x$
  - Für $x \geq 10$ gilt $F(x) = 1$.

Damit:

$$F(x) = \begin{cases} 0 & \text{falls } x \leq 0, \\ 0.1x & \text{falls } 0 < x < 10, \\ 1 & \text{falls } x \geq 10. \end{cases}$$

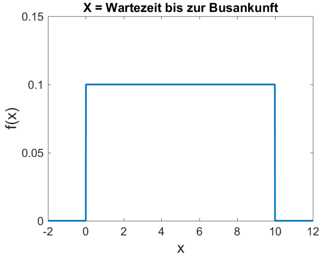


Bild: PDF der Wartezeit

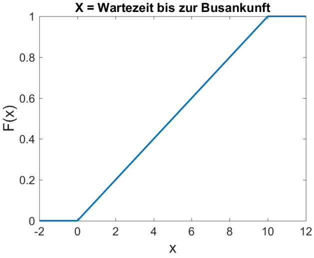


Bild: CDF der Wartezeit

Beispiel: Gleichverteilung III

- Mit welcher Wahrscheinlichkeit wartet man höchstens 3 Minuten?
  $P(X \leq 3) = F(3) = 0.1 \cdot 3 = 0.3 = 30\%$

- Mit welcher Wahrscheinlichkeit wartet man mindestens 2 Minuten?
  $P(X \geq 2) = 1 - P(X < 2) = 1 - F(2) = 1 - 0.1 \cdot 2 = 0.8 = 80\%$

- Mit welcher Wahrscheinlichkeit wartet man zwischen 5 und 9 Minuten?
  $P(5 \leq X \leq 9) = F(9) - F(5) = 0.1 \cdot 9 - 0.1 \cdot 5 = 0.4 = 40\%$

- Unter welchem Wert liegen 90% aller Wartezeiten?
  gesucht: t mit $P(X \leq t) = 0.9 \Rightarrow 0.1t = 0.9 \Rightarrow t = 9$ (Minuten)
  t nennt man auch 90%-Quantil.

Quantile

Definition

Gegeben sei eine (diskrete oder stetige) Zufallsvariable X und $p \in (0;1)$. $x_p \in \mathbb{R}$, für das

$$F(x_p) = p$$

gilt, heißt *p*-Quantil von X. Ein Quantil zu $p = 0.5$ heißt Median.

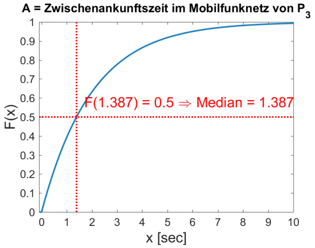


Bild: Bestimmung des Medians für eine stetige ZV

Bemerkungen:

- $p\%$ aller Werte die X annimmt sind kleiner als das $p\%$-Quantil.
- Für stetige ZV existiert für jedes $p \in (0;1)$ ein eindeutiges p-Quantil $x_p = F^{-1}(p)$.
- Für diskrete ZV muss ein Quantil nicht für jedes $p \in (0;1)$ existieren (Sprungstellen)!

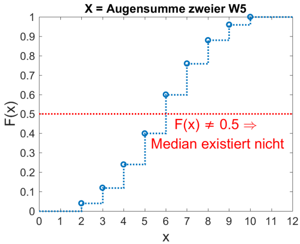


Bild: Für eine diskrete ZV existiert nicht jedes Quantil

Beispiele zum Mitdenken

Wir betrachten das Mobilfunknetz von P_3 dessen Zwischenankunftszeit A in Konstanz die folgende Verteilungsfunktion hat:

$$A(t) = \begin{cases} 1 - e^{-\frac{t}{2}} & \text{falls } t \geq 0, \\ 0 & \text{falls } t < 0. \end{cases}$$

- Nach welcher Zeit kommt mit 60%-iger Wahrscheinlichkeit spätestens wieder eine Anforderung im Netz von P_3 an?
  Lösung: Wir suchen $x_{0.6}$ mit
  $$F(x_{0.6}) = 1 - e^{-\frac{x_{0.6}}{2}} = 0.6 \Rightarrow x_{0.6} = -2 \cdot \ln(0.4) = 1.832$$
  Mit 60%iger Wahrscheinlichkeit kommt also spätestens nach 1.832 Sekunden wieder eine Anforderung im Netz an.
- Geben Sie den Median von A an.
  Lösung: Wir suchen $x_{0.5}$ mit
  $$F(x_{0.5}) = 1 - e^{-\frac{x_{0.5}}{2}} = 0.5 \Rightarrow x_{0.5} = -2 \cdot \ln(0.5) = 1.387$$
  Mit 50%iger Wahrscheinlichkeit kommt also spätestens nach 1.387 Sekunden wieder eine Anforderung im Netz an.

Erwartungswert und Varianz stetiger Zufallsvariablen

Erinnerung: Erwartungswert und Varianz einer diskreten ZV X:

$$\mathrm{E}(X) = \mu_X = \sum_{x_i \in T} x_i p_i \quad \text{und} \quad \mathrm{Var}(X) = \sum_{x_i \in T} (x_i - \mu_X)^2 p_i$$

Für eine stetige ZV ersetzt man die Summe durch ein Integral.

Definition

Für eine stetige Zufallsvariable X definiert man

- den Erwartungswert, $\mathrm{E}(X)$ oder μ_X als
$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x$$
- die Varianz, $\mathrm{Var}(X)$ oder σ_X^2 als
$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) \, \mathrm{d}x$$
- die Standardabweichung von X als $\sigma_X = \sqrt{\mathrm{Var}(X)}$

Beispiele zum Mitdenken

Betrachten Sie die Zufallsvariable X mit der Dichte

$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & x \leq 0 \end{cases}$$

Berechnen Sie für X Verteilungsfunktion, Erwartungswert, Varianz und Standardabweichung.

$$\begin{aligned}
F(x) &= \begin{cases} \int_{-\infty}^{x} f(t)\ \mathrm{d}t = \int_0^x 2t\ \mathrm{d}t = \left[t^2\right]_0^x = x^2 & \text{für } 0 \leq x \leq 1 \\ 0 & \text{sonst.} \end{cases} \\
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\ \mathrm{d}x = \int_0^1 x \cdot 2x\ \mathrm{d}x = \int_0^1 2x^2\ \mathrm{d}x \\
&= \left[\tfrac{2}{3}x^3\right]_0^1 = \tfrac{2}{3} - 0 = \tfrac{2}{3} \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \int_{-\infty}^{\infty} x^2 f(x)\ \mathrm{d}x - (\tfrac{2}{3})^2 = \int_0^1 x^2 \cdot 2x\ \mathrm{d}x - \tfrac{4}{9} \\
&= \int_0^1 2x^3\ \mathrm{d}x - \tfrac{4}{9} = \left[\tfrac{2}{4}x^4\right]_0^1 - \tfrac{4}{9} = \tfrac{1}{2} - 0 - \tfrac{4}{9} = \tfrac{9}{18} - \tfrac{8}{18} = \tfrac{1}{18} \\
\sigma_X &= \frac{1}{\sqrt{18}}
\end{aligned}$$

Diskrete und stetige Zufallsvariablen: Allgemeines

	diskrete ZV	**stetige ZV**
X	$X : \Omega \to \mathbb{R}, \quad \omega \mapsto X(\omega)$	
Ω	endlich oder abzählbar unendlich	überabzählbar unendlich
Verteilung	$p_i = P(X = x_i)$	-
Verteilungsdichte	-	$f(x) = F'(x)$ $P(X = x) = 0$
Verteilungsfunktion	$F(x) = P(X \leq x)$	
	$F(x) = \sum\limits_{i:x_i \leq x} p_i$	$F(x) = \int\limits_{-\infty}^{x} f(t)\mathrm{d}t$

Bemerkung: Manche Autoren betrachten auch für diskrete Zufallsvariablen die Verteilungsdichte. Diese ist überall 0, bis auf die Werte x_i aus dem Träger von X, an denen $f(x_i) = p_i$ ist.

Diskrete und stetige Zufallsvariablen: Kennwerte

	diskrete ZV	**stetige ZV**
X	$X : \Omega \to \mathbb{R}, \quad \omega \mapsto X(\omega)$	
p-Quantil $p \in (0;1)$	$x_p = F^{-1}(p)$, existiert für manche p-Werte	$x_p = F^{-1}(p)$, existiert für alle p-Werte
Erwartungswert	$\mathrm{E}(X) = \mu_X = \sum\limits_{x_i \in T} x_i p_i$	$\mathrm{E}(X) = \mu_X = \int\limits_{-\infty}^{\infty} x f(x)\, \mathrm{d}x$
Varianz	$\mathrm{Var}(X) = \sum\limits_{x_i \in T} (x_i - \mu_X)^2 p_i$	$\mathrm{Var}(X) = \int\limits_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\, \mathrm{d}x$
	$Var(X) = E(X^2) - E(X)^2$	
Standardabweichung	$\sigma_X = \sqrt{\mathrm{Var}(X)}$	

Diskrete und stetige Zufallsvariablen: Intervallwahrscheinlichkeiten

Für eine Zufallsvariable X berechnet sich die Intervallwahrscheinlichkeit, also die Wahrscheinlichkeit, dass X einen Wert in einem bestimmten Intervall annimmt, je nach Typ:

	allgemein	diskrete Zufallsvariable	stetige Zufallsvariable
$P(a < X \leq b)$	$F(b) - F(a)$	$\sum\limits_{i:a<x_i\leq b} p_i$	$\int\limits_a^b f(t)\mathrm{d}t$
$P(X \leq b)$	$F(b)$	$\sum\limits_{i:x_i\leq b} p_i$	$\int\limits_{-\infty}^b f(t)\mathrm{d}t$
$P(X > a)$	$1 - F(a)$	$\sum\limits_{i:x_i>a} p_i$	$\int\limits_a^\infty f(t)\mathrm{d}t$

Abschnitt 2

Wichtige stetige Verteilungen

Die Gleichverteilung

Bekanntes Beispiel: Ein Bus fährt pünktlich alle 10 Minuten. Ein Fahrgast hat als Wartezeit die ZV X, die mit gleicher Wahrscheinlichkeit alle reellen Werte von 0 bis 10 annehmen kann.

Modell:

- X ist gleichverteilt auf dem Intervall $[0, 10]$
- Verteilungsdichtefunktion: $f(x) = 0.1$
- Verteilungsfunktion: $F(X) = 0.1x$

Definition

Eine Zufallsvariable X, die alle Werte des rellen Intervalls $[a; b]$ mit der gleichen Wahrscheinlichkeit annehmen kann, heißt gleichverteilt (manchmal auch rechteckverteilt:) $X \sim U(a, b)$. Es gilt:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{sonst} \end{cases} \qquad F(x) = \begin{cases} 0 & \text{für } x < a \\ \frac{x-a}{b-a} & \text{für } a \leq x \leq b \\ 1 & \text{für } x > b \end{cases}$$

Fakten zur Gleichverteilung

Für eine gleichverteilte Zufallsvariable X, $X \sim U(a, b)$ gilt

Wertebereich $X \in [a, b]$

Parameter $a, b \in \mathbb{R}$: minimaler und maximaler Wert, den X annehmen kann

Verteilungsdichte $f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$

Verteilungsfunktion $F(x) = P(X \leq x) = \begin{cases} 0 & \text{für } x < a \\ \frac{x-a}{b-a} & \text{für } a \leq x \leq b \\ 1 & \text{für } x > b \end{cases}$

Erwartungswert $\mathrm{E}[X] = \frac{a+b}{2}$

Varianz $\mathrm{Var}[X] = \frac{(b-a)^2}{12}$

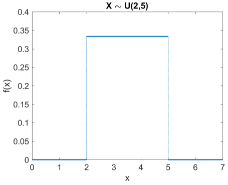


Bild: Verteilungsdichtefunktion

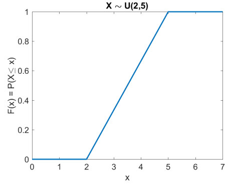


Bild: Verteilungsfunktion

Twitternutzung

Bemerkung: Die Gleichverteilung wird verwendet, wenn man die wahre Verteilung einer Zufallsvariable nicht kennt, oder wenn diese einen maximal zufälligen Wert annimmt (Zufallszahlengeneratoren verwenden defaultmäßig die Gleichverteilung).

Beispiel:

- Eve hat liest Meldungen in Ihrer Twitter-Timeline.
- Nicht alle Meldungen interessieren sie in gleichem Maße, daher ist die Zeit, die sie für das Lesen eines Beitrags (und ggf. der Kommentare) benötigt, eine völlig zufällige Zeit aus dem Intervall zwischen 5 Sekunden und 5 Minuten.
- Fragen:
  - Wie ist die Zufallsvariable verteilt, die dieses Szenario modelliert?
    Lösung: Verwende $X \sim U(5, 300)$ für die Lesezeit in Sekunden
  - Was ist die Wahrscheinlichkeit, dass Eve höchsten 1 Minute zum Lesen eines Tweets benötigt?
    Lösung: $P(X \leq 60) = F(60) = \frac{60-5}{300-5} = \frac{55}{295} = \frac{11}{59} = 0.1864$
  - Was ist die Wahrscheinlichkeit, dass Eve mehr als 1 Minute braucht?
    Lösung: $P(X > 60) = 1 - P(X \leq 60) = 0.8136$

Die Exponentialverteilung

Bekanntes Beispiel: Für das Netz des Mobilfunk-Providers P_3 in Konstanz folgt die Zwischenankunftszeit A, d.h. die Zeit die zwischen zwei Anforderungen (z.B. Anrufe, Datenpakete, SMS, ...) liegt, der Exponentialverteilung mit Parameter $\lambda = 0.5 \text{ sec}^{-1}$.
Die mittlere Zeit zwischen zwei Anforderungen beträgt $\frac{1}{\lambda} = \frac{1}{0.5} = 2$ sec.

Definition

Eine Zufallsvariable X, welche nur Werte > 0 mit Erwartungswert $\frac{1}{\lambda}$ annehmen kann und die Zeit zwischen zwei Ereignissen beschreibt, heißt exponentialverteilt: $X \sim exp(\lambda)$, falls gilt:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Bemerkung: Die Exponentialverteilung wird als Modell für die Dauer von zufälligen Zeitintervallen benutzt, wie z. B. Zeit zwischen zwei Anrufen, Lebensdauer von Bauteilen, Maschinen und Geräten, ...

Fakten zur Exponentialverteilung

Für eine exponentialverteilte Zufallsvariable X, $X \sim exp(\lambda)$ gilt

Wertebereich	$X \in \mathbb{R}_0^+$
Parameter	$\lambda \in \mathbb{R}^+$: Ankunftsrate der Ereignisse
Verteilungsdichte	$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$
Verteilungsfunktion	$P(X \leq x) = F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$
Erwartungswert	$\mathrm{E}[X] = \frac{1}{\lambda}$
Varianz	$\mathrm{Var}[X] = \frac{1}{\lambda^2}$

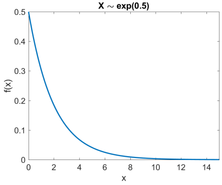


Bild: Verteilungsdichtefunktion

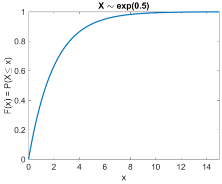


Bild: Verteilungsfunktion

Beispiele zum Mitdenken: Laufschuhe

- Ein Paar Laufschuhe hält im Mittel 18 Monate, falls diese täglich benutzt werden. Die Zeit, die Laufschuhe nutzbar sind, bevor sie kaputt gehen, ist eine exponentialverteilte Zufallsvariable.
- Fragen:
  - Wie ist die Zufallsvariable verteilt, welche die Haltbarkeit in Monaten von Laufschuhen bei täglicher Nutzung angibt?
    Lösung: Verwende $X \sim exp(\frac{1}{18})$
  - Mit welcher Wahrscheinlichkeit hält ein Paar täglich genutzte Laufschuhe länger als 15 Monate?
    Lösung: $P(X > 15) = 1 - F(15) = 1 - (1 - e^{-\frac{1}{18}\cdot 15}) = e^{-\frac{5}{6}} = 0.435$
  - Mit welcher Wahrscheinlichkeit hält ein Paar täglich genutzte Laufschuhe zwischen 12 und 15 Monaten?
    Lösung: $P(12 < X < 15) = F(15) - F(12) =$
    $(1 - e^{-\frac{5}{6}}) - (1 - e^{-\frac{1}{18}\cdot 12}) = e^{-\frac{2}{3}} - e^{-\frac{5}{6}} = 0.0788$
  - Wie lange halten 75% der täglich genutzten Laufschuhe höchstens?
    Lösung: Gesucht ist $x_{0.75}$ in Monaten mit
    $F(x_{0.75}) = 1 - e^{-\frac{1}{18}\cdot x_{0.75}} = 0.75 \;\Rightarrow x_{0.75} = -18 \cdot \ln(0.25) = 24.953$

Ungetrübter Musikgenuss I

Beispiel: Bob (1.85 m groß) ist ein großer Musikfan. Besonders genießt er (Rock) Konzerte. Allerdings hasst er es, wenn er keinen freien Blick auf die Bühne hat.

Frage: Mit welcher Wahrscheinlichkeit ist der vor Bob stehende Mann (vereinfachende Annahmen) mindestens genauso groß wie Bob und versperrt ihm damit die Sicht?

Lösung: Die Größe eines zufällig ausgewählten (deutschen) Mannes in cm ist eine normalverteilte Zufallsvariable mit Erwartungswert 180.3 cm und Standardabweichung 7.17 cm ([Griffith, 2014]).

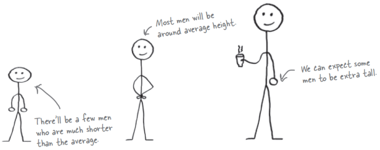


Bild: Die Größen von Männern [Griffith, 2014]

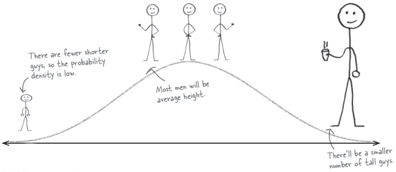


Bild: Die Dichteverteilung der Männergrößen [Griffith, 2014]

Die Normalverteilung I

Definition

Eine Zufallsvariable X heißt normalverteilt mit den Parametern μ und σ, $X \sim N(\mu, \sigma)$, wenn sie die Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

besitzt. Die zugehörige Wahrscheinlichkeitsverteilung heißt Normalverteilung oder auch Gauß-Verteilung. Der Graph der Dichtefunktion wird Gauß'sche Glockenkurve genannt. Die Parameter μ und σ sind Erwartungswert bzw. Standardabweichung von X.

Bemerkungen:

- Die Normalverteilung ist die wichtigste, weil „natürlichste" Verteilung: die meisten Werte der ZV nehmen den Erwartungswert an, es gibt jedoch ungefähr gleich viele kleinere und größere Werte.
- Viele natürlichen (Biologie, Chemie, Medizin, ...) und technischen Phänomene (Fehler, Füllstände, Größen, ...) sind normalverteilt.

Die Normalverteilung II

Weitere Bemerkungen:

- μ beschreibt die Verschiebung der Glocke auf der x-Achse relativ zum Ursprung.
- Die Fläche unter der Glockenkurve ist immer gleich 1 (unabhängig von μ und σ).
  - Je größer σ, desto breiter und niedriger ist die Glocke,
  - Je kleiner σ, desto schmaler und höher ist die Glocke.

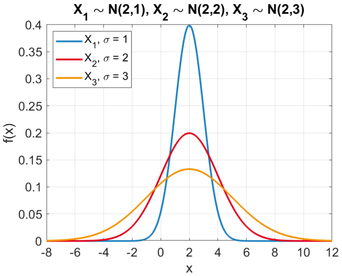


Bild: Verteilungsdichte für Normalverteilungen mit unterschiedlichem Mittelwert und unterschiedlichen Varianzen

Achtung: Die Verteilungsfunktion der Normalverteilung

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

kann nur numerisch berechnet werden, da f keine Stammfunktion besitzt.

Fakten zur Normalverteilung

Für eine normalverteilte Zufallsvariable X, $X \sim N(\mu, \sigma)$ gilt

Wertebereich $X \in \mathbb{R}$

Parameter $\mu \in \mathbb{R}$: Erwartungswert (Ortsparameter), $\sigma \in \mathbb{R}_0^+$: Standardabweichung (Skalierungsparameter)

Verteilungsdichte $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Verteilungsfunktion $F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$

Erwartungswert $\mathrm{E}[X] = \mu$

Varianz $\mathrm{Var}[X] = \sigma^2$

$X \sim N(0,1)$ heißt standardnormalverteilt mit den Parametern $\mu = 0$ und $\sigma = 1$, (siehe Bilder)

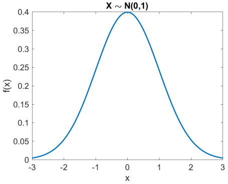


Bild: Verteilungsdichtefunktion

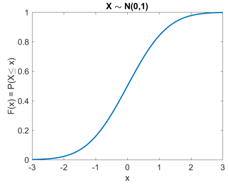


Bild: Verteilungsfunktion

Die Standardnormalverteilung

Definition

Eine Zufallsvariable Z heißt standardnormalverteilt, wenn sie normalverteilt mit den Parametern $\mu = 0$ und $\sigma = 1$, $Z \sim N(0,1)$, ist. Ihr Dichte- und Verteilungsfunktion sind gegeben durch

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$

und

$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt.$$

Verwendung:

- **Jede beliebige normalverteilte Zufallsvariable** kann standardisiert, d.h. mittels geeigneter linearer Umformungen in $Z \sim N(0,1)$ transformiert, ihre Wahrscheinlichkeiten also mit Hilfe von $\Phi(z)$ berechnet werden!
- $\Phi(z)$ ist tabelliert (Papier bzw. Tools)!

Lösungswege des 21. Jahrhunderts

Beispiele:

1. Mit welcher Wahrscheinlichkeit ist ein zufällig ausgewählter Mann höchstens 1.75 m groß ist?
2. Welche Größe überschreiten 95% der Männer nicht?

Lösung: (Verwende Annahme, dass Größe $\sim N(180.3, 7.17)$ ist)

Excel
1. `NORM.VERT(175;180,3;7,17;WAHR)`
2. `NORM.INV(0,95;180,3;7,17)`

MATLAB
1. `normcdf(175,180.3,7.17)`
2. `norminv(0.95,180.3,7.17)`

TR (Texas Instruments)
1. `DISTR` $\rightarrow$ `Normalcdf`: $\mu = 180.3, \sigma = 7.17$, `LOWERbnd` $= -1E99$, `UPPERbnd` $= 175$
2. `DISTR` $\rightarrow$ `invNormal`: `area` $= 0.95, \mu = 180.3, \sigma = 7.17$

Ergebnis
1. $P(X \leq 175) = 0.23$
2. $x_{0.95} = 192.09$ cm

Ungetrübter Musikgenuss II

Frage: Mit welcher Wahrscheinlichkeit steht vor Bob (1.85 m groß) ein Mann, der mindestens genauso groß ist, wie er selbst?
Modell: Die Größe eines zufällig ausgewählten Mannes ist eine Zufallsvariable $X \sim N(180.3, 7.17)$.

Lösung: $P(X \geq 185) = 1 - P(X < 185) = 1 - P(X \leq 185) = 1 - F(185) \overset{\text{TR}}{=} 1 - 0.7454 = 0.2546 \approx 25.5\%$

Weitere Fragen:

- Mit welcher Wahrscheinlichkeit wird Dave (1.80 m groß) durch einem vor ihm stehenden Mann der Blick zur Bühne versperrt?

  Lösung: $P(X \geq 180) = 1 - P(X < 180) = 1 - P(X \leq 180) = 1 - F(180) \overset{\text{TR}}{=} 1 - 0.483 = 0.517 \approx 51.7\%$

- Welche Größe wird von 90% aller deutschen Männer nicht überschritten?

  Lösung: 90%-Quantil $x_{0.90} \overset{\text{TR}}{=} 189.489$ cm

Beispiele zum Mitdenken

Sie kaufen zum Schnäppchenpreis elektronische Bauteile ein. Daher ist der elektrische Widerstand der Bauteile eine normalverteilt Zufallsvariable X mit $\mu = 200\ \Omega$ und $\sigma = 10\ \Omega$.

Fragen:

- Wie viel Prozent der Bauteile haben mindestens 190 Ω?
  Lösung: $P(X \geq 190) = 1 - P(X < 190) = 1 - P(X \leq 190) = 1 - F(190) \overset{\text{TR}}{=} 1 - 0.157 = 0.841$
- Welcher Widerstandswert wird nur von 2% aller Bauteile überschritten?
  Lösung: Gesucht ist das 98%-Quantil von X.
  Berechnung: $x_{0.98} \overset{\text{TR}}{=} 220.54$ Antwort: ca. 98% der Bauteile haben einen Widerstand unter 220.54 Ω, bzw. ca. 2% der Bauteile überschreiten den Wert von 220.54 Ω nicht.

68–95–99.7-Regel

Motivation: Oft ist es interessant zu einer normalverteilten Zufallsvariable nicht nur den Mittelwert, sondern auch für jene Werte, die in einem symmetrisch um den Mittelwert gelegenen Intervall liegen.

68–95–99.7-Regel / 3σ-Grenzen

Bei **jeder** normalverteilten Zufallsvariable X (mit beliebigem Mittelwert μ und Standardabweichung σ) ist die Wahrscheinlichkeit, dass X

- einen Wert zwischen $\mu - \sigma$ und $\mu + \sigma$ annimmt, etwa 68.3%,
- einen Wert zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$ annimmt, etwa 95.5%,
- einen Wert zwischen $\mu - 3\sigma$ und $\mu + 3\sigma$ annimmt, etwa 99.7%.

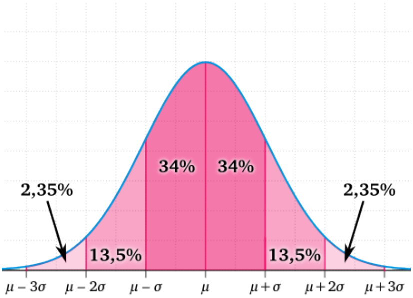
Bemerkungen:

- Eine normalverteilte ZV X nimmt fast immer einen Werte zwischen den 3σ-Grenzen an.
- Bei sehr vielen Messungen einer normalverteilten ZV X werden nur ca. 3 von 1000 Werten außerhalb der 3σ-Grenzen liegen.

Wie groß sind deutsche Frauen? Quelle: MαtheGuru

- Die Körpergröße aller Menschen (pro Geschlecht betrachtet) ist normalverteilt.
- Statistik aus dem Jahr 2006: die Größe deutscher Frauen ist normalverteilt mit $\mu = 165.4$ cm, $\sigma = 4.5$ cm.

- Anwendung der 68-95-99.7-Regel:
  - 68% aller deutschen Frauen sind zwischen 160.9cm ($\mu - \sigma$) und 169.9cm ($\mu + \sigma$) groß
  - 95% aller deutschen Frauen sind zwischen 156.4cm ($\mu - 2\sigma$) und 174.4cm ($\mu + 2\sigma$) groß
  - 99.7% aller deutschen Frauen sind zwischen 151.9cm ($\mu - 3\sigma$) und 178.9cm ($\mu + 3\sigma$) groß



Quelle: MαtheGuru

Dexter und die Liebes-Achterbahn I

Dexter's innovative Geschäftsidee:
Vermiete eine Achterbahn für Trauungen.

Problem:

- TÜV-Zulassung: Maximal 380 Pfund (190 kg) pro Wagen
- Das Gewicht der Bräute (in Pfund) ist eine ZV $X \sim N(150, 20)$
- Das Gewicht der Bräutigame (in Pfund) ist eine ZV $Y \sim N(190, \sqrt{500})$
- Mit welcher Wahrscheinlichkeit ist das Gewicht des Brautpaares kleiner gleich 380 Pfund?

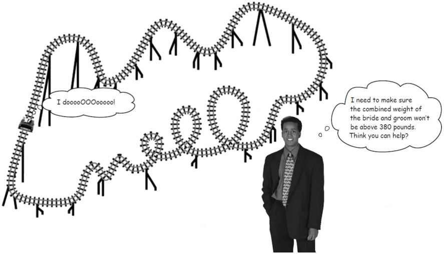


Bild: Dexter's Idee und sein Problem [Griffith, 2014]

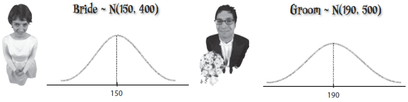


Bild: Bekannte Gewichte[Griffith, 2014]

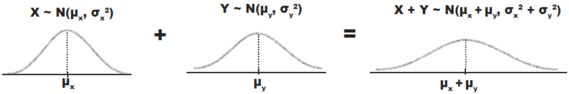
Gesucht: $P(X + Y \leq 380)$

Additionssatz der Normalverteilung

Satz

Für X und Y unabhängig und normalverteilte ZV mit $X \sim N(\mu_X, \sigma_X)$ bzw. $Y \sim N(\mu_Y, \sigma_Y)$ gilt: $X + Y$ ist ebenfalls normalverteilt mit Erwartungswert $\mu_X + \mu_Y$ und Standardabweichung $\sqrt{\sigma_X^2 + \sigma_Y^2}$:

$$X + Y \sim N(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}) :$$



Quelle: [Griffith, 2014]

Bemerkung:

- Dass sich Erwartungswerte und Varianzen addieren, gilt für alle unabhängigen Zufallsvariablen.
- Dass $X + Y$ wieder normalverteilt ist, ist eine Besonderheit der Normalverteilung, meist ändert sich die Verteilung der Summe (Gegenbeispiel: Würfel und Summe zweier Würfel).

Dexter und die Liebes-Achterbahn II

Gesucht: $P(X + Y \leq 380)$, mit $X \sim N(150, 400)$ und $Y \sim N(190, \sqrt{500})$.

Lösung:

- $X + Y \sim N(150 + 190, \sqrt{400 + 500}) = N(340, 30)$
- $P(X + Y \leq 380) = F_{X+Y}(380) \stackrel{\text{TR}}{=} 0.909$

Ein Brautpaar kann also mit einer Wahrscheinlichkeit von 90.9% in der Achterbahn getraut werden.

Neue Frage: Wie sieht es mit rein weiblichen Brautpaaren aus?

Lösung:

- $X + X \sim N(150 + 150, \sqrt{400 + 400}) = N(300, 28.28)$
- $P(X + X \leq 380) = F_{X+X}(380) \stackrel{\text{TR}}{=} 0.9977$

Ein weibliches Brautpaar kann also mit einer Wahrscheinlichkeit von 99.8% in der Achterbahn getraut werden.

Die Normalverteilung als Näherung der Binomialverteilung

Original X_B sei eine binomialverteilte ZV mit $X_B \sim Bin(n, p)$, mit $\mu = np$ und $\sigma^2 = np(1-p)$ und Verteilung $F_B(x)$.

Näherung durch eine normalverteilte ZV X_N mit Parametern $\mu = np$ und $\sigma = \sqrt{np(1-p)}$ und Verteilung $F_N(x)$.

Es gilt

$$F_B(x) \approx F_N(x + 0.5)$$

Bedingung np und $n(1-p)$ müssen „groß genug" sein, $n \cdot p \cdot (1-p) \gtrapprox 9$.

Bemerkungen:

- Der Faktor „$+0.5$" ist eine Stetigkeitskorrektur, die dafür sorgt, dass die diskrete Binomialverteilung besser durch die stetige Normalverteilung approximiert wird (siehe nächste Folie)
- Verwende die Näherung, wenn n sehr groß wird und große Binomialkoeffizienten bzw. Summen per Hand ausgerechnet werden müssten.

Stetigkeitskorrektur in Bildern [Griffith, 2014]

Beispiel: Berechnung von $P(X < 6)$

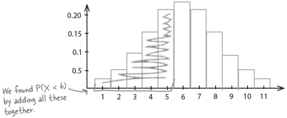


Bild: Exakt mit $X \sim Bin(12, 0.5)$

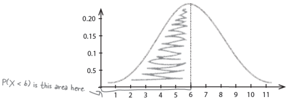


Bild: Approximativ mit $X \overset{a}{\sim} N(6, 3)$

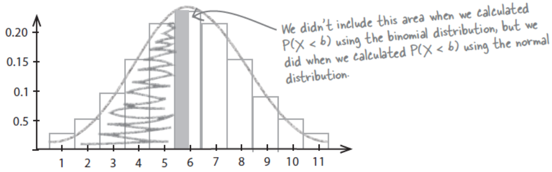


Bild: Fehler ohne Stetigkeitskorrektur

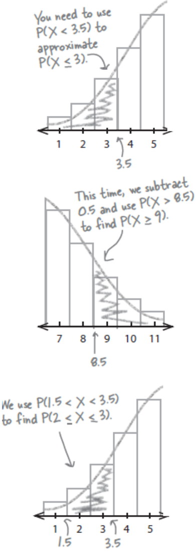


Bild: Fallunterscheidung

Bemerkung: Welche Stetigkeitskorrektur ($+0.5$ vs. -0.5) den Fehler minimiert, hängt von der gesuchten Wahrscheinlichkeit ab.

Beispiele zum Mitdenken

Bei einer digitalen Übertragung von 10 Millionen Bits ist die Wahrscheinlichkeit, dass ein Bit fehlerhaft übertragen wird, immer gleich 10^{-5}. Wie groß ist die Wahrscheinlichkeit, dass unter 10 Millionen Bits mehr als 90 Bits fehlerhaft übertragen werden?

Lösung:

- Verwende ZV X = Anzahl der fehlerhaft übertragenen Bits.
- $X \sim Bin(10^7, 10^{-5}), n = 10^7, p = 10^{-5}$, Verteilung $F_B(x)$
- $n \cdot p \cdot (1-p) = 99.999 > 9 \Rightarrow$ Näherung durch Normalverteilung $F_N(x)$ mit $\mu = np = 100$, $\sigma = \sqrt{np(1-p)} = 9.99995$ möglich:
  - $P(X > 90) = 1 - F_B(90) \approx 1 - F_N(90.5) \overset{\text{TR}}{=} 1 - 0.1586 \approx 82.89\%$
  - `MATLAB: 1 - normcdf(90.5,100,9.99995)`
- exakte Lösung: 82.86 %
  MATLAB: `1-cdf(makedist('Binomial','N',10^7,'p',10^(-5)),90)`
- Zeitersparnis in `MATLAB`: Faktor 20 bzw. Faktor 2;
  Implementierungsabhängig: 0.00003 bzw. 0.003 vs. 0.005 Sekunden
  (Mittelwerte aus 50 Berechnungen)

Die Normalverteilung als Näherung der Poisson-Verteilung

Original X_P sei eine poissonverteilte ZV mit $X_P \sim Po(\lambda)$, mit $\mu = \lambda$ und $\sigma^2 = \lambda$ und Verteilung $F_P(x)$.

Näherung durch eine normalverteilte ZV X_N mit Parametern $\mu = \lambda$ und $\sigma = \sqrt{\lambda}$ und Verteilung $F_N(x)$.

Es gilt

$$F_P(x) \approx F_N(x + 0.5)$$

Bedingung λ muss „groß genug“ sein, $\lambda \gtrapprox 9$.

Bemerkungen:

- Der Faktor „$+0.5$“ ist wieder eine Stetigkeitskorrektur, die dafür sorgt, dass die diskrete Poisson-Verteilung besser durch die stetige Normalverteilung approximiert wird.
- Verwende die Näherung, wenn n sehr groß wird und große Summen per Hand ausgerechnet werden müssten; außerdem für schnellere Berechnung von Wahrscheinlichkeiten in Intervallen

Beispiele zum Mitdenken

Angenommen, die Anzahl von Feinstaubteilchen pro m^3 Luft in der Konstanzer Innenstadt ist poissonverteilt mit dem Erwartungswert 20 (Teilchen pro m^3). Wie groß ist die Wahrscheinlichkeit, weniger als 15 Feinstaubteilchen pro m^3 Luft vorzufinden?

Lösung:

- Verwende ZV $X =$ Anzahl der Feinstaubteilchen pro m^3 Luft.
- $X \sim Po(20), \lambda = 20$, Verteilung $F_P(x)$
- $\lambda = 20 > 9 \Rightarrow$ Näherung durch Normalverteilung $F_N(x)$ mit $\mu = \lambda = 20$, $\sigma = \sqrt{\lambda} = 4.472$ möglich
  - $P(X < 15) = P(X \leq 14) = F_P(14) \approx F_N(14) \stackrel{\text{TR}}{=} 10.93\%$
  - MATLAB: `normcdf(14,20,4.472)`
- exakte Lösung: 10.49 %
  MATLAB: `cdf(makedist('Poisson','lambda',20),14)`
- Zeitersparnis in `MATLAB`: Faktor 10 bzw. Faktor 1.02; Implementierungsabhängig: 0.00003 bzw. 0.00222 vs. 0.00229 Sekunden (Mittelwerte aus 50 Berechnungen)

Verwendete oder empfohlene Literatur I

[Diez et al., 2015] Diez, D. M., Barr, C. D., and Çetinkaya Rundel, M. (2015).
OpenIntroStatistics.
openintro.org.
Online verfügbar unter www.openintro.org.

[Downey, 2014] Downey, A. B. (2014).
Think Stats - Exploratory Data Analysis in Python.
O'Reilly.
Online verfügbar unter www.greenteapress.com.

[Griffith, 2014] Griffith, D. (2014).
Head First Statistics / Statistik von Kopf bis Fuß.
O'Reilly.

Verwendete oder empfohlene Literatur II

[Haslwanter, 2016] Haslwanter, T. (2016).
An Introduction to Statistics with Python.
Springer.
Python Notebooks auf github verfügbar.

[Papula, 2016] Papula, L. (2016).
Mathematik für Ingenieure und Naturwissenschaftler, Band 3.
Springer Vieweg, 7. Auflage.
Als eBook in der HTWG-Bibliothek verfügbar.

[Teschl and Teschl, 2014] Teschl, G. and Teschl, S. (2014).
Mathematik für Informatiker: Band 2: Analysis und Statistik.
Springer Vieweg, 3. Auflage.
Als eBook in der HTWG-Bibliothek verfügbar.