Normalverteilung

Master Practical Course “Data Analysis with Python”
(WiSe 2016/17)

Robert Müller, Christian Lemke, Max Wagner, Mattes Wieben

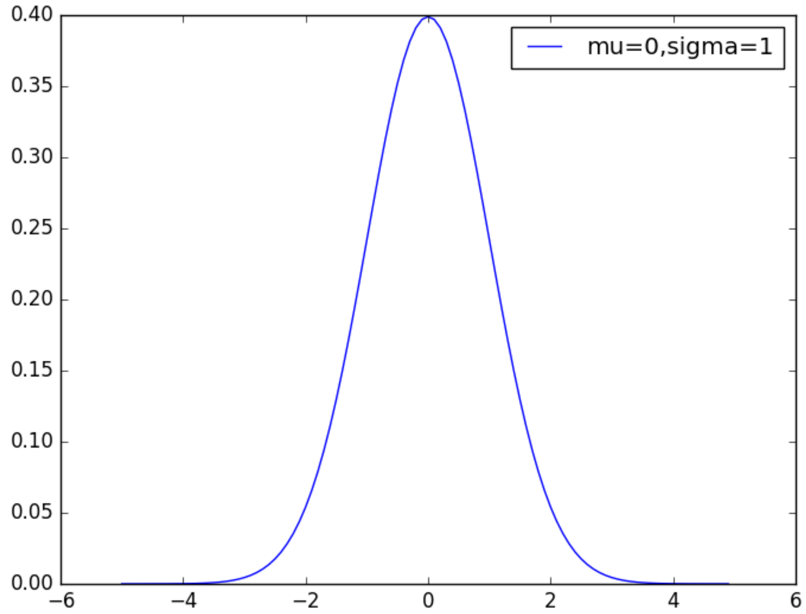
Lehr- und Forschungseinheit für Programmier- und Modellierungssprachen
Institut für Informatik
Ludwig-Maximilians-Universität München

06. Dezember 2016

Agenda

Allgemeine Grundlagen

- Beschreiben von Zufallsvariablen
- Bekannt als Gauß-Kurve
- Geprägt durch de Moivre, Laplace/Poisson und Gauß
- Approximiert Binomialverteilung



Mathematische Grundlagen

Sei X eine diskrete Zufallsvariable, wobei A die Menge der möglichen Werte beschreibt, die X annehmen kann, dann gilt:

- Erwartungswert $\mu(X) = \sum_{x \in A}(x * P(X = x))$
- Varianz $V(X) = \sum_{x \in A}((x - \mu)^2 * P(X = x))$
- Standardabweichung $\sigma(X) = \sqrt{V(X)}$

Ein Münzwurf wäre zum Beispiel eine diskrete Verteilung.

Beispiele

Table: Münzwurf

Ergebnis	Kopf	Zahl
Wahrscheinlichkeit	$1/2$	$1/2$

Table: Würfel

Ergebnis	1	2	3	4	5	6
Wahrscheinlichkeit	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

Stetige Wahrscheinlichkeitsverteilungen

Für stetige Verteilungen wäre die Wahrscheinlichkeit jedoch immer nahe 0.
$\rightarrow$ Daher werden diese Verteilungen in einer Dichtefunktion angegeben

Dichtefunktion

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

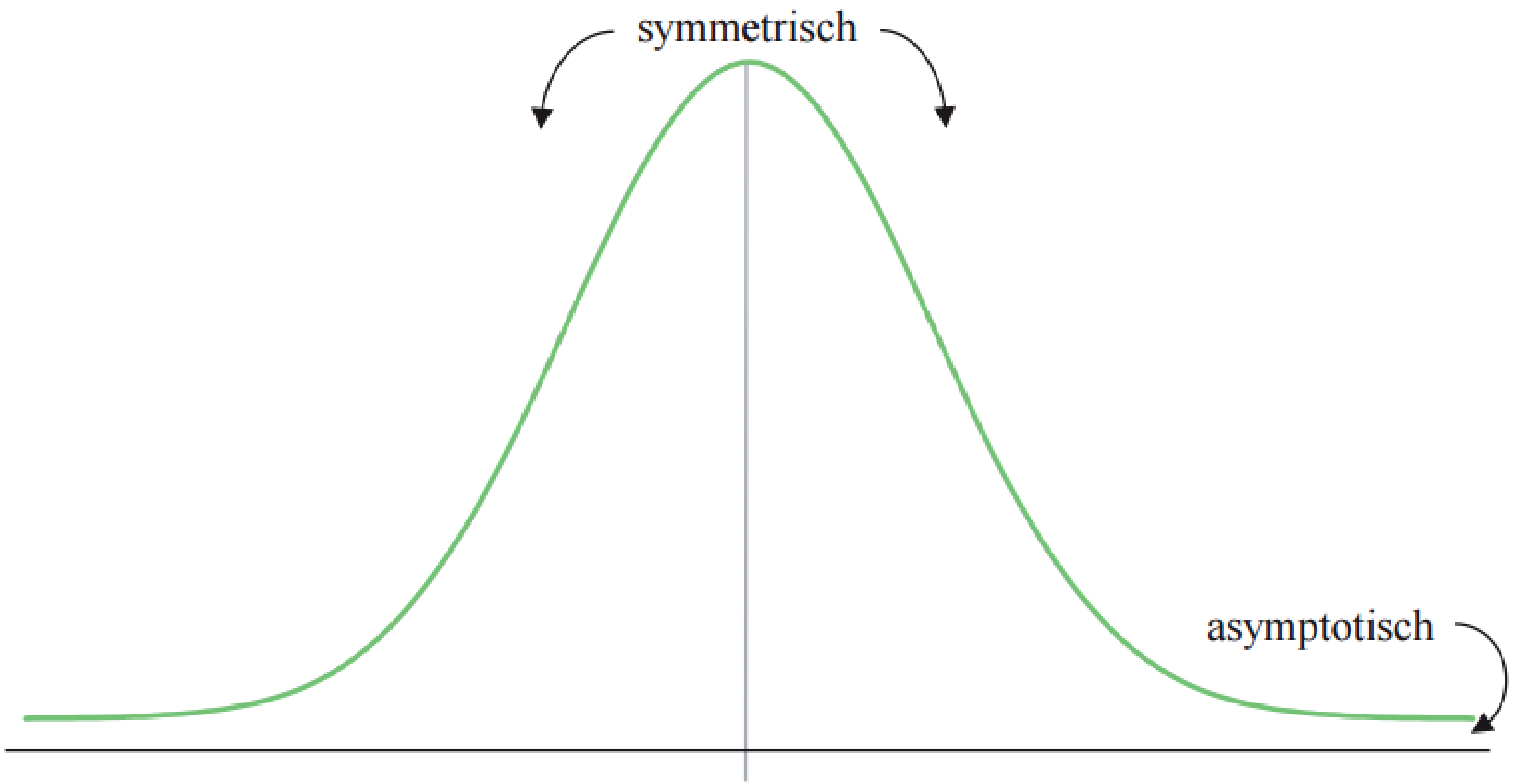


Figure: Gauß-Kurve[1]

[1] Franz Kronthaler: Statistik angewandt (2014) S. 110

Dichtefunktion

Wahrscheinlichkeitsdichtefunktion

Wahrscheinlichkeitsdichtefunktionen lösen das Problem, dass bei stetigen Funktionen keine Wahrscheinlichkeiten (außer 0) für diskrete Werte angegeben werden können.

- Die Fläche A für ein Interval I unter einer Dichtefunktion für eine Zufallsvariable X gibt an, wie hoch die Wahrscheinlichkeit ist, dass der Wert von X im Interval I liegt.

$$P(x \in [y, z]) = \int_y^z f(x \big| \mu, \sigma) dt \qquad (2)$$

wobei $f(x \big| \mu, \sigma)$ die Dichtefunktion von X ist.

Dichtefunktion und Standardabweichung

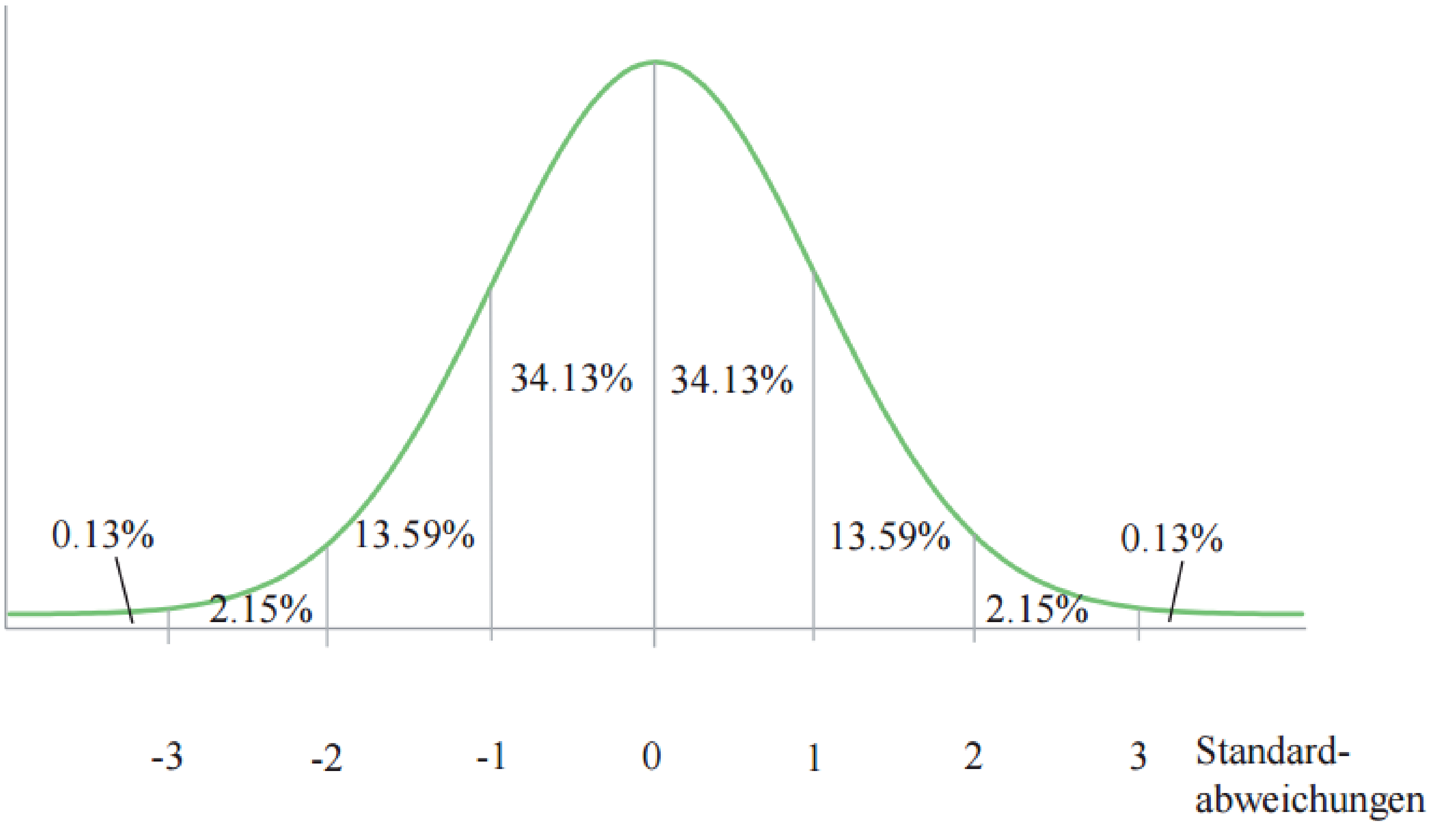


Figure: 68-95-99,7-Regel[2]

[2] vgl.: Franz Kronthaler: Statistik angewandt (2014) S. 111 [geändert]

Dichtefunktion und Standardabweichung

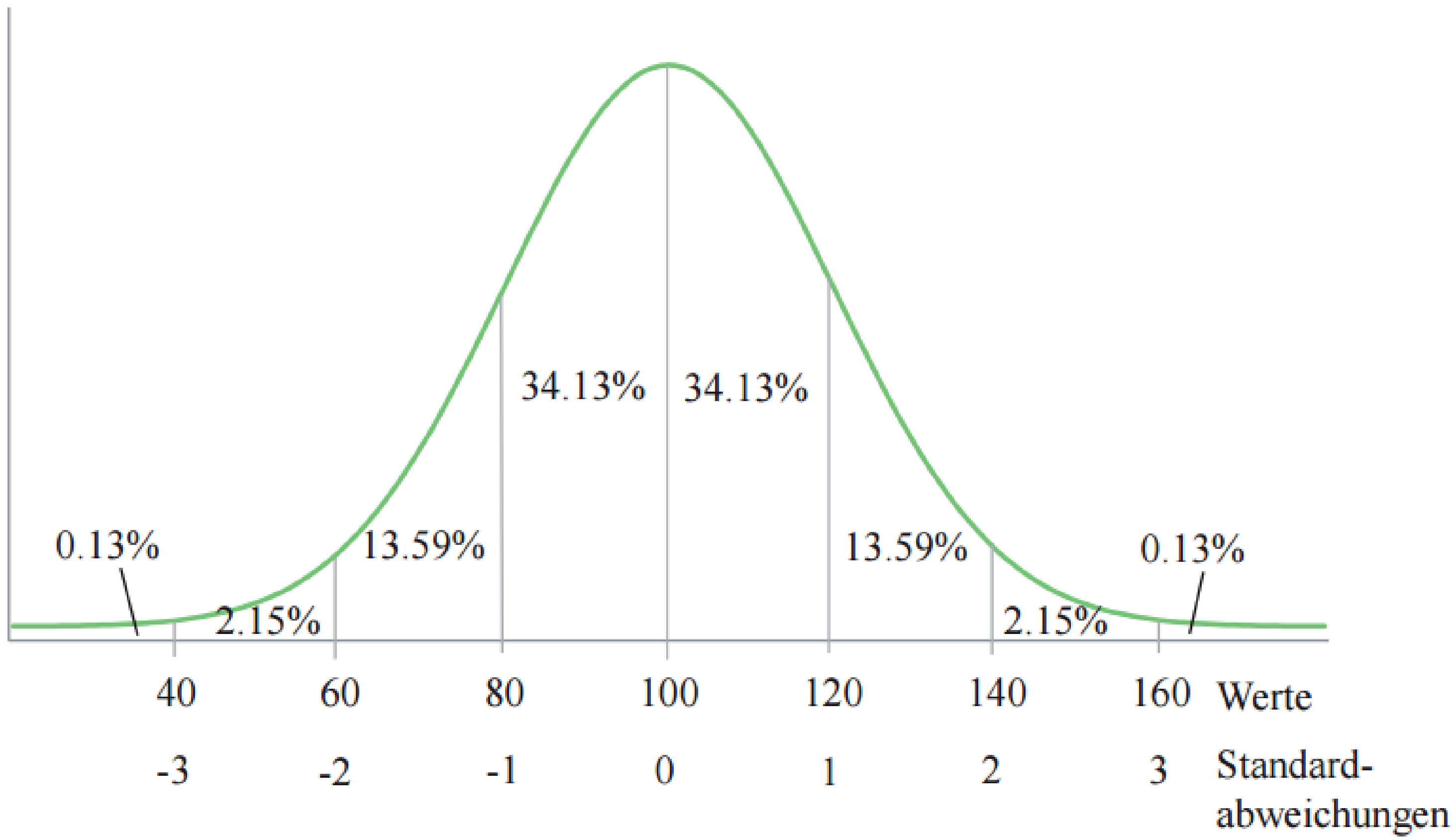
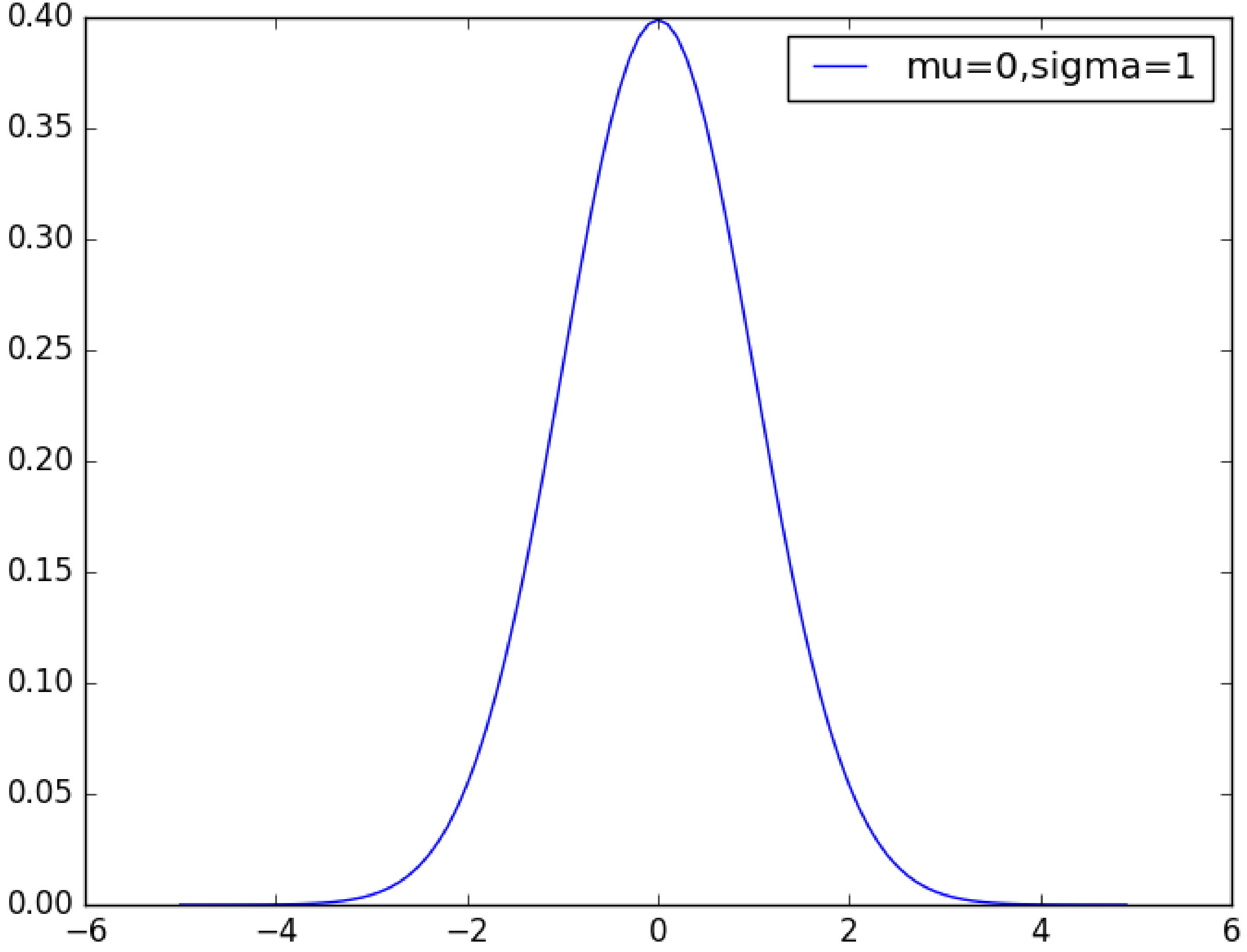


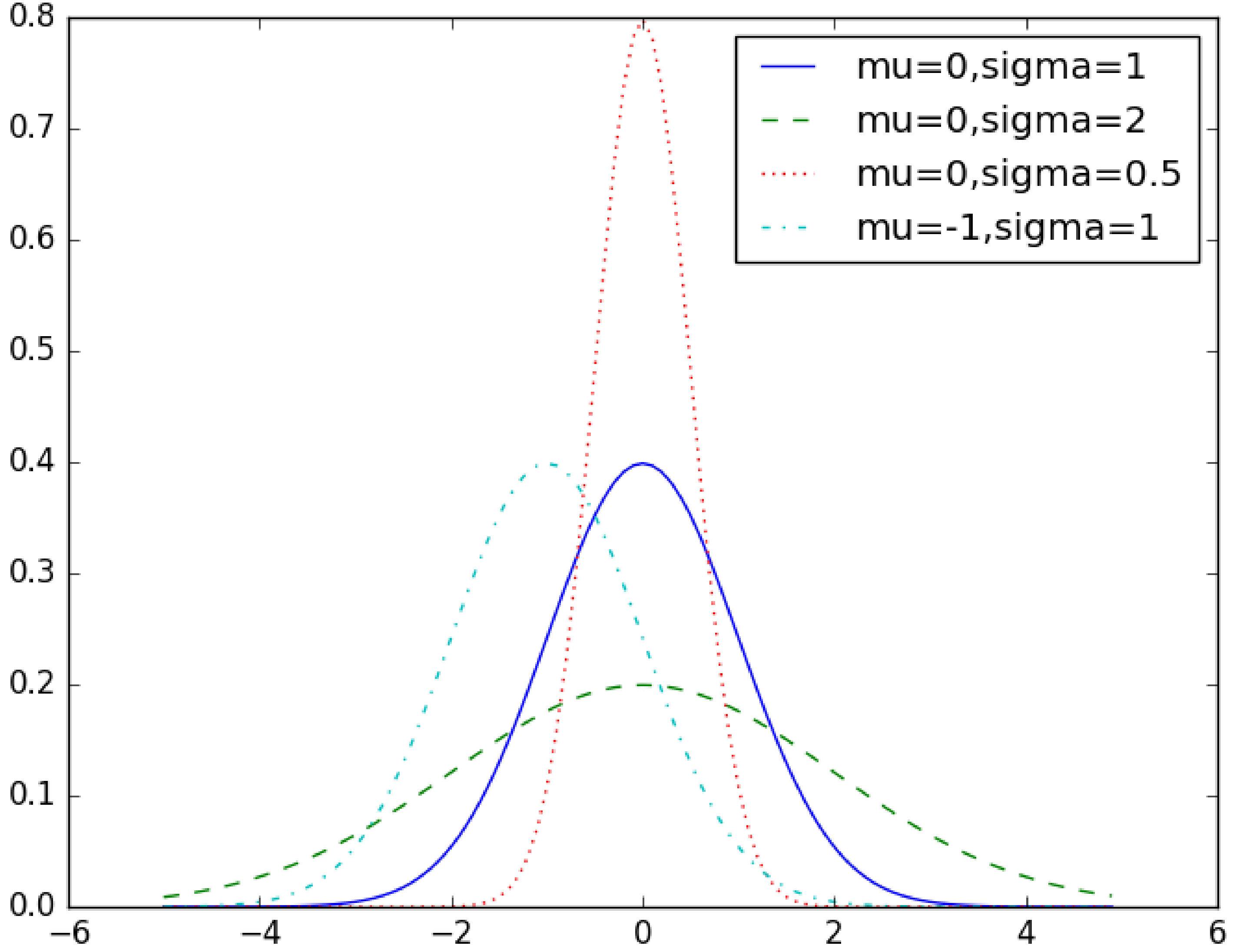
Figure: Flächen der Normalverteilung[3]

[3]Franz Kronthaler: Statistik angewandt (2014) S. 111

Standardnormalverteilung

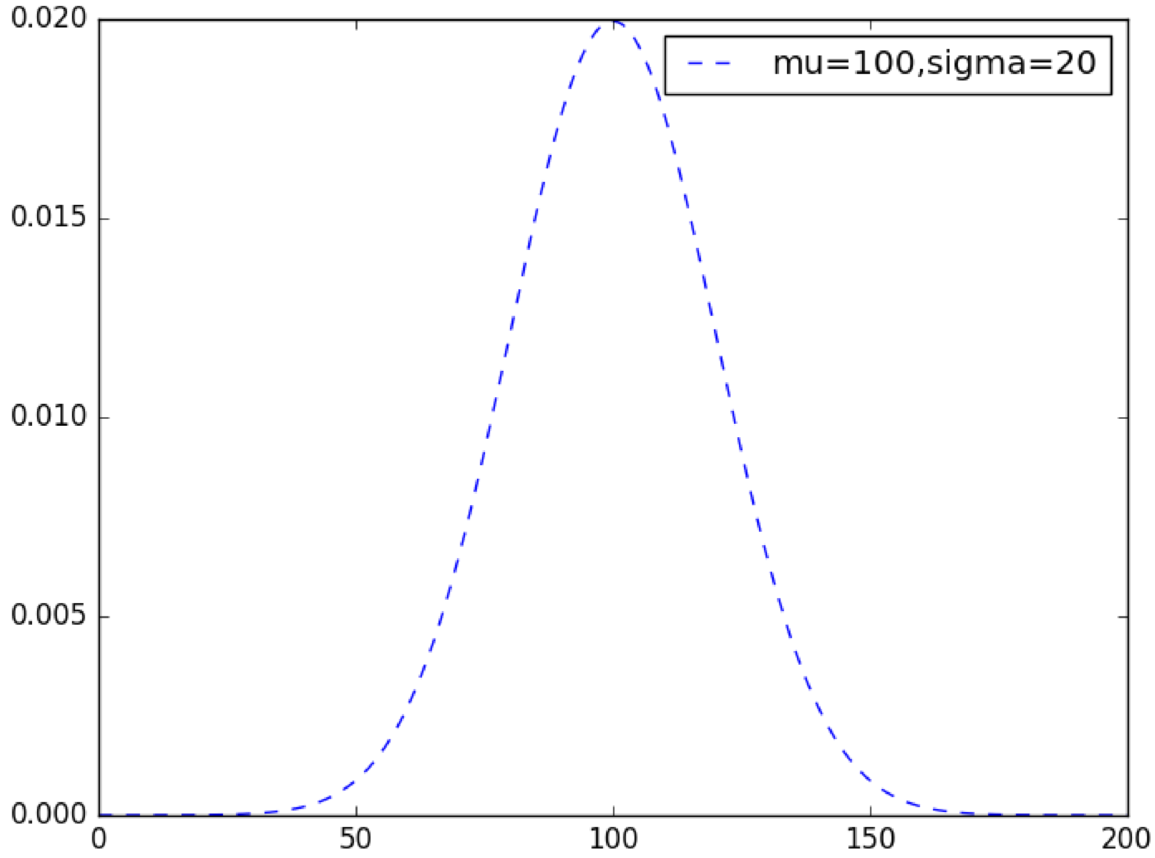


Verschiedene Dichtefunktionen normalverteilter Zufallsvariablen



Z-Transformation

Für jede normalverteilte Zufallsvariable X gilt: $Z := \frac{X-\mu}{\sigma}$,
wobei X dann in die standardnormalverteilte Zufallsvariable Z transformiert wurde.



Standardnormalverteilung-Tabelle

$$\Phi_{0,1}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2} dt \quad (3)$$

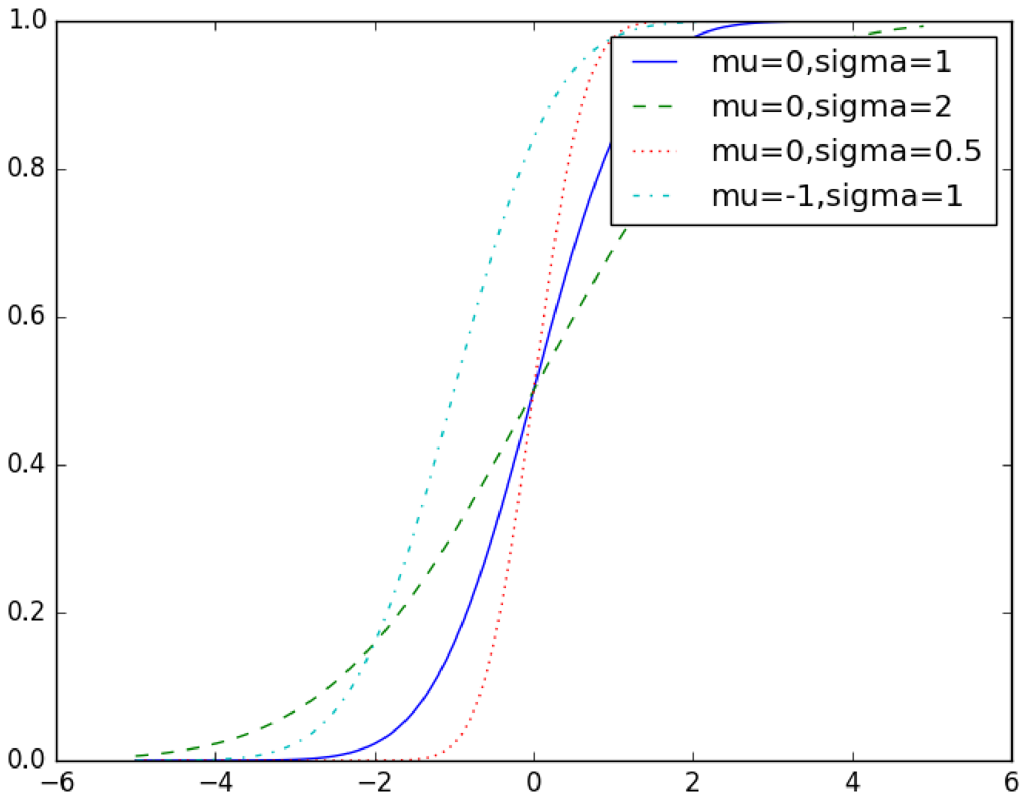
z	0	0,01	0,02	0,03	0,04	0,05
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884

Figure: Wahrscheinlichkeiten[4]

[4] vgl.: `https://de.wikipedia.org/wiki/Tabelle_Standardnormalverteilung`

Verschiedene Verteilungsfunktionen normalverteilter Zufallsvariablen

- $f(x)$ gibt an, wie hoch die Wahrscheinlichkeit ist, dass der Wert einer Zufallsvariablen X kleiner oder gleich x ausfällt.
- existiert für jedes positive μ und jedes positive σ



Anwendung in der Datanalyse

Anwendung in der Datanalyse

- Anzahl der Bilder, die in einem Jahr erstellt wurden
- Anzahl der Bilder pro Künstler
- Anzahl der Bilder pro Land
- Anzahl der Tags pro Bild
- Anzahl der Taggungen pro Bild

- Füllmenge von Lebensmitteln
- Körpergröße/Schuhgröße von Menschen
- IQ von Menschen
- Trinkgeld eines Bar-Mitarbeiters pro Tag
- Jahresniederschlag in München

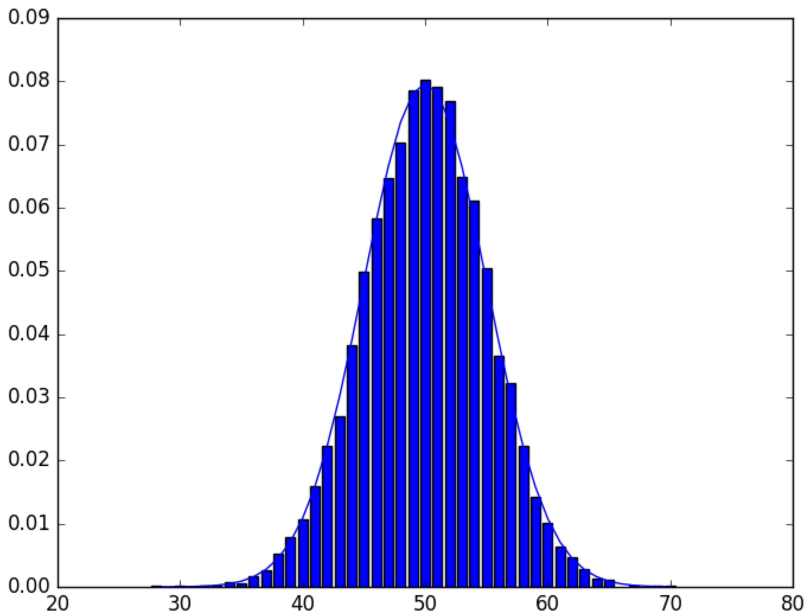
Zentraler Grenzwertsatz

$$X = \frac{1}{n}(x_1 + ... + x_n) \tag{4}$$

Voraussetzungen:

- n ist groß.
- μ und σ sind für $x_1 ... x_n$ ca. gleich groß

Folge: $\mu(X) \approx \mu(x_1 ... x_n)$ und $\sigma(X) \approx \sigma(x_1 ... x_n)$



Zusammenfassung

- Beschreibung von Zufallsvariablen
- Dichte- und Verteilungsfunktion
- Standardnormalverteilung: $\mu = 0$ und $\sigma = 1$.
- Z-Transformation
- Grundlage für zentralen Grenzwertsatz

Quellen

- Franz Kronthaler: Statistik angewandt (2014)
- Joel Grus: Data Science from Scratch (2015)
- Hans-Otto Georgii: Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik (2009)
- matheguru.com/stochastik/31-normalverteilung.html