Stochastik für die Informatik, Vorlesung 14

Inhalt

- Normalverteilung
- Zentraler Grenzwertsatz
- Normalapproximation der Binomialverteilung

Lernziele

- Die Normalverteilung kennen, mit ihnen rechnen können, ihr Auftreten kennen
- Den zentralen Grenzwertsatz und seine wichtigsten Implikationen kennen
- Die Binomialverteilung mit Hilfe der Normalverteilung approximieren können

Vorkenntnisse Stoff der bisherigen Vorlesungen, insbesondere zum Thema Zufallsvariablen und Verteilungen, Integral- und Differentialrechnung

Wichtige Beispiele: Normalverteilung/Gaußverteilung

(Def. 6.4). Eine Zufallsvariable X heißt normalverteilt bzw. Gauß-verteilt zu den Parametern $\mu \in \mathbb{R}, \sigma^2 > 0$, falls X die Dichte f_X hat, mit

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(Satz 6.6) Für die Normalverteilung gilt

$$\mathbb{E}[X] = \mu, \mathbb{V}(X) = \sigma^2.$$

- Verteilungsfunktion: keine geschlossene Form.
- Notation oft: $\varphi_{\mu,\sigma}$ oder φ_{μ,σ^2} für die Dichte, Φ_{μ,σ^2} für die Verteilungsfunktion.
- Standardnormalverteilung: Normalverteilung mit den Parameteren $\mu = 0, \sigma^2 = 1$.
- Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$.

Wichtige Beispiele: Normalverteilung/Gaußverteilung

(Def. 6.4). Eine Zufallsvariable X heißt normalverteilt bzw. Gauß-verteilt zu den Parametern $\mu \in \mathbb{R}, \sigma^2 > 0$, falls X die Dichte f_X hat, mit

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(Satz 6.6) Für die Normalverteilung gilt

$$\mathbb{E}[X] = \mu, \mathbb{V}(X) = \sigma^2.$$

- Verteilungsfunktion: keine geschlossene Form.
- Notation oft: $\varphi_{\mu,\sigma}$ oder φ_{μ,σ^2} für die Dichte, Φ_{μ,σ^2} für die Verteilungsfunktion.
- Standardnormalverteilung: Normalverteilung mit den Parameteren $\mu = 0, \sigma^2 = 1$.
- Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$.

Wichtige Beispiele: Normalverteilung/Gaußverteilung

(Def. 6.4). Eine Zufallsvariable X heißt normalverteilt bzw. Gauß-verteilt zu den Parametern $\mu \in \mathbb{R}, \sigma^2 > 0$, falls X die Dichte f_X hat, mit

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(Satz 6.6) Für die Normalverteilung gilt

$$\mathbb{E}[X] = \mu, \mathbb{V}(X) = \sigma^2.$$

- Verteilungsfunktion: keine geschlossene Form.
- Notation oft: $\varphi_{\mu,\sigma}$ oder φ_{μ,σ^2} für die Dichte, Φ_{μ,σ^2} für die Verteilungsfunktion.
- Standardnormalverteilung: Normalverteilung mit den Parameteren $\mu = 0, \sigma^2 = 1$.
- Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$.

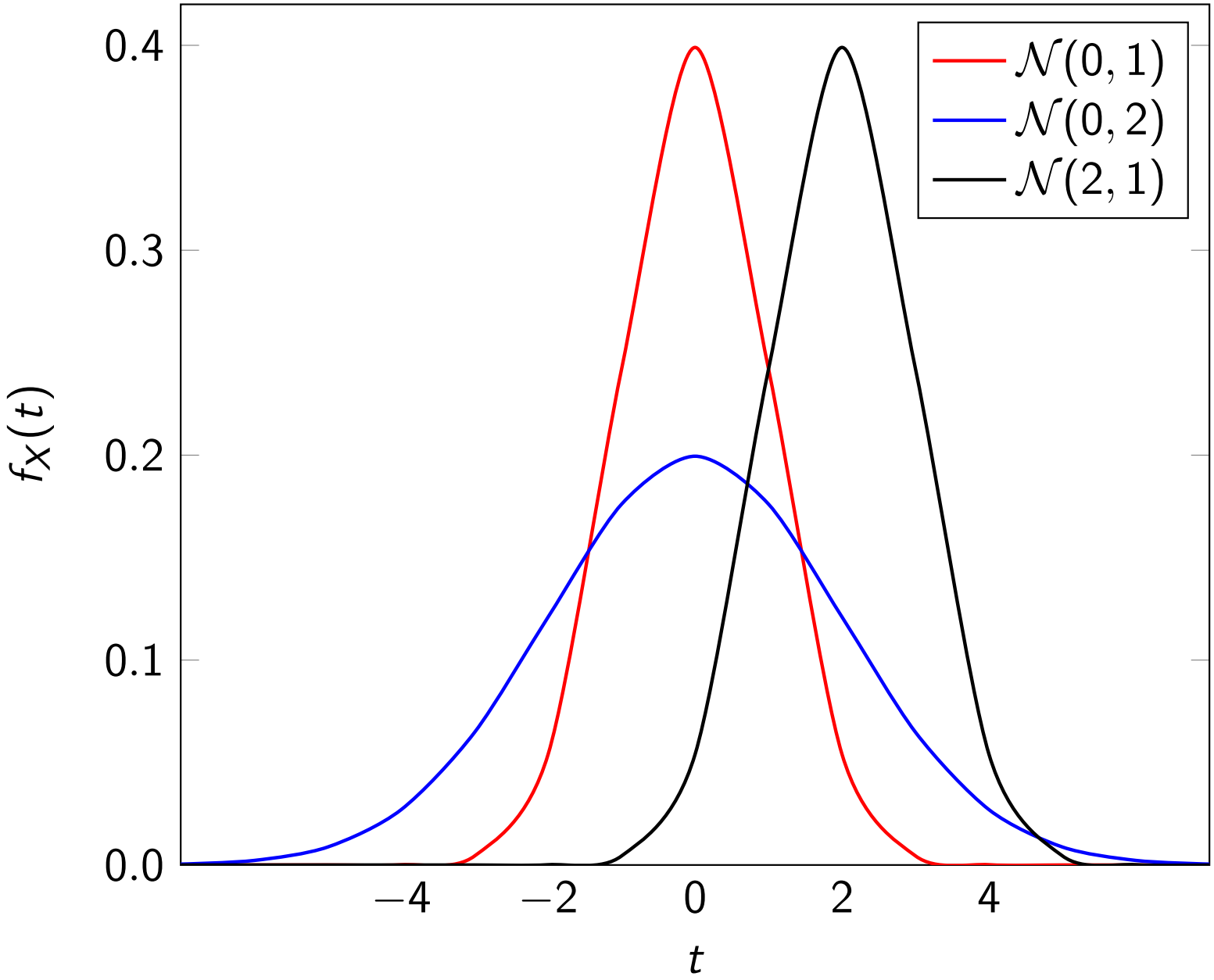


Abbildung: Dichte der Normalverteilung für verschiedene Parameterwerte

Normalverteilung: Standardisierung

(Satz 6.7) Sei $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann ist

$$Y := \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

- Damit können beliebige normalverteilte Zufallsvariablen auf standardnormalverteilte Zufallsvariablen "transformiert" werden.
- Die Standardnormalverteilung $\Phi_{0,1}$ ist in Tabellen verfügbar.
- (Beweis)

- Beispiel 6.6: Sei $X \sim \mathcal{N}(-1.3, 4)$. Berechne $\mathbb{P}(X > 0)$.
- Beispiel 6.7: Sei $X \sim \mathcal{N}(-1.3, 4)$. Berechne $\mathbb{P}(X > -2)$.
- (Beispiel 6.8 Standardintervalle)

Normalverteilung: Tabelle

$\Phi_{0,1}(x) = \mathbb{P}(X \leq x)$ für eine standardnormalverteilte Zufallsvariable X

x	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
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	0,5636	0,5675	0,5714	0,5753
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	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	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	0,8389
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
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964

Einschub: Unabhängigkeit von allgemeinen Zufallsvariablen

(Erinnerung: Satz 4.1). Zwei diskrete Zufallsvariablen X und Y sind genau dann unabhängig, wenn für alle $x \in X(\Omega), y \in Y(\Omega)$ gilt

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y).$$

Problem: Für Zufallsvariablen mit Dichte sind diese Wahrscheinlichkeiten sind immer gleich 0!

Allgemeiner:

(Definition) Zwei (beliebige) Zufallsvariablen X und Y sind genau dann unabhängig, wenn für alle $x \in X(\Omega), y \in Y(\Omega)$ gilt

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y).$$

- Für diskrete Zufallsvariablen ist das äquivalent zur bisherigen Definition.
- Gemeinsame Verteilung
- Faltungsformel

Einschub: Unabhängigkeit von allgemeinen Zufallsvariablen

(Erinnerung: Satz 4.1). Zwei diskrete Zufallsvariablen X und Y sind genau dann unabhängig, wenn für alle $x \in X(\Omega), y \in Y(\Omega)$ gilt

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y).$$

Problem: Für Zufallsvariablen mit Dichte sind diese Wahrscheinlichkeiten sind immer gleich 0!

Allgemeiner:

(Definition) Zwei (beliebige) Zufallsvariablen X und Y sind genau dann unabhängig, wenn für alle $x \in X(\Omega), y \in Y(\Omega)$ gilt

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y).$$

- Für diskrete Zufallsvariablen ist das äquivalent zur bisherigen Definition.
- Gemeinsame Verteilung
- Faltungsformel

Normalverteilung: Weitere wichtige Eigenschaften

(Satz 6.11) Seien X und Y normalverteilt. Dann sind X und Y genau dann unabhängig, wenn sie unkorreliert sind.

- Für normalverteilte Zufallsvariablen reicht es zur Überprüfung der Unabhängigkeit also, zu beweisen dass $\text{cov}(X, Y) = 0$ ist.
- Für alle anderen Verteilungen stimmt das normalerweise nicht!

(Satz 6.8) Seien X und Y unabhängige normalverteilte Zufallsvariablen, $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Dann ist $X + Y$ wieder normalverteilt, und zwar mit Parametern $\mu_1 + \mu_2$ und $\sigma_1^2 + \sigma_2^2$.

- Auch diese Eigenschaft gilt nicht grundsätzlich für beliebige Verteilungen
- Allgemein: Faltungsformel
- Falls X und Y abhängig sind, gilt der Satz nicht.
- Bem. Mehrdimensionale Normalverteilung.

Normalverteilung: Weitere wichtige Eigenschaften

(Satz 6.11) Seien X und Y normalverteilt. Dann sind X und Y genau dann unabhängig, wenn sie unkorreliert sind.

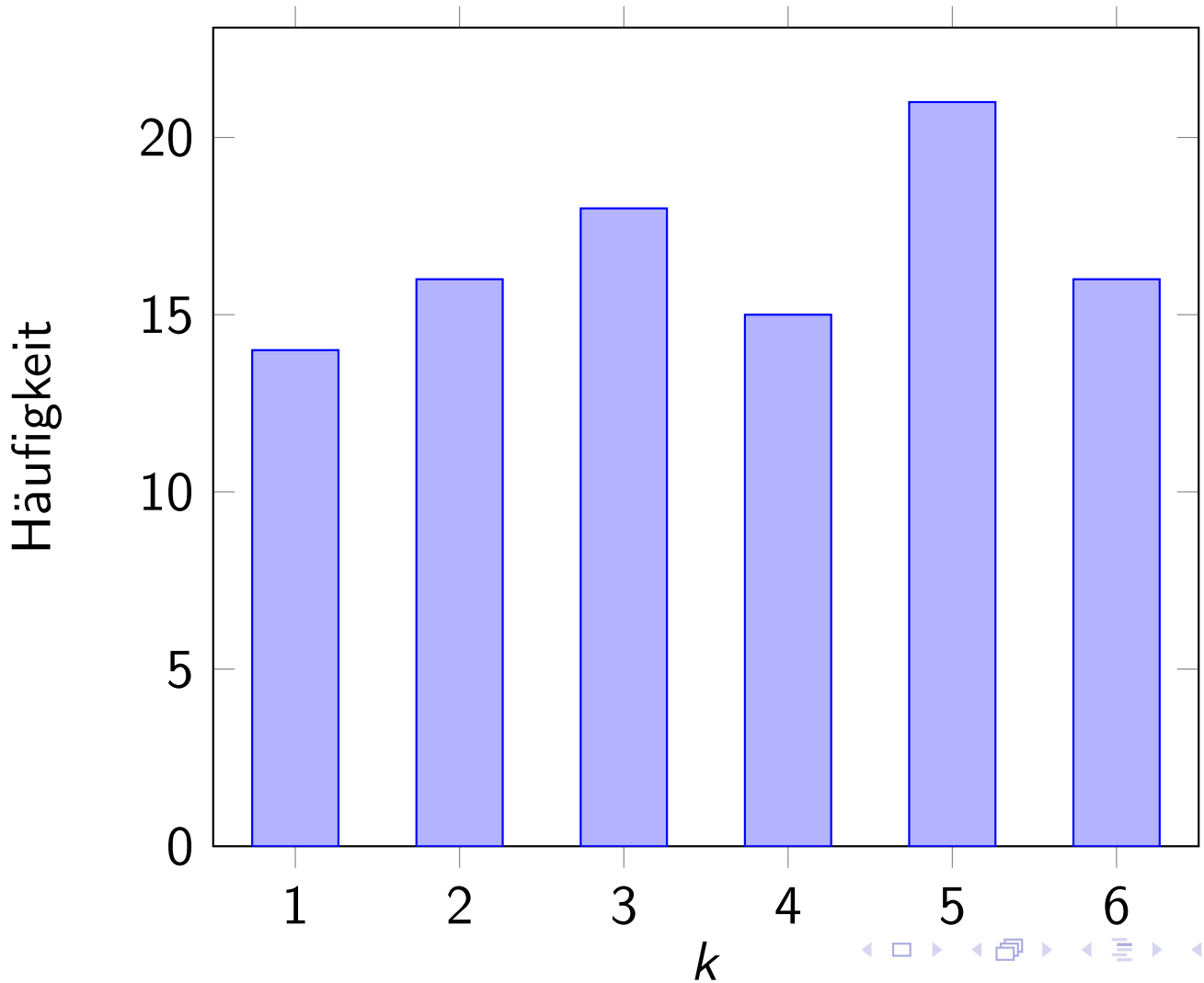
- Für normalverteilte Zufallsvariablen reicht es zur Überprüfung der Unabhängigkeit also, zu beweisen dass $\text{cov}(X, Y) = 0$ ist.
- Für alle anderen Verteilungen stimmt das normalerweise nicht!

(Satz 6.8) Seien X und Y unabhängige normalverteilte Zufallsvariablen, $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Dann ist $X + Y$ wieder normalverteilt, und zwar mit Parametern $\mu_1 + \mu_2$ und $\sigma_1^2 + \sigma_2^2$.

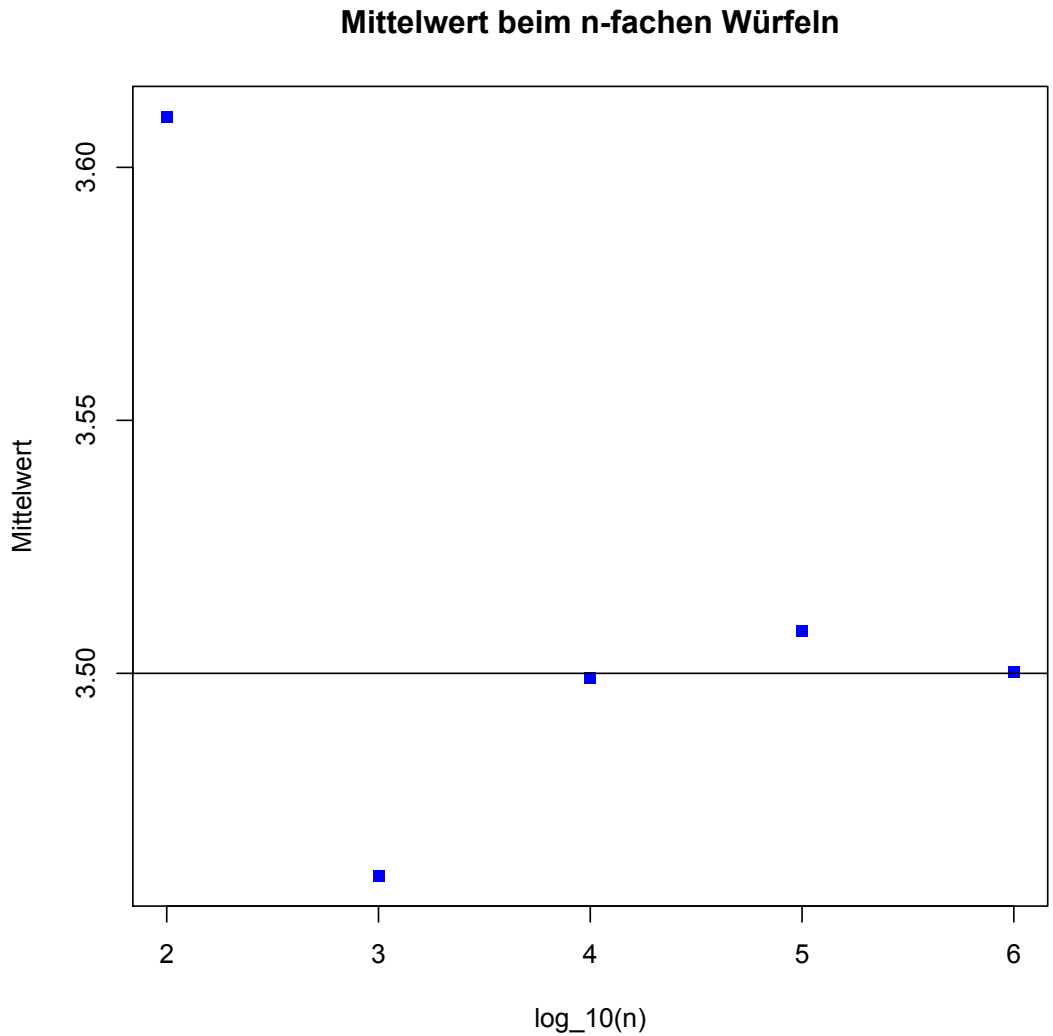
- Auch diese Eigenschaft gilt nicht grundsätzlich für beliebige Verteilungen
- Allgemein: Faltungsformel
- Falls X und Y abhängig sind, gilt der Satz nicht.
- Bem. Mehrdimensionale Normalverteilung.

Kapitel 7: Grenzwertsätze

Beispiel aus Kapitel 5: Ergebnis beim 100-fachen Würfeln. Fairer Würfel, X =Ergebnis eines Wurfs.
Simulation: 100x Würfeln, y_i Ergebnis des $i-$ten Wurfs.



Mittelwert beim n-fachen Würfeln



Für große n nähert sich der beobachtete Mittelwert $\frac{1}{n}\sum_{i=1}^{n} y_i$ dem Erwartungswert an.

Gesetz der großen Zahlen

(Satz 7.1: Gesetz der großen Zahlen). Sei $(X_i)_{i\in\mathbb{N}}$ eine Folge von unabhängigen, identisch verteilten Zufallsvariablen auf $(\Omega, \mathbb{P})$ mit $\mathbb{V}(X_i) < \infty$. Dann gilt

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X_1].$$

- "identisch verteilt" bedeutet dabei, dass alle X_i *dieselbe* Verteilung haben. Insbesondere gilt $\mathbb{E}[X_1] = ... = \mathbb{E}[X_n]$ und $\mathbb{V}(X_1) = ... = \mathbb{V}(X_n)$.
- (Beweis mit Hilfe der Chebyshev-Ungleichung)
- (Bem. Art der Konvergenz)

Gesetz der großen Zahlen: Approximation

Aus dem Gesetz der großen Zahlen wissen wir, dass für unabhängige, identisch verteilte Zufallsvariablen $(X_i)_{i\in\mathbb{N}}$ gilt:

$$\sum_{i=1}^{n} X_i \approx n \cdot \mathbb{E}[X_1].$$

- Grundpfeiler der Statistik: Mittel über Messwerte als Schätzer für den Erwartungswert.
- Verbesserung der Approximation? eine Aussage über den Fehler?

Zentraler Grenzwertsatz

(Satz 7.2: Zentraler Grenzwertsatz) Sei (X_i) eine Folge von unabhängigen, identisch verteilten Zufallsvariablen auf $(\Omega, \mathbb{P})$, mit $\mathbb{E}[X_1] = \mu, \mathbb{V}(X_1) = \sigma^2 \in (0, \infty)$. Dann gilt für alle $x \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \leq x\Big) = \Phi_{0,1}(x).$$

- $\Phi_{0,1}(x)$ ist die Verteilungsfunktion der Standardnormalverteilung
- (ohne Beweis)
- Dieser Satz gilt unabhängig von der Verteilung der X_i! Die Normalverteilung ist der universelle Limes.

Zentraler Grenzwertsatz

- Der zentrale Grenzwertsatz besagt, dass die Zufallsvariable

$$Y := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

  ungefähr (für große n) standardnormalverteilt ist.

- Verbesserte Approximation durch Umformung der obigen Gleichung:

$$\sum_{i=1}^{n} X_i \approx n \cdot \mathbb{E}[X_1] + \sqrt{n} \cdot \sigma \cdot Y,$$

  wobei Y eine standardnormalverteilte Zufallsvariable ist. $n \cdot \mathbb{E}[X_i]$ ist die Information aus dem Gesetz der großen Zahlen, $\sqrt{n} \cdot \sigma \cdot Y$ die Information aus dem zentralen Grenzwertsatz.

Beispiel 7.2: Binomialverteilung

Sei $Z \sim \text{Bin}(0.4, 20)$.

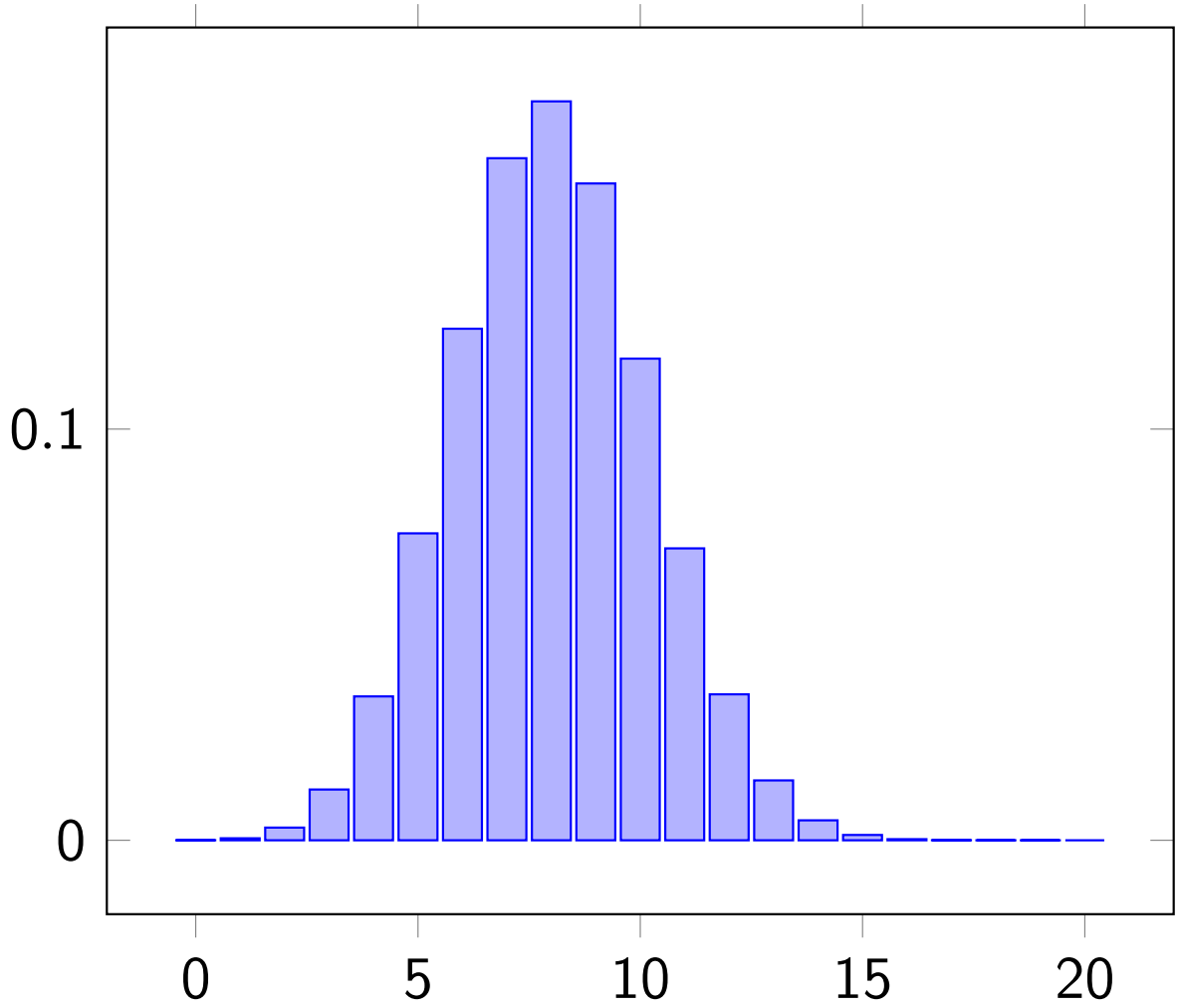
Wie in Kapitel 3 können wir schreiben:

$$Z = \sum_{i=1}^{20} X_i,$$

wobei die X_i unabhängige, identisch verteilte Bernoulli-Variablen mit Parameter $p = 0.4$ sind, also ist

$$\mathbb{E}[X_1] = p = 0.4, \quad \text{und} \quad \mathbb{V}(X_1) = p(1-p) = 0.4 \cdot 0.6 = 0.24.$$

Binomialverteilung mit $p = 0.4, n = 20$



Beispiel 7.2, Fortsetzung

Wir können nun den zentralen Grenzwertsatz auf

$$Z = \sum_{i=1}^{20} X_i$$

anwenden, und erhalten

$$Z \approx n \cdot \mathbb{E}[X_1] + \sqrt{n} \cdot \sigma \cdot Y = 20 \cdot 0.4 + \sqrt{20 \cdot 0.24}\, Y \approx 8 + 2.2\, Y,$$

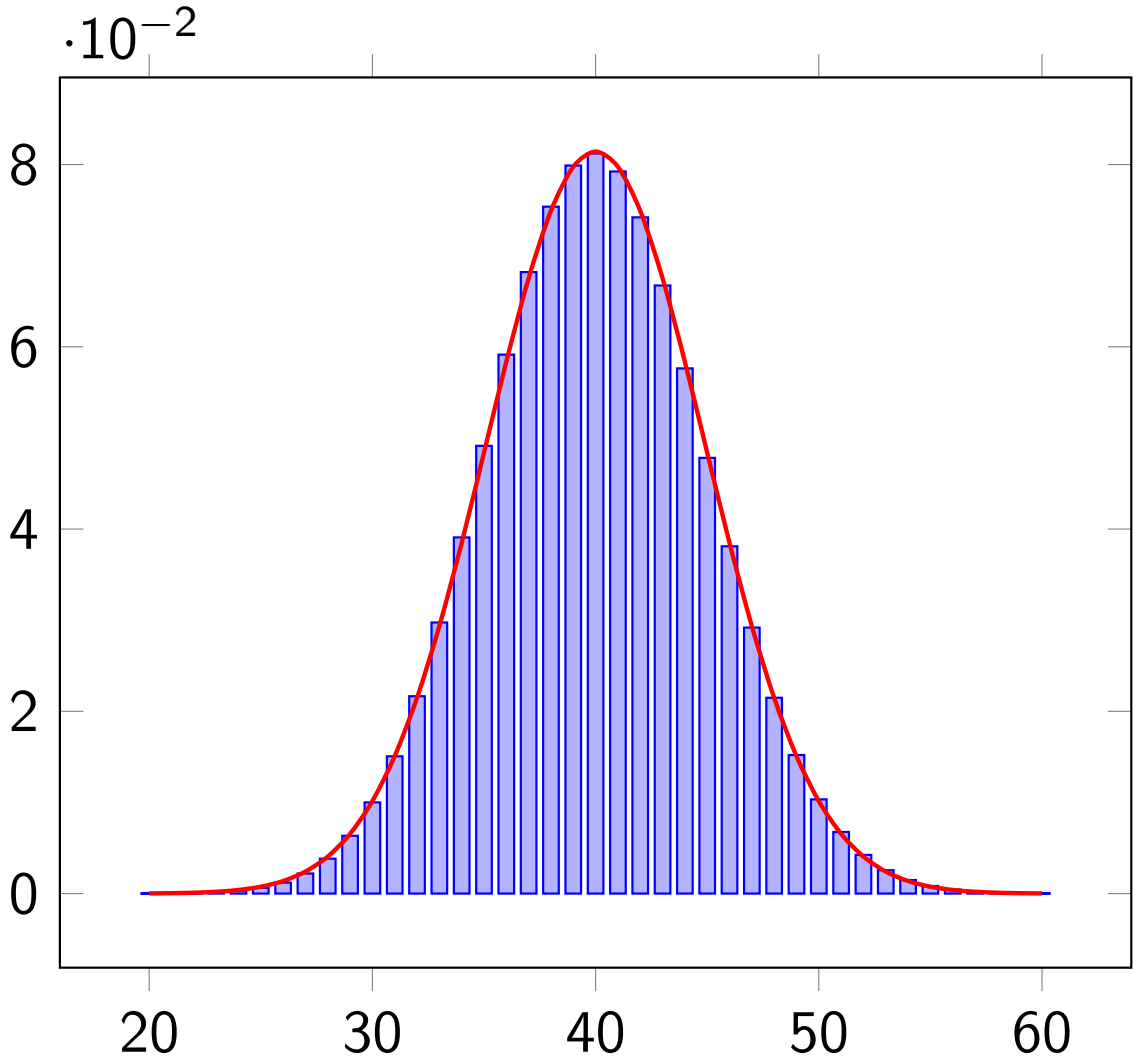
wobei $Y \sim \mathcal{N}(0,1)$ ist. Beachte:

$$8 + \sqrt{20 \cdot 0.24}\, Y \sim \mathcal{N}(8, 4.8).$$

Verbesserung der Approximation: n größer wählen. Für $n = 100$:

$$Z \approx 40 + 4.9\, Y \sim \mathcal{N}(40, 24)$$

Binomialverteilung mit $p = 0.4, n = 100$, Normalverteilungsdichte $\mathcal{N}(40, 24)$



Anwendung: Normalapproximation der Binomialverteilung

Sei $Z \sim \mathrm{Bin}(n, p)$. Dann ist für n hinreichend groß die Zufallsvariable

$$\frac{Z - \mathbb{E}[Z]}{\sqrt{\mathbb{V}(Z)}} = \frac{Z - np}{\sqrt{np(1-p)}}$$

annähernd normalverteilt, also gilt für $a, b \in \{0, \ldots, n\}$

$$\mathbb{P}(a \leq Z \leq b) \approx \Phi_{0,1}\Big(\frac{b - np}{\sqrt{np(1-p)}}\Big) - \Phi_{0,1}\Big(\frac{a - np}{\sqrt{np(1-p)}}\Big).$$

- Herleitung aus zentralem Grenzwertsatz: $Z = \sum_{i=1}^{n} X_i$, für (X_i) unabhängig, Bernoulli-verteilt.
- Faustregel: Die Approximation ist gut (stimmt bis auf 2-3 Nachkommastellen), falls $np \geq 5$ und $n(1-p) \geq 5$ erfüllt sind.

Anwendung: Normalapproximation der Binomialverteilung

Etwas genauere Approximation:
(Satz 7.3) Sei $Z \sim \text{Bin}(n, p)$. Dann gilt für $a, b \in \{0, ..., n\}$,

$$\mathbb{P}(a \leq Z \leq b) \approx \Phi_{0,1}\Big(\frac{b + 1/2 - np}{\sqrt{np(1-p)}}\Big) - \Phi_{0,1}\Big(\frac{a - 1/2 - np}{\sqrt{np(1-p)}}\Big).$$

- 1/2-Korrektur aus Übergang diskret-stetig:

$$\mathbb{P}(4 \leq Z \leq 8) = \mathbb{P}(3.5 \leq Z \leq 8.5)$$

- (Beispiel 7.3)

Zentraler Grenzwertsatz: Zusammenfassung

- Der zentrale Grenzwertsatz besagt, dass die Zufallsvariablen

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

  ungefähr (für große n) standardnormalverteilt ist.
- Der zentrale Grenzwertsatz gilt universell, also egal welche Verteilung die X_i haben (solange sie unabhängig und identisch verteilt mit endlicher Varianz sind).
- Eine Zufallsvariable, welche als Summe von unabhängigen, identisch verteilten Zufallsvariablen geschrieben werden kann, ist (nach Reskalierung) ungefähr normalverteilt
- Wann immer viele unabhängige Ergebnisse aufsummiert werden, so ist das Ergebnis nach Reskalierung ungefähr normalverteilt.
- Wichtige Grundlage für die Statistik