Stochastik für die Informatik, Vorlesung 15

Inhalt

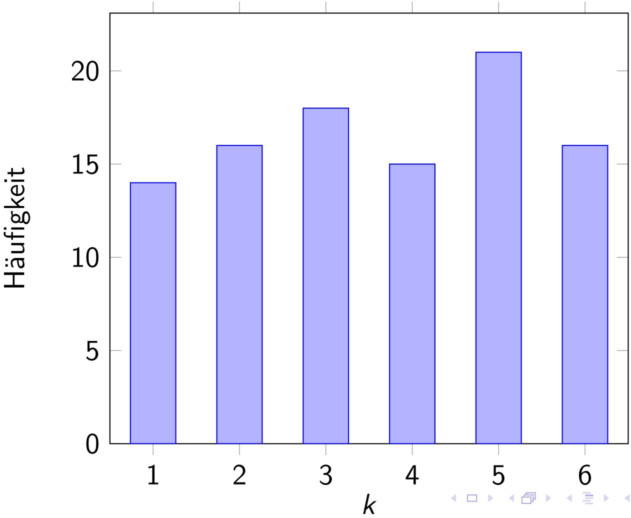
- Gesetz der groen Zahlen und zentraler Grenzwertsatz
- Normalapproximation der Binomialverteilung

Lernziele

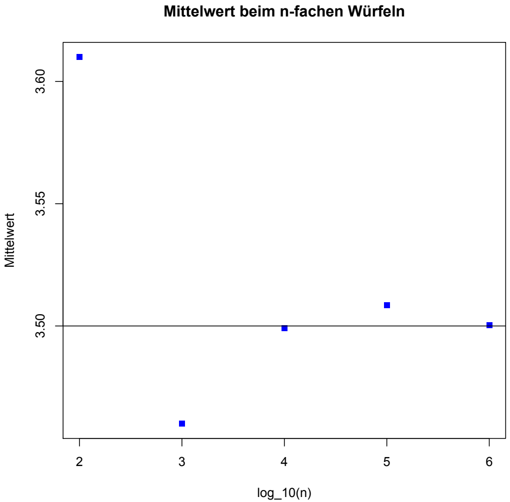
- Das Gesetz der groen Zahlen kennen
- Den zentralen Grenzwertsatz und seine wichtigsten Implikationen kennen
- Die Binomialverteilung mit Hilfe der Normalverteilung approximieren können

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

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

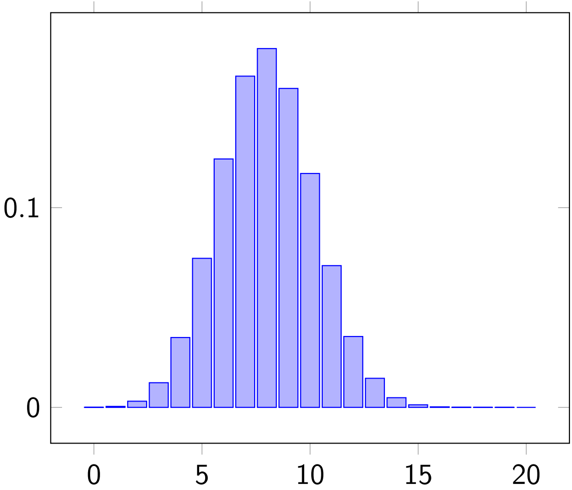
Wie in Kapitel 3 können wir schreiben:

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

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

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

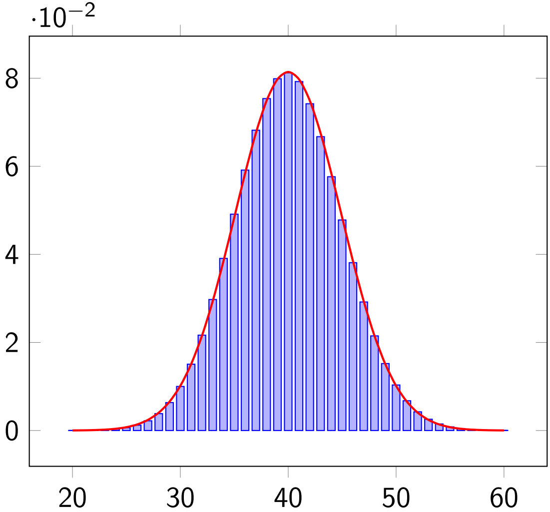
wobei $Y \sim \mathcal{N}(0,1)$ ist. Beachte:

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

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 $a, b \in \{0, \dots, n\}$

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

- 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 $np \ge 5$ und $n(1-p) \ge 5$ erfüllt sind.

Anwendung: Normalapproximation der Binomialverteilung

Etwas genauere Approximation:
(Satz 7.3) Sei $Z \sim \mathrm{Bin}(n, p)$. 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

Beispiel: Wahlergebnisse

Anzahl Wahlbüros, welche eine bestimmte Prozentzahl für die Partei gemeldet haben. x_i : Prozentzahl, welches Wahlbüro Nr. i für Partei X gemeldet hat. Geplottet sind die Häufigkeiten der gemeldeten Prozentzahlen.

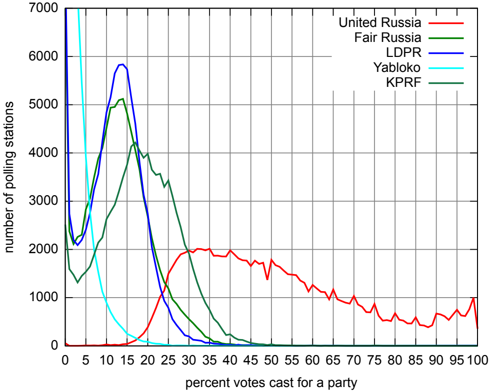


Bild: Wikipedia, Russian legislative election 2011 von Gritzko. Lizenziert unter CC BY-SA 3.0 über Wikimedia Commons, (siehe auch http://en.gazeta.ru/news/2011/12/13/a_3926402.shtml)

Beispiel: Wahlergebnisse

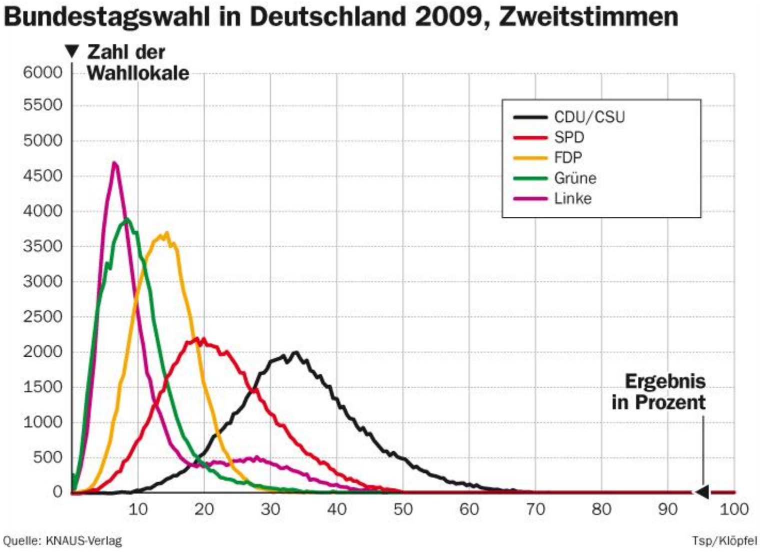


Bild: Tagesspiegel, http://www.tagesspiegel.de/wissen/wahlbetrug-erkennen-wir-glauben-gauss/8822834.html

Kapitel 8: Parameterschätzung

Wahrscheinlichkeitstheorie: Allgemeine Theorie angewandt auf konkrete Modelle.

Woher kommt das konkrete Modell?

Statistik: Durch Analyse und Interpretation von Daten aus Beobachtungen können Modelle aufgestellt, getestet und kalibriert werden

Statistik: Grundproblem

Gegeben: Große Anzahl von Messwerten (Daten) $x_1, ..., x_n$ mit $x_i \in \mathbb{R}$

Allgemeines Ziel der Statistik: Aufstellen eines mathematischen Modells, welches diese Daten beschreibt, und welches mit wahrscheinlichkeitstheoretischen Methoden untersucht werden kann.

Grundaufgaben:

- Schätzer, Bestimmung von Kenngrößen
- Konfidenzintervalle
- Hypothesentests

Grundannahmen der Statistik:

Betreffend der gemessenen Daten $x_1, \ldots, x_n$ gehen wir von einer der beiden (sich nicht ausschließenden) Grundannahmen aus:

- Die gemessenen Daten sind einzelne Realisierungen von (unabhängigen, identisch verteilten) Zufallsvariablen $X_1, \ldots, X_n$
- Die gemessenen Daten stellen eine Stichprobe aus einer (noch viel größeren) Population dar.

Unter dieser Prämisse will man mittels der Stichprobe Aussagen über die zugrundeliegende Zufallsvariablen bzw. über die gesamte Population machen

Beschreibende Statistik

Beispiel 8.1: Eine Messreihe. Messung der Zeit bis zur Betriebsbereitschaft eines elektronischen Geräts (in Sekunden)

Messung Nr.	1	2	3	4	5	6	7	8
Wert	10.9	6.8	9.5	6.9	8.2	3.4	6.2	8.6

Messung Nr.	9	10	11	12	13	14
Wert	5.3	10.7	8.1	8.0	8.9	10.7

- Wie können solche Daten geeignet dargestellt werden?
- Welche Informationen können aus diesen Daten abgelesen werden?
- Um welche "Art" von Daten handelt es sich hier, und inwiefern sind sie mit unseren Grundannahmen kompatibel?
- Welche weiterführenden Fragestellungen ergeben sich möglicherweise?

Häufigkeiten

(Def. 8.1) Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten). Sei $x \in \mathbb{R}$. Die absolute Häufigkeit von x ist

$$H(x) := |\{i : x_i = x\}|,$$

d.h. sie gibt an, wie oft der Wert x im Vektor $(x_1, ..., x_n)$ vorkommt. Die relative Häufigkeit von x ist

$$h(x) := \frac{H(x)}{n}.$$

- $H(x) \in \mathbb{N}_0, h(x) \in [0, 1]$.
- (Beispiel 8.1)
- (Bem. stetige und diskrete Merkmale)

(Def. 8.2) Ein Histogramm der Daten ist ein Plot der Funktion $x \mapsto H(x)$ oder $x \mapsto h(x)$, oder, im Falle einer Einteilung in Klassen, der Funktion $A \mapsto H(A)$ bzw. $A \mapsto h(A)$.

Kenngrößen von Daten

(Def. 8.4) Sei $(x_1, \dots, x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Das empirische Mittel von $(x_1, \dots, x_n)$ ist definiert als

$$\bar{\mu}_n(x_1, \dots, x_n) = \bar{\mu}_x := \frac{1}{n} \sum_{i=1}^{n} x_i$$

(Def. 8.5) Sei $(x_1, \dots, x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Der Median von $(x_1, \dots, x_n)$ ist definiert als der Wert in der Mitte der geordneten Liste. Falls n gerade ist, wird der Durchschnitt der beiden mittleren Werten gebildet.

(Def. 8.6) Sei $(x_1, \dots, x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Die empirische Varianz von $(x_1, \dots, x_n)$ ist definiert als

$$\bar{\sigma}_n^2(x_1, \dots, x_n) = \bar{\sigma}_x^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\mu}_x)^2$$

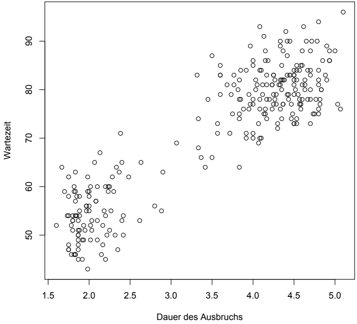
Kenngrößen von Daten

- Empirisches Mittel: Durchschnittswert
- Empirische Varianz: Maß für die Streuung

R-Befehle:

- mean() empirisches Mittel
- var() empirische Varianz
- median() Median
- sort() Liste aufsteigend sortieren
- hist() Zeichnet Histogramm.

Beispiel 8.13: R-Datensatz



- Daten von zwei gleichzeitig gemessenen Größen: Paare von Messwerten $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Grundannahme: Realisierungen zweier Zufallsvariablen X und Y.

Beispiel 8.13: R-Datensatz

Beispieldatensatz in R: Old faithful Geysir, Yellowstone

- Aufrufen mit Befehl `faithful`
- Dauer des Ausbruchs und Wartezeit zwischen Ausbrüchen in Minuten
- 272 Datenpaare
- Beispiel: Berechnung von empirischem Mittel und empirischer Varianz von Dauer und Wartezeit:

```
> data<-faithful
> x<-faithful$eruptions
> y<-faithful$waiting
> mx<-mean(x)
> my<-mean(y)
> vx<-var(x)
> vy<-var(y)
```

- Ergebnis: $\bar{\mu}_x = 3.487783, \bar{\mu}_y = 70.89706$, $\bar{\sigma}_x^2 = 1.302728, \bar{\sigma}_y^2 = 184.8233$

Parameterschätzung

Grundprinzip der Parameterschätzung: Aus den gemessenen Daten die Kenngrößen der (unbekannten) zugrundeliegenden Verteilung schätzen. Dafür müssen gewisse Annahmen getroffen werden (z.B. Unabhängigkeit).

Es gibt viele verschiedene Methoden für die Parameterschätzung. Wir lernen klassische Schätzer für wichtige Kenngrößen kennen, sowie die Methode der "Maximum Likelihood"-Schätzung.

Schätzfunktion

(Def. 8.7) Seien $X_1, ..., X_n$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$. Eine Schätzfunktion zur Stichprobengröße n ist eine Funktion

$$\theta_n : \mathbb{R}^n \to \mathbb{R},\ (X_1, ..., X_n) \mapsto \theta_n(X_1, ..., X_n).$$

- Für praktische Zwecke sollte eine Schätzfunktion einen Zusammenhang mit einem Parameter oder einer Kenngröße der Verteilung der $X_1, ..., X_n$ haben
- Ist dieser Parameter unbekannt, so erhält man, nach Messung der Daten $x_1, ..., x_n$ als Realisierungen von $X_1, ..., X_n$ einen Schätzer oder Schätzwert für den Parameter.
- Damit die Schätzfunktion und der Schätzer nützlich sind, sollten sie gewisse günstige Eigenschaften haben.
- (Bem. Statistisches Modell)

Klassische Schätzer

Beispiel 8.3: Empirisches Mittel. Seien $(x_1, \ldots, x_n) \in \mathbb{R}^n$ Messwerte. Das empirische Mittel (vgl. Def. 8.4) ist definiert als

$$\bar{\mu}_x = \bar{\mu}_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Sind die x_i Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$, so gilt mit dem Gesetz der großen Zahlen

$$\bar{\mu}_n(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mathbb{E}[X_i] = \mu$$

- $\bar{\mu}_x$ ist ein Schätzwert für den Erwartungswert der zugrundeliegenden Zufallsvariablen.

Klassische Schätzer

Beispiel 8.4: Empirische Varianz. Seien $(x_1, ..., x_n) \in \mathbb{R}^n$ Messwerte. Die empirische Varianz (vgl. Def. 8.6) ist definiert als

$$\bar{\sigma}_x^2 = \bar{\sigma}_n^2(x_1, ..., x_n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\mu}_n(x_1, ..., x_n))^2$$

- $\bar{\sigma}_x^2$ ist ein Schätzwert für die Varianz der zugrundeliegenden Zufallsvariablen.

Eigenschaften von Schätzern

(Def. 8.7) Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor von Daten, welche als Realisierungen von identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit Parameter θ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$. Sei θ_n eine Schätzfunktion zur Stichprobengröße n.

- θ_n ist ein erwartungstreuer Schätzer für θ, falls

$$\mathbb{E}[\theta_n(X_1, ..., X_n)] = \theta$$

  ist (engl: unbiased).
- θ_n ist ein konsistenter Schätzer für θ, falls

$$\lim_{n\to\infty} \theta_n(X_1, ..., X_n) = \theta$$

  gilt.
- θ_n ist ein effizienter Schätzer für θ, falls

$$\lim_{n\to\infty} \mathbb{V}(\theta_n(X_1, ..., X_n)) = 0$$

  gilt.
- (Beispiele)