Gesetz der grossen Zahlen
Zentraler Grenzwertsatz

Peter Büchel

HSLU I

Stat: Block 04

Unabhängigkeit und i.i.d. Annahme

- Wichtig für kommende Theorie: i.i.d.-Annahme
- Wenn Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind und alle *dieselbe* Verteilung haben, dann schreibt man

$$X_1, \ldots, X_n \text{ i.i.d.}$$

- Abkürzung i.i.d.: **i**ndependent, **i**dentically **d**istributed
- Was heisst das?

Repetition Unabhängigkeit

- Zwei Zufallsvariablen X_1 und X_2 heissen *stochastisch unabhängig*, falls die W'keiten $P(X_2 = x_2)$ nicht von den W'keiten $P(X_1 = x_1)$ abhängen und umgekehrt

- Beispiel: Werfen faire Münze zweimal nacheinander
  - ZV X_1: Zahl werfen 0, Kopf werfen 1 beim 1. Wurf
  - ZV X_2: Zahl werfen 0, Kopf werfen beim 2. Wurf
  - W'keit Kopf zu werfen im 1. Wurf hat keinen Einfluss auf die W'keit Zahl zu werfen im 2. Wurf
  - Dies allerdings nur richtig, wenn Münze ideal
  - Reale Münze: Durch Aufprall minimalste Veränderungen
  - Diese haben Einfluss auf die Wurfw'keit für Kopf (oder Zahl) beim nächsten Wurf
  - Veränderungen aber so klein, dass sie vernachlässigbar sind

Beispiel für Abhängigkeit

- In Topf 20 Lose mit 5 Gewinnen
- Ziehen zweimal hintereinander *ohne Zurücklegen*
- ZV X_1: Gewinn beim ersten Ziehen 1, Niete 0
- ZV X_2: Gewinn beim zweiten Ziehen 1, Niete 0
- Diese beiden ZV sind *nicht* stochastisch unabhängig
- Ziehen beim ersten Ziehen ein Gewinnlos: W'keit, dass $X_1 = 1$ eintrifft:
$$P(X_1 = 1) = \frac{5}{20}$$

- Bei 2. Ziehung fehlt ein Gewinn: W'keit dann zu gewinnen:

$$P(X_2 = 1) = \frac{4}{19}$$

- Ziehen ersten Ziehung Niete: W'keit bei der 2. Ziehung zu gewinnen:

$$P(X_1 = 0) = \frac{15}{20} \qquad \text{und} \qquad P(X_2 = 1) = \frac{5}{19}$$

- $P(X_2 = 1)$ hängt also von $P(X_1 = x_1)$ ab

- Die ZV sind also nicht stochastisch unabhängig

Gleiche Verteilung

- $X_1, \ldots, X_n$ haben dieselbe Verteilung
- Beispiel: $X_1, \ldots, X_n$ sind alle normalverteilt mit dem gleichen μ und σ
- Beispiel: X_i bezeichnet das i-te Los und hat den Wert 1 bei einem Gewinn, sonst 0
- Also ist $X_i \sim$Bernoulli(π) und $X_1, \ldots, X_n$ i.i.d., da ein Gewinn unabhängig von den anderen Losen ist.
- Etwas salopp für i.i.d.: Es wird dasselbe unter denselben Bedingungen mehrmals gemessen

Empirirische Illustration Gesetz der grossen Zahlen

- Betrachten zwei Situationen:
  - Werfe *10 Würfel*

  - Werfe *40 Würfel*

- X_i: Augenzahl des i-ten Würfels
- Erwartungswert:

$$\mu = \mathrm{E}[X_i] = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

- In einem Durchgang wird einmal mit allen 10 und einmal mit allen 40 Würfeln gewürfelt
- Notieren *Augensumme* für $n \in \{10, 40\}$

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

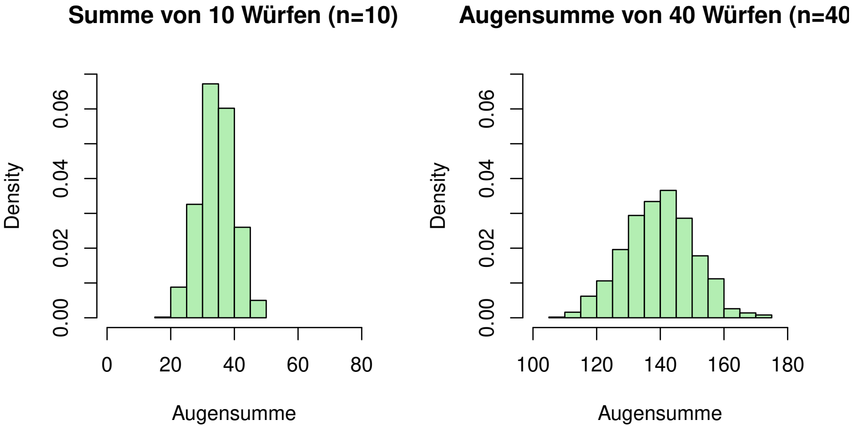
- Notieren *mittlere Augenzahl* für $n \in \{10, 40\}$:

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Führen Versuch 1000mal durch
- Siehe Jupyter-Notebook: `gesetz_grosse_zahlen_1.ipynb`

Summe

- Skizze:



Feststellung: Erwartungswert

- Für $n = 10$ Würfe:
  - Summenzahlen konzentrieren sich um 35
  - Etwa das 10-fache des Erwartungswertes von 3.5
- Für $n = 40$ Würfe:
  - Summenzahlen konzentrieren sich um 140
  - Etwa das 40-fache des Erwartungswertes von 3.5
- Vermutung:

$$E(S_n) = n\mu$$

Feststellungen: Streuung

- Streuung bei $n = 40$ grösser als bei $n = 10$
- Genauer: Streuung bei $n = 40$ *doppelt so gross* wie bei $n = 10$
- Vermutung: Vervierfachung der Würfe verdoppelt die Streuung

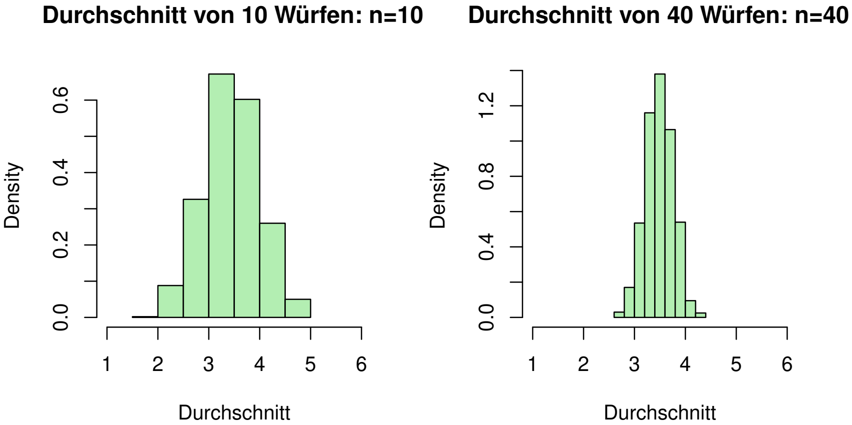
$$\sigma_{S_{40}} = 2\sigma_{S_{10}}$$

- Oder: Vervierfachung der Würfe vervierfacht die Varianz:

$$\operatorname{Var}(S_{40}) = 4\operatorname{Var}(S_{10})$$

Durchschnitt

- Skizze:



Feststellungen

- Durchschnitte für $n = 10$ und $n = 40$ konzentrieren sich um 3.5
- Vermutung:

$$\mathrm{E}(\overline{X_n}) = \mu$$

- Streuung bei $n = 40$ kleiner als bei $n = 10$
- Genauer: Streuung bei $n = 40$ *halb so gross* wie bei $n = 10$
- Vermutung: Vervierfachung der Würfe verdoppelt die Streuung

$$\sigma_{\overline{X}_{40}} = \frac{1}{2}\sigma_{\overline{X}_{10}}$$

- Oder: Vervierfachung der Würfe vervierfacht die Varianz:

$$\mathrm{Var}(\overline{X}_{40}) = \frac{1}{4}\,\mathrm{Var}(\overline{X}_{10})$$

Warum wird Streuung kleiner?

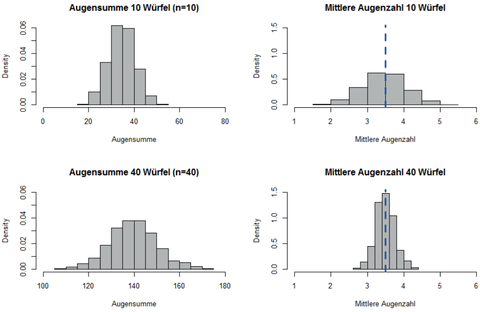
- Histogramm $n = 10$: Durchschnitte um 1.5 kommen vor
- Histogramm $n = 40$: Durchschnitte um 1.5 kommen *nicht* vor
- Grund: Durchschnitt von 1.5 (sehr viele Einer) ist für $n = 40$ sehr viel unwahrscheinlicher als für $n = 10$
- Gleiches gilt für Durchschnitte um 5 (sehr viele Sechser oder hohe Zahlen)

Simulationsresultate: Zusammenfassung

Die Anzahl Summanden wird 4 Mal so gross

$\sigma_{S_{10}}$

$\sigma_{\overline{X}_{10}}$



$\sigma_{S_{40}} = 2 \cdot \sigma_{S_{10}}$

$\sigma_{\overline{X}_{40}} = \frac{1}{2} \cdot \sigma_{\overline{X}_{10}}$

Standardabweichung der Summe verdoppelt sich.

Standardabweichung des arithm. Mittels halbiert sich.

Schlussfolgerung

- Je grösser n, desto grösser wird die Streuung der *Augensumme*
- Für *durchschnittliche* Augenzahl wird Streuung aber kleiner
- Man ist immer *näher am Erwartungswert* (hier 3.5)
- Intuitiv: *Wenn man über viele Beobachtungen mittelt, wird man immer genauer*
- D.h. für n sehr gross ist das arithm. Mittel $\overline{X}_n$ sehr nahe am Erwartungswert
- Diese Aussage heisst *Gesetz der grossen Zahlen* (GGZ)

Kennzahlen von S_n (ohne Herleitung)

Kennzahlen von S_n

Für $X_1, X_2, \ldots, X_n$ i.i.d gilt

$$\mathrm{E}(S_n) = n\mu$$

$$\mathrm{Var}(S_n) = n\sigma_X^2$$

$$\sigma(S_n) = \sqrt{n}\sigma_X$$

mit

$$\mu = \mathrm{E}(X_i) \qquad \text{und} \qquad \sigma_X = \sigma_{X_i}$$

Bemerkung: Da $X_1, X_2, \ldots, X_n$ i.i.d. haben alle gleiche μ und σ

Kennzahlen von $\overline{X}_n$ (ohne Herleitung)

Kennzahlen von $\overline{X}_n$

Für $X_1, X_2, \ldots, X_n$ i.i.d gilt

$$E(\overline{X}_n) = \mu$$

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma_X^2}{n}$$

$$\sigma(\overline{X}_n) = \frac{\sigma_X}{\sqrt{n}}$$

Standardabweichung von $\overline{X}_n$ heisst *Standard-Fehler* des arith. Mittels

Beispiel: Wartezeiten

- Bus fährt alle 8 Minuten
- Gehen *zufällig* an Haltestelle (ohne auf Uhr zu schauen)
- ZV X: Wartezeiten an einer Bushaltestelle
- X uniform verteilt:
$$X \sim \text{Unif}(0, 8)$$
- Wie gross ist der Erwartungswert für die gesamte Wartezeit an 20 Tagen?

Lösung

- ZV X_i: Wartezeit am i-ten Tag
- ZV $X_1, \ldots, X_{20}$ i.i.d
- Es gilt (siehe Folie 43 Block 3):

$$\mu = \mathrm{E}(X_i) = \frac{a+b}{2} = \frac{8+0}{2} = 4$$

- Es gilt:

$$\mathrm{E}(S_{20}) = 20\mu = 20 \cdot 4 = 80$$

- Man muss insgesamt mit einer Wartezeit von 80 Minuten rechnen

Gesetz der grossen Zahlen

Für $n \to \infty$ geht die Streuung gegen null. Es gilt das *Gesetz der grossen Zahlen:* Falls $X_1, \ldots, X_n$ i.i.d., dann

$$\overline{X}_n \longrightarrow \mu \quad \text{für} \quad n \to \infty$$

- Standardabweichung des arith. Mittels (*Standardfehler*) ist *nicht* proportional zu $1/n$ $\quad\to\quad$ Nimmt nur mit Faktor $1/\sqrt{n}$ ab:

$$\sigma_{\overline{X}_n} = \frac{1}{\sqrt{n}}\sigma_X$$

- Um den Standardfehler zu halbieren, braucht man also *viermal* so viele Beobachtungen
- Dies nennt man auch das $\sqrt{n}$-*Gesetz*

Illustration ZGWS: Akkumulation von Messfehlern

- Betrachten eine Messung, die aus der *Summe* von mehreren Einzelmessungen besteht
- Beispiel: Auf einer Baustelle wird täglich die Arbeitsdauer eines Arbeiters gemessen, um die totale Zeit für seinen Arbeitsauftrag zu bestimmen
- Arbeiter gehen *zufällig* nach Hause (schauen nicht auf Uhr)
- Jede Einzelmessung werde gerundet, also liegt der Messfehler einer Einzelmessung zwischen -0.5 und 0.5 (Stunden)
- Modellierung des Messfehler U_j der j-ten Messung mit einer Uniformen Verteilung mit Parametern $a = -0.5$ und $b = 0.5$

Illustration: Akkumulation von Messfehlern

- Betrachten akkumulierten Fehler über die gesamte Summe der Arbeitszeiten eines Arbeiters
- $U_1 + U_2$: Summe der Messfehler des ersten und zweiten Arbeitstages
- Tabelle

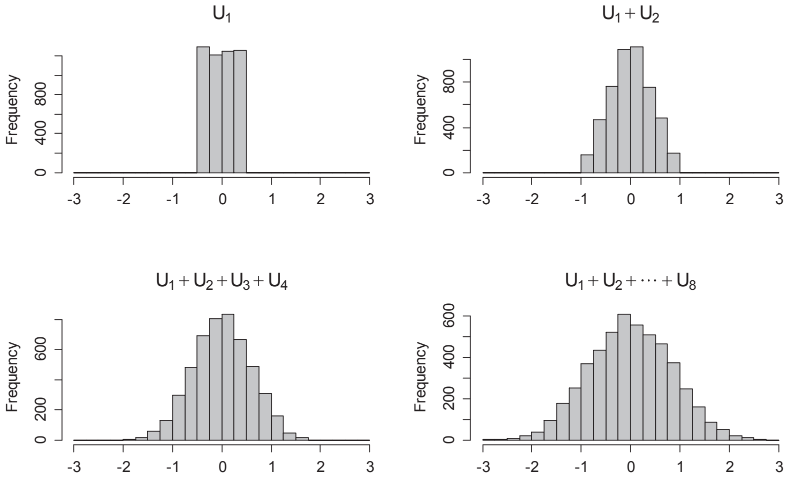
Anzahl Messungen	Messfehler
1 Messung	U_1
2 Messung	$U_1 + U_2$
4 Messung	$U_1 + U_2 + U_3 + U_4$
8 Messung	$\sum_{j=1}^{8} U_j$

- Annahme: Alle U_j voneinander unabhängig und

$$U_j \sim \mathrm{Unif}(-0.5, 0.5)$$

- Jetzt: Situation einer Grossbaustelle mit 5000 Arbeitern
- Für jeden der 5000 Arbeiter Werte U_j simuliert, wobei j-ter Arbeitstag
- U_j für alle Arbeiter i.i.d. (gerechtfertigt?)
  - Wohl eher nicht, da wohl oft mehrere Arbeiter gemeinsam gehen
- Siehe Jupyter-Notebook: `Beispiel_ZGWS_1.ipynb`

Histogramme von simulierten Messfehlern



Form ähnelt immer mehr der Glockenkurve der Normalverteilung

Feststellungen

- Histogramm links oben:
  - Balkenbreite: 0.25 (15 Minuten)
  - Balkenhöhen bei allen Balken in etwa gleich
  - Auch zu erwarten: Arbeiter gehen unabhängig voneinander ohne auf die Uhr zu schauen
- Histogramm rechts oben:
  - Summe der Messfehler: Zwischen -1 und 1
  - Balkenhöhen nicht mehr bei allen Balken
  - Grund: Balken von -1 bis -0.75 kleiner, da dies Arbeiter sind, die zweimal hintereinander etwa eine halbe Stunde vor der vollen Stunde gegangen sind
  - Zu erwarten: Arbeiter geht einmal vor, einmal nach der vollen Stunde

- Histogramm links unten:
  - Summe der Messfehler: Zwischen -2 und 2
  - Balken von -2 bis -1.75 sehr klein: Arbeiter sind, die *zufällig* viermal hintereinander etwa eine halbe Stunde vor der vollen Stunde gegangen sind
  - Eher unwahrscheinlich
  - Zu erwarten: Arbeiter geht einmal vor, einmal nach der vollen Stunde
- Wichtig: *Form* des Histogrammes geht für mehr Arbeitstage immer mehr gegen eine Normalverteilung
- Dies ist so (ohne Herleitung)
- Überraschend:
  - U_j: Uniform verteilt
  - $S_n = U_1 + \ldots + U_n$: Annähernd normalverteilt für grosse n

Zentraler Grenzwertsatz

- Kennzahlen von S_n und $\overline{X}_n$ bereits ermittelt

- Wie aber sind S_n und $\overline{X}_n$ *verteilt*?

- Beispiel mit den Messfehlern in Bezug auf Arbeitszeit:

> S_n ist die Summe von uniform verteilten Zufallsvariablen (Messfehlern) und ist approximiert normalverteilt

- Dies gilt allgemein (ohne Herleitung):
  - $X_1, \ldots, X_n$ i.i.d. (irgendeine Verteilung)
  - Dann ist S_n und $\overline{X}_n$ normalverteilt

- Dies ist die Aussage ders Zentralen Grenzwertssatzes (ZSWS)

- Allgemein gilt der sehr bedeutende:

> **Zentraler Grenzwertsatz**
>
> - Falls $X_1, \ldots, X_n$ i.i.d mit Erwartungswert μ und Varianz σ^2, dann gilt
>
> $$S_n \approx \mathcal{N}(n\mu, n\sigma_X^2) \qquad \overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$
>
> - Approximation wird im Allgemeinen mit grösserem n besser
> - Approximation besser, je näher Verteilung von X_i bei der Normal-Verteilung $\mathcal{N}(\mu, \sigma_X^2)$ ist

- Bemerkung: Wie gross muss n sein, damit die Approximation „gut" ist?
  - Dies kann man allgemein nicht sagen und kommt auf die Problemstellung an

Beispiel: Wartezeit vorher

- Wie gross ist die W'keit, dass die totale Wartezeit in diesen 20 Tagen grösser als 85 Minuten ist?
- Wie gross ist die W'keit, dass man durchschnittlich weniger als 3 Minuten warten muss?

Lösung

- ZV X_i: Wartezeit am i-ten Tag
- Es gilt (schon gesehen, Folie 43, Block 3):

$$\mu = \mathrm{E}(X_i) = \frac{a+b}{2} = \frac{0+8}{2} = 4$$

- Und:

$$\mathrm{Var}(X_i) = \frac{(b-a)^2}{12} = \frac{(8-0)^2}{12} = \frac{64}{12} = \frac{16}{3}$$

- Standardabweichung:

$$\sigma_{X_i} = \sqrt{\mathrm{Var}(X_i)} = \sqrt{\frac{16}{3}} = 2.309$$

- Für die gesamte Wartezeit gilt:

$$S_{20} \approx \mathcal{N}(20\mu, 20\sigma_X^2) = \mathcal{N}\left(20 \cdot 4, 20 \cdot \frac{16}{3}\right) = \mathcal{N}\left(80, \frac{320}{3}\right)$$

- Gesucht W'keit:

$$P(S_{20} \geq 85) = 1 - P(S_{20} \leq 85)$$

```python
import numpy as np
from scipy.stats import norm

1 - norm.cdf(x=85, loc=80, scale=np.sqrt(320/3))
## 0.3141493185879607
```

- W'keit, dass gesamte Wartezeit grösser 85 Minuten ist, ist 0.31415

- Für die durchschnittliche Wartezeit gilt:

$$\overline{X}_{20} \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{20}\right) = \mathcal{N}\left(4, \frac{16/3}{20}\right) = \mathcal{N}\left(4, \frac{16}{60}\right) = \mathcal{N}\left(4, \frac{4}{15}\right)$$

- Gesucht W'keit:

$$P(\overline{X}_{20} \leq 3)$$

```python
import numpy as np
from scipy.stats import norm

norm.cdf(x=3, loc=4, scale=np.sqrt(4/15))
## 0.02640375570805679
```

- W'keit, dass durchschnittliche Wartezeit in 20 Tagen kleiner als 3 Minuten ist, ist 0.026

Zentraler Grenzwertsatz: Roulette

- 18 rote Felder, 18 schwarze Felder, 1 grünes Feld

- Spieler setzt CHF 1 auf rot
- Gewinn des Casinos im *i*-ten Spiel sei X_i
- $X_i = \begin{cases} 1 & \text{W'heit } \frac{19}{37} \quad \text{18 schwarz, 1 grün} \\ -1 & \text{W'heit } \frac{18}{37} \quad \text{18 rot} \end{cases}$
- *Frage*: Was ist die W'keit, dass das Casino Gewinn macht, wenn 10'000 (unabhängige) Spiele betrachtet werden?

Lösung

- Totaler Gewinn nach n Spielen: S_n
- $\mathrm{E}(X_i) = 1 \cdot \frac{19}{37} + (-1) \cdot \frac{18}{37} = \frac{1}{37}$, d.h. Casino leicht im Vorteil
- $\mathrm{E}(X_i^2) = \frac{19}{37} + \frac{18}{37} = 1$
- $\mathrm{Var}(X_i) = E[X_i^2] - (E[X_i])^2 = 1 - \left(\frac{1}{37}\right)^2 = 0.99927 \approx 1$

Lösung

- Erwartungswert:

$$E(S_n) = n \cdot E(X_i) = 10'000 \cdot \frac{1}{37} \approx 270.27$$

- Varianz:

$$\text{Var}(S_n) = n \cdot \text{Var}(X_i) = 10'000 \cdot 0.999927 \approx 9992.7$$

- Standardabweichung:

$$\Rightarrow \sigma_{S_n} = \sqrt{9992.7} \approx 99.96$$

- Annahme: $n = 10000$ gross
- Normalverteilung mit diesem Erwartungswert und dieser Varianz

$$P[S_n > 0] = 1 - P[S_n < 0]$$

- Mit **Python**:

```
from scipy.stats import norm

1 - norm.cdf(x=0, loc=270.27, scale=99.96)
## 0.9965722325091758
```

- Durch den *leichten Vorteil* des Casinos und die *vielen Spiele* reduziert sich das Verlustrisiko sehr stark!
- Wenn Anzahl Spiele erhöht wird, verstärkt sich dieser Effekt und das Casino macht mit hoher W'keit einen (grossen) Gewinn