Gesetz der grossen Zahlen
Zentraler Grenzwertsatz

Peter Büchel

HSLU I

Stoc: Block 04

Lineare Transformation einer Zufallsvariablen

- Fall einer linearen Transformation

$$Y = a + bX \qquad (a,\ b \in \mathbb{R})$$

- *Eigenschaften Lineare Transformationen*

i) $\mathrm{E}(Y) = \mathrm{E}(a + bX) = a + b\,\mathrm{E}(X)$

ii) $\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X), \quad \sigma_Y = |b|\sigma_X$

iii) $\alpha -$ Quantil von $Y = q_Y(\alpha) = a + bq_X(\alpha)$

iv) $f_Y(y) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right)$

Beispiel: Transformation von Grad zu Fahrenheit

- Messung einer Temperatur in Grad Celsius gemessen und Standardabweichung des Messfehlers auf dieser Skala bekannt:

$$\sigma_C = \frac{1}{3}{}^\circ\mathrm{C}$$

- Nun: Temperatur aber nicht in Grad Celsius, sondern in Grad Fahrenheit angeben

- Wie gross ist die Standardabweichung σ_F des Messfehlers, wenn Temperatur in Grad Fahrenheit angeben?

Lösung

- Umrechnung Temperatur T_C in Grad Celsius in die Temperatur T_F in Grad Fahrenheit:

$$T_F = \frac{9}{5} \cdot T_C + 32$$

- Also: Lineare Transformation

$$T_F = b \cdot T_C + a \qquad \text{mit} \qquad b = \frac{9}{5} \quad \text{und} \quad a = 32$$

- Daher ist die Standardabweichung in Grad Fahrenheit

$$\sigma_F = b \cdot \sigma_C = \frac{9}{5}\sigma_C = \frac{9}{5} \cdot \frac{1}{3} = \frac{3}{5}$$

Unabhängigkeit und 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 $\mathrm{i.i.d.}$ steht für: **i**ndependent, **i**dentically **d**istributed

- 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.

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 $n \in \{10, 40\}$

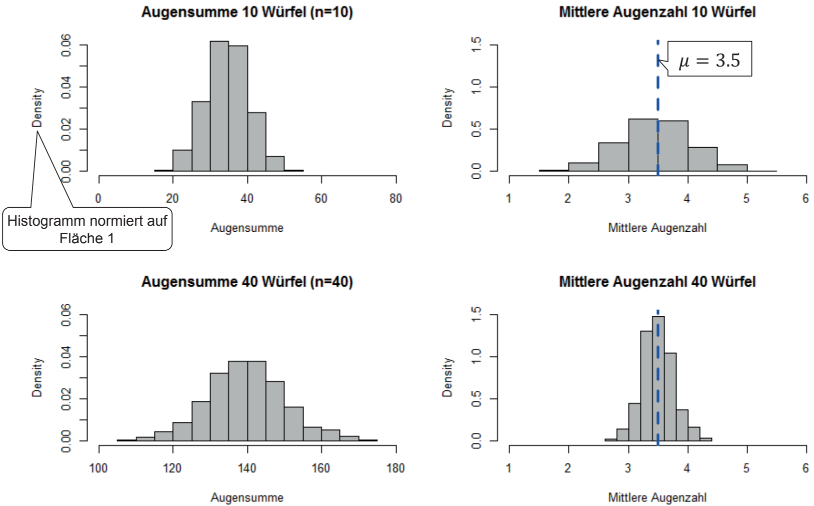
$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

- 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

Simulationsresultate (je 1000 Durchgänge)

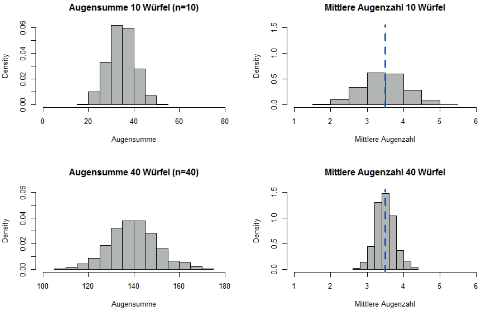


Simulationsresultate (je 1000 Durchgänge)

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

Kennzahlen von S_n

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

$$\begin{aligned}
\mathrm{E}(S_n) &= \mathrm{E}\left(X_1 + X_2 + \ldots + X_n\right) &&= \sum_{i=1}^{n} \mathrm{E}(X_i) &&= n\mu \\
\mathrm{Var}(S_n) &= \mathrm{Var}\left(X_1 + X_2 + \ldots + X_n\right) &&= \sum_{i=1}^{n} \mathrm{Var}(X_i) &&= n\sigma_X^2 \\
\sigma(S_n) &= && &&= \sqrt{n}\sigma_X
\end{aligned}$$

Kennzahlen von $\overline{X}_n$

Kennzahlen von $\overline{X}_n$

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

$$\begin{aligned}
E(\overline{X}_n) &= E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu \\
\operatorname{Var}(\overline{X}_n) &= \operatorname{Var}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i) = \frac{1}{n^2} n\sigma_X^2 = \frac{\sigma_X^2}{n} \\
\sigma(\overline{X}_n) &= \frac{\sigma_X}{\sqrt{n}}
\end{aligned}$$

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

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
- 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	$\sum_{j=1}^{8} U_j$

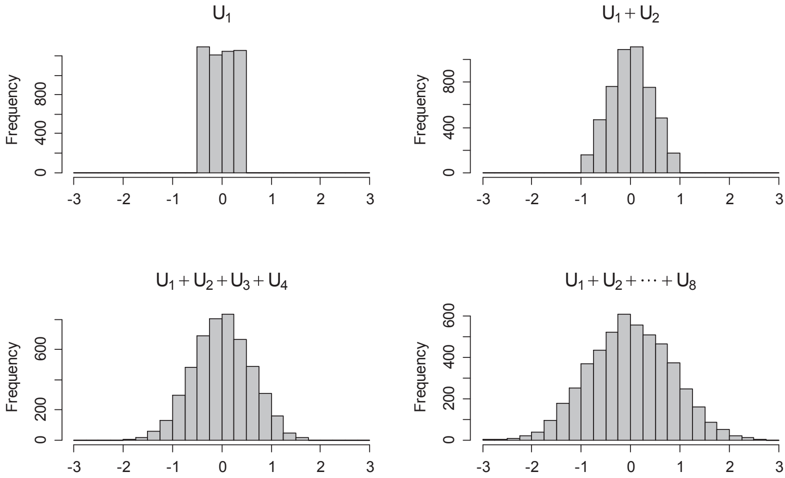
- Annahme: Alle U_j voneinander unabhängig sind und

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

- Jetzt: Situation einer Grossbaustelle mit 5000 Arbeitern

- Für jeden der 5000 Arbeiter werden die Werte U_j simuliert, wobei j den Arbeitstag bezeichnet

Histogramme von simulierten Messfehlern



Form ähnelt immer mehr der Glockenkurve der Normalverteilung

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

- 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

Zentraler Grenzwertsatz: Anmerkungen

- *Binomialverteilung* $\approx$ *Normalverteilung für n gross* und π nicht zu klein (da die Binomialverteilung eine Summe von vielen Bernoulli-Verteilungen ist)

- *Poissonverteilung* $\approx$ *Normalverteilung für* λ *gross* (da Poissonverteilung eine Summe von vielen anderen Poissonverteilungen ist)

Normalapproximation Binomialverteilung

- Wie geht man dann konkret vor?

- Wenn $X \sim \text{Bin}(n, \pi)$, dann gilt

$$\text{E}[X] = n\pi, \quad \text{Var}(X) = n\pi(1-\pi)$$

- Als Approximation Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ mit

$$\mu = n\pi, \quad \sigma^2 = n\pi(1-\pi)$$

- Also:

$$P[X \leq x] \approx \Phi\left(\frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

Beispiel: Normalapproximation

- Wie gross ist die W'keit, dass bei 10 000 Würfen mit einer Münze maximal 5100 mal Kopf erscheint?
- Anzahl Würfe mit Kopf ist $\text{Bin}(10000, 0.5)$-verteilt
- Diese Verteilung mit einer Normalverteilung approximieren:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

mit

$$\mu = 10000 \cdot 0.5 = 5000$$

und

$$\sigma^2 = 10000 \cdot 0.5 \cdot (1 - 0.5) = 2500$$

- Von Interesse ist

$$P(X \leq 5100)$$

- Mit

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

norm.cdf(x=5100, loc=5000, scale=np.sqrt(2500))

## 0.9772498680518208
```

- Vergleichen mit Verteilung:

$$X \sim \mathrm{Bin}(10\,000, 0.5)$$

- Dann gilt

$$P(X \leq 5100) \approx 0.98$$

- Mit

```
from scipy.stats import binom
import numpy as np

binom.cdf(k=5100, n=10000, p=0.5)

## 0.9777871004771368
```

- Die Übereinstimmung ist also sehr gut

Normalapproximation Poissonverteilung

- Wenn $X \sim \text{Pois}(\lambda)$, dann ist

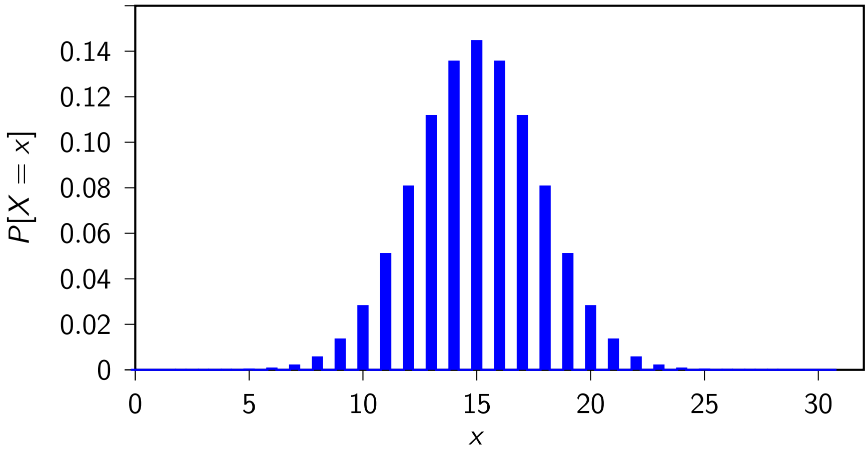
$$\mathrm{E}[X] = \lambda, \quad \mathrm{Var}(X) = \lambda$$

- Approximation durch Normalverteilung mit $\mu = \lambda,\ \sigma^2 = \lambda$

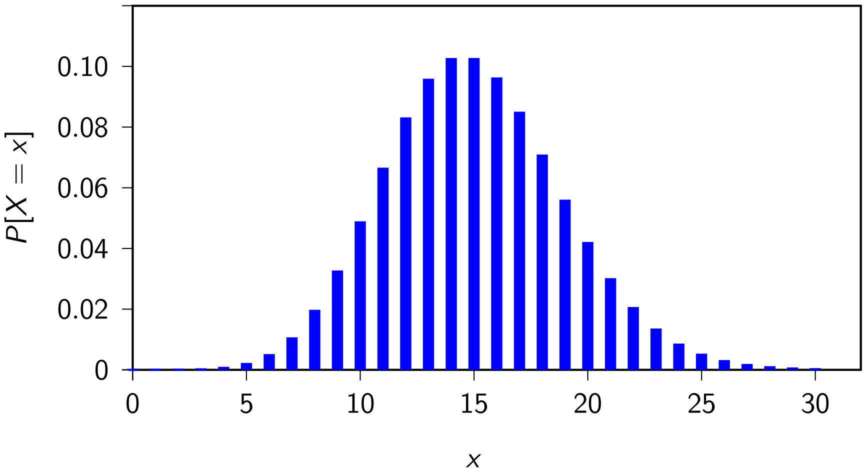
- Also:

$$P[X \leq x] \approx \Phi\left(\frac{x - \lambda}{\sqrt{\lambda}}\right)$$

Binomialverteilung ($n = 30, \pi = 0.5$)



Poissonverteilung ($\lambda = 15$)



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
- $X_i = \begin{cases} 1 & \text{W'heit } \frac{19}{37} \quad \text{18 schwarz, 1 grün} \\ -1 & \text{W'heit } \frac{18}{37} \quad \text{18 rot} \end{cases}$
- Totaler Gewinn nach n Spielen ist S_n

Zentraler Grenzwertsatz: Roulette

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

Zentraler Grenzwertsatz: Roulette

- *Frage*: Was ist die W'keit, dass das Casino Gewinn macht, wenn 10'000 (unabhängige) Spiele betrachtet werden?
- $\mathrm{E}[S_n] = n \cdot \mathrm{E}[X_i] = 10'000 \cdot \frac{1}{37} \approx 270.27$
- $\mathrm{Var}(S_n) = n \cdot \mathrm{Var}[X_i] = 10'000 \cdot 0.999927 \approx 9992.7$
- $\Rightarrow \sigma_{S_n} = \sqrt{9992.7} \approx 99.96$

Zentraler Grenzwertsatz: Roulette

- Annahme: Normalverteilung mit diesem Erwartungswert und dieser Varianz

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

- Mit :

```
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

Fehlerrechnung: Systematische und zufällige Fehler

- Messungen physikalischer Grössen sind grundsätzlich *fehlerbehaftet*
- Man erhält Messwerte, die vom wahren Wert mehr oder weniger abweichen.
- Unterscheidung zwischen *systematischen* und *zufälligen* Fehlern
- Systematische Fehler rühren von der Unvollkommenheit der Messgeräte her
- Beispiel: Funktionsfehler und Eichfehler oder Unvollkommenheit der Messverfahren

Fehlerrechnung: Systematische und zufällige Fehler

- Beispiele von *systematischen Messfehlern*:
  - Bei Kurzschluss der Eingänge zeigt ein Voltmeter nicht mehr 0 V an (Nullpunktsfehler)
  - Durch den Innenwiderstand eines Voltmeters sind gemessene Strom- oder Spannungswerte stets zu klein
  - Bei einem Pendelversuch wird durch die Luft- und Lagerreibung die Schwingung gedämpft, wodurch die Frequenz der Schwingung verringert wird
- Systematische Fehler sollten nach Möglichkeit vermieden oder kleingehalten werden.
- Sie sind jedoch nicht Gegenstand einer Fehlerrechnung

Zufällige Fehler

- Zufällige Fehler: vor allem durch die Naturgesetze selber wie
  - aufgrund der statistischen Natur von Kernzerfällen
  - durch Ungeschicklichkeit beim Messen
  - durch statistisch schwankende äussere und innere Einflüsse (Druck, Temperatur, Luftfeuchtigkeit)
- Messwerte streuen um *arithmetischen Mittelwert*

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Je grösser die Messreihe ist (also je grösser n), um so näher liegt der Mittelwert am wahren Wert und umso kleiner wird der Fehler des Mittelwertes

- Der sogenannte *Standardfehler* ist also:

$$s_{\overline{x}_n} = \frac{s_x}{\sqrt{n}}$$

Darstellung von Fehlern

- Standardfehler ist *absoluter Fehler*, den man im Zusammenhang mit dem arithmetischen Mittel einer Messreihe angibt:

$$\overline{x}_n \pm s_{\overline{x}_n}$$

- *Beispiel:* Umlaufzeit eines Plattentellers mit *absolutem Fehler* des Mittelwertes:

$$T = (1.817 \pm 0.012)\mathrm{s}$$

- Der *relative Fehler* wird folgendermassen angegeben:

$$\overline{x}_n \pm \frac{s_{\overline{x}_n}}{\overline{x}_n} \cdot 100\,\%$$

- *Beispiel:* Umlaufzeit eines Plattentellers mit *relativem Fehler* des Mittelwertes:

$$T = 1.817\,\mathrm{s} \pm \frac{0.012s}{1.817s} = 1.817\,\mathrm{s} \pm 0.66\,\%$$