5. Zentraler Grenzwertsatz

Lernziele:

- Sie verstehen die Bedeutung der Normalverteilung als Grenzverteilung von Summen unabhängiger, identisch verteilter Zufallsvariablen mit existierenden Varianzen.
- Sie können den zentralen Grenzwertsatz anwenden, um Wahrscheinlichkeiten von Summen bzw. Mittelwerten unabhängiger, identisch verteilter Zufallsvariablen näherungsweise zu berechnen.

Literatur:

- Teschl Band 2, Kap. 29.2
- Arens et al., Kap 39.3 ab S. 1489
- Zucchini, Kap. 7.5

5.1 Zentraler Grenzwertsatz (ZGWS)

Situation:

Für n unabhängige identisch verteilte (i.i.d.) ZV X_i seien zwar der Erwartungswert μ und die Varianz σ^2 bekannt, aber nicht die Verteilung.

Kann man dann Wahrscheinlichkeitsaussagen für X_i machen?

Beispiel 5.1.1:

$n =$

Eine Kfz-Versicherung hat 25000 Versicherungsverträge abgeschlossen.

X_i

Der jährliche Schaden pro Vertrag ist eine ZV mit Erwartungswert $\mu = 320$ und Standardabweichung $\sigma = 540$.

Mit welcher Wahrscheinlichkeit liegt die Gesamtschadenssumme in einem Jahr über 8,3 Millionen?

Ges.: $P\left(\sum_{i=1}^{n} X_i > 8.3 \cdot 10^6\right)$

Verteilung von X_i unbekannt, aber $\sum X_i$ ist nach ZGWS näherungsweise normalverteilt!

Geg.: $E[X_i] = 320$, $Var[X_i] = 540^2$

$\overset{\text{ZGWS}}{\Longrightarrow} \quad \sum_{i=1}^{n} X_i \sim N_{n \cdot 320,\, n \cdot 540^2}$

$E\left[\sum_i X_i\right] = \sum_i E[X_i] = n \cdot E[X_i] = 25000 \cdot 320$

$Var\left[\sum_i X_i\right] \overset{\text{unabh.}}{=} \sum_i Var[X_i] = 25000 \cdot 540^2$

$$\Rightarrow P\left(\sum_i X_i > 8.3\cdot 10^6\right) = 1 - P\left(\sum X_i \leq 8.3\cdot 10^6\right)$$

$$= 1 - \text{pnorm}\left(8.3\cdot 10^6,\ 25000\cdot 320,\ \text{sqrt}(25000)\cdot 540\right)$$

$$\approx 0.02\,\%$$

$$\bar{X} = \frac{1}{n}\sum_i X_i \overset{\text{ZGWS}}{\sim} N_{E[\bar{X}],\,Var[\bar{X}]}$$

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n}\cdot n\,\underbrace{E[X_i]}_{=\mu} = \mu$$

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_i X_i\right] \overset{\text{unabh.}}{=} \frac{1}{n^2}\cdot n\,\underbrace{Var[X_i]}_{\sigma^2} = \frac{1}{n}\cdot\sigma^2$$

Beispiel 5.1.2:
Um die Rechengeschwindigkeit eines Computers zu messen, wird die Zeit T, die der Computer für die Abarbeitung eines Musterprogramms benötigt, gestoppt. Durch n-malige Zeitmessung und Durchschnittsbildung als Schätzwert für T wird der Einfluß von Messfehlern abgefangen.
Wie viele Messungen n müssen durchgeführt werden, wenn man annimmt, dass die Messungen i.i.d. ZV mit Erwartungswert $\mu = R$ [s] und Standardabweichung $\sigma = 2$ [s] sind und zu 95% sicher gehen will, dass die Schätzung eine Genauigkeit von ± 0.5 [s] aufweist?

Geg.: $E[X_i] = R$, $Var[X_i] = 4$ $\overset{ZGWS}{\Longrightarrow}$ $\bar{X} \sim N_{R,\frac{4}{n}}$

bzw. $\underbrace{\frac{\bar{X}-R}{2}\sqrt{n}}_{Z} \sim N_{0,1}$

Ges.: n mit $\underbrace{P(-0.5 \leq \bar{X} - R \leq 0.5)}_{P(R-0.5 \leq \bar{X} \leq R+0.5)} \geq 0.95$

Standardisierung

$$P\left(\frac{-0.5}{2}\sqrt{n} \le \underbrace{\frac{\bar{X}-R}{2}\sqrt{n}}_{Z \sim N_{0,1}} \le \frac{0.5}{2}\sqrt{n}\right) \ge 0.95$$

$$P\left(Z \le \frac{\sqrt{n}}{4}\right) - P\left(Z \le -\frac{\sqrt{n}}{4}\right) = 2 \cdot \underbrace{P\left(Z \le \frac{\sqrt{n}}{4}\right)}_{\Phi\left(\frac{\sqrt{n}}{4}\right)} - 1 \ge 0.95$$

$$\Phi\left(\frac{\sqrt{n}}{4}\right) \ge 0.975 \qquad \Big| \ \Phi^{-1}(\ )$$

$$\Rightarrow \quad \frac{\sqrt{n}}{4} \ge \underbrace{\Phi^{-1}(0.975)}_{\text{qnorm}(0.975) \approx 1.96} \quad \Rightarrow \quad \sqrt{n} \ge 4 \cdot 1.96 \quad \Rightarrow \quad n \ge 62$$

Satz 5.1: Zentraler Grenzwertsatz

Seien X_i $(i = 1, \ldots, n)$ unabhängige identisch verteilte (i.i.d.) ZV mit Erwartungswert μ und Varianz σ^2.

Dann gilt für hinreichend große n und $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ näherungsweise

Stichprobe

$$\sum_{i=1}^{n} X_i \sim N_{n\mu, n\sigma^2} \quad \text{und} \quad \frac{\sum X_i - n\mu}{\sqrt{n\sigma}} \sim N_{0,1}$$

$$\bar{X} \sim N_{\mu, \frac{\sigma^2}{n}} \quad \text{und} \quad \underbrace{\frac{\bar{X} - \mu}{\sigma}\sqrt{n}}_{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}} \sim N_{0,1}$$

Bemerkung: Der Satz gilt sogar allgemeiner, wenn die X_i abhängig und nicht identisch verteilt sind, vorausgesetzt kein X_i ist deutlich dominanter als die anderen.

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\cdot\sum_{i=1}^{n} \underbrace{E[x_i]}_{\mu} = \mu$$

$$Var[\bar{x}] = Var\left[\frac{1}{n}\cdot\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2} \sum_{i=1}^{n} \underbrace{Var[x_i]}_{\sigma^2} = \frac{\sigma^2}{n}$$

x_i stochastisch unabh.

Wesentlich bei den Voraussetzungen des ZGWS ist, dass die X_i nicht normalverteilt sein müssen, damit $\sum_{i=1}^{n} X_i$ bzw. $\bar{X}$ bei hinreichend großem n normalverteilt sind.

Faustregel für die Größe von n:

Je schiefer die Verteilung der X_i, desto größer muss n sein.

- $n > 30$, falls die unbekannte Verteilung ohne markanten Ausreißer, aber schief ist (z.B. Exponentialverteilung).
- $n > 15$, falls die unbekannte Verteilung annähernd symmetrisch ist (z.B. Binomialverteilung).
- $n \leq 15$, falls die unbekannte Verteilung annähernd normalverteilt ist.

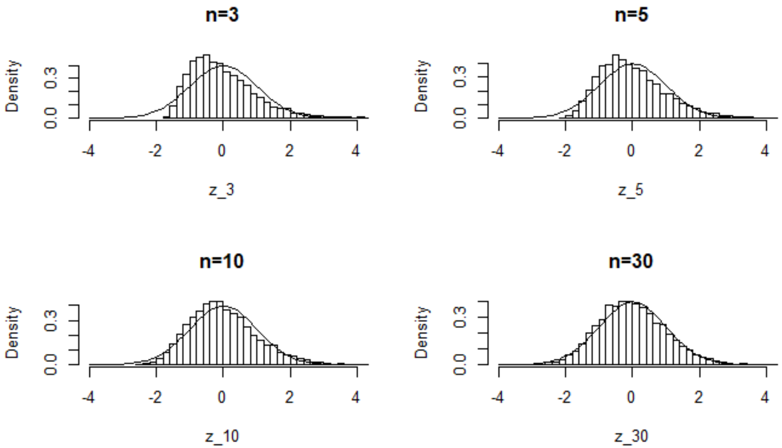

Abbildung: Wahrscheinlichkeitsverteilungen für die standardisierte Summe von n exponentialverteilten ZV mit $\lambda = 1$

Anwendung des ZGWS: Aufgabentypen

Seien X_i i.i.d. ZV mit Erwartungswert μ und Varianz σ^2, aber unbekannter Verteilung. Dann sind $Z_1 = \frac{\sum X_i - n\mu}{\sqrt{n}\sigma}$ und $Z_2 = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ näherungsweise standardnormalverteilt.

- Es lassen sich Wahrscheinlichkeiten für $\sum X_i$, $\bar{X}$, Z_1 bzw. Z_2 berechnen.

  Beispiel 5.1.1 und Übung 7.2

- Es lässt sich n bestimmen, so dass zu vorgegebener Schranke k und Wahrscheinlichkeit p gilt:
  $P(Z_i > k) \geq p$ bzw. $P(-k \leq Z_i \leq k) \geq p$

  Beispiel 5.1.2, Übung 7.1 und 7.3

5.2 Stichprobenverteilungen für normalverteilte Grundgesamtheiten

Literatur: Teschl, Mathematik für Informatiker, Band 2, Kapitel 30.1 + 2

In der schließenden Statistik wird zur Untersuchung von hinreichend großen Grundgesamtheiten eine vergleichsweise kleine Stichprobe vom Umfang n betrachtet und davon auf die Grundgesamtheit geschlossen. Die Stichprobe ist dann eine Folge $X_1, \ldots, X_n$ von i.i.d. Zufallsvariablen.

Definition 5.1: Stichprobenmittel

Die Stichprobenfunktion $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist eine erwartungstreue Schätzfunktion für den Erwartungswert μ, d. h. $E[\bar{X}] = \mu$.

Definition 5.2: Stichprobenvarianz

Die Stichprobenfunktion $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$ ist eine erwartungstreue Schätzfunktion für die Varianz σ^2, d. h. $E[S^2] = \sigma^2$.

Beweis: $E[S^2] = \sigma^2$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \underset{\substack{\text{Verschiebungs-}\\ \text{satz}}}{=} \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\cdot\bar{X}^2\right)$$

$$E[S^2] = \frac{1}{n-1}E\left[\sum X_i^2 - n\cdot\bar{X}^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n}E[X_i^2] - n\cdot E[\bar{X}^2]\right)$$

$$\overset{(*)}{=} \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \frac{1}{n-1}\left(\overbrace{n\sigma^2-\sigma^2}^{(n-1)\cdot\sigma^2}\right) = \sigma^2$$

NR.: Es gilt $\underbrace{Var[X]}_{\sigma^2} = E[X^2] - (\underbrace{E[X]}_{\mu})^2 \Longrightarrow \underset{(*)}{E[X^2] = \sigma^2+\mu^2}$

$$Var[\bar{X}] = E[\bar{X}^2] - (E[\bar{X}])^2 \Longrightarrow E[\bar{X}^2] = \frac{\sigma^2}{n} + \mu^2$$

$\frac{\sigma^2}{n}$ μ

Seien X_i $(i = 1, \ldots, n)$ unabhängige normalverteilte ZV mit Erwartungswert μ und Varianz σ^2. Dann gilt:

siehe Parameterschätzung und Hypothesentests

(1) $\frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N_{0,1}$ bei bekannter Varianz σ^2

(2) $\underbrace{\frac{(n-1)S^2}{\sigma^2}}_{\text{standardisiert}} \sim \chi^2_{n-1}$

(3) $\frac{\bar{X} - \mu}{S} \sqrt{n} \sim t_{n-1}$ bei unbekannter Varianz

Beispiel 5.2.1:
Die Bearbeitungsdauer eines Servers für einen bestimmten Auftragstyp ist normalverteilt mit Erwartungswert $\mu = 20$ [s] und Standardabweichung $\sigma = 3$ [s].
Anhand einer Stichprobe von $n = 15$ Aufträgen soll ermittelt werden, mit welcher Wahrscheinlichkeit die Stichprobenvarianz S^2 größer als 12 [s^2] ist.

Ges.: $P(S^2 > 12)$

Voraussetzung: $\underbrace{X_i}_{\text{Stichprobenwerte}} \sim N_{20,9} \Longrightarrow \frac{(n-1)S^2}{\sigma^2} = \frac{14 \cdot S^2}{9} \sim \chi^2_{14}$

$$\Rightarrow P(S^2 > 12) = P\left(\frac{14}{9}S^2 > \frac{14 \cdot 12}{9}\right) = 1 - P\left(\frac{14}{9}S^2 \leq \frac{56}{3}\right)$$

$$= 1 - \text{pchisq}\left(\frac{56}{3}, 14\right) \approx 17.8\,\%$$