Zenraler Grenzwertsatz

Anton Klimovsky

Zentraler Grenzwertsatz[1]. Konvergenz in Verteilung. Normalapproximation.

In diesem Abschnitt beschäftigen wir uns mit der folgenden Frage.

[1] Everyone believes in the [normal] law of errors: the mathematicians, because they think it is an experimental fact; and the experimenters, because they suppose it is a theorem of mathematics. –Gabriel Lippmann.

Frage: Wie sieht die Verteilung einer Summe der ZVen S_n für großes n aus?

Eine Antwort liefert der zentrale Grenzwertsatz. Dieser ist eine Präzisierung des GGZ.

Beispiel (Soziologische Umfrage). Sei p der Anteil von allen Wählern, die einen bestimmten Kandidaten unterstützen. Der Anteil p ist unbekannt bevor das endgültige Wahlergebnis bekannt ist. Um trotzten eine Idee zu haben darüber, wie groß p ist, machen wir eine Umfrage bei n "zufällig ausgewählten" Wählern[2].

[2] Aus Kostengründen sollen wir n so klein wie möglich halten. Andererseits muss n so gewählt werden, dass wir mit einer hohen Wahrscheinlichkeit eine "gute" Abschätzung von p bekommen.

Wir modellieren die "zufällig ausgewählten" Wählern als n unabhängig und gleichverteilt ausgewählte Personen aus der Gesamtpopulation. D.h. die Antwort jeder ausgewählten Person ist eine unabhängige Bernoulli ZV X_i mit Erfolgswahrscheinlichkeit p und Varianz $\sigma^2 = p(1-p)$.

Wir wollen den empirischen Mittelwert $\bar{X}_n$ als eine Abschätzung von p benutzen. Die Ungleichung von Tschebyscheff liefert

$$\mathbb{P}\{|\bar{X}_n - p| \geq \varepsilon\} \leq \frac{p(1-p)}{n\varepsilon^2}. \tag{1}$$

Die linke Seite von der Ungleichung (1) sieht schon mal gar nicht so schlecht aus. Allerdings gibt es ein Problem mit der Ungleichung (1). Nämlich, dass das p unbekannt ist. Dieses Problem können wir mit der folgenden Beobachtung umgehen

$$\max_{p\in[0,1]} p(1-p) = 1/4. \tag{2}$$

Somit

$$\mathbb{P}\{|\bar{X}_n - p| \geq \varepsilon\} \leq \frac{1}{4n\varepsilon^2}. \tag{3}$$

Z.b. erhalten wir aus (3) für $\varepsilon = 0,1$ und $n = 100$

$$\mathbb{P}\{|\bar{X}_n - p| \geq \varepsilon\} \leq \frac{1}{4 \cdot 100 \cdot (0,1)^2} = 0,25. \tag{4}$$

In Worten sagt uns (4) folgendes. Bei einer Stichprobe vom Umfang $n = 100$ ist die Wahrscheinlichkeit, dass unsere empirische

Abschätzung $\bar{X}_n$ ein Fehler größer $0,1$ hat, kleiner als $0,25$ ist. Anderes ausgedrückt wird in einem viertel aller Fälle ein Fehler von mindestens 10% gemacht[3].

[3] Natürlich gilt dies unter Annahme, dass X_i u.i.v. Bernoulli ZVen sind. Diese Annahme soll sehr wohl bezweifelt werden! In Wirklichkeit können ZV X_i Abhängig sein und müssen nicht unbedingt die gleiche Verteilung haben!

Vorbereitungen

Lemma 0.1. *Sei* $\{X_i\}_{i=1}^{\infty}$ *eine Folge von u.i.v. ZV mit* $\mathrm{Var}[X_1] = \sigma^2 < \infty$ *und* $\mathbb{E}[X] = \mu$. *Es gilt*

$$\mathbb{E}[S_n] = n\mu \tag{5}$$
$$\mathrm{Var}[S_n] = n\sigma^2 \tag{6}$$

Lemma 0.2 (Standardisierung)**.** *Gegeben ist eine ZV X. Definiere*

$$Y := \frac{X - \mathbb{E}[X]}{\mathrm{Var}[X]}. \tag{7}$$

Dann gilt $\mathbb{E}[Y] = 0$ *und* $\mathrm{Var}[Y] = 1$.[4]

[4] Diese werte vom Erwartungswert $= 0$ und der Varianz $= 1$ heißen Standardwerte.

Zentraler Grenzwertsatz: Formulierung

Frage. Wie sieht die Verteilung einer standardisierten Summe S_n von u.i.v. ZVen **für großes** $n \in \mathbb{N}$ aus?

Theorem 0.1 (zentraler Grenzwertsatz)**.** *Sei* $\{X_i\}_{i=1}^{\infty}$ *eine Folge von u.i.v. ZV mit* $\mathrm{Var}[X_1] = \sigma^2 < \infty$ *und* $\mathbb{E}[X] = \mu$. *Sei* $\sigma^2 \neq 0$. *Dann gilt für alle* $x \in \mathbb{R}$

$$\boxed{\mathbb{P}\left\{\frac{S_n - \mu n}{\sigma\sqrt{n}} \leq x\right\} \underset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2} \mathrm{d}y =: \Phi(x).} \tag{8}$$

In Wörtern besagt der zentrale Grenzwertsatz, dass die Verteilungsfunktion der standardisierten Summe S_n (linke Seite) gegen die *Verteilungsfunktion der Normalverteilung*[5] mit dem Erwartungswert 0 und Varianz 1 (rechte Seite) konvergiert, als n groß wird.

[5] Die Verteilungsfunktion ist nicht in einer geschlossenen Form darstellbar. Es gibt aber sehr detaillierte Tabellen bzw. Standardbefehle in ihrer Lieblings-Programmiersprache, die diese Funktion präzise ausrechnen.

Normalapproximation. Der zentrale Grenzwertsatz ist eine relativ allgemeine Aussage. Wir haben nichts außer Unabhängigkeit, Identischverteiltheit und Existenz der endlichen Varianz für die Summanden $\{X_i\}_{i=1}^{\infty}$ angenommen. Unter diesen Voraussetzungen sagt uns der zentrale Grenzwertsatz, dass die standardisierte Summe S_n ist im Grenzwert $n \to \infty$ *standard-nomalverteilt*. Genauer gesagt gilt folgendes:

1. Die Verteilung von S_n konzentriert sich um $n\mu$. (Dies kennen wir schon aus dem Gesetz der großen Zahlen.)

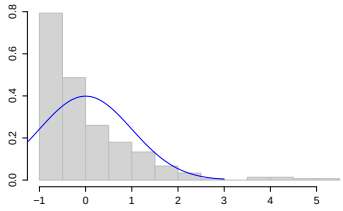

Abbildung 1: Histogramm einer standardisierten Stichprobe von 300 Realisierungen aus der Exponentiell-Verteilung (graue Balken, **linkssteil**) verglichen mit der Standardnormalverteilung (blaue Kurve).

2. Die Schwankungen (= Fluktuationen) von S_n um $n\mu$ sind von der Größenordnung $\sqrt{n}$.

3. Die asymptotische ($n \to \infty$) Verteilung dieser Fluktuationen ist Gauß'sch.

Dies gilt (unter den genannten Voraussetzungen) unabhängig von der Verteilung der Summanden!

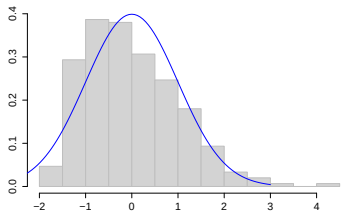

Abbildung 2: Histogramm der standardisierten Summe von **5** unabh. Exponential-Verteilten ZV (graue Balken, **linksteil**) verglichen mit der mit der Standardnormalverteilung (blaue Kurve)

Informelle Prozedur der Normalapproximation. Seien die Voraussetzungen vom Theorem 0.1 erfüllt. Für n groß genug kann man die Wahrscheinlichkeit $\mathbb{P}\{S_n \leq x\}$ *approximativ* ausrechnen in dem man S_n *approximativ* als Gauß'sch betrachtet! Und zwar wie folgt.

1. Berechne den Erwartungswert $n\mu$ und die Varianz $n\sigma^2$ von S_n.

2. Berechne den *standardisierten Wert* von x

$$z := \frac{x - n\mu}{\sigma\sqrt{n}}. \tag{9}$$

3. Benutze die folgende Approximation[6]:

$$\mathbb{P}\{S_n \leq x\} \approx \Phi(z). \tag{10}$$

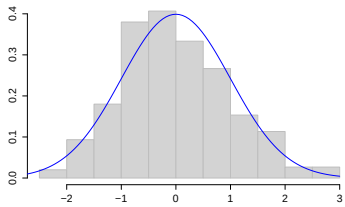

Abbildung 3: Histogramm der standardisierten Summe von **20** unabh. exponential-verteilten ZV (graue Balken, **etwas linksteil**) verglichen mit der mit der Standard-Normalverteilung (blaue Kurve)

Bemerkung. ZGS kann zusammenbrechen falls einer von den Voraussetzungen schwer verletzt sind:

- Es kann sein, dass die ZVen $\{X_i\}_{i=1}^{\infty}$ abhängig sind.
- Es kann sein, dass die Varianz von X_1 unendlich[7] ist.
- Es kann sein, dass die ZVen $\{X_i\}_{i=1}^{\infty}$ nicht identisch-verteilt sind.

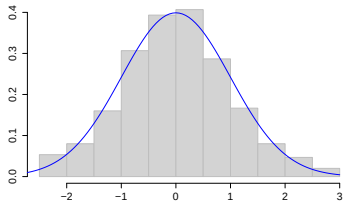

Abbildung 4: Histogramm der standardisierten Summe von **50** unabh. exponential-verteilten ZV (graue Balken, **fast Symmetrisch**) verglichen mit der mit der Standard-Normalverteilung (blaue Kurve)

Die Normalapproximation kann außerdem schlecht sein, falls n klein ist, oder falls die Verteilung von X_1 asymmetrisch oder multimodal oder diskret ist, oder falls x weit weg von der "Hauptmasse" der Verteilung von X_1 ist.

Beispiel (Soziologische Umfrage: Fortsetzung). Nun wollen wir $\mathbb{P}\{|\bar{X}_n - p| \geq \varepsilon\}$ mit Hilfe vom ZGS (approximativ!) abschätzen. Wir verfahren, wie oben bei der Normalapproximation beschrieben ist. Erstens wollen wir die Wahrscheinlichkeit, für die wir uns interessieren, über die Wahrscheinlichkeit auf der linken Seite vom (10) darstellen. Es gilt

$$\bar{X}_n - p = \frac{1}{n}\sum_{i=1}^{n}(X_n - p) = \sum_{i=1}^{n} Y_i, \tag{11}$$

[6] Vorsicht: dies ist eine Faustregel. Keine rigorose Aussage. Sehe folgende Vorbehalte zur Anwendbarkeit!

[7] Ein anderer Trivialfall, der nicht vom ZGS abgedeckt ist, ist der Fall mit $\sigma^2 = 0$. Was bedeutet dies für die ZV X_1?

wobei $Y_i := \frac{1}{n}(X_i - p)$. Beachte, dass $\mathbb{E}[Y_i] = 0$. Ferner gilt

$$\begin{aligned}
&\mathbb{P}\{|\bar{X}_n - p| \geq \varepsilon\} = \mathbb{P}\{|\sum_{i=1}^{n} Y_i| \geq \varepsilon\} \\
&= \mathbb{P}\{\sum_{i=1}^{n} Y_i \geq \varepsilon\} + \mathbb{P}\{\sum_{i=1}^{n} Y_i \leq -\varepsilon\} \\
&\approx \text{[Symmetrie der Normalverteilung]} \approx 2\mathbb{P}\{\sum_{i=1}^{n} Y_i \geq \varepsilon\} \\
&= 2(1 - \mathbb{P}\{\sum_{i=1}^{n} Y_i \leq \varepsilon\}).
\end{aligned} \tag{12}$$

Nun verfahren wir nach der oben beschriebenen Normalapproximation-Prozedur:

1. $\mathbb{E}[Y_1] = 0$, $\mathrm{Var}[Y_1] = p(1-p)/n^2$.
2. $z = \varepsilon\sqrt{\frac{n}{p(1-p)}}$.
3. $\mathbb{P}\{\sum_{i=1}^{n} Y_i \leq \varepsilon\} \approx \Phi(z) = \Phi(\varepsilon\sqrt{\frac{n}{p(1-p)}}) \geq \Phi(2\varepsilon\sqrt{n})$, da $p(1-p) \leq \frac{1}{4}$ ist.

Nun folgt aus (12) und 3.

$$\mathbb{P}\{|\bar{X}_n - p| \geq \varepsilon\} \lessapprox 2(1 - \Phi(2\varepsilon\sqrt{n})) \tag{13}$$

Zum Vergleich mit (4) setzen wir $n = 100$ und $\varepsilon = 0,1$ in (13) ein und bekommen

$$\mathbb{P}\{|\bar{X}_{100} - p| \geq 0,1\} \lessapprox 2 - 2\Phi(2 \cdot 0,1 \cdot \sqrt{100}) \approx 0,046, \tag{14}$$

wobei in der letzten Gleichung haben wir den ungefähren numerischen Wert[8] der Funktion $\Phi(x)$ an der Stelle $x = 2 \cdot 0,1 \cdot \sqrt{100} = 2$ eingesetzt ($\Phi(2) \approx 0,9772$).

Example 0.1 (de Moivre-Laplace Approximation der Binomialverteilung)**.** *Sei S_n die Binomialverteilung mit den Parametern $n \in \mathbb{N}$ und $p \in (0,1)$. In diesem Fall ist X_1 Bernoulli-verteilt mit der Erfolgswahrscheinlichkeit p. Wir erinnern uns an die Momente der Bernoulli ZV: $\mathbb{E}[X_1] = p$ und $\mathrm{Var}[X_1] = p(1-p)$. Somit gilt nach ZGS (Theorem 0.1) für hinreichend[9] großes $n \in \mathbb{N}$*

$$\mathbb{P}\{k \leq S_n \leq l\} \approx \Phi\left(\frac{l - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right), \tag{15}$$

wobei $k, l \in \mathbb{Z}_+$.

Die Konvergenzart in (8) hat einen Namen.

[8] Zum numerischen Ausrechnen von $\Phi(x)$ gibt es entsprechende Befehle in Ihrer Lieblings-Programmiersprache.

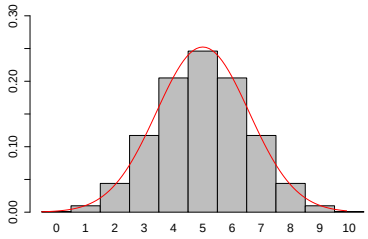

Abbildung 5: Histogramm der Binomial-Verteilung (die graue Balken) für $n = 10$ und $p = 0,5$ und die Dichte der Normalverteilung (die rote Glockenkurve) mit $\mu = 5,5$ und $\sigma = \sqrt{2,5}$.

[9] Für nicht "allzu großes" $n \in \mathbb{N}$ ist

$$\begin{aligned}
&\mathbb{P}\{k \leq S_n \leq l\} \\
&\approx \Phi\left(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)
\end{aligned}$$

eine etwas bessere Approximation als (15). Genau dies haben wir auf der Abbildung 5 gemacht ($\mu = 10 \cdot 0,5 + 0,5 = 5,5$).

Definition 0.1 (Verteilungskonvergenz). *Sei $\{X_n\}$ eine Folge von ZVen und X_∞ eine ZV. Wir sagen, dass $\{X_n\}$ gegen X_∞ konvergiert, falls*

$$\lim_{n\to\infty} F_{X_n}(x) = F_{X_\infty}(x), \quad x \in S \subset \mathbb{R}, \tag{16}$$

wobei F_{X_n} und F_{X_∞} die Verteilungsfunktionen von der ZVen X_n und X_∞ sind und $S \subset \mathbb{R}$ die Menge aller Punkten auf $\mathbb{R}$ ist, wo die Verteilungsfunktion F_{X_∞} stetig ist.

NOTATION. Konvergiert X_n gegen X_∞ in Verteilung, so schreiben wir:

$$X_n \xrightarrow[n\to\infty]{\mathrm{D}} X_\infty. \tag{17}$$

BEMERKUNG. Verteilungskonvergenz ist lediglich eine Aussage über die Verteilungen von den ZVen und nicht über ihre Realisierungen.

Elementare Poisson-Approximation: seltene Ereignisse

Wir erinnern uns an die folgende Aussage. Sei $S_n \sim \mathrm{Bin}(n, \lambda/n)$, für $\lambda \in \mathbb{R}_+$ und $n \in \mathbb{N}$ groß genug[10]. Dann gilt

$$\lim_{n\to\infty} \mathbb{P}\{S_n = k\} = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}, \tag{18}$$

Demzufolge $S_n \xrightarrow[n\to\infty]{\mathrm{D}} Y$, wobei $Y \sim \mathrm{Pois}(\lambda)$. (Warum?)

[10] So dass $\lambda/n \in [0, 1]$.

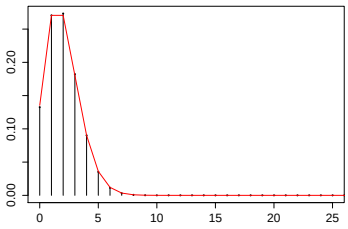

Abbildung 6: Die Wahrscheinlichkeitsfunktion der Binomial-Verteilung (schwarze Säulen) für $n = 100$ und $p = 5/100$ und die Wahrscheinlichkeitsfunktion der Poisson-Verteilung (rote Kurve) mit $\mu = 5$. Sehr gute Approximation!

BEMERKUNG. Die Poisson-Approximation funktioniert in der Situation, wo wir viele Summanden haben, die mit einer **kleinen Wahrscheinlichkeit** $O(n^{-1})$ nicht Null sind. Allerdings sind diese Summanden selber $O(1)$. Bei der Normal-Approximation sind die Summanden klein $O(n^{-1/2})$, dafür dürfen sie alle $\neq 0$ sein.

FRAGE. Wie sieht die Poisson-Verteilung für großes λ aus?

ANTWORT. Die Normalapproximation suggeriert:

$$\frac{\mathrm{Pois}(\lambda) - \lambda}{\sqrt{\lambda}} \xrightarrow[\lambda\to\infty]{\mathrm{D}} \mathcal{N}(0, 1). \tag{19}$$

Dies ist in der Tat der Fall (s. Übung).

Example 0.2 (Stirling-Formel). *Frage: Wie schnell wächst $n!$, wenn $n \to \infty$? Antwort: Sei $\{X_i\}_{i=1}^\infty$ eine Folge von unabhängigen* Pois(1)*-verteilten ZVen. Es gilt*

$$S_n = \sum_{i=1}^n X_i \sim \mathrm{Pois}(n). \tag{20}$$

Nach (19) *können wir die Normalapproximation anwenden.*

Seien ZV $N \sim \mathrm{Pois}(n)$ *und* $X \sim \mathcal{N}(\mu, \sigma^2)$*, wobei* σ^2 *und* μ *vom Teil (a) sind. Aus (a) folgt (Normalapproximation!)*

$$\mathbb{P}\{N = n\} = \mathbb{P}\{n - 0.5 \leq N \leq n + 0.5\} \underset{n\to\infty}{\sim} \mathbb{P}\{n - 0.5 \leq X \leq n + 0.5\}. \tag{21}$$

Um die Stirling-Formel herzuleiten, vergleiche die linke und die rechte Seite von (21)*.*

[XXX]

Literatur

Götz Kersting and Anton Wakolbinger. *Elementare Stochastik*. Springer, 2010.