PTS Vorlesung – Kap 4

4 Stichproben / Sampling

4.1 Stichprobenmittel und zentraler Grenzwertsatz

Angenommen, wir ziehen eine Stichprobe aus einer Population und führen an den Elementen eine Reihe von Messungen aus (z.B. Gewicht, Dauer usw.). Angenommen, die Größen, die wir messen, sind verteilt mit Mittelwert μ und Varianz σ^2.

Diese Messungen modellieren wir als eine Folge von Zufallsvariablen $X_1, X_2, \dots, X_n$ die uiv* sind

Das n-te Stichprobenmittel (sample mean) ist die ZV

$$\overline{X}_n := \frac{\sum_{i=1}^{n} X_i}{n}$$

* uiv = unabhängig und identisch verteilt
iid = independent and identically distributed

Wir wissen

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot n\cdot \mu = \mu$$

$$Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\cdot n\cdot \sigma^2 = \frac{1}{n}\sigma^2$$

Das heißt

- der Erwartungswert / das Mittel bleibt gleich
- die Standardabweichung geht gegen 0.

Das ist der Grund für das Gesetz der großen Zahlen:

$$P[\,|\bar{X}_n - \mu| > \varepsilon\,] \longrightarrow 0$$

für alle möglichen Schranken ε.

Welche Form hat die Verteilung der $\bar{X}_n$?

Problem: $\bar{X}_n$ wird immer dünner wegen der Division durch n.

Untersuche statt dessen

$$Y_i := \frac{X_i - \mu}{\sigma}.$$

Dann ist $E[Y_i] = 0$, $Var(Y_i) = Var\left(\frac{X_i-\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X_i) = 1$

Die X_i waren uiv, deshalb sind auch die Y_i uiv.

Sei $U_n := \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} = \sqrt{n}\,\bar{Y}_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}$.

Statt durch n, teilen wir durch $\sqrt{n}$!

Dann ist

- $E[U_n] = \sqrt{n}\cdot E[\bar{Y}_n] = 0$
- $Var[U_n] = Var[\sqrt{n}\,\bar{Y}_n] = n\,Var[\bar{Y}_n] =$

$$= n\,Var\left(\frac{\bar{X}_n - \mu}{\sigma}\right) = \frac{n}{\sigma^2} Var(\bar{X}_n) = \frac{n}{\sigma^2}\cdot\frac{\sigma^2}{n} = 1.$$

Der zentrale Grenzwertsatz (ZGW) / The Central Limit Theorem (CLT)

Der ZGW besagt, dass die Verteilungen der U_n (cdfs of the U_n) gegen die Verteilung $N(0,1)$ konvergiert.

f.a. $x \in \mathbb{R}$

Theorem (Lindeberg-Lévy) [ZGW/CLT]

Seien X_i uiv (iid) ZVen mit Mittelwert μ und Varianz σ^2 und seien

$$U_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$$

Φ := Verteilung von $N(0,1)$ (d.h. $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$)

Dann gilt

$$\lim_{n \to \infty} P[U_n \leq x] = \Phi(x)$$

Interpretation und Anwendung des ZGW

Seien X_i uiv ZVen mit Mittel μ und Varianz σ^2

Sei $S_n := \sum_{i=n}^{n} X_i$ die Summe der ersten n X_i.

Der ZGW sagt, dass für große n die normalisierten Summen

$$\frac{1}{\sqrt{n}\sigma}(S_n - n\mu) = \frac{(S_n - n\mu)}{\sqrt{n}\sigma} \sim N(0,1)$$

annähernd standard-normalverteilt sind.

$Var(S_n) = n\sigma^2$

$\Rightarrow \sigma_{S_n} = \sqrt{n}\sigma$

Daraus kann man schließen, dass ungefähr (approximately)

$$S_n \sim N(n\mu, n\sigma^2)$$

Die Approximation ist am besten in der Nähe des Mittels $n\mu$.

Wahrscheinlichkeiten der Werte von S_n können dann approximiert werden durch die Wahrscheinlichkeiten einer $N(n\mu, n\sigma^2)$-RV.

Beispiel 64: Eine Versicherungsgesellschaft hat 25.000 Kunden. Die Gesellschaft betrachtet die jährlichen Forderungen eines Kunden als ZV. Dabei hat sie festgestellt

- der Mittelwert der Forderungen ist $\mu = € 320$
- die Standardabweichung ist $\sigma = € 540$

Wie groß ist die Wahrscheinlichkeit, dass die gesamten jährlichen Forderungen $>$ € 8,3 Mio sind?

Sei F_i die jährliche Forderung von Kunde i, und sei

$S_n = \sum_{i=1}^{n} F_i$ die jährliche Summe an Forderungen, $n = 25.000$

$\overline{F}_n = \frac{1}{n} S_n$ der Durchschnitt der Forderungen.

Wir fragen nach $P[S_n > s]$, $s = 8{,}3$ Mio

$$S_n \approx N(25.000 \cdot 320, 25.000 \cdot 540^2)$$

Mit dem ZGW schließen wir, dass

$$S_n \sim N(n\cdot\mu, n\sigma^2) \quad \text{approx}$$

Daher ist

$$P[S_n > s] = P\left[\frac{S_n - n\mu}{\sqrt{n}\sigma} > \frac{s - n\mu}{\sqrt{n}\sigma}\right]$$

$$\approx P\left[Z > \frac{s - n\mu}{\sqrt{n}\sigma}\right] = 1 - \Phi\left(\frac{s - n\mu}{\sqrt{n}\sigma}\right)$$

Wir haben

$$n\cdot\mu = 25.000 \cdot 320$$
$$= 0{,}25 \cdot 10^5 \cdot 32 \cdot 10$$
$$= \frac{1}{4} \cdot 32 \cdot 10^6$$
$$= 8 \cdot 10^6$$

$$s - n\mu = 8{,}3 \cdot 10^6 - 8 \cdot 10^6$$
$$= 3 \cdot 10^5$$

$$\sqrt{n}\cdot\sigma = \sqrt{25.000} \cdot 540$$
$$= \sqrt{2{,}5} \cdot 10^2 \cdot 5{,}4 \cdot 10^2$$

$$\frac{s - n\mu}{\sqrt{n}\cdot\sigma} = \frac{3}{\sqrt{2{,}5}\cdot 5{,}4} \cdot \frac{10^5}{10^4}$$
$$= 0{,}351 \cdot 10 = 3{,}51$$

Also ist

$$P[S_n > s] = 1 - \Phi(3{,}51) = 1 - 0.9998 = 0.0002$$

Wenn wir R benutzen wollen, reicht es, zu wissen, dass

$$S_n \sim N(n\cdot\mu, n\sigma^2)$$

und wir fragen nach

$$P[S_n > s] = 1 - P[S_n \leq s].$$

Das können wir berechnen durch Aufrufe

$$1 - \text{dnorm}(s, n\cdot\mu, \sqrt{n}\cdot\sigma)$$

Beachte, dass R als Eingabe die Standardabweichung verlangt, nicht die Varianz

bzw

$$\text{dnorm}(s, n\cdot\mu, \sqrt{n}\cdot\sigma, \text{lower.tail} = \text{FALSE})$$

Normalverteilung und Binomialverteilung

Korollar: Seien X_i unabhängige Bernoulli(p)-verteilte ZVen. Dann gilt

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{n}\cdot\sqrt{p\cdot(1-p)}} \longrightarrow N(0,1)$$

Faustregeln (Rules of Thumb):

Eine Binomial(n,p)-Verteilung liegt nahe bei

- $N(np, np(1-p))$ falls $np > 5$ und $n(1-p) > 5$
- Poisson(np) falls $n > 20$ und $np < 5$ oder $n(1-p) < 5$

Beispiel 65: In einem Flugzeug haben 150 Passagiere Platz. Auf einer stark frequentierten Route nehmen nur 30 % der gebuchten Passagiere tatsächlich den Flug. Wenn 450 Tickets pro Flug verkauft werden, wie groß ist die Wahrscheinlichkeit, dass der Flug überbucht ist?

Sei P die Zahl der Passagiere, die tatsächlich den Flug nehmen.

Dann ist P eine binomial verteilte RV mit

$$n = 450 \qquad p = 0{,}3$$

Sei $s = 150$ die Zahl der verfügbaren Sitze.

Der Flug ist überbucht, wenn gilt

$$P > 150$$

Wir approximieren P durch eine ZV $X \sim N(np, np(1-p))$. Dann ist

$$P[P > s] = P[X > s + 0.5]$$

Anpassung, wenn eine diskrete in eine stetige Verteilung übersetzt wird *)

$$= P\left[\frac{X - np}{\sqrt{n}\sqrt{p(1-p)}} > \frac{s + 0{,}5 + np}{\sqrt{n}\sqrt{p(1-p)}}\right] = 1 - \Phi\left(\frac{s + 0{,}5 + np}{\sqrt{n}\sqrt{p(1-p)}}\right)$$

$$= 1 - \Phi(1{,}59) = 1 - 0{,}944 = 0{,}056 = 5{,}6\,\%$$

Alternativ können wir in R den folgenden Aufruf eingeben

```
pnorm( s + 0,5 , np , √n · p · (1-p), lower.tail = FALSE)
```

*) Kontinuitätskorrektur (continuity correction)

Beispiel 69: Meinungsumfrage

Angenommen 40 % der Bevölkerung unterstützen eine bestimmte Partei bei einer Wahl.

Wir wählen eine zufällige Stichprobe aus 150 Personen.

1) Was sind Erwartungswert und Varianz der Zahl des Unterstützes in der Stichprobe?
2) Wie groß ist die Wahrscheinlichkeit, dass mehr als 50 % der befragten Personen die Partei unterstützen?

Beispiel 69: Meinungsumfrage

Angenommen 40 % der Bevölkerung unterstützen eine bestimmte Partei bei einer Wahl.

Wir wählen eine zufällige Stichprobe aus 150 Personen.

1) Was sind Erwartungswert und Varianz der Zahl des Unterstützes in der Stichprobe?
2) Wie groß ist die Wahrscheinlichkeit, dass mehr als 50 % der befragten Personen die Partei unterstützen?

Sei X_i die Antwort der i-ten Person, mit $1 =$ „Ja“ und $0 =$ „Nein“.

$\Rightarrow X_i \sim \text{Bernoulli}(p)$ mit $p = 0{,}4$

Sei $Y := \sum_{i=1}^{n} X_i \Rightarrow Y \sim \text{Bin}(n, p)$, mit $n = 150$

$\Rightarrow E[Y] = n \cdot p = 150 \times 0{,}4 = 60$,

$\text{Var}(Y) = n \cdot p \cdot (1-p) = 150 \times 0{,}4 \times 0{,}6 = 150 \times 0{,}24 = 36$

$\Rightarrow E[Y] = n \cdot p = 150 \times 0{,}4 = 60$,

$Var(Y) = n \cdot p \cdot (1-p) = 150 \times 0{,}4 \times 0{,}6 = 150 \times 0{,}24 = 36$

Wir überprüfen die Faustregeln:

$n \times p = 60 > 5$, $\quad n \times (1-p) = 150 \times 0{,}6 = 90 > 5$

$\Rightarrow$ Approximiere Y durch $N(60, 36)$

Wir wollen wissen

$$P[Y > 75] \; ?$$

Wie können wir das berechnen?

1) Verwende die Binomialverteilung $\quad Binom(150; 0{,}4)$

In R erhalten wir

$P[Y > 75] =$ pbinom(75, 150, 0.4, lower.tail = FALSE)

$= 1 - P[Y \leq 75] =$ pbinom(75, 150, 0.4)

$= 0{,}005225$

2) Approximiere Y durch ein $Y' \sim N(60, 36)$

In R erhalten wir

$P[Y > 75] = P[Y > \boxed{75{,}5}]^{*} \simeq P[Y' > 75{,}5]$

$= 1 - \Phi_{60,36}(75{,}5)$

$=$ 1 - pnorm(75.5, 60, 36)

$=$ pnorm(75.5, 60, 36, lower.tail = FALSE)

*Stetigkeitskorrektur / continuity correction

3) Approximiere durch Normalverteilung, standardisiere und schlage nach in Tabelle

Transformiere y' zu $Z \sim N(0,1)$

$$P[y' > 75{,}5] = P\left[\frac{y'-60}{6} > \frac{75{,}5-60}{6}\right]$$

$$\approx P\left[Z > \frac{75{,}5-60}{6}\right] = P\left[Z > \frac{15{,}5}{6}\right]$$

$$= P[Z > 2{,}583] = 1 - \Phi(2.583)$$

$$\approx 1 - \Phi(2.58) = 1 - 0.9951 = 0.0049$$

4.2 Stichproben-Varianz (Sample Variance)

Wenn wir eine Größe messen, betrachten wir dies als die Auswertung einer Zufallsvariablen X. Führen wir mehrere Messungen durch, dann betrachten wir dies als Auswertung von ZVen $X_1, \dots, X_n$, die u.i.v. (i.i.d.) sind und dieselbe Verteilung wie X haben.

Wie können wir den Mittelwert der Verteilung von X schätzen, d.h., $E[X]$?

Der Durchschnitt (average) der X_i, also $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, sollte ein guter Schätzwert sein.

Wie können wir überprüfen, dass dies konzeptuell richtig ist?

Unverzerrte Schätzer (Unbiased Estimators)

Seien $X, X_1, \dots, X_i, \dots$ u.i.v. ZVen

Wenden wir ein Funktion $F(x_1, \dots, x_n)$ an auf $X_1, \dots, X_n$, erhalten wir eine neue ZV $F(X_1, \dots, X_n)$.

Ein Beispiel ist $\bar{X}_n$, das definiert ist durch

$$F(x_1, \dots, x_n) = \frac{1}{n}(x_1 + \dots + x_n) = \bar{x}_n$$

Definition: Seien $X_1, \dots, X_n$ u.i.v ZVen, $F: \mathbb{R}^n \to \mathbb{R}$, und θ ein Parameter (etwa Mittelwert oder Varianz) der Verteilung der X_i.

Wenn

$$E[F(X_1, \dots, X_n)] = \theta$$

dann ist $F(X_1, \dots, X_n)$ ein unverzerrter oder erwartungstreuer Schätzer für θ.

Beispiele: (1) Der Durchschnitt $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist ein unverzerrter Schätzer für den Mittelwert μ

(2) Der durchschnittliche quadratische Abstand (average squared mean)

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

ist ein unverzerrter Schätzer für die Varianz σ^2
(Beachte: wir benutzen μ, nicht $\bar{X}_n$.)

Beweis:

i) $E[\bar{X}_n] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot n \cdot \mu = \mu$

ii) Zur Erinnerung: $Var(X) = E[(X-\mu)^2]$. Also

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2] = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu)^2] = \frac{1}{n} \cdot n \cdot \sigma^2 = \sigma^2$$

Schätzung der Varianz

Betrachte die Funktion

$$T^2(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad * \qquad \text{mit } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Dann rechnet man nach, dass

$$E\left[T^2(\xi_1,\dots,\xi_n)\right] = \frac{n-1}{n}\,\mathrm{Var}(\xi).$$

Dieser Schätzer ist verzerrt!

Aber

$$\frac{n}{n-1}\,T^2(x_1,\dots,x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 =: S^2$$

ist unverzerrt! Dies ist die Stichproben-Varianz (sample variance)

---

* T^2 ist eine missbräuchliche Schreibweise, als Gegenüberstellung zu S^2

Wir können die Qualität einer Schätzung bestimmen, wenn wir wissen, wie die Zufallsvariable, mit der wir schätzen, verteilt ist

Oft nehmen wir an, dass unsere ξ_i $N(\mu,\sigma^2)$-verteilt sind.

Dann ist

$\bar{\xi}_n$ $N(\mu, \frac{1}{n}\sigma^2)$-verteilt.

Was ist die Verteilung von

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2 \quad ?$$

Man kann zeigen, das $\frac{n-1}{\sigma^2}S^2$ verteilt ist wie die Summe $Z_1^2+\dots+Z_{n-1}^2$ von $n-1$ unabhängigen $N(0,1)$-Variablen $Z_1,\dots,Z_{n-1}$.

Dies ist die χ^2_{n-1}-Verteilung, d.h., die Chi-Quadrat-Verteilung

(chi-square distribution)

Zusammenspiel der Normal- und der Chi-Quadrat-Verteilung

Theorem 67: Seien $X_1, \dots, X_n$ unabhängig und $N(\mu, \sigma^2)$-verteilt. Dann gilt

(Stichproben-Mittel, Stichproben-Varianz)

- $\bar{X}$, S^2 sind unabhängig
- $\bar{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right)$
- $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

($n-1$: Freiheitsgrade (degrees of freedom))

Wenn wir die Varianz nicht kennen? t-Verteilung!

Wir wissen, dass

$$X_i \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma^2} \sim N(0,1)$$

Was geschieht, wenn wir σ mit $S = \sqrt{S^2}$ ersetzen?

$$\sqrt{n}\,\frac{\bar{X}-\mu}{S} = \frac{\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}}{\sqrt{\frac{1}{n-1}}\,\sqrt{\frac{(n-1)S^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Eine ZV

$$T_n = \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}}$$

hat eine t-Verteilung mit n Freiheitsgraden, geschrieben $T_n \sim t_n$

Die t-Verteilung: Definition

Seien Z und χ^2_n unabhängige ZVen mit

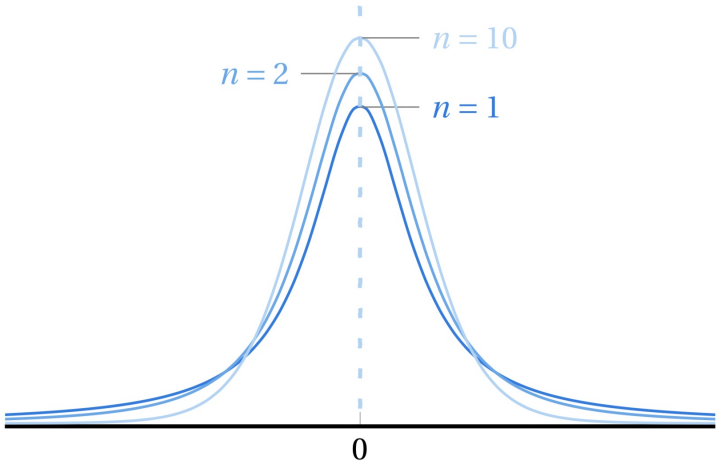
- $Z \sim N(0,1)$
- $\chi^2_n \sim \chi^2_n$

Dann ist die ZV

$$T_n = \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}}$$

t-verteilt mit n Freiheitsgraden, geschrieben $T_n \sim t_n$

Hierdurch sind auch Verteilungsfunktion und Dichte der t-Verteilung definiert



The density function of T_n for $n = 1, 2, 10$.

Eigenschaften der t-Verteilung

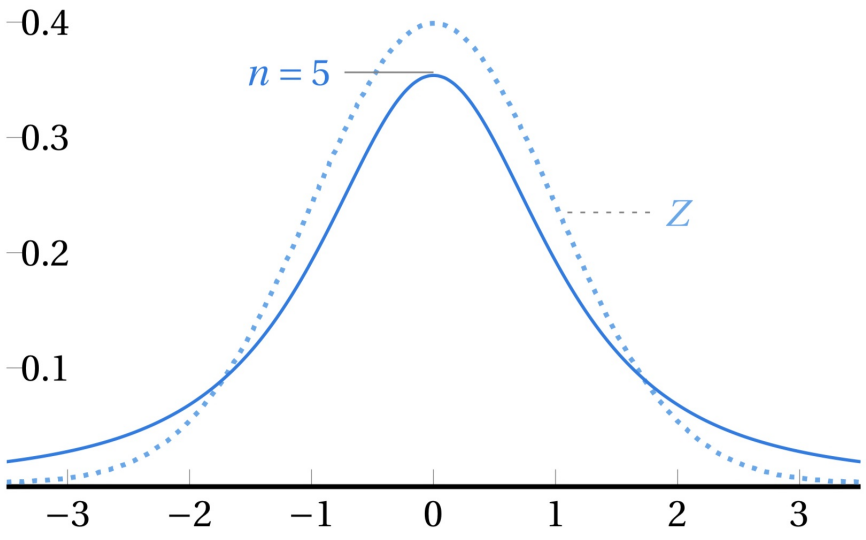
- glockenförmig (bell-shaped), aber breiter als die Normalverteilung,
  mit dickeren Enden (thicker tails) und
  niedrigerem Gipfel (lower peak)

- Parameter

$$\mu = \begin{cases} \text{undefiniert} & \text{für } n=1 \\ 0 & \text{für } n>1 \end{cases} \qquad \sigma^2 = \begin{cases} \text{undefiniert} & \text{für } n=1,2 \\ \frac{n}{n-2} & \text{für } n>2 \end{cases}$$

- $T_n \longrightarrow Z$ die t-Verteilung konvergiert $N(0,1)$

praktisch nur für kleine Stichproben ($n \leq 30$) relevant
für größere n ist der Unterschied vernachlässigbar.



The density function of T_5 (solid) and Z (dotted).

$T_n \longrightarrow Z$ die t-Verteilung konvergiert $N(0,1)$

Warum?

$$T_n = \frac{Z}{\sqrt{\frac{\chi^2_n}{n-1}}}$$

$E[Z^2] = Var(Z) + E[Z]^2 = 1 + 0 = 1$

Seien $Z_i \sim N(0,1)$ u.i.v. $\Rightarrow \sum_{i=1}^{n} Z_i^2 = \chi^2_n$

Gesetz der großen Zahlen:

$$\Rightarrow \frac{\chi^2_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Z_i^2 \longrightarrow E[Z^2] = 1$$

$$\Rightarrow \sqrt{\frac{\chi^2_n}{n}} \longrightarrow 1 \qquad \Rightarrow \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}} \longrightarrow Z$$