Mathe III

Ralf Herbrich

Parameterschätzung

Überblick

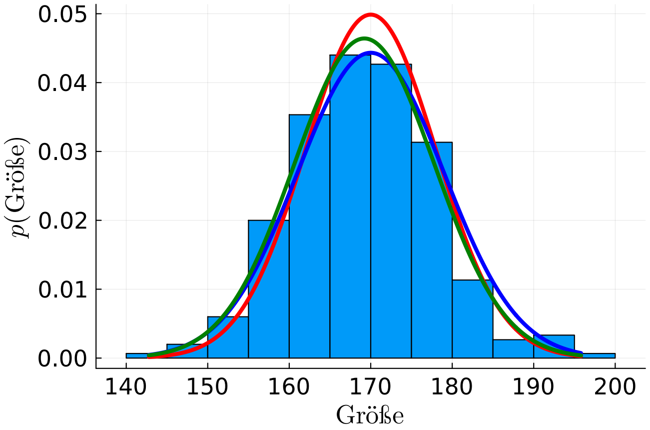
1. Parametrische Schätzer
2. Eigenschaften von Schätzern
3. Verteilung von Schätzern
4. Maximum Likelihood Prinzip
5. *Method of Moments*

Überblick

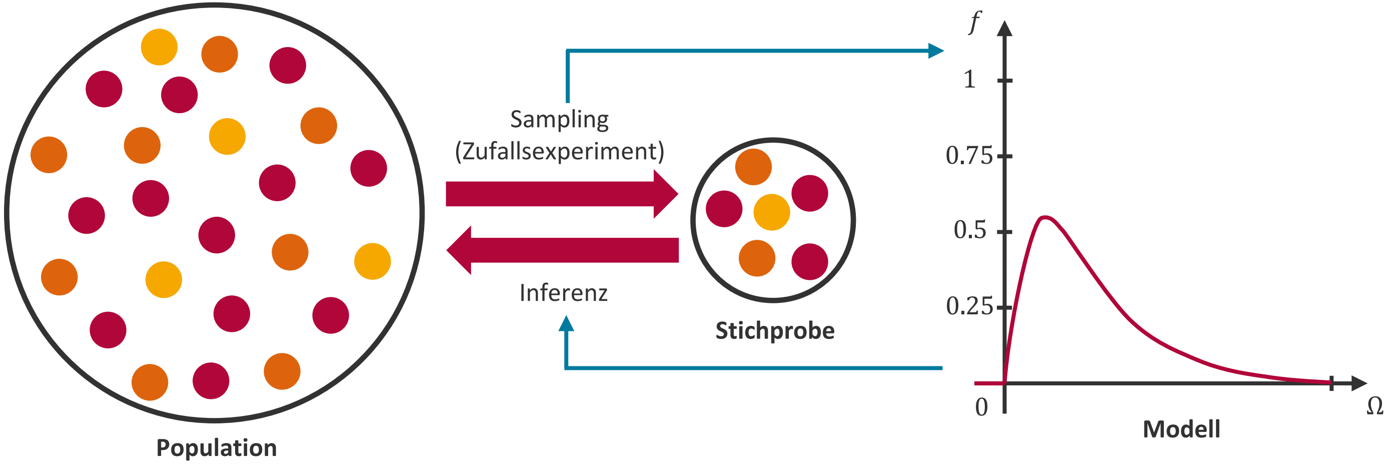
Motivation: Parameterschätzung

- **Beispiel (Größenverteilung)**. Wir bauen ein neues Vorlesungsgebäude am HPI und müssen entscheiden, welche Stühle wir für das neue Vorlesungsgebäude kaufen. Das hängt von der Verteilung der Körpergröße unserer Studys ab. Wie sind die Größen verteilt?

- **Ansatz**: Wir betrachten eine Stichprobe $\boldsymbol{x}$ von Beobachtungen $x_1, \ldots, x_{300}$ und fragen 300 Studys nach ihrer Größe. Dabei ergibt sich folgendes Histogramm der Verteilung der Größen.
- **Frage**: Wie können wir von der Stichprobe auf alle 800 aktuellen Studys (und zukünftige Studys) schließen?
- **Idee**: Wir machen Verteilungsannahmen mit 2 freien Parametern
  - Normalverteilungsannahme mit $\mu = 170, \sigma = 8$
  - Normalverteilungsannahme mit $\mu = 170, \sigma = 9$
  - Normalverteilungsannahme mit $\mu = \overline{\boldsymbol{x}}, \sigma = \sqrt{V[\boldsymbol{x}]}$



Stichprobe und Modell



- Das **stochastische Modell** beschreibt das Zufallsexperiment der Stichprobenwahl.
- Wir nutzen dieses, um von Stichprobenmerkmalen auf die Populationsmerkmale zu schließen mit Hilfe von **Schätzern**!
- Beim Aufstellen eines Modells trifft man Annahmen – stimmen diese nicht, sind die Ergebnisse hinfällig.

Parametrisches Modell & Schätzer

- **Definition (Parametrisches Modell).** Ein parametrisches Modell ist ein Tripel $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ bestehend aus einer Ergebnismenge Ω, einem Ereignissystem $\mathcal{F}$ und einer Menge $\{P_\theta \mid \theta \in \Theta\}$ von Wahrscheinlichkeitsverteilungen auf $(\Omega, \mathcal{F})$, die mit einer Indexmenge $\Theta \subset \mathbb{R}^d$ für ein $d \in \mathbb{N}$ indiziert ist. Man spricht von einem d-parametrigen Modell.
  - **Beispiel (Parametrisches Modell).** Wenn wir wissen, dass alle n unabhängigen Beobachtungen normalverteilt mit unbekanntem Erwartungswert und Varianz von 4 sind, dann ist

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{\boxed{(X_1, \ldots, X_n) \to \prod_{i=1}^{n} \mathcal{N}(X_i; \mu, 4)} \,\middle|\, \mu \in \mathbb{R}\right\}\right)$$

    Produktverteilung wegen angenommener Unabhängigkeit

    ein-parametriges Modell.

- **Definition (Schätzer).** Für ein parametrisches Modell $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ heißt eine beliebige Zufallsvariable T von Ω nach Θ ein Schätzer.
  - **Beispiel (Schätzer).** Für das oben genannte Modell sind dies zwei mögliche Schätzer

$$T_1(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Benutzt die Daten

$$T_2(X_1, \ldots, X_n) = 500$$

Ignoriert die Daten

Beispiel: Parametrisches Modell & Schätzer

- **Beispiel (Club)**. Wir beobachten 200 Wartende vor einem Club, bei dem die Türsteherin mit einer unbekannten Wahrscheinlichkeit einen Gast einlässt. Welches parametrische Modell bietet sich an, um die unbekannte Wahrscheinlichkeit zu schätzen und welchen Schätzer soll man wählen?
- **Lösung**. Sei X_i eine Bernoulli-verteilte Zufallsvariable, $X_i \sim \mathrm{Bern}(p)$ mit der Bedeutung

$$X_i = \begin{cases} 0 & \text{Gast } i \text{ wird nicht reingelassen} \\ 1 & \text{Gast } i \text{ wird reingelassen} \end{cases}$$

Alle X_i sind paarweise unabhängig. Dann ist die Stichprobe eine Realisierung der $X_1, \dots, X_{200}$ und das ein-parametrische Modell ist

$$\left(\{0,1\}^{200}, \mathcal{P}(\{0,1\}^{200}), \left\{(X_1, \dots, X_{200}) \to \prod_{i=1}^{200} \mathrm{Bern}(X_i; p) \,\middle|\, p \in [0,1]\right\}\right)$$

Produktverteilung wegen angenommener Unabhängigkeit

Ein möglicher Schätzer für p ist die Zufallsvariable

$$T(X_1, \dots, X_{200}) = \frac{1}{200} \sum_{i=1}^{200} X_i$$

Empirischer Anteil von Gästen, die in den Club eingelassen wurden

Überblick

Eigenschaften von Schätzern

- **Beobachtung**: Wenn ein Schätzer $T: \Omega \to \Theta$ eine Zufallsvariable ist, dann hat er auch eine Verteilung P_T.
- **Frage**: Wie wähle ich unter verschiedenen Schätzern aus?
- **Idee**:
  1. Schätzer sind „im Mittel" korrekt (Erwartungstreue)
  2. Schätzer machen im Erwartungswert über alle Stichproben den kleinsten Fehler (Effizienz)

- **Definition (Mittlerer Quadratischer Schätzfehler)**. Für ein parametrisches Modell $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ und einen Schätzer $T: \Omega \to \Theta$ ist der mittlere quadratische Fehler (*MSE*) definiert als

$$\mathrm{MSE}[T; \theta] = E_{X \sim P_\theta}[(T(X) - \theta)^2]$$

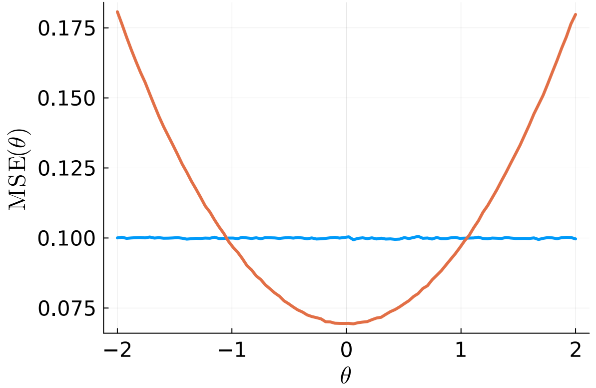
- **Bemerkungen (Mittlerer Quadratischer Schätzfehler)**
  - Der MSE bewertet die Qualität einer Schätzfunktion **im Mittel** über viele Stichproben!
  - Der MSE ist eine Güteeigenschaft des Schätzprozesses, nicht einer einzelnen Schätzung!

Beispiele von Mittleren Quadratischen Schätzfehlern

- **Beispiel (Erwartungswert).** Seien $X_1, \dots, X_{10} \sim \mathcal{N}(\theta, 1)$ und die Schätzer

$$T_1(X_1, \dots, X_{10}) = \frac{1}{10} \sum_{i=1}^{10} X_i$$

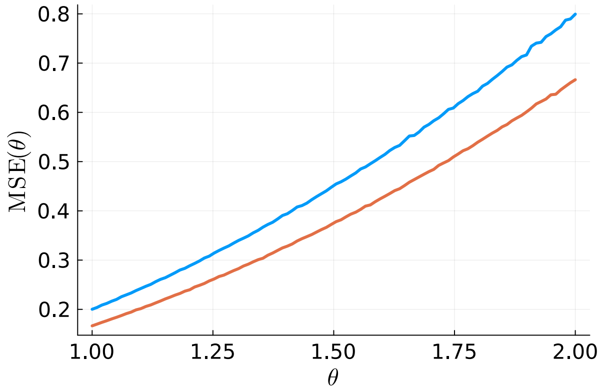
$$T_2(X_1, \dots, X_{10}) = \frac{1}{12} \sum_{i=1}^{10} X_i$$



- **Beispiel (Varianz).** Seien $X_1, \dots, X_{10} \sim \mathcal{N}(0, \theta)$ und die Schätzer

$$T_1(X_1, \dots, X_{10}) = \frac{1}{10} \sum_{i=1}^{10} X_i^2$$

$$T_2(X_1, \dots, X_{10}) = \frac{1}{12} \sum_{i=1}^{10} X_i^2$$



$n = 500\,000$ Stichproben zur Approximation des Erwartungswertes

Verzerrungs-Varianz Zerlegung (*bias-variance decomposition*)

- **Satz (Verzerrungs-Varianz Zerlegung des Schätzfehlers).** Für ein parametrisches Modell $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ und einen Schätzer $T: \Omega \to \Theta$ gilt für den mittleren quadratischen Schätzfehler $\mathrm{MSE}[T;\theta]$, dass

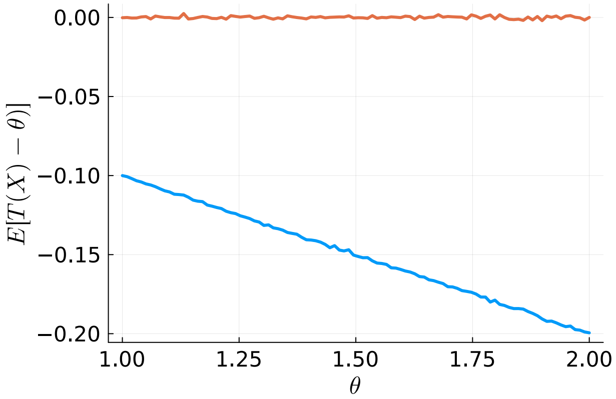
$$\mathrm{MSE}[T;\theta] = \underbrace{\left(\theta - E_{X\sim P_\theta}[T(X)]\right)^2}_{\text{Verzerrung (\textit{bias})}} + \underbrace{V_{X\sim P_\theta}[T(X)]}_{\text{Varianz (\textit{variance})}}$$

- **Beweis**: Wir benutzen die Linearität des Erwartungswertes und $\mu = E_{X\sim P_\theta}[T(X)]$

$$\begin{aligned}
\mathrm{MSE}[T;\theta] &= E_{X\sim P_\theta}[(T(X)-\theta)^2] \\
&= E_{X\sim P_\theta}[(T(X)\underbrace{-\mu+\mu}_{=0}-\theta)^2] \\
&= E_{X\sim P_\theta}\left[\left((T(X)-\mu)-(\theta-\mu)\right)^2\right] \\
&= E_{X\sim P_\theta}[(T(X)-\mu)^2 - 2\cdot(T(X)-\mu)(\theta-\mu) + (\theta-\mu)^2] \\
&= \underbrace{E_{X\sim P_\theta}[(T(X)-\mu)^2]}_{=V_{X\sim P_\theta}[T(X)]} - 2(\theta-\mu)\cdot\underbrace{E_{X\sim P_\theta}[(T(X)-\mu)]}_{=0} + (\theta-\mu)^2
\end{aligned}$$

Erwartungstreue

- **Definition (Erwartungstreue eines Schätzers).** Für ein parametrisches Modell $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ und einen Schätzer $T: \Omega \to \Theta$ ist der Schätzer erwartungstreu wenn er eine Verzerrung von Null hat, d.h. für alle $\theta \in \Theta$ gilt
$$E_{X\sim P_\theta}[T(X)] = \theta$$
- **Bemerkungen (Erwartungstreue eines Schätzers)**
  - Erwartungstreue heißt, dass der Schätzer „im Mittel" über seine Anwendung korrekt ist.
  - Erwartungstreue Schätzer haben **nicht** immer den kleinsten MSE!
  - Die empirische Varianz (als Schätzer der Varianz)
$$V[\boldsymbol{x}] = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{\boldsymbol{x}})^2$$
ist **nicht** erwartungstreu, aber
$$\hat{V}[\boldsymbol{x}] = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{\boldsymbol{x}})^2$$
ist erwartungstreu!



$n = 1\,000\,000$ Stichproben zur Approximation des Erwartungswertes

Erwartungstreue der korrigierten empirischen Varianz

- **Beweis**: Zuerst zeigen wir, dass für beliebige $x_1, \dots, x_n$ gilt

$$\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2 = \sum_{i=1}^{n} x_i^2 - n \cdot (\bar{\boldsymbol{x}})^2$$

Seien $X_1, \dots, X_n$ identisch unabhängig verteilte Zufallsvariablen mit $E[X_i] = \mu$ und $V[X_i] = \sigma^2$ und sei $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Dann gilt

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

Unabhängigkeit

$$= n \cdot \left(E[X_1^2] - E[(\bar{X})^2]\right)$$

Varianzzerlegung: $V[X] = E[X^2] - (E[X])^2$

$$= n \cdot \left(V[X_1] + (E[X_1])^2 - (V[\bar{X}] + (E[\bar{X}])^2)\right)$$

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

$E[\bar{X}] = E[X]$

$V[\bar{X}] = \frac{V[X]}{n}$

$$\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2$$

$$= \sum_{i=1}^{n}\left(x_i^2 - 2x_i \cdot \bar{\boldsymbol{x}} + (\bar{\boldsymbol{x}})^2\right)$$

$$= \sum_{i=1}^{n} x_i^2 - 2 \cdot n(\bar{\boldsymbol{x}})^2 + n \cdot (\bar{\boldsymbol{x}})^2$$

$\sum_{i=1}^{n} x_i = n \cdot (\bar{\boldsymbol{x}})$

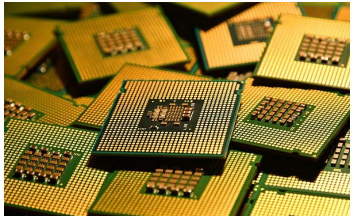
Überblick

1. Parametrische Schätzer
2. Eigenschaften von Schätzern
3. **Verteilung von Schätzern**
4. Maximum Likelihood Prinzip
5. *Method of Moments*

Verteilung des empirischen Mittelwerts

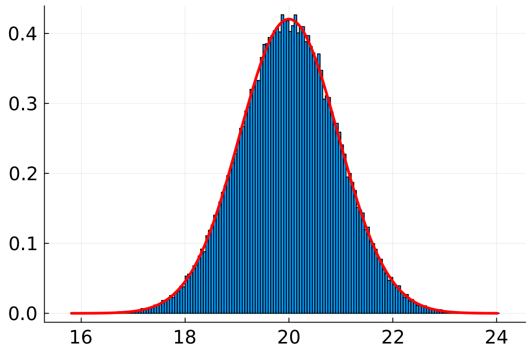
- **Beobachtung**: Wir haben gesehen, dass der empirische Mittelwert
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
einer unabhängig und identisch verteilten Stichprobe $X_1, \dots, X_n$ ein möglicher Schätzer des Erwartungswertes ist, wobei $X_i \sim P_X$ für alle $i = 1, \dots, n$.
- **Frage**: Da die Stichprobe aber eine Zufallsvariable ist, ist auch der empirische Mittelwert $\bar{X}_n$ eine Zufallsvariable. Wie ist $\bar{X}_n$ verteilt?
- **Antwort**: Der zentrale Grenzwertsatz liefert die Antwort, denn
$$\bar{X}_n \xrightarrow{d} \mathcal{N}\left(E[X], \frac{V[X]}{n}\right)$$
- **Bemerkungen (Verteilung des empirischen Mittelwerts)**
  - In der Praxis gilt diese Approximation der Verteilung als verlässlich für $n > 30$.
  - Je näher die ursprüngliche Verteilung P_X an einer unimodalen Verteilung ist, um so besser ist die Approximation schon für kleine n.

Verteilung von $\bar{X}_n$ und $\hat{V}[X_1, \ldots, X_n]$ bei Normalverteilung

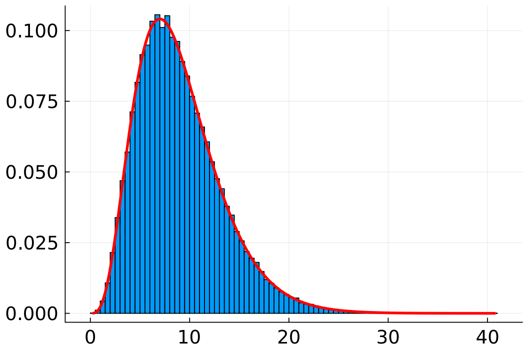


- **Beispiel (CPU-Rechenzeit)**. Eine verteilte Rechenaufgabe wird auf 10 Maschinen verteilt. Die Rechenzeiten pro Maschine sind unabhängig normalverteilt mit einem Erwartungswert von 20 Sekunden und einer Standardabweichung von 3 Sekunden. Wie sind der empirische Mittelwert $\bar{X}_{10}$ und die korrigierte Varianz $\hat{V}[X_1, \ldots, X_{10}]$ verteilt.
- **Lösung**. Wir simulieren 100 000 Stichproben der Größe 10 und plotten die Histogramme

$P_{\bar{X}_{10}}$



$P_{\hat{V}[X_1,\ldots,X_{10}]}$



Verteilung von $\bar{X}_n$ und $\hat{V}[X_1, \ldots, X_n]$ bei Normalverteilung

- **Satz (Verteilung von $\bar{X}_n$ und $\hat{V}[X_1, \ldots, X_n]$).** Seien $X_1, \ldots, X_n$ unabhängige normalverteilte Zufallsvariablen mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$ für alle $i = 1, \ldots, n$. Dann sind der Mittelwert $\bar{X}_n$ und die korrigierte Varianz $\hat{V}[X_1, \ldots, X_n]$ folgendermaßen verteilt

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{n-1}{\sigma^2} \cdot \hat{V}[X_1, \ldots, X_n] = \sum_{i=1}^{n} \frac{(X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2(n-1)$$

- **Beweis** (von Teil 2): Sei $Y_i = X_i - \mu$. Dann wissen wir, dass $\bar{Y}_n = \bar{X}_n - \mu$ und

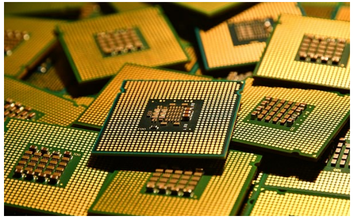
$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n \cdot (\bar{Y})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n \cdot (\bar{X}_n - \mu)^2$$

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2}_{\sim \chi^2(n)} = \sum_{i=1}^{n}\frac{(X_i - \bar{X}_n)^2}{\sigma^2} + \underbrace{\left(\frac{(\bar{X}_n - \mu)}{\frac{\sigma}{\sqrt{n}}}\right)^2}_{\sim \chi^2(1)}$$

Satz. Seien $X_1 \sim \chi^2(n_1)$ und $X_2 \sim \chi^2(n_2)$ zwei unabhängige Zufallsvariablen. Dann gilt

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

Verteilung von $\bar{X}_n$ und $\hat{V}[X_1, \ldots, X_n]$ bei Normalverteilung



- **Beispiel (CPU-Rechenzeit).** Eine verteilte Rechenaufgabe wird auf 10 Maschinen verteilt. Die Rechenzeiten pro Maschine sind unabhängig normalverteilt mit einem Erwartungswert von 20 Sekunden und einer Standardabweichung von 3 Sekunden. Wie wahrscheinlich ist es, dass die empirische korrigierte Varianz größer ist als 12?
- **Lösung**. Sei $S^2 := \hat{V}[X_1, \ldots, X_{10}]$ die empirische korrigierte Varianz, dann wissen wir, dass $\frac{9}{9} \cdot S^2 \sim \chi^2(9)$ und daher

$$P(S^2 > 12) = 1 - P(S^2 \leq 12)$$

$$= 1 - \chi^2(12; 9)$$

```
pchi(12,9)
```

$$= 0.2133 \approx 21.3\%$$

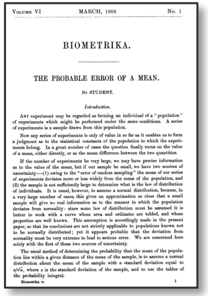
Verteilung des empirischen normierten Mittelwertschätzers

- **Korollar (Verteilung des empirischen normierten Mittelwertschätzers).** Seien $X_1, \ldots, X_n$ unabhängige normalverteilte Zufallsvariablen mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$ für alle $i = 1, \ldots, n$. Dann ist der normierte Mittelwertschätzer t-verteilt mit $n-1$ Freiheitsgraden:

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\hat{V}[X_1, \ldots, X_n]}{n}}} \sim t(n-1)$$

- **Bemerkungen (des empirischen normierten Mittelwertschätzers)**
  - Folgt direkt aus der Definition der t-Verteilung und der Verteilung von $\hat{V}[X_1, \ldots, X_n]$
  - Die t-Verteilung wurde von William Sealy Gosset, der für die Guinness Brauerei arbeitete, entdeckt und unter dem Pseudonym „Student" veröffentlicht (daher nennt man die Verteilung auch „Student t-Verteilung").
  - Das Ergebnis erlaubt es, auch bei unbekannter Varianz (und daher empirisch geschätzter Varianz) die Verteilung des Schätzers zu kennen.
  - Die t-Verteilung konvergiert in Wahrscheinlichkeit zur Standardnormalverteilung!

William Sealy Gosset (1876 – 1937)



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

Viel Spaß bis zur nächsten Vorlesung!