Mathe III

Ralf Herbrich

Parameterschätzung

Überblick

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

Überblick

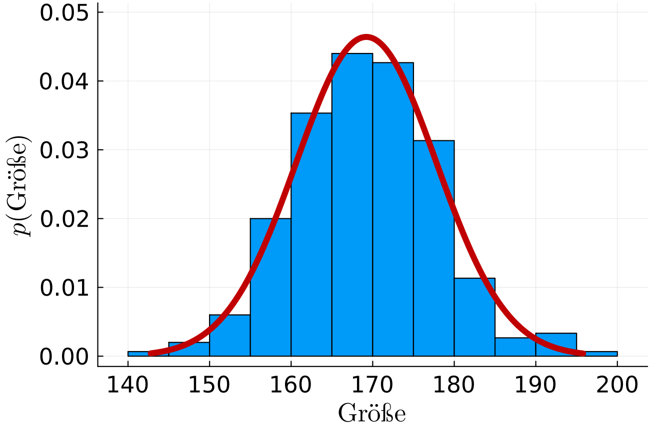
1. Parametrische Schätzer
2. Eigenschaften von Schätzern
3. Verteilung von Schätzern
4. **Maximum *Likelihood* Schätzer**
5. *Method of Moments*

Wiederholung: 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?

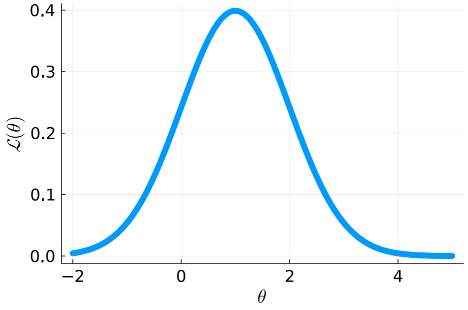
- **Ansatz**: Wir betrachten eine Stichprobe $\boldsymbol{x}$ von Beobachtungen $x_1, \dots, x_{300}$ und fragen 300 Studys nach ihrer Größe. Dabei ergibt sich folgendes Histogramm der Verteilung der Größen.
- **Beobachtung**: Jede Wahrscheinlichkeitsverteilung P_θ eines parametrischen Modells $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ weist der Stichprobe eine (Zähl)dichte $p_\theta(x_1, \dots, x_{300})$ zu!
- **Idee**: Wir finden den Parameter θ, unter dem die Stichprobe $x_1, \dots, x_{300}$ am wahrscheinlichsten ist! Wenn $X_i \sim \mathcal{N}(\theta_1, 8)$ dann ist dieser Schätzer gegeben durch

$$T(X_1, \dots, X_{300}) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

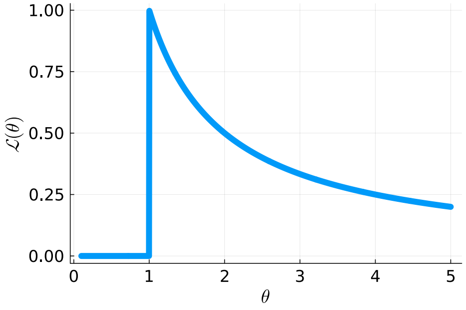


- **Definition (*Likelihood*).** Ist $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ ein parametrisches Modell und sei $x \in \Omega$ eine Stichprobe, so heißt die Funktion $\mathcal{L}\colon \Theta \times \Omega \to [0, +\infty)$ mit $\mathcal{L}(\theta, x) = p_\theta(x)$ die zugehörige *Likelihood*, wobei p_θ die (Zähl)dichte von P_θ ist.
- **Beispiel (*Likelihood*).** Betrachten wir die Beobachtung $x = 1$

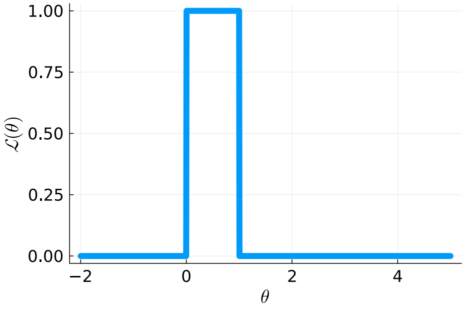
$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{N}(\theta, 1) \mid \theta \in \mathbb{R}\})$



$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{U}(0, \theta) \mid \theta \in \mathbb{R}\})$



$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{U}(\theta, \theta + 1) \mid \theta \in \mathbb{R}\})$



- **Bemerkungen (*Likelihood*)**
  - Die *Likelihood* ist nicht normiert (summiert/integriert sich nicht zu 1) über θ!
  - Die *Likelihood* ist hat nicht immer ein eindeutiges Maximum.

Maximum *Likelihood* Schätzer

- **Definition (Maximum *Likelihood* Schätzer).** Ein Schätzer $T_{\mathrm{MLE}}: \Omega \to \Theta$ heißt ein Maximum *Likelihood* Schätzer (*maximum likelihood estimator (MLE)*) wenn

$$T_{\mathrm{MLE}}(x) = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta, x)$$

- **Bemerkungen (Maximum *Likelihood* Schätzer)**
  - Die *Likelihood* hat nicht immer ein eindeutiges Maximum (T_{MLE} ist nicht eindeutig).
  - Das Maximum *Likelihood* Prinzip wurde sporadisch von Gauss und Laplace benutzt aber etabliert zwischen 1912 und 1922 von Sir Ronald Fisher.
- **Wie berechnet man einen MLE**?
  - **Trick**: Bestimme das Maximum des Logarithmus der *Likelihood* weil aus Produkten Summen werden, der Logarithmus monoton steigend ist und der Wertebereich numerisch besser in Gleitkommazahlen abzubilden ist.
  - Die Nullstelle der ersten Ableitung der *log-likelihood* ist ein möglicher Kandidat
  - Zur Überprüfung, dass es sich um ein Maximum handelt, muss noch die zweite Ableitung geprüft werden (muss negativ sein!)

Sir Ronald Fisher (1890 – 1962)

$$\log_b(x \cdot y) = \log_b(x) + \log_b(y)$$

MLE für die Bernoulliverteilung

- **Satz (MLE der Bernoulliverteilung)**. Seien $X_1, \ldots, X_n$ unabhängig und $X_i \sim \mathrm{Bern}(p)$. Dann ist der Maximum *Likelihood* Schätzer T_{MLE} für p gegeben durch

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

- **Beweis**: Wir betrachten die log-likelihood Funktion

$$\log\left(\mathcal{L}(p,(x_1,\ldots,x_n))\right) = \log\left(\prod_{i=1}^{n} p^{x_i}\cdot(1-p)^{1-x_i}\right) = \sum_{i=1}^{n} x_i \cdot \log(p) + \sum_{i=1}^{n}(1-x_i)\cdot\log(1-p)$$

Die erste Ableitung bezüglich p ist dann

$$\frac{d\log\left(\mathcal{L}(p,(x_1,\ldots,x_n))\right)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-p}$$

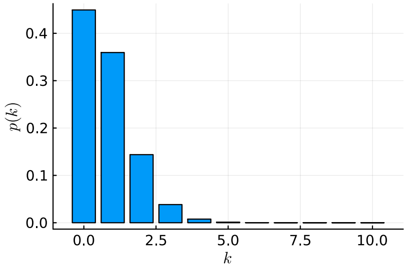
Für die Nullstelle $\hat{p} = T_{\mathrm{MLE}}(X_1, \ldots, X_n)$ der ersten Ableitung gilt dann

$$0 = \frac{\sum_{i=1}^{n} x_i}{\hat{p}} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-\hat{p}} \Leftrightarrow \frac{(1-\hat{p})}{\sum_{i=1}^{n}(1-x_i)} = \frac{\hat{p}}{\sum_{i=1}^{n} x_i} \Leftrightarrow \boxed{\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

MLE für die Poissionverteilung

- **Satz (MLE der Poissionverteilung).** Seien $X_1, \ldots, X_n$ unabhängig und $X_i \sim \mathrm{Pois}(\lambda)$. Dann ist der Maximum *Likelihood* Schätzer T_{MLE} für λ gegeben durch

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



- **Beweis**: Wir betrachten die log-likelihood Funktion

$$\log\left(\mathcal{L}(\lambda, (x_1, \ldots, x_n))\right) = \log\left(\prod_{i=1}^{n} \exp(-\lambda) \cdot \frac{\lambda^{x_i}}{x_i!}\right) = -n \cdot \lambda + \sum_{i=1}^{n} \left(x_i \cdot \log(\lambda) - \log(x_i!)\right)$$

Die erste Ableitung bezüglich λ ist dann

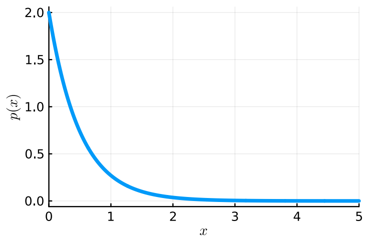
$$\frac{d \log\left(\mathcal{L}(\lambda, (x_1, \ldots, x_n))\right)}{d\lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$

Für die Nullstelle $\hat{\lambda} = T_{\mathrm{MLE}}(X_1, \ldots, X_n)$ der ersten Ableitung gilt dann

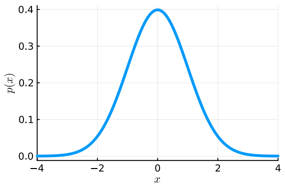
$$0 = -n + \frac{\sum_{i=1}^{n} x_i}{\hat{\lambda}} \quad \Leftrightarrow \quad n \cdot \hat{\lambda} = \sum_{i=1}^{n} x_i \quad \Leftrightarrow \quad \boxed{\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

MLE für die Exponentialverteilung

- **Satz (MLE der Exponentialverteilung).** Seien $X_1, \dots, X_n$ unabhängig und $X_i \sim \mathcal{E}(\lambda)$. Dann ist der Maximum *Likelihood* Schätzer T_{MLE} für λ gegeben durch
$$T_{\text{MLE}}(X_1, \dots, X_n) = \frac{n}{\sum_{i=1}^n X_i}$$
- **Beweis**: Wir betrachten die log-likelihood Funktion
$$\log\left(\mathcal{L}(\lambda, (x_1, \dots, x_n))\right) = \log\left(\prod_{i=1}^n \lambda \cdot \exp(-\lambda \cdot x_i)\right) = n \cdot \log(\lambda) - \sum_{i=1}^n \lambda \cdot x_i$$
Die erste Ableitung bezüglich λ ist dann
$$\frac{d \log\left(\mathcal{L}(\lambda, (x_1, \dots, x_n))\right)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$
Für die Nullstelle $\hat{\lambda} = T_{\text{MLE}}(X_1, \dots, X_n)$ der ersten Ableitung gilt dann
$$0 = \frac{n}{\hat{\lambda}} - \sum_{i=1}^n x_i \quad \Leftrightarrow \quad \hat{\lambda} \cdot \sum_{i=1}^n x_i = n \quad \Leftrightarrow \quad \boxed{\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i}}$$



MLE für den Erwartungswert der Normalverteilung



- **Satz (MLE der Normalverteilung).** Seien $X_1, \ldots, X_n$ unabhängig und $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Dann ist der Maximum *Likelihood* Schätzer T_{MLE} für μ gegeben durch

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

- **Beweis**: Wir betrachten die log-likelihood Funktion

$$\log\left(\mathcal{L}(\mu, (x_1, \ldots, x_n))\right) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(x_i - \mu)^2}{\sigma^2}\right)\right) = -\frac{n}{2} \cdot \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{1}{2} \cdot \frac{(x_i - \mu)^2}{\sigma^2}$$

Die erste Ableitung bezüglich μ ist dann

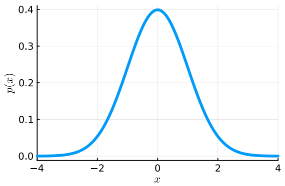
$$\frac{d \log\left(\mathcal{L}(\mu, (x_1, \ldots, x_n))\right)}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

Für die Nullstelle $\hat{\mu} = T_{\text{MLE}}(X_1, \ldots, X_n)$ der ersten Ableitung gilt dann

$$0 = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu}) \quad \Leftrightarrow \quad \sum_{i=1}^{n} x_i = n \cdot \hat{\mu} \quad \Leftrightarrow \quad \boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

MLE für die Varianz der Normalverteilung

- **Satz (MLE der Normalverteilung).** Seien $X_1, \dots, X_n$ unabhängig und $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Dann ist der Maximum *Likelihood* Schätzer T_{MLE} für σ^2 gegeben durch



$$T_{\text{MLE}}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

Nicht erwartungstreu, wie die korrigiert empirische Varianz $\hat{V}[X_1, \dots, X_n]$

- **Beweis**: Wir betrachten die log-likelihood Funktion

$$\log\left(\mathcal{L}\left(\sigma^2, (x_1, \dots, x_n)\right)\right) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(x_i - \mu)^2}{\sigma^2}\right)\right) = -\frac{n}{2} \cdot \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{1}{2} \cdot \frac{(x_i - \mu)^2}{\sigma^2}$$

Die erste Ableitung bezüglich σ^2 ist dann

$$\frac{d \log\left(\mathcal{L}\left(\sigma^2, (x_1, \dots, x_n)\right)\right)}{d\sigma^2} = -\frac{n}{2 \cdot \sigma^2} + \frac{1}{2} \cdot \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^4}$$

Für die Nullstelle $\hat{\sigma}^2 = T_{\text{MLE}}(X_1, \dots, X_n)$ der ersten Ableitung gilt dann

$$0 = -\frac{n}{2 \cdot \hat{\sigma}^2} + \frac{1}{2} \cdot \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\hat{\sigma}^4} \Leftrightarrow n \cdot \hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \Leftrightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Überblick

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

Motivation: Warteschlangen

- **Beispiel (Warteschlangen).** Sie fahren jeden Morgen die gleiche Strecke zur Arbeit und kommen auf dem Weg zur Arbeit an einer Ampelkreuzung vorbei. Über die letzten 5 Tage notieren Sie sich die Wartezeiten $X_1, \dots, X_5$ und beobachten die folgende Verteilung. Wenn Sie annehmen, dass die Wartezeit gleichverteilt zwischen 0 und θ ist, was ist dann Ihre MLE-Schätzung für θ?
  - **Lösung**: Wir betrachten die log-likelihood Funktion

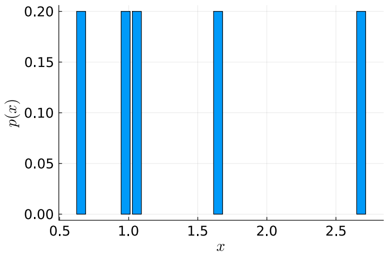
$$\log\left(\mathcal{L}(\theta,(x_1,\dots,x_n))\right) = \log\left(\prod_{i=1}^{n}\frac{1}{\theta}\cdot\mathbb{I}_{[0,\theta]}(x_i)\right) = \begin{cases}-n\cdot\log(\theta) & \theta > \max(x_1,\dots,x_n)\\ -\infty & \text{sonst}\end{cases}$$

Die erste Ableitung (im relevanten Bereich) bezüglich θ ist dann

$$\frac{d\log\left(\mathcal{L}(\theta,(x_1,\dots,x_n))\right)}{d\theta} = -\frac{n}{\theta} < 0$$

Damit hat die erste Ableitung keine Nullstelle, ist monoton fallend und wird daher maximiert für den kleinsten zulässigen Wert von θ, d.h. $T_{\text{MLE}}(x_1,\dots,x_5) = 2.68$!

- **Bemerkung (MLE der Gleichverteilung)**
  - Der MLE ist sehr anfällig für einzelne Ausreißer in den Daten!

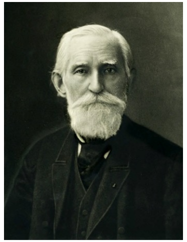


Momente einer Verteilung

- **Definition (k-tes Moment).** Sei X eine reelle Zufallsvariable. Dann heißt $E[X^k]$ das k-te Moment von X, $k \in \{0,1,2,\dots\}$. Gegeben eine Stichprobe $x_1, \dots, x_n$, ist das k-te empirische Moment definiert als

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k$$

- **Bemerkungen (Momente)**
  - Die Verteilungsfunktion einer Zufallsvariable ist eindeutig durch alle Momente $\{E[X^k]\}_{k\in\mathbb{N}}$ bestimmt.
  - Wir haben schon das 0-te, 1-te und 2-te Moment kennengelernt unter dem Namen
    1. Normierung: $E[X^0] = 1$
    2. Erwartungswert: $E[X^1]$
    3. Varianzzerlegung: $E[X^2] = V[X] + (E[X^1])^2$
  - Momente waren auch für den Beweis des zentralen Grenzwertsatzes hilfreich.



Pafnuty Chebyshev (1821 – 1894)

Methods of Moments

- **Definition (*Methods of Moments* (MoM) Schätzer).** Sei $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\prod_{i=1}^n P_\theta \mid \theta \in \Theta\})$ ein parametrisches Modell, das eine unabhängige und identische Verteilung P_θ annimmt. Ein Schätzer $T_{\mathrm{MoM},k}: \mathbb{R}^n \to \Theta$ heißt ein *Methods of Moments* Schätzer, wenn er diejenige Verteilung berechnet, bei welcher das k-te Moment dem k-ten empirischen Moment entspricht

$$T_{\mathrm{MoM},k}(x_1, \dots, x_n) \in \left\{ \theta \in \Theta \;\middle|\; E_{X\sim P_\theta}[X^k] = \frac{1}{n}\sum_{i=1}^n x_i^k \right\}$$

- **Beispiel (Warteschlangen).** Für $k = 1$ und $X \sim \mathcal{U}(0, \theta)$ wissen wir, dass

$$E[X^1] = \frac{\theta}{2}$$

Daher ist der $\hat{\theta} = T_{\mathrm{MoM},1}(x_1, \dots, x_n)$ gegeben durch

$$\frac{\hat{\theta}}{2} = \frac{1}{n}\sum_{i=1}^n x_i \quad \Leftrightarrow \quad \hat{\theta} = \frac{2}{n}\sum_{i=1}^n x_i$$

- **Bemerkung (*Methods of Moments* Schätzer der Gleichverteilung)**
  - Der MoM Schätzer ist nicht anfällig für einzelne Ausreißer in den Daten!

Karl Pearson
(1857 – 1936)

Methods of Moments Beispiel

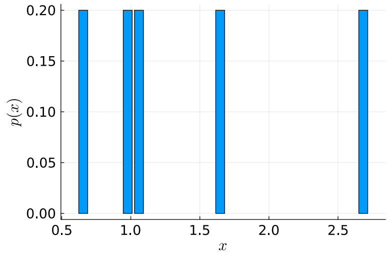
- **Beispiel (Warteschlangen).** Welchen MoM Schätzer bekommen wir für $k = 2$ und $X \sim \mathcal{U}(0, \theta)$?
- **Lösung**: Wir wissen aus den Eigenschaften der Gleichverteilung, dass

$$E[X] = \frac{\theta}{2} \qquad V[X] = \frac{\theta^2}{12} \quad \Rightarrow \quad E[X^2] = V[X] + (E[X])^2 = \frac{\theta^2}{12} + \frac{\theta^2}{4} = \frac{\theta^2}{3}$$

Daher ist der $\hat{\theta} = T_{\mathrm{MoM},2}(x_1, \dots, x_n)$ für $k = 2$ gegeben durch

$$\frac{\hat{\theta}^2}{3} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \hat{\theta} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} x_i^2}$$

- **Bemerkung (Schätzer der Gleichverteilung).** Die Schätzer für unsere Daten
  - $T_{\mathrm{MLE}}(x_1, \dots, x_5) = 2.68$
  - $T_{\mathrm{MoM},1}(x_1, \dots, x_5) = 2.81$
  - $T_{\mathrm{MoM},2}(x_1, \dots, x_5) = 2.73$



Methods of Moments für die Normalverteilung

- **Beispiel (MoM Schätzer für den Erwartungswert).** Welchen MoM Schätzer bekommen wir für den Erwartungswert μ für $k = 1$ und $X \sim \mathcal{N}(\mu, \sigma^2)$?
- **Lösung**. Wir wissen aus den Eigenschaften der Normalverteilung $\mathcal{N}(\mu, \sigma^2)$, dass

$$E[X] = \mu \quad \Leftrightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Beispiel (MoM Schätzer für die Varianz).** Welchen MoM Schätzer bekommen wir für die Varianz σ^2 für $k = 2$ und $X \sim \mathcal{N}(\mu, \sigma^2)$?
- **Lösung**. Wir wissen aus den Eigenschaften der Normalverteilung $\mathcal{N}(\mu, \sigma^2)$, dass

$$E[X] = \mu \qquad V[X] = \sigma^2 \quad \Rightarrow \quad E[X^2] = V[X] + (E[X])^2 = \sigma^2 + \mu^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\hat{\mu}^2 + \hat{\mu}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} 2x_i \cdot \hat{\mu}\right) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Wann benutzen wir welchen Schätzer?

- **Maximum *Likelihood* Schätzer (MLE)**
  - Hat per Definition immer die maximale *Likelihood*
  - Ist per Definition immer im erlaubten Parameterraum
  - Berücksichtigt alle verfügbare Information
- ***Methods of Moments* Schätzer**
  - Ist auch berechenbar, wenn der MLE nicht analytisch lösbar ist
  - Ist oft ein guter Startpunkt für die numerische MLE-Lösung
  - Ist auch schon bei sehr kleinen Stichprobengrößen einsetzbar
  - Ist per Definition für $k = 1$ erwartungstreu
- **Bemerkungen (Verteilungsannahme)**
  - Parameterschätzer wie MLE und MoM liefern unterschiedliche Ergebnisse, je nachdem, welche Verteilungsannahme getroffen wurde.
  - Schätzer wie MLE und MoM sind nicht dafür geeignet zu entscheiden, ob eine bestimmte Annahme der unterliegenden Verteilung korrekt ist!

Viel Spaß bis zur nächsten Vorlesung!