Eigenschaften des Schätzers $\bar{X}$ für den unbekannten Erwartungswert μ

- Plausibles Vorgehen: Schätze den unbekannten Mittelwert der Grundgesamtheit durch den Mittelwert aus der Stichprobe
- $\bar{X}$ ist erwartungstreuer Schätzer
- $\bar{X}$ hat die Standardabweichung $\sigma/\sqrt{n}$
  Diese wird häufig als Standardfehler bezeichnet
- „Um die Schätzgenauigkeit zu verdoppeln, ist eine Erhöhung des Stichprobenumfangs n um den Faktor 4 nötig“

Wahlumfrage

Gegeben sei eine Stichprobe der wahlberechtigten Bundesbürger. Geben Sie einen erwartungstreuen Schätzer des Anteils der rot-grün Wähler an.
Grundgesamtheit: Dichotomes Merkmal

$$\tilde{X} = \begin{cases} 1 & \text{rot/grün: ja} \\ 0 & \text{rot/grün: nein} \end{cases}$$

Der Mittelwert π von $\tilde{X}$ ist der Anteil der rot/grün-Wähler in der Grundgesamtheit.
Stichprobe $X_1, \ldots, X_n$ vom Umfang n:

$$X_i = \begin{cases} 1 & i\text{-te Person wählt rot/grün} \\ 0 & \text{sonst} \end{cases}$$

Aus den Überlegungen zum arithmetischen Mittel folgt, dass

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ein erwartungstreuer Schätzer für den hier betrachteten Parameter π ist. Also verwendet man die relative Häufigkeit in der Stichprobe, um den wahren Anteil π in der Grundgesamtheit zu schätzen.

Bedeutung der Erwartungstreue:

Erwartungstreue ist ein schwaches Kriterium!
Betrachte die offensichtlich unsinnige Schätzfunktion

$$T_2 = g_2(X_1, \ldots, X_n) = X_1,$$

d.h. $T_2 = 100\%$, falls der erste Befragte rot-grün wählt und $T_2 = 0\%$ sonst.
Die Schätzfunktion ignoriert fast alle Daten, ist aber erwartungtreu:

$$\mathbb{E}(T_2) = \mathbb{E}(X_1) = \mu$$

Deshalb betrachtet man zusätzlich die Effizienz eines Schätzers

Effizienz

Beispiel Wahlumfrage: Gegeben sind zwei erwartungstreue Schätzer (n sei gerade):

$$\begin{aligned} T_1 &= \frac{1}{n} \sum_{i=1}^{n} X_i \\ T_2 &= \frac{1}{n/2} \sum_{i=1}^{n/2} X_i \end{aligned}$$

Was unterscheidet formal T_1 von dem unsinnigen Schätzer T_2, der die in der Stichprobe enthaltene Information nicht vollständig ausnutzt? Vergleiche die Schätzer über ihre Varianz, nicht nur über den Erwartungswert!

Effizienz

Wenn n so groß ist, dass der zentrale Grenzwertsatz angewendet werden kann, dann gilt approximativ

$$\frac{1}{\sqrt{n}} \frac{\sum_{i=1}^{n}(X_i - \pi)}{\sqrt{\pi(1-\pi)}} = \frac{\sum_{i=1}^{n} X_i - n \cdot \pi}{\sqrt{n}\sqrt{\pi(1-\pi)}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim \mathrm{N}(0;1)$$

und damit

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\pi; \frac{\pi(1-\pi)}{n}\right).$$

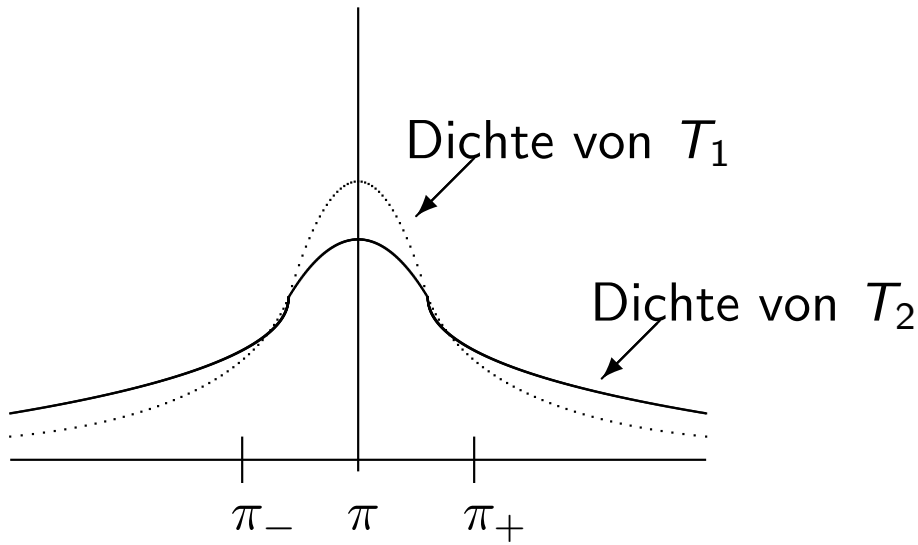
Effizienz

Analog kann man zeigen:

$$T_2 = \frac{1}{n/2} \sum_{i=1}^{n/2} X_i \sim \mathrm{N}\left(\pi, \frac{\pi(1-\pi)}{n/2}\right).$$

T_1 und T_2 sind approximativ normalverteilt, wobei T_1 eine deutlich kleinere Varianz als T_2 hat.
T_1 und T_2 treffen beide im Durchschnitt den richtigen Wert π. T_1 schwankt aber weniger um das wahre π, ist also „im Durchschnitt genauer“.

Effizienz



Für jeden Punkt $\pi_+ > \pi$ ist damit $P(T_1 > \pi_+) < P(T_2 > \pi_+)$
und für jeden Punkt $\pi_- < \pi$ ist $P(T_1 < \pi_-) < P(T_2 < \pi_-)$.

Effizienz

Es ist also die Wahrscheinlichkeit, mindestens um $\pi_+ - \pi$ bzw. $\pi - \pi_-$ daneben zu liegen, bei T_2 stets größer als bei T_1. Umgekehrt gesagt: Ein konkreter Wert ist damit verlässlicher, wenn er von T_1, als wenn er von T_2 stammt.
Diese Überlegung gilt ganz allgemein: Ein erwartungstreuer Schätzer ist umso besser, je kleiner seine Varianz ist.

$$\text{Var}(T) = \text{Erwartete quadratische Abweichung von } T \text{ von } \underbrace{\mathbb{E}(T)}_{=\vartheta\,!}$$

Je kleiner die Varianz, umso mehr konzentriert sich die Verteilung eines erwartungstreuen Schätzers um den wahren Wert. Dies ist umso wichtiger, da der Schätzer den wahren Wert i.A. nur selten exakt trifft.

Beste Schätzer

- Gegeben seien zwei erwartungstreue Schätzfunktionen T_1 und T_2 für einen Parameter ϑ. Gilt

$$\text{Var}_\vartheta(T_1) \leq \text{Var}_\vartheta(T_2) \text{ für alle } \vartheta$$

und

$$\text{Var}_{\vartheta^*}(T_1) < \text{Var}_{\vartheta^*}(T_2) \text{ für mindestens ein } \vartheta^*$$

so heißt T_1 *effizienter als* T_2.

- Eine für ϑ erwartungstreue Schätzfunktion T heißt *UMVU-Schätzfunktion* für ϑ (*u*niformly *m*inimum *v*ariance *u*nbiased), falls

$$\text{Var}_\vartheta(T) \leq \text{Var}_\vartheta(T^*)$$

für alle ϑ und für alle erwartungstreuen Schätzfunktionen T^*.

UMVU-Schätzer

- *Inhaltliche Bemerkung:* Der (tiefere) Sinn von Optimalitätskriterien wird klassischerweise insbesondere auch in der *Gewährleistung von Objektivität* gesehen.
- Ist $X_1, \ldots, X_n$ eine i.i.d. Stichprobe mit $X_i \sim \mathrm{N}(\mu, \sigma^2)$, dann ist
  - $\bar{X}$ UMVU-Schätzfunktion für μ und
  - S^2 UMVU-Schätzfunktion für σ^2.

Verzerrte Schätzer

- Ist $X_1, \ldots, X_n$ mit $X_i \in \{0, 1\}$ eine i.i.d. Stichprobe mit $\pi = P(X_i = 1)$, dann ist die relative Häufigkeit $\bar{X}$ UMVU-Schätzfunktion für π.
- Bei nicht erwartungstreuen Schätzern macht es keinen Sinn, sich ausschließlich auf die Varianz zu konzentrieren.
- Z.B. hat der unsinnige Schätzer $T = g(X_1, \ldots, X_n) = 42$, der die Stichprobe nicht beachtet, Varianz 0.

MSE

- Man zieht dann den sogenannten *Mean Squared Error*

$$\mathrm{MSE}_\vartheta(T) = \mathbb{E}_\vartheta(T - \vartheta)^2$$

zur Beurteilung heran. Es gilt

$$\mathrm{MSE}_\vartheta(T) = \mathrm{Var}_\vartheta(T) + (\mathrm{Bias}_\vartheta(T))^2.$$

Der MSE kann als Kompromiss zwischen zwei Auffassungen von Präzision gesehen werden: möglichst geringe systematische Verzerrung (Bias) und möglichst geringe Schwankung (Varianz).

Asymptotische Erwartungstreue

* Eine Schätzfunktion heißt asymptotisch erwartungstreu, falls

$$\lim_{n \to \infty} E(\hat{\theta}) = \theta$$

bzw.

$$\lim_{n \to \infty} \text{Bias}(\hat{\theta}) = 0$$

gelten.

* Abschwächung des Begriffs der Erwartungstreue: Gilt nur noch bei einer unendlich großen Stichprobe.
* Erwartungstreue Schätzer sind auch asymptotisch erwartungstreu.
* Sowohl S^2 als auch $\tilde{S}^2$ sind asymptotisch erwartungstreu.

Konsistenz

- Für komplexere Modelle ist oft die Erwartungstreue der Verfahren ein zu restriktives Kriterium. Man fordert deshalb oft nur, dass sich der Schätzer wenigstens für große Stichproben gut verhält. Hierzu gibt es v.a. zwei verwandte aber „etwas“ unterschiedliche Kriterien.
- Ein Schätzer heißt (MSE-)konsistent oder konsistent im quadratischen Mittel, wenn gilt

$$\lim_{n \to \infty} (\mathrm{MSE}(T)) = 0.$$

Beispiel: $\bar{X}$

Der MSE von $\bar{X}$ ist gegeben durch

$$\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) + \text{Bias}^2(\bar{X}) = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n} \to 0.$$

$\bar{X}$ ist also ein MSE-konsistente Schäter für den Erwartungswert. Anschaulich bedeutet die Konsistenz, dass sich die Verteilung des Schätzers für wachsenden Stichprobenumfang n immer stärker beim richtigen Wert „zusammenzieht“. Er trifft also für unendlich große Stichproben praktische sicher den wahren Wert. (Dies gilt als eine Minimalanforderung an statistische Verfahren.)

Maximum Likelihood

Das Maximum-Likelihood-Prinzip Sie wissen als Wirt, dass heute die Lokalparteien ihre Busausflüge unternehmen: Es werden Busse mit je 100 Personen von der jeweiliger Partei organisiert. Bus I: 85% Partei A, 15% Partei B
Bus II: 15% Partei A, 85% Partei B
Bus fährt vor, anhand Stichprobe ermitteln, ob Bild von ... von der Wand genommen werden soll oder nicht.
Stichprobe von 10 Personen ergibt 80% Anhänger der Partei A, welche Partei: wohl A, aber B nicht ausgeschlossen bei unglücklicher Auswahl.
Warum: A ist plausibler, da die Wahrscheinlichkeit, ungefähr den in der Stichprobe beobachteten Wert zu erhalten (bzw. erhalten zu haben) bei Bus I wesentlich größer ist als bei Bus II.

ML-Prinzip

Aufgabe: Schätze den Parameter ϑ eines parametrischen Modells anhand einer i.i.d. Stichprobe $X_1, \ldots, X_n$ mit der konkreten Realisation $x_1, \ldots, x_n$.
Idee der Maximium-Likelihood (ML) Schätzung für diskrete Verteilungen:

- Man kann für jedes ϑ die Wahrscheinlichkeit ausrechnen, genau die Stichprobe $x_1, \ldots, x_n$ zu erhalten:

$$P_\vartheta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P_\vartheta(X_i = x_i)$$

- Je größer für ein gegebenes ϑ_0 die Wahrscheinlichkeit ist, die konkrete Stichprobe erhalten zu haben, umso plausibler ist es, dass tatsächlich ϑ_0 der wahre Wert ist (gute Übereinstimmung zwischen Modell und Daten).

Beispiel

I.i.d. Stichprobe vom Umfang $n = 5$ aus einer $B(10, \pi)$-Verteilung:

$$6 \quad 5 \quad 3 \quad 4 \quad 4$$

Wahrscheinlichkeit der Stichprobe für gegebenes π:

$$\begin{aligned} P(X_1 = 6, \ldots, X_5 = 4 || \pi) &= P(X_1 = 6 || \pi) \cdot \ldots \cdot P(X_5 = 4 || \pi) \\ &= \binom{10}{6} \pi^6 (1-\pi)^4 \cdot \ldots \cdot \binom{10}{4} \pi^4 (1-\pi)^6. \end{aligned}$$

„$P(\ldots || \pi)$ Wahrscheinlichkeit, wenn π der wahre Parameter ist“

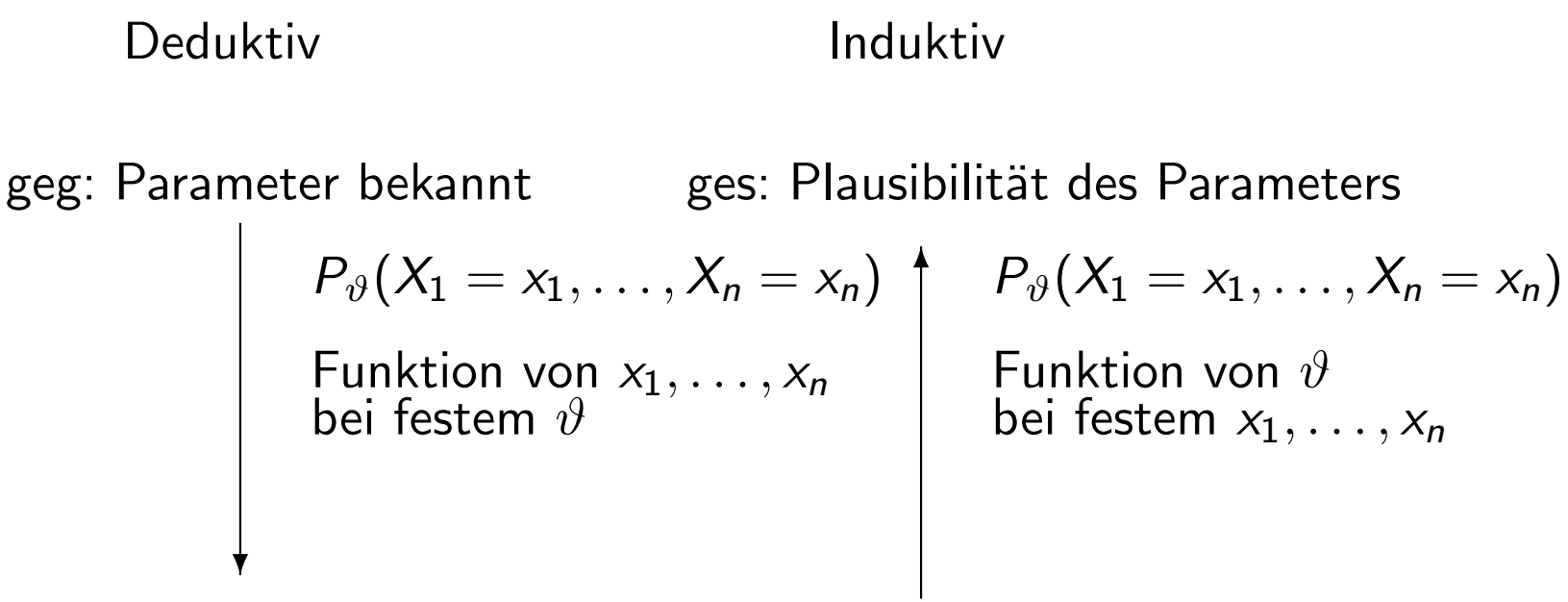
Wahrscheinlichkeit für einige Werte von π:

π	$P(X_1 = 6, \ldots, X_5 = 4 \mid \pi)$
0.1	0.0000000000001
0.2	0.0000000227200
0.3	0.0000040425220
0.4	0.0003025481000
0.5	0.0002487367000
0.6	0.0000026561150
0.7	0.0000000250490
0.8	0.0000000000055
0.9	0.0000000000000

Man nennt daher $L(\vartheta) = P_\vartheta(X_1 = x_1, \ldots, X_n = x_n)$, nun als Funktion von ϑ gesehen, die *Likelihood* (deutsch: Plausibilität, Mutmaßlichkeit) von ϑ gegeben die Realisation $x_1, \ldots, x_n$. Derjenige Wert $\hat{\vartheta} = \hat{\vartheta}(x_1, \ldots, x_n)$, der $L(\vartheta)$ maximiert, heißt *Maximum-Likelihood-Schätzwert*; die zugehörige Schätzfunktion $T(X_1, \ldots, X_n)$ *Maximum-Likelihood-Schätzer*

$P_\vartheta(X_1 = x_1, \ldots, X_n = x_n)$:

- Deduktiv (Wahrscheinlichkeitsrechnung): ϑ bekannt, $x_1, \ldots, x_n$ zufällig („unbekannt“).
- Induktiv (Statistik): ϑ unbekannt, $x_1, \ldots, x_n$ bekannt.



Dies liefert denselben Schätzwert $\hat{\vartheta}$ und erspart beim Differenzieren die Anwendung der Produktregel.
Der Logarithmus ist streng monoton wachsend. Allgemein gilt für streng monoton wachsende Funktionen g: x_0 Stelle des Maximums von $L(x)$ $\Longleftrightarrow$ x_0 auch Stelle des Maximums von $g(L(x))$.

Definition ML

Gegeben sei die Realisation $x_1, \ldots, x_n$ einer i.i.d. Stichprobe. Die Funktion in ϑ

$$L(\vartheta) = \begin{cases} \prod_{i=1}^{n} P_\vartheta(X_i = x_i) & \text{falls } X_i \text{ diskret} \\ \prod_{i=1}^{n} f_\vartheta(x_i) & \text{falls } X_i \text{ stetig.} \end{cases}$$

heißt *Likelihood* des Parameters ϑ bei der Beobachtung $x_1, \ldots, x_n$. Derjenige Wert $\hat{\vartheta} = \hat{\vartheta}(x_1, \ldots, x_n)$, der $L(\vartheta)$ maximiert, heißt *Maximum-Likelihood-Schätzwert*; die zugehörige Schätzfunktion $T(X_1, \ldots, X_n)$ *Maximum-Likelihood-Schätzer*.

Stetige Verteilungen

- In diesem Fall verwendet man die Dichte

$$f_\vartheta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\vartheta(x_i)$$

als Maß für die Plausibilität von ϑ.

- Für die praktische Berechnung maximiert man statt der Likelihood typischerweise die Log-Likelihood

$$l(\vartheta) = \ln(L(\vartheta)) = \ln \prod_{i=1}^{n} P_\vartheta(X_i = x_i) = \sum_{i=1}^{n} \ln P_\vartheta(X_i = x_i)$$

bzw.

$$l(\vartheta) = \ln \prod_{i=1}^{n} f_\vartheta(x_i) = \sum_{i=1}^{n} \ln f_\vartheta(x_i).$$

$$X_i = \begin{cases} 1 & \text{falls Rot/Grün} \\ 0 & \text{sonst} \end{cases}$$

Verteilung der X_i: Binomialverteilung $B(1, \pi)$ (Bernoulliverteilung)

$$\begin{aligned} P(X_i = 1) &= \pi \\ P(X_i = 0) &= 1 - \pi \\ P(X_i = x_i) &= \pi^{x_i} \cdot (1 - \pi)^{1 - x_i} \ , \quad x_i \in \{0; 1\}. \end{aligned}$$

Hier ist π der unbekannte Parameter, der allgemeine mit ϑ bezeichnet wird.

Likelihood:

$$
\begin{aligned}
L(\pi) &= P(X_1 = x_1, \ldots, X_n = x_n) \\
&= \prod_{i=1}^{n} \pi^{x_i} (1-\pi)^{1-x_i} \\
&= \pi^{\sum_{i=1}^{n} x_i} \cdot (1-\pi)^{n - \sum_{i=1}^{n} x_i}
\end{aligned}
$$

Logarithmierte Likelihood:

$$l(\pi) = \ln(P(X_1 = x_1, \ldots, X_n = x_n)) = \sum_{i=1}^{n} x_i \cdot \ln(\pi) + (n - \sum_{i=1}^{n} x_i) \cdot \ln(1-\pi)$$

Ableiten (nach π):

$$\frac{\partial}{\partial \pi} l(\pi) = \frac{\sum_{i=1}^{n} x_i}{\pi} + \frac{n - \sum_{i=1}^{n} x_i}{1-\pi} \cdot (-1)$$

Nullsetzen und nach π auflösen ergibt:

$$\frac{\partial}{\partial \pi} l(\pi) = 0 \iff \frac{\sum_{i=1}^{n} x_i}{\pi} = \frac{n - \sum_{i=1}^{n} x_i}{1 - \pi}$$

$$\iff (1 - \pi) \sum_{i=1}^{n} x_i = n \cdot \pi - \pi \sum_{i=1}^{n} x_i$$

$$\iff \sum_{i=1}^{n} x_i = n \cdot \pi$$

also

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

Also ist $\bar{X}$ der Maximum-Likelihood-Schätzer für π.

ML-Schätzung bei Normalverteilung

- Bestimme die Likelihoodfunktion

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}(\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \frac{1}{2\pi^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{aligned}$$

- Bestimme die Log-Likelihoodfunktion

$$\begin{aligned} l(\mu, \sigma^2) &= \ln(L(\mu, \sigma^2)) \\ &= \ln(1) - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

ML-Schätzung bei Normalverteilung

- Ableiten und Nullsetzen der Loglikelihoodfunktion

$$\begin{aligned}
\frac{\partial l(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{2\sigma^2} 2 \cdot \sum_{i=1}^{n} (x_i - \mu) = 0 \\
\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0
\end{aligned}$$

ML-Schätzung bei Normalverteilung

- Auflösen der beiden Gleichungen nach μ und σ^2

  Aus der ersten Gleichung erhalten wir

$$\sum_{i=1}^{n} x_i - n\mu = 0 \quad \text{also} \quad \hat{\mu} = \bar{x}.$$

  Aus der zweiten Gleichung erhalten wir durch Einsetzen von $\hat{\mu} = \bar{x}$

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

  also

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

Fazit

- Der ML-Schätzer $\hat{\mu} = \bar{X}$ für μ stimmt mit dem üblichen Schätzer für den Erwartungswert überein.
- Der ML-Schätzer $\hat{\sigma}^2 = \tilde{S}^2$ für σ^2 ist verzerrt, d.h. nicht erwartungstreu.

Einige allgemeine Eigenschaften von ML-Schätzern

- ML-Schätzer $\hat{\theta}$ sind im Allgemeinen nicht erwartungstreu.
- ML-Schätzer $\hat{\theta}$ sind asymptotisch erwartungstreu.
- ML-Schätzer $\hat{\theta}$ sind konsistent (und meist in einem asymptotischen Sinne effizient).