VL7: Die Maximum-Likelihood-Methode

Die Maximum-Likelihood-Methode ist eine Standard-Methode der Statistik, um die Modell-Parameter eines statistischen Modells zu bestimmen. Maximum Likelihood Schätzer haben in der Regel sehr gute statistische Eigenschaften. Die Funktionsweise der Methode versteht man am besten an einem

Beispiel: Gegeben seien n Zufallszahlen $x_1, x_2, \cdots, x_n$. Wir wissen, dass diese Zahlen mit Hilfe einer Normalverteilung generiert worden sind, kennen aber nicht den Mittelwert μ und die Standardabweichung σ. Welche Werte von μ und σ passen 'am besten' zu den gegebenen Zahlen $x_1, \cdots, x_n$? Und welches Kriterium nehmen wir für 'am besten'?

Lösung: Die Aussage: "Die x_i sind normalverteilt mit Mittelwert μ und Standardabweichung σ" ist äquivalent zu

$$\mathsf{Prob}\big[\, x_i \in [\tilde{x}_i, \tilde{x}_i + d\tilde{x}_i) \,\big] \;=\; \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\tilde{x}_i-\mu)^2}{2\sigma^2}}\, d\tilde{x}_i \tag{1}$$

Wir nehmen an, dass wir unabhängige Zufallszahlen haben, und bekommen dann aus (1)

$$\begin{aligned}
&\mathsf{Prob}\Big[\, x_1 \in [\tilde{x}_1, \tilde{x}_1 + d\tilde{x}_1),\; x_2 \in [\tilde{x}_2, \tilde{x}_2 + d\tilde{x}_2),\; \cdots,\; x_n \in [\tilde{x}_n, \tilde{x}_n + d\tilde{x}_n) \Big] \;= \\
&\overset{\text{unabh.}}{=}\; \mathsf{Prob}\big[\, x_1 \in [\tilde{x}_1, \tilde{x}_1 + d\tilde{x}_1) \,\big]\; \mathsf{Prob}\big[\, x_2 \in [\tilde{x}_2, \tilde{x}_2 + d\tilde{x}_2) \,\big]\; \cdots\; \mathsf{Prob}\big[\, x_n \in [\tilde{x}_n, \tilde{x}_n + d\tilde{x}_n) \,\big] \\
&\overset{(1)}{=}\; \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\tilde{x}_i-\mu)^2}{2\sigma^2}}\, d\tilde{x}_i \\
&=\; (2\pi\sigma^2)^{-\frac{n}{2}}\; \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\tilde{x}_i - \mu)^2 \Big\}\, d\tilde{x}_1 \cdots d\tilde{x}_n
\end{aligned} \tag{2}$$

Jetzt haben wir ja konkrete Werte für die Zufallszahlen, nämlich $x_1, \cdots, x_n$. Diese Werte können wir auf der rechten Seite von (2) für die $\tilde{x}_i$ einsetzen. Die $d\tilde{x}_i$ sind Intervallbreiten, etwa $d\tilde{x}_i = 0.1$ oder $d\tilde{x}_i = 0.01$. Wir werden gleich sehen, dass diese Intervallbreiten für die eigentliche Rechnung keine Rolle spielen, aber um eine konkrete Vorstellung zu haben, wählen wir etwa $d\tilde{x}_i = 0.01$ und wir lassen die Tilde weg, schreiben einfach dx_i. Dann bekommen wir eine Funktion, die nur noch von μ und σ abhängt,

$$L(\mu, \sigma) \;:=\; (2\pi\sigma^2)^{-\frac{n}{2}}\; \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big\}\, dx_1 \cdots dx_n \tag{3}$$

Diese Funktion heisst **Likelihood-Funktion**, sie gibt also die W'keit an, dass, unter der Annahme, die Zufallszahlen sind von einer Normalverteilung mit Mittelwert μ und Standard-abweichung σ generiert worden, dass sich dann konkret Werte in den Intervallen $[x_i, x_i + dx_i)$

realisieren. Also ist es plausibel, das μ und das σ dann so zu wählen, dass diese W'keit maximal wird,

$$L(\mu,\sigma) \stackrel{!}{\to} \max \qquad \Rightarrow \qquad \mu = \mu_{\mathrm{ML}},\ \sigma = \sigma_{\mathrm{ML}} \tag{4}$$

Die so gewonnenen Werte von μ und σ heissen dann die **Maximum-Likelihood-Schätzer** von μ und σ.

In unserem konkreten Beispiel können wir das Maximum der Likelihood-Funktion (3) analytisch in geschlossener Form berechnen, das machen wir gleich in dem Theorem 7.1 weiter unten. In vielen Anwendungssituationen ist jedoch eine analytische Berechnung in geschlossener Form nicht möglich und man ist auf numerische Prozeduren angewiesen. Anstatt die Likelihood-Funktion L selber zu maximieren, betrachtet man dann typischerweise den Logarithmus der Likelihood-Funktion $\log L$. Da der Logarithmus eine streng monoton steigende Funktion ist, gilt

$$(\mu,\sigma) \ \text{ maximiert } \ L(\mu,\sigma) \qquad \Leftrightarrow \qquad (\mu,\sigma) \ \text{ maximiert } \ \log L(\mu,\sigma)$$

Das macht man jetzt aus folgendem Grund: Das L ist typischerweise durch ein Produkt gegeben und die Anzahl der Faktoren in diesem Produkt ist gleich der Anzahl der verfügbaren Daten. Wenn man etwa eine Finanz-Zeitreihe hat mit täglichen Schlusskursen der letzten 10 Jahre, dann sind das etwa $N = 2500$ Daten und das L ist dann im wesentlichen

$$L \ \sim \ \mathrm{prob}^N \ = \ \mathrm{prob}^{2500}$$

Auf dem Computer ist die Zahl prob^{2500} entweder 0 oder unendlich, je nachdem, ob das prob kleiner oder grösser ist als 1 (probieren Sie das ruhig mal aus), also das ist numerisch nicht handhabbar. In Gegensatz dazu ist

$$\log L \ \sim \ \log\big[\,\mathrm{prob}^N\,\big] \ = \ N \log[\mathrm{prob}] \ = \ 2500\, \log[\mathrm{prob}]$$

und das ist numerisch unproblematisch.

Kehren wir jetzt zu unserem konkreten Beispiel zurück. Auch für die analytische Rechnung ist es günstig, den Logarithmus zu betrachten,

$$\begin{aligned}
\log L(\mu,\sigma) \ &= \ \log\bigg[\,(2\pi\sigma^2)^{-\frac{n}{2}}\ \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\Big\}\,dx_1\cdots dx_n\bigg] \\
&= \ -\frac{n}{2}\log[2\pi\sigma^2] \ - \ \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 \ + \ \log[dx_1\cdots dx_n] \\
&= \ -\,n\,\log\sigma \ - \ \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 \ + \ \mathrm{const}
\end{aligned} \tag{5}$$

wobei die Grösse

$$\mathrm{const} \ := \ -\frac{n}{2}\log[2\pi] \ + \ \log[dx_1\cdots dx_n]$$

eine von μ und σ unabhängige Zahl ist. Es ist also hinreichend, die Funktion

$$\log \tilde{L}(\mu,\sigma) \ := \ -\,n\,\log\sigma \ - \ \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 \tag{6}$$

zu betrachten. Wie oben bereits angekündigt, stellen wir also fest, dass die Intervallbreiten $dx_1, \cdots, dx_n$ für die eigentliche Rechnung irrelevant sind. Jetzt gilt das folgende

Theorem 7.1: Das $\log \tilde{L}(\mu, \sigma)$ aus Gleichung (6) wird maximiert durch

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i =: \mu_{\mathrm{ML}}(x_1, \cdots, x_n) \tag{7}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 =: \sigma^2_{\mathrm{ML}}(x_1, \cdots, x_n) \tag{8}$$

mit $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i = \mu_{\mathrm{ML}}(x_1, \cdots, x_n)$.

Beweis: Notwendige Bedingung für ein Maximum ist

$$\frac{\partial}{\partial \mu} \log \tilde{L}(\mu, \sigma) = 0 \tag{9}$$

$$\frac{\partial}{\partial \sigma} \log \tilde{L}(\mu, \sigma) = 0 \tag{10}$$

Nun ist

$$\frac{\partial}{\partial \mu} \log \tilde{L}(\mu, \sigma) = + \frac{2}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

also folgt aus Gleichung (9)

$$\sum_{i=1}^{n} (x_i - \mu) = \sum_{i=1}^{n} x_i - n\,\mu = 0$$

oder

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

Für die Ableitung nach σ erhalten wir

$$\frac{\partial}{\partial \sigma} \log \tilde{L}(\mu, \sigma) = - \frac{n}{\sigma} + \frac{2}{2\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

und Gleichung (10) liefert

$$\frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{n}{\sigma}$$

oder

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

Damit ist das Theorem bewiesen. ■

Ein wichtiger Begriff in der Statistik ist der der Erwartungstreue. Bei Schätzern möchte man immer gerne haben, dass sie erwartungstreu sind. Was heisst das jetzt genau? Der Maximum-Likelihood-Schätzer für den Mittelwert μ ist gegeben durch

$$\mu_{\mathrm{ML}}(x_1, \cdots, x_n) \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i \tag{11}$$

Wenn die x_i's jetzt nicht irgendwelche Daten sind, sondern wirklich durch Simulation generierte normalverteilte Zufallszahlen sind, dann sollte, wenn wir jetzt etwa 1 Million mal solche n Zufallszahlen $(x_1, \cdots, x_n)$ simulieren und dann jeweils die Grösse (11) berechnen (für ein festes n, etwa $n = 30$), dann sollte der Mittelwert dieser 1 Millionen μ_{ML} dann auch wirklich gegen das tatsächliche μ konvergieren. Das ist genau dann der Fall, wenn der Erwartungswert der μ_{ML}'s, gegeben durch

$$\mathsf{E}\big[\, \mu_{\mathrm{ML}}(x_1, \cdots, x_n) \,\big] \;=\; \int_{\mathbb{R}^n} \mu_{\mathrm{ML}}(x_1, \cdots, x_n) \, \prod_{i=1}^{n} \Big\{ \, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \; \frac{dx_i}{\sqrt{2\pi\sigma^2}} \, \Big\} \tag{12}$$

gleich dem tatsächlichen μ ist. Das führt dann also zu der folgenden

Definition 7.2: Die Maximum-Likelihood-Schätzer (7) und (8) aus dem Theorem 7.1 heissen erwartungstreu, wenn sie die folgende Eigenschaft haben:

$$\mathsf{E}\big[\, \mu_{\mathrm{ML}}(x_1, \cdots, x_n) \,\big] \;=\; \mu \tag{13}$$
$$\mathsf{E}\big[\, \sigma^2_{\mathrm{ML}}(x_1, \cdots, x_n) \,\big] \;=\; \sigma^2 \tag{14}$$

Dabei sind die Erwartungswerte gemäss Gleichung (12) zu berechnen. Es gilt nun das folgende

Theorem 7.3: Wir betrachten die Maximum-Likelihood-Schätzer (7) und (8) aus dem Theorem 7.1. Dann gilt:

a) Der Schätzer μ_{ML} ist erwartungstreu.

b) Der Schätzer σ^2_{ML} ist nur asymptotisch, im Limes $n \to \infty$, erwartungstreu. Genauer gilt:

$$\mathsf{E}\big[\, \sigma^2_{\mathrm{ML}}(x_1, \cdots, x_n) \,\big] \;=\; \frac{n-1}{n} \, \sigma^2 \tag{15}$$

und der modifizierte Schätzer

$$s^2 \;:=\; \frac{n}{n-1} \sigma^2_{\mathrm{ML}} \;=\; \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{16}$$

ist erwartungstreu.

Beweis: Wir kürzen ab:

$$p_{\mu,\sigma}(x) \;:=\; \tfrac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Das $p_{\mu,\sigma}(x)$ ist also die Dichte der Normalverteilung mit Mittelwert μ und Standardabweichung σ. Da das eine W'keitsdichte ist, gilt

$$\textstyle\int_{\mathbb{R}} p_{\mu,\sigma}(x)\, dx \;=\; 1 \tag{17}$$

Weiterhin haben wir die folgenden Integrale: Der Mittelwert einer $p_{\mu,\sigma}$-verteilten Zufallszahl ist μ meint:

$$\int_{\mathbb{R}} x\, p_{\mu,\sigma}(x)\, dx \;=\; \mu \tag{18}$$

Die Varianz einer $p_{\mu,\sigma}$-verteilten Zufallszahl ist σ^2 meint:

$$\int_{\mathbb{R}} (x-\mu)^2\, p_{\mu,\sigma}(x)\, dx \;=\; \sigma^2$$

oder

$$\begin{aligned}
\int_{\mathbb{R}} x^2\, p_{\mu,\sigma}(x)\, dx &= \int_{\mathbb{R}} (x-\mu+\mu)^2\, p_{\mu,\sigma}(x)\, dx \\
&= \int_{\mathbb{R}} \big\{\, (x-\mu)^2 \,+\, 2(x-\mu)\mu \,+\, \mu^2 \,\big\}\, p_{\mu,\sigma}(x)\, dx \\
&= \sigma^2 \,+\, 0 \,+\, \mu^2 \;=\; \sigma^2 \,+\, \mu^2
\end{aligned} \tag{19}$$

Damit bekommen wir:

$$\begin{aligned}
\mathsf{E}\big[\, \mu_{\mathrm{ML}}(x_1,\cdots,x_n) \,\big] &= \int_{\mathbb{R}^n} \mu_{\mathrm{ML}}(x_1,\cdots,x_n) \prod_{k=1}^{n} \Big\{\, e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \,\frac{dx_k}{\sqrt{2\pi\sigma^2}} \,\Big\} \\
&= \int_{\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} x_i \prod_{k=1}^{n} \Big\{\, e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \,\frac{dx_k}{\sqrt{2\pi\sigma^2}} \,\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}^n} x_i \prod_{k=1}^{n} \Big\{\, e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \,\frac{dx_k}{\sqrt{2\pi\sigma^2}} \,\Big\}
\end{aligned}$$

mit

$$\begin{aligned}
\int_{\mathbb{R}^n} x_i \prod_{k=1}^{n} \Big\{\, e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \,\frac{dx_k}{\sqrt{2\pi\sigma^2}} \,\Big\} &= \int_{\mathbb{R}} p_{\mu,\sigma}(x_1)\, dx_1 \;\cdots\; \int_{\mathbb{R}} x_i\, p_{\mu,\sigma}(x_i)\, dx_i \;\cdots\; \int_{\mathbb{R}} p_{\mu,\sigma}(x_n)\, dx_n \\
&\overset{(17,18)}{=} 1 \times \cdots \times 1 \times \mu \times 1 \times \cdots \times 1 \\
&= \mu
\end{aligned}$$

Also,

$$\begin{aligned}
\mathsf{E}\big[\, \mu_{\mathrm{ML}}(x_1,\cdots,x_n) \,\big] &= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}^n} x_i \prod_{k=1}^{n} \Big\{\, e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \,\frac{dx_k}{\sqrt{2\pi\sigma^2}} \,\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \;=\; \mu
\end{aligned}$$

und μ_{ML} ist erwartungstreu. Für die Berechnung von $\mathsf{E}[\sigma^2_{\mathrm{ML}}]$ machen wir zunächst ein paar Umformungen:

$$\begin{aligned}
\sigma^2_{\mathrm{ML}}(x_1,\cdots,x_n) &= \frac{1}{n}\sum_{i=1}^{n} (x_i-\bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \big\{\, x_i^2 \,-\, 2x_i\bar{x} \,+\, \bar{x}^2 \,\big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \,-\, 2\bar{x}\,\bar{x} \,+\, \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \,-\, \bar{x}^2
\end{aligned} \tag{20}$$

mit

$$\begin{aligned}
\bar{x}^2 &= \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i \right\}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i \, \frac{1}{n} \sum_{j=1}^{n} x_j = \frac{1}{n^2} \sum_{i,j=1}^{n} x_i \, x_j \\
&= \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i=j}}^{n} x_i \, x_j \; + \; \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} x_i \, x_j \\
&= \frac{1}{n^2} \sum_{i=1}^{n} x_i^2 \; + \; \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} x_i \, x_j \qquad (21)
\end{aligned}$$

Wir setzen (21) in (20) ein und bekommen

$$\begin{aligned}
\sigma^2_{\mathrm{ML}}(x_1, \cdots, x_n) &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 \; - \; \bar{x}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^2 \; - \; \frac{1}{n^2} \sum_{i=1}^{n} x_i^2 \; - \; \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} x_i \, x_j \\
&= \frac{n-1}{n^2} \sum_{i=1}^{n} x_i^2 \; - \; \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} x_i \, x_j \qquad (22)
\end{aligned}$$

Damit erhalten wir

$$\begin{aligned}
\mathsf{E}\big[\, \sigma^2_{\mathrm{ML}}(x_1, \cdots, x_n) \,\big] &= \int_{\mathbb{R}^n} \sigma_{\mathrm{ML}}(x_1, \cdots, x_n) \; \prod_{k=1}^{n} \left\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \; \frac{dx_k}{\sqrt{2\pi\sigma^2}} \right\} \\
&\overset{(22)}{=} \int_{\mathbb{R}^n} \left\{ \frac{n-1}{n^2} \sum_{i=1}^{n} x_i^2 \; - \; \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} x_i \, x_j \right\} \; \prod_{k=1}^{n} \left\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \; \frac{dx_k}{\sqrt{2\pi\sigma^2}} \right\} \\
&= \frac{n-1}{n^2} \sum_{i=1}^{n} \int_{\mathbb{R}^n} x_i^2 \; \prod_{k=1}^{n} \left\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \; \frac{dx_k}{\sqrt{2\pi\sigma^2}} \right\} \\
&\quad - \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \int_{\mathbb{R}^n} x_i \, x_j \; \prod_{k=1}^{n} \left\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \; \frac{dx_k}{\sqrt{2\pi\sigma^2}} \right\} \qquad (23)
\end{aligned}$$

mit den Integralen

$$\begin{aligned}
\int_{\mathbb{R}^n} x_i^2 \; \prod_{k=1}^{n} \left\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \; \frac{dx_k}{\sqrt{2\pi\sigma^2}} \right\} &= \int_{\mathbb{R}} p_{\mu,\sigma}(x_1) \, dx_1 \; \cdots \; \int_{\mathbb{R}} x_i^2 \; p_{\mu,\sigma}(x_i) \, dx_i \; \cdots \; \int_{\mathbb{R}} p_{\mu,\sigma}(x_n) \, dx_n \\
&\overset{(17,19)}{=} 1 \times \cdots \times 1 \times (\sigma^2 + \mu^2) \times 1 \times \cdots \times 1 \\
&= \sigma^2 + \mu^2
\end{aligned}$$

und für $i \neq j$

$$
\begin{aligned}
&\int_{\mathbb{R}^n} x_i\, x_j \prod_{k=1}^{n} \Big\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \, \frac{dx_k}{\sqrt{2\pi\sigma^2}} \Big\} \\
&\quad = \quad \textstyle\int_{\mathbb{R}} p_{\mu,\sigma}(x_1)\, dx_1 \;\cdots\; \int_{\mathbb{R}} x_i\; p_{\mu,\sigma}(x_i)\, dx_i \;\cdots\; \int_{\mathbb{R}} x_j\; p_{\mu,\sigma}(x_j)\, dx_j \;\cdots\; \int_{\mathbb{R}} p_{\mu,\sigma}(x_n)\, dx_n \\
&\quad \overset{(17,18)}{=} \quad 1 \times \cdots \times \mu \times \cdots \times \mu \times \cdots \times 1 \\
&\quad = \quad 1^{n-2} \;\times\; \mu^2 \;=\; \mu^2
\end{aligned}
$$

Das setzen wir in (23) ein und bekommen

$$
\begin{aligned}
\mathsf{E}\big[\, \sigma^2_{\mathrm{ML}}(x_1, \cdots, x_n) \,\big] \;&=\; \frac{n-1}{n^2} \sum_{i=1}^{n} \int_{\mathbb{R}^n} x_i^2 \prod_{k=1}^{n} \Big\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \, \frac{dx_k}{\sqrt{2\pi\sigma^2}} \Big\} \\
&\qquad - \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \int_{\mathbb{R}^n} x_i\, x_j \prod_{k=1}^{n} \Big\{ e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \, \frac{dx_k}{\sqrt{2\pi\sigma^2}} \Big\} \\
&=\; \frac{n-1}{n^2} \sum_{i=1}^{n} (\sigma^2 + \mu^2) \;-\; \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \mu^2 \\
&=\; \frac{n-1}{n} \, (\sigma^2 + \mu^2) \;-\; \frac{n^2-n}{n^2} \, \mu^2 \\
&=\; \frac{n-1}{n} \, \sigma^2 \;+\; \frac{n-1}{n} \, \mu^2 \;-\; \frac{n-1}{n} \, \mu^2 \\
&=\; \frac{n-1}{n} \, \sigma^2
\end{aligned}
$$

Also ist $\mathsf{E}[\sigma^2_{\mathrm{ML}}]$ nur im Limes $n \to \infty$ erwartungstreu und das Theorem ist bewiesen. ■