Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

Heinrich-Heine-Universität Düsseldorf

8. Dezember 2010

Teil V

Schließende Statistik

1. Parameterschätzung
   - Erwartungstreue und Konsistenz
   - Maximum-Likelihood Schätzer für diskrete Zufallsvariable
   - Rückfangexperiment
   - Maximum-Likelihood Schätzer für stetige Zufallsvariable
   - Dichte der gemeinsamen Verteilung
   - ML-Schätzer für eine Exponentialverteilung

Parameterschätzung

- Aus einer Population wird eine Stichprobe entnommen
- zu dieser Stichprobe werden Daten erhoben
- mit diesen Daten wird ein Parameter geschätzt

- Häufig misst man den interessierenden Parameter direkt
- Statistische Sicht
  - Zufallsvariable X_j gibt des Ergebnis des j-ten Experiments an
  - Für ein konkretes Experiment existiert ein Zufallsanteil, der zu einer Abweichung vom wahren Wert führt
  - Im Mittel kommt der wahre Wert heraus (sonst hat der Versuchsaufbau einen "systematischen Fehler")
- Das bedeutet: Wir schätzen den Erwartungswert von X_1

Schätzung des Erwartungswerts

Seien $X_1, \ldots, X_n$ reelle, unabhängige Zufallsvariable, die alle dieselbe Verteilung haben. Sei $\mu = E(X_1) = \cdots = E(X_n)$. Dann ist

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

das arithmetische Mittel der Messwerte

$\overline{X}$ ist ein Schätzer für μ

Anforderungen an Schätzer

Ein Schätzer sollte idealerweise zwei Eigenschaften haben:

- Erwartungstreue: Kein systematischer Fehler
- Konsistenz: Mehr Versuche führen zu genaueren Ergebnissen

Erwartungstreue

- Der Schätzer ist selber eine Zufallsvariable. Er heißt *erwartungstreu*, wenn sein Erwartungswert gleich dem zu schätzenden Parameter ist
- Das arithmetische Mittel ist ein erwartungstreuer Schätzer für den Erwartungswert
- Das heißt $E(\overline{X}) = \mu$

Konsistenz

- Ein Schätzer heißt *konsistent*, wenn für große Stichprobenumfänge die Differenz zwischen dem geschätzten und dem wahren Parameter beliebig klein wird
- Ein Schätzer, der nicht konsistent ist, taugt nicht viel. Dagegen ist Erwartungstreue eine Eigenschaft, auf die man notfalls verzichten kann
- Das arithmetische Mittel ist ein konsistenter Schätzer für den Erwartungswert
- Diese Aussage ist eine Variante des Gesetzes von der großen Zahl

Schätzung der Varianz

- Die empirische Varianz ist ein erwartungstreuer, konsistenter Schätzer für die Varianz.
- *Erinnerung:* Die empirische Varianz ist definiert als

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

- Wenn wir statt des Nenners $n-1$ den Nenner n benutzt hätten, dann wäre der Schätzer zwar noch konsistent, aber nicht mehr erwartungstreu.

Maximum-Likelihood Schätzer

- "Likelihood" = Wahrscheinlichkeit
- Ein Parameter θ soll geschätzt werden. Ein Datensatz liege vor
- Der *Maximum-Likelihood Schätzwert* ist derjenige Wert von θ, für den der vorliegende Datensatz die größte Wahrscheinlichkeit besitzt.

ML-Schätzer für diskrete Zufallsvariable

- Es sei Θ eine Menge von Parameterwerten. Zu jedem Parameterwert $\theta \in \Theta$ gebe es eine diskrete Verteilung P_θ
- Von der Zufallsvariablen X sei bekannt, dass ihre Verteilung gleich einem der P_θ ist. Dieses θ soll geschätzt werden
- Für jede mögliche Realisierung x von X bezeichnet man die Abbildung

$$L_x(\theta) = P_\theta(X = x)$$

  als *Likelihood-Funktion* zu x
- Die Likelihood-Funktion ist die Wahrscheinlichkeit des beobachteten Ergebnisses für den Parameterwert θ

Theorie des ML-Schätzers, Fortsetzung

- Der Parameterwert $T(x)$, für den gilt

$$L_x(T(x)) = \max\{L_x(\theta) \mid \theta \in \Theta\}$$

  heißt *Maximum-Likelihood Schätzwert* von θ zur Realisierung x
- Wenn es mehrere solche Parameterwerte $T(x)$ gibt, ist jeder von ihnen ein Maximum-Likelihood Schätzer
- Also

$$P_{T(x)}(X = x) = \max_{\theta \in \Theta} P_\theta(X = x)$$

- Man bezeichnet den Schätzwert oft mit $\hat{\theta}$

ML-Schätzer für Erfolgswahrscheinlichkeit

- Ein ja/nein-Experiment wird n-mal wiederholt
- Der zu schätzende Parameter θ ist die Erfolgswahrscheinlichkeit im Einzelfall
- Es seien k Erfolge beobachtet worden
- Dann ist die Likelihood-Funktion gleich

$$L_k(\theta) = \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k}$$

- Wir diskutieren ihren Logarithmus

$$g(\theta) = \ln \binom{n}{k} + k \cdot \ln(\theta) + (n-k) \cdot \ln(1-\theta)$$

- Also

$$g'(\theta) = \frac{k}{\theta} - \frac{n-k}{1-\theta}$$

ML-Schätzer für Erfolgsw'keit, Fortsetzung

- Suche Nullstellen von

$$g'(\theta) = \frac{k}{\theta} - \frac{n-k}{1-\theta}$$

$$= \frac{k \cdot (1-\theta) - (n-k) \cdot \theta}{\theta \cdot (1-\theta)} = \frac{k - n\theta}{\theta \cdot (1-\theta)}$$

- Die Nullstelle ist bei $\theta = \frac{k}{n}$
- Der ML-Schätzer für die Erfolgswahrscheinlichkeit ist

$$\hat{p} = \frac{k}{n}$$

Beispiel: Wartezeiten

- Wartezeit in getaktetem Modell soll geschätzt werden
- Beabachtung: einmal Erfolg im ersten Versuch, einmal im zweiten, fünfmal in einem späteren als dem zweiten
- Zufallsvariable verteilt gemäß geometrischer Verteilung G_p
- p ist zu schätzen

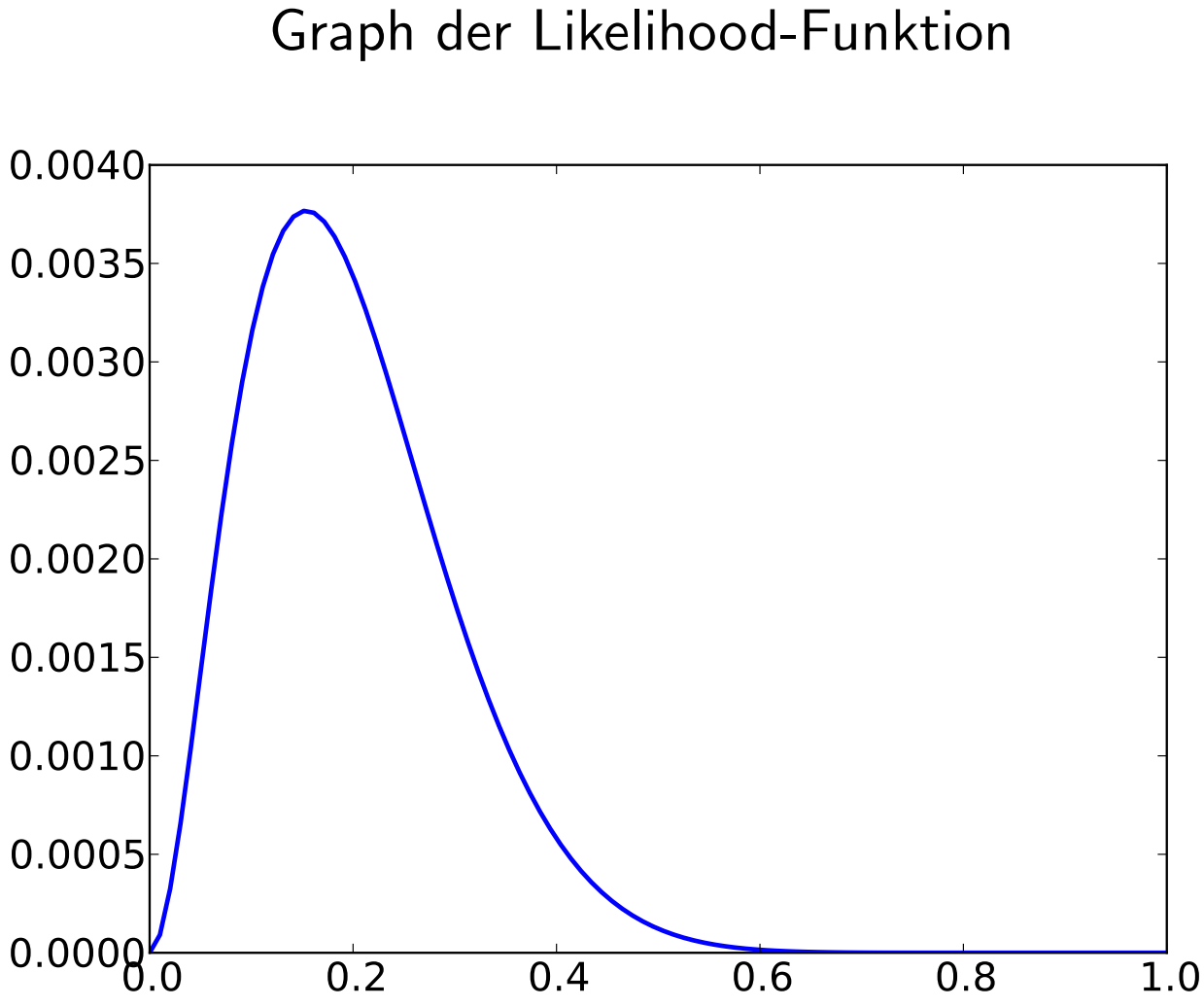
$$
\begin{aligned}
P(X = 1) &= p \\
P(X = 2) &= p \cdot (1 - p) \\
P(X \geq 3) &= 1 - (p + p \cdot (1 - p)) \\
&= 1 - 2p + p^2
\end{aligned}
$$

Wartezeiten, Fortsetzung

$$\begin{aligned} L(p) &= P(X_1 = 1, X_2 = 2, X_3 \geq 3, \ldots, X_7 \geq 3) \\ &= p \cdot p \cdot (1-p) \cdot \left(1 - 2p + p^2\right)^5 \end{aligned}$$

- Analytische Behandlung ist schwierig, der Graph zeigt ein Maximum bei 0.15
- Also $\hat{p} = 0.15$

Wartezeiten, Graph



Rückfangexperiment

- In einem Gewächshaus befinden sich mehrere Hundert Mücken. Wir wollen ihre Anzahl schätzen
- Dazu fangen wir 160 Tiere und markieren sie
- Wir lassen die markierten Tiere frei und warten, bis sie sich mit den anderen vermischt haben
- Wir fangen noch einmal 100 Tiere
- Aus der Zahl der markierten unter den gefangenen Tieren wollen wir die Anzahl aller Mücken schätzen
- Wenn N die Anzahl aller Mücken ist, dann ist

$$p = \frac{160}{N}$$

  die Wahrscheinlichkeit, dass eine einzelne Mücke markiert ist
- Die Anzahl der markierten Mücken unter denen des zweiten Fangs ist $B_{100,\,p}$-verteilt

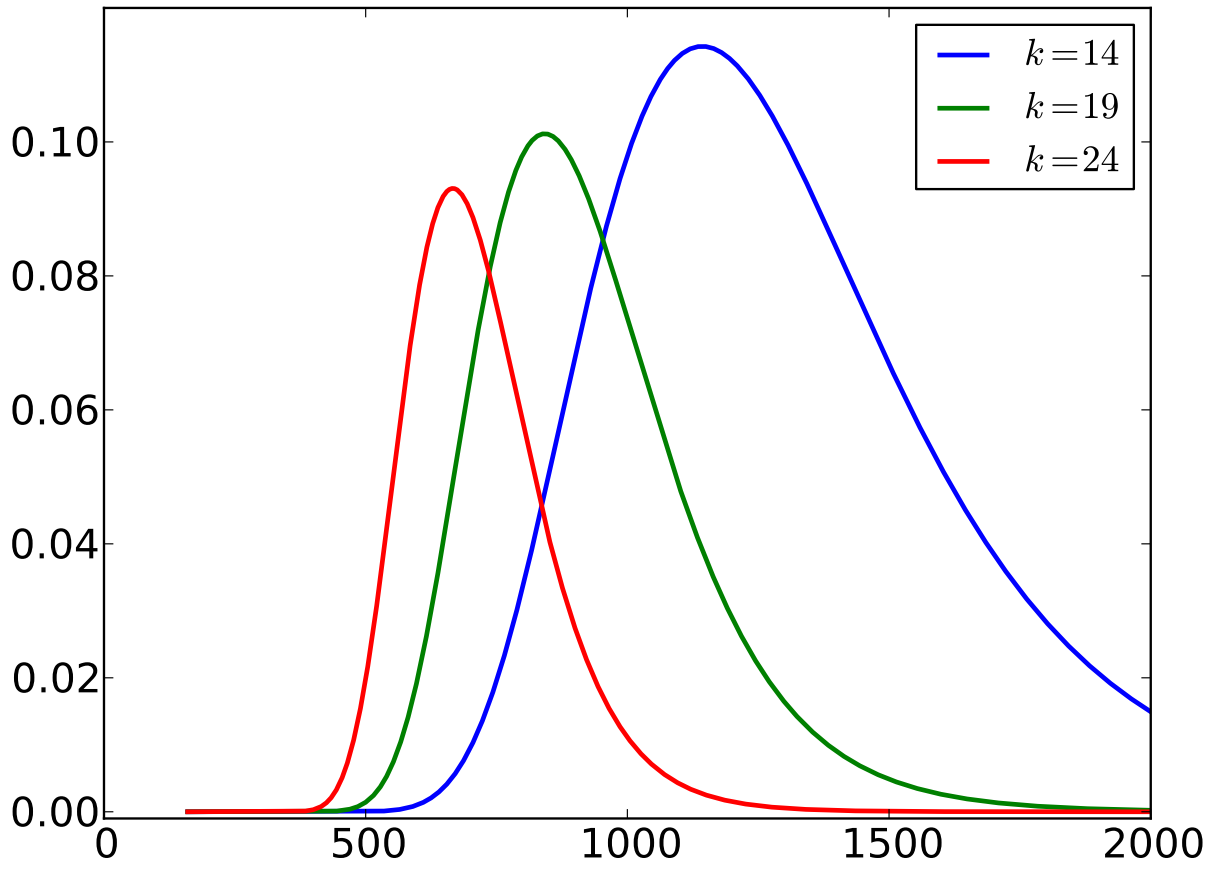
Rückfangexperiment, Fortsetzung

- Der Parameter $\theta =$ "Anzahl der Mücken" soll geschätzt werden
- $\Theta = \{160, 161, 162, \dots\}$
- Die Zufallsvariable X zählt die markierten Tiere im zweiten Fang
- X ist $B_{100,\,p}$ verteilt für $p = \frac{160}{N}$. Dabei ist N der zu schätzende Parameter
- k ist die Realisierung, also die Anzahl der markierten Mücken im zweiten Fang
- Die Likelihood-Funktion ist

$$\begin{aligned} L_k(N) &= B_{100,\,160/N}(k) \\ &= \binom{100}{k} \cdot \left(\frac{160}{N}\right)^k \cdot \left(1 - \frac{160}{N}\right)^{100-k} \end{aligned}$$

Rückfangexperiment, Graph

Likelihood-Funktion für verschiedene Werte von k



Rückfangexperiment, Fortsetzung

- Zweistufiges Vorgehen
  - Schätze zuerst für den Rückfang die Erfolgswahrscheinlichkeit p dafür, eine markierte Mücke zu fangen
  - Rechne das dann um in den Schätzer für N
- Für den Rückfang $n = 100$

$$\hat{p} = \frac{k}{100}$$

- Für den Rückfang $\hat{p} = \frac{160}{\hat{N}}$

$$\hat{N} = \frac{160}{\hat{p}} = \frac{16\,000}{k}$$

Rückfangexperiment, konkrete Zahlen

Für die folgenden Werte von k erhält man die folgenden Schätzwerte

Realisierung k	Schätzwert $\hat{N}$
14	1143
19	842
24	667

Aus zweistelligen Messdaten kann man aber nur zwei gültige Stellen bekommen

Realisierung k	Schätzwert $\hat{N}$
14	1100
19	840
24	670

Rückfangexperiment, Stabilität

- Wie ändert sich $\widehat{N}$, wenn k sich um 1 ändert?
- Die Antwort auf diese Frage gibt die Ableitung von

$$h(x) = \frac{16\,000}{x}$$

$$h'(x) = -\frac{16\,000}{x^2}$$

- Also z. B.

$$h'(19) = -44.32$$

- Eine gefangene Mücke mehr oder weniger ändert den Schätzwert um ungefähr 45 Mücken

ML-Schätzer für stetige Zufallsvariable

- Idee: Ersetze die Wahrscheinlichkeit $P(X = x)$ durch die Dichte $f(x)$
- Es sei Θ eine Menge von Parameterwerten. Zu jedem $\theta \in \Theta$ gebe es eine stetige Verteilung P_θ mit Dichte f_θ
- Von der Zufallsvariablen X sei bekannt, dass ihre Verteilung gleich einem der P_θ ist. Dieses θ soll geschätzt werden
- Für jede mögliche Realisierung x von X bezeichnet man die Abbildung

$$L_x(\theta) = f_\theta(x)$$

als *Likelihood-Funktion* zu x

ML-Schätzer für stetige Zufallsvariable, Fortsetzung

- Der Parameter $T(x)$, für den gilt

$$L_x(T(x)) = \max\{L_x(\theta) \mid \theta \in \Theta\}$$

  heißt *Maximum-Likelihood Schätzwert* von θ zur Realisierung x
- Wenn es mehrere solche Parameterwerte $T(x)$ gibt, ist jeder von ihnen ein Maximum-Likelihood Schätzer
- Also

$$P_{T(x)}(X = x) = \max_{\theta \in \Theta} f_\theta(x)$$

- Man bezeichnet den Schätzwert wieder mit $\hat{\theta}$

ML-Schätzer für stetige Zufallsvariable, Beispiel

- Die Zufallsvariable X sei $N(\mu, 1)$-verteilt
- μ soll geschätzt werden, ausgehend von Realisierung x
- Die Dichte ist

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

- Für gegebene Realisierung x suche μ mit größtem Wert
- Der Bruch ist konstant. Der zweite Faktor wird logarithmiert

$$g(\mu) = -\frac{(x-\mu)^2}{2}$$

- Diese Zahl wird maximal für $\mu = x$
- Der ML-Schätzer ist $\hat{\mu} = x$

Unabhängige Zufallsvariable

- Gegeben: unabhängige Zufallsvariable $X_1, X_2, \ldots, X_n$
- Dichte von X_j ist f_j
- Gesucht: Dichte der gemeinsamen Verteilung
- Im diskreten Fall entspricht dies

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n) \end{aligned}$$

- Die *Dichte der gemeinsamen Verteilung* ist

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n)$$

ML-Schätzer für eine Exponentialverteilung

- Die Lebensdauer von Glühbirnen ist näherungsweise exponentialverteilt, besitzt also eine Dichte

$$f_\lambda(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

  Der Parameter λ ist zu schätzen.
- Dazu werden die Lebensdauern von n Glühbirnen bestimmt.

ML-Schätzer zur Exponentialverteilung, Fortsetzung

- Stochastisches Modell: $X_1, \ldots, X_n$ unabhängige Zufallsvariable mit Dichte f_λ. Die Dichte der gemeinsamen Verteilung muss verwendet werden
- Dann ist die Likelihood-Funktion zur Realisierung $(x_1, \ldots, x_n)$ gegeben durch

$$\begin{aligned} g(\lambda) &= f_\lambda(x_1) \cdot f_\lambda(x_2) \cdots f_\lambda(x_n) \\ &= \lambda^n \cdot e^{-\lambda \cdot x_1} \cdots e^{-\lambda \cdot x_n} = \lambda^n \cdot \exp(-\lambda \cdot (x_1 + \cdots + x_n)) \end{aligned}$$

- Es ist zweckmäßig, den Logarithmus von g zu betrachten

$$h(\lambda) = \ln(g(\lambda)) = n \cdot \ln(\lambda) - \lambda \cdot (x_1 + \cdots + x_n)$$

- h hat dieselben Maximalstellen wie g, lässt sich aber leichter nach λ differenzieren.

ML-Schätzer zur Exponentialverteilung, Fortsetzung

- $h(\lambda) = n \cdot \ln(\lambda) - \lambda \cdot (x_1 + \cdots + x_n)$, also

$$h'(\lambda) = \frac{n}{\lambda} - (x_1 + \cdots + x_n)$$

- Einzige kritische Stelle

$$\lambda_0 = \frac{n}{x_1 + \cdots + x_n}$$

- Dort muss das Maximum liegen.
- Begründung: Das ist die einzige kritische Stelle, und die Randwerte sind

$$\lim_{\lambda \to 0} h(\lambda) = -\infty \quad \text{und} \quad \lim_{\lambda \to \infty} h(\lambda) = -\infty$$

ML-Schätzer zur Exponentialverteilung, Zahlenbeispiel

Die folgenden Brenndauern sind gemessen worden (in Stunden)

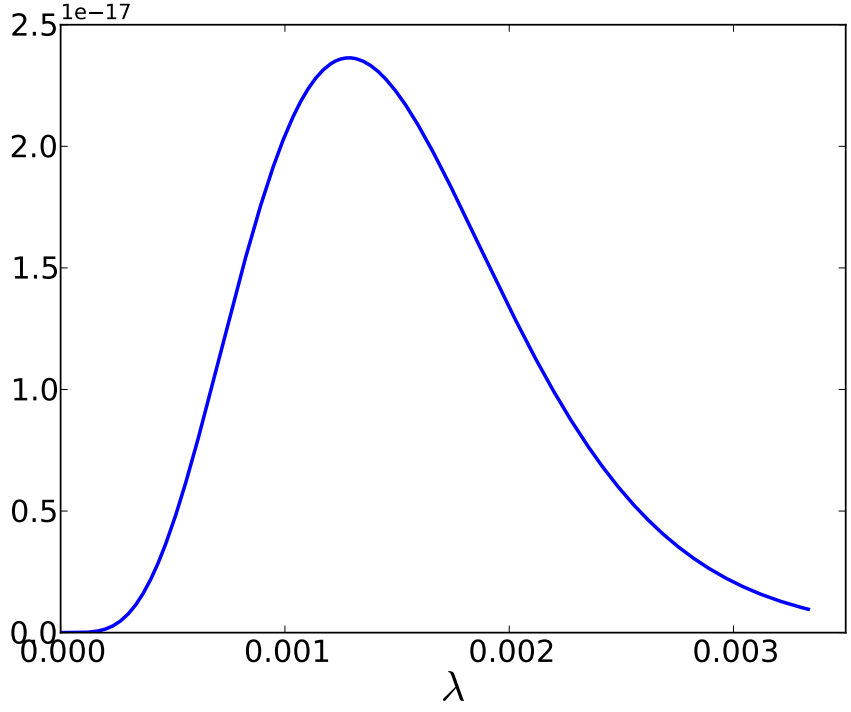
$$x_1 = 740, \quad x_2 = 800, \quad x_3 = 760, \quad x_4 = 600, \quad x_5 = 990$$

Dann ist der Maximum-Likelihood Schätzer für λ gleich

$$\hat{\lambda} = \frac{5}{740 + 800 + 760 + 600 + 990} = 0.001285$$

Die mittlere Brenndauer beträgt $1/0.001285 = 778$ Stunden.

Likelihood-Funktion im Glühlampenbeispiel



Graph von

$$h(\lambda) = \lambda^5 e^{-\lambda(740+800+760+600+990)}$$