Introduction to Maximum Likelihood Inference

Last modified: 2026-09-29 00:05:17 (PDT)

Overview of maximum likelihood estimation

These notes are derived primarily from (Dobson and Barnett 2018, chaps. 1–5), with some material from (McLachlan and Krishnan 2007) and (Casella and Berger 2002).

The likelihood function

Definition 0.1 (Likelihood of a single observation)**.** Let X be a random variable, with observed value x, and let $\mathrm{p}_\Theta(X = x)$ be a probability model for the distribution of X, with parameter vector Θ. The **likelihood** of the parameter value θ, for model $\mathrm{p}_\Theta(X = x)$ and data $X = x$, is the probability of the event $X = x$ when $\Theta = \theta$:

$$\mathcal{L}(\theta) \stackrel{\text{def}}{=} \mathrm{p}_\theta(X = x)$$

For a continuous random variable, the likelihood is the probability density of X at x instead.

Example 0.1 (Likelihood of one Bernoulli observation)**.** If $X \sim \mathrm{Ber}(\pi)$, then $\mathrm{p}_\pi(X = x) = \pi^x(1-\pi)^{1-x}$ for $x \in \{0, 1\}$. If we observe $x = 1$, the likelihood is $\mathcal{L}(\pi) = \pi$: for example, $\mathcal{L}(0.2) = 0.2$ and $\mathcal{L}(0.7) = 0.7$, so the observation $x = 1$ is more likely under $\pi = 0.7$ than under $\pi = 0.2$.

Definition 0.2 (Likelihood of a dataset)**.** Let $\tilde{x} \stackrel{\text{def}}{=} x_1, \ldots, x_n$ be a dataset with corresponding random vector $\tilde{X}$, and let $\mathrm{p}_\Theta(\tilde{X} = \tilde{x})$ be a probability model for the distribution of $\tilde{X}$, with unknown parameter vector Θ. The **likelihood** of the parameter value θ, for model p_Θ and data $\tilde{X} = \tilde{x}$, is the *joint probability* (or joint density) of $\tilde{X} = \tilde{x}$ when $\Theta = \theta$:

$$\begin{aligned}\mathcal{L}(\theta) &\overset{\text{def}}{=} \mathrm{p}(\tilde{X} = \tilde{x} \mid \Theta = \theta) \\ &= \mathrm{p}(X_1 = x_1, \dots, X_n = x_n \mid \Theta = \theta)\end{aligned}$$

Notation for the likelihood function

Sources write the likelihood function in several ways, all meaning the same function:

- $\mathcal{L}(\theta)$;
- $\mathcal{L}(\tilde{x}; \theta)$ or $\mathcal{L}(\theta; \tilde{x})$;
- $\mathcal{L}_{\tilde{x}}(\theta)$ or $\mathcal{L}_{\theta}(\tilde{x})$;
- $\mathcal{L}(\tilde{x} \mid \theta)$.

These notes mostly write $\mathcal{L}(\theta)$, leaving the data implicit, to emphasize that the likelihood is a function of the parameters, with the data held fixed at their observed values.

Theorem 0.1 (Likelihood of an independent sample)**.** *For mutually independent data* $X_1, \dots, X_n$*:*

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathrm{p}(X_i = x_i \mid \theta) \tag{1}$$

Proof

Proof.

$$\begin{aligned}\mathcal{L}(\theta) &\overset{\text{def}}{=} \mathrm{p}(X_1 = x_1, \dots, X_n = x_n \mid \theta) && \text{(definition of likelihood)} \\ &= \prod_{i=1}^{n} \mathrm{p}(X_i = x_i \mid \theta) && \text{(definition of mutual independence)}\end{aligned}$$

□

Definition 0.3 (Likelihood components)**.** For a dataset $\tilde{x}$ of mutually independent observations, the **likelihood component** (or **likelihood factor**) of observation $X_i = x_i$ is the likelihood of that observation alone:

$$\mathcal{L}_i(\theta) \overset{\text{def}}{=} \mathrm{p}(X_i = x_i \mid \theta)$$

Theorem 0.2 (Dataset likelihood as a product of observation likelihoods)**.** *For mutually independent data* $\tilde{x} \overset{def}{=} x_1, \dots, x_n$*, the likelihood of the dataset is the product of the*

observations' likelihood components:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathcal{L}_i(\theta)$$

Proof

Proof. By Theorem 0.1, $\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathrm{p}(X_i = x_i \mid \theta)$, and by Definition 0.3, each factor $\mathrm{p}(X_i = x_i \mid \theta)$ is $\mathcal{L}_i(\theta)$. □

Exercise 0.1 (Likelihood of binary outcomes with one event probability)**.** A binary outcome Y with event probability π has

$$\begin{aligned}
\Pr(Y = 1) &= \pi \\
\Pr(Y = 0) &= 1 - \pi \\
\Pr(Y = y) &= \pi^y (1 - \pi)^{1-y}, \quad y \in \{0, 1\}.
\end{aligned}$$

Let $\tilde{y} \overset{\text{def}}{=} (y_1, \ldots, y_n)$ be a dataset of mutually independent binary outcomes, all with the same event probability π: $Y_i \sim_{\perp\!\!\!\perp} \mathrm{Ber}(\pi)$. Write the likelihood of $\tilde{y}$.

Solution

Solution 0.1.

$$\begin{aligned}
\mathcal{L}(\pi; \tilde{y}) &= \prod_{i=1}^{n} \mathcal{L}_i(\pi) && \text{(product of likelihood components)} \\
&= \prod_{i=1}^{n} \Pr(Y_i = y_i) && \text{(definition of likelihood components)} \\
&= \prod_{i=1}^{n} \pi^{y_i} (1 - \pi)^{1-y_i} && \text{(Bernoulli PMF)} \\
&= \pi^{\sum_{i=1}^{n} y_i} (1 - \pi)^{\sum_{i=1}^{n} (1-y_i)} && \text{(product of powers of a common base)} \\
&= \pi^{\sum_{i=1}^{n} y_i} (1 - \pi)^{n - \sum_{i=1}^{n} y_i} && \left(\textstyle\sum_{i=1}^{n} 1 = n\right)
\end{aligned}$$

The maximum likelihood estimate

Definition 0.4 (Maximum likelihood estimate)**.** The **maximum likelihood estimate** (MLE) of a parameter vector Θ, written $\hat{\theta}_{\text{ML}}$, is the value of Θ that maximizes the likelihood:

$$\hat{\theta}_{\text{ML}} \stackrel{\text{def}}{=} \arg\max_{\Theta} \mathcal{L}(\Theta) \tag{2}$$

Example 0.2 (MLE for one Bernoulli observation)**.** In Example 0.1, the likelihood of the observation $x = 1$ is $\mathcal{L}(\pi) = \pi$ for $\pi \in [0, 1]$. This function is increasing, so it is maximized at the upper edge of the parameter space: $\hat{\pi}_{\text{ML}} = 1$.

Finding the maximum of a function

Definition 0.5 (Critical point)**.** A **critical point** of a differentiable function f is an input value x_0 where $f'(x_0) = 0$; for a function of a vector, a point where the gradient is the zero vector.

From calculus: if $f(x)$ is differentiable, its maximum over an interval of input values can occur only at an endpoint of the interval or at a critical point. At a critical point x_0, $f''(x_0) < 0$ is sufficient for x_0 to be a local maximum, but not necessary: $f(x) = -x^4$ has a maximum at $x_0 = 0$, where $f''(0) = 0$. For a function of a vector, a negative definite Hessian matrix at a critical point is sufficient for a local maximum.

Directly maximizing the likelihood function for independent data

To find the maximizer of the likelihood function, we solve $\mathcal{L}'(\theta) = 0$ for θ. For mutually independent data, Equation 1 gives:

$$\begin{aligned} \mathcal{L}'(\theta) &= \frac{\partial}{\partial\theta}\mathcal{L}(\theta) \\ &= \frac{\partial}{\partial\theta}\prod_{i=1}^{n} \mathrm{p}(X_i = x_i \mid \theta) \end{aligned} \tag{3}$$

Equation 3 is the derivative of a product of n factors, which takes $n - 1$ applications of the product rule and produces n terms. The log-likelihood avoids this work.

The log-likelihood function

Definition 0.6 (Log-likelihood)**.** The **log-likelihood** of parameter value θ, for model $\mathrm{p}_\Theta(\tilde{X})$ and data $\tilde{X} = \tilde{x}$, is the natural logarithm of the likelihood:

$$\ell \stackrel{\text{def}}{=} \log\{\mathcal{L}(\tilde{x}|\theta)\} \tag{4}$$

Example 0.3 (Log-likelihood of one Bernoulli observation)**.** In Example 0.1, $\mathcal{L}(\pi) = \pi$, so $\ell(\pi) = \log \pi$; for example, $\ell(0.7) = \log 0.7 \approx -0.357$.

Theorem 0.3 (Maximize the log-likelihood instead of the likelihood)**.** *If $\mathcal{L}(\theta) > 0$ for every θ, the likelihood and log-likelihood have the same maximizers:*

$$\arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta \ell(\theta)$$

Proof

Proof. The natural logarithm is strictly increasing on $(0, \infty)$, so for any θ_1 and θ_2, $\mathcal{L}(\theta_1) \geq \mathcal{L}(\theta_2)$ if and only if $\log \mathcal{L}(\theta_1) \geq \log \mathcal{L}(\theta_2)$, that is, $\ell(\theta_1) \geq \ell(\theta_2)$. So θ^* satisfies $\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta)$ for every θ if and only if it satisfies $\ell(\theta^*) \geq \ell(\theta)$ for every θ. □

Theorem 0.4 (Log-likelihood of an independent sample)**.** *For mutually independent data $X_1, \ldots, X_n$:*

$$\ell(\theta) = \sum_{i=1}^{n} \log \mathrm{p}(X_i = x_i \mid \theta) \tag{5}$$

If the X_i also share a common distribution $\mathrm{p}(X = x \mid \theta)$, each term is $\log \mathrm{p}(X = x_i \mid \theta)$.

Proof

Proof.

$$\begin{aligned}
\ell(\theta) &\stackrel{\text{def}}{=} \log \mathcal{L}(\theta) && \text{(definition of log-likelihood)} \\
&= \log \prod_{i=1}^{n} \mathrm{p}(X_i = x_i \mid \theta) && \text{(likelihood of an independent sample)} \\
&= \sum_{i=1}^{n} \log \mathrm{p}(X_i = x_i \mid \theta) && \text{(log of a product is a sum of logs)}
\end{aligned}$$

With a common distribution, $\mathrm{p}(X_i = x_i \mid \theta) = \mathrm{p}(X = x_i \mid \theta)$. □

Theorem 0.5 (Derivative of the log-likelihood function for iid data)**.** *For* iid *data:*

$$\ell'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log \mathrm{p}(X = x_i \mid \theta) \tag{6}$$

Proof

Proof.

$$\begin{aligned}
\ell'(\theta) &= \frac{\partial}{\partial \theta} \ell(\theta) && \text{(notation for the derivative)} \\
&= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log \mathrm{p}(X = x_i \mid \theta) && \text{(log-likelihood of an iid sample)} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log \mathrm{p}(X = x_i \mid \theta) && \text{(derivative of a sum is the sum of derivatives)}
\end{aligned}$$

Unlike Equation 3, each term involves only one observation, so no product rule is needed. □

Exercise 0.2 (Log-likelihood of binary outcomes with one event probability)**.** Write the log-likelihood of $\tilde{y}$ from Exercise 0.1.

Solution

Solution 0.2. Starting from the likelihood in Solution 0.1:

$$\begin{aligned}
\ell(\pi; \tilde{y}) &= \log\left\{\pi^{\sum_{i=1}^{n} y_i} (1-\pi)^{n - \sum_{i=1}^{n} y_i}\right\} && \text{(log of the likelihood)} \\
&= \left(\sum_{i=1}^{n} y_i\right) \log\{\pi\} + \left(n - \sum_{i=1}^{n} y_i\right) \log\{1-\pi\} && \text{(log of a product; log of a power)} \\
&= \left(\sum_{i=1}^{n} y_i\right) (\log\{\pi\} - \log\{1-\pi\}) + n \log\{1-\pi\} && \text{(collect the terms in } \textstyle\sum_{i=1}^{n} y_i\text{)} \\
&= \left(\sum_{i=1}^{n} y_i\right) \log\left\{\frac{\pi}{1-\pi}\right\} + n \log\{1-\pi\} && \text{(log of a quotient)} \\
&= \left(\sum_{i=1}^{n} y_i\right) \mathrm{logit}(\pi) + n \log\{1-\pi\} && \text{(definition of logit)}
\end{aligned}$$

The score function

Definition 0.7 (Score function)**.** The **score function** of a statistical model $\mathrm{p}(\tilde{X} = \tilde{x})$ is the gradient (vector of first derivatives) of the model's log-likelihood with respect to the parameters:

$$\ell' \overset{\text{def}}{=} \frac{\partial}{\partial \theta} \ell(\tilde{x}|\theta) \tag{7}$$

We often omit the arguments $\tilde{x}$ and θ, writing $\ell' \overset{\text{def}}{=} \ell'(\tilde{x} \mid \theta) \overset{\text{def}}{=} \ell'(\theta)$. Some sources write U or S for the score function instead of ℓ'; for example, Dobson and Barnett (2018) write U. These notes use ℓ', which keeps U and S free for other uses and needs no extra symbol to memorize.

Exercise 0.3 (Score function of a Bernoulli variable)**.** Derive the score function for a single Bernoulli random variable X. In other words, differentiate the log-likelihood of a single Bernoulli random variable X with respect to the event probability parameter π. Simplify as much as possible.

Solution

Solution 0.3. With $n = 1$, Solution 0.2 gives $\ell = x \log\{\pi\} + (1 - x) \log\{1 - \pi\}$, so:

$$\begin{aligned}
\ell' &\overset{\text{def}}{=} \frac{\partial}{\partial \pi} \ell && \text{(definition of the score)} \\
&= \frac{\partial}{\partial \pi} (x \log\{\pi\} + (1 - x) \log\{1 - \pi\}) && \text{(Bernoulli log-likelihood)} \\
&= x \frac{\partial}{\partial \pi} \log\{\pi\} + (1 - x) \frac{\partial}{\partial \pi} \log\{1 - \pi\} && \text{(linearity of differentiation)} \\
&= x \frac{1}{\pi} - (1 - x) \frac{1}{1 - \pi} && \text{(derivative of log; chain rule)} \\
&= \frac{x(1 - \pi) - (1 - x)\pi}{\pi(1 - \pi)} && \text{(common denominator)} \\
&= \frac{x - x\pi - \pi + x\pi}{\pi(1 - \pi)} && \text{(expand the numerator)} \\
&= \frac{x - \pi}{\pi(1 - \pi)} && \text{(cancel } x\pi\text{)} \\
&= \frac{x - \mathrm{E}[X]}{\mathrm{Var}(X)} && (\mathrm{E}[X] = \pi \text{ and } \mathrm{Var}(X) = \pi(1 - \pi))
\end{aligned}$$

Exercise 0.4 (Score function of a Poisson variable)**.** Derive the score function for a single Poisson random variable X, with respect to its mean parameter λ.

Solution

Solution 0.4. The log-likelihood of one Poisson observation is $\ell = x\log\{\lambda\} - \lambda - \log\{x!\}$, so:

$$\begin{aligned}
\ell' &\stackrel{\text{def}}{=} \frac{\partial}{\partial\lambda}(x\log\{\lambda\} - \lambda - \log\{x!\}) && \text{(definition of the score)}\\
&= x\frac{\partial}{\partial\lambda}\log\{\lambda\} - \frac{\partial}{\partial\lambda}\lambda - \frac{\partial}{\partial\lambda}\log\{x!\} && \text{(linearity of differentiation)}\\
&= \frac{x}{\lambda} - 1 - 0 && \text{(derivatives of } \log\lambda\text{, } \lambda\text{, and a constant)}\\
&= \frac{x-\lambda}{\lambda} && \text{(common denominator)}\\
&= \frac{x - \mathrm{E}[X]}{\mathrm{Var}(X)} && (\mathrm{E}[X] = \mathrm{Var}(X) = \lambda)
\end{aligned}$$

Exercise 0.5 (Score function of a Gaussian variable)**.** Derive the score function for a single Gaussian random variable X, with respect to the mean parameter μ, treating the variance σ^2 as known.

Solution

Solution 0.5.

$$\begin{aligned}
\ell' &\stackrel{\text{def}}{=} \frac{\partial}{\partial\mu}\ell && \text{(definition of the score)}\\
&= \frac{\partial}{\partial\mu}\left(\frac{-1}{2}\left(\log\{2\pi\sigma^2\} + \frac{(x-\mu)^2}{\sigma^2}\right)\right) && \text{(Gaussian log-likelihood)}\\
&= \frac{-1}{2}\left(\frac{\partial}{\partial\mu}\log\{2\pi\sigma^2\} + \frac{\partial}{\partial\mu}\frac{(x-\mu)^2}{\sigma^2}\right) && \text{(linearity of differentiation)}\\
&= \frac{-1}{2}\left(0 + \frac{-2(x-\mu)}{\sigma^2}\right) && \text{(constant; chain rule)}\\
&= \frac{x-\mu}{\sigma^2} && \text{(simplify)}\\
&= \frac{x - \mathrm{E}[X]}{\mathrm{Var}(X)} && (\mathrm{E}[X] = \mu \text{ and } \mathrm{Var}(X) = \sigma^2)
\end{aligned}$$

Exercise 0.6 (Score function of an exponential variable)**.** Derive the score function for a single exponential random variable X, with respect to the mean parameter μ.

Solution

Solution 0.6. The exponential density with mean μ is $\mathrm{p}(X = x) = \mu^{-1}e^{-x/\mu}$ for $x > 0$, so $\ell = -\log\{\mu\} - \frac{x}{\mu}$, and:

$$\begin{aligned}
\ell' &\stackrel{\text{def}}{=} \frac{\partial}{\partial\mu}\ell && \text{(definition of the score)} \\
&= \frac{\partial}{\partial\mu}(-\log\{\mu\}) - \frac{\partial}{\partial\mu}\frac{x}{\mu} && \text{(exponential log-likelihood; linearity)} \\
&= -\frac{1}{\mu} + \frac{x}{\mu^2} && \text{(derivatives of } \log\mu \text{ and } \mu^{-1}\text{)} \\
&= \frac{x-\mu}{\mu^2} && \text{(common denominator)} \\
&= \frac{x - \mathrm{E}[X]}{\mathrm{Var}(X)} && (\mathrm{E}[X] = \mu \text{ and } \mathrm{Var}(X) = \mu^2)
\end{aligned}$$

In all four examples (Exercise 0.3, Exercise 0.4, Exercise 0.5, and Exercise 0.6), the score function with respect to the mean turned out to be:

$$\ell' = \frac{x - \mathrm{E}[X]}{\mathrm{Var}(X)}$$

This pattern is no coincidence. With the mean as the parameter, each of these four models is a one-parameter *natural* (or linear) exponential family, whose log-density is linear in x; for every such family, the score with respect to the mean is $(x - \mathrm{E}[X])/\mathrm{Var}(X)$. Other members of the broader exponential family, such as the Weibull distribution with known shape $k \neq 1$, do not have this form. Exponential-family distributions share many special properties (Hogg et al. 2019, sec. 6.7; Dobson and Barnett 2018, chap. 3).

Definition 0.8 (Score equation)**.** The **score equation** (also called the estimating equation) is the equation $\ell'(\tilde{\theta}) = \mathbf{0}_{p\times 1}$, which sets the score function to zero.

Example 0.4 (Score equation of a Bernoulli sample)**.** For n independent Bernoulli observations with $r = \sum_i y_i$ successes, the score is $\sum_{i=1}^{n}(y_i - \pi)/(\pi(1-\pi)) = (r - n\pi)/(\pi(1-\pi))$ (summing Exercise 0.3 over the observations), so for $0 < r < n$ the score equation $(r - n\pi)/(\pi(1-\pi)) = 0$ has the single solution $\pi = r/n$.

Information matrices

Definition 0.9 (Hessian)**.** The **Hessian matrix** of the log-likelihood function is the matrix of its second derivatives with respect to the parameters:

$$\ell'' \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{\theta}} \frac{\partial}{\partial \tilde{\theta}^\top} \ell(\tilde{x}|\tilde{\theta}) \tag{8}$$

The Hessian is named after the mathematician Otto Hesse.

Theorem 0.6 (Elements of the Hessian matrix)**.** *If $\tilde{\theta}$ is a $p \times 1$ vector, then the Hessian is a $p \times p$ matrix, whose ijth entry is:*

$$\ell''_{ij} = \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \ell(\tilde{X} = \tilde{x}|\tilde{\theta}) \tag{9}$$

Proof

Proof. $\frac{\partial}{\partial \tilde{\theta}^\top} \ell$ is the $1 \times p$ row vector whose jth entry is $\frac{\partial}{\partial \theta_j} \ell$. Differentiating each entry with respect to the $p \times 1$ vector $\tilde{\theta}$ gives a $p \times 1$ column of derivatives per entry, so $\frac{\partial}{\partial \tilde{\theta}} \frac{\partial}{\partial \tilde{\theta}^\top} \ell$ is $p \times p$, with ijth entry $\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \ell$. □

Theorem 0.7 (Hessian is the derivative of the transposed score)**.**

$$\ell''(\tilde{x} \mid \tilde{\theta}) = \frac{\partial}{\partial \tilde{\theta}} \big(\ell'(\tilde{x} \mid \tilde{\theta}) \big)^\top$$

Proof

Proof. By Definition 0.7, $\ell' = \frac{\partial}{\partial \tilde{\theta}} \ell$, so $(\ell')^\top = \frac{\partial}{\partial \tilde{\theta}^\top} \ell$, and $\frac{\partial}{\partial \tilde{\theta}} (\ell')^\top = \frac{\partial}{\partial \tilde{\theta}} \frac{\partial}{\partial \tilde{\theta}^\top} \ell = \ell''$ by Definition 0.9. □

Definition 0.10 (Observed information matrix)**.** The **observed information matrix**, written I, is the negative of the Hessian of the log-likelihood:

$$I \stackrel{\text{def}}{=} -\ell''(\tilde{x}|\tilde{\theta}) \tag{10}$$

Definition 0.11 (Expected information)**.** The **expected information matrix**, also called the **Fisher information matrix** or just the **information matrix**, is written $\mathcal{I}$, and is the expected value of the observed information matrix, with the data $\tilde{X}$ treated as random:

$$\mathcal{I}(\tilde{\theta}) \stackrel{\text{def}}{=} \mathrm{E}\Big[I(\tilde{X} \mid \tilde{\theta})\Big] \tag{11}$$

Example 0.5 (Information for Poisson data)**.** For $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$, one observation's score is $x/\lambda - 1$ (Exercise 0.4), whose derivative with respect to λ is $-x/\lambda^2$. Summing over the observations (Theorem 0.5), the Hessian is $\ell'' = -\sum_{i=1}^n x_i/\lambda^2$, so:

$$\begin{aligned}
I(\lambda) &= \frac{\sum_{i=1}^n x_i}{\lambda^2} && \text{(observed information: negative Hessian)} \\
\mathcal{I}(\lambda) &= \mathrm{E}\left[\frac{\sum_{i=1}^n X_i}{\lambda^2}\right] && \text{(expected information)} \\
&= \frac{\sum_{i=1}^n \mathrm{E}[X_i]}{\lambda^2} && \text{(linearity of expectation)} \\
&= \frac{n\lambda}{\lambda^2} && (\mathrm{E}[X_i] = \lambda) \\
&= \frac{n}{\lambda} && \text{(cancel one factor of } \lambda)
\end{aligned}$$

Theorem 0.8 (Mean and variance of the score)**.** *Suppose the set of possible data values does not depend on $\tilde{\theta}$, and the order of differentiation with respect to $\tilde{\theta}$ and integration over the data can be exchanged. Then the score has mean zero, and its variance is the expected information:*

$$\mathrm{E}\Big[\ell'(\tilde{X} \mid \tilde{\theta})\Big] = \mathbf{0}_{p\times 1} \tag{12}$$

$$\mathcal{I}(\tilde{\theta}) = \mathrm{Cov}\Big(\ell'(\tilde{X} \mid \tilde{\theta})\Big) = \mathrm{E}\Big[\ell' \ell'^{\top}\Big] \tag{13}$$

Proof

Proof. We give the proof for a scalar θ and a continuous $\tilde{X}$ with density $\mathrm{p}(\tilde{x} \mid \theta)$; the vector case applies the same steps to each pair of entries, and a discrete $\tilde{X}$ replaces integrals with sums.
For Equation 12:

$$
\begin{aligned}
\mathrm{E}[\ell'] &= \int \left(\frac{\partial}{\partial \theta} \log \mathrm{p}(\tilde{x} \mid \theta) \right) \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x} && \text{(definition of expectation)} \\
&= \int \frac{\frac{\partial}{\partial \theta} \mathrm{p}(\tilde{x} \mid \theta)}{\mathrm{p}(\tilde{x} \mid \theta)} \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x} && \text{(chain rule for log)} \\
&= \int \frac{\partial}{\partial \theta} \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x} && \text{(cancel } \mathrm{p}(\tilde{x} \mid \theta)\text{)} \\
&= \frac{\partial}{\partial \theta} \int \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x} && \text{(exchange derivative and integral)} \\
&= \frac{\partial}{\partial \theta} 1 && \text{(a density integrates to 1)} \\
&= 0
\end{aligned}
$$

For Equation 13, differentiate the identity $0 = \int \left(\frac{\partial}{\partial \theta} \log \mathrm{p}(\tilde{x} \mid \theta) \right) \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x}$ from the first line with respect to θ:

$$
\begin{aligned}
0 &= \int \frac{\partial}{\partial \theta} \left[\left(\frac{\partial}{\partial \theta} \log \mathrm{p}(\tilde{x} \mid \theta) \right) \mathrm{p}(\tilde{x} \mid \theta) \right] d\tilde{x} && \text{(exchange derivative and integral)} \\
&= \int \left(\frac{\partial^2}{\partial \theta^2} \log \mathrm{p}(\tilde{x} \mid \theta) \right) \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x} + \int \left(\frac{\partial}{\partial \theta} \log \mathrm{p}(\tilde{x} \mid \theta) \right) \frac{\partial}{\partial \theta} \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x} && \text{(product rule)} \\
&= \mathrm{E}[\ell''] + \int \left(\frac{\partial}{\partial \theta} \log \mathrm{p}(\tilde{x} \mid \theta) \right)^2 \mathrm{p}(\tilde{x} \mid \theta)\, d\tilde{x} && \left(\frac{\partial}{\partial \theta} \mathrm{p} = \left(\frac{\partial}{\partial \theta} \log \mathrm{p} \right) \mathrm{p}\right) \\
&= \mathrm{E}[\ell''] + \mathrm{E}[\ell'^2] && \text{(definition of expectation)}
\end{aligned}
$$

So $\mathcal{I}(\theta) = \mathrm{E}[-\ell''] = \mathrm{E}[\ell'^2]$, and because $\mathrm{E}[\ell'] = 0$, $\mathrm{E}[\ell'^2] = \mathrm{E}[\ell'^2] - (\mathrm{E}[\ell'])^2 = \mathrm{Var}(\ell')$. □

Example 0.6 (Checking the information equality for Poisson data)**.** For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Pois}(\lambda)$, the score is $\ell' = \sum_{i=1}^n X_i/\lambda - n$ (Example 0.5). Then:

$$
\begin{aligned}
\mathrm{E}[\ell'] &= \frac{\sum_{i=1}^n \mathrm{E}[X_i]}{\lambda} - n && \text{(linearity of expectation)} \\
&= \frac{n\lambda}{\lambda} - n = 0 && (\mathrm{E}[X_i] = \lambda) \\
\mathrm{Var}(\ell') &= \frac{\sum_{i=1}^n \mathrm{Var}(X_i)}{\lambda^2} && \text{(variance of a sum of independent variables; constants)} \\
&= \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda} && (\mathrm{Var}(X_i) = \lambda)
\end{aligned}
$$

which matches $\mathcal{I}(\lambda) = n/\lambda$ from Example 0.5.

Example 0.7 (When the support depends on the parameter)**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Uniform}(0, \theta)$. The likelihood is $\mathcal{L}(\theta) = \theta^{-n}$ for $\theta \geq \max_i x_i$ (and 0 otherwise), so on that range $\ell(\theta) = -n \log \theta$ and $\ell'(\theta) = -n/\theta$. The score is a nonzero constant, so $\mathrm{E}[\ell'] = -n/\theta \neq 0$: Equation 12 fails, because the set of possible data values, $(0, \theta)$, depends on θ.

Sources disagree on the symbols for the observed and expected information (Table 1).

Table 1: Notation for information matrices in several sources

Source	Observed information	Expected information
These notes	I	$\mathcal{J}$
Dobson and Barnett (2018)	$-U'$	$\mathfrak{I}$
Dunn and Smyth (2018)	$\mathfrak{I}$	$\mathcal{J}$
McLachlan and Krishnan (2007)	I	$\mathcal{J}$
Wood (2017)	$\hat{I}$	$\mathcal{J}$

Asymptotic distribution of the maximum likelihood estimate

Theorem 0.9 (Central limit theorem for MLEs)**.** *For* iid *data from a correctly specified model satisfying regularity conditions (including those of Theorem 0.8, an identifiable parameter, a true parameter value in the interior of the parameter space, and a positive definite information matrix), a consistent solution $\hat{\theta}_{ML}$ of the score equation exists, and for large n it has approximately a Gaussian distribution, centered at the true parameter value $\tilde{\theta}$, with covariance matrix equal to the inverse of the expected information:*

$$\hat{\theta}_{ML} \mathrel{\dot{\sim}} \mathrm{N}\left(\tilde{\theta}, \left(\mathcal{J}(\tilde{\theta})\right)^{-1}\right) \tag{14}$$

i Proof

Proof. The proof is beyond the scope of these notes; see (Lehmann 1999, Theorem 7.5.2, p. 501) and (Newey and McFadden 1994). □

These conditions guarantee a consistent root of the score equation; that root is the global maximizer of the likelihood under further conditions, for example when the log-likelihood is strictly concave, as for Poisson data, whose Hessian is negative for every λ (Example 0.5).

Example 0.8 (Approximate distribution of the Poisson MLE)**.** For $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$, setting the score $\sum_{i=1}^{n} X_i/\lambda - n$ (Example 0.6) to zero gives $\hat{\lambda}_{\text{ML}} = \bar{X}$, and

$\mathcal{I}(\lambda) = n/\lambda$ (Example 0.5), so Theorem 0.9 says that for large n:

$$\hat{\lambda}_{\text{ML}} \mathrel{\dot{\sim}} \text{N}\left(\lambda, \frac{\lambda}{n}\right)$$

In this example the mean and variance are exact: $\text{E}\left[\bar{X}\right] = \lambda$ and $\text{Var}\left(\bar{X}\right) = \lambda/n$. The approximation is in the Gaussian shape.

Theorem 0.9 involves the unknown $\tilde{\theta}$, so to use it we estimate $\mathcal{I}(\tilde{\theta})$, by either the expected information at the MLE, $\mathcal{I}(\hat{\theta}_{\text{ML}})$, or the observed information at the MLE, $I(\tilde{x}; \hat{\theta}_{\text{ML}})$. Either way, the estimated standard error of the kth entry of $\hat{\theta}_{\text{ML}}$ is:

$$\widehat{\text{SE}}\left(\hat{\theta}_k\right) = \sqrt{\left[\left(\hat{\mathcal{I}}\right)^{-1}\right]_{kk}}$$

where $\hat{\mathcal{I}}$ is whichever estimate of $\mathcal{I}(\tilde{\theta})$ we chose.

Using the observed information is often more convenient, and there are settings where it is provably better by some criteria (Efron and Hinkley 1978).

Quantifying uncertainty about MLEs

Confidence intervals for MLEs

Definition 0.12 (Wald confidence interval)**.** The approximate $100(1-\alpha)\%$ **Wald confidence interval** for the kth entry θ_k of a parameter vector is

$$\hat{\theta}_k \pm z_{1-\alpha/2} \times \widehat{\text{SE}}\left(\hat{\theta}_k\right)$$

where $\hat{\theta}_k$ is the kth entry of $\hat{\theta}_{\text{ML}}$, $\widehat{\text{SE}}\left(\hat{\theta}_k\right)$ is its estimated standard error, and z_β is the β quantile of the standard Gaussian distribution.

By Theorem 0.9, $(\hat{\theta}_k - \theta_k)/\widehat{\text{SE}}\left(\hat{\theta}_k\right)$ has approximately a standard Gaussian distribution in large samples, so the Wald interval is an approximate confidence interval for θ_k. For a 95% interval, $z_{0.975} \approx 1.96$.

Wald tests

Definition 0.13 (Wald test)**.** The **Wald test** of $H_0 : \theta_k = \theta_{k,0}$ uses the test statistic

$$Z \overset{\text{def}}{=} \frac{\hat{\theta}_k - \theta_{k,0}}{\widehat{\text{SE}}\left(\hat{\theta}_k\right)}$$

which, by Theorem 0.9, has approximately a standard Gaussian distribution under H_0 in large samples. For q constraints $H_0 : \tilde{\theta}_{(q)} = \tilde{\theta}_{(q),0}$ on a $q \times 1$ subvector, the Wald statistic is $\left(\hat{\tilde{\theta}}_{(q)} - \tilde{\theta}_{(q),0}\right)^\top \left(\hat{V}_{(q)}\right)^{-1} \left(\hat{\tilde{\theta}}_{(q)} - \tilde{\theta}_{(q),0}\right)$, where $\hat{V}_{(q)}$ is the corresponding $q \times q$ block of $\left(\hat{\mathcal{J}}\right)^{-1}$; it has approximately a χ^2_q distribution under H_0.

Example 0.9 (Wald interval and test for a Poisson rate)**.** For $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$, $\hat{\lambda}_{\text{ML}} = \bar{x}$ and $\mathcal{J}(\lambda) = n/\lambda$ (Example 0.8), so $\widehat{\text{SE}}\left(\hat{\lambda}\right) = \sqrt{\bar{x}/n}$. With $n = 13$ and $\bar{x} = 72/13$, the 95% Wald interval for λ, and the Wald test of $H_0 : \lambda = 4$, are:

```
n_obs <- 13
xbar_obs <- 72 / 13
se_rate <- sqrt(xbar_obs / n_obs)
z_wald <- (xbar_obs - 4) / se_rate
c(
  lower = xbar_obs - qnorm(0.975) * se_rate,
  upper = xbar_obs + qnorm(0.975) * se_rate,
  z = z_wald,
  p_value = 2 * pnorm(-abs(z_wald))
)
#>     lower     upper         z   p_value
#> 4.2591657 6.8177574 2.3570226 0.0184221
```

Likelihood ratio tests for MLEs

Theorem 0.10 (Wilks' theorem)**.** *Suppose the conditions of Theorem 0.9 hold, and a null hypothesis H_0 imposes q constraints on the parameter vector $\tilde{\theta}$. Let $\hat{\theta}_{ML}$ be the unrestricted MLE and $\hat{\theta}_0$ the MLE under H_0. Then, if H_0 is true, as $n \to \infty$:*

$$\Lambda \overset{def}{=} 2\left(\ell(\hat{\theta}_{ML}) - \ell(\hat{\theta}_0)\right) \overset{d}{\to} \chi^2_q$$

Proof

Proof. We sketch the argument for a scalar parameter and the point null hypothesis $H_0 : \theta = \theta_0$, which imposes $q = 1$ constraint. Then the restricted MLE is $\hat{\theta}_0 = \theta_0$, and

$\Lambda = 2\big(\ell(\hat{\theta}_{\mathrm{ML}}) - \ell(\theta_0)\big)$. For the full proof, see (Wilks 1938) or (Dobson and Barnett 2018, sec. 5.7).
Step 1: expand the log-likelihood around the MLE. The MLE maximizes ℓ at an interior point, so $\ell'(\hat{\theta}_{\mathrm{ML}}) = 0$. A second-order Taylor expansion of $\ell(\theta_0)$ around $\hat{\theta}_{\mathrm{ML}}$ gives:

$$\begin{aligned}
\ell(\theta_0) &\approx \ell(\hat{\theta}_{\mathrm{ML}}) + \ell'(\hat{\theta}_{\mathrm{ML}})(\theta_0 - \hat{\theta}_{\mathrm{ML}}) + \frac{1}{2}\ell''(\hat{\theta}_{\mathrm{ML}})(\theta_0 - \hat{\theta}_{\mathrm{ML}})^2 && \text{(Taylor expansion)} \\
&= \ell(\hat{\theta}_{\mathrm{ML}}) + \frac{1}{2}\ell''(\hat{\theta}_{\mathrm{ML}})(\theta_0 - \hat{\theta}_{\mathrm{ML}})^2 && (\ell'(\hat{\theta}_{\mathrm{ML}}) = 0)
\end{aligned}$$

Rearranging:

$$\begin{aligned}
\Lambda &= 2\big(\ell(\hat{\theta}_{\mathrm{ML}}) - \ell(\theta_0)\big) && \text{(definition of } \Lambda) \\
&\approx -\ell''(\hat{\theta}_{\mathrm{ML}})(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2 && \text{(substitute the expansion)} \\
&= I(\hat{\theta}_{\mathrm{ML}})(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2 && \text{(observed information is the negative Hessian)}
\end{aligned}$$

Step 2: replace the observed information by the expected information. Under the regularity conditions, $I(\hat{\theta}_{\mathrm{ML}})/\mathcal{I}(\theta_0) \to 1$ in probability, by the law of large numbers and the consistency of $\hat{\theta}_{\mathrm{ML}}$, so:

$$\Lambda \approx \mathcal{I}(\theta_0)(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2$$

Step 3: recognize a squared standard Gaussian variable. Define $Z \overset{\text{def}}{=} \sqrt{\mathcal{I}(\theta_0)}\,(\hat{\theta}_{\mathrm{ML}} - \theta_0)$, so that $\Lambda \approx Z^2$. By Theorem 0.9, $\hat{\theta}_{\mathrm{ML}} \mathrel{\dot\sim} \mathrm{N}(\theta_0, \mathcal{I}(\theta_0)^{-1})$ under H_0, so $Z \mathrel{\dot\sim} \mathrm{N}(0,1)$, and therefore $\Lambda \approx Z^2$ has approximately a χ^2_1 distribution.
With q constraints, the same argument in matrix form makes Λ approximately a sum of q squared, independent standard Gaussian variables, which has a χ^2_q distribution. □

Equivalently, in terms of nested models: if a full model M_1 has p free parameters, and a nested model $M_0 \subset M_1$, obtained by imposing q constraints on M_1, has $p_0 = p - q$ free parameters, then when M_0 is true, $\Lambda = 2\big(\ell_{M_1}(\hat{\theta}_{\mathrm{ML}}) - \ell_{M_0}(\hat{\theta}_0)\big)$ converges in distribution to χ^2_q.

Example 0.10 (Likelihood ratio test for a Poisson rate)**.** For $X_1, \ldots, X_n \mathrel{\sim_{\text{iid}}} \mathrm{Pois}(\lambda)$ and $H_0 : \lambda = \lambda_0$, the log-likelihood is $\ell(\lambda) = n\bar{x}\log\lambda - n\lambda - \sum_i \log x_i!$ and $\hat{\lambda}_{\mathrm{ML}} = \bar{x}$, so:

$$\begin{aligned}
\Lambda &= 2(\ell(\bar{x}) - \ell(\lambda_0)) && \text{(definition of } \Lambda) \\
&= 2(n\bar{x}\log\bar{x} - n\bar{x} - n\bar{x}\log\lambda_0 + n\lambda_0) && \text{(substitute; the } \log x_i! \text{ terms cancel)} \\
&= 2n\left(\bar{x}\log\frac{\bar{x}}{\lambda_0} - \bar{x} + \lambda_0\right) && \text{(factor out } n\text{; log of a quotient)}
\end{aligned}$$

For example, with $n = 13$, $\bar{x} = 72/13$, and $\lambda_0 = 4$:

```r
n_obs <- 13
xbar_obs <- 72 / 13
lambda0 <- 4
lr_stat <- 2 * n_obs *
  (xbar_obs * log(xbar_obs / lambda0) - xbar_obs + lambda0)
c(
  statistic = lr_stat,
  p_value = pchisq(lr_stat, df = 1, lower.tail = FALSE)
)
#>  statistic    p_value
#> 6.86082566 0.00881058
```

See also (Dobson and Barnett 2018, sec. 5.7) and https://online.stat.psu.edu/stat504/Lesson02.

Exact and approximate tests

Table 2: Exact tests that assume Gaussian outcomes, and their approximate, large-sample counterparts based on maximum likelihood. p is the number of regression coefficients.

Inference goal	Exact test (Gaussian outcomes)	Null distribution	Approximate test (MLE)	Approximate null distribution
Single coefficient	t-test	T_{n-p}	Wald z-test	$Z \sim N(0,1)$
Linear combination of coefficients	t-test	T_{n-p}	Wald z-test	$Z \sim N(0,1)$
Nested models (q constraints)	Partial F-test	$F_{q,\, n-p}$	Likelihood ratio test or Wald test	χ^2_q
One-sample mean	One-sample t-test	T_{n-1}	z-test	$Z \sim N(0,1)$
Two-sample means	Two-sample t-test (pooled variance)	$T_{n_1+n_2-2}$	z-test	$Z \sim N(0,1)$
K-group means (ANOVA)	F-test	$F_{K-1,\, n-K}$	Likelihood ratio test or Wald test	χ^2_{K-1}

The t and F distributions are defined in Statistical Inference. The exact tests assume $Y_i \overset{\perp\!\!\!\perp}{\sim} N(\mu_i, \sigma^2)$, with a common variance. The approximate tests hold asymptotically, for any model that is correctly specified and satisfies the regularity conditions of Theorem 0.9, Gaussian or not.

Prediction intervals

Definition 0.14 (Prediction interval)**.** A $100(1-\alpha)\%$ **prediction interval** for a future random quantity Y^* is a pair of statistics L and U, computed from the observed data, such that $\Pr(L \leq Y^* \leq U) = 1-\alpha$, where the probability accounts for the randomness of both the observed data and Y^*.

Suppose $X_1, \ldots, X_n \sim_{\text{iid}} \mathrm{N}(\mu, \sigma^2)$ with σ^2 known, and we want to predict the mean $\bar{X}^*$ of m new observations from the same distribution, independent of the first n. The MLE of μ is $\hat{\mu} = \bar{X}$, and the prediction error $\bar{X}^* - \hat{\mu}$ is a difference of independent Gaussian variables, so:

$$\begin{aligned} \mathrm{Var}\big(\bar{X}^* - \hat{\mu}\big) &= \mathrm{Var}\big(\bar{X}^*\big) + \mathrm{Var}(\hat{\mu}) && \text{(variance of a difference of independent variables)} \\ &= \frac{\sigma^2}{m} + \frac{\sigma^2}{n} && \text{(variance of a sample mean)} \end{aligned}$$

and $\bar{X}^* - \hat{\mu} \sim \mathrm{N}\big(0, \sigma^2\big(\frac{1}{m} + \frac{1}{n}\big)\big)$. So a $100(1-\alpha)\%$ prediction interval for $\bar{X}^*$ is:

$$\hat{\mu} \pm z_{1-\alpha/2}\, \sigma \sqrt{\frac{1}{m} + \frac{1}{n}}$$

Usually $m = 1$. The term $1/n$ accounts for the uncertainty in $\hat{\mu}$, and becomes negligible when n is much larger than m.

Example: maximum likelihood for tropical cyclones in Australia

Adapted from (Dobson and Barnett 2018, sec. 1.6.5).

Data

Table 3 records the number of tropical cyclones in northeastern Australia during 13 November-to-April cyclone seasons, from 1956/57 to 1968/69 (Dobson and Barnett 2018, sec. 1.6.5). Figure 1 graphs the number of cyclones by season. Let X_i represent the number of cyclones in season i, and x_i its observed value.

```
cyclones <- tibble::tibble(
  years = c(
    "1956/7", "1957/8", "1958/9", "1959/60", "1960/1", "1961/2", "1962/3",
    "1963/4", "1964/5", "1965/6", "1966/7", "1967/8", "1968/9"
  ),
```

```
  season = 1:13,
  number = c(6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4)
)
cyclones |> knitr::kable()
```

Table 3: Number of tropical cyclones during each November-to-April season in northeastern Australia (Dobson and Barnett 2018, sec. 1.6.5)

years	season	number
1956/7	1	6
1957/8	2	5
1958/9	3	4
1959/60	4	6
1960/1	5	6
1961/2	6	3
1962/3	7	12
1963/4	8	7
1964/5	9	4
1965/6	10	2
1966/7	11	6
1967/8	12	7
1968/9	13	4

Exploratory analysis

Suppose we want to learn how many cyclones to expect per season.

Figure 1 shows no obvious trend, and no obvious correlation between adjacent seasons, so we assume that the seasons' counts are mutually independent. We also assume that they are identically distributed, with a common distribution $\Pr(X = x)$; the expression has no index i, because the distribution is the same for every season.

Figure 2 shows the empirical distribution of the counts.

Table 4 provides summary statistics.

```
n <- nrow(cyclones)
sumx <- sum(cyclones$number)
xbar <- mean(cyclones$number)
tibble::tibble(
  seasons = n,
```

```
cyclones |>
  dplyr::mutate(years = factor(years, levels = years)) |>
  ggplot2::ggplot() +
  ggplot2::aes(x = years, y = number, group = 1) +
  ggplot2::geom_point() +
  ggplot2::geom_line() +
  ggplot2::xlab("Season") +
  ggplot2::ylab("Number of cyclones") +
  ggplot2::expand_limits(y = 0) +
  ggplot2::theme(axis.text.x = ggplot2::element_text(vjust = .5, angle = 45))
```

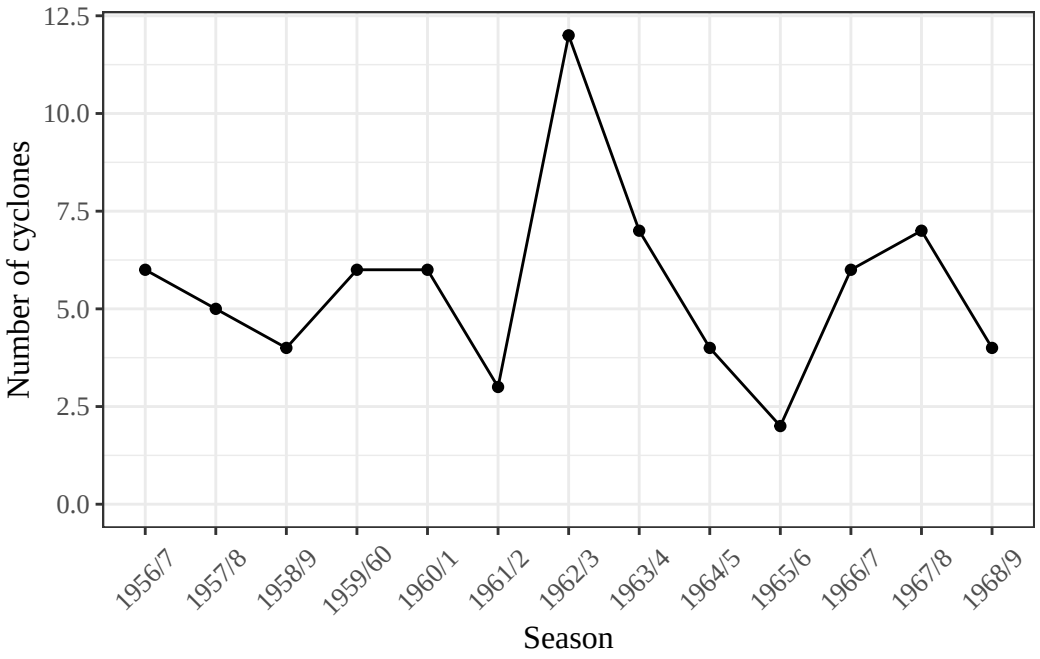


Figure 1: Number of tropical cyclones per season in northeastern Australia, 1956/57 to 1968/69

```
cyclones |>
  ggplot2::ggplot() +
  ggplot2::aes(x = number) +
  ggplot2::geom_bar() +
  ggplot2::expand_limits(x = 0) +
  ggplot2::xlab("Number of cyclones") +
  ggplot2::ylab("Count (number of seasons)")
```

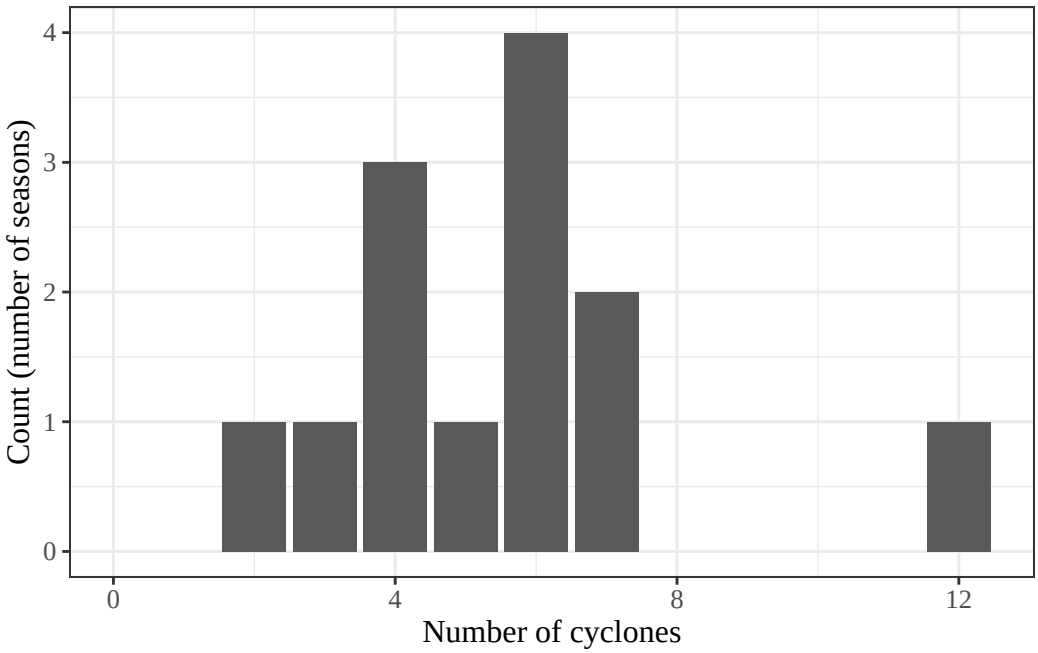


Figure 2: Bar plot of cyclones per season

```
  total = sumx,
  mean = xbar,
  variance = var(cyclones$number)
) |>
  knitr::kable(digits = 2)
```

Table 4: Summary statistics for the `cyclones` data

seasons	total	mean	variance
13	72	5.54	6.1

Model

We want to estimate $\Pr(X = x)$; that is, $\Pr(X = x)$ is our estimand.

We could estimate $\Pr(X = x)$ for each value of x in $0, 1, 2, ...$ separately ("nonparametrically"), using the fraction of our data with $X_i = x$; but then we would be estimating an infinite set of parameters from 13 observations, and each estimate would be imprecise. A parametric model, with a few parameters, will probably do better.

Exercise 0.7 (Choosing a distribution family)**.** What parametric family of probability distributions might we use to model this empirical distribution?

Solution

Solution. The Poisson family. The data are counts, which can take any non-negative integer value. The binomial family also models counts, but a binomial count has an upper limit (its number of trials), and there is no natural upper limit on the number of cyclones in a season.

Exercise 0.8 (The Poisson probability mass function)**.** Write down the Poisson distribution's probability mass function.

Solution

Solution.

$$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, 2, ...\} \tag{15}$$

Estimating the model parameters using maximum likelihood

We can estimate the parameter λ using maximum likelihood estimation.

Exercise 0.9 (Likelihood of one observation)**.** Write down the likelihood of a single observation x, according to the Poisson model.

Solution

Solution.

$$\begin{aligned} \mathcal{L}(\lambda; x) &\stackrel{\text{def}}{=} \Pr(X = x) && \text{(definition of likelihood)} \\ &= \frac{\lambda^x e^{-\lambda}}{x!} && \text{(Poisson PMF)} \end{aligned}$$

Exercise 0.10 (Model parameters)**.** Write down the vector of parameters in the model.

Solution

Solution. There is only one parameter, λ:

$$\theta = (\lambda)$$

Exercise 0.11 (Mean and variance)**.** Write down the population mean and variance of a single observation from the model, as functions of the parameters.

Solution

Solution.

- Population mean: $\mathrm{E}[X] = \lambda$.
- Population variance: $\mathrm{Var}(X) = \lambda$.

The sample mean and variance in Table 4 are of similar size, as the model implies.

Exercise 0.12 (Likelihood of the dataset)**.** Write down the likelihood of the full dataset.

Solution

Solution.

$$
\begin{aligned}
\mathcal{L}(\lambda; \tilde{x}) &\overset{\text{def}}{=} \Pr(\tilde{X} = \tilde{x}) && \text{(definition of likelihood)} \\
&= \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_{13} = x_{13}) && \text{(write out the vector)} \\
&= \prod_{i=1}^{13} \Pr(X_i = x_i) && \text{(independence)} \\
&= \prod_{i=1}^{13} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} && \text{(Poisson PMF)}
\end{aligned}
$$

Exercise 0.13 (Graphing the likelihood)**.** Graph the likelihood as a function of λ.

Solution

Solution.

```r
# `cyclones` is defined earlier on the page:
# nolint next: object_usage_linter.
lik <- function(lambda, y = cyclones$number, n = length(y)) {
  lambda^sum(y) * exp(-n * lambda) / prod(factorial(y))
}

lik_plot <-
  ggplot2::ggplot() +
  ggplot2::geom_function(fun = lik, n = 1001) +
  ggplot2::xlim(min(cyclones$number), max(cyclones$number)) +
  ggplot2::ylab("likelihood") +
  ggplot2::xlab("lambda")

print(lik_plot)
```

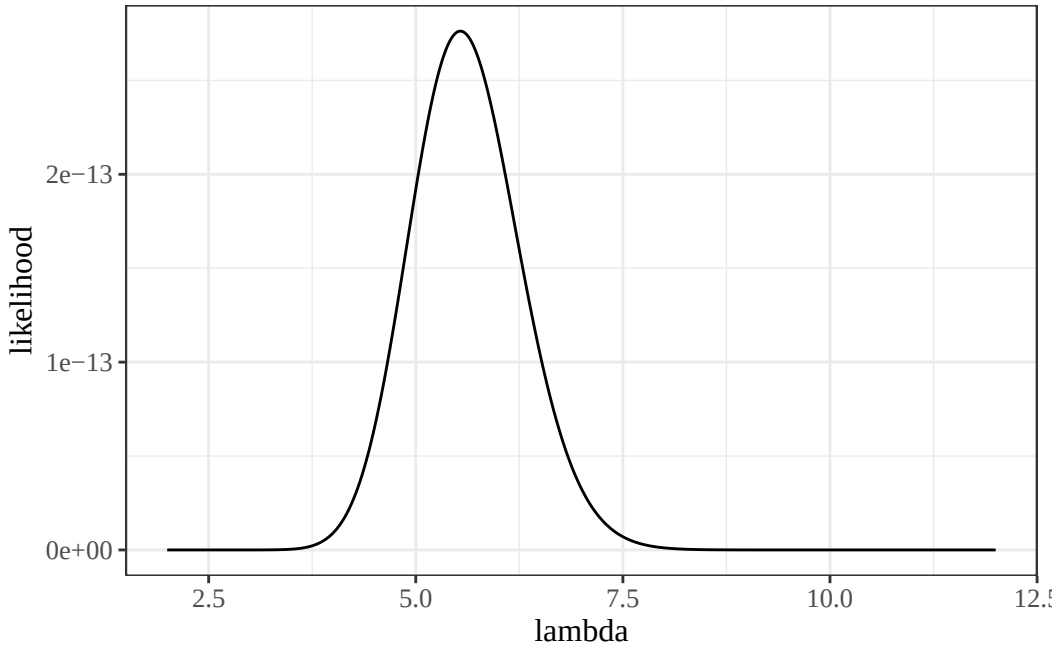


Figure 3: Likelihood of the cyclone data

Exercise 0.14 (Log-likelihood of the dataset)**.** Write down the log-likelihood of the full dataset.

Solution

Solution.

$$
\begin{aligned}
\ell(\lambda; \tilde{x}) &\stackrel{\text{def}}{=} \log\{\mathcal{L}(\lambda; \tilde{x})\} && \text{(definition of log-likelihood)} \\
&= \log\left\{\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right\} && \text{(likelihood of the dataset)} \\
&= \sum_{i=1}^{n} \log\left\{\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right\} && \text{(log of a product)} \\
&= \sum_{i=1}^{n} (\log\{\lambda^{x_i}\} + \log\{e^{-\lambda}\} - \log\{x_i!\}) && \text{(log of a product and of a quotient)} \\
&= \sum_{i=1}^{n} (x_i \log\{\lambda\} - \lambda - \log\{x_i!\}) && \text{(log of a power; } \log e^a = a\text{)} \\
&= \left(\sum_{i=1}^{n} x_i\right) \log\{\lambda\} - n\lambda - \sum_{i=1}^{n} \log\{x_i!\} && \text{(split the sum)}
\end{aligned}
$$

Exercise 0.15 (Graphing the log-likelihood)**.** Graph the log-likelihood as a function of λ.

Solution

Solution.

```r
# `cyclones` is defined earlier on the page:
# nolint next: object_usage_linter.
loglik <- function(lambda, y = cyclones$number, n = length(y)) {
  sum(y) * log(lambda) - n * lambda - sum(lfactorial(y))
}

ll_plot <- ggplot2::ggplot() +
  ggplot2::geom_function(fun = loglik, n = 1001) +
  ggplot2::xlim(min(cyclones$number), max(cyclones$number)) +
  ggplot2::ylab("log-likelihood") +
  ggplot2::xlab("lambda")
ll_plot
```

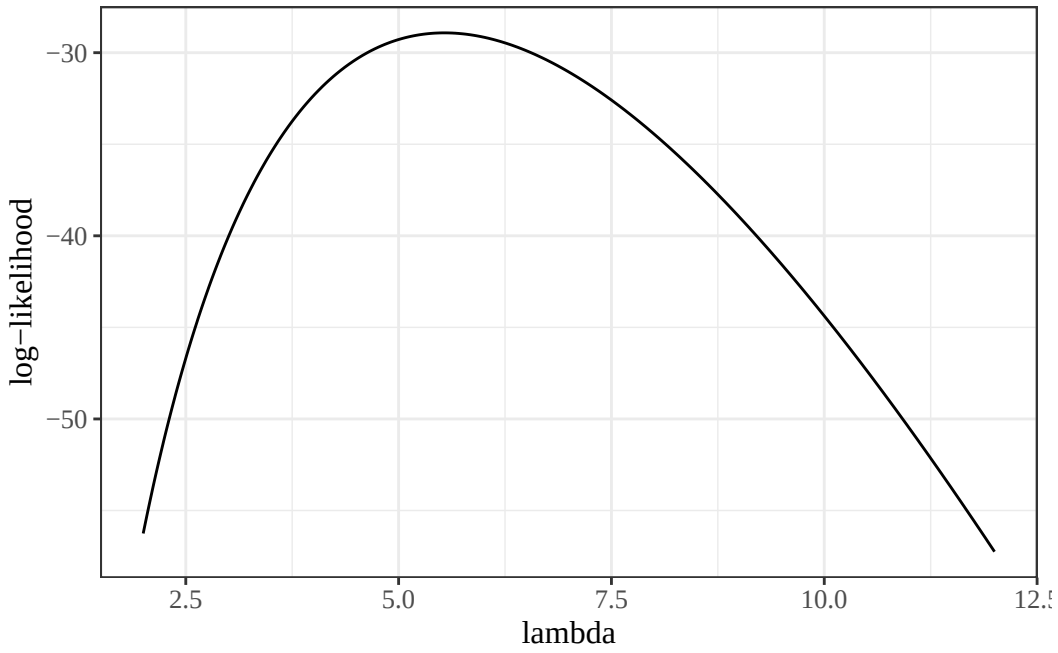


Figure 4: Log-likelihood of the cyclone data

The score function

Exercise 0.16 (Score function of the dataset)**.** Derive the score function for the dataset.

Solution

Solution. The score function is the first derivative of the log-likelihood from Exercise 0.14:

$$\begin{aligned}
\ell'(\lambda;\tilde{x}) &\stackrel{\text{def}}{=} \frac{\partial}{\partial\lambda}\left(\left(\sum_{i=1}^{n} x_i\right)\log\{\lambda\} - n\lambda - \sum_{i=1}^{n}\log\{x_i!\}\right) && \text{(definition of the score)}\\
&= \left(\sum_{i=1}^{n} x_i\right)\frac{\partial}{\partial\lambda}\log\{\lambda\} - n\frac{\partial}{\partial\lambda}\lambda - \frac{\partial}{\partial\lambda}\sum_{i=1}^{n}\log\{x_i!\} && \text{(linearity of differentiation)}\\
&= \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\lambda} - n - 0 && \text{(derivatives of } \log\lambda\text{, } \lambda\text{, and a constant)}\\
&= \frac{n\bar{x}}{\lambda} - n && \left(\textstyle\sum_{i=1}^{n} x_i = n\bar{x}\right)
\end{aligned}$$

For the cyclone data, $n = 13$ and $n\bar{x} = 72$, so $\ell'(\lambda;\tilde{x}) = 72/\lambda - 13$.

Exercise 0.17 (Graphing the score function)**.** Graph the score function.

Solution

Solution.

```r
# `cyclones` is defined earlier on the page:
# nolint next: object_usage_linter.
score <- function(lambda, y = cyclones$number, n = length(y)) {
  (sum(y) / lambda) - n
}

ggplot2::ggplot() +
  ggplot2::geom_function(fun = score, n = 1001) +
  ggplot2::xlim(min(cyclones$number), max(cyclones$number)) +
  ggplot2::ylab("l'(lambda)") +
  ggplot2::xlab("lambda") +
  ggplot2::geom_hline(yintercept = 0, col = "red")
```

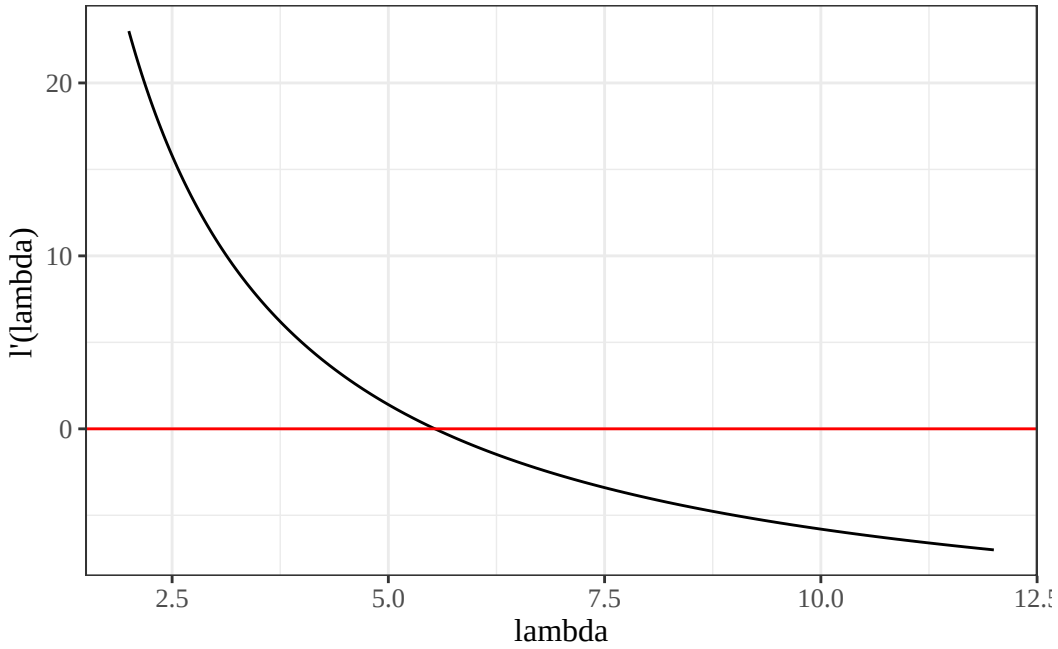


Figure 5: Score function of the cyclone data

The Hessian

Exercise 0.18 (Hessian)**.** Derive the Hessian.

Solution

Solution. With one parameter, the Hessian is the second derivative of the log-likelihood, which is the derivative of the score from Exercise 0.16:

$$\begin{aligned}
\ell''(\lambda; \tilde{x}) &= \frac{\partial}{\partial \lambda}\left(\frac{n\bar{x}}{\lambda} - n\right) && \text{(differentiate the score)} \\
&= n\bar{x}\frac{\partial}{\partial \lambda}\frac{1}{\lambda} - \frac{\partial}{\partial \lambda}n && \text{(linearity of differentiation)} \\
&= -\frac{n\bar{x}}{\lambda^2} && \left(\tfrac{d}{d\lambda}\lambda^{-1} = -\lambda^{-2};\ n \text{ is constant}\right)
\end{aligned}$$

For the cyclone data, $\ell''(\lambda; \tilde{x}) = -72/\lambda^2$.

Exercise 0.19 (Graphing the Hessian)**.** Graph the Hessian.

Solution

Solution.

```r
# `cyclones` is defined earlier on the page:
# nolint next: object_usage_linter.
hessian <- function(lambda, y = cyclones$number, n = length(y)) {
  -sum(y) / (lambda^2)
}

ggplot2::ggplot() +
  ggplot2::geom_function(fun = hessian, n = 1001) +
  ggplot2::xlim(min(cyclones$number), max(cyclones$number)) +
  ggplot2::ylab("l''(lambda)") +
  ggplot2::xlab("lambda") +
  ggplot2::geom_hline(yintercept = 0, col = "red")
```

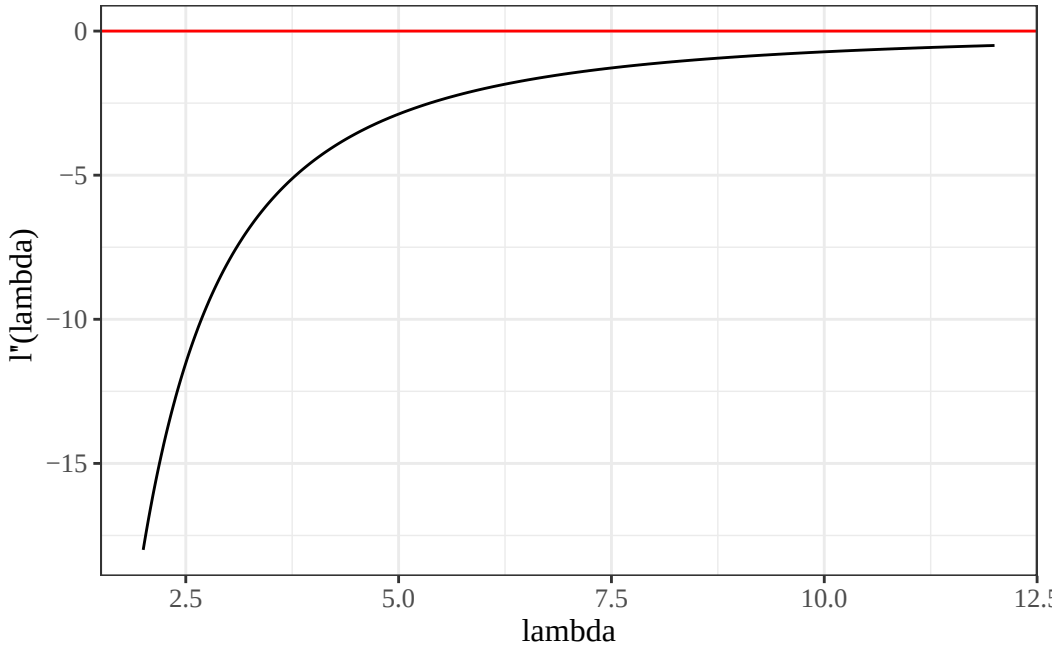


Figure 6: Hessian of the cyclone data's log-likelihood

Exercise 0.20 (Score equation)**.** Write the score equation.

Solution

Solution.

$$\ell'(\lambda; \tilde{x}) = 0$$

Finding the MLE analytically

In this example, we can find the MLE of λ by solving the score equation algebraically.

Exercise 0.21 (Solving the score equation)**.** Solve the score equation from Exercise 0.20 for λ, using the score from Exercise 0.16.

Solution

Solution.

$$\begin{aligned} 0 &= \frac{n\bar{x}}{\lambda} - n && \text{(score of the dataset)} \\ n &= \frac{n\bar{x}}{\lambda} && \text{(add } n \text{ to both sides)} \\ n\lambda &= n\bar{x} && \text{(multiply both sides by } \lambda > 0) \\ \lambda &= \bar{x} && \text{(divide both sides by } n) \end{aligned}$$

Call this solution of the score equation $\tilde{\lambda}$ for now:

$$\tilde{\lambda} \stackrel{\text{def}}{=} \bar{x}$$

Exercise 0.22 (Checking the second derivative)**.** Confirm that the Hessian $\ell''(\lambda; \tilde{x})$ is negative when evaluated at $\tilde{\lambda}$, using Exercise 0.18.

Solution

Solution.

$$\begin{aligned} \ell''(\tilde{\lambda}; \tilde{x}) &= -\frac{n\bar{x}}{\tilde{\lambda}^2} && \text{(Hessian of the log-likelihood)} \\ &= -\frac{n\bar{x}}{\bar{x}^2} && \text{(substitute } \tilde{\lambda} = \bar{x}) \\ &= -\frac{n}{\bar{x}} && \text{(cancel one factor of } \bar{x}) \\ &< 0 && (n > 0 \text{ and } \bar{x} > 0) \end{aligned}$$

Exercise 0.23 (Identifying the MLE)**.** Draw conclusions about the MLE of λ.

Solution

Solution. Since $\ell''(\tilde{\lambda}; \tilde{x}) < 0$, $\tilde{\lambda}$ is a local maximizer of the log-likelihood. Moreover, $\ell''(\lambda; \tilde{x}) = -n\bar{x}/\lambda^2 < 0$ for every $\lambda > 0$, so the log-likelihood is strictly concave, and a local maximizer of a strictly concave function is its unique global maximizer. So $\tilde{\lambda}$ maximizes ℓ, and therefore $\mathcal{L}$:

```
mle <- mean(cyclones$number)
```

$$\hat{\lambda}_{\text{ML}} = \bar{x} = 5.538$$

Exercise 0.24 (Graphing the MLE)**.** Graph the log-likelihood with the MLE superimposed.

Solution

Solution.

```
mle_data <- tibble::tibble(x = mle, y = loglik(mle))
ll_plot +
  ggplot2::geom_point(data = mle_data, ggplot2::aes(x = x, y = y), col = "red")
```

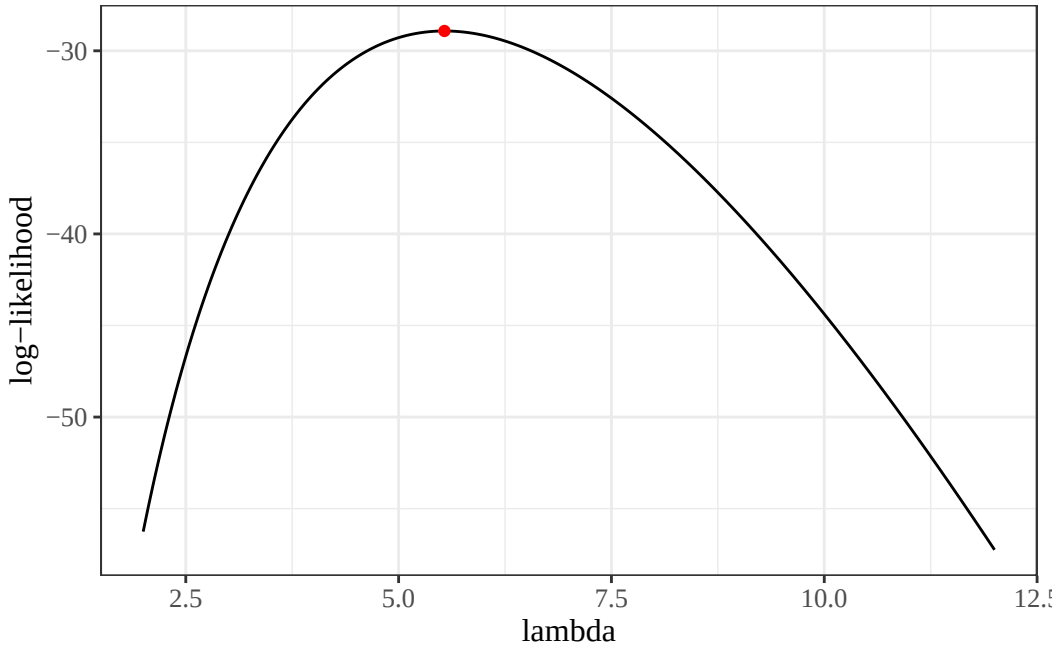


Figure 7: Log-likelihood of the cyclone data, with the MLE marked in red

Figure 8 graphs the observed information, $I(\lambda; \tilde{x}) = -\ell''(\lambda; \tilde{x})$.

Finding the MLE using the Newton-Raphson algorithm

Iterative maximization

When we cannot solve the score equation $\ell'(\theta) = 0$ algebraically, we can search for its solution numerically (Dobson and Barnett 2018, chap. 4).

Definition 0.15 (Newton-Raphson algorithm)**.** The **Newton-Raphson algorithm** for solving the score equation $\ell'(\theta) = 0$ starts from an initial guess $\hat{\theta}^*$ and repeats the update

$$\begin{aligned}\hat{\theta}^* &\leftarrow \hat{\theta}^* - \left(\ell''\left(\tilde{x}; \hat{\theta}^*\right)\right)^{-1} \ell'\left(\tilde{x}; \hat{\theta}^*\right) \\ &= \hat{\theta}^* + \left(I\left(\tilde{x}; \hat{\theta}^*\right)\right)^{-1} \ell'\left(\tilde{x}; \hat{\theta}^*\right)\end{aligned}$$

until the estimate stops changing.

```r
obs_inf <- function(...) -hessian(...) # nolint: object_usage_linter.
ggplot2::ggplot() +
  ggplot2::geom_function(fun = obs_inf, n = 1001) +
  ggplot2::xlim(min(cyclones$number), max(cyclones$number)) +
  ggplot2::ylab("I(lambda)") +
  ggplot2::xlab("lambda") +
  ggplot2::geom_hline(yintercept = 0, col = "red")
```

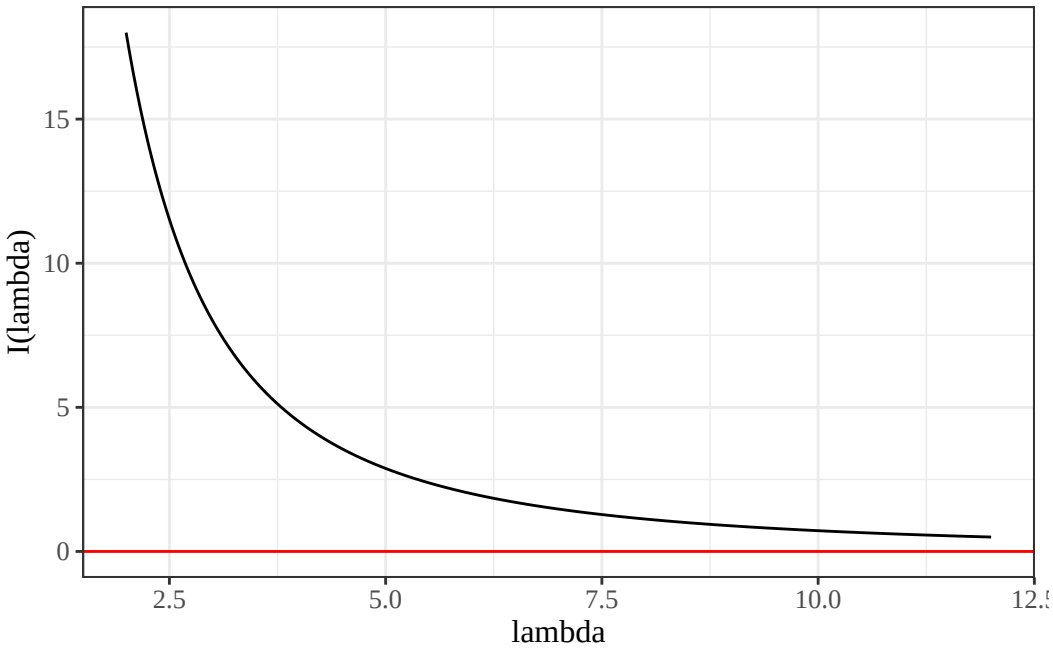


Figure 8: Observed information of the cyclone data

The second line of the update uses $I = -\ell''$ (observed information).

The update comes from approximating the score function near $\hat{\theta}^*$ by its first-order Taylor polynomial:

$$\begin{aligned} \ell'(\theta) &\approx \ell'^*(\theta) \\ &\stackrel{\text{def}}{=} \ell'(\hat{\theta}^*) + \ell''(\hat{\theta}^*)(\theta - \hat{\theta}^*) \end{aligned}$$

The approximate score function $\ell'^*(\theta)$ is linear in θ, so the approximate score equation $\ell'^*(\theta) = 0$ is easy to solve:

$$\begin{aligned} 0 &= \ell'(\hat{\theta}^*) + \ell''(\hat{\theta}^*)(\theta - \hat{\theta}^*) && \text{(set } \ell'^*(\theta) = 0\text{)} \\ -\ell'(\hat{\theta}^*) &= \ell''(\hat{\theta}^*)(\theta - \hat{\theta}^*) && \text{(subtract } \ell'(\hat{\theta}^*)\text{)} \\ -\left(\ell''(\hat{\theta}^*)\right)^{-1}\ell'(\hat{\theta}^*) &= \theta - \hat{\theta}^* && \text{(multiply by } \left(\ell''(\hat{\theta}^*)\right)^{-1} \text{ on the left)} \\ \theta &= \hat{\theta}^* - \left(\ell''(\hat{\theta}^*)\right)^{-1}\ell'(\hat{\theta}^*) && \text{(add } \hat{\theta}^*\text{)} \end{aligned}$$

and the solution becomes the next guess.

Definition 0.16 (Fisher scoring)**. Fisher scoring** (also called the method of scoring) is the Newton-Raphson algorithm with the observed information $I(\tilde{x}; \hat{\theta}^*)$ in the update replaced by the expected information $\mathcal{I}(\hat{\theta}^*)$:

$$\hat{\theta}^* \leftarrow \hat{\theta}^* + \left(\mathcal{I}\left(\hat{\theta}^*\right)\right)^{-1}\ell'\left(\tilde{x}; \hat{\theta}^*\right)$$

The expected information is sometimes simpler to compute than the observed information.

Example 0.11 (Fisher scoring for Poisson data)**.** For $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$, the score is $\ell'(\lambda) = n\bar{x}/\lambda - n$ and the expected information is $\mathcal{I}(\lambda) = n/\lambda$ (Example 0.5), so one Fisher scoring step from any $\hat{\lambda}^* > 0$ gives:

$$\begin{aligned} \hat{\lambda}^* + \left(\mathcal{I}(\hat{\lambda}^*)\right)^{-1}\ell'(\hat{\lambda}^*) &= \hat{\lambda}^* + \frac{\hat{\lambda}^*}{n}\left(\frac{n\bar{x}}{\hat{\lambda}^*} - n\right) && \text{(substitute } \mathcal{I} \text{ and } \ell'\text{)} \\ &= \hat{\lambda}^* + \bar{x} - \hat{\lambda}^* && \text{(distribute } \hat{\lambda}^*/n\text{)} \\ &= \bar{x} && \text{(cancel } \hat{\lambda}^*\text{)} \end{aligned}$$

So Fisher scoring reaches the MLE $\bar{x}$ in one step, while Newton-Raphson needs several (Table 5).

Definition 0.17 (Empirical information matrix)**.** For mutually independent observations, the **empirical information matrix** is (McLachlan and Krishnan 2007):

$$I_e(\theta; \tilde{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell_i' \ell_i'^{\top} - \frac{1}{n} \ell' \ell'^{\top}$$

where ℓ_i' is the score of the ith observation's log-likelihood, and $\ell' = \sum_{i=1}^{n} \ell_i'$.

For iid data, $\frac{1}{n} I_e(\theta; \tilde{x})$ is the sample covariance matrix (with divisor n) of the observations' scores, so it estimates the covariance matrix of one observation's score, $\mathcal{I}(\theta)/n$ (Theorem 0.8). The empirical information needs only first derivatives, so it can be easier to compute than the observed information, and it can replace the observed information in the Newton-Raphson update.

Applying Newton-Raphson to the cyclone data

Example 0.12 (Finding the MLE using the Newton-Raphson algorithm)**.** We found the MLE $\hat{\lambda} = \bar{x}$ by solving the score equation $\ell'(\lambda) = 0$ algebraically (Exercise 0.21). If we could not have solved it, we could instead start from an initial guess such as $\hat{\lambda}^* = 3$ and apply the Newton-Raphson algorithm.

```
cur_lambda_est <- 3
```

From Exercise 0.16 and Exercise 0.18, the score function and Hessian are:

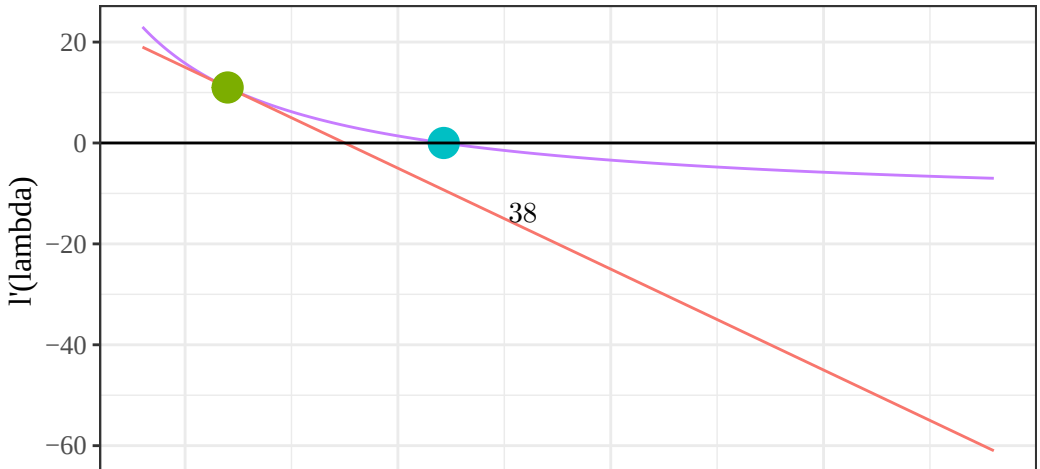
$$\ell'(\lambda; \tilde{x}) = \frac{72}{\lambda} - 13$$
$$\ell''(\lambda; \tilde{x}) = -\frac{72}{\lambda^2}$$

So the first-order Taylor approximation of the score function around $\hat{\lambda}^*$ is:

$$\begin{aligned} \ell'(\lambda) &\approx \ell'^*(\lambda) \\ &\stackrel{\text{def}}{=} \ell'(\hat{\lambda}^*) + \ell''(\hat{\lambda}^*)(\lambda - \hat{\lambda}^*) \\ &= \left(\frac{72}{\hat{\lambda}^*} - 13\right) + \left(-\frac{72}{\left(\hat{\lambda}^*\right)^2}\right)(\lambda - \hat{\lambda}^*) \end{aligned}$$

Figure 9 compares the score function and the approximate score function at $\hat{\lambda}^* = 3$.

```r
# score(), hessian() and loglik() are defined earlier on the page
# nolint start: object_usage_linter.
approx_score <- function(lambda, lhat, ...) {
  score(lambda = lhat, ...) +
    hessian(lambda = lhat, ...) * (lambda - lhat)
}
# nolint end

point_size <- 5

plot1 <- ggplot2::ggplot() +
  ggplot2::geom_function(
    fun = score,
    ggplot2::aes(col = "score function"),
    n = 1001
  ) +
  ggplot2::geom_function(
    fun = approx_score,
    ggplot2::aes(col = "approximate score function"),
    n = 1001,
    args = list(lhat = cur_lambda_est)
  ) +
  ggplot2::geom_point(
    size = point_size,
    ggplot2::aes(
      x = cur_lambda_est, y = score(lambda = cur_lambda_est),
      col = "current estimate"
    )
  ) +
  ggplot2::geom_point(
    size = point_size,
    ggplot2::aes(x = xbar, y = 0, col = "MLE")
  ) +
  ggplot2::xlim(min(cyclones$number), max(cyclones$number)) +
  ggplot2::ylab("l'(lambda)") +
  ggplot2::xlab("lambda") +
  ggplot2::geom_hline(yintercept = 0)

print(plot1)
```



Approximating the score function by a linear function is equivalent to approximating the log-likelihood by a second-order Taylor polynomial (Figure 10):

$$\ell^*(\lambda) \stackrel{\text{def}}{=} \ell(\hat{\lambda}^*) + (\lambda - \hat{\lambda}^*)\ell'(\hat{\lambda}^*) + \frac{1}{2}\ell''(\hat{\lambda}^*)(\lambda - \hat{\lambda}^*)^2$$

Solving the approximate score equation $\ell'^*(\lambda) = 0$ gives the next estimate:

$$\begin{aligned}\lambda &= \hat{\lambda}^* - \ell'(\hat{\lambda}^*) \cdot \left(\ell''(\hat{\lambda}^*)\right)^{-1} \\ &= 4.375\end{aligned}$$

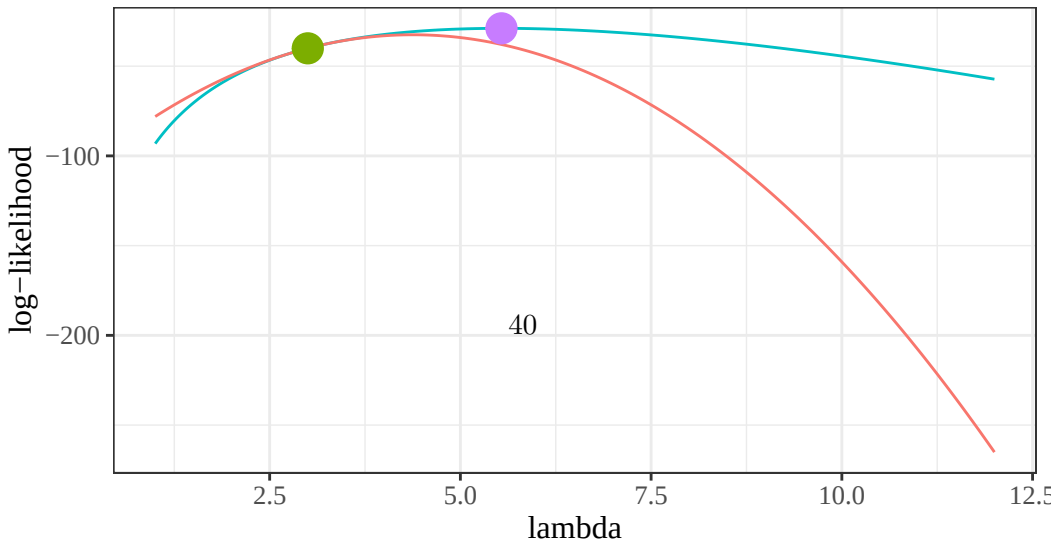
```
new_lambda_est <-
  cur_lambda_est - score(cur_lambda_est) / hessian(cur_lambda_est)
```

We update $\hat{\lambda}^* \leftarrow 4.375$ and repeat the process (Figure 12).

We repeat this process until the log-likelihood stops changing (Table 5).

```
cur_lambda_est <- 3 # restart from the initial guess
tolerance <- 10^-4
max_iter <- 100
nr_info <- tibble::tibble(
  iteration = 0,
  lambda = cur_lambda_est,
  `log(likelihood)` = loglik(cur_lambda_est),
  score = score(cur_lambda_est),
  hessian = hessian(cur_lambda_est)
)

for (cur_iter in 1:max_iter) {
  new_lambda_est <-
    cur_lambda_est - score(cur_lambda_est) / hessian(cur_lambda_est)

  diff_loglik <- loglik(new_lambda_est) - loglik(cur_lambda_est)

  nr_info <- nr_info |>
    dplyr::bind_rows(
      tibble::tibble(
        iteration = cur_iter,
        lambda = new_lambda_est,
        `log(likelihood)` = loglik(new_lambda_est),
        score = score(new_lambda_est),
```

```r
# nolint start: object_usage_linter.
approx_loglik <- function(lambda, lhat, ...) {
  loglik(lambda = lhat, ...) +
    score(lambda = lhat, ...) * (lambda - lhat) +
    1 / 2 * hessian(lambda = lhat, ...) * (lambda - lhat)^2
}
# nolint end

plot_loglik <- ggplot2::ggplot() +
  ggplot2::geom_function(
    fun = loglik,
    ggplot2::aes(col = "log-likelihood"),
    n = 1001
  ) +
  ggplot2::geom_function(
    fun = approx_loglik,
    ggplot2::aes(col = "approximate log-likelihood"),
    n = 1001,
    args = list(lhat = cur_lambda_est)
  ) +
  ggplot2::geom_point(
    size = point_size,
    ggplot2::aes(
      x = cur_lambda_est, y = loglik(lambda = cur_lambda_est),
      col = "current estimate"
    )
  ) +
  ggplot2::geom_point(
    size = point_size,
    ggplot2::aes(x = xbar, y = loglik(xbar), col = "MLE")
  ) +
  ggplot2::xlim(min(cyclones$number) - 1, max(cyclones$number)) +
  ggplot2::ylab("log-likelihood") +
  ggplot2::xlab("lambda")

print(plot_loglik)
```



```
plot2 <- plot1 +
  ggplot2::geom_point(
    size = point_size,
    ggplot2::aes(x = new_lambda_est, y = 0, col = "new estimate")
  ) +
  ggplot2::geom_segment(
    arrow = grid::arrow(),
    linewidth = 2,
    alpha = .7,
    ggplot2::aes(
      x = cur_lambda_est,
      y = approx_score(lhat = cur_lambda_est, lambda = cur_lambda_est),
      xend = new_lambda_est,
      yend = 0,
      col = "update"
    )
  )
print(plot2)
```

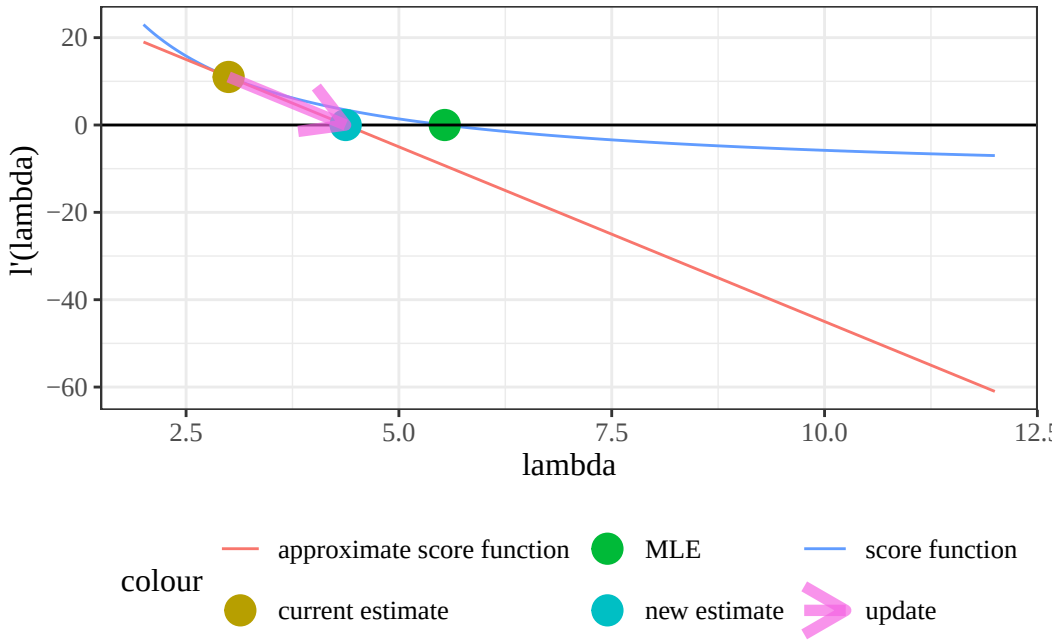


Figure 11: The first Newton-Raphson update: the new estimate is where the approximate score function crosses zero

```
plot2 +
  ggplot2::geom_function(
    fun = approx_score,
    ggplot2::aes(col = "new approximate score function"),
    n = 1001,
    args = list(lhat = new_lambda_est)
  ) +
  ggplot2::geom_point(
    size = point_size,
    ggplot2::aes(
      x = new_lambda_est, y = score(lambda = new_lambda_est),
      col = "new estimate"
    )
  )
```

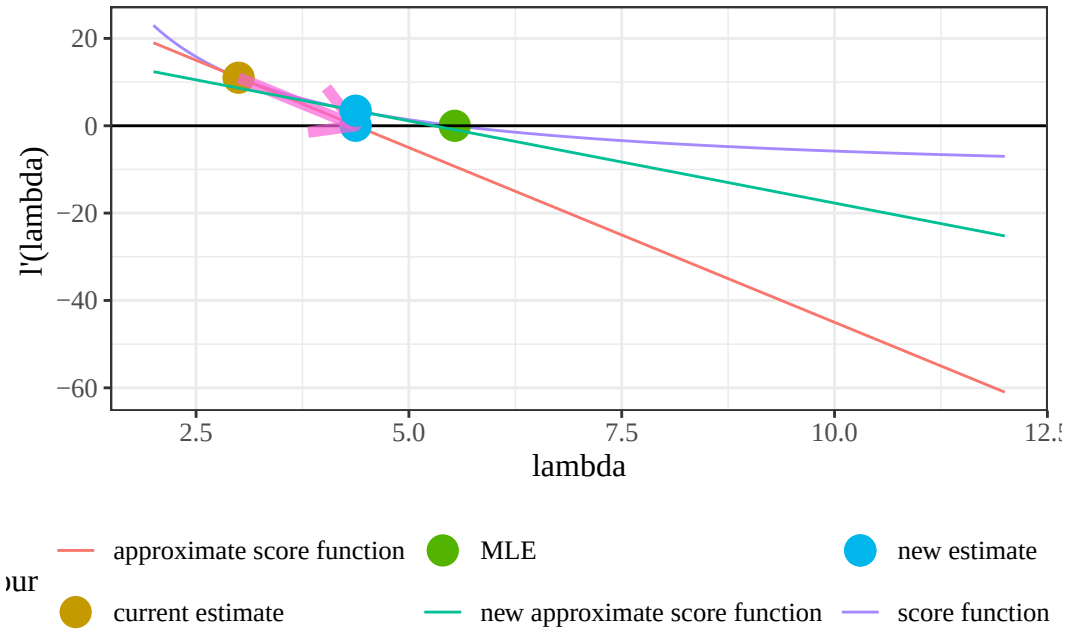


Figure 12: The approximate score function at the updated estimate

```
        hessian = hessian(new_lambda_est),
        `diff(loglik)` = diff_loglik
      )
    )

  cur_lambda_est <- new_lambda_est

  if (abs(diff_loglik) < tolerance) {
    break
  }
}

nr_info |> knitr::kable(digits = 5)
```

Table 5: Convergence of the Newton-Raphson algorithm to the MLE for the cyclone data

iteration	lambda	log(likelihood)	score	hessian	diff(loglik)
0	3.00000	-40.0610	11.00000	-8.00000	NA
1	4.37500	-30.7708	3.45714	-3.76163	9.29018
2	5.29405	-28.9897	0.60016	-2.56895	1.78110
3	5.52768	-28.9176	0.02537	-2.35639	0.07210
4	5.53844	-28.9175	0.00005	-2.34724	0.00014
5	5.53846	-28.9175	0.00000	-2.34722	0.00000

The final estimate matches the closed-form MLE, $\bar{x} = 5.53846$ (Exercise 0.23).

Maximum likelihood for univariate Gaussian models

Suppose $X_1, \ldots, X_n \sim_{\text{iid}} \mathrm{N}(\mu, \sigma^2)$, and let $x_1, \ldots, x_n$ be the observed values. The parameter vector is $\tilde{\theta} = (\mu, \sigma^2)$. We treat σ^2, rather than σ, as the second parameter, and differentiate with respect to σ^2 directly.

By Theorem 0.1, the likelihood is:

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

and by Theorem 0.4, the log-likelihood is:

```
ll_plot +
  ggplot2::geom_segment(
    data = nr_info,
    arrow = grid::arrow(),
    alpha = .7,
    ggplot2::aes(
      x = lambda,
      xend = dplyr::lead(lambda),
      y = `log(likelihood)`,
      yend = dplyr::lead(`log(likelihood)`),
      col = factor(iteration)
    )
  ) +
  ggplot2::labs(col = "iteration")
```

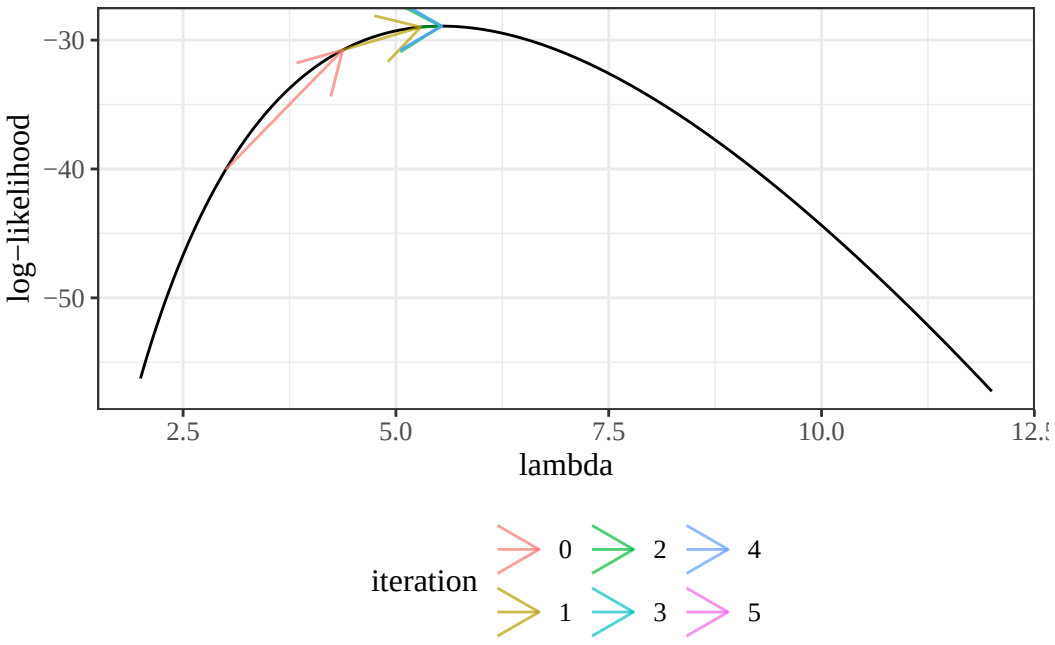


Figure 13: Newton-Raphson steps toward the MLE of the Poisson model (Equation 15) for the cyclone data

$$\begin{aligned}
\ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \left(-\frac{1}{2} \log\{2\pi\sigma^2\} - \frac{(x_i - \mu)^2}{2\sigma^2} \right) && \text{(log of each Gaussian density)} \\
&= -\frac{n}{2} \log\{2\pi\} - \frac{n}{2} \log\{\sigma^2\} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 && \text{(split the sum; log of a product)}
\end{aligned}$$

The score function

The score function is the vector of the two partial derivatives:

$$\ell'(\mu, \sigma^2) = \begin{pmatrix} \frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) \\ \frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) \end{pmatrix}$$

For the first entry, only the last term of ℓ depends on μ:

$$\begin{aligned}
\frac{\partial}{\partial \mu} \ell &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial \mu} (x_i - \mu)^2 && \text{(linearity of differentiation)} \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} -2(x_i - \mu) && \text{(chain rule)} \\
&= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right) && \text{(simplify; split the sum)}
\end{aligned}$$

For the second entry, write σ^2 as a single variable v, so that ℓ contains $-\frac{n}{2} \log v$ and $-\frac{1}{2} v^{-1} \sum_i (x_i - \mu)^2$:

$$\frac{\partial}{\partial \sigma^2} \ell = -\frac{n}{2} (\sigma^2)^{-1} + \frac{1}{2} (\sigma^2)^{-2} \sum_{i=1}^{n} (x_i - \mu)^2 \quad \text{(derivatives of } \log v \text{ and } v^{-1}\text{)}$$

MLE of μ

Setting $\frac{\partial}{\partial \mu} \ell = 0$:

$$\begin{aligned}
0 &= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right) && \text{(score for } \mu\text{)} \\
n\mu &= \sum_{i=1}^{n} x_i && \text{(multiply by } \sigma^2\text{; add } n\mu\text{)} \\
\mu &= \bar{x} && \text{(divide by } n\text{)}
\end{aligned}$$

This solution does not depend on σ^2. The second derivative is

$$\frac{\partial^2 \ell}{\partial \mu^2} = \frac{\partial}{\partial \mu} \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right) = -\frac{n}{\sigma^2} < 0,$$

so for every fixed σ^2, ℓ is maximized over μ at $\bar{x}$, and $\hat{\mu}_{\mathrm{ML}} = \bar{x}$.

MLE of σ^2

Definition 0.18 (Profile log-likelihood)**.** Split a parameter vector into (ψ, λ), and for each fixed value of ψ let $\widehat{\lambda}(\psi)$ maximize $\ell(\psi, \lambda)$ over λ. The **profile log-likelihood** of ψ is

$$\ell_p(\psi) \stackrel{\text{def}}{=} \ell\big(\psi, \widehat{\lambda}(\psi)\big).$$

Example 0.13 (Profile log-likelihood of a Gaussian variance)**.** In the Gaussian model, $\hat{\mu} = \bar{x}$ maximizes ℓ over μ for every value of σ^2 (Section), so the profile log-likelihood of σ^2 is

$$\begin{aligned} \ell_p(\sigma^2) &= \ell(\bar{x}, \sigma^2) && \text{(definition of the profile log-likelihood)} \\ &= -\frac{n}{2} \log\{2\pi\} - \frac{n}{2} \log\{\sigma^2\} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 && \text{(Gaussian log-likelihood at } \mu = \bar{x}) \end{aligned}$$

Setting $\frac{\partial}{\partial \sigma^2} \ell = 0$:

$$\begin{aligned} 0 &= -\frac{n}{2} (\sigma^2)^{-1} + \frac{1}{2} (\sigma^2)^{-2} \sum_{i=1}^{n} (x_i - \mu)^2 && \text{(score for } \sigma^2) \\ \frac{n}{2} (\sigma^2)^{-1} &= \frac{1}{2} (\sigma^2)^{-2} \sum_{i=1}^{n} (x_i - \mu)^2 && \text{(add } \tfrac{n}{2}(\sigma^2)^{-1}) \\ \sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 && \text{(multiply by } 2(\sigma^2)^2/n) \end{aligned}$$

Substituting the maximizer $\mu = \bar{x}$, which does not depend on σ^2, maximizes the profile log-likelihood of Example 0.13, and gives:

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

The profile log-likelihood, $\ell_p(\sigma^2) = -\frac{n}{2}\log\{2\pi\} - \frac{n}{2}\log\{\sigma^2\} - \frac{n\hat{\sigma}^2_{\mathrm{ML}}}{2\sigma^2}$, increases for $\sigma^2 < \hat{\sigma}^2_{\mathrm{ML}}$ and decreases for $\sigma^2 > \hat{\sigma}^2_{\mathrm{ML}}$, because its derivative, $\frac{n}{2}(\sigma^2)^{-2}(\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2)$, has the sign of $\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2$. So $(\bar{x}, \hat{\sigma}^2_{\mathrm{ML}})$ is the global maximizer, provided the x_i are not all equal.

Differentiating with respect to σ^2 as a single variable, rather than with respect to σ, keeps the algebra short. Replacing σ^2 with the precision $\tau \stackrel{\text{def}}{=} 1/\sigma^2$ and differentiating with respect to τ can be shorter still, because τ enters the log-likelihood as $\frac{n}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^n (x_i - \mu)^2$. By the invariance of maximum likelihood estimates, $\hat{\tau}_{\mathrm{ML}} = 1/\hat{\sigma}^2_{\mathrm{ML}}$.

This MLE divides by n, so it is a biased estimator of σ^2 (bias of the divide-by-n estimator).

Second derivatives

The remaining second derivatives are:

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial(\sigma^2)^2} &= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^n (x_i - \mu)^2\right) && \text{(differentiate the score for } \sigma^2\text{)} \\
&= \frac{n}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}\sum_{i=1}^n (x_i - \mu)^2 && \text{(power rule)} \\
\frac{\partial^2 \ell}{\partial\mu\,\partial\sigma^2} &= \frac{\partial}{\partial\mu}\left(\frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^n (x_i - \mu)^2\right) && \text{(only this term depends on } \mu\text{)} \\
&= -(\sigma^2)^{-2}\sum_{i=1}^n (x_i - \mu) && \text{(chain rule)}
\end{aligned}$$

At the MLE, $\sum_{i=1}^n (x_i - \bar{x}) = 0$ and $\sum_{i=1}^n (x_i - \bar{x})^2 = n\hat{\sigma}^2$ (writing $\hat{\sigma}^2$ for $\hat{\sigma}^2_{\mathrm{ML}}$), so:

$$\begin{aligned}
\left.\frac{\partial^2 \ell}{\partial(\sigma^2)^2}\right|_{\mathrm{MLE}} &= \frac{n}{2}(\hat{\sigma}^2)^{-2} - (\hat{\sigma}^2)^{-3} n\hat{\sigma}^2 && \text{(substitute)} \\
&= -\frac{n}{2}(\hat{\sigma}^2)^{-2} && \text{(simplify)} \\
\left.\frac{\partial^2 \ell}{\partial\mu\,\partial\sigma^2}\right|_{\mathrm{MLE}} &= 0 && \text{(substitute)}
\end{aligned}$$

Information matrix and standard errors

Collecting the second derivatives at the MLE, the observed information is

$$I(\hat{\mu}, \hat{\sigma}^2) = \begin{bmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2(\hat{\sigma}^2)^2} \end{bmatrix}$$

Its diagonal entries are positive and its off-diagonal entries are zero, so it is positive definite, consistent with the MLE being a maximum. The inverse of a diagonal matrix inverts each diagonal entry, so:

$$\left(I(\hat{\mu}, \hat{\sigma}^2)\right)^{-1} = \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2(\hat{\sigma}^2)^2}{n} \end{bmatrix}$$

By Theorem 0.9, the estimated standard errors are $\widehat{\mathrm{SE}}(\hat{\mu}) = \hat{\sigma}/\sqrt{n}$ and $\widehat{\mathrm{SE}}(\hat{\sigma}^2) = \hat{\sigma}^2\sqrt{2/n}$, and the two estimates are approximately uncorrelated in large samples.

See also (Casella and Berger 2002, Example 7.2.12).

Example: hormone therapy study

This example fits a Gaussian model to real data by maximum likelihood, and then uses simulation to examine the properties of maximum likelihood estimation for that model.

Data

The "heart and estrogen/progestin study" (HERS) was a clinical trial of hormone therapy for prevention of recurrent heart attacks and death among 2,763 post-menopausal women with existing coronary heart disease (CHD) (Hulley et al. 1998).

The trial was conducted at 20 US clinical centers. Participants were randomized to receive either conjugated equine estrogens (0.625 mg/day) plus medroxyprogesterone acetate (2.5 mg/day) or a matching placebo (Hulley et al. 1998). Women were followed for an average of 4.1 years (Hulley et al. 1998).

The primary outcome was nonfatal myocardial infarction or CHD death (Hulley et al. 1998).

We model the distribution of fasting glucose among HERS participants who do not have diabetes and do not exercise.

The HERS data are distributed with Vittinghoff et al. (2012) on the book's companion website:

```
# one unbroken string, so that link checkers test the whole URL:
url <- "https://regression.ucsf.edu/sites/g/files/tkssra16191/files/wysiwyg/home/data/hersda
hers <- haven::read_dta(url)
```

Table 6: The first rows of the HERS dataset

```
hers |> head()
#> # A tibble: 6 x 37
#>      HT   age raceth nonwhite smoking drinkany exercise physact globrat poorfair
#>   <dbl> <dbl>  <dbl>    <dbl>   <dbl>    <dbl>    <dbl>   <dbl>   <dbl>    <dbl>
#> 1     0    70      2        1       0        0        0       5       3        0
#> 2     0    62      2        1       0        0        0       1       3        0
#> 3     1    69      1        0       0        0        0       3       3        0
#> 4     0    64      1        0       1        1        0       1       3        0
#> 5     0    65      1        0       0        0        0       2       3        0
#> 6     1    68      2        1       0        1        0       3       3        0
#> # i 27 more variables: medcond <dbl>, htnmeds <dbl>, statins <dbl>,
#> #   diabetes <dbl>, dmpills <dbl>, insulin <dbl>, weight <dbl>, BMI <dbl>,
#> #   waist <dbl>, WHR <dbl>, glucose <dbl>, weight1 <dbl>, BMI1 <dbl>,
#> #   waist1 <dbl>, WHR1 <dbl>, glucose1 <dbl>, tchol <dbl>, LDL <dbl>,
#> #   HDL <dbl>, TG <dbl>, tchol1 <dbl>, LDL1 <dbl>, HDL1 <dbl>, TG1 <dbl>,
#> #   SBP <dbl>, DBP <dbl>, age10 <dbl>
```

The `rmb` R package includes the same file, which these notes use so that rendering does not depend on the website (Table 6):

```
hers <- rmb::hers |> haven::zap_labels()
```

To keep the likelihood graphs readable, we use only the first 100 eligible participants (Figure 14); with the whole subset, the likelihood would be too concentrated to graph clearly.

```
n_obs <- 100

data1 <-
  hers |>
  dplyr::filter(
    diabetes == 0,
    exercise == 0
  ) |>
  head(n_obs)

glucose_data <- data1$glucose
```

The histogram is irregular, as histograms of 100 observations often are, with a somewhat longer right tail than left tail. A Gaussian model is a rough but usable starting point.

```r
plot1 <-
  data1 |>
  ggplot2::ggplot() +
  ggplot2::aes(x = glucose) +
  ggplot2::geom_histogram(
    ggplot2::aes(y = ggplot2::after_stat(density)),
    bins = 20
  ) +
  ggplot2::xlab("Fasting glucose (mg/dL)")

print(plot1)
```

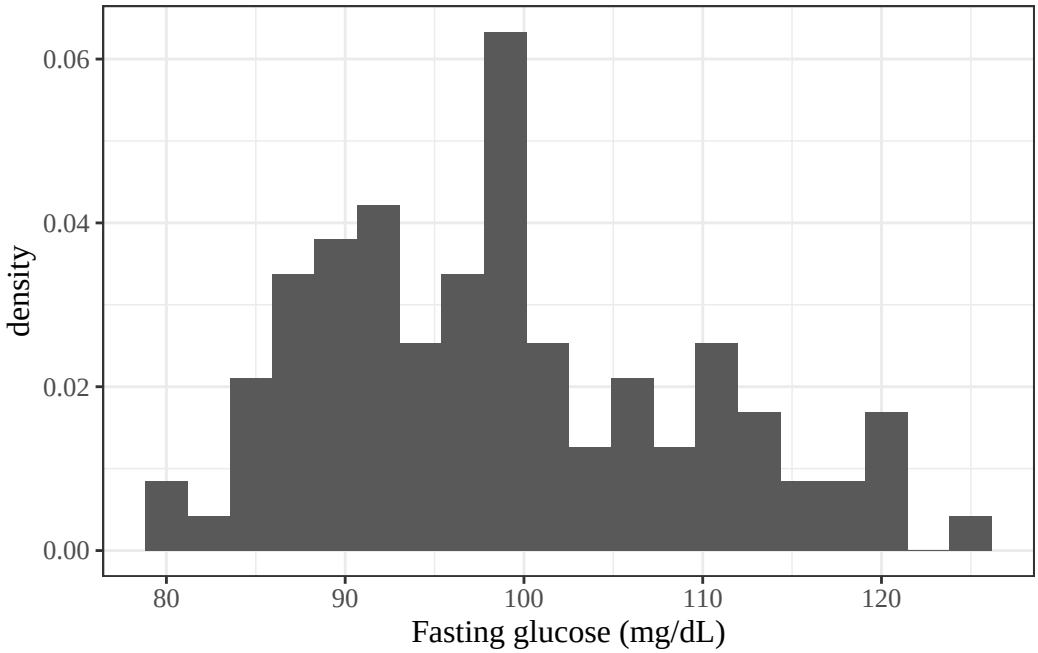


Figure 14: Fasting glucose among 100 HERS participants without diabetes who do not exercise

Maximum likelihood estimates

By the Gaussian MLEs, $\hat{\mu}_{\text{ML}} = \bar{x}$ and $\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_i (x_i - \bar{x})^2$:

```
mu_hat <- mean(glucose_data)
sigma_sq_hat <- mean((glucose_data - mu_hat)^2)
sigma_hat <- sqrt(sigma_sq_hat)
c(mu_hat = mu_hat, sigma_sq_hat = sigma_sq_hat)
#>       mu_hat sigma_sq_hat
#>       98.660      104.744
```

Figure 15 superimposes the fitted Gaussian density on the histogram.

```
plot1 +
  ggplot2::geom_function(
    fun = function(x) dnorm(x, mean = mu_hat, sd = sigma_hat),
    col = "red"
  )
```

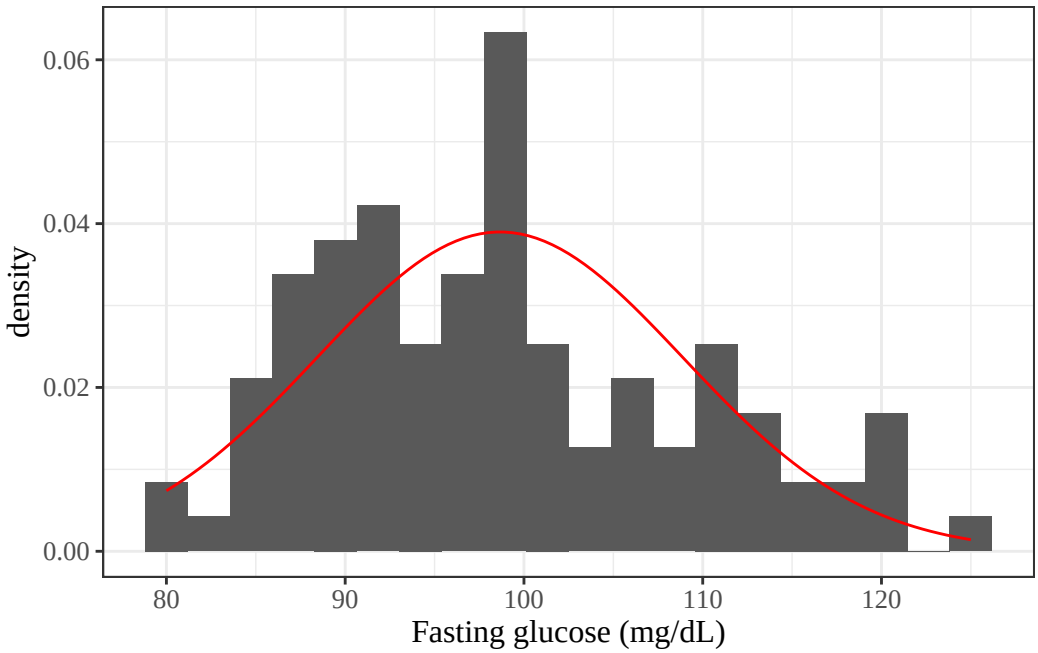


Figure 15: Fasting glucose, with the fitted Gaussian density in red

The fitted curve follows the overall shape of the histogram, but it underestimates the frequency of values between 110 and 122 mg/dL, consistent with the histogram's longer right tail.

Likelihood and log-likelihood functions

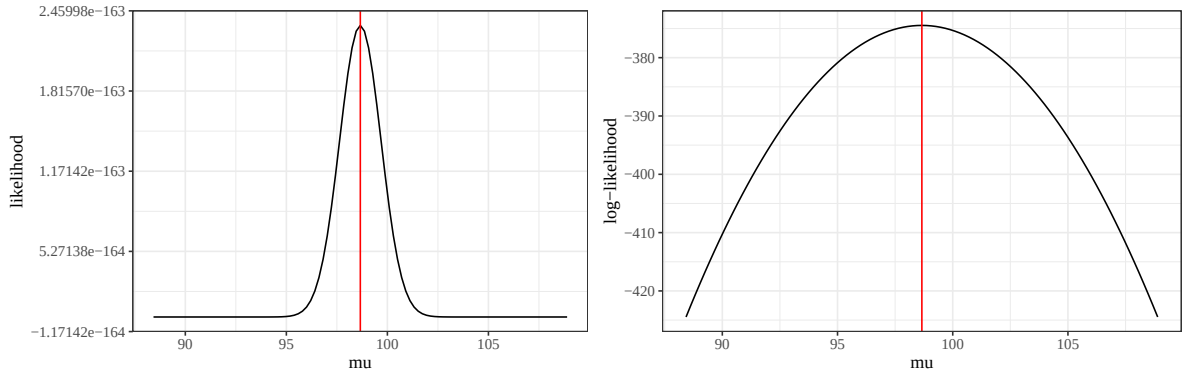
It is numerically better to compute the log-likelihood first and exponentiate it to get the likelihood, because a product of 100 densities can underflow to zero:

```
loglik <- function(mu, sigma, x) {
  n <- length(x)
  normalizing_constant <- -n / 2 * log(2 * pi * sigma^2)
  # written with sum(x), sum(x^2) so that `mu` can be a vector of values:
  kernel <- -1 / (2 * sigma^2) * (sum(x^2) - 2 * sum(x) * mu + n * mu^2)
  normalizing_constant + kernel
}

lik <- function(...) exp(loglik(...))
```

Figure 16 graphs the likelihood and log-likelihood as functions of μ, with σ fixed at $\hat{\sigma}_{\text{ML}}$.

```
ggplot2::ggplot() +
  ggplot2::geom_function(
    fun = lik,
    args = list(sigma = sigma_hat, x =
  ) +
  ggplot2::xlim(mu_hat + c(-1, 1) * si
  ggplot2::xlab("mu") +
  ggplot2::ylab("likelihood") +
  ggplot2::geom_vline(xintercept = mu_h
```

```
ggplot2::ggplot() +
  ggplot2::geom_function(
    fun = loglik,
    args = list(sigma = sigma_hat, x = glucose_data)
  ) +
  ggplot2::xlim(mu_hat + c(-1, 1) * sigma_hat) +
  ggplot2::xlab("mu") +
  ggplot2::ylab("log-likelihood") +
  ggplot2::geom_vline(xintercept = mu_hat, col = "red"
```



(a) Likelihood

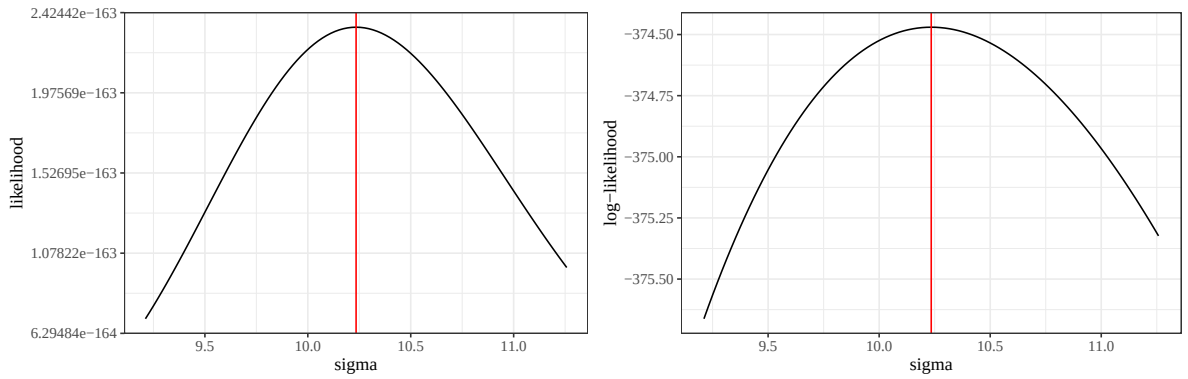
(b) Log-likelihood

Figure 16: Likelihood and log-likelihood of the HERS glucose data as functions of μ, with $\sigma = \hat{\sigma}_{\text{ML}}$; the red line marks $\hat{\mu}_{\text{ML}}$

Figure 17 graphs them as functions of σ, with μ fixed at $\hat{\mu}_{\text{ML}}$.

```r
ggplot2::ggplot() +
  ggplot2::geom_function(
    fun = lik,
    args = list(mu = mu_hat, x = glucos
  ) +
  ggplot2::xlim(sigma_hat * c(0.9, 1.1)
  ggplot2::geom_vline(xintercept = sigr
  ggplot2::xlab("sigma") +
  ggplot2::ylab("likelihood")
```

```r
ggplot2::ggplot() +
  ggplot2::geom_function(
    fun = loglik,
    args = list(mu = mu_hat, x = glucose_data)
  ) +
  ggplot2::xlim(sigma_hat * c(0.9, 1.1)) +
  ggplot2::geom_vline(xintercept = sigma_hat, col = "r
  ggplot2::xlab("sigma") +
  ggplot2::ylab("log-likelihood")
```



(a) Likelihood

(b) Log-likelihood

Figure 17: Likelihood and log-likelihood of the HERS glucose data as functions of σ, with $\mu = \hat{\mu}_{\text{ML}}$; the red line marks $\hat{\sigma}_{\text{ML}}$

Log-likelihood surface

Figure 18 graphs the log-likelihood over both parameters at once.

```
n_points <- 25
mu_grid <- seq(94, 104, length.out = n_points)
sigma_grid <- seq(7, 15, length.out = n_points)
lliks <- outer(mu_grid, sigma_grid, loglik, x = glucose_data)

plotly::plot_ly(
  type = "surface",
  x = ~mu_grid,
  y = ~sigma_grid,
  # plot_ly() expects rows of z to correspond to y values,
  # so the matrix is transposed;
  # see https://stackoverflow.com/questions/69472185
  z = ~ t(lliks)
) |>
  plotly::layout(
    scene = list(
      xaxis = list(title = "mu"),
      yaxis = list(title = "sigma"),
      zaxis = list(title = "log-likelihood")
    )
  )
```

(This graph is interactive; see the HTML version of this page to view it.)

Figure 18: Log-likelihood of the HERS glucose data as a function of μ and σ (interactive: drag to rotate)

Standard errors by sample size

By Section , the estimated standard error of $\hat{\mu}_{\text{ML}}$ is

$$\widehat{\text{SE}}(\hat{\mu}) = \sqrt{\left[(I(\hat{\mu}, \hat{\sigma}^2))^{-1}\right]_{11}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

which shrinks in proportion to $1/\sqrt{n}$ (Figure 19).

```r
se_mu_hat <- function(n, sigma) sigma / sqrt(n)
ggplot2::ggplot() +
  ggplot2::geom_function(fun = se_mu_hat, args = list(sigma = sigma_hat)) +
  ggplot2::scale_x_log10(
    limits = c(10, 10^5), name = "Sample size",
    labels = scales::label_comma()
  ) +
  ggplot2::ylab("Standard error of mu-hat (mg/dL)")
```

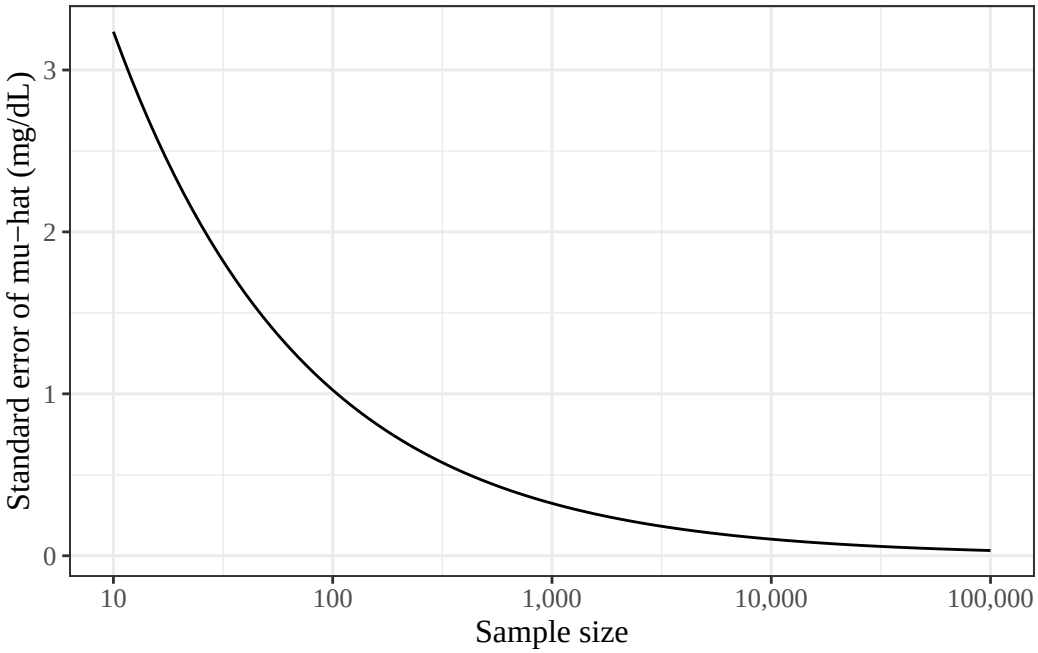


Figure 19: Standard error of $\hat{\mu}_{\text{ML}}$ as a function of sample size, with $\sigma = \hat{\sigma}_{\text{ML}}$

Power

Suppose we test the null hypothesis $H_0 : \mu = \mu_0$, with $\mu_0 = 95$ mg/dL, at significance level $\alpha = 0.05$, and suppose for simplicity that σ is known, equal to $\hat{\sigma}_{\text{ML}}$. Then under H_0, $\bar{X} \sim \text{N}(\mu_0, \sigma^2/n)$, and the test rejects H_0 when $\bar{x}$ falls outside the non-rejection interval

$$\mu_0 \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

```
mu0 <- 95
se <- se_mu_hat(n = n_obs, sigma = sigma_hat)
margin <- qnorm(0.975) * se
bounds <- mu0 + c(lower = -1, upper = 1) * margin
bounds
#>   lower   upper
#> 92.9941 97.0059
```

Definition 0.19 (Power)**.** The **power** of a hypothesis test against an alternative parameter value is the probability that the test rejects the null hypothesis when the parameter equals that alternative value.

For this test, under $\mu = \mu_1$, $\bar{X} \sim \text{N}(\mu_1, \sigma^2/n)$, so:

$$\text{power}(\mu_1) = \Phi\left(\frac{\mu_0 - z_{1-\alpha/2}\,\sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(\frac{\mu_0 + z_{1-\alpha/2}\,\sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right)$$

where Φ is the standard Gaussian CDF. For example, the power against $\mu_1 = 100$ mg/dL is:

```
power <- function(n, null, alt, sigma) {
  se <- sigma / sqrt(n)
  lower <- null - qnorm(0.975) * se
  upper <- null + qnorm(0.975) * se
  pnorm((lower - alt) / se) + pnorm((upper - alt) / se, lower.tail = FALSE)
}
mu1 <- 100
power(n = n_obs, null = mu0, alt = mu1, sigma = sigma_hat)
#> [1] 0.99828
```

Figure 20 graphs the power as a function of the sample size.

The alternative μ_1 should be chosen before seeing the data, as a difference worth detecting. Power computed at $\mu_1 = \hat{\mu}$ ("observed power") is a function of the p-value, so it adds no information about the data already analyzed (Hoenig and Heisey 2001).

```r
ggplot2::ggplot() +
  ggplot2::geom_function(
    fun = power,
    args = list(null = mu0, alt = mu1, sigma = sigma_hat),
    n = 199
  ) +
  ggplot2::xlim(c(2, 200)) +
  ggplot2::ylim(0, 1) +
  ggplot2::ylab("Power") +
  ggplot2::xlab("n")
```

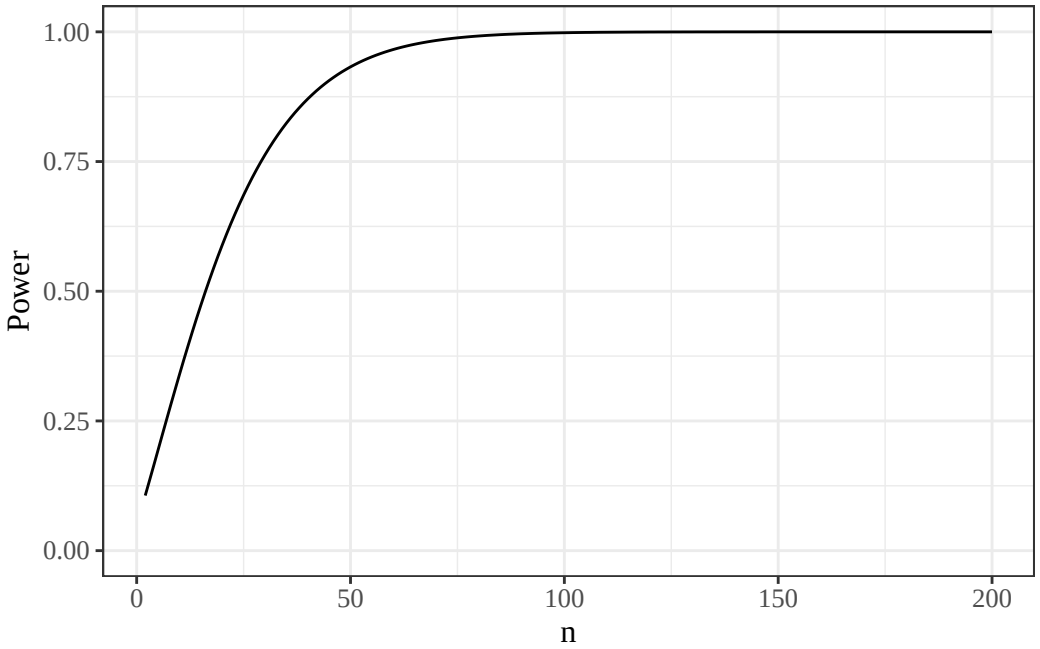


Figure 20: Power of the test of $H_0 : \mu = 95$ against $\mu_1 = 100$ mg/dL, by sample size

Simulation

To check how maximum likelihood estimation behaves for this model, we simulate many datasets from a Gaussian distribution whose parameters equal the HERS estimates, analyze each one, and summarize the results.

`do_one_sim()` simulates and analyzes one dataset: it computes $\hat{\mu}$, its estimated standard error, a 95% t-based confidence interval for μ, and the t-test of $H_0 : \mu = \mu_0$.

```
do_one_sim <- function(n, mu, mu0, sigma2, return_data = FALSE) {
  # generate data
  x <- rnorm(n = n, mean = mu, sd = sqrt(sigma2))

  # analyze data
  est <- mean(x)
  se_est <- sd(x) / sqrt(n)
  confint <- est + c(-1, 1) * se_est * qt(0.975, df = n - 1)
  tstat <- abs(est - mu0) / se_est
  pval <- 2 * pt(q = tstat, df = n - 1, lower.tail = FALSE)

  results <- tibble::tibble(
    mu_hat = est,
    sigma_hat = sd(x),
    se_hat = se_est,
    confint_left = confint[1],
    confint_right = confint[2],
    tstat = tstat,
    pval = pval,
    confint_covers = dplyr::between(mu, confint[1], confint[2]),
    test_rejects = pval < 0.05
  )

  if (return_data) {
    list(data = x, results = results)
  } else {
    results
  }
}
```

To check `do_one_sim()`, we compare its output with `stats::t.test()` on the same simulated data:

```r
set.seed(1)
sim_output <- do_one_sim(
  n = 100, mu = mu_hat, mu0 = 80, sigma2 = sigma_sq_hat,
  return_data = TRUE
)
t_test <- t.test(sim_output$data, mu = 80)

dplyr::bind_rows(
  `do_one_sim()` = sim_output$results |>
    dplyr::select(mu_hat, se_hat, confint_left, confint_right, pval),
  `t.test()` = tibble::tibble(
    mu_hat = unname(t_test$estimate),
    se_hat = t_test$stderr,
    confint_left = t_test$conf.int[1],
    confint_right = t_test$conf.int[2],
    pval = t_test$p.value
  ),
  .id = "source"
)
#> # A tibble: 2 x 6
#>   source       mu_hat se_hat confint_left confint_right     pval
#>   <chr>         <dbl>  <dbl>        <dbl>         <dbl>    <dbl>
#> 1 do_one_sim()   99.8  0.919         98.0          102. 4.23e-39
#> 2 t.test()       99.8  0.919         98.0          102. 4.23e-39
```

The two rows agree.

`do_n_sims()` repeats the simulation `n_sims` times, with a different random seed for each dataset:

```r
do_n_sims <- function(n_sims = 1000, ...) {
  lapply(seq_len(n_sims), function(i) {
    set.seed(i)
    do_one_sim(...) |>
      dplyr::mutate(sim_number = i, .before = dplyr::everything())
  }) |>
    dplyr::bind_rows()
}

sim_results <- do_n_sims(
  n_sims = 1000,
  n = 100, mu = mu_hat, mu0 = 0.9 * mu_hat, sigma2 = sigma_sq_hat
```

```
)
sim_results
#> # A tibble: 1,000 x 10
#>    sim_number mu_hat sigma_hat se_hat confint_left confint_right tstat     pval
#>         <int>  <dbl>     <dbl>  <dbl>        <dbl>         <dbl> <dbl>    <dbl>
#>  1          1   99.8      9.19  0.919         98.0          102.  11.9  6.73e-21
#>  2          2   98.3     11.9   1.19          96.0          101.   8.04 1.93e-12
#>  3          3   98.8      8.76  0.876         97.0          101.  11.4  1.05e-19
#>  4          4   99.6      9.35  0.935         97.8          102.  11.6  3.61e-20
#>  5          5   99.0      9.67  0.967         97.1          101.  10.5  7.56e-18
#>  6          6   98.6     10.6   1.06          96.5          101.   9.23 5.26e-15
#>  7          7  100.       9.81  0.981         98.1          102.  11.5  6.05e-20
#>  8          8   97.7     11.0   1.10          95.5           99.9  8.07 1.68e-12
#>  9          9   98.1      9.81  0.981         96.2          100.   9.50 1.38e-15
#> 10         10   97.3      9.63  0.963         95.4           99.2  8.79 4.71e-14
#> # i 990 more rows
#> # i 2 more variables: confint_covers <lgl>, test_rejects <lgl>
```

`summarize_sim()` compares the simulation results with the true data-generating parameters:

```
summarize_sim <- function(sim_results, mu, sigma2, n) {
  true_se <- sqrt(sigma2 / n)
  tibble::tibble(
    bias_mu_hat = mean(sim_results$mu_hat) - mu,
    sd_mu_hat = sd(sim_results$mu_hat),
    true_se = true_se,
    bias_se_hat = mean(sim_results$se_hat) - true_se,
    coverage = mean(sim_results$confint_covers),
    power = mean(sim_results$test_rejects)
  )
}

sim_summary <- summarize_sim(
  sim_results,
  mu = mu_hat, sigma2 = sigma_sq_hat, n = 100
)
sim_summary
#> # A tibble: 1 x 6
#>   bias_mu_hat sd_mu_hat true_se bias_se_hat coverage power
#>         <dbl>     <dbl>   <dbl>       <dbl>    <dbl> <dbl>
#> 1    -0.00499     0.996    1.02    -0.00113    0.959     1
```

Across 1000 simulated datasets:

- the average error of $\hat{\mu}$ is -0.005 mg/dL, small relative to its standard error of 1.02 mg/dL, consistent with $\hat{\mu}$ being unbiased;
- the standard deviation of the $\hat{\mu}$ values, 0.996, is close to the true standard error;
- the 95% confidence intervals covered the true μ in 95.9% of datasets, close to their nominal 95%;
- the test of $H_0 : \mu = 0.9\,\hat{\mu}$ rejected in 100% of datasets: with $n = 100$, a 10% difference in the mean is easy to detect.

Changing the sample size, the true μ, or σ^2 in `do_n_sims()` shows how these properties depend on them.

Practice exercises

Exercise 0.25 (Binomial likelihood (adapted from Dobson and Barnett (2018), Chapter 3))**.** Let $Y \sim \text{Binomial}(n, \pi)$, so that

$$\text{p}(Y = y \mid \pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y \in \{0, 1, \dots, n\}.$$

Assume $0 < y < n$ so the MLE lies in the interior of $(0, 1)$.
(a) Write the log-likelihood $\ell(\pi; y)$ for a single observation y.
(b) Derive the score function $\ell'(\pi; y) \stackrel{\text{def}}{=} \frac{\partial}{\partial \pi} \ell(\pi; y)$.
(c) Set the score equal to zero and solve for $\hat{\pi}_{ML}$. Confirm that $\hat{\pi}_{ML} = y/n$.
(d) Compute the second derivative $\ell''(\pi; y)$ and verify that it is negative, confirming a maximum.

Solution

Solution. **(a)**

$$\begin{aligned} \ell(\pi; y) &= \log\left\{ \binom{n}{y} \pi^y (1-\pi)^{n-y} \right\} \\ &= \log\left\{ \binom{n}{y} \right\} + y \log\{\pi\} + (n-y) \log\{1-\pi\} \end{aligned}$$

The first term does not depend on π, so it drops out of every derivative with respect to π.
(b)

$$\begin{aligned}\ell'(\pi; y) &= \frac{\partial}{\partial \pi}\left[\log\left\{\binom{n}{y}\right\} + y\log\{\pi\} + (n-y)\log\{1-\pi\}\right] \\ &= \frac{y}{\pi} - \frac{n-y}{1-\pi} \\ &= \frac{y(1-\pi) - (n-y)\pi}{\pi(1-\pi)} \\ &= \frac{y - n\pi}{\pi(1-\pi)}\end{aligned}$$

(c)
Setting $\ell'(\pi; y) = 0$:

$$\begin{aligned}0 &= \frac{y - n\pi}{\pi(1-\pi)} \\ y &= n\pi \\ \hat{\pi}_{ML} &= \frac{y}{n}\end{aligned}$$

(d)

$$\begin{aligned}\ell''(\pi; y) &= \frac{\partial}{\partial \pi}\left[\frac{y}{\pi} - \frac{n-y}{1-\pi}\right] \\ &= -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2}\end{aligned}$$

Since $y \geq 0$, $n-y \geq 0$, and y and $n-y$ are not both zero, $\ell''(\pi; y) < 0$ for every $\pi \in (0, 1)$. So ℓ is strictly concave, and its critical point $\hat{\pi}_{ML} = y/n$ is its global maximum.

Exercise 0.26 (Gaussian log-likelihood (adapted from Kleinbaum et al. (2014), Chapter 5))**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \mathrm{N}(\mu, \sigma^2)$.
(a) Write the likelihood $\mathcal{L}(\mu, \sigma^2; \tilde{x})$ for the observed data $\tilde{x} = (x_1, \ldots, x_n)$.
(b) Write the log-likelihood $\ell(\mu, \sigma^2; \tilde{x})$. Show that it can be written as

$$\ell(\mu, \sigma^2; \tilde{x}) = -\frac{n}{2}\log\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

(c) Derive the MLE $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$.

Solution

Solution. **(a)**

$$\mathcal{L}(\mu, \sigma^2; \tilde{x}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

(b)

$$\begin{aligned} \ell(\mu, \sigma^2; \tilde{x}) &= \log\{\mathcal{L}(\mu, \sigma^2; \tilde{x})\} \\ &= -\frac{n}{2} \log\{2\pi\sigma^2\} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

(c)
Deriving $\hat{\mu}_{ML}$:

$$\begin{aligned} \frac{\partial}{\partial \mu} \ell &= \frac{\partial}{\partial \mu} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ &= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right) \end{aligned}$$

Setting this to zero: $\hat{\mu}_{ML} = \bar{x} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i$.
Deriving $\hat{\sigma}^2_{ML}$:

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

Setting this to zero and solving:

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

Substituting $\hat{\mu}_{ML} = \bar{x}$:

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

Note: this maximum likelihood estimator is a biased estimator of σ^2; the unbiased sample variance divides by $n - 1$.

Exercise 0.27 (Score at the MLE (adapted from Dobson and Barnett (2018), Chapter 3))**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$.
(a) Write the log-likelihood $\ell(\lambda; \tilde{x})$.
(b) Derive the score function $\ell'(\lambda; \tilde{x})$.
(c) Show that $\hat{\lambda}_{ML} = \bar{x}$, and verify that the score equals zero at the MLE.
(d) Provide an intuitive interpretation: why does the score being zero at $\hat{\lambda}_{ML}$ make sense?

Solution

Solution. **(a)**

$$\begin{aligned}
\ell(\lambda; \tilde{x}) &= \log\left\{\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right\} \\
&= \sum_{i=1}^{n} (x_i \log\{\lambda\} - \lambda - \log\{x_i!\}) \\
&= \left(\sum_{i=1}^{n} x_i\right) \log\{\lambda\} - n\lambda - \sum_{i=1}^{n} \log\{x_i!\}
\end{aligned}$$

(b)

$$\begin{aligned}
\ell'(\lambda; \tilde{x}) &= \frac{\partial}{\partial \lambda}\left[\left(\sum_{i=1}^{n} x_i\right) \log\{\lambda\} - n\lambda\right] \\
&= \frac{\sum_{i=1}^{n} x_i}{\lambda} - n \\
&= \frac{n\bar{x}}{\lambda} - n
\end{aligned}$$

(c)
Setting $\ell'(\lambda; \tilde{x}) = 0$:

$$\begin{aligned}
\frac{n\bar{x}}{\lambda} &= n \\
\hat{\lambda}_{ML} &= \bar{x}
\end{aligned}$$

If $\bar{x} > 0$ (equivalently, at least one $x_i > 0$), then

$$\begin{aligned}
\ell'(\bar{x}; \tilde{x}) &= \frac{n\bar{x}}{\bar{x}} - n \\
&= n - n \\
&= 0.
\end{aligned}$$

If instead all $x_i = 0$, then $\bar{x} = 0$ and $\hat{\lambda}_{ML} = 0$ is a boundary value. In that case, the score formula $\ell'(\lambda; \tilde{x}) = \frac{n\bar{x}}{\lambda} - n$ is not defined at $\lambda = 0$, so the usual interior verification $\ell'(\hat{\lambda}_{ML}; \tilde{x}) = 0$ does not apply.

(d)

The score measures the rate of change of the log-likelihood. When it equals zero, increasing or decreasing λ slightly would not improve the fit; we are at a "flat" point. Intuitively, $\hat{\lambda}_{ML} = \bar{x}$ is the value of λ that makes the expected count per observation (λ) exactly equal to the observed average count ($\bar{x}$).

Exercise 0.28 (Standard error of an MLE (adapted from Dobson and Barnett (2018), Chapter 5))**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$.

(a) Derive the Hessian $\ell''(\lambda; \tilde{x}) = \frac{\partial}{\partial[}2]\lambda\ell(\lambda; \tilde{x})$.

(b) Derive the observed information $I(\lambda; \tilde{x}) = -\ell''(\lambda; \tilde{x})$.

(c) Evaluate $I(\hat{\lambda}_{ML}; \tilde{x})$.

(d) Give an approximate 95% confidence interval for λ using the asymptotic normal distribution of the MLE.

Solution

Solution. **(a)**

From Exercise 0.27, $\ell'(\lambda; \tilde{x}) = \frac{n\bar{x}}{\lambda} - n$.

$$\begin{aligned}\ell''(\lambda; \tilde{x}) &= \frac{\partial}{\partial\lambda}\left[\frac{n\bar{x}}{\lambda} - n\right] \\ &= -\frac{n\bar{x}}{\lambda^2}\end{aligned}$$

(b)

$$I(\lambda; \tilde{x}) = -\ell''(\lambda; \tilde{x}) = \frac{n\bar{x}}{\lambda^2}$$

(c)

Assuming $\bar{x} > 0$ (i.e., at least one $x_i > 0$), substituting $\hat{\lambda}_{ML} = \bar{x}$:

$$I(\hat{\lambda}_{ML}; \tilde{x}) = \frac{n\bar{x}}{\bar{x}^2} = \frac{n}{\bar{x}}$$

(d)

By the asymptotic theory of MLEs:

$$\hat{\lambda}_{ML} \dot{\sim} \text{N}\big(\lambda,\ (\mathcal{J}(\lambda))^{-1}\big)$$

We estimate this asymptotic variance using the observed information evaluated at the MLE. Since $I(\hat{\lambda}_{ML}; \tilde{x})^{-1} = \bar{x}/n$:

$$\mathrm{SE}\big(\hat{\lambda}_{ML}\big) \approx \sqrt{\frac{\bar{x}}{n}}$$

An approximate 95% CI for λ is:

$$\hat{\lambda}_{ML} \pm 1.96 \times \sqrt{\frac{\bar{x}}{n}} = \bar{x} \pm 1.96\sqrt{\frac{\bar{x}}{n}}$$

Exercise 0.29 (Exponential MLE (adapted from Dobson and Barnett (2018), Chapter 3))**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Exponential}(\mu)$, so that

$$\mathrm{p}(X = x \mid \mu) = \frac{1}{\mu} e^{-x/\mu}, \quad x > 0, \quad \mu > 0.$$

(a) Write the log-likelihood $\ell(\mu; \tilde{x})$ for the observed data $\tilde{x} = (x_1, \ldots, x_n)$.
(b) Derive the score function $\ell'(\mu; \tilde{x}) = \frac{\partial}{\partial \mu} \ell(\mu; \tilde{x})$.
(c) Set the score equal to zero and show that $\hat{\mu}_{ML} = \bar{x}$. Compute the second derivative and verify this critical point is a maximum.
(d) Derive the observed information $I(\mu; \tilde{x}) = -\ell''(\mu; \tilde{x})$, evaluate it at $\hat{\mu}_{ML}$, and give an approximate 95% confidence interval for μ.

Solution

Solution. **(a)**

$$\begin{aligned}
\ell(\mu; \tilde{x}) &= \log\left\{ \prod_{i=1}^{n} \frac{1}{\mu} e^{-x_i/\mu} \right\} \\
&= \sum_{i=1}^{n} \log\left\{ \frac{1}{\mu} e^{-x_i/\mu} \right\} \\
&= \sum_{i=1}^{n} \Big(-\log\{\mu\} - \frac{x_i}{\mu} \Big) \\
&= -n \log\{\mu\} - \frac{1}{\mu} \sum_{i=1}^{n} x_i \\
&= -n \log\{\mu\} - \frac{n\bar{x}}{\mu}
\end{aligned}$$

(b)

$$\begin{aligned}\ell'(\mu;\tilde{x}) &= \frac{\partial}{\partial\mu}\left(-n\log\{\mu\} - \frac{n\bar{x}}{\mu}\right)\\ &= -\frac{n}{\mu} + \frac{n\bar{x}}{\mu^2}\\ &= \frac{n}{\mu^2}(\bar{x}-\mu)\end{aligned}$$

(c)
Setting $\ell'(\mu;\tilde{x}) = 0$:

$$\begin{aligned}0 &= \frac{n}{\mu^2}(\bar{x}-\mu)\\ \hat{\mu}_{ML} &= \bar{x}\end{aligned}$$

The second derivative is:

$$\begin{aligned}\ell''(\mu;\tilde{x}) &= \frac{\partial}{\partial\mu}\left(-\frac{n}{\mu} + \frac{n\bar{x}}{\mu^2}\right)\\ &= \frac{n}{\mu^2} - \frac{2n\bar{x}}{\mu^3}\end{aligned}$$

Evaluated at $\hat{\mu}_{ML} = \bar{x}$:

$$\ell''(\bar{x};\tilde{x}) = \frac{n}{\bar{x}^2} - \frac{2n\bar{x}}{\bar{x}^3} = \frac{n}{\bar{x}^2} - \frac{2n}{\bar{x}^2} = -\frac{n}{\bar{x}^2} < 0$$

Since the second derivative is negative, $\hat{\mu}_{ML} = \bar{x}$ is a maximum.
(d)
The observed information is:

$$I(\mu;\tilde{x}) = -\ell''(\mu;\tilde{x}) = \frac{2n\bar{x}}{\mu^3} - \frac{n}{\mu^2}$$

Evaluated at $\hat{\mu}_{ML} = \bar{x}$:

$$I(\hat{\mu}_{ML};\tilde{x}) = \frac{2n\bar{x}}{\bar{x}^3} - \frac{n}{\bar{x}^2} = \frac{2n}{\bar{x}^2} - \frac{n}{\bar{x}^2} = \frac{n}{\bar{x}^2}$$

So $\text{SE}(\hat{\mu}_{ML}) \approx \sqrt{I(\hat{\mu}_{ML};\tilde{x})^{-1}} = \frac{\bar{x}}{\sqrt{n}}$.
An approximate 95% CI for μ is:

$$\hat{\mu}_{ML} \pm 1.96 \times \frac{\bar{x}}{\sqrt{n}} = \bar{x} \pm \frac{1.96\bar{x}}{\sqrt{n}}$$

Note: the exponential distribution has $\text{Var}(X) = \mu^2$, so $\text{SE}(\hat{\mu}_{ML}) = \mu/\sqrt{n}$, which is estimated by $\bar{x}/\sqrt{n}$. More generally, the standard error of a sample mean is $\text{SD}(X)/\sqrt{n}$; here that reduces to $\mu/\sqrt{n}$ because $\text{SD}(X) = \mu$ for the exponential distribution.

References

Casella, George, and Roger Berger. 2002. *Statistical Inference*. 2nd ed. Cengage Learning. https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/.

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. https://doi.org/10.1201/9781315182780.

Dunn, Peter K, and Gordon K Smyth. 2018. *Generalized Linear Models with Examples in R*. Vol. 53. Springer. https://doi.org/10.1007/978-1-4419-0118-7.

Efron, Bradley, and David V Hinkley. 1978. "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information." *Biometrika* 65 (3): 457–83. https://doi.org/10.1093/biomet/65.3.457.

Hoenig, John M., and Dennis M. Heisey. 2001. "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *The American Statistician* 55 (1): 19–24. https://doi.org/10.1198/000313001300339897.

Hogg, Robert V., Elliot A. Tanis, and Dale L. Zimmerman. 2019. *Probability and Statistical Inference*. Tenth edition. Pearson.

Hulley, Stephen, Deborah Grady, Trudy Bush, et al. 1998. "Randomized Trial of Estrogen Plus Progestin for Secondary Prevention of Coronary Heart Disease in Postmenopausal Women." *JAMA : The Journal of the American Medical Association* (Chicago, IL) 280 (7): 605–13. https://doi.org/10.1001/jama.280.7.605.

Kleinbaum, David G, Lawrence L Kupper, Azhar Nizam, K Muller, and ES Rosenberg. 2014. *Applied Regression Analysis and Other Multivariable Methods*. 5th ed. Cengage Learning. https://www.cengage.com/c/applied-regression-analysis-and-other-multivariable-methods-5e-kleinbaum/9781285051086/.

Lehmann, E. L. 1999. *Elements of Large-Sample Theory*. Springer Texts in Statistics. Springer. https://doi.org/10.1007/b98855.

McLachlan, Geoffrey J, and Thriyambakam Krishnan. 2007. *The EM Algorithm and Extensions*. 2nd ed. John Wiley & Sons. https://doi.org/10.1002/9780470191613.

Newey, Whitney K, and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*, edited by Robert Engle and Dan McFadden, vol. 4. Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4.

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1353-0.

Wilks, Samuel S. 1938. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics* 9 (1): 60–62. https://doi.org/10.1214/aoms/1177732360.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with r*. 2nd ed. Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279.