Introduction to Maximum Likelihood Inference

Last modified: 2026-11-12 22:02:32 (PST)

Overview of maximum likelihood estimation

These notes are derived primarily from [1, Ch. 1–5], with some material from [2] and [3].

The likelihood function

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

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

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 is a Bernoulli variable, $X \sim \text{Ber}(\pi)$, then

$$\mathbb{p}_\pi(X = x) = \pi^x(1 - \pi)^{1-x}$$

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 $\underset{\sim}{x} \stackrel{\text{def}}{=} x_1, \ldots, x_n$ be a dataset with corresponding random vector $\underset{\sim}{X}$, and let $\mathbb{p}_\Theta(\underset{\sim}{X} = \underset{\sim}{x})$ be a probability model for the distribution of $\underset{\sim}{X}$, with unknown parameter vector $\underset{\sim}{\Theta}$. The **likelihood** of the parameter value θ, for

model $\mathbb{p}_\Theta$ and data $\underset{\sim}{X} = \underset{\sim}{x}$, is the *joint probability* (or joint density) of $\underset{\sim}{X} = \underset{\sim}{x}$ when $\Theta = \theta$:

$$\begin{aligned}\mathcal{L}(\theta) &\overset{\text{def}}{=} \mathbb{p}(\underset{\sim}{X} = \underset{\sim}{x} \mid \Theta = \theta) && \text{(definition of the likelihood)}\\ &= \mathbb{p}(X_1 = x_1, \dots, X_n = x_n \mid \Theta = \theta) && \text{(write } \underset{\sim}{X} = \underset{\sim}{x} \text{ componentwise)}\end{aligned}$$

Source: Wikipedia contributors [4].

i Notation for the likelihood function

These notes mostly write $\mathcal{L}(\theta)$, leaving the data implicit. The notation choices page explains why, and lists the other ways sources write the likelihood function.

Exercise 0.1 (Likelihood of an independent sample)**.** For mutually independent data $X_1, \dots, X_n$, write the likelihood $\mathcal{L}(\theta)$ in terms of the distributions of the individual X_i.

Solution

Solution 0.1.

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

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

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

i Proof

Proof. This is the solution to Exercise 0.1. □

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

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

Exercise 0.2 (Dataset likelihood from likelihood components)**.** For mutually independent data $\underset{\sim}{x} \stackrel{\text{def}}{=} x_1, \ldots, x_n$, use Theorem 0.1 to write the likelihood of the dataset in terms of the observations' likelihood components.

Solution

Solution 0.2.

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^{n} \mathbb{p}(X_i = x_i \mid \theta) && \text{(likelihood of an independent sample)} \\ &= \prod_{i=1}^{n} \mathcal{L}_i(\theta) && \text{(definition of likelihood components)} \end{aligned}$$

Theorem 0.2 (Dataset likelihood as a product of observation likelihoods)**.** *For mutually independent data* $\underset{\sim}{x} \stackrel{def}{=} x_1, \ldots, x_n$*, 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. This is the solution to Exercise 0.2. □

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:

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

Source: Wikipedia contributors [5].

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: $\hat{\pi}_{\text{ML}} = 1$.

Watch

5MinuteAI explains maximum likelihood estimation and works examples in five minutes each, following the University of Washington's CSE 312 [6].

Read

- Hogg et al. [7, Sec. 6.4], pp. 264–276 covers maximum likelihood and method of moments estimation.
- Casella and Berger [3, Sec. 7.2.2], pp. 315–323 covers maximum likelihood estimators.
- Rice [8, Sec. 8.5], pp. 267–284 covers the method of maximum likelihood.

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 the derivative of f is zero, or, for a function of a vector, a point $\underset{\sim}{x}_0$ where the gradient f' of f is the zero vector:

$$f'(x_0) = 0 \quad \text{(scalar input)}$$
$$f'(\underset{\sim}{x}_0) = \underset{\sim}{0} \quad \text{(vector input)}$$

Source: Wikipedia contributors [9].

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:

$$\begin{aligned} \mathcal{L}'(\theta) &= \frac{\partial}{\partial\theta}\mathcal{L}(\theta) && \text{(notation for the derivative)} \\ &= \frac{\partial}{\partial\theta}\prod_{i=1}^{n}\mathbb{p}(X_i = x_i \mid \theta) && \text{(likelihood of mutually independent data)} \end{aligned} \tag{3}$$

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 $\mathbb{p}_\Theta(\underset{\sim}{X})$ and data $\underset{\sim}{X} = \underset{\sim}{x}$, is the natural logarithm of the likelihood:

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

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

Exercise 0.3 (Comparing likelihoods through log-likelihoods)**.** Suppose $\mathcal{L}(\theta) > 0$ for every θ. Determine how the parameter values that maximize the likelihood $\mathcal{L}$ relate to the parameter values that maximize the log-likelihood ℓ.

Solution

Solution 0.3. Let θ^* be a parameter value. 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. So a parameter value θ^* maximizes $\mathcal{L}$ if and only if it maximizes ℓ.

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. This is the solution to Exercise 0.3. □

Exercise 0.4 (Log-likelihood of an independent sample)**.** For mutually independent data $X_1, \dots, X_n$, use Theorem 0.1 to write the log-likelihood $\ell(\theta)$ as a sum over the observations. What does each term become if the X_i also share a common distribution $\mathbb{p}(X = x \mid \theta)$?

Solution

Solution 0.4.

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

With a common distribution,

$$\mathbb{p}(X_i = x_i \mid \theta) = \mathbb{p}(X = x_i \mid \theta),$$

so each term is $\log \mathbb{p}(X = x_i \mid \theta)$.

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

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

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

Proof

Proof. This is the solution to Exercise 0.4. □

Exercise 0.5 (Derivative of the log-likelihood for iid data)**.** For iid data, use Theorem 0.4 to write the derivative $\ell'(\theta)$ of the log-likelihood as a sum over the observations.

Solution

Solution 0.5.

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

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

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

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

Proof

Proof. This is the solution to Exercise 0.5. □

The score function

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

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

Source: Wikipedia contributors [10].

Remark (Notation for the score function)

Remark 0.1 (Notation for the score function). We often omit the arguments $\underset{\sim}{x}$ and θ, writing $\ell' \overset{\text{def}}{=} \ell'(\underset{\sim}{x} \mid \theta) \overset{\text{def}}{=} \ell'(\theta)$. Some sources write U or S for the score function instead of ℓ'; for example, Dobson and Barnett [1] write U. The notation choices page explains why these notes use ℓ'.

Remark (Why the score function is called "score")

Remark 0.2 (Why the score function is called "score"). Statisticians call the score function by that name because of its origin in genetics. Fisher [11] gave each family in a study of linkage a "score", which summarized what the family showed about the linkage parameter. He compared that scoring rule with an "ideal score", the derivative of the log of the probability of the data as the linkage parameter varies, which is the score function of these notes [10]. For independent observations, the log-likelihood is a sum with one contribution from each of them, so its derivative is also a sum. Rao [12], who had worked on estimating linkage parameters in Fisher's genetics laboratory, based a test on the score function, with one score for each parameter, and called it a score test [13]. The name then spread to the derivative itself.
In other settings, "score" refers to a different quantity. These notes call a classifier's raw output the log-odds, never a "score", so that "score" always refers to the derivative of the log-likelihood.

Exercise 0.6 (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 0.6. The log-likelihood of one Poisson observation is $\ell = x\log\{\lambda\} - \lambda - \log\{x!\}$, so:

$$
\begin{aligned}
\ell' &\overset{\text{def}}{=} \frac{\partial}{\partial\lambda}\ell && \text{(definition of the score)} \\
&= \frac{\partial}{\partial\lambda}(x\log\{\lambda\} - \lambda - \log\{x!\}) && \text{(Poisson log-likelihood)} \\
&= \frac{\partial}{\partial\lambda}(x\log\{\lambda\}) - \frac{\partial}{\partial\lambda}\lambda - \frac{\partial}{\partial\lambda}\log\{x!\} && \text{(linearity of differentiation)} \\
&= \frac{\partial}{\partial\lambda}(x\log\{\lambda\}) - \frac{\partial}{\partial\lambda}\lambda - 0 && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial\lambda}(x\log\{\lambda\}) - \frac{\partial}{\partial\lambda}\lambda && \text{(drop the zero term)} \\
&= x\frac{\partial}{\partial\lambda}\log\{\lambda\} - \frac{\partial}{\partial\lambda}\lambda && \text{(constant multiple rule)} \\
&= x\lambda^{-1} - \frac{\partial}{\partial\lambda}\lambda && \text{(derivative of } \log\lambda\text{)} \\
&= x\lambda^{-1} - 1 && \text{(derivative of } \lambda \text{ with respect to itself)} \\
&= \frac{x}{\lambda} - 1 && \text{(write the negative power as a fraction)} \\
&= \frac{x}{\lambda} - \frac{\lambda}{\lambda} && \text{(write both terms over the common denominator)} \\
&= \frac{x-\lambda}{\lambda} && \text{(combine the fractions)} \\
&= \frac{x - \mathbb{E}[X]}{\lambda} && (\mathbb{E}[X] = \lambda) \\
&= \frac{x - \mathbb{E}[X]}{\mathrm{Var}(X)} && (\mathrm{Var}(X) = \lambda)
\end{aligned}
$$

Theorem 0.6 (Score function of a Poisson variable)**.** *Let* $X \sim \text{Pois}(\lambda)$*, with mean* λ *and observed count* x*. The score* ℓ' *of the single observation* x *is the derivative of its log-likelihood* ℓ *with respect to* λ*. It equals the deviation of* x *from the mean* $\mathbb{E}[X]$*, divided by the variance* $\mathrm{Var}(X)$*:*

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

Proof

Proof. This is the solution to Exercise 0.6, worked in Solution 0.6. □

Exercise 0.7 (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 0.7.

$$
\begin{aligned}
\ell' &\stackrel{\text{def}}{=} \frac{\partial}{\partial \mu}\ell && \text{(definition of the score)} \\
&= \frac{\partial}{\partial \mu}\left(\frac{-1}{2}\left(\log\{2\pi\sigma^2\} + \frac{(x-\mu)^2}{\sigma^2}\right)\right) && \text{(Gaussian log-likelihood)} \\
&= \frac{-1}{2}\frac{\partial}{\partial \mu}\left(\log\{2\pi\sigma^2\} + \frac{(x-\mu)^2}{\sigma^2}\right) && \text{(constant multiple rule)} \\
&= \frac{-1}{2}\left(\frac{\partial}{\partial \mu}\log\{2\pi\sigma^2\} + \frac{\partial}{\partial \mu}\frac{(x-\mu)^2}{\sigma^2}\right) && \text{(linearity of differentiation)} \\
&= \frac{-1}{2}\left(0 + \frac{\partial}{\partial \mu}\frac{(x-\mu)^2}{\sigma^2}\right) && \text{(derivative of a constant)} \\
&= \frac{-1}{2}\left(0 + \frac{1}{\sigma^2}\frac{\partial}{\partial \mu}(x-\mu)^2\right) && \text{(constant multiple rule)} \\
&= \frac{-1}{2}\left(0 + \frac{1}{\sigma^2}\cdot 2(x-\mu)\frac{\partial}{\partial \mu}(x-\mu)\right) && \text{(chain rule, outer function } u^2\text{)}
\end{aligned}
$$

The inner derivative is:

$$
\begin{aligned}
\frac{\partial}{\partial \mu}(x-\mu) &= \frac{\partial}{\partial \mu}x - \frac{\partial}{\partial \mu}\mu && \text{(linearity of differentiation)} \\
&= 0 - \frac{\partial}{\partial \mu}\mu && (x \text{ does not depend on } \mu) \\
&= 0 - 1 && \text{(derivative of } \mu \text{ with respect to itself)} \\
&= -1 && \text{(subtract)}
\end{aligned}
$$

Substituting the inner derivative back in:

$$
\begin{aligned}
\ell' &= \frac{-1}{2}\left(0 + \frac{1}{\sigma^2} \cdot 2(x-\mu) \cdot (-1)\right) && \text{(inner derivative is } -1\text{)} \\
&= \frac{-1}{2}\left(0 + \frac{-2(x-\mu)}{\sigma^2}\right) && \text{(multiply)} \\
&= \frac{-1}{2} \cdot \frac{-2(x-\mu)}{\sigma^2} && \text{(drop the zero term)} \\
&= \frac{x-\mu}{\sigma^2} && \text{(multiply)} \\
&= \frac{x-\mathbb{E}[X]}{\sigma^2} && (\mathbb{E}[X] = \mu) \\
&= \frac{x-\mathbb{E}[X]}{\mathrm{Var}(X)} && (\mathrm{Var}(X) = \sigma^2)
\end{aligned}
$$

Theorem 0.7 (Score function of a Gaussian variable)**.** *Let $X \sim \mathrm{N}(\mu, \sigma^2)$, with mean μ, known variance σ^2, and observed value x. The score ℓ' of the single observation x is the derivative of its log-likelihood ℓ with respect to μ. It equals the deviation of x from the mean $\mathbb{E}[X]$, divided by the variance $\mathrm{Var}(X)$:*

$$\ell' = \frac{x-\mathbb{E}[X]}{\mathrm{Var}(X)} = \frac{x-\mu}{\sigma^2}$$

Proof

Proof. This is the solution to Exercise 0.7, worked in Solution 0.7. □

Exercise 0.8 (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 0.8. The exponential density with mean μ is

$$\mathbb{p}(X = x) = \mu^{-1} e^{-x/\mu}$$

for $x > 0$, so $\ell = -\log\{\mu\} - \frac{x}{\mu}$, and:

$$
\begin{aligned}
\ell' &\stackrel{\text{def}}{=} \frac{\partial}{\partial \mu}\ell && \text{(definition of the score)} \\
&= \frac{\partial}{\partial \mu}\left(-\log\{\mu\} - \frac{x}{\mu}\right) && \text{(exponential log-likelihood)} \\
&= \frac{\partial}{\partial \mu}(-\log\{\mu\} - x\mu^{-1}) && \text{(write } \frac{x}{\mu} \text{ as } x\mu^{-1}\text{)} \\
&= \frac{\partial}{\partial \mu}(-\log\{\mu\}) - \frac{\partial}{\partial \mu}(x\mu^{-1}) && \text{(linearity of differentiation)} \\
&= -\frac{\partial}{\partial \mu}\log\{\mu\} - x\frac{\partial}{\partial \mu}\mu^{-1} && \text{(constant multiple rule)} \\
&= -\mu^{-1} - x\frac{\partial}{\partial \mu}\mu^{-1} && \text{(derivative of } \log \mu\text{)} \\
&= -\mu^{-1} - x(-\mu^{-2}) && \text{(power rule)} \\
&= -\mu^{-1} + x\mu^{-2} && \text{(multiply)} \\
&= -\frac{1}{\mu} + \frac{x}{\mu^2} && \text{(write the negative powers as fractions)} \\
&= -\frac{\mu}{\mu^2} + \frac{x}{\mu^2} && \text{(write both terms over the common denominator)} \\
&= \frac{-\mu + x}{\mu^2} && \text{(combine the fractions)} \\
&= \frac{x - \mu}{\mu^2} && \text{(reorder the numerator)} \\
&= \frac{x - \mathbb{E}[X]}{\mu^2} && (\mathbb{E}[X] = \mu) \\
&= \frac{x - \mathbb{E}[X]}{\mathrm{Var}(X)} && (\mathrm{Var}(X) = \mu^2)
\end{aligned}
$$

Theorem 0.8 (Score function of an exponential variable)**.** *Let X be an exponential random variable with mean μ and observed value $x > 0$. The score ℓ' of the single observation x is the derivative of its log-likelihood ℓ with respect to μ. It equals the deviation of x from the mean $\mathbb{E}[X]$, divided by the variance $\mathrm{Var}(X)$:*

$$\ell' = \frac{x - \mathbb{E}[X]}{\mathrm{Var}(X)} = \frac{x - \mu}{\mu^2}$$

Proof

Proof. This is the solution to Exercise 0.8, worked in Solution 0.8. □

Definition 0.8 (Score equation)**.** The **score equation** is the equation that sets the score function to zero:

$$\ell'(\underset{\sim}{\theta}) = \underset{\sim}{0}$$

The score equation is one example of an **estimating equation**, an equation whose solution defines an estimator. Estimating equations in general need not come from a likelihood.

Information matrices

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

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

Source: Wikipedia contributors [14].

The Hessian is named after the mathematician Otto Hesse.

Exercise 0.9 (Entries of the Hessian)**.** If $\underset{\sim}{\theta}$ is a $p \times 1$ vector, find the dimensions of the Hessian $\frac{\partial}{\partial \underset{\sim}{\theta}} \frac{\partial}{\partial \underset{\sim}{\theta}^\top} \ell$ and an expression for its ijth entry.

Solution

Solution 0.9. $\frac{\partial}{\partial \underset{\sim}{\theta}^\top} \ell$ 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 $\underset{\sim}{\theta}$ gives a $p \times 1$ column of derivatives per entry, so $\frac{\partial}{\partial \underset{\sim}{\theta}} \frac{\partial}{\partial \underset{\sim}{\theta}^\top} \ell$ is $p \times p$, with ijth entry $\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \ell$.

Theorem 0.9 (Elements of the Hessian matrix)**.** *If $\underset{\sim}{\theta}$ is a $p \times 1$ vector, then the Hessian is a $p \times p$ matrix, whose ijth entry is:*

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

Proof

Proof. This is the solution to Exercise 0.9. □

Exercise 0.10 (Hessian from the score)**.** Express the Hessian ℓ'' as a derivative of the score function ℓ'.

Solution

Solution 0.10. By Definition 0.7, $\ell' = \frac{\partial}{\partial \underset{\sim}{\theta}}\ell$, so $(\ell')^\top = \frac{\partial}{\partial \underset{\sim}{\theta}^\top}\ell$. Then:

$$\begin{aligned}\frac{\partial}{\partial \underset{\sim}{\theta}}(\ell')^\top &= \frac{\partial}{\partial \underset{\sim}{\theta}}\frac{\partial}{\partial \underset{\sim}{\theta}^\top}\ell && \text{(transpose of the score)}\\ &= \ell'' && \text{(definition of the Hessian)}\end{aligned}$$

Theorem 0.10 (Hessian is the derivative of the transposed score)**.**

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

Proof

Proof. This is the solution to Exercise 0.10. □

Definition 0.10 (Expected information matrix (Fisher information matrix, information matrix))**.** 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 negative Hessian of the log-likelihood, with the data $\underset{\sim}{X}$ treated as random:

$$\mathcal{I}(\underset{\sim}{\theta}) \stackrel{\text{def}}{=} \mathbb{E}[-\ell''(\underset{\sim}{X} \mid \underset{\sim}{\theta})] \tag{10}$$

Source: Wikipedia contributors [15].

Definition 0.11 (Observed information matrix)**.** The **observed information matrix**, written I, is the negative of the Hessian of the log-likelihood, evaluated at the observed data:

$$I \stackrel{\text{def}}{=} -\ell''(\underset{\sim}{x}|\underset{\sim}{\theta}) \tag{11}$$

Remark (The expected information is a function of the parameter alone)

Remark 0.3 (The expected information is a function of the parameter alone). The expected information is a function of the parameter alone, so it is the quantity that the observed information estimates from one data set.

Remark (The observed information is the sample counterpart of the expected information)

Remark 0.4 (The observed information is the sample counterpart of the expected information). The observed information matrix is the sample counterpart of the expected information, which averages the same quantity over possible data sets.

Example 0.4 (Information for Poisson data)**.** For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$, one observation's score is $x/\lambda - 1$ (Theorem 0.6), 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:

$$
\begin{aligned}
I(\lambda) &= -\ell'' && \text{(observed information: negative Hessian)} \\
&= -\left(-\frac{\sum_{i=1}^n x_i}{\lambda^2}\right) && \text{(substitute the Hessian)} \\
&= \frac{\sum_{i=1}^n x_i}{\lambda^2} && \text{(two negatives make a positive)} \\
\mathcal{I}(\lambda) &= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{\lambda^2}\right] && \text{(expected information)} \\
&= \frac{\mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\lambda^2} && \text{(factor the constant } \tfrac{1}{\lambda^2} \text{ out of the expectation)} \\
&= \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{\lambda^2} && \text{(linearity of expectation)} \\
&= \frac{\sum_{i=1}^n \lambda}{\lambda^2} && (\mathbb{E}[X_i] = \lambda) \\
&= \frac{n\lambda}{\lambda^2} && \text{(sum of } n \text{ copies of } \lambda) \\
&= \frac{n}{\lambda} && \text{(cancel one factor of } \lambda)
\end{aligned}
$$

Exercise 0.11 (Derivative of a density through its logarithm)**.** Let $\mathbb{p}(\underset{\sim}{x} \mid \theta)$ be a density that is positive and differentiable in a scalar parameter θ. Express $\frac{\partial}{\partial\theta}\mathbb{p}(\underset{\sim}{x} \mid \theta)$ in terms of $\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x} \mid \theta)$.

Solution

Solution 0.11. By the chain rule, with outer function $\log u$ and inner function $u = \mathbb{p}(\underset{\sim}{x} \mid \theta)$:

$$\frac{\partial}{\partial\theta} \log \mathbb{p}(\underset{\sim}{x} \mid \theta) = \left(\frac{d}{du} \log u\right)\Big|_{u=\mathbb{p}(\underset{\sim}{x}|\theta)} \frac{\partial}{\partial\theta}\mathbb{p}(\underset{\sim}{x} \mid \theta).$$

The inner derivative $\frac{\partial}{\partial\theta}\mathbb{p}(\underset{\sim}{x} \mid \theta)$ is the quantity we want, so only the outer derivative needs working out:

$$\frac{d}{du} \log u = \frac{1}{u} \quad \text{(derivative of the natural logarithm)}$$

Plugging the outer derivative back in:

$$\begin{aligned} \frac{\partial}{\partial\theta} \log \mathbb{p}(\underset{\sim}{x} \mid \theta) &= \frac{1}{\mathbb{p}(\underset{\sim}{x} \mid \theta)} \frac{\partial}{\partial\theta}\mathbb{p}(\underset{\sim}{x} \mid \theta) && \text{(outer derivative is } 1/u) \\ \frac{\partial}{\partial\theta}\mathbb{p}(\underset{\sim}{x} \mid \theta) &= \left(\frac{\partial}{\partial\theta} \log \mathbb{p}(\underset{\sim}{x} \mid \theta)\right)\mathbb{p}(\underset{\sim}{x} \mid \theta) && \text{(multiply both sides by } \mathbb{p}(\underset{\sim}{x} \mid \theta) > 0) \end{aligned}$$

Lemma 0.1 (Derivative of a density through its logarithm)**.** *Let $\mathbb{p}(\underset{\sim}{x} \mid \theta)$ be a density that is positive and differentiable in a scalar parameter θ. Then its derivative with respect to θ is the derivative of its logarithm times the density:*

$$\frac{\partial}{\partial\theta}\mathbb{p}(\underset{\sim}{x} \mid \theta) = \left(\frac{\partial}{\partial\theta} \log \mathbb{p}(\underset{\sim}{x} \mid \theta)\right)\mathbb{p}(\underset{\sim}{x} \mid \theta)$$

Proof

Proof. This is the solution to Exercise 0.11, worked in Solution 0.11. □

Exercise 0.12 (Mean of the score)**.** Let $\underset{\sim}{X}$ be continuous with density $\mathbb{p}(\underset{\sim}{x} \mid \theta)$ for a scalar parameter θ. Suppose that $\underset{\sim}{X}$ has a range $\mathcal{R}(\underset{\sim}{X})$ that does not depend on θ, and the order of differentiation with respect to θ and integration over the data can be exchanged. Find the mean of the score.

Solution

Solution 0.12. The step that rewrites the derivative of $\log \mathbb{p}(\underset{\sim}{x} \mid \theta)$ uses Lemma 0.1:

$$
\begin{aligned}
\mathbb{E}[\ell'] &= \int_{\mathcal{R}(\underset{\sim}{X})} \ell' \, \mathbb{p}(\underset{\sim}{x} \mid \theta) \, d\underset{\sim}{x} && \text{(definition of expectation)} \\
&= \int_{\mathcal{R}(\underset{\sim}{X})} \left(\frac{\partial}{\partial \theta} \log \mathbb{p}(\underset{\sim}{x} \mid \theta) \right) \mathbb{p}(\underset{\sim}{x} \mid \theta) \, d\underset{\sim}{x} && \text{(definition of the score)} \\
&= \int_{\mathcal{R}(\underset{\sim}{X})} \frac{\frac{\partial}{\partial \theta} \mathbb{p}(\underset{\sim}{x} \mid \theta)}{\mathbb{p}(\underset{\sim}{x} \mid \theta)} \mathbb{p}(\underset{\sim}{x} \mid \theta) \, d\underset{\sim}{x} && \text{(chain rule for log)} \\
&= \int_{\mathcal{R}(\underset{\sim}{X})} \frac{\partial}{\partial \theta} \mathbb{p}(\underset{\sim}{x} \mid \theta) \, d\underset{\sim}{x} && \text{(cancel } \mathbb{p}(\underset{\sim}{x} \mid \theta)\text{)} \\
&= \frac{\partial}{\partial \theta} \int_{\mathcal{R}(\underset{\sim}{X})} \mathbb{p}(\underset{\sim}{x} \mid \theta) \, d\underset{\sim}{x} && \text{(exchange derivative and integral)} \\
&= \frac{\partial}{\partial \theta} 1 && \text{(a density integrates to 1)} \\
&= 0 && \text{(derivative of a constant)}
\end{aligned}
$$

Theorem 0.11 (Mean of the score)**.** *Let $\underset{\sim}{X}$ be a continuous random vector with density $\mathbb{p}(\underset{\sim}{x} \mid \theta)$ for a scalar parameter θ. Suppose the set of possible data values does not depend on θ, and the order of differentiation with respect to θ and integration over the data can be exchanged. Then the score ℓ', the derivative of the log-likelihood with respect to θ, has expectation zero:*

$$\mathbb{E}[\ell'] = 0$$

Proof

Proof. This is the solution to Exercise 0.12, worked in Solution 0.12. It builds on Lemma 0.1. □

Exercise 0.13 (Second moment of the score)**.** Under the assumptions of Exercise 0.12, show that $\mathbb{E}[\ell'^2] = \mathcal{I}(\theta)$, the expected information.

Solution

Solution 0.13.
By Theorem 0.11, $0 = \int_{\mathcal{R}(\underset{\sim}{X})} \left(\frac{\partial}{\partial \theta} \log \mathbb{p}(\underset{\sim}{x} \mid \theta) \right) \mathbb{p}(\underset{\sim}{x} \mid \theta) \, d\underset{\sim}{x}$ for every θ. Differentiate this identity with respect to θ; the step that substitutes for $\frac{\partial}{\partial \theta} \mathbb{p}(\underset{\sim}{x} \mid \theta)$ uses Lemma 0.1:

$$
\begin{aligned}
\frac{\partial}{\partial\theta}0 &= \frac{\partial}{\partial\theta}\int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} && \text{(differentiat}\\
0 &= \frac{\partial}{\partial\theta}\int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} && \text{(derivative o}\\
&= \int_{\mathcal{R}(\underset{\sim}{X})}\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\mathbb{p}(\underset{\sim}{x}\mid\theta)\right]d\underset{\sim}{x} && \text{(exchange d}\\
&= \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial^2}{\partial\theta^2}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} + \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\frac{\partial}{\partial\theta}\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} && \text{(product rul}\\
&= \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial^2}{\partial\theta^2}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} + \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} && \left(\frac{\partial}{\partial\theta}\mathbb{p} = \left(\frac{\partial}{\partial\theta}\right.\right.\\
&= \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial^2}{\partial\theta^2}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} + \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)^2\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} && \text{(product of}\\
&= \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log \mathbb{p}(\underset{\sim}{X}\mid\theta)\right] + \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)^2\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} && \text{(definition o}\\
&= \mathbb{E}[\ell''] + \int_{\mathcal{R}(\underset{\sim}{X})}\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{x}\mid\theta)\right)^2\mathbb{p}(\underset{\sim}{x}\mid\theta)\,d\underset{\sim}{x} && \text{(definition o}\\
&= \mathbb{E}[\ell''] + \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log \mathbb{p}(\underset{\sim}{X}\mid\theta)\right)^2\right] && \text{(definition o}\\
&= \mathbb{E}[\ell''] + \mathbb{E}[\ell'^2] && \text{(definition o}
\end{aligned}
$$

Rearranging:

$$
\begin{aligned}
\mathbb{E}[\ell'^2] &= -\mathbb{E}[\ell''] && \text{(subtract } \mathbb{E}[\ell''] \text{ from both sides)}\\
&= \mathbb{E}[-\ell''] && \text{(move the constant factor } -1 \text{ into the expectation)}\\
&= \mathbb{E}[I] && \text{(observed information is the negative Hessian)}\\
&= \mathcal{I}(\theta) && \text{(definition of expected information)}
\end{aligned}
$$

Theorem 0.12 (Second moment of the score)**.** *Under the assumptions of Theorem 0.11, the second moment* $\mathbb{E}[\ell'^2]$ *of the score* ℓ' *is the expected information* $\mathcal{I}(\theta)$*, the expected negative second derivative of the log-likelihood:*

$$\mathbb{E}[\ell'^2] = \mathcal{I}(\theta)$$

Proof

Proof. This is the solution to Exercise 0.13, worked in Solution 0.13. It builds on Theorem 0.11 and Lemma 0.1. □

Theorem 0.13 (Mean and variance of the score)**.** *Suppose the set of possible data values does not depend on $\underset{\sim}{\theta}$, and the order of differentiation with respect to $\underset{\sim}{\theta}$ and integration over the data can be exchanged. Then the score has mean zero, and its variance is the expected information:*

$$\mathbb{E}[\ell'(\underset{\sim}{X} \mid \underset{\sim}{\theta})] = \mathbf{0}_{p\times 1} \tag{12}$$

$$\begin{aligned} \mathcal{I}(\underset{\sim}{\theta}) &= \mathrm{Cov}(\ell'(\underset{\sim}{X} \mid \underset{\sim}{\theta})) \\ &= \mathbb{E}\left[\ell'\ell'^{\top}\right] \end{aligned} \tag{13}$$

Proof

Proof. We give the proof for a scalar θ and a continuous $\underset{\sim}{X}$ with density $\mathbb{p}(\underset{\sim}{x} \mid \theta)$; the vector case applies the same steps to each pair of entries, and a discrete $\underset{\sim}{X}$ replaces integrals with sums.
Equation 12 follows from Theorem 0.11. For Equation 13, Theorem 0.12 gives $\mathcal{I}(\theta) = \mathbb{E}[\ell'^2]$, and because $\mathbb{E}[\ell'] = 0$,

$$\begin{aligned} \mathbb{E}[\ell'^2] &= \mathbb{E}[\ell'^2] - (\mathbb{E}[\ell'])^2 \\ &= \mathrm{Var}(\ell'). \end{aligned}$$

□

Example 0.5 (Checking the information equality for Poisson data)**.** For $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$, the score is $\ell' = \sum_{i=1}^{n} X_i/\lambda - n$ (Example 0.4). Then:

$$
\begin{aligned}
\mathbb{E}[\ell'] &= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{\lambda} - n\right] && \text{(Poisson score)} \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{\lambda}\right] - \mathbb{E}[n] && \text{(linearity of expectation)} \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{\lambda}\right] - n && \text{(the expectation of a constant is that constant)} \\
&= \frac{\mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\lambda} - n && \text{(factor the constant } \tfrac{1}{\lambda} \text{ out of the expectation)} \\
&= \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{\lambda} - n && \text{(linearity of expectation)} \\
&= \frac{\sum_{i=1}^n \lambda}{\lambda} - n && (\mathbb{E}[X_i] = \lambda) \\
&= \frac{n\lambda}{\lambda} - n && \text{(sum of } n \text{ copies of } \lambda) \\
&= n - n && \text{(cancel } \lambda) \\
&= 0 && \text{(subtract)} \\
\mathrm{Var}(\ell') &= \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i}{\lambda} - n\right) && \text{(Poisson score)} \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i}{\lambda}\right) && \text{(subtracting a constant does not change a variance)} \\
&= \frac{\mathrm{Var}\left(\sum_{i=1}^n X_i\right)}{\lambda^2} && (\mathrm{Var}(aY) = a^2\mathrm{Var}(Y)) \\
&= \frac{\sum_{i=1}^n \mathrm{Var}(X_i)}{\lambda^2} && \text{(variance of a sum of independent variables)} \\
&= \frac{\sum_{i=1}^n \lambda}{\lambda^2} && (\mathrm{Var}(X_i) = \lambda) \\
&= \frac{n\lambda}{\lambda^2} && \text{(sum of } n \text{ copies of } \lambda) \\
&= \frac{n}{\lambda} && \text{(cancel one factor of } \lambda)
\end{aligned}
$$

which matches $\mathcal{I}(\lambda) = n/\lambda$ from Example 0.4.

Example 0.6 (When the support depends on the parameter)**.** Let $X_1, \dots, X_n \sim_{\text{iid}} \text{Uniform}(0, \theta)$. 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 $\mathbb{E}[\ell'] = -n/\theta \neq 0$: Equation 12 fails, because the set of possible data values, $(0, \theta)$, depends on θ.

Remark (The score is the negative derivative of the surprisal)

Remark 0.5 (The score is the negative derivative of the surprisal). For a continuous $\underset{\sim}{X}$ with density $\mathbb{p}(\underset{\sim}{x} \mid \theta)$, the surprisal of the data value $\underset{\sim}{x}$ under parameter value θ is $-\log \mathbb{p}(\underset{\sim}{x} \mid \theta)$, which is minus the log-likelihood $-\ell(\theta)$. So the score is minus the derivative of the surprisal with respect to θ. By Theorem 0.13, the expected information is the expected squared derivative of the surprisal, with the expectation taken over the data $\underset{\sim}{X}$ when the parameter is θ: $\mathcal{I}(\theta) = \mathbb{E}[\ell'^2]$.

Definition 0.12 (Kullback–Leibler divergence between two parameter values)**.** Let $\underset{\sim}{X}$ be a continuous random vector with density $\mathbb{p}(\underset{\sim}{x} \mid \theta)$ for a scalar parameter θ, and let $\underset{\sim}{\theta}_0$ be the true parameter value. The Kullback–Leibler divergence of the model with parameter θ from the model with parameter θ_0 is the expectation, under θ_0, of the logarithm of the ratio of the two densities. As a function of θ, it is written

$$D_{\mathrm{KL}}(\theta_0 \,\|\, \theta) \stackrel{\text{def}}{=} \mathbb{E}_{\theta_0}\left[\log \frac{\mathbb{p}(\underset{\sim}{X} \mid \theta_0)}{\mathbb{p}(\underset{\sim}{X} \mid \theta)}\right].$$

Example 0.7 (Divergence between two Poisson rates)**.** For one observation $X \sim \mathrm{Pois}(\lambda_0)$, $\log \mathbb{p}(x \mid \lambda) = x \log \lambda - \lambda - \log x!$, so

$$\begin{aligned} D_{\mathrm{KL}}(\lambda_0 \,\|\, \lambda) &= \mathbb{E}_{\lambda_0}\left[X \log \frac{\lambda_0}{\lambda} - \lambda_0 + \lambda\right] \\ &= \lambda_0 \log \frac{\lambda_0}{\lambda} - \lambda_0 + \lambda. \end{aligned}$$

Exercise 0.14 (Derivative of the divergence)**.** Under the assumptions of Theorem 0.11, show that the derivative of the divergence of Definition 0.12 is

$$\frac{\partial}{\partial \theta} D_{\mathrm{KL}}(\theta_0 \,\|\, \theta) = -\mathbb{E}_{\theta_0}[\ell'(\theta)],$$

where $\ell'(\theta)$ is the score evaluated at θ, and the expectation is taken with the data generated under θ_0.

Solution

Solution 0.14. Write $\ell(\theta) = \log \mathbb{p}(\underset{\sim}{X} \mid \theta)$.

$$\begin{aligned}
D_{\mathrm{KL}}(\theta_0 \,\|\, \theta) &= \mathbb{E}_{\theta_0}[\ell(\theta_0) - \ell(\theta)] && \text{(log of a ratio is a difference of logs)} \\
&= \mathbb{E}_{\theta_0}[\ell(\theta_0)] - \mathbb{E}_{\theta_0}[\ell(\theta)] && \text{(linearity of expectation)} \\
\frac{\partial}{\partial\theta} D_{\mathrm{KL}}(\theta_0 \,\|\, \theta) &= -\frac{\partial}{\partial\theta} \mathbb{E}_{\theta_0}[\ell(\theta)] && \text{(the first term does not depend on } \theta) \\
&= -\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] && \text{(exchange derivative and expectation)} \\
&= -\mathbb{E}_{\theta_0}[\ell'(\theta)] && \text{(definition of the score)}
\end{aligned}$$

Exercise 0.15 (Second derivative of the divergence)**.** Under the assumptions of Exercise 0.14, show that the second derivative of the divergence of Definition 0.12 is

$$\frac{\partial^2}{\partial\theta^2} D_{\mathrm{KL}}(\theta_0 \,\|\, \theta) = \mathbb{E}_{\theta_0}[-\ell''(\theta)],$$

where $\ell''(\theta)$ is the Hessian of the log-likelihood evaluated at θ.

Solution

Solution 0.15.

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2} D_{\mathrm{KL}}(\theta_0 \,\|\, \theta) &= \frac{\partial}{\partial\theta}\Big(-\mathbb{E}_{\theta_0}[\ell'(\theta)]\Big) && \text{(@sol-kl-slope)} \\
&= -\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\ell'(\theta)\right] && \text{(exchange derivative and expectation)} \\
&= -\mathbb{E}_{\theta_0}[\ell''(\theta)] && \text{(the Hessian is the derivative of the score)} \\
&= \mathbb{E}_{\theta_0}[-\ell''(\theta)] && \text{(move the sign inside the expectation)}
\end{aligned}$$

Theorem 0.14 (The expected information is the curvature of the Kullback–Leibler divergence)**.** *Under the assumptions of Theorem 0.11, the second derivative at $\theta = \theta_0$ of the divergence $D_{\mathrm{KL}}(\theta_0 \,\|\, \theta)$ of Definition 0.12 is the expected information:*

$$\frac{\partial^2}{\partial\theta^2} D_{\mathrm{KL}}(\theta_0 \,\|\, \theta)\Big|_{\theta=\theta_0} = \mathcal{I}(\theta_0)$$

Proof

Proof. By Solution 0.15, the second derivative is $\mathbb{E}_{\theta_0}[-\ell''(\theta)]$ at every θ. Setting $\theta = \theta_0$ gives the expectation, under the true parameter θ_0, of minus the Hessian of the log-likelihood, which is $\mathcal{J}(\theta_0)$ by Definition 0.10. □

Example 0.8 (Curvature of the divergence for Poisson data)**.** Differentiating the divergence of Example 0.7 twice with respect to λ gives $\frac{\partial}{\partial\lambda} D_{\mathrm{KL}}(\lambda_0 \,\|\, \lambda) = 1 - \lambda_0/\lambda$ and $\frac{\partial^2}{\partial\lambda^2} D_{\mathrm{KL}}(\lambda_0 \,\|\, \lambda) = \lambda_0/\lambda^2$, so at $\lambda = \lambda_0$ the second derivative is $1/\lambda_0$, which is the expected information of one observation (Example 0.4). A finite-difference second derivative of the divergence, summed directly over the Poisson probabilities, agrees:

```
rate_true <- 3
kl_div <- function(rate) {
  x <- 0:200
  log_ratio <- dpois(x, rate_true, log = TRUE) - dpois(x, rate, log = TRUE)
  sum(dpois(x, rate_true) * log_ratio)
}
h <- 1e-3
kl_values <- c(
  kl_div(rate_true - h),
  kl_div(rate_true),
  kl_div(rate_true + h)
)
c(
  second_difference = sum(kl_values * c(1, -2, 1)) / h^2,
  expected_information = 1 / rate_true
)
#>    second_difference expected_information
#>             0.333333             0.333333
```

These notes write the observed information as I and the expected information as $\mathcal{J}$. Sources disagree on these symbols. The Notation Across Sources page of the macros site lists them side by side.

Asymptotic distribution of the maximum likelihood estimate

Theorem 0.15 (Central limit theorem for MLEs)**.** *For* iid *data from a correctly specified model satisfying regularity conditions (including those of Theorem 0.13, 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 $\underset{\sim}{\theta}$, with covariance matrix equal to the inverse of the expected information:*

$$\hat{\theta}_{ML} \dot{\sim} \mathrm{N}\Big(\underset{\sim}{\theta}, \big(\mathcal{I}(\underset{\sim}{\theta})\big)^{-1}\Big) \tag{14}$$

i Proof

Proof. The proof is beyond the scope of these notes; see [16], Theorem 7.5.2, p. 501 and [17]. □

i Remark (Consistent roots and global maximizers)

Remark 0.6 (Consistent roots and global maximizers). 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 0.4).

Example 0.9 (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 0.5) to zero gives $\hat{\lambda}_{\text{ML}} = \bar{X}$, and $\mathcal{I}(\lambda) = n/\lambda$ (Example 0.4), so Theorem 0.15 says that for large n:

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

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

Theorem 0.15 involves the unknown $\underset{\sim}{\theta}$, so to use it we estimate $\mathcal{I}(\underset{\sim}{\theta})$, by either the expected information at the MLE, $\mathcal{I}(\hat{\theta}_{\text{ML}})$, or the observed information at the MLE, $I(\underset{\sim}{x}; \hat{\theta}_{\text{ML}})$. Either way, the estimated standard error of the kth entry of $\hat{\theta}_{\text{ML}}$ is:

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

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

Remark (Observed versus expected information)

Remark 0.7 (Observed versus expected information). Using the observed information is often more convenient, and there are settings where it is provably better by some criteria [18].

Read

- Hogg et al. [7, Sec. 6.6], pp. 285–289 covers asymptotic distributions of maximum likelihood estimators.
- Rice [8, Sec. 8.5.2], pp. 274–278 covers large-sample theory for maximum likelihood estimators.

Quantifying uncertainty about MLEs

Confidence intervals for MLEs

Definition 0.13 (Wald confidence interval)**.** The approximate $100(1-\alpha)\%$ **Wald confidence interval** for the kth entry θ_k of a parameter vector is

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

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

By Theorem 0.15, $(\hat{\theta}_k - \theta_k)/\widehat{\mathrm{SE}}\left(\hat{\theta}_k\right)$ 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, $z_{0.975} \approx 1.96$.

Read

- Casella and Berger [3, Sec. 10.4.1], pp. 496–498 covers approximate maximum likelihood intervals.
- Rice [8, Sec. 8.5.3], pp. 279–284 covers confidence intervals from maximum likelihood estimators.

Wald tests

Definition 0.14 (Wald test)**.** The **Wald test** of $H_0 : \theta_k = \theta_{k,0}$ uses the test statistic

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

Source: Wikipedia contributors [19].

Definition 0.15 (Wald statistic for several constraints)**.** For q constraints $H_0 : \underset{\sim}{\theta}_{(q)} = \underset{\sim}{\theta}_{(q),0}$ on a $q \times 1$ subvector, the **Wald statistic** is $\left(\hat{\underset{\sim}{\theta}}_{(q)} - \underset{\sim}{\theta}_{(q),0}\right)^{\top} \left(\hat{V}_{(q)}\right)^{-1} \left(\hat{\underset{\sim}{\theta}}_{(q)} - \underset{\sim}{\theta}_{(q),0}\right)$, where $\hat{V}_{(q)}$ is the corresponding $q \times q$ block of $\left(\hat{\mathcal{J}}\right)^{-1}$.

Example 0.10 (Wald interval and test for a Poisson rate)**.** For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$, $\hat{\lambda}_{\text{ML}} = \bar{x}$ and $\mathcal{J}(\lambda) = n/\lambda$ (Example 0.9), 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
```

Remark (Null distributions of the Wald statistics)

Remark 0.8 (Null distributions of the Wald statistics). By Theorem 0.15, the statistic Z of the Wald test has approximately a standard Gaussian distribution under H_0 in large samples. The Wald statistic for q constraints (Definition 0.15) has approximately a χ^2_q distribution under H_0.

Likelihood ratio tests for MLEs

Definition 0.16 (Likelihood ratio statistic)**.** Suppose a null hypothesis H_0 imposes q constraints on the parameter vector $\underset{\sim}{\theta}$. Let $\hat{\theta}_{\mathrm{ML}}$ be the unrestricted MLE and $\hat{\theta}_0$ the MLE under H_0. The **likelihood ratio statistic** is

$$\Lambda \stackrel{\text{def}}{=} 2\Big(\ell(\hat{\theta}_{\mathrm{ML}}) - \ell(\hat{\theta}_0)\Big)$$

Source: Wikipedia contributors [20].

Exercise 0.16 (Quadratic approximation to the log-likelihood)**.** Let θ be a scalar parameter, and suppose the MLE $\hat{\theta}_{\mathrm{ML}}$ maximizes ℓ at an interior point of the parameter space, where ℓ is twice differentiable. Use a second-order Taylor expansion around $\hat{\theta}_{\mathrm{ML}}$ to approximate $\ell(\theta_0)$ in terms of $\ell(\hat{\theta}_{\mathrm{ML}})$ and the Hessian $\ell''(\hat{\theta}_{\mathrm{ML}})$.

Solution

Solution 0.16. 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:

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

Lemma 0.2 (Quadratic approximation to the log-likelihood)**.** *Let θ be a scalar parameter with log-likelihood ℓ, and suppose the MLE $\hat{\theta}_{ML}$ maximizes ℓ at an interior point of the parameter space, where ℓ is twice differentiable with Hessian ℓ''. Then the log-likelihood at any nearby value θ_0 is approximately quadratic around $\hat{\theta}_{ML}$:*

$$\ell(\theta_0) \approx \ell(\hat{\theta}_{ML}) + \frac{1}{2}\ell''(\hat{\theta}_{ML})(\theta_0 - \hat{\theta}_{ML})^2$$

Proof

Proof. This is the solution to Exercise 0.16, worked in Solution 0.16. □

Exercise 0.17 (Likelihood ratio statistic for a point null hypothesis)**.** With a scalar parameter θ, the point null hypothesis $H_0 : \theta = \theta_0$ imposes $q = 1$ constraint. Use Exercise 0.16 to approximate the likelihood ratio statistic Λ in terms of the observed information $I(\hat{\theta}_{\mathrm{ML}})$.

Solution

Solution 0.17. Under H_0, the only allowed value is θ_0, so the restricted MLE is $\hat{\theta}_0 = \theta_0$. Substituting the approximation from Lemma 0.2:

$$
\begin{aligned}
\Lambda &= 2\Big(\ell(\hat{\theta}_{\mathrm{ML}}) - \ell(\hat{\theta}_0)\Big) && \text{(definition of } \Lambda) \\
&= 2\Big(\ell(\hat{\theta}_{\mathrm{ML}}) - \ell(\theta_0)\Big) && (\hat{\theta}_0 = \theta_0) \\
&\approx 2\left(\ell(\hat{\theta}_{\mathrm{ML}}) - \left(\ell(\hat{\theta}_{\mathrm{ML}}) + \frac{1}{2}\ell''(\hat{\theta}_{\mathrm{ML}})(\theta_0 - \hat{\theta}_{\mathrm{ML}})^2\right)\right) && \text{(substitute the expansion)} \\
&= 2\left(\ell(\hat{\theta}_{\mathrm{ML}}) - \ell(\hat{\theta}_{\mathrm{ML}}) - \frac{1}{2}\ell''(\hat{\theta}_{\mathrm{ML}})(\theta_0 - \hat{\theta}_{\mathrm{ML}})^2\right) && \text{(distribute the minus sign)} \\
&= 2\left(-\frac{1}{2}\ell''(\hat{\theta}_{\mathrm{ML}})(\theta_0 - \hat{\theta}_{\mathrm{ML}})^2\right) && \text{(cancel } \ell(\hat{\theta}_{\mathrm{ML}})) \\
&= -\ell''(\hat{\theta}_{\mathrm{ML}})(\theta_0 - \hat{\theta}_{\mathrm{ML}})^2 && \text{(multiply by 2)} \\
&= -\ell''(\hat{\theta}_{\mathrm{ML}})(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2 && ((a-b)^2 = (b-a)^2) \\
&= I(\hat{\theta}_{\mathrm{ML}})(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2 && \text{(observed information is the negative Hessian)}
\end{aligned}
$$

Lemma 0.3 (Likelihood ratio statistic for a point null hypothesis)**.** *Under the conditions of Lemma 0.2, consider the point null hypothesis $H_0 : \theta = \theta_0$, which imposes $q = 1$ constraint. Then the likelihood ratio statistic Λ is approximately the observed information $I(\hat{\theta}_{ML})$ at the MLE $\hat{\theta}_{ML}$ times the squared distance from the MLE to θ_0:*

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

Proof

Proof. This is the solution to Exercise 0.17, worked in Solution 0.17. It builds on Lemma 0.2. □

Exercise 0.18 (Standardizing an approximately Gaussian estimate)**.** Suppose $\hat{\theta}_{\mathrm{ML}} \dot{\sim} \mathrm{N}(\theta_0, \mathcal{J}(\theta_0)^{-1})$, and define $Z \overset{\text{def}}{=} \sqrt{\mathcal{J}(\theta_0)}\,(\hat{\theta}_{\mathrm{ML}} - \theta_0)$. Find the approximate distri-

butions of Z and of Z^2.

Solution

Solution 0.18. Z is a linear function of $\hat{\theta}_{\text{ML}}$, so it is approximately Gaussian, with approximate mean and variance:

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\sqrt{\mathcal{I}(\theta_0)}\,(\hat{\theta}_{\text{ML}} - \theta_0)\right] && \text{(definition of } Z\text{)} \\
&= \sqrt{\mathcal{I}(\theta_0)}\,\mathbb{E}\left[\hat{\theta}_{\text{ML}} - \theta_0\right] && \text{(factor the constant } \sqrt{\mathcal{I}(\theta_0)} \text{ out of the expectation)} \\
&= \sqrt{\mathcal{I}(\theta_0)}\left(\mathbb{E}\left[\hat{\theta}_{\text{ML}}\right] - \mathbb{E}[\theta_0]\right) && \text{(linearity of expectation)} \\
&= \sqrt{\mathcal{I}(\theta_0)}\left(\mathbb{E}\left[\hat{\theta}_{\text{ML}}\right] - \theta_0\right) && \text{(the expectation of a constant is that constant)} \\
&\approx \sqrt{\mathcal{I}(\theta_0)}\,(\theta_0 - \theta_0) && \text{(approximate mean of } \hat{\theta}_{\text{ML}}\text{)} \\
&= \sqrt{\mathcal{I}(\theta_0)} \cdot 0 && \text{(subtract)} \\
&= 0 && \text{(multiply by zero)}
\end{aligned}$$

$$\begin{aligned}
\text{Var}(Z) &= \text{Var}\left(\sqrt{\mathcal{I}(\theta_0)}\,(\hat{\theta}_{\text{ML}} - \theta_0)\right) && \text{(definition of } Z\text{)} \\
&= \mathcal{I}(\theta_0)\text{Var}\left(\hat{\theta}_{\text{ML}}\right) && (\text{Var}(aY + b) = a^2\text{Var}(Y)) \\
&\approx \mathcal{I}(\theta_0)\mathcal{I}(\theta_0)^{-1} && \text{(approximate variance of } \hat{\theta}_{\text{ML}}\text{)} \\
&= 1 && \text{(a number times its inverse is 1)}
\end{aligned}$$

So $Z \mathrel{\dot{\sim}} \text{N}(0, 1)$. The square of a standard Gaussian variable has a χ^2_1 distribution, so Z^2 has approximately a χ^2_1 distribution.

Lemma 0.4 (Standardizing an approximately Gaussian estimate)**.** *Suppose the MLE $\hat{\theta}_{ML}$ is approximately Gaussian with mean θ_0 and variance $\mathcal{I}(\theta_0)^{-1}$, the inverse of the expected information: $\hat{\theta}_{ML} \mathrel{\dot{\sim}} \text{N}(\theta_0, \mathcal{I}(\theta_0)^{-1})$. Define $Z \overset{def}{=} \sqrt{\mathcal{I}(\theta_0)}\,(\hat{\theta}_{ML} - \theta_0)$. Then Z is approximately standard Gaussian, so Z^2 is approximately χ^2_1:*

$$Z^2 \mathrel{\dot{\sim}} \chi^2_1$$

Proof

Proof. This is the solution to Exercise 0.18, worked in Solution 0.18. □

Theorem 0.16 (Wilks' theorem)**.** *Suppose the conditions of Theorem 0.15 hold, and a null hypothesis H_0 imposes q constraints on the parameter vector $\underset{\sim}{\theta}$. Then, if H_0 is true, as $n \to \infty$, the likelihood ratio statistic Λ satisfies:*

$$\Lambda \xrightarrow{d} \chi^2_q$$

Source: Wikipedia contributors [21].

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. For the full proof, see [22] or [1, Sec. 5.7]. By Lemma 0.3, $\Lambda \approx I(\hat{\theta}_{\mathrm{ML}})(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2$. 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 $\hat{\theta}_{\mathrm{ML}}$; this sketch assumes that step rather than proving it. So, with Z as in Lemma 0.4:

$$\begin{aligned}
\Lambda &\approx I(\hat{\theta}_{\mathrm{ML}})(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2 && \text{(quadratic approximation to } \Lambda) \\
&\approx \mathcal{I}(\theta_0)(\hat{\theta}_{\mathrm{ML}} - \theta_0)^2 && (I(\hat{\theta}_{\mathrm{ML}})/\mathcal{I}(\theta_0) \to 1) \\
&= \left(\sqrt{\mathcal{I}(\theta_0)}\,(\hat{\theta}_{\mathrm{ML}} - \theta_0)\right)^2 && \text{(write the product as a square)} \\
&= Z^2 && \text{(definition of } Z)
\end{aligned}$$

By Theorem 0.15, $\hat{\theta}_{\mathrm{ML}} \mathrel{\dot\sim} \mathrm{N}(\theta_0, \mathcal{I}(\theta_0)^{-1})$ under H_0, so by Lemma 0.4, $\Lambda \approx Z^2$ 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; this sketch does not derive the matrix form. □

Remark (Wilks' theorem for nested models)

Remark 0.9 (Wilks' theorem for nested models). 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 0.11 (Likelihood ratio test for a Poisson rate)**.**
For $X_1, \ldots, X_n \sim_{\mathrm{iid}} \mathrm{Pois}(\lambda)$ and $H_0 : \lambda = \lambda_0$, the log-likelihood is $\ell(\lambda) = n\bar{x}\log\lambda - n\lambda - \sum_{i=1}^n \log x_i!$ and $\hat{\lambda}_{\mathrm{ML}} = \bar{x}$, so:

$$
\begin{aligned}
\Lambda &= 2(\ell(\bar{x}) - \ell(\lambda_0)) && \text{(definition of } \Lambda\text{)} \\
&= 2\left(\left(n\bar{x}\log\bar{x} - n\bar{x} - \sum_{i=1}^{n}\log x_i!\right) - \left(n\bar{x}\log\lambda_0 - n\lambda_0 - \sum_{i=1}^{n}\log x_i!\right)\right) && \text{(substitute } \ell(\bar{x}) \text{ and } \ell(\lambda_0)\text{)} \\
&= 2\left(n\bar{x}\log\bar{x} - n\bar{x} - \sum_{i=1}^{n}\log x_i! - n\bar{x}\log\lambda_0 + n\lambda_0 + \sum_{i=1}^{n}\log x_i!\right) && \text{(distribute the minus sign)} \\
&= 2(n\bar{x}\log\bar{x} - n\bar{x} - n\bar{x}\log\lambda_0 + n\lambda_0) && \text{(the } \log x_i! \text{ terms cancel)} \\
&= 2n(\bar{x}\log\bar{x} - \bar{x} - \bar{x}\log\lambda_0 + \lambda_0) && \text{(factor out } n\text{)} \\
&= 2n(\bar{x}\log\bar{x} - \bar{x}\log\lambda_0 - \bar{x} + \lambda_0) && \text{(reorder the terms)} \\
&= 2n(\bar{x}(\log\bar{x} - \log\lambda_0) - \bar{x} + \lambda_0) && \text{(factor } \bar{x} \text{ out of the two log term} \\
&= 2n\left(\bar{x}\log\frac{\bar{x}}{\lambda_0} - \bar{x} + \lambda_0\right) && \text{(log of a quotient)}
\end{aligned}
$$

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

```
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 [1, Sec. 5.7] and https://online.stat.psu.edu/stat504/Lesson02.

Read

- Casella and Berger [3, Sec. 10.3.1], pp. 488–491 covers asymptotic distribution of likelihood ratio tests.
- Rice [8, Sec. 9.4], pp. 339–340 covers generalized likelihood ratio tests.

Exact and approximate tests

Table 1: 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 population mean	One-sample t-test	T_{n-1}	z-test	$Z \sim N(0,1)$
Two-sample population means	Two-sample t-test (pooled variance)	$T_{n_1+n_2-2}$	z-test	$Z \sim N(0,1)$
K-group population means (ANOVA)	F-test	$F_{K-1,\, n-K}$	Likelihood ratio test or Wald test	χ^2_{K-1}

> **Remark** (Exact and approximate tests)
>
> *Remark* 0.10 (Exact and approximate tests). 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 0.15, Gaussian or not.

Prediction intervals

Definition 0.17 (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 the interval from L to U contains Y^* with probability $1-\alpha$, where the probability accounts for the randomness of both the observed data and Y^*:

$$\Pr(L \leq Y^* \leq U) = 1 - \alpha$$

Source: Wikipedia contributors [23].

Suppose $X_1, \dots, X_n \sim_{\text{iid}} \mathrm{N}(\mu, \sigma^2)$ with σ^2 known, and we want to predict the sample 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 difference $\bar{X}^* - \hat{\mu}$ (the negative of the prediction error) 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.

i Remark (Prediction intervals versus confidence intervals)

Remark 0.11 (Prediction intervals versus confidence intervals). A confidence interval covers a fixed parameter, such as the population mean μ. A prediction interval (Definition 0.17) covers a random quantity, such as a new outcome, or, in the example above, the sample mean $\bar{X}^*$ of m new observations. Its probability accounts for the randomness of the new observations, not only of the observed data.
In the Gaussian example above, the standard error of $\hat{\mu} = \bar{X}$ is $\sigma/\sqrt{n}$, so the two intervals, the Wald confidence interval for μ and the prediction interval for $\bar{X}^*$, are

$$\begin{aligned} &\hat{\mu} \pm z_{1-\alpha/2}\, \sigma \sqrt{\frac{1}{n}} && \text{(confidence interval for } \mu) \\ &\hat{\mu} \pm z_{1-\alpha/2}\, \sigma \sqrt{\frac{1}{m} + \frac{1}{n}} && \text{(prediction interval for } \bar{X}^*) \end{aligned}$$

The prediction interval is always the wider of the two, because its variance has the extra term σ^2/m, the variance of the new observations' sample mean. As n grows, the width of the confidence interval shrinks to 0, but the width of the prediction interval shrinks only to $2z_{1-\alpha/2}\, \sigma/\sqrt{m}$: more data pins down μ, but cannot remove the randomness of the new observations. The same contrast holds in regression, between a confidence interval for the conditional mean $\mathbb{E}[Y \mid X = x]$ and a prediction interval for a new outcome Y at $X = x$.

Example: maximum likelihood for tropical cyclones in Australia

Adapted from [1, Sec. 1.6.5].

Data

Table 2 records the number of tropical cyclones in northeastern Australia during 13 November-to-April cyclone seasons, from 1956/57 to 1968/69 [1, 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 2: Number of tropical cyclones during each November-to-April season in northeastern Australia [1, 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.

```
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))
```

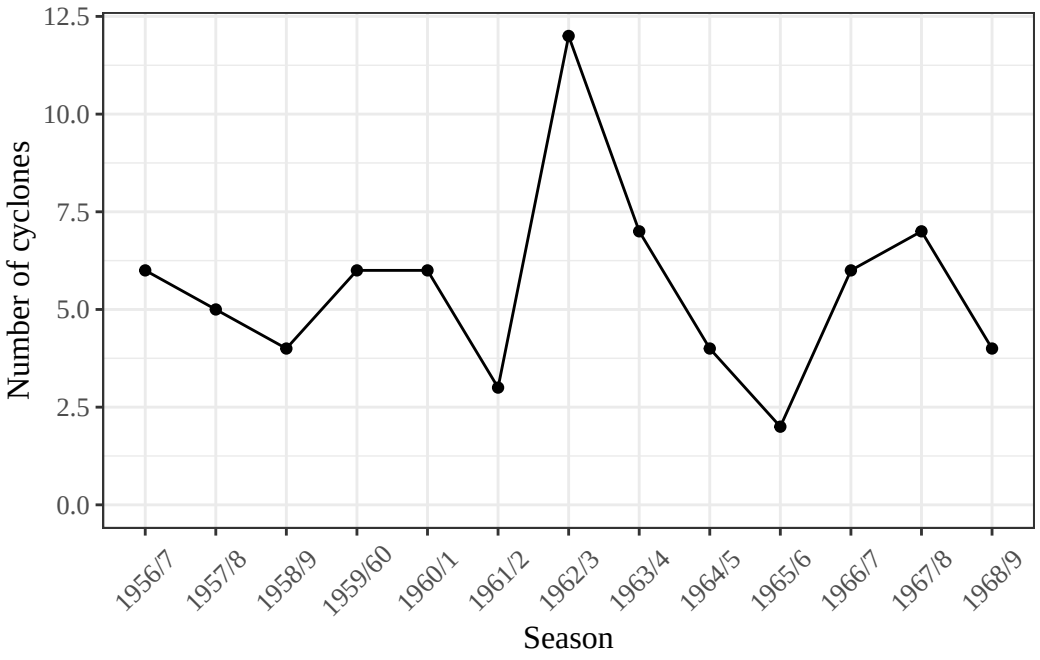


Figure 1: Number of tropical cyclones per season in northeastern Australia, 1956/57 to 1968/69

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 3 provides summary statistics.

```r
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)")
```

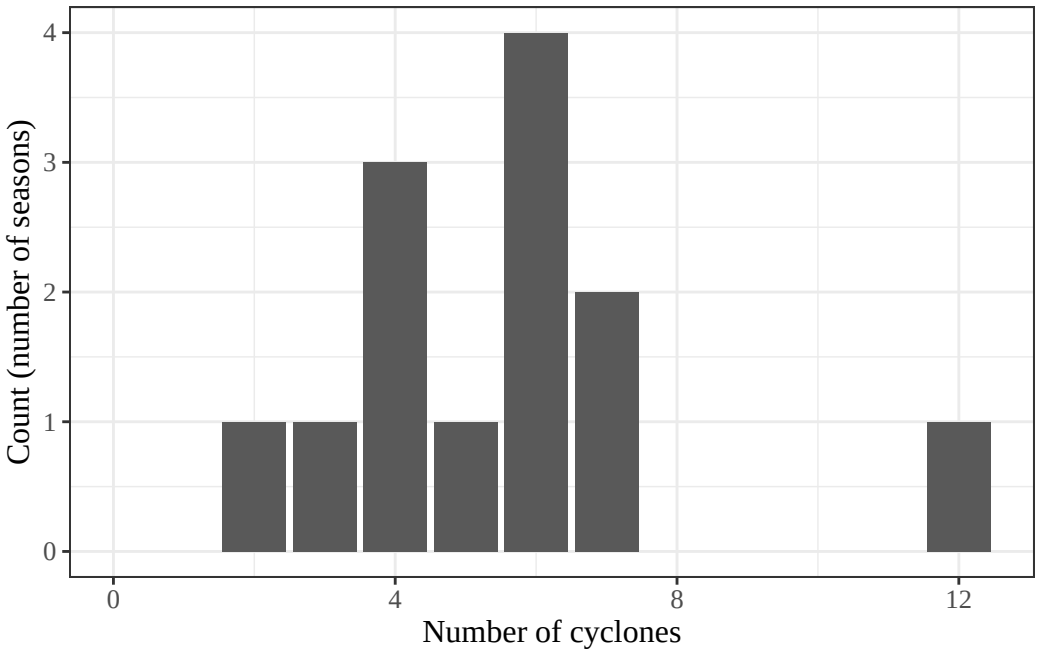


Figure 2: Bar plot of cyclones per season

```r
n <- nrow(cyclones)
sumx <- sum(cyclones$number)
xbar <- mean(cyclones$number)
tibble::tibble(
  seasons = n,
  total = sumx,
  mean = xbar,
  variance = var(cyclones$number)
) |>
  knitr::kable(digits = 2)
```

Table 3: 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 0.1 (Choosing a distribution family)**.** What parametric family of probability distributions might we use to model this empirical distribution?

Solution

Solution 0.1. 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 0.2 (The Poisson probability mass function)**.** Write down the Poisson distribution's probability mass function.

Solution

Solution 0.2.

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

Estimating the model parameters using maximum likelihood

We can estimate the parameter λ using maximum likelihood estimation.

Exercise 0.3 (Likelihood of one observation)**.** Write down the likelihood of a single observation x, according to the Poisson model.

Solution

Solution 0.3.

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

Exercise 0.4 (Model parameters)**.** Write down the vector of parameters in the model.

Solution

Solution 0.4. There is only one parameter, λ:

$$\theta = (\lambda)$$

Exercise 0.5 (Mean and variance)**.** Write down the population mean and variance of a single observation from the model, as functions of the parameters.

Solution

Solution 0.5.

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

The sample mean and variance in Table 3 are of similar size, as the model implies.

Exercise 0.6 (Likelihood of the dataset)**.** Write down the likelihood of the full dataset.

Solution

Solution 0.6.

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

Exercise 0.7 (Graphing the likelihood)**.** Graph the likelihood as a function of λ.

Solution

Solution 0.7.

```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)
```

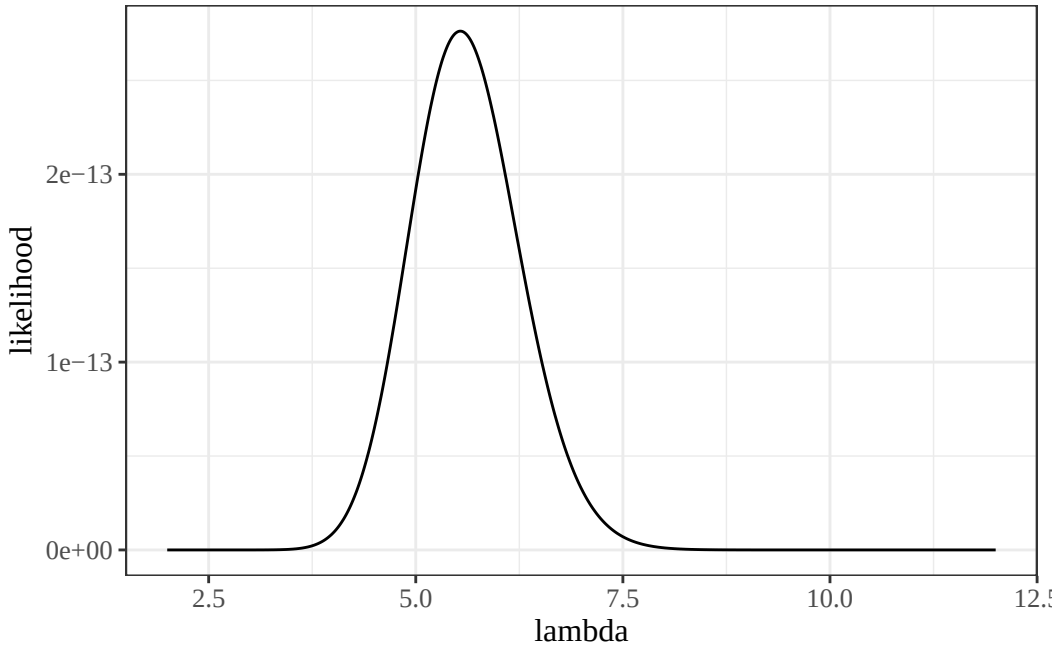


Figure 3: Likelihood of the cyclone data

Exercise 0.8 (Log-likelihood of the dataset)**.** Write down the log-likelihood of the full dataset.

Solution

Solution 0.8.

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

Exercise 0.9 (Graphing the log-likelihood)**.** Graph the log-likelihood as a function of λ.

Solution

Solution 0.9.

```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
```

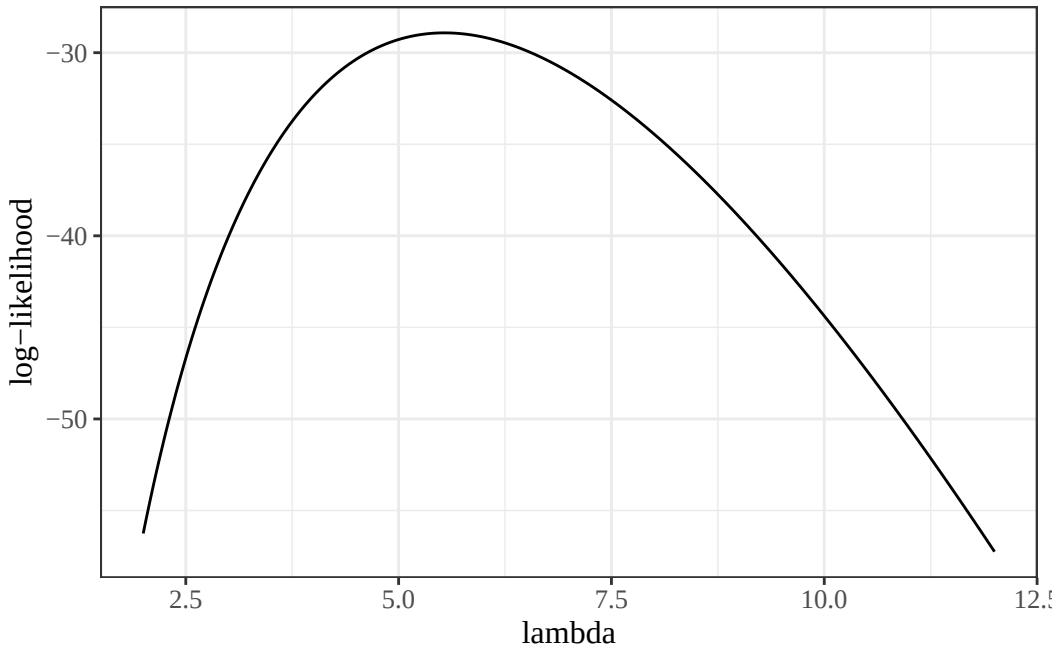


Figure 4: Log-likelihood of the cyclone data

The score function

Exercise 0.10 (Score function of the dataset)**.** Derive the score function for the dataset.

Solution

Solution 0.10. The score function is the first derivative of the log-likelihood from Exercise 0.8:

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

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

Exercise 0.11 (Graphing the score function)**.** Graph the score function.

Solution

Solution 0.11.

```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")
```

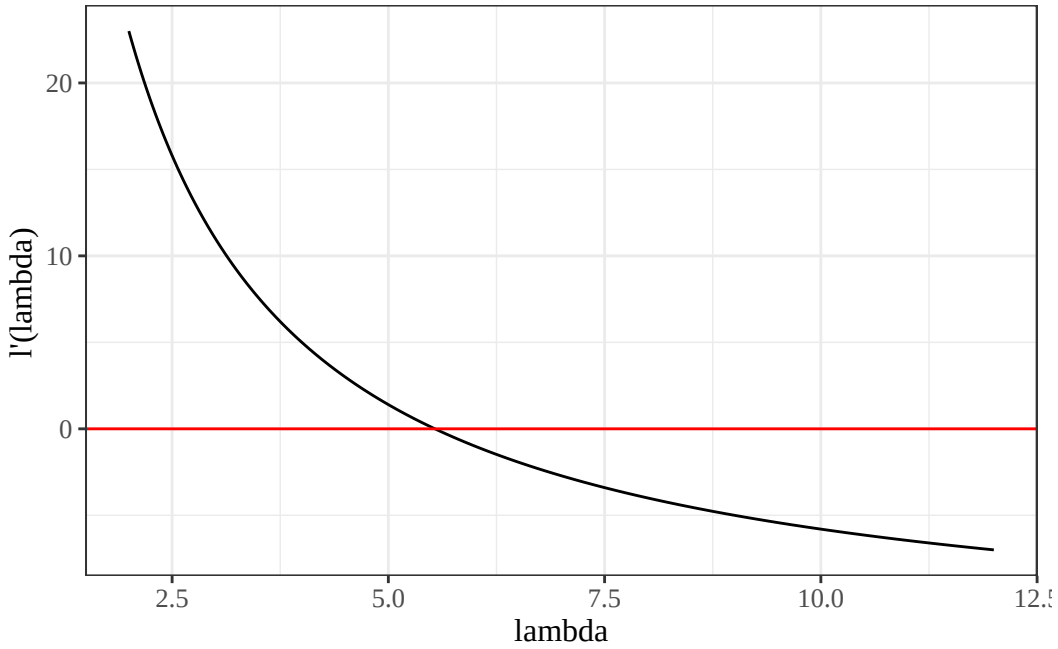


Figure 5: Score function of the cyclone data

The Hessian

Exercise 0.12 (Hessian)**.** Derive the Hessian.

Solution

Solution 0.12. With one parameter, the Hessian is the second derivative of the log-likelihood, which is the derivative of the score from Exercise 0.10:

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

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

Exercise 0.13 (Graphing the Hessian)**.** Graph the Hessian.

Solution

Solution 0.13.

```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")
```

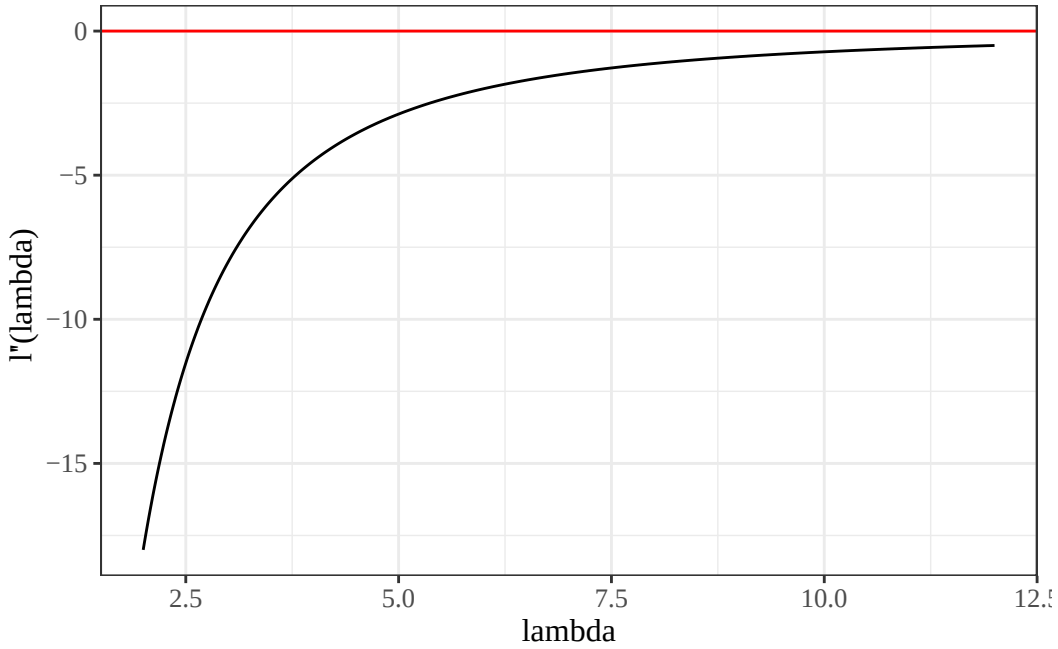


Figure 6: Hessian of the cyclone data's log-likelihood

Exercise 0.14 (Score equation)**.** State the score equation.

Solution

Solution 0.14.

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

Finding the MLE analytically

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

Exercise 0.15 (Solving the score equation)**.** Solve the score equation from Exercise 0.14 for λ, using the score from Exercise 0.10.

Solution

Solution 0.15.

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

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

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

Exercise 0.16 (Checking the second derivative)**.** Confirm that the Hessian $\ell''(\lambda; \underset{\sim}{x})$ is negative when evaluated at $\tilde{\lambda}$, using Exercise 0.12.

Solution

Solution 0.16.

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

Exercise 0.17 (Identifying the MLE)**.** Draw conclusions about the MLE of λ.

Solution

Solution 0.17. Since $\ell''(\tilde{\lambda}; \underset{\sim}{x}) < 0$, $\tilde{\lambda}$ is a local maximizer of the log-likelihood. Moreover, $\ell''(\lambda; \underset{\sim}{x}) = -n\bar{x}/\lambda^2 < 0$ 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)
mle
#> [1] 5.53846
```

$$\begin{aligned}\hat{\lambda}_{\text{ML}} &= \bar{x} \\ &= 5.538\end{aligned}$$

Exercise 0.18 (Graphing the MLE)**.** Graph the log-likelihood with the MLE superimposed.

Solution

Solution 0.18.

```r
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")
```

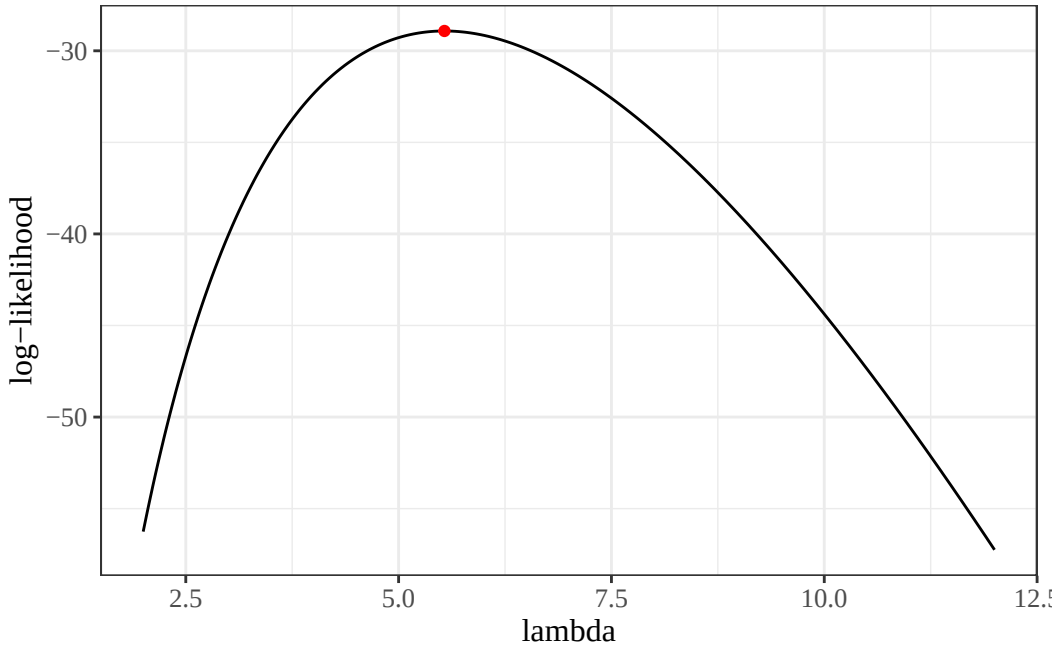


Figure 7: Log-likelihood of the cyclone data, with the MLE marked in red

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

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 [1, Ch. 4].

Definition 0.1 (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

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

```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")
```

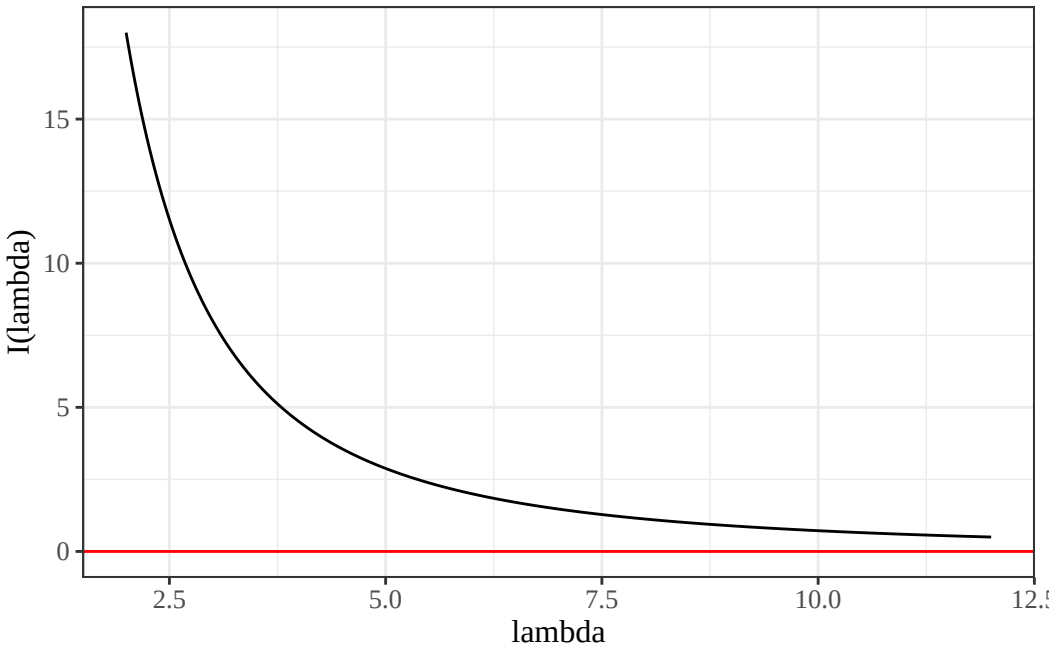


Figure 8: Observed information of the cyclone data

until the estimate stops changing.

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

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

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

Example 0.1 (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 0.4), 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 4).

Definition 0.3 (Empirical information matrix)**.** For mutually independent observations, the **empirical information matrix** is [2]:

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

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

For iid data, $\frac{1}{n} I_e(\theta; \underset{\sim}{x})$ 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 0.13). 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 0.2 (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 0.15). 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
cur_lambda_est
#> [1] 3
```

From Exercise 0.10 and Exercise 0.12, the score function and Hessian are:

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

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

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

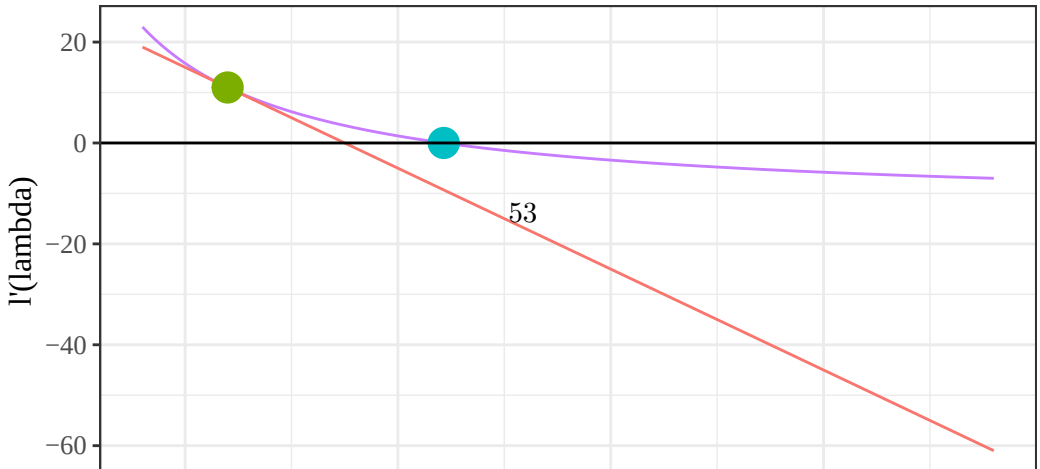
Figure 9 compares the score function and the approximate score function at $\hat{\lambda}^* = 3$.

```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)
new_lambda_est
#> [1] 4.375
```

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

We repeat this process until the log-likelihood stops changing (Table 4).

```
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)

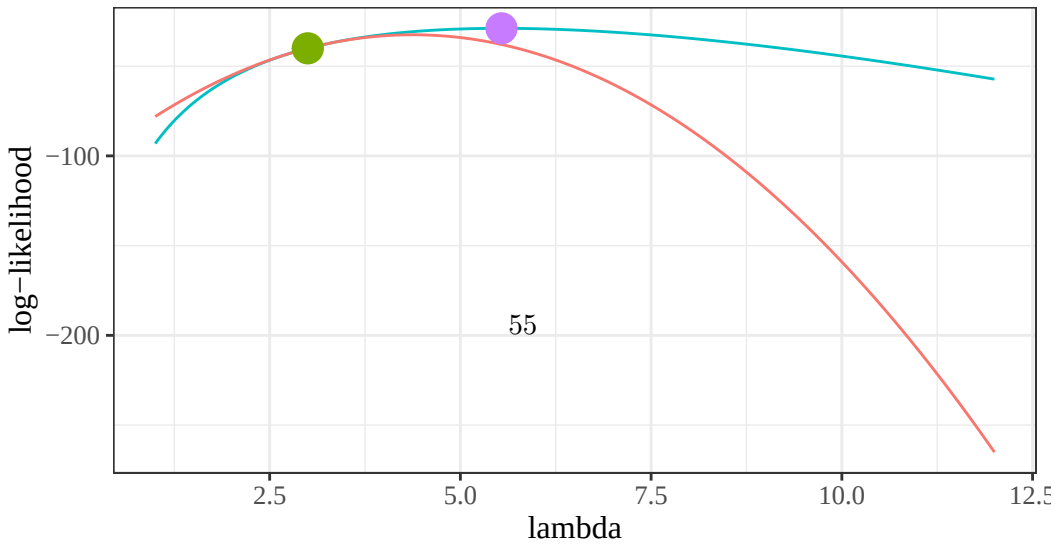
  diff_loglik <- loglik(new_lambda_est) - loglik(cur_lambda_est)

  nr_info <- nr_info |>
    dplyr::bind_rows(
      tibble::tibble(
        iteration = cur_iter,
        lambda = new_lambda_est,
```

```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)
```

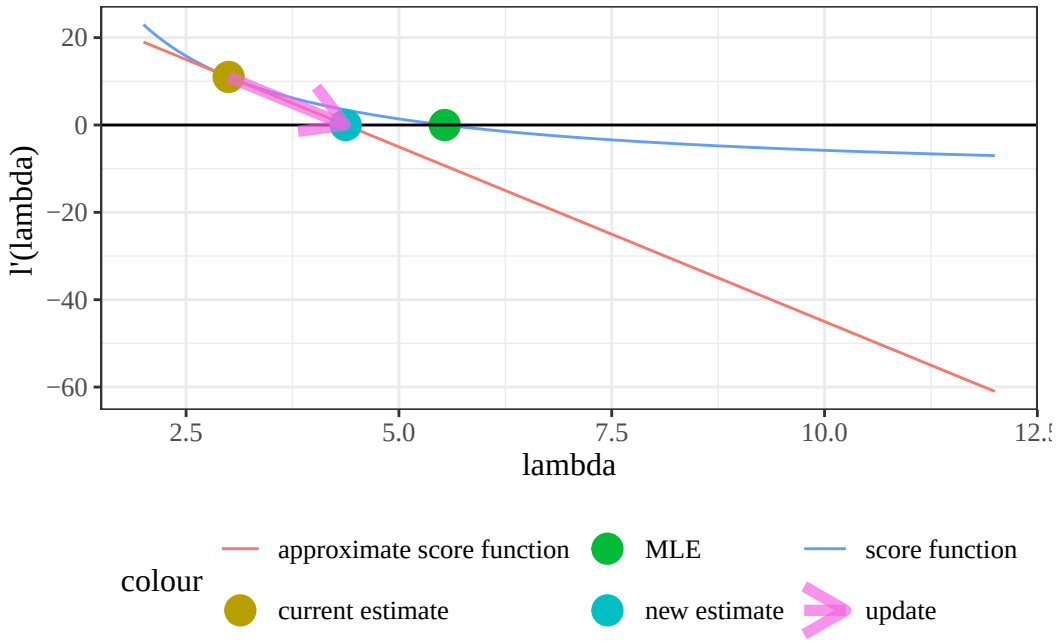


Figure 11: The first Newton-Raphson update: the new estimate is where the approximate score function crosses zero

```
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"
    )
  )
```

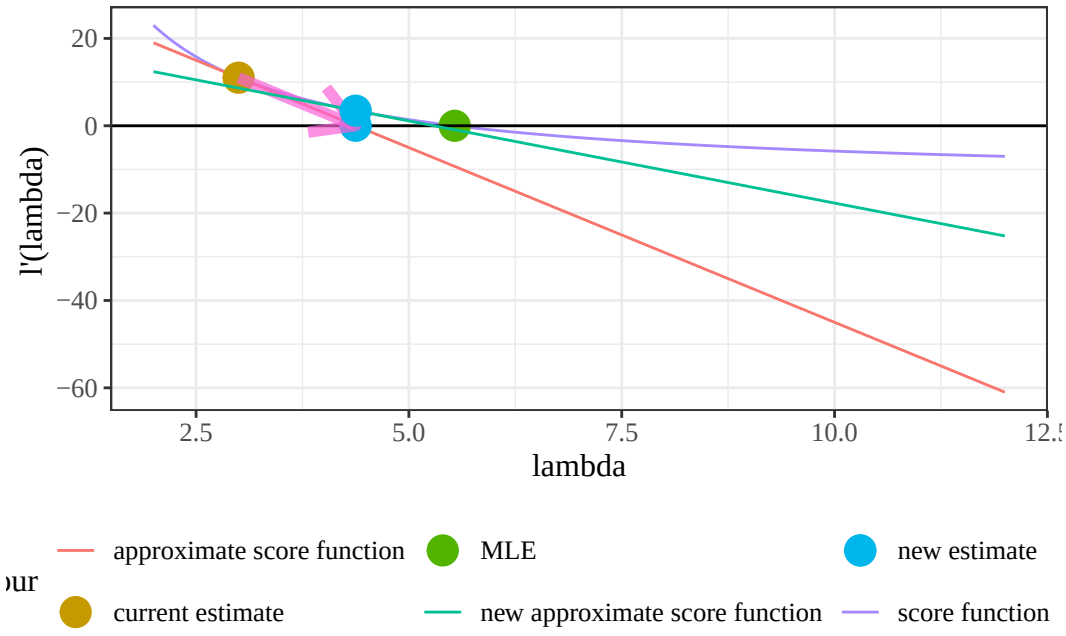


Figure 12: The approximate score function at the updated estimate

```
        `log(likelihood)` = loglik(new_lambda_est),
        score = score(new_lambda_est),
        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 4: 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 0.17).

Maximum likelihood for a Bernoulli model without covariates

The likelihood

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

$$\begin{aligned} &\Pr(Y = 1) = \pi && \text{(definition of the event probability)} \\ &\Pr(Y = 0) = 1 - \pi && \text{(complement rule)} \\ &\Pr(Y = y) = \pi^y (1 - \pi)^{1-y}, \quad y \in \{0, 1\} && \text{(combine the two cases)} \end{aligned}$$

```
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")
```

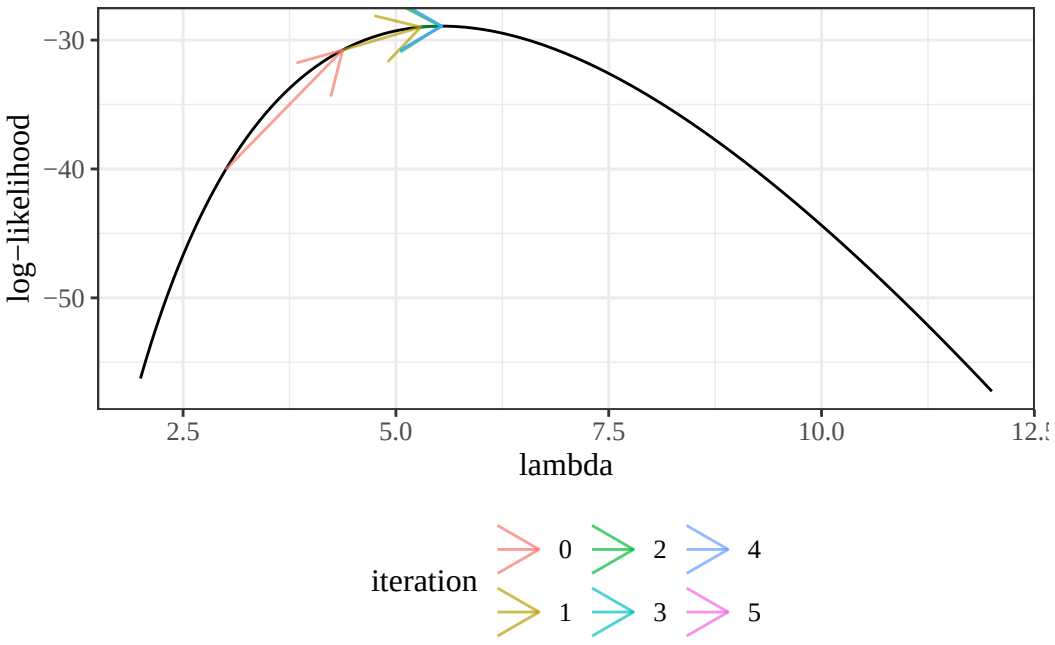


Figure 13: Newton-Raphson steps toward the MLE of the Poisson model (Equation 15) for the cyclone data

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

Solution

Solution 0.1.

$$
\begin{aligned}
\mathcal{L}(\pi; \underset{\sim}{y}) &= \prod_{i=1}^{n} \mathcal{L}_i(\pi) && \text{(product of likelihood components)} \\
&= \prod_{i=1}^{n} \Pr(Y_i = y_i) && \text{(definition of likelihood components)} \\
&= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} && \text{(Bernoulli PMF)} \\
&= \left(\prod_{i=1}^{n} \pi^{y_i}\right)\left(\prod_{i=1}^{n}(1-\pi)^{1-y_i}\right) && \text{(regroup the factors of the product)} \\
&= \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{\sum_{i=1}^{n}(1-y_i)} && \text{(product of powers of a common base)} \\
&= \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{\sum_{i=1}^{n} 1 - \sum_{i=1}^{n} y_i} && \text{(distribute the sum over the terms in the exponent)} \\
&= \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n - \sum_{i=1}^{n} y_i} && \left(\textstyle\sum_{i=1}^{n} 1 = n\right) \\
&= \pi^{s_{\underset{\sim}{y}}}(1-\pi)^{n - s_{\underset{\sim}{y}}} && \left(\text{definition of } s_{\underset{\sim}{y}} \overset{\text{def}}{=} \textstyle\sum_{i=1}^{n} y_i\right)
\end{aligned}
$$

Theorem 0.1 (Likelihood of binary outcomes with one event probability)**.** *Let $Y_1, \ldots, Y_n$ be mutually independent binary random variables, each with the same event probability π: $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Ber}(\pi)$. For observed outcomes $\underset{\sim}{y} \overset{def}{=} (y_1, \ldots, y_n)$, with each $y_i \in \{0, 1\}$, the likelihood $\mathcal{L}(\pi; \underset{\sim}{y})$ of π depends on the data only through the number of outcomes equal to 1, which is the sum of the elements $s_{\underset{\sim}{y}} \overset{def}{=} \sum_{i=1}^{n} y_i$ of $\underset{\sim}{y}$:*

$$\mathcal{L}(\pi; \underset{\sim}{y}) = \pi^{s_{\underset{\sim}{y}}}(1-\pi)^{n - s_{\underset{\sim}{y}}}$$

Proof

Proof. This is the solution to Exercise 0.1, worked in Solution 0.1. □

Log-likelihood contributions

Exercise 0.2 (Log-likelihood contribution of one binary outcome)**.** Let:

- $Y_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi)$ for $i = 1, \ldots, n$, with no covariates;
- $y_i \in \{0, 1\}$ be the observed value of Y_i.

Derive the log-likelihood contribution $\ell_i(\pi)$ of observation i, the log of its likelihood component.

Solution

Solution 0.2.

$$\begin{aligned}
\ell_i(\pi) &= \log\{\Pr(Y_i = y_i)\} && \text{(log of the likelihood component)} \\
&= \log\{\pi^{y_i}(1-\pi)^{1-y_i}\} && \text{(Bernoulli PMF)} \\
&= \log\{\pi^{y_i}\} + \log\{(1-\pi)^{1-y_i}\} && \text{(log of a product)} \\
&= y_i \log\{\pi\} + (1-y_i)\log\{1-\pi\} && \text{(log of a power)}
\end{aligned}$$

Theorem 0.2 (Log-likelihood contribution of one binary outcome)**.** *If $Y_i \sim \text{Ber}(\pi)$ has observed value $y_i \in \{0, 1\}$, its log-likelihood contribution is:*

$$\ell_i(\pi) = y_i \log\{\pi\} + (1-y_i)\log\{1-\pi\}$$

Proof

Proof. This is the solution to Exercise 0.2, worked in Solution 0.2. □

Exercise 0.3 (Log-likelihood contribution of one binary outcome in terms of the log-odds)**.** In the setting of Exercise 0.2, write the log-likelihood contribution $\ell_i(\pi)$ using the log-odds $\eta \stackrel{\text{def}}{=} \text{logit}(\pi)$.

Solution

Solution 0.3. Starting from Theorem 0.2:

$$
\begin{aligned}
\ell_i(\pi) &= y_i \log\{\pi\} + (1 - y_i)\log\{1 - \pi\} && \text{(the contribution from the theorem above)} \\
&= y_i \log\{\pi\} + \log\{1 - \pi\} - y_i \log\{1 - \pi\} && \text{(distribute } \log\{1 - \pi\}\text{)} \\
&= y_i \log\{\pi\} - y_i \log\{1 - \pi\} + \log\{1 - \pi\} && \text{(reorder the terms)} \\
&= y_i (\log\{\pi\} - \log\{1 - \pi\}) + \log\{1 - \pi\} && \text{(collect the terms in } y_i\text{)} \\
&= y_i \log\left\{\frac{\pi}{1 - \pi}\right\} + \log\{1 - \pi\} && \text{(log of a quotient)} \\
&= y_i \eta + \log\{1 - \pi\} && \text{(definition of } \eta \stackrel{\text{def}}{=} \text{logit}(\pi)\text{)}
\end{aligned}
$$

Corollary 0.1 (Log-likelihood contribution of one binary outcome in terms of the log-odds)**.** *If* $Y_i \sim \text{Ber}(\pi)$ *has observed value* $y_i \in \{0, 1\}$*, its log-likelihood contribution is, with the log-odds* $\eta \stackrel{def}{=} \text{logit}(\pi)$*:*

$$\ell_i(\pi) = y_i \eta + \log\{1 - \pi\}$$

Proof

Proof. This is the solution to Exercise 0.3, worked in Solution 0.3. It follows from Theorem 0.2. □

The log-likelihood of the sample

Exercise 0.4 (Log-likelihood of binary outcomes with one event probability)**.** Derive the log-likelihood of $\underset{\sim}{y}$ from Exercise 0.1.

Solution

Solution 0.4. Starting from the likelihood in Theorem 0.1, where $s_{\underset{\sim}{y}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} y_i$:

$$
\begin{aligned}
\ell(\pi; \underset{\sim}{y}) &= \log\{\pi^{s_{\underset{\sim}{y}}}(1-\pi)^{n-s_{\underset{\sim}{y}}}\} && \text{(log of the likelihood)} \\
&= \log\{\pi^{s_{\underset{\sim}{y}}}\} + \log\{(1-\pi)^{n-s_{\underset{\sim}{y}}}\} && \text{(log of a product)} \\
&= s_{\underset{\sim}{y}}\log\{\pi\} + \big(n - s_{\underset{\sim}{y}}\big)\log\{1-\pi\} && \text{(log of a power)} \\
&= s_{\underset{\sim}{y}}\log\{\pi\} + n\log\{1-\pi\} - s_{\underset{\sim}{y}}\log\{1-\pi\} && \text{(distribute } \log\{1-\pi\}\text{)} \\
&= s_{\underset{\sim}{y}}\log\{\pi\} - s_{\underset{\sim}{y}}\log\{1-\pi\} + n\log\{1-\pi\} && \text{(reorder the terms)} \\
&= s_{\underset{\sim}{y}}(\log\{\pi\} - \log\{1-\pi\}) + n\log\{1-\pi\} && \text{(collect the terms in } s_{\underset{\sim}{y}}\text{)} \\
&= s_{\underset{\sim}{y}}\log\Big\{\frac{\pi}{1-\pi}\Big\} + n\log\{1-\pi\} && \text{(log of a quotient)} \\
&= s_{\underset{\sim}{y}}\eta + n\log\{1-\pi\} && \text{(definition of } \eta \overset{\text{def}}{=} \text{logit}(\pi)\text{)}
\end{aligned}
$$

Theorem 0.3 (Log-likelihood of binary outcomes with one event probability)**.** *Let* $Y_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi)$ *for* $i = 1, \dots, n$*, with observed outcomes* $\underset{\sim}{y} \overset{def}{=} (y_1, \dots, y_n)$*, each* $y_i \in \{0, 1\}$*. The log-likelihood* $\ell(\pi; \underset{\sim}{y})$ *of the event probability* π*, written with the log-odds* $\eta \overset{def}{=} \text{logit}(\pi)$*, is:*

$$\ell(\pi; \underset{\sim}{y}) = s_{\underset{\sim}{y}}\eta + n\log\{1-\pi\}$$

Proof

Proof. This is the solution to Exercise 0.4, worked in Solution 0.4. It builds on Theorem 0.1. □

Score contributions

Example 0.1 (Precision of a Bernoulli variable)**.** If Y is a Bernoulli variable, $Y \sim \text{Ber}(\pi)$ with $\pi \in (0, 1)$, then $\text{Var}(Y) = \pi(1-\pi)$, so the precision of Y is $\tau = 1/(\pi(1-\pi))$.

Exercise 0.5 (Score function of a Bernoulli variable)**.** Let $X \sim \text{Ber}(\pi)$ have observed value $x \in \{0, 1\}$.
Derive the score of x. In other words, differentiate the log-likelihood of x with respect to the event probability parameter π, starting from the log-likelihood in terms of the log-odds $\eta \overset{\text{def}}{=} \text{logit}(\pi)$ in Corollary 0.1.

Solution

Solution 0.5. With $y_i = x$, Corollary 0.1 gives $\ell = x\eta + \log\{1-\pi\}$, so:

$$\begin{aligned}
\ell' &\stackrel{\text{def}}{=} \frac{\partial}{\partial\pi}\ell && \text{(definition of the score)}\\
&= \frac{\partial}{\partial\pi}(x\eta + \log\{1-\pi\}) && \text{(log-likelihood in terms of the log-odds)}\\
&= \frac{\partial}{\partial\pi}(x\eta) + \frac{\partial}{\partial\pi}\log\{1-\pi\} && \text{(linearity of differentiation)}\\
&= x\frac{\partial}{\partial\pi}\eta + \frac{\partial}{\partial\pi}\log\{1-\pi\} && \text{(constant multiple rule)}
\end{aligned}$$

The inner derivative in the second term is:

$$\begin{aligned}
\frac{\partial}{\partial\pi}(1-\pi) &= \frac{\partial}{\partial\pi}1 - \frac{\partial}{\partial\pi}\pi && \text{(linearity of differentiation)}\\
&= 0 - 1 && \text{(derivatives of a constant and of } \pi)\\
&= -1 && \text{(subtract)}
\end{aligned}$$

The second term is then:

$$\begin{aligned}
\frac{\partial}{\partial\pi}\log\{1-\pi\} &= \frac{1}{1-\pi}\frac{\partial}{\partial\pi}(1-\pi) && \text{(chain rule and derivative of log)}\\
&= \frac{1}{1-\pi}(-1) && \text{(inner derivative is } -1)\\
&= -\frac{1}{1-\pi} && \text{(multiply by } -1)
\end{aligned}$$

The derivative of the log-odds is:

$$\begin{aligned}
\frac{\partial}{\partial\pi}\eta &= \frac{\partial}{\partial\pi}\log\left\{\frac{\pi}{1-\pi}\right\} && \text{(definition of } \eta)\\
&= \frac{\partial}{\partial\pi}(\log\{\pi\} - \log\{1-\pi\}) && \text{(log of a quotient)}\\
&= \frac{\partial}{\partial\pi}\log\{\pi\} - \frac{\partial}{\partial\pi}\log\{1-\pi\} && \text{(linearity of differentiation)}\\
&= \frac{1}{\pi} - \left(-\frac{1}{1-\pi}\right) && \text{(derivative of log, and the second term above)}\\
&= \frac{1}{\pi} + \frac{1}{1-\pi} && \text{(subtracting a negative)}\\
&= \frac{1-\pi}{\pi(1-\pi)} + \frac{\pi}{\pi(1-\pi)} && \text{(write both terms over the common denominator)}\\
&= \frac{1}{\pi(1-\pi)} && \text{(add the numerators)}
\end{aligned}$$

Substituting both derivatives back in:

$$\begin{aligned}
\ell' &= x\left(\frac{1}{\pi(1-\pi)}\right) + \left(-\frac{1}{1-\pi}\right) && \text{(substitute the two derivatives)} \\
&= \frac{x}{\pi(1-\pi)} + \left(-\frac{1}{1-\pi}\right) && \text{(multiply } x \text{ into the numerator)} \\
&= \frac{x}{\pi(1-\pi)} - \frac{1}{1-\pi} && \text{(adding a negative is subtracting)} \\
&= \frac{x}{\pi(1-\pi)} - \frac{\pi}{\pi(1-\pi)} && \text{(multiply the numerator and denominator of the second term by } \pi) \\
&= \frac{x-\pi}{\pi(1-\pi)} && \text{(combine the fractions)}
\end{aligned}$$

Exercise 0.6 (Score of a Bernoulli variable in terms of deviation and precision)**.** In the setting of Exercise 0.5, let $\mathbb{d} \overset{\text{def}}{=} x - \mathbb{E}[X]$ be the deviation of x from the mean of X, and let $\tau \overset{\text{def}}{=} 1/\text{Var}(X)$ be the precision of X. Write the score using $\mathbb{d}$ and τ.

Solution

Solution 0.6. Starting from Solution 0.5:

$$\begin{aligned}
\ell' &= \frac{x-\pi}{\pi(1-\pi)} && \text{(score)} \\
&= \frac{x-\pi}{\text{Var}(X)} && (\text{Var}(X) = \pi(1-\pi)) \\
&= \frac{x-\mathbb{E}[X]}{\text{Var}(X)} && (\mathbb{E}[X] = \pi) \\
&= \frac{\mathbb{d}}{\text{Var}(X)} && \text{(definition of } \mathbb{d}) \\
&= \tau\mathbb{d} && \text{(definition of } \tau)
\end{aligned}$$

Theorem 0.4 (Score function of a Bernoulli variable)**.** *Let $X \sim \text{Ber}(\pi)$, with event probability π and observed value $x \in \{0, 1\}$. The score ℓ' of the single observation x is the derivative of its log-likelihood ℓ with respect to π. It equals the precision $\tau = 1/\text{Var}(X)$ of X times the deviation $\mathbb{d} \overset{def}{=} x - \mathbb{E}[X]$ of x from the mean $\mathbb{E}[X]$:*

$$\ell' = \tau\mathbb{d} = \frac{\mathbb{d}}{\text{Var}(X)} = \frac{x-\pi}{\pi(1-\pi)}$$

Proof

Proof. The score as $(x-\pi)/(\pi(1-\pi))$ is the solution to Exercise 0.5, worked in Solution 0.5. Writing the score as $\mathrm{d}/\mathrm{Var}(X)$ and then as $\tau\mathrm{d}$ is the solution to Exercise 0.6, worked in Solution 0.6. The proof builds on Corollary 0.1. □

Example 0.2 (Score of one Bernoulli observation)**.** For a Bernoulli variable with event probability π and observed value $x \in \{0, 1\}$, the score in Theorem 0.4 is $1/\pi$ when $x = 1$ and $-1/(1-\pi)$ when $x = 0$. When the event occurs, the log-likelihood increases as π increases. When the event does not occur, the log-likelihood decreases as π increases.

In all four examples (Bernoulli score, Theorem 0.6, Theorem 0.7, and Theorem 0.8), the score function with respect to the mean turned out to be:

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

Remark (The score function in natural exponential families)

Remark 0.1 (The score function in natural exponential families). 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 $(x - \mathbb{E}[X])/\mathrm{Var}(X)$. 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 [1, Ch. 3], [7, Sec. 6.7].

Exercise 0.7 (Score contribution of one binary outcome)**.** In the setting of Exercise 0.2, derive the score contribution $\ell_i'(\pi) \overset{\text{def}}{=} \frac{\partial}{\partial \pi}\ell_i(\pi)$ of observation i.

Solution

Solution 0.7. The score contribution is the score of the single observation y_i, so Theorem 0.4 with $x = y_i$ gives

$$\begin{aligned} \ell_i'(\pi) &= \frac{\partial}{\partial \pi}\ell_i(\pi) && \text{(definition of the score contribution)} \\ &= \frac{y_i - \pi}{\pi(1-\pi)} && \text{(score of a Bernoulli variable)} \end{aligned}$$

Exercise 0.8 (Score contribution in terms of deviation and precision)**.** In the setting of Exercise 0.7, let:

- $\mathbb{d}_i \stackrel{\text{def}}{=} y_i - \pi$ be the deviation of y_i from the mean π of Y_i;
- $\tau \stackrel{\text{def}}{=} 1/\text{Var}(Y_i)$ be the precision of Y_i.

Write the score contribution using $\mathbb{d}_i$ and τ.

Solution

Solution 0.8. Starting from Solution 0.7:

$$\begin{aligned}
\ell_i'(\pi) &= \frac{y_i - \pi}{\pi(1-\pi)} && \text{(score contribution)} \\
&= \frac{y_i - \pi}{\text{Var}(Y_i)} && (\text{Var}(Y_i) = \pi(1-\pi)) \\
&= \frac{y_i - \mathbb{E}[Y_i]}{\text{Var}(Y_i)} && (\mathbb{E}[Y_i] = \pi) \\
&= \frac{\mathbb{d}_i}{\text{Var}(Y_i)} && (\text{definition of } \mathbb{d}_i) \\
&= \tau \mathbb{d}_i && (\text{definition of } \tau)
\end{aligned}$$

Theorem 0.5 (Score contribution of one binary outcome)**.** *If $Y_i \sim \text{Ber}(\pi)$ has observed value $y_i \in \{0, 1\}$ and precision $\tau = 1/\text{Var}(Y_i) = 1/(\pi(1-\pi))$, its score contribution is the precision times the deviation $\mathbb{d}_i \stackrel{def}{=} y_i - \pi$ of y_i from its mean π:*

$$\ell_i'(\pi) = \tau \mathbb{d}_i$$

The precision τ is the same for every i, because all Y_i share the event probability π.

Proof

Proof. The score contribution as $(y_i - \pi)/(\pi(1-\pi))$ is the solution to Exercise 0.7, worked in Solution 0.7. The form in terms of the deviation is the solution to Exercise 0.8, worked in Solution 0.8. □

The score function of the sample

Exercise 0.9 (Sample mean of the deviations of a binary sample)**.** In the setting of Exercise 0.7, let:

- $\bar{y} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} y_i$ be the sample mean of the observed outcomes;
- $\bar{\mathbb{d}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{d}_i$ be the sample mean of the deviations $\mathbb{d}_i$.

Write $\bar{\mathbb{d}}$ using $\bar{y}$ and π.

Solution

Solution 0.9.

$$
\begin{aligned}
\bar{\mathbb{d}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{d}_i && \text{(definition of } \bar{\mathbb{d}}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} (y_i - \pi) && \text{(definition of } \mathbb{d}_i\text{)} \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi\right) && \text{(distribute the sum over the difference)} \\
&= \frac{1}{n}\left(s_{\underset{\sim}{y}} - \sum_{i=1}^{n} \pi\right) && \text{(definition of } s_{\underset{\sim}{y}}\text{)} \\
&= \frac{1}{n}(n\bar{y} - n\pi) && (s_{\underset{\sim}{y}} = n\bar{y} \text{ and } \textstyle\sum_{i=1}^{n} \pi = n\pi) \\
&= \frac{1}{n} \cdot n(\bar{y} - \pi) && \text{(factor out } n\text{)} \\
&= \bar{y} - \pi && (\tfrac{1}{n} \cdot n = 1)
\end{aligned}
$$

Exercise 0.10 (Score function of a binary sample)**.** In the setting of Exercise 0.9, assemble the score function $\ell'(\pi)$ of the whole sample from the score contributions, and write it using $\bar{\mathbb{d}}$ and τ.

Solution

Solution 0.10.

$$\begin{aligned}
\ell'(\pi) &= \sum_{i=1}^{n} \ell_i'(\pi) && \text{(score of an independent sample is the sum of the score contributions)} \\
&= \sum_{i=1}^{n} \tau \mathbb{d}_i && \text{(score contribution of a binary outcome)} \\
&= \tau \sum_{i=1}^{n} \mathbb{d}_i && (\tau \text{ is the same for every } i) \\
&= \tau \cdot n\bar{\mathbb{d}} && (\textstyle\sum_{i=1}^{n} \mathbb{d}_i = n\bar{\mathbb{d}}, \text{ from the definition of } \bar{\mathbb{d}}) \\
&= n\tau\bar{\mathbb{d}} && \text{(reorder the factors)}
\end{aligned}$$

Theorem 0.6 (Score function of a binary sample)**.** *If* $Y_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi)$ *for* $i = 1, \ldots, n$*, with observed sample mean* $\bar{y}$*, precision* $\tau = 1/(\pi(1-\pi))$*, and sample mean* $\bar{\mathbb{d}} \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{d}_i$ *of the deviations* $\mathbb{d}_i = y_i - \pi$*, the score function is:*

$$\ell'(\pi) = n\tau\bar{\mathbb{d}} = n\tau(\bar{y} - \pi)$$

Proof

Proof. The first equality is the solution to Exercise 0.10, worked in Solution 0.10. The second is the solution to Exercise 0.9, worked in Solution 0.9. □

Solving the score equation

Exercise 0.11 (Solving the score equation of a binary sample)**.** In the setting of Exercise 0.10, assume $\bar{y} \in (0, 1)$: the sample has at least one event and at least one non-event. Solve the score equation $\ell'(\pi) = 0$ for $\pi \in (0, 1)$.

Solution

Solution 0.11. For $\pi \in (0, 1)$, $\tau = 1/(\pi(1-\pi))$ is finite and positive, and $n > 0$, so $n\tau \neq 0$.

$$
\begin{aligned}
0 &= \ell'(\pi) && \text{(score equation)} \\
0 &= n\tau(\bar{y} - \pi) && \text{(score function of a binary sample)} \\
0 &= \bar{y} - \pi && \text{(divide both sides by } n\tau \neq 0\text{)} \\
\pi &= \bar{y} && \text{(add } \pi \text{ to both sides)}
\end{aligned}
$$

Since $\bar{y} \in (0, 1)$, this solution lies in $(0, 1)$.

Theorem 0.7 (Solution of the score equation of a binary sample)**.** *If* $Y_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi)$ *for* $i = 1, \ldots, n$ *and the observed sample mean satisfies* $\bar{y} \in (0, 1)$*, then the score equation* $\ell'(\pi) = 0$ *has exactly one solution in* $(0, 1)$*: the sample mean* $\pi = \bar{y}$*.*

Proof

Proof. This is the solution to Exercise 0.11, worked in Solution 0.11. □

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

Hessian contributions

Exercise 0.12 (Hessian contribution of one binary outcome)**.** In the setting of Exercise 0.7, derive the Hessian contribution $\ell_i''(\pi) \overset{\text{def}}{=} \frac{\partial}{\partial \pi}\ell_i'(\pi)$ of observation i. Start from the score contribution written as a difference of two fractions, $\ell_i'(\pi) = y_i/\pi - (1-y_i)/(1-\pi)$, the form in Solution 0.5 before its fractions are combined.

Solution

Solution 0.12.

$$\begin{aligned}
\ell_i''(\pi) &= \frac{\partial}{\partial \pi}\ell_i'(\pi) && \text{(definition of the Hessian contribution)}\\
&= \frac{\partial}{\partial \pi}\left(\frac{y_i}{\pi} - \frac{1-y_i}{1-\pi}\right) && \text{(score contribution as a difference of fractions)}\\
&= \frac{\partial}{\partial \pi}(y_i\pi^{-1} - (1-y_i)(1-\pi)^{-1}) && \text{(write each fraction as a negative power)}\\
&= y_i\frac{\partial}{\partial \pi}\pi^{-1} - (1-y_i)\frac{\partial}{\partial \pi}(1-\pi)^{-1} && \text{(linearity and constant multiple rule)}
\end{aligned}$$

The first derivative is:

$$\frac{\partial}{\partial \pi}\pi^{-1} = -\pi^{-2} \qquad \text{(power rule)}$$

The second derivative needs the chain rule:

$$\begin{aligned}
\frac{\partial}{\partial \pi}(1-\pi)^{-1} &= -(1-\pi)^{-2}\frac{\partial}{\partial \pi}(1-\pi) && \text{(chain rule and power rule)}\\
&= -(1-\pi)^{-2}(-1) && (\frac{\partial}{\partial \pi}(1-\pi) = -1, \text{ as in the score exercise})\\
&= (1-\pi)^{-2} && \text{(multiply)}
\end{aligned}$$

Substituting both derivatives back in:

$$\begin{aligned}
\ell_i''(\pi) &= y_i(-\pi^{-2}) - (1-y_i)(1-\pi)^{-2} && \text{(substitute the two derivatives)}\\
&= -\frac{y_i}{(\pi)^2} - \frac{1-y_i}{(1-\pi)^2} && \text{(write the negative powers as fractions)}
\end{aligned}$$

Theorem 0.8 (Hessian contribution of one binary outcome)**.** *If* $Y_i \sim \text{Ber}(\pi)$ *has observed value* $y_i \in \{0, 1\}$*, its Hessian contribution is:*

$$\ell_i''(\pi) = -\frac{y_i}{(\pi)^2} - \frac{1-y_i}{(1-\pi)^2}$$

Proof

Proof. This is the solution to Exercise 0.12, worked in Solution 0.12. □

Lemma 0.1 (Derivative of the Bernoulli precision)**.** *If Y is a Bernoulli variable, $Y \sim \text{Ber}(\pi)$ with $\pi \in (0,1)$ and precision $\tau = 1/(\pi(1-\pi))$, then the derivative of τ with respect to π is*

$$\frac{\partial}{\partial \pi}\tau = -(\tau)^2(1-2\pi)$$

Proof

Proof. The derivative needs the chain rule:

$$\begin{aligned}
\frac{\partial}{\partial \pi}\tau &= \frac{\partial}{\partial \pi}(\pi(1-\pi))^{-1} && \text{(definition of } \tau \text{ as a negative power)} \\
&= \frac{\partial}{\partial \pi}\left(\pi - (\pi)^2\right)^{-1} && \text{(distribute } \pi\text{)} \\
&= -\left(\pi - (\pi)^2\right)^{-2}\frac{\partial}{\partial \pi}\left(\pi - (\pi)^2\right) && \text{(chain rule and power rule)} \\
&= -\left(\pi - (\pi)^2\right)^{-2}(1-2\pi) && \text{(linearity and power rule)} \\
&= -(\pi(1-\pi))^{-2}(1-2\pi) && \text{(factor } \pi - (\pi)^2 = \pi(1-\pi)\text{)} \\
&= -(\tau)^2(1-2\pi) && \text{(definition of } \tau\text{)}
\end{aligned}$$

□

Example 0.4 (Derivative of the precision at $\pi = 1/4$)**.** At $\pi = 1/4$, the precision is

$$\begin{aligned}
\tau &= \frac{1}{\pi(1-\pi)} && \text{(definition of } \tau\text{)} \\
&= \frac{1}{\frac{1}{4}\cdot\frac{3}{4}} && \text{(substitute } \pi = 1/4\text{)} \\
&= \frac{16}{3} && \text{(multiply the fractions and take the reciprocal)}
\end{aligned}$$

so by Lemma 0.1,

$$\begin{aligned}
\frac{\partial}{\partial \pi}\tau &= -(\tau)^2(1-2\pi) && \text{(@lem-bernoulli-precision-deriv)} \\
&= -\left(\left(\frac{16}{3}\right)\right)^2\left(1-\frac{1}{2}\right) && \text{(substitute } \tau = 16/3 \text{ and } \pi = 1/4\text{)} \\
&= -\frac{256}{9}\cdot\frac{1}{2} && \text{(square and subtract)} \\
&= -\frac{128}{9} && \text{(multiply the fractions)}
\end{aligned}$$

The precision falls as π rises toward $1/2$, where it is smallest, which matches a derivative below zero.

Exercise 0.13 (Hessian contribution of one binary outcome in terms of the precision and deviation)**.** In the setting of Exercise 0.12, derive the Hessian contribution $\ell_i''(\pi)$ starting from the simplified score contribution $\ell_i'(\pi) = \tau \mathbb{d}_i$ in Theorem 0.5, with the precision $\tau = 1/(\pi(1-\pi))$ and the deviation $\mathbb{d}_i \overset{\text{def}}{=} y_i - \pi$. Both τ and $\mathbb{d}_i$ depend on π, while the observed value y_i does not.

Solution

Solution 0.13.

$$\begin{aligned}
\ell_i''(\pi) &= \frac{\partial}{\partial \pi} \ell_i'(\pi) && \text{(definition of the Hessian contribution)} \\
&= \frac{\partial}{\partial \pi} (\tau \mathbb{d}_i) && \text{(score contribution)} \\
&= \left(\frac{\partial}{\partial \pi} \tau \right) \mathbb{d}_i + \tau \frac{\partial}{\partial \pi} \mathbb{d}_i && \text{(product rule)}
\end{aligned}$$

The derivative of the deviation is:

$$\begin{aligned}
\frac{\partial}{\partial \pi} \mathbb{d}_i &= \frac{\partial}{\partial \pi} (y_i - \pi) && \text{(definition of } \mathbb{d}_i) \\
&= \frac{\partial}{\partial \pi} y_i - \frac{\partial}{\partial \pi} \pi && \text{(linearity of differentiation)} \\
&= 0 - 1 && \text{(derivatives of a constant and of } \pi) \\
&= -1 && \text{(subtract)}
\end{aligned}$$

The derivative of the precision is given by Lemma 0.1:

$$\frac{\partial}{\partial \pi} \tau = -(\tau)^2 (1 - 2\pi)$$

Substituting both derivatives back in:

$$\begin{aligned}
\ell_i''(\pi) &= \Big(-(\tau)^2 (1 - 2\pi) \Big) \mathbb{d}_i + \tau(-1) && \text{(substitute the two derivatives)} \\
&= -\tau - (1 - 2\pi)(\tau)^2 \mathbb{d}_i && \text{(multiply by } -1 \text{ and reorder the terms)}
\end{aligned}$$

Corollary 0.2 (Hessian contribution of one binary outcome in terms of the precision and deviation)**.** *If* $Y_i \sim \text{Ber}(\pi)$ *has observed value* $y_i \in \{0, 1\}$ *and precision* $\tau = 1/(\pi(1-\pi))$,

its Hessian contribution is, with the deviation $\mathbb{d}_i \overset{def}{=} y_i - \pi$*:*

$$\ell_i''(\pi) = -\tau - (1-2\pi)(\tau)^2\mathbb{d}_i$$

i Proof

Proof. This is the solution to Exercise 0.13, worked in Solution 0.13. It follows from Theorem 0.8. □

Exercise 0.14 (Hessian contribution of one binary outcome as a squared deviation)**.** In the setting of Exercise 0.13, starting from the Hessian contribution $\ell_i''(\pi)$ in Corollary 0.2, write $\ell_i''(\pi)$ with the precision $\tau = 1/(\pi(1-\pi))$ as a multiple of $(\tau)^2$ times the square of the deviation $\mathbb{d}_i \overset{\text{def}}{=} y_i - \pi$. Hint: because $y_i \in \{0,1\}$, $(y_i)^2 = y_i$.

Solution

Solution 0.14. Starting from Corollary 0.2, factor out $(\tau)^2$. The reciprocal of the precision is $1/\tau = \pi(1-\pi)$, so:

$$\begin{aligned}
\ell_i''(\pi) &= -\tau - (1-2\pi)(\tau)^2\mathbb{d}_i && \text{(the contribution from the corollary)}\\
&= -(\tau)^2\frac{1}{\tau} - (1-2\pi)(\tau)^2\mathbb{d}_i && (\tau = (\tau)^2/\tau)\\
&= -(\tau)^2(\pi(1-\pi) + (1-2\pi)\mathbb{d}_i) && \text{(factor out } -(\tau)^2)
\end{aligned}$$

The factor in parentheses is:

$$\begin{aligned}
\pi(1-\pi) + (1-2\pi)\mathbb{d}_i &= \pi - (\pi)^2 + (1-2\pi)(y_i - \pi) && \text{(distribute } \pi\text{; definition of } \mathbb{d}_i)\\
&= \pi - (\pi)^2 + y_i - \pi - 2\pi y_i + 2(\pi)^2 && \text{(distribute } 1-2\pi)\\
&= y_i - 2\pi y_i + \left(2(\pi)^2 - (\pi)^2\right) + (\pi - \pi) && \text{(reorder and group the terms)}\\
&= y_i - 2\pi y_i + (\pi)^2 + 0 && (2a - a = a \text{ and } a - a = 0)\\
&= y_i - 2\pi y_i + (\pi)^2 && (a + 0 = a)\\
&= (y_i)^2 - 2\pi y_i + (\pi)^2 && (y_i \in \{0,1\}, \text{ so } y_i = (y_i)^2)\\
&= ((y_i - \pi))^2 && \text{(square of a difference)}\\
&= (\mathbb{d}_i)^2 && \text{(definition of } \mathbb{d}_i)
\end{aligned}$$

Substituting the factor back in:

$$\ell_i''(\pi) = -(\tau)^2(\mathbb{d}_i)^2 \qquad \text{(substitute } (\mathbb{d}_i)^2)$$

Corollary 0.3 (Hessian contribution of one binary outcome as a squared deviation)**.** *If* $Y_i \sim \mathrm{Ber}(\pi)$ *has observed value* $y_i \in \{0,1\}$*, its Hessian contribution, written with the precision* $\tau = 1/(\pi(1-\pi))$ *and the deviation* $\mathbb{d}_i \overset{def}{=} y_i - \pi$*, is:*

$$\ell_i''(\pi) = -(\tau)^2(\mathbb{d}_i)^2$$

Proof

Proof. This is the solution to Exercise 0.14, worked in Solution 0.14. It follows from Corollary 0.2. □

The Hessian of the sample

Exercise 0.15 (Hessian of a binary sample)**.** In the setting of Exercise 0.10, assemble the Hessian $\ell''(\pi)$ of the whole sample from the Hessian contributions.

Solution

Solution 0.15.

$$\begin{aligned}
\ell''(\pi) &= \sum_{i=1}^{n} \ell_i''(\pi) && \text{(the derivative of a sum is the sum of the derivatives)} \\
&= \sum_{i=1}^{n} \left(-\frac{y_i}{(\pi)^2} - \frac{1-y_i}{(1-\pi)^2} \right) && \text{(Hessian contribution of a binary outcome)} \\
&= -\frac{1}{(\pi)^2} \sum_{i=1}^{n} y_i - \frac{1}{(1-\pi)^2} \sum_{i=1}^{n} (1-y_i) && \text{(distribute the sum over the terms; the denominators do not de} \\
&= -\frac{1}{(\pi)^2} s_{\underset{\sim}{y}} - \frac{1}{(1-\pi)^2} \sum_{i=1}^{n} (1-y_i) && \text{(definition of } s_{\underset{\sim}{y}}\text{)} \\
&= -\frac{n\bar{y}}{(\pi)^2} - \frac{1}{(1-\pi)^2} (n - n\bar{y}) && (s_{\underset{\sim}{y}} = n\bar{y} \text{ and } \textstyle\sum_{i=1}^{n} 1 = n) \\
&= -\frac{n\bar{y}}{(\pi)^2} - \frac{n(1-\bar{y})}{(1-\pi)^2} && \text{(factor out } n\text{)} \\
&= -n \left(\frac{\bar{y}}{(\pi)^2} + \frac{1-\bar{y}}{(1-\pi)^2} \right) && \text{(factor out } -n\text{)}
\end{aligned}$$

Theorem 0.9 (Hessian of a binary sample)**.** *If* $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Ber}(\pi)$ *for* $i = 1, \dots, n$*, with observed sample mean* $\bar{y}$*, the Hessian of the log-likelihood is the* 1×1 *matrix whose only entry is:*

$$\ell''(\pi) = -n\left(\frac{\bar{y}}{(\pi)^2} + \frac{1-\bar{y}}{(1-\pi)^2}\right)$$

Proof

Proof. This is the solution to Exercise 0.15, worked in Solution 0.15. □

Checking that the Hessian is negative definite

Exercise 0.16 (Negative definiteness of the Hessian of a binary sample)**.** In the setting of Exercise 0.15, show that the Hessian $\ell''(\pi)$ is negative definite for every $\pi \in (0, 1)$.

Solution

Solution 0.16. A 1×1 matrix is negative definite exactly when its only entry is negative, so it is enough to show $\ell''(\pi) < 0$.
Fix $\pi \in (0, 1)$. Then:

- $(\pi)^2 > 0$ and $(1-\pi)^2 > 0$;
- $\bar{y} \geq 0$ and $1 - \bar{y} \geq 0$, because $y_i \in \{0, 1\}$;
- $\bar{y}$ and $1 - \bar{y}$ are not both zero, because they sum to 1.

So $\bar{y}/(\pi)^2 + (1-\bar{y})/(1-\pi)^2$ is a sum of two nonnegative terms, at least one of them positive. It is therefore positive, and with $n > 0$:

$$\ell''(\pi) = -n\left(\frac{\bar{y}}{(\pi)^2} + \frac{1-\bar{y}}{(1-\pi)^2}\right) < 0$$

This holds for every $\pi \in (0, 1)$.

Theorem 0.10 (Negative definiteness of the Hessian of a binary sample)**.** *If* $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Ber}(\pi)$ *for* $i = 1, \dots, n$*, the Hessian* $\ell''(\pi)$ *of the log-likelihood is negative definite for every* $\pi \in (0, 1)$*, whatever the observed outcomes. Hence the log-likelihood is strictly concave on* $(0, 1)$*.*

Proof

Proof. This is the solution to Exercise 0.16, worked in Solution 0.16. It builds on Theorem 0.9. □

The maximum likelihood estimate

Exercise 0.17 (Maximum likelihood estimate of a binary sample)**.** In the setting of Exercise 0.11, use Theorem 0.10 to find the maximum likelihood estimate $\hat{\pi}_{\text{ML}}$ of π.

Solution

Solution 0.17. By Theorem 0.10, ℓ is strictly concave on $(0,1)$. A strictly concave function that has a critical point is maximized there, and nowhere else. With $\bar{y} \in (0,1)$, Theorem 0.7 says the only critical point of ℓ in $(0,1)$ is $\bar{y}$. So ℓ is maximized at $\pi = \bar{y}$.

Theorem 0.11 (Maximum likelihood estimate of a binary sample)**.** *If* $Y_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi)$ *for* $i = 1, \dots, n$ *and the observed sample mean satisfies* $\bar{y} \in (0,1)$*, the maximum likelihood estimate of* π *is the sample mean:*

$$\hat{\pi}_{ML} = \bar{y}$$

Proof

Proof. This is the solution to Exercise 0.17, worked in Solution 0.17. It builds on Theorem 0.7 and Theorem 0.10. □

Maximum likelihood for univariate Gaussian models

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

By the likelihood of independent data, 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 the log-likelihood of independent data, the log-likelihood is:

$$
\begin{aligned}
\ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \log\left\{(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right\} && \text{(log of each Gaussian density)} \\
&= \sum_{i=1}^{n}\left(\log\left\{(2\pi\sigma^2)^{-1/2}\right\} + \log\left\{\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right\}\right) && \text{(log of a product)} \\
&= \sum_{i=1}^{n}\left(-\frac{1}{2}\log\{2\pi\sigma^2\} + \log\left\{\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right\}\right) && \text{(log of a power)} \\
&= \sum_{i=1}^{n}\left(-\frac{1}{2}\log\{2\pi\sigma^2\} - \frac{(x_i-\mu)^2}{2\sigma^2}\right) && \text{(log undoes exp)} \\
&= \sum_{i=1}^{n}\left(-\frac{1}{2}\log\{2\pi\sigma^2\}\right) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} && \text{(distribute the sum over the terms)} \\
&= -\frac{n}{2}\log\{2\pi\sigma^2\} - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} && \text{(sum of } n \text{ identical terms)} \\
&= -\frac{n}{2}\log\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(factor the constant } \tfrac{1}{2\sigma^2} \text{ out of the sum)} \\
&= -\frac{n}{2}(\log\{2\pi\} + \log\{\sigma^2\}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(log of a product)} \\
&= -\frac{n}{2}\log\{2\pi\} - \frac{n}{2}\log\{\sigma^2\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(distribute } -\tfrac{n}{2}\text{)}
\end{aligned}
$$

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:

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\ell &= \frac{\partial}{\partial\mu}\left(-\frac{n}{2}\log\{2\pi\} - \frac{n}{2}\log\{\sigma^2\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(Gaussian log-likelihood)}\\
&= \frac{\partial}{\partial\mu}\Big(-\frac{n}{2}\log\{2\pi\}\Big) + \frac{\partial}{\partial\mu}\Big(-\frac{n}{2}\log\{\sigma^2\}\Big) + \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(linearity of differentiation)}\\
&= 0 + 0 + \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(derivative of a constant)}\\
&= \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(drop the zero terms)}\\
&= -\frac{1}{2\sigma^2}\frac{\partial}{\partial\mu}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(constant multiple rule)}\\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial\mu}(x_i-\mu)^2 && \text{(derivative of a sum is the sum of d}\\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)\frac{\partial}{\partial\mu}(x_i-\mu) && \text{(chain rule, outer function } u^2\text{)}
\end{aligned}
$$

The inner derivative is:

$$
\begin{aligned}
\frac{\partial}{\partial\mu}(x_i-\mu) &= \frac{\partial}{\partial\mu}x_i - \frac{\partial}{\partial\mu}\mu && \text{(linearity of differentiation)}\\
&= 0 - \frac{\partial}{\partial\mu}\mu && (x_i \text{ does not depend on } \mu)\\
&= 0 - 1 && \text{(derivative of } \mu \text{ with respect to itself)}\\
&= -1 && \text{(subtract)}
\end{aligned}
$$

Substituting the inner derivative back in:

$$
\begin{aligned}
\frac{\partial}{\partial \mu}\ell &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)\cdot(-1) && \text{(inner derivative is } -1\text{)} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(x_i-\mu) && \text{(multiply)} \\
&= -\frac{1}{2\sigma^2}\cdot(-2)\sum_{i=1}^{n}(x_i-\mu) && \text{(factor the constant } -2 \text{ out of the sum)} \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) && \text{(multiply the constants)} \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu\right) && \text{(distribute the sum over the terms)} \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) && \text{(sum of } n \text{ copies of } \mu\text{)}
\end{aligned}
$$

For the second entry, write the population variance σ^2 as a single variable v, so that the log-likelihood ℓ contains $-\frac{n}{2}\log v$ and $-\frac{1}{2}v^{-1}S$, where the sum of squared deviations $S = \sum_{i=1}^{n}\left(\left(\mathbb{D}_{\mu}(x_i)\right)^2\right)$ from the mean parameter μ does not depend on σ^2:

$$
\begin{aligned}
\frac{\partial}{\partial \sigma^2}\ell &= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\{2\pi\} - \frac{n}{2}\log\{\sigma^2\} - \frac{1}{2}(\sigma^2)^{-1}S\right) && \text{(Gaussian log-likelihood)} \\
&= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\{2\pi\}\right) + \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\{\sigma^2\}\right) + \frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2}(\sigma^2)^{-1}S\right) && \text{(linearity of differentiation)} \\
&= 0 + \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\{\sigma^2\}\right) + \frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2}(\sigma^2)^{-1}S\right) && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\{\sigma^2\}\right) + \frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2}(\sigma^2)^{-1}S\right) && \text{(drop the zero term)} \\
&= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\{\sigma^2\}\right) + \frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2}S(\sigma^2)^{-1}\right) && \text{(reorder the factors)} \\
&= -\frac{n}{2}\frac{\partial}{\partial \sigma^2}\log\{\sigma^2\} - \frac{1}{2}S\frac{\partial}{\partial \sigma^2}(\sigma^2)^{-1} && \text{(constant multiple rule)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} - \frac{1}{2}S\frac{\partial}{\partial \sigma^2}(\sigma^2)^{-1} && \text{(derivative of } \log v\text{)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} - \frac{1}{2}S\left(-(\sigma^2)^{-2}\right) && \text{(power rule for } v^{-1}\text{)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} + \frac{1}{2}S(\sigma^2)^{-2} && \text{(multiply)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2}S && \text{(reorder the factors)}
\end{aligned}
$$

MLE of μ

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

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

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

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu^2} &= \frac{\partial}{\partial \mu}\left(\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)\right) && \text{(differentiate the score for } \mu\text{)} \\
&= \frac{1}{\sigma^2}\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{n} x_i - n\mu\right) && \text{(constant multiple rule)} \\
&= \frac{1}{\sigma^2}\left(\frac{\partial}{\partial \mu}\sum_{i=1}^{n} x_i - \frac{\partial}{\partial \mu}(n\mu)\right) && \text{(linearity of differentiation)} \\
&= \frac{1}{\sigma^2}\left(0 - \frac{\partial}{\partial \mu}(n\mu)\right) && (\textstyle\sum_{i=1}^{n} x_i \text{ does not depend on } \mu) \\
&= \frac{1}{\sigma^2}\left(0 - n\frac{\partial}{\partial \mu}\mu\right) && \text{(constant multiple rule)} \\
&= \frac{1}{\sigma^2}(0 - n) && \text{(derivative of } \mu \text{ with respect to itself)} \\
&= \frac{1}{\sigma^2}(-n) && \text{(subtract)} \\
&= -\frac{n}{\sigma^2} && \text{(multiply)} \\
&< 0, && (n > 0 \text{ and } \sigma^2 > 0)
\end{aligned}
$$

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

The profile log-likelihood

Definition 0.1 (Profile log-likelihood)**.** Split a parameter vector into (ψ, λ), and for each fixed value of ψ let $\hat{\lambda}(\psi)$ maximize $\ell(\psi, \lambda)$ over λ. The **profile log-likelihood** of ψ is

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

Example 0.1 (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 (Gaussian MLEs), so the profile log-likelihood of σ^2 is

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

MLE of σ^2

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

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

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

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

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

> **Remark** (Differentiating with respect to the variance or the precision)
>
> *Remark* 0.1 (Differentiating with respect to the variance or the precision). 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, $\hat{\tau}_{\text{ML}} = 1/\hat{\sigma}^2_{\text{ML}}$.

> **Remark** (The Gaussian variance MLE is biased)
>
> *Remark* 0.2 (The Gaussian variance MLE is biased). 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{\partial}{\partial \sigma^2}\Big(-\frac{n}{2}(\sigma^2)^{-1}\Big) + \frac{\partial}{\partial \sigma^2}\left(\frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(linearity of differentiation)} \\
&= \frac{\partial}{\partial \sigma^2}\Big(-\frac{n}{2}(\sigma^2)^{-1}\Big) + \frac{\partial}{\partial \sigma^2}\left(\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(\sigma^2)^{-2}\right) && \text{(reorder the factors)} \\
&= -\frac{n}{2}\frac{\partial}{\partial \sigma^2}(\sigma^2)^{-1} + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\frac{\partial}{\partial \sigma^2}(\sigma^2)^{-2} && \text{(constant multiple rule)} \\
&= -\frac{n}{2}(-(\sigma^2)^{-2}) + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\frac{\partial}{\partial \sigma^2}(\sigma^2)^{-2} && \text{(power rule for } v^{-1}\text{)} \\
&= -\frac{n}{2}(-(\sigma^2)^{-2}) + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(-2(\sigma^2)^{-3}) && \text{(power rule for } v^{-2}\text{)} \\
&= \frac{n}{2}(\sigma^2)^{-2} - \sum_{i=1}^{n}(x_i-\mu)^2(\sigma^2)^{-3} && \text{(multiply the constants)} \\
&= \frac{n}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(reorder the factors)} \\
\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} &= \frac{\partial}{\partial \mu}\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{\partial}{\partial \mu}\Big(-\frac{n}{2}(\sigma^2)^{-1}\Big) + \frac{\partial}{\partial \mu}\left(\frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(linearity of differentiation)} \\
&= 0 + \frac{\partial}{\partial \mu}\left(\frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial \mu}\left(\frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2\right) && \text{(drop the zero term)} \\
&= \frac{1}{2}(\sigma^2)^{-2}\frac{\partial}{\partial \mu}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(constant multiple rule)} \\
&= \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(x_i-\mu)^2 && \text{(derivative of a sum is the sum of derivatives)} \\
&= \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}-2(x_i-\mu) && \text{(chain-rule result from the score for } \mu\text{)} \\
&= \frac{1}{2}(\sigma^2)^{-2}\cdot(-2)\sum_{i=1}^{n}(x_i-\mu) && \text{(factor the constant } -2 \text{ out of the sum)} \\
&= -(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu) && \text{(multiply the constants)}
\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:

$$\begin{aligned}
\left.\frac{\partial^2 \ell}{\partial (\sigma^2)^2}\right|_{\mathrm{MLE}} &= \frac{n}{2}(\hat{\sigma}^2)^{-2} - (\hat{\sigma}^2)^{-3}\sum_{i=1}^n (x_i - \bar{x})^2 && \text{(evaluate at } \mu = \bar{x},\ \sigma^2 = \hat{\sigma}^2\text{)} \\
&= \frac{n}{2}(\hat{\sigma}^2)^{-2} - (\hat{\sigma}^2)^{-3} n\hat{\sigma}^2 && \text{(substitute } \textstyle\sum_{i=1}^n (x_i - \bar{x})^2 = n\hat{\sigma}^2\text{)} \\
&= \frac{n}{2}(\hat{\sigma}^2)^{-2} - n(\hat{\sigma}^2)^{-3}\hat{\sigma}^2 && \text{(reorder the factors)} \\
&= \frac{n}{2}(\hat{\sigma}^2)^{-2} - n(\hat{\sigma}^2)^{-2} && \text{(cancel one factor of } \hat{\sigma}^2\text{)} \\
&= -\frac{n}{2}(\hat{\sigma}^2)^{-2} && \text{(combine like terms)} \\
\left.\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2}\right|_{\mathrm{MLE}} &= -(\hat{\sigma}^2)^{-2}\sum_{i=1}^n (x_i - \bar{x}) && \text{(evaluate at } \mu = \bar{x},\ \sigma^2 = \hat{\sigma}^2\text{)} \\
&= -(\hat{\sigma}^2)^{-2} \cdot 0 && \text{(substitute } \textstyle\sum_{i=1}^n (x_i - \bar{x}) = 0\text{)} \\
&= 0 && \text{(multiply by zero)}
\end{aligned}$$

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 the asymptotic distribution of the MLE, 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 [3], 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) [24].

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 [24]. Women were followed for an average of 4.1 years [24].

The primary outcome was nonfatal myocardial infarction or CHD death [24].

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. [25] 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)
```

The `rmb` R package includes the same file, which these notes use so that rendering does not depend on the website (Table 5):

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
```

Table 5: The first rows of the HERS dataset

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

```
summary(glucose_data)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    80.0    91.5    98.0    98.7   105.0   125.0
```

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.

Maximum likelihood estimates

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

```
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.

```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)
```

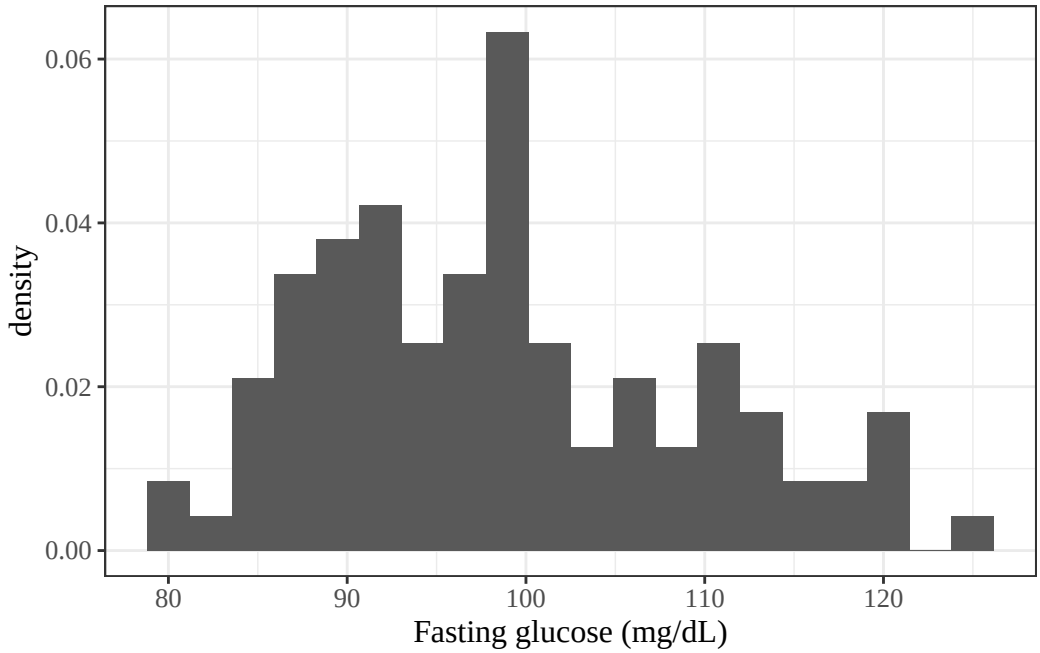


Figure 14: Fasting glucose among 100 HERS participants without diabetes who do not exercise

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

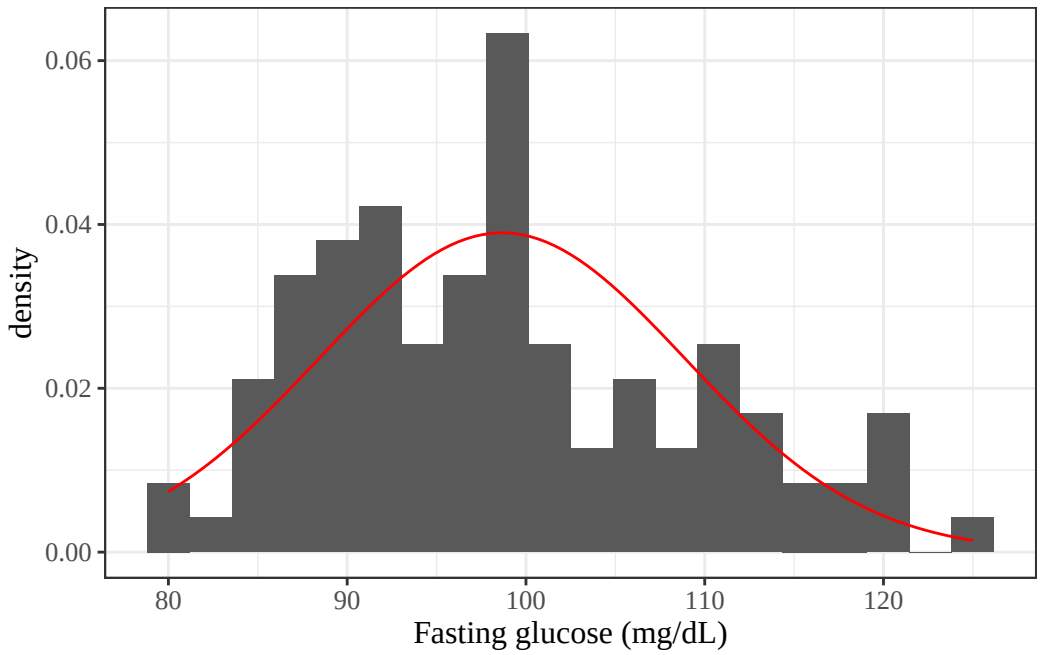


Figure 15: Fasting glucose, with the fitted Gaussian density in red

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(...))

# log-likelihood at the MLEs:
loglik(mu = mu_hat, sigma = sigma_hat, x = glucose_data)
#> [1] -374.47
```

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

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

Log-likelihood surface

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

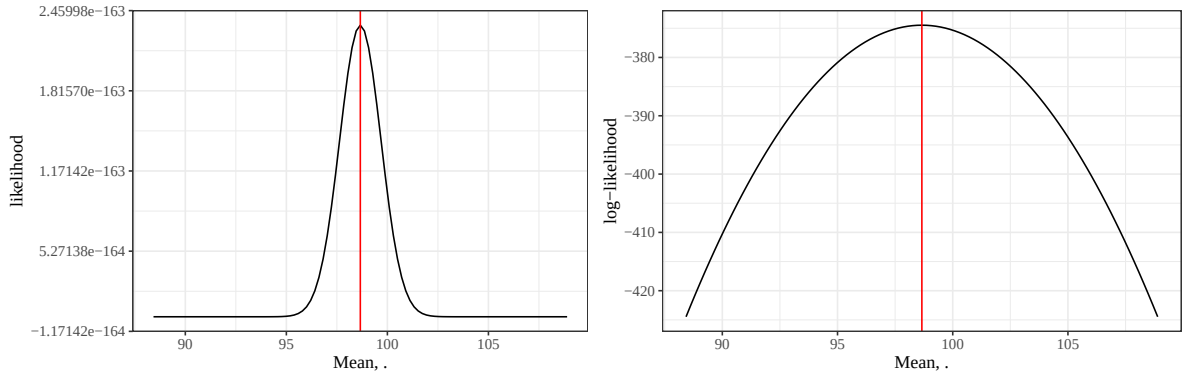
Standard errors by sample size

By the covariance matrix of the Gaussian MLEs, the estimated standard error of $\hat{\mu}_{\text{ML}}$ is

$$\begin{aligned}
\widehat{\text{SE}}(\hat{\mu}) &= \sqrt{\left[(I(\hat{\mu}, \hat{\sigma}^2))^{-1}\right]_{11}} && \text{(estimated standard error from the inverse information)} \\
&= \sqrt{\frac{\hat{\sigma}^2}{n}} && \text{(first diagonal entry of the inverse information)} \\
&= \frac{\hat{\sigma}}{\sqrt{n}} && \text{(square root of a quotient)}
\end{aligned}$$

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

```r
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("Mean,  ") +
  ggplot2::ylab("log-likelihood") +
  ggplot2::geom_vline(xintercept = mu_hat, col = "red"
```



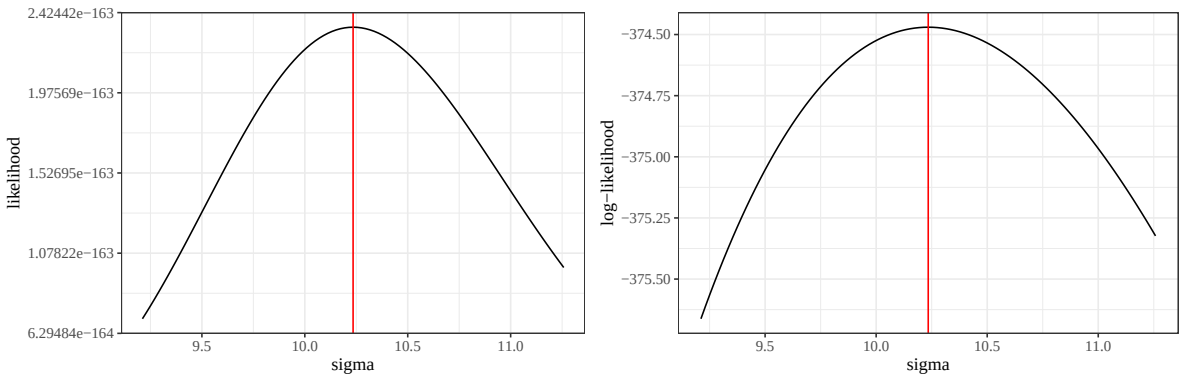
(a) Likelihood

(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 $\hat{\mu}_{\text{ML}}$

```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 $\hat{\sigma}_{\text{ML}}$

```r
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 = "Mean,  "),
      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)

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

```
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)")
```

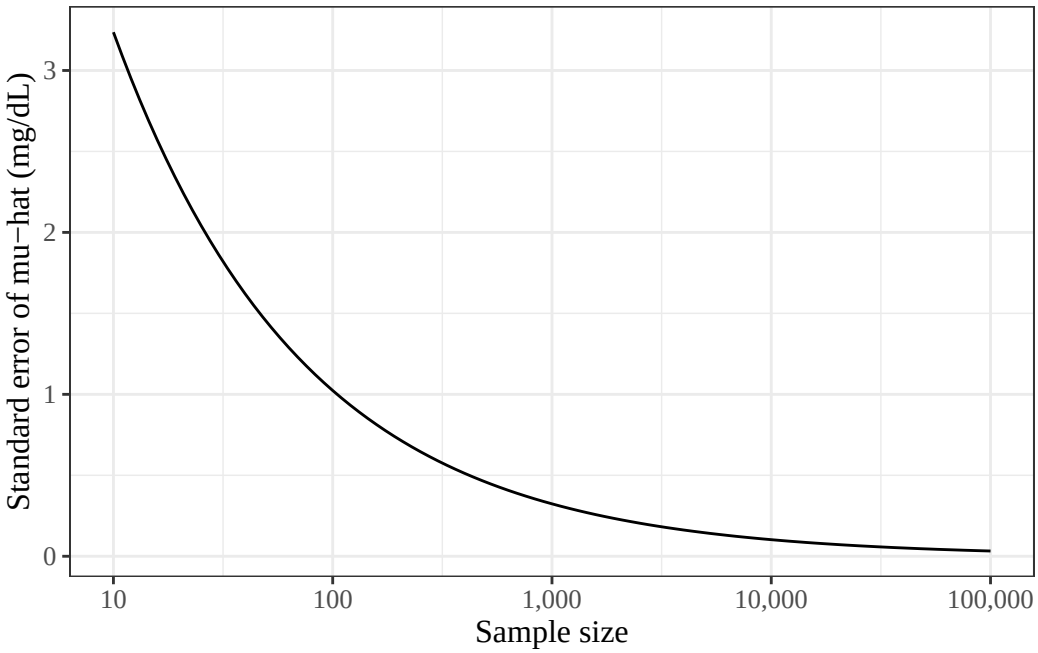


Figure 19: Standard error of $\hat{\mu}_{\text{ML}}$ as a function of sample size, with $\sigma = \hat{\sigma}_{\text{ML}}$

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

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

```r
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 0.1 (Power)**.** The **power** of a test of a null hypothesis H_0 about a parameter θ, against an alternative value θ_1, is the probability that the test rejects H_0 when θ equals θ_1:

$$\text{power}(\theta_1) \stackrel{\text{def}}{=} \Pr(\text{reject } H_0 \mid \theta = \theta_1).$$

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:

```r
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.

i Remark (Choosing the alternative, and observed power)

Remark 0.1 (Choosing the alternative, and observed power). 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 [26].

```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("Sample size, n")
```

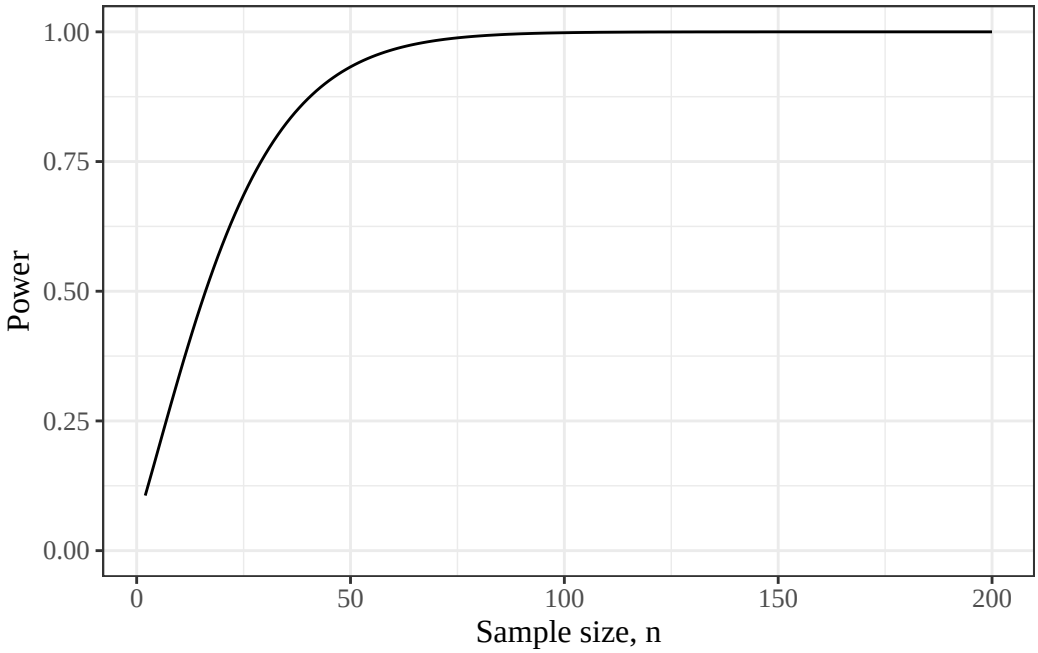


Figure 20: Power of the test of $H_0 : \mu = 95$ against $\mu_1 = 100$ mg/dL, by sample size

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
  }
}

# one small example dataset:
set.seed(0)
do_one_sim(n = 10, mu = 0, mu0 = 0, sigma2 = 1)
#> # A tibble: 1 x 9
```

```
#>   mu_hat sigma_hat se_hat confint_left confint_right tstat  pval confint_covers
#>    <dbl>     <dbl>  <dbl>        <dbl>         <dbl> <dbl> <dbl> <lgl>
#> 1  0.359      1.21  0.381       -0.503          1.22 0.942 0.371 TRUE
#> # i 1 more variable: test_rejects <lgl>
```

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

```
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:

```
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 sample mean of the estimation error of $\hat{\mu}$ over these datasets 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 population 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 0.1 (Binomial likelihood (adapted from Dobson and Barnett [1], Chapter 3))**.** Let $Y \sim \text{Binomial}(n, \pi)$, so that

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

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 0.1. **(a)**

$$
\begin{aligned}
\ell(\pi; y) &= \log\left\{ \binom{n}{y} \pi^y (1-\pi)^{n-y} \right\} && \text{(log of the binomial PMF)} \\
&= \log\left\{ \binom{n}{y} \right\} + \log\{\pi^y\} + \log\{(1-\pi)^{n-y}\} && \text{(log of a product)} \\
&= \log\left\{ \binom{n}{y} \right\} + y\log\{\pi\} + (n-y)\log\{1-\pi\} && \text{(log of a power)}
\end{aligned}
$$

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

(b)

$$
\begin{aligned}
\ell'(\pi; y) &= \frac{\partial}{\partial \pi}\left[\log\left\{ \binom{n}{y} \right\} + y\log\{\pi\} + (n-y)\log\{1-\pi\} \right] && \text{(log-likelihood from (a))} \\
&= \frac{\partial}{\partial \pi}\log\left\{ \binom{n}{y} \right\} + \frac{\partial}{\partial \pi}[y\log\{\pi\}] + \frac{\partial}{\partial \pi}[(n-y)\log\{1-\pi\}] && \text{(linearity of differentiation)} \\
&= 0 + \frac{\partial}{\partial \pi}[y\log\{\pi\}] + \frac{\partial}{\partial \pi}[(n-y)\log\{1-\pi\}] && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial \pi}[y\log\{\pi\}] + \frac{\partial}{\partial \pi}[(n-y)\log\{1-\pi\}] && \text{(drop the zero term)} \\
&= y\frac{\partial}{\partial \pi}\log\{\pi\} + (n-y)\frac{\partial}{\partial \pi}\log\{1-\pi\} && \text{(constant multiple rule)} \\
&= y\frac{1}{\pi} + (n-y)\frac{\partial}{\partial \pi}\log\{1-\pi\} && \text{(derivative of log)} \\
&= y\frac{1}{\pi} + (n-y)\frac{1}{1-\pi}\frac{\partial}{\partial \pi}(1-\pi) && \text{(chain rule)}
\end{aligned}
$$

The inner derivative is:

$$
\begin{aligned}
\frac{\partial}{\partial \pi}(1-\pi) &= \frac{\partial}{\partial \pi}1 - \frac{\partial}{\partial \pi}\pi && \text{(linearity of differentiation)} \\
&= 0 - \frac{\partial}{\partial \pi}\pi && \text{(derivative of a constant)} \\
&= 0 - 1 && \text{(derivative of } \pi \text{ with respect to itself)} \\
&= -1 && \text{(subtract)}
\end{aligned}
$$

Substituting the inner derivative back in:

$$
\begin{aligned}
\ell'(\pi; y) &= y\frac{1}{\pi} + (n-y)\frac{1}{1-\pi}(-1) && \text{(inner derivative is } -1\text{)} \\
&= y\frac{1}{\pi} - (n-y)\frac{1}{1-\pi} && \text{(multiply by } -1\text{)} \\
&= \frac{y}{\pi} - \frac{n-y}{1-\pi} && \text{(multiply)} \\
&= \frac{y(1-\pi)}{\pi(1-\pi)} - \frac{(n-y)\pi}{\pi(1-\pi)} && \text{(write both terms over the common denominator)} \\
&= \frac{y(1-\pi) - (n-y)\pi}{\pi(1-\pi)} && \text{(combine the fractions)} \\
&= \frac{y - y\pi - n\pi + y\pi}{\pi(1-\pi)} && \text{(expand the numerator)} \\
&= \frac{y - n\pi}{\pi(1-\pi)} && \text{(cancel } y\pi\text{)}
\end{aligned}
$$

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

$$
\begin{aligned}
0 &= \frac{y - n\pi}{\pi(1-\pi)} && \text{(score from (b))} \\
0 &= y - n\pi && \text{(multiply both sides by } \pi(1-\pi)\text{)} \\
y &= n\pi && \text{(add } n\pi \text{ to both sides)} \\
\hat{\pi}_{ML} &= \frac{y}{n} && \text{(divide both sides by } n\text{)}
\end{aligned}
$$

(d)

$$
\begin{aligned}
\ell''(\pi; y) &= \frac{\partial}{\partial \pi}\Big[\frac{y}{\pi} - \frac{n-y}{1-\pi}\Big] && \text{(score from (b))} \\
&= \frac{\partial}{\partial \pi}\big[y\pi^{-1} - (n-y)(1-\pi)^{-1}\big] && \text{(write the fractions as negative powers)} \\
&= \frac{\partial}{\partial \pi}\big[y\pi^{-1}\big] - \frac{\partial}{\partial \pi}\big[(n-y)(1-\pi)^{-1}\big] && \text{(linearity of differentiation)} \\
&= y\frac{\partial}{\partial \pi}\pi^{-1} - (n-y)\frac{\partial}{\partial \pi}(1-\pi)^{-1} && \text{(constant multiple rule)} \\
&= y(-\pi^{-2}) - (n-y)\frac{\partial}{\partial \pi}(1-\pi)^{-1} && \text{(power rule)} \\
&= y(-\pi^{-2}) - (n-y)(-(1-\pi)^{-2})\frac{\partial}{\partial \pi}(1-\pi) && \text{(chain rule, outer function } u^{-1}\text{)} \\
&= y(-\pi^{-2}) - (n-y)(-(1-\pi)^{-2})(-1) && \text{(inner derivative from (b) is } -1\text{)} \\
&= -y\pi^{-2} - (n-y)(1-\pi)^{-2} && \text{(multiply)} \\
&= -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2} && \text{(write the negative powers as fractions)}
\end{aligned}
$$

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 0.2 (Gaussian log-likelihood (adapted from Kleinbaum et al. [27], Chapter 5))**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$.
(a) Write the likelihood $\mathcal{L}(\mu, \sigma^2; \underset{\sim}{x})$ for the observed data $\underset{\sim}{x} = (x_1, \ldots, x_n)$.
(b) Write the log-likelihood $\ell(\mu, \sigma^2; \underset{\sim}{x})$ in simplified form, as the sum of a term that depends on the data only through n and a term that depends on the data only through $\sum_{i=1}^{n}(x_i - \mu)^2$.
(c) Derive the MLE $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$.

Solution

Solution 0.2. **(a)**

$$
\begin{aligned}
\mathcal{L}(\mu, \sigma^2; \underset{\sim}{x}) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} && \text{(likelihood of an iid sample)} \\
&= \left(\prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \right) \left(\prod_{i=1}^{n} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right) && \text{(regroup the factors of the product)} \\
&= (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} && \text{(product of } n \text{ copies of a power)} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{ \sum_{i=1}^{n} -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} && \text{(product of exponentials)} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} && \text{(factor the constant out of the sum)}
\end{aligned}
$$

(b)

$$
\begin{aligned}
\ell(\mu, \sigma^2; \underset{\sim}{x}) &= \log\{ \mathcal{L}(\mu, \sigma^2; \underset{\sim}{x}) \} && \text{(definition of the log-likelihood)} \\
&= \log\left\{ (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \right\} && \text{(likelihood from (a))} \\
&= \log\{ (2\pi\sigma^2)^{-n/2} \} + \log\left\{ \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \right\} && \text{(log of a product)} \\
&= -\frac{n}{2} \log\{ 2\pi\sigma^2 \} + \log\left\{ \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \right\} && \text{(log of a power)} \\
&= -\frac{n}{2} \log\{ 2\pi\sigma^2 \} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 && \text{(log undoes exp)}
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{\partial}{\partial \mu}\ell &= \frac{\partial}{\partial \mu}\left[-\frac{n}{2}\log\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(log-likelihood from (b))} \\
&= \frac{\partial}{\partial \mu}\left[-\frac{n}{2}\log\{2\pi\sigma^2\}\right] + \frac{\partial}{\partial \mu}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(linearity of differentiation)} \\
&= 0 + \frac{\partial}{\partial \mu}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial \mu}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(drop the zero term)} \\
&= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \mu}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(constant multiple rule)} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(x_i-\mu)^2 && \text{(derivative of a sum is the sum of derivatives)} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)\frac{\partial}{\partial \mu}(x_i-\mu) && \text{(chain rule, outer function } u^2\text{)}
\end{aligned}
$$

The inner derivative is:

$$
\begin{aligned}
\frac{\partial}{\partial \mu}(x_i-\mu) &= \frac{\partial}{\partial \mu}x_i - \frac{\partial}{\partial \mu}\mu && \text{(linearity of differentiation)} \\
&= 0 - \frac{\partial}{\partial \mu}\mu && (x_i \text{ does not depend on } \mu) \\
&= 0 - 1 && \text{(derivative of } \mu \text{ with respect to itself)} \\
&= -1 && \text{(subtract)}
\end{aligned}
$$

Substituting the inner derivative back in:

$$
\begin{aligned}
\frac{\partial}{\partial \mu}\ell &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)\cdot(-1) && \text{(inner derivative is } -1\text{)} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(x_i-\mu) && \text{(multiply)} \\
&= -\frac{1}{2\sigma^2}\cdot(-2)\sum_{i=1}^{n}(x_i-\mu) && \text{(factor the constant } -2 \text{ out of the sum)} \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) && \text{(multiply the constants)} \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu\right) && \text{(distribute the sum over the terms)} \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) && \text{(sum of } n \text{ copies of } \mu\text{)}
\end{aligned}
$$

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

Deriving $\hat{\sigma}^2_{ML}$:

$$
\begin{aligned}
\frac{\partial}{\partial \sigma^2}\ell &= \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(log-likelihood from (b))} \\
&= \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}(\log\{2\pi\} + \log\{\sigma^2\}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(log of a product)} \\
&= \frac{\partial}{\partial \sigma^2}\left[-\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\right] && \text{(distribute } -\tfrac{n}{2}\text{)} \\
&= \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\{2\pi\} - \frac{n}{2}\log\{\sigma^2\} - \frac{1}{2}(\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(write } \tfrac{1}{2\sigma^2} \text{ as } \tfrac{1}{2}(\sigma^2)^{-1}\text{)} \\
&= \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\{2\pi\}\right] + \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\{\sigma^2\}\right] + \frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}(\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(linearity of differentiation)} \\
&= 0 + \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\{\sigma^2\}\right] + \frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}(\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\{\sigma^2\}\right] + \frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}(\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2\right] && \text{(drop the zero term)} \\
&= \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\{\sigma^2\}\right] + \frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(\sigma^2)^{-1}\right] && \text{(reorder the factors)} \\
&= -\frac{n}{2}\frac{\partial}{\partial \sigma^2}\log\{\sigma^2\} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\frac{\partial}{\partial \sigma^2}(\sigma^2)^{-1} && \text{(constant multiple rule)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\frac{\partial}{\partial \sigma^2}(\sigma^2)^{-1} && \text{(derivative of log)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(-(\sigma^2)^{-2}) && \text{(power rule)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(\sigma^2)^{-2} && \text{(multiply)} \\
&= -\frac{n}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(reorder the factors)}
\end{aligned}
$$

Setting this to zero and solving:

$$
\begin{aligned}
0 &= -\frac{n}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(set the derivative to zero)}\\
\frac{n}{2}(\sigma^2)^{-1} &= \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(add } \tfrac{n}{2}(\sigma^2)^{-1} \text{ to both sides)}\\
n(\sigma^2)^{-1}(\sigma^2)^2 &= (\sigma^2)^{-2}(\sigma^2)^2\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(multiply both sides by } 2(\sigma^2)^2)\\
n\sigma^2 &= (\sigma^2)^{-2}(\sigma^2)^2\sum_{i=1}^{n}(x_i-\mu)^2 && ((\sigma^2)^{-1}(\sigma^2)^2 = \sigma^2)\\
n\sigma^2 &= \sum_{i=1}^{n}(x_i-\mu)^2 && ((\sigma^2)^{-2}(\sigma^2)^2 = 1)\\
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 && \text{(divide both sides by } n)
\end{aligned}
$$

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 0.3 (Score at the MLE (adapted from Dobson and Barnett [1], Chapter 3))**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$.
(a) Write the log-likelihood $\ell(\lambda; \underset{\sim}{x})$.
(b) Derive the score function $\ell'(\lambda; \underset{\sim}{x})$.
(c) Find the MLE $\hat{\lambda}_{ML}$, and evaluate the score at the MLE.
(d) Provide an intuitive interpretation of the value of the score at $\hat{\lambda}_{ML}$.

Solution

Solution 0.3. **(a)**

$$
\begin{aligned}
\ell(\lambda; \underset{\sim}{x}) &= \log\left\{\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right\} && \text{(log of the likelihood of an iid sample)} \\
&= \sum_{i=1}^{n} \log\left\{\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right\} && \text{(log of a product)} \\
&= \sum_{i=1}^{n} (\log\{\lambda^{x_i} e^{-\lambda}\} - \log\{x_i!\}) && \text{(log of a quotient)} \\
&= \sum_{i=1}^{n} (\log\{\lambda^{x_i}\} + \log\{e^{-\lambda}\} - \log\{x_i!\}) && \text{(log of a product)} \\
&= \sum_{i=1}^{n} (x_i \log\{\lambda\} + \log\{e^{-\lambda}\} - \log\{x_i!\}) && \text{(log of a power)} \\
&= \sum_{i=1}^{n} (x_i \log\{\lambda\} - \lambda - \log\{x_i!\}) && \text{(log undoes exp)} \\
&= \sum_{i=1}^{n} x_i \log\{\lambda\} - \sum_{i=1}^{n} \lambda - \sum_{i=1}^{n} \log\{x_i!\} && \text{(distribute the sum over the terms)} \\
&= \left(\sum_{i=1}^{n} x_i\right) \log\{\lambda\} - \sum_{i=1}^{n} \lambda - \sum_{i=1}^{n} \log\{x_i!\} && \text{(factor } \log\{\lambda\} \text{ out of the first sum)} \\
&= \left(\sum_{i=1}^{n} x_i\right) \log\{\lambda\} - n\lambda - \sum_{i=1}^{n} \log\{x_i!\} && \text{(sum of } n \text{ copies of } \lambda\text{)}
\end{aligned}
$$

(b)

$$
\begin{aligned}
\ell'(\lambda; \underset{\sim}{x}) &= \frac{\partial}{\partial \lambda}\left[\left(\sum_{i=1}^{n} x_i\right)\log\{\lambda\} - n\lambda - \sum_{i=1}^{n}\log\{x_i!\}\right] && \text{(log-likelihood from (a))} \\
&= \frac{\partial}{\partial \lambda}\left[\left(\sum_{i=1}^{n} x_i\right)\log\{\lambda\}\right] - \frac{\partial}{\partial \lambda}[n\lambda] - \frac{\partial}{\partial \lambda}\left[\sum_{i=1}^{n}\log\{x_i!\}\right] && \text{(linearity of differentiation)} \\
&= \frac{\partial}{\partial \lambda}\left[\left(\sum_{i=1}^{n} x_i\right)\log\{\lambda\}\right] - \frac{\partial}{\partial \lambda}[n\lambda] - 0 && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial \lambda}\left[\left(\sum_{i=1}^{n} x_i\right)\log\{\lambda\}\right] - \frac{\partial}{\partial \lambda}[n\lambda] && \text{(drop the zero term)} \\
&= \left(\sum_{i=1}^{n} x_i\right)\frac{\partial}{\partial \lambda}\log\{\lambda\} - n\frac{\partial}{\partial \lambda}\lambda && \text{(constant multiple rule)} \\
&= \left(\sum_{i=1}^{n} x_i\right)\lambda^{-1} - n\frac{\partial}{\partial \lambda}\lambda && \text{(derivative of log)} \\
&= \left(\sum_{i=1}^{n} x_i\right)\lambda^{-1} - n && \text{(derivative of } \lambda \text{ with respect to itself)} \\
&= \frac{\sum_{i=1}^{n} x_i}{\lambda} - n && \text{(write the negative power as a fraction)} \\
&= \frac{n\bar{x}}{\lambda} - n && \left(\textstyle\sum_{i=1}^{n} x_i = n\bar{x}\right)
\end{aligned}
$$

(c)
The MLE solves $\ell'(\lambda; \underset{\sim}{x}) = 0$, so evaluating the score there afterwards is a check on the algebra. Setting $\ell'(\lambda; \underset{\sim}{x}) = 0$:

$$
\begin{aligned}
0 &= \frac{n\bar{x}}{\lambda} - n && \text{(score from (b))} \\
\frac{n\bar{x}}{\lambda} &= n && \text{(add } n \text{ to both sides)} \\
n\bar{x} &= n\lambda && \text{(multiply both sides by } \lambda\text{)} \\
\hat{\lambda}_{ML} &= \bar{x} && \text{(divide both sides by } n\text{)}
\end{aligned}
$$

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

$$
\begin{aligned}
\ell'(\bar{x}; \underset{\sim}{x}) &= \frac{n\bar{x}}{\bar{x}} - n && \text{(score from (b) at } \lambda = \bar{x}\text{)} \\
&= n - n && \text{(cancel } \bar{x}\text{, since } \bar{x} > 0\text{)} \\
&= 0 && \text{(subtract)}
\end{aligned}
$$

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 $\ell'(\lambda; \underset{\sim}{x}) = \frac{n\bar{x}}{\lambda} - n$ is not defined at $\lambda = 0$, so the usual interior verification $\ell'(\hat{\lambda}_{ML}; \underset{\sim}{x}) = 0$ 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 sample average count ($\bar{x}$).

Exercise 0.4 (Standard error of an MLE (adapted from Dobson and Barnett [1], Chapter 5))**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$.

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

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

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

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

Solution

Solution 0.4. **(a)**

From Exercise 0.3, $\ell'(\lambda; \underset{\sim}{x}) = \frac{n\bar{x}}{\lambda} - n$.

$$
\begin{aligned}
\ell''(\lambda; \underset{\sim}{x}) &= \frac{\partial}{\partial\lambda}\left[\frac{n\bar{x}}{\lambda} - n\right] && \text{(Poisson score)} \\
&= \frac{\partial}{\partial\lambda}[n\bar{x}\lambda^{-1} - n] && \text{(write } \tfrac{n\bar{x}}{\lambda} \text{ as } n\bar{x}\lambda^{-1}\text{)} \\
&= \frac{\partial}{\partial\lambda}[n\bar{x}\lambda^{-1}] - \frac{\partial}{\partial\lambda}n && \text{(linearity of differentiation)} \\
&= \frac{\partial}{\partial\lambda}[n\bar{x}\lambda^{-1}] - 0 && \text{(derivative of a constant)} \\
&= \frac{\partial}{\partial\lambda}[n\bar{x}\lambda^{-1}] && \text{(drop the zero term)} \\
&= n\bar{x}\frac{\partial}{\partial\lambda}\lambda^{-1} && \text{(constant multiple rule)} \\
&= n\bar{x}(-\lambda^{-2}) && \text{(power rule)} \\
&= -n\bar{x}\lambda^{-2} && \text{(multiply)} \\
&= -\frac{n\bar{x}}{\lambda^2} && \text{(write the negative power as a fraction)}
\end{aligned}
$$

(b)

$$
\begin{aligned}
I(\lambda; \underset{\sim}{x}) &= -\ell''(\lambda; \underset{\sim}{x}) && \text{(observed information is the negative Hessian)} \\
&= -\left(-\frac{n\bar{x}}{\lambda^2}\right) && \text{(Hessian from (a))} \\
&= \frac{n\bar{x}}{\lambda^2} && \text{(two negatives make a positive)}
\end{aligned}
$$

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

$$
\begin{aligned}
I(\hat{\lambda}_{ML}; \underset{\sim}{x}) &= \frac{n\bar{x}}{\bar{x}^2} && \text{(observed information at } \lambda = \bar{x}) \\
&= \frac{n}{\bar{x}} && \text{(cancel one factor of } \bar{x})
\end{aligned}
$$

(d)
By the asymptotic theory of MLEs:

$$
\hat{\lambda}_{ML} \mathrel{\dot{\sim}} \mathrm{N}\Big(\lambda,\ (\mathcal{I}(\lambda))^{-1}\Big)
$$

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

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

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 0.5 (Exponential MLE (adapted from Dobson and Barnett [1], Chapter 3))**.** Let $X_1, \ldots, X_n \ \sim_{\text{iid}}\ \text{Exponential}(\mu)$, so that

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

Write the log-likelihood $\ell(\mu; \underset{\sim}{x})$ for the observed data $\underset{\sim}{x} = (x_1, \ldots, x_n)$.

Solution

Solution 0.5.

$$
\begin{aligned}
\ell(\mu; \underset{\sim}{x}) &= \log\left\{\prod_{i=1}^{n} \frac{1}{\mu} e^{-x_i/\mu}\right\} && \text{(log of the likelihood of an iid sample)} \\
&= \sum_{i=1}^{n} \log\left\{\frac{1}{\mu} e^{-x_i/\mu}\right\} && \text{(log of a product)} \\
&= \sum_{i=1}^{n} \left(\log\left\{\frac{1}{\mu}\right\} + \log\left\{e^{-x_i/\mu}\right\}\right) && \text{(log of a product)} \\
&= \sum_{i=1}^{n} \left(-\log\{\mu\} + \log\left\{e^{-x_i/\mu}\right\}\right) && \text{(log of a reciprocal)} \\
&= \sum_{i=1}^{n} \left(-\log\{\mu\} - \frac{x_i}{\mu}\right) && \text{(log undoes exp)} \\
&= \sum_{i=1}^{n} (-\log\{\mu\}) - \sum_{i=1}^{n} \frac{x_i}{\mu} && \text{(distribute the sum over the terms)} \\
&= -n\log\{\mu\} - \sum_{i=1}^{n} \frac{x_i}{\mu} && \text{(sum of } n \text{ identical terms)} \\
&= -n\log\{\mu\} - \frac{1}{\mu} \sum_{i=1}^{n} x_i && \text{(factor the constant } \tfrac{1}{\mu} \text{ out of the sum)} \\
&= -n\log\{\mu\} - \frac{n\bar{x}}{\mu} && \left(\textstyle\sum_{i=1}^{n} x_i = n\bar{x}\right)
\end{aligned}
$$

Exercise 0.6 (Score function of the exponential)**.** In the setting of Exercise 0.5, derive the score function $\ell'(\mu; \underset{\sim}{x}) = \frac{\partial}{\partial \mu} \ell(\mu; \underset{\sim}{x})$.

Solution

Solution 0.6.

$$
\begin{aligned}
\ell'(\mu; \underset{\sim}{x}) &= \frac{\partial}{\partial \mu}\left(-n\log\{\mu\} - \frac{n\bar{x}}{\mu}\right) && \text{(log-likelihood)} \\
&= \frac{\partial}{\partial \mu}(-n\log\{\mu\} - n\bar{x}\mu^{-1}) && \text{(write } \tfrac{n\bar{x}}{\mu} \text{ as } n\bar{x}\mu^{-1}\text{)} \\
&= \frac{\partial}{\partial \mu}(-n\log\{\mu\}) - \frac{\partial}{\partial \mu}(n\bar{x}\mu^{-1}) && \text{(linearity of differentiation)} \\
&= -n\frac{\partial}{\partial \mu}\log\{\mu\} - n\bar{x}\frac{\partial}{\partial \mu}\mu^{-1} && \text{(constant multiple rule)} \\
&= -n\mu^{-1} - n\bar{x}\frac{\partial}{\partial \mu}\mu^{-1} && \text{(derivative of log)} \\
&= -n\mu^{-1} - n\bar{x}(-\mu^{-2}) && \text{(power rule)} \\
&= -n\mu^{-1} + n\bar{x}\mu^{-2} && \text{(multiply)} \\
&= -\frac{n}{\mu} + \frac{n\bar{x}}{\mu^2} && \text{(write the negative powers as fractions)} \\
&= -\frac{n\mu}{\mu^2} + \frac{n\bar{x}}{\mu^2} && \text{(write both terms over the common denominator)} \\
&= \frac{-n\mu + n\bar{x}}{\mu^2} && \text{(combine the fractions)} \\
&= \frac{n\bar{x} - n\mu}{\mu^2} && \text{(reorder the numerator)} \\
&= \frac{n}{\mu^2}(\bar{x} - \mu) && \text{(factor out } \tfrac{n}{\mu^2}\text{)}
\end{aligned}
$$

Exercise 0.7 (Maximum likelihood estimate of the exponential mean)**.** In the setting of Exercise 0.5, solve the score equation $\ell'(\mu; \underset{\sim}{x}) = 0$ with the score from Exercise 0.6, to show that $\hat{\mu}_{ML} = \bar{x}$.

Solution

Solution 0.7. Setting $\ell'(\mu; \underset{\sim}{x}) = 0$:

$$
\begin{aligned}
0 &= \frac{n}{\mu^2}(\bar{x} - \mu) && \text{(score)} \\
0 &= \bar{x} - \mu && \text{(multiply both sides by } \tfrac{\mu^2}{n}\text{)} \\
\hat{\mu}_{ML} &= \bar{x} && \text{(add } \mu \text{ to both sides)}
\end{aligned}
$$

Exercise 0.8 (Second derivative of the exponential log-likelihood)**.** In the setting of Exercise 0.5, compute the second derivative $\ell''(\mu; \underset{\sim}{x})$ of the log-likelihood.

Solution

Solution 0.8. The second derivative is:

$$\begin{aligned}
\ell''(\mu; \underset{\sim}{x}) &= \frac{\partial}{\partial \mu}\left(-\frac{n}{\mu} + \frac{n\bar{x}}{\mu^2}\right) && \text{(score, before factoring)} \\
&= \frac{\partial}{\partial \mu}(-n\mu^{-1} + n\bar{x}\mu^{-2}) && \text{(write the fractions as negative powers)} \\
&= \frac{\partial}{\partial \mu}(-n\mu^{-1}) + \frac{\partial}{\partial \mu}(n\bar{x}\mu^{-2}) && \text{(linearity of differentiation)} \\
&= -n\frac{\partial}{\partial \mu}\mu^{-1} + n\bar{x}\frac{\partial}{\partial \mu}\mu^{-2} && \text{(constant multiple rule)} \\
&= -n(-\mu^{-2}) + n\bar{x}\frac{\partial}{\partial \mu}\mu^{-2} && \text{(power rule for } \mu^{-1}\text{)} \\
&= -n(-\mu^{-2}) + n\bar{x}(-2\mu^{-3}) && \text{(power rule for } \mu^{-2}\text{)} \\
&= n\mu^{-2} - 2n\bar{x}\mu^{-3} && \text{(multiply)} \\
&= \frac{n}{\mu^2} - \frac{2n\bar{x}}{\mu^3} && \text{(write the negative powers as fractions)}
\end{aligned}$$

Exercise 0.9 (Check the critical point is a maximum)**.** In the setting of Exercise 0.5, use the second derivative from Exercise 0.8 to verify that the critical point $\hat{\mu}_{ML}$ from Exercise 0.7 is a maximum.

Solution

Solution 0.9. Evaluated at $\hat{\mu}_{ML} = \bar{x}$:

$$\begin{aligned}
\ell''(\bar{x}; \underset{\sim}{x}) &= \frac{n}{\bar{x}^2} - \frac{2n\bar{x}}{\bar{x}^3} && \text{(second derivative at } \mu = \bar{x}\text{)} \\
&= \frac{n}{\bar{x}^2} - \frac{2n}{\bar{x}^2} && \text{(cancel one factor of } \bar{x}\text{)} \\
&= -\frac{n}{\bar{x}^2} && \text{(combine like terms)}
\end{aligned}$$

Since $n > 0$ and $\bar{x}^2 > 0$, the second derivative is negative, so $\hat{\mu}_{ML} = \bar{x}$ is a maximum.

Exercise 0.10 (Observed information for the exponential mean)**.** In the setting of Exercise 0.5, use the second derivative from Exercise 0.8 to derive the observed information $I(\mu; \underset{\sim}{x}) = -\ell''(\mu; \underset{\sim}{x})$.

Solution

Solution 0.10. The observed information is:

$$\begin{aligned} I(\mu; \underset{\sim}{x}) &= -\ell''(\mu; \underset{\sim}{x}) && \text{(observed information is the negative Hessian)} \\ &= -\left(\frac{n}{\mu^2} - \frac{2n\bar{x}}{\mu^3}\right) && \text{(second derivative)} \\ &= -\frac{n}{\mu^2} + \frac{2n\bar{x}}{\mu^3} && \text{(distribute the minus sign)} \\ &= \frac{2n\bar{x}}{\mu^3} - \frac{n}{\mu^2} && \text{(reorder the terms)} \end{aligned}$$

Exercise 0.11 (Observed information at the MLE)**.** In the setting of Exercise 0.5, evaluate the observed information from Exercise 0.10 at $\hat{\mu}_{ML}$ from Exercise 0.7.

Solution

Solution 0.11. Evaluated at $\hat{\mu}_{ML} = \bar{x}$:

$$\begin{aligned} I(\hat{\mu}_{ML}; \underset{\sim}{x}) &= \frac{2n\bar{x}}{\bar{x}^3} - \frac{n}{\bar{x}^2} && \text{(observed information at } \mu = \bar{x}) \\ &= \frac{2n}{\bar{x}^2} - \frac{n}{\bar{x}^2} && \text{(cancel one factor of } \bar{x}) \\ &= \frac{n}{\bar{x}^2} && \text{(combine like terms)} \end{aligned}$$

So $\text{SE}(\hat{\mu}_{ML}) \approx \sqrt{I(\hat{\mu}_{ML}; \underset{\sim}{x})^{-1}} = \frac{\bar{x}}{\sqrt{n}}$.

Exercise 0.12 (Approximate confidence interval for the exponential mean)**.** In the setting of Exercise 0.5, use the observed information at the MLE from Exercise 0.11 to give an approximate 95% confidence interval for μ.

Solution

Solution 0.12. 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 the standard error reduces to $\mu/\sqrt{n}$ because $\text{SD}(X) = \mu$ for the exponential distribution.

Further reading

- Josh Starmer's StatQuest video *Maximum Likelihood, clearly explained!!!* introduces the idea of maximum likelihood estimation [28].

References

[1] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, 4th ed. CRC press, 2018. doi: 10.1201/9781315182780.

[2] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, 2nd ed. John Wiley & Sons, 2007. doi: 10.1002/9780470191613.

[3] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[4] Wikipedia contributors, "Likelihood function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Likelihood_function

[5] Wikipedia contributors, "Maximum likelihood estimation — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

[6] 5MinuteAI, "Probability & statistics." Video playlist on YouTube. Accessed: Oct. 12, 2026. [Online]. Available: https://www.youtube.com/playlist?list=PLeB45KifGiuHesi4PALNZSYZFhViVGQJK

[7] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman, *Probability and statistical inference*, Tenth edition. Boston: Pearson, 2019.

[8] J. A. Rice, *Mathematical statistics and data analysis*, 3rd ed. Belmont, CA: Cengage Learning, 2007.

[9] Wikipedia contributors, "Critical point (mathematics) — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Critical_point_%28mathematics%29

[10] Wikipedia contributors, "Informant function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Informant_function

[11] R. A. Fisher, "The detection of linkage with 'dominant' abnormalities," *Annals of Eugenics*, vol. 6, no. 2, pp. 187–201, 1935, doi: 10.1111/j.1469-1809.1935.tb02227.x.

[12] C. R. Rao, "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 44, no. 1, pp. 50–57, 1948, doi: 10.1017/S0305004100023987.

[13] A. K. Bera and Y. Bilias, "Three score and 15 years (1948–2023) of Rao's score test: A brief history." arXiv:2406.19956, 2024. doi: 10.48550/arXiv.2406.19956.

[14] Wikipedia contributors, "Hessian matrix — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Hessian_matrix

[15] Wikipedia contributors, "Fisher information — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Fisher_information

[16] E. L. Lehmann, *Elements of large-sample theory*. in Springer texts in statistics. New York: Springer, 1999. doi: 10.1007/b98855.

[17] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," in *Handbook of econometrics*, vol. 4, R. Engle and D. McFadden, Eds., Elsevier, 1994, pp. 2111–2245. doi: 10.1016/S1573-4412(05)80005-4.

[18] B. Efron and D. V. Hinkley, "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information," *Biometrika*, vol. 65, no. 3, pp. 457–483, 1978, doi: 10.1093/biomet/65.3.457.

[19] Wikipedia contributors, "Wald test — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Wald_test

[20] Wikipedia contributors, "Likelihood-ratio test — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Likelihood-ratio_test

[21] Wikipedia contributors, "Wilks' theorem — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Wilks%27_theorem

[22] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938, doi: 10.1214/aoms/1177732360.

[23] Wikipedia contributors, "Prediction interval — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Prediction_interval

[24] S. Hulley *et al.*, "Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women," *JAMA : the journal of the American Medical Association*, vol. 280, no. 7, pp. 605–613, 1998, doi: 10.1001/jama.280.7.605.

[25] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*, 2nd ed. Springer, 2012. doi: 10.1007/978-1-4614-1353-0.

[26] J. M. Hoenig and D. M. Heisey, "The abuse of power: The pervasive fallacy of power calculations for data analysis," *The American Statistician*, vol. 55, no. 1, pp. 19–24, 2001, doi: 10.1198/000313001300339897.

[27] D. G. Kleinbaum, L. L. Kupper, A. Nizam, K. Muller, and E. Rosenberg, *Applied regression analysis and other multivariable methods*, 5th ed. Cengage Learning, 2014. Available: https://www.cengage.com/c/applied-regression-analysis-and-other-multivariable-methods-5e-kleinbaum/9781285051086/

[28] J. Starmer, "Maximum likelihood, clearly explained." StatQuest, video on YouTube, 2017. Available: https://www.youtube.com/watch?v=XepXtl9YKwc