Chapter 4
Likelihood-based inference
4.1 Likelihoods

Data \( \mathbf{x} = \{x_1, \ldots, x_n\} \), joint distribution of \( \mathbf{x} \) depends on unknown \( \theta \).
Likelihood is density (or probability if \( x_i \) is discrete) of the data \( x \) conditional on the parameter \( \theta \), i.e.

\[
f(\mathbf{x}|\theta).
\]

Function of \( \theta \) for fixed \( \mathbf{x} \), so denote the likelihood function by \( L(\theta; \mathbf{x}) \):

\[
L(\theta; \mathbf{x}) = f(\mathbf{x}|\theta).
\]

If \( x_1, \ldots, x_n \) are independent, then
\[
f(\mathbf{x}|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \ldots \times f(x_n|\theta),
\]
and so

\[
L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).
\]

Used for point and interval estimation, and hypothesis testing.
Score statistics, Fisher information and the Cramer-Rao minimum variance bound

The score statistic is defined to be \( \frac{\partial}{\partial \theta} l(\theta; x) = \frac{\partial}{\partial \theta} \log f(x|\theta). \)

\(X\): unobserved value of \(x\). Define the random variable

\[
\frac{\partial}{\partial \theta} l(\theta; X) = \frac{\partial}{\partial \theta} \log f(X|\theta).
\]

Transformation of a r.v. \(X\), where transformation is derivative, w.r.t. \(\theta\), of the log of the density of \(X\).

N.B. We treat \(l(\theta; X)\) as a function of the random data \(X\), evaluated at the true value of \(\theta\), rather than a function of the parameter \(\theta\) for fixed data \(x\).
\[
\frac{\partial}{\partial \theta} l(\theta; \mathbf{X}) &= \frac{\partial}{\partial \theta} \log L(\theta; \mathbf{X}) \\
&= \left\{ \frac{\partial}{\partial \theta} L(\theta; \mathbf{X}) \right\} \times \frac{1}{L(\theta; \mathbf{X})} = \left\{ \frac{\partial}{\partial \theta} f(\mathbf{X}|\theta) \right\} \times \frac{1}{f(\mathbf{X}|\theta)}. 
\]

\[
E \left\{ \frac{\partial}{\partial \theta} l(\theta; \mathbf{X}) \right\} = \int \left\{ \frac{\partial}{\partial \theta} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x} \\
= \int \left\{ \frac{\partial}{\partial \theta} f(\mathbf{x}|\theta) \right\} \times \frac{1}{f(\mathbf{x}|\theta)} f(\mathbf{x}|\theta) d\mathbf{x} \\
= \frac{\partial}{\partial \theta} \int f(\mathbf{x}|\theta) d\mathbf{x} \\
= \frac{\partial}{\partial \theta} 1 = 0.
\]
Expected value of the derivative of the log-likelihood at the true value of $\theta$ is 0.

Consider example of $X \sim \text{exp}(\text{rate} = \theta)$. Then $\ell(\theta; X) = \log \theta - \theta X$ and

$$\frac{\partial}{\partial \theta} \ell(\theta; X) = \frac{1}{\theta} - X,$$

so

$$E \left\{ \frac{\partial}{\partial \theta} \ell(\theta; X) \right\} = \int \left( \frac{1}{\theta} - x \right) \theta \exp(-\theta x) dx$$

$$= \frac{1}{\theta} \int \theta \exp(-\theta x) dx - \int x \theta \exp(-\theta x) dx$$

$$= \frac{1}{\theta} - \frac{1}{\theta} = 0.$$
However, the expected value of the derivative of the log-likelihood evaluated at the *wrong* value of $\theta$, say $\theta^*$, is not 0. For example,

$$\frac{\partial}{\partial \theta} l(\theta; X) \bigg|_{\theta=\theta^*} = \frac{1}{\theta^*} - X,$$

with

$$E \left\{ \frac{\partial}{\partial \theta} l(\theta; X) \bigg|_{\theta=\theta^*} \right\} = \int \left( \frac{1}{\theta^*} - x \right) \theta \exp(-\theta x) dx$$

$$= \frac{1}{\theta^*} - \frac{1}{\theta},$$

which is non-zero for $\theta^* \neq \theta$. 
To derive an expression for the variance of $\frac{\partial}{\partial \theta} l(\theta; \mathbf{X})$, we note that

$$0 = \int \frac{\partial}{\partial \theta} l(\theta; \mathbf{x}) f(\mathbf{x}|\theta) d\mathbf{x}$$

$$\Rightarrow 0 = \frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta} l(\theta; \mathbf{x}) f(\mathbf{x}|\theta) d\mathbf{x}$$

$$\Rightarrow 0 = \int \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x} + \int \left\{ \frac{\partial}{\partial \theta} l(\theta; \mathbf{x}) \right\} \left\{ \frac{\partial}{\partial \theta} f(\mathbf{x}|\theta) \right\} d\mathbf{x}$$

$$\Rightarrow 0 = \int \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x}$$

$$+ \int \left\{ \frac{\partial}{\partial \theta} l(\theta; \mathbf{x}) \right\} \left\{ \frac{\partial}{\partial \theta} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x}$$

$$\Rightarrow E \left[ \left\{ \frac{\partial}{\partial \theta} l(\theta; \mathbf{X}) \right\}^2 \right] = -E \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; \mathbf{X}) \right\}.$$
\[ E \left[ \left\{ \frac{\partial}{\partial \theta} l(\theta; X) \right\}^2 \right] = -E \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; X) \right\} \]

Since \( E\{ \frac{\partial}{\partial \theta} l(\theta; X) \} = 0 \), we have

\[ Var \left\{ \frac{\partial}{\partial \theta} l(\theta; X) \right\} = -E \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; X) \right\} . \]

The term \( -E \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; X) \right\} \) is known as the Fisher information which we will denote by \( \mathcal{I}_E(\theta) \):

\[ \mathcal{I}_E(\theta) \equiv -E \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; X) \right\} . \]
Fisher information: measure of amount of information a sample size of $n$ contains about $\theta$. For independent observations $X_1, \ldots, X_n$,

$$l(\theta; X) = \sum_{i=1}^{n} \log f(X_i | \theta),$$

$$\mathcal{I}_E(\theta) = -nE \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; X_i) \right\},$$

hence Fisher information is proportional to sample size.

• Example. Consider $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with $\sigma^2$ known. Then

$$-E \left\{ \frac{\partial^2}{\partial \theta^2} l(\theta; X) \right\} = -E \left\{ \frac{\partial^2}{\partial \theta^2} \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \theta)^2 \right\} = \frac{n}{\sigma^2},$$

Fisher information is $n/\sigma^2$. As $\sigma^2$ decreases, the observations more likely to be close to $\theta$, so data more informative about $\theta$. 
Fisher information can be used to give a bound on the variance of an estimator. Let $T(X)$ be an unbiased estimator, with $X_1, \ldots, X_n$ independent. Then it is possible to prove that

$$\text{Var}(T) \geq \frac{1}{\mathcal{I}_E(\theta)}.$$ 

This is known as the **Cramér-Rao minimum variance bound**.
Asymptotic normality

For large $n$, the distribution of the m.l.e $\hat{\theta}$ is approximately normal, with

$$\hat{\theta} \sim N\{\theta, \mathcal{I}_E(\theta)^{-1}\}.$$ 

Thus for large $n$, the m.l.e. $\hat{\theta}$ is \textit{approximately} unbiased, and achieves the Cramer-Rao minimum variance bound.
In the multivariate case with \( \theta = (\theta_1, \ldots, \theta_d) \) we have

\[
\mathcal{I}_E(\theta) = \begin{pmatrix}
  e_{1,1}(\theta) & \cdots & e_{1,d}(\theta) \\
  \vdots & \ddots & \vdots \\
  e_{d,1}(\theta) & \cdots & e_{d,d}(\theta)
\end{pmatrix},
\]

with

\[
e_{i,j}(\theta) = E \left\{ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta) \right\}.
\]

So for large \( n \), the distribution of the m.l.e of \( \theta \) is approximately multivariate normal:

\[
\hat{\theta} \sim N_d(\theta, \mathcal{I}_E(\theta)^{-1}),
\]
Example: normally distributed data

Consider $X_1, \ldots, X_n$ with $X_i \sim N(\theta_1, \theta_2)$, with both $\theta_1$ and $\theta_2$ unknown. We write $\theta = (\theta_1, \theta_2)^T$.

$$l(\theta; x) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \theta_2 - \frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2,$$

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\mathcal{I}_E(\theta) = \begin{pmatrix} \frac{n}{\theta_2^2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{pmatrix}.$$
For large $n$, the approximate distribution of $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)^T$ is

$$
\begin{pmatrix}
\hat{\theta}_1 \\
\hat{\theta}_2
\end{pmatrix}
\sim
N\left\{ 
\begin{pmatrix}
\theta_1 \\
\theta_2
\end{pmatrix},
\begin{pmatrix}
\frac{\theta_2}{n} & 0 \\
0 & \frac{2\theta_2^2}{n}
\end{pmatrix}
\right\}
$$
Confidence intervals based on asymptotic normality

Suppose we want a $100(1 - \alpha)\%$ confidence interval for any particular element of $\theta$, say $\theta_j$. For suitably large $n$, we have

$$\hat{\theta}_j \sim N(\theta_j, \gamma_{j,j}),$$

where $\gamma_{j,j}$ is the $\{j, j\}$ element of $\mathcal{I}_E(\theta)^{-1}$. This then gives us an approximate interval as

$$(\hat{\theta}_j - z_{1 - \frac{\alpha}{2}} \sqrt{\gamma_{j,j}}, \hat{\theta}_j + z_{1 - \frac{\alpha}{2}} \sqrt{\gamma_{j,j}}),$$
\( \theta \) unknown, so approximate \( \mathcal{I}_E(\theta) \) by observed information matrix

\[
\mathcal{I}_O(\theta) = \begin{pmatrix}
-\frac{\partial^2}{\partial \theta_1^2} l(\theta) & \cdots & -\frac{\partial^2}{\partial \theta_1 \partial \theta_d} l(\theta) \\
\vdots & & \vdots \\
-\frac{\partial^2}{\partial \theta_d \partial \theta_1} l(\theta) & \cdots & -\frac{\partial^2}{\partial \theta_d^2} l(\theta)
\end{pmatrix},
\]
evaluated at \( \theta = \hat{\theta} \).
\( \tilde{\gamma}_{i,j} \): the \( i, j \)th element of the inverse of \( \mathcal{I}_O(\theta) \), we use

\[
(\hat{\theta}_j - z_{1-\frac{\alpha}{2}} \sqrt{\tilde{\gamma}_{j,j}}, \hat{\theta}_j + z_{1-\frac{\alpha}{2}} \sqrt{\tilde{\gamma}_{j,j}}),
\]

as an approximate confidence interval. Since we know that \( \hat{\theta} \to \theta \) as \( n \to \infty \), with probability 1, we would expect \( \mathcal{I}_O(\theta) \) to be similar to \( \mathcal{I}_E(\theta) \) for large sample sizes.
4.2 Profile Likelihood

- RV $X$, density function $f$, parameters $\theta = \{\theta_1, \ldots, \theta_d\}$
- Given $x = (x_1, \ldots, x_n)$, only want inferences about *subset* of $\theta$.
- Partition $\theta$ into $\theta = (\theta_1, \theta_2)$ with $\theta_1$ the parameters of direct interest.
- $\theta_2$, the parameters not of direct interest are known as *nuisance parameters*. 
Example: $X \sim N(\mu, \sigma^2)$ with both $\mu$ and $\sigma^2$ unknown, though we may only be interested in the mean parameter $\mu$.

Can use asymptotic distribution of m.l.e. to derive confidence intervals for individual parameters.

Will now consider an alternative form of likelihood function which in some cases can produce more accurate confidence intervals.
Partitioning $\theta = (\theta_1, \theta_2)$, profile log-likelihood function for $\theta_1$ is

$$l_p(\theta_1; x) = \max_{\theta_2} l(\theta).$$  \hspace{1cm} (1)

To get the profile log-likelihood function for $\theta_1$:

1. Treat $\theta_1$ as a constant in $l(\theta; x)$.
2. Find the maximum likelihood estimate $\hat{\theta}_2$ in terms of the data $x$ and $\theta_1$.
3. Plug in this expression for $\hat{\theta}_2$ into the full log-likelihood $l(\theta; x)$ to get the profile log-likelihood $l_p(\theta_1; x)$.

- Writing $\theta = (\theta_i, \theta_{-i})$, plotting $l_p(\theta_i)$ gives us profile of log-likelihood surface viewed from $\theta_i$ axis.
- If $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ maximises $l(\theta)$, then $\hat{\theta}_1$ maximises $l_p(\theta_1)$ and $\hat{\theta}_2$ maximises $l_p(\theta_2)$.
- Useful exploratory tool; allows you to plot a likelihood $l_p(\theta_i)$ for a single parameter $\theta_i$.
- Can be used to derive more accurate confidence intervals.
Example 1

\[ X_1, \ldots, X_n \sim N(\mu, \sigma^2) \text{ i.i.d.} \]

\[ l(\mu, \sigma^2; \mathbf{x}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \]  \hfill (2)

Fixing \( \mu \), the MLE of \( \sigma^2 \) is \( \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \). Substituting this back into the full log-likelihood \( l(\mu, \sigma^2; \mathbf{x}) \), we get

\[ l_p(\mu; \mathbf{x}) = -\frac{n}{2} \log \left\{ \frac{1}{n} \sum (x_i - \mu)^2 \right\} - \frac{n}{2}. \]  \hfill (3)

Fixing \( \sigma^2 \), the MLE of \( \mu \) is \( \bar{x} \). The profile log-likelihood for \( \sigma^2 \) is

\[ l_p(\sigma^2; \mathbf{x}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2. \]  \hfill (4)
Inference using the deviance function

- Can construct CI for $\theta$ based on asymptotic normality of MLE. Alternative approach: use **deviance function**.

- For arbitrary $\theta^*$,

$$D(\theta^*) = 2\{l(\hat{\theta}; x) - l(\theta^*; x)\}. \quad (5)$$

$\hat{\theta}$ maximises log-likelihood, so $D(\theta^*) \geq 0$.

- If $D(\theta^*)$ is small, then $l(\theta^*)$ must be close to $l(\hat{\theta})$, which suggests that $\theta^*$ is a plausible estimate for the true unknown value of $\theta$.

- A confidence interval (or region if $\theta$ is a vector) could then be of the form

$$C = \{\theta^* : D(\theta^*) \leq c\}, \quad (6)$$

for some suitable value of $c$. 
With data \( x_1, \ldots, x_n \), for sufficiently large \( n \), it can be shown that at the true value of \( \theta \), \( D(\theta) \sim \chi^2_d \), where \( d \) is the dimensionality of \( \theta \).

An approximate \((1 - \alpha)\) confidence region for \( \theta \) is then given by

\[
C_\alpha = \{ \theta^* : D(\theta^*) \leq c_\alpha \},
\]

with \( c_\alpha \) the \((1 - \alpha)\) percentage point of the \( \chi^2_d \) distribution.

Usually more accurate than asymptotic normality approximation, may require greater computational effort.
Profile likelihood and the deviance function

- $\theta = (\theta_1, \theta_2)$, with $\theta_1$ a $k$-dimensional subset of $\theta$. **Profile deviance:**

$$D_p(\theta_1^*) = 2\{l(\hat{\theta}; x) - l_p(\theta_1^*; x)\},$$

with $\hat{\theta}$ the maximum likelihood estimator of $\theta$.

- Based on a sample of size $n$, with $n$ sufficiently large,

$$D_p(\theta_1) \sim \chi^2_k.$$  

- Can obtain a confidence interval for any element $\theta_i$ as

$$C_\alpha = \{\theta_i^*: D_p(\theta_i^*) \leq c_\alpha\},$$

again, with $c_\alpha$ the $(1 - \alpha)$ percentage point of the $\chi^2_1$ distribution.

- This will often be more accurate than the interval

$$\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\psi_{i,i}}$$

stated earlier.
Example: leukaemia data

- Leukaemia patients given drug, 6-mercaptopurine (6-MP), and the number of days \( t_i \) until freedom from symptoms is recorded:


A * denotes an observation censored at that time.

- Weibull model:

\[
f_T(t) = \alpha \beta (\beta t)^{\alpha-1} \exp\{- (\beta t)^\alpha\}\]   \hspace{1cm} (12)

for \( t > 0 \). \( \alpha = 1 \) gives exponential distribution.

- For censored data

\[
P(T > t) = \exp\{- (\beta t)^\alpha\}. \]   \hspace{1cm} (13)

\( d \): no. of uncensored observations, \( \sum_u \log t_i \): sum of all logs of the uncensored observations.
\[ l(\alpha, \beta; \mathbf{x}) = d \log \alpha + \alpha d \log \beta + (\alpha - 1) \sum_u \log t_i - \beta^\alpha \sum_{i=1}^n t_i^\alpha. \tag{14} \]

Treat \( \alpha \) as fixed, and find MLE of \( \beta \) as function of data and \( \alpha \).

\[ \hat{\beta} = \left( \frac{d}{\sum_{i=1}^n t_i^\alpha} \right)^{\frac{1}{\alpha}}. \tag{15} \]

The profile log-likelihood of \( \alpha \) is then given by

\[
\begin{align*}
\ell_p(\alpha) &= l(\alpha, \hat{\beta}) \\
&= d \log \alpha + \alpha d \log \left( \frac{d}{\sum_{i=1}^n t_i^\alpha} \right)^{\frac{1}{\alpha}} + (\alpha - 1) \sum_u \log t_i - d
\end{align*}
\]
Finding the full MLE \((\hat{\alpha}, \hat{\beta})\) cannot be done analytically, so numerical methods have to be used.

To construct the confidence interval, only need \(\hat{\alpha}\) that maximises \(l_p(\hat{\alpha})\), as \(l_p(\hat{\alpha}) = l(\hat{\alpha}, \hat{\beta})\).

For a 95% confidence interval, the 95th percentage point of the \(\chi^2_1\) distribution is 3.841. The confidence interval is then given by

\[
C_{0.05} = \{ \alpha^* : D_p(\alpha^*) \leq 3.841 \} \tag{16}
\]

\[
= [\alpha^* : 2\{l_p(\hat{\alpha}) - l_p(\alpha^*)\} \leq 3.841] \tag{17}
\]

\[
= \{ \alpha^* : l_p(\alpha^*) > l_p(\hat{\alpha}) - 3.841/2 \}. \tag{18}
\]

Numerically, we estimate the MLE \(\hat{\alpha}\) to be 1.35, with \(l_p(\hat{\beta}) = -41.66\).

From the graph, we can then read off the 95% confidence interval for \(\alpha\) as \((0.73, 2.2)\).

This contains the value 1, so the simpler exponential distribution is plausible for this dataset.
Example: machine component failure

- Level of corrosion $w$ in a machine component recorded and component tested until a failure is observed, at time $t$.
- Denote each observation by $(w_i, t_i)$, where $w_i$ is the level of corrosion, and $t_i$ is the failure time.
- Possible model: $T \sim \text{Exponential}(\lambda)$ distribution, with $\lambda$ a function of the corrosion level $w$:

$$\lambda = \alpha w^\beta. \quad (19)$$

$w$ treated as fixed, i.e. model distribution of the failure time conditional on the corrosion.
- $\beta = 0$ implies same expected time to failure, $\alpha^{-1}$ for all components, regardless of the corrosion level $w$. 
The density of a single observation \((w, t)\) is given by

\[
f_T(t) = \alpha w^\beta \exp\{-\alpha w^\beta t\}.
\]  

(20)

\[
l(\alpha, \beta; x) = n \log \alpha + \beta \sum_{i=1}^{n} \log w_i - \alpha \sum_{i=1}^{n} w_i^\beta t_i.
\]  

(21)

We can derive an expression for the profile log-likelihood of \(\beta\): Treating \(\beta\) as fixed, we obtain the MLE of \(\alpha\) as

\[
\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} w_i^\beta t_i}.
\]  

(22)

We then substitute this expression for \(\alpha\) in the full log-likelihood \(l(\alpha, \beta)\) to get the profile log-likelihood for \(\beta\):

\[
l_p(\beta; x) = n \log \left( \frac{n}{\sum_{i=1}^{n} w_i^\beta t_i} \right) + \beta \sum_{i=1}^{n} \log w_i - n.
\]  

(23)
Numerically, estimate $\hat{\beta} = 0.473$, with $l_p(\hat{\beta}; x) = -20.01$.

From graph, read off 95% confidence interval for $\beta$ as $(0.11, 0.95)$.

Doesn’t contain zero, and so there is clear evidence that $\beta \neq 0$

For comparison, compute confidence interval for $\beta$ using normal approximation.

Observed information matrix is given by

$$\left( \begin{array}{cc} -\frac{\partial^2 l}{\partial \alpha^2} & -\frac{\partial^2 l}{\partial \alpha \partial \beta} \\ -\frac{\partial^2 l}{\partial \alpha \partial \beta} & -\frac{\partial^2 l}{\partial \beta^2} \end{array} \right) = \left( \begin{array}{cc} n\alpha^{-2} & \sum w_i^\beta t_i \log w_i \\ \sum w_i^\beta t_i \log w_i & \alpha \sum w_i^\beta t_i (\log w_i)^2 \end{array} \right)$$

(24)
Obtain $\hat{\alpha}$ by substituting $\beta = 0.473$ into formula, gives $\hat{\alpha} = 1.099$.

Substitute $\alpha = 1.099$, $\beta = 0.473$ into observed information matrix, invert to get

$$V = \begin{pmatrix} 0.0534 & -0.0241 \\ -0.0241 & 0.0442 \end{pmatrix}.$$  \hspace{1cm} (25)

CI for $\beta$ using asymptotic normality is

$$\hat{\beta} \pm 1.96 \times 0.0442^{0.5},$$  \hspace{1cm} (26)

which gives $(0.0611, 0.8849)$. 