University of Glasgow
School of Mathematics & Statistics

Statistical Inference: Level M

Chapter 3
Parametric Inference (Part 2)
Interval Estimation and Multiparameter Models

February 2021

Dr Benn Macdonald
Room 225, Maths & Stats Building
Benn.Macdonald@glasgow.ac.uk

3.5 Interval Estimation using Likelihood

As we have already seen, point estimates obtained from different samples will generally take different values, and the point estimate calculated from a particular sample will not generally be equal to the unknown parameter (though we might anticipate it being fairly close). Therefore, instead of relying on one number as our estimate, we might try to use the sample data to identify a range of plausible values for the unknown parameter. Such a range of values is called an **interval estimate**. We clearly wish to calculate an interval estimate that is very likely to contain the true value of the population parameter.

In order to obtain an interval estimate of θ, we shall use the relative likelihood function, $R(\theta)$, which is defined by

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{MLE})}$$

Since $\hat{\theta}_{MLE}$ is defined to be that value of θ that maximises $L(\theta)$, then $R(\theta)$ must also reach its maximum at $\hat{\theta}_{MLE}$.

We can use the relative likelihood to define an interval estimate for θ. A range of plausible values for θ consists of all θ such that:

$$L(\theta) \geq p \times L(\hat{\theta}_{MLE})$$

i.e.

$$\frac{L(\theta)}{L(\hat{\theta}_{MLE})} \geq p$$

i.e.

$$R(\theta) \geq p$$

For p between 0 and 1, this is known as a $100p\%$ likelihood interval for θ.

It is often easier to determine likelihood intervals using the log relative likelihood function, defined by:

$$\begin{aligned} r(\theta) &= \log_e\{R(\theta)\} \\ &= \log_e\left\{\frac{L(\theta)}{L(\hat{\theta}_{MLE})}\right\} \\ &= \log_e\{L(\theta)\} - \log_e\{L(\hat{\theta}_{MLE})\} \\ &= \ell(\theta) - \ell(\hat{\theta}_{MLE}) \end{aligned}$$

In terms of $r(\theta)$, a $100p\%$ likelihood interval for θ is defined by

$$r(\theta) \geq \log_e(p)$$

In particular, a 50% likelihood interval is defined by:

$$r(\theta) \geq \log_e(0.5) = -0.693.$$

The Newton-Raphson method can be used to determine each of the bounds of the likelihood interval by setting:

$$g(\theta) = r(\theta) - \log_e(0.5) = 0$$

and by using appropriate starting estimates for the lower and upper bounds of the interval.

Example 7 continued - air conditioning failures

Data were recorded on the time (intervals in service-hours) between the failures of the air-conditioning equipment in a Boeing 720 aircraft. The 24 observations were:

```
50, 44, 102, 72, 22, 39, 3, 15, 197, 188, 79, 88,
46, 5, 5, 36, 22, 139, 210, 97, 30, 23, 13, 14.
```

From earlier in Chapter 3 we saw that if we assume an exponential distribution for the data with parameter θ (Expo(θ)), $\theta > 0$:

Likelihood:

$$L(\theta) \propto \prod_{i=1}^{n} \theta \; e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

Log-likelihood:

$$\ell(\theta) = n \log_e \theta - \theta \sum_{i=1}^{n} x_i$$

$$\hat{\theta}_{MLE} = \frac{1}{\bar{x}} = 0.016$$

The relative likelihood function is defined by:

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta}_{MLE})} \\ &= \frac{\theta^n e^{-\theta \sum x_i}}{\hat{\theta}_{MLE}^n e^{-\hat{\theta}_{MLE} \sum x_i}} = \frac{\theta^{24} e^{-1539\theta}}{0.016^{24} e^{-0.016 \times 1539}} \end{aligned}$$

A plot of the relative likelihood function is displayed in Figure 1. This illustrates $\hat{\theta}_{MLE} = \frac{1}{\bar{x}} = 0.016$.

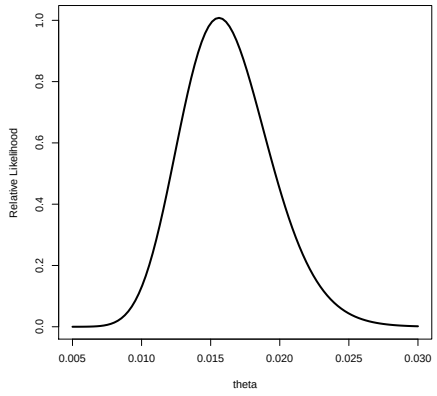

Figure 1: Relative likelihood function for air conditioning failures data.

The log relative likelihood is given by:

$$r(\theta) = \log_e \frac{L(\theta)}{L(\hat{\theta}_{MLE})}$$

As mentioned previously, a 50% likelihood interval is defined by:

$$r(\theta) \geq \log_e(0.5) = -0.693.$$

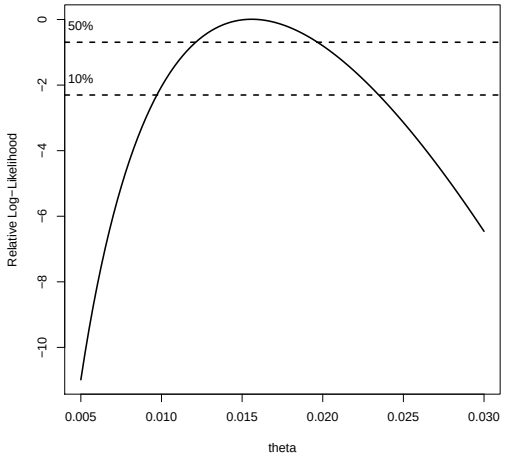

Figure 2: Relative log-likelihood function for air conditioning failures data with horizontal lines for 10% and 50% likelihood intervals.

Figure 2 shows a 10% and a 50% interval for θ in this example.

For example, it appears as though **a 50% interval for θ is approximately:**

The Newton-Raphson method introduced in Section 3.3.4 of this chapter can be used to determine each of the bounds of the 50% likelihood interval more accurately.

In this case the following iterative algorithm is used to find the bounds (B) for the interval, where θ_L is the lower bound and θ_U is the upper bound:

$$\theta_B^{(j+1)} = \theta_B^{(j)} - \frac{g(\theta_B^{(j)})}{g'(\theta_B^{(j)})}$$

with $g(\theta_B) = r(\theta) - \log_e(0.5)$, where:

$g(\theta_B) =$

$g'(\theta_B) =$

To start the Newton-Raphson algorithm, initial values can be estimated from a plot of the relative log-likelihood (Figure 2) i.e. for a lower bound start with 0.011 and for the upper bound start at 0.019. The results for θ_L and θ_U are provided in Tables 1 and 2.

Iteration	$\theta^{(j)}$	$g(\theta^{(j)})$	$g'(\theta^{(j)})$	$\frac{g(\theta^{(j)})}{g'(\theta^{(j)})}$
0	0.011	-0.573	642.82	-0.0009
1	0.012	-0.023	461	-0.0001
2	0.012	-0.023	461	-0.0001

Table 1: Newton-Raphson iterative algorithm to obtain the lower bound for an interval estimate around θ.

Iteration	$\theta^{(j)}$	$g(\theta^{(j)})$	$g'(\theta^{(j)})$	$\frac{g(\theta^{(j)})}{g'(\theta^{(j)})}$
0	0.019	0.232	-275.84	-0.0008
1	0.020	-0.076	-339	0.0002
2	0.020	-0.076	-339	0.0002

Table 2: Newton-Raphson iterative algorithm to obtain the upper bound for an interval estimate around θ.

So, in Example 7, a **50% likelihood interval for $\hat{\theta}_{MLE}$ is:**

$$(0.012, 0.020)$$

This means it is likely that the unknown population parameter θ lies in this range. It is likely that the average failure rate is in the range 0.012 to 0.020.

At this stage, we have not identified any particular reason for using $100p\%$ likelihood intervals for one value of p rather than another. In principle, $100p\%$ intervals for any p in the range 0 to 1 could be used. Notice that, as p increases the width of the $100p\%$ likelihood interval decreases.

3.6 Confidence Intervals in a Normal Model

We will now show that, in the Normal model, likelihood estimation is equivalent to another, classical approach that produces interval estimates (known as **confidence intervals**) with a more specific interpretation. This means that some values of p give more intuitively attractive interval estimates than others.

3.6.1 MLE for a Normal Model

Assume that $X_1, X_2, \ldots, X_n$, are independent $N(\mu, \sigma^2)$, where the standard deviation, σ, of the distribution is assumed to be known and always to be the same (no matter what the true population mean might be).

This means that μ is the only unknown parameter in the model, and we may use likelihood to obtain a point estimate of it in the usual way.

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$$

So

$$L(\mu; x_1, \ldots, x_n) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$$

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

where K is a constant.

$$\ell'(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

when,

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

i.e. when,

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

Now,

$$\ell''(\mu) = -\frac{n}{\sigma^2} < 0$$

for all possible μ

So $\hat{\mu}_{MLE} = \bar{x}$

3.6.2 Likelihood Interval for a Normal Model

In order to obtain likelihood intervals for μ, we must continue to treat σ as known and produce the log relative likelihood function for μ :

$$\begin{aligned} r(\mu) &= \ell(\mu) - \ell(\hat{\mu}) \\ &= -\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i-\mu)^2 - \sum_{i=1}^{n}(x_i-\bar{x})^2\right\} \\ &= -\frac{n}{2\sigma^2}\{\mu^2 - 2\mu\bar{x} + \bar{x}^2\} \\ &= -\frac{n}{2\sigma^2}(\mu-\bar{x})^2 \end{aligned}$$

A $100p\%$ likelihood interval $(0 < p < 1)$ for μ has the form:

$$\{\mu : r(\mu) \geq \log_e p\}$$

i.e.

$$\left\{\mu : \frac{n}{2\sigma^2}(\bar{x}-\mu)^2 \leq -\log_e p\right\}$$

i.e.

$$\left\{\mu : -\sqrt{-2\log_e p} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq \sqrt{-2\log_e p}\right\}$$

3.6.3 Likelihood Interval to Confidence Interval

Using properties of the Normal distribution, we can investigate this interval estimate in further detail. From our previous work on point estimation earlier in tutorial sheet 2, we know that:

$$\mathrm{E}\{\bar{X}\} = \mu \text{ and } \mathrm{var}\{\bar{X}\} = \tfrac{\sigma^2}{n}$$

If we assume that $X_1, X_2, \ldots, X_n$ are independent random variables, that are all Normally distributed, then it can also be shown (Probability M) that $\bar{X}$ is Normally distributed. So,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

The only random quantity in the formula for Z is $\bar{X}$. This means that any probability statement we make about Z is equivalent to a probability statement about $\bar{X}$.

The likelihood interval we obtained for μ can be written in the form:

$$\left\{-\sqrt{-2\log_e p} \leq Z \leq \sqrt{-2\log_e p}\right\}$$

where $Z \sim N(0,1)$.

We can find the probability of this event, as follows:

$$\begin{aligned}
& P(-\sqrt{-2\log_e p} \leq Z \leq \sqrt{-2\log_e p}) \\
= \; & P(Z \leq \sqrt{-2\log_e p}) - P(Z \leq -\sqrt{-2\log_e p}) \\
= \; & \phi(\sqrt{-2\log_e p}) - \phi(-\sqrt{-2\log_e p}) \\
= \; & \phi(\sqrt{-2\log_e p}) - \left[1 - \phi(\sqrt{-2\log_e p})\right] \\
= \; & 2\phi(\sqrt{-2\log_e p}) - 1
\end{aligned}$$

where $\phi()$ is the cdf of the standard normal distribution.

This is known as the coverage probability of the interval estimate. In the case of Normal models, p can be set to control the coverage probability.

For example,

$$\phi(1.96) = 0.975$$

so

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

This means that we obtain a coverage probability of 0.95 (or 95%) if we set:

$$\sqrt{-2\log_e p} = 1.96$$

i.e.

$$p = \exp\left[\frac{-(1.96)^2}{2}\right] = 0.1465$$

With this value of p, we obtain the interval estimate:

$$\{-1.96 \leq Z \leq 1.96\}$$

$$\left\{-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right\}$$

$$\left\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right\}$$

This is known as a **confidence interval for** μ.

Since it has **95% coverage**, it is called a **95% confidence interval**. Implicitly, the interval is based on a probability statement about the sample data, in particular about $\bar{X}$.

3.6.3.1 Interpretation of 95% Confidence Intervals

A 95% confidence interval will not always contain the true value of the parameter. In fact, on average only 95% of such intervals will do so.

On 95% of the occasions on which a 95% confidence interval is calculated from sample data, it will contain the true value of the parameter.

To illustrate this, Figure 3 shows 100 realisations of a 95% confidence interval for a given population mean μ. The 95% confidence intervals have been constructed from 100 random samples each of size 25 from the same population. If we randomly choose one realisation, the probability is 95% that

we choose an interval that contains the true population mean. This diagram demonstrates the coverage property of confidence intervals. Averaging over many samples, 95% of the 95% confidence intervals constructed using the formula will capture the true population mean. And 5% will not!

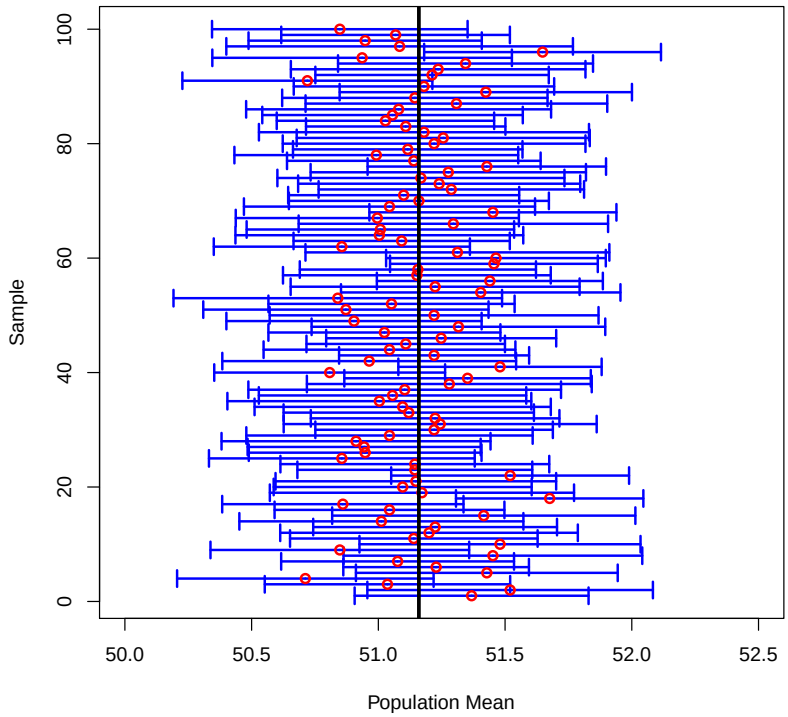

Figure 3: 100 95% confidence intervals for the population mean based on samples of size 25.

The beginning of section 3.6.3 illustrated that a 95% confidence interval is the same as the 14.65% likelihood interval defined by:

$$\{\mu : r(\mu) \geq \log_e(0.1465) = -1.92\}$$

Likelihood intervals do not generally coincide with confidence intervals; this is a feature of only a small number of models. We will now outline a general method for constructing confidence intervals using an approach that is not directly based on likelihood.

3.6.4 One-sample t-interval and t-test

A pivotal function is a function of the data, $\mathbf{X}$, and a parameter of interest, θ, which, when regarded as a random variable calculated at θ_T (the true value of θ), has a probability distribution whose form does not depend on any unknown parameter. We usually denote a pivotal function by $PIV(\theta_T, \mathbf{X})$.

We found a pivotal function for μ_T above (assuming σ is known):

$$PIV(\mu_T, \mathbf{X}) = \frac{\bar{X} - \mu_T}{\sigma/\sqrt{n}} = Z$$

This pivotal function has an $N(0,1)$ distribution whatever the value of μ_T.

Now assume that $X_1, X_2, \ldots, X_n$ are independent random variables, each with a $N(\mu_T, \sigma_T^2)$ distribution, and that we wish to estimate the population mean μ_T, but that both μ_T and σ_T (the population standard deviation) are unknown. This is now a two-parameter model and we will look at finding the maximum likelihood estimate for multiparameter cases shortly.

The pivotal function for μ_T found previously is not a pivotal function in this case since it depends on the unknown parameter σ. A pivotal function for μ_T in this case is found by replacing σ_T with its estimator, s (the sample standard deviation).

It can be shown that:

$$t = \frac{\bar{X} - \mu_T}{s/\sqrt{n}} \sim t(n-1)$$

where $t(n-1)$ is the Student's t distribution with $n-1$ degrees of freedom. See the Probability (level M) course.

All t distributions are symmetric and unimodal with expected value 0, just like the $N(0,1)$ distribution. t distributions are less peaked and more spread out than the $N(0,1)$ distribution.

t is identical to Z, except that the known value σ_T is replaced by an estimate s. This further source of uncertainty causes the distribution to be more spread out. As $n \to \infty$, then the $t(n-1)$ distribution tends towards the $N(0,1)$ distribution.

Having identified a suitable pivotal function (if one exists), then we can construct a confidence interval for our parameter of interest. There is no general method for deriving a pivotal function.

It is possible to produce $100c\%$ confidence intervals for any value of c in the range $0 < c < 1$.

Let $t_{1-\frac{(1-c)}{2}}(n-1)$ denote the value of a t random variable such that:

$$P(t \leq t_{1-\frac{(1-c)}{2}}(n-1)) = 1 - \frac{(1-c)}{2}.$$

For example, let $t_{0.975}(n-1)$ denote the value of a t random variable such that: $P(t \leq t_{0.975}(n-1)) = 0.975, c = 0.95$.

Since the $t(n-1)$ distribution is symmetric around 0, there is probability 0.95 (or 95%) that:

$$-t_{0.975}(n-1) \leq t = \frac{\bar{X} - \mu_T}{s/\sqrt{n}} \leq t_{0.975}(n-1)$$

i.e.

$$\mu_T \leq \bar{X} + t_{0.975}(n-1)s/\sqrt{n}$$

and

$$\mu_T \geq \bar{X} - t_{0.975}(n-1)s/\sqrt{n}$$

simultaneously.

This means that **a 95% confidence interval for the population mean**, $\boldsymbol{\mu_T}$ (when σ_T is unknown), is:

$$\left(\bar{x} - t_{0.975}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{0.975}(n-1)\frac{s}{\sqrt{n}}\right)$$

i.e.

$$\left(\bar{x} \pm t_{0.975}(n-1)\frac{s}{\sqrt{n}}\right)$$

Intervals with 95% coverage are obtained using $t_{0.975}()$ and intervals with 99% coverage are obtained using $t_{0.995}()$. The greater the coverage required, i.e. the larger the value of c that is used, the wider the confidence interval becomes.

Example 9 - Mean Annual Temperature in New Haven

As a simple example consider the following 60 year record of mean annual temperature in New Haven, Connecticut (in °F).

```
49.9 52.3 49.4 51.1 49.4 47.9 49.8 50.9 49.3 51.9 50.8 49.6 49.3 50.6
48.4 50.7 50.9 50.6 51.5 52.8 51.8 51.1 49.8 50.2 50.4 51.6 51.8 50.9
48.8 51.7 51.0 50.6 51.7 51.5 52.1 51.3 51.0 54.0 51.4 52.7 53.1 54.6
52.0 52.0 50.9 52.6 50.2 52.6 51.6 51.9 50.5 50.9 51.7 51.4 51.7 50.8
51.9 51.8 51.9 53.0
```

The summary statistics:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3069.6}{60} = 51.16, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1.602$$

A 95% confidence interval for the population mean annual temperature μ_T (when σ_T is unknown) is:

Conclusion

It can therefore be concluded that the population mean annual temperature is highly likely to lie in the range:

with a point estimate for $\hat{\mu}$ of 51.16°F.

3.6.4.1 Hypothesis Testing

The same pivotal function that was used above may also be used to carry out hypothesis tests about the parameter μ.

Suppose that we wish to investigate the following hypotheses about the population mean μ_T:

$H_0 : \mu_T = 51.2$

$H_1 : \mu_T \neq 51.2$

Under H_0:

$$\frac{\bar{X} - \mu_T}{s/\sqrt{n}} \sim t(n-1)$$

This means that t is a suitable test statistic. We reject H_0 in favour of H_1 when t is too far away from 51.2 to be consistent with H_0.

Adopting a significance level of $\alpha = 0.05$, we reject H_0 in favour of H_1 if:

$$|t| > t_{0.975}(n-1)$$

This very famous statistical procedure is known as a **one-sample t-test**.

The conclusions drawn from the one-sample t-test (with a two-sided alternative hypothesis) are guaranteed to agree with those from the confidence interval previously discussed, in the following sense. The null hypothesis $H_0 : \mu_T = 51.2$ is rejected at a significance level of $\alpha = 0.05$ if and only if a $100(1-\alpha)\%$ confidence interval for μ_T does not include the value 51.2.

This is the parametric equivalent of a **Wilcoxon Signed Ranks** test. The one-sample t-test requires all the assumptions of the Wilcoxon Signed Ranks test plus the assumption of Normality. Should the Normality assumption not seem reasonable, then the Wilcoxon Signed Ranks test should be used in preference to a procedure based on the t distribution.

The function `t.test()` can be used in `R` to perform this test and produce both a p-value for the hypothesis test and a confidence interval.

The `R` command and results are displayed below.

```
t.test(temp, mu=51.2)

One Sample t-test

data:  temp
t = -0.2448, df = 59, p-value = 0.8074
alternative hypothesis: true mean is not equal to 51.2
95 percent confidence interval:
 50.83306 51.48694
sample estimates:
mean of x
    51.16
```

Conclusion

Since the p-value is > 0.05 at 0.807, we do not reject H_0 and conclude that there is insufficient evidence that the population mean is different from 51.2^oF.

(Note: the interpretation of the confidence interval (50.8, 51.5) has been discussed on pages 13 and 14. Since the confidence interval contains the value 51.2, it provides the same conlusion as the p-value that there is insufficient evidence that the population mean is different from 51.2^oF.)

3.6.5 Two-sample t-interval and t-test

The above results can be extended to compare two populations, the parametric equivalent of a **Mann-Whitney test**.

Denote the sample values for sample 1 by $X_1, X_2, \ldots, X_m$ and for sample 2 by $Y_1, Y_2, \ldots, Y_n$, with sample means $\bar{X}$ and $\bar{Y}$ and sample standard deviations s_1 and s_2. Let μ_1 and μ_2 denote the corresponding population means and let σ_1 and σ_2 denote the population standard deviations.

We make the following assumptions:

- All recorded values are independent observations from their respective populations.
- The distribution of the variable of interest is the same in both populations, except, possibly, for a difference in the population means. In particular, this requires that the population standard deviations are equal, i.e. $\sigma_1 = \sigma_2 (= \sigma_T$, say)
- The distribution of the variable of interest is Normal in both the populations.

With the above assumptions, a pivotal function for the difference in population means, $\mu_1 - \mu_2$, is derived as follows.

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_T^2}{m}\right)$$

independently of

$$\bar{Y} \sim N\left(\mu_2, \frac{\sigma_T^2}{n}\right)$$

It can then be shown that (this will not be proved here):

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma_T^2\right)$$

i.e.

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\sigma_T^2}} \sim N(0, 1)$$

Since we typically do not know the true value σ_T^2, we obtain the pivotal function:

$$PIV(\mu_1 - \mu_2; X, Y) = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2}} \sim t(m + n - 2)$$

The second assumption suggests that we could obtain a better estimate of the common standard deviation, σ, by pooling the information from the two samples. The resulting, pooled estimate of σ is usually denoted s_p, where:

$$s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m + n - 2} = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2}{m + n - 2}$$

A complete proof of this result will not be considered here.

It follows from this result that a **symmetric 95% confidence interval for a difference in population means, $\boldsymbol{\mu_1 - \mu_2}$,** is given by:

$$(\bar{X} - \bar{Y}) \pm t_{0.975}(m + n - 2)\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}$$

If this interval contains the value 0, then it is plausible that $\mu_1 - \mu_2 = 0$, i.e. it is plausible that $\mu_1 = \mu_2$. We may then conclude that there is insufficient evidence that the two population means are significantly different. If the confidence interval does not contain the value 0, then it is not plausible that $\mu_1 = \mu_2$; and hence there is evidence that the two population means are different.

Note: In $t(m+n-2)$ the degrees of freedom $m+n-2$ is determined by the total sample size $m + n$ and then subtracting the number of parameters we are estimating, which in this case is 2 for $\hat{\mu_1}$ and $\hat{\mu_2}$.

Example 3 (from Chapter 2 cont..) - Preferred Room Temperatures

In a controlled environment laboratory, 10 men and 10 women were tested individually to determine the room temperature (^{0}F) they found to be most comfortable. The following results were obtained:

Women (X)	75	77	78	79	77	73	78	79	78	80
Men (Y)	74	72	77	76	76	73	75	73	74	75

Let μ_1, μ_2 denote the population mean preferred temperatures for women and men respectively.

A 95% confidence interval for the difference in population means $\mu_1 - \mu_2$:

$$\bar{X} = \sum_{i=1}^{10} \frac{X_i}{10} = 77.4, \quad \bar{Y} = \sum_{i=1}^{10} \frac{Y_i}{10} = 74.5$$

$$s_p^2 = \frac{\sum_{i=1}^{10}(X_i - \bar{X})^2 + \sum_{j=1}^{10}(Y_j - \bar{Y})^2}{10 + 10 - 2} = 3.3833$$

$$(\bar{X} - \bar{Y}) \pm t_{0.975}(m + n - 2)\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}$$

$$(77.4 - 74.5) \pm t_{0.975}(18)\sqrt{3.3833\left(\frac{1}{10} + \frac{1}{10}\right)}$$

$$2.9 \pm 2.101 \times 0.8226$$

$$2.9 \pm 1.728$$

$$(1.17, 4.63)$$

A two-sample t-test can be performed in R using the command `t.test()`. The results from R are displayed below. In this case the hypotheses are:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

```
t.test(data~grp2, var.equal=TRUE)

Two Sample t-test

data:  data by grp2
t = 3.5254, df = 18, p-value = 0.002416
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.171787 4.628213
sample estimates:
mean in group F mean in group M
           77.4            74.5
```

Conclusion

Since the confidence interval (1.17, 4.63) does not contain zero and the p-value is < 0.05 at 0.002, there is a statistically significant difference between the mean preferred room temperatures of males and females. The mean difference is highly likely to lie between 1.17 and 4.63^{o}F.

3.7 Intervals based on Approximate Confidence

The results in the previous section produced confidence intervals with exact coverage properties within normal models. However, we can use results based on the large sample properties of Maximum Likelihood Estimators to provide confidence intervals for a range of statistical models. We will use these results without proof in this section and these results will be returned to later and in further courses.

3.7.1 Wilks Intervals

When θ is set to its true value θ_T then, approximately,

$$2[\ell(\hat{\theta}_{MLE}) - \ell(\theta_T)] \sim \chi^2_1$$

This means that the quantity $2[\ell(\hat{\theta}_{MLE}) - \ell(\theta_T)]$ is an approximate **pivotal quantity** for θ. The large sample distribution of $2[\ell(\hat{\theta}_{MLE}) - \ell(\theta_T)]$ is approximately χ^2.

A confidence interval stems from the result that

$$P\{2[\ell(\hat{\theta}_{MLE}) - \ell(\theta_T)] \leq \chi^2_1(c)\} \approx c,$$

where $\chi^2_1(c)$ denotes the cth quantile of the χ^2_1 distribution i.e. a χ^2_1 random variable will be less than $\chi^2_1(c)$ with probability c.

An approximate $100c\%$ confidence interval for θ is then defined by

$$\{\theta : 2[\ell(\hat{\theta}_{MLE}) - \ell(\theta)] \leq \chi^2_1(c)\}.$$

Notice that we can also define the confidence interval in terms of the relative log-likelihood function as

$$\{\theta : -2r(\theta) \leq \chi^2_1(c)\}$$

$$\{\theta : r(\theta) \geq -\frac{1}{2}\chi^2_1(c)\}$$

This approximation is valid provided the sample size (i.e. the number of independent observations) is large.

For 95% confidence this result is:

$$\{\theta : -2r(\theta) \leq \chi^2_1(0.95)\}$$

$$\{\theta : -2r(\theta) \leq 3.84\}$$

$$\{\theta : r(\theta) \geq -1.92\}$$

Confidence intervals obtained by this method usually have to be found by numerical search. Such intervals are sometimes known as **Wilks** intervals, after the person who discovered the distributional result.

*NB: Compare this result to that on page 10 for likelihood/confidence intervals.

Example 7 continued - air conditioning failures

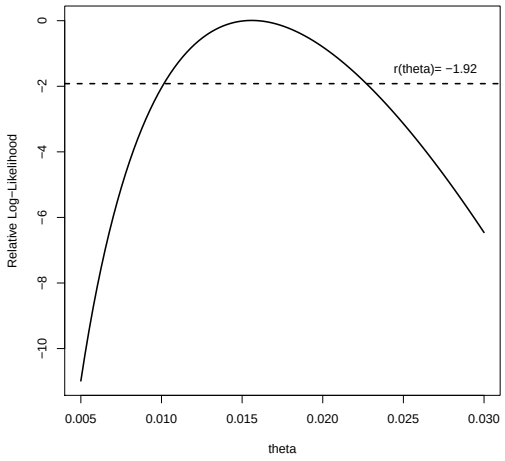

Figure 4: Relative log likelihood for the air conditioning failures data.

It can be seen from Figure 4 that the Wilks interval is approximately:

To obtain the Wilks interval with approximate confidence 0.95 the Newton-Raphson algorithm can be used. This is very similar to the routine described for calculating $100p\%$ likelihood intervals in Section 3.5. Starting values can be estimated from Figure 4 for the lower and upper bounds of the interval i.e. 0.010 and 0.024 respectively.

In this case the following iterative algorithm is used to find the bounds (B) for the interval, where θ_L is the lower bound and θ_U is the upper bound:

$$\theta_B^{(j+1)} = \theta_B^{(j)} - \frac{g(\theta_B^{(j)})}{g'(\theta_B^{(j)})}$$

with $g(\theta_B) = r(\theta) + 1.92$.

The Newton-Raphson algorithm converged in two steps in `R` to the Wilks interval with approximate confidence 0.95 of:

$$(0.010, 0.023)$$

It is therefore highly likely that $\hat{\theta}_{MLE}$ lies in the range $(0.010, 0.023)$.

Invariance Properties

Wilks intervals have attractive invariance properties. Suppose θ and β are parameters and $\beta = g(\theta)$ where g is some strictly increasing or decreasing function. If (θ_a, θ_b) is a Wilks interval for θ then the Wilks interval that would be obtained by re-parameterizing in terms of β would have end-points $g(\theta_a)$ and $g(\theta_b)$. A related appealing property of the Wilks intervals is that all the parameter values inside the interval have a higher likelihood than those values outside the interval.

3.7.2 Wald Intervals

This approach is based on another distributional approximation. This states that when θ_T is the true value of θ then, approximately,

$$\hat{\theta}_{MLE} \sim N(\theta_T, 1/k(\mathbf{x})).$$

The large sample distribution of a maximum likelihood estimate $\hat{\theta}$ is approximately normal with mean θ_T and variance $1/k(\mathbf{x})$. This result holds as the sample size tends to infinity.

Recall that $k(\mathbf{x})$ denotes the sample information, defined as $-l''(\hat{\theta}_{MLE})$. This means that, approximately,

$$\frac{\hat{\theta}_{MLE} - \theta_T}{\sqrt{1/k(\mathbf{x})}} \sim N(0,1)$$

and so the quantity $(\hat{\theta}_{MLE} - \theta_T)/\sqrt{1/k(\mathbf{x})}$ is an approximate pivotal quantity for θ.

This means that

$$P\{-z \leq \frac{\hat{\theta}_{MLE} - \theta_T}{\sqrt{1/k(\mathbf{x})}} \leq z\} \approx c,$$

where z denotes the point in a standard normal distribution below which lies probability $1 - (1-c)/2$, namely $\Phi^{-1}(1 - \frac{(1-c)}{2})$.

An approximate 100c% confidence interval for $\boldsymbol{\theta}$ is therefore given by

$$(\hat{\theta}_{MLE} - z\ \sqrt{1/k(\mathbf{x})},\ \hat{\theta}_{MLE} + z\ \sqrt{1/k(\mathbf{x})}).$$

Typically $c = 0.95$ and hence $z = \Phi^{-1}(0.975) = 1.96$.

These are called **Wald** intervals, again after the person who proposed this approach.

Note:

$$\text{var}\{\hat{\theta}_{MLE}\} \approx \frac{1}{k(\mathbf{x})}$$

The intervals are usually very easy to obtain relative to Wilks intervals, but their properties are less satisfactory for finite sample sizes. In particular the

intervals are not invariant, so that different parameterizations of a model lead to fundamentally different intervals. A related property is that the intervals are always symmetric, so that unless the log likelihood is symmetric we will actually end up including some parameter values in the interval that are *less likely* than some values outside the interval. These properties do not fit well with the general likelihood approach.

Sometimes however, the Wald interval is so much easier to obtain than the Wilks interval that it is sensible to use the approach, and in this case the intervals are generally at their best if a parameterization is used with respect to which the likelihood is reasonably symmetric over the width of the confidence interval.

Example 7 continued - air-conditioning failures

From earlier in the chapter we know that:

Log-likelihood:

$$\ell(\theta) = n \log_e \theta - \theta \sum_{i=1}^{n} x_i$$

$$\ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$

$$\hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

$$\ell''(\theta_{MLE}) = -\frac{n}{\theta^2}$$

and this is < 0 for all $\theta > 0$.

So, we have

$$k(\mathbf{x}) = -l''(\hat{\theta}_{MLE}) = \frac{n}{\hat{\theta}^2}$$

Since $\hat{\theta}_{MLE} = 0.016$, the standard error of $\hat{\theta}_{MLE}$ is

The 0.975 quantile of the standard normal distribution is 1.96.

An approximate 95% confidence interval for θ is then

Induced confidence intervals/ confidence sets

If A is a $100c\%$ confidence interval for a parameter θ and g is a strictly increasing or strictly decreasing function of θ then $B = \{g(\theta^*) : \theta^* \in A\}$ is a $100c\%$ confidence interval for $g(\theta)$. The proof is almost trivial. If A is one of the $100c\%$ of sets containing the true θ then B contains the true $g(\theta)$, while if A is one of the $100(1-c)\%$ of confidence sets not containing the true θ then B will not include the true $g(\theta)$. This establishes that B has the correct coverage probability.

If the strict monotonicity/unique image property of g does not hold then we can not generally construct an interval with the correct coverage probability. In this case any A containing the true θ would still result in a B containing the true $g(\theta)$, but now it is also possible for some A not containing the true θ to induce a set B which does include the true $g(\theta)$. i.e. in this case B has too high a coverage probability.

For example, if $[a, b]$ is a 95% confidence interval for θ then $[\log(a), \log(b)]$ is a 95% CI for $\log(\theta)$, while $[1/b, 1/a]$ is a 95% CI for $1/\theta$.

3.8 Problems with more than one parameter

3.8.1 An example with two parameters

So far only single parameter models have been considered, but the method of likelihood works in the same way when a model involves several parameters.

Example 9 continued - Annual mean temperatures in New Haven

As a simple example consider the 60 year record of mean annual temperature in New Haven, Connecticut example again.

A normal probability model may be reasonable for these data, as indicated previously. We would then treat the data as observations of independent identically distributed random variables X_i each with p.d.f.

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

where μ and σ are unknown parameters.

From page 6,

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

The log likelihood is then

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

A plot of the log-likelihood function is displayed in Figure 5 and a contour plot, Figure 6, is also a useful way of representing the function.

We can identify the maximum likelihood estimates simply by examining this function in Figure 6 carefully: $\hat{\mu} = 51.16$, $\hat{\sigma} = 1.266$.

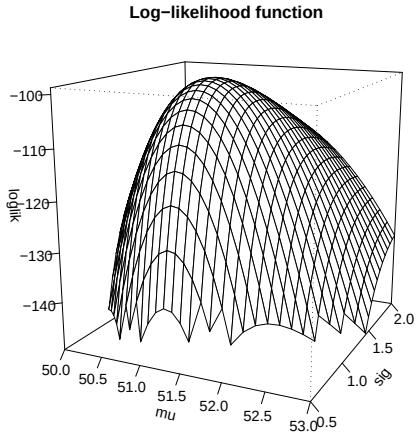

Figure 5: Log-likelihood function for New Haven Temperature Data.

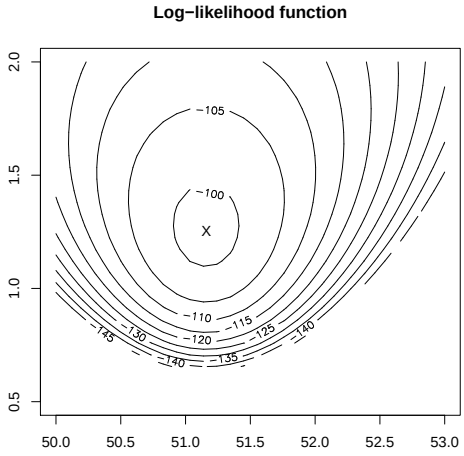

Figure 6: Contour plot of log-likelihood function for New Haven Temperature Data.

Maximization of $\ell()$ follows the same approach as in the single parameter case. **First differentiate ℓ w.r.t. each of the parameters:**

Now we set both these derivatives to zero and **solve** the resulting pair of simultaneous equations **for μ and σ.**

We must also check that the estimates are **maximum** likelihood estimates. This involves evaluating the second derivative matrix, or **Hessian (H)**, of $\ell()$ w.r.t. each of the parameters at $\hat{\mu}, \hat{\sigma}$.

That is

$$\mathbf{H} = \left(\begin{array}{cc} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{array} \right) \Bigg|_{\hat{\mu}, \hat{\sigma}}$$

For the turning point at $\hat{\mu}, \hat{\sigma}$ to be a maximum, $\mathbf{H}$ must be *negative definite*. This means that the quantity $a^T \mathbf{H} a$ is negative for every vector a. (In other words, the second derivative in every direction is negative.)

Therefore

$$\left(\begin{array}{cc} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{array} \right) =$$

So

$$\mathbf{H} = \left(\begin{array}{cc} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{array} \right) \Bigg|_{\hat{\mu}, \hat{\sigma}}$$

It is easy to see that, for any vector $a = (a_1, a_2)^T$, the quantity $a^T \mathbf{H} a = -(a_1^2 \ \frac{n}{\hat{\sigma}^2} + a_2^2 \ \frac{2n}{\hat{\sigma}^2})$. The matrix $\mathbf{H}$ is therefore negative definite and so in this case we do have *maximum* likelihood estimates.

Figure 7 compares the data (histogram) and estimated model (smooth curve). The fitted model looks reasonably good.

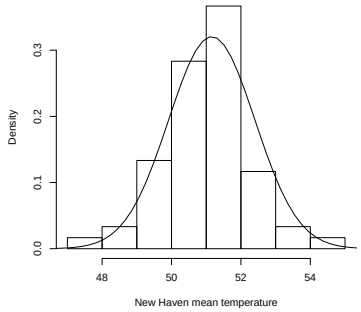

Figure 7: Data and fitted model for New Haven temperature data.

3.8.2 Maximising likelihood functions numerically

It was shown earlier in Chapter 3 that sometimes the likelihood functions cannot be solved algebraically, see Section 3.3.4. We introduced the Newton-Raphson algorithm to solve the likelihood functions numerically.

These ideas can be extended to vector parameters. For **vector parameters** in the Newton-Raphson method, good initial estimates are required, as in the one parameter case, for each of the parameters in the model. The following formula is then used to find successively better estimates; the iterations proceed until convergence.

The iterative formula is:

$$\hat{\boldsymbol{\theta}}^{(j+1)} = \hat{\boldsymbol{\theta}}^{(j)} - \mathbf{H}^{-1}\mathbf{g}.$$

where,

$$\mathbf{g} = \begin{pmatrix} \left.\frac{\partial \ell}{\partial \theta_1}\right|_{\hat{\boldsymbol{\theta}}^j} \\ \left.\frac{\partial \ell}{\partial \theta_2}\right|_{\hat{\boldsymbol{\theta}}^j} \\ \cdot \\ \cdot \end{pmatrix} \quad \text{and} \quad \mathbf{H} = \begin{pmatrix} \left.\frac{\partial^2 \ell}{\partial \theta_1^2}\right|_{\hat{\boldsymbol{\theta}}^j} & \left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right|_{\hat{\boldsymbol{\theta}}^j} & \cdot & \cdot \\ \left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right|_{\hat{\boldsymbol{\theta}}^j} & \left.\frac{\partial^2 \ell}{\partial \theta_2^2}\right|_{\hat{\boldsymbol{\theta}}^j} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

In practice the `optim()` function in `R` is a very powerful command that can be used for numerical optimisation.

3.9 Confidence regions

Our results for Wilks and Wald intervals in the one parameter case can be extended to the multiparameter case.

3.9.1 Wilks Confidence Regions

A $100p\%$ likelihood region can be defined in terms of the relative log-likelihood function,

$$r(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})$$

An approximate 100c% Wilks confidence region for $\boldsymbol{\theta}$ is then defined by

$$\{\boldsymbol{\theta} : -2r(\boldsymbol{\theta}) \leq \chi^2_k(c)\}$$

where k is the number of parameters in the model, or, equivalently,

$$\{\boldsymbol{\theta} : r(\boldsymbol{\theta}) \geq -\frac{1}{2}\chi^2_k(c)\}$$

For example, in example 9 on the annual mean temperature data from New Haven we have two parameters, $\hat{\mu}$ and $\hat{\sigma}$, therefore $k = 2$, and hence,

$$\chi^2_2(0.95) = 5.99$$

So the threshold is $-5.99/2 = -3.00$ to 2 decimal places.

3.9.2 Wald confidence regions

We can use the idea of quadratic approximation to define a **Wald** confidence region.

We create a quadratic approximation around the MLE itself.

The quadratic approximation is

$$\ell(\boldsymbol{\theta}) \simeq \ell(\hat{\boldsymbol{\theta}}_{MLE}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{g} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})$$

where $\mathbf{g}$ denotes the vector of first derivatives of ℓ evaluated at $\hat{\boldsymbol{\theta}}_{MLE}$ and $\mathbf{H}$ is the usual Hessian matrix of second derivatives, evaluated again at $\hat{\boldsymbol{\theta}}_{MLE}$.

However, since $\hat{\boldsymbol{\theta}}_{MLE}$ maximises the log-likelihood, the first derivatives at $\hat{\boldsymbol{\theta}}_{MLE}$ are $\mathbf{0}$ i.e. $\mathbf{g} = \mathbf{0}$.

So,

$$\ell(\boldsymbol{\theta}) \simeq \ell(\hat{\boldsymbol{\theta}}_{MLE}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})$$

By subtracting

$$\ell(\hat{\boldsymbol{\theta}}_{MLE})$$

from both sides, we can write this as

$$r(\boldsymbol{\theta}) \simeq \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})$$

The Wilks confidence region

$$\{\boldsymbol{\theta} : r(\boldsymbol{\theta}) \geq -\frac{1}{2}\chi^2_k(c)\}$$

can therefore be approximated by the Wald region

$$\{\boldsymbol{\theta} : \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE}) \geq -\frac{1}{2}\chi^2_k(c)\}$$

or, equivalently,

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{K}(\mathbf{x}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE}) \leq \chi^2_k(c)\}$$

where $\mathbf{K}(\mathbf{x})$ denotes the sample information matrix, namely $-\mathbf{H}$.

Geometrically, this defines an ellipsoid in k-dimensional space.

3.9.2.1 Confidence interval for linear functions

Sometimes we do not need a confidence region for the entire set of parameters in the model, but we would like a confidence interval for a particular linear combination of parameters, say $\boldsymbol{b}^T\boldsymbol{\theta}$.

The distributional approximation above, extends our earlier result for 1-parameter problems.

$$\hat{\boldsymbol{\theta}}_{MLE} \sim N_k(\boldsymbol{\theta}, \mathbf{K^{-1}}(\mathbf{x}))$$

It then follows that

$$\boldsymbol{b}^T\hat{\boldsymbol{\theta}}_{MLE} \sim N(\boldsymbol{b}^T\boldsymbol{\theta}, \boldsymbol{b}^T\mathbf{K^{-1}}(\mathbf{x})\boldsymbol{b})$$

An approximate pivotal function for $\boldsymbol{b}^T\boldsymbol{\theta}$ is then given by

$$\frac{\boldsymbol{b}^T\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{b}^T\boldsymbol{\theta}}{\sqrt{\boldsymbol{b}^T\mathbf{K^{-1}}(\mathbf{x})\boldsymbol{b}}}$$

An **approximate 95% confidence interval for $\boldsymbol{b}^T\boldsymbol{\theta}$** is then given by

$$\boldsymbol{b}^T\hat{\boldsymbol{\theta}}_{MLE} \pm z\sqrt{\boldsymbol{b}^T\mathbf{K^{-1}}(\mathbf{x})\boldsymbol{b}}$$

where, $z = \Phi^{-1}(1 - \frac{(1-c)}{2}) = \Phi^{-1}(0.975) = 1.96$

Example 10 - Comparing two population proportions

In a study reported in the *Lancet* (March 7th, 1998, pp. 700-708) patients with acute myeloid leukemia in first remission were randomly allocated to receive one of two treatments. The patients had been treated with chemotherapy and were then given no further treatment (NFT) or a bone-marrow transplant (BMT). Later it was noted whether or not each of the patients suffered a relapse. For patients in the 10-14 age range, 15 out of 50 patients who received a BMT suffered a relapse, whereas this happened for 26 out of 50 NFT patients.

What can we say about the probabilities of relapse in the two patient groups?

An appropriate model for the data is

$$X_1 \sim \text{Bi}(50, \theta_1)$$
$$X_2 \sim \text{Bi}(50, \theta_2)$$

where X_1 and X_2 denote the numbers of patient suffering a relapse in the BMT and NFT groups respectively.

The likelihood function is

$$L(\theta_1, \theta_2; \mathbf{x}) = K_1\theta_1^{15}(1-\theta_1)^{35}K_2\theta_2^{26}(1-\theta_2)^{24}$$

where K_1 and K_2 are constants.

The log-likelihood function is

$$\ell(\theta_1, \theta_2; \mathbf{x}) = 15 \log_e(\theta_1) + 35 \log_e(1 - \theta_1) + 26 \log_e(\theta_2) + 24 \log_e(1 - \theta_2) + K$$

The likelihood equations are

$$\frac{15}{\theta_1} - \frac{35}{1 - \theta_1} = 0$$
$$\frac{26}{\theta_2} - \frac{24}{1 - \theta_2} = 0$$

The maximum likelihood esimates are

$$\hat{\theta}_1 = 15/50 = 0.30$$
$$\hat{\theta}_2 = 26/50 = 0.52$$

This is confirmed by the Hessian matrix, **H**

$$\begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_2^2} \end{pmatrix} =$$

which is clearly negative-definite at all values of (θ_1, θ_2) in $0 < \theta_i < 1, i = 1, 2$. The sample information matrix evaluates this at the MLE's.

Notice that since $15 = 50\hat{\theta}_1$ and $26 = 50\hat{\theta}_2$

and similarly for the term in $\hat{\theta}_2$.

The sample information matrix is therefore

and its inverse is

$\mathbf{K^{-1}} = (-\mathbf{H})^{-1}$ is the variance covariance matrix of $\hat{\theta}_1$ and $\hat{\theta}_2$ and so the diagonal entries are the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$.

Notice that the variance of the sample proportions has been automatically generated by the likelihood procedure.

Numerically,

If we now wish to estimate $\theta_1 - \theta_2$ then we can use the result on linear combinations, with $b = (1, -1)^T$, to form **a 95% confidence interval for $\boldsymbol{\theta_1 - \theta_2}$**. This gives

$$\boldsymbol{b}^T\hat{\boldsymbol{\theta}}_{MLE} \pm 1.96\sqrt{\boldsymbol{b}^T\mathbf{K^{-1}}(\mathbf{x})\boldsymbol{b}}$$