# Inference

Simon Wood

## Contents

**1 Introduction** **2**
1.1 Some examples . . . 2
1.2 Books and Software . . . 3

**2 Likelihood** **4**
2.1 The Likelihood Function . . . 4

**3 Maximum Likelihood Estimation** **6**
3.1 Further examples . . . 8

**4 Numerical likelihood maximization** **10**
4.1 Single parameter example . . . 11
4.2 Vector parameter example . . . 13
4.3 Newton's method: problems and extensions . . . 15
4.4 Numerical maximization in S . . . 15
4.4.1 The basics of S . . . 15
4.4.2 Maximizing likelihoods with `optim` . . . 18
4.5 "Completely Numerical" calculations . . . 19

**5 Properties of Maximum Likelihood Estimators** **20**
5.1 Invariance . . . 20
5.2 Properties of the expected log-likelihood . . . 21
5.3 Consistency . . . 23
5.4 Large sample distribution of $\hat{\boldsymbol{\theta}}$ . . . 24
5.5 Example . . . 25
5.6 What to look for in a good estimator: minimum variance unbiasedness . . . 25

**6 Hypothesis tests** **26**
6.1 The generalized likelihood ratio test (GLRT) . . . 27
6.1.1 Example: the bone marrow data . . . 28
6.1.2 Simple example: Geiger counter calibration . . . 29
6.2 Why, in the large sample limit, is $2\lambda \sim \chi^2_r$ under $H_0$? . . . 30
6.3 What to look for in a good testing procedure: power etc. . . . 32
6.3.1 Power function example . . . 32

**7 Interval Estimation** **34**
7.1 Intervals based on GLRT inversion . . . 35
7.1.1 Simple single parameter example: bacteria model . . . 35
7.1.2 Multi-parameter example: AIDS epidemic model . . . 36
7.2 Intervals for functions of parameters . . . 38
7.3 Intervals based on $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1})$ . . . 39
7.3.1 Wald interval example: AIDS again . . . 40
7.4 Induced confidence intervals/ confidence sets . . . 40

**8 Assumptions of the large sample results** **40**

# 1 Introduction

Consider the situation in which you have some data and a statistical model describing how the data were generated, but that the model has some parameters the values of which are not known. Parametric statistical inference is concerned with deciding which values of these parameters are consistent with the data. This usually involves asking one of three related questions:

1. What value(s) of the parameter(s) are most consistent with the data?
2. Is some specified restriction on the value(s) that the parameter(s) can take consistent with the data?
3. What range of values of the parameter(s) is consistent with the data?

The methods for answering these questions are known as *point estimation*, *hypothesis testing* and *interval estimation*, respectively. Usually the statistical model is an attempt to describe the real circumstances which generated the data, and it is hoped that what we learn about the model by statistical inference tells us something about the reality that the model is trying to describe. Sometimes this hope is well founded. In this course we will simply assume that models are correct and develop theory on the basis of this assumption — in other courses you will cover the important topic of *model checking*.

## 1.1 Some examples

Thanks to Jim Kay for some of these...
Point Estimation

- Lightbulbs don't last for ever and customers and manufacturers need to know how long any particular type of bulb lasts on average, and what the range of lifetimes is. To find out a sample of lightbulbs can be taken and left on until they fail or exceed the available time for the experiment, with the failure times recorded. Assuming that we can write down a probability model for the failure times (an exponential distribution is not a bad model for tungsten filament bulbs) we would want to find the best estimate of the mean failure time of all bulbs.
- At the beginning of epidemics of new diseases (e.g. HIV/AIDS, SARS, BSE, vCJD) there is often considerable uncertainty about the underlying rate of increase in new cases of the disease, but it is important to try and estimate it. Typically records of the number of new cases will be available at regular intervals, but these will be quite variable. One simple model of early increase of the disease says that the number of cases $X_i$ in month $t_i$ is a Poisson random variable with mean $\lambda_i = \alpha e^{\beta t_i}$, where $\alpha$ is the initial number of cases and $\beta$ the rate of increase parameter. Given data, what would be the best estimates of these parameters? If we could answer this then short term forecasting would be possible.

Hypothesis testing

- After the Chernobyl nuclear disaster it was necessary to monitor radio-caesium levels in sheep grazing land. Spot radioactivity measurements are taken at randomly selected points. It is assumed that the measurements can be treated as observations of a random variable, whose mean corresponds to the overall radio-caesium radioactivity. It is of interest to know whether the observations are consistent with a mean level at or above the 'danger' level, or whether they provide sufficiently strong evidence that the mean level is below this, that the area can be considered 'safe'.
- Data were collected at the Ohio State University Bone Marrow Transplant Unit to compare two methods of bone marrow transplant using 23 patients suffering from non-Hodgkin's Lymphoma. The patients were each randomly allocated to one of two treatments. The *allogenic* treatment consisted of a transplant from a matched sibling donor. The *Autogenic* treatment consisted of removing the patient's marrow, 'cleaning it' and returning it after a high dose of chemotherapy. For each patient the time of death or relapse is recorded, but for patients who did well, only a

time before which the patient had not died or relapsed is recorded (the data are known as 'right censored').

| Treatment | Time (Days) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Allogenic | 28 | 32 | 49 | 84 | 357 | 933* | 1078* | 1183* | 1560* | 2114* | 2144* | |
| Autogenic | 42 | 53 | 57 | 63 | 81 | 140 | 176 | 210* | 252 | 476* | 524 | 1037* |

Here a * indicates a censored observation (data are from *Survival Analysis* Klein and Moeschberger).

A reasonable model of these data is that they follow exponential distributions with parameters $\theta_l$ and $\theta_u$ respectively (mean survival times are $1/\theta_{u/l}$). Medically the interesting question is whether the data are consistent with $\theta_l = \theta_u$ or whether separate parameters are required for the two groups. i.e. do the data provide evidence for a difference in survival times between the two groups or not?

Interval Estimation

- Hubble's law states that the further away a galaxy is the faster it is moving away from us. i.e. if $V$ is velocity (relative to us) and $d$ is distance from us then $V = \theta d$. $1/\theta$ is approximately the age of the universe. It is not easy to measure the distance to a galaxy, although its velocity is a little easier, hence the direct observations of $V$ and $d$ are best viewed as observations of random variables the means of which are related by Hubble's law. Astronomers would like to know what range of values of $\theta$ are consistent with the observed data, thereby obtaining bounds on the age of the universe (Of course there is another school of thought which holds that this quantity is known exactly, but this is beyond the scope of this course).

- A study examining the cosmetic effects of radiotherapy on women undergoing treatment for early stage breast cancer examined patients every 4-6 months or so and assessed whether or not they showed signs of moderate to severe breast retardation ( termed 'cosmetic deterioration'). Because of the infrequent examination the data consist of interval within which it is known that cosmetic deterioration set in. For women who showed no deterioration before dying, dropping out or the end of the study, only the time before which there was definitely no deterioration is known.

  Time to cosmetic deterioration or censoring (months).
  (0,7], (0,8], (0,5],(4,11], (5,11], (5,12], (6,10], (7,16], (7,14] (11,15],(11,18], 15*, 17*, (17,25], (17,25], 18*, (19,35], (18,26], 22*, 24*, 24*, (25,37], (26,40], (27,34], 32*, 33*, 34*, (36,44], (36,48], 36*, 36*, (37,44], 37*, 37*, 37*, 38*, 40*, 45*, 46*, 46*, 46*, 46*, 46*, 46*, 46*, 46*

  (*'s indicate right censored data. The data are from same source as the bone marrow data). Again the data can be modelled as being observations from an exponential distribution, with mean $(1/\theta)$. What range of $\theta$ values are consistent with the data — i.e. what is the range of average onset times that are consistent with the data?

Note the common theme in all the above examples — we have some data and a model of how the data were generated which has some parameters of unknown value. In each case we want to know about which values of the parameters are consistent with the data. This is what parametric statistical inference is about.

## 1.2 Books and Software

Silvey (1975) *Statistical Inference*, Chapman and Hall and Cox and Hinkley (1974) *Theoretical Statistics*, Chapman and Hall are both worth a look (and also cover material in subsequent Inference courses).

The S statistical language will be used in these notes. Either the commercial version S-PLUS or the free version, R, (see `cran.r-project.org`) should work.

# 2 Likelihood

Who said:

> First, we would not accept a treaty that would not have been ratified, nor a treaty that I thought made sense for the country.

was it (a) Tony Blair, (b) Ronald Reagan or (c) George W. Bush?

If you take apart the reasoning behind arriving at an answer it goes something like this: Could it be Tony Blair? Whatever his failings he can think on his feet and he's a trained lawyer, it's improbable that he would make such a slip, so it's unlikely to be him. Reagan? Everything was scripted for him and read from an auto-cue: unless the auto-cue broke down or he had a really dumb speech writer it's improbable that he would have said this, so it's unlikely to be him. George W. Bush often says things like "One year ago today, the time for excuse-making has come to an end": it's quite probable he would say such a thing, so it's likely that he is the author of this remark. Of the three possibilities Dubya is the most likely.

This type of reasoning lies behind the statistical idea of likelihood. The key idea is as follows. Suppose that we have some data and a probability model describing how the data were generated which has some parameters the values of which are not known ...

> **Parameter values which make the data appear relatively probable according to the model are more *likely* to be correct than parameter values which make the data appear relatively improbable according to the model.**

Similarly, subject to some caveats covered later, *models* which make the data appear relatively probable are more likely to be correct than models under which the data appear improbable. By 'relatively probable' is meant either having a relatively high probability (discrete data), or a relatively high probability density (continuous data).

To tediously labour the point: in the political quote example, the datum was the quote, the model was 'one of three politicians said it' and the unknown 'parameter' was the name of the politician. The data appeared most probable when the parameter took the value 'George W. Bush', so this is the most likely value of the parameter (and, as it happens, the correct one). Now let's formalize this idea.

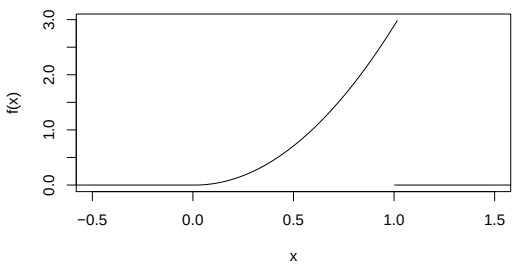
## 2.1 The Likelihood Function

Suppose you have observations of random variables $X_1, X_2 \ldots X_n$ and can write down their joint p.d.f. or joint p.m.f., $f(x_1, x_2, \ldots, x_n; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters of the probability model $f$. You can simply plug the observed values into $f(\cdot)$ and treat the result as a function of $\boldsymbol{\theta}$ alone. Joint p.d.f.s or p.m.f.s with the data plugged in in this way are known as "likelihoods" of the parameters. The easiest way to see how it works is to apply the idea to an example.

Suppose that we have $n$ data $\tilde{x}_i$ which we model as observations of independent random variables $X_i$ with p.d.f.

$$f(x_i) = \begin{cases} (\beta+1)x_i^{\beta} & 0 < x_i < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\beta$ is an unknown parameter. The p.d.f. is shown here for the case $\beta = 2$:



Since the $X_i$'s are independent, their joint p.d.f. is simply the product of their marginal p.d.f.'s:

$$f(x_1, x_2, \ldots, x_n; \beta) = \prod_{i=1}^{n} (\beta + 1) x_i^{\beta} \quad 0 < x_i < 1 \text{ for all } i$$

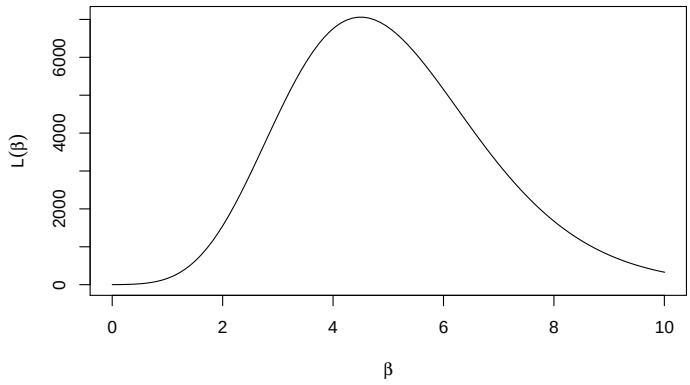
Plugging the observations, $\tilde{x}_i$, (of $X_i$) into this gives a result that indicates how probable observations are according to the model. Values of $\beta$ are not likely to be correct if they cause the model to suggest that what actually happened is improbable. A value of $\beta$ which causes the observed data to be probable under the model is much more likely to be correct (i.e to be the value that actually generated the data).

The joint p.d.f., $f(\cdot)$, with the observed data plugged in and considered as a function of $\beta$ is called the **likelihood function** (or simply the **likelihood**) of $\beta$:

$$L(\beta) = \prod_{i=1}^{n} (\beta + 1) \tilde{x}_i^{\beta}$$

$\ldots L$ will be relatively large for likely values of $\beta$ (values that make the observed data appear relatively probable under the model), and small for unlikely values of $\beta$ (values that make the observed data relatively improbable according to the model). The most likely value of $\beta$ is the one that maximises $L$. Here I have been quite careful to distinguish the dummy variables, $x_i$ that are arguments of the joint p.d.f. and the actual observed values $\tilde{x}_i$. Most textbooks are not so careful, and just write $x_i$ for both: once you are used to likelihood this doesn't cause confusion, and usually I won't bother being so careful.

To make the likelihood less abstract, suppose that we have 10 observations ($\tilde{x}_i$'s): 0.68, 0.75, 0.96, 0.90, 0.96, 0.62, 0.95, 0.98, 0.73, 0.91. Then we can plot $L$ as a function of $\beta$:




From the plot ...

1. What is the most likely value of the $\beta$?

2. Is the hypothesis $\beta = 0.1$ likely relative to the alternative that $\beta$ takes some other value?

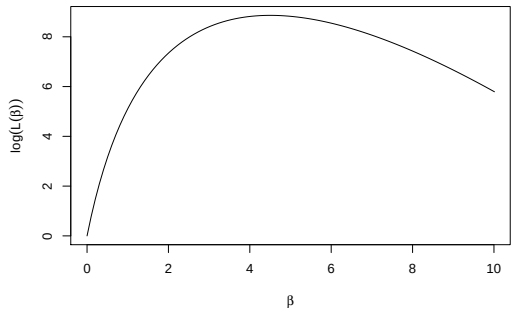
3. What range of values of $\beta$ are consistent with the data - i.e. have a reasonable likelihood? One way to answer this might be to find the range of values that have a likelihood that is at least 10% of the maximum likelihood.

In the rest of the course we'll develop well founded formal methods for approaching these types of question using likelihood — at the moment the important thing is to be clear about the basic principles.

# 3 Maximum Likelihood Estimation

Likelihood provides a good general approach to point estimation. Using the magnitude of the likelihood to judge how consistent parameter values are with the data, the parameter values which maximize the likelihood function would be judged to be most consistent with the data. Maximum likelihood estimation consists of finding the values of the parameters that maximize the likelihood and using these as the best estimates of the parameters. As we have seen, the approach is intuitively appealing, but as we will see later, it also leads to very general methods with some very good properties.

To actually find maximum likelihood estimates we don't have to resort to plotting $L$, as was done in section 2.1, but can find the $\beta$ that maximises $L$ mathematically. As an illustration let's continue with the example from section 2.1. The process is made easier if we note that $\log(L)$ will be maximised by the same value of $\beta$ that maximises $L$ (whichever $\beta$ gives the biggest $L$ value, must automatically give the biggest $\log(L)$ value). The following plot of $\log(L)$ against $\beta$ illustrates this:




... the shape of the plot is very different, but the maximum is at the same value of $\beta$. Define $l \equiv \log(L)$. Repeated application of the rules $\log(AB) = \log(A) + \log(B)$* and $\log(A^b) = b\log(A)$ yields:

$$l(\beta) = \log(L) = \sum_{i=1}^{n} \{\log(\beta + 1) + \beta \log(x_i)\} = n \log(\beta + 1) + \beta \sum_{i=1}^{n} \log(x_i)$$

To find the maximum of $l$ w.r.t. $\beta$ we need to find the value of $\beta$ at which $\mathrm{d}l/\mathrm{d}\beta = 0$.

$$\frac{\mathrm{d}l}{\mathrm{d}\beta} = \frac{n}{\beta + 1} + \sum_{i=1}^{n} \log(x_i)$$

and setting this to 0 implies that:

$$\hat{\beta} = -\frac{n}{\sum \log(x_i)} - 1$$

Use this expression to obtain the M.L.E. of $\beta$ from the 10 data given in section 2.1:

In this case the plot of $l$ vs. $\beta$ makes it clear that the estimate is a *maximum* likelihood estimate, but often this should be checked. Differentiating the log-likelihood again yields

$$\frac{\mathrm{d}^2 l}{\mathrm{d}\beta^2} = \frac{-n}{(1 + \beta)^2}$$

*Which generalizes to:

$$\log\left(\prod_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \log(x_i)$$

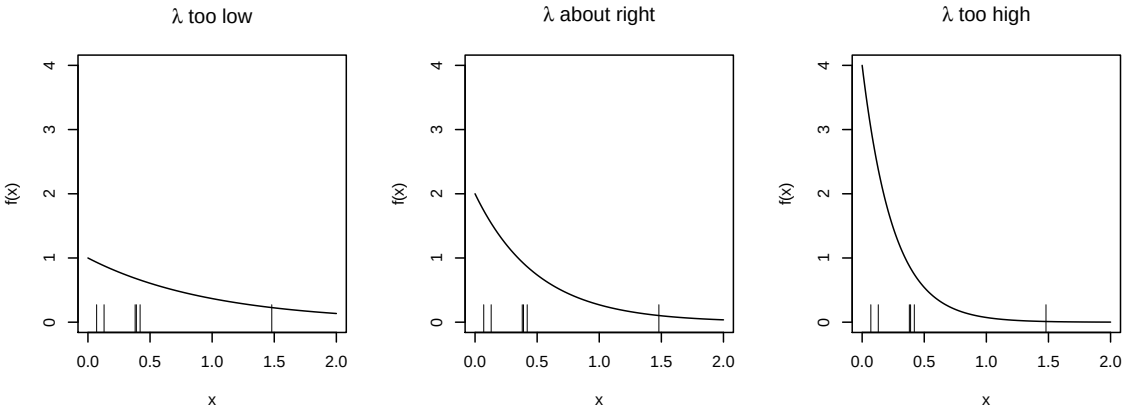
which is clearly negative at $\hat{\beta}$ (or indeed any valid $\beta$ value) indicating that $l(\hat{\beta})$ is indeed the *maximum* of the likelihood.

Now try a complete example. At relatively small time scales the time between calls arriving at a telephone exchange can be quite well modelled by independent exponential random variables, $X_i$, having p.d.f.

$$f(x_i) = \begin{cases} \lambda e^{-\lambda x_i} & x_i \geq 0 \\ 0 & x_i < 0 \end{cases}$$

Suppose that you have observations of the time in seconds between 6 calls, 1.48, 0.13, 0.42, 0.39, 0.38, 0.07. Use these to obtain a maximum likelihood estimate of $\lambda$. (First form $L$ then $log(L)$ and then find the maximum w.r.t. $\lambda$ — it's usually best to only plug in the actual numbers right at the end)

Having worked through an example of the mechanics of maximum likelihood estimation, it's worth revisiting what it's trying to do once more. The example that you have just done provides a useful way of illustrating this. The following 3 plots show the p.d.f. of the exponential for 3 different $\lambda$ values, with the 6 data that you have just used shown on the x axis.




On the left $\lambda$ is set very low, which means that all the data have rather low associated probability densities - the data don't seem all that probable under the model relative to the next panel. In the middle panel the 5 fairly low valued data get quite high probability densities while the remaining point still gets assigned a reasonable probability density, so overall the data look quite probable under the model, suggesting that this $\lambda$ is quite likely. In the final panel $\lambda$ is too high: the lower 5 values still have quite high probability densities, but now the probability density of the final datum is close to zero, which makes the data set look rather improbable according to the model. The mathematics you have just been through simply formalizes this approach.

Maximum likelihood estimation works in exactly the same way when we have several unknown parameters to estimate, the only difference being that we have to find the maximum of the likelihood w.r.t. several parameters, rather than just one. Don't forget that it also works just as well for discrete data as

for continuous data: all that changes is that p.m.f.s take the place of p.d.f.s.

## 3.1 Further examples

**Bone marrow**

Now consider again the bone marrow transplant data from section 1.1, concentrating just on the autogenic sample. Recall that a possible model for the time to relapse or death is that the data are observations (possibly censored) of independent exponential random variables, $T_i$, with p.d.f.

$$f(t_i) = \theta e^{-\theta t_i}, \quad t_i \geq 0.$$

The value of $\theta$ is unknown and must be estimated from the data, but in this case there is no 'obvious' estimator because some of the data are 'censored' observations — we only know that relapse or death occurred *after* some known time. Let $u$ be the set of indices of data that were uncensored (i.e. actual relapse or death times) and $c$ be the set of indices of data that were censored. We know the p.d.f. for the uncensored observations, but we also need to find the probabilities for the censored data. These are easily obtained ...

$$\Pr[T_i > t_i] = \int_{t_i}^{\infty} \theta e^{-\theta t} dt = \left[-e^{-\theta t}\right]_{t_i}^{\infty} = e^{-\theta t_i}$$

The likelihood is now made up of the product of the joint probability of the censored data and the joint p.d.f. of the uncensored data, under the model. i.e.

$$\begin{aligned} L(\theta) &= \prod_{i \in u} \theta e^{-\theta t_i} \times \prod_{i \in c} e^{-\theta t_i} \\ &= \theta^{n_u} \prod_{i=1}^{n} e^{-\theta t_i} \end{aligned}$$

where $n_u$ is the number of uncensored observations. As usual it is more convenient to work with the log-likelihood

$$l(\theta) = n_u \log(\theta) - \sum_{i=1}^{n} \theta t_i$$

Find the maximum likelihood estimate for $\theta$ and evaluate it, given that $\sum t_i = 3111$, and hence obtain the expected relapse (or death) time.

Applying the same estimator to the allogenic data gives an expected relapse time of about 1912 days — apparently much better, but there is a great deal of variability between individuals in this study, and to be sure that the apparent difference reflects more than a result of this variability we would need to use the hypothesis testing methods covered a little later in the course.

**A two parameter example**

So far only single parameter models have been considered, but the method works in the same way for more parameters. 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
```

A normal probability model may be reasonable for these data. i.e. we could try treating the data as observations of i.i.d. r.v.s $X_i$ each with p.d.f.

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

where $\mu$ and $\sigma$ are unknown parameters. As usual the log likelihood will be given by:

$$l(\mu, \sigma) = \sum_{i=1}^{n} \log(f(x_i))$$

that is

$$l(\mu, \sigma) = \sum_{i=1}^{n} \left[ -\log(\sqrt{2\pi}) - \log(\sigma) - (x_i - \mu)^2/(2\sigma^2) \right].$$

As is often the case with log-likelihoods, there is a constant in this expression, $-\sum \log(\sqrt{2\pi})$, which serves only to shift the log-likelihood function down by the same amount for all parameter values. Because it is not affecting the location of the m.l.e., or indeed any aspect of the shape of the log-likelihood function, our results will not be altered at all by simply dropping this term from the log-likelihood. For this reason constants which do not depend on the parameters are usually dropped from log-likelihoods, without comment, and the resulting relative log likelihood function is still generally referred to simply as 'the log likelihood function'. To save ink, this convention will generally be followed here ...

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

Maximization of $l$ follows the same approach as in the single parameter case. First differentiate $l$ w.r.t. the parameters,

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

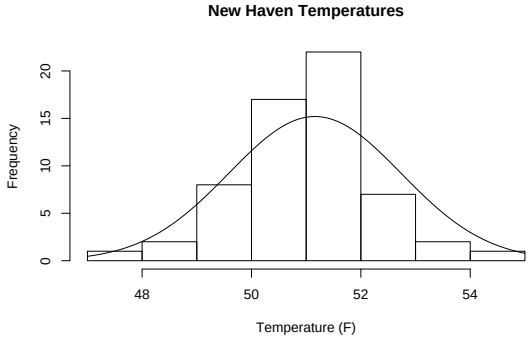
then set both the resulting expressions to zero and solve the resulting pair of simultaneous equations for $\hat{\mu}$ and $\hat{\sigma}$:

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

$$\mathbf{H} = \left( \begin{array}{cc} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l}{\partial \sigma \partial \mu} & \frac{\partial^2 l}{\partial \sigma^2} \end{array} \right)\Bigg|_{\hat{\mu},\hat{\sigma}} = \left( \begin{array}{cc} -\frac{n}{\sigma^2} & -\frac{2}{\sigma^3}\sum(x_i-\mu) \\ -\frac{2}{\sigma^3}\sum(x_i-\mu) & \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum(x_i-\mu)^2 \end{array} \right)\Bigg|_{\hat{\mu},\hat{\sigma}} = \left( \begin{array}{cc} -\frac{n}{\sigma^2} & 0 \\ 0 & \frac{-2n}{\hat{\sigma}^2} \end{array} \right).$$

For the turning point at $\hat{\mu}, \hat{\sigma}$ to be a maximum requires that $\mathbf{H}$ is *negative definite* — that is all its eigenvalues must be negative. The eigen-values of a diagonal matrix *are* the elements on the diagonal[†] so in this case we do have *maximum* likelihood estimates.

Plugging the data into the estimators yields $\hat{\mu} \simeq 51.2$ and $\hat{\sigma} \simeq 1.58$. The following plot compares the data (histogram) and estimated model (smooth curve).




The fitted model looks plausible, but perhaps underestimates in the centre of the data, while overestimating in the 'shoulders' of the distribution. This suggests that it might be worth trying a model somehow based on the t-distribution. However the p.d.f. of the t-distribution is much less easy to handle than that of the normal, which brings us to the next topic: numerical methods for maximizing likelihoods.

# 4 Numerical likelihood maximization

For most interesting models it is not possible to obtain simple closed form expressions for the m.l.e.s and numerical methods are used instead. Fortunately, some very simple, general and reliable methods are available. One of the best is **Newton's method**, which is based on the fact that by Taylor's Theorem we can always approximate a (sufficiently smooth) function by a quadratic, and we can always find the turning points of a quadratic.

Here are the steps of the method for maximizing a log likelihood $l(\theta)$ w.r.t. $\theta$ (which may be a single parameter or a parameter vector).

1. Start with an initial parameter guess $\hat{\theta}_0$ and set index $k = 0$.
2. Approximate $l(\theta)$ with a quadratic, by making a Taylor expansion around $\hat{\theta}_k$.
3. Find $\hat{\theta}_{k+1}$ to maximize this quadratic.
4. If $\partial l/\partial \theta|_{\hat{\theta}_{k+1}} \approx 0$ then stop, returning $\hat{\theta}_{k+1}$ as the m.l.e. Otherwise increase $k$ by one and return to step 2.

In the **single parameter** case step 2 is as follows. First write $\theta = \hat{\theta}_k + \Delta$ and then

$$l(\theta) \simeq l(\hat{\theta}_k) + \Delta \left.\frac{\partial l}{\partial \theta}\right|_{\hat{\theta}_k} + \frac{1}{2}\Delta^2 \left.\frac{\partial^2 l}{\partial \theta^2}\right|_{\hat{\theta}_k}.$$

[†]The eigenvectors are vectors with all elements zero, except for one, that element being 1: you can easily prove this yourself.

Differentiating w.r.t. $\Delta$ gives

$$\frac{\partial l}{\partial \Delta} \simeq \left.\frac{\partial l}{\partial \theta}\right|_{\hat{\theta}_k} + \Delta \left.\frac{\partial^2 l}{\partial \theta^2}\right|_{\hat{\theta}_k}$$

while setting the differential to zero and solving yields

$$\hat{\Delta} = -\left(\left.\frac{\partial^2 l}{\partial \theta^2}\right|_{\hat{\theta}_k}\right)^{-1} \left.\frac{\partial l}{\partial \theta}\right|_{\hat{\theta}_k}$$

which maximizes the quadratic approximation to $l$. Hence step 3 is

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \hat{\Delta}.$$

For **vector parameters** in general, step 2 is

$$l(\boldsymbol{\theta}) \simeq l(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\Delta}^T \mathbf{g} + \frac{1}{2}\boldsymbol{\Delta}^T \mathbf{H} \boldsymbol{\Delta}$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}_k + \boldsymbol{\Delta}$,

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

Differentiating w.r.t. each element of $\boldsymbol{\Delta}$ results in

$$\begin{pmatrix} \frac{\partial l}{\partial \Delta_1} \\ \frac{\partial l}{\partial \Delta_2} \\ . \\ . \end{pmatrix} = \mathbf{g} + \mathbf{H}\boldsymbol{\Delta}.$$

Setting each of these derivatives to zero and solving the resulting system of equations gives

$$\hat{\boldsymbol{\Delta}} = -\mathbf{H}^{-1}\mathbf{g}$$

as the maximizer of the quadratic approximation to the log-likelihood. Hence in this case step 3 amounts to

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - \mathbf{H}^{-1}\mathbf{g}.$$

## 4.1 Single parameter example

As an example of using Newton's method to maximize a likelihood in the single parameter case, consider an experiment on anti-biotic efficacy. A 1 litre culture of $5 \times 10^5$ cells (this figure known quite accurately) is set up and dosed with antibiotic. After 2 hours and every subsequent hour up to 14 hours after dosing 0.1ml of the culture is removed and the live bacteria in this sample counted under a microscope. The data are:

| Sample hour, $t_i$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Live bacteria count, $y_i$ | 35 | 33 | 33 | 39 | 24 | 25 | 18 | 20 | 23 | 13 | 14 | 20 | 18 |

A simple model for the sample counts, $y_i$, is that their expected value is

$$E(Y_i) = 50e^{-\delta t_i},$$

where $\delta$ is an unknown 'death rate' parameter (per hour) and $t_i$ is the sample time in hours. Given the sampling protocol, it is reasonable to assume that the actual counts are observations of independent Poisson random variables with this mean. The parameter $\delta$ must be estimated. Note that this example is different in kind from the bone marrow survival example — the main source of variability in this case is the sampling and not the random timing of individual deaths (because the population is so large in this case).

Writing $\mu_i \equiv E(Y_i)$, we have that the probability function for $Y_i$ is

$$f(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \quad \text{where} \quad \mu_i = 50e^{-\delta t_i}.$$

Therefore the log likelihood is

$$l(\delta) = \sum_{i=1}^{n} [y_i \log(\mu_i) - \mu_i - \log(y_i!)]$$

Dropping the $\delta$ independent constant $\sum \log(y_i!)$ and substituting in the model for $\mu_i$ gives

$$l(\delta) = \sum_{i=1}^{n} y_i[\log(50) - \delta t_i] - \sum_{i=1}^{n} 50e^{-\delta t_i}.$$

The derivatives are then

$$\frac{\partial l}{\partial \delta} = -\sum_{i=1}^{n} y_i t_i + \sum_{i=1}^{n} 50 t_i e^{-\delta t_i}$$

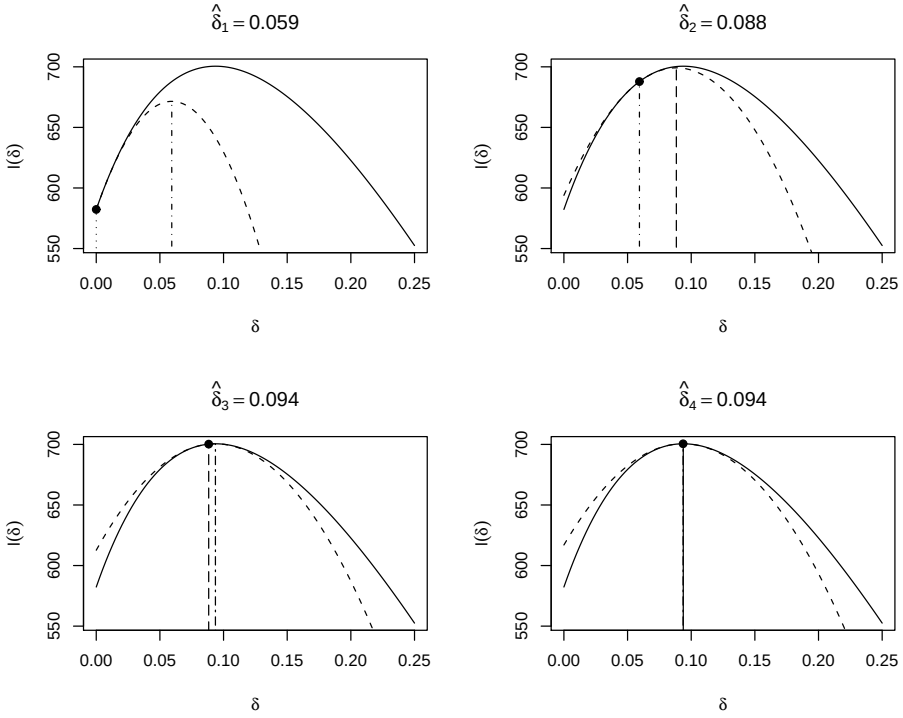
and

$$\frac{\partial^2 l}{\partial \delta^2} = -50 \sum_{i=1}^{n} t_i^2 e^{-\delta t_i}.$$

The presence of the $t_i$ term in $e^{-\delta t_i}$ makes it impossible to find a closed form solution for $\partial l/\partial \delta = 0$, so numerical methods are required here.

**Exercise** Given that $\sum y_i = 315$, $\sum t_i = 104$, $\sum y_i t_i = 2192$ and $\sum t_i^2 = 1014$, and starting from a guess $\hat{\delta}_0 = 0$, apply one step of Newton's method to obtain $\hat{\delta}_1$.

In fact it only takes 4 steps of Newton's algorithm to maximize the likelihood as the following plot shows. In each panel the continuous curve is the log likelihood, and the dashed curve is the quadratic approximation obtained by expanding the log-likelihood around the point marked •. As the iterations progress you can see how the maximum of the quadratic approximation gets closer to the maximum of the log-likelihood, until they eventually coincide. It follows of course that the $\hat{\delta}_k$ become ever closer to the m.l.e. as iteration progresses.




## 4.2 Vector parameter example

Now consider an example with a vector parameter. The following data are reported AIDS cases in Belgium in the early stages of the epidemic there.

| Year (19—) | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| New cases | 12 | 14 | 33 | 50 | 67 | 74 | 123 | 141 | 165 | 204 | 253 | 246 | 240 |

One important question early in such epidemics is whether control measures are beginning to have an impact, or whether the disease is continuing to spread essentially unchecked. A simple model for unchecked growth leads to an 'exponential increase' model (similar in structure to the bacteria model). The model says that the number of cases, $y_i$, are observations of independent Poisson r.v.s, with expected values

$$\mu_i = \alpha e^{\beta t_i}$$

where $t_i$ is the number of years since 1980. Proceeding broadly as in the bacteria example leads to a log likelihood

$$l(\alpha, \beta) = \sum_{i=1}^{n} y_i[\log(\alpha) + \beta t_i] - \sum_{i=1}^{n} \alpha e^{\beta t_i}$$

so the gradient vector is

$$\begin{pmatrix} \partial l/\partial \alpha \\ \partial l/\partial \beta \end{pmatrix} = \begin{pmatrix} \sum y_i/\alpha - \sum \exp(\beta t_i) \\ \sum y_i t_i - \alpha \sum t_i \exp(\beta t_i) \end{pmatrix}$$

and the Hessian is

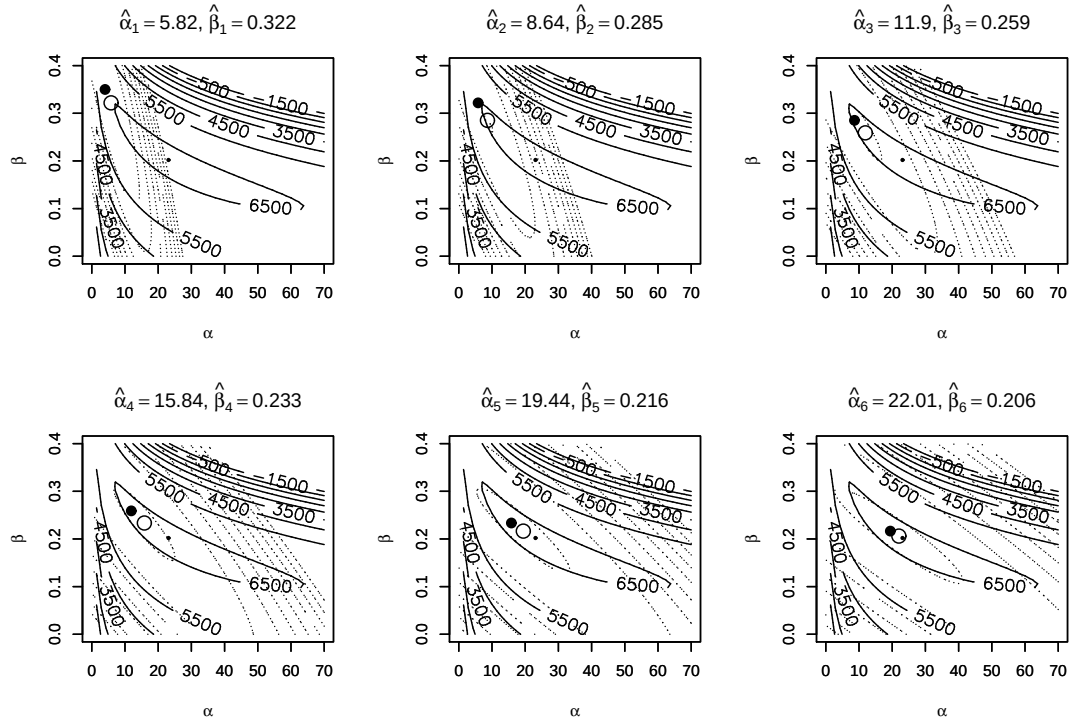
$$\begin{pmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 l}{\partial \theta_2^2} \end{pmatrix} = \begin{pmatrix} -\sum y_i/\alpha^2 & -\sum t_i e^{\beta t_i} \\ -\sum t_i e^{\beta t_i} & -\alpha \sum t_i^2 e^{\beta t_i}. \end{pmatrix}$$

A swift glance at the expression for the gradients should be enough to convince you that numerical methods will be required to find the m.l.e.s of the parameters. Starting from an initial guess $\hat{\alpha}_0 = 4$,

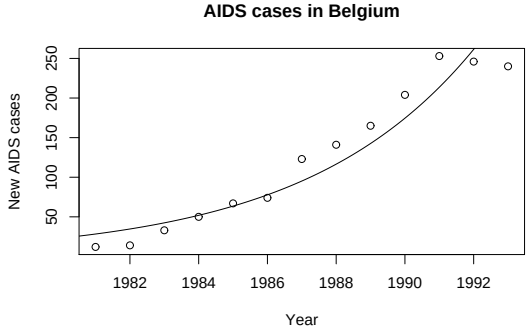
$\hat{\beta}_0 = .35$, here is the first Newton iteration.

$$\begin{pmatrix} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} 4 \\ .35 \end{pmatrix} \Rightarrow \mathbf{g} = \begin{pmatrix} 88.4372 \\ 1850.02 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} -101.375 & -3409.25 \\ -3409.25 & 154567 \end{pmatrix}$$
$$\Rightarrow \mathbf{H}^{-1}\mathbf{g} = \begin{pmatrix} -1.820 \\ 0.028 \end{pmatrix}$$
$$\Rightarrow \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{pmatrix} - \mathbf{H}^{-1}\mathbf{g} = \begin{pmatrix} 5.82 \\ 0.322 \end{pmatrix}$$

After a number of further iterations the likelihood is maximized at $\hat{\alpha} = 23.1$, $\hat{\beta} = 0.202$. The following figure illustrates the first 6 steps of the Newton method for this case. In each panel the continuous labelled contours show the log likelihood. The dashed unlabelled contours (which are at the same levels as the continuous) show the quadratic approximation to the likelihood based on the Taylor expansion at the point •. The maximum of the approximating quadratic is shown as ∘ in each case, and the true maximum is marked with a point.




The original data with the curve giving the expected cases per year according to the estimate model is shown here ...



The fit doesn't look too bad, but the points are perhaps not scattered around the curve as much as you would expect if the differences between data and model were no more than random variability — perhaps the exponential model was too pessimistic and a model which showed some slowing down would be better supported by the data (but perhaps, with 13 data, this is over-interpretation).

## 4.3 Newton's method: problems and extensions

For a likelihood that is continuous to second derivative and has a maximum within the allowable parameter space, Taylor's theorem guarantees that Newton's method will find that maximum *provided the initial parameter estimates are close enough to the m.l.e.*. Unfortunately 'close enough' varies from problem to problem. If the initial estimates are too far from the m.l.e. then the Newton iteration can diverge, with estimates actually getting further from the m.l.e. Furthermore, if the log-likelihood is an awkward shape then it can happen that the Hessian is not negative definite over some regions of parameter space, and hence the quadratic approximation has no unique maximum. In fact both divergence and indefinite Hessians can occur for the AIDS model log-likelihood shown in the previous section, if the initial parameter guess is unfortunate.

In practice two simple steps usually prevent these problems leading to actual failure. The first is to treat $\boldsymbol{\Delta}$ not as defining the step to take, but merely as defining the *direction* in the parameter space along which to search for better parameter values. Evaluating the log-likelihood at a few locations along $\boldsymbol{\Delta}$, from $\boldsymbol{\theta}_k$, usually locates parameter values with higher likelihood, even if stepping all the way to $\boldsymbol{\theta}_k + \boldsymbol{\Delta}$ would decrease the likelihood. The second step only comes into play if the first fails (e.g. if the Hessian is indefinite — i.e. some of its eigenvalues are positive): in this case one can treat $\mathbf{g}$ as the direction in which to search for better parameters. Unless the m.l.e. has already been located, a sufficiently small step in this direction is guaranteed to increase the likelihood.

Given that Newton type methods will almost always be performed by computer, a possible 'enhancement' is to estimate the required derivatives numerically, rather than working out exact expressions for them. For an appropriate choice of small interval $h$,

$$\frac{\partial l}{\partial \theta} \simeq \frac{l(\theta + h) - l(\theta)}{h}$$

and equivalent formulae can be obtained for second derivatives. With careful choice of $h$ these *finite difference approximations* can be very accurate. For example, if $l$ can be calculated to 14 significant figures, then the derivative can usually be obtained to 7 significant figures.

Newton type methods based on finite difference approximate derivatives can be very effective, but only if finite difference intervals $h$ are very carefully chosen ... a topic best left to experts. Fortunately, there are plenty of expert implementations freely available, which make it easy to maximize most likelihoods without having to calculate any derivatives for yourself, as we shall now see.

## 4.4 Numerical maximization in S

The S statistical computing language has built in routines for numerical maximization of likelihoods (or any other function, in fact). To use these facilities you need to spend a little time learning some new things, and revising some old things about the language. There are two implementations of S in widespread use. S-PLUS is the commercial implementation and is available in the labs here. R (`cran.r-project.org`) is a free version written by university statisticians: it tends to be a faster and some of its facilities, including those for maximization, are arguably a bit better than the S-PLUS versions.

### 4.4.1 The basics of S

S stores all data and variables in named objects. You create objects by assigning a value to them using the **assignment** operator `<-`. Here are some examples. Anything written after a `#` character is a comment.

```
a<-1            # create a variable 'a' with value 1
a<-3.5          # give 'a' the value 3.5
```

```
b<-c(1,3.1,4.7)  # create an array called 'b' with elements 1, 3.1 and 4.7
b[2]<- -1.3      # assign the second element of 'b' the value -1.3
d<-1:4           # create an array 'd' of length 4 containing 1,2,3,4
M<-matrix(0,3,2) # create a 3 row, 2 column matrix of 0's called M
M[3,1]<-2        # assign 3rd row of column 1 of M the value 2
M[2,]<-c(3,1)    # set the second row of M to values 3,1
M[,2]<-b         # set the second column of M to the values in b
```

To see what is contained in an object, simply type its name. For example

```
> M<-matrix(1:6,2,3)
> M
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

where > is the S 'command prompt' (indicates that S is waiting for you to type in a command).

Sometimes it is useful to create objects known as **lists**, which are simple named lists of other objects. For example, following on from the previous examples

```
my.list<-list(vec=b,mat=M) # create a list called 'my.list'
```

would create a list containing 2 items `vec` is an array of the 3 numbers that were in `b` and `mat` is a copy of the matrix `M`. The items in the list are accessed by using the list name followed by a `$` character followed by the item name. For example `my.list$vec`.

**Array arithmetic** works almost as it does in Minitab, except that no `let` command is needed and `=` is replaced by `<-`. Here are some examples, with the results printed too.

```
> a<-1:3
> b<-a/2           # divide each element of a by 2
> b
[1] 0.5 1.0 1.5
> d<-a*b           # multiply arrays a and b element by element - result in d
> d
[1]  0.5  2.0 4.5
> d<-a^2*exp(b)   # d_i = a_i^2 exp(b_i) for i=1..3
> d
[1]  1.648721 10.873127 40.335202
```

**Matrix arithmetic** is also straightforward. For example if $\mathbf{M}$ is an $n \times n$ matrix and $\mathbf{b}$ is an $n$-vector (i.e. an array of length $n$) then

```
y<-M%*%b
qf.1<-t(b)%*%M%*%b
x<-solve(M,b)
qf.2<-t(b)%*%solve(M,b)
```

form $\mathbf{y} = \mathbf{Mb}$, $QF_1 = \mathbf{b}^T\mathbf{Mb}$, $\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$ and $QF_2 = \mathbf{b}^T\mathbf{M}^{-1}\mathbf{b}$ respectively. `%*%` performs matrix multiplication, `t()` transposes a matrix and `solve(M,b)` forms $\mathbf{M}^{-1}\mathbf{b}$ (by solving $\mathbf{Mx} = \mathbf{b}$ for $\mathbf{x}$).

**Functions** are one of the real strengths of S. Hundreds of built-in functions are available for performing all sorts of data manipulation and statistical tasks, but you can also write your own functions (e.g. log likelihood functions). `solve()`, `t()` and `matrix()` are examples of functions that have already been used above. Functions have a name, take *arguments* supplied between brackets `()` and return an object of some sort as a result, which you can assign to a named object if you wish (if you don't assign the result of a function to something then it will usually be printed and will then be lost). Some functions don't return an object, but produce a 'side-effect' such as a plot. Here are some useful functions

```
> a<-1:4          # create array 1,2,3,4
> s.a<-sum(a)     # sum the array and store result in 's.a'
> s.a             # print value of 's.a'
[1] 10
> prod(a)         # use prod() function to form product of elements of a
[1] 24
> dpois(a,4)      # probability of elements of a if Poi(4) r.v.s
[1] 0.07326256 0.14652511 0.19536681 0.19536681
> ppois(a,3)      # Pr[A<=a] if A~Poi(3) using function for c.d.f of Poi
[1] 0.1991483 0.4231901 0.6472319 0.8152632
```

More complicated functions may have many possible arguments, most of which have default values which they will take if you don't specify a value. For example the `plot()` function can take optional arguments `xlab`, `ylab` and `main` for the axes labels and titles, but will simply use its own defaults if you don't supply any. e.g.

```
x.a<-1:20;y.a<-x+rnorm(20) # simulate straight line data with N(0,1) errors
plot(x.a,y.a)              # plot with default annotation
plot(x.a,y.a,xlab="x",ylab="y",main="A plot about nothing") # plot with annotation
```

Notice how the later arguments have been supplied in the form `name=something` — the naming of arguments lets you supply some arguments and omit others without confusing S.

Writing your own functions is straightforward. Suppose that you want to write a function called `my.func` which will take two parameters $b_1$ and $b_2$, say, two n-vectors $\mathbf{x}$ and $\mathbf{y}$ and form

$$\sum_{i=1}^{n} x_i^{b_1} \exp(-b_2 y_i)$$

here is how to do it ...

```
my.func<-function(b,x,y)   # create a function 'my.func' defined by code between { and }
{ res<-sum(x^b[1]*exp(-b[2]*y)) # calculate required quantity
  res                           # return value
}
```

Once created the function can be called like any other. For example:

```
b<-c(1.5,1);x<-runif(20);y<-runif(20) # create test data
my.func(b,x,y)
```

Here's the function in action ...

```
> b<-c(1.5,1);x<-runif(20);y<-runif(20) # create test data
> my.func(b,x,y)                        # call function
[1] 4.783403
```

Of course the following would work just as well:

```
> params<-c(1.5,1);x.data<-runif(20);response<-runif(20)
> my.func(b=params,x=x.data,y=response)
[1] 4.783403
```

and for this simple function we don't *have* to use the argument names: `my.func(params,x.data,response)` would also have worked.

**Help** is available by typing `?` followed by the function you want help with. e.g. `?plot`.

### 4.4.2 Maximizing likelihoods with `optim`

`optim` is a useful general function for maximizing functions that is available in both S-PLUS[‡] and R. To use it you must first define the function that you want to maximize. The first argument of this function must be the vector of parameters with respect to which you want to maximize, but the function may have as many further arguments as you need.

As a simple example let's create a function the maximum of which we know:

```
boring<-function(b,y)
{ -sum((b[1]-y)^2)}
```

this function is simply

$$-\sum_i (b_1 - y_i)^2$$

which is easily shown to be maximized when $b_1 = \bar{y}$. Now let `optim` maximize the function:

```
> y<-runif(20)                    # simulate some data
> b.0<-1                          # starting guess at parameter
> opt.res<-optim(par=b.0,fn=boring,method="BFGS",control=list(fnscale= -1),y=y) # maximize
> opt.res$par                     # returned maximizing parameter
[1] 0.5445544
> mean(y)                         # c.f. known maximizer
[1] 0.5445544
```

The arguments of `optim` are as follows. `par` is the array of initial parameter values, corresponding to the parameters of your function; `fn` is the name of the function that you want to maximize; `method` selects one of several maximization methods: "BFGS"[§] is a variant on the Newton method known as a "quasi-Newton" method; `control` is a list containing any of a bewildering number of optional control parameters — `fnscale` should be set to -1 to make the function perform maximisation rather than minimization. After all the `optim` arguments, you then supply the (named) arguments for your function, just `y` in this case. If you need more information see `?optim`.

#### AIDS in Belgium again

As a practical example let's use `optim` to find the m.l.e.s for the Belgium AIDS data model. First enter the data.

```
y<-c(12,14,33,50,67,74,123,141,165,204,253,246,240)
t<-1:13
```

Now write a function to evaluate the log likelihood

```
ll<-function(b,y,t)  # b[1] is alpha, b[2] is beta
{ sum(y*(log(b[1])+b[2]*t) - b[1]*exp(b[2]*t))
}
```

and then call `optim` to maximize the likelihood.

```
> aids.fit<-optim(c(10,.1),ll,control=list(fnscale=-1),y=y,t=t)
> aids.fit
$par:
[1] 23.1265075  0.2021064
$value:
[1] 6602.287
```

[‡]To use `optim` in S-PLUS you must first load the add-on 'MASS' library — type `library(MASS)` at the command prompt to do this. In R `optim` is built in.

[§]which I remember as "Big Friendly Giant Steps".

Notice how simple this is — there is not even any need to calculate derivatives. As we will see shortly, the Hessian matrix at the m.l.e.s is often required when using likelihood, but `optim` will also return an estimate of this for you, if you set its argument `hessian` to `TRUE`. Although not required, it is possible to supply derivatives to `optim`, and if this is done the algorithm will usually converge more quickly and the estimated Hessian will be more accurate.

## 4.5 "Completely Numerical" calculations

Sometimes the likelihood itself is actually difficult or tedious to write down explicitly. For example, since there is no closed form expression for the integral of the normal p.d.f. we cannot write down an explicit expression for the likelihood for interval data on normal r.v.s. Similarly the expression for the p.d.f. of a t-distribution is unpleasantly complicated. However, S and other statistics packages have built in functions for evaluating many p.d.f.s and c.d.f.s, including the c.d.f. for the normal and the p.d.f. for the t-distribution. As a consequence it is very straightforward to write an S function to evaluate the log-likelihood in such cases, even though the likelihood may be awkward to write down.

To illustrate this point let's return to the New Haven temperature data from section 3.1, but try modelling them with a t-distribution, which puts more probability in the tails of the distribution than the normal. To motivate the model, first note that the original model could have been re-written as

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1).$$

Standard transformation theory tells us that if the p.d.f. of $Z_i$ is $f_z(z_i)$ then the p.d.f. of $X_i$ will be $f_z((x_i-\mu)/\sigma)/\sigma$, which written out in full is the normal p.d.f. used in section 3.1.

To try and overcome the deficiencies of the original model let's try the model

$$Z_i = \frac{X_i - \mu}{\sigma} \sim t_m$$

where $m$ is the degrees of freedom of the t-distribution. If $f_{t,m}$ is the p.d.f. of the t-distribution with $m$ degrees of freedom, then standard transformation theory says that the p.d.f. of $X_i$ is now

$$f_x(x_i) = f_{t,m}((x_i-\mu)/\sigma)/\sigma$$

so that the log-likelihood is

$$l(\sigma,\mu,m) = \sum_{i=1}^{n} \log(f_x(x_i))$$

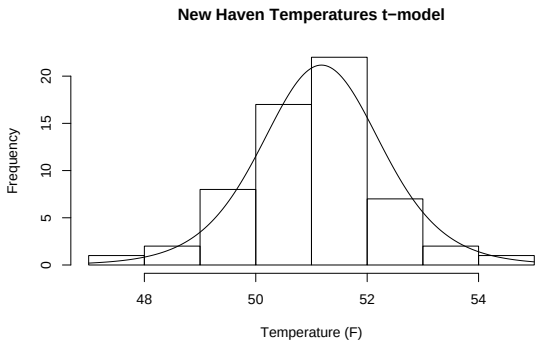
which is easily evaluated in S. A suitable function is:

```
logL<-function(b,x,df)              # df is m
{ z<-(x-b[1])/exp(b[2])             # transform data. b[1] is mu, exp(b[2]) is sigma
  sum(log(dt(z,df=df)/exp(b[2])))   # sum the logs of the p.d.f. of x
}
```

where `b[2]` is the log of $\sigma$ — a trick to ensure that the estimate of $\sigma$ cannot become negative. $m$/`df` has not been included in the parameter array `b` because it is a discrete parameter, and `optim` only deals with continuous parameters. To find the m.l.e. for $m$, `optim` can simply be called for a range of $m$ values, and the one resulting in the highest maximized likelihood chosen. A few trials soon established that $\hat{m} = 8$.

Here is the S code to maximize the likelihood and plot the results.

```
b<-optim(c(50,log(2)),logL,control=list(fnscale=-1),df=8,x=nhtemp)
mu.hat<-b$par[1];sigma.hat<-exp(b$par[2])                    # extract estimates
hist(nhtemp,xlab="Temperature (F)",main="New Haven Temperatures t-model")
temp<-seq(47,55,length=200)                 # sequence of temperatures to predict over
lines(temp,n*dt((temp-mu.hat)/sigma.hat,df=8)/sigma.hat)
```




The new model does look better than the old one, and reflecting this, the likelihood has also increased. Now $\hat{\mu} \simeq 51.2$ as before, but $\hat{\sigma} \simeq 1.1$. However, the main point to note is that the model has been estimated without actually having to write out the likelihood except in the most general terms.

# 5 Properties of Maximum Likelihood Estimators

The preceding sections should have given you a feel for how very general the method of maximum likelihood estimation is, and how straightforward it is to use with modern statistical software. In this section we'll consider some of the theoretical properties of maximum likelihood estimates, which tend to strengthen the case for using this method. In most cases only outline proofs will be given: detailed enough to provide understanding of how each result comes about, without going into the mass of detail often required for a fully rigorous proof.

## 5.1 Invariance

Consider an observation $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ of a vector of random variables with joint p.m.f. or p.d.f. $f(\mathbf{x}, \theta)$, where $\theta$ is a parameter with m.l.e. $\hat{\theta}$. If $\beta$ is a parameter such that $\beta = g(\theta)$ where $g$ is any function, then the maximum likelihood estimate of $\beta$ is $\hat{\beta} = g(\hat{\theta})$, and this property is known as *invariance*. So, when working with maximum likelihood estimation, we can adopt whatever parameterization is most convenient for performing calculations, and simply transform back to the most interpretable parameterization at the end.

Invariance holds for any $g$, but a proof is easiest for the case in which $g$ is a one to one function so that $g^{-1}$ is well defined. In this case $\theta = g^{-1}(\beta)$ and maximum likelihood estimation would proceed by maximizing the likelihood

$$L(\beta) = f(\mathbf{x}, g^{-1}(\beta))$$

w.r.t. $\beta$. But we know that the maximum of $f$ occurs at $f(\mathbf{x}, \hat{\theta})$, by definition of $\hat{\theta}$ so it must be the case that the maximum occurs when $\hat{\theta} = g^{-1}(\hat{\beta})$, i.e.

$$\hat{\beta} = g(\hat{\theta})$$

is the m.l.e. of $\beta$. Note that invariance holds for vector parameters as well.

**Exercise:** Suppose that weight (in kg) of newborn baby girls can be modelled as independent observations on a random variable $X$ with p.d.f.

$$f(x) = \sqrt{\delta/\pi} e^{-\delta(x-\alpha)^2}$$

where $\alpha$ and $\delta$ are positive parameters. Based on a sample of weights of newborn babies m.l.e.s $\hat{\alpha} = 3.5$ and $\hat{\delta} = 2$ are obtained. What is the m.l.e. for the standard deviation of the weight of a newborn baby girl?

## 5.2 Properties of the expected log-likelihood

The key to proving and understanding the large sample properties of maximum likelihood estimators lies in obtaining some results for the expectation of the log-likelihood and then using the convergence in probability of the log-likelihood to its expected value which results from the law of large numbers. In this section, some simple properties of the expected log likelihood are derived.

Let $x_1, x_2, \ldots, x_n$ be independent observations from a p.d.f. $f(x, \theta)$ where $\theta$ is an unknown parameter with true value $\theta_0$. Treating $\theta$ as unknown, the log-likelihood for $\theta$ is

$$l(\theta) = \sum_{i=1}^{n} \log[f(x_i, \theta)] = \sum_{i=1}^{n} l_i(\theta)$$

where $l_i$ is the log-likelihood given only the single observation $x_i$. Treating $l$ as a function of random variables $X_1, X_2, \ldots, X_n$ means that $l$ is itself a random variable (and the $l_i$ are independent random variables). Hence we can consider expectations of $l$ and its derivatives.

**Result 1:**

$$E_0\left(\left.\frac{\partial l}{\partial \theta}\right|_{\theta_0}\right) = 0$$

Where the subscript on the expectation is to emphasize that the expectation is w.r.t. $f(x, \theta_0)$. The proof goes as follows (where it is to be taken that all differentials are evaluated at $\theta_0$, and there is sufficient regularity that the order of differentiation and integration can be exchanged)

$$\begin{aligned} E_0\left(\frac{\partial l_i}{\partial \theta}\right) = E_0\left(\frac{\partial}{\partial \theta}\log[f(X,\theta)]\right) &= \int \frac{1}{f(x,\theta_0)}\frac{\partial f}{\partial \theta} f(x,\theta_0)dx \\ &= \int \frac{\partial f}{\partial \theta}dx = \frac{\partial}{\partial \theta}\int f dx \\ &= \end{aligned}$$

That the same holds for $l$ follows immediately.

Result 1 has the following obvious consequence:

**Result 2:**

$$\text{var}\left(\left.\frac{\partial l}{\partial \theta}\right|_{\theta_0}\right) = E_0\left[\left(\left.\frac{\partial l}{\partial \theta}\right|_{\theta_0}\right)^2\right]$$

and it can further be shown that

**Result 3:**

$$\mathcal{I} \equiv E_0\left[\left(\left.\frac{\partial l}{\partial \theta}\right|_{\theta_0}\right)^2\right] = -E_0\left[\left.\frac{\partial^2 l}{\partial \theta^2}\right|_{\theta_0}\right]$$

where $\mathcal{I}$ is referred to as the **information** about $\theta$ contained in the data. The terminology refers to the fact that if the data tie down $\theta$ very closely (and accurately) then the log likelihood will be sharply peaked in the in the vicinity $\theta_0$ (i.e. high $\mathcal{I}$), whereas data containing little information about $\theta$ will lead to an almost flat likelihood and low $\mathcal{I}$.

The proof of result 3 is simple. For a single observation, result 1 says that

$$\int \frac{\partial \log(f)}{\partial \theta} f dx = 0$$

Differentiating again w.r.t. $\theta$ yields

$$\int \frac{\partial^2 \log(f)}{\partial \theta^2} f + \frac{\partial \log(f)}{\partial \theta} \frac{\partial f}{\partial \theta} dx$$

but

$$\frac{\partial \log(f)}{\partial \theta} = \frac{1}{f} \frac{\partial f}{\partial \theta}$$

and so

which is

$$E_0 \left[ \left. \frac{\partial^2 l_i}{\partial \theta^2} \right|_{\theta_0} \right] = -E_0 \left[ \left( \left. \frac{\partial l_i}{\partial \theta} \right|_{\theta_0} \right)^2 \right]$$

The result follows very easily (given the independence of the $l_i$).

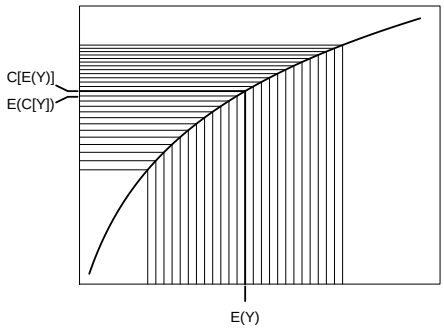
Now notice that result 1 says that the expected log likelihood has a turning point at $\theta_0$, while since $\mathcal{I}$ is non-negative, result 3 indicates that this turning point is a maximum. So the expected log likelihood has a maximum at the true parameter value. Unfortunately results 1 and 3 don't establish that this maximum is the global maximum of the expected log likelihood, but a slightly more involved proof shows that this is in fact the case.

**Result 4:** $E_0[l(\theta_0)] \geq E_0[l(\theta)] \ \ \forall \ \theta$

The proof is based on Jensen's inequality, which says that if $c$ is a concave function (i.e. has negative second derivative) and $Y$ is a random variable, then

$$E[c(Y)] \leq c(E[Y]).$$

The inequality is almost a statement of the obvious as the following figure illustrates:




Now consider the concave function log and the random variable $f(X,\theta)/f(X,\theta_0)$. Jensen's inequality implies that

$$E_0 \left[ \log \left( \frac{f(X,\theta)}{f(X,\theta_0)} \right) \right] \leq \log \left[ E_0 \left( \frac{f(X,\theta)}{f(X,\theta_0)} \right) \right].$$

Consider the right hand side of the inequality.

$$E_0 \left( \frac{f(X,\theta)}{f(X,\theta_0)} \right) = \int \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0) dx = \int f(x,\theta) dx = 1.$$

So, since $\log(1) = 0$ the inequality becomes

$$E_0\left[\log\left(\frac{f(X,\theta)}{f(X,\theta_0)}\right)\right] \leq 0$$

$$\Rightarrow E_0[\log(f(X,\theta))] \leq E_0[\log(f(X,\theta_0))]$$

from which the result follows immediately.

The above results were derived for continuous $X$, but also hold for discrete $X$: the proofs are almost identical, but with $\sum_{all\ x_i}$ replacing $\int dx$. Note also that although the results presented here were derived assuming that the data were independent observations from the same distribution, this is in fact much more restrictive than is necessary, and the results hold more generally. Similarly the results generalize immediately to vector parameters. In this case result 3 is:

**Result 3 (vector parameter)**

$$\boldsymbol{\mathcal{I}} \equiv E_0\begin{pmatrix} \left(\frac{\partial l}{\partial \theta_1}\right)^2 & \frac{\partial l}{\partial \theta_1}\frac{\partial l}{\partial \theta_2} & . \\ \frac{\partial l}{\partial \theta_2}\frac{\partial l}{\partial \theta_1} & \left(\frac{\partial l}{\partial \theta_2}\right)^2 & . \\ . & . & . \end{pmatrix} = -E_0\begin{pmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & . \\ \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_1^2} & . \\ . & . & . \end{pmatrix} \quad (1)$$

## 5.3 Consistency

Maximum likelihood estimators are often not unbiased, but under quite mild regularity conditions they are *consistent*. This means that as the sample size on which the estimate is based tends to infinity, the maximum likelihood estimator tends in probability to the true parameter value. Consistency therefore implies asymptotic¶ unbiasedness, but it actually implies slightly more than this — for example that the variance of the estimator is decreasing with sample size.

Formally if $\theta_0$ is the true value of parameter $\theta$ and $\hat{\theta}_n$ is its m.l.e. based on $n$ observations $x_1, x_2, \ldots, x_n$ then consistency means that

$$\Pr[|\hat{\theta}_n - \theta_0| < \epsilon] \to 1$$

as $n \to \infty$ for any positive $\epsilon$.

To see why m.l.e.s are consistent, consider an outline proof for the case of a single parameter, $\theta$, estimated from independent observations $x_1, x_2, \ldots, x_n$ on a random variable with p.m.f. or p.d.f. $f(x, \theta_0)$. The log-likelihood in this case will be

$$l(\theta) \propto \frac{1}{n}\sum_{i=1}^{n} \log(f(x_i, \theta))$$

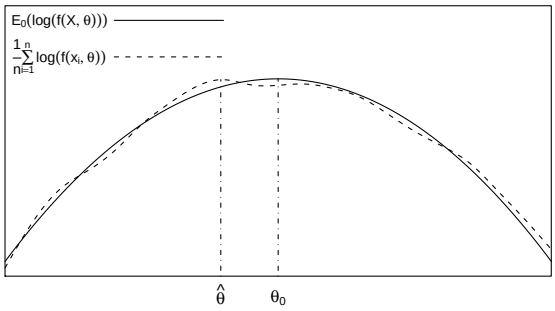
where the factor of $1/n$ is introduced purely for later convenience. We need to show that in the large sample limit $l(\theta)$ achieves its maximum at the true parameter value $\theta_0$, but in the previous section it was shown that the expected value of the log likelihood for a single observation attains its maximum at $\theta_0$. The law of large numbers tells us that as $n \to \infty$, $\sum_{i=1}^{n} \log[f(X_i, \theta)]/n$ tends (in probability) to $E_0[\log(f(X, \theta))]$. So in the large sample limit we have that

$$l(\theta_0) \geq l(\theta)$$

i.e. that $\hat{\theta}$ is $\theta_0$.

To show that $\hat{\theta} \to \theta_0$ in some well ordered manner as $n \to \infty$ requires that we assume some regularity (for example we need at least to be able to assume that if $\theta_1$ and $\theta_2$ are 'close' then so are $l(\theta_1)$ and $l(\theta_2)$), but in the vast majority of practical situations such conditions hold. A picture can help illustrate how the argument works:

¶'asymptotic' here meaning 'as sample size tends to infinity'.




... as the sample size tends to infinity the dashed curve, proportional to the log likelihood, tends in probability to the solid curve, $E_0[\log(f(X,\theta))]$, which has its maximum at $\theta_0$, hence $\hat{\theta} \to \theta_0$.

For simplicity of presentation, the above argument dealt only with a single parameter and data that were independent observations of a random variable from one distribution. In fact consistency holds in much more general circumstance: for vector parameters, and non-independent data that do not necessarily all come from the same distribution.

## 5.4 Large sample distribution of $\hat{\boldsymbol{\theta}}$

To obtain the large sample distribution of the m.l.e. $\hat{\theta}$ we make a Taylor expansion of the derivative of the log likelihood around the true parameter $\theta_0$ and evaluate this at $\hat{\theta}$.

$$\left.\frac{\partial l}{\partial \theta}\right|_{\hat{\theta}} \simeq \left.\frac{\partial l}{\partial \theta}\right|_{\theta_0} + \left(\hat{\theta} - \theta_0\right) \left.\frac{\partial^2 l}{\partial \theta^2}\right|_{\theta_0}$$

and from the definition of the m.l.e. the left hand side must be zero, so we have that

$$\left(\hat{\theta} - \theta_0\right) \simeq \frac{\partial l/\partial\theta|_{\theta_0}}{-\,\partial^2 l/\partial\theta^2|_{\theta_0}}$$

with equality, in the large sample limit (by consistency of $\hat{\theta}$). Now the top of this fraction has expected value zero and variance $\mathcal{I}$ (see section 5.2), but it is also made up of a sum of i.i.d. random variables, $\partial l_i/\partial\theta$, so that by the central limit theorem as $n \to \infty$ its distribution will tend to $N(0,\mathcal{I})$. By the law of large numbers we also have that as $n \to \infty$, $-\partial^2 l/\partial\theta^2\big|_{\theta_0} \to \mathcal{I}$ (in probability). So in the large sample limit $(\hat{\theta} - \theta_0)$ is distributed as an $N(0,\mathcal{I})$ r.v. divided by $\mathcal{I}$. i.e. in the limit as $n \to \infty$

$$\left(\hat{\theta} - \theta_0\right) \sim N(0, \mathcal{I}^{-1}).$$

The result generalizes to vector parameters:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1})$$

in the large sample limit. Again the result holds generally and not just for the somewhat restricted form of the likelihood which we have assumed here.

Usually, of course, $\boldsymbol{\mathcal{I}}$ will not be known any more than $\boldsymbol{\theta}$ is and will have to be estimated by plugging $\hat{\boldsymbol{\theta}}$ into the expression for $\boldsymbol{\mathcal{I}}$. In fact, often the *empirical information matrix*, which is just the negative of the hessian $(-\mathbf{H})$ of the log-likelihood evaluated at the m.l.e., is an adequate approximation to the information matrix $\boldsymbol{\mathcal{I}}$ itself (this follows from the law of large numbers).

## 5.5 Example

In an experiment designed to assess the mean lifetime of a batch of lightbulbs, 100 were selected at random and left on for 1000 hours or until they failed, whichever was the shorter. At the end of the experiment 90 bulbs were still alight and 10 had failed. The failure times (in hours) were:

$$11\ 51\ 59\ 118\ 168\ 257\ 263\ 396\ 485\ 933$$

A reasonable model for the failure times $T_i$ is that they are observations of exponential random variables with p.d.f.

$$f(t_i) = \lambda e^{-\lambda t_i}$$

The likelihood of $\lambda$ is the product of the joint p.d.f. of the 10 observed failure times and the probability of the 90 other bulbs lasting more than 1000 hours, i.e.

$$L(\lambda) = \prod_{i=1}^{10} \lambda e^{-\lambda t_i} \times \prod_{i=11}^{100} e^{-\lambda 1000}$$

leading to a log likelihood

$$l(\lambda) = 10 \log(\lambda) - \lambda \left( \sum_{i=1}^{10} t_i + 90000 \right).$$

Differentiating gives

$$\frac{\partial l}{\partial \lambda} = \frac{10}{\lambda} - \sum_{i=1}^{10} t_i - 90000, \qquad \frac{\partial^2 l}{\partial \lambda^2} = \frac{-10}{\lambda^2}.$$

Setting $\partial l/\partial \lambda = 0 \Rightarrow \hat{\lambda} = 10/(\sum t_i + 90000) = 10/(92741) = 1.078 \times 10^{-4}$. It follows by invariance that the m.l.e. of the mean failure time is $1/\hat{\lambda} = 9274.1$.

From the large sample distributional results it follows that the approximate standard deviation of $\hat{\lambda}$ is given by

$$\sigma_{\hat{\lambda}} \simeq \left( -E\left[ \frac{\partial^2 l}{\partial \lambda^2} \right] \right)^{-1/2} = \frac{\lambda}{\sqrt{10}}$$

which is estimated as $\hat{\lambda}/\sqrt{10} \simeq 0.341 \times 10^{-4}$. Indeed since the large sample results say that the distribution of $\hat{\lambda}$ is approximately normal, we can set approximate 95% confidence limits on $\lambda$ of

$$1.08 \pm 1.96 \times .34 \times 10^{-4} = (0.41 \times 10^{-4}, 1.75 \times 10^{-4}).$$

This corresponds to a 95% confidence interval for the mean lifetime of $(5726, 24178)$ (obtained by taking the reciprocal of the confidence limits on $\lambda$). What two changes could you make to the study in order to narrow this interval?

Note that while confidence intervals calculated in this way are usually acceptable, they do have some strange properties. For example if we had parameterized the likelihood in terms of the mean lifetime then, by invariance, the m.l.e. for the mean lifetime would have come out exactly as before, but the confidence interval would not. i.e. confidence intervals calculated in this manner are not invariant — instead, the answer you get depends on the parameterization you choose to use. This seems somewhat unsatisfactory, and we will see later how the situation can be improved.

## 5.6 What to look for in a good estimator: minimum variance unbiasedness

How should we judge how good an estimation procedure is? One important theoretical result which helps is the **Cramer-Rao lower bound** on the variance of an unbiased estimator. Let $\theta$ be a parameter and $\hat{\theta}$ an **unbiased estimator** of $\theta$, meaning that $E(\hat{\theta}) = \theta$. The Cramer-Rao result says that the variance

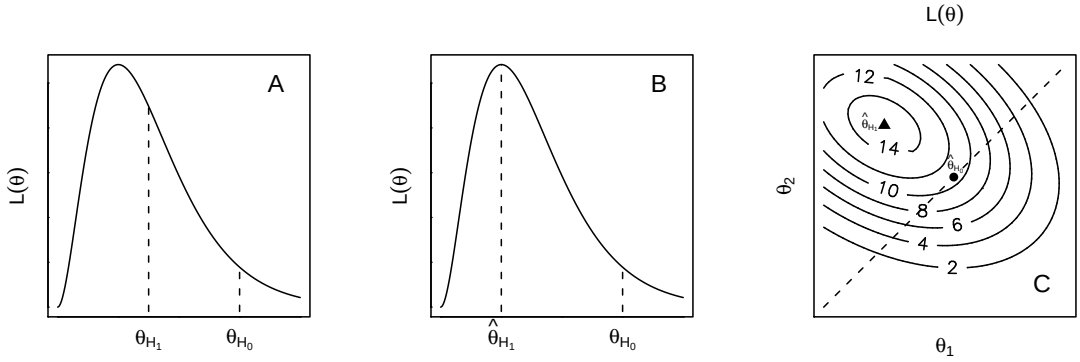
of $\hat{\theta}$ can not be smaller than $\mathcal{I}^{-1}$ - the inverse of the information about $\theta$ ($\boldsymbol{\mathcal{I}}^{-1}$ in the vector parameter case)[‖]. This result offers rather strong support for the method of maximum likelihood estimation, for as we have seen, in the large sample limit m.l.e.'s are unbiased and have exactly $\mathcal{I}^{-1}$ variance.

# 6 Hypothesis tests

Suppose that you have data $x_1, x_2, \ldots, x_n$ and a probability model with unknown parameters $\boldsymbol{\theta}$ describing how these data were generated. Further suppose that you have some preconceived notion of the values that the parameters might take, or more generally some restrictions on the values that they might take, and that you want to test this **null hypothesis** ($H_0$) against some **alternative hypothesis** ($H_1$) about the values that the parameters can take. In the situation described, a likelihood function can be written down for $\boldsymbol{\theta}$ and it seems natural to use this likelihood to compare the relative plausibility of the hypotheses.

Generally hypotheses fall into two main classes. *Simple hypotheses* completely specify the values taken by the parameters. For example, in a single parameter problem $H_0 : \theta = 2.5$ is an example of a simple hypothesis. *Composite hypotheses* are hypotheses that allow the parameter(s) to take a range of possible values. For example, in a one parameter problem $H_1 : \theta > 2.5$ is an example of a composite hypothesis — a range of $\theta$ values are consistent with it. A two parameter example of a composite hypothesis is $H_0 : \theta_1 = \theta_2$ — there is a range of values of $\theta_1$ and $\theta_2$ satisfying the hypothesis.

The figure shows examples of the 3 possible combinations of class of null and alternative hypothesis.




A illustrates a situation in which there is a single parameter, $\theta$, and we wish to test $H_0 : \theta = \theta_{H_0}$ against the alternative $H_1 : \theta = \theta_{H_1}$. In this case both hypotheses are simple and it is straightforward to compare their likelihoods. The ratio of the likelihoods $\Lambda = L(\theta_{H_1})/L(\theta_{H_0})$ would be a sensible statistic to use for the comparison. In fact this situation, while of considerable theoretical interest, is not often encountered in practice.

B illustrates the comparison of a simple null hypothesis $H_0 : \theta = \theta_{H_0}$ and a composite alternative $H_1 : \theta \neq \theta_{H_0}$. Here the likelihood for the null is well defined, but under the alternative $\theta$ can take almost any value so what should be used to measure the likelihood for the alternative? The obvious quantity to use is the maximum of the likelihood under the alternative $\hat{\theta}_{H_1}$ i.e. the likelihood at the m.l.e. Then the (generalized) likelihood ratio $\Lambda = L(\hat{\theta}_{H_1})/L(\theta_{H_0})$ is a sensible measure of the the relative plausibility of the hypotheses.

C illustrates the comparison of two composite hypotheses in a two parameter situation, $H_0 : \theta_1 = \theta_2$ vs. $H_1$ : no restriction. The contour plot is the likelihood, and under $H_0$ the parameters would have to lie on the dashed line. The obvious way to compare the hypotheses now, is to compare the maximum likelihood achievable under $H_0$ (i.e $L$ at •) with the maximum likelihood achievable under $H_1$ ($L$ at ▲). So in this case the generalized likelihood ratio statistic would be $\Lambda = L(\hat{\theta}_{H_1})/L(\hat{\theta}_{H_0})$

[‖]It is always possible to come up with a *biased* estimator with a lower variance — e.g. I can decide to adopt $37\frac{3}{4}$ yards exactly as my estimator of the distance to the moon: this has commendably low variance (zero), but rather high bias.

In each testing situation some variation on the generalized likelihood ratio statistic provides a reasonable measure of the relative plausibility of the hypotheses, with 'high' values supporting $H_1$ and 'low' values supporting $H_0$. At first sight it might seem appropriate to simply accept $H_1$ if $\Lambda > 1$ and $H_0$ otherwise, but there are two reasons why this is not done.

1. In many situations there is some prior reason to favour $H_0$. Often it is favoured because it is the simpler hypothesis, and we would like to have the simplest model consistent with the data*. In other circumstances there may be more pressing considerations: for example in a radioactive contamination monitoring programme it makes sense to accept the null hypothesis that contamination is at dangerous levels, unless the data provide strong evidence to the contrary.

2. Usually the null hypothesis states that the data were generated by a more restricted version of the model assumed under the alternative hypothesis. In this circumstance it is always the case that $L(\hat{\boldsymbol{\theta}}_{H_1}) \geq L(\hat{\boldsymbol{\theta}}_{H_0}) \Rightarrow \Lambda \geq 1$, and in fact $\Lambda > 1$ almost surely. Hence accepting $H_1$ if $\Lambda > 1$ would mean (almost) always accepting $H_1$, even if $H_0$ is true†.

For these reasons it is usual to accept $H_0$ unless $\Lambda$ is so large that, in effect, the evidence against $H_0$ has become too strong to ignore. The null hypothesis is presumed true until it is proven otherwise beyond reasonable doubt. The strength of the evidence against $H_0$ (and for $H_1$) is judged using a **p-value**.

> The p-value is the probability of the test statistic taking a value at least as favourable to $H_1$ as that actually observed, *if $H_0$ is true.*

Learn this and understand it. The p-value is an attempt to measure the consistency of the data with the null hypothesis.

1. Low p-values imply that the data are improbable if $H_0$ is true. Since the data actually happened, this suggests rejecting $H_0$ in favour of $H_1$.

2. High p-values suggest that the data are quite probable under $H_0$: since data and hypothesis appear consistent, $H_0$ is accepted.

How low should a p-value be in order to cause rejection of $H_0$? There is no 'correct' answer to this, and many statisticians prefer not to make one up, simply quoting the p-value instead. If a value, $\alpha$, is chosen, such that $H_0$ will be rejected if the p-value is $\leq \alpha$ then $\alpha$ is known as the **significance level** of the test. Traditional choices for $\alpha$ are 0.05, 0.01, 0.005 and 0.001, with 0.05 being the most common.

The usefulness of p-values as a measure of the evidence against $H_0$ stems from their general applicability: a p-value of 0.04 has the same meaning whatever hypothesis is being tested and whatever method is being used for the test and that meaning is well defined.

## 6.1 The generalized likelihood ratio test (GLRT)

Sometimes it is possible to find the exact distribution of $\Lambda$ (or more usually $\lambda = \log \Lambda$), and hence calculate a p-value exactly. But in most circumstances it is not possible and we have to use approximate p-values based on large sample results. Fortunately a very general result exists for the log of the generalized likelihood ratio test statistic.

Consider an observation $\mathbf{x}$ on a random vector of dimension $n$ with p.d.f. (or p.m.f.) $f(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector. Suppose that we want to test

$$H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0} \quad \text{vs.} \quad H_1 : \mathbf{R}(\boldsymbol{\theta}) \neq \mathbf{0}$$

where $\mathbf{R}$ is a vector valued function of $\boldsymbol{\theta}$ such that $H_0$ imposes $r$ restrictions on the parameter vector. If $H_0$ is true then in the limit as $n \to \infty$

$$2\lambda = 2(l(\hat{\boldsymbol{\theta}}_{H_1}) - l(\hat{\boldsymbol{\theta}}_{H_0})) \sim \chi^2_r \tag{2}$$

*This approach fits well some quite well accepted ideas in the philosophy of science.

†This argument applies for any test statistic, not just $\Lambda$.

where $l$ is the log-likelihood function, $\hat{\boldsymbol{\theta}}_{H_1}$ is the m.l.e. of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_{H_0}$ is the value of $\boldsymbol{\theta}$ satisfying $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$ which maximizes the likelihood (i.e. the restricted m.l.e.). This result is the basis for calculating approximate p-values.

### 6.1.1 Example: the bone marrow data

Recall the bone marrow transplant survival data from section 1.1 ...

| Treatment | Time (Days) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Allogenic | 28 | 32 | 49 | 84 | 357 | 933* | 1078* | 1183* | 1560* | 2114* | 2144* | |
| Autogenic | 42 | 53 | 57 | 63 | 81 | 140 | 176 | 210* | 252 | 476* | 524 | 1037* |

The data are times $t_i$ from treatment to relapse or death, except for the data marked $*$, which are observations that $t_i > t_i^*$. The medically interesting question is whether these data provide evidence that the average relapse rate differs between the groups (equivalently, that the mean relapse time differs between the groups). A possible model for these data is that

$$f(t_i) = \begin{cases} \theta_{al} e^{-\theta_{al} t_i} & t_i > 0 \text{ \& Allogenic group} \\ \theta_{au} e^{-\theta_{au} t_i} & t_i > 0 \text{ \& Autogenic group} \\ 0 & \text{otherwise} \end{cases}$$

where the parameters $\theta_{au}$ and $\theta_{al}$ are the 'relapse rates' for each group. To test the medical question statistically we can test

$$H_0 : \theta_{au} = \theta_{al} \quad \text{versus} \quad H_1 : \theta_{au} \neq \theta_{al}$$

using a generalized likelihood ratio test. To calculate the test statistic we must find the maximized likelihood under each of the two hypotheses. Following section 3.1 the log likelihood is

$$l(\theta_{au}, \theta_{al}) = n_{au,uc} \log(\theta_{au}) - \theta_{au} \sum_{\text{au group}} t_i + n_{al,uc} \log(\theta_{al}) - \theta_{al} \sum_{\text{al group}} t_i$$

and this must be maximized to find $\hat{\boldsymbol{\theta}}_{H_1}$ (the $uc$ subscript indicates 'uncensored'). Under the restriction imposed by $H_0$ we can replace all occurrences of $\theta_{au}$ and $\theta_{al}$ in the log likelihood by a single parameter $\theta \equiv \theta_{au} = \theta_{al}$, in which case we could simplify the log likelihood to

$$l(\theta) = n_{uc} \log(\theta) - \theta \sum_{\text{all cases}} t_i.$$

The likelihood under $H_0$ must also be maximized to find $\hat{\boldsymbol{\theta}}_{H_0}$.

As we saw in section 3.1 we could maximize these log- likelihoods by hand calculation, but following the principle that you shouldn't waste brain power on something you can more easily get a machine to do, we can also use `optim` in S. I'll enter the censored data as negative numbers, so that they can easily be identified, and then ignore the sign when calculating the log likelihood. I'll also get `optim` to maximize the likelihoods w.r.t. the logs of the $\theta$s — invariance says that this is legitimate and it will ensure that the $\theta$ estimates come out positive, as they should. So the data are entered as ...

```
allo<-c(28 , 32 , 49 , 84 , 357 , -933 , -1078 , -1183 , -1560 , -2114 , -2144)
auto<-c(42 , 53 , 57 , 63 , 81 , 140 , 176 , -210 , 252 , -476 , 524 , -1037)
```

and here is a function suitable for calculating the log likelihood under $H_1$ and $H_0$ (depending on whether it is supplied with a single parameter or a 2-vector).

```
bm.ll<-function(p,allo,auto)
# work out bone marrow log-lik. Code censored obs. as -ve
{ p<-exp(p)                              # optimize w.r.t log(theta) since theta +ve
  n.al.uc<-sum(allo>0)                                  # number uncensored for allo
  n.au.uc<-sum(auto>0)                                  # number uncensored for auto
  if (length(p)==1) theta.au<-theta.al<-p               # under H_0
```

```
  else {theta.au<-p[1];theta.al<-p[2]}                    # under H_1
  ll<-n.au.uc*log(theta.au) - theta.au*sum(abs(auto)) +  # the log-likelihood
      n.al.uc*log(theta.al) - theta.al*sum(abs(allo))
}
```

It is now straightforward to evaluate $l(\hat{\boldsymbol{\theta}}_{H_1})$, $l(\hat{\boldsymbol{\theta}}_{H_0})$, $\lambda$ and hence the p-value.

```
> m1<-optim(c(-7,-7),bm.ll,control=list(fnscale=-1),method="BFGS",allo=allo,auto=auto)
> m0<-optim(-7,bm.ll,control=list(fnscale=-1),method="BFGS",allo=allo,auto=auto)
> lambda<-m1$value-m0$value
> p.value<-1-pchisq(2*lambda,df=1);p.value
[1] 0.001699133
```

(`pchisq(x,df=r)` returns $\Pr[X \leq x]$ where $X \sim \chi^2_r$.) So there is quite strong evidence to reject the null here: the treatments really differ here.

What do the results tell us? Under $H_1$ the m.l.e.s of the parameters are $(0.0029, 0.0005)$ and under $H_0$: $(0.0011, 0.0011)$. So the data suggest that the experimental treatment has a higher relapse rate, and the hypothesis test indicates that this difference is real. It also looks clinically significant so the new treatment does not look like a good prospect when the other treatment is possible. (The conventional treatment requires that a matched donor can be found and that they undergo a painful operation themselves, to extract the donated marrow. This is not always possible.)

### 6.1.2 Simple example: Geiger counter calibration

A Geiger counter (radioactivity meter) is calibrated using a source of known radioactivity. The counts recorded by the counter, $x_i$, over 200 1 second intervals are recorded ...

```
 8 12  6 11  3  9  9  8  5  4  6 11  6 14  3  5 15 11  7  6  9  9 14 13
 6 11  . . . . . . . . . . . . . . . . . . . . . .  9  8  5  8  9 14 14
```

The sum of the counts $\sum_{i=1}^{200} x_i = 1800$. The counts can be treated as observations of i.i.d. Poi($\theta$) r.v.s. with p.m.f.

$$f(x_i, \theta) = \frac{\theta^{x_i} e^{-\theta}}{x_i!} \quad x_i \geq 0, \theta \geq 0.$$

If the Geiger counter is functioning correctly then $\theta = 10$, and to check this we would test

$$H_0 : \theta = 10 \quad \text{versus} \quad H_1 : \theta \neq 10.$$

Suppose that we choose to test at a significance level of 5% (i.e. $H_0$ will be rejected if the p-value is $\leq 0.05$). The test can be performed using a generalized likelihood ratio test. The log likelihood is

$$l(\theta) = \log(\theta) \sum_{i=1}^{n} x_i - n\theta$$

(where I have simply dropped the $\theta$-independent constant $\sum \log(x_i!)$).

In this case $H_0$ completely specifies the value of $\theta$, so the m.l.e. under $H_0$ is 10, and the maximized log likelihood under $H_0$ is simply

$$l(10) = \log(10) \sum_{i=1}^{n} x_i - n \times 10 = 2144.653$$

Find the maximum of the log likelihood under $H_1$.

Now calculate the log likelihood ratio test statistic $2\lambda$.

Given that a $\chi^2_1$ r.v. has a 5% chance of being greater than or equal to 3.84, would you accept or reject $H_0$? What does this imply about the Geiger counter?

Finally, given the form of the m.l.e., what was the point of recording the counts in 200 1 - second intervals rather than recording the count in 1 200 second interval?

## 6.2 Why, in the large sample limit, is $2\lambda \sim \chi^2_r$ under $\mathrm{H}_0$?

To simplify matters, first suppose that the parameterization is such that $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{\gamma} \end{pmatrix}$ where $\boldsymbol{\psi}$ is $r$ dimensional and the null hypothesis can be written $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$. In principle it is always possible to re-parameterize a model so that the null has this form[‡].

Now let the unrestricted m.l.e. be $\begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix}$ and $\begin{pmatrix} \boldsymbol{\psi}_0 \\ \hat{\boldsymbol{\gamma}}_0 \end{pmatrix}$ be the m.l.e. under the restrictions defining the null hypothesis. The key to making progress is to be able to express $\hat{\boldsymbol{\gamma}}_0$ in terms of $\hat{\boldsymbol{\psi}}$, $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\psi}_0$. This is possible in general in the large sample limit, provided that the null hypothesis is true, so that $\hat{\boldsymbol{\psi}}$ is close to $\boldsymbol{\psi}_0$. Taking a Taylor expansion of the log likelihood around the unrestricted m.l.e. $\hat{\boldsymbol{\theta}}$ yields

$$l(\boldsymbol{\theta}) \simeq l(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \mathbf{H} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) \tag{3}$$

where $H_{i,j} = -\partial^2 l / \partial\theta_i \partial\theta_j \big|_{\hat{\boldsymbol{\theta}}}$. Exponentiating this expression the likelihood can be written

$$L(\boldsymbol{\theta}) \simeq L(\hat{\boldsymbol{\theta}}) \exp\left[ -\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \mathbf{H} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) /2 \right].$$

i.e. the likelihood can be approximated by a function proportional to the p.d.f. of an $N(\hat{\boldsymbol{\theta}}, \mathbf{H}^{-1})$ random variable. Now it is a standard result from probability that if part of $\boldsymbol{\theta}$ is fixed in the normal p.d.f. then the remainder of $\boldsymbol{\theta}$ has a normal p.d.f., but with a different mean. Specifically if

$$\begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{\gamma} \end{pmatrix} \sim N\left( \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\psi\psi} & \boldsymbol{\Sigma}_{\psi\gamma} \\ \boldsymbol{\Sigma}_{\gamma\psi} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{pmatrix} \right)$$

[‡]Of course, to use the result no re-parameterization is necessary — it's only being done here for theoretical convenience when deriving the result.

then the conditional mean is

$$E(\boldsymbol{\gamma}|\boldsymbol{\psi}) = \hat{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\gamma\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}).$$

and hence the conditional likelihood will be approximately gaussian in form with this conditional mean as the gaussian mean. Since we know that the normal p.d.f. (Gaussian) attains its maximum at its mean it follows that the m.l.e. for $\boldsymbol{\gamma}$ given that $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ must be

$$\hat{\boldsymbol{\gamma}}_0 \simeq \hat{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\gamma\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1}(\boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}}). \tag{4}$$

If the null hypothesis is true then in the large sample limit $\hat{\boldsymbol{\psi}} \to \boldsymbol{\psi}_0$ (in probability) so that the approximate likelihood tends to the true likelihood and we can expect (4) to hold for the maximizers of the exact likelihood.

Expressing (4) in terms of a partitioning of $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$, is not as useful as having the results in terms of the equivalent partitioning of $\mathbf{H}$ itself. Writing $\boldsymbol{\Sigma}\mathbf{H} = \mathbf{I}$ in partitioned form,

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\psi\psi} & \boldsymbol{\Sigma}_{\psi\gamma} \\ \boldsymbol{\Sigma}_{\gamma\psi} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{pmatrix} \begin{pmatrix} \mathbf{H}_{\psi\psi} & \mathbf{H}_{\psi\gamma} \\ \mathbf{H}_{\gamma\psi} & \mathbf{H}_{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

and multiplying out, results in four matrix equations of which two are useful:

$$\boldsymbol{\Sigma}_{\psi\psi}\mathbf{H}_{\psi\psi} + \boldsymbol{\Sigma}_{\psi\gamma}\mathbf{H}_{\gamma\psi} = \mathbf{I} \tag{5}$$

$$\boldsymbol{\Sigma}_{\psi\psi}\mathbf{H}_{\psi\gamma} + \boldsymbol{\Sigma}_{\psi\gamma}\mathbf{H}_{\gamma\gamma} = \mathbf{0}. \tag{6}$$

Re-arranging (6) while noting that by symmetry $\mathbf{H}_{\psi\gamma}^T = \mathbf{H}_{\gamma\psi}$ and $\boldsymbol{\Sigma}_{\psi\gamma}^T = \boldsymbol{\Sigma}_{\gamma\psi}$[§], yields

$$-\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi} = \boldsymbol{\Sigma}_{\gamma\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1}$$

and hence

$$\hat{\boldsymbol{\gamma}}_0 = \hat{\boldsymbol{\gamma}} + \mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0). \tag{7}$$

For later use it's also worth eliminating $\boldsymbol{\Sigma}_{\psi\gamma}$ from (5) and (6), which results in

$$\boldsymbol{\Sigma}_{\psi\psi}^{-1} = \mathbf{H}_{\psi\psi} - \mathbf{H}_{\psi\gamma}\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}. \tag{8}$$

Now provided that the null hypothesis is true, so that $\hat{\boldsymbol{\psi}}$ is close to $\boldsymbol{\psi}_0$, we can re-use the expansion (3) and write the log-likelihood at the restricted m.l.e. as

$$l(\boldsymbol{\psi}_0, \hat{\boldsymbol{\gamma}}_0) \simeq l(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) - \frac{1}{2}\begin{pmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{pmatrix}^T \mathbf{H} \begin{pmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{pmatrix}.$$

Hence

$$2\lambda = 2(l(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\psi}_0, \hat{\boldsymbol{\gamma}}_0)) \simeq \begin{pmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{pmatrix}^T \mathbf{H} \begin{pmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{pmatrix}.$$

Substituting for $\hat{\boldsymbol{\gamma}}_0$ from (7) and writing out $\mathbf{H}$ in partitioned form gives

$$2\lambda \simeq \begin{pmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \end{pmatrix}^T \begin{pmatrix} \mathbf{H}_{\psi\psi} & \mathbf{H}_{\psi\gamma} \\ \mathbf{H}_{\gamma\psi} & \mathbf{H}_{\gamma\gamma} \end{pmatrix} \begin{pmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \end{pmatrix}$$

and a short routine slog results in

$$2\lambda \simeq (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \left[\mathbf{H}_{\psi\psi} - \mathbf{H}_{\psi\gamma}\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}\right](\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0).$$

But given (8), this means that

$$2\lambda \simeq (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \boldsymbol{\Sigma}_{\psi\psi}^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \tag{9}$$

[§]and of course remembering that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.

Now if $H_0$ is true then as $n \to \infty$ this expression will tend towards exactness as $\hat{\boldsymbol{\psi}} \to \boldsymbol{\psi}_0$. Furthermore, by the law of large numbers and (1), $\mathbf{H} \to \boldsymbol{\mathcal{I}}$ as $n \to \infty$ (recall that in this section $\mathbf{H}$ is the negative second derivative matrix), which means that $\boldsymbol{\Sigma}$ tends to $\boldsymbol{\mathcal{I}}^{-1}$, and hence $\boldsymbol{\Sigma}_{\psi\psi}$ tends to the covariance matrix of $\hat{\boldsymbol{\psi}}$ (see section 5.4). Hence by the asymptotic normality of the m.l.e. $\hat{\boldsymbol{\psi}}$

$$2\lambda \sim \chi^2_r$$

Having proved the asymptotic distribution of $\lambda$ you might at this point be wondering why it was worth bothering when we could simply have used the right hand side of (9) directly as the test statistic. This approach is indeed possible and is known as the Wald test, but it suffers from the disadvantage that at finite sample sizes the magnitude of the test statistic depends on how we choose to parameterize the model. The GLRT on the other hand is invariant to the parameterization we choose to use, irrespective of sample size. This invariance seems much more satisfactory — we don't generally want our statistical conclusions to depend on details of how we set the model up which could never be detected by observing data from the model.

## 6.3 What to look for in a good testing procedure: power etc.

In testing procedures in which a significance level is chosen and the null hypothesis is either accepted or rejected, two errors are possible.

1. A **Type I error** is rejection of the null hypothesis when it is true.
2. A **Type II error** is acceptance of the null hypothesis when it is false.[¶]

A good testing procedure should keep the probabilities of both types of error low, but there is a trade-off between keeping the type I error probability low and keeping the type II error probability low. Generally a test procedure which tries to avoid falsely rejecting $H_0$ will require a great deal of evidence in order to reject $H_0$ — but this is bound to lead to a quite high probability of not rejecting $H_0$ when it is false.

To move beyond generalities error probabilities must be calculated. By construction of p-values, the probability of a type one error is simply the significance level of a test. For example, when testing at the 5% level we reject $H_0$ whenever the test statistic is as supportive of $H_1$ as will only be seen in 5% of datasets generated under $H_0$: this means that we will reject $H_0$ for 5% of datasets for which it is true.

So the probability of a type I error is set by the chosen significance level, which suggests that given a choice of testing procedures we should select the one that gives the lowest probability of a type II error given our chosen significance level. Since $1 - \Pr[\text{Type II error}]$ is the probability of correctly rejecting the null when it is false, it is known as the **power** of a test. So at a given significance level, we would generally choose the most powerful test available: this principle is often referred to as the Neyman-Pearson approach to hypothesis testing.

If both hypotheses are simple then the power is usually quite easy to calculate, and it is then possible to prove that the most powerful test will always be based on the likelihood ratio test statistic. Although derived only for simple hypotheses, this result, known as the **Neyman-Pearson lemma**, provides part of the theoretical support for using the generalized likelihood ratio test statistic.

In most interesting cases $H_1$ is not simple and then the power depends on the true parameter values: in this circumstance it is usual to calculate the power function for the test procedure. The following simple example illustrates the principle.

### 6.3.1 Power function example

An often overlooked consequence of globalization is that many dragons can now only find work providing outdoor heating in cafe's and bars often at well below the minimum wage. Low pay, poor diet and long hours take their toll and whereas dragon breath at exit should have a temperature of 666 Celsius plus or minus one degree it is common to find sadly worn out specimens barely able to manage 661C. The

[¶]If like me you find it hard to remember such uselessly non-descriptive labels as 'Type I' and 'Type II' then you may find the phrase '12-ra-ra-ho-ho' helpful.

following data from the Dragon Action Foundation Trust is for a 7068 year old dragon found working in an Amsterdam coffee shop, where the flame temperature was measured on a considerable number of consecutive days.

664.0 662.8 663.2 664.0 663.6 662.2 663.2 662.4 662.8 663.5 662.2 662.2 663.5

A normal distribution $N(\mu, 1)$ is an appropriate model, where $\mu$ is an unknown parameter. Interest focuses on testing:

$$H_0 : \mu = 666 \quad \text{versus} \quad H_1 : \mu \neq 666$$

The p.d.f. is

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

and so the log likelihood (dropping constants) is

$$l(\mu) = -\sum_{i=1}^{n} (x_i - \mu)^2/2$$

which is readily shown to be maximized at $\hat{\mu} = \bar{x}$. Writing $\mu_0$ for 666, the log likelihood ratio statistic is therefore

$$2\lambda = \sum_{i=1}^{n} (x_i - \mu_0)^2 - (x_i - \bar{x})^2 = n(\bar{x} - \mu_0)^2 = (\sqrt{n}\bar{x} - \sqrt{n}\mu_0)^2$$

But from known properties of the normal, under $H_0$, $\sqrt{n}\bar{x} \sim N(\sqrt{n}\mu_0, 1) \Rightarrow \sqrt{n}\bar{x} - \sqrt{n}\mu_0 \sim N(0, 1)$ and so if $H_0$ is true

$$2\lambda \sim \chi^2_1$$

exactly, in this case. Testing at the 5% level, we would accept the null if

$$2\lambda = n(\bar{x} - 666)^2 < 3.84.$$

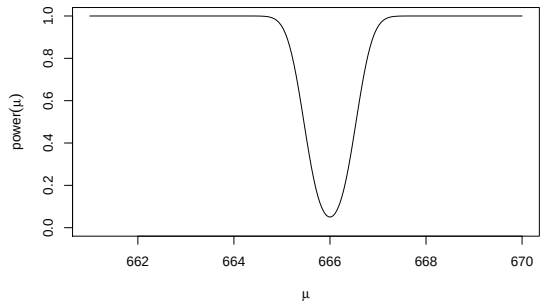
We now need to work out the probability of accepting the null for any value of $\mu$ other than 666, i.e. what is

$$\beta = \Pr[n(\bar{x} - 666)^2 < 3.84|\mu]$$

when $\mu$ can be any value? A little tedious manipulation provides the answer

$$\begin{aligned}
\beta(\mu) &= \Pr[|\sqrt{n}\bar{x} - \sqrt{n}666| < \sqrt{3.84}|\mu] \\
&= \Pr[\sqrt{n}\bar{x} - \sqrt{n}666 < \sqrt{3.84}|\mu] - \Pr[\sqrt{n}\bar{x} - \sqrt{n}666 < -\sqrt{3.84}|\mu] \\
&= \Pr[\sqrt{n}\bar{x} < \sqrt{n}666 + \sqrt{3.84}|\mu] - \Pr[\sqrt{n}\bar{x} < \sqrt{n}666 - \sqrt{3.84}|\mu] \\
&= \Phi(\sqrt{n}666 + \sqrt{3.84} - \sqrt{n}\mu) - \Phi(\sqrt{n}666 - \sqrt{3.84} - \sqrt{n}\mu)
\end{aligned}$$

where $\Phi$ is the c.d.f. of $N(0, 1)$. The power function is simply $1 - \beta(\mu)$ and is shown here:



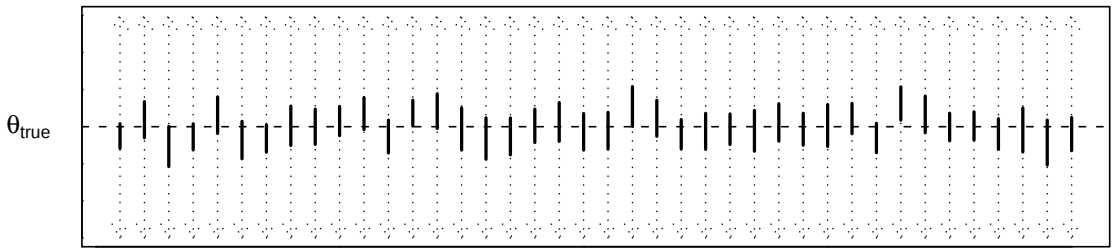

Notice the intuitively obvious property that the power function increases to 1 as $|\mu - \mu_0|$ increases: i.e. ity is always easier to reject grossly wrong hypotheses, as opposed to slightly wrong ones.

# 7 Interval Estimation

Having established how to find the parameter values that best fit the data, and how to test whether particular parameter values (or other restrictions on the parameters) are consistent with data, consider the question of what *range* of parameter values might be consistent with data. An obvious way of defining such a range is to find the set of numbers that would have been accepted as a null hypothesis about the parameter's value, at some specified significance level, $\alpha$. Such a range of values is known as a $100(1-\alpha)\%$ **confidence interval** for the parameter.

To put this mathematically: consider data $x_1, x_2, \ldots x_n$ and a probability model for these data that states that they are observations of random variables with joint p.d.f. $f(\mathbf{x}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters. Further suppose that we have some procedure for testing $H_0 : \theta_i = \theta_0$ and choose to accept $H_0$ for p-values greater than significance level $\alpha$ and to reject it otherwise. The range of values of $\theta_0$ which would cause $H_0$ to be accepted is a $100(1-\alpha)\%$ confidence interval for $\theta_i$.

A 'thought experiment' helps explain the key property of these intervals. Suppose that you have an infinite sequence of replicate sets of data, each generated from the same true model, for which $\theta = \theta_{\text{true}}$. Now suppose that for each replicate set of data you use a test procedure to test the null hypotheses that $\theta$ is each value on the real line, at the 5% level. For each replicate you'll end up with a set of accepted and a set of rejected $\theta$ values. The following figure is an attempt to illustrate this for the first 40 replicates. The vertical axis is the $\theta$ axis. For each replicate the thick line is the range of accepted $\theta$ values (the 95% confidence interval), while the dotted line shows the rejected values.




Now by construction, a hypothesis test operating with a 5% significance level will reject the true null hypothesis with probability 0.05 and accept it with probability 0.95. i.e in 5% of the replicates it will reject the correct value $\theta_{\text{true}}$ and in 95% of the replicates it will accept $\theta_{\text{true}}$. Hence 95% of the confidence intervals will include $\theta_{\text{true}}$ while 5% will not (which is exactly what has turned out to happen in this sample of 40 replicates).

This $(1-\alpha)$ probability of a $100(1-\alpha)\%$ confidence interval including the true parameter is the defining characteristic of a confidence interval. Always remember that it is the interval which is the random quantity here, not the parameter [‖]. This general process for constructing intervals based on testing procedures is known as **test inversion**.

More generally (for example when dealing with vector parameters) it may be more useful to consider a $100c\%$ **confidence set**: the set of parameter (vectors) that would be accepted by a hypothesis testing procedure conducted at the $1-c$ significance level. Often in single parameter contexts the confidence set is simply all points within the confidence interval, but in some complicated situations a confidence set may not be continuous, and for vector parameters, intervals for each component of the vector often provide only crude information about the confidence set.

[‖] although the Bayesian approach to statistics, which you'll meet in later inference courses effectively reverses this!

## 7.1 Intervals based on GLRT inversion

Consider finding a $100c\%$ confidence interval for a parameter $\theta_i$: the ith component of a parameter vector $\boldsymbol{\theta}$ with log-likelihood function $l(\boldsymbol{\theta})$. Conceptually we could test

$$H_0 : \theta_i = \theta_0 \quad \text{versus} \quad H_1 : \theta_i \neq \theta_0$$

for every value of $\theta_0$ using a GRLT test at the $1 - c$ significance level. The confidence set would consist of all the $\theta_0$ values accepted on this basis. Usually the confidence set would be a continuous set so that the smallest and largest elements of the set would yield a confidence interval for $\theta_i$. In practice we must find all values $\theta_0$ such that

$$2\lambda = 2[l(\hat{\boldsymbol{\theta}}_{H_1}) - l(\hat{\boldsymbol{\theta}}_{H_0})] \leq \chi^2_1(c) \tag{10}$$

where $\chi^2_1(\alpha)$ denotes the $\alpha$ quantile of the $\chi^2_1$ distribution (i.e. a $\chi^2_1$ r.v. will be less than $\chi^2_1(\alpha)$ with probability $\alpha$).

Recasting (10) we would search for the values of $\theta_0$ such that the value of the log-likelihood maximized under $H_0$ is at least the value of the log-likelihood maximized under $H_1$ minus $\chi^2_1(\alpha)/2$. i.e. we search for the $\theta_0$ values that imply

$$l(\hat{\boldsymbol{\theta}}_{H_0}) > l(\hat{\boldsymbol{\theta}}_{H_1}) - \chi^2_1(c)/2$$

This approach applies equally well to finding confidence sets for parameter vectors — all that will change is the degrees of freedom of the $\chi^2$ distribution, which is given by the number of parameters involved.

Confidence sets/intervals obtained by this method usually have to be found by numerical search. Such intervals are sometimes known as **Wilks** intervals/sets. Wilks intervals have nice invariance properties. Suppose $\theta$ and $\beta$ are parameters and $\beta = g(\theta)$ where $g$ is some strictly increasing or decreasing function. If $[\theta_a, \theta_b]$ is a Wilks interval for $\theta$ then the Wilks interval that would be obtained by re-parameterizing in terms of $\beta$ 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.

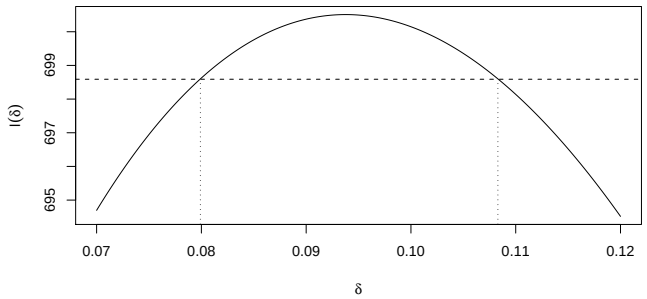
### 7.1.1 Simple single parameter example: bacteria model

Recall the bacteria culture example from section 4.1. In that example there was only a single unknown parameter, $\delta$, and in such cases the GLRT interval is very easy to calculate, since the maximized likelihood under $H_0$ is simply the likelihood evaluated at the hypothesized value for $\delta$ (i.e. there is no need to perform any maximization).

Writing a function to evaluate the log likelihood and maximizing w.r.t. $\delta$ yields $\hat{\delta} = 0.0937$ and $l(\hat{\delta}) = 700.51$. To obtain a 95% CI for $\delta$ we simply evaluate $\chi^2_1(0.95)$ e.g.

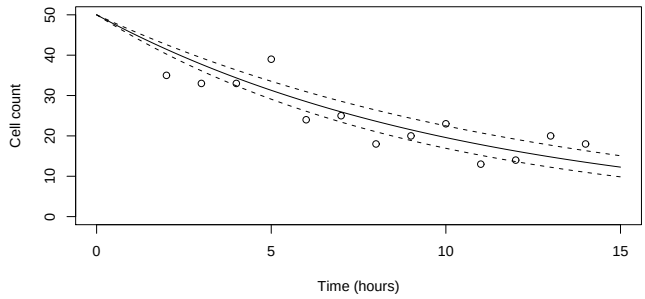
```
> qchisq(0.95,1)
[1] 3.841459
```

and then find the interval of $\delta$ values such that $l(\delta) \geq 700.51 - 3.84/2 = 698.59$. This is easily done by plotting the log likelihood against $\delta$ ...



where the continuous curve is the log-likelihood and values of $\delta$ for which the log-likelihood is above the horizontal line form the approximate 95% confidence set. The approximate 95% confidence interval lies between the two dashed vertical lines and is (0.0799,0.1083) — given the function that evaluates the likelihood it is very easy to search for these values numerically.

The m.l.e. model and the models implied by the limits of the confidence interval are overlaid on a plot of the data in the following figure.




### 7.1.2 Multi-parameter example: AIDS epidemic model

Now consider obtaining an approximate 95% confidence set for the parameters $\alpha$ and $\beta$ of the model for AIDS in Belgium from section 4.2. In this case we need to find all the values of $\alpha_0$ and $\beta_0$ that would lead to

$$H_0 : \alpha = \alpha_0, \beta = \beta_0$$

being accepted when tested against an alternative of no restriction on the parameters at the 0.05 significance level. That is we need to find all the $\alpha_0$ and $\beta_0$ values that imply that

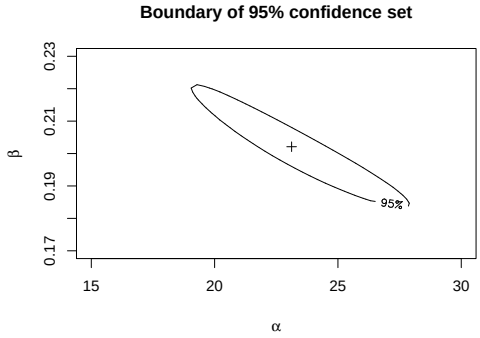
$$l(\hat{\alpha}, \hat{\beta}) - l(\alpha_0, \beta_0) < \chi^2_2(0.95)/2$$

where $\hat{\alpha}, \hat{\beta}$ are the m.l.e.s under $H_1$. Again it will be necessary to search for the parameter values numerically. A simple approach is to evaluate the log-likelihood over a grid of $\alpha_0$, $\beta_0$ values and then produce a contour plot of the log-likelihood. From the contour plot the region for which the likelihood is greater than $l(\hat{\alpha}, \hat{\beta}) - \chi^2_2(0.95)/2$ can be read off.

In fact in S things are even simpler, since you can provide the `contour` routine with the levels at which contours should be drawn and can hence request a single contour at $l(\hat{\alpha}, \hat{\alpha}) - \chi^2_2(0.95)/2$. The following illustrates this (assuming that the data are already read in and the log likelihood is given by a function `ll`.).

```
er<-optim(c(10,.1),ll,control=list(fnscale=-1),method="BFGS",y=y,t=t)
boundary<-er$value-qchisq(0.95,df=2)/2  # threshold for inclusion in confidence set
m<-50                          # m by m grid will be produced
a<-seq(15,30,length=m)         # alpha values at which to evaluate ll
b<-seq(0.17,0.23,length=m)     # beta values at which to evaluate ll
log.l<-matrix(0,m,m)           # matrix to hold evaluated log-likelihood
for (i in 1:m) for (j in 1:m) log.l[i,j]<-ll(c(a[i],b[j]),y,t) # evaluate log-lik
contour(a,b,log.l,levels=boundary) # produce contour plot with 1 contour at boundary
points(er$par[1],er$par[2],pch=3)  # add location of m.l.e.
```

The results are as follows, with the continuous curve illustrating the boundary of the confidence set for $\alpha, \beta$, with every point within the curve lying within the set. (While the given code will run in both S-PLUS and R, the plot was actually produced in R which has some extra labelling facilities not available in S-PLUS.)




**Exercise:** only one line of the code would need modification to produce a 99% C.I. Write out the modified line.

Now suppose that a marginal 95% confidence interval is required, for $\beta$, say. That is an interval is required which will include the true $\beta$ with probability 0.95 irrespective of the value of $\alpha$. Simply reading off the $\beta$ limits of the joint confidence set for $\alpha$ and $\beta$ will not generally give an interval with the correct coverage probability**.

Again the interval for $\beta$ is obtained by asking what values of $\beta_0$ would lead to acceptance of

$$H_0 : \beta = \beta_0$$

when tested against an alternative of no restrictions on either parameter? That is we seek the $\beta_0$ values such that

$$l(\hat{\alpha}_{H_0}, \beta_0) > l(\hat{\alpha}, \hat{\beta}) - \chi^2_1(0.95)/2.$$

Obviously the calculations are less straightforward this time, since the likelihood has to be maximized w.r.t. $\alpha$ to find $\hat{\alpha}_{H_0}$ in order to decide whether to accept or reject any particular $\beta_0$. To do this I re-wrote the log-likelihood function so that `optim` only optimizes w.r.t. $\alpha$ while $\beta$ is supplied as a fixed argument.

```
llb<-function(a,y,t,b)
# log likelihood for exponential growth epidemic model with fixed beta
{ sum(y*(log(a)+b*t) - a*exp(b*t))
}
```

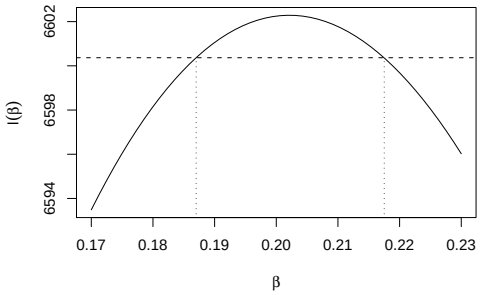
Then I evaluated $l(\hat{\alpha}, \hat{\beta}) - \chi^2_1(0.95)/2$

```
> er<-optim(c(10,.1),ll,control=list(fnscale=-1),method="BFGS",y=y,t=t)
> bound<-er$value-qchisq(0.95,df=1)/2
> bound
[1] 6600.367
```

Finally I looped through a set of trial $\beta$ values, maximizing the log likelihood w.r.t. $\alpha$ for each, and plotted the maximized log likelihood against $\beta$

```
m<-50;b<-seq(0.17,0.23,length=m)   # sequence of trial beta's
log.lb<-0                          # array for restricted likelihoods
for (i in 1:m)                     # loop through trial values
log.lb[i]<-optim(.2,llb,control=list(fnscale=-1),method="BFGS",y=y,t=t,b=b[i])$value
plot(b,log.lb,type="l")            # plot the maximized (restricted) log likelihood
abline(bound,0,lty=2)              # plot line at acceptance/rejection boundary
```

**The joint confidence set should include the true $\alpha$ *and* $\beta$ with probability 0.95 — to achieve this it will almost certainly have to include the true $\beta$ with a probability greater than 0.95.




Reading the graph very carefully or searching numerically gives an approximate 95% CI for $\beta$ of (0.1870, 0.2175), and these limits have been added to the above plot. Notice how this marginal interval is narrower than would have been obtained by (wrongly) taking the lower and upper $\beta$ values within the joint confidence set.

## 7.2 Intervals for functions of parameters

Sometimes interest focuses not on finding intervals for the parameters themselves, but on an interval for some function of the parameters. An example illustrates this. Consider again the allogenic/autogenic bone marrow transplant model from section 6.1.1, and suppose that a 90% CI is required for $\theta_{al} - \theta_{au}$ the difference in rates between the two treatments. This is easily approached by considering testing

$$H_0 : \theta_{al} - \theta_{au} = \delta_0$$

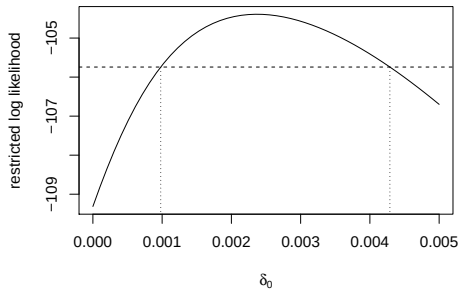
against the alternative of no restriction on the parameters, where $\delta_0$ is some hypothesized difference in rates. Working out the range of $\delta_0$ values that would be accepted at the 10% level gives a 90% CI for the difference in rates. So we seek the $\delta_0$ values which will imply that

$$l(\hat{\theta}_{al}, \hat{\theta}_{au}) - l(\hat{\theta}^*_{al}, \hat{\theta}^*_{al} + \delta_0) < \chi^2_1(0.9)/2$$

where $\hat{\theta}^*_{al}$ is the m.l.e. of $\theta_{al}$ under the restriction imposed by $H_0$. To actually work out this range I re-wrote the function for the bone marrow log-likelihood in such a way that `optim` would maximize only w.r.t. $\theta_{al}$ while taking $\theta_{au} = \theta_{al} + \delta_0$ where $\delta_0$ is supplied as fixed parameter.

```
bm.llr<-function(p,allo,auto,diff)
# work out bone marrow log-lik. Code censored obs. as -ve
{ p<-exp(p)             # optimize w.r.t log(theta) since theta +ve
  n.al.uc<-sum(allo>0)                                  # number uncensored for allo
  n.au.uc<-sum(auto>0)                                  # number uncensored for auto
  theta.al<-p;theta.au<-theta.al+diff                   # under H_0
  if (theta.au<1e-200) theta.au<-1e-200                 # cludge to avoid theta.au <=0
  ll<-n.au.uc*log(theta.au) - theta.au*sum(abs(auto)) +  # the log-likelihood
      n.al.uc*log(theta.al) - theta.al*sum(abs(allo))
}
```

I then used `optim` to maximize this log likelihood for each of a sequence of $\delta_0$ values and plotted the value of the maximized log likelihood against $\delta_0$.




Again careful reading from the plot, or a simple numerical search yields an approximate 90% CI of (0.00098,0.00429), illustrated by the vertical dotted lines on the plot. Note that using the straightforward approach given here requires some care, since large negative values of $\delta_0$ lead to problems with $\log(\theta_{au})$!

## 7.3 Intervals based on $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1})$

In section 5.4 it was shown that in the large sample limit the m.l.e.s are normally distributed around the true parameter values $\boldsymbol{\theta}_0$. i.e.

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1}).$$

We can use standard normal distribution theory to obtain alternative approximate confidence intervals based on this result. For clarity, concentrate on the single parameter case.

$$\hat{\theta} \sim N(\theta_0, \mathcal{I}^{-1}) \Rightarrow \frac{\hat{\theta} - \theta_0}{\sqrt{\mathcal{I}^{-1}}} \sim N(0,1)$$

Working at the $\alpha$ significance level, we would usually accept any *hypothesized* $\theta_0$ (against a two sided alternative) if

$$\left| \frac{\hat{\theta} - \theta_0}{\sqrt{\mathcal{I}^{-1}}} \right| \leq \Phi^{-1}(1 - \alpha/2)$$

where $\Phi^{-1}$ denotes the inverse c.d.f. of the $N(0,1)$ distribution. Hence we accept all $\theta_0$ in the interval

$$\hat{\theta} \pm \Phi^{-1}(1 - \alpha/2)\sqrt{\mathcal{I}^{-1}}.$$

Of course usually the empirical information matrix (the negative of the hessian of the log-likelihood) would be used in place of the information matrix itself, and depending on circumstances we might use quantiles of a $t$ distribution rather than the standard normal.

In the vector parameter case the intervals will be much the same except that the reciprocal of the information would be replaced by the appropriate element from the leading diagonal of the inverse of the information matrix.

These so called **Wald** 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.

### 7.3.1 Wald interval example: AIDS again

Consider estimating a 95% CI for the parameter $\beta$ for the AIDS model example using the Wald approach. To do this requires that we evaluate the hessian matrix (second derivative matrix) of the log likelihood: `optim` will actually estimate this for us if specify `hessian=TRUE` in its argument list. Here is the S code to evaluate the Wald interval.

```
> # maximize AIDS log-likelihood and hessian at the m.l.e.
> m1<-optim(c(10,.1),ll,control=list(fnscale=-1),method="BFGS",hessian=TRUE,y=y,t=t)
> V<-solve(-m1$hessian) # find the inverse of -H, the empirical information
> b.hat<-m1$par[2]      # m.l.e of beta
> sigma.b<-sqrt(V[2,2]) # s.d. of estimator of beta
> b.hat-1.96*sigma.b;b.hat+1.96*sigma.b # interval limits
[1] 0.1868727
[1] 0.2173361
```

The 95% CI (0.1869,0.2173) is very close to the previous Wilks estimated interval of (0.1870, 0.2175), largely because the log likelihood is almost symmetric w.r.t. $\beta$ over this sort of range.

## 7.4 Induced confidence intervals/ confidence sets

If $A$ is a $c$% confidence set for a parameter $\theta$ and $g$ is a strictly increasing or strictly decreasing function of $\theta$ then $B = \{g(\theta^*) : \theta^* \in A\}$ is a $c$% confidence set for $g(\theta)$. The proof is almost trivial. If $A$ is one of the $c$% of sets containing the true $\theta$ then $B$ contains the true $g(\theta)$, while if $A$ is one of the $(100 - c)$% of confidence sets not containing the true $\theta$ then $B$ will not include the true $g(\theta)$. This establishes that $B$ has the correct coverage probability. In the vector parameter case the same argument holds provided that every single point in the parameter space of $\boldsymbol{\theta}$ results in a unique single point in the space of $g(\boldsymbol{\theta})$.

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 $\theta$ would still result in a $B$ containing the true $g(\theta)$, but now it is also possible for some $A$ not containing the true $\theta$ 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 99% confidence interval for $\theta$ then $[\log(a), \log(b)]$ is a 95% CI for $\log(\theta)$, while $[1/b, 1/a]$ is a 99% CI for $1/\theta$.

# 8 Assumptions of the large sample results

Implicit in the derivations of the large sample properties of maximum likelihood estimation were some assumptions which it is now as well to state explicitly.

1. The parameter space is finite dimensional - i.e. $\boldsymbol{\theta}$ is finite dimensional **and** the true $\boldsymbol{\theta}$ lies in the interior of the space, not on a boundary.
2. The probability distributions defined by any two distinct $\boldsymbol{\theta}$ are distinct.
3. The first three derivatives of the log-likelihood w.r.t. $\boldsymbol{\theta}$ exist in the neighbourhood of the true parameter value (and in that neighbourhood the magnitude of the third derivative divided by the sample size is bounded above by a function of the data, whose expectation exists!)
4. Result 3 from section 5.2 holds.

The properties are generally required to justify the Taylor expansions used, for example. The assumptions are sufficient, but not always necessary, so the results may sometimes hold even if one or more of the assumptions are violated (but in such cases they need different justification).