Session 4

Statistical Estimation

Sampling distribution 4-2
Properties of Estimators 4-2
The method of Maximum Likelihood 4-3
Estimation with the Normal Distribution 4-5
Invariance Principle 4-6
Interval Estimation 4-7
Confidence Interval for a Normal Mean 4-7
Properties of the Normal Distribution 4-7
Sampling Distribution of the Mean 4-7
Sampling Distribution of the Variance 4-8
Practical 4 4-9
4. Statistical Estimation

If we have rejected the stated value of the null hypothesis in favour of a composite alternative then we have no single value for the parameter. It is natural, particularly in cases where the parameter has a physical meaning that we might wish to have a best estimate of the value. There have been many suggestions as to how to proceed to define “best” in statistical estimation. Since our objective is to establish values for the unknown parameters of a distribution this is often referred to as “fitting the model” to the data.

Sampling Distribution
In order to establish what might be a good estimator we must establish its repeated sampling properties. To do this we consider the estimators sampling distribution. The estimator is a function of random variables and so is itself a random variable and has a distribution. Using the repeated sampling principle we consider each value as the result of a complete rerun of the experiment or sample under identical conditions followed by estimation of the parameter. This results in a sequence of parameter estimates which will generate a probability distribution, the sampling distribution. The standard deviation of the sampling distribution has an important role so is given a special name the standard error.

Properties of Estimators
In the early development of the theory there was much emphasis on unbiasedness i.e. that the expectation of the parameter estimate was equal to the true value in the repeated sampling sense. This insistence is not always sensible as the parameterisation of the Likelihood may be arbitrary leading to different estimates for different parameterisations that may result in different fitted models. It is also possible to construct examples in which the only unbiased estimator is very inefficient (in the sense of not using all the information in the data). However, we would like to have a result similar to the WLLN, that as $n \to \infty$ the estimate converges in probability to the true value.
This property is called *consistency*. Another property of estimators is *efficiency* defined in terms of the variance of the sampling distribution with high efficiency associated with low variance. We have already seen the we can readily establish the properties of the sample mean as an estimator of the population mean, it is unbiased and has variance $\sigma^2/n$. This measure of efficiency is only sensible when we also insist on unbiasedness so becomes redundant if we relax this insistence. One measure that can be useful is *mean square error* (MSE) defined as follows. Define $\hat{\theta}$ to be an estimator for $\theta$ then:

$$MSE = E(\hat{\theta} - \theta)^2$$

Since our focus of attention has been on the Likelihood it is natural to base our estimation on the Likelihood. The previous results used the idea that data values with high Likelihood provide the strongest support for the model, so that a natural principle to consider is to estimate the parameter by the value that has greatest support from the data. This then leads us to the method of *maximum Likelihood*.

**The Method of Maximum Likelihood**

From the above we consider a parameter value has greater support when the Likelihood is greater. Therefore when looking for our best estimate of an unknown parameter value we might consider the value of the parameter that results in the largest possible value for the Likelihood. This is called the *Maximum Likelihood* estimate. For a general Likelihood of a parameter $\theta$, $L(y | \theta)$ then the maximum likelihood estimate $\hat{\theta}$ is defined as the (admissible) value of $\theta$ such that:

$$L(y | \hat{\theta}) > L(y | \theta)$$

Consider the ESP example with $N=25$ and $y=5$

This corresponds to a likelihood based on the Binomial distribution $B(N,p)$:

$$L(y | p) = ^{25}_C p^5 (1 - p)^{20}$$

The binomial coefficient can be omitted since it will not affect the position of the maximum (we can always ignore an additive or multiplicative constant).
As is very common, it is easier to deal with the logarithm of the likelihood:

\[
\log L(y \mid p) = k + 5\log p + 20\log(1 - p)
\]

(Note: in mathematics \(\log\) is by default the natural or Naperian logarithm or more formally the logarithm to base \(e\), this becomes important when using computer software in which \(\log\) may mean \(\log\) to base 10 by default.

Technically we should write \(\ln(.)\) but most textbooks still use \(\log(.)\).

When the maximum is at a point interior to the admissible range of values, we can solve the problem of finding the maximum likelihood estimate by setting the first derivative to zero and solving for \(p\).

\[
\frac{\partial \log L(y \mid p)}{\partial p} = \frac{\partial}{\partial p} \left( 5\log p + 20\log(1 - p) \right) = \frac{5}{p} - \frac{20}{1 - p}
\]

setting this to zero and solving gives

\[
5(1 - \hat{p}) = 20\hat{p}
\]

\[
\hat{p} = 5 / 20
\]

Considering this in a general setting. For a single observation, \(y\), from a binomial distribution \(B(N, \theta)\) then \(\hat{\theta} = y / N\) the sample proportion.

This conforms to the “commonsense” estimate, indeed we would be rather surprised if we did not get this answer. So, this may seem like hard work to get a trivial answer. The advantage is that it is universally applicable particularly in those examples where our commonsense deserts us.
Since we will be using the ML (maximum likelihood) to obtain the MLE (maximum likelihood estimate) and the corresponding maximum likelihood
value extensively, it is useful to build this into our consideration of the
likelihood function. It is convenient (and useful) to consider the relative
likelihood RL defined as

\[ RL(y \mid \theta) = \frac{L(y \mid \theta)}{L(y \mid \hat{\theta})} \]

This looks very similar to the likelihood ratio (LR) used previously. The
important difference is that whereas the LR was considered as a function of \( y \)
the RL is considered a function of \( \theta \). This can easily cause confusion when
first trying to understand the theory, but is unavoidable. When the meaning is
clear we can use \( RL(\theta) \) to emphasise that it is being considered as a function
of \( \theta \) but ultimately we will wish to consider the repeated sampling properties of
the methodology, so that considering these objects as functions of \( y \) becomes
imperative.

Estimation with the Normal Distribution

Most commonly (though not universally) the Normal (or Gaussian) distribution
is parameterised by its mean, \( \mu \), and variance, \( \sigma^2 \), defined in the usual way.
However, for now, we will use a scale parameter \( \varphi = \sigma^2 \). The p.d.f. (probability
density function) is then

\[ f(y) = \frac{1}{\sqrt{2\pi\varphi}} \exp\left\{ -\frac{(y - \mu)^2}{2\varphi} \right\}, -\infty < y < \infty \]

(Note that \( \exp(x) \) is another way of writing \( e^x \))

Consider the usual situation where we have a (random) sample of values
\( y_1, y_2, \ldots, y_n \).

To obtain ML estimates of \( \mu \) and \( \varphi \) we maximise the log-likelihood:

\[
L = f(y_1)f(y_2)\ldots f(y_n) \\
\log L = \log f(y_1) + \log f(y_2) + \cdots + \log f(y_n)
\]

noting that
\[
\log f(y) = k - \frac{1}{2} \log \phi - \frac{1}{2} \frac{(y - \mu)^2}{\phi}
\]

we get (ignoring the constant)

\[
\log L = -\frac{n}{2} \log \phi - \frac{1}{2} \sum_i (y_i - \mu)^2 \phi
\]

which we differentiate

\[
\frac{\partial \log L}{\partial \mu} = - \sum_i \frac{(y_i - \mu)}{\phi}
\]

\[
\frac{\partial \log L}{\partial \phi} = - \frac{n}{2\phi} + \frac{1}{2} \sum_i \frac{(y_i - \mu)^2}{\phi^2}
\]

giving MLE’s

\[
\hat{\mu} = \frac{1}{n} \sum_i y_i = \bar{y}
\]

\[
\hat{\phi} = \frac{1}{n} \sum_i (y_i - \bar{y})^2
\]

Note that the latter is not the usual estimate of variance that has divisor \((n-1)\). It is easy to show that this ML estimator of \(\sigma^2\) is biased whereas with divisor \((n-1)\) is is unbiased. Usually we will adopt this modification to the MLE.

If we had parameterised the distribution using \(\sigma\) rather than \(\varphi\) we would obtain effectively the same answer since

\[
\hat{\sigma} = \sqrt{\hat{\phi}}
\]

This is an illustration of the invariance principle.

**Invariance Principle**

Consider a transformation of the parameter \(\theta\) to \(\varphi\) by \(\varphi = g(\theta)\). Then the MLE’s are such that:

\[
\hat{\varphi} = g(\hat{\theta})
\]

Strictly, we have to restrict this to “sensible” transformations that have a proper inverse, but in practice these are the only transformations we would consider. For an informal proof, it is clear that we have not altered the values that the likelihood can attain, so that same maximum value pertains at a value
of $\phi$ that is the transformed value of the MLE of $\theta$. This is an important principle as the parameterisation is arbitrary in many cases.

**Interval Estimation**

*Hypothesis tests* take account of the uncertainty in the data but do not provide us with an estimate of the parameter is the null hypothesis is rejected, while a point estimation gives an estimate of the parameter but no measure of uncertainty in the data. A combined approach gives a range of value for the parameter that are well-supported by the data. This range is called a *confidence interval*.

**Confidence Interval for a Normal Mean**

This idea is easiest to develop with the Normal distribution. To simplify ideas we will consider the (unlikely) case that $\sigma^2$ is known. As we have seen the MLE is $\bar{y}$. In order to measure uncertainty we need the *repeated sampling* principle, that is the we need the *sampling distribution* of $\bar{Y}$.

**Properties of the Normal Distribution**

It is not difficult to show the following

1. If $Y \sim N(\mu, \sigma^2)$ then $a + bY \sim N(a + b\mu, b^2 \sigma^2)$
2. If $Y_1$ and $Y_2$ have (any) Normal distributions then $Y_1 + Y_2$ also has the Normal distribution.

**Sampling Distribution of the Mean**

From the above properties we can see that the sampling distribution of $\bar{Y}$ is the Normal distribution. We derived the expectation and variance of the sample mean previously for any distribution, so that

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$$

So the *standard error* is $\frac{\sigma}{\sqrt{n}}$.

Now we can make probability statements about $\bar{Y}$ such as:
\[ \Pr(\mu - 1.96 \frac{\sigma}{\sqrt{n}} < \bar{Y} < \mu + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95 \]

We can rearrange the inequality to give the equivalent statement

\[ \Pr(\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95 \]

This we can interpret as a statement about \( \mu \) as a confidence interval for \( \mu \). We say we are 95% confident that \( \mu \) lies in the interval:

\[ \left( \bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \]

**Sampling Distribution of the Variance**

Consider \( z_1, z_2, \ldots, z_N \) to be independent random variables from the standard Normal distribution \( N(0,1) \). If we combine these to form:

\[ \chi^2 = z_1^2 + z_2^2 + \cdots + z_N^2 \]

then the distribution of this can be found. The distribution is called the chi-squared (\( \chi^2 \)) distribution. The distribution depends on \( N \) called the "degrees of freedom". See the Dists Excel spreadsheet for a graphical display of this distribution.

Now consider the sample variance. At its heart is the sum:

\[ \sum_{i=1}^{N} (y_i - \bar{y})^2 \]

To standardize we need to divide the \( y_i \) by \( \sigma \) giving

\[ \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{\sigma^2} = \sum_{i=1}^{N} (z_i - \bar{z})^2 \]

This is similar to the definition of \( \chi^2 \) but the subtraction of \( \bar{z} \) from each \( z_i \) makes the degrees of freedom (\( N-1 \)). Therefore

\[ \frac{(N - 1)s^2}{\sigma^2} \sim \chi^2_{N-1} \]

This result would allow us to derive confidence limits and tests of the sample variance. Functions to compute probabilities for this distribution are available in Excel as the CHIDIST and CHIINV functions.
Practical 4

1. Try out the Sampling Distribution demonstration on the RVLS web site.
   (a) starting with the Normal distribution verify that the sampling distribution on the mean is also Normal even in very small samples.
   (b) using other distributions, verify that as the sample size increases the sampling distribution of the mean becomes closer to Normality.
   (c) Which distribution is worst in terms of the sampling distribution of the mean deviating from Normality.
   (d) Verify the formula for the s.d. of sampling distribution of the mean, and that this works for any distribution. Note choose a convenient sample size e.g. 16.
   (e) Examine the sampling distributions of other statistics. For the case of the Normal distribution verify that the sample median is an unbiased estimator of the mean but not as efficient as the sample mean.

2. Try out the Confidence Interval demonstration. Use different sample sizes to see the effect on the CI. By taking many large samples verify that the method gives the correct long run properties.

3. Examine the likelihoods in exercise 2 in practical 3.
   Find the maximum likelihood estimates.
   Do you see any problems with these MLE?

4. Find the MLE for exercise 3 in practical 3. Compare the result with the MLE for the Binomial distribution.

5. Examine the shape of the relative likelihood function, using the sheet RL, with varying N and Y.