1 Inference, probability and estimators

The rest of the module is concerned with *statistical inference* and, in particular the classical approach. We will cover the following topics over the next few weeks.

- What is statistical inference, what is the classical approach and how does it differ from other approaches?
- Estimators and their properties.
- The method of moments for estimation.
- Likelihood and maximum likelihood estimation.

Therefore we begin with a summary of the essential elements of statistical inference and a brief description of the classical approach.

1.1 Statistical Inference

If you look in the literature you can find many different definitions for statistics and the process of statistical inference. For example, Barnett (*Comparative Statistical Inference*, John Wiley & Sons, Chichester, 1982) suggests defining *statistics* as:

> *... the study of how information should be employed to reflect on, and give guidance for action in, a practical situation involving uncertainty*

We first consider examples of the kind of situations where we expect statistical inference to be important.

Example 1. Suppose you are a farmer who has to choose between two fertilizers (A and B). before placing a large order for either you obtain samples of each and apply A to 5 'test' plots and B to 5 plots of a certain crop. Then at the end of the season you measure the yield of the crop in *kgs* in each plot. The results are:

Fertiliser A: 12.5, 14.4, 15.1, 11.9, 13.1

Fertiliser B: 12.6, 13.2, 10.4, 10.9, 12.8

Now for fertilisers A and B the average yields are 13.4 and 11.98, respectively. Does this tell us that A is better than B and that we should place a large order for A? Well, it might suggest this but the data also tell us that 'yield' is something that varies from plot to plot even if the same fertilizer is used. If we repeated the experiment we would surely obtain different measurements and, plausibly, could come to a different conclusion regarding which was better. Thus it is possible that B may be better. If we decide to choose A then we may be making the wrong decision.

Statistical inference is the process whereby we will quantify how strong the evidence is that A is really better than B. Having done this we can then decide whether to order A, or whether, perhaps, we should do further experiments before making a decision.

It's not difficult to think of all kinds of scenarios that are analogous to example 1: testing effectiveness of different drugs, comparing lifetimes of different brands of a component. Every day we have to make potentially important decisions in the face of limited evidence and statistics gives us a set of tools for trying to do this.

See lectures for the 'Monty Hall' game.

1.2 The concept of probability

A key element of statistical inference is *probability theory*. Probability theory is used to construct the models from which observations are assumed to arise (see later) but also to express the degree of certainty in the conclusions. In example 1 can we associate a probability with the conclusion "A is better than B"?

You should recall from Statistics II the notion of confidence interval (e.g a 95% confidence interval). Suppose, in the context of example 1, we knew that the measurements for fertilizer B follow a $N(\mu_B, 1)$ distribution, where μ is unknown. Then we could have calculated a 95% CI as

$$(11.98 - 1.96\tfrac{1}{\sqrt{5}}, 11.98 - 1.96\tfrac{1}{\sqrt{5}})$$

i.e. (11.1, 12.8). This is an example of a probability being used to express the degree of certainty one might have in a conclusion regarding the range of values which μ_B could plausibly take.

Exercise. Revise Stats II notes on how to calculate a 95% CI for the difference $\mu_A - \mu_B$ assuming that the observations for A and B are distributed as

$N(\mu_A, \sigma^2)$ and $N(\mu_B, \sigma^2)$ when $\sigma^2 = 1$ and when σ^2 is unknown. (Later in Statistics VI we cover this material again in detail.) How can you use the CI to help you decide whether A is really better than B?

The "95%" in the above construction is indeed a probability. But of what event? In the next section we examine the interpretation of probabilities and, in particular, its interpretation in the *classical* or $frequentist$ philosophy of inference.

1.3 The intepretations of probability

To a pure *frequentist*, the only meaningful definition of probability of an event is the frequency with which it occurs over a *long sequence of independent trials.* Thus statements such as:

- the probability that a '6' is scored when a fair die is rolled is $\frac{1}{6}$
- the probability that a 'H' is uppermost when a fair coin is tossed is $\frac{1}{2}$

would be perfectly meaningful to a frequentist because there is a clearly defined, repeatable experiment in each case. The probability has an interpretation as a frequency.

A frequentist would not recognize, for example:

- the 'probability' that I voted Labour in the last election
- the 'probability' that Lord Lucan is still alive
- the 'probability' that Celtic will win the Scottish Premier League this year

as true probabilities. These deal with propositions that must either be true or false. There is no sequence of *identical* repeatable experiments for which these probabilities are frequencies. For the last example, there is only 1 2004-2005 football season. Even though the SPL is played every year, the teams and their personnel change from year to year.

Most people would agree that these probabilities - which express the degree of belief in a proposition regarding the world and are known as *subjective probabilities* - *are* meaningful. For example, a bookmaker, when setting the odds of Celtic winning the SPL this year, must consider his subjective probability that they will win. Nevertheless, for this course we will use only the

frequentist interpretation of probability. The role of other interpretations of probability in inference will be considered in later courses (Statistical Inference in 3rd Year).

Let's go back to example 1, where we calculated a 95% CI for μ_B to be (11.1, 12.8). The 95% must be interpretable as the frequency of occurrence of some event over a sequence of independent trials. Which event? Which experiment?

Here the repeatable *experiment* or trial is:

> Sow 5 plots with the crop, fertilise them with fertiliser B and measure the yield of each plot. (We assume that on any experiment the observations $Y_1, Y_2, ..., Y_5$ are drawn from a $N(\mu_B, 1)$ where μ_B is unknown but is identical for all trials).

What is the event whose frequency in a sequence of identical experiments is 0.95? Is it the event that $\mu_B \in (11.1, 12.8)$. No, it can't be. Since μ_B never changes then this event either has frequency 0 or 1. The event that occurs with frequency 0.95 is:

> The observed sample mean lies within $\pm 1.96\frac{1}{\sqrt{5}}$ of the unknown value μ_B.

Put another way, the frequency with which the CI we calculate will contain μ_B is 0.95. The classical argument then goes that since the there is no reason to think that our particular experiment is 'atypical' then we can be 95% confident that CI (11.1, 12.8) contains the value μ_B. The difficult issue of interpreting frequentist probabilities in inference will considered next year!

1.4 The main elements of frequentist (classical) inference

Having motivated things with a specific example, we now introduce the central concepts of classical inference and the main terminology that we will use throughout the rest of the module.

1.4.1 The experiment

At the heart of classical inference is the concept of an *experiment*. This is a procedure that can be repeated independently many times where the

outcome of the experiment is subject to uncertainty. We can usually think of experiments as measuring some real-valued quantity, and therefore the outcome of the experiment can be considered to be a random variable, X. Typical examples of an experiment are:

- Select an individual randomly (with replacement) from a large population and determine whether they carry a certain gene.
- Select a component randomly from a large population and measure the lifetime.
- Observe the night sky and measure the time elapsed before the first meteor appears.

1.4.2 The statistical model

The statistical model is simply the distribution that we assume for the observation X. Usually we shall specify the distribution using its probability density function (p.d.f.) if the experiment yields a continuous measurement such as a mass, a time, a height etc. or probability mass function (p.m.f.) if the experiment measures a discrete quantity such as a count. In either case we shall denote the p.d.f. or p.m.f. by $f_X(x)$.

In proposing a sensible statistical model we need to use our physical intuition and judgement. For the first of the three experiments described above we should suggest that $X \sim Bernoulli(p)$ where p is the proportion of gene carriers in the population. For the second experiment we might reasonably suggest that $X \sim \Gamma(\alpha, \beta)$. The same choice could be made for the 3rd experiment, although a simpler model, $X \sim Exp(\beta)$ might be suggested.

The choice of model is an art in statistics - and data should always be examined to see if it conforms to the selected model.

See lecture for typical distributions used to model experimental outcomes.

The statistical model is also referred to as the *population distribution* of the measurement X. This is because we think of the outcome of the experiment as a random draw from a very large population over which the the probability function of X is given by $f_X(x)$.

Normally when we carry out an experiment we take more than just a single measurement. We would repeat the 'basic' experiment several times and generate a sequence of measurements $X_1, X_2, ..., X_n$. In the above examples

we might select n components randomly or measure the intra-arrival times for n meteors.

When we take multiple measurements we can often (though not always - see later) assume that the observations are independent random variables all with distribution f_X. We can therefore write down the multivariate p.d.f. or p.m.f. of $\underline{X} = (X_1, X_2, ..., X_n)$ as

$$f_{\underline{X}}(x_1, x_2, ..., x_n) = f_X(x_1) \times f_X(x_2) \times ... \times f_X(x_n)$$

When we have independence of the measurements we refer to them as a *random sample of size n*. This is equivalent to $X_1, X_2, ..., X_n$ being i.i.d. with probability function f_X. In many, but not all, situations we encounter in this course the experiment will consist of a random sample of size n from some distribution.

1.4.3 Model parameters

Usually, although we might be sure of the family from which the distribution of X comes, we don't know what the distribution's parameters are (p, α and β above). The main point of doing the experiment is (usually) to find out more about the parameters or some other fixed characteristic of the population distribution such as its *mean* or *variance*. We call this process *estimation*. In example 1 we wanted to estimate $\mu_A - \mu_B$ to help us choose between the 2 fertilizers.

Note that the moments of a distribution will be functions of its parameters, so there is a close between estimating parameters and estimating population moments.

Exercise: make sure you know the mean and variance of the Gamma, Exponential, Normal, Binomial, and Poisson distributions in terms of their parameters.

We will exploit the relationship between parameters and moments when we look at *method-of-moments* estimation later in the course.

1.4.4 Estimation

Now let's consider the general situation where we carry out an experiment that yields a random sample $X_1, ..., X_n$ from some population distribution and we wish to use the data to estimate a quantity θ which could be a population moment (*e.g.* mean or variance) or a parameter. Since our estimate

depends on the data, it must be some function of the random sample and we can write it as

$$\hat{\theta} = g(X_1, ..., X_n)$$

We refer to $\hat{\theta}$ as an *estimator* of θ. The key thing to note is that, being a function of $X_1, ..., X_n$ (which are all random variables) $\hat{\theta}$ is itself a random variable (and will therefore vary from experiment to experiment).

Sometimes there will be a very natural choice for the function g. Suppose we wish to estimate the population mean (usually called μ) from a random sample $X_1, ..., X_n$. Then a natural choice is the *sample mean*

$$\hat{\mu} = g(X_1, ..., X_n) = \bar{X} = \frac{\sum X_i}{n}.$$

Suppose we were asked to estimate the population variance σ^2 from the random sample. Then a natural estimator to use here would be the *sample variance*:

$$\hat{\sigma}^2 = g(X_1, ..., X_n) = S^2 = \frac{1}{n-1}\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\}$$

What are the properties of $\bar{X}$ and S^2 that make them sensible choices as estimators for μ and σ^2 respectively? We examine this in the following section.

2 Properties of Estimators

We retain the notation of the last chapter. We conduct an experiment which yields a random sample $X_1, ..., X_n$ from a population. We wish to estimate some population quantity θ using an estimator $\hat{\theta} = g(\underline{X})$. How can we decide whether this choice of estimator is sensible? In the next two sections we consider two important properties that we can use.

2.1 Bias of an estimator

Remember that an estimator $\hat{\theta} = g(\underline{X})$ is a random variable. We refer to its distribution as its *sampling distribution*. The sampling distribution of $g(\underline{X})$ tells us how g will vary over a large number of independent experiments, and its properties will determine whether $g(\underline{X})$ is a sensible estimator to use. Thus it's going to be important to be able to derive properties of the distribution g from the population distribution F_X.

Now if $g(\underline{X})$ is a sensible estimator for θ then the values it typically takes should be close to θ in some sense. If we conduct a number of identical experiments and generate an estimate $g(\underline{X})$ for each then the values should be scattered around the true value of θ if the estimator is going to be any use.

One measure used to characterise a 'typical' or central value from a distribution is the expectation. If $E(g(\underline{X}))$ were close to θ this might be seen as a favourable property. This leads naturally the concept of *bias*. We define the bias of an estimator (for θ) $g(\underline{X})$ to be:

$$Bias(g(\underline{X})) = E(g(\underline{X}) - \theta).$$

We see that bias measures the difference between the expectation of g and the thing it is meant to be estimating, θ. Bias can be negative or positive. If $Bias(g) = 0$ then this means that the expectation of g is precisely θ and we say that g is *unbiased*.

Example 2.1 Let us return to the estimation of the population mean μ from random sample $X_1, ..., X_n$. What is the bias of the estimator $g(\underline{X}) = \bar{X}$?

Solution:

Note that this is true so long as the population mean μ exists - it doesn't matter what the precise form of the distribution F_X is. A harder problem is to show that the sample variance, S^2 is an unbiased estimator of the population variance σ^2. This is a question on the tutorial sheets.

Sometimes we know more about the distribution of the sample mean than merely its expectation. If the values in our random sample have a normal distribution, $X_i \sim N(\mu, \sigma^2)$ then the sample mean $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Suppose now our observations $X_1, ..., X_n$ were a random sample from a $\Gamma(\alpha, \beta)$ distribution. (This could arise if our observations were lifetimes of components or survival times of individuals.) Can you identify the precise form of the distribution of $\bar{X}$ in this case? (Hint: It's another Gamma distribution, but what are the parameters.)

Bias is not the whole story. Just because an estimator is unbiased it doesn't have to be useful. Consider the following picture that shows the values of 10 independent estimates of θ, for each of 3 different estimators $g_1(\underline{X})$, $g_2(\underline{X})$, $g_3(\underline{X})$.

Rank the estimators in order of usefulness.

It follows that an estimator that is unbiased could be of little value while one that is biased could actually be very useful. We need to consider other characteristics of estimators that are a better guide to how useful they are in practice.

2.2 Mean square error and efficiency

Whenever we use calculate an estimate $g(\underline{X})$ of the θ there will be some random error (equal to $g(\underline{X}) - \theta$). For a good estimator the magnitude of this random error should be small on average. We could quantify how accurate $g(\underline{X})$ is by looking at $E(|g(\underline{X}) - \theta)|)$ but the modulus function is not particularly nice to work with when trying to calculate expectations. A better measure to work with is the *mean square error (MSE)* of $g(\underline{X})$. This is defined to be

$$MSE(g(\underline{X})) = E((g(\underline{X}) - \theta)^2)$$

Note that if $g(\underline{X})$ is unbiased then $E(g(\underline{X})) = \theta$ so that $MSE(g(\underline{X})) = Var(g(\underline{X}))$. There is a particular instance of a general rule that states that

$$MSE(g(\underline{X})) = Bias^2(g(\underline{X})) + Var((g(\underline{X})).$$

You are asked to prove this in the tutorial sheets.

Example. We have seen that sample mean $\bar{X}$ is an unbiased estimator of the population mean μ. What is its MSE? It follows from the equation above that $MSE(\bar{X}) = Var(\bar{X}) = \frac{\sigma^2}{n}$. (You should be able to prove this!!)

Can we work out the mean-square error of S^2 as an estimator for σ^2. Yes, but we require to know something about higher order population moments (see Tutorials.)

We can use the MSE as a measure to compare different estimators. If $g_1(\underline{X})$ and $g_2(\underline{X})$ are estimators of θ then we say that g_1 is more efficient that g_2 if $MSE(g_1) \leq MSE(g_2)$ for all θ. To illustrate the concept of efficiency consider a random sample (X_1, X_2) from a population with mean μ and variance σ^2. Now consider the following two estimators $g_1(X_1, X_2) = \bar{X}$ and

$g_2(X_1, X_2) = \frac{X_1}{3} + \frac{2X_2}{3}$ as estimators of μ. Clearly both of these are unbiased ($E(g_1) = E(g_2) = \mu$.) It follows that $MSE(g_1) = Var(g_1) = \frac{\sigma^2}{2}$. What about $MSE(g_2)$?

$$MSE(g_2) = Var(g_2) = \frac{\sigma^2}{9} + \frac{4\sigma^2}{6} = \frac{5\sigma^2}{9} > MSE(g_1).$$

From this we see that g_1 is more efficient that g_2. This is not always the case.

2.2.1 Example

In an experiment a bacteriologist must measure the temperature, θ of a growth chamber. He has two thermometers of differing quality. The first returns a random quantity $T_1 = \theta + E_1$ where E_1 is a random error with mean zero and variance σ_1^2. The second returns $T_2 = \theta + E_2$ where E_2 has zero mean and variance σ_2^2. Suppose further that $E1$ and $E2$ are independent. He decides to combine the two measurements via the estimator:

$$\hat{\theta} = aT_1 + (1 - a)T_2$$

where $a \geq 0$. How should a be chosen to give the most efficient estimator of this form?

Solution. We need to calculate the MSE of $\hat{\theta}$ as a function of a. Now, for all a, $E(\hat{\theta}) = \theta$ (check this!!). Therefore (using $MSE = Bias^2 + Var$) we have that

$$\begin{aligned} MSE(\hat{\theta}) = Var(\hat{\theta}) &= a^2\sigma_1^2 + (1 - a)^2\sigma_2^2 \\ &= a^2(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 a + \sigma_2^2. \end{aligned}$$

This is a quadratic in a and is minimised when

$$a = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Therefore, as σ_2^2 increases relative to σ_1^2 the weight assigned to T_1 in our estimate increases.

Suppose now that the bacteriologist considers a more general estimator of the form $\hat{\theta} = aT_1 + bT_2$, where $a + b$ doesn't have to be equal to unity. This means that $\hat{\theta}$ might be biased. Find:

- the bias of $\hat{\theta}$ as a function of a and b;

- the MSE of $\hat{\theta}$ as a function of a and b;
- the values of a and b which minimise the MSE.

Solution. It is straightforward to show that

$$Bias(\hat{\theta}(a,b)) = \theta(a+b-1)$$

and that

$$Var(\hat{\theta}(a,b)) = a^2\sigma_1^2 + b^2\sigma_2^2$$

Using $MSE = Bias^2 + Var$ we can show that

$$MSE(\hat{\theta}(a,b)) = a^2\sigma_1^2 + b^2\sigma_2^2 + \theta^2(a+b-1)^2.$$

This is a quadratic function of a and b. We can use standard techniques from multivariate calculus to show that it has a mimimum value at

$$a^* = \frac{\theta^2\sigma_2^2}{(\sigma_1^2\sigma_2^2 + \theta^2(\sigma_1^2+\sigma_2^2))}$$

$$b^* = \frac{\theta^2\sigma_1^2}{\sigma_1^2\sigma_2^2 + \theta^2(\sigma_1^2+\sigma_2^2)}$$

Note that this solution gives us a biased estimator since $a^* + b^* < 1$ for any value of θ. This choice of a and b will give a smaller MSE than the optimal unbiased estimator. There are two things to note here:

- the values a^* and b^* depend on θ which is unknown (that's why you're doing the experiment!)
- if θ is large in comparison to σ_1^2 and σ_2^2 then there will be little difference between the optimal estimator (defined by a^*, b^*) and the optimal unbiased estimator. (See tutorial sheets for further discussion of this.)

Note: The second derivative test in 2 dimensions

Suppose that $f : R^2 \to R$ is a twice continuously differentiable function and that $f(x,y)$ has a critical point at (x_0, y_0).

- (1) If

$$\frac{d^2}{dx^2}f(x_0,y_0)\frac{d^2}{dy^2}f(x_0,y_0) - \left(\frac{d^2}{dxdy}f(x_0,y_0)\right)^2 > 0$$

then either $\frac{d^2}{dx^2}f(x_0,y_0)$ and $\frac{d^2}{dy^2}f(x_0,y_0)$ are *both* positive or *both* negative.

- Furthermore, if $\frac{d^2}{dx^2}f(x_0, y_0) > 0$ (or $\frac{d^2}{dy^2}f(x_0, y_0) > 0$) then $f(x, y)$ has a LOCAL MINIMUM at (x_0, y_0). If $\frac{d^2}{dx^2}f(x_0, y_0) < 0$ (or $\frac{d^2}{dy^2}f(x_0, y_0) < 0$) then $f(x, y)$ has a LOCAL MAXIMUM at (x_0, y_0).
- (2) If
$$\frac{d^2}{dx^2}f(x_0, y_0)\frac{d^2}{dy^2}f(x_0, y_0) - \left(\frac{d^2}{dxdy}f(x_0, y_0)\right)^2 < 0$$
then $f(x, y)$ has a SADDLE POINT at (x_0, y_0).
- (3) If
$$\frac{d^2}{dx^2}f(x_0, y_0)\frac{d^2}{dy^2}f(x_0, y_0) - \left(\frac{d^2}{dxdy}f(x_0, y_0)\right)^2 = 0$$
then the test is inconclusive

2.3 Consistency

In essence, an estimator is *consistent* if the larger the size of the random sample, the better the estimate is. However, we need to express this idea mathematically. Suppose $\underline{X}$ is a random sample of size n ang $g(\underline{X})$ is an estimator of θ. Then we say that g is consistent if, for all δ,

$$\lim_{n\to\infty} P(|g(\underline{X}) - \theta| > \delta) = 0.$$

Put another way this means that as n increases the probability that the modulus of your error is bigger than δ tends to 0 (for any δ). Clearly, the smaller you make δ then the larger n would need to be to ensure, say, 99% probability of having an error smaller than δ.

To show that an estimator is consistent from the above definition looks a but tricky. However we can use Tchebychev's inequality to relate consistency to the MSE.

Since

$$P(|g(\bar{X}) - \theta| > \delta) < \frac{E(g(\underline{X}) - \theta)^2)}{\delta^2} = \frac{MSE(g(\underline{X}))}{\delta^2}$$

by T's Inequality, it follows that g is consistent if we can show that $MSE(g(\underline{X}))) \to 0$ as $n \to \infty$. Equivalently (since $MSE = Var + Bias^2$) we just need to check that the bias and variance of an estimator both tend to zero as the sample size increases.

2.3.1 Example

For a random sample $\underline{X}$ of size n from a population with mean μ and variance σ^2 the sample mean $\bar{X}$ is a consistent estimator of μ. Why?

.

2.4 Most efficient estimators and the Cramer-Rao Lower Bound (CRLB)

How can we decide whether or not a give estimator $g(\underline{X})$ is the most efficient possible (i.e. the one with the smallest MSE)? In general this is very difficult to do. However, we are sometimes able to do this for the particular case of an *unbiased* estimator.

2.4.1 The Cramer-Rao Lower Bound

Let $\underline{X}$ be a random sample of size n from a distribution with pmf or pdf $f_X(x;\theta)$, parameterised by θ. Suppose that:

- $g(\underline{X})$ is an unbiased estimator for θ;
- the range of X does not depend on θ.

Then

$$Var(g(\underline{X})) \geq \frac{1}{nE[(\frac{\partial \log(f_X(X;\theta)}{\partial\theta})^2]} = -\frac{1}{nE[\frac{\partial^2 \log(f_X(X;\theta)}{\partial\theta^2}]}$$

This gives a way of testing whether or not a given unbiased estimator might be the most efficient unbiased estimator. Its proof will be encountered later in 3rd year. In the meantime we illustrate its use with a simple example.

2.4.2 Estimating a bionomial proportion

A gambler is concerned that a coin may not be fair and conducts an experiment to estimate the probability $p \in (0,1)$ of obtaining a head (H) in which he tosses the coin m times. He considers the outcome of this experiment as a random sample $\underline{X} = (X_1, X_2, ..., X_n)$ of size m from a Bernoulli(p) distribution where $X_i = 1$ if the i^{th} toss results in 'H' and is 0 otherwise.

Is there an obvious estimator for p? Yes, let $\hat{p} = g(\underline{X}) = \frac{\sum X_i}{m}$ (*i.e.* the number of heads over the total number of tosses). We now consider the following questions.

- Find the bias, variance and MSE of $\hat{p}$.
- Is $\hat{p}$ a *consistent* estimator?
- is $\hat{p}$ the most efficient estimator of p?

See lectures for solutions:

2.5 Simple random samples from finite populations

Up until now we've looked at mainly random samples where the observations are i.i.d.. Independence of measurements cannot always be guaranteed. One situation where dependence naturally arises is when our experiment involves taking a simple random sample of size n without replacement from a population of size N, and measuring some property of each Y. We assume that all subsets of size n are equiprobable to be chosen. Suppose the n values are $Y_1, Y_2, ..., Y_n$. Now, the Y_i all have the same *marginal* distribution. (What is this?) Why are they not *independent* random variables?

Suppose that we wish to use the sample to estimate the population mean:

$$\mu = \frac{1}{N}\sum_{1}^{n} y_i$$

Then a natural estimator to consider is $\bar{Y} = \frac{1}{n}\sum Y_i$. Since $E(Y_i) = \mu$ it follows (check this!) that $\bar{Y}$ is unbiased. However, as we see, its MSE is *not equal* to $\frac{\sigma^2}{n}$ where

$$\sigma^2 = Var(Y_i) = \frac{1}{N}\sum_{1}^{N}(y_i - \mu)^2.$$

To work out the MSE we need to be able to calculate the variance of $\bar{Y}$ over repeated sampling. This is done as follows.

$$\begin{aligned} Var(\bar{Y}) &= E((\bar{Y}-\mu)^2) \\ &= \frac{1}{n^2}E[(\sum_{j=1}^{n}(Y_j-\mu))^2] \\ &= \frac{1}{n^2}(\sum_{j=1}^{n}Var(Y_1)+\sum_{j_1\neq j_2}\sum Cov(Y_{j_1},Y_{j_2})) \\ &= \frac{1}{n^2}(nVar(Y_1)+n(n-1)Cov(Y_1,Y_2)) \end{aligned}$$

Now

$$Cov(Y_1,Y_2)=\frac{1}{N(N-1)}\sum_{j_1\neq j_2}\sum(y_{j_1}-\mu)(y_{j_2}-\mu)$$

Note that

$$\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}(y_{i_1}-\mu)(y_{i_2}-\mu)=0$$

This implies that

$$Cov(Y_1,Y_2)=\frac{1}{N(N-1)}\sum_{j_1\neq j_2}\sum(y_{j_1}-\mu)(y_{j_2}-\mu)=-\frac{1}{N-1}Var(Y_1).$$

It follows that

$$\begin{aligned} Var(\bar{Y}) &= \frac{1}{n}(1-\frac{n-1}{N-1})Var(Y_1) \\ &= (1-\frac{n-1}{N-1})\frac{\sigma^2}{n} \\ &= \frac{(N-n)}{(N-1)}\frac{\sigma^2}{n} \end{aligned}$$

Since $\bar{Y}$ is unbiased this gives us the MSE also. We can see that MSE is less than that which we would get if we carried out random sampling with replacement. If we replace then the observations are independent and the MSE is $\frac{\sigma^2}{n}$ as before. *Note.*

- The factor $\frac{N-n}{N-1}$ is sometimes referred to as the *finite population correction factor.*

- If N is very large compared to n then this factor is close to 1. Moreover, $Cov(Y_1, Y_2)$ is close to 0 in this case. When we are taking a sample from a very large population, then we can treat the observations as being independent.

A good example of where this arises is in opinion polls. Suppose we select 1000 people randomly without replacement from a population of several million and ask them whether they intend to vote Conservative in the next election. Let the observations be $Y_1, ..., Y_{1000}$ where $Y_i = 1$ if they intend to vote Conservative and 0 otherwise. Then, $Y_i \sim Bernoulli(p)$ where p is the proportion of Conservative voters in the whole population. Since N is large compared to n we can assume that the Y_i are independent in which case the sum of the Y_i (i.e. the number of Conservative voters in the sample) is $Bin(1000, p)$.

3 Constructing estimators

We've looked at properties of some estimators without describing a systematic approach to constructing them. In this section we will look at some ways of constructing estimators. The first of these has been met before (if you did Statistics II!). This is what we call method of moments estimation. The second approach is very different (even though it may lead to the same estimator in some situations) and is called maximum likelihood estimation.

3.1 Method of moments estimation (MME)

Consider an experiment that yields an i.i.d. set of observations $X_1, ..., X_n$ from a distribution with density or mass function $f_X(x;\theta)$, where θ is unknown. We will often consider cases where θ is a 1-dimensional vector. Sometimes it will be a vector that has more than a single component. For example if $X \sim \Gamma(\alpha, \beta)$ then the dimension of $\theta = (\alpha, \beta)$ is 2. How does MME operate in order to come up with an estimate for θ? Put simply it estimates θ so as to make the moments for the values in the sample match the 'theoretical' moments of the population distribution. It is best seen by example.

3.1.1 Estimating p for the coin-tossing experiment

Recall that the results of the experiment in which a coin was tossed n times gives a random sample $X_1, ..., X_n$ from a $Bernoulli(p)$ distribution (where a 'H' is signified by $X_i = 1$. The obvious estimator is $\hat{p} = g(\underline{X}) = \frac{\sum X_i}{n}$. How could we have derived this systematically using MME. For the $Bernoulli(p)$ distribution, the first moment is:

$$E(X) = \mu(p) = p.$$

We equate this 'population' moment to the sample mean, $\bar{X} = \frac{\sum X_i}{n}$ and solve for the parameter p to obtain the estimator

$$\hat{p} = \bar{X}$$

More generally the method of moments estimator for a 1-dimensional parameter θ is found by solving the equation

$$\mu(\theta) = \bar{X}$$

3.1.2 MME for λ in Exp(λ) distribution

You believe that the lifetime of a certain kind of lightbulb follows an $Exp(\lambda)$ distribution, where λ is unknown. To estimate λ you randomly select n lightbulbs and measure their lifetimes. Call these $X_1, ..., X_n$. What is the MME for λ. To obtain the MME we must identify the population mean, set it equal to the sample mean and solve for λ. This gives

$$\mu(\lambda) = \frac{1}{\lambda} = \bar{X}.$$

From this it follows that the MME for λ is given by

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

In the two examples above we have constructed the method-of-moments *estimator* for a parameter. In other words it is a *random variable* which can be expressed as a function of the values in the random sample. When we carry out the experiment, we obtain the observed values $x_1, \ldots, x_n$ and plug these values into the expression for the MME to obtain the estimate for our experiment. In classical statistics it is vital to recognise the distinction between random variables that vary from experiment to experiment, and the realised values for a particular experiment. The former are denoted by upper case letters (X_i), and the latter by lower case (x_i). All frequentist probabilities relate to the former and are not to be considered as probabilities conditional on the realised values.

3.1.3 MME for more than a single unknown parameter

When there is more than a single unknown parameter we can still do method of moments estimation. However it cannot be done by just solving a single equation. Suppose that instead of the $Exp(\lambda)$ distribution, you believe that the light-bulb lifetimes are distributed as a $\Gamma(\alpha, \beta)$ distribution. (N.B. This is a generalisation of the exponential case since $Exp(\lambda) \sim \Gamma(1, \lambda)$.) We can do this by MME but we need to find two simultaneous equations in α and β to solve. To do this we will need to consider moments of higher order than just the 1st. For the Gamma distribution we have:

$$\begin{aligned} E(X) &= \mu(\alpha, \beta) = \frac{\alpha}{\beta} \\ Var(X) &= \sigma^2(\alpha, \beta) = \frac{\alpha}{\beta^2} \end{aligned}$$

We can generate equations for the MME estimators $(\hat{\alpha}, \hat{\beta})$ using the sample mean and variance $\bar{X}$ and S^2. That is, we solve

$$\begin{aligned} \bar{X} &= \frac{\alpha}{\beta} \\ S^2 &= \frac{\alpha}{\beta^2} \end{aligned}$$

for α and β.

You can check (see tutorial sheet) that this give MMEs

$$\begin{aligned} \hat{\alpha} &= \frac{\bar{X}^2}{S^2} \\ \hat{\beta} &= \frac{\bar{X}}{S^2} \end{aligned}$$

More generally we can do method-of-moments estimation for distributions that have arbitrarily many parameters. Suppose our unknown parameters are $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Then we would need k simultaneous equations in order to solve for the parameters. Therefore we consider all moments up to order k and equate the k^{th} sample moment with the k^{th} population moment. This generates k simulataneous equations of the form

$$\frac{1}{n}\sum X_i^r = \mu_r = E(X^r), r = 1, \ldots, k$$

which we then solve to get the estimators. Equivalently we could work with central moments $E(X - \mu)^r$ equated with their sample values. In ths course we shall not consider k greater than 2, in which case it is sufficient to work with the sample mean and variance as above.

3.1.4 Disadvantages of method of moments

While the method of moments has many advantages (it's generally simple to apply) it also has some disadvantages. In particular, in some circumstances it can give results that are clearly nonsensical. Consider the following example.
Example Suppose that you conduct an experiment in which you observe $X_1, ..., X_n$ where the X_i are i.i.d. from a $U(0, \theta)$, where $\theta > 0$. Find the MME of θ.
Solution. Since we know that $E(X) = \frac{\theta}{2}$ it's straightforward to obtain the MME as $\hat{\theta} = 2\bar{X}$. Now suppose that for the particular experiment $n = 3$

and $(X_1, X_2, X_3) = (0.1, 0.2, 0.9)$. Then in this case we would obtain $\hat{\theta} = 2\bar{x} = 0.8$. However, the data are not consistent with this value of θ. In particular, we cannot have a value of 0.9 as an observation from the $U(0, 0.8)$ distribution.

It is also the case that MME's may not always be particularly efficient. For $\hat{\theta} = 2\bar{X}$ we can show (you should do this for yourself!) that $MSE(\hat{\theta}) = \frac{\theta^2}{3n}$. It follows that the MSE decreases like n^{-1} as the sample size n increases. Later we will discuss an alternative approach that gives an estimator whose MSE decreases more rapidly with n.

A discrete example Suppose that an insurance company records the numbers of claims made on a given class of policy over a number of years. For the data the sample mean and variance are 12.5 and 10.0 respectively. You believe that the number of claims varies from year to year as a $Bin(n, p)$ with n and p unknown.

To estimate n and p using MME we equate the sample mean and variance with the binomial mean and variance to obtain:

$$\begin{aligned} np &= 12.5 \\ np(1-p) &= 10.0. \end{aligned}$$

Solving these gives $\hat{p} = \frac{5}{25}$ and $\hat{n} = 62.5$. Of course we should round the latter estimate to 62 or 63 since n is constrained to be an integer.

What would have happened if the sample mean and variance had been 12.5 and 16.0 respectively? In this case when you solve the equations you will get nonsensical answers ($\hat{p} = -0.28$)! Since the variance of the binomial cannot be greater than the mean, then when we apply MME to data where the sample variance is *bigger* than the sample mean we come unstuck! Whenever you come across discrete data where this is the case then it probably does not come from a binomial distribution. If the sample mean and sample variance are approximately equal then the Poisson distribution is a possible choice.

3.2 Likelihood and Maximum Likelihood Estimation

In this section we introduce one of the key concepts in statistics - the *likelihood*. We illustrate it with an example. Suppose I have 2 biased dice. For die number 1, the frequencies with which the numbers 1 to 6 arise are, respectively:

0.1, 0.1, 0.1, 0.1, 0.1, 0.5.

For die number 2, the frequencies are:

0.5, 0.1, 0.1, 0.1, 0.1, 0.1.

I roll one of the dice and score a 6. Which die do you think I rolled?

This is an example of estimating an unknown parameter (the number of die rolled) from data (the score on the roll). Most people would guess that the die rolled was number 1, since that yields a 6 with probability 0.5, while the other die only yields a 6 with probability 0.1. Put another way, the score 6 is *much more likely* if the chosen die is number 1, compared with number 2. This in essence is the idea underlying maximum likelihood estimation of a parameter θ from an experiment. We estimate θ to be the value that makes the data most likely.

3.2.1 The likelihood.

We introduce the ideas in a very general way before illustrating them for particular examples. Suppose we carry out an experiment which can give various outcomes with probabilities governed by an unknown parameter θ. Suppose *in our particular experiment* we observe data y. Then the likelihood is a function which takes all the possible parameter values θ and is defined to be:

$$L(\theta; y) = Pr(y|\theta).$$

In other words is tells us how 'likely' the data are for different values of θ. It is important to stress that we will be interest in how $L(\theta; y)$ varies with θ - the experimental observations y will typically be fixed. Calculating likelihoods will require you to be familiar with the rules of probability, the use of probability functions and cumulative distribution functions. Having constructed a likelihood, we obtain the maximum likelihood estimate of θ as the value of θ that maximises the likelihood.

3.2.2 Likelihood for i.i.d. observations from a discrete distribution

We begin with the case where our experiment produces observations $Y_1, ..., Y_n$ where the Y_i are i.i.d. samples from as discrete distribution with probability mass function $f_Y(y;\theta)$. Suppose the realised values in the experiment are $y_1, \dots y_n$. The construction of the likelihood is straightforward here. We have

$$\begin{aligned} L(\theta; y) &= Pr(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|\theta) \\ &= f_Y(y_1;\theta) \times f_Y(y_2;\theta) \times \cdots \times f_Y(y_n;\theta) \\ &= \prod_{i=1}^{n} f_Y(y_i;\theta) \end{aligned}$$

The above joint probability decomposes as a product since the observations are all independent of each other. Having obtained the likelihood we obtain the MLE by maximising $L(\theta)$ with respect to θ. This will require us to use basic calculus. Often it is more convenient to maximise the logarithm of the likelihood

$$l(\theta; y) = \log(L(\theta; y) = \sum_{i=1}^{n} \log(f_Y(y_i;\theta)).$$

Note that since log is a monotonic increasing function, the value of θ that maximises $l(\theta)$ also maximises $L(\theta)$ and *vice versa*.

We illustrate this now for some basic examples.

Example. A gambler tosses a coin n times and records the outcome of each toss in order to estimate the unknown probability of a H, $p \in (0, 1)$. Suppose that r of the tosses results in 'H'. He supposes that the results of the tosses are independent events and that the probability of 'H' is the same for each toss. Let y denote the observations, so that y is sequence of length n consisting of r H's and $n - r$ T's. It follows that

$$L(p; y) = p^r(1 - p)^{n-r}.$$

Now the log-likelihood is given by

$$l(p; y) = r \log p + (n - r) \log(1 - p).$$

Differentiating with respect to p and equating to zero gives

$$\frac{dl}{dp} = \frac{r}{p} - \frac{n - r}{1 - p} = 0.$$

When we solve for p we obtain

$$\hat{p} = \frac{r}{n}.$$

You should also check that this value of p gives a maximum of the log-likelihood. This can be done by evaluating the 2nd derivative with respect to p:

$$\frac{d^2l}{dp^2} = -\frac{r}{p^2} - \frac{n-r}{(1-p)^2} < 0, \forall p \in (0,1).$$

This means that $\hat{p} = \frac{r}{n}$ is indeed the maximum likelihood estimate of p.

Note. The maximum likelihood estimator is R/n where R is the random variable denoting the number of H's in n trials. R follows a $Bin(n,p)$ distribution and we can use this fact to work out its sampling properties (e.g. bias and MSE).

We see that for this example, the MLE is exactly the same as the MME calculated earlier.

Example. An astronomer counts the number of meteors in a section of the night sky for n consecutive nights obtaining data $x_1, \ldots, x_n$. She believes that these values are a random sample (i.i.d.) from a $Poisson(\lambda)$ distribution where λ is unknown. What is the MLE of λ?

Solution It is again simple to construct the likelihood since the observations are assumed to be i.i.d.. We have

$$L(\lambda) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$$

Taking the log, differentiating, equating to zero and solving for λ, we obtain

$$\hat{\lambda} = \frac{\sum x_i}{n} = \bar{x}.$$

We again see that the MLE of λ is the same as we would obtain using method of moments. This is not always the case (see later examples).

3.2.3 Likelihood for i.i.d. observations from a continuous distribution

Suppose now our experiment consists of i.i.d. observations $X_1, \ldots, X_n$ from a continuous distribution with density function $f_X(x;\theta)$. This would be the case if our experiment involved measuring lifetimes of people or components, weights, event times etc. etc.. How do we construct the likelihood in this case. At first sight there is a problem. If we assume that our observations are measured to arbitrary high precision (an infinite number of decimal places), then the 'probability of the data' is equal to zero for any θ. However let's suppose that we can only measure a physical quantity to some finite precision (i.e. within $\pm\frac{\delta}{2}$) for some δ and that we observe the values $x_1, \ldots, x_n$. Then we consider this event as $A = A_1 \cap A_2 \cap \cdots \cap A_n$, where A_i is the event $X_i \in (x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2})$. Since the X_i are independent random variables then the A_i are a set of pairwise independent events. Moreover we have

$$P(A_i) \approx f_X(x_i;\theta)\delta.$$

It follows that, since the likelihood as the probability of the data given θ,

$$L(\theta;x) \approx Pr(A|\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)\delta^n,$$

with the approximation becoming increasingly accurate as δ decreases (i.e. the precision increases). Now, the positive factor δ^n does not depend on θ and has no bearing on the value of θ that maximises $L(\theta)$. Therefore, we can omit it from the the above expression and define our likelihood to be

$$L(\theta;x) = \prod_{i=1}^{n} f_X(x_i;\theta).$$

3.2.4 MLE for the λ in $Exp(\lambda)$

Suppose we observe $X_1, ..., X_n$ i.i.d. from an $Exp(\lambda)$. What is the MLE of λ? The p.d.f. is $f_X(x;\lambda) = \lambda e^{-\lambda x}$. For given $\underline{x} = (x_1, ..., x_n)$, the likelihood is

$$L(\lambda) = \prod_{i=1}^{n} f_X(x_i;\lambda) = \lambda^n e^{-\lambda \sum x_i}$$

Taking logs and maximising you should find that the maximum likelihood estimate is $\hat{\lambda} = \frac{1}{\bar{x}}$. This is a question in the tutorials. How does this estimate compare with the method-of-moments estimate?

3.3 Calculus won't always help you.

Up until now we've looked at likelihoods that can be maximised using methods from calculus. However, in some situations alternative approaches are necessary. Consider the case where we observe $X_1, ..., X_n$ from a Uniform(0, θ) distribution, where $\theta > 0$. What is the MLE of θ? Well, we know that $f(x;\theta) = \frac{1}{\theta}$,if $0 < x < \theta$, and is 0, otherwise. Now the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

So long as $\theta > x_i$ for all i then the right-hand-side of this equation will be equal to $\frac{1}{\theta^n}$. However, if $x_i \geq \theta$ then the factor $f_X(x_i;\theta)$ vanishes in the likelihood in which case the whole product is zero.

It is *easy* to find the value of θ that maximises L if we *graph* the likelihood function:

It follows that the maximum likelihood estimator of θ is

$$\hat{\theta} = X_{(n)} = \max\{X_1, X_2, ..., X_n\}.$$

Problem: Find the Bias and MSE of the MLE for θ. Compare your results with the corresponding ones for the MME. Which estimator do you think is better?

Solution: The main steps are

- Find the sampling distribution of $\hat{\theta} = X_{(n)} = \max\{X_1, X_2, ..., X_n\}$.
- Then obtain its mean and variance to calculate the Bias and MSE.

4 Likelihood estimation continued

4.1 Practical sampling situations

Maximum likelihood estimation is very flexible as we can use it whenever we can construct a likelihood for the observations we obtain in our experiment. Remember that we can think of the likelihood $L(\theta; \underline{x})$ (loosely) as being the probability of the obervations, $\underline{x}$ as a function of the parameter(s) θ.

In the examples we've seen so far the observations have taken the form of random samples $X_1, ..., X_n$ where the X_i are discrete or continuous random variables. There are many practical situations that do not conform to such a simple sampling model and we need to use our skills in probability to construct likelihoods in these cases also.

4.1.1 An example of censored observations.

The lifetimes (in months) of a certain brand of lightbulbs are believed to follow an Exp(λ) distribution where λ is unknown. To estimate λ you take a random sample of 20 bulbs and set them in continuous operation at $t = 0.0$ and subsequently measure the times at which each bulb fails. You have promised to give your estimate of λ no later than 2 months after the start of the experiment. At the end of the 2 month period, 10 bulbs have failed at times $t_1, t_2, ..., t_{10}$ with $\sum_{i=1}^{10} t_i = 7.3$. The remaining 10 bulbs are still operational. Can we construct a likelihood and estimate λ?

Solution. We assume that the lifetimes of bulbs in the sample are *independent* of each other. Therefore, 'the probability of the data given λ' naturally can be represented as a product of factors - one corresponding to the experimental result for each of the 20 bulbs. For the bulbs failing at $t_i < 2$ months,$1 \leq i \leq 10$, the probability is $f_X(t_i; \lambda)\delta = \lambda e^{-\lambda t_i}\delta$. Here δ represents the finite precision to which the times are measured. It can be omitted from the likelihood since it doesn't depend on λ.

Now for a bulb which doesn't fail, the probability of the observation is the probability that a lifetime exceeds 2 months, i.e.:

$$P(X > 2) = 1 - F_X(2; \lambda) = 1 - (1 - e^{-2\lambda}) = e^{-2\lambda}$$

Each bulb surviving the 2-month period contributes a similar factor to the likelihood. Therefore we obtain as our likelihood:

$$L(\lambda) = \prod_{i=1}^{10} \lambda e^{-\lambda t_i} \times \prod_{i=11}^{20} e^{-2\lambda} = \lambda^{10} e^{-\lambda(20+\sum t_i)}$$

Taking logs we obtain the log-likelihood:

$$l(\lambda) = 10 \log(\lambda) - \lambda(20 + \sum t_i).$$

Differentiating and equating to zero (you should be able to do this!!) we find that this is maximised by

$$\hat{\lambda} = \frac{10}{(20 + \sum t_i)}.$$

This answer is intuitively plausible. The denominator is the total number of months for which bulbs were operational during the experiment. Therefore the estimate of λ has the form (number of failures)/(total months of operation) and seems a natural estimator of a failure 'rate'.

4.2 Invariance of MLEs under 1-1 transformations

There is a very useful property of maximum likelihood estimators that we can sometimes exploit to simplify the problem of finding the MLE of a parameter. Let $\phi = h(\theta)$ be a 1-1 mapping. Then we could parameterise our model in terms of ϕ instead of θ. (For example, sometimes you will see the exponential distribution parameterised using the mean $\mu = \frac{1}{\lambda}$.) Now if $\hat{\theta}$ is the MLE of θ then it follows that the MLE of ϕ satisfies

$$\hat{\phi} = h(\hat{\theta}).$$

How can this simplify things? Well let's go back to the light bulb experiment of the last section and suppose that you don't measure*any* times precisely but only whether each bulb is operational at $t = 2$, noting that 10 out of the 20 have failed before this time. For this experiment, the obervations can be considered to be a random sample from a Bernoulli(p) distribution (with '1' denoting the event that a given bulb survives beyond $t = 2$). The parameter p is related to λ by $p = e^{-2\lambda}$, this being the probability of a random lifetime exceeding 2 months.

Now it is easy to check that the MLE of p is $\hat{p} = \frac{1}{2}$. By the invariance of MLEs under 1-1 transformation, we have that $\hat{p} = e^{-2\hat{\lambda}}$ in which case we have that $\hat{\lambda} = -\frac{\log p}{2}$.

4.3 MLEs for parameters in Normal distribution.

Suppose that $\underline{X}$ is a random sample of size n from $N(\mu, \sigma^2)$. What are the MLEs for μ and σ^2?

The likelihood function is given by

$$\begin{aligned} L(\mu, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\ logL(\mu, \sigma^2) &= -n\log(\sqrt{2\pi}) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned}$$

Differentiate the log-likelihood function with respect to μ and σ^2 to get the following system of equations:

$$\frac{d}{d\mu}\log L(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \quad (1)$$

$$\frac{d}{d\sigma^2}\log L(\mu, \sigma^2) = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0. \quad (2)$$

The first of these equations can be solved to get $\hat{\mu} = \bar{X}$.

Next, substitute the estimate $\bar{X}$ for μ into the second equation, and solve for σ^2.

We get $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Q. Is $\log L(\mu, \sigma^2)$ *maximized* at $\hat{\mu} = \bar{X}$, $\hat{\sigma^2} = \frac{n-1}{n}S^2$?

Since $\log L(\mu, \sigma^2)$ is a function of *two* variables, we have to check various conditions of the *2-dimensional* 2nd derivative test: We now **apply** this test.

Note that:

$$\begin{aligned}
\frac{d^2}{d\mu^2}\log L(\hat{\mu},\hat{\sigma}^2) &= \frac{-n}{\hat{\sigma}^2} < 0 \\
\frac{d^2}{d(\sigma^2)^2}\log L(\hat{\mu},\hat{\sigma^2}) &= \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{n}{2\hat{\sigma}^4} - \frac{n\hat{\sigma}^2}{\hat{\sigma}^6} = \frac{-n}{2\hat{\sigma}^4} < 0 \\
\frac{d^2}{d\mu d\sigma^2}\log L(\hat{\mu},\hat{\sigma}^2) &= -\frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\bar{x}) = 0
\end{aligned}$$

Therefore,

$$\frac{d^2}{d\mu^2}\log L(\hat{\mu},\hat{\sigma}^2)\cdot\frac{d^2}{d(\sigma^2)^2}\log L(\hat{\mu},\hat{\sigma}^2) - \left(\frac{d^2}{d\mu d\sigma^2}\log L(\hat{\mu},\hat{\sigma}^2)\right)^2 = \frac{n^2}{2\hat{\sigma}^6} > 0.$$

So, $L(\mu,\sigma^2)$ is *maximized* at $\hat{\mu} = \bar{X}, \hat{\sigma}^2 = \frac{n-1}{n}S^2$, and this is an absolute maximum (since there is only one critical point).

Note that the MLE of σ^2 is *biased* (since $E(S^2) = \sigma^2$).

4.3.1 Another example with censored observations

Suppose that a random sample of size n is taken from the normal distribution, $N(\mu, 1)$, but only the *signs* of the observations are recorded.

Q. If k observations are negative (and $n-k$ are positive), what is the MLE for μ?

We can find the MLE provided we can write down the likelihood function. If $x \sim N(\mu,1)$, then the probability that X is negative is given by

$$P(X<0) = P(\frac{X-\mu}{1} < -\mu) = P(Z < -\mu) = \Phi(-\mu)$$

where $Z \sim N(0,1)$ and $\Phi(x)$ is the distribution function for the standard normal distribution. Therefore,

$$\begin{aligned}
L(\mu) = L(\mu; x_1, x_2, ..., x_n) &= (\Phi(-\mu))^k(1-\Phi(-\mu))^{n-k} \\
l(\mu) &= k\log(\Phi(-\mu)) + (n-k)\log(1-\Phi(-\mu)) \\
\frac{d}{d\mu}l(\mu) &= \frac{-k\phi(-\mu)}{\Phi(-\mu)} + \frac{(n-k)\phi(-\mu)}{1-\Phi(-\mu)}
\end{aligned}$$

Solve $\frac{d}{d\mu}l(\mu) = 0$ to get $\Phi(-\mu) = \frac{k}{n}$.
Therefore, $\hat{\mu} = -\Phi^{-1}(\frac{k}{n})$. (Need to check that $\log L(\mu)$ is maximized at $\hat{\mu} = k/n$.)

To actually find the numerical value of $\hat{\mu}$ you need to use tables.

Example: If $n = 10, k = 4$, then $\hat{\mu} = 0.25$.

4.4 Numerical methods

For the examples that we've seen likelihoods and log-likelihoods can be maximised using analytic methods - (calculus, or considering graphs etc.). However this won't always work. Consider the following (highly artificial) example.

Consider a population of sealed boxes each containing a random number of lead weights, X, where $X \sim Poisson(\lambda)$. If a box contains 2 or more weights it will sink in water. Otherwise it will float. You select a random sample of 10 boxes, 5 of which float. What is the MLE of λ?

Solution. The MLE satisfies then equation

$$2(1 + \hat{\lambda})e^{-\hat{\lambda}} - 1 = 0.$$

See lecture notes for a derivation of this.

Now this equation cannot be solved analytically. We require to use a numerical algorithm to obtain the solution.

4.4.1 Newton-Raphson algorithm

The Newton-Raphson algorithm is a simple way of finding the root of an equation of the form $f(x) = 0$ where f is differentiable, given an initial guess of the root, x_0. The idea is to suppose that the graph of f approximates to a straight line in a neighbourhood of the root and use this to come up with an improved guess x_1. It is best understood from the following diagram:

From the diagram we have that

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Now we can apply the same procedure to x_1 to obtain an even better guess. In this way we can generate a sequence of estimates $x_1, x_2, \ldots, x_i, \ldots$ where

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}.$$

We repeat the process until the values of x_i converge.

Now let's do this for the 'boxes' example. We need to solve $f(\lambda) = 0$ where $f(\lambda) = 2(1+\lambda)e^{-\lambda} - 1$. Now $f'(\lambda) = -2\lambda e^{-\lambda}$ and we obtain the recursion:

$$\lambda_{i+1} = \lambda_i + \frac{2(1+\lambda_i)e^{-\lambda_i} - 1}{2\lambda_i e^{-\lambda_i}}.$$

We also need an initial guess. Now since 5 of the observed values of X are less than or equal to 1, and the other 5 are greater than or equal to 2 we could guess that the mean is around 1.5. Since the mean of the Poisson is λ then $\lambda_0 = 1.5$ might not be a bad guess. This gives the sequence of estimates:

The Newton-Raphson algorithm is fairly flexible. So long as the initial guess is sufficiently close to the root it will ultimately find it. There are several tutorial problems where the algorithm is used.

4.5 Important properties of likelihood.

Here we summarise some of the important properties of likelihood that make a very useful tool for constructing estimators.

1. If there exists a most efficient unbiased estimator for a parameter θ that attains the Cramer-Rao lower bound, then it must be the Maximum Likelihood Estimator (MLE). Therefore if we are trying to find a 'good' estimator for θ then it makes sense to try and find the MLE.

2. Maximum likelihood estimation can be applied whenever we are able to write down the likelihood (i.e. 'the probability of the observations given θ'). It is in the construction of likelihoods that your skills in probability theory are vital!

3. When the equation
$$\frac{dl}{d\theta} = 0$$
cannot be solved analytically we can nevertheless use numerical methods to solve it and identify the maximum likelihood estimates.

4. MLEs are invariant under 1-1 transformations. That is to say if h is a 1-1 mapping and $\hat{\theta}$ is the MLE of θ then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

5. *Asymptotic distribution of the MLE.* Suppose the data $X_1, ..., X_n$ are a random sample from a distribution whose range does not depend on the parameter θ. Then it is (usually) the case that $\hat{\theta}$ is approximately normally distributed with mean θ, and variance equal to the Cramer-Rao Lower Bound, so long as the sample size n is large. This means that MLEs are asymptotically *normal*, *unbiased* and *most efficient*.