THE DESIGN AND MIXED-MODEL ANALYSIS OF EXPERIMENTS

As the title suggests this course is about experiments — how to design them and how to analyse them. Its features are:

- the standard range of designs used in agriculture will be covered
- determining the number of replicates will be described
- the analyses will based on linear models, not nonlinear models, and fixed versus random terms will be discussed
- the model selection aspect of the analysis will be emphasized — that is, what model best describes the behaviour of the data
- the detailed analysis of treatment differences will be included
- Genstat will be used to perform the analyses — its facilities for analyzing designed experiments are outstanding
- the theory underlying the analysis carried out in Genstat will be covered — it will be a matrix approach to the theory with a heavy emphasis on projection matrices.

In this first week and I am going to revise a number of topics to make sure that we all have a basic understanding. I am going to assume that you are thoroughly familiar with the basic matrix algebra and the use of summation notation: $\sum_{i=1}^{n}\ldots$ If unsure try it out on a small arithmetical example that you can easily verify. For example to show that $\sum_i ax_i = a\sum_i x_i$, verify for $a = 2$ and $x_1 = 1$ and $x_1 = 3$: $(2\times 1)+(2\times 3) = 2\times(1+3)$.

I. Statistical inference

I.A Overview of the process

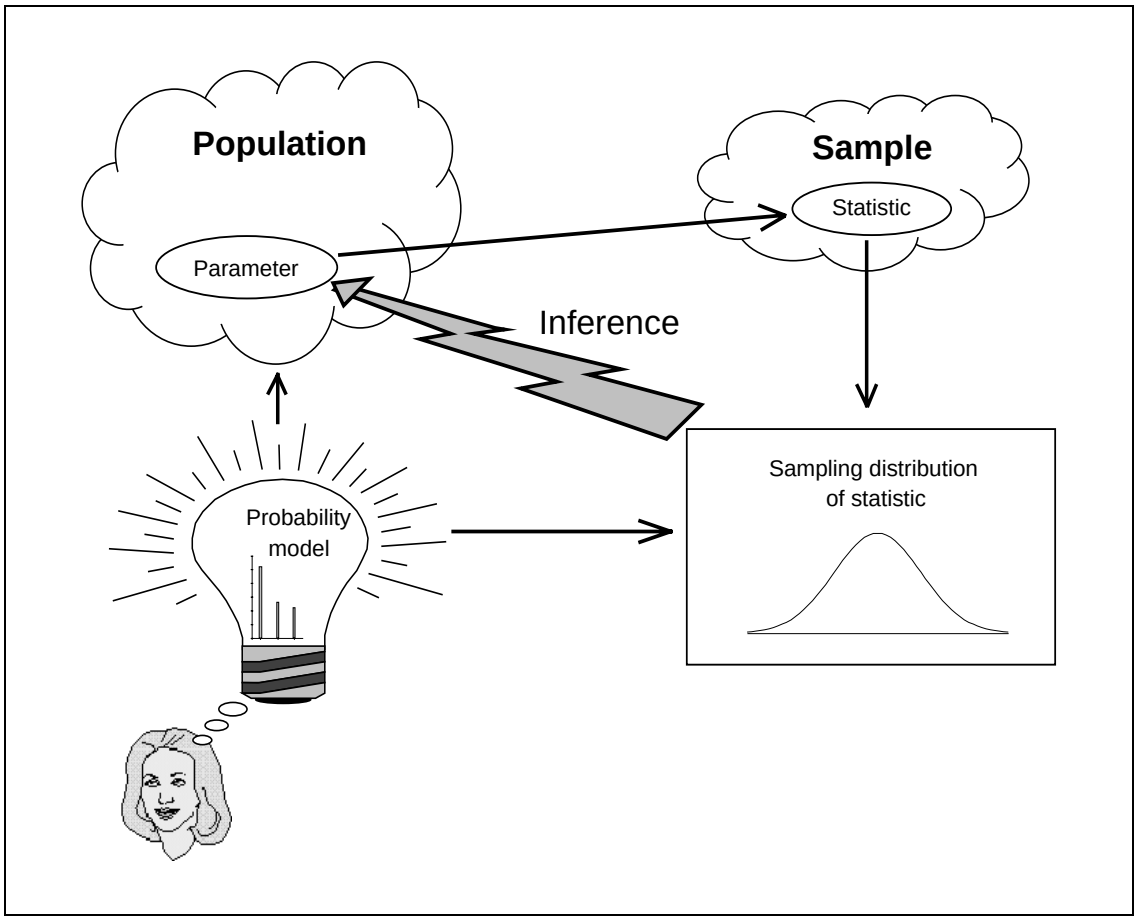
Statistical inference involves the following process, which is illustrated in the figure below:

1. *Specify the parameter of interest:* the **parameter** is a numeric value calculated from the values of the observed variable for all members of the population.
2. *Set up a probability model for the situation:* the **probability model** is a *representation* of the relative frequencies with which the values of the variable occur in the population.
3. *Obtain a sample, identify a statistic to estimate the parameter and compute the statistic:* the **statistic** is a numeric value calculated from the observed values of the variable for the members of the sample only.

4. *Derive the sampling distribution of the statistic:* The **sampling distribution** of a statistic gives the probability with which values of the statistic will occur when a population is repeatedly sampled; the statistic is computed, for each sample, from the values of the variable that is measured. In deriving the sampling distribution, one assumption we make is that the probability model adequately describes the population distribution of the values of the variable.
5. *Make inference about the parameter:* Use the sampling distribution of the statistic, under the probability model, to quantify the uncertainty in inferences or conclusions about the parameter.

So the key elements of this process are that we must rely on a sample of the population to draw our conclusions about the entire population, and this imposes an element of uncertainty on our conclusions. The sampling distribution, which is based on a probability model, is used to assess this uncertainty.

The process underlying statistical inference



In this lecture, we are not going to do full-blown statistical inference. In fact we are going to avoid determining the sampling distribution itself. However, we will examine the mean of the sampling distribution. Further, our statistical inference is going to be rather crude in that only a point estimate is going to be used to make inference.

I.B Unbiased estimation for continuous random variables

Suppose I want to estimate the weight of steers in a particular herd. I obtain a random sample of 12 steers and I am going to use the sample mean to estimate the population mean. We are in the situation of wanting to make a statistical inference because we want to draw a conclusion about the population based on a sample. Firstly, we recognize that the parameter of interest is the population mean. Secondly, we need to set up a probability model for our variable. Formally stated, we have to identify the random variable involved and a probability distribution that represents or 'models' the distribution of values of the random variables in the population.

Definition I.1: A **continuous random variable**, *Y* say, is defined as the set of values from a sample space that consists of all real numbers within an interval on the real line, the sample space having a probability distribution function defined on it. ■

Definition I.2: The **sample space** is the set of all possible outcomes in making the prescribed observations. ■

Definition I.3: A **probability distribution function** is a function that specifies the probabilities of values of the random variable. ■

So the sample space for our steer weights is some interval on the real line and a steer weight can be any value in this interval, although the accuracy of our measuring device may limit the values we obtain. As we have a continuous random variable, we require a continuous probability distribution function. Any continuous function, say $f(y)$, can be a continuous probability distribution function provided that

$$\int_{-\infty}^{\infty} f(y)\,dy = 1 \text{ and } f(y) \geq 0 \text{ for } -\infty < y < \infty$$

For a continuous random variable, the probability that *Y* lies in the interval a to b is written as

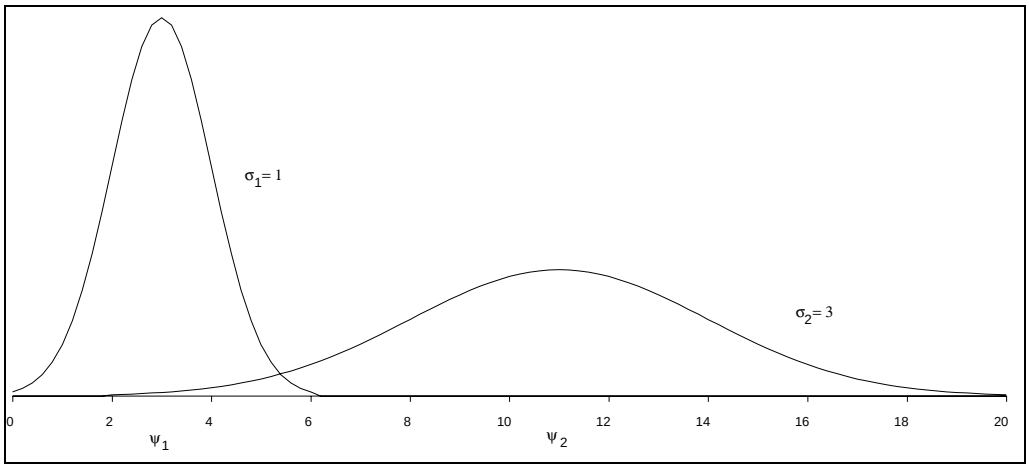
$$P\{a \leq Y \leq b\} = \int_{a}^{b} f(y)\,dy$$

Note the use of *Y* for the name of the random variable and *y* for a value of the random variable.

A very commonly used continuous probability distribution function is the normal distribution. Its formula is:

$$\mathrm{f}(y) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left\{\frac{-(y-\psi)^2}{2\sigma^2}\right\}$$

A plot of this function is given below.



So in employing this distribution as a model we are saying that the population distribution of steer weights for the whole herd displays a shape similar to these diagrams — it is symmetric and bell-shaped. Perhaps it is known from previous experience that weight tends to be normally distributed, so this seems like a good choice. Clearly, to determine the particular normal distribution that is to apply in our example, the values of ψ and σ^2 (or σ) must be specified. We require an estimator of ψ (and perhaps σ). Here we concentrate on the estimator of ψ. But what does ψ represent? The following definition and theorems will tell us.

The expected value is the mean of the variable Y in a population — it is a population parameter.

Definition I.4: The **expected value**, $E[Y] = \psi_Y$, of a continuous random variable Y whose population distribution is described by $f(y)$ is given by

$$E[Y] = \int_{-\infty}^{\infty} y f(y)\, dy$$

■

That is, $E[Y] = \psi_Y$ is the mean in a population whose distribution is described by $f(y)$.

Theorem I.1: Let Y be a continuous random variable with probability distribution function $f(y)$. The expected value of a function $u(Y)$ of the random variable is

$$E[u(Y)] = \int_{-\infty}^{\infty} u(y) f(y)\, dy .$$

Proof: not given ■

Note that any function of a random variable is itself a random variable. This theorem tells us how to get the expected value of the function. It will now be used to find $E\left[a\times v(Y)+b\right]$ by setting $u(Y)=a\times v(Y)+b$.

Theorem I.2: $E\left[a\times v(Y)+b\right]=aE\left[v(Y)\right]+b$

Proof:

For a continuous random variable, we have from theorem I.1

$$\begin{aligned}E\left[a\times v(Y)+b\right] &= \int_{-\infty}^{\infty}\left\{a\times v(Y)+b\right\}f(y)\,dy \\ &= \int_{-\infty}^{\infty}\left\{a\times v(Y)\right\}f(y)\,dy+\int_{-\infty}^{\infty}bf(y)\,dy \\ &= a\int_{-\infty}^{\infty}v(Y)f(y)\,dy+b\int_{-\infty}^{\infty}f(y)\,dy \\ &= aE\left[v(Y)\right]+b\end{aligned}$$

■

In particular, $E\left[aY+b\right]=aE\left[Y\right]+b$.

What about the spread in the heights? A measure of spread is the standard deviation which is computed as the square root of the variance.

Definition I.5: The **variance**, $\text{var}\left[Y\right]=\sigma_Y^2$, of any random variable Y is defined to be

$$\text{var}\left[Y\right]=E\left[\left(Y-\psi_Y\right)^2\right]=E\left[\left(Y-E\left[Y\right]\right)^2\right]$$

■

That is variance is the mean in the population of the squares of the deviations of the observed values from the population mean — it measures how far on average observations are from the mean in the population. It is also a population parameter.

Theorem I.3: The variance, $\text{var}\left[Y\right]=\sigma_Y^2$, of a continuous random variable Y whose population distribution is described by $f(y)$ is given by

$$\text{var}\left[Y\right]=\int_{-\infty}^{\infty}\left(y-\psi_Y\right)^2 f(y)\,dy.$$

Proof: This is a straight forward application of theorem I.1 where $u(Y)=\left(Y-\psi_Y\right)^2$. ■

A useful theorem in the context of the normal distribution follows.

Theorem I.4: For a random variable Y that is normally distributed, $E[Y] = \psi_Y = \psi$ and $\text{var}[Y] = \sigma_Y^2 = \sigma^2$.

Proof: The proof, which is not given, uses the definition I.4 for $E[Y]$ and theorem I.3 for $\text{var}[Y]$ for a continuous variable and the formula for the normal probability distribution function. ■

The importance of this theorem is that it tells us that the population mean and population variance for a normal population are equal to ψ and σ^2 from the distribution function. So we want to estimate ψ and we have a sample $y_1, y_2, \ldots, y_n$. The obvious estimator of ψ is the sample mean $\bar{Y} = \sum_{i=1}^{n} Y_i / n$. Note we call the formula that tells us how to estimate a parameter an **estimator**. The value obtained by substituting the sample values into the formula is called the **estimate**.

It is common practice to denote the estimator as the parameter with a caret over it. For example, $\hat{\psi} = \bar{Y}$ means that the estimator of ψ is $\bar{Y}$; in this notation $\hat{\psi}$ also stands for the estimate so that $\hat{\psi} = \bar{y}$ means that an estimate of ψ is $\bar{y}$.

But, is $\bar{Y}$ a good estimator of ψ? What do we mean by good? One property of a good estimator is that it be unbiased.

Definition I.6: An **unbiased estimator** is one for which the expected value of the estimator is equal to the parameter being estimated. ■

In our case, this implies that

$$E\left[\bar{Y}\right] = \psi_{\bar{Y}} = \psi$$

To work out the $E\left[\bar{Y}\right]$ in order to show this result, we need to consider the joint distribution of the sample values and some rules for manipulating expected values.

To derive the joint distribution of sample values, first consider a sample consisting of two observations. Each of these observations will provide us with the value of a random variable. Denote by Y_1 and Y_2 the random variables corresponding to the first and second observations. The probability distribution function of both these random variables is $p(y)$. That is, the distribution of values for the first observation is specified by $p(y)$, as is the distribution of the second observation. Now what is the joint distribution of the two observations? If we assume that the value of the first observation is independent of the value of the second observation, and the fact that we have a random sample supports this assumption, then the probability is the product of the probabilities of getting each value. That is, for two observations y_1 and y_2 from random variables Y_1 and Y_2, the joint probability distribution function is:

$$p(y_1, y_2) = p(y_1)\,p(y_2)$$

Generalizing to a random sample of n, we have an observation, y_i, from each of n random variables, Y_i, $i = 1,\ldots,n$, with joint probability distribution:

$$p(y_1, y_2,\ldots, y_n) = p(y_1)\,p(y_2)\ldots p(y_n)$$

We call $(Y_1, Y_2,\ldots, Y_n)$ a **continuous multivariate random variable**, it involving many variables. Also, $p(y_1, y_2,\ldots, y_n)$ is called a **continuous multivariate probability distribution function**. The probability distribution function for a random sample is a rather special multivariate probability distribution function in that it is the product of the univariate probability distribution functions — in general, they are not.

Theorem I.5: Let $(Y_1, Y_2,\ldots, Y_n)$ be a continuous multivariate random variable with multivariate probability distribution function $f(y_1, y_2,\ldots, y_n)$. The expected value of a function $u(Y_1, Y_2,\ldots, Y_n)$ of the random variable is

$$E\left[u(Y_1, Y_2,\ldots, Y_n)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} u(y_1, y_2,\ldots, y_n)\,f(y_1, y_2,\ldots, y_n)dy_1dy_2\ldots dy_n.$$

Proof: not given. ■

Theorem I.6: Let $(Y_1, Y_2,\ldots, Y_n)$ be a continuous multivariate random variable with multivariate probability distribution function $f(y_1, y_2,\ldots, y_n)$, c_j, $j = 1,\ldots,t$ be t constants and $u_j(Y_1, Y_2,\ldots, Y_n)$, $j = 1,\ldots,t$ be t functions of the multivariate random variable. The expected value of $\sum_{j=1}^{t} c_j u_j(Y_1, Y_2,\ldots, Y_n)$ is:

$$E\left[\sum_{j=1}^{t} c_j u_j(Y_1, Y_2,\ldots, Y_n)\right] = \sum_{j=1}^{t} c_j E\left[u_j(Y_1, Y_2,\ldots, Y_n)\right]$$

Proof: The proof is not given. ■

We are now in a position to work out $E\left[\bar{Y}\right]$ for a continuous random variable.

Theorem I.7: The expected value of the mean of random sample of n observations from a population with probability distribution function $f(y)$ is given by $E\left[\bar{Y}\right] = \psi$.

Proof: As explained above, the random sample $y_1, y_2,\ldots, y_n$ has a continuous multivariate probability distribution function $f(y_1, y_2,\ldots, y_n) = f(y_1)\,f(y_2)\ldots f(y_n)$. The sample mean is an observation of the following function of the random variables:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \sum_{i=1}^{n} Y_i \Big/ n$$

In the notation of theorem I.6, $c_j = 1/n$ and $u_j(Y_1, Y_2, \ldots, Y_n) = Y_j$. Clearly, $E\left[u_j(Y_1, Y_2, \ldots, Y_n)\right] = E\left[Y_j\right] = \psi$. Now,

$$E\left[\bar{Y}\right] = E\left[\sum_{i=1}^{n} Y_i \Big/ n\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[Y_i\right] = \frac{1}{n}\sum_{i=1}^{n} \psi = \frac{1}{n} \times n\psi = \psi$$

■

Our estimator is unbiased! This means that if we repeatedly took samples from our population, the mean of the sample means would be equal to the population mean.

Note that this theorem does not state that the observations have to come from a normal distribution. It is only required that they be an independent random sample from the same population.

In the case of the normal, we know that the population mean is equal to ψ from the distribution function.

Another desirable property for an estimator is that it be minimum-variance. That is, of all possible estimators, the one used gives the smallest variance of the estimates obtained from repeated samples. We will not determine if our estimator is minimum variance. Another question that will be left at this point is that of the sampling distribution of the sample mean which would be required to go on with the inference process. We are essentially stopping at the point where we have a point estimator, rather than going on to consider confidence intervals and hypothesis tests. However, we have determined that the mean of the sampling distribution is the population mean and that it would be desirable for the variance of the sampling distribution to be as small as possible.

What about estimating σ^2? Natural estimators are:

$$S_n^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n} \text{ and } S_{n-1}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}.$$

Which one should be used?

Theorem I.8: The expected value of S_n^2 calculated from a random sample of n observations from a population with probability distribution function $f(y)$ is given by $E\left[S_n^2\right] = \frac{n-1}{n}\sigma^2$.

Proof: not given. ■

The estimator S_n^2 is biased.

Theorem I.9: The expected value of S_{n-1}^2 calculated from a random sample of n observations from a population with probability distribution function $f(y)$ is given by $E\left[S_{n-1}^2\right] = \sigma^2$.

Proof: Note that $S_{n-1}^2 = \frac{n}{n-1} S_n^2$, that $n/n-1$ is a constant and that S_n^2 is a function of $Y_1, Y_2, \ldots, Y_n$. Consequently, we can use theorem I.2 as follows:

$$E\left[S_{n-1}^2\right] = E\left[\frac{n}{n-1} S_n^2\right] = \frac{n}{n-1} E\left[S_n^2\right] = \frac{n}{n-1} \times \frac{n-1}{n} \sigma^2 = \sigma^2$$

■

That is, S_{n-1}^2 is an unbiased estimator of σ^2 and so is the preferred estimator.

I.C Summary

In this chapter, we have

- looked at the definition and meaning of the expectation and variance of a random variable;
- investigated the expectation of a linear function of a random variable;
- derived the probability distribution function to be used to describe a random sample;
- examined the expectation of a linear combination of functions of random variables;
- discussed the properties of estimators: unbiasedness and minimum variance, particularly for the mean and variance.