BiostatIntro2008

Biostatistics for Genetics and Genomics

Birmingham AL

July 2008

W. J. Ewens

These notes are quite extensive. They will be used as the basis for the "Introduction to Biostatistics" lecture. The main points discussed in the lecture are available on the slides for the lecture. The actual material to be discussed in the lecture will depend on the preferences of the class members.

1 Probability and Statistics

1.1 Introduction: What is "Statistics"?

Statistics is the science of making inferences, or inductions, about reality based on data in whose generation **chance** has played some part. This explains why statistics is important in biology, medicine and, in particular, in genetics and genomics. In these areas many chance mechanisms are at work. In genetics one of two genes, essentially chosen at random, is transmitted from a parent to any offspring. Crossing over (recombination) is a random chance phenomenon that affects the genome of any individual (at the time of conception). Perhaps more important, chance mechanisms arise through the **sampling** procedure. The data we use to make statistical inferences are usually derived from a random sample of individuals. A different sample would yield different data and perhaps even lead us to a different conclusion. This means that a study of **probability theory** is necessary for an understanding of the concepts and processes of statistics.

Logically, then, a study of probability theory should come **before** a study of statistics. However, since statistical procedures provide the motivation for the probability theory considered, it is natural to start with a statistical question and then go back to the probability theory necessary to handle it.

1.2 An example

Suppose we have a drug that we know, from long experience, cures a patient with some specific illness in 70 % of cases. A new drug is proposed as having a higher cure rate than the present one. We now give an example of the result of a drug trial aimed at assessing this claim.

Suppose that, in a drug trial, the new drug is given to 1,000 people suffering from the illness, and that of these, 741 are cured. Do we have significant evidence that this new drug is better than the current one?

The key thing to note now is the following. **We have no way of**

answering this question objectively unless we first answer the question: if the new drug is equally effective as the current one, how likely is it that, by chance, 741 or more people given the new drug will be cured?

This question can only answered by a probability theory calculation, which shows that if the cure probability of the new drug is 0.7, the probability that it will cure 741 individuals out of 1000 is about 0.0023. The fact that this probability is quite small might make us feel that we have good evidence that the new drug truly is better than the current one, since if in fact it is equally effective as the current one, something quite unlikely occurred in our drug trial.

If the above probability had been (say) .26 instead of .0023, we might feel that we do *not* have sufficient evidence that the new drug is superior to the current one. Thus the numerical value of the calculated probability is crucial to our decision.

Randomness came into this example through the sampling procedure. If a different sample of 1,000 people had been given the new drug, a different number of people cured would almost certainly have arisen. Also, how a person responds to a drug today might differ, for random reasons unknown to us, from how he/she might respond in six months' time.

We now look more carefully at the relation between probability and statistics.

1.3 The relation between probability and statistics

The conclusion reached in the example in the preceding section is a *statistical inference*. That is, using some observed data and the appropriate corresponding probability calculation, we reached some conclusion about the proposed new drug.

It is important to note the "direction" of this inference, and to observe that it goes in the other direction from a probability calculation. This is illustrated in the statements below. The key thing to note is that the statistical inference made depends entirely on the

result (the values 0.0023 in the first example above) of a probability calculation.

The probability calculation. Under the assumption that the new drug has the same cure rate(0.7) as the currently used drug, the probability that 741 or more out of 1000 individuals given the new drug will be cured is quite small (.0023).

The statistical inference, When the new drug was given to 1000 individuals, 741 were cured. Based on the above probability calculation, it is reasonable to conclude that the new drug has a higher cure rate than the current drug.

We shall refer to this drug trial several times later, in the Statistics part of these notes, both with respect to estimation and with respect to hypothesis testing.

The drug trial example illustrates the general principle that *every* statistical inference that we ever carry out depends on some probability calculation, and that no valid statistical inference can be made without first carrying out the probability calculation appropriate to that inference. Later we will consider some statistical inference procedures that are so complicated that we will have to take on trust the probability theory behind the inference made. Nevertheless, in an advanced statistics class one would learn what this complicated probability theory is.

Thus a study of probability theory, no matter how brief, is essential for understanding statistics. This is why we start with an introduction to some areas of probability theory.

2 Probability theory: discrete random variables

2.1 Probability distributions and parameters

We start with the concepts of random variables and parameters.

Random variables are of two types, *discrete* and *continuous*. A discrete random variable usually comes from counting and thus, usually, takes one or other of the values 0, 1, 2, ... The example above illustrates this: before the experiment is started, the number of patients who will cured by a drug being tested in a drug trial is a random variable. We do not know before the experiment what value it will take, but we know it must be one of the numbers 0, 1, 2, ..., 1000. That is, this number is a discrete random variable.

By contrast, a continuous random variable usually comes from a measurement, for example the (random) reduction in blood pressure that a patient will have after taking some drug. This reduction in blood pressure can take any value in some range of values. It is a random variable, since before the patient takes the drug, it is not known what the reduction in blood pressure will be.

The theory for discrete random variables is simpler than that for continuous random variables, so we start with the definition, and the associated theory, of a discrete random variable.

For our purposes, a good definition of a discrete random variable is the following. A *discrete random variable* is a numerical quantity that in some future experiment that involves some degree of randomness will take one value from some discrete set of possible values. An example is given above: the number of patients who will be cured is, before the drug trial, a random variable that will take one of the values $0, 1, 2, 3, \ldots, 1,000$.

By convention, random variables are written as upper case symbols, for example X. The *probability distribution* of a discrete random variable X is the set of values that this random variable can take, together with their associated probabilities. Probabilities are numbers between zero and one inclusive that always add to one when summed over all possible values of the random variable.

The probability distribution is often presented in the form of a table, listing the possible values that the random variable can take together with the probabilities of each value. If the respective possible values of a random variable X are v_1, v_2, ... , v_k, and these values have respective probabilities p_1, p_2, ... , p_k, then the generic form of this table is as follows:-

Possible values of X	v_1	v_2	...	v_k
Associated probabilities	p_1	p_2	...	p_k

(1)

For example, if we plan to toss a fair coin twice, and the random variable X is the number of heads that eventually turn up, the probability distribution of this random variable can be presented as follows:

Possible values of X	0	1	2
Associated probabilities	.25	.50	.25

(2)

We show later how the probabilities in (2) are calculated.

In practice, the probabilities p_1, p_2, ... , p_k associated with the possible values of the random variable of interest are often unknown. The "thumbtack" example below illustrates this.

There is another frequently used method of presenting a probability distribution. This is by using a diagram, in which the possible values of the random variable are indicated on the horizontal axis, and their corresponding probabilities by the heights of vertical bars above each respective possible value (see Figure 1 below). This mode of presentation of a probability distribution is feasible only if the probabilities p_1, p_2, ... , p_k are known.

A third method of describing a probability distribution is by a formula. Although this is the most complex way of describing a probability distribution, it is the appropriate method for any theoretical work. An example is given in (4) below.

An important point about probability distributions is that one may not use any probability distribution just because one feels like it. For one given situation one probability distribution is appropriate

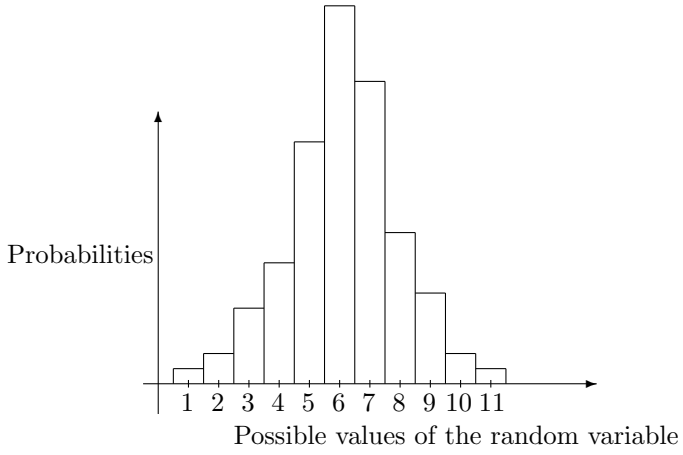

Figure 1:

and another situation another is appropriate. We discuss this further below.

A probability distribution will often contain a *parameter*. For us, a parameter is an unknown number. Almost all of statistics consists of making inferences about parameters.

As an example, when a thumbtack is thrown in the air, it will land either "point up" or "point down". Because of the physical nature of a thumbtack, the probability that it will land "point up" is unknown to us. Thus this probability is a parameter. We follow standard statistical notation and denote this probability by a Greek letter, say θ. Suppose that we plan to toss this thumbtack twice, and focus on the number of times (0, 1 or 2) that will land "point up". This number is random variable, since we do not know in advance what value it will take, so we denote it by X. The probability distribution of X can be shown to be as follows:

Possible values of X	0	1	2
Associated probabilities	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2

(3)

Even though we do not know the numerical value of θ, so that we do not know the various probabilities in (3), we still think of (3) as

a probability distribution.

A more important example, having the same mathematical properties as the thumbtack example, arises in genetics through the phenomenon of recombination (crossing-over). There is some probability that a recombination event occurs between two gene loci when genetic material is passed on from parent to child. This probability depends on the physical distance of these loci on a chromosome. If (as is usual) we do not know this distance, we do not know the probability of a recombination event, so we denote it by θ. The probabilities in (3) would then give the respective probabilities of 0, 1 and 2 recombination events in two transmissions of genetic material.

2.2 The binomial distribution

The drug trial example, the thumbtack example, and the recombination example, are all cases where the *binomial* distribution arises. It is without doubt the most frequently used discrete distribution in biostatistics. Thus we discuss this distribution in detail, doing so in abstract terms.

The binomial distribution arises if all four of the following requirements hold.

(i) The number n of trials is fixed in advance, and does not depend on the outcomes of the trials as they proceed.

(ii) Each trial results in one of two outcomes, which we conventionally call "success" and "failure". (In the thumbtack example, we could call "point up" a success and "point down" a failure.)

(iii) The various trials must be independent.

(iv) The probability of success must be the same on all trials.

If the probability of success on any trial is unknown, then it is denoted in these notes by θ. We call θ the parameter, and n the index, of the binomial distribution. The total number of successes in the n trials is a binomial random variable, which we denote by X. The probability that X assumes some specific value v is given

by the formula

$$\binom{n}{v}\theta^v(1-\theta)^{n-v}, \quad v = 0, 1, 2, \ldots, n. \tag{4}$$

We do not prove this result here. The notation $\binom{n}{v}$ is sometimes spoken as "n choose v": it is the number of different orderings in which the v successes can arise in the n trials. The probability distribution in (3) is the special case of the binomial distribution (4) when $n = 2$. The probability distribution in (2) is an even more special case of the binomial distribution, applying when $n = 2$ and $\theta = 1/2$.

As stated also in connection with the probability distribution (3), the quantity θ in (4) is often an unknown to us: that is, it is a *parameter.* One of the aims of statistics is to *estimate* this parameter, and another is to *test hypotheses* about this parameter. For both these operations we have to use data, that is the observed number of successes once our experiment is conducted. More on this later.

2.3 Other discrete distributions

There are *many* other discrete probability distributions besides the binomial. For example, if we roll a die, there are six possible outcomes on each roll (trial), not just two (as in the binomial). Next, there are many cases where one of the assumptions in the binomial distribution (that the probability of "success" is the same on each trial) does not hold. In some cases the outcome of the successive trials are not independent.

These comments are made to emphasize the point made above, that any given probability distribution applies only in some specific well-defined situation. Although we do not discuss any discrete distribution in these notes other than the binomial, there are many books which indicate which distribution is appropriate for any given research circumstance.

2.4 The mean of a discrete random variable

The mean of a random variable is often confused with the concept of an average, and it is important to keep the distinction between the two concepts clear. The mean of a discrete random variable whose probability distribution is as given in (1) is defined as

$$\text{mean} \;=\; v_1 p_1 + v_2 p_2 + \cdots + v_k p_k. \tag{5}$$

It can be shown that the mean is at the center of gravity, or the knife-edge balance point, of a probability distribution. There are four important points to make regarding the mean of a discrete random variable.

(i) If a probability distribution involves an unknown parameter, then the mean of this distribution will usually involve this parameter, and thus will also be unknown to us. For example, the mean of a random variable having the binomial distribution (4) can be shown to be $n\theta$. If we do not know the numerical value of θ, we do not know the numerical value of this mean.

(ii) The notation μ is often used for a mean. This is because the mean of a discrete random variable is often unknown to us, that is it is a parameter, and is thus denoted by a Greek letter. Of all the estimation and hypothesis testing procedures in Statistics, estimation of, and testing hypotheses about, an unknown mean are among the most important.

(iii) An alternative name for the mean of a random variable is the "expected value" of that random variable, and this leads to a second frequently used notation, namely $E(X)$, for the expected value, or mean, of the random variable X.

(iv) The word "average" is *not* an alternative for the word "mean", and has a quite different interpretation from that of "mean." This distinction will be discussed in detail later.

2.5 The variance of a discrete random variable

A quantity of importance equal to that of the mean of a random variable is its *variance*. The variance (denoted by σ^2) of the discrete

random variable X is defined, using the notation of the probability distribution (1), as

$$\text{variance} \;=\; (v_1 - \mu)^2 p_1 + (v_2 - \mu)^2 p_2 + \cdots + (v_k - \mu)^2 p_k, \qquad (6)$$

where μ is the mean of the distribution, defined in (5). In informal terms, the variance of a random variable measures the "spread-outness" of its probability distribution around, or relative to, the mean of the distribution. Thus the variance of the random variable having the right-hand side probability distribution in Figure 2 is larger than the variance of the random variable having the left-hand side probability distribution in Figure 2.

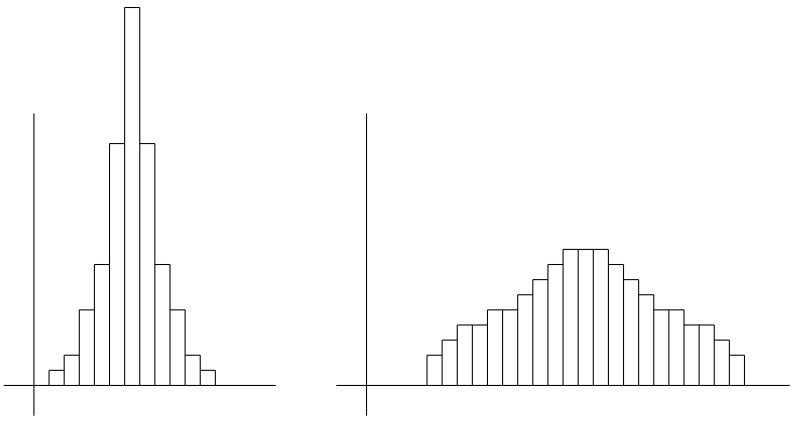

Figure 2:

There are four important points to note concerning the variance of a discrete random variable are as follows.

(i) The variance, like the mean, is often unknown to us. It is therefore denoted by a Greek letter (σ^2).

(ii) A quantity that is often more useful than the variance of a probability distribution is the *standard deviation* of that random variable. This is defined as the positive square root of the variance, and (naturally enough) is denoted by σ.

(iii) The probability distribution of a random variable with a small variance is closely concentrated around the mean. Thus the ob-

served value of a random variable with a small variance should be quite close to the mean. We use this idea in statistics when *estimating* a mean, as discussed below.

(iv) The variance of the binomial distribution (4) is $n\theta(1-\theta)$. This implies, for example, that if a fair coin ($\theta = 1/2$) is tossed $n = 10,000$ times, the variance of the number of heads that will appear is $10,000 \times \frac{1}{2} \times \frac{1}{2} = 2,500$. From note (i) above, the standard deviation of the number of heads is thus 50.

2.6 The two standard deviation rule

The two standard deviation rule is a rough rule of thumb which states that it is about 95% likely that a random variable will lie within two standard deviations its mean. As an example, suppose that a fair coin is to be tossed 10,000 times. Since the coin is fair, $\theta = 1/2$ and from note (i) of Section 2.4, the mean number of heads to appear is 5,000. From note (iv) in Section 2.5 the standard deviation of the number of heads to appear is 50. Thus it is about 95% likely, if a fair coin is to be tossed 10,000 times, that the number of heads to appear will be between 4,900 and 5,100. This rule can be used, as we see below, to assess the accuracy of the estimate of a parameter.

2.7 The proportion of successes

In the binomial context the *proportion* of successes (rather than the number of successes) is often of main interest. This proportion is also a *random variable*: before we do our experiment we do not know what value it will take. We therefore denote this by an upper case symbol, namely P. The properties of P are needed for statistical purposes, as discussed below. The two most important properties of P are that its the mean is θ and its variance is $\theta(1-\theta)/n$. We will need these properties in Section 6.3.2 and also in Section 7.1.

3 Probability theory: continuous random variables

3.1 Introduction

Some random variables, by their nature, are discrete, such as the number of individuals cured in the drug trial described above. Other random variables, by contrast, are continuous. Measurements such as height, weight and blood pressure are all of this type. A continuous random variable can take any value in some continuous range of values.

Probabilities for continuous random variables are not allocated to specific values, but rather are allocated to ranges of values. Every continuous random variable has an associated *density function* $f(x)$, and the probability that the random variable takes a value in some given range of values is obtained by integrating the density function over that range. This integration leads to an area under the density function curve. An example is shown in Figure 3. We do not pursue the (calculus) details of this any further here.

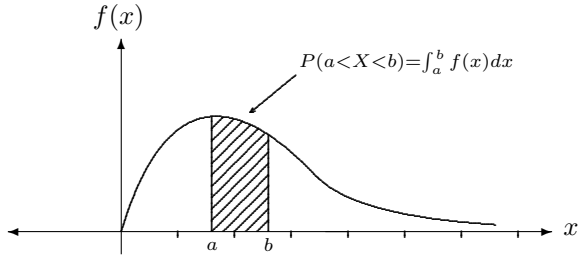

Figure 3: $P(a < X < b) = \int_a^b f(x)dx$.

3.2 The mean and variance of a continuous random variable

The mean μ and variance σ^2 of a continuous random variable have formal mathematical definitions parallel to those for a discrete random variable. Because these definitions involve calculus operations, they are not described here. The main idea is that, as with a discrete random variable, the mean is at the "center of gravity", or the

"knife-edge balance point," of the density function, and the variance is a measure of the "spread-out-ness" of the density function around the mean. In other words, the mean and variance have the same general interpretations for a continuous random variable as they do for a discrete random variable.

3.3 The normal distribution

By far the most important continuous random variable is one having the *normal*, or *Gaussian*, distribution. The (continuous) random variable X has a normal distribution if its density function is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty. \tag{7}$$

This formula incorporates the mean μ and the variance σ^2 into the mathematical form of the normal distribution.

Strictly speaking there is a family of normal distributions, each member of the family corresponding to some specific (μ, σ^2) combination. A particularly important member of this family is the normal distribution for which $\mu = 0$ and $\sigma^2 = 1$, whose density function is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < +\infty. \tag{8}$$

This density function is graphed in Figure 4. This is sometimes called the *standard normal* distribution. Published tables and charts for the normal distribution always refer to this specific normal distribution.

It can be shown that the probability that a random variable having the standard normal distribution takes a value between -2 and $+2$ is approximately 0.95. The values ± 2 are the basis of the "two standard deviation rule", mentioned above, and will arise again several times below. A more accurate calculation shows that the probability that a random variable having the standard normal distribution takes a value between -1.96 and $+1.96$ is almost exactly 0.95. The more accurate values ± 1.96 will also arise several times below.

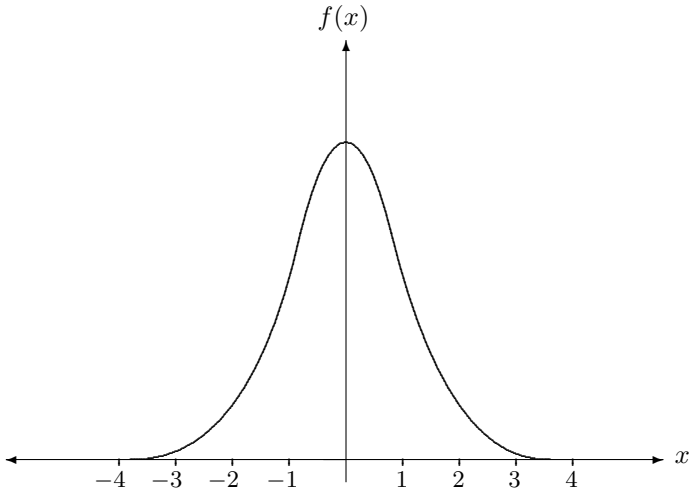

Figure 4: The density function for the standard normal distribution with $\mu = 0$, $\sigma = 1$.

4 Probability theory for many random variables

4.1 Introduction

In almost every application of statistical methods we deal with the analysis of many observations. For example, if we wish to estimate the mean blood pressure of individuals having some disease, we would measure the blood pressure of many individuals with this disease. Thus before our experiment we are concerned with many random variables, that is with the various reductions in blood pressures of the individuals planned to be in the experiment.

We denote the number of individuals in the experiment by n. Before the experiment is conducted, the blood pressures of these individuals are unknown to use, that is they are random variables. We therefore denote them (before the experiment), using upper case letters, by $X_1, X_2, \ldots, X_n$. That is, before the experiment we denote blood pressure of the first person in the experiment as X_1, of the second person as X_2, and of the last (n^{th}) person as X_n. We will assume that these random variables are *independent* and that they all have the same probability distribution. We denote the mean of this dis-

tribution by μ and the variance of this distribution by σ^2. Since we do not know the exact form of this probability distribution – if we did, why would we even do this experiment? – we do not know the values of this mean and variance.

4.2 Averages and their properties

It would be natural, in the blood pressure example, once the experiment is finished, to use the average of the observed blood pressure readings in the n individuals in the experiment and to *estimate* the mean blood pressure by this average. To find out properties of this estimate, we have to consider the situation before the experiment, when we are dealing with random variables, and in particular to examine the properties of the average of the random variables $X_1, X_2, \ldots, X_n$. This average is itself a *random variable*, which we write as $\bar{X}$, defined by

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{9}$$

The two key results that we will need for statistical operations is that the mean of $\bar{X}$ is

$$\text{mean of } \bar{X} = \mu \tag{10}$$

and the variance of $\bar{X}$ is

$$\text{variance of } \bar{X} = \frac{\sigma^2}{n}. \tag{11}$$

5 Regression

An important area of science is that where we attempt to find the way in which one quantity (for example the growth height of a plant in a greenhouse) depends on another (for example the amount of water given to the plant during its early growth period). In this example the growth height is a random variable, since we do not know in advance of the experiment what value it will take. On the other hand the amount of water is non-random, since we can choose this at any level that we like. We therefore write the growth height in upper case (as Y) and the amount of water in lower case (as x).

The simplest model is that the mean of Y is a linear function of x. The constants in this linear function are unknown to us, that is they are parameters, so we write

$$\text{mean of } Y = \alpha + \beta x. \tag{12}$$

This is sometimes called "the regression model of Y on x". Two important statistical operations for this model are to estimate the parameters α and β and to test hypotheses about these parameters. These procedures are discussed later in these notes. To carry out these operations it is necessary to use at least 5 or 10 plants in the experiment. Thus we consider, before the experiment, the random heights $Y_1, Y_2, \ldots, Y_n$ of n plants, as well as the respective amounts of water $x_1, x_2, \ldots, x_n$ that we propose to give to the plants, and write equation (12) in more detail as

$$\text{mean of } Y_i = \alpha + \beta x_i,\ i = 1, 2, \ldots, n. \tag{13}$$

There are several strategic aspects about the choice of $x_1, x_2, \ldots, x_n$. One clearly is that we would not choose these all to be the same, since if we were to do that it would be impossible to assess how the amount of water affects the growth height. One reasonable procedure is to choose $x_1, x_2, \ldots, x_n$ in an arithmetic progression, for example with $x_1 = 1, x_2 = 2, \ldots, x_n = n$.

Since each Y_i is a random variable, it has a variance (as well as a mean). This variance is of course unknown to us, and the simplest regression model assumes that $Y_1, Y_2, \ldots, Y_n$ all have the same unknown variance σ^2. This model is considered later in these notes, but it should be noted that more complicated regression models than this are often necessary, and can be found in the statistical literature.

6 Statistics

6.1 Introduction

So far we have considered probability theory, an activity relevant to the time *before* we do some experiment, when we do not know what the outcome of the experiment will be. We now change direction 180 degrees, and consider the situation *after* the experiment has been

completed. This experiment will yield data, and Statistics is the operation of using these data to make various inferences about some unknown parameter. The data in any experiment are the observed values of random variables, so these statistical inferences procedure must be based on some appropriate probability calculation.

There are two main areas of statistical inference, namely estimation of parameters and testing hypotheses about parameters. In these notes we consider only three cases: estimation of, and testing hypotheses about, a binomial parameter θ, estimation of, and testing hypotheses about, the mean μ of some probability distribution, and estimation of, and testing hypotheses about, the parameters α and β in the regression model (13). This is done, first, because these are important examples, and second because further examples become quite complicated.

6.2 Notation

In these notes upper case notation is used for random variables. For example, in the binomial case the (random) number of successes was denoted by X, and the (random) proportion of successes by P. In the blood pressure example, the (random) blood pressures of the individuals to be in the experiment were denoted by $(X_1, X_2, \ldots, X_n)$. The convention is that after the experiment is completed, we use the corresponding lower case notation for the actually observed values for our these quantities. Thus in the binomial case we denote the actual observed number of successes once the relevant experiment is completed by x, and the observed proportion of successes by p. In the blood pressure example we denote the observed blood pressures of the n individuals in the experiment after it is carried out by $(x_1, x_2, \ldots, x_n)$. In the regression model (13) we denote the eventual growth heights of the plants, once the experiment is concluded, by $y_1, y_2, \ldots, y_n$.

6.3 Estimation of parameters

6.3.1 Introduction: unbiasedness, precision and the 95% confidence interval

We have two aims in estimating any parameter. First, we want to estimate it in an "unbiased" way (this term will be defined shortly). Second, we want to have some idea of how precise our estimate is. The concept of the mean of a random variable is used to assess unbiasedness, and the concept of the variance of a random variable and the two standard deviation rule of Section 2.6 are used to address the concept of precision, using the so-called 95% confidence interval. We illustrate these ideas concretely below in estimating a binomial parameter and in estimation the mean of a probability distribution. In doing this we will introduce the important terms *estimator* and *estimate*.

6.3.2 Estimation of a binomial parameter θ

The estimation of a binomial parameter θ is carried out by using the theory of the binomial distribution, and relies on the theory described (briefly) in Section 2.7. How is this theory used in practice?

It is natural to estimate the parameter θ by the observed proportion p of successes. We call this quantity, which in any concrete case is a number derived from the data, the *estimate* of θ. What are the properties of this estimate? The theory of Section 2.7 states that the mean of the (random) proportion P of successes before the experiment is carried out is θ. Thus P is said to be an *unbiased estimator* of θ, and correspondingly p is said to be an *unbiased estimate* of θ. In loose terms, in using an unbiased estimate we are "shooting at the desired target".

The precision of this estimate is determined by the variance of the estimator P. It was shown in Section 2.7 that this variance is $\theta(1-\theta)/n$, where n is the sample size. The standard deviation of P is thus $\sqrt{\theta(1-\theta)/n}$. The two standard deviation rule states that, to a close approximation, the probability that P is within $2\sqrt{\theta(1-\theta)/n}$ of θ is about 95%. We invert this statement to say that the probability that θ is within $2\sqrt{\theta(1-\theta)/n}$ of p is about 95%. Since θ is unknown, we

estimate the standard deviation $\sqrt{\theta(1-\theta)/n}$ of P by $\sqrt{p(1-p)/n}$, and then say that, to a sufficient approximation,

$$\text{Probability}(p - 2\sqrt{p(1-p)/n} < \theta < p + 2\sqrt{p(1-p)/n}) \approx 0.95. \tag{14}$$

This gives us some idea of the precision of the estimate of θ. The range of values from $p-2\sqrt{p(1-p)/n}$ to $p+2\sqrt{p(1-p)/n}$ is called a 95% confidence interval for θ.

Example. In the drug trial referred to above, 741 out of 1000 individuals given the drug were cured. We then *estimate* θ, the unknown probability that a person will be cured by this drug by $p = 741/1000 = 0.741$. Second, equation (14) gives an approximate 95% confidence interval for θ as

$$.741 - \sqrt{.741 \times .259/1000} \text{ to } .741 + \sqrt{.741 \times .259/1000},$$

which evaluates to the interval .713 to .769. In research papers this result is often written, not too correctly, as

$$\theta = .741 \pm .028.$$

6.3.3 Estimation of a mean μ

Suppose that we wish to estimate μ, the mean blood pressure in the example considered earlier. Equation (10) shows the mean of $\bar{X}$ is μ, so that $\bar{X}$ is an *unbiased estimator* of μ. Because of this, the average $\bar{x}$, defined by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{15}$$

is an *unbiased estimate* of μ. In loose terms, in using $\bar{x}$ to estimate μ we are "shooting at the desired target".

To find out the properties of this estimate, it is necessary to consider the properties of the corresponding estimator $\bar{X}$. First, since the mean value of $\bar{X}$ is μ (from equation (10)), $\bar{x}$ is an unbiased estimate of μ. Second, we have to consider the precision of this estimate. Again we have to consider properties of $\bar{X}$, in this case its variance, namely σ^2/n. The two standard deviation rule states that

$$\text{Probability } (\mu - \frac{2\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{2\sigma}{\sqrt{n}} \approx 0.95. \tag{16}$$

If we knew σ^2, an approximate 95% confidence interval for μ would be found from this by an inversion procedure similar to that described above for estimating a binomial parameter, yielding

$$\bar{x} - \frac{2\sigma}{\sqrt{n}} \text{ to } \bar{x} + \frac{2\sigma}{\sqrt{n}}. \tag{17}$$

Unfortunately, we almost always do not know σ^2, and we have to estimate it from the data. It turns out that the best estimate is s^2, defined by

$$s^2 = \frac{x_1^2 + x_2^2 + \ldots x_n^2 - n\bar{x}^2}{n-1}. \tag{18}$$

From this, we get an (even more) approximate 95% confidence interval for μ as

$$\bar{x} - \frac{2s}{\sqrt{n}} \text{ to } \bar{x} + \frac{2s}{\sqrt{n}}. \tag{19}$$

In these expressions, s is the square root of s^2.

We now have a method for not only estimating the mean (in this case, of blood pressure) from the data, but also for using the data to give a 95% confidence interval for the mean.

6.3.4 Estimation of the regression parameter β

Of the three parameters α, β and σ^2 in the regression model (13), the most important is β. This is because β estimates the extra growth height per unit increase in the amount of water given to a plant. The estimate b of β is given by the formula

$$b = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{20}$$

This can be shown to be an unbiased estimate, since the corresponding estimator B has mean β. The variance of B is

$$\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \tag{21}$$

and although σ^2 is unknown, it can be estimated from the data. We denote the estimate of σ^2 by s_r^2, (the suffix "r" indicates "regression", to emphasize that the formula for s_r^2 differs from that in (18)).

We do not give the details of this formula since it is quite complicated. From this a 95% confidence interval for β can be found: this confidence is

$$b - \frac{2s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \text{ to } b + \frac{2s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

The form of the expression in (21) shows that the precision of the estimate of β depends on the values $x_1, x_2, \ldots, x_n$. If the x_i values are chosen all to be equal, the denominator in (21) is zero, so that the variance of B is infinite. This implies that the experiment cannot provide any information about β, and confirms the common-sense statement made earlier.

6.3.5 Final comment

While the estimation of a binomial probability θ and the mean μ of a distribution are examples of "natural" estimators, in more complicated cases there might not be an obvious "natural" estimator of a parameter - the estimation of a variance, and (perhaps) the estimation of the parameter β in a regression model, are examples of this. Thus we need a more systematic approach to finding estimators of parameters. Fortunately there is general statistical procedure which does this – the so-called "maximum likelihood" theory. However this theory is complicated, so we do not go into the details of it in these notes, other than noting that all the estimates given above are derived from this theory.

6.4 Testing hypotheses

6.4.1 General principles

Classical statistical hypothesis testing involves the test of a *null hypothesis* against an *alternative hypothesis*. The procedure consists of five steps, the first four of which are completed before the data to be used for the test are gathered, and they all relate to probability calculations. They are used to set up the statistical inference process taken in the final step. These steps are as follows, and are illustrated by the drug example discussed above and also by a "coin" example.

Step 1 is to declare the null hypothesis (usually written as H_0) and the alternative hypothesis (usually written as H_1).

Example 1. We consider again the drug test referred to earlier. We wish to test whether a proposed new drug for curing some disease has a higher cure rate than the drug that we are currently using. Suppose that we have been using the current drug for so long that we can say with acceptable accuracy that the probability that a person using this drug is cured is 0.7.

We write θ for the (unknown) probability that a person using the new drug gives will be cured. Since the drug is a new one, θ is a *parameter*, that is, an unknown constant. The hypothesis-testing procedure is a test about the value of θ. The null hypothesis H_0 is that the new drug has the same probability of cure as the current drug, that is that $\theta = 0.7$. This is typical of a the null hypothesis, which can be thought of as claiming that nothing interesting is happening, that is, in this example, that the new drug is no better than the currently used one. The alternative hypothesis H_1 of practical interest in this example is that the new drug is *better than* the currently used one. That is, the alternative hypothesis claims that $\theta > 0.7$. Because of the nature of the alternative hypothesis, we call it this a *one-sided up* test.

Example 2. Suppose that we wish to test whether a certain coin is fair. If it is unbiased we have no idea whether the bias would be towards heads or tails. Here the null hypothesis H_0 is that the coin is fair ("nothing interesting happening"), and the alternative hypothesis H_1 is that the coin is biased one way or the other. This implies that the alternative hypothesis is *two-sided* and that our eventual test will be *two-sided*. In algebraic terms, if θ is the probability of head on the coin, the alterative hypothesis is $\theta \neq 0.5$.

There are two important things to say about the choice of null and alternative hypotheses.

(i) The choice of null and alternative hypotheses should be made before the data are seen. To decide on a hypothesis as a result of the data is to introduce a bias into the procedure, invalidating any

conclusion that might be drawn from it.

(ii) The choice of the alternative hypothesis is decided from the context. In the drug example we are interested *only* in the possibility that the new drug is better than the current one. This leads to the alternative hypothesis that the probability of a cure using the new drug exceeds 0.7. This implies that in the drug example the eventual test that we carry out is one-sided. Only probabilities for the new drug exceeding 0.7 are of interest. In the coin example, since we has no *a priori* view in which way the coin might be biased, it is natural to carry out a two-sided test.

Step 2. Since the decision as to whether H_0 or H_1 is accepted will be made on the basis of data derived from some random process, it is possible that an incorrect decision will be made. That is, it is possible that we will reject H_0 when it is true (a *Type I error*, or a false positive) and it is also possible that we do not reject H_0 when in fact it is false (a *Type II error*, or a false negative). In the drug case a Type I error (false positive) would arise if we claimed that the new drug is better than the current one when in fact it is not. A Type II error (false negative) would arise if we did not claim that the new drug was better than the current one when in fact it is.

When testing a simple null hypothesis against a simple alternative it is not possible, with a fixed predetermined sample size, to ensure that the probabilities of making a Type I error and a Type II error are both arbitrarily small. This difficulty is resolved in practice by observing that there is often an asymmetry in the implications of making the two types of error. In the drug example, there might be more concern about making the false positive claim (that the new drug is better than the current one when it is not), and less concern about making the false negative conclusion (that it is not better than the current one when in fact it is). For this reason a procedure frequently adopted is to fix the numerical value of the Type I error at some acceptably low level (usually 1% or 5%), and not to attempt to control the numerical value of the Type II error.

Step 2 of the hypothesis testing procedure consists in choosing the numerical value for the Type I error, usually 5% or 1%.

Step 3. The next step consists in determining a *test statistic.* This is the quantity calculated from the data whose numerical value leads to acceptance or rejection of the null hypothesis. In the drug example a reasonable test statistic is the total number of individuals cured in the trial of the new drug. (It is equivalent to use the proportion of individuals cured in this trial, provided that appropriate changes are made to the calculations given below.) In the coin example example a reasonable test statistic is the total number of heads (or, equivalently, the proportion of tosses giving heads) that appear when the coin is tossed.

In tests more complicated than the drug or coin examples the choice of test statistic is not at all obvious in advance. Examples of these include the test statistic used in a more complicated drug trial that uses a placebo, the "t" statistic and the chi-square statistic. These are all be considered later.

Step 4. The next step consists in determining which observed values of the test statistic will lead to rejection of H_0. This choice is made so as to ensure that the test has the numerical value for the Type I error chosen in Step 2.

In the drug example, the total number of individuals who will be cured is not known before the trial of the new drug begins. This number is thus a random variable, so we denote it by X. The null hypothesis $\theta = 0.7$ will be rejected in favor of the alternative if the observed value x of X, once the trial is conducted, is sufficiently large, that is, if x is greater than or equal to some *significance point* k. If for example the Type I error is chosen as 1%, k is found from the requirement that the probability that the null hypothesis is rejected if it is true be 0.01. In algebraic terms, this requirement can be written

$$\text{Prob}(X \geq k \,\text{when}\, \theta = 0.7) = 0.01. \tag{22}$$

Given the sample size n, the value of k can be found in the binomial distribution. (Computer packages are now available to do this calculation, and show that $k = 734$.)

In the coin case the null hypothesis would be rejected if x, the observed number of heads, is sufficiently small *or* is sufficiently large. The calculation of "how small" and "how large" can also be carried out using the binomial distribution.

Step 5. Steps 1 - 4 above are all carried out before the relevant experiment (the drug trial, the coin tossing) is carried out. They are all concerned with probability concepts and calculations. The final step in the testing procedure is the only statistical one. It consists of obtaining the data, determining whether the observed value of the test statistic is equal to or more extreme than the relevant significance point (for example the value 734 for the drug trial) calculated in Step 4, and to reject the null hypothesis if it is. This statistical step is easy to carry out once the first four steps have been completed.

6.4.2 *P*-Values

A testing procedure equivalent to that just described involves the calculation of a so-called *P-value*. Here Step 4 in the above sequence, the calculation of a significance point k, is not carried out. Instead, once the data are obtained, we calculate the null hypothesis probability of obtaining the observed value of the test statistic or one more extreme in the direction indicated by the alternative hypothesis. This probability is called the P-value. If the P-value is *less than* the chosen Type I error, the null hypothesis is rejected. Otherwise the null hypothesis is not rejected. This procedure always leads to a conclusion identical to that based on the significance point approach.

For example, suppose as above that in the drug trial we give the new drug to 1000 individuals and that of these, 741 are cured. The P-value associated with the observed number 741 is the probability that a random variable having a binomial distribution with $\theta = 0.7, n = 1000$ takes a value 741 or more. This P-value can be found by using the binomial distribution, and is about 0.0023. Since this is less than the chosen Type I error of 0.01, we reject the null hypothesis (in agreement with the conclusion reached using the significance point $k = 734$).

6.5 Summary and the "cookbook" approach

In this section we summarize the drug trial procedure described in the previous section, and then describe the "cookbook" approach to the procedure. This is done because in all further examples in these notes only the "cookbook" procedure is described, since the background probability theory is quite complicated in those procedures.

In *Step 1* we set up the null hypothesis, namely that the new drug has the same cure rate as the currently used drug, namely 0.7. We also set up the natural alternative for this situation, that the new drug has a higher cure rate than the current one. We recognize that we might reject the null hypothesis even though it is true, and in *Step 2* we designate the probability that we are willing to accept for making this error (usually chosen as either 1% or 5%). In *Step 3* we decide on the "test statistic" that we will calculate from our data in order to decide between the null and alternative hypotheses. In the drug example the test statistic is simply the number of individuals cured in the trial of the new drug. In more complicated experiments complicated probability theory is needed to determine what should be used as test statistic.

Step 4 consists of calculating which values of the test statistic will lead us to reject the null hypothesis. In the drug example, this means determining how many individuals taking the new drug must be cured for us to conclude that we have significant evidence that the new drug is better than the current one. This step requires a probability theory calculation. In general, the calculations needed in more complicated tests of hypothesis can be quite difficult.

Step 5 is the only statistical step. It consists of getting the data and carrying out the test as prescribed in Steps 1–4.

The P-value approach is an alternative to Step 4, and this also often requires a difficult probability calculation.

In the rest of these notes we adopt a "cookbook" approach. The quantity to be calculated from the data is given and the values of this quantity that lead to rejection of the null hypothesis is also given. However, the background probability theory is not discussed.

This approach is taken because the probability theory behind these tests is quite complicated.

7 Other hypothesis testing examples

7.1 Comparing two binomial parameters

The drug example in the previous section, given to illustrate general hypothesis testing concepts, is an example of the test about the value of a single binomial parameter. A frequently used test is that for the equality of two binomial parameters. An example is given below, again in the drug trial context.

Suppose that we wish to test whether a new drug is effective in curing some illness. We plan to give the new drug to n_1 individuals and a placebo to n_2 individuals, so that the total number of individuals in the trial is $n_1 + n_2 = n$. Let θ_1 be the probability that a person is cured if given the new drug and θ_2 be the probability that a person is cured if given the placebo. Step 1 in the hypothesis testing procedure is to set up the null and the alternative hypotheses. Here the null hypothesis H_0 is

$$H_0 : \theta_1 = \theta_2 \ (= \theta, \text{unspecified}),$$

which in effect claims that the drug is ineffective (i.e. no better than the placebo). The natural alternative hypothesis is

$$H_1 : \theta_1 > \theta_2,$$

which in effect claims that the drug is effective. In Step 2 we choose the Type I error rate, or the probability of a false positive conclusion. Suppose that we choose 1% for this probability.

In Step 3 we choose the test statistic. This is most conveniently done by forming the "two-by-two" table given below.

	Cured	*Not cured*	Total
Drug	y_{11}	y_{12}	n_1
Placebo	y_{21}	y_{22}	n_2
Total	c_1	c_2	n

Thus in this table the number of people who were given the new drug and who were cured is denoted by y_{11}, the number of individuals who were given the new drug and who were not cured is denoted by y_{12}, and the total number of individuals who were given the new drug is denoted by n_1. Similar definitions hold for individuals given the placebo.

The analysis of a "two-by-two" table is not simple. If $n_1 \neq n_2$, it makes no sense to compare y_{11} with y_{21}, since this is not a "level playing field" comparison. Instead we should compare the two *proportions* y_{11}/n_1 and y_{21}/n_2. (This is one reason why it is necessary to consider probability theory relating to proportions in Section 2.7.) It can be shown, after a long probability theory calculation involving the proportion of successes in a binomial situation, that the appropriate test statistic is

$$\frac{\big(y_{11}y_{22} - y_{12}y_{21}\big)\sqrt{n}}{\sqrt{c_1 c_2 n_1 n_2}}. \tag{23}$$

If the Type I error of 1%, a simple approximate procedure shows that any value of the quantity (23) in excess of 2.326 would lead to rejection of the null hypothesis. The probability theory behind this rule is that, for example, if the new drug is not effective, there is approximately a 1% probability that we would declare it to be effective. If the Type I error was chosen as 5%, any value of the quantity (23) in excess of 1.645 would lead to rejection of the null hypothesis.

While this testing procedure is an approximate one, it is quite accurate when the sample sizes n_1 and n_2 are large, meaning in practice about 15 or more. When the sample sizes n_1 and n_2 are small, one can use a completely accurate procedure, namely *Fisher's exact test.* This procedure is not discussed in detail here.

The above is an example of a one-sided test. In some cases we might wish to carry out a *two-sided* test. As an example, suppose that we are curious as to whether there is a difference in "handedness" between males and females. We have no prior reason to believe that males are more or less likely to be left-handed than are females. This means that our test will be *two-sided.* We could take the two rows in the 2×2 table to correspond respectively to males and females, and the two columns to correspond to left- and right-handed, and

the data in the four positions in the table give, for some sample of individuals, the numbers of left-handed males, of right-handed males, of left-handed females, and of right-handed females.

In this example the null hypothesis claims that the probabilities that a male is left-handed is the same as corresponding probability for females. The alternative hypothesis claims that these probabilities are different, but there is no prior direction to this alternative hypothesis.

Again we calculate the observed value of the quantity in (23). If the Type I error was chosen to be 5%, we would reject the null hypothesis if the observed value quantity (23) is less than or equal to -1.96 or is greater than or equal to $+1.96$. These values are chosen so that if the null hypothesis (that the probability that a male is left-handed is the same as the probability that a female is left-handed) is true, there is only a 5% probability that we will claim it to be false.

In the case of a *two-sided* test an equivalent procedure is possible, namely to calculate the *square* of the quantity (23), and (if the Type I error is 5%) to reject the null hypothesis if this squared quantity is greater than or equal to $(1.96)^2 = 3.841$. This procedure is quite convenient, since the squaring operation removes the square roots in (23). That is, we reject the null hypothesis if the statistic

$$\frac{\left(y_{11}y_{22} - y_{12}y_{21}\right)^2 n}{c_1 c_2 n_1 n_2} \tag{24}$$

exceeds the value 3.841. This quantity is an example of a *chi-square* statistic.

7.2 Tests on means

7.2.1 The one-sample t test

A classic test in statistics concerns the unknown mean μ of a normal distribution with unknown variance σ^2. This is demonstrated by the following blood pressure example.

Suppose that there is concern that individuals with a certain illness tend to have unduly elevated blood pressures. Suppose that for normal healthy individuals, the probability distribution of blood pressure is known to be a normal distribution with mean 124. The null hypothesis is that the blood pressure of individuals having this illness is also a normal distribution with mean 124. The alternative hypothesis in this case is that the mean blood pressure for individuals with this illness exceeds 124. The declaration of this null hypothesis and this alternative completes Step 1 of the hypothesis testing procedure.

In Step 2 we declare our chosen Type I error. Suppose that we choose a value 5%. That is, we wish our procedure to be such that if individuals with the illness do indeed have the same probability distribution of blood pressure as normal healthy individuals, we want to limit to 5% the probability that we make the (in this case) false claim that they have a probability distribution with some higher mean.

The third step is to choose a test statistic on the basis of which we will either accept the null hypothesis or reject it. Suppose that we measure the blood pressures of n individuals with this illness, and record values $x_1, x_2, \ldots, x_n$. If the null hypothesis is true, each of these is the observed value of a random variable having a normal distribution with mean 124. However, the variance of this distribution is not known, and must be estimated from the data. This is done using the quantity s^2 defined in equation (18), and repeated here for convenience:

$$s^2 \;=\; \frac{x_1^2 + x_2^2 + \ldots x_n^2 - n\bar{x}^2}{n-1}. \tag{25}$$

Having done this we use as our test statistic the quantity t, defined by

$$t = \frac{(\bar{x} - 124)\sqrt{n}}{s}, \tag{26}$$

where s is the square root of s^2.

In Step 4 of the hypothesis testing procedure we have to decide what values of t lead us to reject the null hypothesis. Clearly for

this example sufficiently large positive values of t will lead us to do this. How large t has to be for the null hypothesis to be rejected depends on the number of "degrees of freedom" for t (in this case $n-1$) and the Type I error chosen. Charts are available that give the appropriate values for a wide choice of values of n and the Type I error chosen.

The final step (Step 5) of the hypothesis testing procedure is the only statistical one, and in it we get the data and do the test as prescribed above.

This procedure can be generalized in two ways. First, the test described above is an example of a test of the null hypothesis $\mu = 124$ against the one-sided alternative hypothesis $\mu > 124$. Of course the value 124 refers to the specific "blood pressure" example. Suppose that in general we wish to carry out the test of the null hypothesis $\mu = \mu_0$, for any given value μ_0. The appropriate test statistic is

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s}. \tag{27}$$

The second generalization is that the test concerning blood pressure, and the more general test where the alternative hypothesis is $\mu > \mu_0$, is a one-sided up test. This is so because, for example in the blood pressure case, the concern is that the mean blood pressure of individuals having the illness exceeds 124. Suppose instead that the concern had been that the mean blood pressure of these individuals is less than 124. Then the alternative hypothesis is "one-sided down." We still use t as the test statistic, but now we would reject the null hypothesis if t is significantly large *negative*. Following on from this, suppose we had a case where the two-sided alternative hypothesis $\mu \neq \mu_0$) is appropriate. Then we would reject the null hypothesis if the observed value t were significantly large positive or negative. t charts can be used to carry out this "two-sided" test.

All t statistics can be thought of as "signal to noise" ratios. For the t statistic in (27) the signal is the numerator expression $\bar{x} - \mu$. The larger the size of this quantity the less the data support the null hypothesis. However this difference on its own is not enough, and it has to be divided by an estimate of the standard deviation

of $\bar{X} - \mu$, namely $s/\sqrt{n}$, which can be thought of as the noise. This interpretation of a t statistic is a very helpful one.

The use of a t statistic is justified only when it can be assumed that the data analyzed (in the above example the blood pressures) have a normal distribution, since the values in t charts are calculated under this assumption. This matter is taken up below.

7.2.2 Two-sample t-test

A situation arising very often in practice is that where we test for the equality of two means. Suppose that we have observations of some measured quantity (for example blood pressure) from one group (individuals with some illness) and also observations from individuals in a "control" group (individuals not having the illness). One way of testing for a difference between the means of of the two groups is to do a two-sample t test. This is described below, again using the blood pressure example.

Suppose that the data are the observed values $x_1, x_2, \ldots, x_m$ of the blood pressures of m individuals with the illness and the observed values $y_1, y_2, \ldots, y_n$ of the blood pressures of n individuals who do not have the illness. If we carry out a two sample t test described below, we implicitly assume that these are the observed values of random variables $X_1, X_2, \ldots, X_m$ having a normal distribution with mean μ_1 and random variables $Y_1, Y_2, \ldots, Y_n$ having a normal distribution with mean μ_2. In the procedure described below it is also assumed that the variances of these two normal distributions are equal. If we are not willing to make these various assumptions, alternative procedures (not described here) are available.

The null hypothesis in any two-sample t test is that $\mu_1 = \mu_2$. The alternative hypothesis depends on the context. In the blood pressure example above, the natural alternative hypothesis of interest might be the "one-sided up" hypothesis $\mu_1 > \mu_2$. In other contexts the natural alternative hypothesis might be $\mu_1 < \mu_2$ and in other contexts again it might be $\mu_1 \neq \mu_2$.

Under the assumptions described above, the appropriate test statis-

tic is t, defined by

$$t = \frac{(\bar{x} - \bar{y})\sqrt{mn}}{s\sqrt{m+n}}. \tag{28}$$

In this expression, $\bar{x}$ and $\bar{y}$ are defined by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_m}{m}, \tag{29}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}, \tag{30}$$

and s is defined as the square root of the quantity s^2, defined by

$$s^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{m + n - 2}. \tag{31}$$

For the one-sided up test discussed above, significantly large positive values of t lead to the rejection of the null hypothesis. How large positive t has to be before the null hypothesis is rejected depends on two quantities, the Type I error rate chosen for the test and the number of degrees of freedom of the test, which in this case is $n + m - 2$. The Type I error in this case is the error in claiming that $\mu_1 > \mu_2$ if in fact $\mu_1 = \mu_2$.

If the alternative hypothesis had been $\mu_1 < \mu_2$, the test required is *one-sided down*. Here we would reject the null hypothesis if the observed value t is sufficiently large and *negative*. If the alternative hypothesis had been $\mu_1 \neq \mu_2$, the test is two-sided. The test statistic t is still used in the testing procedure, and both significantly large positive values of the observed value of t and also significantly large negative values of the observed value of t lead to the rejection of the null hypothesis.

In the t statistic (28) the signal is the difference $\bar{x} - \bar{y}$, which is an estimate of the extent to which μ_x differs from μ_y, and in (28) this is divided by an appropriate measure of noise.

7.2.3 Testing for means: the "paired" two-sample t-test

An important case of the two-sample t test arises if $n = m$ and the observations x_i and $y_i, i = 1, 2, \ldots, m$ can be logically paired, for example by using brother-brother pairs, where one brother in each

pair has the illness and the other does not. The "paired" t-test case is carried out by using the differences $d_i = x_i - y_i, i = 1, 2, \ldots, n$ in blood pressures of the various brothers in each pair, and basing the test entirely on these differences. This reduces the test to a one-sample t-test with test statistic t defined in (32).

$$t = \frac{\bar{d}\sqrt{n}}{s_d}, \tag{32}$$

Here $\bar{d}$ is the average of the d_i values and s_d^2 is given by

$$s_d^2 = \frac{d_1^2 + d_2^2 + \cdots + d_n^2 - n\bar{d}^2}{n-1}. \tag{33}$$

The t chart with $n-1$ degrees of freedom is used to evaluate the significance of the observed value of t.

7.2.4 The normal distribution assumption.

It was assumed in the testing procedures described above that the blood pressure of an individual taken at random has a normal distribution. The use of t charts is not appropriate if this assumption is not correct, since the significance points in the t chart were calculated under the normal distribution assumption. In many cases in practice this assumption might be unreasonable, and this leads to the introduction of alternative testing procedures that do not rely on the normality and equal variances assumptions. These are now discussed.

7.3 Non-parametric alternatives to the unpaired and paired t-tests

7.3.1 Introduction

Hypothesis testing methods that do not rely on the any assumption about the probability distribution that the data used in the testing procedures come from are called *non-parametric*, or (perhaps more accurately) *distribution-free*. We now describe two non-parametric procedures that are sometimes used as alternatives to the two-sample t-test.

7.3.2 The permutation test

The two-sample permutation test is often used in the analysis of microarray data, currently of much interest in bioinformatics, so we now discuss it in some detail.

For this test, as for the two-sample t test considered above, there are two groups of observations, m observations $x_1, x_2, \ldots, x_m$ in the first group and n observations $y_1, y_2, \ldots, y_n$ in the second. The null hypothesis tested in the permutation procedure is that the probability distribution of the x's is identical to that of y's. Various alternative hypotheses are possible: here we assume, to be concrete, that the alternative hypothesis is that the two distributions are identical in all ways except that the "x" distribution is shifted towards higher values compared to the "y" distribution.

When the null hypothesis is true, all possible $\binom{m+n}{m}$ distinct allocations of the $x_1, x_2, \ldots, x_m$ values and the $y_1, y_2, \ldots, y_n$ values, in which m randomly chosen values are taken as the "x"s and the remaining n values are taken as "y"s are equally likely. The permutation test is based on this fact, and the procedure is carried out as follows.

For each one of the $\binom{m+n}{m}$ possible allocations described above, we calculate the average of the values allocated as "x"s and the the average of the values allocated as "y"s, and then take the difference of each of these averages. This procedure will result in $\binom{m+n}{m}$ differences, exactly one of which corresponds to the correct allocation of values to the "x" group and values to the "y" group. If a Type I error α is chosen, the null hypothesis is rejected if the correct difference value is among the largest $100\alpha\%$ of all the $\binom{m+n}{m}$ difference values.

In practice, $\binom{m+n}{m}$ might be a very large number, perhaps several billion, and in this case a random sample of perhaps 10,000 different allocations might be chosen, one of which is the correct allocation, and again if a Type I error α is chosen, the null hypothesis is rejected if the correct difference value is among the largest $100\alpha\%$ of all the difference values calculated.

Clearly the permutation procedure is computationally intensive, and is suited to computer calculations. Indeed it is rarely carried out other than by using a computer package.

7.3.3 The Mann–Whitney test

The *Mann–Whitney* test (sometimes called the *Wilcoxon two-sample* test) is another frequently used non-parametric alternative to the two-sample t-test. We therefore assume here the same random variables as for that test. (These are also the same random variables as for the permutation test just described.) The null hypothesis tested is that the "x" distribution is identical to the "y" distribution.

In the Mann-Whitney test the observed values $x_1,\ x_2, \ldots,\ x_m$ and $y_1,\ y_2, \ldots,\ y_n$ of the $m + n$ random variables are jointly listed in increasing order, and each observation is associated with its rank in this list. Thus each observation is associated with one of the numbers $1, 2, \ldots, m + n$. (If ties exist a slightly amended procedure is used. We ignore this complication here.) The test statistic is the sum of the ranks of the observations in the first, or "x", group. The sum of the ranks of all $m+n$ observations is $(m+n)(m+n+1)/2$. The observations in the first group provide a fraction $m/(m + n)$ of all observations, and proportionality arguments then show that when the null hypothesis (that the x distribution and the y distribution are identical) is true, the permutation mean value of the sum of the ranks of the observations in the first group is the fraction $m/(m+n)$ of the sum $(m + n)(m + n + 1)/2$ of all ranks. Thus this mean value is $m(m+n+1)/2$. A more advanced calculation shows that the null hypothesis permutation variance of the sum of the ranks for the first group in the sample is $mn(m+n+1)/12$. It is also possible to show that this sum has very close to a normal distribution. If the null hypothesis is true, we would expect the observed sum of ranks to be reasonably close to its permutation mean $m(m+n+1)/2$. A formal testing procedure is available to test if this can be taken as being the case, but we do not give the details of the procedure here.

7.3.4 Testing the hypothesis $\beta = 0$ in the regression model

It was stated earlier that the most interesting parameter in the regression model (13) is β, and correspondingly the most interesting

test of hypothesis in this model is the test of the null hypothesis $\beta = 0$. This is because, in the plant example, if $\beta = 0$ the amount of water given to a plant has no effect on its growth height. Assuming that plant growth has a normal distribution, the test statistic is another form of t, in this case

$$t = \frac{b\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{s_r}, \tag{34}$$

and this is, again, a signal to noise ratio. The signal in this case is b, and large values of b do not support the null hypothesis. The noise is $s_r/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, which (see (21)) is an estimate from the data of the standard deviation of B.

8 The Analysis of Variance (ANOVA)

NOTE: These notes are a direct continuation of the "Biostatistics for Genetics and Genomics" notes.

8.1 Introduction

Perhaps the most frequently used hypothesis testing procedure in statistics is that of the Analysis of Variance (ANOVA). The simplest ANOVA, that is the "one-way ANOVA", can be regarded as a generalization of the two-sample equal variance t-test, and we approach ANOVA through this generalization.

As described in the Biostatistics for Genetics and Genomics notes, the two-sample t-test (see the statistic (28)) tests for equality of the means of two groups. We change notation now to one more suitable for that in ANOVA by replacing the notation for the observations $y_1, y_2, \ldots y_m$ in the first group by $x_{11}, x_{12}, \ldots x_{1m}$ and the observations $x_1, x_2, \ldots x_n$ in the second group by $x_{21}, x_{22}, \ldots x_{2n}$. If this is done, the model for the two-sample t test can be written as

$$X_{ij} = \mu_i + E_{ij},\ i = 1, 2, \tag{35}$$

where the X_{ij} are assumed to be independent and the E_{ij} are assumed to be $\text{NID}(0, \sigma^2)$ random variables. Here we follow the notation used in the Biostatistics for Genetics and Genomics notes of using upper case notation for random variables and the corresponding lower case notation for the observed values of these random variables once the relevant experiment is done.

The null hypothesis being tested is $\mu_1 = \mu_2$. This model is also often written in the form

$$X_{ij} = \mu + \alpha_i + E_{ij},\ i = 1, 2. \tag{36}$$

In this model we can think of μ as an overall mean and α_j as a deviation from this overall mean characteristic of group j. In this form the model is *overparameterized*. There are three parameters in the model (μ, α_1 and α_2) when only two (μ_1 and μ_2)are necessary. This overparameterization can be overcome by requiring that α_1 and α_2 satisfy the requirement $m\alpha_1 + n\alpha_2 = 0$, and we always assume that this requirement is imposed. It might seem to be a roundabout approach to write the model (35) in the form (36), with the condition

$m\alpha_1 + n\alpha_2 = 0$ imposed, but for several reasons it is convenient to do so. The test of hypothesis is identical in the two representations of the model.

8.2 From t to F: Sums of Squares and the F Statistic

We start by deriving a test procedure equivalent to the two-sample two-sided t-test described in the Introductory Statistics notes. This test is carried out by using $|t|$, the absolute value of the statistic (28), as test statistic. It is equivalent to use t^2 as test statistic, since this is a monotonic function of $|t|$. Straightforward algebraic manipulation shows that if the square of the two-sample t statistic (28) is written as F, then

$$F = \frac{B}{W}(m + n - 2), \tag{37}$$

where, in our new notation,

$$\bar{x}_1 = \sum_{j=1}^{m} x_{1j}/m, \quad \bar{x}_2 = \sum_{j=1}^{n} x_{2j}/n, \quad \bar{\bar{x}} = (m\bar{x}_1 + n\bar{x}_2)/(m + n),$$

$$B = \frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)^2 = m(\bar{x}_1 - \bar{\bar{x}})^2 + n(\bar{x}_2 - \bar{\bar{x}})^2 \tag{38}$$

and

$$W = \sum_{j=1}^{m}(x_{1j} - \bar{x}_1)^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2. \tag{39}$$

B is the *between group sum of squares* (more precisely called the *among group sum of squares*), and W is called the *within group sum of squares*.

The sum of B and W can be shown to be

$$\sum_{i=1}^{m}(x_{1i} - \bar{\bar{x}})^2 + \sum_{i=1}^{n}(x_{2i} - \bar{\bar{x}})^2. \tag{40}$$

This is called the *total sum of squares*. The total number of degrees of freedom is $m + n - 1$, and this is one less than the total number of observations. Just as the total sum of squares is split up into a between group component and a within group component, so also

the total number of degrees of freedom is split up into two components, 1 degree of freedom between groups and $m+n-2$ degrees of freedom within groups.

The two main components of the statistic F are B and W. Our aim is to test for significant differences between the means of the two groups, and the component B measures this difference by the difference between the group averages $\bar{x}_1$ and $\bar{x}_2$. The value of $\bar{x}_1 - \bar{x}_2$ will tend to be large when the means of the two groups differ. However, the significance of any such difference must be measured relative to the variation within groups. The component W measures this variation, and is unaffected by any difference in means between the two groups. Large observed values of F arise when the observed variation between groups is large compared with the observed variation within groups, and sufficiently large observed values of F give significant evidence that a difference exists between the two means. The ANOVA procedure makes this precise, as follows.

It can be shown that, when the null hypothesis $\mu_1 = \mu_2$ is true and when the standard ANOVA assumptions listed above are met, the random variables corresponding to the quantity F in (37) has the F distribution with 1 and $m+n-2$ degrees of freedom. The test of the null hypothesis is then carried out by referring the observed value of F to tables of significance points of the F distribution with 1 and $m+n-2$ degrees of freedom. For any desired Type I error, the values of F that lead to rejection of the null hypothesis can be found from widely available tables.

This procedure demonstrates the two key steps in any ANOVA. The first is the subdivision of the observed total sum of squares into several components (in the above case the observed values of B and W), each measuring some meaningful component of variation. The second step is the comparison of these components to test some hypothesis, using for each comparison the appropriate F statistic.

The two-group comparison above generalizes immediately to a test for the equality of the means of any number of groups, and then to a hierarchy of further ANOVA tests. below we describe the simplest of these, of the "one-way fixed effects " ANOVA test.

8.3 One-way Fixed Effects ANOVA

The one-way ANOVA test is a direct generalization of the two-sample t-test to the case of an arbitrary number g of groups, in which the null hypothesis is that the means of all the groups are equal. This generalization will require a further change of notation: instead of writing the two group sizes in a two-sample t test as m and n, we write the g group sizes in an ANOVA as $n_1, n_2, \ldots, n_g$. We also write the n_i observations in group i as $x_{i1}, x_{i2}, \ldots, x_{in_i}$ for $i = 1, 2, \ldots, g$. The model generalizing (35) is

$$X_{ij} = \mu_i + E_{ij}, \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, g, \tag{41}$$

and the model generalizing (36) is

$$X_{ij} = \mu + \alpha_i + E_{ij}, \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, g. \tag{42}$$

In both models the E_{ij} are assumed to be NID$(0, \sigma^2)$. The model (42), like the model (36), is overparameterized, and the overparameterization is overcome by requiring that $\sum n_i \alpha_i = 0$. This overparameterized model is often more convenient to use than is (41), and the test of hypothesis is the same in both models.

The null hypothesis in the ANOVA model (42) is that $\alpha_1 = \alpha_2 = \cdots = \alpha_g = 0$. The parameter μ is not involved in the test. It should be noted that despite the appearance of the word "variance" in the expression "Analysis of Variance," the hypothesis tested in this (and any) ANOVA procedure is a test about *means*.

The test procedure is a direct extension of the two-group procedure described in the Biostatistics for Genetics and Genomics notes. The total sum of squares $\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2$ is subdivided into the between group sum of squares B and the within group sum of squares W. These are the direct generalizations of the two-group values in (38) and (39), and are defined by

$$B = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{\bar{x}})^2, \tag{43}$$

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \tag{44}$$

Here $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ and $\bar{\bar{x}} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} x_{ij}/N$, where N is defined as $\sum_{i=1}^{g} n_i$. Just as the total sum of squares is split up into two

components, so also the total degrees of freedom $N-1$ is subdivided into two components, the between group degrees of freedom $g-1$, and the within group degrees of freedom $N-g$.

We define the *between group mean square* as the between group sum of squares divided by the between group degrees of freedom, and the *within group mean square* as the within group sum of squares divided by the within group degrees of freedom. The test statistic F is the ratio of these two mean squares, or equivalently

$$F = \frac{B}{W} \times \frac{N-g}{g-1}. \tag{45}$$

The test statistic F in (45) is the direct generalization of that in (37), and has the F distribution with $g-1, N-g$ degrees of freedom when the null hypothesis is true and the standard ANOVA assumptions are all met. The test is therefore carried out by referring the observed value of F to significance points of this F distribution.

There are various ways in which the ANOVA assumptions might not be met. The first of these arises if the random variables X_{ij} do not have a normal distribution. The procedure is fairly *robust* against non-normality, so that the F statistic has approximately the F distribution for non-normal data, provided that the non-normality is not extreme. Non-parametric alternatives generalizing the Mann–Whitney test, and also the permutation test procedure, are available if non-normality appears to be extreme. As with the t-test, the ANOVA test is also fairly robust when the variances in the various groups differ, at least when the group sizes are equal.