

Statistics Notes

Birmingham Short Course

Warren J. Ewens

Introduction

Statistics is the science of analyzing data in whose generation chance has played some part. This explains why Statistics is important in genetics. In genetics there are many chance mechanisms at work. For example, the random transmission of one of two genes at a locus from parent to offspring introduces a chance mechanism into many areas of genetics, especially evolutionary genetics and the genetics of disease genes. Second, data are usually derived some random sample of individuals. A different sample would almost certainly yield different data, so that the sampling process introduces a second chance element.

Several of the examples given below are taken from the genetics context, first because they introduce some important problems in that area, and second because they are concrete and usually quite straightforward.

An example: the “linkage analysis” case

(This example will be discussed from time to time in these notes.)

A “marker” locus is a genetic locus whose location in the genome is known, and for which we know the genotype (i.e. the two genes at the locus) for any given individual.

Consider a parent who is heterozygous M_1M_2 at some marker locus. If no further information is given, the probability that this parent transmits the same gene to two nominated offspring (that is, either M_1 to both offspring or M_2 to both offspring) is 0.5. It can however be shown that, if both offspring are affected by some disease, and the marker locus is linked to (i.e. close on the same chromosome to) the disease locus, the probability that this parent transmits the same gene to the two affected offspring *exceeds* 0.5.

Suppose we get data from 100 families, in each of which there is an M_1M_2 parent, an “uninformative” M_2M_2 parent, and two affected children. Suppose that in these 100 families, the M_1M_2 parent passes on the same gene to the two affected offspring in 61 cases. Do we have evidence that the marker locus is linked to the disease locus?

We cannot give any objective answer to this question unless we first ask: “Assuming that the marker locus is NOT linked to the disease locus, what is the probability that the parent passes on the same gene to the two affected offspring in 61 (or even more) cases?”

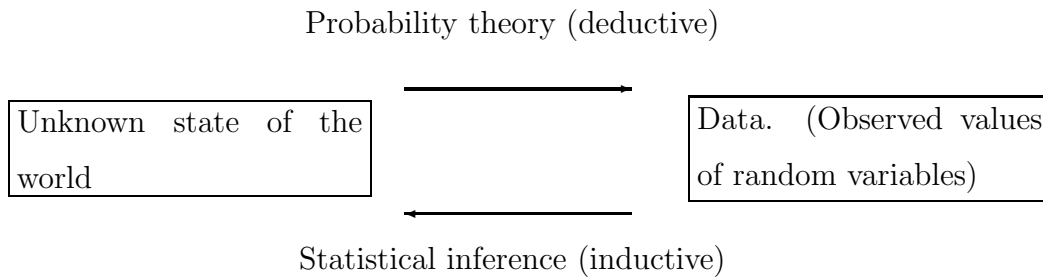
If, under this assumption, this probability is very small, the observation that the same gene was passed on 61 times out of 100 gives strong evidence that the marker locus IS linked to the disease locus. If this probability is quite large, the observation does not give strong evidence that marker locus is linked to the disease locus. Under the “no linkage” assumption the actual probability is about 0.0179, and this is small enough to give significant evidence that the marker and disease loci are linked.

The claim that the data value 61 supports the hypothesis that the marker locus is linked to the disease locus is a statement of *statistics*. No statistical statement can be made without first making some *probability* calculation - in the above case the calculation leading to the value 0.0179 - upon which the statistical statement is based. This leads us to make a closer examination of the relation between probability and statistics.

The Relation Between Probability and Statistics

The concluding comment in the previous section makes it important to consider in more detail the relationship between probability and statistics.

Suppose that we decide, having observed 61 cases out of 100 where the two affected children in a family share the same marker locus gene, and having made the probability calculation 0.0179 discussed above, that there is significant evidence that the marker locus is linked to the disease locus. This conclusion is a typical example of a *statistical inference*. It is important to note that the “direction” of such an inference is the opposite to that of a probability calculation. This is illustrated in the following diagram.



In a probability calculation we make some assumption (in the above example, that the marker locus is unlinked to the disease locus), and then calculate the probability of the observed data (or something more extreme) *under this assumption*. In this case the probability calculation yields the value 0.0179. In statistical inference we start with observed data, (in this case the observed number 61), and on the basis of an appropriate probability calculation, in this case the one leading to the value 0.0179, we make a statement about whether the assumption used in the probability calculation is reasonable.

Every statistical inference depends on some probability calculation, and no valid statistical inference can be made without first carrying out the probability calculation appropriate to that inference. Thus a study of probability theory is essential for an understanding of statistics. Probability theory is important also on its own account, quite apart from its underpinning of statistics. This is particularly true in genetics. Thus these notes start with a brief discussion of probability theory.

Probability: One Discrete Random Variable

Definition: one discrete random variable

It is convenient to consider separately the cases of discrete and continuous random variables. In this section we give informal definitions for discrete random variables, probability distributions, and parameters, rather than the formal definitions often found in statistics textbooks. Continuous random variables will be considered in a later section.

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. In practice the possible values of a discrete random variable often consist of the numbers $0, 1, 2, 3, \dots, n$, for some number n .

As a matter of notational convention, a random variable is always denoted by an upper case Roman letter. We use the letter X in these notes for this purpose.

The concept of a random variable arises because we have to force ourselves to think of the situation *before* we do some experiment. Suppose I plan to toss a coin 100 times tomorrow. To me, today, the number of heads that I will get, tomorrow, is a random variable - I cannot know, today, what its value will be. So, today, it is correct to denote this number by X .

Later, when we consider statistics, we will consider the *observed* value of a random variable, once the relevant experiment has been done. The notation used for this is the lower case letter corresponding to the upper case letter for the random variable. So tomorrow, after I have tossed the coin 100 times, I can denote the number of heads I actually observed by x . Thus it makes sense, after I have tossed the coin tomorrow, to say “ $x = 54$ ”, but it does not make sense today, before I toss it, to say $X = 54$. This latter statement “does not compute”.

Probability distributions and parameters

There will always be some set of possible values for a random variable. For example, in the coin tossing case just described, the possible values for the random variable X are $0, 1, 2, \dots, 100$.

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. We denote the probability that the random variable X takes the value x by $P(X = x)$.

For example, if we plan to toss a fair coin twice tomorrow, and the random variable X is

the number of heads that eventually turn up tomorrow, the probability distribution of X is:

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

Here $P(X = 0) = .25$, $P(X = 1) = .5$, $P(X = 2) = .25$.

In practice, the probabilities associated with the possible values of the random variable of interest are often unknown. For example, if the probability π of a head on each of the two tosses is unknown to us, then the probabilities for 0, 1, or 2 heads when this coin is tossed twice are unknown. Nevertheless we can still write down the probability distribution of the number of heads that will appear in terms of π as

Possible values of X	0	1	2	(2)
Associated probabilities	$(1 - \pi)^2$	$2\pi(1 - \pi)$	π^2	

Thus in this case,

$$P(X = 0) = (1 - \pi)^2, P(X = 1) = 2\pi(1 - \pi), P(X = 2) = \pi^2. \quad (3)$$

In this distribution π is a *parameter*, that is, some unknown constant. Since in research, which is where we use statistics, we are delving into the unknown, almost all interesting probability distributions contain unknown parameters, and much of statistical inference concerns estimating, and testing hypotheses about, these parameters.

As a matter of notation, parameters are often denoted by Greek letters, and this convention is followed in these notes (as with π above).

The disease and marker locus case discussed above provides an example of an unknown parameter. If θ is the so-called recombination fraction between disease and marker loci, then the test that these loci are unlinked can be rephrased as the test that the unknown parameter θ takes the value $1/2$, (since this is the value of θ for unlinked loci). Although it might seem an awkward or roundabout way of doing it, it is usually convenient to cast any test of hypothesis in the form of a test about the value of an unknown parameter. Examples of these tests are discussed below.

Independence

The concept of independence is an important one in probability and statistics. Formal definitions of independence of events and independence of random variables are given in statistics textbooks. Here we give an informal definition. That is, we take the word “independent” to have its common everyday meaning: Two or more events are independent if the outcome of one event does not affect in any way the probability of any other event. Two or more discrete random variables are independent if the value of one does not affect in any way the probabilities associated with the possible values of any other random variable.

As an example, if a coin is to be tossed many times, the event “ i th toss will give head, j th toss will give head” are usually assumed to be independent when $i \neq j$. On the other hand the heights of two brothers would not be taken as independent, due to their common parentage and environment.

The binomial distribution

There are many important discrete probability distributions that arise time and again in applications of probability and statistics. Here we focus only on the most important of these, the binomial distribution.

The binomial distribution arises if, and only if, all four of the following requirements hold. First, we plan to conduct some fixed number n of trials. (By “fixed” we mean fixed in advance, and not, for example, determined by the outcomes of the trials as they occur.) Second, each trial must result in one of two possible outcomes. (The two outcomes are often called, for convenience, “success” and “failure”, and we use this terminology here.) Third, the various trials must be independent. Finally, the probability of success must be the same on all trials. In conformity with the notational convention given above, this probability is here denoted by the Greek letter π .

The random variable of interest is the total number X of successes in the n trials. The

probability distribution of X is given by the (binomial distribution) formula

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, \dots, n. \quad (4)$$

The coefficient $\binom{n}{x}$ is often spoken as “ n choose x ”: it is the number of different orderings of x objects of one kind and $n - x$ of another. π is called the parameter, and n the index, of this distribution.

The probabilities in (??) (equivalently in (??)) are binomial distribution probabilities for the case $n = 2$.

There are three important comments to make concerning the binomial distribution. First, the numerical value of the parameter π in (??) is often unknown. In many applications of statistics we want to make a test of hypothesis about the numerical value of this parameter. An example of this is given below illustrating the general principles of hypothesis testing methods (pages 20 to 23).

Second, to carry out these tests of hypothesis we often want to calculate probabilities of the form: “What is the probability that a random variable having a binomial distribution with parameter π and index n takes a value x or more?” (In the case $\pi = 1/2$, $n = 100$, $x = 61$, this probability was given above as 0.0179.) Probabilities of this type are hard to calculate when n is large, and they are often approximated by the “normal approximation to the binomial”, discussed later.

Finally, one must be careful when using a binomial distribution that the four defining conditions above all hold.

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 X is defined as

$$\sum_x xP(X = x), \quad (5)$$

the summation being over all possible values that the random variable X can take. As an example, the mean of a random variable having the binomial distribution (??) is

$$\sum_{x=0}^n x \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad (6)$$

and this can be shown to be $n\pi$.

There are several remarks to make regarding the mean of a discrete random variable.

- (i) The notation μ is often used for a mean. An alternative name for the mean of a random variable is the “expected value” of that random variable, leading to a second notation, $E(X)$, for the expected value, or mean, of the random variable X .
- (ii) Following on from this, in many practical situations the mean μ of a discrete random variable X is unknown to us, because we do not know the numerical values of the probabilities $P(X = x)$. That is to say, μ is a parameter, and this is why we use Greek notation for it.
- (iii) Estimating a mean, and testing hypotheses about the value of a mean, are among the most important of statistical operations. Two important examples of tests of hypotheses about means are t -tests and ANOVA, described below.
- (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 again later.

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 by

$$\sigma^2 = \sum_x (x - \mu)^2 P(X = x), \quad (7)$$

the summation being taken over all possible values of X .

There are several points to note concerning the variance of a discrete random variable.

- (i) The variance has the standard notation σ^2 , anticipated above.
- (ii) The variance is a measure of the dispersion of the probability distribution of the random variable around its mean. Thus a random variable with a small variance is likely to be close to its mean.
- (iii) A quantity that is often more useful than the variance of a probability distribution is the *standard deviation*. This is defined as the positive square root of the variance, and (naturally enough) is denoted by σ .
- (iv) The variance, like the mean, is often unknown to us.
- (v) The variance of the binomial distribution (??) is $n\pi(1 - \pi)$.

Probability: One Continuous Random Variable

Definition

Some random variables, by their nature, are discrete, such as the number of heads in n tosses of a coin. Other random variables, by contrast, are continuous. Measurements such as height and blood pressure are of this type. We use the same notation as for a discrete random variable and denote a continuous random variable by X . Continuous random variables can take any value in some continuous range of values. Here we denote this range by (L, H) , (L = lowest, H = highest, possible values), and use this notation throughout this section.

Probabilities for continuous random variables are not allocated to specific values, but rather are allocated to intervals, or ranges, of values. The probability that a continuous random variable takes some specified numerical value is zero.

Each random variable X has an associated *density function* $f(x)$, and the probability that the random variable takes a value in some given interval is obtained by integrating this density function over that interval. For example,

$$\text{Prob}(a < X < b) = \int_a^b f(x) dx. \tag{8}$$

Because the probability that a continuous random variable takes some specified numerical value is zero, the three probabilities $\text{Prob}(a \leq X < b)$, $\text{Prob}(a < X \leq b)$, and $\text{Prob}(a \leq X \leq b)$ are also given by the right-hand side in (??).

As a particular case of equation (??),

$$\int_L^H f(x) dx = 1. \quad (9)$$

This equation simply states that a random variable must take some value in its range of values.

The mean and variance of a continuous random variable

The mean μ and variance σ^2 of a continuous random variable X having range (L, H) and density function $f(x)$ are defined respectively by

$$\mu = \int_L^H x f(x) dx \quad (10)$$

and

$$\sigma^2 = \int_L^H (x - \mu)^2 f(x) dx. \quad (11)$$

These definitions are the natural analogues of the corresponding definitions for a discrete random variable, and the remarks about the mean and the variance of a continuous random variable are very similar to those of a discrete random variable given above. In particular, a continuous random variable with a small variance is likely to be close to its mean.

The normal distribution

There are many continuous probability distributions relevant to the statistics. We discuss the most important one in this section.

The (continuous) random variable X has a *normal*, or *Gaussian*, distribution if its range (i.e. set of possible values) is $(-\infty, \infty)$ and density function $f(x)$ given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (12)$$

It can be shown that the mean of this distribution is μ and its variance σ^2 , and these parameters are built into the functional form of the distribution, as (??) shows. A random variable having this distribution is said to be an $N(\mu, \sigma^2)$ random variable.

A particularly important normal distribution is the one for which $\mu = 0$ and $\sigma^2 = 1$. This is sometimes called the *standardized normal* distribution, and this name arises for the following reason. Suppose that a random variable X has the normal distribution (??), that is, with arbitrary mean μ and arbitrary variance σ^2 . Then the “standardized” random variable Z , defined by $Z = (X - \mu)/\sigma$, has a normal distribution with mean 0, variance 1.

One of the uses of this standardization is the following. If X is a random variable having a normal distribution with mean 6 and variance 16, we cannot find $\text{Prob}(5 < X < 8)$ by integrating the function in (??) in closed form. However, for the standardized variable Z , accurate approximations of such integrals are widely available in tables. These tables may be used, in conjunction with the standardization procedure, to find probabilities for a random variable having any normal distribution. Thus in the above example, we can rewrite $\text{Prob}(5 < X < 8)$ in terms of the standard normal Z by

$$\begin{aligned} \text{Prob}(5 < X < 8) &= \text{Prob}\left(\frac{5-6}{\sqrt{16}} < \frac{X-6}{\sqrt{16}} < \frac{8-6}{\sqrt{16}}\right) \\ &= \text{Prob}(-0.25 < Z < 0.5), \end{aligned} \tag{13}$$

and this probability is found from standardized normal tables as 0.2902.

One of the many uses of the normal distribution is to provide approximations for probabilities for various discrete random variables. Perhaps the most important is the normal approximation to the binomial. If the number of trials n in the binomial distribution (??) is very large, the normal distribution with mean $n\pi$ and variance $n\pi(1 - \pi)$ can be used to provide a very good approximation for binomial probabilities.

As one example of this approximation, if X is binomial with parameters p and n , and a is any integer,

$$\text{Prob}(X \geq a) \approx \text{Prob}(X^* \geq a - 1/2),$$

where X^* is a random variable having a normal distribution with $\mu = n\pi$ and $\sigma^2 = n\pi(1 - \pi)$. The term $\frac{1}{2}$ is known as the “continuity correction”, and it arises whenever discrete probabilities are approximated by a continuous distribution.

Thus the probability that a binomial random variable with parameter $\pi = 1/2$ and $n = 100$ is 61 or more is approximately equal to the probability that a normal random variable with mean 50 and variance 25 is 60.5 or more. The Z -ing procedure described above shows that this is the probability that a standard normal random variable is 2.1 or more. Tables of the standard normal distribution show that this probability is 0.0179. This is the value given in the “linkage” example above, and indeed this is how this probability was approximated.

The two-standard deviation rule

The “the two-standard deviation rule” is a rough rule of thumb that is surprisingly accurate in many cases. It states that if a random variable X has mean μ and standard deviation σ , then

$$\text{Prob}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95. \quad (14)$$

We will use this rule in some of the statistical calculations given below.

Probability: Many Random Variables

Introduction

Almost every application of statistical methods requires the analysis of many observations. For example, if we wish to test whether a certain die is fair, we would plan to roll it many times, and thus plan to get many observations, before making our assessment. This leads to the consideration of the probability theory for many random variables. Many of the calculations needed in statistics depend on this theory. This theory is quite complicated, and except in one important case we do not go into the details of it here.

Since we are now considering many random variables, the notation “ X ” is no longer sufficient for us. We denote the first random variable by X_1 , the second by X_2 , and so on.

As an example, suppose we plan to roll a die n times. We denote by X_1 as the (random) number that will turn up on the first roll; \dots , by X_n the (random) number that will turn up on the n -th roll.

The die example introduces a further concept. We would reasonably assume that X_1, X_2, \dots, X_n all have the same probability distribution, since it is the same die that is being rolled each time. Further, we would also reasonably assume that the various values X_1, X_2, \dots, X_n are all independent of each other. Random variables which are independent and have the same probability distribution are said to be *iid* (independently and identically distributed). We use this terminology often.

Returning to the die example, suppose for the moment that the die is fair. Then the number turning up on any one roll of this die is a random variable, whose possible values are 1, 2, \dots , 6, each with probability $1/6$. Application of equation (??) shows that the mean of this random variable is $1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 3.5$, and application of (??) shows, after some calculation (try it!), that the variance of this random variable is $35/12$.

We now turn to the *average* of n *iid* random variables X_1, X_2, \dots, X_n . This average is denoted by \bar{X} , and is defined by

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \tag{15}$$

The most important thing to say about \bar{X} is that it is a *random variable*. As such, it must be distinguished carefully from the mean, which is a parameter.

The fact that the average is a random variable is easy to see in the die-rolling case - we just do not know what this average will be. This should make it easier to accept that in general, whatever the context, \bar{X} is a *random variable*.

By contrast, in the die case, if the die is fair, we know that the mean is 3.5, as calculated above. (If the die were unfair we would *not* know what the mean is.) It “does not compute” to even talk about the numerical value of \bar{X} .

Since \bar{X} is a random variable it has a probability distribution, and in particular a mean and a variance. These must be related in some way to the mean and the variance of each of X_1, X_2, \dots, X_n . The general theory of many random variables is that if X_1, X_2, \dots, X_n are *iid*, and that their (common) mean is μ and their (common) variance σ^2 , then the mean and the variance of the random variable \bar{X} are given by

$$\text{mean of } \bar{X} = \mu, \quad \text{variance of } \bar{X} = \frac{\sigma^2}{n}. \quad (16)$$

In the case of a fair die, each X_i has mean 3.5 and variance 35/12, as given above, so the (for example) the average of the numbers to turn up on 100 rolls of the die is a random variable with mean 3.5 and variance 35/1200. This small variance implies that once we roll the die 100 times, it is very likely that the average of the numbers to eventually turn up will be very close to 3.5. This is no more than what intuition suggests, but now we have a concrete quantification for this variability. This idea will be used in Statistics, to which we now turn.

Statistical Inference

Introduction

Statistics, the method of analyzing data in whose generation chance has played some part, consists of two main areas: estimation (of the numerical values of parameters) and testing hypotheses (usually about the numerical values of parameters). Both operations are used extensively in genetics. This section gives an introduction to estimation and hypothesis testing ideas.

Notation

All the probability theory discussed above concerns the situation *before* we do our experiment. Thus it concerns *random variables*, since before we do our experiment we do not know

what values we will get. Recall that we denote random variables by upper case letters.

We now suppose that our experiment has been done. If before the experiment we had been considering some single random variable X , we denote the actually observed value of this random variable once the experiment has been carried out by x . Thus if the experiment concerned the tossing of a coin 100 times, and before the experiment the random variable X was the number of heads we would see, it makes sense after the experiment to say $x = 59$. It does *not* make sense, before the experiment, to say $X = 59$.

If before the experiment we had been considering several random variables X_1, X_2, \dots, X_n , we denote the actually observed value of these random variable once the experiment has been carried out by x_1, x_2, \dots, x_n . Thus if the experiment consisted of the rolling of a die $n = 3$ times, it makes sense to say, after the experiment has been done, that $x_1 = 5, x_2 = 3, x_3 = 3$. it does *not* make sense to say, before the experiment has been done, that $X_1 = 5, X_2 = 3, X_3 = 3$.

Estimation methods

Introduction

Much of the theory concerning estimation of parameters is essentially the same for both discrete and continuous random variables, so in this section we consider only discrete random variables.

Let X be a discrete random variable having an unknown probability distribution $P(X = x; \theta)$, the distribution depending, as the notation indicates, on some parameter θ whose numerical value is unknown. How should we estimate this parameter from the observed value x of X , once the experiment has been done?

The observed value x on its own will usually not be sufficient to allow good estimation. For example, if we have a possibly biased die and we want to estimate the mean of the number to turn up, the fact that (say) on one roll of the die we got the observed value $x = 5$

is of very little value in estimating this mean.

We must consider repeating the experiment that generated this value an (ideally large) number of times. That is, before the experiment, we must conceptualize n random variables X_1, X_2, \dots, X_n . Here X_i is the conceptual value of the random variable in i -th repetition of this experiment. We assume throughout these notes that the random variables X_1, X_2, \dots, X_n are *iid*.

The notation θ is a generic one and is usually replaced by some other notation if it is more appropriate to do so. For example, if θ is the mean of a probability distribution, is usual to replace the notation θ by μ .

An *estimator* of θ is some function of the random variables X_1, X_2, \dots, X_n which is used to estimate that parameter. It is often written as $\hat{\theta}(X_1, X_2, \dots, X_n)$, a notation that emphasizes this dependence on X_1, X_2, \dots, X_n . Because of this dependence, $\hat{\theta}(X_1, X_2, \dots, X_n)$ is itself a random variable. For convenience we generally use the shorthand notation $\hat{\theta}$ for it.

Once the experiment is completed we will have the observed values x_1, x_2, \dots, x_n of these random variables. The quantity $\hat{\theta}(x_1, x_2, \dots, x_n)$, calculated from these is called the *estimate* of θ . We could denote it $\hat{\theta}_{\text{obs}}$, the suffix “obs” indicating that this is the value we happened to observe for this random variable once the experiment is completed. However a more specific notation is often used, as indicated in the next section.

Estimation of a mean

Perhaps the most important estimation procedure is that of estimating a mean. We discuss this in the present section.

Suppose that we wish to estimate the mean height μ of adult males. We plan to take a random sample of $n = 5$ adult males to do this. Before we take the sample, the five heights that we will get are random variables, and we denote them X_1, X_2, \dots, X_5 . Suppose that we decide that our estimator of μ will be the average of these. Then $\hat{\mu}(X_1, X_2, \dots, X_5) = \bar{X} = (X_1 + X_2 + \dots + X_5)/5$.

Suppose now that we have taken our sample of five adult males. We then have five

observed height values x_1, x_2, \dots, x_5 . Then the estimate of μ will be the average of these five observed heights, once the data are obtained, and would be denoted \bar{x} .

Thus the *estimator* $\hat{\mu}$ of μ is \bar{X} and the *estimate* $\hat{\mu}_{\text{obs}}$ of μ is \bar{x} . Note the use of the two different words *estimator* and *estimate*.

An estimator $\hat{\theta}$ is said to be an *unbiased* estimator of θ if its mean value $E(\hat{\theta})$ is equal to θ . Whether or not an estimator is unbiased is found from the theory of many random variables, and is not discussed further here. Standard estimators found in textbooks are usually unbiased. The first component in equations (??) shows that, for any probability distribution, \bar{X} is an unbiased estimator of μ .

Since the variance of \bar{X} decreases as the sample size n increases (see the second component in equations (??)), the estimator \bar{X} has a probability distribution more and more closely concentrated around μ as the sample size increases. This implies that the observed average \bar{x} , once the sample is taken, is more and more likely to be close to the unknown mean as the sample size n increases. This is in accordance with common sense.

However we often do not have large samples, and in any event a quantitative statement about how close \bar{x} is likely to be to μ depends on the *variance* of the distribution of the random variables involved. Unfortunately, in practice, this variance is usually not known. This implies that, before we can assess how close the observed average \bar{x} is likely to be to μ , we have to consider the question of the estimation of a variance.

Estimation of a variance

Just as the variance of a random variable is as important as its mean, the estimation of the variance of a random variable is as important as the estimation of its mean. In this section we discuss this estimation procedure, which is not as intuitive as estimation of a mean. Here we just give the final result, without the theory leading up to it.

Given n independent random variables X_1, X_2, \dots, X_n , all having the same probability

distribution with variance σ^2 , an unbiased estimator of σ^2 is S^2 , defined by

$$S^2 = \frac{X_1^2 + X_2^2 + \cdots + X_n^2 - n\bar{X}^2}{n-1}. \quad (17)$$

Correspondingly, given the observed values x_1, x_2, \dots, x_n of these random variables once the experiment of interest is completed, our estimate of σ^2 is s^2 , defined by

$$s^2 = \frac{x_1^2 + x_2^2 + \cdots + x_n^2 - n\bar{x}^2}{n-1}. \quad (18)$$

Putting it together

The two-standard-deviation rule (of thumb) given in (??), together with the formulae for the mean and variance of an average given in (??), show that to a close approximation,

$$\text{Prob} \left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + \frac{2\sigma}{\sqrt{n}} \right) \approx 0.95. \quad (19)$$

This is mathematically identical to saying

$$\text{Prob} \left(\bar{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{2\sigma}{\sqrt{n}} \right) \approx 0.95. \quad (20)$$

Normally we do not know the value of σ , but we use the fact that S^2 is an unbiased estimator of σ^2 (see previous section) to say, approximately, that

$$\text{Prob} \left(\bar{X} - \frac{2S}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{2S}{\sqrt{n}} \right) \approx 0.95. \quad (21)$$

All of this thinking is done before the experiment is carried out. We now carry out our experiment, get an observed value \bar{x} of \bar{X} and an observed value s^2 of S^2 . We then say: “We estimate μ by the value \bar{x} , and from (??) we are approximately 95% certain that μ is between $\bar{x} - \frac{2s}{\sqrt{n}}$ and $\bar{x} + \frac{2s}{\sqrt{n}}$.”

By doing this we not only have an unbiased estimate of μ , but also some idea of the precision of that estimate.

Estimation of a binomial parameter π

The estimation of a binomial parameter π is carried out by using the theory of the binomial distribution. Let X have the binomial distribution parameter π and index n . The fact that

the mean of X is $n\pi$ can be shown to imply that the mean value of X/n is π , and the fact that the variance of X is $n\pi(1 - \pi)$ can be shown to imply that the variance of X/n is $\pi(1 - \pi)/n$. Thus $\hat{\pi} = X/n$ is an unbiased estimator of π whose variance is approximately $\frac{X}{n}(1 - \frac{X}{n})/n$.

Thus if x successes were obtained when the n trials are carried out, the estimate of π is x/n and the estimate of the variance of $\hat{\pi}$ is

$$\frac{x(n - x)}{n^3}. \tag{22}$$

The procedure in the previous section using the two-standard deviation rule can then be used to give us, not only an estimate of π , but also two limits between which we are approximately 95% certain that π lies. An approximate 95% confidence interval for π calculated from this is

$$\left(\frac{x}{n} - 2\sqrt{\frac{x(n - x)}{n^3}}, \frac{x}{n} + 2\sqrt{\frac{x(n - x)}{n^3}} \right). \tag{23}$$

Hypothesis testing

General principles of hypothesis testing

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 relate to probabilistic calculations that set up the statistical inference process. We illustrate these steps by using the linkage example discussed above.

Step 1. The first step is to declare the null hypothesis H_0 and the alternative hypothesis H_1 . In the linkage example the null hypothesis is that disease and marker loci are unlinked. In terms of the recombination fraction θ between disease and marker loci, the null hypothesis claims that $\theta = 1/2$.

The alternative hypothesis is that disease and marker loci are linked. In terms of the recombination fraction θ , the alternative hypothesis claims that $\theta < 1/2$.

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.

A hypothesis can be *simple* or *composite*. A simple hypothesis specifies the numerical values of all unknown parameters in the probability distribution of interest. A composite alternative does not specify all numerical values of all the unknown parameters. In the linkage example, the null hypothesis “ $\theta = 0.5$ ” is simple while the alternative hypothesis “ $\theta < 0.5$ ” is composite. It is also *one-sided* ($\theta < 0.5$) as opposed to *two-sided* ($\theta \neq 0.5$). The “one-sided” alternative is the natural choice in the linkage case, but in other situations the natural alternative hypothesis might be two-sided.

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, to reject H_0 when it is true (a *Type I error* or *false positive*), or to accept H_0 when it is false (a *Type II error* or *false negative*). 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 dilemma is usually resolved in practice by observing that there is often an asymmetry in the implications of making the two types of error. In the linkage analysis case, for example, there might be more concern about making the false positive claim of a linkage between the disease and marker loci when they are in fact unlinked (a false positive statement), and less concern about making the conclusion that the two loci are not linked when in fact they are (the false negative statement). For this reason a procedure frequently adopted is to fix the numerical value of the Type I error, often denoted by α , 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.

For the test of a simple null hypothesis against a simple alternative, again with a fixed

sample size, the choice of the Type I error implicitly determines the numerical value of the Type II error, or equivalently of the *power* of the test, defined as the probability of rejecting the null hypothesis when the alternative is true.

When the alternative hypothesis is composite, Step 2 again consists of choosing the numerical value α of the Type I error. In this case there is no unique Type II error and thus no unique value for the power of the test. If the alternative hypothesis leaves the value of a parameter unspecified, the power of the test depends on the value of this parameter. In this case there is a *power curve*, giving the probability that the null hypothesis is rejected as a function of that parameter.

Step 3. Step 3 in the procedure consists of determining a *test statistic*. This is the quantity calculated from the data and whose numerical value leads to acceptance or rejection of the null hypothesis. In the linkage analysis example, one possible test statistic is the total number X of transmissions of the same gene from the heterozygous parent to his/her two affected offspring. There is a substantial body of statistical theory associated with the optimal choice of test statistic. We do not discuss this here - all test statistics discussed will be in some sense “optimal.”

Step 4. Step 4 consists of determining those observed values of the test statistic that 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. We illustrate this step with the linkage analysis example. Suppose that the total number X of transmissions of the same gene from the heterozygous parent to his/her two affected offspring is chosen as the test statistic. Then because this number should be unusually large when the alternative hypothesis is true, we will reject the null hypothesis if the observed value x of X 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 5%, K is

found from the requirement

$$\begin{aligned} \text{Prob (null hypothesis is rejected when it is true)} \\ = \text{Prob}(X \geq K | \pi = 0.5) = 0.05. \end{aligned} \quad (24)$$

For example, when $\pi = .5$ and $n = 100 =$ number of families considered, binomial tables show that the value of K satisfying this equation is $K = 59$.

For very large sample sizes, a normal approximation to the binomial might be employed. For example, if the number of families is 1,000, binomial tables are not available and a normal approximation to the binomial would be used.

Note that all the above steps are carried out before the data are seen. The procedures that they specify are independent of the data to be observed.

Step 5. The final step in the testing procedure is to obtain the data, and to determine whether the observed value of the test statistic is equal to or more extreme than the significance point calculated in Step 4, and to reject the null hypothesis if it is. Thus if for example in the sample of 100 families there were 61 cases where the informative parent passed on the same gene to his/her affected offspring, then with a Type I error of 5%, we would reject the null hypothesis and claim that we had significant evidence of linkage between disease and marker loci, since the observed number 61 exceeds the significance point $K = 59$ calculated above.

P-values

A testing procedure equivalent to that just described involves the calculation of a so-called *P-value*. In this procedure Step 4 in the above sequence, the calculation of the significance point K , is not carried out. Instead, once the data are obtained, we calculate the probability, assuming that the null hypothesis is true, 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. This procedure always leads to a conclusion identical to that based on the

significance point approach. For example, the calculations given above show that if $\pi = 0.5$, the probability of observing 61 or more families where the informative parent passed on the same gene to his/her affected offspring is 0.0179. This is then the P -value associated with the observed number 61. (This is the same probability as that shown on page 3.) If the Type I error had been chosen to be 0.05 the null hypothesis would then be rejected, since the P -value is less than this value. This conclusion agrees with that obtained using the significance point 59 calculated above.

Further examples of hypothesis testing

There are many statistical hypothesis testing procedures that are used frequently in science generally and in genetics in particular. Here we describe two of these.

Example 1. The two-sample t -test. A frequently-occurring case of testing composite hypotheses arises when we take observations from two groups, for example (in an expression array context) the expression levels of a certain gene in tissue of one type (i.e. group 1) and in tissue of some other type (i.e. group 2). Before the experiment is conducted the observations in each group are random variables. We assume that all random variables in group 1 are independent, and have a normal distribution with mean μ_1 and variance σ^2 . Similarly we assume that and all random variables in group 2 are independent, and have a normal distribution with mean μ_2 and variance σ^2 . The null hypothesis specifies that the means μ_1 and μ_2 of the two distributions are equal, but does not specify what the common value is. It also makes no specification about the (common) variance σ^2 . The alternative hypothesis depends on the context: depending on the context, it will be one-sided up ($\mu_1 > \mu_2$), one-sided down ($\mu_1 < \mu_2$) or two-sided ($\mu_1 \neq \mu_2$). In all three cases the alternative hypothesis is composite.

Suppose that we eventually take n_1 observations from group 1 and n_2 observations from group 2. The values of the n_1 observations in group 1 are denoted $x_{11}, x_{12}, \dots, x_{1n_1}$ and the

values of the n_2 observations from group 2 are denoted $x_{21}, x_{22}, \dots, x_{2n_2}$. The theory of composite tests shows that under the (normal distribution and equal variance) assumptions made, the appropriate test statistic is t , defined by

$$t = \frac{(\bar{x}_1 - \bar{x}_2)\sqrt{n_1 n_2}}{s\sqrt{n_1 + n_2}}, \quad (25)$$

with $\bar{x}_1 = \sum_{j=1}^{n_1} x_{1j}/n_1$, $\bar{x}_2 = \sum_{j=1}^{n_2} x_{2j}/n_2$, and with s defined implicitly from

$$s^2 = \frac{\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2 + \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2}{n_1 + n_2 - 2}. \quad (26)$$

The quantity $n_1 + n_2 - 2$ in the denominator of s^2 is the so-called “degrees of freedom” of s^2 . It is a central concept and we return to it again later. Under the assumptions made, the null hypothesis probability distribution of t is well known (as the t distribution with $n_1 + n_2 - 2$ degrees of freedom), and significance points of this distribution are widely available in t tables. This enables a convenient assessment of the significance of the observed value of t . If the alternative hypothesis is $\mu_1 > \mu_2$, a sufficiently large *positive* value of t leads to rejection of the null hypothesis. If the alternative hypothesis is $\mu_1 < \mu_2$, a sufficiently large *negative* value of t leads to rejection of the null hypothesis. If the alternative hypothesis is $\mu_1 \neq \mu_2$, a sufficiently large *positive or negative* value of t leads to rejection of the null hypothesis.

In practice it may not always reasonably be assumed that the random variables have normal distributions or that the variances of the two groups are equal. These two assumptions were made in the above t -test procedure, and the significance points of the t distribution as given in t tables were calculated assuming that both assumptions hold. Thus in practice it is not appropriate to use this t -test if one or both of these assumptions does not hold. This problem leads to the introduction of alternative testing procedures that do not rely on the normality and equal variances assumptions. The normal distribution assumption is discussed in the following section. The equal variance assumption causes complications that are not discussed here.

The ANOVA testing procedure is a direct generalization of the two-sided two-sample t test procedure described above. Before discussing it, it is convenient to consider the two-sided two-sample t test a little further.

First, the null hypothesis (of equal means) is rejected in the two-sided two-sample t test of the numerical value of t , whether positive or negative, is large enough. This is equivalent to rejecting the null hypothesis if t^2 is large enough. It can be shown, after some algebra, that

$$t^2 = \frac{n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2}{s^2}, \quad (27)$$

where $\bar{\bar{x}}$ is the overall average $(x_{11} + x_{21} + \cdots + x_{1n_1} + x_{21} + \cdots + x_{2n_2})/(n_1 + n_2)$.

The numerator in (27) can be thought of as a measure of the variation *between* the two groups. It is zero if and only if the averages of the observations in the two groups are equal. This would arise, for example, for the (very artificial) data values

Group 1: $x_{11} = 4, x_{12} = 7, x_{13} = 5, x_{14} = 4$, (average = $\bar{x}_1 = 5$),

Group 2: $x_{21} = 3, x_{22} = 8, x_{23} = 4$, (average = $\bar{x}_2 = 5$).

The quantity s^2 in the denominator in (27) is a measure of the variation *within* the two groups, as measured by the squares of the differences of the observations within any one group from the average in that group (see the definition of s^2 in (27)). If all the observations within any one group are equal, there is then zero variation within groups. This denominator would thus be zero with the (very artificial) data values

Group 1: $x_{11} = 4, x_{12} = 4, x_{13} = 4, x_{14} = 4$, (average = $\bar{x}_1 = 4$),

Group 2: $x_{21} = 6, x_{22} = 6, x_{23} = 6$, (average = $\bar{x}_2 = 6$).

In practice, with “real” data, neither the numerator nor the denominator in t^2 is likely to be zero.

The t test can then be described by saying that the null hypothesis is rejected if the variation *between* groups is sufficiently large relative to the variation *within* groups. The

ANOVA procedure follows exactly the same philosophy. We now turn to the details of what ANOVA is and how it works.

Example 2. ANOVA (the Analysis of variance)

ANOVA is one of the most frequently used statistical method for analyzing scientific data. Many complicated forms of ANOVA exist; here we consider only the simplest of these, namely the *one-way* ANOVA.

ANOVA considers data from any number of groups, not just two groups as in the t test described above. We denote the number of groups by k and the number of observations in group i by n_i . The data are then denoted as follows:

Group 1: $x_{11}, x_{12}, \dots, x_{1n_1}$ (average \bar{x}_1),

Group 2: $x_{21}, x_{22}, \dots, x_{2n_2}$ (average \bar{x}_2),

.....

Group k : $x_{k1}, x_{k2}, \dots, x_{kn_k}$ (average \bar{x}_k).

The grand overall average, namely

$$\frac{x_{11} + x_{12} + \dots + x_{1n_1} + x_{21} + x_{22} + \dots + x_{2n_2} + \dots + x_{k1} + x_{k2} + \dots + x_{kn_k}}{n_1 + n_2 + \dots + n_k}$$

is denoted $\bar{\bar{x}}$.

The variation *between groups* is the direct analogue of the numerator in (??). This is called the *between group* sum of squares (BGSS), and is defined by

$$\text{BGSS} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \dots + n_k(\bar{x}_k - \bar{\bar{x}})^2. \quad (28)$$

This component of the total variation has $k - 1$ degrees of freedom associated with it.

The variation *within groups* is the direct analogue of the numerator of s^2 (defined in (??)). This is the *within group* sum of squares (WGSS), defined by

$$\text{WGSS} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \quad (29)$$

The degrees of freedom of this quantity is the direct analogue of the corresponding degree of freedom value for s^2 in the t^2 statistic, namely $n - k$, where $n = n_1 + n_2 + \dots + n_k$.

Finally, the test statistic used in ANOVA, denoted F , is the direct extension of t^2 , and is defined by

$$F = \frac{BGSS/(k-1)}{WGSS/(n-k)}. \quad (30)$$

Notice that in the case of the t^2 procedure, where $k = 2$, the term $k - 1$ in the above is 1, and this is why it does not appear in the t^2 statistic.

What we have done above is to *analyze*, that is *subdivide*, the total variability in the observations into a between group and a within group component, and considered their *relative* values. More exactly, we will claim that the means of the k groups differ if the between group sum of squares is sufficiently large relative to the within group sum of squares, as measured by the value of F .

Tables of significance points of F are available, and this allows objective testing of the null hypothesis that the means of all k groups are the same.

Some general comments about hypothesis testing

Non-parametric (distribution-free) tests

The two-sample t -test procedure and also the ANOVA procedure described above both rely on the assumption that the observations in the test have a normal distribution. In many cases this is not a reasonable assumption. When it is not reasonable, a possible approach is to use a “non-parametric”, or equivalently a “distribution-free”, test. There are several non-parametric alternatives to the t -test and several non-parametric alternatives to the ANOVA test. One of the alternatives to the t -test is the *permutation test*. This test assumes equal variances in the two groups, and is carried out as follows.

When the null hypothesis is true, all possible $\binom{m+n}{m}$ distinct permutations of the data, for which m randomly chosen data values are taken as the “observations” for the first group and the remaining n taken as the “observations” for the second group, are equally likely. For each such permutation we calculate the value of the t statistic. (One of the values of t will be that arising for the permutation corresponding to the observed data.) If the alternative

hypothesis claims (for example) that the mean of the first group exceeds that of the second group, then using a Type I error 5%, the null hypothesis is rejected if the observed value of t is among the largest positive 5% of these permutation t values. If $\binom{m+n}{m}$ is too large to allow all possible combinations to be computed, a random sample of perhaps 10,000 combinations might be used instead.

When the two variances may not be assumed to be equal, further difficulties arise. One way around these is to use so-called “bootstrap” methods. These are not described here.

Multiple testing

A problem that can arise in the statistical analysis of linkage tests is that of *multiple testing*. We illustrate this issue with an example.

Suppose that we consider 1000 marker locus and test for linkage of each marker locus to a purported disease locus, as described above for each such test. If we choose a Type I error of 5% for each such test, and if the various linkage tests are independent, the experiment-wide Type I error is $1 - (.95)^{1000}$, and this is essentially equal to 1. Thus despite having started with a reasonable Type I error of 5% for each marker locus, the *experiment-wise* Type I error, namely the probability of claiming linkage to at least one marker locus, given that all marker loci are unlinked to the disease locus, is nowhere near 5%. Another way of viewing this problem is to say that even if no marker locus is linked to the disease locus, the procedure will nevertheless produce approximately $(.05) \cdot (1000) = 50$ false positive results. There appears to be little doubt that many “significant” findings of linkage in the literature are in fact no more than such false positive conclusions.

One approach to this problem is to attempt to obtain a reasonable experiment-wise Type I error. If we formulate a collection of (say) 1000 null hypotheses, null hypothesis j stating that marker j is unlinked to the disease locus, an experiment-wise Type I error of 5% could be achieved by using a Type I error of approximately 0.005% for each marker locus. However, this implies a quite stringent requirement for the rejection of any one of the null hypotheses, and might lead us to miss some marker loci that really are linked to the disease locus. This

is an example of the *multiple testing problem*.