Biostatistics for Genetics and Genomics

These slides provide a sketch outline of the material in the biostatistics handout notes. There are many further details in the notes which are not discussed in these slides.

Overview: What is Statistics?

Statistics is the activity of analyzing data in whose generation *chance,* or *randomness*, has played some part.

Thus statistics is important in genetics and genomics (and in the biological sciences generally), since chance events almost always arise in the generation of genetic and genomic data.

Because statistics analyses data in whose generation chance, or randomness, plays some part, it relies on a knowledge of probability theory. The relation between probability and statistics is shown on the next slide.

Probability is a *deductive* activity: we assume something, and deduce some consequence of it. Example: *Assuming that* the (unknown) cure rate for this new drug is 0.7, the probability that in a drug trial it cures 741 or more people out of 1000 is 0.0023, that is, it is pretty small.

Statistics is an *inductive* activity: Example: In the drug trial the new drug cured 741 out of 1000 people in the drug trial. *Based on the above probability calculation*, we may reasonably claim that the cure rate exceeds 0.7. Making this conclusion is a statistical activity, and it relies on the probability calculation above.

In general, every statistical procedure relies on some probability calculation, so we start with some probability theory.

First, some definitions.

Discrete random variables and their probability distributions

A discrete random variable is one which can only take some discrete set of values, which we denote here by

$$v_1, \ v_2, \ \dots, \ v_k.$$

The probability distribution of a discrete random variables consists of the above listing of the possible values of the discrete random variable together with their respective probabilities

$$p_1, \ p_2, \ \dots, \ p_k.$$

In practice these probabilities are often unknown to us, (so we have to use algebraic symbols for them). This implies that quantities defined in terms of them, for example the mean and the variance of the probability distribution (see later slides), are also unknown to us.

The binomial distribution

This is an important and frequently arising probability distribution. It arises when *all four* of the following conditions hold:-

1. We conduct a *fixed* number (n) of trials.

2. Each trial gives rise to one of two possible outcomes (which we conventionally call success or failure).

3. The probability of success, denoted by θ, is the same on all trials.

4. The outcomes of the various trials are independent.

What is θ? It is a *parameter*, that is, some unknown numerical value - see examples soon. In these slides and the accompanying notes, parameters (and only parameters) are always denoted by Greek letters.

(This is a notational convenience to help you remember if something is a parameter or is something else.)

The random variable in the binomial distribution is the total number of successes. Its possible values are 0, 1,…, n. The probability that it takes the value v *is*

$$\frac{n!}{v!(n-v)!}\theta^v(1-\theta)^{n-v}, \qquad v = 0,\ 1,\ …,\ n.$$

There are many other discrete probability distributions besides the binomial. However, it is the only one considered in this lecture.

Suppose that a random variable can take the possible values

$$v_1, \ v_2, \ \ldots, \ v_k$$

with respective probabilities

$$p_1, \ p_2, \ \ldots, \ p_k.$$

Then the *mean* (μ) of this random variable is defined as

$$\mu = v_1 p_1 + v_2 p_2 + \ldots + v_k p_k.$$

We use Greek notation for a mean because it is a parameter, that is a (usually unknown) quantity.

Notes on the mean.

1. Another expression for the mean is "the expected value".

2. A mean is a *totally different thing* from an average. *Never* confuse the two (although in the scientific literature the two are almost invariably confused). An average is defined later.

3. The mean is usually unknown to us. Much of statistics is involved with (i) estimating a mean and (ii) testing hypotheses about a mean.

4. The mean of the binomial distribution is $n\theta$.

Suppose that a random variable can take possible values

$$v_1, \ v_2, \ \ldots, \ v_k$$

with respective probabilities

$$p_1, \ p_2, \ \ldots, \ p_k.$$

Then the variance of this random variable is defined as

$$\sigma^2 \ = \ (v_1 - \mu)^2 \, p_1 \ + \ (v_2 - \mu)^2 \, p_2 \ + \ \ldots \ + \ (v_k - \mu)^2 \, p_k$$

We use Greek notation for a variance because it is a parameter, that is a (usually unknown) quantity.

The variance describes the "spread-out-ness" of a probability distribution relative to its mean – see Figure 2 on page 11 of the notes..

Other features:-

1.   The square root of the variance (denoted σ) is called the standard deviation of the probability distribution. It is often more useful than the variance itself.

2.   The variance (and thus the standard deviation) of a probability distribution is usually unknown to us.

3.   The variance of the binomial distribution is $n\theta(1-\theta)$.

The two-standard-deviation-rule

This is a fairly accurate rule of thumb. It states that many random variables are approximately 95% likely to lie within two of its standard deviations of its mean.

Example: The number of heads to arise when a fair coin is tossed 10,000 times has mean 5,000, variance 2,500, and thus standard deviation 50. Thus the probability that the number of heads will be between 4900 and 5100 is about 95%.

The proportion of successes

So far we have focused on the *number* of successes in the n binomial trials. In many situations we are more interested in the *proportion* of successes in the n trials.

This proportion is also a random variable. It has mean θ and variance $\theta(1-\theta)/n$.

Continuous random variables and their density functions

A continuous random variable can take any value in some continuous range of values. Continuous random variables are usually measurements, e.g. blood pressure. Associated with any continuous random variable is its (usually unknown) density function $f(x)$. The probability that the continuous random variable takes a value in the range (a,b) is the integral of the density function over that range. See page 13 of hand-out notes.

The mean and the variance of a continuous random variable are defined calculus operations that are not given here. The main things to remember are (i) that the mean has the same general interpretation as the mean in a discrete distribution (as the "center of gravity" of the distribution), (ii) the variance measures the "spread-out-ness" of the probability distribution, and (iii) in practice we rarely know a mean or a variance. (Later we consider *estimating* a mean and *testing hypotheses* about a mean.

The most important continuous probability distribution is the *normal*, or *Gaussian*, distribution, for which

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

is density function. The important particular case when μ = 0, σ = 1 is called the *standard normal distribution.* Its density function is shown on page 15 of the hand-out notes. All published tables for the normal distribution refer to this standard normal distribution. Any calculation relating to an arbitrary normal distribution can be reduced to a calculation from the standard normal distribution – details are not given in the notes.

Many random variables

If we wanted to test the hypothesis that the mean blood pressure of individuals with a certain disease is 133, we would not take a sample of just one person with the disease – we would take a sample of as many people with the disease as we reasonably could. We always denote the sample size by n. This means that we have to consider the probability theory for *many* random variables.

We think of the situation *before* we get our data. We conceptualize about the n blood pressure readings we will get, and denote these as
$$X_1, \ X_2, \ \ldots, \ X_n.$$
These are *random variables* – at this stage, before we draw our sample, we do not know what values they will take. We have followed the standard convention of always denoting random variables in upper case.

We can also conceptualize about the *average* of these quantities. (See next slide.)

This average is $(X_1 + X_2 + \ldots + X_n)/n$, and is denoted by $\overline{X}$.

This quantity is ALSO a random variable. It is NOT the same as the mean. In fact we often use it as the so-called *estimator* of the mean.

The three key properties of $\overline{X}$ are as follows.

1. If the mean blood pressure is μ, the mean of $\overline{X}$ is also μ. So we say that $\overline{X}$ is an *unbiased* estimator of μ.

2. If the variance of blood pressure is σ^2, the variance of $\overline{X}$ is σ^2/n. This decreases as *n* increases, showing that as *n* increases, the estimate of μ should get closer and closer to μ.

3. When *n* is large, $\overline{X}$ has a normal distribution (to a very close approximation).

Regression models

An important area in science is the question of how one thing depends on another. For example, we might ask how the growth height of a plant depends on the amount of water given to the plant during the growth period. Here the growth height is a random variable – we do not know in advance of the experiment what it will be, but the amount of water can be, and normally would be, chosen in advance of the experiment.

We therefore denote the growth height of some plant in upper case (as Y) and the amount of water in lower case (as x). The simplest regression model is that the mean of Y is a linear function of x. In algebraic terms,

$$\text{Mean of } Y = \alpha + \beta x.$$

Statistics

So far we have been contemplating the situation *before* our experiment. Thus we have been thinking of random variables and their properties.

We now do our experiment. We now have the observed values of these random variables, for example the observed number of people cured by the drug, the observed blood pressures of the n people in the blood pressure example, or the observed growth heights of n plants given different amounts of water.

Notational convention

We denote random variables by upper case letters. A standard convention is to denote the observed values of these random variables after the experiment is carried out by the corresponding lower case letters. In the binomial case we denote the observed number of successes by x and the observed proportion of successes by p. In the blood pressure example, we denote the actually observed blood pressures by x_1, x_2, …, x_n. In the plants example we denote the observed growth heights by y_1, y_2, …, y_n.

The two main operations of statistics are to use the data to (i) estimate (unknown) parameters and (ii) to test hypotheses about (unknown) parameters. To see how we do this appropriately, we have to consider properties of the random variables before the experiment was conducted. Thus in doing statistics we are continually referring back to probability theory.

Estimation of the binomial parameter θ

We estimate θ by the observed proportion p of successes. This is a natural thing to do, but it also has theoretical support: the mean of the (random) proportion P is θ, so in using p to estimate θ we are "shooting at the right target".

More formally, p is an unbiased estimate of θ.

The standard deviation of P, together with the two standard deviation rule, is used to get an approximate "95% confidence interval for θ". This interval goes from

$$p - 2\sqrt{p(1-p)/n}$$

to

$$p + 2\sqrt{p(1-p)/n}$$

Estimate of a mean μ

We estimate the mean of some distribution by the average $\overline{x}$ of n observed values from this distribution. This estimate is unbiased, since the mean of the corresponding random variable $\overline{X}$ is μ. The approximate 95% confidence interval for μ is given on page 21 of the notes (equation (19)).

In the regression case the most interesting parameter is β, since this gives the mean growth height per unit amount of water. We estimate β from the observed growth heights y_1, y_2, …, y_n and the corresponding amounts of water x_1, x_2, …, x_n by *b*, defined by

$$b = \frac{\sum\limits_{i=1}^{n} y_i (x_i - \bar{x})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}$$

Hypothesis testing

There are five steps in any hypothesis testing procedure. These are:-

1. State the null and alternative hypotheses (before the data are seen).
2. Choose a Type I error (the false positive rate).
3. Determine which test statistic will be used.
4. Determine which values of the test statistic will lead you to reject the null hypothesis.
5. Get the data and carry out the test.

There are many complicated details involved in these steps. See pages 22-26 of the handout notes. These pages also describe an example concerning a test of hypothesis about a binomial parameter θ.

The test of the equality of two binomial parameters is described on pages 28-30 of the handout notes.

Tests on means

The one-sample t test.

Here we wish to test the null hypothesis that the mean of some probability distribution is some specified numerical value, denoted by μ_0. In practice the variance of this distribution is not known, and has to be estimated (by s^2, see equation (25) on page 31 of the notes).

The test statistic is t, defined by

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

Think of this as a "signal-to-noise" ratio. The signal is $(\bar{x} - \mu_0)$ and the noise is s.

The two-sample t test

Here we have observations x_1, x_2, .., x_m from a distribution with mean μ_1 and observations y_1, y_2, .., y_m from a distribution with mean μ_2. The null hypothesis claims that $\mu_1 = \mu_2$. The signal in the relevant t statistic is the difference

$$\bar{x} - \bar{y}$$

between the two sample averages. (See equation (28) on page 34 for more details.)

Non-parametric (distribution-free) tests

These are used when we don't wish to assume that our observations come from a normal distribution. (Use of any t test DOES make the normal distribution assumption.) The two tests that are described are non-parametric alternatives to the two-sample t $test$, namely the permutation test and the Mann-Whitney test.

Hypothesis testing in regression

If the amount of water given to a plant has no effect on its growth height, then $\beta = 0$.

Thus a test of the (null) hypothesis that the amount of water has no effect on plant height is the same as a test of the (null) hypothesis $\beta = 0$. This test is described on pages 37-38 of the notes.