

## Chapter 5

# Hypothesis Testing

Major factors effecting metabolism, measured as oxygen consumption, are body mass and ambient temperature. In mammals, the relationship of metabolism to body mass has been established for many species at various temperatures. These relationships can be used to predict or hypothesize the metabolism of a species that has never before been measured. Assume that these relationships are used to predict (hypothesize) a metabolism of 10 units for the newly discovered species, *Fattus rattus*. Assume further that you measure the oxygen consumption of a sample of *Fattus rattus* under the same conditions as used for the prediction. You find that in a sample of 150 animals the mean oxygen consumption is 12 with a standard deviation of 2.

These data may then be used to test the **null hypothesis** (abbreviated as  $H_0$ ) that the 150 oxygen consumption values from the animals sampled came from a population having a mean of 10. The word "null" indicates that the hypothesis is one of "no difference" between the parameter of the population (the mean oxygen consumption) and the predicted or hypothetical oxygen consumption. The **alternate hypothesis** (abbreviated as  $H_a$ ) is that the mean oxygen consumption of this species is some value other than 10. The null and alternate hypotheses for our example problem can be written in short-hand notation as follows.

$$H_0: \mu = 10$$

$$H_a: \mu \neq 10$$

Statistically one can generally test only the null hypothesis and not the alternate. Basically the statistical procedure is to estimate the probability that  $H_0$  is true. Thus, in our example an appropriate statistical test is used to estimate the probability of drawing from a population having a mean oxygen consumption of 10 a random sample of 150 individual oxygen consumption values with a mean of 12. If the probability is very large of obtaining a sample of 150 individuals with a mean of 12 from a population with a mean of 10, then it may be reasonably concluded that  $H_0$  is true. If so,  $H_0$  is accepted and  $H_a$  is rejected. Biologically, this would lead to the conclusion that the species *Fattus rattus* has an oxygen consumption not statistically different from that projected for any mammal of that size measured at that temperature. Alternatively, if the probability of obtaining a sample of 150 individuals with a mean of 12 from a population with a mean of 10 is very small then it may be reasonably concluded that the  $H_0$  is false. In this case, it is the  $H_a$  that is accepted and the  $H_0$  is rejected. Biologically, this leads to the conclusion that the species *Fattus rattus* has an oxygen consumption statistically different than that projected for a mammal of that size measured at that temperature.

The question remains as to what constitutes a small enough probability to conclude that the  $H_0$  is false. The probability conventionally used in statistics for this conclusion is 0.05 or less. That is, if the appropriate statistical test of  $H_0$  results in a probability that is 0.05 or less, then  $H_0$  is rejected and  $H_a$  is accepted. Alternatively, if the appropriate statistical test of  $H_0$  results in a probability that is larger than 0.05, then  $H_0$  is accepted and  $H_a$  is rejected. This is a key rule in statistics and one that all students must know. So memorize the following rule:

$$P > 0.05$$

accept  $H_0$ , reject  $H_a$

no significant difference

$$P < 0.05$$

reject  $H_0$ , accept  $H_a$

significant difference

One important thing to remember in hypothesis testing is that a hypothesis is never proved only supported or refuted. The reason for this is that there is always the possibility of making a mistake. In fact, there are two potential mistakes that can be made: we could accept the  $H_a$  when in fact  $H_0$  is true (**Type I Error**) or we could do the opposite and accept the  $H_0$  when in fact the  $H_a$  is true (**Type II Error**). In fact, there is always the change that a sample will be very different from the population from which it was collected so the first type of error, Type I, can never be completely avoided. It is therefore necessary to determine how certain we want to be that a Type I Error is not being made. As with most things, there is a trade-off. By decreasing the likelihood of making a Type I Error you increase the likelihood of making a Type II Error, or accepting  $H_0$  when  $H_a$  is true.

By convention, scientists use 5% as the level of error that is deemed acceptable. When stated in this way, we say that the significance level ( $\alpha$ ) is 5%. Regions of a frequency distribution that are beyond the 5% significance level lead to rejection of  $H_0$ . In hypothesis testing we are always assuming the  $H_0$  is true. For instance, we could calculate the probability of obtaining a sample with a particular  $\bar{x}$  from a population with a given  $\mu$ , assuming that the sample came from that population, or  $H_0: \bar{x} = \mu$ . To do this one could calculate the normal deviate,  $Z$ , using estimates of central tendency and dispersion of the population based on the sample. The appropriate calculation of  $Z$  would be

$$Z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad [5.1]$$

As described in Chapter 4, the  $s_{\bar{x}}$  is the standard error of the mean which can be estimated as the standard deviation divided by the square root of the sample size (Equation 3.4). Recall that the hypothetical mean oxygen consumption, or  $\mu$ , was 10 while the measured mean,  $\bar{x}$ , was 12 with a standard deviation ( $s$ ) of 2 based on a sample of 150. Then,

$$Z = \frac{(12 - 10)}{\frac{2}{\sqrt{150}}} = \frac{2}{0.16} = 12.5$$

The probability associated with a  $Z$  of 12.5 is very low; the probability is so low in fact that a  $Z$  of 12.5 is not even in most tables (Table 1, Appendix B). So the probability would be reported as being less than 0.001, written as  $P(Z > 12.5) < 0.001$ . Thus, the probability is less than 0.001 that the null hypothesis of  $\mu=10$  is true. In other words, there is a very small probability that we would have a sample with a  $\bar{x}=12$  that came from a population with a  $\mu=10$ . It is not impossible, of course, but based on the 0.05 rule, we would reject the null hypothesis and accept the alternate hypothesis. Biologically, this leads us to the conclusion that the species *Fattus rattus* has an oxygen consumption higher than that projected for a mammal of that size measured at that temperature. Since we are ultimately interested in what data tell us about the biological phenomenon being investigated, it is important to remember to always interpret any statistical analysis in terms of the biology and not just the statistics.

The calculation of  $Z$  requires knowledge of the parametric standard deviation, which typically is not known. However, the statistic standard deviation is considered a good estimator of the parametric standard deviation when the sample size is 120 or larger. However, for most biological situations, the sample sizes are not this large and  $t$  distribution must be used instead of the standardized normal distribution ( $Z$  distribution). The test statistic computed is  $t$  instead of  $Z$ , and  $t$  is computed as

$$t = \frac{(\bar{x} - \mu)}{s_x} \quad [5.2]$$

The  $t$  distribution has different shapes for different degrees of freedom,  $df$  (Figure 5.1). For hypotheses concerning a single mean, the degrees of freedom are equal to the sample size minus 1, or  $df = n - 1$ . The  $t$  distribution has a greater concentration of values around the mean and in the tails than does the normal distribution at low degrees of freedom, but it tends to resemble the normal distribution more closely as  $n$  increases. This can be seen in Figure 5.1 which shows the  $t$  distributions for  $df$  equal to 4, 8, 50, and infinity. The  $t$  and  $Z$  distributions are basically the same at sample sizes of about 120 so one typically uses  $t$  when  $n < 120$  and  $Z$  when  $n > 120$ . A table of  $t$  values and associated probabilities is located in Appendix B (Table 2), and an abbreviated version is provided below for convenience (Table 5.1).

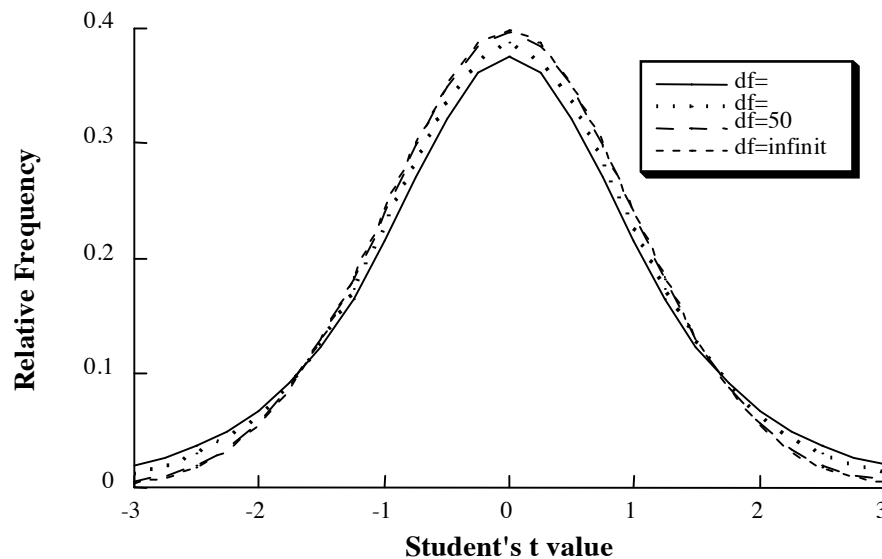


Figure 5.1. The  $t$  distribution for degrees of freedom of 4, 8, 50 and infinity. The distribution at infinity is the same as for a normal deviate,  $Z$ .

Table 5.1. Cumulative t distribution for relatively few degrees of freedom (see Table 2, Appendix B for expanded version). The body of the table contains t values at different degrees of freedom (*df*, left column) at decreasing cumulative one-tailed,  $\alpha(1)$ , and two-tailed,  $\alpha(2)$ , probabilities.

|             |       |       |       |       |       |       |       |       |        |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| $\alpha(2)$ | 0.3   | 0.25  | 0.2   | 0.15  | 0.1   | 0.05  | 0.02  | 0.01  | 0.001  |
| $\alpha(1)$ | 0.15  | 0.13  | 0.1   | 0.08  | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
| <b>df</b>   |       |       |       |       |       |       |       |       |        |
| 5           | 1.156 | 1.301 | 1.476 | 1.699 | 2.015 | 2.571 | 3.365 | 4.032 | 6.869  |
| 6           | 1.134 | 1.273 | 1.440 | 1.650 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959  |
| 7           | 1.119 | 1.254 | 1.415 | 1.617 | 1.895 | 2.365 | 2.998 | 3.499 | 5.408  |
| 8           | 1.108 | 1.240 | 1.397 | 1.592 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041  |
| 9           | 1.100 | 1.230 | 1.383 | 1.574 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781  |
| 10          | 1.093 | 1.221 | 1.372 | 1.559 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587  |

The construction of a t distribution can be imagined in a similar fashion as was described for the standardized normal distribution. Imagine again constructing a huge known population with some variable, say length, which has a known  $\mu$  and  $\sigma$ . Then imagine taking a sample of a fixed size (e.g. 5) from this population, measuring the length of each individual, computing the mean for the sample, and then calculating the t using Eqn 5.1. Imagine doing this a zillion, zillion times, and then plotting on the Y-axis the frequency of times that a particular t score on the X-axis is observed. If the values on the Y-axis, the frequencies, are divided by the total number of t's obtained (e.g. zillion, zillion) then that axis becomes relative frequency. The resulting plot would be identical to the one shown for  $df=4$  in Figure 5.1.

Returning again to metabolic measurements, assume that the measurements of metabolism were made on only nine individuals of *Fattus rattus*, and that the mean was determined to be 12 with a standard deviation of 2. The Z test could not be used since the parametric standard deviation is not known and the sample size is below 120. In this case, the t test (5.3) would have to be used to test  $H_0: \mu=10$  and  $H_a: \mu \neq 10$  as follows where the test statistic,  $t$ , has 8 degrees of freedom.

$$t = \frac{(12 - 10)}{\left( \frac{2}{\sqrt{9}} \right)} = \frac{2}{0.67} = 3.0$$

The t table (Table 5.1 and Table 2, Appendix B) shows that at  $df=8$  a  $t$  value of 3 lies between probability values of 0.01 and 0.02 for a two-tailed test, indicated with the row labeled  $\alpha(2)$ , and between 0.005 and 0.01 for a one-tailed test, row labeled  $\alpha(1)$ .

For either one-tailed or two-tailed test, the probability of the test statistic  $t$  is below the 0.05 level. Thus, the null hypothesis would be rejected and the alternate hypothesis of  $\mu \neq 10$  accepted. Biologically, the t test, like the preceding Z test, leads to the conclusion that the species *Fattus rattus* has an oxygen consumption different from that projected for a mammal of that size measured at that temperature.

## Two-tailed vs. one-tailed tests

The previous example is a two-tailed test since the alternate hypothesis,  $H_a$ , was simply that  $\mu$  did not equal 10 (it could be either higher or lower). The  $t$  table (Table 5.1) indicates that at  $df=8$  the  $t$  value associated with a probability of 0.05 for a two-tailed test [ $\alpha(2)$ ] is 2.306. This  $t$  value is referred to as the *critical value* since it is the value against which the calculated value of  $t$  is compared. If the calculated value of  $t$  is equal to or greater than the critical one then the probability that the null hypothesis is valid is less than 0.05.

In many cases the interest lies in whether  $\mu$  is significantly larger or smaller than a particular value. In these cases we employ one-tailed or one-sided hypothesis. For example, if *Fattus rattus* is from a very cold climate, it could be hypothesized to have a higher metabolism than projected based on other mammals. For this one-tailed test, the null and alternate hypotheses are stated as:

$$H_0: \mu \leq 10 \qquad H_a: \mu > 10$$

Here the  $H_0$  states that the mean oxygen consumption will be less than or equal to 10, while the  $H_a$  states that the mean will be greater than 10. In our example problem we would also reject this  $H_0$  since the computed  $t$  value of 3.0 is greater than the critical one-tailed  $t$  value of 1.860 listed in Table 5.1 when  $df=8$ .

A one-tailed test requires a smaller difference between a hypothetical value and the sample mean to accept the alternative hypothesis. This is because the *critical t* is always lower for a one-tailed test. This is illustrated in Figure 5.2 with a  $t$  distribution for  $df=8$ . From Table 5.1 the *critical t* at the 0.05 level for a one-tailed test is 1.860 while it is 2.360 for a two-tailed test. The one-tailed test the 0.05 level is marked on the right side only at the *critical t*=1.860, and 5% of the area under the entire curve is above  $t=1.860$  (Figure 5.2). For the two-tailed test, the 5% is made up of 2.5% on both tails (two-tailed!) and the critical level is either +2.360 or -2.360.

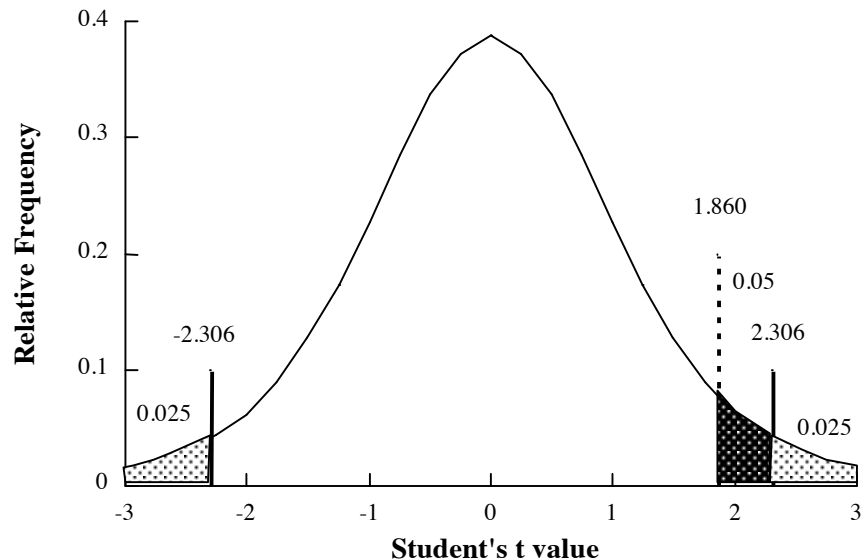


Figure 5.2. Distribution of  $t$  ( $df=8$ ) showing the critical regions (shaded areas) for a one-sided and two-sided test at  $\alpha=0.05$ .

Typically, experimental biologists "want" to accept the alternative hypothesis, which is easier to do with a one-sided test. To avoid this bias it is important that one generates the hypothesis on theoretical grounds before observing any data.

Remember, hypotheses are two-tailed (two-sided) if the null or alternative hypothesis contains no indication of direction of difference. Conversely, the hypotheses are one-tailed (one-sided) if the null or alternative hypothesis contains indication of direction of difference (greater than or lesser than).

### Confidence Limits

Although the sample mean is the best estimate of  $\mu$ , it is still only an estimate. The accuracy of the estimate can be expressed by computing confidence intervals, and typically 95% confidence intervals are computed since the 0.05 level is used for acceptance or rejection of the hypotheses. The 95% confidence intervals are computed as

$$\bar{x} \pm (s_{\bar{x}})(t_{\alpha(2)=0.05, df=n-1}) \quad [5.4]$$

where  $\bar{x}$  and  $s_{\bar{x}}$  are the sample mean and standard error, and the  $t$  value is associated with a probability of 0.05 when  $\alpha(2)$  and  $df$  is equal to  $n-1$ . Note that it is always the two-tailed  $t$  value which is used for confidence interval calculations since limits are being set to either side of the mean. The quantity,

$$\bar{x} - (s_{\bar{x}})(t_{\alpha(2)=0.05, df=n-1}) \quad [5.5]$$

is called the *lower 95% confidence limit* while the quantity,

$$\bar{x} + (s_{\bar{x}})(t_{\alpha(2)=0.05, df=n-1}) \quad [5.6]$$

is the *upper 95% confidence limit*. This confidence interval allows one to be 95% confident that the real parametric mean,  $\mu$ , lies between the two limits.

In our example, the sample mean and standard error are 12 and 0.67 based on a sample of 9. The  $t$  value for 8 degrees of freedom with a probability of 0.05 at  $\alpha(2)$  is 2.306; thus, the 95% confidence interval for this sample is computed as

$$12 \pm (0.67)(2.306) = 12 \pm 1.54 = 10.46 - 13.54$$

Therefore, we are 95% confident that the real parametric mean,  $\mu$ , of oxygen consumption of *Fattus rattus*, lies between 10.46 and 13.54.

It is obvious from examining Eqn 5.5 that the smaller the standard error,  $s_{\bar{x}}$  the smaller the confidence interval, or the more accurate the estimate of  $\mu$ . The calculation of the standard error,  $s_{\bar{x}}$ , also indicates that the larger the sample size the smaller the confidence interval. Thus, a parametric estimate based on a large sample is more accurate than one based on a small sample. Besides being mathematically correct, it makes intuitive sense that the larger the sample the more accurate the estimate.

## Problem Set – Hypothesis Testing

For each of the following problems conduct an hypothesis test and determine the 95% confidence intervals of the sample. You should be able to perform these calculations by hand. Use the descriptive statistics excel template to check your answers. You will not turn in any of these problems, but we will go over the answers next lecture period.

- 5.1) The following data are the lengths of the menstrual cycle in days in a random sample of 15 human females. Determine if the mean length of human menstrual cycles is equal to a lunar month (a lunar month is 29.5 days).

Data: 26, 24, 29, 33, 25, 26, 23, 30, 31, 30, 28, 27, 29, 26, 28

- 5.2) The following data are osmotic concentrations ( $\text{mmol}\cdot\text{g}^{-1}$ ) of the blood of an arthropod. Determine if the mean osmotic concentration of the blood of this animal is the same as the osmotic concentration of the water in which it lives. (The environmental osmotic concentration is  $0.64 \text{ mmol}\cdot\text{g}^{-1}$ .)

Data: 0.66, 0.69, 0.68, 0.71, 0.70, 0.68, 0.66, 0.67, 0.63.

- 5.3) Body temperatures ( $^{\circ}\text{C}$ ) were obtained from a sample of eight inter-tidal crabs exposed to air at  $26.2^{\circ}\text{C}$ . Determine if the mean body temperature of this species of crab under these conditions is less than  $26.2^{\circ}\text{C}$ .

Data: 25.8, 24.6, 26.1, 24.9, 25.1, 25.3, 24.0, 24.5.

- 5.4) Twelve men were placed on a weight-reducing diet. The changes in weight (Kg) exhibited by the men were as follows. Determine whether the diet is successful.

Data: -1.2, -1.4, -1.0, -0.4, -0.3, -0.8, 0.5, 0.1, -0.9, -1.8, 0.0, -2.1.