8. PAIRED-SAMPLE INFERENCE

Two-sample vs. Paired-sample Designs

As mentioned in a previous lab, there are two ways of comparing means of two treatments. In the two-sample design, independent observations are assigned to one or the other treatment. The two random samples of individuals are therefore from separate populations, and our goal is to compare the means of these two populations, μ_1 and μ_2. In the "split plot" or paired-sample design, both treatments are applied to each independent unit (e.g., patient, or field plots) in the random sample. As before, we are interested in comparing population means of two treatments but the two measurements made on the same patient (or in each field plot) are no longer independent.

The difference between these two approaches is crucial, and affects the statistical method used to test for treatment effects. When a paired design is used, the two measurements made on each individual at the end of the experiment must be reduced to a single number: the change, or difference *d*, between the two measurements. We then use the familiar one-sample methods to estimate the mean difference μ_d and/or test hypotheses about the mean difference. Paired-sample inference is a straightforward extension of one-sample methods learned in a previous lab exercise. Methods for dealing with two-sample experiments were covered in the previous lab exercise.

Confidence Interval for a Mean Difference

The confidence interval for the mean difference μ_d between paired measurements is obtained in the same way as that for a single population mean. We simply treat our sample of differences for what it is: a random sample from a single population. Thus, for paired data the 95% confidence interval for the mean difference is:

$$\bar{d} - t_{0.05(2),\nu}\, s_{\bar{d}} \le \mu_d \le \bar{d} + t_{0.05(2),\nu}\, s_{\bar{d}}$$

where μ_d is the parameter for the mean difference between measurements, $\bar{d}$ is the sample mean difference, $s_{\bar{d}}$ is the standard error of the sample mean difference, ν is the degrees of freedom (n - 1). As before, this interval assumes that the data are from a normally distributed population. If the data are not from a normal population then the computer confidence interval is approximate, and is expected to be accurate only when *n* is large (by the Central Limit Theorem).

Hypothesis Testing for a Difference

The **paired-sample t-test** is appropriate for testing Ho: $\mu_d = 0$ vs. Ha: $\mu_d \neq 0$ (and corresponding one-tailed hypotheses) when the population of differences *d* has a **normal** distribution. Standard methods should therefore be applied to the random sample of *d* values to test the validity of this assumption. The one-sample *t*-statistic is our measure of discrepancy between the sample mean $\bar{d}$ and the value of μ_d stated in the null hypothesis:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

If *d* has a normal distribution, then *t* has a *t*-distribution *n*−1 degrees of freedom, where *n* is the sample size (number of independent observations).

What if *d* does not have a normal distribution in the population? If *n* is large then the distribution of $\bar{d}$ is nevertheless approximately normal (by the Central Limit Theorem) and we may still use the paired-sample *t*-test as above. If *d* is not normally distributed and sample size is not large, then the best approach is to use a **non-parametric** test (also called **rank** test) instead. These methods assume only that the data are a random sample from a continuous distribution, but this distribution need not be normal.

The **Wilcoxon signed rank test** (also called the **Wilcoxon paired-sample test**) is the most powerful non-parametric analogue of the paired *t*-test. Its power is about 95% of that of a paired *t*-test under ideal conditions. See your textbook for a worked example of this test. Briefly, the test is carried out as follows. First, the absolute values of the differences *d* are ranked. Two sums are then computed. The first sum, T_+, is the sum of the ranks corresponding to positive values of *d*. The second sum, T_-, is the sum of the ranks corresponding to negative values of *d*. With a two-tailed test of "Ho: no difference between treatments; Ha: a difference between treatments exists", the null hypothesis is rejected if the smaller of the two sums is less than or equal to the critical value (in Zar the critical values are provided in Appendix Table B12). **Notice that the statement of the null and alternative hypotheses do not refer to the mean, μ_d.** This is because, strictly speaking, the Wilcoxon signed rank test is not a comparison of means. Rather, it compares the rank sums T_+ and T_-, which, under the null hypothesis, should be roughly equal.

The **sign test** is even simpler than the Wilcoxon signed rank test, and is really just an application of the familiar binomial test. We record whether the differences *d* are positive or negative. Under the null hypothesis of no difference, the number of positive *d*-values should be roughly equal to the number of negative *d*-values. Let *p* be the proportion of differences that are positive. Under the null hypothesis of no difference, Ho: $p = 0.5$, whereas Ha: $p \neq 0.5$ under the alternative hypothesis. This test is really only used as a last resort because it less powerful even than the Wilcoxon signed rank test, and is much less powerful than the paired *t*-test.

Using the Program

To carry out the paired sample t-test or its non-parametric analogue you will need to enter both measurements for each individual in separate columns on the same row. Then create a new variable computed as the difference between the paired measurements. Then proceed as in the earlier lab exercise on one-sample tests.

Problems

1. Before proceeding with further research into the mechanisms regulating erythrocyte pH in toads (*Bufo marinus*), scientists compared two methods of measuring intracellular pH to determine whether or not the methods give the same results. Arterial blood (0.8 ml) was collected from a random

sample of 37 toads. Each sample was equally divided and erythrocyte pH in each aliquot was determined either by a freeze-thaw (FT) method or a method involving C^{14}-labelled 5,5-dimethyl-2, 4-oxazolidinedione (DMO). The data are stored in the file **toads.jmp** on the shared drive. Each row corresponds to a different toad.

a) Test whether the two methods of measuring erythrocyte pH give the same results on average, using the most powerful test available. Show all steps.

b) What assumption is required in (a)? Visually examine the data for departures for this assumption. Is your assumption met? Explain.

c) Carry out a test of your assumption given in (b). Do the results of the test match your visual interpretation? Recommend a strategy for testing differences between the two methods on the basis of your results.

d) Calculate the 95% confidence interval for the difference between means. Based on your evaluation in (b) and (c), is the interval likely to be accurate? Explain.

2. Scientists studying the effect of slash burning examined the diversity of spiders in clear-cut areas of coastal forests. The number of species of spiders was measured at 27 sites of equal size (1.4 ha). The sites were then burned. Four years later the number of spider species at each site was measured again. The results are stored in the file **spider.jmp** on the shared drive.

a) When examining changes in spider diversity between the period before burning and four years after burning, is a one-tail test or a two-tail test most appropriate?

b) Was there a significant change in number of species of spiders between the two sampling periods? Explain how you chose your method for testing.

c) Suggest an improved experimental design to determine the effects of burning on diversity of spiders in clear-cuts.

d) Comment on the advantages and disadvantages of a paired sampling design such as the one used here over a two-sample design in which the experimenter simply compare spider diversity in burned plots with those in other plots not burned?

e) Compute the 95% confidence interval for the change in number of species. Is this interval likely to be accurate? Explain.

3. In insect species whose females mate with multiple males (*polyandrous*), male seminal fluid contains toxins that increase the proportion of fertilizations a male obtains relative to other males mating with the same female. However, these toxins reduce the survival of females. Experiments have shown that over multiple generations females evolve defenses to prevailing male toxins, but that males forever evolve new toxins.

The result is a long-term “arms race” between the sexes. Researchers have postulated that this process in polyandrous insect species should speed the rate at which sterility barriers evolve between different populations of that species, increasing the rate at which new species are formed. In contrast, sterility barriers should evolve more slowly in insect species whose females mate only once (*monandrous*), since no arms race between the sexes occurs, yielding a lower rate at which new species are formed. To test this idea, Arnqvist et al. (2000, *Proc. Natl. Acad. Sci.* 97:10460–10464) compared the total numbers of species in 25 pairs of insect taxa. Each pair consisted of two closely related “clades” (a clade is a group of species all of which share a common ancestor). One of the clades of each pair contained only polyandrous species, whereas all of the species in the other clade of the pair were monandrous. The number of species in each pair of clades is provided in the file **conflict & speciation.jmp**. These data were taken directly from Table 1 in Arnqvist et al. (2000).

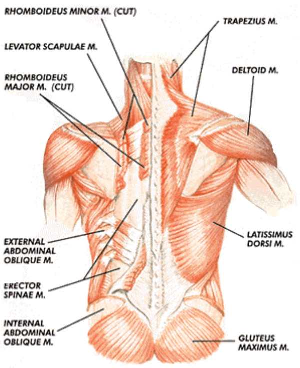
a) Using these data, test whether the number of species in polyandrous clades is significantly different from the number in monandrous clades. Use a two-tailed test. Justify your choice of method by testing appropriate assumptions.

b) Repeat the exercise in (a) using the **log** number of species in clades instead. How did this affect the best procedure for testing the hypotheses? Does your conclusion differ? [We will be investigating the use of data transformations like the logarithm more thoroughly in a later lab exercise.]

c) Carry out a test of the same hypotheses using the **sign test** (a.k.a., the **binomial test**) in JMP IN. How do your results compare with those from the previous tests?

d) Were the authors justified in concluding that polyandrous taxa of insect species have more species than related monandrous taxa?

4. The human species is polymorphic for the ACE gene (angiotensin-converting enzyme, functioning in human skeletal muscle). Two alleles (alternative states of the gene) are present. The “I” allele carries an insertion of 287 base pairs not present in the “D” allele. This longer “I” allele leads to lower enzyme activity and enhanced endurance under intense exercise training. In a recent study, researchers measured training-related changes in the mechanical efficiency of human skeletal muscle (energy used per unit power output) in “II” and “DD” type individuals. Thirty “II” type individuals were randomly sampled from a population of young Caucasion male army recruits. Thirty “DD” type individuals were sampled from the same population. Measurements of mechanical efficiency of skeletal muscles were made on all 60 individuals before and after an 11-week programme of aerobic physical training. Neither the subjects nor the staff knew the genotypes of the 60 individuals (i.e., the study was ‘double blind’). The change in mechanical efficiency for the 60 males are provided in the data file **ace.jmp**.


RHOMBOIDEUS MINOR M. (CUT)
LEVATOR SCAPULAE M.
RHOMBOIDEUS MAJOR M. (CUT)
TRAPEZIUS M.
DELTOID M.
EXTERNAL ABDOMINAL OBLIQUE M.
ERECTOR SPINAE M.
INTERNAL ABDOMINAL OBLIQUE M.
LATISSIMUS DORSI M.
GLUTEUS MAXIMUS M.

a) With these data, test whether change in mechanical efficiency of muscle was different between the “II” and “DD” groups of males.

b) What are your assumptions in (a)? Test these assumptions.