

## Chapter 5

### Schistosomiasis: An Introduction to Randomization and Permutation Tests

Schistosomiasis (shis-tuh-soh-mahy-uh-sis) is a disease occurring in humans caused by a parasitic flatworm. Schistosomiasis affects about 200 million people world wide and is a serious problem in sub-Saharan Africa, South America, China, and Southeast Asia. The disease can cause death, but more commonly results in chronic and debilitating symptoms, caused primarily by the body's immune reaction to parasite eggs lodged in the liver, spleen, and intestines.

Currently there is one drug, praziquantel, in common use for treatment of schistosomiasis; it is cheap and effective. However many organizations are worried about relying on a single drug to treat a serious disease which affects so many people worldwide. Drug resistance may have prompted a 1990s outbreak in Senegal, where cure rates were low. A promising drug that, in theory, might treat the disease is K11777, which is under evaluation as a drug candidate for another parasitic infection (Chagas' disease).

In this lab, we will analyze a study where the researchers wanted to find out whether K11777 had any activity against schistosome worms. In one phase of the study, 10 female laboratory mice and 10 male laboratory mice were deliberately infected with the schistosome parasite. Within each sex, 5 mice were randomly assigned to a treatment (Trt) of K11777 injected in a water solution whereas the other 5 mice formed a control (Ctl) group injected with an equal volume of plain water. The injections commenced at day 7 from infection and were given through day 35. At day 49, the researchers euthanized the mice and measured both numbers of eggs and numbers of worms in the mice livers, both expected to be lower if the drug was effective.

Table 5.1 gives the worm count data of each mouse. The data shows the treatment reduced the number of worms in both females and males. The mean number of worms changed from 12.00 to 4.40 for the females and from 29.00 to 6.60 for the males.

	Female Trt	Female Ctl	Male Trt	Male Ctl
	1	16	3	31
	2	10	5	26
	2	10	9	28
	10	7	10	13
	7	17	6	47
MEAN	4.40	12.00	6.60	29.00

**Table 5.1:** Worm count data for the schistosomiasis study. Treatment is a “long-course” regimen of K11777 injected with water solution from days 7-35 from infection. Data from [1].

**[On Your Own:]**

1. Calculate appropriate summary statistics for each group of data and produce an individual value plot to compare number of worms for both the male and female mice.

The descriptive analysis above clearly points to a positive effect: K11777 appears to reduce the viability of the parasitic worms in the mice in this sample. But we know that descriptive analysis is usually only a first step in ascertaining whether an effect is real; we often want to know if chance alone could explain the effect.

### **5.1 Nonparametric and Randomization Tests**

In previous work, you have encountered two facets of statistical inference: significance tests and confidence intervals. In the context of significance tests, you have learned about null and alternative hypotheses, p-values, and one- and two-side alternatives. It is likely that every hypothesis test you have conducted in the past involved a normal, t or F distribution. Hypothesis tests that involve theoretical distributions are called parametric tests, meaning that we are testing a parameter (such as  $H_0: \mu = 0$ ) from a specific distribution. These parametric tests require that several distributional assumptions are met, such as independent and identically distributed samples, equal variances for each group, and the sample mean is normally distributed (remember that the Central Limit Theorem states that with large enough sample sizes, sample means always tend to follow the normal distribution).

In this study of schistosomiasis, the variances of each group do not look equivalent for each of the four worm counts. In addition, there are only 5 observations in each group. This is much less than the Central Limit Theorem suggests and we can not be confident that the sample averages are normally distributed. **Nonparametric tests** are hypothesis tests that do not require distributional assumptions. These tests are often useful in studies with very small sample sizes. In this lab we introduce one form of nonparametric statistical inference known as randomization methods. We will describe two kinds of inference using randomization methods, namely randomization hypothesis tests and randomization confidence intervals.

**[Key Concept]** Nonparametric (and randomization) tests do not require the same complex theoretical assumptions as the t-tests or F-tests and can be very useful when there is a small sample size.

The basic concepts of randomization tests will be introduced in a setting where subjects (experimental units) are allocated completely at random to a treatment or control group. Using a significance test, we will decide if an observed treatment effect (i.e. observed mean difference between Trt and Ctl) is “real” or if “random chance alone” could plausibly explain the observed effect. The null hypothesis is stating that “random chance alone” is the reason for the observed effect. The natural alternative hypothesis will be one-sided in this initial discussion; we want to show that the true Trt mean is less than the

true Ctl mean. Next, we will enlarge the discussion to consider modifications needed to deal with two-sided alternatives.

Finally, we will consider confidence intervals, where a confidence interval for a parameter is defined as a set of “plausible values” for the parameter in a sense to be made clear, but related to the previous discussion of significance tests. The relationship between this definition and previous definitions of confidence intervals will also be discussed.

## 5.2 Statistical Inference through a Randomization Test

Inferential procedures take the form of significance tests or confidence intervals, but any such procedure rests upon the *Fundamental Question for Inference*: “What would happen if we did this many times?” Let’s unpack this question in the context of the female mouse study before us. We have observed a mean difference of  $7.6 = 12.00 - 4.40$  between control and treatment groups. While we expect this large a difference reflects the drug’s efficacy, let’s consider another explanation, one which the statistically literate will recognize as a so-called “null hypothesis,”

**[Key Concept]** Every statistical inference procedure is based upon asking the question “How does what we observed in our sample compare to what would happen if we repeated the process many times?”

The null hypothesis acts as something of a “devil’s advocate” position. The devil’s reasoning goes this way: Each female mouse was destined to have a certain number of worms in the liver, this number varies from mouse to mouse, and the only explanation for the mean difference we observed was that our random allocation randomly placed larger numbers in the control group and smaller numbers in the treatment group in such a way that the observed mean difference was 7.6 worms. Whether or not the drug is effective, there clearly is variability in the responses of mice to the infestation of schistosomes. Each group exhibits this variability and even if the drug is ineffectual, some mice do better than others.

The devil’s reasoning thus leads us to our randomization test and the interpretation of the fundamental question for inference. Under the devil’s reasoning whether a mouse is labeled treatment or control has no relevancy to predicting his or her outcome, and we ask the fundamental question this way for this context: If we randomly allocated our 10 mice to treatment and control groups many times, what proportion of the time would we have observed a mean difference as big as or bigger than 7.6? This long-run proportion is a probability that statisticians call the **p-value** of the randomization test.

### **[On Your Own:]**

2. To get a feel for the concept of p-value, put the 10 female worm counts onto 10 index cards. Shuffle the cards, and then draw out 5 cards at random (without replacement). Call these 5 cards the treatment group and the 5 un-picked cards the control group. Under the devil’s advocate position (i.e. the null hypothesis),

this allocation mimics precisely what actually happened in our experiment, since the only cause for group differences under the devil's reasoning was random allocation.

Calculate the mean of the 5 cards representing the treatment group and the mean of the 5 cards representing the control group. What is the mean difference between control and treatment groups that you obtained in your allocation?

3. If you were to do another random allocation, would you get the same difference in means? Explain.
4. Now, perform 9 more random allocations, each time computing and writing down the difference in mean worm count between control group and treatment group. Make a dotplot of the 10 differences. What proportion of these differences is 7.6 or larger?
5. What if you performed the simulation many times? Do you expect the proportion of times the mean differences exceeds 7.6 to be large or small? Explain.

As mentioned above, this long-range proportion is a probability that we call the p-value of the randomization test, and we here formally give the definition:

**[Key Concept]** Assuming that there is nothing creating group differences except the random allocation process, the **p-value** of a randomization test is the probability one would obtain a group difference as big or bigger than the group difference actually observed in the experiment.

**[Key Concept]** Since the p-value is a probability, the computing of a p-value requires conceptually these steps:

- Repeat the random allocation process a number of times,  $N$ .
- Record, each time, whether the group difference exceeds or is the same as the difference observed in the actual experiment; let  $X$  be the number of such occurrences.
- Compute the proportion of times the difference exceeds the observed difference, that is compute  $X/N$ .

The p-value is a long-range relative proportion, which we commonly refer to as a probability. There are two conceptual ways to compute this probability:

- Write down the sample space of all possibilities in an equally-likely model and then compute the proportion of the sample space for which the mean difference is 7.6 or greater.
- Perform a simulation study of the random allocation scheme for a large number of iterations and compute the proportion of the iterations where the difference in means is 7.6 or greater. This approach will only give an approximation to the p-value, but is usually the preferred choice because the sample space will be prohibitively large. We usually pick a round number such as 1,000 or 10,000

iterations and approximate the p-value accordingly. The approximate p-value we get is often referred to as the **empirical p-value** to respect the fact that it is actually an approximation.

There are instances where mathematical methods allow us to obtain an exact p-value without the necessity of a simulation, but computing is cheap and easy and empirical p-values are almost always adequate.

### 5.3 Performing a randomization test using a computer simulation

While physical simulations as we did with the index cards help us understand the process of computing an empirical p-value, we will use computer software to give us an efficient way of producing a simulation based upon a large number of iterations. If you only are simulating 10 random allocations, it is just as easy to use index cards as the computer. However, the advantage of a computer simulation is that 1,000 random allocations can be conducted in almost the same amount of time it takes to simulate 10 samples. In the following steps, you will develop a program to simulate the random allocation we conducted with the index cards.

#### [On Your Own:]

6. Insert that data into a statistical software package and randomly allocate each of the 10 female worm counts to either a treatment or control group.
7. Calculate the average difference between the treatment and the control sample.
8. Write a program, function, or macro to run more than one iteration.
9. Run the macro/program 1000 times. Count the number of simulations that resulted in a mean difference greater than or equal to 7.6 and report the empirical p-value. The p-value is the number of iterations where the mean difference was greater than or equal to 7.6 divided by the total number of iterations. Create a histogram of the 1000 mean differences and comment on the shape.
10. Based upon Question 9, assuming that the treatment is not more effective than the placebo about how frequently would one obtain a mean difference as large as or larger than 7.6 by random allocation alone?
11. Does your answer to 10 lead you to believe the devil's advocate position (aka, null hypothesis) or does it lead you to believe that K11777 has a positive, inhibitory effect on the schistosome worm in female mice? Explain!

Based upon a simulation with 1000 reps, the author obtained a p-value of  $30/1000 = 0.03$ ; your results should be close to this value. This says that random allocation alone would produce a mean group difference as large as or larger than 7.6 only about 3% of the time, which suggests that something other than chance is at work to explain the difference in group means. Since the only other distinction between the groups is the presence or absence of treatment, we can conclude that the treatment causes an improvement, i.e., a reduction, in worm counts.

## 5.4 Two-Sided Tests

The direction of the alternative hypothesis is derived from the research hypothesis. In the instance of the K11777 study, we enter the study expecting a reduction in worm count and hoping the data will bear out this expectation. It is our expectation, hopes, or interest that drives the alternative hypothesis and the randomization calculation. Occasionally, we enter a study without a firm direction in mind for the alternative, in which case we entertain a two-sided alternative. The p-value then must take into account extreme values of the test statistics in either direction and so we define the p-value in a more general way than our previous definition, a way that allows for a wide variety of significance testing situations:

**[Key Concept]** The *p-value* of a randomization test is the probability one would obtain a group difference as extreme as or more extreme than the group difference actually observed in the experiment, assuming that there is nothing creating group differences except the random allocation process.

This definition encompasses the earlier definition if we interpret “extreme” to mean either “greater than” or “less than,” depending on direction. But in the two-sided case, extreme encompasses both directions. In the K11777 example, we observed a mean difference of 7.6 between control and treatment groups. Thus the two-sided p-value calculation will count up all instances among the 1000 replications where the randomly allocated mean difference is either  $\geq 7.6$  or  $\leq -7.6$ .

12. Run the simulation study again to find the empirical p-value for testing the 2-sided hypothesis test to determine if there is a difference between the treatment and control for female mice. Was the number of simulations resulting in a difference greater than or equal to 7.6 identical to the number of simulations resulting in a difference less than or equal to -7.6? Explain!

## 5.5 Link to Conventional Terminology and Notation.

The devil’s advocate position is usually called the **null hypothesis** and is denoted  $H_0$ . The research hypothesis (e.g., treatment causes a reduction in worm count) is called the **alternative hypothesis** and is denoted  $H_a$ . Alternative hypotheses can be “one-sided, greater than” (which was tested in the simulation above), “one-sided, less-than” (e.g., treatment causes an increase in worm count), or “two-sided” (e.g., treatment is different, in one direction or the other, from control).

Sometimes we imagine having some threshold p-value at or below which we will agree to reject the null hypothesis and conclude in favor of the alternative. This threshold value is called a **significance level** and is usually denoted by the Greek letter  $\alpha$ . Common values are  $\alpha = .05$  or  $\alpha = .01$ , but the value will depend heavily upon context and the researcher’s assessment of acceptable risks in stating bad conclusions. When the study’s p-value is

less than or equal to this significance level, we will state that the results are **statistically significant at level  $\alpha$** . If you see the phrase “statistically significant” without a specification of  $\alpha$ , the writer is most likely to assume  $\alpha = .05$ , for reasons of history and convention alone.

### [On Your Own:]

13. Run a simulation using the data from the male mice to decide if you feel that K11777 inhibits schistosome viability (i.e. reduces worm count) in male mice. Describe the results by including a histogram of the simulation results, the p-value, and a summary statement that indicates what your conclusion is about the research question of schistosome viability.

For the female simulation study, you should have obtained a 1-sided p-value of about .034; for the male study a p-value of close to 0.010. Given equal sample sizes for each study and given the smaller p-value for the male study, and given the larger absolute difference in mean for the male study than for the female study, we are led to state that the efficacy of K11777 for treating shistosomiasis is stronger in male mice than in female mice. For either sex it is natural to follow up a significance test, which provides the degree of evidence to support an alternative hypothesis, with a confidence interval.

## 5.6 From Significance Tests to Confidence Intervals

A **confidence interval** gives a range of plausible values for some parameter. This range is around an observed estimate of the parameter, an estimate based upon the data. To the range of values we attach our “confidence” that the true parameter lies in the range. Thus, a confidence interval gives an estimate of where we think the parameter is and how precisely we have it pinned down.

Confidence intervals for a single parameter can be one-sided, either “less than” or “greater than”, or they can be two-sided. If  $d$  represents some parameter, these three types of confidence intervals will lead to a confidence statement in one of these forms:

We are  $100(1-\alpha)\%$  confident that  $d \leq U$  (a one-sided, “less than” interval).

We are  $100(1-\alpha)\%$  confident that  $d \geq L$  (a one-sided, “greater than” interval).

We are  $100(1-\alpha)\%$  confident that  $L \leq d \leq U$  (a two-sided interval).

For example, if  $\alpha = .05$  we would be making 95% confidence intervals.

If you have encountered the concept of confidence intervals before, it is most likely in the context of having sampled from some population with a desire to estimate some numerical characteristic, i.e., **parameter**, for the population, such as the population mean or median. But, in the current context, we have not sampled from a population, we have simply performed an experiment on a convenient and available collection of mice.

**[Key Concept]** It is the experimental set up that allows us to both define a parameter and gives us a basis for constructing a confidence interval.

The assumption we make is that the treatment effect is *additive*, meaning that there is a parameter, call it  $d$ , such that the response for an experimental unit in the presence of the control is  $d$  greater (less if  $d$  were negative) than what it would be in the presence of treatment. In our example,  $d$  is the expected average reduction in worm counts when K11777 is applied to a group of mice. That is, if  $X_i$  represents the number of worms for the  $i$ -th mouse in the absence of K11777 and  $Y_i$  represents the number of worms for the  $i$ -th mouse in the presence of K11777, then our (additivity) assumption is  $X_i = Y_i + d$  for all  $i$ . In the original experiment, a natural estimate for  $d$  would be the difference in the sample means, namely 7.6.

Under this additivity assumption and notation the randomization test above can also be written as:

$$H_0: d = 0 \text{ versus } H_a: d > 0 \tag{5.1}$$

The randomization tests conducted earlier, for either sex of mouse, concluded that the alternative hypothesis was more consistent with the data. To better understand the concept of a confidence interval, we focus on the concept for *plausible (and implausible)* values of the parameter  $d$ . We will find the set of plausible values through the use of a set of two-sided alternative hypotheses. Suppose we consider a significance test for the situation:

$$H_0: d = d_0 \text{ versus } H_a: d \neq d_0 \tag{5.2}$$

for some specific value of the parameter  $d_0$ .

**[Key Concept]** If we reject  $H_0$  in favor of  $H_a$  then we will call the value  $d_0$  *implausible*; if we fail to reject  $H_0$  we call  $d_0$  *plausible*.

In Question 12 above, Table 5.2 was used to observe a p-value of .056 for the two-sided hypothesis test for female mice. Since the p-value is greater than  $\alpha = 0.05$ , the author failed to reject  $H_0: d = 0$ . Thus 0 is a plausible value for a two-sided confidence interval. 0 will be in the two-sided confidence interval.

Difference	-8.8	-7.6	-6.4	-5.6	-5.2	-4.8	-4.4	-4	-3.6	-3.2	-2.8	-2.4
Count	6	25	12	24	26	11	50	38	8	16	19	20

Difference	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	2	2.4
Count	71	73	32	44	39	27	39	44	27	64	60	9

Difference	2.8	3.2	3.6	4	4.4	4.8	5.2	5.6	6.4	7.6	8.8
Count	12	19	13	28	49	13	22	20	15	19	6

**Table 5.2:** Results of 1000 simulations of using a Minitab macro. Difference is the list of observed differences of female ctl – female trt. Count is the frequency of each observed difference. The highlighted cells show that 56 out of 1000 simulations were at least as extreme as what was observed in our sample of 10 female mice.

*[Note:]* By failing to reject  $H_0$ , we conclude based on the sample of 10 female mice and the author’s 1000 simulations, the mean number of worms in the female control group is equal to the mean number of worms in the female treatment group. Thus for the equation,  $X_i = Y_i + d_0$ ,  $d_0 = 0$  is plausible. However, since .056 is an empirical p-value, others may observe a slightly different value. If by chance a researcher found a p-value less than or equal to 0.05, 0 would be an implausible value. This shows that 1000 simulations is often not enough. If researchers run 10,000 simulations, the p-value will tend to be more accurate.

This simulation only works for testing  $H_0: d = 0$  versus  $H_a: d \neq 0$ . The rest of this section will discuss how to generalize the hypothesis test in (5.2) to other values of  $d_0$ . Intuitively, values of  $d_0$  close to the observed treatment effect of 7.6 will be plausible and those far to the right or left of the observed treatment effect will be implausible. Finding the precise transition points will lead to numbers L and U for which the set of values  $L \leq d \leq U$  is our set of plausible values, and if we have chosen  $\alpha$  as the significance level for our randomization tests this set of values forms a  $100(1 - \alpha)\%$  confidence interval for  $d$ . We interpret this interval by saying we are  $100(1 - \alpha)\%$  confident that  $L \leq d \leq U$ .

Suppose we let  $x_1, x_2, \dots, x_n$  represent the control sample and let  $y_1, y_2, \dots, y_m$  represent treatment sample for female mice. If, under the null hypothesis, the  $i^{th}$  mouse accumulates  $d_0$  fewer worms using the treatment, then if the control mouse represented by datum  $x_1$  would have randomly been assigned to the treatment group instead it would have produced a reading of  $x_1 - d_0$  worms instead of  $x_1$  worms. Thus, if we subtract  $d_0$  from each x-value we would have adjusted the control sample to a set of values that would be the control mice’s readings in the presence of the treatment. Thus, we can perform the randomization test for the situation in line (5.2) by applying the `schistosome.mtb` simulation to the two columns “female ctl -  $d_0$ ” and “female trt”. If the null is rejected by the test, we conclude that the treatment effect is not equal to  $d_0$ , which would make  $d_0$  an implausible value and  $d_0$  would not be in the confidence interval.

For example, to test  $H_0: d = 1$  replace the five data values `female ctl` = (16, 10, 10, 7, 17) with `female ctl - 1` = (15, 9, 9, 6, 16) and use your simulation program to conduct 1000 simulations. Note that the cutoff values will now be  $|11 - 4.4| = 6.6$  and  $-6.6$  instead of 7.6 and -7.6. The author observed a p-value of .059, thus we fail to reject  $H_0: d = 1$  and 1 is in the range of plausible values.

**[On Your Own:]**

14. Modify the `female ctl` data and run a simulation to test the two-sided hypothesis  $H_0: d = -1$  vs.  $H_0: d \neq -1$ . What p-value did you observe?

Output from running this simulation for multiple values of  $d_0$  are shown in Table 5.3 below. The p-value transitions from .020 to .052 between  $d_0=-1$  and  $d_0=0$ . Thus, L, the lower value of my 95% confidence interval is somewhere between -1 and 0. The p-value also transitions between  $d_0=15$  and  $d_0=16$ , which suggests U is between 15 and 16.

$d_0$	-1	0	1	2	...	14	15	16	17	18
p-value	.020	.056	.059	.080	....	.081	.060	.015	.010	.040

**Table 5.3:** Results of 1000 simulations for various values of  $d_0$ . 1000 simulations is often not reliable enough to show accurate estimates of the true p-values. Answers will vary. In addition, L somewhere between -1 and 0 is not very accurate. Instead of testing only integer values of  $d_0$ , values of  $d_0$  are often tested in steps of .1 or .01.

**[On Your Own:]**

15. Instead of just integers, find more specific estimates of L and U for a 95% two-sided confidence interval for female mice to the nearest 10<sup>th</sup> of a worm. You may need to increase the number of simulations to 10,000 to get more specific estimates.
16. Estimate L for a 99% one-sided, “greater than” confidence interval for male mice (i.e. find all plausible integer values of  $d_0$  for the 1-sided hypothesis test  $H_0: d = d_0$  versus  $H_a: d > d_0$ .) Make an interpretive statement about the interval.

This method lacks some precision; pinning down L and U can be rather tedious. But we hope the logic of finding a confidence interval is clear and with persistence-and a good simulation program, one can find L and U to a good level of precision. Garthwaite [3] provides an algorithm that is more efficient (but less transparent) that obtains L and U quite precisely, but we introduced the cruder methods here to highlight the logic of confidence intervals for the randomization model.

**5.7 Lab Summary**

In this study, the units (mice) have probably not been selected at random from some larger population, but most assuredly the mice have been randomly allocated to treatment or control groups. The randomization hypothesis tests concluded that for both male and female mice, we could plausibly rule out the explanation of chance alone for the group differences. Thus, one-sided tests in this study design allows us to establish a cause-effect relationship between treatment and response (e.g. we can conclude that K11777 did reduce the average number of worms infected with the schistosome parasite). However, since our sample was not randomly selected, we should be hesitant to conclude that this result holds for a large population of mice.

Our sample showed an average reduction of 7.6 worms in female mice and an average reduction of 12.4 worms in male mice. Confidence intervals are often created in addition

to hypothesis tests to estimate the parameter (e.g. the mean reduction in worm count) and determine how precisely we have it pinned down.

Accurate randomization tests typically require a large number of simulations. Statistical software programs such as R and SAS typically can run simulations more quickly and efficiently than Minitab. Confidence intervals in particular provide only crude interval estimates unless simulation studies include at least 10,000 runs.

## Chapter 5B: A Closer Look at Randomization Tests

### 5B.1 Permutation Tests and Randomization Tests

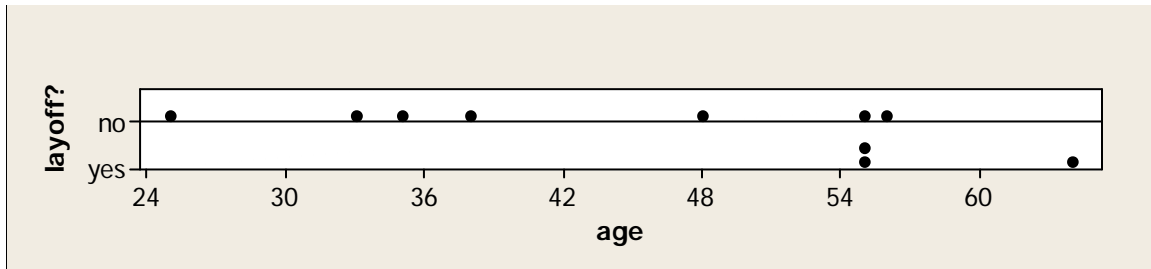
The random allocation of experimental units (e.g., mice) to groups provides the basis for statistical inference in a randomized, comparative experiment. In the schistosomiasis K11777 treatment study we used a significance test to ascertain whether cause and effect is at work, and used a confidence interval to estimate the size of the asserted treatment effect. In the context of the random allocation study design, we called our significance test a randomization test.

In the age discrimination study below, we apply the same inferential procedures in a context devoid of random allocation. In this context, it is common to call the significance test a **permutation test** rather than a randomization test. More importantly, in this context, the results of the test cannot be used to claim cause and effect. A researcher should exhibit more caution in the interpretation of results.

#### Age discrimination study

In this example, we use a data set from *Statistics in Action* [4]. Westvaco is a company that produces paper products. In 1991, Robert Martin was working in the engineering department of the company's envelope division when he was laid off in Round 2 of several rounds of layoffs by the company. He sued the company, claiming to be the victim of age discrimination. The ages of the 10 workers involved in Round 2 were: 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64. The ages of the three people laid off were 55, 55, and 64.

Figure 5B.1 shows comparative dotplots for age by layoff category, which gives an impression that Robert Martin may have a case: it certainly appears as if older workers were more likely to be laid off. But we know enough about variability to be cautious.



**Figure 5B.1: Dotplot of age vs layoff.** Age of worker by whether he or she was laid off. Data come from *Statistics in Action* [4] and refer to a reduction in force program by the Westvaco Corporation.

**[On Your Own:]**

1. Run a permutation test using ‘age’ as the response variable and ‘layoff?’ as the group identifier to decide if a mean difference as large as represented in the data gives statistically significant data (use  $\alpha = .05$ ) that the mean age for those laid off exceeds the mean age for those not laid off.
2. One major advantage of randomization/permutation tests, over classical methods, is that they easily allow the use of test statistics other than the mean.
  - a. Modify the program/macro you created in question 1 to measure a difference in group medians instead of a difference in means for the Westvaco age data. Report the p-value and compare your results to question 1.
  - b. One might also wonder if the treatment might change the variability in layoff groups. Modify the macro you created in question 1 to test if the variances of each layoff group are equal. Report the p-value and state your conclusion in both cases.

Using 100,000 replications in exercise 1, the author obtained an empirical p-value of 0.04951, barely reaching the magical status of “statistically significant at the .05 level” and giving some evidence to the effect that older workers were laid off at a higher rate. Since there was no random allocation, statistical significance does not give us the right to assert that old age is *causing* a difference in response to being laid off.

Think about the reasoning. In the random-allocation setting of the schistosomiasis K11777 treatment study, only two things distinguish the two groups: (1) random allocation and (2) one group gets treatment, the other not. If we can rule out (1) random allocation as a plausible reason for difference in response (which a small p-value does) then only reason (2) remains to explain the response difference.

In randomization tests, the null hypothesis  $H_0: d = 0$  asserts that random allocation alone accounts for the observed difference. In the age discrimination case, group assignment is not at random (i.e., layoffs are *not* determined by a randomization scheme). The null hypothesis in this context becomes: The observed difference could be explained *as if* by random allocation alone. That is, we proceed as any practicing social scientist must when working with observational data. We “imagine” an experiment in which workers are randomly allocated to the layoff group and then determine if the observed average

difference in the ages of laid off workers and those not laid off is “significantly” larger than would be expected to occur by chance from a randomized, comparative experiment. When one is able to reject this null hypothesis, as we did in problem 1 at the .05 level, one is left with competing alternatives. While age could be the cause for the difference—hence proving an allegation of age discrimination—there are a myriad other possibilities (i.e, confounding or lurking variables), such as the educational levels of the workers, their competence to do the job, ratings on past performance evaluations, etc.

Rejecting this “as if by random allocation” hypothesis in the non-randomized context can be a *useful step* toward establishing causality, but it can never establish causality.

### **Bird nest study**

The data set labeled *birdnest* represents data collected by Grinnell College student, Amy Moore, in the spring of 1999 for a class project. Each record in the data set represents data for a species of North American passerine birds. Passerines are “perching birds,” and include many families of familiar small birds (e.g., sparrows, warblers, etc), and some larger species like crows and ravens, but does not include hawks, owls, water fowl, wading birds, woodpeckers, etc. Moore took all North American passerines (“perching birds”) for which complete evolutionary data were available, which comprised 99 species of the 470 total passerines in North America. (Part of her study used this evolutionary information.) One hypothesis of interest was about the relationship of body size to type of nest. Body size was measured as average length of the species. Nest type was categorized into either closed or open. Nests come in a variety of types (see the *Nesttype* variable). In this data set the variable “closed” refers to nests with only a small opening to the outside, such as the tree cavity of many woodpeckers or the pendant style nest of an oriole. “Open” nests include the cup shape of the American robin.

### **[On Your Own:]**

3. Moore suspected that closed nests tend to be built by larger birds, but we will treat the alternative here as two-sided, since the evidence for her suspicion was based upon scanty evidence.
  - a. Use comparative dotplots or boxplots and summary statistics to describe the relationship between average body length and the nest type (the *closed?* variable). (Note: *closed?* = 1 for closed nests; *closed?* = 0 for open nests.) Does it appear that Moore’s initial suspicion is born out by the data?
  - b. Now, run a permutation test using a two-sided alternative to determine if type of nest varies by body length. and interpret your results. Be sure to state your conclusions in the context of the problem and address how random allocation and random sampling (or lack of either) impacts your conclusions.

## **5B.2 Permutation and Randomization Tests for Different Study Designs**

The ideas we have developed in this unit can be extended to other study designs. We illustrate that first with a very basic, two-variable design called a **matched pairs design**.

In a matched pairs design, each experimental unit provides both measurements in a study with two treatments (one of which could be a control). Conversely, in the completely randomized situation of the Schistosomiasis K11777 treatment study, half the units were assigned to control and half to treatment; no mouse received both treatments. Let us begin with a concrete example.

Grinnell College students Anne Tillema and Anna Tekippe wanted to do an experiment to find out the effect that music may have on a person's level of relaxation. They hypothesized that fast songs would increase pulse rate more than slow songs. The file called *music* contains the data from their experiment. They decided to use a person's pulse rate as an operational definition of the person's level of relaxation. They decided to compare pulse rates for two selections of music—one representing a fast selection and one a slow selection. For the fast selection they chose “Beyond” by Nine-inch Nails and for the slow selection they chose Rachmaninoff's “Vocalise”. They recruited 28 student subjects for the experiment.

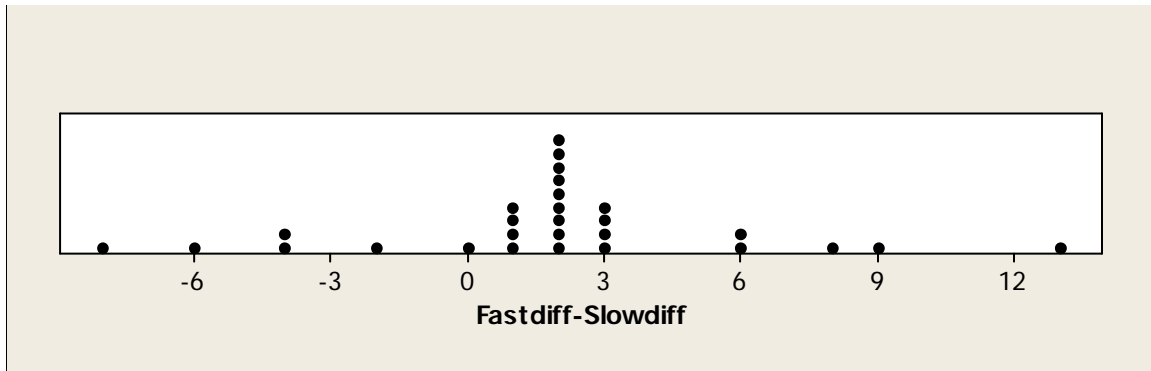
Anne and Anna came up with the following design. For their fundamental question there are two treatments involved: (1) listening to the fast selection and (2) listening to the slow. They could have randomly allocated 14 to hear the fast selection and 14 to hear the slow selection, but a more efficient approach turned out to have each subject provide both measurements. That is, each subject listened to both selections, giving rise to two, so-called “matched pairs” of data values for each subject. Randomization came into play in that it was decided by a “coin flip” whether each given subject would listen to the fast selection first and the slow second or in the opposite order. Specifically, by coin flips, about half experienced the following procedure:

[one minute of rest] > [measure pulse (“prepulse”)] > [listen to fast selection for 2 minutes; measure pulse for second minute (“fastsong”)] > [rest for one minute] > [listen to slow selection for 2 minutes; measure pulse for second minute (“slowsong”)].

The other half experienced the procedure the same way except that they heard the slow selection first and the fast selection second.

For analysis, each subject gives us two measurements: (1) their fast song pulse minus their pre-pulse, and (2) their slow song pulse minus their pre-pulse. In the data file, these two measurements are called *Fastdiff* and *Slowdiff*, respectively.

Figure 5B.2 shows a dotplot of the 28 *Fastdiff*-minus-*Slowdiff* values. Notice that positive numbers predominate and the mean difference is 1.857, all suggesting that the fast song does indeed effect a heightened response (pulse rate) compared to the slow song. We need to confirm this suspicion with a randomization test.



**Figure 5B.2: Dotplot of Fastdiff-Slowdiff.** Dotplot of difference by subject for their fast song pulse increase minus their slow song pulse increase. The mean difference is 1.857 beats per minute and most values are positive, all of which suggest that the fast song increases pulse more.

**[On Your Own:]**

- Anne and Anna decided to use a 1-sided test to show that fast music increased pulse rate more than slow music before they looked at the data. Why is it important to determine the direction of the test before looking at the data?

To perform a randomization test, we mimic the randomization procedure of the study design. Here, the randomization gave an order to which the subject heard the song. So randomization is within pairs of measurements on each subject. That is, accepting the null hypothesis that the order makes no difference, then the only factor that affects a difference is this random allocation process. So to compute a p-value we ask ourselves how frequently would one obtain an observed difference as big as 1.857 or bigger.

**[On Your Own:]**

- Create a simulation to test the Music data. Write a program/macro that will randomly multiply a 1 or a -1 to each observed difference. Thus randomly assigning an order (Fastdiff-Slowdiff or Slowdiff-Fastdiff). Then, for each iteration, calculate the mean difference for each group. Be sure to state your conclusions in the context of the problem and address how random allocation and random sampling (or lack of either) impacts your conclusions.

The p-value is the number of times your simulation found a difference in means greater than or equal to 1.857143.

Simulations will show a one-sided p-value of about 0.01. The author obtained an empirical p-value of 0.011 with 1,000 replications. Since this p-value represents the likelihood of a difference of this magnitude being attributable to random allocation, we conclude that the fast song does cause an elevated increase in pulse rate.

One note of caution: this type of randomization DOES NOT average out confounding variables such as a great love and excitement for “Nine Inch Nails” by some students or a complete detachment and boredom with this band by others (i.e., “musical taste” is still a possible confounder that randomizing the order of listening cannot randomize away).

This will always be a caveat in this type of study, since we are rather crudely letting the one “Nine Inch Nails” song “represent” fast songs.

### Twins study

6. In a 1990 study by Suddath et al, reported in Ramsey and Schafer [2], researchers used magnetic resonance imaging to measure the volume of various regions of the brain for a sample of 15 monozygotic twins, where one twin was affected with schizophrenia and the other was unaffected. The twins were from North America and comprise 8 male pairs, 7 female, and ranged in age from 25 to 44 at the time of the study. The sizes in volume  $\text{cm}^3$  of the hippocampus are in the file called *twins*.
  - a. The explanatory variable here is whether or not a twin is affected by schizophrenia and clearly random allocation is not involved at all. This is an observational study, not an experiment. Nevertheless, would the conceptual design principle of an imaginary “random allocation” or would “matched pairs” seem more appropriate in considering the analysis of these data?
  - b. Use appropriate graphics and summary statistics to describe the difference in brain volume for affected and unaffected twins.
  - c. Use the appropriate permutation test to ascertain if the conclusion about differences in brain volume drawn in (b) is the result of schizophrenia or if it could be explained as a chance difference. Report your p-value and summarize your conclusion.

### 5B.3 Relationship Between the Randomization Test and t-test

If you have seen two-sample inference in previous settings it is likely to have been in what Ernst [5] calls the **population model**, which he distinguishes from the **randomization model**. In a randomization model experimental units are subjected to treatments with some kind of random allocation scheme. In a population model, experimental units are selected at random from population(s). In one simple case, one may be interested in comparing two, separate populations—say, by comparing two population means. One can take two, independent, simple random samples and use the classical two-sample t-test to make the comparison. This process clearly differs from examples such as the schistosomiasis example, where the subjects are not selected at random, but two (dependent) samples are formed from a collection of available experimental units that are divided into two groups completely at random.

R.A. Fisher, perhaps the pre-eminent statistician of the twentieth century, introduced the randomization test in the context of a two-group, randomly allocated experiment in his famous 1935 book, *Design of Experiments* [6]. He also acknowledged there the impracticality of carrying out a randomization test with any but miniscule data sets because of the computational intensity of the computation because, obviously, 1935 predates modern computing. Indeed, Efron and Tibshirani [8] describe permutation tests as “... a computer-intensive statistical technique that predates computers.” Fisher went

on to illustrate and assert that the classical two-sample t-test (for independent samples) approximates the randomization test very well. In 1937, E.J.G. Pitman gave a mathematical derivation of Fisher's assertion [7]. Pitman's result thus provided a computationally simple way to approximate the randomization test—the conceptually appropriate test for a randomization model—which was the t-test's only legitimacy in such situations.

Over time, approximations to the randomization tests using classical and computationally tractable methods have been published. Ernst [5] cites references to this literature. As one example, which Ernst attributes to Fisher, we can use the paired t-test to approximate the randomization test in the matched pairs setting.

7. Using the t-test, compute the two-sided p-value for the bird nest study above and compare the results to what you found with the randomization test.
8. Using the t-test, compute the one-sided p-value for the music study above and compare the results to what you found with the randomization test.

As a simple example, using 100,000 replications of the randomization process in the schistosomiasis example, the author obtained an empirical p-value (one-sided) of 0.0289. The one-sided p-value from the two-sample t-test is 0.0096. The results are relatively close and would most likely lead to similar conclusions about the research hypothesis, although the level of evidence would be a bit stronger if one just used the approximate t-test. We are fortunate that in this age of modern computing, we no longer have to routinely make the choice to compromise and use the t-test to approximate the randomization test.

### References cited

1. Maha-Hamadien Abdulla, Kee-Chong Lim, Mohammed Sajid, James H McKerrow, and Conor R Caffrey, "Schistosomiasis Mansonii: Novel Chemotherapy Using a Cysteine Protease Inhibitor," *PLoS Medicine*, v4,n1, Jan. 2007: On line: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=17214506> NOTE: Professor Conor R Caffrey provided the raw data upon request.
2. Fred L Ramsey and Daniel W Schafer, *The Statistical Sleuth*, Duxbury, 2002, Pacific Grove, CA
3. Paul H Garthwaite, "Confidence Intervals from Randomization Tests," *Biometrics*, v52, Dec 1996, 1387-1393.
4. Ann E Watkins, Richard L Scheaffer, and George W Cobb, *Statistics in Action*, Key Curriculum Press, 2004, Emeryville, CA.
5. Michael D Ernst, "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, v19, n4, 2004, 676-685.

6. R.A. Fisher, *Design of Experiments*, Oliver and Boyd, 1935, Edinburgh.
7. E.J.G. Pitman, "Significance Tests Which May be Applied to Samples from any Population," *J. Royal Statistical Society, Supp. 4*, 1937, 119-130.
8. Bradley Efron and Robert Tibshirani, *An Introduction to the Bootstrap*, CRC Press, 1993, p. 202