5 HYPOTHESIS TESTING: ALPHA, BETA, AND POWER

Objectives

- Understand the concepts of statistical errors, sample variability, and effect size.
- Explore the interplay among alpha, beta, effect size, sample variability, and the power of test.

Suggested Preliminary Exercise: Statistical Distributions

INTRODUCTION

Much research in ecology involves making statistical tests of one kind or another. We frequently want to know if two or more populations differ from one another with respect to some parameter that they share. For example, are trees from one forest "significantly" larger than those from another? Do older rabbits have thicker coats than younger rabbits? Is the species diversity of the restored prairie different from that of the degraded one? These comparisons generally involve estimating the value of a parameter in each population using data obtained through sampling. Typically, these estimates are compared using a statistical test to identify a difference or lack thereof.

Sampling and Uncertainty

Because sampling always involves some uncertainty (with sampling we are never entirely sure that we have properly estimated the true value of the parameter for the population), we have to consider the possibility that any difference that we see between two estimated parameters could result from sampling flukes. That is, the populations we sampled don't actually differ, but we drew unrepresentative samples by chance that give the incorrect appearance of a difference. This is a **Type I** statistical error. The probability of committing a Type I error is called **alpha (α)**.

Alternatively, the populations we are interested in may actually be different, but from some fluke in sampling we drew two samples that showed no differences. This is a **Type II** error. The probability of committing a Type II error is known as **beta (β)**. Type II errors may occur because the actual difference between the populations (the "effect size") is small and the variability in our samples obscures the difference and prevents us from detecting it. We are obviously more likely to detect a difference between populations the more precisely we have estimated the parameters in each of them (perhaps because we sampled each population well) or where the difference is substantial enough to detect despite the variability in samples.

Thus, the major challenge in performing a statistical test is simple: Ensure that you don't commit a type I or II error and thereby confidently detect any differences that might exist (Sokal and Rohlf, 1981). You can guard against committing a Type I error by using an appropriately stringent α level, say, 0.05 or lower. Guarding against a Type II error can be more problematic. A test that will detect differences if they exist, regardless of the sample variability and the effect size, is said to be have high statistical **power**. Power = $1 - \beta$, so a low β (probability of missing an important difference) equates to a high power of the test. Statistical power of the test is an important concept because ensuring that a given test has high power means that it will accomplish what you hope it will: that is, it will detect differences should they exist. All too often we regard a lack of difference, as indicated by a nonsignificant result on a test, to reflect no real difference between populations, when it may actually be the result of a poorly designed study (too variable or too small a sample to detect a subtle difference that nonetheless exists).

Choosing acceptable α and β values is worth additional consideration. Standard biological literature generally sets α to 0.05 and β at 0.2. In many cases, it may make sense to use other values. If the goal is to detect important differences, perhaps doing so at the risk of an increased level of false detections, then designing a test using high α and a low β (high power) would be advisable. This might be the case, for example, in looking for trends in a population of an endangered species. You want to quickly detect any declines in the species so you can step in and do something about them, but you are comfortable exploring some false reports of declines should they occur. On the other hand, if wrongly detecting a difference is very costly, then you might want to use a low α to guard against committing a Type I error. The important message is that "statistical significance" is only relative to the levels of α and β that you consider to be reasonable and that you set in advance.

The purpose of this exercise is to enable you to explore the interplay among α, β, effect size, sample variability, and the power of test. If you clearly understand the trade-offs among these parameters, you will greatly enhance your ability to design appropriate sampling schemes for detecting differences, should they exist, among populations. As always, save your work frequently to disk.

Summary of Type I and Type II Errors, and Power

Suppose we sample coat thickness of two populations of rabbits. The **null hypothesis (H_0)** is that the groups do not differ in coat thickness. We hope to gather evidence to reject the null hypothesis at a given probability level (α). If H_0 is in fact true and the populations do not actually differ in coat thickness, but you reject H_0 and conclude that the populations are different, you have committed a Type I error. If H_0 is false and the populations have different coat thicknesses but you fail to reject the H_0, you have committed a Type II error; your sampling lacked power to detect actual differences. **Power** is the probability of rejecting H_0 when it is in fact false.

	Reject H_0	**Fail to reject H_0**
H_0 is true:	Type I error (α)	Correct decision. Other ideas?
H_0 is false:	Correct decision. Nobel Prize!	Type II error (β)

INSTRUCTIONS

ANNOTATION

A. Set up and sample two model populations.

1. Open a new spreadsheet and set up column headings as shown in Figure 1.

	A	B	C	D	E	F	G
1							
2		TYPE I ERRORS			TYPE II ERRORS and POWER		
3							
4		mean	sd			mean	sd
5	Pop 1 =>	50	5		Pop 1 =>	45	5
6	Pop 2 =>	50	5		Pop 2 =>	50	5
7	Effect size =	0			Effect size =	5	
8							
9	Individual	Pop 1	Pop 2		Individual	Pop 1	Pop 2
10	1				1		
11	2				2		
12	3				3		
13	4				4		
14	5				5		
15	6				6		
16	7				7		
17	8				8		
18	9				9		
19	10				10		
20							
21	mean				mean		
22	std				std		
23							
24	α=	t-test:	significant?		α=	t-test:	significant?
25	0.05				0.05		
26							
27							
28	trial	t-test	significant?		trial	t-test	significant?

Figure 1

Generate the α symbol by typing an "a." Select the letter in the formula bar and change the font to symbol font.

2. Enter the values shown in cells B5–C6.

We'll start by exploring Type I errors in columns A, B, and C. We'll make a statistical comparison of two populations (columns B and C) that have identical means and variances. Enter 50 in cells B5 and B6 to indicate a mean value of the population, say, height. Enter a standard deviation of 5 in cells C5 and C6. Thus, both populations have the same mean (μ) and standard deviation (σ^2) in height (of course, you don't really know these are the true means and variances of the populations; you will sample individuals to estimate these parameters).

3. In cell B7, calculate the **effect size** as the difference between the means of the two populations. Save your work.

Enter the formula **=ABS(B5–B6)** in cell B7. In this case, the effect size is 0.

4. In cell B10, enter the formula **=NORMINV (RAND(),B5, C5)**. Copy the formula down to cell B19.

Now we will "sample" 10 individuals from population 1 by generating random measurements as if they came from a population with a normal height distribution. We can use the **NORMINV** function and **RAND** function to do this. The **NORMINV** function returns the inverse of the normal cumulative distribution for the specified mean and standard deviation, and has the form **NORMINV(probability,mean,standard_dev)**. The B10 formula tells Excel to draw a random probability (the **RAND()** portion of the formula) from a normal distribution with a mean height given in cell B5 and a standard deviation given in cell C5; Excel will convert that random probability into a value (height) from that distribution.

5. In cells C10–C19, obtain 10 samples from population 2.

Obtain heights of individuals for 10 individuals drawn at random from Population 2. We used the formula **=NORMINV(RAND(),B6,C6)** in cells C10–C19 (following the procedure in Step 4).

6. In cells B21 and C21, enter a formula to calculate the mean of your sample for populations 1 and 2, respectively.

In cell B21 we used the formula **=AVERAGE(B10:B19)**.
In cell C21 we used the formula **=AVERAGE(C10:C19)**.

7. In cells B22 and C22, enter a formula to calculate the standard deviation of your sample for population 1 and 2, respectively. Save your work.

Enter the formula **=STDEV(B10:B19)** in cell B22.
Enter the formula **=STDEV(C10:C19**) in cell C22.

B. Conduct a* t*-test to determine if samples from populations 1 and 2 differ in height.

1. Enter 0.05 in cell A25.

In cell A25, you need to specify what α will be. By convention, α = 0.05 is used. Remember that α is the probability of committing a Type I error—rejecting the null hypothesis when the null hypothesis is in fact true. In the next step you will generate a *t*-test statistic and a probability associated with that test statistic. If the test statistic has a probability that is less than or equal to the α level you have selected, you would conclude that the two populations are different. If the test statistic has a probability that is greater than the α level you have selected, you would conclude that the populations are not statistically different. You can set α to any level you like (although α > 0.15 will raise eyebrows). For now, we will use the conventional α = 0.05, and will change α levels later in the exercise.

2. In cell B25, use the **TTEST** function to conduct a *t*-test on the two population sample means.

Enter the formula **=TTEST(B10:B19,C10:C19,2,2) in cell B25.**
Now that you have determined what kind of Type I error rate you can live with, you're ready to perform a *t*-test to compare the sample means of the two populations. The **TTEST** formula returns the *probability* associated with a Student's *t*-Test (it does not return the value of the test statistic itself). You will use **TTEST** to determine whether the two samples are likely to have come from two underlying populations that have the same mean. The **TTEST** formula has the form **TTEST(array1,array2,tails,type).** Array1 is the first data set (or the 10 individuals sampled from population 1), Array 2 is the second data set (or the 10 individuals sampled from population 2), tails refers to whether you want to conduct a one- or two-tailed test (choose 2), and type is the kind of *t*-test to perform (for now, choose two-sample equal variance).

3. In cell C25, enter an **IF** formula to return a 0 if your *t*-test statistic is greater than alpha, and a 1 if your *t*-test statistic is less than alpha.

Enter the formula **=IF(B25>A25,0,1)** in cell C25.
Now that you have a test statistic probability, you need to compare it to the α level you've chosen. If the probability of the test statistic is <0.05 (your α level), you would conclude that the two populations are different. If the test statistic probability is >0.05 you would conclude the populations are not different (or, more correctly, the samples failed to show differences). The **IF** formula returns one value if a condition you specify is true, and another value if the condition you specify is false. It has the syntax **IF(logical_test,value_if_true,value_if_false)**. A score of 1 indicates that the two populations are statistically different; a score of 0 indicates they are not statistically different. Based on your test, what conclusions can you make about the two populations?

C. Run 100 sampling trials.

1. Set up a linear series from 1 to 100 in cells A29–A128.

Enter **1** in cell A29.
Enter the formula **=1+A29** in cell A30. Copy this formula down to cell A128.
A value of $\alpha = 0.05$ means that if you ran your *t*-test on samples (new samples) over and over again, about 5 times in 100 you would conclude that the two populations are different when in fact they are identical. We'll prove that to ourselves by running a number of trials in which we randomly draw 10 individuals from each population, calculate their means, run a *t*-test, and determine if the two populations are statistically different or not.

2. Under Trial 1 in cells B29–C29, re-enter by hand the results you obtained in cells B25 and C25.

Now that you've run your first trial and recorded your results, you are ready to run 99 more trials.

3. Switch to Manual Calculation.

Under Tools | Options | Calculation, select Manual Calculation.

4. Write a macro to run 99 more trials and record results in cells B30–C128.

Open the macro program and assign a shortcut key (refer to Exercise 2 for details on building macros). In Record mode, perform the following tasks:

- Select Tools | Macro | Record New Macro. Name your macro and assign it a shortcut key. For example, you might name your macro Type_I and assign it the shortcut "control t". Every keystroke you now make will be recorded as part of the macro.
- Press F9, the calculate key, to obtain new random samples from Population 1 and Population 2.
- Use your mouse to highlight cells B25 and C25, the new *t*-test statistic probability and significance result, and open Edit | Copy.
- Highlight cell B28, then go to Edit | Find. A dialog box will appear. You want to leave the Find What box completely blank, and search by columns. Click the Find Next button, then Close. Excel will move your cursor to the next blank cell in column B.
- Select Edit | Paste Special | Paste Values.
- You're finished. Select Tools | Macro | Stop Recording. Now when you press your shortcut key 99 times, your new results will automatically fill into the appropriate cells. Run your macro until you have results from 100 trials.

5. Save your work.

Our first five results looked like Figure 2; yours will very likely look different.

	A	B	C
28	**trial**	**t-test**	**significant?**
29	1	0.602499	0
30	2	0.910298643	0
31	3	0.70263163	0
32	4	0.810947176	0
33	5	0.099062869	0

Figure 2

D. Calculate Type I error rate.

1. Set up new headings as shown in Figure 3:

	A	B	C
130	**number of tests showing**		
131	**significant differences = >**		
132	**probability of Type I error = >**		

Figure 3

Switch back to automatic calculation, and visually inspect the *t*-test probabilities you obtained in your trials. Most of the results should indicate that the two populations are not statistically different from each other. Occasionally, however—about 5 times in 100—you will conclude that the two populations are different even though they have exactly the same mean height (μ) and standard deviation (σ^2). These are Type I errors. By a sampling fluke, you concluded the populations were different when in fact they are not.

2. In cell C131, use the **SUM** function to count the number of Type I errors committed.

We used the formula **=SUM(C29:C128)**.

3. In cell C132, calculate the Type I error rate as the number of Type 1 errors divided by 100 trials.

We used the formula **= C131/100**.
Your answer should be somewhat close to 0.05 because you established a Type I error rate of 0.05 in cell A25.

4. Save your work, and answer Question 1 at the end of the exercise.

E. Type II errors and power.

1. Enter values shown in cells F5–G6 (see Figure 1).

Now let's switch gears and think about Type II errors, which we'll deal with in Columns E, F, and G. Let's assume that the two populations really have different underlying distributions in terms of height. In cell F5, enter 45 to indicate that population 1 has an average height (μ) of 45 mm and a standard deviation (σ^2) of 5 mm (entered in cell G5). In cell F6, enter 50 to indicate that population 2 has an average height (μ) of 50 mm and a standard deviation (σ^2) of 5 mm (entered in cell G6). The effect size is entered in cell F7 as **=ABS(F5-F6)**. Although the effect size may seem small, these differences in height might be biologically meaningful, and you'd like to know this.

2. Calculate the effect size in cell F7.

3. Enter 0.05 in cell E25.

Set $\alpha = 0.05$ in cell E25.

4. Obtain samples from your population, and run 100 trials as you did earlier. You will need to create a new macro to keep track of results from 100 trials in cells F29–G128.

You'll sample from these populations, calculate a *t*-test, determine if you conclude the two populations are statistically different or not, and run 100 trials in total. Your spreadsheet columns E, F, and G should look like columns A, B, and C in appearance, although you will be sampling from different populations. In case you get stuck, the formulae we used are given at the top of the next page:

- F10 – F19 **=NORMINV(RAND(),F5,G5)**
- G10 – G19 **=NORMINV(RAND(),F6,G6)**
- F21 **=AVERAGE(F10:F19)**
- G21 **=AVERAGE(G10:G19)**
- F22 **=STDEV(F10:F19)**
- G22 **=STDEV(G10:G19)**
- F25 **=TTEST(F10:F19,G10:G19,2,2)**
- G25 **=IF(F25>E25,0,1)**

5. Set up headings as shown in Figure 4.

	E	F	G
130	**number of tests NOT showing**		
131		**significant differences =>**	
132		**probability of Type II error = β=>**	
133		**Power = 1 - β =>**	

Figure 4

6. In cell G131, use the **COUNTIF** formula to count the number of tests *not* showing a significant difference.

Remember that the populations really are different biologically, and we're trying to determine if they are different based on our samples. The **COUNTIF** formula counts the number of cells within a range that meet a given criterion. It has the syntax **COUNTIF**(**range**,**criteria**). We used the formula **=COUNTIF(G29:G128,0)** to count the number of times our *t*-test was not significant. These are the Type II errors. By a sampling fluke, you concluded that the populations are not different when in fact they are.

7. Calculate the probability of a Type II error (β) in cell G132.

Remember that a Type II error is falsely concluding that the two populations are similar when in fact they are different. Enter the formula **=G131/100** in cell G132. Is your Type II error rate acceptable, or is it too high for your liking?

8. Calculate power as 1– β in cell G133.

Enter the formula **=1– G132** in cell G133.
Scientists usually calculate the **power** of their design to detect differences assuming that they really exist, rather than reporting the probability of a Type II error. Remember that power is simply 1– β.

9. Save your work, and answer Questions 2–6.

QUESTIONS

1. If you change α in cell A25 to 0.1, approximately how many Type I errors are you likely to make if you run 100 trials again? How many Type I errors are you likely to commit if you set α to 0.01?
2. How does decreasing the standard deviation of the two populations affect Type II error rates and power? Enter 1 in cells G5 and G6. Press F9, the calculate key, 20 times and examine the significance of your 20 *t*-tests in cells F25 and G25. Keep track of the number of Type II errors out of 20 trials.
3. How does increasing the standard deviation of the two populations affect Type II error rates and power? Enter 10 in cells G5 and G6. Press F9 20 times and keep track of the number of Type II errors out of 20 trials.
4. How does effect size influence Type II error rates? Enter 45 in cell F5 and enter 55 in cell F6 (effect size = 10). Enter 5 in cells G5 and G6. Press the F9 key 20 times and keep track of the number of Type I and Type II errors out of 20 trials.
5. Does changing the α level in cell E25 affect β or power? Clear your macro results in cells F29–F128 and run 100 trials with varying α levels. Interpret your results.

6. How does sample size affect Type I and Type II error rates? Set cells B5–B6 and cells F5–G6 back to their original values. Then, develop a new model with population sizes of 1000 individuals, and compare the Type I and Type II error rates for populations of size 10 (currently modeled) with your new populations.

LITERATURE CITED AND FURTHER READINGS

Johnson, D. H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63(3): 763–772.

Sokal, R. R. and F. J. Rohlf. 1981. *Biometry*, 2nd Edition. W. H. Freeman, New York.

Taylor, B. L. and T. Gerrodette. 1993. The uses of statistical power in conservation biology: The vaquita and northern spotted owl. *Conservation Biology* 7: 489–500.