

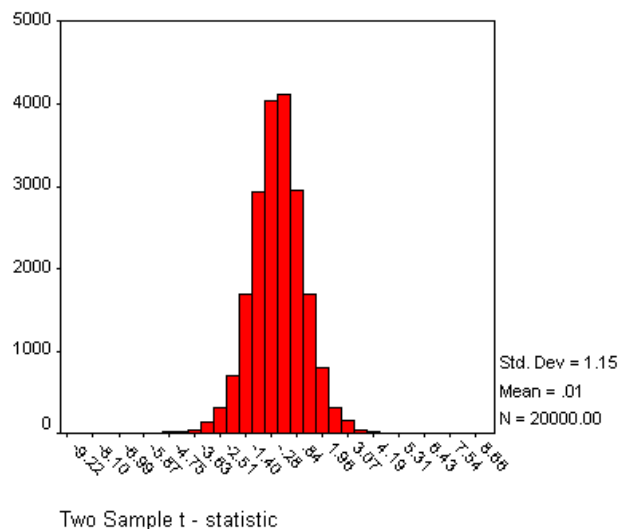
Biology 458 Biometry

Lab 5 - Risk Analysis, Robustness, and Power

I. Risk Analysis

The process of statistical hypothesis testing involves estimating the probability of making errors when, after the examination of quantitative data, we conclude that the tested null hypothesis is false or not false. These 'errors' can be thought of as the "risks" of making particular kinds of mistakes when basing a decision on the examination of quantitative data. To determine or estimate 'risks' (the probability of Type I and Type II errors), we follow a formal hypothesis testing procedure in which we insure or assume that our data meet certain conditions (e.g., independence, normality, and homogeneity of variances for parametric tests and independence and continuity for non-parametric tests). We state a decision rule prior to applying our test which amounts to specifying α , the significance level for our test, but also is synonymous with setting our Type I error rate and defining our region of rejection (i.e., the values of the test statistic for which we will reject the null hypothesis). Since we control or specify our Type I error rate prior to performing a test, if we meet the assumptions of the test, then the actual frequency of Type I errors we would make, if we repeated our experiment many times under the same conditions, will equal to the nominal rate we specify.

For example, the histogram below represents 20,000, 2-sample t - tests computed from data that were derived by sampling 5 subjects from each of two 2 underlying populations with exactly identical means (9.0) and variances (2.0). Note that this histogram is the sampling distribution of the two-sample t - statistic for samples of size $n_1 = n_2 = 5$, and that the distribution has a mean suspiciously close to 0, and a variance suspiciously close to 1.0.



If our t -test is working well and we perform an upper one-tailed test with $\alpha = 0.05$, then 5% of our t -values would fall in the region of rejection even when the null hypothesis is exactly true! In the example above, 20 sets of 1000 t -tests were conducted and one would expect that $5\% \times 1000 = 50$ tests to be significant at the 5% level. In fact the average number of Type I errors (significant t -tests even when the null hypothesis was true) was 50.8 ± 2.6359 (se).

II. Robustness

In many cases, we may be in error when we assume that our data follow a particular distribution or when we claim that the variances in our treatment groups are equal. When this situation arises, the distribution of the statistic that we use in testing our hypothesis may differ from that expected when the assumptions of the test are met. For example, if we use a t -test to compare a sample estimate of the mean to a theoretical value, the test statistic does not have an exact t -distribution if the population of values of our random variable, x , is not normally distributed. Similarly, if we are performing a separate groups t -test and assume that the treatment group variances are equal when in fact they are not, our estimates of the probability of making Type I and Type II errors will be distorted.

Therefore, the consequence of not meeting the assumptions of normality and homogeneity of variances is that our estimates of the probability of making Type I or Type II errors will be distorted. For instance, we may conclude from a t -test that the chance we are making a Type I error when we reject the null hypothesis is 5%. But, if x is not a normally distributed random variable, the true probability of making a Type I error might be more than 5% because the test statistic does not follow exactly the t -distribution with the specified degrees of freedom.

The ability of a specific statistical hypothesis test to provide accurate estimates of the probability of Type I and Type II errors, even when the underlying assumptions are violated, is called **robustness**. Some hypothesis tests are more robust to deviations from certain underlying assumptions than others. The type and magnitude of the deviation of the data from the assumptions required by a test is often important in choosing the appropriate statistical test to apply. Tests of hypotheses are used in many situations where the underlying assumptions are violated. Therefore, robustness is a desirable property.

The following table summarizes the results of a large number of computer simulations to determine the consequences of violating the assumptions of normality and equality of variances in the two sample t -test. For each simulation, samples were drawn from two underlying populations with exactly equal means and either equal or unequal variances. Since the population means were known to be equal, anytime we reject the null hypothesis we are committing a Type I error. The values given in this table are the number of Type I errors (out of 1000 replicate simulation runs) when the null hypothesis was tested versus an upper 1-tailed alternative hypothesis. At an $\alpha = 0.05$ level, we would expect $0.05 \times 1000 = 50$ Type I errors in each instance, if the test was performing

according to expectations. Since 20 - replicate runs of 1000 simulations were performed for each table entry, the standard error of each entry is also presented. Values that are substantially lower than 50 or greater than 50 indicate that the test is actually being performed at an α level other than the nominal 5% we intended. In other words we either make to many Type I errors (reject when true) or too few (fail to reject when false). The simulations were performed on populations that were either normally or uniformly distributed with means equal to 9.0. In the case where variances were equal, variances both = 4.0. For the case of unequal variances $var_1 = 4.0$, $var_2 = 16.0$.

Assessment of the Robustness of the t - test to violations of the assumptions of Normality and Equality of Variances (values in table are number of Type I errors in 1000 trials \pm se)

Sample Sizes	Equal Variances ($var_1 = var_2 = 4.0$)		Unequal Variances ($var_1 = 4.0, var_2 = 16.0$)			
	n1,n2		pooled variance estimate		separate variance estimate	
	normal	uniform	normal	uniform	normal	uniform
	A	B	C	D	E	F
5,5	49.5 \pm 1.4	50.6 \pm 1.5	54.2 \pm 2.1	71.1 \pm 2.1	48.8 \pm 1.8	61.1 \pm 1.8
20,20	48.6 \pm 1.7	50.4 \pm 1.5	51.8 \pm 1.3	80.7 \pm 1.9	46.7 \pm 1.5	84.6 \pm 2.0
100,100	53.1 \pm 1.5	52.6 \pm 1.3	50.4 \pm 1.5	138.9 \pm 2.3	51.1 \pm 1.8	141.3 \pm 2.9
5,20	53.4 \pm 2.0	51.5 \pm 1.6	10.0 \pm 0.8	15.8 \pm 0.9	47.4 \pm 1.9	71.35 \pm 1.7
20,100	52.4 \pm 1.4	51.3 \pm 1.7	6.7 \pm 0.7	19.7 \pm 0.9	50.4 \pm 1.6	106.7 \pm 2.9
5,100	49.4 \pm 1.9	50.4 \pm 1.4	0.90 \pm 0.2	2.9 \pm 0.4	49.3 \pm 1.3	78.9 \pm 2.1
20,5	-	-	131.4 \pm 2.0	163.9 \pm 2.5	46.9 \pm 1.9	65.3 \pm 1.8
100,20	-	-	144.4 \pm 2.2	203.9 \pm 2.6	53.6 \pm 1.9	81.8 \pm 1.6
100,5	-	-	185.9 \pm 2.7	227.7 \pm 2.3	44.9 \pm 1.4	63.0 \pm 2.0

From this table we see that when variances are equal, the t - test performed well regardless of whether the population was normally distributed or the sample sizes were equal (columns **A** and **B**). However, when the variances are unequal the test only performs well if the population is normally distributed and sample sizes are equal (first three rows in column **C**). If we use the Satterthwaite correction with unequal variances, then the t - test also performs well for unequal variances whether or not sample sizes are equal (data column **E**). In virtually all instances, the t - test does not perform well when the underlying population is non-normal and variances are unequal (data column **F**).

II. Power

Power is a measure of the ability of a statistical test to detect an experimental effect that is actually present. Power is an estimate of the probability of rejecting the null-hypothesis (H_0) if a specified alternative hypothesis (H_a) is actually correct. Power is equal to one minus the probability of making a Type II error, β , (failing to reject H_0 when it is false): $\text{power} = 1 - \beta$; thus, the smaller the Type II error, the greater the power and, therefore, the greater the sensitivity of the test. The level of power will depend on several factors: 1) the magnitude of the difference between H_0 and H_a , which is also called the "effect size," 2) the amount of variability in the underlying population(s) (the variances), 3) the sample size, n , and 4) the level of significance chosen for the test, α or the probability of a Type I error. For more information on statistical power analysis see the supplemental lecture notes on [power](#) and [more about power analysis](#).

As long as we are committed to making decisions in the face of incomplete knowledge, as every scientist is, we cannot avoid making Type I and Type II errors. We can, however, try to minimize the chances of making them. We directly control the probability of making a Type I error by our selection of α , the significance level of our test. By setting a region of rejection, we are taking a risk that a certain proportion of the time (for example 5%, when $\alpha = 0.05$) we will obtain values of our test statistic that would lead us to reject the null hypothesis (fall in the region of rejection) even when the null hypothesis is true.

How can we reduce the probability of making a Type II error? One obvious way is to increase the size of the region of rejection. In other words, increase α . Of course, we do so at the cost of an increased probability of making Type I errors. Every researcher must strike a balance between the two types of error. Other ways to reduce the probability of making a Type II error are to increase the sample size or to reduce the variability in the data. Increasing sample size is simple, but reducing variability in the observations is also an important means to improve the power of tests. The process of "experimental design" is one in which research effort is allocated to insure the most powerful test for the effort expended is actually conducted.

The probability of making a Type II error and the power of a statistical test are more difficult to determine because they require one to specify a quantitative alternative hypothesis. However, if one can specify a reasonable alternative hypothesis, guides for calculating the power of a test, or the sample size necessary to achieve a desired level of power are now becoming more widely available. The most comprehensive book on statistical power analysis is by Cohen (1977), but many web-based power calculators and statistical packages now include functions to permit the determination of power and sample size. [Click here](#) to view the available web-based power calculators or see the table below for links to power calculators for t -tests. [Click here](#) to link to the website where you can download the program PS, a Windows based power calculator.

III. Exercises

Calculate power or sample size for the following tests: Use the program PS (note PS gives power and sample size values assuming a two tailed test, to get 1-tailed results for a test at the $\alpha = 0.05$ level, use an $\alpha = 0.1$), a web based calculator, or the methods outlined in your text to solve these problems.

A. Given that the variance in arsenic concentration in drinking water is 8 ppb, what is the power of a test based on 10 samples to determine if arsenic levels exceeds the public health standard of 5 ppb by 2 ppb? Assume that the test is performed at $\alpha = 0.05$. What sample size would be necessary to have $\alpha = \beta = 0.05$? Plot a curve for power versus sample size for this example.

B. An ecologist is contemplating a study on the effects of ice plant on seedling growth in a native plant. Two treatments are anticipated: treatment 1 will examine growth of the native plant at locations where ice plant has not previously grown, and treatment 2 will examine growth of the native at locations from which ice plant has been removed. Given a preliminary estimate that plants reach an average height of 35 cm and have a variance of 20 cm, what sample size is necessary to detect a 25% reduction in height caused by ice plant with power of 80%? (Assume that the test is to be performed at the $\alpha = 0.05$ level). Would it be better to use an independent groups t -test or a paired t -test? (hint: using the same preliminary estimates calculate the sample size required to achieve the desired power for both a separate groups and a paired t -test). What consequence would there be to the independent group t -test to using unequal sample sizes? (Hint: use the Iowa calculator to answer this question.)

Perform the power and sample size calculations requested in parts A and B, and turn in a brief write-up of your results.

Further Instructions on completing the Lab

Take note that the power calculator "PS" always assumes that you are performing a 2-tailed test. Therefore, to perform a test with $\alpha = 0.05$, you need to enter $\alpha = 0.1$ in the appropriate box for α . Also, the parameter m in "PS" is the ratio of sample size in the two treatments when you are performing an independent groups t -test. You get different results depending on which sample size you put in the numerator or denominator. Therefore, avoid using "PS" to address questions about unequal samples sizes. Use Russ Lenth's the web-based power calculator at the University of Iowa.

When doing the problems read them carefully and extract information on variability, effect size, Type I error rate, if the problem should be a one tailed test, etc. so that you can enter values in the calculators to compute power or sample size. When completing

your write-up state which calculator you used and what values you entered so that I can tell where you went wrong. Also, for part B of the exercise to determine the effect of unequal sample sizes on power, keep the total sample size constant and vary allocation between treatments. For example, if $n_1 = 4$ and $n_2 = 4$, compare the power to $n_1 = 6$ and $n_2 = 2$, so the total sample size remains 8.

Links to Web Based Power and Sample Size Calculators for t-tests	
1 - sample tests	http://stat.ubc.ca/~rollin/stats/ssize/index.html
2 - sample tests	http://stat.ubc.ca/~rollin/stats/ssize/index.html
1 and 2 sample tests	http://www.math.uiowa.edu/~rlenth/Power/index.html