

CHAPTER FIVE

HYPOTHESIS TESTING FOR TWO OR MORE SAMPLES

Choosing the Correct Test

In the prior unit we looked at the logic of hypothesis testing and how a one-sample hypothesis test compares a sample to a hypothesized value to see if there is a significant difference. Hypothesis testing in itself is a challenging concept to learn, but once it is embraced, it is far easier to learn about the different tests.

One of the biggest challenges researchers face is in deciding which hypothesis test technique to use. This requires answering four questions:

1. What is the level of measurement?
2. Can a parametric test be used?
3. How many samples are involved?
4. If two or more samples are involved, are the cases related or independent?

1. What is the level of measurement? The level of measurement for a variable can be Nominal, Ordinal or Scale (i.e., Interval or Ratio). This topic was addressed in Unit 1, and it is important because certain variable types can only be tested with certain hypothesis tests. The one sample hypothesis test for means, from unit 4, compared a sample mean to a hypothesized value (which requires the variable to be Interval or Ratio); similarly the one sample hypothesis test for proportions compares a sample proportion to a hypothesized value (which requires the variable to be categorical, and can be Nominal or sometimes Ordinal).

2. Can a parametric test be used? When the hypothesis involves looking at means or other population computations (known as parameters), we use what is known as a parametric test, and this test has a set of assumptions that must be met (one of which is typically that the data is normally distributed). If the assumptions for the parametric test cannot be met, then we can use an alternate approach known as a nonparametric test (not as powerful but little to no assumptions to worry about). As with unit 4, this unit will only look at the parametric tests.

3. How many samples are involved? The most basic hypothesis test involves looking at a single sample and computing a value from that sample to compare to a hypothesized value (which was addressed in unit 4). More commonly, we look at two samples to compare a difference or a relationship in some measured value (e.g., compare the mean age for males vs. females); for these

cases, we will use one of several t-tests. Sometimes you will see three or more samples for which we assess the differences in some measured value (e.g., compare the mean age for married vs. single vs. divorced vs. widowed); for these cases, we will use an Analysis of Variance (ANOVA).

4. If two or more samples are involved, are the cases related or independent? If multiple samples are used, it is possible that values from one sample are linked to the other(s), such as in a pre/post test (e.g., a diet with weight before vs. after). There are hypothesis tests designed for independent samples and ones designed for related samples, so it is important that the right test be used for a given situation.

T-test for Independent Samples

Supposed you wanted to know if men or women could do more sit-ups. Or if homeschooled students do score better on SATs than those in a traditional classroom do. Or if people with dogs tend to live longer than those without dogs. In each case you can conduct a t-test for independent means to determine if a significant difference exists. In each case there is a dependent variable (the scale variable) from which we compute a sample mean, and an independent variable (the categorical variable) which has the two subgroups.

Conducting this hypothesis test is essentially like computing two confidence intervals to see if they overlap or are clearly different. In the case of SATs, if a sample of homeschooled students averaged 545 Math and a sample of traditional students averaged 490 Math, you could use the sample sizes and standard deviations to compute the margins of error. Then you would have a confidence interval for each group, and if the low point for the homeschooled were still higher than the high point for traditional students, the two intervals would not overlap and there would clearly be a significant difference. But if the intervals did overlap, you would not be able to conclude a difference exists since it is possible the means are the same.

Figure 1: Screen display from **Hypothesis Tests Two Samples.xls**

Two Sample Hypothesis Test for the Mean (Independent Samples)				
Level of Significance	0.05			
	Homeschoolers	Traditional		
Sample Mean	545	490		
Sample Size	25	30		
Sample Standard Deviation	100	100		
Standard Error (Computed)	27.08			
Test Statistic (Computed)	2.03			
	Lower Crit Value	Upper Crit Value	p-Value	Decision
Two-Tailed Test H1: ($\mu_1 < \mu_2$)	-2.0057	2.0057	0.0473	Reject the null hypothesis
Upper-Tail Test H1: ($\mu_1 > \mu_2$ OR $\mu_2 < \mu_1$)	n/a	1.6741	0.0235	Reject the null hypothesis
Lower-Tail Test H1: ($\mu_1 < \mu_2$ OR $\mu_2 > \mu_1$)	-1.6741	n/a	0.9765	Do not reject the null hypothesis
If you REJECT the null hypothesis, conclude that H1 is true.				
If you DO NOT REJECT the null hypothesis, there is insufficient evidence to conclude that H1 is true.				
NEVER conclude that the null hypothesis is true (i.e., we CANNOT ACCEPT the null).				

In the above display from Hypothesis_Tests_Two_Samples.xls, we are comparing SAT Math scores for a sample of homeschooled students and a sample of traditional students.

Chapter Five: Hypothesis Testing for Two or More Samples

The hypotheses would be written as:

Ho: The mean Math SAT of homeschoolers equals that of traditional students.

Ha: The mean Math SAT of homeschoolers does not equal that of traditional students.

If you had reason to believe one group would excel over the other, you could word the hypothesis to reflect that (e.g., The mean Math SAT of homeschoolers is greater than that of traditional students). Once you determine the correct hypothesis test to use, you must determine the direction of the test. If you are testing whether or not a difference exists in the two means, you would use a two-tailed test. If you are testing whether the first group mentioned has a mean that is greater than the second group, you would use an upper-tailed test. And if you are testing whether the first group mentioned has a mean that is less than the second group, you would use a lower-tailed test. So if you wished to test whether homeschoolers scored higher, you would use an upper-tailed test if the homeschoolers were in the first group, and a lower-tailed test if they were in the second group (the results are identical, so it doesn't matter which one you put them in, as long as you choose the appropriate direction).

So in this case, if we are conducting a two-tailed test, we get a p-value of .0473, which is less than our .05 significance level, so the null hypothesis is rejected and we can conclude a difference in the mean Math SATs for the homeschoolers and traditional students. If we were conducting a one-tailed test to determine if the homeschoolers scored higher, then we would have a p-value of .0235 which is less than .05, so the null hypothesis is rejected, and we can conclude that the mean Math SATs is higher for homeschoolers than for traditional students.

And so it seems that once you choose the correct hypothesis test and load the data correctly, the rest is similar to the procedures for a one-sample test (i.e., choose the direction, decide to reject or not, and state the conclusion).

What if the data is categorical instead of scale, and we are measuring proportions instead of means? Just as we were able to compare a single sample's proportion to a hypothesized value, we can compare the proportions from two samples to each other. Perhaps you wish to know whether the proportion of men who smoke is greater than that of women? Or if there is a difference in the proportion of homeschoolers who go on to college vs. traditional students? Or if those with an MBA get a better response when applying for an open position than those without an MBA? In each case there is a dependent variable (a categorical variable) from which we compute a sample proportion, and an independent variable (also a categorical variable) which has the two subgroups.

Chapter Five: Hypothesis Testing for Two or More Samples

Figure 2: Screen display from **Hypothesis Tests Two Samples.xls**

Two Sample Hypothesis Test for the Proportion				
Level of Significance	0.05			
	MBA	non-MBA		
Number of Successes	56	90		
Sample Size	100	200		
Proportion (Computed)	0.56	0.45		
Z Test Statistic (Computed)	1.80			
	Lower Crit Value	Upper Crit Value	p-Value	Decision
Two-Tailed Test H1: (P1 <> P2)	-1.9600	1.9600	0.0723	Do not reject the null hypothesis
Upper-Tail Test H1: (P1 > P2 or P2 < P1)	n/a	1.6449	0.0362	Reject the null hypothesis
Lower-Tail Test H1: (P1 < P2 or P2 > P1)	-1.6449	n/a	0.9638	Do not reject the null hypothesis
If you REJECT the null hypothesis, conclude that H1 is true.				
If you DO NOT REJECT the null hypothesis, there is insufficient evidence to conclude that H1 is true.				
NEVER conclude that the null hypothesis is true (i.e., we CANNOT ACCEPT the null).				

In the above display from Hypothesis_Tests_Two_Samples.xls, we are comparing the positive responses received by MBA grads applying for a senior management job opportunity in their field vs. non-MBAs. Since we would have a reasonable expectation that those with an MBA would do better, the hypotheses would be written as:

Ho: The proportion of MBA grads who get a positive response in applying for a senior management job equals that of non-MBA grads.

Ha: The proportion of MBA grads who get a positive response in applying for a senior management job is greater than that of non-MBA grads.

Since we are testing if the first group (the MBA grads) had a greater value than the second group (the non-MBAs), we conduct an upper-tailed test. The p-value is .0362 which is less than .05, and so we reject the null hypothesis and conclude that MBA grads have a greater success rate in getting interviews for senior management job opportunities. If we didn't have reason to believe one group would excel over the other (which might happen if we were comparing MBA grads from two different schools of similar quality), you would word the hypothesis to reflect that (e.g., There is a difference in the proportion of MBA grads who get a positive response in applying for a senior management job as compared to non-MBAs) and thus use a two-tailed test. If this were a two-tailed test, the p-value of .0723 which is greater than .05 would result in our not rejecting the null hypothesis; we would then state that there is insufficient evidence to conclude a difference in the proportions of MBA grads getting senior management interviews vs. non-MBAs.

T-test for Paired Samples

While the t-test for independent samples gives us a way to test for significant differences between two groups, it often takes a large difference or a large sample size for the difference to be considered significant enough to reject the null hypothesis. A way to make the test more powerful and increase the chances of finding significant differences is to use paired samples where possible. A paired sample is a one-to-one linking between the data in the two groups such that each value in group 1 corresponds to a single value in group 2.

Whereas the t-test for independent samples computed a mean $\pm$ a margin of error within each group and compared the two, the t-test for paired samples computes the differences within each pair of observations and analyzes the differences. While the spread of data in each subgroup could be quite large, the differences in the paired samples are likely to be small numbers, and so you could potentially reject the null with minor differences as long as a large majority of the differences are in the same direction (i.e., mostly positive or mostly negative).

The most common occurrence is a pre/post test in which we would compare someone's before and after scores, such as one's weight before and after a diet. Another possibility occurs when the same item is measured under two circumstances, such as having each house in a sample appraised by the same two appraisers. Yet another possibility involves manually pairing values based on known information (e.g., pairing husbands with wives, or identical twins, or just two people that are similar in factors the research finds critical). Pairing the samples together makes the test more powerful, but it is not always practical to do and it involves a lot of planning since both members of a pair must be collected. What makes a paired samples test more powerful is the ability to reject the null with a much smaller sample than the independent samples test and with smaller differences.

Figure 3: Screen display from **Hypothesis Tests Two Samples.xls**

Two Sample Hypothesis Test for the Mean (Paired Samples)																								
Item	Appraiser 1	Appraiser 2	Level of Significance	0.05																				
1	135	128																						
2	110	105	Standard Error (Computed)	4.40																				
3	131	119	Test Statistic (Computed)	3.30																				
4	142	140																						
5	105	98																						
6	130	123																						
7	131	127																						
8	110	115																						
9	125	122																						
10	149	145																						
11																								
12																								
13																								
14																								
15																								
16																								
17																								
18																								
19																								
20																								
21																								
22																								
<table border="1"> <thead> <tr> <th></th><th>Lower Crit Value</th><th>Upper Crit Value</th><th>p-Value</th><th>Decision</th></tr> </thead> <tbody> <tr> <td>Two-Tailed Test H1: ($\mu_1 < \mu_2$), ($\mu_1 > \mu_2$)</td><td>-2.2622</td><td>2.2622</td><td>0.0092</td><td>Reject the null hypothesis</td></tr> <tr> <td>Upper-Tail Test H1: ($\mu_1 > \mu_2$), ($\mu_1 > \mu_2$ OR $\mu_2 < \mu_1$)</td><td>n/a</td><td>1.8331</td><td>0.0046</td><td>Reject the null hypothesis</td></tr> <tr> <td>Lower-Tail Test H1: ($\mu_1 < \mu_2$), ($\mu_1 < \mu_2$ OR $\mu_2 > \mu_1$)</td><td>-1.8331</td><td>n/a</td><td>0.9954</td><td>Do not reject the null hypothesis</td></tr> </tbody> </table>						Lower Crit Value	Upper Crit Value	p-Value	Decision	Two-Tailed Test H1: ($\mu_1 < \mu_2$), ($\mu_1 > \mu_2$)	-2.2622	2.2622	0.0092	Reject the null hypothesis	Upper-Tail Test H1: ($\mu_1 > \mu_2$), ($\mu_1 > \mu_2$ OR $\mu_2 < \mu_1$)	n/a	1.8331	0.0046	Reject the null hypothesis	Lower-Tail Test H1: ($\mu_1 < \mu_2$), ($\mu_1 < \mu_2$ OR $\mu_2 > \mu_1$)	-1.8331	n/a	0.9954	Do not reject the null hypothesis
	Lower Crit Value	Upper Crit Value	p-Value	Decision																				
Two-Tailed Test H1: ($\mu_1 < \mu_2$), ($\mu_1 > \mu_2$)	-2.2622	2.2622	0.0092	Reject the null hypothesis																				
Upper-Tail Test H1: ($\mu_1 > \mu_2$), ($\mu_1 > \mu_2$ OR $\mu_2 < \mu_1$)	n/a	1.8331	0.0046	Reject the null hypothesis																				
Lower-Tail Test H1: ($\mu_1 < \mu_2$), ($\mu_1 < \mu_2$ OR $\mu_2 > \mu_1$)	-1.8331	n/a	0.9954	Do not reject the null hypothesis																				
If you REJECT the null hypothesis, conclude that H1 is true.																								
If you DO NOT REJECT the null hypothesis, there is insufficient evidence to conclude that H1 is true.																								
NEVER conclude that the null hypothesis is true (i.e., we CANNOT ACCEPT the null).																								

The hypothesis being tested here is in regards to a difference in appraisals of 10 homes by 2 appraisers.

Ho: There is no difference in the mean appraised values of homes given by appraiser 1 and 2.

Ha: There is a difference in the mean appraised values of homes given by appraiser 1 and 2.

Note that all prices shown in the data are in thousands of dollars. It doesn't matter if these are high-priced homes or if they are located in slums; since the same homes are being evaluated by both appraisers, all we care about is how much they differ on each home. With independent samples, you would have to assume that the two groups were randomly assigned and unbiased so that the results don't unfairly portray a difference in appraisers, and you would need a large enough sample size to support your findings, but paired samples eliminate this concern. The real problem is that paired samples data doesn't just happen; while you can retrieve data from independent samples already collected, paired samples data needs advance planning (in this case you would have to set it up for both appraisers to appraise the same homes on similar weather days and within a day or two of each other so that the conditions of the homes don't change).

Looking at the output here, we would use a two-tailed test since we are testing a difference. The p-value of .0092 is less than .05, so we would reject the null hypothesis and conclude that there is a difference in the appraised values given by the two appraisers. While the average difference in appraised values is quite small, what makes this significant is that in 9 out of 10 cases, appraiser 1 had higher values. The differences could have been \$1000 per house and it would still be significant if 9 out of 10 went in the same direction.

Analysis of Variance

When we have two samples to test, the obvious tool is the t-test, and you need only decide if you are testing means or proportions and if the samples are independent or paired. When you have three or more samples, you could still opt to use t-tests, comparing two groups at a time, but doing this is not very efficient and is error-prone. If you could imagine comparing five groups, A – E, and you tested two at a time (A vs. B, A vs. C, ... D vs. E), you would have to conduct 10 separate t-tests, with each one having an inherent degree of error based on the significance level. In doing 10 such t-tests, the probability of incurring a type 1 error is likely to exceed 50%, possibly a lot higher, so you are almost guaranteed to get false results. What we need is a way to compare multiple groups without increasing the probability of error. The Analysis of Variance (ANOVA) is such a test.

While the math behind the ANOVA can be quite cumbersome and often frightening, the computerized version is actually quite simple. The test is always non-directional since we only test if there is or is not a difference. And there is no limit as to the number of groups being compared (even though the Excel file provided here is limited to five columns).

In the example shown here, we are testing the difference in how students perform when given a quiz in different formats. Assume a class of 18 students takes a 20-question quiz; 6 of them take a True/False quiz, 6 take a Multiple Choice quiz and 6 take a Fill-in the Blank quiz. Note that each quiz addresses the same 20 questions in the same order. The 18 students were randomly assigned the quizzes. The results are as follows:

Chapter Five: Hypothesis Testing for Two or More Samples

Figure 4: Screen display from **ANOVA.xls**

Analysis of Variance				
T/F	M/C	Fill-In		
15	13	9		
18	15	11		
17	16	13		
19	17	15		
19	19	14		
20	16	16		

SUMMARY				
Groups	Count	Sum	Average	Variance
T/F	6	108.000	18.000	3.200
M/C	6	96.000	16.000	4.000
Fill-In	6	78.000	13.000	6.800
0	0	0.000		
0	0	0.000		

ANOVA					
Source of Variation	SS	df	MS	F	F crit
Treatment	76.000	2	38.000	8.143	3.682
Error	70.000	15	4.667		
Total	146.000	17			

Null hypothesis:	$\mu_1 = \mu_2 = \mu_3$
Alternate hypothesis:	Not all the means are equal
Significance level:	0.05
p-Value	0.0040
Decision:	Reject the null hypothesis

Mean Differences Post-Hoc Tests				
Comparison	Difference of Means	Lower Conf. Limit	Upper Conf. Limit	Difference in Treatment Means???
T/F to M/C	2.000	-0.658	4.658	- no difference -
T/F to Fill-In	5.000	2.342	7.658	*** Means are different ***
M/C to Fill-In	3.000	0.342	5.658	*** Means are different ***

Ho: There is no difference in the mean quiz scores in the T/F, Multiple Choice and Fill-in formats.

Ha: Not all of the means are equal in the three quiz formats.

With a p-value of .0040 which is less than .05, we reject the null hypothesis and conclude that a difference exists somewhere...but where. The post hoc test that follows tests every possible pair (essentially doing the equivalent of a t-test for each pair). When comparing T/F to Multiple Choice, there was a difference of 2.00 in the mean quiz scores, and it was found to be not significant. When comparing T/F to Fill-in, there was a difference of 5.00 which was found to be significant. And when comparing Multiple Choice to Fill-in, there was a difference of 3.00 which was also found to be significant. So essentially, we find that Fill-in-the-blank quizzes result in significantly lower scores than either Multiple Choice or True/False.

Had there been five groups, there would have been ten different paired comparisons in the post-hoc test, and all ten comparisons would be handled in a single ANOVA. The sample sizes in each group did not have to be equal (they were in this example but it is not a requirement and is not common). The ANOVA table has a lot of information in it, but it is mainly intermediate computations, and the only critical information is the p-value that is derived from it.

CHAPTER FIVE KNOWLEDGE ASSESSMENT

Hypothesis Testing for Two or More Samples

Discussion Questions

DISCUSSION QUESTION 1

PROPORTIONALLY SPEAKING: If a random sample of 10 people found that 9 were pro-life (i.e., 90%), while another random sample of 1000 people from the same population found that 550 were pro-life (i.e., 55%), which would you find to be more significant? Why?

DISCUSSION QUESTION 2

PROBING THE POLLSTERS: When pollsters report results of a poll, they often include the margin of error but not much more. If candidate X has 54% support of the voters with a margin of error of 3%, that means the candidate is predicted to have between 51% and 57% support in the election. Before jumping to conclusions, what other information would you like the pollster to provide?

Practice Problems: Real Estate

Solutions are provided to practice problems so you can check your work.

Use the Real_Estate.xls file which consists of 100 homes purchased in 2007 and appraised in 2008. It includes variables regarding the number of bedrooms, number of bathrooms, whether the house has a pool or garage, the age, size and price of the home, what the house is constructed from, how far it is to the city center, and the appraisals from two agents.

PRACTICE PROBLEM 1:

It has been noted that swimming pools do nothing to increase value of a home. Using the sample of 100 homes, conduct a hypothesis test for two independent samples to determine if the mean sales differs for homes with and without a pool. Use a .05 significance level. *Note that this requires an assumption that homes with a pool vs. those without are not much different otherwise (if those with a pool are in better locations, made of better materials, are newer and larger, then those homes will be worth more and it won't have a thing to do with a pool. So for the sake of simplicity, let's assume that the homes with vs. without a pool have similar variability in them.*

PRACTICE PROBLEM 2:

How different can two real estate agents be in their appraisals? In the sample of 100 homes (in the data file), conduct a hypothesis test for paired samples and test if there is a difference in the mean appraisal prices given by these agents on the same homes. Use a .05 significance level.

CHAPTER FIVE KNOWLEDGE ASSESSMENT

Hypothesis Testing for Two or More Samples

PRACTICE PROBLEM 3:

If people are going to spend a lot on their homes by constructing them out of brick, are they going to take the plunge and get a swimming pool too? Using the sample of 100 homes (in the data file), conduct a hypothesis test of proportions to determine if the proportion of homes made of brick are more likely to have a swimming pool vs. homes made of other materials. Use a .05 significance level.

PRACTICE PROBLEM 4:

You might expect that homes with more bedrooms are worth more since they are probably larger, but is there more to the value (like location, construction, age, etc.)? Using the sample of 100 homes (in the data file), conduct a hypothesis test using Analysis of Variance to determine if there is a difference in the mean sale price of homes with 2 bedrooms vs. 3 bedrooms vs. 4 bedrooms vs. 5 bedrooms. Use a .05 significance level. *Since there are homes made of varying sizes at different locations and made of different material for this sample, it would be reasonable to assume that location and construction are not factors in this test.*

Assigned Problems: Student Data

Use the Student_Data.xls file which consists of 200 MBA at Whatsamattu U. It includes variables regarding their age, gender, major, GPA, Bachelors GPA, course load, English speaking status, family, weekly hours spent studying.

ASSIGNED PROBLEM 1:

The battle of the sexes lives on still today. Since admission standards do not address gender whatsoever, there should be an equally diverse group of men and women in school, but do they perform equally well. Using the sample of 200 students, conduct a hypothesis test for two independent samples to determine if the mean GPA differs for men and women. Use a .05 significance level.

ASSIGNED PROBLEM 2:

Can a student keep up their grade performance at the next level? Is a strong GPA at the Bachelors level a good predictor of a strong GPA at the Masters level, or are GPAs naturally going to decline since graduate school is tougher, or will GPAs automatically be higher in graduate school because of the 3.00 requirement to graduate and the treatment of a C as subpar instead of average? Using the sample of 200 students (in the data file), conduct a hypothesis test for paired samples and test if there is a difference in the mean GPA from the Bachelors to the Masters programs. Use a .05 significance level.

CHAPTER FIVE KNOWLEDGE ASSESSMENT

Hypothesis Testing for Two or More Samples

ASSIGNED PROBLEM 3:

Given the reasons why people get their Masters, you surmise that men are more likely to declare a major than women. Using the sample of 200 students (in the data file), conduct a hypothesis test of proportions to determine if the proportion of women with “no major” is greater than the proportion of men with “no major”. Use a .05 significance level.

ASSIGNED PROBLEM 4: You have probably heard that if you want something done, give it to a busy person. So is one’s employment status a factor in their academic performance? Using the sample of 200 students (in the data file), conduct a hypothesis test using Analysis of Variance to determine if there is a difference in the mean GPA for those who are unemployed vs. work part-time vs. work full-time.

T
R
A
I
S
-
K
E
N
D
A
L
L
,
A
L
I
C
E

4
0
5
1
T
S