

CHAPTER FOUR

HYPOTHESIS TESTING FOR ONE SAMPLE

The Logic of Hypothesis Testing

How can we prove a coin is perfectly fair? And can a doctor find you perfectly healthy? And can you prove your weight has not changed at all? The answer is NO in all cases. We can never prove equality or lack of change, but we can find evidence to the contrary. In medical tests, a physician looks for problems and in the absence of any being detected, they tell you that you tested negative and failed to find anything wrong with you. With a coin, if you flip five heads and five tails, you are not convinced of anything regarding that coin; likewise if you flip nine heads and one tail, you are not convinced but are starting to get suspicious. But if you flip 90 heads and 10 tails, you may rightly believe something is wrong, while 60 heads and 40 tails may not be quite enough to sway you.

If you improved your SAT by 10 points using a SAT improvement course, would you be satisfied? I doubt it, considering scores range from 200 to 800 on each section and are in 10-point increments. But if you improved your ACT by 10, you would be probably be thrilled (the ACT scale is from 0 to 36). Points are relative. There is a threshold for everyone such that one will draw a conclusion with little chance of error.

If you step on and off a scale, you notice the weights vary. Why? Do you lose weight merely by stepping on and off? What if you weighed in at 165, stepped off the scale, stepped back on and registered 164 – would you claim that you lost a pound from step aerobics by stepping off / on the scale? Or if you exercised for 20 minutes and got on the scale to see that same one pound loss, would you then feel better at attributing the weight loss to that exercise? At what point do you consider it a weight loss significant enough to declare it was attributed to exercise and not to the imbalance of the scale? If you lose 5 pounds in 5 days you might be convinced, but 3 pounds in 5 days may be shifts in water weight combined with scale imbalance. Where is the fine line?

When we have a theory about a parameter (the average is ..., the proportion is ..., etc.), we can test that theory via a hypothesis test. If the results are far enough from the hypothesized number, we can conclude that with minimal error, the statement must be true. If it isn't far enough, we state that we just lack sufficient evidence to draw such a conclusion. Just like with weight loss, which is a form of a hypothesis test, we look for evidence (often in the form of data) to make our case, and then make a decision after reviewing the evidence.

Chapter Four: Hypothesis Testing for One Sample

One of the best examples of hypothesis testing can be found in the judicial system. Here the null hypothesis (H_0) is that the defendant is innocent, while the alternate hypothesis (H_a) is that the defendant is guilty. No one is ever found innocent, because we can never prove anyone innocent (no matter what evidence you have about who did the crime, you will never know if the defendant was or wasn't involved somehow, so short of being in a coma for a year or so, we can never prove his innocence with evidence). Can you imagine how hard it would be if you had to be proven innocent? All of the examples of honesty cannot prove exclusive innocence, while one example of dishonesty surely proves guilt. This is why we assume innocence until proven guilty, but we never actually declare innocence. We either prove guilt, or we lack enough evidence to conclusively prove guilt; and the choice of verdicts are GUILTY or NOT GUILTY (meaning that there is insufficient evidence to find the defendant guilty). INNOCENT is not an option for jurors thankfully; they only judge if the evidence is sufficient or not to judge as GUILTY. All hypothesis tests are designed to either prove H_a (the alternate hypothesis) is true or have insufficient evidence to prove H_a is true. But let's continue with this example.

Anyway, the prosecutor's job is to gather evidence in the form of fingerprints, witnesses, blood, etc. and try to convince the jury that there is enough evidence to convict the defendant. The jury has a significance level that serves as a cutoff for error (known as *reasonable doubt*). The jury either decides that sufficient evidence exists beyond reasonable doubt and therefore rejects the null that he is innocent and finds the defendant guilty (meaning the chance of convicting by mistake is very small), or the jury finds that there is not sufficient evidence to convict and therefore finds the defendant not guilty. Note that a "not guilty" verdict does not mean they find the defendant innocent, because the evidence is still there, but the evidence wasn't enough to convince them beyond reasonable doubt (i.e., too great a risk of error of sending an innocent person to prison). This is like showing a 2-pound weight loss on the scale – yes you might have actually lost weight, but you are not convinced beyond reasonable doubt of that fact; you don't declare that you haven't lost weight at all, but that there is insufficient evidence to prove you lost weight.

And how is *reasonable doubt* assessed? In hypothesis testing, we call it a significance level, and the norm is to set it to .05 or 5%. This is essentially your tolerance for error, meaning that you are willing to reject the null by mistake 5% of the time. The value can be increased or decreased at the will of the researcher, depending on the circumstances. In a jury trial, would you want to convict a person knowing that there is a 5% probability you might be convicting an innocent person? That might be too high. But if you were screening airline passengers for weapons, would you care about having the metal detector go off by mistake and having to search a suspect? Chances are, you wouldn't mind having that error occur more often if it meant saving lives (especially since there is no real cost in searching someone who is unarmed and innocent, but there is a heavy cost for not searching someone who is armed and dangerous. To help understand this further, we need to discuss Type 1 and Type 2 errors.

What bothers people most – convicting an innocent man or acquitting a guilty man? Both are detestable, but our system is predicated on the former being worse. When you reject the null and the null is true, you have committed a Type 1 error; the probability of such an error is the significance level which you choose for the hypothesis test (note that the judge doesn't give you the measure of reasonable doubt, as it is hard to put an exact number on it). So if the significance level is at 2%, then should the defendant be innocent, there is a 2% probability of his being convicted by mistake (since the significance level = the probability of a Type 1 error). When you fail to reject the null and it is false, you have committed a Type 2 error (this

is the equivalent of a guilty defendant being acquitted); the probability of such an error is unknown and cannot be computed, but we do know that as the Type 1 error increases, the Type 2 error decreases by some unknown amount, and vice versa. Thus, you can indirectly raise or lower the probability of a Type 2 error by decreasing or increasing the significance level, respectively.

So let's see if we can fix this judicial system to the satisfaction of all. Since we don't like to see innocent men go to jail, we would like to see the probability of a Type 1 error = 0, so we can set the significance level to 0, BUT that would mean everyone must be acquitted, as it is the only way to guarantee that no innocent person is ever convicted. As a result, the probability of a Type 2 error grows to 100% because all guilty people will go free. And what about the problem of acquitting the guilty – for this we need to make sure the probability of a Type 2 error = 0 so that never happens, but the only way that is possible is to convict everyone, which means setting the significance level or the probability of a Type 1 error = 100%. So we essentially stop convicting innocent people by setting the significance level to 0% and we stop acquitting the guilty by setting the significance level to 100% -- problem solved! Obviously not. It is impossible to eliminate both errors, and to eliminate one of the two means the other occurs 100% of the time. Thus, the best solution is to allow the possibility of both errors (which is what our court system currently does). We may not like it, but it is truly the best solution. We just decided as a nation that convicting the innocent is worse, so we set our level of reasonable doubt as a low one, and while it means a greater chance of a guilty person going free, it means the prosecutor needs to be better prepared with evidence before going to trial.

You do have some input, however, in that you can raise or lower the significance level a bit, depending on which of the two errors is more serious. In the court system, the Type 1 error is more serious, so you might set the significance level to .01, making it only a 1% chance of convicting an innocent person. If you were in charge of airport security and the null hypothesis is that the passenger is unarmed, a Type 1 error would mean having to frisk a person in error (because the alarm sounded) and a Type 2 error would mean letting an armed man walk through. Of course a Type 2 error is worse here, so the best thing to do is raise the significance level to maybe .10 (10% probability that an unarmed person is searched), thus making it very unlikely that an armed person gets by. Likewise on a medical test, it is far better that a person tests positive by mistake and gets further testing than tests negative and goes home, only to potentially die unnecessarily; here the Type 2 error is worse so you would likely increase the significance level. If neither error is truly a big deal (like if you were testing the preference for peanut butter), the rule of thumb is to just leave the significance level at 5%.

What determines if you reject or don't reject the null? If you did the test using formulas and paper, you would rely on the test statistic and tables, but in this high tech world, you will find that all statistical software produces a p-value in the output for a hypothesis test. A p-value is merely the probability of making a type one error if you decide to reject the null. Going back to the judicial example, it is the probability that the defendant is truly innocent should you decide to convict him. While the significance level is your tolerance for being wrong, the p-value is the likelihood you actually are wrong, the risk you would be taking. When the p-value is smaller than the significance level, then the risk of error is less than your tolerance for error and so you reject the null hypothesis; when the p-value is larger than the significance level, the risk is too great for your tolerance and so you don't reject the null hypothesis.

The output you see from statistics software typically includes a p-value (also shown as a sig. value on some). This p-value is the risk of error from the hypothesis test, while the significance level that you set

Chapter Four: Hypothesis Testing for One Sample

is the tolerance for error. If the p-value is less than the significance level, then your risk for error is within your tolerance level and you reject the null; in the judicial example it means you feel very confident of your decision to convict. If the p-value is greater than the significance level, then your risk is too great and you fail to reject the null (thus acquitting the defendant due to lack of evidence). The p-value of .009, for example, means that if you convict the defendant, there is a 0.9% chance you are wrong and he is innocent, but as this is very small, you go ahead and convict; a p-value of .25 means that if you convict the defendant, there is a 25% chance you are wrong and he is innocent, but since this is very large, you choose not to convict. The cutoff is determined by the significance level you set (typically 5% in most studies). Thus, a low p-value is typically the goal in research so you can actually prove something with evidence.

Hypothesis tests can be done on virtually any population measurement, but the most common tests are about means and proportions. A hypothesis test is essentially the same as comparing a confidence interval against a theorized value. Every election poll shows a percentage of people responding a certain way, and knowing that it takes 50.1% to win the election, if the sample proportion +/- the margin of error is completely above 50%, then you are X% confident that person will win (and if it is completely below 50%, then you are X% confident that person will lose). Hypothesis tests for proportions work the same way; the null hypothesis might be that the population proportion = .50, and with the sample proportion computed, you would use the test results to determine if the sample proportion is significantly different from .50. Significantly different really means that the margin of error is smaller than the distance to .50. If it is significantly different, you reject the null hypothesis and conclude the alternate hypothesis is true. If it is not significantly different, you fail to reject the null hypothesis and state that there is insufficient evidence to prove the alternate is true. If the Bush-Kerry poll showed Bush with 52% of the vote in Florida, you might fail to reject the null – this doesn't mean they are tied, but it does mean the results are still inconclusive (the media calls this a swing state).

Figure 1: Screen display from **Hypothesis Tests One Sample.xls**

One Sample Hypothesis Test for the Mean				
Null Hypothesis	$\mu =$	500		
Level of Significance		0.05		
Sample Size		30		
Sample Mean		535		
Standard Deviation		100		
Test Statistic (Computed)		1.92		
Direction of Test	Lower Crit Value	Upper Crit Value	p-Value	Decision
Two-Tailed Test HA: $\mu \neq 500$	-1.9600	1.9600	0.0552	Do not reject the null hypothesis
Upper-Tail Test HA: $\mu > 500$	n/a	1.6449	0.0276	Reject the null hypothesis
Lower-Tail Test HA: $\mu < 500$	-1.6449	n/a	0.9724	Do not reject the null hypothesis
If you REJECT the null hypothesis, conclude that HA is true.				
If you DO NOT REJECT the null hypothesis, there is insufficient evidence to conclude that HA is true.				
NEVER conclude that the null hypothesis is true (i.e., we CANNOT ACCEPT the null).				

In the example shown here, we are testing if a sample of 30 students at Whatsamatta U. has a mean GMAT score that differs from the national average of 500. The null hypothesis is that the mean = 500. The

Chapter Four: Hypothesis Testing for One Sample

sample mean is 535, and the known standard deviation is 100. There are 3 options of p-values to use, based on the direction of the test. The hypotheses can be two-tailed (non-directional), upper-tailed or lower-tailed. A two-tailed test is one in which the null hypothesis has an equals sign and the alternate has a “not equals” sign; you are testing if the sample mean *is different from* the hypothesized value. An upper-tailed test is one in which the alternate hypothesis tests if the sample mean *is greater than* the hypothesized value, and a lower-tailed test is one in which the alternate hypothesis tests if the sample mean *is less than* the hypothesized value. By conducting an upper-tailed or lower-tailed test (also called directional tests), the p-value is cut in half and it is easier to reject the null hypothesis. While this sounds great, it is also dangerous and irresponsible to set up the test as directional unless you have good reason to do so. For example, if you are testing a diet pill, it is reasonable to test if the average weight has declined. It would also be understandable to test if the mean SAT score at a top-tier school is above the national average. If you wanted to study the grade performance of students who drink Pepsi, it would not be reasonable to test if grades are higher or lower than a norm. The best decision is to use a non-directional approach unless you have justification to pick a direction. For this particular problem, though, we are testing if the GMAT is different from 500, so we would use a two-tailed test, and the resulting p-value is .0552, which is greater than the significance level of .05; thus the null hypothesis is not rejected and there is insufficient evidence to conclude that the mean GMAT at Whatsamatta U. differs from the national mean of 500.

Suppose instead that we had been testing if Whatsamatta U. has standards which exceed the national average, and so we wanted to test if the mean was greater than 500. In that case we would have used an upper-tailed test, and with the resulting p-value of .0276 which is less than .05, we would have rejected the null hypothesis and concluded that the mean GMAT was greater than the national average of 500.

Figure 2: Screen display from [Hypothesis Tests One Sample.xls](#)

One Sample Hypothesis Test for the Proportion				
Null Hypothesis	$P =$	0.15		
Level of Significance		0.05		
Number of Successes		9		
Sample Size		100		
Sample Proportion		0.09	(computed from Successes / Sample Size)	
Z Test Statistic (Computed)		-1.68		
Direction of Test	Lower Crit Value	Upper Crit Value	p-Value	Decision
Two-Tailed Test HA: $P \neq 0.15$	-1.9600	1.9600	0.0929	Do not reject the null hypothesis
Upper-Tail Test HA: $P > 0.15$	n/a	1.6449	0.9536	Do not reject the null hypothesis
Lower-Tail Test HA: $P < 0.15$	-1.6449	n/a	0.0464	Reject the null hypothesis
If you REJECT the null hypothesis, conclude that HA is true.				
If you DO NOT REJECT the null hypothesis, there is insufficient evidence to conclude that HA is true.				
NEVER conclude that the null hypothesis is true (i.e., we CANNOT ACCEPT the null).				

Now let's look at how proportions are testing. Suppose a baseball player made errors 15% of the time last year, so he went to a training camp to improve his defense. This season he had 100 opportunities to make a play and committed 9 errors. Test the hypothesis that the error rate is now less than 15%. The

null hypothesis entered here would be that $P = .15$. The sample size is 100 and the number of successes is 9 (note that a success is not a good thing or bad thing – it is just whatever is being measured). Leaving the significance level at .05, we choose to do a lower-tail test to determine if $P < .15$. The p-value is .0464 which is less than .05, and so we reject the null and conclude that the proportion of errors has decreased from last season.

Z-tests vs. t-tests

Hypothesis tests for means can use one of two approaches – a z-test or a t-test. To use a z-test, you must know the population standard deviation (which can be obtained from historical data or prior analysis). You could also use the z-test when you don't know the population standard deviation provided that your sample is large enough (at least 30); the rationale here is that with a large sample, the sample standard deviation tends to be similar to the population standard deviation. If you don't know the population standard deviation and your sample is not large enough, fear not because you can use a t-test. The t-test helps to deal with the bias from the sample standard deviation. While there are slightly different formulas and different tables in the textbook, those using Excel will find the only difference between the two tests is negligible.

Closing Notes

Remember that the aim of a study is to prove the alternate hypothesis (often called the research hypothesis). Researchers spend a lot of time and money trying to prove the alternate is true, as there is no value in failing to reject the null and being where they began. Prosecutors don't want to go to trial unless they can convict, or else it is a waste of time and money. Since the object is to reject the null and conclude that the alternate is true, researchers do what they can to help their odds. A directional test will cut the p-value and help the chances of rejection, but it must be appropriate to do so. Raising the level of significance gives a wider berth for rejecting the null, but can you really tolerate that much error. Lastly, increasing the sample size will make a small difference more significant (it is always acceptable to take an additional sample if your p-value is close to the rejection mark, but it may not be convenient to do).

If a governor had 60% support in a poll, it may not be statistically significant (e.g., 3 out of 5 surveyed support him), but if he had 51%, it may be statistically significant (e.g., 51,000 out of 100,000 support him). The percents can be misleading; the sample size plays a big factor in determining if the results are truly significant. So don't jump to conclusions based on a percent sounding significant or sounding close.

Hypothesis testing is truly a shift in the way you think, but it makes sense. Once you learn to set up the hypotheses properly, things get easier. Of course, the results are completely meaningless if the data is not valid or reliable. Don't cut corners just to get the results you want. Would you want to try drugs that were improperly tested by an anxious researcher who just wanted to get published? Even if a researcher proves a drug to cure what ails you, a small percentage of studies will incur an error of that type, which is why reputable pharmaceutical companies will re-test, as the aim is to be sure. Never let ethics take a backseat to desired results. Hypothesis testing should open your eyes to how election polling is done, how medical tests are conducted, how psychological studies are done, etc., and should give you a greater appreciation for the results.

I suggest you also read Chapter 24 of <http://stat-www.berkeley.edu/users/stark/SticiGui/Text/> for an excellent write-up on hypothesis testing.

CHAPTER FOUR KNOWLEDGE ASSESSMENT

Hypothesis Testing for One Sample

Discussion Questions

DISCUSSION QUESTION 1

TRUSTING THE RESULTS: If you were to read the results of a study showing that daily use of a certain exercise machine resulted in an average 10-pound weight loss, what more would you want to know about the numbers in addition to the average? (Hint: Do you think everyone who used the machine lost 10 pounds?)

DISCUSSION QUESTION 2

WHAT'S IN A SAMPLE?: Why is it important that a sample be random and representative when conducting hypothesis testing?

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:

Assume that in 2000, the average sales price of a home was \$205,000. Has this increased in seven years? Using the sample of 100 homes, conduct a one-sample hypothesis test to determine if the mean sales price of a home is greater than \$205,000. Use a .05 significance level.

PRACTICE PROBLEM 2:

Assume you read in an advertisement that the average home in a community is 5 years. Is this really true? You took a sample of 100 homes (in the data file). Conduct a one-sample hypothesis test to determine if the mean age is different from 5 years. Use a .05 significance level.

PRACTICE PROBLEM 3:

You have heard from many folks in the real estate business that fewer than half of the homes are made of brick, and you decide to put data to the test. You took a sample of 100 homes (in the data file). Conduct a one-sample hypothesis test to determine if the proportion of homes made of brick is less than 50%. Use a .05 significance level.

CHAPTER FOUR KNOWLEDGE ASSESSMENT

Hypothesis Testing for One Sample

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:

It is pretty common across most schools to find the grades at the MBA level divided between A's and B's. As such, you expect the mean GPA to be around 3.50. Using the sample of 200 MBA students, conduct a one-sample hypothesis test to determine if the mean GPA is different from 3.50. Use a .05 significance level.

ASSIGNED PROBLEM 2:

Assume you read in the Whatsamatta U website that the average age of their MBA students is 45. Is this really true or have they failed to update this correctly? You think it is far less because there have been a lot more students going straight from their Bachelors to their Masters since the economy is so bad. You took a sample of 200 students (in the data file). Conduct a one-sample hypothesis test to determine if the mean age is less than 45. Use a .05 significance level.

ASSIGNED PROBLEM 3:

You have heard from idle chatter that most students don't declare a major in their MBA programs. You took a sample of 200 students (in the data file). Conduct a one-sample hypothesis test to determine if the proportion without a major is greater than 50%. Use a .05 significance level.

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

4
0
5
1
T
S