STAT 109 Lecture Notes

7 Hypothesis Testing

7.1 Introduction to Hypothesis Testing

Definition 1: A hypothesis is a statement that is assumed to be true, but can be experimentally tested to verify its accuracy. (This should sound familiar to the contexts in which you've seen this word before - probably middle school and high school science classes.) In particular, the statement that we assume to be true is called the null hypothesis. It is denoted by H_0 (pronounced "H-sub-zero" or "H-naught") and it specifies a population model parameter of interest and proposes a value for that parameter. With this we identify (1) the parameter we are interested in and (2) the specific hypothesized value for that parameter (what we *think* is true about it, possibly based on previous experimental knowledge).

$$H_0: \textit{ parameter : hypothesized value;}$$

When we have a particular outcome in mind that we'd like to experimentally test the accuracy of, we may define an alternate hypothesis, denoted H_a (pronounced "H-sub-A"), which contains the values of the parameter that we consider plausible if we reject the null hypothesis. This is going to be a range of values.

Example 7.1: The Department of Motor Vehicles (DMV) in a large city claimed that 80% of candidates pass driving tests, but a reporter's survey of 90 randomly selected local teens found only 61 who passed. Does this finding suggest that the passing rate for teenagers is lower than what the DMV reported?

"I'll assume the passing rate for teenagers is the same as the DMV's overall rate of 80%, unless there's strong evidence that it's lower."

Note : We never "accept" the null hypothesis, we only "reject" or "not reject" a null hypothesis.

Types of Hypotheses

For population proportions: $\begin{array}{lll} H_0: p \geq p_0 & H_0: p \leq p_0 & H_0: p = p_0 \\ H_a: p < p_0 & H_a: p > p_0 & H_a: p \neq p_0 \end{array}$

For population means: $\begin{array}{lll} H_0: \mu \geq \mu_0 & H_0: \mu \leq \mu_0 & H_0: \mu = \mu_0 \\ H_a: \mu < \mu_0 & H_a: \mu > \mu_0 & H_a: \mu \neq \mu_0 \end{array}$

Alternative hypotheses come in three different flavors, and so hypothesis testing comes in three different flavors:

1. **One-tailed, right-tailed test:** Guess that the parameter is bigger than the null hypothesis
2. **One-tailed, left-tailed test:** Guess that the parameter is smaller than the null hypothesis
3. **Two-tailed test:** Guess that the parameter is different from (not equal to) the null hypothesis [and we don't care if it's bigger or smaller, just that it's different]

Example 7.2: Identify the null hypothesis, alternative hypothesis, and the type of hypothesis test that should be used for each experiment described below.

1. the DMV example above.

2. A study by the National Association of Realtors found that the average age of a person buying a second home as a vacation home was 44 years. A real estate agent wants to estimate the average age of vacation home buyers in her area. She randomly selects 20 such buyers and their mean age is 49.2 years old. Is the mean in her area higher than in the general population?

3. A 1996 report claimed that at least 90% of all American homes have at least one smoke detector. A city's fire department has been running a public service campaign about smoke detectors and wonders if this effort raised the local level above the 90% national rate.

4. There are supposed to be 20% orange M&M's in a bag. Suppose a bag of 122 has only 21 orange ones. Does this contradict the company's 20% claim?

5. In 2001 a national statistic reported that about 3% of all births produced twins. Is the rate of twin births the same among young mothers?

6. Long-term research suggests that the mean pH of rainfall in the Shenandoah Mountains is 4.9. Researchers plan to choose a random sample of storms from the past decade because they are concerned that the rain has recently become more acidic in the area.

P-values

Definition 2: The <u>P-value</u> is the probability that the observed statistical value, $\hat{p}$ or $\overline{x}$, (or an even more extreme value) occur if the null hypothesis model is correct.

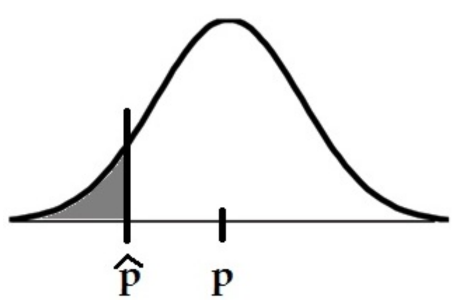
We will use the sampling distribution to determine these probabilities. You may remember from Chapter 5 that when the sample size is large, the sampling distribution is always normal. For smaller sample sizes, we can use the following conditions:

- If we are considering the proportion related to a qualitative variable, the sampling distribution is normal as long as $np \geq 5$ and $nq \geq 5$.
- If we are considering the mean related to a quantitative variable x, the sampling distribution is normal if either the underlying distribution of x is normal, or if $n \geq 30$.

Before you try to compute the P-value using the normal model, you should consider the above conditions to make sure it is actually normal!

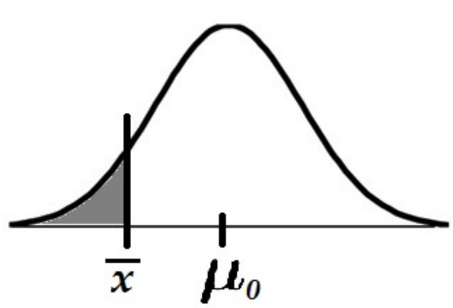
One-sided tails: when H_a is a single inequality. Suppose $H_0 : p \geq p_0$ and $H_A : p < p_0$. Because our alternate hypothesis is a less-than inequality, we are testing whether the true proportion is smaller than the null hypothesis value. Our P-value will be the probability that a value less than or equal to $\hat{p}$ occurs, as in the picture below. This means we will be computing (in the calculator)

$$P = \text{normalcdf}\left(-10 \wedge 99, \hat{p}, p_0, \sqrt{\frac{p_0 q_0}{n}}\right).$$



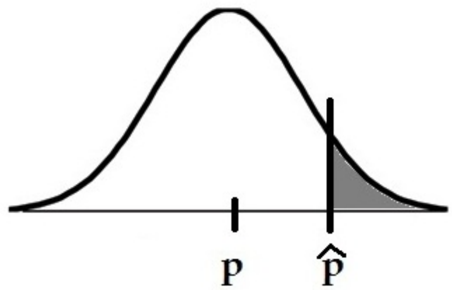
For quantitative variables:

$$P = \text{normalcdf}\left(-10 \wedge 99, \overline{x}, \mu_0, \frac{\sigma}{\sqrt{n}}\right).$$

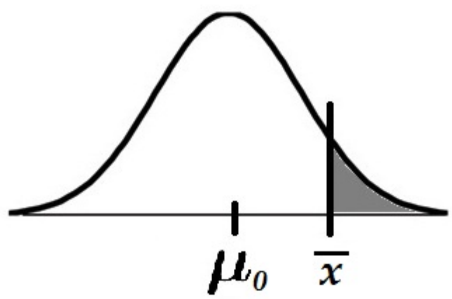


Now suppose $H_0 : p \leq p_0$ and $H_A : p > p_0$. Because our alternate hypothesis is a greather-than inequality, we are testing whether the true proportion is larger than the null hypothesis value. Our P-value will be the probability that a value greather than or equal to $\hat{p}$ occurs, as in the picture below. This means we will be computing (in the calculator)

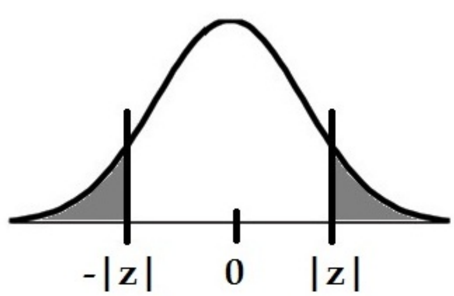
$$P = \text{normalcdf}\left(\hat{p}, 10 \wedge 99, p_0, \sqrt{\frac{p_0 q_0}{n}}\right).$$



$$P = \text{normalcdf}\left(\overline{x}, 10 \wedge 99, \mu_0, \frac{\sigma}{\sqrt{n}}\right).$$



Two-sided tails: Suppose $H_0 : p = p_0$ and $H_A : p \neq p_0$. Because our alternate hypothesis is a not-equal inequality, we are testing whether the true proportion is far from the null hypothesis value *in either direction*. Our P-value will be the probability that a value *as extreme as* $\hat{p}$ occurs (as far as $\hat{p}$ is from p **in either direction**), as in the picture below. Because of the symmetry of the normal distribution, you can compute a one-sided tail (as in the paragraphs above) and then multiply by two to get the P-value.



Type I and Type II Errors

Either the null hypothesis is true or it isn't (in reality). Either our test leads us to reject the null hypothesis or it doesn't. When our test matches reality, we are happy. When it doesn't, we've made an error. These errors are called Type I (false negative) and Type II (false positive).

		Reality	
		H_0 True	H_0 False
Test	Reject H_0	Type I Error	Yay!
Conclusion	Fail to reject H_0	Yay!	Type II Error

Let α be the probability we make a Type I error and β be the probability we make a Type II error. As α goes up, β goes down, and vice versa. Often α is called the level of significance. We will control α. If in the context of our testing, a Type I error would have severe consequences, we want α to be as small as possible (usually $\alpha = 0.01$). If the consequences are not that severe, we'll choose slightly higher α (like $\alpha = 0.05$ or $\alpha = 0.10$).

Draw a Conclusion

Example 7.3: Continuing with the DMV example, what can the reporter conclude and how might they explain the P-value for their story?

"Because the P-value of 0.002 is very low, I reject the null hypothesis. The survey data provides strong evidence that the passing rate for teenagers taking the driving test is lower than 80%.

If the passing rate for teenage driving candidates were actually 80%, we would expect to see success rates this low in only about 1 in 500 samples (.2%). This seems quite unlikely, casting doubts that the DMV's stated success rates applies to teens."

The conclusion in a hypothesis test is always a statement about the null hypothesis. This statement will state either that we reject of that we fail to reject the null hypothesis. This should always be stated in context.

- If $P < \alpha$, we reject the null hypothesis because there is strong evidence that the alternate hypothesis holds.
- If $P \geq \alpha$, we fail to reject the null hypothesis because we have found no strong evidence that the alternate hypothesis holds.

Example 7.4: For each of the tests described below, state the hypotheses, then use the given P value and α level to state a conclusion.

1. A study by the National Association of Realtors found that the average age of a person buying a second home as a vacation home was 44 years. A real estate agent wants to estimate the average age of vacation home buyers in her area. She randomly selects 20 such buyers and their mean age is 49.2 years old. Is the mean in her area higher than in the general population? Assuming that the population standard deviation is 9.4 years, the P value is 0.00668. Use $\alpha = 0.01$.

2. The Association of American Medical Colleges reports that 46% of students who apply to medical school were admitted. One small college boasts that they prepare their students particularly well, as 568 students from last year's graduating class applied to medical school and 291 were admitted. Do the have a right ot brag? You conduct an appropriate hypothesis test with $\alpha = 0.05$ and find $P = 0.00617$. What do you conclude?

3. Long-term research suggests that the pH of rainfall in the Shenandoah Mountains is approximately normal with mean 4.9 and standard deviation 0.6. Researchers plan to choose a random sample of storms from the past decade because they are concerned that the rain has recently become more acidic in the area. The mean pH of the 30 storms selected is 2.8. The P value of the test is 3.277×10^{-82}. Use $\alpha = 0.05$.

4. Google first became a publicly traded company in August 2004. In 2005, the mean daily volume of the stock was $\mu = 10.7$ million shares with a standard deviation of $\sigma = 5.7$ million shares. A random sample of 35 trading days in 2008 had a mean of 5.97 million shares traded per day. The distribution of the variable "daily shares traded" is skewed right with outliers to the right. Is the sampling distribution normal? If so, with what mean and standard deviation? Perform a hypothesis test to determine if the trade volume was different in 2008 than it was in 2005 with $\alpha = 0.05$ and $P = 0.000000914$.

Hypothesis Testing Summary

1. Identify the question you are trying to answer. This is the alternative hypothesis. Write down H_a explicitly.

   Examples: "Has the pass rate increased since 1995?" "Has the average time to complete a college degree changed over the past 50 years?" "Has the portion of the population who oppose marriage equality decreased?"

2. Identify a previously known value associated with your question. This may be based on "old" data (when looking at changes over time) or just a different survey/experiment (if you're replicating a study to see if the original results are valid). This is the null hypothesis. Write down H_0 explicitly.

   Remember: We conduct the hypothesis test under the assumption that H_0 is true. In other words, we assume the null hypothesis is true and the answer to our alternate hypothesis question is "No, it's the same." Then we will conduct our test to see if we have evidence that the null hypothesis is false.

3. Choose an α-level. (Usually this is given in the problem, but in the real world, you would choose it for your own study.) Common values: $\alpha = 0.05, \alpha = 0.01, \alpha = 0.10$

4. Conduct the experiment/survey/study. Compute the test statistic ($\widehat{p}$ or $\overline{x}$ depending on whether the variable of interest is qualitative or quantitative). (Again, in this class, this information will be given in the problem.)

5. Compute the P-value. (We'll talk about this in more detail in the next section.)

6. Write down the conclusion for your test. Your conclusion is either "reject H_0" or "fail to reject H_0."

   Remember: You **reject** H_0 if you find evidence that H_0 is false. That is, if your test statistic has a low probability (relative to α) of occuring under the null hypothesis. This means $P < \alpha$. You **fail to reject** H_0 if you do not find any strong evidence that H_0 is false, ie, if $P \geq \alpha$.

7.2 and 7.4: Computing P-values for Hypothesis Tests

Important facts from Section 7.1: We are interested in the probability that our sample statistic occurs *under the assumption that H_0 is true.* We call this probability P.

H_0 gives us a value for μ (if our variable is quantitative) or p (if our variable is qualitative). Since we are talking about the probability that a **sample statistic** takes on a sample value, we know that information comes from the **sampling distribution**. Furthermore, we know from Section 5.4 what the sampling distribution is.

- If x is quantitative, if we know the population standard deviation σ, and if our null hypothesis is $H_0 : \mu = \mu_0$, then the sampling distribution is $N(\mu_0, \sigma/\sqrt{n})$.
- If x is qualitative and $H_0 : p = p_0$, then the sampling distribution is $N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$.

Once we identify H_0 and H_a, we know if we are performing a one-tailed test (left or right) or a two-tailed test. Use the normal model to find the associated P-value. (See the figures on pages 3-4 of this handout.)

Example 7.5: A researcher claims that more than 30% of US smartphone owners user their phone while watching TV. In a random sample of 150 adults, 38% said they used their phone while watching TV. Is there enough evidence to support the researcher's claim at $\alpha = 0.05$?

Example 7.6: Suppose the average time to complete a pit stop in a professional car race is 13 seconds, with a standard deviation of 0.19 seconds. One particular crew claims they can do it faster (in less time). A random sample of 32 pit stops by this crew had an average time of 12.9 seconds. Is there enough evidence to support the crew's claim at $\alpha = 0.01$?

Example 7.7: Employees at a construction company claim that the average salary for the company's mechanical engineers is less than the average salary for engineers at competing firms. The competition pays an average salary of \$68,000. A random sample of 20 mechanical engineers at this particular company had an average salary of \$66,900. Assume the population is normally distributed with a standard deviation of \$5500. Test the employee's claim at $\alpha = 0.05$.

Example 7.8: During the '90s about 63% of high school graduates enrolled in college. The pollsters hope to estimate the percentage of this year's seniors planning to attend college with a margin of error no greater than 4%. What size sample would suffice if they to have 90% confidence in the estimate?

The pollsters randomly select five cities in upstate New York and then randomly select one high school in each city. The guidance office at each of the chosen school is instructed to ask 100 randomly selected seniors what their current plans are, and to report the results back to the pollsters. The data collected from the five schools are summarized in the table.

Plans	Count
College	289
Employment	112
Military	26
Other (travel, parenting, etc.)	51
Undecided/No response	22

Construct a 90% confidence interval for the percentage of seniors planning to go to college this year. Explain carefully what your interval means.

During the '90s about 4.5% of high school seniors enlisted in the military. Do these data suggest that the percentage who enlist is different this year? Use $\alpha = 0.05$ to test an appropriate hypothesis.