Lecture 23

Nancy Pfenning Stats 1000

Chapter 11: Testing Hypotheses About Proportions

Recall: last time we presented the following examples:

1. In a group of 371 Pitt students, 42 were left-handed. Is this significantly lower than the proportion of all Americans who are left-handed, which is .12?

2. In a group of 371 students, 45 chose the number seven when picking a number between one and twenty "at random". Does this provide convincing statistical evidence of bias in favor of the number seven, in that the proportion of students picking seven is significantly higher than $1/20 = .05$?

3. A university has found over the years that out of all the students who are offered admission, the proportion who accept is .70. After a new director of admissions is hired, the university wants to check if the proportion of students accepting has changed significantly. Suppose they offer admission to 1200 students and 888 accept. Is this evidence of a change from the status quo?

Each example mentions a possible value for p, which would indicate no difference/no change/status quo. The **null hypothesis** H_0 states that p equals this "traditional" value.

In contrast to the null hypothesis, each example suggests that an alternative may be true: a significance test problem always pits an **alternative hypothesis** H_a against H_0. H_a proposes that the proportion differs from the "traditional" value p_0—H_a rocks the boat/upsets the apple cart/ marches to a different drummer. A key difference among our three examples is the direction in which H_a refutes H_0. In the first, it is suggested that the proportion of all Pitt students who are left-handed is *less than* the proportion for adults in the U.S., which is .12. In the second, we wonder if the proportion of students picking the number seven is significantly *more than* .05. In the third, we inquire about a difference *in either direction* from the stated proportion of .70. We can list our null and alternative hypotheses as follows:

1. $H_0 : p = .12 \qquad H_a : p < .12$
2. $H_0 : p = .05 \qquad H_a : p > .05$
3. $H_0 : p = .70 \qquad H_a : p \neq .70$

In general, we have $H_0 : p = p_0$ vs. $H_a : p \left\{ \begin{array}{c} < \\ > \\ \neq \end{array} \right\} p_0$

Note that your textbook may have expressed the first two null hypotheses as $H_0 : p \geq .12$ and $H_0 : p \leq .05$. These expressions serve well as logical opposites to the alternative hypotheses, but our strategy to carry out a test will be to assume H_0 is true, which means we must commit to a single value p_0 at which to center the hypothesized distribution of $\hat{p}$. Thus, we will write $H_0 : p = p_0$ in these notes.

Alternatives with $<$ or $>$ signs are called *one-sided alternatives*; with $\neq$ they are *two-sided*. When in doubt, a two-sided alternative should be used, because it is more general.

Note: In statistical inference, we draw conclusions about unknown *parameters*. Thus, H_0 and H_a are statements about a parameter (p), not a statistic ($\hat{p}$). We can't argue about $\hat{p}$; its value has been measured and taken as fact.

Note: Just as "success" and "failure" in binomial settings lost their connotations of favorable and unfavorable, H_a may or may not be a desired outcome. It can be something we hope or fear or simply suspect is true. However, because p_0 is a traditionally accepted value, we'll stick with H_0 unless there is convincing evidence to the contrary: H_0 is "innocent until proven guilty".

How can we produce evidence to refute H_0? By using what probabilty theory tells us about the behavior of the R.V. sample proportion $\hat{p}$: it is centered at p, has spread $\sqrt{\frac{p(1-p)}{n}}$, and for large enough n its shape is normal.

Our strategy will be to determine if the observed value $\hat{p}$ is just too unlikely to have occurred if $H_0 : p = p_0$ were true. If the probability of such an outcome (called the P-value) is too small, then we'll reject H_0 in favor of H_a.

The **P-value** of a test about a proportion p is the probability, computed assuming that $H_0 : p = p_0$ is true, that the test statistic ($\hat{p}$) would take a value at least as extreme—that is, as low or as high or as different—as the one observed. The smaller the P-value, the stronger the evidence against H_0.

Since Example 1 has $H_a : p < .12$, the P-value is the probability of a sample proportion of left-handers *as low* as $\frac{42}{371} = .113$ or lower, coming from a population where the proportion of left-handers is .12. Based on what we learned about the sampling distribution of $\hat{p}$, we know that $\hat{p}$ here, assuming H_0 is true, has mean $p = .12$, standard error $\sqrt{\frac{.12(.88)}{371}}$, and an approximately normal shape since $371(.12) \approx 45$ and $371(.88) \approx 326$ are both greater than 10. [Also, we have in mind a much larger population of Pitt students, certainly more than $10(371) = 3710$.]

$$\text{P-value} = P(\hat{p} \leq .113) \approx P(Z \leq \frac{.113 - .12}{\sqrt{\frac{.12(.88)}{371}}}) = P(Z \leq -.41) = .3409$$

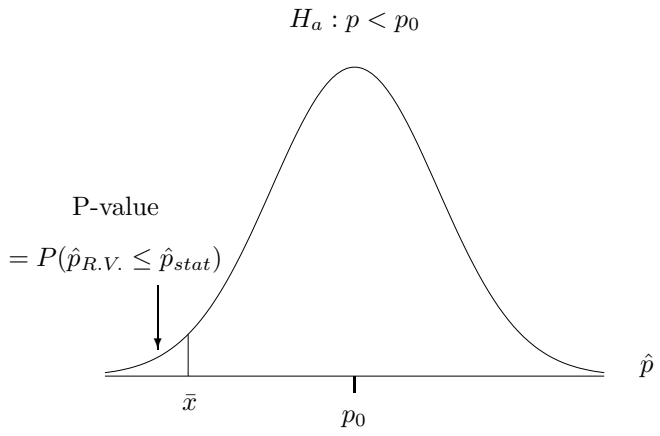
Note: For confidence intervals, since $\hat{p}$ has standard deviation $\sqrt{\frac{p(1-p)}{n}}$ with p unknown, we estimated it with $s.e.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Now, we carry out our test assuming $H_0 : p = p_0$ is true, so the standard deviation of $\hat{p}$ is $\sqrt{\frac{p_0(1-p_0)}{n}}$ and the test statistic is $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

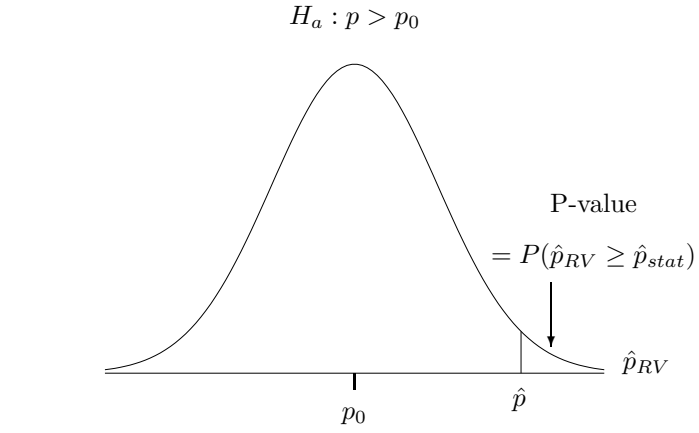
Since it's not at all unlikely (probability about 34%) for a random sample of 371 from a population with proportion .12 of left-handers to have a sample proportion of only .113 left-handers, we have no cause to reject $H_0 : p = .12$. The proportion of left-handers at Pitt may well be the same as for the whole country, .12.

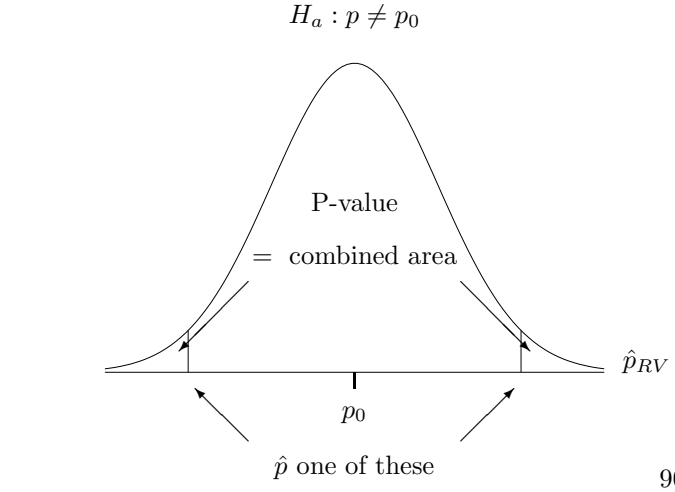
Because we rely on standard normal tables to determine the P-value, we transform from an observed value $\hat{p}$ to a standardized value $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. The way to compute the P-value depends on the form of H_a, as illustrated below, first in terms of $\hat{p}$, then in terms of z.

P-value for Tests of Significance about p

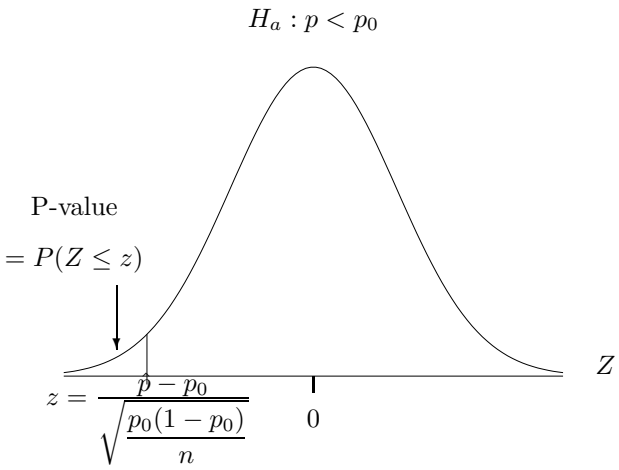
Observed

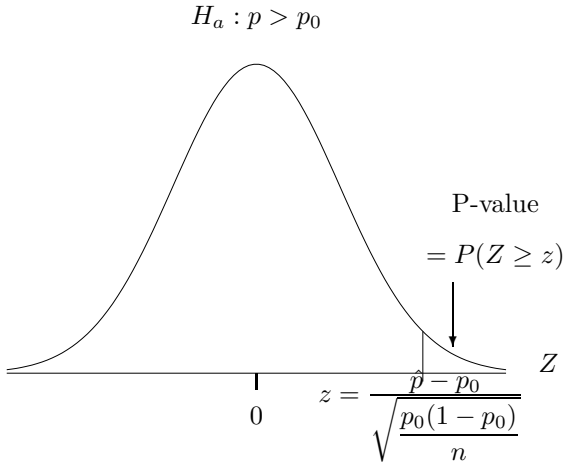


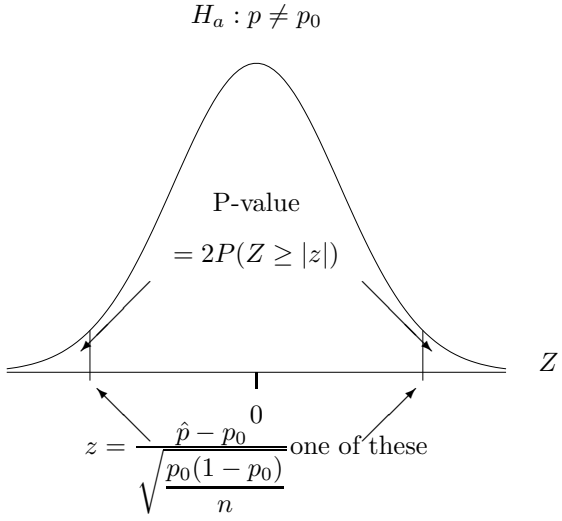




Standardized







Summary of Test of Significance about p

Say a simple random sample of size n is drawn from a large population with unknown proportion p of successes. We measure $\hat{p} = \frac{X}{n}$ and carry out the test as follows:

1. Set up $H_0 : p = p_0$ vs. $H_a : p \left\{ \begin{array}{c} < \\ > \\ \neq \end{array} \right\} p_0$
2. Verify that the population is at least 10 times the sample size, and that $np_0 \geq 10$ and $n(1-p_0) \geq 10$. Then calculate standardized test statistic $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
3. Find P-value $= P(Z_{R.V.} \leq z_{\text{statistic}})$ for $H_a : p < p_0$
   $= P(Z_{R.V.} \geq z_{\text{statistic}})$ for $H_a : p > p_0$
   $= 2P(Z_{R.V.} \geq |z_{\text{statistic}}|)$ for $H_a : p \neq p_0$
4. Determine if the results are statistically significant: if the P-value is "small", reject H_0 in favor of H_a, and say the data are "statistically significant"; otherwise, we have failed to produce convincing evidence against H_0. [For specified α, reject H_0 if P-value $< \alpha$.]
5. State conclusion in context of the particular problem.

Example

Let's follow these steps to solve the second problem. In a group of 371 students, 45 chose the number seven when picking a number between one and twenty "at random". Does this provide convincing statistical evidence of bias in favor of the number seven, in that the proportion of students picking seven is significantly higher than $1/20 = .05$? First calculate $\hat{p} = \frac{45}{371} = .12$.

1. $H_0 : p = .05 \qquad H_a : p > .05$
2. We have in mind a very large population of all students. We check that $371(.05) = 19$ and $371(.95) = 352$ are both greater than 10. Next, $z = \frac{.12-.05}{\sqrt{\frac{.05(.95)}{371}}} = 6.19$
3. P-value $= P(Z \geq 6.19) = P(Z \leq -6.19) \approx 0$
4. Since the P-value is very small, we reject H_0 and say the results are statistically significant.
5. There is very strong evidence of bias in favor of the number seven.

Example

I also suspected bias in favor of the number seventeen. In a group of 371 students, 25 chose the number seventeen when picking a number between one and twenty "at random". Does this provide convincing statistical evidence of bias in favor of the number seventeen, in that the proportion of students picking seventeen is significantly higher than $1/20 = .05$? First calculate $\hat{p} = \frac{25}{371} = .067$.

1. $H_0 : p = .05 \qquad H_a : p > .05$
2. $z = \frac{.067-.05}{\sqrt{\frac{.05(.95)}{371}}} = 1.50$
3. P-value $= P(Z \geq 1.50) = P(Z \leq -1.50) = .0668$
4. We could call this a borderline P-value. Next lecture, we'll discuss guidelines for how small the P-value should be in order to reject H_0, and we'll solve the third example. Often, a cut-off probability α is set in advance, in which case we reject H_0 if the P-value is less than α.

Lecture 24

Testing Hypotheses About Proportions

Last time, we learned the steps to carry out a test of significance:

1. Set up $H_0 : p = p_0$ vs. $H_a : p \left\{ \begin{array}{c} < \\ > \\ \neq \end{array} \right\} p_0$

2. In order to verify that the underlying distribution is approximately binomial, check that the population is at least 10 times the sample size. In order to justify use of a normal approximation to binomial proportion, check that $np_0 \geq 10$ and $n(1-p_0) \geq 10$. Calculate standardized test statistic $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

3. Find P-value $= P(Z_{R.V.} \leq z_{\text{statistic}})$ for $H_a : p < p_0$
   $= P(Z_{R.V.} \geq z_{\text{statistic}})$ for $H_a : p > p_0$
   $= 2P(Z_{R.V.} \geq |z_{\text{statistic}}|)$ for $H_a : p \neq p_0$

4. Assess significance: if the P-value is "small", reject H_0 in favor of H_a, and say the data are "statistically significant"; otherwise, we have failed to produce convincing evidence against H_0. [For specified α, reject H_0 if P-value $< \alpha$.]

5. State conclusions in context.

Last time we began to solve the following example:

Example

When students are asked to pick a number "at random" from one to twenty, I suspect their selections will show bias in favor of the number seventeen. In a group of 371 students, 25 chose the number seventeen. Does this provide convincing statistical evidence of bias in favor of the number seventeen, in that the proportion of students picking seventeen is significantly higher than $1/20 = .05$?

The null and alternative hypotheses were
$H_0 : p = .05$ $H_a : p > .05$ and so the z-statistic was
$z = \frac{.067-.05}{\sqrt{\frac{.05(.95)}{371}}} = 1.50$ and the P-value was $= P(Z \geq 1.50) = P(Z \leq -1.50) = .0668$
Step 4 says to reject H_0 if the P-value is small. How small is small?

Sometimes, it is decided in advance exactly how small the P-value would have to be to lead us to reject the null hypothesis: a cut-off probability α is prescribed in advance. Then, if the P-value is less than α, we reject H_0, and say the results are **statistically significant at level** α. Otherwise, we do not have sufficient evidence to reject H_0.

Example

Is there evidence of bias in favor of the number seventeen at the $\alpha = .05$ level? The P value is .0668, which is *not* less than .05, so by this criterion it is not small enough to reject H_0. It could be that students didn't have any systematic preference for the number seventeen, and the proportion of seventeens selected was a bit high only by chance.

Example

A university has found over the years that out of all the students who are offered admission, the proportion who accept is .70. After a new director of admissions is hired, the university wants to check if the proportion of students accepting has changed significantly. Suppose they offer admission to 1200 students and 888 accept. Is this evidence at the $\alpha = .05$ level that there

has been a real change from the status quo? How about at the .02 level? First we find that $\hat{p} = \frac{888}{1200} = .73$ is the sample proportion of students who accepted admission.

1. Set up $H_0 : p = .70$ vs. $H_a : p \neq .70$
2. Both conditions are satisfied. $z = \frac{.73-.70}{\sqrt{\frac{.7(.3)}{1200}}} = 2.27$
3. Because of the two-sided alternative, our P-value is $= 2P(Z \geq |2.27|) = 2P(Z \leq -2.27) = 2(.0116) = .0232$
4. Since $.0232 < .05$, we have evidence to reject at the 5% level. But .0232 is not less than .02, so we don't have evidence to reject at the 2% level.
5. If we set out to gather evidence of a change in either direction for overall proportion of students accepting admission, we would say yes with a cutoff of .05, no with a cutoff of .02. Thus, this test is rather inconclusive.

Example

A university has found over the years that out of all the students who are offered admission, the proportion who accept is .70. After a new director of admissions is hired, the university wants to check if the proportion of students accepting has *increased* significantly. Suppose they offer admission to 1200 students and 888 accept. Is this evidence at the $\alpha = .05$ level that there has been a significant increase in proportion of students accepting admission? How about at the .02 level? Again we find that $\hat{p} = \frac{888}{1200} = .73$ is the sample proportion of students who accepted admission.

1. The subtle re-phrasing of the question ("increased" instead of "changed") results in a different alternative hypothesis. $H_0 : p = .7$ vs. $H_a : p > .7$
2. The z statistic is unchanged: $z = \frac{.73-.70}{\sqrt{\frac{.7(.3)}{1200}}} = 2.27$
3. Because of the one-sided alternative, our P-value is $= P(Z \geq 2.27) = P(Z \leq -2.27) = (.0116)$
4. Since $.0116 < .05$, we again have evidence to reject at the 5% level. This time, .0116 is also less than .02, so we also have evidence to reject at the 2% level.
5. If we set out to gather evidence of increased overall proportion of students accepting admission, we would say yes, we have produced evidence of an increase, whether the $\alpha = .05$ or $\alpha = .02$ level is used.

The previous examples demonstrate that

1. It is more difficult to reject H_0 for a two-sided alternative than for a one-sided alternative. In general, the two-sided P-value is twice the one-sided P-value. The one-sided P-value is half the two-sided P-value.

2. It is more difficult to reject H_0 for lower levels of α.

Calculating the P-value in Step 3 gives us the maximum amount of information to carry out our test—we know exactly how unlikely the observed $\hat{p}$ is.

If a cut-off level α is prescribed in advance, then it is possible to bypass the calculation of the P-value in Step 3. Instead, the z-statistic is compared to the **critical value** z^* associated with α. For example, if we have a two-sided alternative and α is set at .05, then the **rejection region** would be where the test-statistic z exceeds 1.96 in absolute value. The disadvantage to this method is that it provides only the bare minimum of information needed to decide whether to reject H_0 or not. We will not employ the rejection region method in this course, but students should be aware of it in case they encounter it in other contexts.

A method that falls somewhere in between those which provide maximum and minimum information is the following: "close in on" the P-value by surrounding the z statistic with neighboring values z^* from the

"infinite" row of Table A.2. The advantage to this method is that it familiarizes us with the use of Table A.2, which will be needed when we carry out hypothesis tests about unknown population mean of a quantitative variable. Note that

$z^* = 1.645$ corresponds to an area of .90 symmetric about zero, so each tail probability, that z takes a value less than -1.645 or greater than +1.645, is .05.

$z^* = 1.960$ corresponds to an area of .95 symmetric about zero, so each tail probability, that z takes a value less than -1.960 or greater than +1.960, is .025.

$z^* = 2.326$ corresponds to an area of .98 symmetric about zero, so each tail probability, that z takes a value less than -2.326 or greater than +2.326, is .01.

$z^* = 2.576$ corresponds to an area of .99 symmetric about zero, so each tail probability, that z takes a value less than -2.576 or greater than +2.576, is .005.

These tail probabilities may be penciled in at the top or bottom ends of the columns in Table A.2 for easy reference.

We will now re-solve some of our earlier examples, using Table A.2 instead of Table A.1.

Example

A university has found over the years that out of all the students who are offered admission, the proportion who accept is .70. After a new director of admissions is hired, the university wants to check if the proportion of students accepting has increased significantly. Suppose they offer admission to 1200 students and 888 accept. Is this evidence at the $\alpha = .05$ level that there has been a significant increase in the proportion of students accepting?

First we found that $\hat{p} = \frac{888}{1200} = .73$ is the sample proportion of students who accepted admission and set up $H_0 : p = .70$ vs. $H_a : p \geq .70$. Next we calculated $z = \frac{.73-.70}{\sqrt{\frac{.7(.3)}{1200}}} = 2.27$.

Our P-value is $= P(Z \geq 2.27)$. According to Table A.2, $z = 2.27$ is between $z^* = 1.960$ and $z^* = 2.326$. Therefore, our p-value, $P(Z \geq 2.27)$, is between .025 and .01, which means it must be less than .05. We can reject H_0 at the 5% level. [Recall: Table A.1 showed the precise P-value to be .0116, which is in fact between .025 and .01.]

Example

A university has found over the years that out of all the students who are offered admission, the proportion who accept is .70. After a new director of admissions is hired, the university wants to check if the proportion of students accepting has changed significantly. Suppose they offer admission to 1200 students and 888 accept. Is this evidence at the $\alpha = .05$ level that there has been a real change (in either direction) from the status quo? First we found that $\hat{p} = \frac{888}{1200} = .73$ is the sample proportion of students who accepted admission, and we set up $H_0 : p = .70$ vs. $H_a : p \neq .70$. Next we calculated $z = \frac{.73-.70}{\sqrt{\frac{.7(.3)}{1200}}} = 2.27$

Because of the two-sided alternative, our P-value is $= 2P(Z \geq |2.27|) = 2P(Z \geq +2.27)$. According to Table A.2, $z = 2.27$ is between $z^* = 1.960$ and $z^* = 2.326$. Therefore, $P(Z \geq 2.27)$ is between .025 and .01, and the P-value, $2P(Z \geq 2.27)$, is between 2(.025) and 2(.01), that is, between .05 and .02. We still can reject H_0 at the 5% level, but not at the 2% level.

Example

In a group of 371 Pitt students, 42 were left-handers, which makes the sample proportion .113. Is this significantly lower than the proportion of Americans who are left-handers, which is .12? Earlier we found the z-statistic to be $\frac{.113-.12}{\sqrt{\frac{.12(.88)}{371}}} = -.41$ and the P-value to be $P(Z \leq -.41)$. Consulting Table A.2, we see that -.41 is less extreme than 1.645, so the P-value is larger than .05. Again, we have failed to produce any evidence against H_0.

Example

When students are asked to pick a number "at random" from one to twenty, I suspect their selections will show bias in favor of the number seventeen. In a group of 371 students, 25 chose the number seventeen. Does this provide convincing statistical evidence of bias in favor of the number seventeen, in that the proportion of students picking seventeen is significantly higher than $1/20 = .05$?

The null and alternative hypotheses were

$H_0 : p = .05 \qquad H_a : p > .05$ and so the z-statistic was

$z = \frac{.067-.05}{\sqrt{\frac{.05(.95)}{371}}} = 1.50$ and the P-value was $= P(Z \geq 1.50)$. Instead of using Table A.1 to find the precise P-value, we note from Table A.2 that 1.50 is less than 1.645, so the tail probability must be greater than $\frac{1-.90}{2} = .05$. Thus, our P-value$= P(Z \geq 1.50)$ is greater than .05 and we do not have convincing evidence of bias. Note: Earlier we found the exact P-value to be $P(Z \leq -1.50) = .0668$, which is indeed greater than .05.

Example

Note: In a previous Example, we began by assuming that the proportion of freshmen taking intro Stats classes is .25. According to survey data, we found the sample proportion of freshmen to be .08. By hand we calculated the probability of a sample proportion this low, coming from a population with proportion .25: it was approximately zero. I characterized this as "virtually impossible" and decided not to believe that the overall proportion of freshmen is .25.

Alternatively, I could use MINITAB to test the hypothesis that population proportion is .25, vs. the "less than" alternative. Since "year" allows for more than two possibilities, it is necessary to use Stat, then Tables, then Tally to count the number of freshmen (35). Then use the Summarized Data option in the 1 Proportion procedure, specifying 445 as the Number of Trials and 35 as the Number of Successes. I opted to "use test and interval based on normal distribution", since that's how I originally solved the problem by hand. The p-value is zero, and I reject the null hypothesis in favor of the alternative. I again conclude that the proportion of freshmen in intro Stats classes (at least in the Fall) is less than .25.

```
Tally for Discrete Variables: Year

 Year  Count
    1     35
    2    257
    3    102
    4     37
other     14
   N=    445
   *=      1

Test and CI for One Proportion

Test of p = 0.25 vs p < 0.25

Sample      X      N  Sample p  95.0% Upper Bound  Z-Value  P-Value
1          35    445  0.078652           0.099642    -8.35    0.000
```

Exercise: *In a previous Exercise, we explored the sampling distribution of sample proportion of females, when random samples are taken from a population where the proportion of females is .5. We noted the sample proportion of females among surveyed Stats students, and calculated by hand the probability of observing such a high sample proportion, if population proportion were really only .5. We used this probability to decide*

whether we were willing to believe that population proportion is in fact .5. For this Exercise, address the same question by carrying out a formal hypothesis test using MINITAB. Be sure to specify the appropriate alternative hypothesis. State your conclusions clearly in context.

Lecture 25

Type I and Type II Error

When we set a cutoff level α in advance for a hypothesis test, we are actually specifying the long-run probability we are willing to take of rejecting a true null hypothesis, which is one of the two possible mistaken decisions that can be made in a hypothesis test setting.

Example

Recall our testing-for-disease example in Chapter 7, in which the probability of a false positive was .015, probability of false negative was .003. All the possibilities for Decision and Actuality are shown in the table below. If we decide to use .015 as our cut-off probability (p-value $<$ .015 means reject H_0; otherwise don't reject), then .015 is the probability of making a **Type I Error**—the probability of rejecting the null hypothesis, even though it is true. That means the probability of correctly accepting a true null hypothesis is $1 - .015 = .985$. In medical situtations, this is the **specificity** of the test.

	Actuality	
⇓Decision	Healthy(H_0 true)	Diseased(H_a true)
Healthy (don't reject H_0)	*correct* prob.=**specificity**=.985	*incorrect*(false neg.) **Type II error**(prob.=.003)
Diseased (accept alt.hyp.)	*incorrect*(false pos.) **Type I error**(prob.=.015)	*correct* prob.=**sensitivity** =**power**=.997

In our example, we were told the probability of a false negative, or **Type II Error**. Thus, the probability of a correct positive for an ill person (called the **sensitivity** of the test) = 1 minus the probabilty of Type II error. Statisticians refer to this probability as the **power** of the test.

In a z test about population proportion p, the probability of a Type II error [incorrectly failing to reject the null hypothesis when the alternative is true] can only be calculated if we are told specifically the actual value of the population proportion. Thus, we need to know the *alternative* proportion which contradicts the null hypothesized proportion. What we do *not* need in order to calculate the probability of Type II error is the value of an observed proportion $\hat{p}$. Our probability is about the test itself, not about the results.

Rather than focusing on making such calculations, we will instead think carefully about the implications of making Type I or Type II errors.

Example

For our medical example above, the probability of incorrectly telling a healthy person that he or she does have AIDS is higher than the probability of incorrectly telling an infected person that he or she does not have AIDS. If a healthy person initially tests positive (Type I error), then the consequence (besides considerable anxiety) is a subsequent, more discerning test, which has a better chance of making the correct diagnosis second time around. If an infected person tests negative (Type II error), then the consequences are more dire, because treatment will be withheld, or at best delayed, and there is the risk of further infecting other individuals. Thus, in this case it makes sense to live with a higher probability of Type I error in order to diminish the probability of Type II error.

Example

Consider the following legal example: the null hypothesis is that the defendant is innocent and the alternative is that the defendant is guilty. The trial weighs evidence as in a hypothesis test in order to decide whether or not to reject the null hypothesis of innocence. What would Type I and II errors signify in this context? A Type I error means rejecting a null hypothesis that is true, in other words finding an innocent person guilty. Most people would agree that this is much worse than committing a Type II error in this context, which would be failing to convict a guilty person.

Dr. Stephen Fienberg of CMU did extensive work for the government is assessing the effectiveness of lie-detector tests. He concluded that probabilities of committing both types of error were so high that he and a panel of investigators recommended discontinuing the use of such tests. **A peek at a brain can unmask a liar** tells about the most recent technology for new sorts of lie detectors to replace the old-fashioned polygraph.

Role of Sample Size

Example

Suppose one demographer claims that there are equal proportions of male and female births in a certain state, whereas another claims there are more males. They use hospital records from all over the state to sample 10,000 recent births, and find 5120 to be males, or $\hat{p} = .512$. They test $H_0 : p = .5$ vs. $H_a : p > .5$ and calculate $z = \frac{.512-.5}{\sqrt{\frac{.5(.5)}{10,000}}} = 2.4$, so the P-value is .0084, quite small. Does this mean (a) they have evidence that the population proportion of male births is much higher than .5; or (b) they have very strong evidence that the population proportion of male births is higher than .5? The interpretation in (b) is the correct one; (a) is not.

Especially when the sample size is large, we may produce very strong evidence of a relatively minor difference from the claimed p_0. Conversely, if n is too small, we may fail to gather evidence about a difference that is quite substantial.

Example

A Statistics recitation instructor suspects there to be a higher proportion of females overall in Stats classes. She observes 12 females in a group of 20 students, so $\hat{p} = .6$. Does this confirm her suspicions? She would test $H_0 : p = .5$ vs. $H_a : p > .5$. First she verifies that $20(.5)$ and $20(1 - .5)$ are both 10, just barely satisfying our condition for a normal approximation. Also, she has in mind a population in the hundreds or even thousands, so the binomial model applies. She calculates $z = \frac{.6-.5}{\sqrt{\frac{.5(.5)}{20}}} = .89$. The P-value is .1867, providing her with no statistical evidence to support her claim. In fact, the population proportion of females really is greater than .5, but this sample size was just too small to prove it. In contrast, a lecture class of 80 students with $\hat{p} = .6$ would produce a z statistic of 1.79 and a P-value of .0367.

Remember that

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(\hat{p} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$$

We reject H_0 for a small P-value, which in turn has arisen from a z that is large in absolute value, on the fringes of the normal curve. There are three components that may result in a z that is large in absolute value, which in turn cause us to reject H_0:

1. What people tend to focus on as the cause of rejecting H_0 is a large difference $\hat{p} - p_0$ between the observed proportion and the proportion proposed in the null hypothesis. This naturally makes z large and the P-value small.

2. A large sample size n, because $\sqrt{n}$ is actually multiplied in the numerator of the test statistic z, brings about a large z and a small P-value. Conversely, a small sample size n may lead to a smaller z and failure to reject H_0, even if it is false (a Type II error).

3. If p_0 is close to .5, then $\sqrt{p_0(1-p_0)}$ is considerably larger than it is for p_0 close to 0 or 1. (For example, $\sqrt{p_0(1-p_0)}$ is .5 for $p_0 = .5$, but it is .1 for $p_0 = .01$ or .09.)

When Hypothsis Tests are *Not* Appropriate

Remember that we carry out a hypothesis test, based on sample data, in order to draw conclusions about the larger population from which the sample was obtained. Hypothesis tests are not appropriate if there is no larger group being represented by the sample.

Example

In 2002, the government requested and won approval for 1228 special warrants for secret wiretaps and searches of suspected terrorists and spies. Is this significantly higher than 934, which was the number of special warrants approved in 2001? Statistical inference is not appropriate here because 1228 and 934 represent entire populations for 2002 and 2001; they are not sample data.

Example

In 2000, $\frac{928{,}000}{1238{,}000} = .75$ of all bachelor's degrees were earned by whites. Is this significantly lower than .86, the proportion of all bachelor's degrees earned by whites in 1981? We would not carry out a significance test, because the given proportions already describe the population.

Example

An internet review of home pregnancy tests reports: "Home pregnancy testing kits usually claim accuracy of over 95% (whatever that may mean). The reality is that the literature contains information on only four kits evaluated as they are intended to be used—by women testing their own urine. The results we have suggest that for every four women who use such a test and are pregnant, one will get a negative test result. It also suggests that for every four women who are not pregnant, one will have a positive test result."

From this information we can identify the probabilities of both Type I and II errors, according to the review, as being 1 in 4, or 25%.

Example

Gonorrhea is a very common infectious disease. In 1999, the rate of reported gonorrhea infections was 132.2 per 100,000 persons. A polymerase chain reaction (PCR) test for gonorrhea is known to have sensitivity 97% and specificity 98%.

What are the probabilities of Type I and Type II Errors? Given the high degree of accuracy of the test, if a randomly chosen person in the U.S. is routinely screened for gonorrhea, and the test comes up positive, what is the probability of actually having the disease?

The null hypothesis would be that someone does not have the disease. A Type I Error would be rejecting the null hypothesis, even though it is true: testing positive when a person does not have the disease. A Type II Error would be failing to reject the null hypothesis, even though it is false: testing negative when a person does have the disease.

A sensitivity of 97% means that if someone has the disease, the probability of correctly testing positive is 97%, and so the probability of testing negative (when someone has the disease) is 3%: this is the probability of a Type II error. A specificity of 98% means that if someone does not have the disease, the probability of correctly testing negative is 98%, and so the probability of

testing positive (when someone does not have the disease) is 2%: this is the probability of a Type I error.

A two-way table makes it easier to identify the probability we are seeking (of having the disease, given that the test is positive). We begin with a total of 100,000 people, of whom 132 have the disease (the remaining 999,868 do not). Sensitivity 97% means 127 of the 132 with gonorrhea test positive. Specificity 98% means 979,871 of the 999,868 people without gonorrhea test negative. The remaining counts can be filled in by subtraction:

	Positive	Negative	Total
Gonorrhea	127	5	132
No Gonorrhea	1,997	97,871	99,868
Total	2,124	97,876	100,000

Of the 2,124 people who test positive, 127 actually have the disease: if someone tests positive, the probability of having the disease is $\frac{127}{2,124} = .06$. Remember, however, that this probability applies to a *randomly* chosen person being screened. If someone is screened because of exhibiting symptoms, the probability is of course higher.

Exercise : Refer to the article **How not to catch a spy: Use a lie detector**, which reports at the bottom of the first column, "Even if the test were designed to catch eight of every 10 spies, it would produce false results for large numbers of people. For every 10,000 employees screened, Fienberg said, eight real spies would be singled out, but 1,598 innocent people would be singled out with them, with no hint of who's a spy and who isn't." Based on this information, set up a two-way table, classifying 10,000 employees as actually being spies or not, and being singled out as a spy by the lie detector or not. Report the probability of a Type I Error and of a Type II Error. If someone is identified by the lie detector as being a spy, what is the probability that he or she is actually a spy?