

## Lecture 20

Nancy Pfenning Stats 1000

### Standardized Statistics

**Recall:** If the underlying population variable  $X$  is normal with mean  $\mu$ , standard deviation  $\sigma$ , then for a random sample of size  $n$ , the random variable  $\bar{X}$  is normal with mean  $\mu$ , standard deviation  $\frac{\sigma}{\sqrt{n}}$ . We used this fact to transform  $\bar{X}$  to a standard normal random variable  $Z$ , and solved for probabilities with normal tables:  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  is normal with mean 0, standard deviation 1. [Note that the spread of  $Z$  is always 1, regardless of sample size  $n$ .]

In situations involving a large sample size  $n$ , sample standard deviation  $s$  is approximately equal to  $\sigma$ , and we can treat  $\frac{\bar{X} - \mu}{s/\sqrt{n}}$  as (approximately) a standard normal variable  $Z$ .

If sample size  $n$  is small,  $s$  may be quite different from  $\sigma$ , and the random variable which we call  $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$  does *not* follow a standard normal distribution. Because of subtracting the expected value of  $\bar{X}$  (that is,  $\mu$ ) from  $\bar{X}$  in the numerator, the distribution of  $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$  is (like  $Z$ ) centered at zero and symmetric. Because of dividing by  $s/\sqrt{n}$  (which is *not* the standard deviation of  $\bar{X}$ ), the standard deviation of  $t$  is *not* fixed at 1 as it is for  $Z$ . Sample standard deviation  $s$  contains less information than  $\sigma$ , so the spread of  $t$  is greater than that of  $Z$ , especially for small sample sizes  $n$ . Since  $s$  approaches  $\sigma$  as sample size  $n$  increases, the  $t$  distribution approaches the standard normal  $Z$  distribution as  $n$  increases. Thus, the spread of sample mean standardized using  $s$  instead of  $\sigma$  depends on the sample size  $n$ . We say the distribution has  $n - 1$  “degrees of freedom”, abbreviated *df*.

Since there are many different  $t$  distributions—one for each *df*—it would take too much space to provide tables for each of them in as much detail as was provided for the standard normal  $z$  in Table A.1. Instead,  $t$  tables are condensed to provide minimal adequate information needed to state useful results.

### Statistical Inference

Statistical inference is the process of inferring something about a larger group (the population) by analyzing data for a part of that group (the sample). There are two general forms of statements we make using statistical inference: (1) confidence intervals; and (2) significance tests. We use these forms of inference in order to answer questions about (a) population proportion  $p$  [for categorical data] or (b) the population mean  $\mu$  [for quantitative data].

[In addition, we can use significance tests to answer questions about relationships between two variables, such as the chi-square test of a relationship between two categorical variables. The chi-square statistic  $\text{chi-square} = \text{sum of all } \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$  is another standardized statistic that follows a known pattern with values and probabilities that can be summarized in a table.]

#### 1. Confidence Interval Questions

- (a) (for  $p$ ) In May, 2000, .56 of 1,012 respondents to an Associated Press survey supported gays’ rights to inherit from their partners. What interval should contain the proportion of *all* Americans who support gays’ rights to inherit? How confident can we be that this interval contains the true proportion  $p$ ?
- (b) (for  $\mu$ ) A random sample of 25 laboratory mice from a large colony was found to have mean weight 33 grams and standard deviation 5 grams. Within what interval does mean weight for *all* colony mice lie? How confident can we be about the correctness of this interval?

#### 2. Significance Test Questions

- (a) (for  $p$ ) In May, 2000, .56 of 1,012 respondents to an Associated Press survey supported gays’ rights to inherit from their partners. Can we conclude that a majority of the population support gays’ rights to inherit?

- (b) (for  $\mu$ ) Researchers are going under the assumption that their lab mice weigh an average of 30 grams, but an assistant feels they actually weigh more. She takes an SRS of 25 mice and finds their mean weight to be 33 grams. [Somehow it is known that weights of all mice in the lab vary normally with standard deviation 5 grams.] If the mean weight were really only 30 grams, how unlikely would it be to get a sample of 25 whose mean weight is as high as 33 grams?

The laws of probability will enable us to answer such questions with precision. But these laws are inapplicable and useless if our data have not been produced correctly. [For example, maybe the lab assistant's selection was biased towards slower, heavier mice, or maybe it was biased towards smaller, cuter mice.] The sample *must* be chosen at random in such a way that it serves as an adequate representative of the entire population. The reliability of our conclusions still depends on conscientious adherence to the basic principles of statistical design presented in Chapters 3 and 4.

## Chapter 10: Estimating Proportions With Confidence

### Probability vs. Confidence

Recall: our **Rules for Sample Proportions** stated that if numerous samples or repetitions of the same size are taken, sample proportion  $\hat{p}$  has mean  $p$ , the true proportion for the population, standard deviation  $\sqrt{\frac{p(1-p)}{n}}$ , and a shape that is approximately normal as long as  $np \geq 10$  and  $n(1-p) \geq 10$ . Because of approximate normality, we can invoke the Empirical Rule: it tells us that the approximate probability is

.68 that  $\hat{p}$  falls within  $\sqrt{\frac{p(1-p)}{n}}$  of  $p$ ;

.95 that  $\hat{p}$  falls within  $2\sqrt{\frac{p(1-p)}{n}}$  of  $p$ ;

.997 that  $\hat{p}$  falls within  $3\sqrt{\frac{p(1-p)}{n}}$  of  $p$ .

If  $\hat{p}$  falls within  $\sqrt{\frac{p(1-p)}{n}}$  of  $p$ , then  $p$  must fall within  $\sqrt{\frac{p(1-p)}{n}}$  of  $\hat{p}$ ! Similarly, if  $\hat{p}$  falls within  $2\sqrt{\frac{p(1-p)}{n}}$  of  $p$ , then  $p$  must fall within  $2\sqrt{\frac{p(1-p)}{n}}$  of  $\hat{p}$ , etc.

But  $p$  is not a random variable like  $\hat{p}$ : its value is not a “numerical outcome of a random phenomenon”, but fixed and unchanging (even if we don't happen to know what it is).

Thus, we cannot talk about the “probability” of  $p$  lying in a certain interval. Instead, if we take a sample of size  $n$  from a population and record the sample proportion  $\hat{p}$  in the category of interest, we can be approximately “68% confident” that the interval  $\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}$  contains the unknown population proportion  $p$ .

Notice that the standard deviation of  $\hat{p}$  is  $\sqrt{\frac{p(1-p)}{n}}$ . Since  $p$  is unknown, this standard deviation cannot be known either, so we estimate it by substituting  $\hat{p}$  for  $p$ : the **standard error** of  $\hat{p}$  is

$$s.e.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

[In general, standard error is calculated from the sample as an estimate for population standard deviation.]

Now, combining the Empirical Rule with the language of confidence and the standard error approximation, we say we are approximately

**68% confident that  $p$  is in the interval  $\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ;**

**95% confident that  $p$  is in the interval  $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ;**

**99.7% confident that  $p$  is in the interval  $\hat{p} \pm 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ .**

The 95% confidence interval is by far the one most commonly seen. When news reports refer to the “margin of error”, they mean the give-or-take around the estimate that results in an interval that captures the unknown parameter with a 95% success rate in the long run, namely  $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ .

Note: If we substitute  $\hat{p} = .5$  into this expression for margin of error, the result equals the “conservative margin of error”  $\frac{1}{\sqrt{n}}$  introduced in Chapter 4 when we first discussed estimating a population proportion based on sample proportion. For values of  $\hat{p}$  further from .5 in either direction (that is, closer to 0 or 1),  $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$  will be considerably smaller than  $\frac{1}{\sqrt{n}}$ , and will let us be more precise in our interval estimate for  $p$ .

### Example

664 teenagers who reported having sex for the first time between 1999 and 2000 were asked where this first encounter took place; 56% said it was at their own or their partner’s home. Assuming those 664 constitute a random sample of all U.S. teens, give an approximate 95% confidence interval for the proportion of all U.S. teens having their first sexual encounter at home.

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .56 \pm 2\sqrt{\frac{.56(.44)}{664}} = .56 \pm .0385 \approx .56 \pm .04 = (.52, .60)$$

[In Chapter 4 we calculated the conservative margin of error to be .0388, quite close to this margin of error because .56 is close to .5.] Again, we are approximately 95% confident that the proportion of all teens having their first sexual encounter at their or their partner’s home is between 56%-4% and 56%+4%, that is, between 52% and 60%.

### Example

An article entitled **Helping stroke victims** reports: “Lowering stroke victims’ body temperature with cooling blankets and other means can significantly improve their chances of survival, researchers say. German researchers who took steps to reduce the temperature of 25 people who had suffered severe strokes found that 14 survived instead of the expected five.”

Based on the information provided, we can set up a 95% confidence interval for the proportion of *all* severe stroke victims who would survive with the cooling blanket treatment. First, the sample proportion of survivors is  $\hat{p} = \frac{14}{25} = .56$ . The 95% confidence interval is

$$.56 \pm 2\sqrt{\frac{.56(.44)}{25}} = .56 \pm .20 = (.36, .75)$$

In order to confirm that chances of survival are “significantly” improved, we note that the expected survival rate is only  $\frac{5}{25} = .20$ , which falls below the entire confidence interval for overall survival rate of those who are treated with cooling blankets.

**Exercise:** Find an article or report that includes mention of sample size and summarizes values of a categorical variable with a count, proportion, or percentage. Based on that information, set up a 95% confidence interval for population proportion in the category of interest.

## Lecture 21

Recall: We used the fact that the probability is 95% for sample proportion  $\hat{p}$  to fall within 2 standard errors of population proportion  $p$  (from the Empirical Rule) in order to construct a 95% confidence interval for unknown population proportion  $p$ , based on a sample proportion  $\hat{p}$  that has been observed.

### Example

Of 5685 respondents in a survey, 4948 confessed to routinely singing in their cars. Give a 95% confidence interval for the proportion of all people who routinely sing in their cars. Since  $\frac{4948}{5685} = .80$ , our 95% confidence interval is  $.80 \pm 2\sqrt{\frac{.8(.2)}{5685}} = .80 \pm .011 \approx (.79, .81)$ . Note that the larger sample size results in a smaller margin of error, and thus a narrower confidence interval.

## Other Levels of Confidence

### Example

1000 husbands and wives were surveyed about the secrets they kept from their spouses; the most common secret, admitted by 48% of the 40% who said they kept secrets (that is, by 19% of the original 1000), was not telling their spouses about the real price of something they bought.

1. Give a 95% confidence interval for the proportion of all spouses who kept a secret about the real price of something they bought.

$$.19 \pm 2\sqrt{\frac{.19(.81)}{1000}} = .19 \pm .0248 \approx (.17, .21)$$

(Notice that the margin of error is smaller this way than it is when we use the easier, more conservative formula  $\frac{1}{\sqrt{n}}$ , because .19 is rather far from .5.)

Unfortunately, the level of precision obtained from the Empirical Rule is not always adequate for our purposes, and so we turn now to normal tables to understand how to obtain a higher level of precision. In fact, 95% probability of being within a certain distance of the mean corresponds to left and right tail areas of .0250, which correspond to  $z$  not quite 2, but 1.96. The margin of error for 95% confidence is not quite  $2\sqrt{\frac{.19(.81)}{1000}} = .0248$  but  $1.96\sqrt{\frac{.19(.81)}{1000}} = .0243$ .

2. If I want to say “I’m *almost positive* that the population proportion  $p$  is in such-and-such an interval”, I may want to set my desired level of confidence at 99% instead of 95%.

First, if I had a standardized score  $z$ , the probability is .99 that  $z$  lies between what values  $-z^*$  and  $+z^*$ ? The ones that have area  $\frac{1.00-.99}{2} = .005$  to the left and right, respectively: According to Table A.1,  $-z^*$  with .005 to the left is between -2.57 and -2.58. An extra decimal digit of accuracy is obtained from the “infinite” row of Table A.2 (recall that the  $t$  distribution with infinite degrees of freedom is equivalent to the standard normal distribution): 99% confidence corresponds to  $z^* = 2.576$ ,  $-z^* = -2.576$ .

The standard normal variable  $Z$  of interest here is standardized sample proportion  $\frac{\hat{p}-p}{s.e.(\hat{p})}$

$$\begin{aligned} .99 &= P(-2.576 < Z < +2.576) \\ &= P(-2.576 < \frac{\hat{p}-p}{s.e.(\hat{p})} < +2.576) \\ &= P(-2.576(s.e.(\hat{p})) < \hat{p} - p < +2.576(s.e.(\hat{p}))) \\ &= P(p - 2.576(s.e.(\hat{p})) < \hat{p} < p + 2.576(s.e.(\hat{p}))) \end{aligned}$$

so .99 is the probability that  $\hat{p}$  lies within  $2.576(s.e.(\hat{p}))$  of  $p$ ,

so we are 99% confident that the interval  $\hat{p} \pm 2.576(s.e.(\hat{p}))$  contains  $p$ ,

i.e. that  $p$  is in the interval

$$.19 \pm 2.576\sqrt{\frac{.19(.81)}{1000}} = .19 \pm .03 = (.16, .22)$$

3. I can narrow the interval by reducing my confidence level: according to Table A.2, 90% confidence corresponds to  $z^* = 1.645$ . I can be 90% confident that  $p$  is in the interval  $\hat{p} \pm 1.645(s.e.(\hat{p}))$  that is, the unknown population proportion is in the interval  $.19 \pm 1.645\sqrt{\frac{.19(.81)}{1000}} = .19 \pm .02 = (.17, .21)$

Note the tradeoff: we have a higher rate of confidence for a wider, less precise interval, and a lower rate of confidence for a narrower, more precise interval.

In general, a level  $C$  confidence interval for any parameter is an interval computed from sample data by a method that has probability  $C$  of producing an interval that contains the true value of the parameter. For now (Chapter 10), the parameter of interest is  $p$ ; in Chapter 12 it will be  $\mu$  or other parameters involving population mean. We want to say our confidence level is  $C$  that

the actual proportion  $p$  lies in a certain interval, in other words that

$p$  lies within a certain distance of  $\hat{p}$ , in other words that

$p$  lies in the interval **estimate  $\pm$  margin of error**

where the estimate is  $\hat{p}$ , and the margin of error depends on confidence  $C$ .  $C$  equals a probability, associated with a standard normal value  $z^*$ . First note that if  $C$  is the area under the standard normal curve between  $-z^*$  and  $+z^*$ , then the regions to the left of  $-z^*$  and to the right of  $+z^*$  each have area  $\frac{1-C}{2}$ . We call  $z^*$  with probability  $\frac{1-C}{2}$  lying to the right under the standard normal curve the **multiplier** that accompanies the confidence level  $C$ . The “infinite” row of Table A.2 provides  $z^*$  values for the four most common confidence levels  $C$ :

1. .90 is the confidence level  $C$  for  $z^* = 1.645$
2. .95 is the confidence level  $C$  for  $z^* = 1.960$
3. .98 is the confidence level  $C$  for  $z^* = 2.326$
4. .99 is the confidence level  $C$  for  $z^* = 2.576$

For a given  $C$ , the approximate margin of error is  $z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ .

**Conditions:** The interval  $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$  is approximately correct as long as the population is at least ten times the sample, and  $n\hat{p}$  and  $n(1-\hat{p})$  are both at least 10. The former guarantees approximate independence of selections; if they were dependent, the standard deviation would change. The latter simply requires a check that there have been at least ten each of “successes” and “failures” observed.

In general

1. A 90% confidence interval for  $p$  is  $\hat{p} \pm 1.645 * s.e.(\hat{p})$
2. A 95% confidence interval for  $p$  is  $\hat{p} \pm 1.960 * s.e.(\hat{p})$
3. A 98% confidence interval for  $p$  is  $\hat{p} \pm 2.326 * s.e.(\hat{p})$
4. A 99% confidence interval for  $p$  is  $\hat{p} \pm 2.576 * s.e.(\hat{p})$

### Example

An article reported that in a random sample of 244 doctors, 184 said they would object to the sale of human organs for transplants. Obtain a 90% confidence interval for the proportion  $p$  of all doctors objecting to such sales.

First we find  $\hat{p} = \frac{184}{244} = .754$ . For  $C=.90$ ,  $z^* = 1.645$ .  $s.e.(\hat{p}) = \sqrt{\frac{.754(.246)}{244}} = .0276$ . A 90% confidence interval for  $p$  is  $.754 \pm 1.645(.0276) = .754 \pm .045 \approx (.71, .80)$ . We are 90% confident that between 71% and 80% of all doctors object to the sale of human organs for transplant.

Caution: the margin of error accounts for random sampling error only; it does not include bias which may result from the selection process, the wording of questions, etc.

### Choosing a Sample Size

Sometimes, before the sample has been taken, we have in mind a particular margin of error that we would like to report in our confidence interval. It is easy enough to take our expression for a conservative margin of error  $m = \frac{1}{\sqrt{n}}$  and turn it around to solve for  $n$  in terms of  $m$ :

$$n = \frac{1}{m^2}$$

Thus, if we desired a margin of error equal to .03, we would take  $n = \frac{1}{.03^2} = 1111$ . Polling organizations often sample roughly 1000 people and report a margin of error close to 3%.

If we desired a margin of error equal to .02, we would take  $n = \frac{1}{.02^2} = 2500$ . Note that as sample size goes up, margin of error goes down.

### Example

A New York Times article entitled **Lawsuits Cast Attention on Passengers' Blood Clots on Long Flights** describes a study published in the New England Journal of Medicine in September 2001. One detail of the study is that for passengers arriving at Charles de Gaulle Airport near Paris, there were 3 cases of pulmonary embolism for 2 million passengers who traveled more than 5000 miles. We should not use this information to set up a confidence interval for the proportion of all passengers traveling more than 5000 miles who would suffer from pulmonary embolism, because the number of "successes" is too small; the distribution of sample proportion wouldn't be normal enough to justify setting up a confidence interval based on normal critical values.

**Exercise:** Here is an excerpt from a Pittsburgh Post-Gazette article entitled **Criminal pasts cited for many city school bus drivers**: State auditors checking the records of a random sample of 100 city bus drivers have found that more than a quarter of them had criminal histories.

The audit also found that 26 of the drivers were never checked for child abuse histories—in Pennsylvania schools, a mandate for all employees and even some volunteers.

In all, the auditors discovered 80 convictions for various offenses among the 100 sampled. Thirty-four of those incidents occurred more than ten years ago, including one rape and four drug offenses.

In Pennsylvania, it's perfectly legal for school officials to hire a bus driver with certain convictions that are more than five years old—but that doesn't mean they should, state Auditor General Robert P. Casey Jr. said yesterday in releasing the report.

"No one convicted of rape should be driving a school bus full of children," said Casey, who also said he was disappointed with the school district's initial response to the audit. "The General Assembly needs to look at this law," he said.

A series of problems last year with school bus drivers—including a February accident that was nearly fatal to an 8-year-old Elliott girl—prompted Casey to take a closer look at Pittsburgh's staff of 750 drivers, he said.

When his office presented their results to school officials about eight months ago, Casey said, "they were very reluctant to do anything about it," and sent him only a brief response outlining what steps were being taken to remedy the problems... Note that the article states that about 25% in a sample of Pittsburgh school bus drivers had criminal records. Report a 98% confidence interval for the proportion of all Pittsburgh school bus drivers with criminal records. One of the conditions for our approximation is not quite met; what is it?

## Lecture 22

### Interpreting Confidence Intervals

#### Example

Suppose the proportion  $p$  of M&Ms that are blue is unknown, and when I take a sample of 75 M&Ms to estimate  $p$ , I get  $\hat{p} = 9/75 = .12$  that are blue. A 95% confidence interval for  $p$  is  $.12 \pm 2\sqrt{\frac{.12(.88)}{75}} = .12 \pm .075 = (.045, .195)$ . Tell whether each of the following is a correct interpretation of this interval:

1. The probability is 95% that the proportion of all M&Ms that are blue is between .045 and .195. No: this is the most common misinterpretation of the interval, and the word "probability" is the problem. Even though it may be unknown, population proportion  $p$  is a fixed parameter, not subject to the laws of probability. Remember that probability is the study of random behavior; it applies to random variables, not to parameters.
2. The probability is 95% that the sample proportion of blue M&Ms is between .045 and .195. No: in fact, the probability is 100% that our *sample* proportion  $\hat{p}$  is in the confidence interval,

because we built the interval around  $\hat{p}$ ! Remember that setting up a confidence interval is a form of statistical inference, a process whereby we use *statistics* to draw conclusions about *parameters*, and so we need to be making a statement about  $p$ , not  $\hat{p}$ .

3. We are 95% confident that the proportion of all M&Ms that are blue is between .045 and .195. Yes.
4. The probability is 95% that the interval we produced, (.045, .195), contains  $p$ . Yes: because sample proportion  $\hat{p}$  varies from sample to sample, the interval built around  $\hat{p}$  varies randomly as long as the sample was random. Thus, the word “probability” does apply to the interval produced.

Picture 100 students each selecting a random sample of 75 M&Ms from a large bowlful and setting up a 95% confidence interval for the proportion  $p$  of all M&Ms that are blue. Roughly 95% of those 100 intervals, that is, 95 of the intervals, should contain  $p$ .

Now imagine the students each randomly selected 75 M&Ms from a huge barrelful instead of a bowlful. Would their confidence intervals be any more or less accurate? No: population size does not enter into our calculations. It is irrelevant as long as it is at least ten times the sample size, and as long as the samples are selected at random.

## Using Confidence Intervals to Guide Decisions

### Example

In a group of 371 college students, 196 wore some type of corrective lenses.

1. Give a 95% confidence interval for the proportion of all college students wearing corrective lenses. Since  $196/371 = .53$ , our interval is

$$.53 \pm 2\sqrt{\frac{.53(.47)}{371}} = .53 \pm .05 = (.48, .58)$$

2. Are you convinced that a majority of students wear corrective lenses? No, because the interval (.48, .58) contains values less than .5, suggesting that the population proportion  $p$  is not necessarily greater than .5.

### Example

In a group of 371 college students, 128 wore *contact* lenses.

1. Give a 95% confidence interval for the proportion of all college students wearing contact lenses. Since  $128/371 = .35$ , our interval is

$$.35 \pm 2\sqrt{\frac{.35(.65)}{371}} = .35 \pm .05 = (.30, .40)$$

2. Are you convinced that a minority of students wear contact lenses? Yes, because the interval (.30, .40) doesn't even come close to containing proportions of .5 or more.

### Example

$32/233 = .137$  of the 233 females in a group of college students wore glasses whereas  $36/138 = .261$  of the 138 males wore glasses. Compare the confidence intervals for population proportions of females and of males wearing glasses in order to decide if these population proportions could be equal.

For the females, a 95% confidence interval for  $p$  is  $.137 \pm 2\sqrt{\frac{.137(.863)}{233}} = .137 \pm .045 = (.092, .182)$ .

For the males, a 95% confidence interval for  $p$  is  $.261 \pm 2\sqrt{\frac{.261(.739)}{138}} = .261 \pm .075 = (.186, .336)$

The males' interval for  $p$  is higher than that of the females', to the point where the intervals share no overlap. It seems doubtful that the proportion of all males wearing glasses is the same as the proportion of all females wearing glasses.

### Example

A 95% confidence interval for the proportion of female students smoking is  $(.081, .167)$  and a 95% confidence interval for the proportion of male students smoking is  $(.062, .170)$ . Is it reasonable to assume that the proportion smoking is the same for females and males? Yes, because the intervals do overlap.

## When Confidence Intervals are *Not* Appropriate

Remember that we set up a confidence interval, based on sample data, in order to draw conclusions about the larger population from which the sample was obtained. Confidence intervals are not appropriate if there is no larger group being represented by the sample.

### Example

In 2000,  $\frac{928,000}{1238,000} = .75$  of all bachelor's degrees were earned by whites. Construct a 95% confidence interval for the proportion of all bachelor's degrees earned by whites. There is no need to construct a confidence interval, because the given proportion already describes the population.

## Two Types of Inference: Confidence Intervals and Hypothesis Tests

In the preceding examples, we examined confidence intervals for population proportion in order to get a feel for whether or not the population proportion could take a hypothetical value. Because this type of conclusion, in the form of a yes-or-no decision, is often quite important, we will now take a more rigorous approach to such problems. The following pairs of problems will help us to highlight the similarities and differences between situations involving confidence intervals and hypothesis tests.

- In a group of 371 Pitt students, 42 were left-handed. Give a 95% confidence interval for the proportion of all Pitt students who are left-handed.
  - 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?
- In a group of 371 students, 45 chose the number seven when picking a number between one and twenty "at random". Give a 95% confidence interval for the proportion of all students who would pick the number seven.
  - 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 choosing seven was significantly higher than  $\frac{1}{20} = .05$ ?
- One year, a university offers admission to 1200 students and 888 accept. Assuming that year is representative of all the recent years, give a 99% confidence interval for the proportion accepting in any given year.
  - 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?



Like the confidence interval problems 1(a), 2(a), and 3(a), the significance test problems 1(b), 2(b), and 3(b) all involve a single categorical variable with two possible values (smoking or not, picking the number seven or not, accepting admission or not). We know the sample size  $n$  and the sample count  $X$  in the category of interest, and so can calculate the sample proportion  $\hat{p} = \frac{X}{n}$  in the category of interest. Based on this sample proportion, we want to draw conclusions about the unknown population proportion  $p$ . In a confidence interval problem, our conclusion takes the form of an interval estimate for  $p$ . In a hypothesis test problem, a hypothetical value for unknown population proportion  $p$  is proposed, and we need to decide whether or not  $p$  really takes that proposed value. We will begin to solve such problems next lecture.