Statistical Inference:
Single Mean and Proportion Problems

Alan B. Gelder
06E:071, The University of Iowa[1]

1 Why do we do statistics, anyway?

Definition. (POPULATION) *A statistical population consists of an entire set that we are interested in studying. This set can be any measurable characteristic about people, places, or things.*

Definition. (SAMPLE) *A statistical sample comprises the portion of the population that is actually measured. The sample is represented by the data available to the researcher.*

The whole reason we do statistics is because we want to discovery the true properties of a population. However, gathering information about every member of a population is often incredibly difficult, expensive, or even impossible. So we rely on samples to make inferences about their corresponding populations. The study of statistics is the study of how to make such inferences about an entire population. It is also the study of the accuracy of these inferences.

Definition. (PARAMETER) *A parameter is a number which describes the distribution of a population.*

Definition. (STATISTIC) *A statistic is a number which describes the distribution of a sample.*

Remember, the P's (population and parameter) and the S's (sample and statistic) go together.

μ, p, σ, and β_i are examples of parameters we will use in this course.
$\overline{x}, \widehat{p}, s$, and b_i are their corresponding statistics.

A key thing to realize is that we typically do not know what the values of the parameters are. In fact, if we knew the actual values of the parameters, then we could ignore probably 99% of the material we will cover in this class. We wouldn't need to

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

make confidence intervals or do hypothesis testing—statistics would be pretty easy. Again, the problem we face is that it is generally very difficult to gather information about every member of a population.

2 Hypothesis Testing

In hypothesis testing, we follow the logic of a proof by contradiction.

The first thing we do is to make two mutually exclusive and exhaustive possibilities. Either something happened or it did not. Either the average height of Americans is 5'6" or it is not. The University of Iowa student population is more than 53% female or it is not.

We call each of the two possibilities a hypothesis. The null hypothesis H_0 always has some form of equality ($=, \geq$, or $\leq$). The alternative hypothesis H_A always has some form of inequality ($\neq, <, >$). H_A can also be called the research hypothesis because it is the idea or theory that we want to test.[2] By using the appropriate signs for the null and alternative hypotheses we can make two groups which cover all of the possible values for the parameter we are interested in knowing more about.

Example hypotheses:

$\mu =$ The mean height of all Americans.
$H_A:$ $\mu \neq 5'6''$
$H_0:$ $\mu = 5'6''$

$p =$ The proportion of female students at the University of Iowa.
$H_A:$ $p > .53$
$H_0:$ $p \leq .53$

We begin by assuming that the null hypothesis H_0 is true (which is why it is always marked by some form of equality: $=, \geq$, or $\leq$). Just like in a court of law, H_0 is innocent until it is proven guilty. Or rather, H_0 is assumed to be true until there is sufficient evidence that it is false. If there is strong evidence against H_0, then that means that there is strong evidence to support H_A. Remember, there are only two possible outcomes: either H_0 or H_A is correct.

The sample is the evidence that we can use to test whether or not H_0 is true. A

[2] H_A is sometimes denoted H_1.

sample is taken from a population, and a population has some kind of distribution. Since we are assuming that H_0 is true, then we will assume that H_0 is a characteristic of the population that the sample is drawn from. For instance, if H_0: $\mu = 5'6''$, like in the example above, then we assume that our sample is drawn from a population where the mean height of Americans is 5'6".

With the distribution for H_0 in mind, we can use a test statistic to identify where our sample lies on the distribution. This will give us an idea of how likely it is that our sample would be selected at random.

For one-proportion problems, the test statistic is

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

The test statistic for one-mean problems is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{s}$$

Four-Steps for Hypothesis Testing

1. **Define**
   - *a*) English definitions for parameters being tested
   - *b*) Hypotheses stated using mathematical symbols
2. **Calculate**
   - *a*) test-statistic
   - *b*) p-value
3. **Decide**
   - *a*) Reject H_0 or fail to reject H_0
   - *b*) Numeric evidence to support decision
4. **Interpret**
   Reject $H_0 \Rightarrow$ "There is sufficient evidence to support H_A."
   Fail to reject $H_0 \Rightarrow$ "There is not sufficient evidence to support H_A."
   [Write H_A out in words. Do not include mathematical symbols like H_A, H_0, p-value, α, etc.]

We will be using these four-steps regularly throughout the semester (**the four-step process would be a great thing to memorize**). The best way to become familiar with the four-step procedure for hypothesis testing is to see it in practice. We

will first do this with one-proportion and one-mean problems.

3 One-Proportion Problems

Proportion problems rely on binary categorical data, such as surveys that have questions with yes/no answers.

Example: Henry Brown is running for governor. In order to be a viable candidate in the primary elections, he must have more than 35% of the votes. In a recent opinion poll, 328 out of 900 registered voters said that they were planning on voting for Henry Brown. Will Mr. Brown be able to get more than 35% of the votes at the primary election? Use $\alpha = .05$.

Four-Step Hypothesis Test

1. a) p = proportion of all voters who will vote for Joseph Brown in the upcoming primary election.
b) $H_A : p > 0.35 \qquad H_0 : p \leq 0.35$

2. a) $\hat{p} = 328/900 = 0.3644 \qquad p_0 = 0.35 \qquad n = 900$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.3644 - 0.35}{\sqrt{\frac{(0.35)(1-0.35)}{900}}} = \frac{0.0144}{0.0159} = 0.9057$$

b) We can find the p-value using the z-table. In order to use the z-table, round the z-statistic to two decimal places: $z = 0.9057 \approx 0.91$. Note that the z-table only has the left side of the normal distribution. This is okay since the normal distribution is symmetric with 0 at the center. Hence, if our z-statistic is positive, we just need to look for -1 times our z-statistic. Find 0.9 on the left hand side of the table (the table only has negative numbers, so look instead for -0.9). Now find 0.01 on the top of the table (the 0.01 accounts for the 1 in 0.91). The table gives us the number .1814 where the -0.9 row and the 0.01 column meet. The number 0.1814 is the area under the normal distribution to the right of the our z-statistic, 0.91. Since H_A has a greater than sign(>), the p-value is the area under the distribution which is greater than (or to the right of) 0.91. Hence, 0.1814 is the p-value.

3. a) Fail to Reject H_0 because (b) p-value $= 0.1814 > 0.05 = \alpha$.

4. There is not sufficient evidence to conclude that Mr. Brown will obtain more than 35% of the vote in the upcoming primary election.

Remember that it is possible that Mr. Brown will still be able to receive more than 35% of the vote. Afterall, over 36% of the voters in the survey were planning on voting for him. However, based on the p-value, there is still too much uncertainty or risk that Mr. Brown will not receive 35% of the vote.

Interpretation of the P-Value
In hypothesis testing, we initially assume that H_0 is true. If we decide that H_A is instead correct and reject H_0, the p-value is the probability that we made a mistake.

In the example above, we had an 18% chance of making a mistake if we rejected H_0 (p-value = 0.1814). However, we were only willing to tolerate a 5% chance of being wrong ($\alpha = 0.05$). Hence, it was too risky to say conclusively that Mr. Brown would be successful in the election.

A Note On Finding P-Values
1. Draw the bell curve and label where your test-statistic is (e.g. z-stat or t-stat).

2. Look up the value for the test-statistic in the table. If the test-statistic is *positive*, then the number from the table is the area under the distribution to the *right* of the test-statistic. If the test-statistic is *negative*, then the number from the table is the area under the distribution to the *left* of the test-statistic.

3. The sign in H_A determines which way you want to shade in order to obtain the p-value. If the sign in H_A is $>$, then the p-value is the area to the right of the test-statistic. If H_A has $<$, then the p-value is the area to the left of the test-statistic. If H_A has $\neq$, then the p-value is two times the number in the table (i.e. the area beyond the positive and negative values of the test-statistic in the tails of the distribution).

4. Use the number from the table and the shading from H_A to help you determine the p-value, keeping in mind that the entire area under the distribution equals 1.

Confidence Intervals for Single Proportion Problems
Hypothesis tests are good for obtaining yes/no answers to a question about the

value of a parameter. Confidence intervals, on the other hand, give us a range for where the parameter is likely to be.

The formula for confidence intervals in one-proportion problems is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

In this formula, z^* can be found at the bottom of the t-table. For instance, z^* for a 95% confidence interval is 1.960, and z^* for a 99% confidence interval is 2.576.

Returning to Mr. Brown's gubernatorial campaign, suppose that we want to predict with 95% accuracy the proportion of the vote that he will receive. This can be done as follows:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.3644 \pm 1.96 * \sqrt{\frac{0.3644(1-0.3644)}{900}}$$

$$= 0.3644 \pm 1.96 * 0.0160 = 0.3644 \pm 0.0314 = (0.3330, 0.3958)$$

Interpretation of the Confidence Interval: With 95% confidence, Mr. Brown will receive between 33.3% and 39.58% of the vote at the primary election.

A confidence interval corresponds to a two-tailed hypothesis test (where H_A uses $\neq$).

4 One-Mean Problems

Example: The ACT scores for a random sample of students attending an elite college are as follows: 28, 31, 30, 29, 33, 32, 34, 31, 32. Is it true that the mean ACT score at this college is lower than 30? Use $\alpha = 0.05$.

Four-Step Hypothesis Test
1. a) μ = The mean ACT score of students attending the elite college.
b) $H_A : \mu < 30 \qquad H_0 : \mu \geq 30$

2. a) $\bar{x} = \frac{28+31+30+...+32}{8} = \frac{280}{9} = 31.1111 \qquad \mu_0 = 30 \qquad n = 9 \qquad s = 1.900$

$$t = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{s} = \frac{\sqrt{9}\,(31.1111 - 30)}{1.9} = \frac{3.3333}{1.9} = 1.7544$$

Refer to the calculator help website for instructions on how to find the standard deviation s on your calculator:
http://www.users.muohio.edu/kuiperc/CalculatorHelp/index.html

b) Unlike the z-table, the t-table only shows the area under the curve for positive t-statistics (the t-distribution is also symmetric about zero). On the left hand side is "df" or degrees of freedom. For one-mean problems, df = n-1. In our example, df = 8. Trace the row for df = 8 to the right until you find the numbers that are closest to the t-statistic (1.7544) from below and above (1.397 and 1.860). Now, follow these two columns up to the top of the table to .10 and .05. The .10 and the 0.05 indicate that the area under the curve to the *right* of the t-statistic is between .05 and .10. However, based on H_A, the p-value is the area to the *left* of the t-statistic.

$$\Rightarrow 1 - .10 = 0.90 < p - value < 0.95 = 1 - .05$$

3. Fail to Reject H_0 since p-value $> 0.90 > 0.05 = \alpha$

4. There is insufficient evidence that the mean ACT score of students at the elite college is less than 30.

Confidence Intervals for Single Mean Problems
The formula for finding confidence intervals for one-mean problems is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

Example: Find a 95% confidence interval for the mean ACT score of students at the elite college:

We'll use the t-table to find t^*. Again, we have 8 degrees of freedom (n=9). Confidence levels are given at the bottom of the table. At the intersection of the column for 95% confidence and the row for 8 df is the number 2.306. Hence, $t^* = 2.306$.

$$31.1111 \pm 2.306 * \left(\frac{1.900}{\sqrt{9}}\right) = 31.1111 \pm 1.4605 = (29.6506, 32.5716)$$

Interpretation of the Confidence Interval: With 95% confidence, the mean ACT score of students at the elite college is between 29.65 and 32.57.

5 Formulas

One-Proportion Hypothesis Test (z-statistic):

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

One-Proportion Confidence Interval:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

One-Mean Hypothesis Test (t-statistic):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{s}$$

One-Mean Confidence Interval:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$