7. Statistical Inference: Significance Tests

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

Last time

- Estimators vs. estimates
- Point estimation vs. interval estimation
- Confidence interval for μ when σ is known
- Confidence interval for π
- Confidence interval for μ when σ is *unknown* (using sample standard deviation s, and t-distribution)
- Assumptions behind confidence interval estimation
- Determining minimum sample sizes for desired interval estimate precision

Significance testing

Two approaches to statistical inference about a population mean μ:

1. Providing an interval estimate of μ
   - We have no *a priori* idea of what μ is
   - The confidence interval represents a range of "likely values" for μ
   - The width of the interval reflects sampling variability in the estimator.
2. Testing hypotheses about μ
   - Begins with a hypothesis about μ
   - Asks whether the sample statistic $\bar{x}$ is consistent with this hypothesis
   - Requires knowledge of the *sampling distribution* of the sample statistic (e.g., $\bar{x}$) to assess how plausible the hypothesis about μ is
   - This approach is called *hypothesis testing* or *significance testing*

Statistical hypotheses

A statistical **hypothesis** is a statement about a population. It is usually a statement that a parameter takes a particular value or falls in some range of values. Hypotheses come from a variety of sources:

- Past experience or "common wisdom"
- Prior research
- Inductively, from a study of past observations
- Deductively, from theory
- A presumption of "no difference," "no change," "no effect", etc.

A **hypothesis test** or **significance test** asks whether the observed data are *consistent* or *inconsistent* with a hypothesis. That is, how plausible is the hypothesis in light of the observed data?

Statistical hypotheses

- The **null hypothesis** ($\mathbf{H_0}$): a statement that a population parameter takes a particular value. The null hypothesis is often the *lack* of a hypothesized finding: e.g., "no effect" or "no difference."

- The **alternative hypothesis** ($\mathbf{H_1}$ or $\mathbf{H_a}$): a statement that the population parameter falls in some alternative range of values (contrary to H_0). The alternative hypothesis is often an "effect" or "difference" that the researcher is testing for.

Statistical hypotheses

Example: I am interested in testing whether the mean number of study hours per week is higher among Vanderbilt undergraduates than the national average. (This is a *directional* hypothesis since I am stating "higher," and not "different from").

- H_0: the mean study hours per week at Vanderbilt is *the same* (or less) than the national average.
- H_1: the mean study hours per week at Vanderbilt is *higher* than the national average.

$$H_0 : \mu_{vu} = \mu_0$$

$$H_1 : \mu_{vu} > \mu_0$$

μ_0 represents the population mean for Vanderbilt undergraduates *if* H_0 *is true*. Here, μ_0 is the national average.

Statistical hypotheses

Example: I believe home lending in my community is racially discriminatory. (That is, minority applicants are *less* likely to be approved for home loans. This is again a directional hypothesis.)

- H_0: the proportion of minority applicants awarded home loans is the *same as* (or greater than) the proportion of white applicants given home loans.
- H_1: the proportion of minority applicants awarded home loans is *lower than* the proportion of white applicants given home loans.

$$H_0 : \pi_m = \pi_0$$

$$H_1 : \pi_m < \pi_0$$

π_0 represents the population proportion of minority applicants awarded home loans *if* H_0 *is true*. Here, π_0 is the proportion of white applicants awarded home loans.

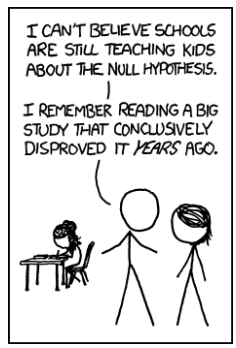
Statistical hypotheses

Note that in both of the above examples H_0 asserts that the population parameter takes a particular value. H_1, on the other hand, asserts that the parameter falls in some *range* of alternative values (e.g., greater than or less than some value).

The logic of hypothesis testing

H_0 is *presumed to hold* unless the evidence strongly suggests otherwise (like a "proof by contradiction").

- A significance test can either *reject* or *not reject* the null hypothesis.
- One should not say we "accept" H_0, or that H_0 is "true," only that the evidence is not sufficient to reject it.
- Important: rejection is in a *statistical sense—even if rejected* there is a chance H_0 is true.



Hypothesis testing in practice

FDA screening: new drugs must be approved for use by the Food and Drug Administration. According to federal law, drugs must be demonstrated to be both safe and effective.

FDA screening starts from the assumption (H_0) that *the drug does not work*. It is up to the manufacturer to demonstrate otherwise: that the evidence is sufficient to reject the null hypothesis with high confidence.

Example 1

The distribution of math SAT scores in the population has a mean of 500 and $\sigma = 100$. You believe that the mean math SAT score in California is *higher* than the national average (assume σ is the same). To test this hypothesis, you randomly sample n=1,600 California math SAT scores. (Assuming we don't have access to the population of California scores).

- Let μ_c represent the population mean SAT in California.
- H_0: $\mu_c = 500$
- H_1: $\mu_c > 500$

The null hypothesis is that $\mu_c = \mu_0 = 500$, i.e., means scores are the same as the national average.

Example 1

We compute the sample mean ($\bar{x}$) from our random sample of California scores. If the sample mean is "sufficiently higher" than 500 we *reject* H_0 in favor of H_1. The question is:

- What value of $\bar{x}$ is "sufficiently higher?"
- What value of $\bar{x}$ would make H_0 seem *implausible*?

Example 1

Remember, in hypothesis testing, we begin with the assumption H_0 is true. In this example, *if* H_0 *were true*, sample means calculated from random samples of size $n = 1,600$ will:

- be normally distributed (per the CLT)
- have a mean of 500
- have a standard error of $\sigma/\sqrt{n} = 100/\sqrt{1600}$, or 2.5

Note: we are assuming σ is known.

Example 1

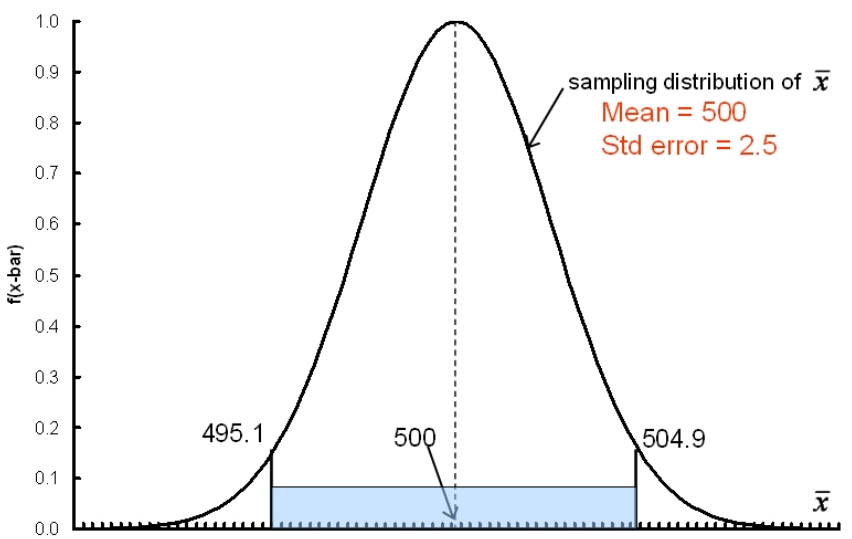


Figure: Distribution of $\bar{x}$ under H_0

Example 1

If H_0 is true, then 95% of the time $\bar{x}$ calculated from a random sample of $n = 1,600$ will fall between:

$$\mu_0 \pm 1.96(\sigma/\sqrt{n})$$

$$500 \pm 1.96(2.5) = (495.1, 504.9)$$

Any realized $\bar{x}$ in this interval wouldn't be that unusual. What about a realized sample mean of $\bar{x} = 507$?

Example 1

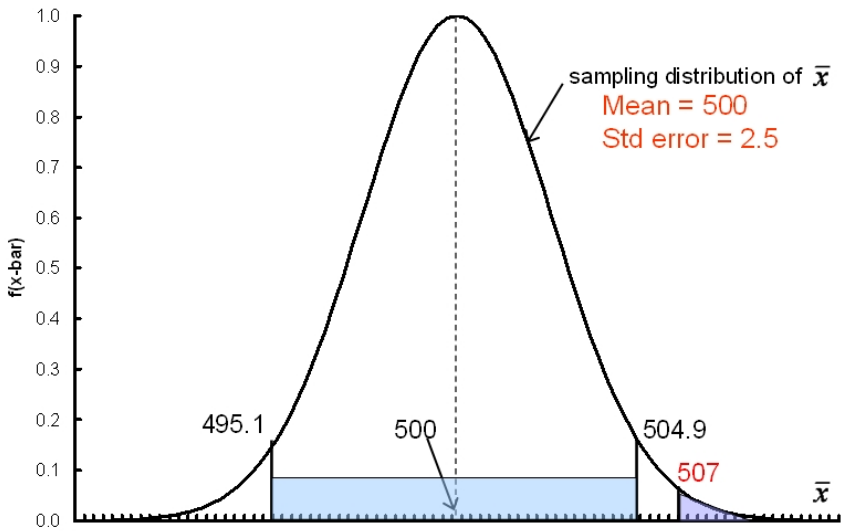


Figure: Distribution of $\bar{x}$ under H_0, and a realized sample mean of 507

Example 1

507 is $(507 - 500)/2.5 = 2.8$ standard errors above the mean. The probability of obtaining a sample mean of 507 or higher, assuming H_0 is true, is:

$$Pr(z > 2.8) = 0.0026$$

In other words, *very unlikely*. The value 2.8 is called a **test statistic**. 0.0026 is called the ***p*-value**.

In Stata: `display 1-normal(2.8)`

Example 1

What about a realized sample mean of $\bar{x} = 502$? 502 is $(502 - 500)/2.5 = 0.8$ standard errors above the mean. The probability of obtaining a sample mean of 502 or higher, assuming H_0 is true, is:

$$Pr(z > 0.8) = 0.2119$$

In other words, not that unlikely. The value 0.8 is our test statistic, and 0.2119 is the p-value.

In Stata: `display 1-normal(0.8)`

p-values

In general the **p-value** is the probability that a test statistic equals the realized value *or a value even more extreme* in the direction predicted by H_1, if H_0 is true.

- Above, the probability of obtaining an $\bar{x}$ of 507 or higher was 0.0026.
- Above, the probability of obtaining an $\bar{x}$ of 502 or higher was 0.2119.
- Intuitively, if we obtain an $\bar{x}$ that would be highly unlikely if H_0 were true (such as 0.0026) then H_0 *was probably not true to begin with.*
- The *less consistent* is the test statistic with H_0, the *smaller* is the p-value.

Significance levels

The researcher decides ahead of time the threshold p-value at which she will conclude the evidence is sufficiently strong against H_0.

- The threshold value is called the **significance level** of the test, or α.
- Some common significance levels are 0.05, 0.01, 0.10
- When $p < 0.05$ we say the result is "significant at the 0.05 level."
- When $p < 0.01$ we say the result is "significant at the 0.01 level," etc.
- The *higher* is α, the "lower the bar" for rejection of H_0.
- The *smaller* is α, the "higher the bar" for rejection.

Significance levels

- If $p < \alpha$, we *reject* H_0 in favor of H_1.
- If $p > \alpha$, we *do not reject* H_0.
- Rejection of H_0 is usually referred to as a **statistically significant** result, effect, or difference.
- Again, even if we cannot reject H_0, this *does not mean it is true*. In other words, we should not say that we "accept" H_0.
- Rejection only means the test statistic would be unlikely if H_0 were true. It is still possible for H_0 to be rejected when true; in fact, with a 0.05 significance level, this will occur in 5/100 random samples.

Hypothesis tests about μ

Guide to the next slide:

- Your test begins from the assumption that H_0 is true, in which case $\bar{x}$ has the null sampling distribution.
- Suppose H_0 is *not* true, but some alternative is (H_A), in which case $\bar{x}$ has the sampling distribution under H_A.
- We will not reject H_0 in the yellow region, *even if* H_0 *is false.*
- More on this later, in Lecture 8.

Hypothesis tests about μ

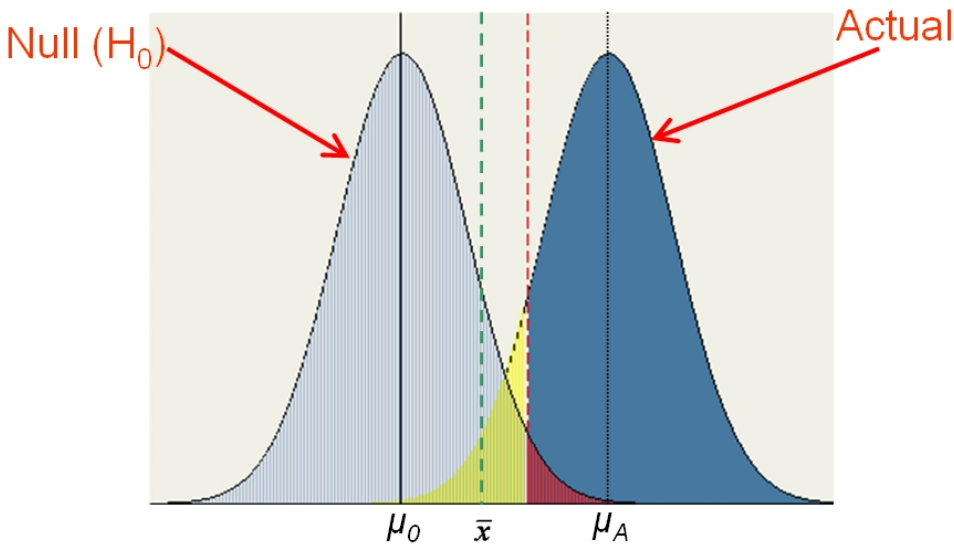


Figure: Distribution of $\bar{x}$ under H_0 and actual distribution

Example 2

Is average adult body temperature in the population really 98.6 degrees? Or is it lower? Assume $\sigma = 0.6$.

$$H_0 : \mu = 98.6$$

$$H_1 : \mu < 98.6$$

We sample 80 adults, and obtain a sample mean of $\bar{x} = 98.4$. Use a significance level of $\alpha = 0.05$ to conduct the test.

Example 2

How likely is it that 80 randomly selected adults would have an $\bar{x} = 98.4$ if μ were actually 98.6 (H_0)? If H_0 were true, then $\bar{x}$ is:

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{98.4 - 98.6}{0.6/\sqrt{80}} = -2.98$$

standard errors below μ (-2.98 is our test statistic). This is pretty unlikely (p=0.001). As $p < \alpha$, we reject H_0. Mean body temperature here is lower than 98.6 and this difference is *statistically significant* at the 0.05 (5%) level.

In Stata: `display normal(-2.98)`

Example 2

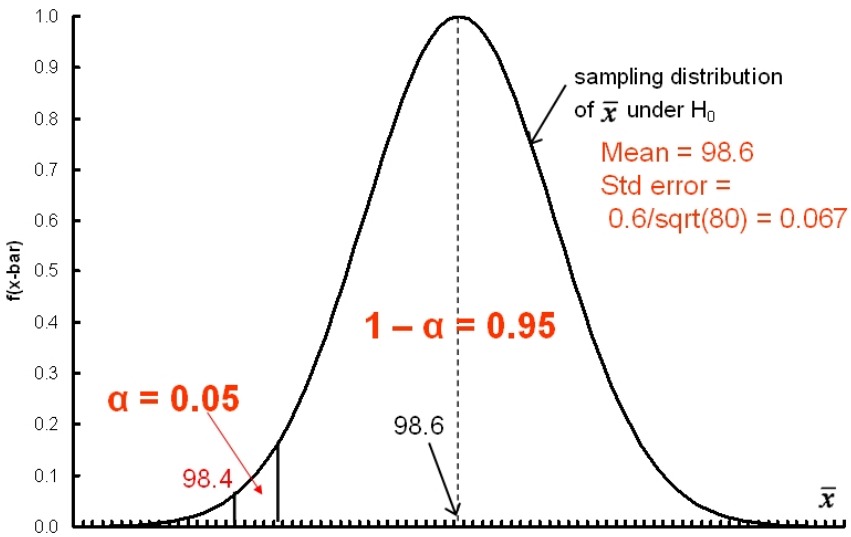


Figure: Distribution of $\bar{x}$ under $H_0 = 98.6$

One- vs. two-sided alternatives

- The alternative hypotheses thus far have been directional, "one-tailed" or "one-sided."
- The tests are constructed so that the rejection region is in one tail of the sampling distribution:
  - Are mean math SAT scores in California *higher* than in the U.S. population?
  - Is mean adult body temperature *lower* than 98.6 degrees?
  - Are *fewer* minority applicants awarded home loans than whites?
- "Two-tailed" or "two-sided" alternatives contain values both below and above the value stated in H_0—they are nondirectional.
- Rejecting H_0 means the parameter is significantly *different* from that stated in H_0, or there is some "effect."

Example 3

Suppose you would like to test whether mean weekly expenditure on alcoholic beverages by State U undergraduates differs from the national average of \$20. (Nationally, $\sigma = 9$, and you assume this is true for State U as well). You decide on a significance level of $\alpha = 0.05$.

$$H_0 : \mu_{stu} = 20$$

$$H_1 : \mu_{stu} \neq 20$$

You take a random sample of $n = 64$ students, and obtain a sample mean of $\bar{x} = \$17$. How likely is it that a random sample of 64 State U students would spend an average amount that *differs by \$3 or more* from μ, if μ were \$20? (I.e. \$17 or less, \$23 or more).

Example 3

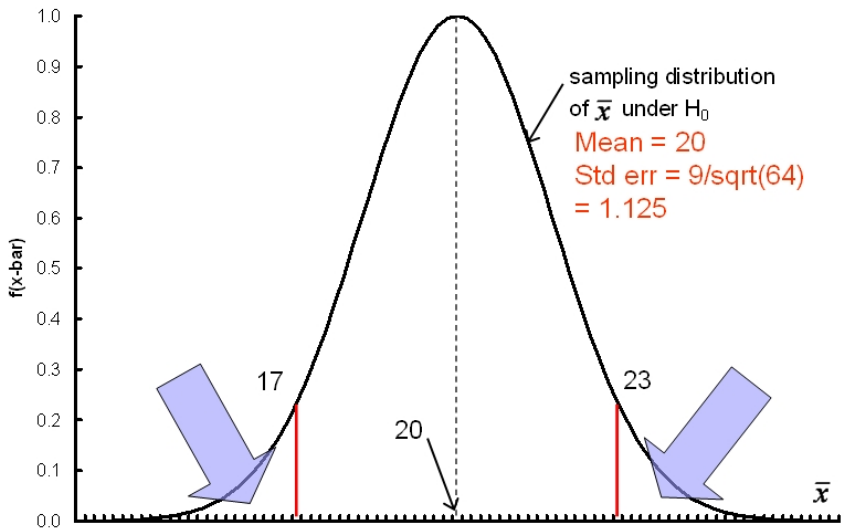


Figure: Two-tailed test example

Example 3

If H_0 is true, the test statistic for $\bar{x}$ is: $\frac{(17-20)}{9/\sqrt{64}}$ = -2.67. Under the standard normal distribution, the probability of obtaining a test statistic this far away from the mean *in either direction* is:

$$2 \times Pr(z < -2.67) = 2 \times 0.0038 = 0.0076$$

Since $p < \alpha$, we reject H_0 in favor of H_1. It is unlikely a random sample of 64 students will have an $\bar{x}$ this far from the mean if $\mu = 20$. Mean weekly alcohol expenditure at State U differs from the national average, and this difference is *statistically significant* at the 0.05 (5%) level.

In Stata: `display 2*normal(-2.67)`

Intervals and hypothesis tests

Confidence intervals can be used to test two-sided hypotheses. Whenever $p > 0.05$ in a two-sided test, a 95% confidence interval for μ necessarily contains the H_0 value of μ.

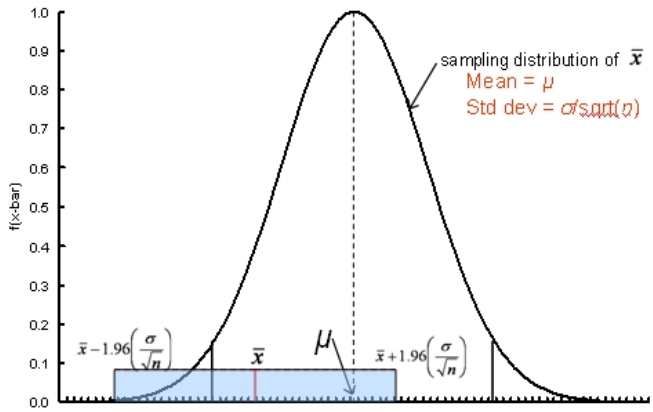


Figure: $p > 0.05$ and confidence interval contains μ_0

Intervals and hypothesis tests

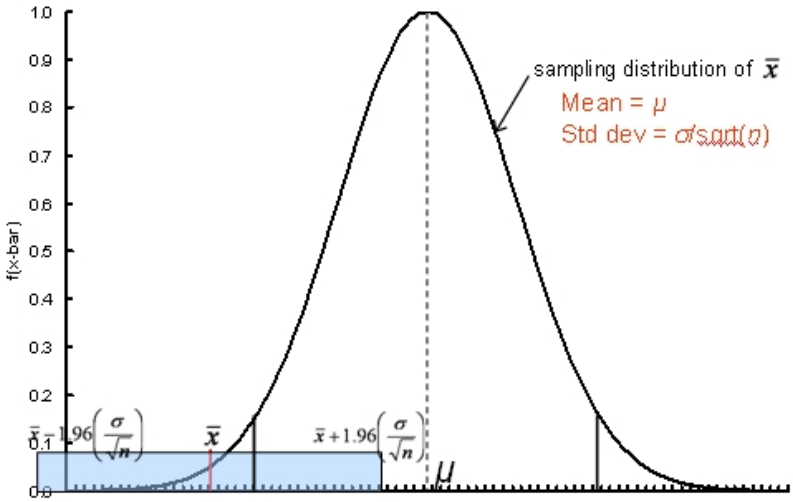


Figure: $p < 0.05$ and confidence interval does not contain μ_0

Intervals and hypothesis tests

Suppose we constructed a 95% interval for mean alcohol expenditure μ in Example 3:

$$17 \pm 1.96\left(\frac{9}{\sqrt{64}}\right) = (14.795, 19.205)$$

Our confidence interval does not support a hypothesized mean of 20—therefore we can reject H_0. In fact, we can reject *any* μ_0 outside the interval (14.795, 19.205). We do *not* reject any μ_0 within this interval.

Womens' height, revisited

Given a sample of $n = 121$ women, $\bar{x} = 64$ and $\sigma = 3$, our 95% confidence interval for the population mean female height was (63.47, 64.53). Could average female height in the population be 63.5 inches?

$$H_0 : \mu = 63.5$$

$$H_1 : \mu \neq 63.5$$

63.5 falls within our confidence interval—thus we cannot reject this null hypothesis. The evidence is consistent with an H_0 of $\mu = 63.5$.

Womens' height, revisited

Given a sample of $n = 121$ women, $\bar{x} = 64$ and $\sigma = 3$, our 95% confidence interval for the population mean female height was (63.47, 64.53). Could average female height in the population be **63** inches?

$$H_0 : \mu = 63$$

$$H_1 : \mu \neq 63$$

63 does **not** fall within our confidence interval—thus we reject H_0. The evidence is **not** consistent with an H_0 of $\mu = 63$.

Intervals and hypothesis tests

- A 95% confidence interval allows you to conduct a two-sided hypothesis test at the $\alpha = 0.05$ significance level.
- A 99% confidence interval allows you to conduct a two-sided hypothesis test at the $\alpha = 0.01$ significance level.
- A $(1 - \alpha)$% confidence interval allows you to conduct a hypothesis two-sided test at the α significance level.

Hypothesis test for μ when σ is unknown

All of the examples thus far have assumed that σ is known. This allowed us to assume a standard normal distribution for the test statistic z used to find a p value (if n is large). Of course, σ is rarely known. When s is used in place of σ the test statistic:

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

has a t distribution with $n - 1$ degrees of freedom. The only implication of this change is that the p-values for the test come from the t-distribution.

Example 4

You are wondering if the average local teen worker is paid above the federal minimum wage of \$7.25/hour. You take a random sample of 20 workers aged 16-19 and calculate a mean hourly wage of $\bar{x} = 7.75$ and sample standard deviation of $s = 1.20$. Your hypotheses are:

$$H_0 : \mu = \$7.25$$

$$H_1 : \mu > \$7.25$$

Example 4

If H_0 is true, the sample mean of \$7.75 is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{7.75 - 7.25}{1.20/\sqrt{20}} = 1.866$$

standard errors above the population mean. Using the t-distribution with 19 degrees of freedom, $p = P(t > 1.866) = 0.039$. We can reject H_0 at the 5% significance level (although not at the 1% level).

In Stata: `display ttail(19, 1.866)`

Hypothesis test for π

Hypothesis tests for population proportions π are conducted in a manner analagous to the tests above. For example, suppose a state's goal is to exceed last year's proficiency rate of 75% in 4th grade mathematics. The state education department tests a random sample of 100 4th graders and finds that 77.5% are proficient in math ($\hat{\pi} = 0.775$). It has decided on a significance level of $\alpha = 0.05$. The null and alternative hypotheses are:

$$H_0 : \pi = 0.75$$

$$H_1 : \pi > 0.75$$

Hypothesis test for π

The test statistic is the number of standard errors the realized value $\hat{\pi}$ is from the assumed mean under H_0. Importantly, the standard error in the denominator is calculated using π_0 (the population proportion under H_0), rather than $\hat{\pi}$, since we care about the distribution of the test statistic when H_0 is true:

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.775 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{100}}} = 0.577$$

Using the standard normal table, the probability of obtaining a test statistic this far above the mean when H_0 is true is $p = P(z > 0.577) = 0.282$. Since this value is not that unlikely (and $p > \alpha$), we cannot reject the null hypothesis that 4th grade proficiency remains at last year's rate of 75%.

In Stata: `display 1-normal(0.577)`

Statistical tests for μ or π using Stata

One can use the `mean` command in Stata to generate a $(1-\alpha)\%$ confidence interval, and then use this interval to conduct a hypothesis test with a α significance level. For example:

```
. mean achmat08

Mean estimation                   Number of obs    =       500
```

	Mean	Std. Err.	[95% Conf.	Interval]
achmat08	56.59102	.4176799	55.77039	57.41165

Statistical tests for μ or π using Stata

Alternatively, the `ttest` command will conduct hypothesis tests for the specified null hypothesis. For example:

```
. ttest achmat08==57.5

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
achmat08 |     500    56.59102    .4176799    9.339608    55.77039    57.41165
------------------------------------------------------------------------------
    mean = mean(achmat08)                                         t =  -2.1763
Ho: mean = 57.5                                 degrees of freedom =       499

   Ha: mean < 57.5               Ha: mean != 57.5                 Ha: mean > 57.5
 Pr(T < t) = 0.0150         Pr(|T| > |t|) = 0.0300          Pr(T > t) = 0.9850
```


Next time

- What to report? Statistical significance, *p*-values, confidence intervals (Romer, 2020; Greenland et al., 2016)
- Practical vs. statistical significance
- Effect size (Kraft, 2020; Hill et al., 2008)
- Type I vs. Type II errors
- Power of the test
- Power analysis