

BIOS 600: Principles of Statistical Inference

Hypothesis Testing

Fall 2012

Reading

A dirty dozen: twelve p-value misconceptions by Steve Goodman.

One-Sided Tests

Please enter in poll everywhere whether the paper you looked up reported two-sided or one-sided results!

Steps in Hypothesis Testing about the Mean

1. Hypothesize a value (μ_0) and set up H_0 and H_A

Steps in Hypothesis Testing about the Mean

1. Hypothesize a value (μ_0) and set up H_0 and H_A
2. Take a random sample of size n and calculate summary statistics (e.g., sample mean and variance)

Steps in Hypothesis Testing about the Mean

1. Hypothesize a value (μ_0) and set up H_0 and H_A
2. Take a random sample of size n and calculate summary statistics (e.g., sample mean and variance)
3. Is it likely that the sample came from a population with mean μ_0 (with $\alpha = 0.05$)

Steps in Hypothesis Testing about the Mean

1. Hypothesize a value (μ_0) and set up H_0 and H_A
2. Take a random sample of size n and calculate summary statistics (e.g., sample mean and variance)
3. Is it **likely** that the sample came from a population with mean μ_0 (with $\alpha = 0.05$)
4. **Draw conclusions** (and **celebrate?**)

Null and Alternative Hypotheses about the Mean

We set up the hypotheses to cover *all* the possibilities for μ and consider three possibilities.

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$
One-sided	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

Two-Sided Tests of Hypotheses

To conduct the hypothesis test, we use what we learned about the sampling distribution of the sample mean \bar{X} . If the underlying population is normally distributed (or n is pretty large), then the random variable

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

has a t_{n-1} distribution, and we can use a *t-test* of our hypothesis by comparing our statistic t to Table A.4 in P&G or by using software (Stata command `ttest`).

Case Study: Ultra Low Dose Contraception

Recall our ultra low dose contraception example. We sampled 50 subjects and obtained $\bar{x} = 0.017$ and $s = 0.008$.

```
. ttesti 50 .017 .008 .02
```

One-sample t test

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
x	50	.017	.0011314	.008	.0147264	.0192736

```

      mean = mean(x)
Ho: mean = .02
      t = -2.6517
      degrees of freedom = 49

      Ha: mean < .02
      Pr(T < t) = 0.0054

      Ha: mean != .02
      Pr(|T| > |t|) = 0.0108

      Ha: mean > .02
      Pr(T > t) = 0.9946

```

Case Study: Ultra Low Dose Contraception

What should we report in a manuscript?

- ▶ Sample size (50)

Case Study: Ultra Low Dose Contraception

What should we report in a manuscript?

- ▶ Sample size (50)
- ▶ Estimated mean

Case Study: Ultra Low Dose Contraception

What should we report in a manuscript?

- ▶ Sample size (50)
- ▶ Estimated mean
- ▶ 95% interval estimate

Case Study: Ultra Low Dose Contraception

What should we report in a manuscript?

- ▶ Sample size (50)
- ▶ Estimated mean
- ▶ 95% interval estimate
- ▶ t-statistic and degrees of freedom

Case Study: Ultra Low Dose Contraception

What should we report in a manuscript?

- ▶ Sample size (50)
- ▶ Estimated mean
- ▶ 95% interval estimate
- ▶ t-statistic and degrees of freedom
- ▶ p-value

Case Study: Ultra Low Dose Contraception

What should we report in a manuscript?

- ▶ Sample size (50)
- ▶ Estimated mean
- ▶ 95% interval estimate
- ▶ t-statistic and degrees of freedom
- ▶ p-value
- ▶ Conclusion from statistical test

Case Study: Ultra Low Dose Contraception

What should we report in a manuscript?

- ▶ Sample size (50)
- ▶ Estimated mean
- ▶ 95% interval estimate
- ▶ t-statistic and degrees of freedom
- ▶ p-value
- ▶ Conclusion from statistical test
- ▶ Thoughtful interpretation based on study and subject matter

Two-Sided Tests of Hypotheses

If σ happens to be known (EXTREMELY RARE), we can calculate

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

and use a *z-test* instead, relying on Table A.3 of P&G or software.

When n is very large (say over 200), the normal and t distributions look very similar, and often the z -test is used instead of the t -test in moderate (30+) to large samples for this reason.

Carrying out the test

You will almost always use software to calculate the test statistic and p-value, but it is important to understand the big concepts behind the scenes. First, think about our test statistic

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

- ▶ $\bar{X} - \mu_0$ makes sense, because we want to look at how far our sample mean is from the hypothesized population mean

Carrying out the test

You will almost always use software to calculate the test statistic and p-value, but it is important to understand the big concepts behind the scenes. First, think about our test statistic

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

- ▶ $\bar{X} - \mu_0$ makes sense, because we want to look at how far our sample mean is from the hypothesized population mean
- ▶ Whether $\bar{X} - \mu_0$ is big depends on the variance (and standard deviation). For example, a difference of $\bar{X} - \mu_0 = 1$ is a small difference if we are looking at weight in g but huge for height in m. This is why we standardize the difference by dividing by the estimated SD of the mean, so t is a measure of how many SDs apart μ_0 and \bar{X} are from each other

Carrying out the test

You will almost always use software to calculate the test statistic and p-value, but it is important to understand the big concepts behind the scenes. Think about our test statistic

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

- ▶ When our test statistic t is large in absolute value and our data are approximately normal (or n is large), our data are not consistent with data from a population with mean μ_0

Carrying out the test

You will almost always use software to calculate the test statistic and p-value, but it is important to understand the big concepts behind the scenes. Think about our test statistic

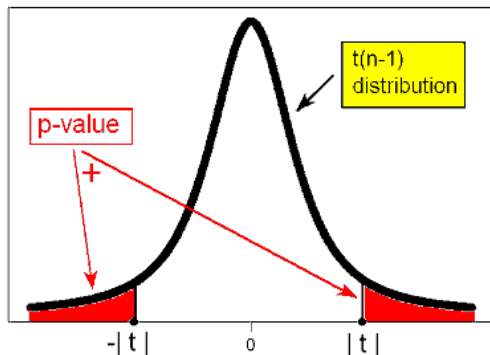
$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

- ▶ When our test statistic t is large in absolute value and our data are approximately normal (or n is large), our data are not consistent with data from a population with mean μ_0
- ▶ When our test statistic t is small in absolute value (and our data are approximately normal or n is large), then our data do not refute the null hypothesis.

Getting the p-value: two-sided test

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$
One-sided	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

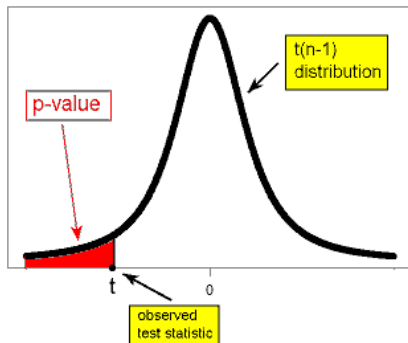
For a **two-sided** test, the p-value is the probability of seeing a t statistic with absolute value as big as, or larger than, what we saw in our data.



Getting the p-value: one-sided test

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$
One-sided	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

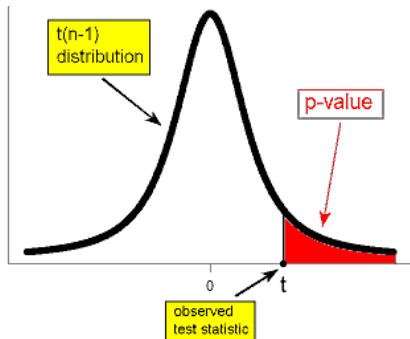
For the **one-sided** test with $H_0 : \mu \geq \mu_0$, the p-value is the probability of seeing a t statistic as small as, or smaller than, what we saw in our data.



Getting the p-value: one-sided test

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$
One-sided	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

For the **one-sided** test with $H_0 : \mu \leq \mu_0$, the p-value is the probability of seeing a t statistic as large as, or larger than, what we saw in our data.



Warning about One-Sided Tests

As P&G note, the choice between a one-sided and a two-sided test can be highly controversial because a one-sided test will have a p-value that is half that of the corresponding two-sided test, due to the symmetry of the t distribution. Sometimes a scientist will (unethically) choose a one-sided test on nonscientific grounds. To protect against this, some journal editors are extremely reluctant to publish studies using one-sided tests.

P-value Defined

The *p-value* is the probability, under the assumption of no effect or no difference (the null hypothesis, H_0), of obtaining a result equal to or more extreme than what was actually observed.

P-value Misused

Most researchers think a p-value of 0.05 means the null hypothesis has a probability of only 5%.
This is wrong!



P-value Misused

Most researchers will see $p = 0.05$ and state that there is a 95% or greater chance that the null hypothesis is incorrect. This is wrong!



P-value Misused

A p-value is calculated *assuming* that H_0 is true. It cannot be used to tell us how likely it is that assumption is correct.



Hypothesis Testing Paradigm

The paradigm of hypothesis testing tries to limit the overall number of incorrect decisions over the long run. While we of course want to know if any one study is showing us something real or a type I or type II error, the paradigm does not give us the tools to determine this.

Example: Chronic Fatigue Syndrome

A randomized controlled trial of hydrocortisone treatment for chronic fatigue syndrome showed a treatment effect with $p=0.06$. The discussion section began, "... hydrocortisone treatment was associated with an improvement in symptoms.... This is the first such study ... to demonstrate improvement [of chronic fatigue syndrome] with a drug treatment."

How do you get from $p=0.06$ to their conclusion?

Example: Chronic Fatigue Syndrome

Other evidence from paper:

- ▶ Magnitude of the effect was small
- ▶ No other endpoints showed improvement
- ▶ No other supporting studies
- ▶ Weak support for proposed biological mechanism

Authors recommended *against* using hydrocortisone because of a risk for adrenal suppression that could outweigh the small beneficial effect.

Do you agree with their conclusion that hydrocortisone is beneficial for treating chronic fatigue syndrome?

What DO You Do with a P-value, Then?

Ideally, you evaluate a p-value in light of other information, such as a proposed biological mechanism, supporting evidence in the literature, size and quality of the study, and size of the purported effect.

I'm Nervous about P-values. What Alternatives are There?

- ▶ Many researchers prefer confidence intervals, which represent the range of effects that are “compatible with the data”

I'm Nervous about P-values. What Alternatives are There?

- ▶ Many researchers prefer confidence intervals, which represent the range of effects that are “compatible with the data”
- ▶ They are sometimes used as a hypothesis test (i.e., reporting results as significant when confidence interval does not include null value) and share many of the properties of p-values we have discussed

I'm Nervous about P-values. What Alternatives are There?

- ▶ Many researchers prefer confidence intervals, which represent the range of effects that are “compatible with the data”
- ▶ They are sometimes used as a hypothesis test (i.e., reporting results as significant when confidence interval does not include null value) and share many of the properties of p-values we have discussed
- ▶ Like p-values, confidence intervals do not offer a mechanism to unite external evidence with that provided by the study at hand

I'm Nervous about P-values. What Alternatives are There?

- ▶ Many researchers prefer confidence intervals, which represent the range of effects that are “compatible with the data”
- ▶ They are sometimes used as a hypothesis test (i.e., reporting results as significant when confidence interval does not include null value) and share many of the properties of p-values we have discussed
- ▶ Like p-values, confidence intervals do not offer a mechanism to unite external evidence with that provided by the study at hand
- ▶ *Bayesian methods* can be used to incorporate current study results with prior knowledge in order to provide a probability a hypothesis is true

Hypothesis Tests and Confidence Intervals

We can also examine a confidence interval to decide whether a proposed value for the population mean is reasonable.

Suppose we are testing $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$ using a significance level of $\alpha = 0.05$. An alternative way to do this is to construct a 95% confidence interval for μ and use the following decision rule.

- ▶ If μ_0 falls outside the confidence interval, reject H_0
- ▶ If μ_0 falls inside the confidence interval, do not reject H_0

Case Study: Ultra Low Dose Contraception

Recall our ultra low dose contraception example. We sampled 50 subjects and obtained $\bar{x} = 0.017$ and $s = 0.008$. We construct a 95% confidence interval as

$$0.017 \pm t_{49,0.025} \frac{.008}{\sqrt{50}}$$

$$0.017 \pm 0.002$$

or

$$(0.015, 0.019)$$

which does not contain the target value of 0.02, so we conclude that the sample is not likely to have come from a population with mean 0.02 μg .

Case Study: Effects of Tobacco Use on Bone-Mineral Density

A study of twin pairs, one of whom was a heavy smoker and the other who was a lighter smoker, was conducted. Bone-mineral density (BMD) of the lumbar spine was measured with the relative difference between twins (heavier-lighter smoker) recorded. In a sample of 41 twin pairs, the sample mean difference was -5 with a standard deviation of 12.8. Is the difference in BMD significant?

Stata: `ttesti 41 -5 12.8 0`

Case Study: Age of BIOS 600 Students

Using the data `age.dta`, test whether the average age of BIOS 600 students is 24 or not.

Stata: `ttest age=24`

Case Study: Parasitic Density in Malaria

Nigerian researchers reported on the parasite density in a population of 136 patients. They reported a mean of 600 parasites per μl and a standard deviation of 300 parasites per μl and would like to know whether this mean is consistent with a mean of 550 parasites per μl obtained from a large population-based study in Asia.



Case Study: Parasitic Density in Malaria

Parasite density is not normally distributed but is heavily skewed.
How might this affect the validity of your test?

Assignment for Next Time

Think about a possible study you would like to conduct in which the hypothesis test of interest concerns a mean in a single sample of data. If you particularly like your study idea, e-mail me before Monday, and I can incorporate it in advance when we talk about how to determine sample size and power of a study.