

# BIOS 600: Principles of Statistical Inference

## Hypothesis Testing

Fall 2012

# Reading

- ▶ Pagano and Gauvreau, Chapter 10
- ▶ Put It to the Test

# Hypothesis Testing Framework

Suppose a case of suspected cheating is brought to the UNC Honor Court. There are two opposing claims.

- ▶ **Student:** I did not cheat on the exam.
- ▶ **Professor:** The student did cheat on the exam.

The Honor Court assumes students are innocent until proven guilty. The professor must provide evidence to support her claim. She explains that she had two different versions of the exam and that the student on three separate problems used numbers from the *other* version of the exam. (oops)

# Hypothesis Testing Framework

The honor court members agree that this would be *extremely unlikely* if it were true that the student did not cheat. They agree that the professor's evidence is very strong and conclude that the student did cheat on the exam. (For the likely sanctions handed down to the student see [the Honor System Website](#).)

# Hypothesis Testing Framework

**Statistical hypothesis testing:** assessing evidence provided by the data in favor of, or against, some claim about the population.

Steps in hypothesis testing:

- ▶ Start with two claims about the population (often about the value of a population parameter or about some potential association between two variables in the population). Call them claim 1 and claim 2.
- ▶ Choose a sample, collect data, and summarize data
- ▶ Figure out how likely it is to see data like what we got, if claim 1 is true.
- ▶ If our data would have been extremely unlikely if claim 1 were true, then we reject claim 1 in favor of claim 2. Otherwise, we cannot reject claim 1.

## Example: Ultra Low Dose Contraception

A certain ultra-low dose oral contraceptive pill is supposed to contain  $0.02 \mu\text{g}$  of estrogen. If the dose is higher, the woman may risk side effects she is trying to avoid, and if the dose is lower, she may get pregnant. The manufacturer wishes to check whether the mean concentration in a large shipment is the needed  $0.02 \mu\text{g}$  or not. A random sample of  $n = 50$  pills is tested, and the sample mean concentration is  $0.017 \mu\text{g}$  with a sample standard deviation of  $0.008 \mu\text{g}$ .

Let's go through the hypothesis testing paradigm.

## Example: Ultra Low Dose Contraception

- ▶ State the claims
  - ▶ **Claim 1:** The shipment is consistent with a population mean of  $0.02 \mu\text{g}$  estrogen.
  - ▶ **Claim 2:** The shipment is not consistent with a population mean of  $0.02 \mu\text{g}$  estrogen.
- ▶ Choose sample, collect, and analyze data
  - ▶ 50 pills were sampled with a sample mean  $\bar{x} = 0.017$  and sample standard deviation  $s = 0.008$
- ▶ Assess likelihood of our data if Claim 1 is true
  - ▶ We'll learn how to do this shortly, but for now assume the probability of getting a result like ours (or even more extreme) is just 0.01 if Claim 1 is true.
- ▶ Conclusion: that's pretty unlikely. We'll reject Claim 1 and assume that something is wrong – the manufacturing procedure is not consistent with one that produces pills at the required  $0.02 \mu\text{g}$  dose.

## Example: Ultra Low Dose Contraception

Now, suppose the probability was relatively large, say 0.50 instead of 0.01. In this case, we would *fail to reject* Claim 1 and state that we do not have evidence to disprove Claim 1. We would not say that evidence leads us to accept Claim 1, though. This is a subtle but important difference.

The concept is the same as that in the US judicial system – we might find someone “not guilty,” but we would not proclaim them “innocent.”





# Hypothesis Testing Steps (Again!)

- ▶ State Claim 1 and Claim 2. Claim 1 states “nothing unusual is happening” and Claim 2 challenges it.
- ▶ Collect relevant data and summarize it.
- ▶ Assess how surprising it would be to see data like that *if Claim 1 is really true*.
- ▶ Draw conclusions.

# Hypothesis testing details: Step 1

Step 1 is to state Claim 1 and Claim 2.

- ▶ Claim 1 is called the *null hypothesis* or  $H_0$ . It states that there is no change from the status quo or no relationship.
- ▶ Claim 2 is called the *alternative hypothesis* or  $H_A$ . It states that there is something going on or some relationship. Usually,  $H_A$  is what we want to check or what we really think is going on.

# Hypothesis testing details: Step 1

Step 1 is to state Claim 1 and Claim 2.

- ▶ Ultra low dose contraception
  - ▶  $H_0$ : The pills are consistent with a population that has mean  $0.2 \mu\text{g}$  estrogen.
  - ▶  $H_A$ : The pills are not consistent with a population that has mean  $0.2 \mu\text{g}$  estrogen.
- ▶ IQ of BIOS 600 students
  - ▶  $H_0$ : BIOS 600 students are not smarter than the average bear



- ▶  $H_A$ : BIOS 600 students are smarter than the average bear

# Practice with Step 1

What are  $H_0$  and  $H_A$  in each case?

- ▶ Forbes magazine reports that the average starting salary of students with business majors is \$48,000. A researcher would like to know whether public health majors have different starting salaries.
- ▶ The CDC estimates that 22% of US adults smoke. A researcher suggests that the smoking rate among Gillings School of Global Health graduate students is lower than the national average.

# More Practice with Step 1

What are  $H_0$  and  $H_A$  in each case?

- ▶ Researchers would like to know whether managed care affects diffusion of psychotropic medications.
- ▶ Researchers would like to know whether a new intervention for informing children in developing countries of their HIV status is associated with different mental health quality of life.
- ▶ An environmental health scientist would like to know whether manganese concentrations in the brains of welders differ from those in the general population.

## Hypothesis testing details: Step 2

Step 2 is to take a sample and summarize the data. We will return to this topic, but the choice of summary statistic will depend on the type of data (e.g., categorical or continuous) as well as the distribution of the data.

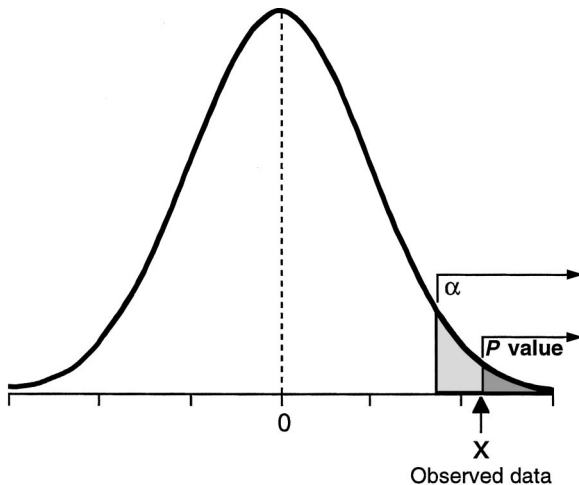
## Hypothesis testing details: Step 3

Step 3 involves assessing the evidence in our data by calculating the probability it is to get data like ours, or more extreme than ours, if  $H_0$  is actually true. This probability will be based on our data summary statistic from Step 2. This probability is often called a *p-value*.

Note that the p-value is a conditional probability,  $Pr(\text{ours or more extreme data} \mid H_0 \text{ true})$ , and it conditions on our null hypothesis being true. It does not provide direct information on  $Pr(H_0 \text{ true})$ .

## Hypothesis testing details: Step 3

Illustration of p-value from one-sided hypothesis test





## Hypothesis testing details: Step 4

Step 4 is drawing conclusions based on our test results. Generally, we reject Claim 1 when the likelihood of seeing our data (or more extreme data) when Claim 1 is true would be relatively small.

Often, we consider a cutpoint (also called a significance level or  $\alpha$  level) of  $\alpha = 0.05$  as defining the p-value as “small.” Using 0.05 as a cutpoint means that when the null hypothesis is true, we would expect to make the wrong decision only 5% of the time. When the p-value  $< \alpha$  we say the results are “statistically significant.”

When the p-value is  $> \alpha$  then we say we have insufficient evidence to reject Claim 1. Sometimes people interpret p-values in the rough range of  $\alpha - 2\alpha$  as marginally significant (so for  $\alpha = 0.05$  that range would be  $0.05 - 0.10$ ).

## Hypothesis testing details: Step 4

Common sense should also be used in drawing conclusions. For example, if there is a very strong body of evidence in favor of the alternative hypothesis and you see  $p = 0.06$ , you wouldn't necessarily use your results to refute all the prior work in an area. This is one reason confidence intervals are often used in place of hypothesis tests.

## Example: Workplace Equity

An employer subscribes to an “equal opportunity” policy, stating it hires employees with children and employees without children equally often in managerial positions. Suppose the current hiring pool is half applicants with children and half applicants without children, with both groups equally well-qualified. The data you have are the last three hires, all of which were applicants without children. How do you carry out a hypothesis test?

## Trial by Jury Analogy

A man is on trial. Did he commit the crime? Evidence is presented as part of the trial by jury.

Jury	Person on Trial	
	Innocent	Guilty
Not Guilty	✓	X
Guilty	X	✓

During the trial, we assume the man is innocent unless he can be proven guilty. The right decisions are finding an innocent man “not guilty” and convicting a guilty man. However, we could make two mistakes: we could wrongly convict an innocent man, or we could declare a guilty man “not guilty.”

Let's translate this into a hypothesis testing framework. Suppose we wish to test that the population mean equals some value, say  $\mu_0$ .

# Trial by Jury Analogy

Test of  $H_0 : \mu = \mu_0$

Jury	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not Guilty	✓	X
Guilty	X	✓

# Trial by Jury Analogy

Test of  $H_0 : \mu = \mu_0$

Analyst	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not reject $H_0$	✓	X
Reject $H_0$	X	✓

# Trial by Jury Analogy

Test of  $H_0 : \mu = \mu_0$

Analyst	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not reject $H_0$	✓	Type II Error
Reject $H_0$	Type I Error	✓

Type I error: rejecting  $H_0$  when it is really true

Type II error: not rejecting  $H_0$  when it is false

Notice these are also conditional probabilities and can be linked to the false positive and false negative rates of our test procedure if we think of it as a screening test.

## Possible errors

- ▶  $\alpha$  is the probability of making a Type I error. These errors, which involve incorrectly changing the status quo, are typically viewed as more severe than Type II errors. The most common value is  $\alpha = 0.05$ , though sometimes smaller values are chosen when multiple tests are being carried out.
- ▶ The probability of making a Type II error is  $\beta$ , which is related to the *power* of the test, given by  $1 - \beta$ .  $\beta$  can also vary but is typically larger, say 0.20 (80% power) or 0.10 (90% power)
- ▶ The *power* of the test is the probability of rejecting  $H_0$  when  $H_0$  is indeed false. (sensitivity)

Let's examine these errors using a cool [applet from the American Statistical Association](#). Type I errors in the applet are dark blue, and type II errors are dark red.



# Steps in Hypothesis Testing about the Mean

1. Hypothesize a value ( $\mu_0$ ) 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  $\mu_0$  (with  $\alpha = 0.05$ )
4. Draw conclusions

# Null and Alternative Hypotheses about the Mean

We set up the hypotheses to cover *all* the possibilities for  $\mu$  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	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

One-sided tests are pretty rare. How rare? You tell me!

Homework: randomly pick a research paper in the most recent issue of the best journal in your field or subfield that carries out at least one hypothesis test. Determine whether the main hypothesis test is one- or two-sided. We'll use our data to estimate the proportion of one-sided tests in the literature!

# Are One-Sided Tests Underutilized?

Usually, you have a pretty good idea what will happen (you do have to get money using some justification!). In this case, why not always do one-sided tests?

# Are One-Sided Tests Underutilized?

Example: CAST (Cardiac Arrhythmia Suppression Trial)

- ▶ New generation of antiarrhythmic agents strongly believed to have fewer side effects with much greater efficacy
- ▶ Due to strength of this belief, one-sided test was designed
- ▶ Recruitment was difficult because many physicians refused to randomize their studies when chance of NOT getting the new drugs was 50%
- ▶ New generation drugs associated with 4x the mortality as status quo
- ▶ P-value was 0.0003 in the wrong direction (i.e., new drugs worse)
- ▶ Fortunately the Data Safety and Monitoring Board stopped the trial quickly

## Reading for Next Time

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