Notes 8: Hypothesis Testing

Julio Garín
Department of Economics

Statistics for Economics

Spring 2012

Introduction

- Why we conduct surveys?
  - We want to estimate a previously unknown characteristic of the population:
    - Fraction of teens that smoke.
    - Fraction of adults unemployed.
    - Median family income.
- Descriptive statistics could be:
  1. Answers in and of themselves.
  2. Intermediate products.
- Is to the last point that we can use statistics to test certain "hypotheses".

Statistical Hypothesis Testing

- The purpose is to test the viability of the *null hypothesis* in the light of experimental data.
- How can we generate hypotheses?
  1. Inductively.
  2. Deductively.
- How the process works?
  - The researcher starts by specifying the *null hypothesis*.
  - Popperian approach: a good hypothesis is one that can falsified.
  - The experiment then can either "reject" or "not reject" the null hypothesis.
- If we reject the null: "statistically significant difference".
- If we cannot reject the null: "statistically insignificant difference".

Specifying Null and Alternative Hypotheses

Specifying a Null Hypothesis: The FDA Approval Process

- New drugs must be approved for use by the Food and Drug Administration (FDA).
- According to Federal law, drugs must be demonstrated to be both safe and effective.
- The FDA starts from the position that the drug does not work.
  - It is up to the manufacturer to show otherwise.

1. μ_1 = expected mortality rate from a particular disease in the absence of the new drug.
2. μ_2 = expected mortality rate when people take the new drug.

- Let $d = \mu_2 - \mu_1$.
- What if $d < 0$?
- Therefore the null hypothesis is

$$H_0 : d = 0$$

Specifying Null and Alternative Hypotheses: Micro Theory

- What is the first lesson from Introductory Economics?
- Workhouse theory of the discipline: theory of demand.
  - What is it?
- Demand theory produces a "testable" prediction between prices and quantity.
  - Null hypothesis (H_0): quantity consumed is independent of price.
  - Alternative hypothesis (H_a): quantity is negatively related to price.

Forms for null and Alternative Hypotheses and Errors

- We will work mainly with four different possibilities for the hypotheses:

$$H_0 : \mu \geq \gamma \qquad H_0 : \mu \leq \gamma \qquad H_0 : \mu = \gamma \qquad H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu < \gamma \qquad H_a : \mu > \gamma \qquad H_a : \mu \neq \gamma \qquad H_a : \mu_1 - \mu_2 \neq 0$$

- What would be some examples?

Hypothesis Test for the Mean.

Testing Hypotheses with Confidence Intervals

- Recall what a $\bar{x}$ is and what it implies for μ.
  - Recall what was the 95% CI for μ.
  - Another way: what are the likely values of μ such that there is only a 5% chance that the true value of μ lies outside this interval?
- Suppose that we have a null hypothesis about the value of μ.
- Two possibilities:
  - The CI contains that value: we cannot reject the null.
  - The CI does not contain the hypothesized value: we can, *with a reasonable degree of statistical certainty*, reject the null.

Example: Two Tailed Test

Suppose that we know with certainty that 25% of US graduate students smoke. In a survey of 29 ND graduate students, 20.68% say they smoke (6 students).

1. Are ND graduate students different from the average graduate students?

From a large survey, we obtain from a sub-sample of 115 teen (16-19) workers the following values:

$$\bar{x} = 8.34 \quad s = 3.20$$

We know that the federal minimum is \$7.15.

1. Are teens working at the minimum wage?

Hypothesis Test With Difference in Means.

Hypothesis Test With Difference in Means

- In many cases, the hypothesis we want to examine requires information from two samples.
  - Compare means across two groups.
  - Compare means over time.
- Recall: Pooled sample variance:

Under the assumption that the two population standard deviations are equal ($\sigma_1 = \sigma_2 = \sigma$), the two sample standard deviation are combined to provide the following *pooled sample variance for the difference between two means*:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Difference in Proportions Over Time I: Drinking Rates

Surveys in 1993 and 2001 show that drinking rates amongst college students have fallen from 84 to 81 percent.
In 1993 we had

$$\bar{p}_1 = 0.84 \quad n_1 = 1,100$$

In 2001 we obtained

$$\bar{p}_2 = 0.81 \quad n_2 = 1,200$$

1. Is this a statistically reliable result? In other words, are we confident that drinking rates have declined, or is there a chance that, because of sampling error, we may have had no change in those rates?

Difference in Proportions Over Time II (Practice): Obesity Rates

Surveys in 1988 and 1999 show that obesity rates amongst kids between 6 and 18 have increased from 11.3 to 15.3 percent. The size of the sample in 1988 was 1,011 kids while in 2001 it was 456.

1. What is a 95% CI on the change in obesity over time? What can you say about the change obesity over time?
2. What is a 99% CI? Does your previous conclusion change?

Testing Hypotheses With a t-test.

Testing Hypotheses With a t-test

- Recall our definition of the 95% CI.
  - How is $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ distributed?
  - What is the expected value of $\frac{\bar{x}-\mu}{s/\sqrt{n}}$?
- Suppose there is a null hypothesis about some value of μ.
  - What would happen if the hypothesized value for μ is correct?
  - What a large of t would imply?

The test statistics for hypothesis tests about a population mean with σ unknown is

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

and this value is typically called the t-statistic.

- As t increases in absolute value, the chances that we would generate a large value for t by chance decline, so we would be less likely to not reject the null hypothesis.
- What is the critical value for t that would make us reject H_0?

How we proceed?

- We construct the t-statistic.
- If we were to draw t at random from a distribution with $n-1$ degrees of freedom, 95% of the time we would generate a value between $\left[-t_{0.025,n-1}, t_{0.025,n-1}\right]$.
- Therefore, if $|t| > t_{0.025,n-1}$, there is only a 5% chance we would draw a number that large at random.
- Therefore, when our t-statistic is larger in absolute value than the $t_{0.025,n-1}$ we conclude the estimate $\bar{x}$ is not centered on our hypothesized value for μ and we reject the null.

Examples I and II

Suppose that we know with certainty that 35% of US graduate students smoke. In a survey of 29 ND graduate students, 20.68% say they smoke (6 students).

1. Are ND graduate students different from the average graduate students?

Surveys in 1993 and 2001 show that drinking rates amongst college students have fallen from 84 to 81 percent.
In 1993 we had

$$\bar{p}_1 = 0.84 \quad n_1 = 1,100$$

In 2001 we obtained

$$\bar{p}_2 = 0.81 \quad n_2 = 1,200$$

1. Has the drinking rate decreased over time?
2. What if we change the level of significance to $\alpha = 0.10$?

Example III: Binge Drinking (Practice)

We have a survey of college students for the year 1999. We define binge drinking:

- Males: 5 or more drinks on one occasion in past 2 weeks.
- Females: 4 or more drinks on one occasion in past 2 weeks.

We obtain the following rates:

- Males: 50%. Females: 40%.
- Freshmen: 42% ($n = 800$). Seniors: 45% ($n = 700$).

1. Test hypothesis that there is no difference in rates across classes.

Example IV: One-Tailed Test

Kardashian Elementary School has 300 students. The principal of the school, Mr. Sorrentino , thinks that the average IQ of students at Kardashian is at least 110. To prove his point, he administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108 with a standard deviation of 10.

1. Based on these results, should Mr. Sorrentino accept or reject his original hypothesis? Assume a significance level of 0.01.

Testing Hypotheses using the p-value.

Testing Hypotheses using the p-value

- It provides with another way of testing the null hypothesis.
  - What the t-statistics is measuring?
  - What large t-statistics suggests?
  - What is the drawback of this test?
- The p-value is the probability that actual value will lie above t or below $-t$.
  - What is the chance you would draw a value "at least this large in absolute value" at random?
- A small p-value gives you confidence that you can reject the null hypothesis.
  - Why?

Examples

From a large survey, we obtain from a sub-sample of 115 teen (16-19) workers the following values:

$$\bar{x} = 8.34 \quad s = 3.20$$

We know that the federal minimum is \$7.15.

1. Use the p-value to test the null hypothesis.

In a 1999 survey of 124 students, Thorpe, Pittenger, and Reed examined self-reported cheating rates amongst college students:

- 70 males (72.9%) reported cheating.
- 54 females (63%) reported cheating.

1. Is there a statistical significant difference in cheating rates across genders?

Example: Poverty and Calcium

According to the Food and Nutrition Board of the National Academy of Sciences, the recommended daily allowance (RDA) of calcium for adults is 800 milligrams (mg). A random sample of 18 people with incomes bellow the poverty line gives an average daily calcium intake of 747.4 and from the population we know $\sigma = 188$.

1. Do the data provide sufficient evidence to conclude that the mean calcium intake of all people with incomes below the poverty level is less than the RDA of 800 mg? Use the three methods covered so far.

Errors in Prediction.

However...

Do not put your faith in what statistics say until you have carefully considered what they do not say. *William W. Watt*

Facing The Possibility of Errors

- As we saw statistical tests can be used to test theoretical hypothesis.
  - Does a new drug work?
  - Have smoking rates changed?
- **Caveat**: these are only statistical tests.
- Example:
  - Consider:

$$H_0 : d = \mu_1 - \mu_2 = 0$$

$$H_a : d = \mu_1 - \mu_2 \neq 0$$

  - What if the t-test is small?

Example I: New Drug

- New drug reduces deaths from stroke by 10%. However a clinical trial cannot reject the null hypothesis that there is no effect.
  - $\bar{x}_1 = 0.4$: death rate in control.
  - $\bar{x}_2 = 0.36$: death rate in treatment.
  - Suppose t-stat is 1.12.

  1. Can we reject the null?

- There are two possible outcomes:
  1. Drug does not work.
  2. Drug does work, but statistical model did not have enough power to detect a statistically significant impact.

Example II: Drinking Coffee While Pregnant

- Public health researcher finds that children whose mother's drink coffee have higher low birth weight rates.
  - $\bar{x}_1 = 0.07$: Low birth weight rate for moms who do not drink coffee.
  - $\bar{x}_2 = 0.08$: Low birth weight rate for moms who drink coffee.
  - Suppose t-stat is 2.13.

1. Can we reject the null?

- There are two possible alternatives:
  1. Coffee actually causes low birth weight.
  2. Statistical fluke - although we reject the null at the 95% confidence level, there is a 5% chance we could be wrong.

Type I and Type II Errors

Table: Errors and Correct Conclusions in Hypothesis Testing

		Population Condition	
		H_o **True**	H_a **True**
Conclusion	**Do not Reject** H_0	Correct Conclusion $(1-\alpha)$	Type II Error (β)
	Reject H_0	Type I Error (α)	Correct Conclusion $(1-\beta)$

- Type I: false positive.
- Type II: false negative.

What is the Probability You Will Make a 'Wrong' Decision?

- Type I error: reject the null when it is in fact true.
  - $H_0 : d = 0$
  - Get a large t-statistic, so reject null.
- There is a chance that, by accident, you will get a large t-stat.
- What is that chance?
  - 1-confidence level.
- Type II error: do not reject the null when it is in fact false.
  - $H_0 : d = 0$
  - Get a small t-statistic, so do not reject null.
- What is the probability this will happen?
  - β
  - $1 - \beta$ is called the *power of the test*.
  - It is a function of the sample size.

As Always, we Face a Trade-off

- What is the trade-off here?
  - Type I error vs. Type II errors.
- Suppose you are concerned about Type I errors, so you increase the size of the confidence interval, therefore increasing the the chance of Type II error.
- Suppose you are concerned about Type II errors, so you increase the size of the test, i.e. increase the chance of Type I error.

Example I: Criminal Court

- Consider a criminal court:
  - H_0 : not guilty.
  - H_a : guilty.
- What is the job of the jury?
  - What would be a Type I error?
  - What would be a Type II error?
- Decision rule: guilt beyond a reasonable doubt.
  - What it means in statistical terms?
    - Very low p-value, high confidence level!

Example II: Mammography

- Low level radiation exam to detect breast cancer growth.
  - H_0 : no breast cancer.
  - H_a : breast cancer.
  - Type I error: false positive.
  - Type II error: false negative.
- Consider the doctor's liability.
  - Suppose a Type II error happens: failed to find a tumor. Sued for malpractice.
  - Suppose a Type I error happens: detect tumor, perform surgery when none was needed. Sued for malpractice.
- For the doctor, what type of error has more 'downside' risk?

Problems

Working late at night, studying for an exam, you are so hungry so that you purchase a ham sandwich out of the vending machines. As you open the sandwich, you see that the ham is green.

1. What are the Type I and II errors associated with eating the sandwich?

Suppose you are like me and you like to walk along railroad tracks. Suddenly hear a loud noise behind you.

1. What are the Type I and II errors?
2. What are the costs of making a Type I and II error and what error do you minimize?

Using Confidence Intervals

TABLE 4—DESCRIPTIVE STATISTICS FOR LEVELS AND CHANGES IN EMPLOYMENT BY STATE, BNW DATA

	Means with standard deviations in parentheses:						Difference-in-differences New Jersey-Pennsylvania (standard error)
	New Jersey			Pennsylvania			
	Before	After	Change	Before	After	Change	
Total payroll hours/35:							
1. Pooled BNW sample	17.5	17.5	−0.1	15.1	15.9	0.8	−0.85
	(5.5)	(5.9)	(3.4)	(4.0)	(5.9)	(3.5)	(0.49)
2. NW subsample	17.7	16.7	−1.0	13.4	12.4	−1.0	−0.05
	(6.1)	(6.3)	(3.3)	(3.8)	(4.9)	(3.5)	(0.61)
3. Original Berman subsample	17.1	19.3	2.1	16.9	20.4	3.4	−1.28
	(3.5)	(4.3)	(2.7)	(3.4)	(4.3)	(2.1)	(0.63)
Nonmanagement employment:							
4. Pooled BNW sample	24.8	28.4	3.6	29.0	31.3	2.2	1.39
	(6.0)	(6.8)	(3.0)	(5.5)	(6.8)	(4.7)	(1.20)

Notes: See text for description of employment variables and samples. Sample sizes are as follows. Row 1: New Jersey 163; Pennsylvania 72. Row 2: New Jersey 114; Pennsylvania 40. Row 3: New Jersey 49; Pennsylvania 23. Row 4: New Jersey 19; Pennsylvania 33.

Card and Krueger (AER, 1995).

Using t-tests

Table 1
Cross-state growth maximum likelihood estimation

Regressor	Dependent variable	
	GGSPW	*GGSP*
Constant	0.178	0.178
	(3.49)	(3.49)
Volatility	0.024	0.024
	(0.17)	(0.17)
Investment share	0.158	0.158
	(5.76)	(5.76)
Labor force growth rate	−0.751	0.249
	(−43.32)	(14.35)
Human capital	0.0012	0.0012
	(3.44)	(3.44)
Initial income	−0.023	−0.023
	(−3.43)	(−3.43)
Log of likelihood function	1915.5	1915.5

Note: Parentheses contain t-statistics. 912 observations are used in the estimation.

Dawson and Stephenson (EL, 1997).

Using t-tests

Table 1

Dependent Variable: Overall Rating

	Ordered probit		*OLS*	
	(a)	*(b)*	*(c)*	*(d)*
ln(Price)	−0.047	−0.061	−0.038	−0.048
	(0.039)**	(0.013)**	(0.038)**	(0.012)**
ln(Price)*Expert		0.171		0.138
		(0.017)**		(0.015)**
Expert		−0.558		−0.448
		(0.001)***		(0.001)***
Constant			2.297	2.337
			(0.000)***	(0.000)***
N	5986	5972	5986	5972
R^2/pseudo-R^2	0.000	0.002	0.001	0.005

Robust *p* -values in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Goldsten et. al. (JWE, 2008).

Brief Introduction to Regression Analysis.

Motivation

- We have developed a number of tests to examine the statistical relationship between variables.
- These tests demonstrate a "statistical association" between two variables.
- In many cases, researchers are only interested in identifying whether there is a statistically significant relationship between variables.
- However, in many instances, researchers want to know more than a statistical relationship they want to know something about the magnitude of the relationship.

Motivation

- Suppose you wanted to estimate the impact of more police on crime.
- Collect data on crime and the size of the police force in 100 cities.
- Find out that the relationship (correlation) is positive.
- Is this casual - do more police actually increase crime?

What is Regression Analysis all About?

- Generate a model that helps measure the impact of x (independent variable) on y (dependent variable).
- The model allows us to measure the strength of the statistical association between two variables.
- This is an accurate estimate ONLY IF we have measured the casual relationship
- In most cases, authors want to estimate the causal impact of 'x on y.'
- In very few cases do they satisfy the necessary assumptions (you'll talk about this more in econometrics).

Linear Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

vs.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- What is ϵ_i?
- Graphically.
- What is the problem then?

What is the Solution to that Problem?

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$