# Statistics Review Chapter 3, with 2.5/2.6

### EC200: Econometrics and Applications

# Learning objectives

- ▶ Understand and use key vocabulary
- ▶ Construct confidence intervals
- Conduct one and two-sided hypothesis tests
  - Using z- and t- distributions
  - Interpret *p*-values

# Statistics Review

## **1** Finite sample properties of estimators

2 Confidence intervals

3 Hypothesis testing

- Overview
- P-values

## What even *is* statistics?

Statistics: use a sample to make an inference about a population.

### Greek

- Letters like  $\beta_1$  are truth
- We estimate them with β<sub>1</sub>
  what we think truth is based on data

## Latin

- Letters like X are actual data from our sample
- We use them to make calculations, like  $\bar{X}$

 $Data \rightarrow Calculation \rightarrow Estimate \xrightarrow{\text{hopefully!}} Truth$ 

Confidence intervals 00000

# Random sampling

Simple random sampling	Definition
Method of choosing a set of observations (sample) from a population, such that each member is <b>equally likely</b> to be included. We label each of $n$ observations as $Y_1, Y_2, \ldots Y_n$	
Independent and identically distributed (i.i.d.)	Definition
When $Y_1, Y_2, \dots Y_n$ are	
<b>1</b> drawn from the same distribution <i>(identical)</i> , and	
<b>2</b> are independent (conditional = marginal distribution)	

With simple random sampling, the random variables  $Y_i$  are *i.i.d.* 

# Finite sample properties of estimators

- ► An estimator of a population parameter is a random variable that depends on sample information, whose value approximates this parameter
- ▶ A specific value of that random variable is an estimate.

### Example 1

Draw a sample of size n from a population, with parameter  $\mu$ . One useful estimator:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

 $\overline{Y}$  is an estimator, and  $\overline{y}$  is the estimate. A sampling distribution is the distribution of an estimator.

#### EC200

Chapter 3: Statistics Review

Confidence intervals 00000 Hypothesis testing 0000000 000000

# Law of large numbers

### Law of large numbers

#### Definition

If  $Y_i$ , i = 1, ..., n is i.i.d, with  $E(Y_i) = \mu_Y$  and if large outliers are unlikely (if  $var(Y_i) = \sigma_Y^2 < \infty$ ), then

$$\bar{Y} \xrightarrow{p} \mu_Y$$

That is,  $\overline{Y}$  "converges in probability" to  $\mu_Y$ . Alternatively, we can say that  $\overline{Y}$  "is consistent" for  $\mu_Y$ 

Confidence intervals 00000 Hypothesis testing 0000000 000000

## Central limit theorem



 $\begin{array}{c} {\rm Confidence\ intervals}\\ {\rm 00000} \end{array}$ 

Hypothesis testing 0000000 000000

# Central limit theorem

### Central limit theorem

#### Definition

- Let  $X_1, X_2, ..., X_n$  be a set of *n* independent random variables with identical distributions with mean  $\mu$  and variance  $\sigma^2$ , and  $\bar{X}$  is the mean of these random variables
- As n becomes large, the distribution of

$$Z = \frac{\bar{X} - \mu_X}{\sigma_{\bar{X}}}$$

approaches the standard normal distribution (is "asymptotically normal")

# Characteristics of point estimators

We evaluate how good an estimator is based on its bias and efficiency:

- Bias: Difference between the expectation of the estimator and the parameter
- Efficiency: Variance of the estimator how much it differs from the true parameter

Bias

## Bias

## Let $\hat{\theta}$ be an estimator of parameter $\theta$ :

Definition

The difference between the expectation of the estimator and the parameter

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$$

The bias of an unbiased estimator is 0.

Confidence intervals 00000 Hypothesis testing 0000000 000000

## Unbiasedness

# $\hat{\theta_1}$ is an unbiased estimator, $\hat{\theta_2}$ is biased:



# Efficiency

- Often, there are several unbiased estimators.
- ► Let  $\hat{\theta_1}$  and  $\hat{\theta_2}$  be two unbiased estimators of  $\theta$ . Then,  $\hat{\theta_1}$  is more efficient than  $\hat{\theta_2}$  if

$$Var(\hat{\theta_1}) < Var(\hat{\theta_2})$$

# Confidence limits for $\mu$

Confidence interval:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where  $z_{\alpha/2}$  is the normal distribution value for the probability of  $\alpha/2$  in each tail

If  $\sigma$  unknown, then use the t distribution instead

Confidence intervals 0 = 000

Hypothesis testing 0000000 000000

# Finding $z_{\alpha/2}$

Consider a 95% confidence interval:



# CI Example

### Example 2

A sample of 11 circuits from a large, normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.

Determine a 95% confidence interval for the true mean resistance of the population.

# CI Example

A sample of 11 circuits from a large, normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms. Find a 95% CI for the true mean resistance of the population.

**1** List what we know:

$$n = 11$$
  $\bar{x} = 2.20$   
 $\sigma = 0.35$   $\alpha = 0.05$ 

population normal

**2** List what we want to find:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

# CI Example

A sample of 11 circuits from a large, normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms. Find a 95% CI for the true mean resistance of the population.

 3 Find the right value of z<sub>α/2</sub>: α = 0.05 ⇒ z<sub>0.05/2</sub> ⇒ P(Z < z<sub>0.025</sub>) = 0.975 ⇒ z<sub>0.025</sub> = 1.96

4 Plug in remaining values:

$$95\% CI = 2.20 \pm 1.96 \frac{0.35}{\sqrt{11}}$$
$$= 2.20 \pm 0.2068$$
$$1.9932 < \mu < 2.4068$$

# Concepts of hypothesis testing

A hypothesis is a claim (assumption) about a population parameter:

- One sample: The mean monthly cell phone bill in Vermont is  $\mu = $52$ .
- ▶ Two sample: The mean monthly cell phone bill in Vermont equals the mean monthly cell phone bill in Massachusetts.

# Setting up hypotheses

- ▶ Null hypothesis  $(H_0)$  states the assumption (numerical) to be tested
- ▶ Alternative hypothesis  $(H_1)$  is the "opposite" of the null
- Determine whether there is enough evidence to reject the null hypothesis.
- Example: The average number of TV sets in U.S. homes equals three  $(H_0: \mu = 3, H_1: \mu \neq 3)$ .



## One-tail tests

In many cases, the alternative hypothesis focuses on one particular direction.

▶ Does fuel additive *increase* gas mileage?

$$\begin{array}{l} H_0: \mu \leq 10.5 \\ H_1: \mu > 10.5 \end{array}$$

Upper-tail test since alternative hypothesis focused on upper tail.

▶ Does cholesterol drug *lower* LDL levels from average of 145?

$$H_0: \mu \ge 145 \\ H_1: \mu < 145$$

Lower-tail test since alternative hypothesis focused on lower tail.

## Two-tail tests

Sometimes, we don't have a specific direction in mind.

Were average U.S. stock market returns affected by Hurricane Katrina, compared to their usual average of 4%?

$$H_0: \mu = 4$$
$$H_1: \mu \neq 4$$

Two-tailed test since we reject if stock returns are very high or very low

# Level of significance, $\alpha$

- Significance level defines the unlikely values of the sample statistic, the rejection region, if the null hypothesis is true
- Designated by  $\alpha$  (level of significance) usually  $\alpha = 0.01, 0.05, 0.10$
- Selected by researcher at beginning
- ▶ Determines the critical value of the test

# Step-by-step

- **1** Set up  $H_0$  and  $H_1$
- **2** Determine *t*-statistic:

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

**3** Compare test statistic to critical value(s) c, depends on  $\alpha$  and one vs. two-sided test

a Upper tail: Reject  $H_0$  if t > c

- **b** Lower tail: Reject  $H_0$  if t < -c
- **c** Two tailed: Reject  $H_0$  if |t| > c
- 4 Reject or do not reject  $H_0$

## Test statistics and critical values

We essentially "convert" our estimate to the t-distribution:

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

If we know  $\sigma$  or as *n* gets large, the *t* distribution converges to a standard normal (z) distribution.



### P-value

The largest significance level at which we could carry out a hypothesis test and still fail to reject the null hypotheses.

- ▶ Also called "observed level of significance"
- Smallest value of  $\alpha$  for which we can reject  $H_0$

# Example: Hypothesis test for mean

### Example 3

A phone industry manager things that customer monthly cell phone bills have increased and now average over \$52 per month.

- ▶ The company wishes to test this claim, so it surveys 150 customers.
- ▶ The average phone bill is \$53.10 per month, with a standard deviation of \$10.
- ► Test the null hypothesis that bills have not increased at the 5% level.

Hypothesis testing 000000 000000

P-values

# Example: Hypothesis test for mean

**1** Write down what we know:

$$\mu_0 = 52 \ s = 10, \ n = 150$$

• 
$$\alpha = 0.05, \, \bar{x} = 53.1$$

**2** Set up hypotheses:

• 
$$H_0: \mu \le 52$$

•  $H_1: \mu > 52 \rightarrow what manager wants to prove$ 

▶ This is an *upper* tail test

# Example: Hypothesis test for mean

- 3 Since we have a upper-tail test, we will reject if we have a t-test statistic greater than  $t_{\alpha}$ .
- **4** Decision rule: Reject  $H_0$  if  $t = \frac{\bar{x} \mu_0}{s/\sqrt{n}} > 1.96$
- **5** Reject or do not reject:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{53.1 - 52}{10/\sqrt{150}} = 1.347$$
  
DO NOT REJECT *H*<sub>0</sub>

# Calculate the p-value

- **3** Convert  $\bar{x}$  to test statistic  $\Rightarrow 1.347$
- 4 Calculate *p*-value

$$P(Z > 1.347) = 1 - F(1.35) = 1 - 0.9115$$
  
= 0.0885

**5** Do not reject, as  $\alpha = 0.05 < 0.0885 = p$ . Can reject only at significance level of 0.0885 or higher.





2 Confidence intervals

- 3 Hypothesis testing
  - Overview
  - P-values