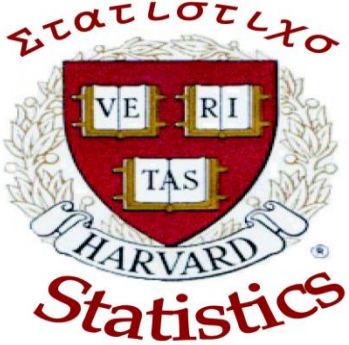



# Stat 104: Quantitative Methods for Economists

## Class 29:Exam 2 Review

# Stat 104 - Roadmap



## Data

### Intro
- Population vs. sample
- Parameters vs. statistics

### Graphs
- Dotplots and Histograms

### Descriptive Stats
- Central tendency
- Variability
- Relative standing
- 2 variable stats

### Basic Probability
- Marginal prob.
- Conditional probability
- Laws of probability
- Probability tables

## Inference

### Discrete Distributions
- Random variable
- E[X], Var[X] & Laws of expectation
- Binomial

### Continuous Distributions
- Density functions
- Uniform distribution
- Normal distribution
- T-distribution

### Central Limit Theorem
- $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
- $\hat{p} \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

### Estimation
- Point estimators
- Confidence intervals
- Levels of confidence: $1-\alpha$

## Analysis

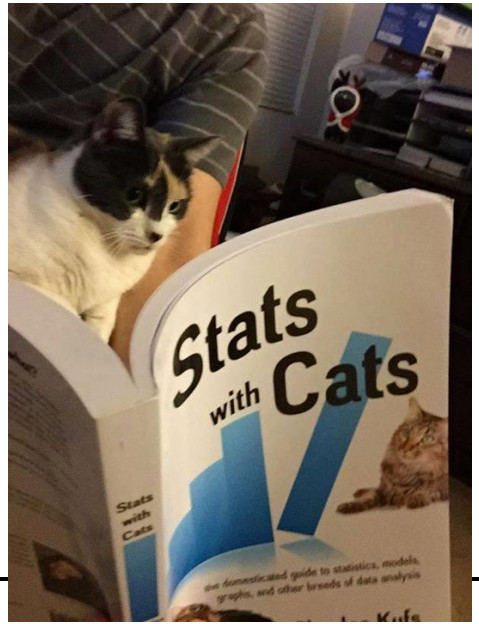
### Hypothesis Testing
- Testing parameters: mu and pi
- Left tailed, right tailed, two-tailed
- Rejection regions approach
- P-values
- Levels of significance
- Type I and Type II errors

### Two Sample Tests
- Compare mean across two populations
- Compare two proportions

### Linear Regression
- Least squares lines: scatterplot
- Multivariate regression
- Dummy variables: 0/1
- Regression Diagnostics

# Exam This Coming Monday

- What's on it? Questions



# Jointly Distributed Random Variables

- See HW5

Suppose the following table represents the joint distribution of X and Y:

| | P(x,y) | 0 | 100 | 300 |
|---|---|---|---|---|
| X | 100 | 0.05 | 0.2 | 0.23 |
| | 200 | 0.3 | 0.2 | 0.02 |

a) What are the mean and standard deviation of X?
b) What are the mean and standard deviation of Y?
c) What is the covariance between X and Y?
d) What is the correlation between X and Y?
e) Calculate E(X+Y).
f) Calculate Var(X+Y).
g) Are X and Y independent? Justify your answer.
h) What is the expected value of Y given X= 250?

# The Central Limit Theorem

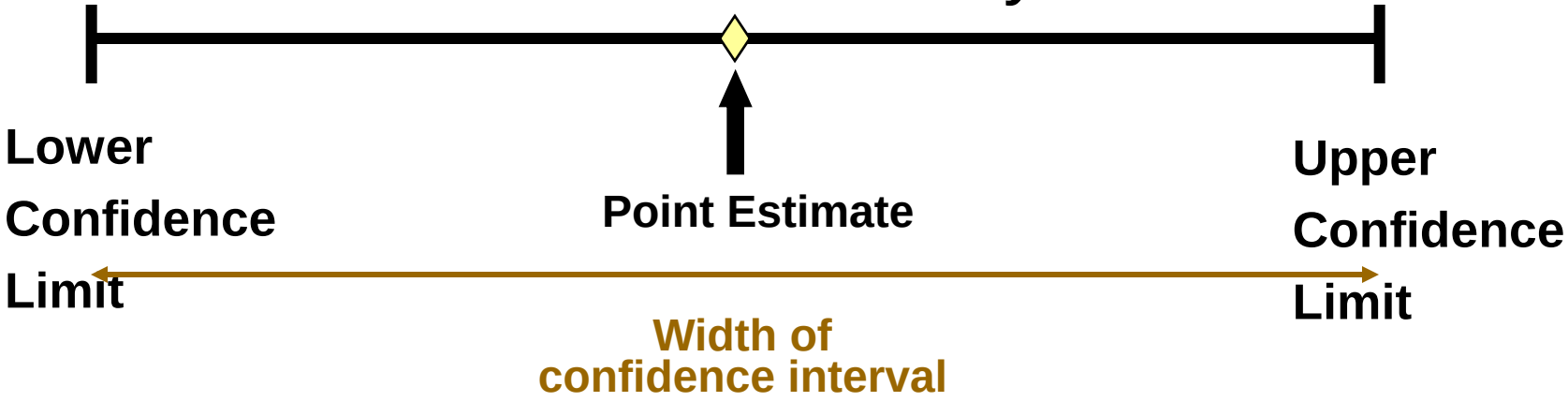
- The central limit theorem is one of the more remarkable results in statistics.
- It says that no matter what the underlying population looks like, the distribution of sample means will follow a normal distribution.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Also clt for proportions

# Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



# The Common Point Estimates

| We can estimate a Population Parameter … | | with a Sample Statistic (a Point Estimate) |
|---|---|---|
| Mean | μ | $\overline{x}$ |
| Proportion | p | $\hat{p}$ |

# One Sample Confidence Intervals

- Large sample mean

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

- Large sample proportion, sample size calc

$$\hat{p} \pm 1.96 \sqrt{\hat{p} \frac{(1-\hat{p})}{n}}$$

$$n = \frac{z^2_{\alpha/2}\, p(1-p)}{e^2}$$

# Two Sample Confidence Intervals

- Difference of two means

$$(\bar{X} - \bar{Y}) \pm 1.96\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$$

- Different of two proportions

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

# Remember the *t* Distribution

- The t distribution looks like the N(0,1) distribution except it has **fatter tails**.
- It is centered at zero and defined by its *degrees of freedom* which equal n-1.

- As the sample size n gets large, the t distribution looks like the N(0,1) distribution.

$$t_{n-1} \xrightarrow{n \to \infty} N(0,1)$$

# Hypothesis Testing

- Basic approach – set up a null hypothesis $H_0$ and alternative $H_a$ ; collect data aiming to show $H_0$ is untrue.
- Two-sided versus one-sided tests
- Reject $H_0$ if P-value < a priori level (e.g. 0.05) or use test statistic approach.
- P(Type I error) = P(reject $H_0$ | $H_0$ is true)

  P(Type II error) = P(not reject $H_0$ | $H_0$ is false)

# Decision Rules for Testing a Population Mean



$$t_{stat} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$ ← Called the test statistic

$H_0 : \mu = \mu_o$ If $|t_{stat}| > 1.96$ *reject* $H_o$

$H_a : \mu \neq \mu_o$

$H_0 : \mu = \mu_o$ If $t_{stat} < -1.64$ *reject* $H_o$

$H_a : \mu < \mu_o$

$H_0 : \mu = \mu_o$ If $t_{stat} > 1.64$ *reject* $H_o$

$H_a : \mu > \mu_o$

if n <30 use t dist for the cut-off values with df=n-1

# Decision Rules for Testing a Proportion



$$T = \frac{(\hat{p} - p_o)}{\sqrt{p_o(1-p_o)/n}}$$

$H_0 : p = p_o$ If $|T| > 1.96$ *reject* $H_o$

$H_a : p \neq p_o$

$H_0 : p = p_o$ If $T < -1.64$ *reject* $H_o$

$H_a : p < p_o$

$H_0 : p = p_o$ If $T > 1.64$ *reject* $H_o$

$H_a : p > p_o$

# Two Sample Hypothesis Tests

- Means

$H_0 : \mu_1 = \mu_2$
$H_a : \mu_1 \neq \mu_2$

$H_0 : \mu_1 = \mu_2$
$H_a : \mu_1 < \mu_2$

$H_0 : \mu_1 = \mu_2$
$H_a : \mu_1 > \mu_2$

- Proportions

$H_0 : p_1 = p_2$
$H_a : p_1 \neq p_2$

$H_0 : p_1 = p_2$
$H_a : p_1 < p_2$

$H_0 : p_1 = p_2$
$H_a : p_1 > p_2$

# Chi Squares Tests and ANOVA

- Chi Square tests are for hypothesis of the form
  $H_0$: $p_1 = a_1$, $p_2 = a_2$, …, $p_k = a_k$
- ANOVA tests

  $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$

# Example

The purpose of statistical inference is to provide information about the

a) sample based upon information contained in the population
b) population based upon information contained in the sample
c) population based upon information contained in the population
d) mean of the sample based upon the mean of the population

# Example

Analysis of variance is used to test:

a) Whether $k$ population variances are all equal.
b) Whether $k$ population standard deviations are all equal.
c) Whether k population means are all equal.
d) Whether $k$ sample means are all equal.

# Example

The Central Limit Theorem is important in statistics because

a) For a large n, it says the population is approximately normally distributed.
b) For any population, it says the sampling distribution of the sample mean is approximately normal, regardless of the sample size.
c) For a large n, it says the sampling distribution of the sample mean is approximately normal, regardless of the shape of the population.
d) For any sized sample, it says the sampling distribution of the sample mean is approximately normal.

# Example

In testing the hypotheses *H*o: $\mu = 50$ vs. *H*a: $\mu > 50$, the following information is known: $n = 64$, $\bar{x} = 53.5$, and $s = 10$. The test statistic equals:

a) 1.96
b) -2.80
c) 2.80
d) -1.96

# Example

A survey claims that 9 out of 10 doctors recommend aspirin for their patients with headaches. To test this claim against the alternative that the actual proportion of doctors who recommend aspirin is less than 0.90, a random sample of 100 doctors' results in 83 who indicate that they recommend aspirin. The value of the test statistic in this problem is approximately equal to:

a) −1.67
b) −2.33
c) −1.96
d) −0.14

# Example

A 95% confidence interval for the true proportion, p, is (0.23, 0.53). What is the best interpretation of the confidence interval?

a) We are 95% confident that the true proportion is 38%.
b) 95% of the time, the true proportion will fall between 23% and 53%.
c) In repeated sampling from this same population and calculating 95% confidence intervals, about 95% of these intervals will contain 38%.
d) In repeated sampling from this same population and calculating 95% confidence intervals, about 95% of these intervals will contain the true proportion.

# Example

A recent study of 750 internet users in Europe found that 35% of internet users were women. What is the 95% confidence interval of the true proportion of women in Europe who use the internet?

a) (.349,.351)
b) (.321,.379)
c) (.316,.384)
d) (.309,.391)

# Example

The National Center for Education would like to estimate the proportion of students who defaulted on their student loans for the state of Arizona. The total sample size needed to construct a 95% confidence interval for the proportion of student loans in default with a margin of error equal to 4% is _____.

a) 336
b) 416
c) 455
d) 601

# Example

) The U.S. Department of Labor and Statistics wanted to compare the results of an unemployment program for the past two months in the U.S. Suppose the proportion of the unemployed two months ago is p2 and the proportion of the unemployed one month ago is p1. A study found a 95% confidence interval for p2–p1 is (-0.0012, 0.003). Give an interpretation of this confidence interval.

a) We are 95% confident that the proportion of the unemployed one month ago is between 0.12% less and 0.3% more than the proportion of the unemployed two months ago.
b) We are 95% confident that the proportion of the unemployed two months ago is between 0.12% less and 0.3% more than the proportion of the unemployed one month ago.
c) We know that 95% of the unemployed two months ago is between 0.12% less and 0.3% more than the unemployed one month ago.
d) We know that 95% of all random samples done on the population will show that the proportion of the unemployed two months ago is between 0.12% less and 0.3% more than the proportion of the unemployed one month ago.
e) We know that 95% of the unemployed one month ago is between 0.12% less and 0.3% more than the unemployed two months ago.

# Example

According to research company comScore, Facebook users spent an average of 405 minutes on the site during the month of January 2012. Assume that this population has a standard deviation of 135 minutes. A random sample of 32 users was selected from this population. What is the probability that the average number of minutes on the site in January was less than 390 minutes?

a) 0.2643
b) 0.3669
c) 0.4801
d) 0.5398

# Example

The Department of Housing and Urban Development (HUD) would like to test the hypothesis that the average size of a newly constructed house in 2010 is different from the average size of a newly constructed house in 2000. The following data summarizes the sample statistics for house sizes, in square feet, for both years. We wish to test $H_o : \mu_{2000} = \mu_{2010}$ $H_a : \mu_{2000} \neq \mu_{2010}$

| | 2000 | 2010 |
|---|---|---|
| Sample mean | 2,180 | 2,390 |
| Sample size | 15 | 12 |
| Sample standard deviation | 300 | 320 |

# Example

```
> tsum.test(mean.x=2180,n.x=15,s.x=300,mean.y=2390,n.y=12,s.y=320)

        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = -1.742, df = 22.98, p-value = 0.09489
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -459.39731   39.39731
sample estimates:
mean of x mean of y
     2180      2390
```

# Example

Which one of the following statements is true?

a) Because the *p*-value is greater than .05, HUD can conclude that the average size of a newly constructed house in 2010 is different from the average size of a newly constructed house in 2000.
b) Because the *p*-value is greater than.05, HUD cannot conclude that the average size of a newly constructed house in 2010 is different from the average size of a newly constructed house in 2000.
c) Because the *p*-value is less than .05, HUD cannot conclude that the average size of a newly constructed house in 2010 is different from the average size of a newly constructed house in 2000.
d) Because the *p*-value is less than .05, HUD can conclude that the average size of a newly constructed house in 2010 is different from the average size of a newly constructed house in 2000.

# Example

You are looking at your calculator before an exam and are faced with the following hypothesis test:

Ho: The amount of power remaining in the battery is equal to a level which is high enough to last the length of the exam.
Ha: The amount of power remaining in the battery is less than the level which is high enough to last the length of the exam.

Which statement best describes a Type I error you could make with these hypotheses?

a) I assume the battery has enough power, when in fact it does not, and the battery dies during the exam.
b) I assume the battery has enough power, when in fact is does, and lasts throughout the exam.
c) I assume the battery will die, when in fact it does have sufficient power, and I replace a good battery with a new one.
d) I assume the battery will die, when in fact it does not have sufficient power, and I replace a bad battery with a new one.
e) I assume the battery has enough power, when in fact it does, but I still bring five extra calculators just in case.

# Example

Three instructors teach different sections of an introductory-level economics class during the fall semester. The number of students in each section is:

| Mankiw | Miron | Sumners |
|---|---|---|
| 100 | 125 | 75 |

The null hypothesis is that all three instructors are equally popular. What is the value of the Chi-square goodness of fit statistic for this data?

# Example

Suppose a 95% confidence interval for the proportion of Americans who exercise regularly is 0.29 to 0.37. Which one of the following statements is FALSE?

A. It is reasonable to say that more than 25% of Americans exercise regularly.
B. It is reasonable to say that more than 40% of Americans exercise regularly.
C. The hypothesis that 33% of Americans exercise regularly cannot be rejected.
D. It is reasonable to say that fewer than 40% of Americans exercise regularly.

# Example

A hypothesis test is done in which the alternative hypothesis is that more than 10% of a population is left-handed. The p-value for the test is calculated to be 0.25. Which statement is correct?

A. We can conclude that more than 10% of the population is left-handed.
B. We can conclude that more than 25% of the population is left-handed.
C. We can conclude that exactly 25% of the population is left-handed.
D. We cannot conclude that more than 10% of the population is left-handed.