ECMT 461 Lecture Slides for Exam 2 Material Spring 2024

Craig Schulman

The following slide decks cover material that will be included on Exam 1.

I reserve the right to amend/expand this material as needed.

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #7

1

AGENDA

- Term Project Proposal
- Sampling Distributions
- Confidence Intervals

2

SAMPLING DISTRIBUTIONS

- The previous discussion dealt with probabilities for specific values of a random variable
- We want to be able to make probability statements regarding various ***sample statistics*** that we calculate from sample data such as the sample mean $\overline{X}$, the sample variance S^2, or a sample proportion $\hat{P}$
- If a random variable X has a constant population mean μ and variance σ^2 it is denoted $X \sim (\mu, \sigma^2)$
- Given a random sample of size n for X, the **Sampling Distribution** is the probability distribution function (PDF) of the various sample statistics

3

SAMPLING DISTRIBUTION OF THE SAMPLE MEAN $\overline{X}$

- For a random variable $X \sim (\mu, \sigma^2)$ – a random variable with a constant mean and constant variance, the sample mean $\overline{X}$ has the distribution properties:

$$\bar{X} = \frac{\sum X_i}{n}$$

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim (0, 1)$$

4

CENTRAL LIMIT THEOREM

- For a random variable $X \sim (\mu, \sigma^2)$, as the sample size n becomes "large" then Z as defined above is approximately Normally distributed.
- Thus, for *any* random variable with a constant mean and variance, we can make statements in probability about the sample mean based on the Normal probability distribution when we have a sufficiently large sample
- Usually, 20 or more observations is a sufficiently large sample to invoke the Central Limit Theorem

5

EXAMPLE

- A random variable X is distributed with a constant mean $\mu = 23$ and variance $\sigma^2 = 400$
- A random sample of size n = 36 is obtained. What is the probability of finding a sample mean $\overline{X} \leq 17$?
- Find the Z-Score for $\overline{X} = 17$

$$Z_0 = \frac{\overline{X}_0 - \mu}{\sqrt{\sigma^2/n}} = \frac{17 - 23}{\sqrt{400/36}} = -1.8$$

- In Excel, use =NORM.S.DIST(-1.8,1) to get $P[Z \leq Z_0] \cong 0.036$

6

EXAMPLE CONT.

- What is the probability of finding a sample mean $\overline{X} \geq 27$?
- Find the Z-Score for $\overline{X} = 27$

$$Z_0 = \frac{27 - 23}{\sqrt{400/36}} = 1.2$$

- In Excel, use =1-NORM.S.DIST(1.2,1) to get $P[Z \geq Z_0] \cong 0.115$
- To find $P[17 \leq \overline{X} \leq 27]$ use
  - =NORM.DIST(1.2,1) – NORM.DIST(-1.8,1) = 0.849

7

NORMAL APPROX. TO BINOMIAL

- Let X be a binomial random variable that equals one (a "success") with probability P and is zero with probability (1 – P)
- The sampling distribution of the sample ***proportion*** $\hat{P}$ is
- n is "large" if: $n\hat{P}(1 - \hat{P}) > 5$

$$\hat{P} = \frac{\sum X_i}{n}$$

$$E(\hat{P}) = P$$

$$Var(\hat{P}) = \frac{P(1-P)}{n}$$

$$Z = \frac{\hat{P} - P}{\sqrt{P(1-P)/n}} \sim N(0,1) \text{ if } n \text{ is "large"}$$

8

EXAMPLE

- In Major League Baseball, the probability that a pitcher throws strictly left-handed is 26% (P = 0.26)
- In a random sample of n = 100 MLB pitchers, what is the probability of finding a sample proportion $\hat{P} \geq 0.33$?

$$Z_0 = \frac{\hat{P} - P}{\sqrt{P(1-P)/n}} = \frac{0.33 - 0.26}{\sqrt{0.26(1-0.26)/100}} \cong 1.596$$

$$P[P \geq \hat{P}] = P[Z \geq Z_0] = 1 - P[Z \leq Z_0] \cong 0.055$$

9

SAMPLING DIST. OF THE SAMPLE VARIANCE S²

- The sampling distribution of the sample variance S^2 is then

$$E(S^2) = \sigma^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

10

STUDENT'S t-DISTRIBUTION

- For any random variable X with a constant mean μ and variance σ^2, the Central Limit Theorem allows us to make probability statements about the sample mean $\bar{X}$ based on the Standard Normal Distribution provided the sample size n is "large"

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ is approximately } N(0,1)$$

- However, to calculate this Z-score, we need to know the population variance σ^2

11

STUDENT'S t-DISTRIBUTION

- The Student's t-distribution was developed to address the issue of an unknown population variance. It is the ratio of a Standard Normal Z-score and the square root of a Chi-Square distribution – using the sampling distribution of the sample variance S^2

$$t_{Stat} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

- Replacing the population standard deviation σ in the Z-score with the sample standard deviation S, we get a t-statistic with n – 1 degrees of freedom

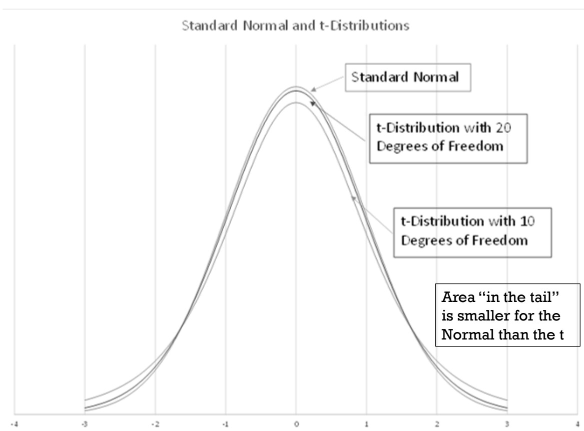
12

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

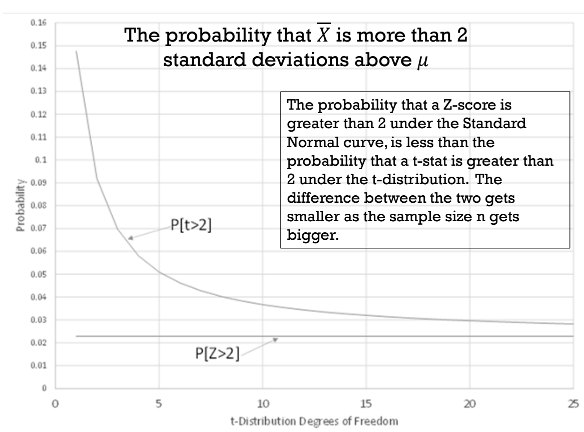
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

$$t = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

13



14



15

NORMAL EXAMPLE

- A random variable X is distributed with a constant mean μ = 23 and variance $\sigma^2 = 400$
- A random sample of size n = 36 is obtained. What is the probability of finding a sample mean $\overline{X} \leq 17$?
- Find the Z-Score for $\overline{X} = 17$

$$Z_0 = \frac{\overline{X}_0 - \mu}{\sqrt{\sigma^2/n}} = \frac{17 - 23}{\sqrt{400/36}} = -1.8$$

- In Excel, use =NORM.S.DIST(-1.8,1) to get $P[Z \leq Z_0] \cong 0.036$

16

SAME #'S BUT UNKNOWN VARIANCE

- A random variable X is distributed with a constant mean μ = 23. A random sample of size n = 36 is obtained. The ***unknown*** variance is estimated to be $S^2 = 400$
- What is the probability of finding a sample mean $\overline{X} \leq 17$?
- Find the t-stat for $\overline{X} = 17$

$$t_{stat} = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} = \frac{17 - 23}{\sqrt{400/36}} = -1.8$$

- In Excel, use =T.DIST(t-stat,degrees of freedom,1)
- =T.DIST(-1.8,35,1), to get $P[t \leq t_{stat}] \cong 0.04$

17

EXAMPLE CONT.

- What is the probability of finding a sample mean $\overline{X} \geq 27$?
- Find the t-stat for $\overline{X} = 27$

$$t_{stat} = \frac{27 - 23}{\sqrt{400/36}} = 1.2$$

- In Excel, use =1-t.DIST(1.2,35,1) to get
  - $P[t \geq t_{stat}] \cong 0.119$, under the Normal this was 0.115
- To find $P[17 \leq \overline{X} \leq 27]$ use
  - =t.DIST(1.2,35,1) – t.DIST(-1.8,35,1) = 0.841

18

CONFIDENCE INTERVALS

- Sample Mean with known population variance
- Sample Proportion (Normal Approximation to the Binomial)
- Sample Mean with unknown population variance
- Sample Variance

19

CONFIDENCE INTERVALS

- For a random variable X~(μ, σ^2), with sample size n large enough to invoke the Central Limit Theorem, assume the population variance σ^2 is known, but the population mean is unknown
- We want an Upper Confidence Limit (UCL) and Lower Confidence Limit (LCL) so that

$$P[LCL \le \mu \le UCL] = 1 - \alpha$$

- Where $(1 - \alpha)$ is the ***Confidence Level*** and α is a chosen ***Significance Level*** between zero and one

20

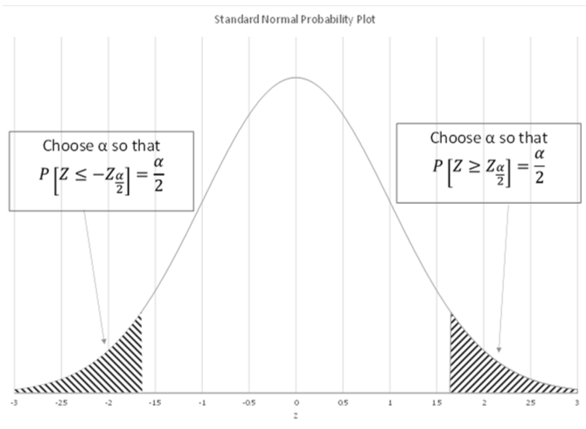
CONFIDENCE INTERVALS

- Start by constructing the Z-Score
- Next, choose α. For example, for a 90% **Confidence Level,** (1 – α) = 90% so α = 10% or 0.1
- Next, split α in half so that α/2% falls below the LCL and above the UCL – in the two tails under the Standard Normal distribution
- Given α, then find $Z_{\alpha/2}$ so that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$P\left[-Z_{\frac{\alpha}{2}} \le Z \le Z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

21



22

CONFIDENCE INTERVALS

- In Excel, suppose we want a 90% confidence level, so we choose α = 10% or α/2 = 5%
- To get $Z_{\frac{\alpha}{2}}$ we can then use =NORM.S.INV$\left(1-\frac{\alpha}{2}\right)$
- =NORM.S.INV(0.95) to get $Z_{\frac{\alpha}{2}} \cong 1.645$
- Because of the symmetry of the Normal $-Z_{\frac{\alpha}{2}} \cong -1.645$

$$P\left[-Z_{\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

23

CONFIDENCE INTERVALS

- Rearranging terms inside the probability statement

$$\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \le \mu \le \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}, or$$

$$\mu \text{ is in the interval: } \bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}, where$$

$$\text{the Upper Confidence Limit: } UCL = \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}},$$

$$\text{the Lower Confidence Limit: } LCL = \bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}, and$$

$$\text{the Margin of Error: } ME = \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$$

24

EXAMPLE

- A random variable X is distributed with an unknown but constant mean μ and known variance $\sigma^2 = 400$
- A random sample of size n = 36 is obtained and the sample mean is calculated to be $\overline{X} = 23$. Find a 90% Confidence Interval for the population mean μ
- $Z_{\frac{\alpha}{2}} \cong 1.645$ so there is a 90% probability that

$$\mu \text{ is in the interval: } \bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} = \frac{20}{\sqrt{36}} 1.645 \cong 5.483$$

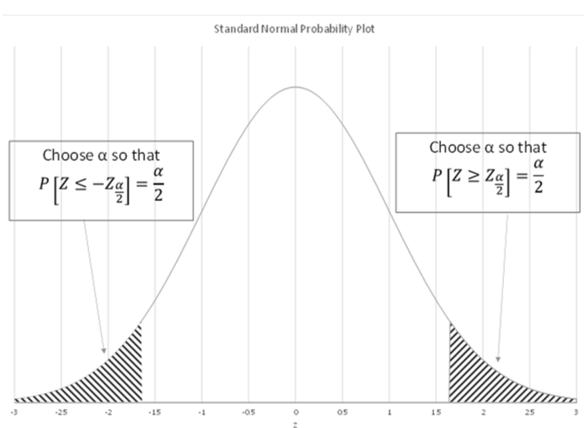
$$23 - 5.483 = 17.517 \leq \mu \leq 28.483 = 23 + 5.483$$

25

???

- What happens to the interval if the confidence level increases from 90% to 95% – the significance level α decreases from 10% to 5%?

26



27

MARGIN OF ERROR

- Note that the Margin of Error (ME) is a product of
- The standard deviation of $\overline{X}$ and a probability measure $Z_{\frac{\alpha}{2}}$

$$\text{the Margin of Error: } ME = \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$$

28

EXAMPLE

- A random variable X is distributed with an unknown but constant mean μ and known variance $\sigma^2 = 400$
- A random sample of size n = 36 is obtained and the sample mean is calculated to be $\overline{X} = 23$. Find a 90% Confidence Interval for the population mean μ
- $Z_{\frac{\alpha}{2}} \cong 1.645$ so there is a 90% probability that

$$\mu \text{ is in the interval: } \bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} = \overline{X} \pm \frac{20}{\sqrt{36}} 1.645 \cong \overline{X} \pm 5.483$$

$$23 - 5.483 = 17.517 \leq \mu \leq 28.483 = 23 + 5.483$$

29

EXAMPLE CONT.

- With a 90% CI, $Z_{\frac{\alpha}{2}} \cong 1.645$ so there is a 90% probability that

$$\mu \text{ is in the interval: } \bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} = \overline{X} \pm \frac{20}{\sqrt{36}} 1.645 \cong \overline{X} \pm 5.483$$

$$23 - 5.483 = 17.517 \leq \mu \leq 28.483 = 23 + 5.483$$

- If we increase our confidence level to 95% then $Z_{\frac{\alpha}{2}} \cong 1.96$ and

$$\mu \text{ is in the interval: } \bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} = \overline{X} \pm \frac{20}{\sqrt{36}} 1.96 \cong \overline{X} \pm 6.533$$

$$23 - 6.533 = 16.467 \leq \mu \leq 29.533 = 23 + 6.533$$

30

ANOTHER EXAMPLE

- Scores on an aptitude test are known to follow a normal distribution with a standard deviation of 32.4 points. A random sample of 12 test scores had a mean score of 189.7 points.
- Based on the sample results, a confidence interval for the population mean is found extending from 171.4 to 208 points. Find the confidence level of this interval.

Margin of Error (ME) ==>$ME = UCL - \bar{X} = 208 - 189.7 = 18.3$

$(or\ ME = \bar{X} - LCL)$

Z-Score ($Z_{\alpha/2}$) ==>$Solve\ ME = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ $18.3 = Z_{\frac{\alpha}{2}} \frac{32.4}{\sqrt{12}} \rightarrow Z_{\frac{\alpha}{2}} \cong 1.96$

Confidence Level ==>$P\left[Z > Z_{\frac{\alpha}{2}} = 1.96\right] = 0.025\ so\ Confidence\ Level = 0.95$

31

SAMPLE PROPORTIONS

- Recall from Lecture 5 that if X is a binomial random variable that equals one (a "success") with probability P and is zero with probability (1 – P)
- The sampling distribution of the sample ***proportion*** $\hat{P}$ is

$$\hat{P} = \frac{\sum X_i}{n}$$

$$E(\hat{P}) = P$$

$$Var(\hat{P}) = \frac{P(1-P)}{n}$$

$$Z = \frac{\hat{P} - P}{\sqrt{P(1-P)/n}} \sim N(0,1)\ if\ n\ is\ "large"$$

32

SAMPLE PROPORTIONS

- Based on this sampling distribution, we can construct a $(1-\alpha)\%$ confidence interval using:

$$\hat{P} \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

$$ME = Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

33

EXAMPLE

- During the 2018 NCAA DI Baseball season, among a random sample of 357 pitchers, 134 (37.5%) were the starting pitcher for more than 5 games. Use the Normal Approximation to the Binomial Distribution to construct a 90% confidence interval (α = 10%) for the population proportion P of pitchers that started more than 5 games.

$$Z_{\alpha/2} \cong 1.645$$

$$ME = 1.645\sqrt{\frac{(0.375)(0.625)}{357}} \cong 0.0421$$

$$LCL = 0.375 - ME \cong 0.3329$$

$$UCL = 0.375 + ME \cong 0.4171$$

34

SAMPLE MEAN – UNKNOWN VARIANCE

- Recall that if we replace the population standard deviation σ in the Z-score with the sample standard deviation S, we get a t-statistic with n − 1 degrees of freedom

$$t_{Stat} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

35

SAMPLE MEAN UNKNOWN VARIANCE

- We can then rearrange this t-stat to construct a $(1 - \alpha)\%$ confidence interval using:

$$\bar{X} \pm \frac{S}{\sqrt{n}} t_{(n-1),\alpha/2}$$

$$ME = \frac{S}{\sqrt{n}} t_{(n-1),\alpha/2}$$

- In Excel, use =T.INV$(1 - \frac{\alpha}{2}, n - 1)$ to get $t_{(n-1),\alpha/2}$
- Use =STDEV.S(array) to calculate the sample standard deviation S

36

EXAMPLE

- Data is collected on a random sample of 35 Texas high schools for Instructional Spending Per Student
- Construct a 95% confidence interval for the population mean

37

SAMPLE MEAN UNKNOWN VARIANCE

	A	B	C	D
1	Instructional Exp Per Student			
2	6130		n	=COUNT(A2:A36)
3	6415		X-Bar	=AVERAGE(A2:A36)
4	5871		Sample Std. Dev.	=STDEV.S(A2:A36)
5	5158		alpha	0.05
6	5796		t(n-1,alpha/2)	=T.INV(1-(D5/2),D2-1)
7	5611		ME	=(D4/SQRT(D2))*D6
8	5466			
9	5481		LCL	=D3-D7
10	6291		UCL	=D3+D7
11	7085			

38

SAMPLE MEAN UNKNOWN VARIANCE

	A	B	C	D
1	Instructional Exp Per Student			
2	6130		n	35
3	6415		X-Bar	5318.971
4	5871		Sample Std. Dev.	1236.490
5	5158		alpha	0.05
6	5796		t(n-1,alpha/2)	2.032
7	5611		ME	424.749
8	5466			
9	5481		LCL	4894.222
10	6291		UCL	5743.721
11	7085			
12	8919			

39

SAMPLE VARIANCE

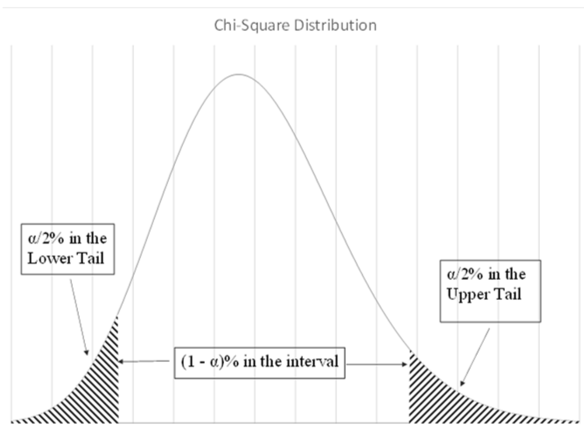
- To construct a confidence interval for the sample variance S^2, we start with the sampling distribution based on the Chi-Square distribution

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

- Recall that the Chi-Square distribution is not symmetric. When we choose a significance level, we need "upper" and "lower" Chi-Square values that leave α/2 in the "tails" to get

$$\frac{(n-1)S^2}{\chi^2_{(n-1),Upper\ Tail}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{(n-1),Lower\ Tail}}$$

40



41

EXAMPLE

- Among a random sample of 61 publicly traded firms, the sample variance S^2 of Debt/Equity ratios is computed to be 38.57. Construct a 90% confidence interval (α=10%) for the population variance σ^2.
- In Excel, use =CHISQ.INV(1 – α/2,n – 1) to get $\chi^2_{Upper} = 79.082$
- Use =CHISQ.INV(α/2,n – 1) to get $\chi^2_{Lower} = 43.188$
- Then calculate to get

$$LCL \cong 29.2633 \quad UCL \cong 53.5844$$

42

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #8

1

HYPOTHESIS TESTING

- Relationship to Confidence Intervals
- Conceptual Issues
  - ***Null Hypothesis***
  - ***Alternative Hypothesis***
  - Significance values and Test Power
- Single Sample Hypothesis Tests

2

CONFIDENCE INTERVALS AND HYPOTHESIS TESTING

- Suppose based on a random sample for the variable X we find a 95% C.I. such that

$$5.5 \leq \mu \leq 6.5$$

- In words, there is a 95% probability that the population mean of X is between 5.5 and 6.5
- Suppose we then pose the Hypothesis that μ is equal to 5.
- This would be considered a "low probability event" and we would reject the Hypothesis

3

THE NULL HYPOTHESIS

- The ***Null Hypothesis,*** denoted H_0, is a statement that we take to be true for purposes of the test
- The Null can be a "specific value" like the previous example stated as H_0: $\mu = 5$ (or any other specific value including zero)
- Or a relationship among the sample statistics of two or more random samples, H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
- We either ***Reject*** or ***Fail to Reject*** a Null hypothesis. A Null is never "accepted"

4

ALTERNATIVE HYPOTHESIS

- The ***Alternative Hypothesis,*** denoted H_A**,** is a statement that, if true, leads us to Reject the Null Hypothesis
- An Alternative Hypothesis can be either one-sided, a.k.a., directional, H_A: $\mu < 5$ or H_A: $\mu > 5$
- Or two-sided, H_a: $\mu \neq 5$
- A two-sided alternative is basically saying the Null is not true

5

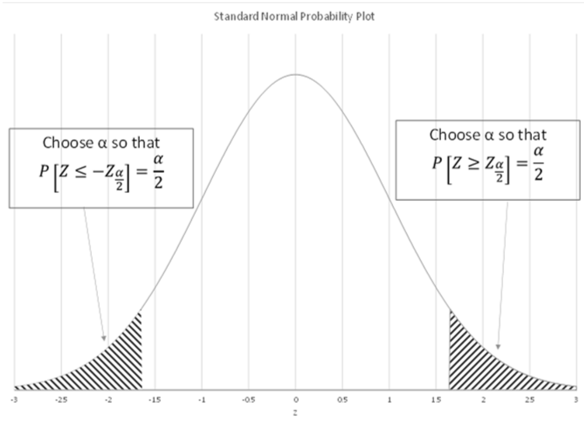
TEST SIGNIFICANCE AND POWER

- We already encountered the Significance Level α with C.I.'s

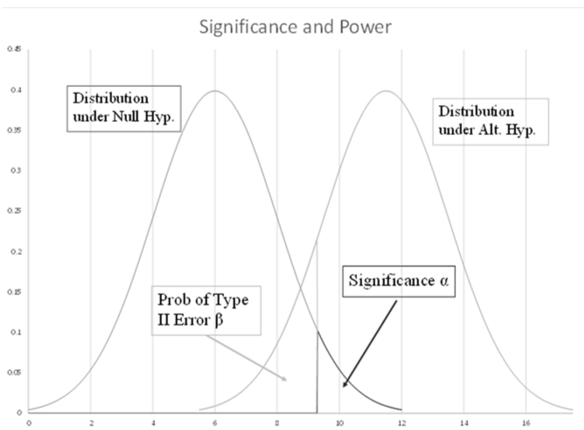
Decision	States of Nature	
	Null Hypothesis is True	Null Hypothesis is False
Fail to Reject H_0	Correct Decision $Prob = 1 - \alpha$	Type II Error $Prob = \beta$
Reject H_0	Type I Error $Prob = \alpha$	Correct Decision $Prob = 1 - \beta$

- The ***Power*** of a test is one minus the probability of a Type II error, denoted $(1 - \beta)$ and depends on the distribution under the Alternative Hypothesis

6



7



8

SINGLE SAMPLE TESTS

- To conduct single sample hypothesis tests, we calculate ***test statistics*** based on the same information that we used to construct C.I.'s
- Suppose X has a known population variance σ^2 and the sample size is n. Then to test H_0: $\mu_x = 7$

$$Z_{calc} = \frac{\bar{X} - 7}{\frac{\sigma}{\sqrt{n}}}$$

9

DECISION RULES

- For <u>one-sided</u> alternative hypotheses

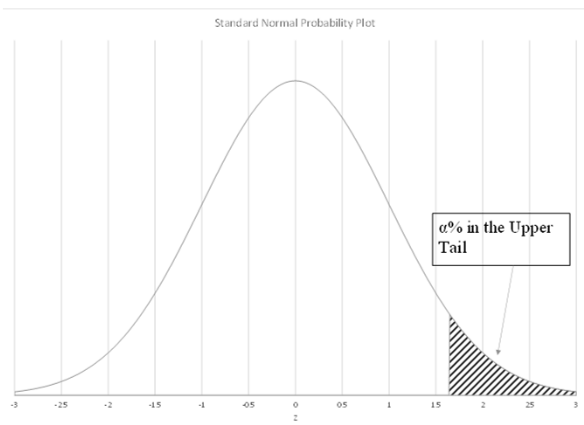
$$For\ H_A:\ \mu_X > 7, Reject\ H_0\ if\ Z_{calc} > Z_\alpha$$

$$otherwise\ Fail\ to\ Reject$$

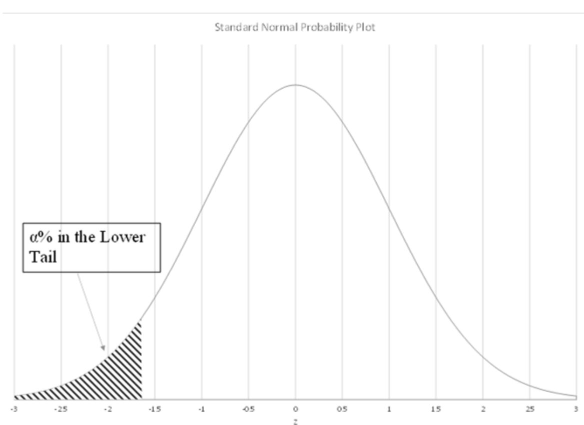
$$For\ H_A:\ \mu_X < 7, Reject\ H_0\ if\ Z_{calc} < -Z_\alpha$$

$$otherwise\ Fail\ to\ Reject$$

10



11



12

EXAMPLE

- A sample taken from a normal population with a **known** variance yields the following information: $\bar{x} = 25, n = 36, \sigma^2 = 81, \alpha = 0.05$
- Provide a test of the null hypothesis that the population mean $\mu = 27.5$ versus the alternative hypothesis that $\mu < 27.5$ at a confidence level $\alpha = 0.05$. First, calculate Z_{calc}

$$Z_{calc} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{25 - 27.5}{9/6} \cong -1.667$$

- For $\alpha = 0.05$, find $Z_\alpha = 1.645$
- Because $Z_{calc} < -Z_\alpha$ *or* $|Z_{calc}| > Z_\alpha$ we **Reject** the Null

13

DECISION RULES

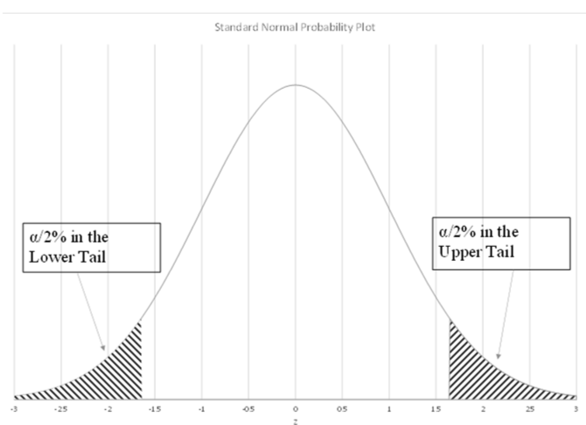
- For a <u>two-sided</u> alternative hypothesis of $H_0: \mu_x = 7$

$$H_A: \mu_X \neq 7$$

$$Reject\ H_0\ if\ Z_{calc} > Z_{\frac{\alpha}{2}}\ or\ Z_{calc} < -Z_{\frac{\alpha}{2}}$$

$$Fail\ to\ Reject\ if\ -Z_{\frac{\alpha}{2}} < Z_{calc} < Z_{\frac{\alpha}{2}}$$

14



15

EXAMPLE

- A sample taken from a normal population with a **known** variance yields the following information: $\bar{x} = 25, n = 36, \sigma^2 = 81, \alpha = 0.05$
- Provide a test of the null hypothesis that the population mean $\mu = 27.5$ versus the alternative hypothesis that $\mu \neq 27.5$ at a confidence level $\alpha = 0.05$. First, calculate Z_{calc}

$$Z_{calc} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{25 - 27.5}{9/6} \cong -1.667$$

- For $\alpha = 0.05$, find $Z_{\alpha/2} = 1.96$
- Because $-Z_{\alpha/2} < Z_{calc} < Z_{\alpha/2}$ we **Fail to Reject** the Null

16

DECISION RULES

- We can also fashion decision rules similar to Confidence Limits based on the same information

$$for\ H_A:\ \mu_X > \mu_0:\ \ \bar{X}_{crit} = \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}}$$
$$Reject\ H_0\ if\ \bar{X} > \bar{X}_{crit}$$
$$for\ H_A:\ \mu_X < \mu_0:\ \ \bar{X}_{crit} = \mu_0 - Z_\alpha \frac{\sigma}{\sqrt{n}}$$
$$Reject\ H_0\ if\ \bar{X} < \bar{X}_{crit}$$
$$for\ H_A:\ \mu_X \neq \mu_0:\ \ \bar{X}_{crit} = \mu_0 \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$
$$Reject\ H_0\ if\ \bar{X} > \mu_0 + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\ , or\ \bar{X} < \mu_0 - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

17

POPULATION PROPORTIONS

- Using the Normal approximation to the Binomial Distribution, given a sample proportion $\hat{P}$ based on a sample size n, to test $H_0:\ P = P_0$ calulate

$$Z_{calc} = \frac{\hat{P} - P_0}{\sqrt{P_0(1 - P_0)/n}}$$

- And apply the same decision rules for one-sided and two-sided alternatives as above
- <u>Note that the standard deviation in the denominator is based on the value of P on the Null Hypothesis</u>

18

EXAMPLE

- In a recent survey of 1000 Americans, 530 said they believed the country is headed in the wrong direction. Provide a test of the hypothesis that the population proportion of Americans who believe the country is heading in the wrong direction is less than or equal to 50% versus the alternative hypothesis that it is greater than 50% with α = 0.02
- H_0: P = 0.5, vs. the alternative H_A: P > 0.5

$$\hat{P} = \frac{530}{1000} = 0.53$$

$$Z_\alpha \cong 2.06$$

$$Z_{calc} = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0.53 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1000}}} = 1.897$$

19

UNKNOWN VARIANCES

- To test the mean of a random variable X with an **unknown** variance, given the sample mean $\bar{X}$, the sample variance S^2, and the sample size n, we calculate a t-statistic, t_{calc}

$$H_0: \mu_X = \mu_0$$

$$t_{calc} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

- Apply the same decision rules substituting $t_{(n-1,\alpha)}$ for Z_α for a one-sided alternative or $t_{(n-1,\alpha/2)}$ for $Z_{\alpha/2}$ for a two-sided alternative

20

EXCEL EXAMPLE

	A	B	C	D	E
1	County	Per Capita Liquor Sales			
2	Coryell	20.61		Sample Size n	=COUNT(B2:B21)
3	Moore	31.31		Sample Mean X-Bar	=AVERAGE(B2:B21)
4	Leon	27.82		Sample Std. Dev. S	=STDEV.S(B2:B21)
5	Burleson	20.5		alpha	0.05
6	Uvalde	35.99		Null Hypothesis H_0: μ =	25
7	Milam	20.13		t-calc	=(E3-E6)/(E4/SQRT(E2))
8	Marion	62.79			
9	Tyler	5.59		Alternative H_A: μ >	25
10	Hale	24.81		$t_{(n-1,\alpha)}$	=T.INV(1-E5,E2-1)
11	San Saba	13.44			
12	Walker	58.56		FAIL TO REJECT	
13	Hartley	51.2			

21

EXCEL EXAMPLE

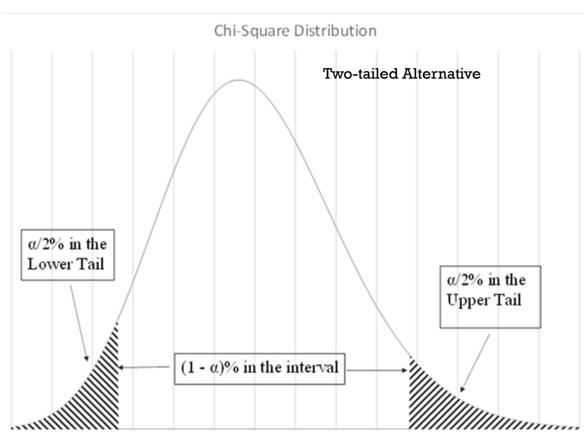
	A	B	C	D	E
1	County	Per Capita Liquor Sales			
2	Coryell	20.61		Sample Size n	20
3	Moore	31.31		Sample Mean X-Bar	29.926
4	Leon	27.82		Sample Std. Dev. S	14.454
5	Burleson	20.5		alpha	5%
6	Uvalde	35.99		Null Hypothesis H_0: μ =	25
7	Milam	20.13		t-calc	1.524
8	Marion	62.79			
9	Tyler	5.59		Alternative H_A: μ >	25
10	Hale	24.81		$t_{(n-1,\alpha)}$	1.729
11	San Saba	13.44			
12	Walker	58.56		FAIL TO REJECT	
13	Hartley	51.2			

22

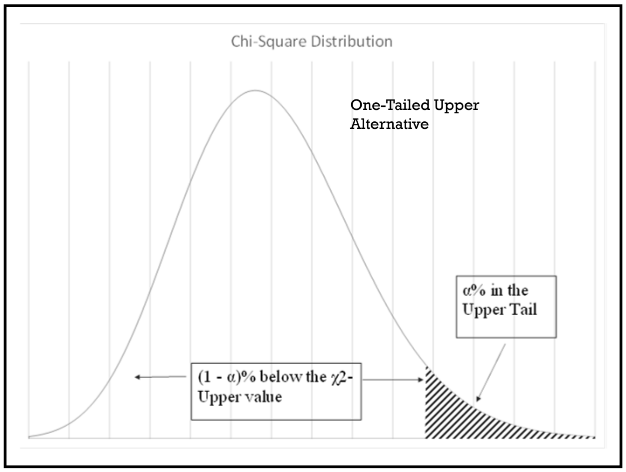
VARIANCE TEST

- To test a specific value hypothesis of the population variance like H_0: $\sigma^2 = \sigma_0^2$, calculate the Chi-Square test statistic
$$\chi^2_{calc} = \frac{(n-1)S^2}{\sigma_0^2}$$
- For the one-sided alternative H_A: $\sigma^2 < \sigma_0^2$, Reject if $\chi^2_{calc} < \chi^2_{(n-1,\alpha)LOWER}$
- For the one-sided alternative H_A: $\sigma^2 > \sigma_0^2$, Reject if $\chi^2_{calc} > \chi^2_{(n-1,\alpha)UPPER}$
- For a two-sided alternative H_A: $\sigma^2 \neq \sigma_0^2$, Reject if
$$\chi^2_{calc} < \chi^2_{(n-1,\frac{\alpha}{2})LOWER} \ or \ \chi^2_{calc} > \chi^2_{(n-1,\frac{\alpha}{2})UPPER}$$

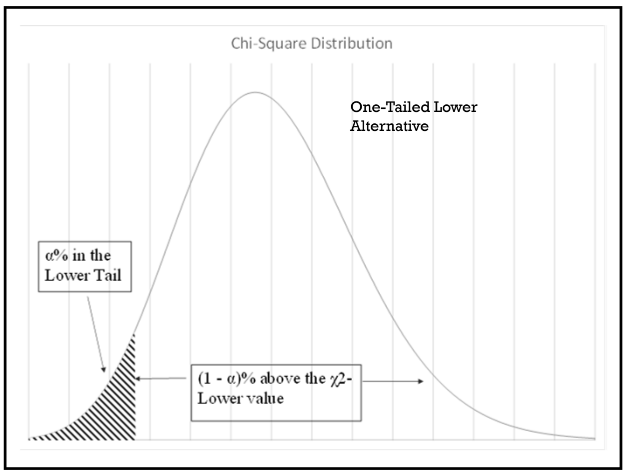
23



24


Chi-Square Distribution
One-Tailed Upper Alternative
α% in the Upper Tail
(1 - α)% below the χ2-Upper value

25


Chi-Square Distribution
One-Tailed Lower Alternative
α% in the Lower Tail
(1 - α)% above the χ2-Lower value

26

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #9

1

TWO-SAMPLE CONFIDENCE INTERVALS

- Relationship to Single-Sample Confidence Intervals
- Dependent Samples
- Independent Samples
  - Known Variances
  - Unknown Variances Assumed Equal
  - Unknown Variances Not Assumed Equal
  - Population Proportions

2

RELATIONSHIP TO SINGLE-SAMPLE CONFIDENCE INTERVALS

- Recall from our discussion of single-sample C.I.'s that the C.I. for a sample statistic such as the sample mean $\overline{X}$, was $\overline{X}$ plus or minus a Margin of Error (ME) and that the ME was the product of the standard deviation of $\overline{X}$ and a probability measure (a Z or t statistic)
- Two-sample C.I.'s are quite similar, relating the *difference* between the sample statistics from two samples, say $\overline{X}$ and $\overline{Y}$, and an ME involving the standard deviation of the *difference* and, again, a probability measure

3

DEPENDENT SAMPLES

- Suppose you have samples for two random variables X and Y where each observation in the X sample can be directly linked to an observation in the Y sample
- Examples
  - Scores from an aptitude test for a sample of students before and after a test prep course
  - Scores on a driving test for a sample of drivers before and after drinking a beer
  - Gas mileage for a sample of autos using regular and premium gasoline
- Interested in the *difference* in means

4

DEPENDENT SAMPLES

- Define the *difference* between each of the n observations in the X and Y samples as $d_i = X_i - Y_i$, then

$$\bar{d} = \frac{\sum d_i}{n} = \frac{\sum (X_i - Y_i)}{n}$$

$$S_d^2 = \frac{\sum \left(d_i - \bar{d}\right)^2}{n-1}$$

$$E\left(\bar{d}\right) = \mu_X - \mu_Y$$

$$Var\left(\bar{d}\right) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) = \frac{\sigma_{\bar{d}}^2}{n}$$

5

DEPENDENT SAMPLES

- We can then construct a $(1-\alpha)\%$ C.I. as

$$\bar{d} \pm \frac{S_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}}$$

$$where\ the\ margin\ of\ error: ME = \frac{S_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}}$$

6

DEPENDENT SAMPLES

- Example using the ERA of a sample of 65 MLB Pitchers Before and After being traded to a new team during a season

A	B	C	D	F	G	H
playerID	ERA Before Trade	ERA After Trade	Difference			
anderni01	3.92	2.11	=B2-C2			
baileho02	4.8	4.3	=B3-C3		Sample Size n	=COUNT(D2:D66)
barneja01	6.86	8.31	=B4-C4		Sample Mean of Diff	=AVERAGE(D2:D66)
barraky01	6.66	2.25	=B5-C5		Sample Std. Dev. S	=STDEV.S(D2:D66)
bauertr01	3.79	6.39	=B6-C6		alpha	0.1
biagijo01	3.78	7.36	=B7-C7			
brachbr01	6.13	3.68	=B8-C8		t(n-1,a/2)	=T.INV(1-(H6/2),H3-1)
brookaa01	5.01	6.18	=B9-C9			
cashnan01	3.83	6.2	=B10-C10		ME	=H8*(H5/SQRT(H3))
chacijh01	5.79	7.36	=B11-C11			
diekmja01	4.75	4.43	=B12-C12		UCL	=H4+H10
dysonsa01	2.47	7.15	=B13-C13		LCL	=H4-H10
fairbpe01	9.35	5.11	=B14-C14			

7

DEPENDENT SAMPLES

- Example using the ERA of a sample of 65 MLB Pitchers Before and After being traded to a new team during a season

A	B	C	D	E	F	G
playerID	ERA Before Trade	ERA After Trade	Difference			
anderni01	3.92	2.11	1.81			
baileho02	4.8	4.3	0.5		Sample Size n	65
barneja01	6.86	8.31	-1.45		Sample Mean of Diff	0.7312
barraky01	6.66	2.25	4.41		Sample Std. Dev. S	3.4596
bauertr01	3.79	6.39	-2.6		alpha	10%
biagijo01	3.78	7.36	-3.58			
brachbr01	6.13	3.68	2.45		t(n-1,a/2)	1.6690
brookaa01	5.01	6.18	-1.17			
cashnan01	3.83	6.2	-2.37		ME	0.7162
chacijh01	5.79	7.36	-1.57			
diekmja01	4.75	4.43	0.32		UCL	1.4474
dysonsa01	2.47	7.15	-4.68		LCL	0.0150
fairbpe01	9.35	5.11	4.24			
fariaja01	2.7	11.42	-8.72			

8

INDEPENDENT SAMPLES

- Suppose you have two *independent* random samples, say variable X and variable Y
- Examples
  - The ERA for a random sample of pitchers from the SEC and another from the ACC
  - Average SAT scores for a random sample a rural schools and another for a random sample of urban schools
  - Scores on an aptitude test from a random sample of Economics majors and another from a random sample of Political Science majors

9

INDEPENDENT SAMPLES – KNOWN VARIANCES

- Let the sample sizes for the two variables be denoted n_X and n_Y and assume the population variances for the two variables are **known**: σ_X^2 and σ_Y^2. Then:

$$E(\bar{X}-\bar{Y}) = \mu_X - \mu_Y$$

$$Var(\bar{X}-\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

- And a $(1-\alpha)$% C.I. is calculated from:

$$(\bar{X}-\bar{Y}) \pm \left(\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right) Z_{\frac{\alpha}{2}}$$

- But in practice we usually don't know the variances

$$ME = \left(\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right) Z_{\frac{\alpha}{2}}$$

10

INDEPENDENT SAMPLES – UNKNOWN VARIANCES ASSUMED EQUAL

- Again let the sample sizes for the two variables be denoted n_X and n_Y but assume the population variances for the two variables are **unknown and equal**: $Var(X) = VAR(Y) = \sigma^2$. Then:

$$E(\bar{X}-\bar{Y}) = \mu_X - \mu_Y$$

$$Var(\bar{X}-\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} = \sigma^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)$$

11

INDEPENDENT SAMPLES – UNKNOWN VARIANCES ASSUMED EQUAL

- Under the assumption that the variances are **unknown and equal**, we use the estimated sample variances S_X^2 and S_Y^2 to calculate a common *pooled* sample variance S_P^2

$$S_P^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

- And a $(1-\alpha)$% C.I. is constructed from

$$\bar{X} - \bar{Y} \pm \left(\sqrt{\frac{S_P^2}{n_X} + \frac{S_P^2}{n_Y}}\right) t_{(n_X+n_Y-2),\frac{\alpha}{2}}$$

- Note the degrees of freedom (DF) for the t distribution probability measure is $(n_X + n_Y - 2)$

12

INDEPENDENT SAMPLES – UNKNOWN VARIANCES ASSUMED EQUAL

- Suppose you have random samples of Expenditures Per Student on 63 Rural high schools, and 47 Suburban high schools. You are asked to construct a 95% C.I. for the difference in mean expenditures between the two samples.

Rural Sample Size	63
Rural Mean	9411.78
Rural Std. Dev	3344.713178
Suburban Sample Size	47
Suburban Mean	8272.21
Suburban Std. Dev	1641.865124
Pooled Variance	7570405
alpha	5%
t(nx+ny-2,a/2)	1.982
ME	1051.183
UCL	2190.75
LCL	88.38

13

INDEPENDENT SAMPLES – UNKNOWN VARIANCES NOT ASSUMED EQUAL

- Again let the sample sizes for the two variables be denoted n_X and n_Y but the population variances for the two variables are **unknown and not assumed to be equal**. Then:

$$E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$$

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

- We can use the sample variances $S_X^2 \; and \; S_Y^2 \; for \; \sigma_X^2 \; and \; \sigma_Y^2$ but this adds an extra degree of uncertainty that we account for by adjusting the degrees of freedom

14

INDEPENDENT SAMPLES – UNKNOWN VARIANCES NOT ASSUMED EQUAL

- With variances that are **unknown and not assumed equal**, we use the estimated sample variances $S_X^2 \; and \; S_Y^2$ to calculate a $(1-\alpha)\%$ C.I. from

$$\bar{X} - \bar{Y} \pm \left(\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\right) t_{\left(m, \frac{\alpha}{2}\right)}$$

- And the degrees of freedom m for the t distribution is calculated from
- We will use procedures in Excel to do the calculations of m

$$m = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_x}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

15

POPULATION PROPORTIONS

- Suppose you have samples for two independent binomial (yes/no) random variables X and Y
- We can use the sample proportions $\hat{P}_X$ *and* $\hat{P}_Y$ to calculate a $(1-\alpha)\%$ C.I. from

$$\hat{P}_X - \hat{P}_Y \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_Y}}$$

- MLB Left vs Right-Handed Pitcher Example

16

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #10

1

DEPENDENT SAMPLES

- We can then construct a $(1-\alpha)\%$ C.I. as

$$\bar{d} \pm \frac{S_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}}$$

$$where\ the\ margin\ of\ error: ME = \frac{S_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}}$$

2

DEPENDENT SAMPLES

A sample of 50 high school students is administered a standardized math exam Before and After a test training course with results provided in the table below. Treat this as a Dependent Sample.

	Before	After	Difference After - Before
Sample Mean	71.83	74.04	2.21
Sample Std. Dev. (S)	9.76	13.91	8.67

Construct a 95% confidence interval (α=5%) for the the mean difference in scores.

a. Margin of Error ME =	
b. Lower Confidence Limit LCL =	
c. Upper Confidence Limit UCL =	

Construct a 90% confidence interval (α=10%) for the the mean difference in scores.

a. Margin of Error ME =	
b. Lower Confidence Limit LCL =	
c. Upper Confidence Limit UCL =	

3

INDEPENDENT SAMPLES – KNOWN VARIANCES

- Let the sample sizes for the two variables be denoted n_X and n_Y and assume the population variances for the two variables are **known**: σ_X^2 *and* σ_Y^2. Then:

$$E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$$

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

- And a $(1-\alpha)\%$ C.I. is calculated from:

$$(\bar{X} - \bar{Y}) \pm \left(\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right) Z_{\frac{\alpha}{2}}$$

- But in practice we usually don't know the variances

$$ME = \left(\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right) Z_{\frac{\alpha}{2}}$$

4

INDEPENDENT SAMPLES – KNOWN VARIANCES

Independent Samples, Known Variances			
The longevity of Samsung smartphones and Apple I-Phones are Normally distributed with known population variances as shown in the table below. Independent samples of 95 Samsung phones and 87 I-Phones are collected to give the sample means shown below.			
	Sample Size	Sample Mean	Population Variance σ^2
Samsung Smart Phone	95	3.8	5.5
Apple I-Phone	87	3.2	4.3
Construct a 95% confidence interval (α=5%) for the the difference in population means $\mu_x - \mu_y$.			
a. Margin of Error ME =			
b. Lower Confidence Limit LCL =			
c. Upper Confidence Limit UCL =			

5

INDEPENDENT SAMPLES – UNKNOWN VARIANCES ASSUMED EQUAL

- Under the assumption that the variances are **unknown and equal**, we use the estimated sample variances S_X^2 *and* S_Y^2 to calculate a common *pooled* sample variance S_P^2

$$S_P^2 = \frac{(n_x - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

- And a $(1-\alpha)\%$ C.I. is constructed from

$$\bar{X} - \bar{Y} \pm \left(\sqrt{\frac{S_P^2}{n_X} + \frac{S_P^2}{n_Y}}\right) t_{(n_X+n_Y-2),\frac{\alpha}{2}}$$

- Note the degrees of freedom (DF) for the t distribution probability measure is $(n_X + n_Y - 2)$

6

INDEPENDENT SAMPLES – UNKNOWN VARIANCES ASSUMED EQUAL

Independent Samples, Unknown Variances Assumed Equal

Among two independent random samples of U.S. Counties, you are provided with the information below regarding the 2021 Unemployment Rate. Assume the **unknown population variances for these samples are equal**.

	Sample 1	Sample 2
Sample Means	4.61	4.35
Sample Standard Deviations	1.69	1.85
Sample Sizes	277	321

Construct a 90% confidence interval (α = 10%) for the difference in two population means (Sample 1 mean minus the Sample 2 mean).

a. Margin of Error ME =	
b. Lower Confidence Limit LCL =	
c. Upper Confidence Limit UCL =	

7

INDEPENDENT SAMPLES – UNKNOWN VARIANCES NOT ASSUMED EQUAL

- With variances that are **unknown and not assumed equal**, we use the estimated sample variances S_X^2 and S_Y^2 to calculate a $(1-\alpha)\%$ C.I. from

$$\bar{X} - \bar{Y} \pm \left(\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\right) t_{\left(m, \frac{\alpha}{2}\right)}$$

- And the degrees of freedom m for the t distribution is calculated from
- We will use procedures in Excel to do the calculations of m

$$m = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_x}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

8

INDEPENDENT SAMPLES – UNKNOWN VARIANCES NOT ASSUMED EQUAL

Independent Samples, Unknown Variances NOT Assumed Equal

Among a random sample of Rural U.S. Counties and a second random sample of Urban U.S. Counties you are provided the sample statistics below related to the Growth Rate in Personal Income. Assume the **unknown population variances for these samples are NOT equal**.

	Rural	Urban
Sample Means	2.19	2.5
Sample Standard Deviations	1.22	1.06
Sample Sizes	119	93
Assumed Degrees of Freedom m	193	

Construct a 95% confidence interval (α = 5%) for the difference in two population means (Rural mean minus the

a. Margin of Error ME =	
b. Lower Confidence Limit LCL =	
c. Upper Confidence Limit UCL =	

9

POPULATION PROPORTIONS

- Suppose you have samples for two independent binomial (yes/no) random variables X and Y
- We can use the sample proportions $\hat{P}_X$ *and* $\hat{P}_Y$ to calculate a $(1-\alpha)$% C.I. from

$$\hat{P}_X - \hat{P}_Y \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_Y}}$$

10

POPULATION PROPORTIONS

Population Proportions		
Among a random sample of major league pitchers from the National League (NL) and a second random sample from the American League (AL), you are provided the sample statistics below related to the proportion of pitchers that allowed more than 3 base-on-balls per nine innings pitched.		
	NL	AL
Sample Proportions	0.653	0.587
Sample Sizes	529	499
8. Construct a 99% confidence interval (α = 1%) for the difference in two population proportions (NL proportion minus the AL proportion).		
a. Margin of Error ME =		
b. Lower Confidence Limit LCL =		
c. Upper Confidence Limit UCL =		

11

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 11

Craig T. Schulman

1

AGENDA

- Two Sample Hypothesis Tests
  - Dependent Samples
  - Independent Samples
    - Known Variances
    - Unknown Variances Assumed Equal
    - Unknown Variances Not Assumed Equal
    - Population Proportions
    - Sample Variances Test
- Using Lecture 9 Example Data

2

TWO-SAMPLE HYPOTHESIS TESTS

- Recall that we formulate a ***Null Hypothesis,*** denoted H_0 that for the purposes of our test, we take to be true
- Set against an ***Alternative Hypothesis,*** denoted H_A
- We construct a calculated **Test Statistic**
- Comparing our Test Statistic to the appropriate Probability Measure, we then
- Reject or Fail to Reject the Null
- Two-Sample tests are typically structured to test the equality of two population parameters

3

DEPENDENT SAMPLES

- Suppose you have samples for two random variables X and Y where each observation in the X sample can be directly linked to an observation in the Y sample
- Examples
  - Scores from an aptitude test for a sample of students before and after a test prep course
  - Scores on a driving test for a sample of drivers before and after drinking a beer
  - Gas mileage for a sample of autos using regular and premium gasoline
- Interested in the *difference* in means

4

DEPENDENT SAMPLES

- Define the *difference* between each of the n observations in the X and Y samples as $d_i = X_i - Y_i$, then

$$H_0: \mu_X - \mu_Y = 0$$

$$t_{calc} = \frac{\bar{d}}{S_d/\sqrt{n}}$$

5

DEPENDENT SAMPLES

- Decision rules

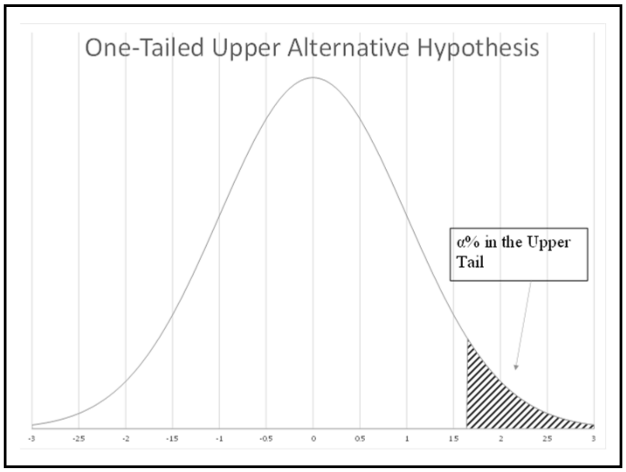
$$H_A: \mu_X - \mu_Y > 0 \quad \text{Reject if} \quad t_{calc} > t_{n-1,\alpha}$$

$$H_A: \mu_X - \mu_Y < 0 \quad \text{Reject if } t_{calc} < 0 \; and \; |t_{calc}| > t_{n-1,\alpha}$$

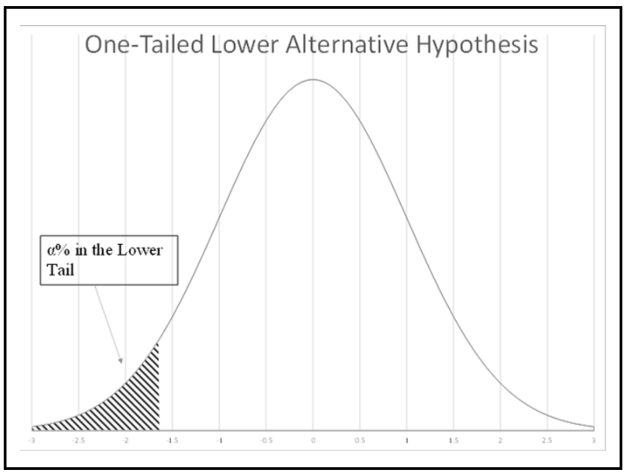
$$H_A: \mu_X - \mu_Y \neq 0 \quad \text{Reject if } |t_{calc}| > t_{n-1,\alpha/2}$$

- MLB Before-After Trade ERA Example

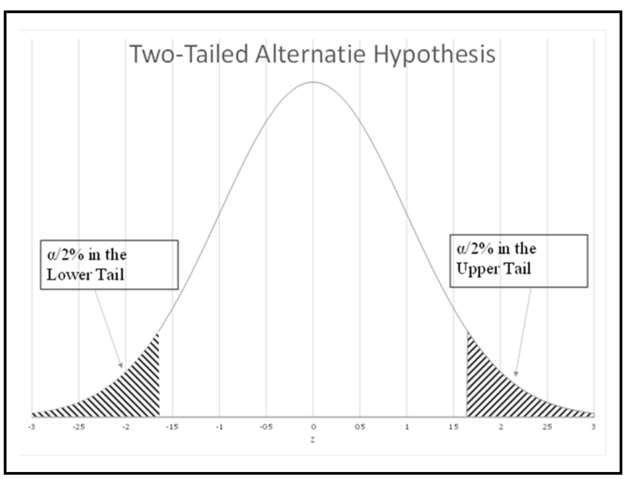
6


One-Tailed Upper Alternative Hypothesis
α% in the Upper Tail

7


One-Tailed Lower Alternative Hypothesis
α% in the Lower Tail

8


Two-Tailed Alternatie Hypothesis
α/2% in the Lower Tail
α/2% in the Upper Tail

9

INDEPENDENT SAMPLES

- Suppose you have two *independent* random samples, say variable X and variable Y
- Examples
  - The ERA for a random sample of pitchers from the SEC and another from the ACC
  - Average SAT scores for a random sample a rural schools and another for a random sample of urban schools
  - Scores on an aptitude test from a random sample of Economics majors and another from a random sample of Political Science majors

10

INDEPENDENT SAMPLES KNOWN VARIANCES

- Let the sample sizes for the two variables be denoted n_X and n_Y and assume the population variances for the two variables are **known**: σ_X^2 *and* σ_Y^2. Then to test $H_0: \mu_X - \mu_Y = 0$

$$H_0: \mu_X - \mu_Y = 0$$

$$Z_{calc} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_y}}}$$

11

INDEPENDENT SAMPLES KNOWN VARIANCES

- Decision rules

$$H_A: \mu_X > \mu_Y \text{ Reject if } Z_{calc} > Z_\alpha$$

$$H_A: \mu_X < \mu_Y \text{ Reject if } Z_{calc} < 0 \textit{ and } |Z_{calc}| > Z_\alpha$$

$$H_A: \mu_X \neq \mu_Y \text{ Reject if } |Z_{calc}| > Z_{\alpha/2}$$

12

INDEPENDENT SAMPLES UNKNOWN VARIANCES ASSUMED EQUAL

- Again, let the sample sizes for the two variables be denoted n_X and n_Y but assume the population variances for the two variables are **unknown and equal**: $Var(X) = VAR(Y) = \sigma^2$. Then we use the estimated sample variances S_X^2 *and* S_Y^2 to calculate a common *pooled* sample variance S_P^2

$$S_P^2 = \frac{(n_x - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

13

INDEPENDENT SAMPLES UNKNOWN VARIANCES ASSUMED EQUAL

- To test the Null that the population means are equal

$$H_0:\ \mu_X - \mu_Y = 0$$

$$t_{calc} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_P^2}{n_X} + \frac{S_P^2}{n_Y}}}$$

14

INDEPENDENT SAMPLES UNKNOWN VARIANCES ASSUMED EQUAL

- Decision rules

$$H_A: \mu_X > \mu_Y \ \text{ Reject if } t_{calc} > t_{(n_X+n_Y-2,\alpha)}$$

$$H_A: \mu_X < \mu_Y \ \text{ Reject if } t_{calc} < 0 \ and \ |t_{calc}| > t_{(n_X+n_Y-2,\alpha)}$$

$$H_A: \mu_X \neq \mu_Y \ \text{ Reject if } |t_{calc}| > t_{(n_X+n_Y-2,\alpha/2)}$$

- High School Expenditures per Student Example

15

INDEPENDENT SAMPLES UNKNOWN VARIANCES ASSUMED EQUAL

- How big does the difference in sample means need to be in order to reject the Null?
- One-Tailed Alternative

$$(\bar{X}-\bar{Y})_{crit} = \pm\left(\sqrt{\frac{S_P^2}{n_X}+\frac{S_P^2}{n_Y}}\right)t_{(n_X+n_Y-2),\alpha}$$

- Two-Tailed Alternative

$$(\bar{X}-\bar{Y})_{crit} = \pm\left(\sqrt{\frac{S_P^2}{n_X}+\frac{S_P^2}{n_Y}}\right)t_{(n_X+n_Y-2),\frac{\alpha}{2}}$$

16

INDEPENDENT SAMPLES – UNKNOWN VARIANCES NOT ASSUMED EQUAL

- Again, let the sample sizes for the two variables be denoted n_X and n_Y but the population variances for the two variables are **unknown and not assumed to be equal**. Then:

$$E(\bar{X}-\bar{Y}) = \mu_X - \mu_Y$$

$$Var(\bar{X}-\bar{Y}) = \frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}$$

- We can use the sample variances S_X^2 *and* S_Y^2 *for* σ_X^2 *and* σ_Y^2 but this adds an extra degree of uncertainty that we account for by adjusting the degrees of freedom

17

INDEPENDENT SAMPLES – UNKNOWN VARIANCES NOT ASSUMED EQUAL

- With variances that are **unknown and not assumed equal**, to test the Null that the population means are equal, we use the estimated sample variances S_X^2 *and* S_Y^2 to calculate

$$H_0\colon\ \mu_X - \mu_Y = 0$$

$$t_{calc} = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}}$$

18

INDEPENDENT SAMPLES – UNKNOWN VARIANCES NOT ASSUMED EQUAL

- Decision rules

$$H_A: \mu_X > \mu_Y \text{ Reject if } t_{calc} > t_{(m,\alpha)}$$

$$H_A: \mu_X < \mu_Y \text{ Reject if } t_{calc} < 0 \text{ and } |t_{calc}| > t_{(m,\alpha)}$$

$$H_A: \mu_X \neq \mu_Y \text{ Reject if } |t_{calc}| > t_{(m,\alpha/2)}$$

- We will use procedures in Excel to do the calculations of the degrees of freedom m

19

POPULATION PROPORTIONS

- Suppose you have samples for two independent binomial (yes/no) random variables X and Y
- To test for equal proportions, we can use the sample proportions $\hat{P}_X$ *and* $\hat{P}_Y$ to calculate

$$H_0: P_X = P_Y$$

$$\textit{Pooled sample Proportion}: P_0 = \frac{n_X\hat{P}_X + n_Y\hat{P}_Y}{n_X + n_Y}$$

$$Z_{calc} = \frac{\hat{P}_X - \hat{P}_Y}{\sqrt{\frac{P_0(1-P_0)}{n_X} + \frac{P_0(1-P_0)}{n_Y}}}$$

20

POPULATION PROPORTIONS

- Decision rules

$$H_A: P_X > P_Y \text{ Reject if } Z_{calc} > Z_\alpha$$

$$H_A: P_X < P_Y \text{ Reject if } Z_{calc} < 0 \text{ and } |Z_{calc}| > Z_\alpha$$

$$H_A: P_X \neq P_Y \text{ Reject if } |Z_{calc}| > Z_{\alpha/2}$$

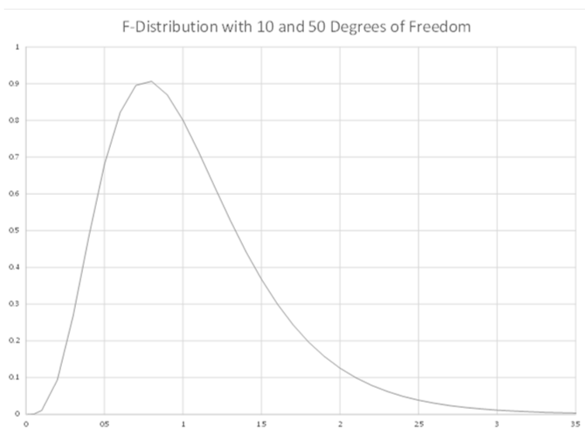
- MLB Left vs Right-Handed Pitcher Example

21

POPULATION VARIANCES

- Given independent random samples for two variables X and Y with sample sizes n_X and n_Y, the sample variances for the two variables have Chi-Square distributions with $(n_X - 1)$ degrees of freedom for S_X^2 and $(n_Y - 1)$ degrees of freedom for S_Y^2
- The ratio of two Chi-Square random variables has an F-Distribution with *numerator* and *denominator* degrees of freedom

22



23

POPULATION VARIANCES

- To test the hypothesis that the two population variances are equal $H_0: \sigma_X^2 = \sigma_Y^2$
- Construct a calculated F-Statistic $F_{calc} = \frac{S_X^2}{S_Y^2}$
- If $F_{calc} > 1$, Reject H_0 if $F_{calc} > F_{(n_X-1, n_y-1, \alpha)}\ Upper$
- If $F_{calc} < 1$, Reject H_0 if $F_{calc} < F_{(n_X-1, n_y-1, \alpha)}\ Lower$

24

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 12

Craig T. Schulman

1

AGENDA

- Two Sample Confidence Interval and Hypothesis Tests Examples
  - Dependent Samples
  - Independent Samples
    - Unknown Variances Assumed Equal
    - Unknown Variances Not Assumed Equal
    - Population Proportions
    - Sample Variances Test
- See the Lecture 12 Example Workbook

2

Dependent Sample

- A sample of 19 Aggie freshmen was administered a driving test before and after drinking a 16-ounce beer. The table below provides the sample means and standard deviations for the driving scores before the beer, after the beer, and the differences in scores.

	Before	After	Difference
Sample Mean	72.7	71.4	1.3
Sample Std. Deviation	15.546	17.489	1.943

- (a) Construct a 90% Confidence Interval for the difference in driving test scores. Show the Margin of Error (ME) the Lower Confidence Limit (LCL) and Upper Confidence Limit (UCL).

3

$$ME = \frac{S_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}}$$

$$UCL = \bar{d} + ME$$

$$LCL = \bar{d} - ME$$

$Given\ \bar{d} = 1.3,\ S_d = 1.943,\ n = 19, \quad \alpha = 10\%, \quad need\ t_{(n-1),\frac{\alpha}{2}}$

n	19
alpha	0.1
t(n-1, a/2)	=T.INV(1-(D11/2),D10-1)

$$ME = \left(\frac{1.943}{\sqrt{19}}\right) 1.734 \cong 0.773$$

$$\text{UCL} = 1.3 + 0.733 \cong 2.073$$

$$\text{LCL} = 1.3 - 0.733 \cong 0.527$$

4

(b) Provide a test of the null hypothesis $H_0: Mean\ difference = 0$ versus $H_A: Mean\ Difference\ \neq 0$. Set the significance level $\alpha = 0.05$, show your test statistic, decision criteria, and state your conclusion.

$$t_{calc} = \frac{\bar{d}}{S_d / \sqrt{n}} \qquad t_{calc} = \frac{1.3}{1.943/\sqrt{19}} \cong 2.916$$

$$H_A: \bar{d} \neq 0 \ \text{ Reject if } |t_{calc}| > t_{n-1,\alpha/2}$$

Two-Tailed	
alpha	0.05
t(n-1, a/2)	=T.INV(1-(D15/2),D10-1)

Reject H_0 *because* $t_{calc} > 2.101$

5

Independent Samples: Unknown variances assumed equal

- Among two independent random samples of Texas Counties, the sample statistics in the table below are computed for Alcohol Sales per person. In the following questions, assume the unknown population variances of these two samples are equal.

	Sample 1	Sample 2
Sample Means	87.5	77.8
Sample Variances	519	687
Sample Sizes	35	42

- (a) Construct a 95% confidence interval (α = 5%) for the difference between the two sample means. Show the Margin of Error (ME) the Lower Confidence Limit (LCL) and Upper Confidence Limit (UCL).

6

$$\bar{X} - \bar{Y} \pm \left(\sqrt{\frac{S_P^2}{n_X} + \frac{S_P^2}{n_Y}} \right) t_{(n_x + n_Y - 2), \frac{\alpha}{2}}$$

Where the pooled variance is calculated as

$$S_P^2 = \frac{(n_x - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

$$Pooled\ Sample\ Variance\ S_P^2 = \frac{(34)(519) + (41)(687)}{75} \cong 610.84$$

$t_{(75,\ 0.025)} = 1.992$

$$ME = t_{(75, 0.025)} * \sqrt{S_P^2 * \left(\frac{1}{35} + \frac{1}{42} \right)} \cong 11.2684$$

$$LCL = \overline{X} - \overline{Y} - ME \cong 9.7 - 11.2683 \cong -1.5684$$
$$UCL = \overline{X} - \overline{Y} + ME \cong 9.7 + 11.2683 \cong 20.9684$$

7

(b) Test the null hypothesis that the two sample means are equal versus the alternative hypothesis that the mean for Sample 1 is larger at a 95% confidence level (α=5%). Show your test statistic, decision criteria, and state your conclusion.

$$t_{calc} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_P^2}{n_X} + \frac{S_P^2}{n_Y}}} \qquad t_{calc} = \frac{9.7}{\sqrt{S_P^2 * \left(\frac{1}{35} + \frac{1}{42} \right)}} \cong 1.7148$$

$H_A: \mu_X > \mu_Y$ Reject if $t_{calc} > t_{(n_X + n_Y - 2, \alpha)}$

Reject if $t_{calc} > t_{(75,\ 0.05)} = 1.665$

Therefore Reject H_0

8

(b) ALTERNATIVE: How large would the difference between the two sample means need to be to reject the null hypothesis that the two sample means are equal versus the alternative hypothesis that the mean for Sample 1 is larger at a 99% confidence level (α=1%).

$$(\bar{X} - \bar{Y})_{crit} = \left(\sqrt{\frac{S_P^2}{n_X} + \frac{S_P^2}{n_Y}} \right) t_{(n_x + n_Y - 2), \alpha} \qquad t_{75,\ 0.01} = 2.377$$

$$\left(\overline{X} - \overline{Y} \right)_{crit} = \left(\sqrt{610.84 \left(\frac{1}{35} + \frac{1}{42} \right)} \right) 2.377 \cong 13.466$$

Because $\left(\overline{X} - \overline{Y} \right) = 9.7 < 13.466, Fail\ to\ Reject\ H_0$

9

Independent Samples: Unknown variances NOT assumed equal

- Among independent random samples of Rural and Urban Texas High Schools, the sample statistics in the table below are computed for a measure of student performance. In the following questions, assume the unknown population variances of these two samples are NOT equal.

	Rural	Urban
Sample Means	4.15	5.78
Sample Standard Deviations	4.47	3.59
Sample Sizes	42	65
Assumed Degrees of Freedom m =	71	

- Test the null hypothesis that the two sample means are equal versus the alternative hypothesis that they are not equal at a 95% confidence level (α=5%). Show your test statistic, decision criteria, and state your conclusion.

10

$$t_{calc} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \qquad t_{calc} = \frac{4.15 - 5.78}{\sqrt{\frac{4.47^2}{42} + \frac{3.59^2}{65}}} \cong -1.9854$$

$$Compare\ t_{calc}\ to\ t\left(m, \frac{\alpha}{2}\right) = t(71, 0.025) \cong 1.9939$$

$$Because\ |t_{calc}| < t\left(m, \frac{\alpha}{2}\right)\ Fail\ to\ Reject\ H_0$$

$$Note:\ If\ you\ had\ used\ t\left(n_X + n_Y - 2, \frac{\alpha}{2}\right) = 1.9828$$

You would have Incorrectly Rejected H_0

11

Independent Samples: Sample Proportions

- During the 2016 Major League Baseball season, among independent random samples of pitchers in the American League and National League, the number and proportion of pitchers born outside the U.S. is provided in the table below. Use the Normal Approximation to the Binomial Distribution to answer the following questions.

	AL	NL
Born Outside U.S.	43	56
Total	209	213
Proportion	20.57%	26.29%

- (a) Construct a 95% confidence interval (α = 5%) for difference between the proportion of pitchers born outside the U.S. between the American League and the National League (American League proportion minus National League proportion). Show the Margin of Error (ME) the Lower Confidence Limit (LCL) and Upper Confidence Limit (UCL).

12

$$\hat{P}_X - \hat{P}_Y \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_Y}} \qquad Z_{\alpha/2} = 1.96$$

$$ME = 1.96\sqrt{\frac{(0.2057)(0.7943)}{209} + \frac{(0.2629)(0.7371)}{213}} \cong 0.0806$$

$$LCL = 0.2057 - 0.2629 - 0.0806 \cong -0.1378$$

$$UCL = 0.2057 - 0.2629 + 0.0806 \cong 0.0234$$

13

(b) Test the null hypothesis that the proportion of pitchers born outside the U.S. in the National League is equal to that of the American League, $H_0\ P_{AL} = P_{NL}$ versus the alternative hypothesis that the American League proportion is smaller, $H_A: P_{AL} < P_{NL}$ at a 95% confidence level (α=5%). Show your test statistic, decision criteria, and state your conclusion.

$$Z_{calc} = \frac{\hat{P}_X - \hat{P}_Y}{\sqrt{\frac{P_0(1-P_0)}{n_X} + \frac{P_0(1-P_0)}{n_Y}}} \qquad P_0 = \frac{n_X P_X + n_y P_Y}{n_X + n_Y}$$

$$P_0 = \frac{(209)(0.2057) + (213)(0.2629)}{209 + 213} \cong 0.2346$$

$$Z_{calc} = \frac{0.2057 - 0.2629}{\sqrt{(0.2346)(0.7654)\left(\frac{1}{209} + \frac{1}{213}\right)}} \cong -1.3865$$

Reject if $Z_{calc} < -Z_\alpha = -1.645$
Therefore, Fail to Reject.

14

Among independent samples of Technology firms and Financial firms, sample sizes and the sample standard deviations of daily stock returns (measured in basis points) are provided in the table below.

	Technology	Financial
Sample Size	103	110
Sample Standard Deviation	27.55	23.45

Test the hypothesis that the population variances across the two samples are equal versus the alternative hypothesis that the variance for Technology firms is higher at a 95% confidence level (α=5%)

15

	Technology	Financial
Sample Size	103	110
Sample Standard Deviation	27.55	23.45

$$F_{calc} = \frac{S_X^2}{S_Y^2} \qquad F_{calc} = \frac{27.55^2}{23.45^2} \cong 1.3802$$

$H_A: \sigma_X^2 > \sigma_Y^2$ Reject if $F_{calc} > F_{(n_X-1, n_y-1, \alpha)}$

$F_{(102,109,0.05)} = 1.3784$

$F_{calc} > F_{(102,109,0.05)}$ *Therefore Reject* H_0

Note incorrectly switching Degrees of Freedom: $F_{(109,102,0.05)} = 1.3811$

16