NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W4: 9/15-17/15

NC STATE UNIVERSITY

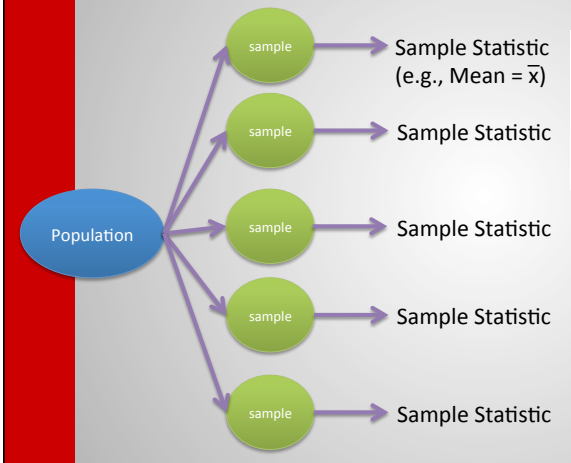
Administrative

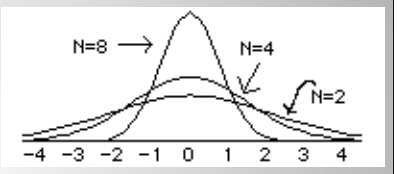
- Moodle usage
  - Good start
  - Please, don't submit any direct solution to any grading related questions
- Likewise, no postings on publicly accessible folders (Web, GitHub, etc.)
  - Applicable your own stuff (grading related materials), and lecture notes/presentations, and additional materials posted on Moodle

NC STATE UNIVERSITY

Key points from 9/10

- Sampling Distribution and Central Limit Theorem





The sd of the sampling distribution of the mean is called the standard error of the mean.

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

CLT: Given a population with μ and σ^2, the sampling distribution of the mean approaches a normal distribution $\bar{x}$ ~ N(mean = μ, and SE = s/√n)

NC STATE UNIVERSITY

Key points from 9/10

- Confidence Intervals
  - possible range of values for the population parameter
- CI of population mean: Computed as the sample mean ± a margin of error
  - This is the critical value corresponding to the middle XY% of the normal distribution times the standard error of the sampling distribution

$$\bar{x} \pm z^* (s)/\sqrt{n}$$

- Simple 4-step procedure to compute CI

NC STATE UNIVERSITY

Today

- Hypothesis Testing
- Significance Testing

NC STATE UNIVERSITY

Student's Preference Example

- Each student, choose a number between 1 and 100.
- Raise your hand if your chosen number is even
- What is the proportion of even numbers (0.5, > 0.5, < 0.5)
- Is it greater than 0.5 (say 0.51)?
  - If yes, can we conclude that CSC-591 students prefer even numbers?
  - We know that this could be just by random chance
- Now, instead if the proportion is 0.9, then can we conclude that CSC-591 students prefer even numbers? (evidence is much grater than 0.51)
- How much evidence do we need before we can make such a conclusion?
- Hypothesis testing helps us figure that out.

NC STATE UNIVERSITY

Recall from our first lecture

- Does a given fertilizer increase crop yield?
  - Collect and analyze agricultural experimental data
- Basic Steps
  - Set your hypothesis first
  - Design an experiment and collect data
  - Choose appropriate model and find parameters
  - Inferential conclusion on a hypothesis of the involved parameters.
- Today we are going to learn
  - Construction of tests for different types of hypotheses

NC STATE UNIVERSITY

Few Definitions

- Hypothesis
  - Is a statement about a population parameter
- Hypothesis testing
  - Based on a sample from population, goal is to decide on which of the two complementary hypothesis is true
- Null and alternative hypothesis
  - Two complementary hypothesis (null and alternative) – both needs to be specified

NC STATE UNIVERSITY

Few Definitions

- Hypothesis test (or test of significance)
  - Is a standard procedure for testing a claim about a property of a population
  - Is a rule that specifies for which sample values H_0 should be rejected or not rejected
- Test statistic
  - Test is specified in terms of test statistic t(X), such that H_0 is accepted if t falls in a certain interval, otherwise alternate hypothesis is accepted

NC STATE UNIVERSITY

Null and Alternate Hypothesis

- Null hypothesis (H_0), is a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.
- Alternative hypothesis (H_1), is a statistical hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters.
- H_0 and H_1 are not treated on equal basis
  - Usually H_1 corresponds to a more serious or complicated situation that you may want to ascertain with very high degree of certainty
- The outcome of the test is most properly formulated as
  - Reject (or do not) reject H_0

NC STATE UNIVERSITY

State the H_0 and H_1 properly

- EX-1: Chemist invents additive to increase the life of automobile battery. If the mean lifetime of the automobile battery without the additive is 40 months, then what are the H_0 and H_1 hypotheses?

H_0: $\mu = 40$
H_1: $\mu > 40$

This test is called *right-tailed*, since the interest is in an increase only.

NC STATE UNIVERSITY

State the H_0 and H_1 properly

- EX-2: A contractor wishes to lower heating bills by using a special type of insulation in houses. If the average of the monthly heating bills is \$70, then what are the H_0 and H_1 hypotheses?

H_0: $\mu = \$70$
H_1: $\mu < \$70$

This test is called *left-tailed*, since the interest is in lower cost.

NC STATE UNIVERSITY

State the H_0 and H_1 properly

- EX-3: A medical researcher is interested in finding out whether a new medication will have any undesirable side effects. The researcher is particularly concerned with the pulse rate of the patients who take the medication. Will the pulse rate increase, decrease, or remain unchanged after a patient takes the medication? Since the researcher knows that the mean pulse rate for the population under study is 82 beats per minute, then what are the H_0 and H_1 hypotheses?

H_0: $\mu = 82$
H_1: $\mu \neq 82$

This test is called *two-tailed*, since the medication can raise or lower the pulse.

NC STATE UNIVERSITY

Summary

Two-tailed test	Right-tailed test	Left-tailed test
$H_0: \mu = k$ $H_1: \mu \neq k$	$H_0: \mu = k$ $H_1: \mu > k$	$H_0: \mu = k$ $H_1: \mu < k$

- Null hypothesis is always stated with the equals sign
- When a researcher conducts a study, he or she is generally looking for evidence to support a claim. Therefore, the claim should be stated as the alternative hypothesis, i.e., using < or > or ≠. Because of this, the alternative hypothesis is sometimes called the research hypothesis.

NC STATE UNIVERSITY

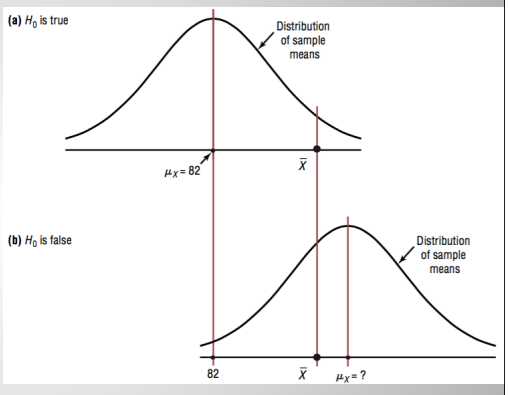
EX-3: Pulse Rate

- After stating the hypothesis
  - designs the study, selects the correct *statistical test,* chooses an appropriate *level of significance,* and formulates a plan for conducting the study.
    - Drug will be administered to a small sample of patients and after allowing a suitable time for the drug to be absorbed, the researcher will measure each person's pulse rate.
- Recall CLT
  - H_0 is true, and the difference between the sample mean and the population mean is due to chance
  - H_0 is false, and the sample came from a population whose mean is not 82 beats per minute but is some other value that is not known

NC STATE UNIVERSITY

Situations in Hypothesis Testing

- The farther away the sample mean is from the population mean, the more evidence there would be for rejecting the null hypothesis. The probability that the sample came from a population whose mean is 82 decreases as the distance or absolute value of the difference between the means increases.
- So, where do we draw the line?

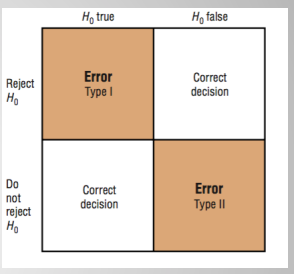


If the difference is significant, then H_0 is rejected

NC STATE UNIVERSITY

Possible Outcomes of a Hypothesis Test

- There 4 possible outcomes
- A type I error occurs if you reject the null hypothesis when it is true.
- A type II error occurs if you do not reject the null hypothesis when it is false.



NC STATE UNIVERSITY

Level of Significance

- The level of significance is the maximum probability of committing a type I error. This probability is symbolized by α. That is, P(type I error) = α.
- *P*(type II error) = β.
  - In most hypothesis-testing situations, β cannot be easily computed; however, α and β are related in that decreasing one increases the other.
- Statisticians generally agree on using three arbitrary significance levels: the 0.10, 0.05, and 0.01 levels.
  - That is, if the null hypothesis is rejected, the probability of a type I error will be 10%, 5%, or 1%, depending on which level of significance is used.

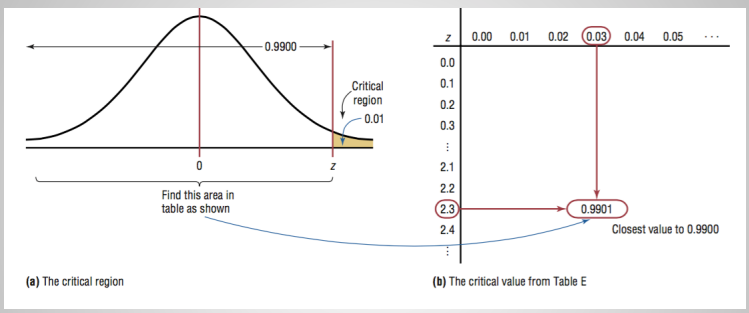
NC STATE UNIVERSITY

Critical Value

- The critical value separates the critical region from the noncritical region. The symbol for critical value is C.V.
- The critical or rejection region is the range of values of the test value that indicates that there is a significant difference and that the null hypothesis should be rejected.
- The noncritical or nonrejection region is the range of values of the test value that indicates that the difference was probably due to chance and that the null hypothesis should not be rejected.
- After a significance level is chosen, a *critical value* is selected from a table for the appropriate test. If a *z* test is used, for example, the *z* table is consulted to find the critical value.

NC STATE UNIVERSITY

Finding Critical Value



- Finding C.V. for α = 0.01 (right-tailed test)

NC STATE UNIVERSITY

Summary

Step 1 Draw the figure and indicate the appropriate area.

a. If the test is left-tailed, the critical region, with an area equal to α, will be on the left side of the mean.

b. If the test is right-tailed, the critical region, with an area equal to α, will be on the right side of the mean.

c. If the test is two-tailed, α must be divided by 2; one-half of the area will be to the right of the mean, and one-half will be to the left of the mean.

Step 2 *a.* For a left-tailed test, use the z value that corresponds to the area equivalent to α in Table E.

b. For a right-tailed test, use the z value that corresponds to the area equivalent to $1 - \alpha$.

c. For a two-tailed test, use the z value that corresponds to $\alpha/2$ for the left value. It will be negative. For the right value, use the z value that corresponds to the area equivalent to $1 - \alpha/2$. It will be positive.

Finding C.V. for specific α values using Z table

NC STATE UNIVERSITY

Practice

- For EX-1 to EX-3, find critical values for α = 0.10 and draw the appropriate figure showing critical region

NC STATE UNIVERSITY

Hypothesis Testing Procedure

Step 1	State the hypotheses and identify the claim.
Step 2	Find the critical value(s) from the appropriate table
Step 3	Compute the test value.
Step 4	Make the decision to reject or not reject the null hypothesis.
Step 5	Summarize the results.

NC STATE UNIVERSITY

z Test

The ***z* test** is a statistical test for the mean of a population. It can be used when $n \geq 30$, or when the population is normally distributed and σ is known.

The formula for the z test is

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

where

$\overline{X}$ = sample mean
μ = hypothesized population mean
σ = population standard deviation
n = sample size

- General form of hypothesis test

$$\frac{\text{(observed value – expected value)}}{\text{standard error}}$$

NC STATE UNIVERSITY

Professors Salaries

- A researcher reports that the average salary of assistant professors is more than $42,000. A sample of 30 assistant professors has a mean salary of $43,260. For α = 0.05, test the claim that assistant professors earn more than $42,000 per year. The standard deviation of the population is $5230.

NC STATE UNIVERSITY

Solution

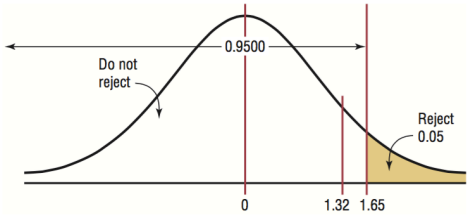
Step 1 State the hypotheses and identify the claim.
H_0: μ = $42,000 and H_1: μ > $42,000 (claim)

Step 2 Find the critical value. Since $\alpha = 0.05$ and the test is a right-tailed test, the critical value is $z = +1.65$.

Step 3 Compute the test value.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\$43{,}260 - \$42{,}000}{\$5230/\sqrt{30}} = 1.32$$

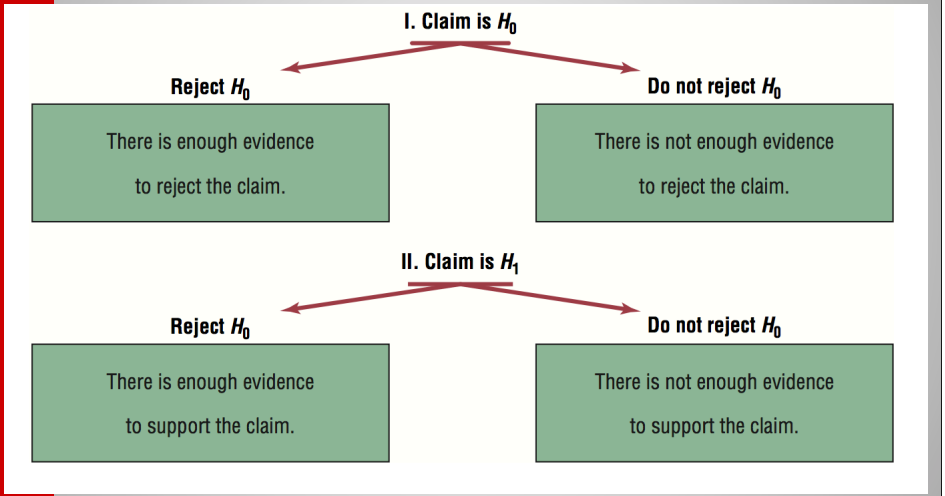
Step 4 Make the decision. Since the test value, +1.32, is less than the critical value, +1.65, and is not in the critical region, the decision is to not reject the null hypothesis. This test is summarized in Figure 8–13.



Step 5 Summarize the results. There is not enough evidence to support the claim that assistant professors earn more on average than $42,000 per year.

NC STATE UNIVERSITY

How to Write Summary Statement



NC STATE UNIVERSITY

t distribution

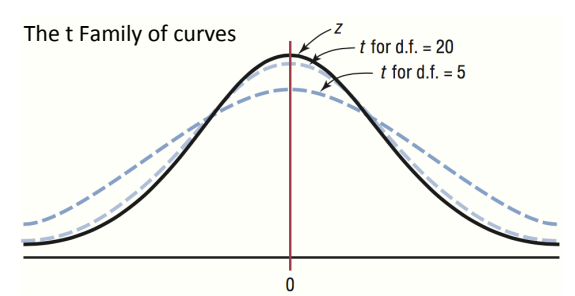
- The *t* distribution is similar to the standard normal distribution in the following ways
  - It is bell-shaped.
  - It is symmetric about the mean.
  - The mean, median, and mode are equal to 0 and are located at the center of the distribution.
  - The curve never touches the *x* axis.
- However, *t* distribution differs from the standard normal distribution in the following ways
  - The variance is greater than 1.
  - The *t* distribution is a family of curves based on the *degrees of freedom*, which is a number related to sample size.
  - As the sample size increases, the *t* distribution approaches the normal distribution.

NC STATE UNIVERSITY

Degrees of freedom

- The degrees of freedom are the number of values that are free to vary after a sample statistic has been computed, and they tell the researcher which specific curve to use when a distribution consists of a family of curves.

For e.g, if the mean of 5 values is 10, then 4 of the 5 values are free to vary. But once 4 values are selected, the fifth value must be a specific number to get a sum of 50, since 50÷5 = 10. Hence, the degrees of freedom are 5-1=4, and this value tells the researcher which *t* curve to use.



NC STATE UNIVERSITY

t Test for a Mean

The ***t* test** is a statistical test for the mean of a population and is used when the population is normally or approximately normally distributed, σ is unknown.

The formula for the *t* test is

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

The degrees of freedom are d.f. $= n - 1$.

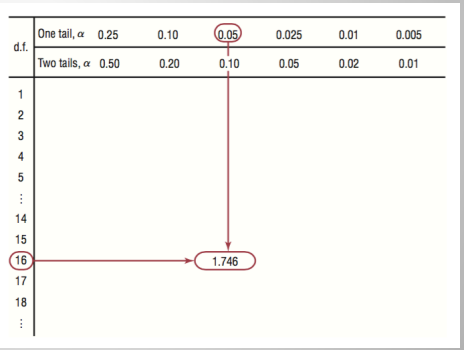
- Like z; the critical values for the *t* test are given in a standard table

NC STATE UNIVERSITY

Example

- Find the critical *t* value for α = 0.05 with d.f. = 16 for a right-tailed *t* test.

Find the 0.05 column in the top row and 16 in the left-hand column. Where the row and column meet, the appropriate critical value is found; it is 1.746.



d.f.	One tail, α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.50	0.20	0.10	0.05	0.02	0.01
1							
2							
3							
4							
5							
⋮							
14							
15							
16				1.746			
17							
18							
⋮							

NC STATE UNIVERSITY

Examples

- Find the critical *t* value for α = 0.01 with d.f. = 22 for a left-tailed test.

  Answer: C.V. = -2.508

- Find the critical values for α = 0.10 with d.f. = 18 for a two-tailed *t* test.
  - Answer: C.V. = +1.734 and -1.734

NC STATE UNIVERSITY

Example

- A medical investigation claims that the average number of infections per week at a hospital in southwestern Pennsylvania is 16.3. A random sample of 10 weeks had a mean number of 17.7 infections. The sample standard deviation is 1.8. Is there enough evidence to reject the investigator's claim at $\alpha = 0.05$?

NC STATE UNIVERSITY

Solution

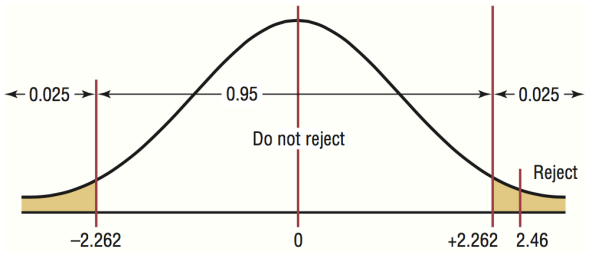
Step 1 H_0: $\mu = 16.3$ (claim) and H_1: $\mu \neq 16.3$.

Step 2 The critical values are $+2.262$ and -2.262 for $\alpha = 0.05$ and d.f. $= 9$.

Step 3 The test value is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{17.7 - 16.3}{1.8/\sqrt{10}} = 2.46$$

Step 4 Reject the null hypothesis since $2.46 > 2.262$. See Figure 8–22.



Step 5 There is enough evidence to reject the claim that the average number of infections is 16.3.

NC STATE UNIVERSITY

Acknowledgements

- J. Lattin, R. Johnson, Rice, Diez, Bluman, Triola, Dekking, Devore, Carlton