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

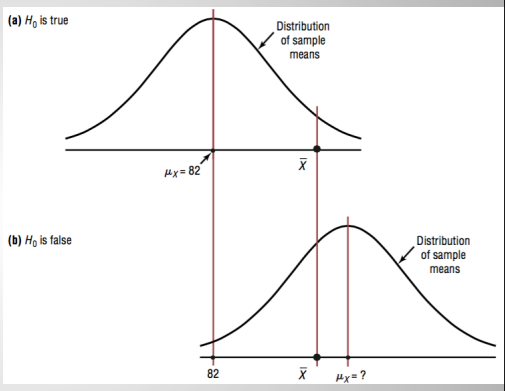
- Changing final exam date is not possible, see policy here
  - https://policies.ncsu.edu/regulation/reg-02-20-14

NC STATE UNIVERSITY

Key points from 9/15

- Null (H_0) and Alternate (H_1) Hypothesis
- Type I and Type II errors

	H_0 true	H_0 false
Reject H_0	**Error** Type I	Correct decision
Do not reject H_0	Correct decision	**Error** Type II



If the difference is significant, then H_0 is rejected

NC STATE UNIVERSITY

Today

- Significance testing
- Testing for variance
- Testing for difference between two means

NC STATE UNIVERSITY

p-Value Method for Hypothesis Testing

The ***P*-value** (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true.

- In other words, the P-value is the actual area under the standard normal curve (or other curve, depending on statistical test used) representing the probability of a particular sample statistic or a more extreme sample statistic occurring if the null hypothesis is true

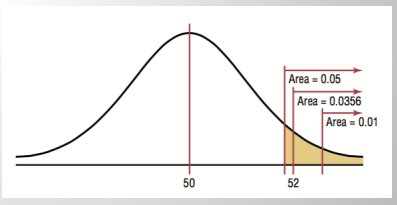
NC STATE UNIVERSITY

Example

- H_0: $\mu = 50$; H_1: $\mu > 50$; Let's say P-value for a statistical test is 0.0356, then probability of getting a sample mean of 52 or grater is 0.0356 if the true population mean is 50 (for given sample size and s.d.

The relationship between the *P*-value and the α value can be explained in this manner.

For $P = 0.0356$, the null hypothesis would be rejected at $\alpha=0.05$ but not at $\alpha=0.01$



NOTE: When the hypothesis test is two-tailed, the area in one tail must be doubled. For a two-tailed test, if $\alpha=0.05$ and the area in one tail is 0.0356, the *P*-value will be 2(0.0356)=0.0712. Now, do you reject H_0?

NC STATE UNIVERSITY

Summary of P-Value Procedure

Step 1	State the hypotheses and identify the claim.
Step 2	Compute the test value.
Step 3	Find the *P*-value.
Step 4	Make the decision.
Step 5	Summarize the results.

NC STATE UNIVERSITY

Example

- A researcher wishes to test the claim that the average cost of tuition and fees at a four-year public college is greater than $5700. She selects a random sample of 36 four-year public colleges and finds the mean to be $5950. The population standard deviation is $659. Is there evidence to support the claim at a 0.05? Use the *P*-value method.

NC STATE UNIVERSITY

Solution

Step 1 State the hypotheses and identify the claim. H_0: = \$5700 and H_1: $\mu >$ \$5700 (claim).

Step 2 Compute the test value.

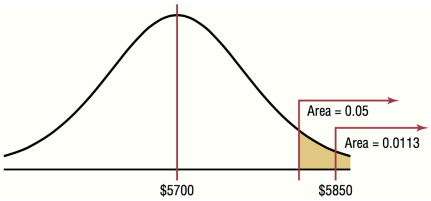
$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{5950 - 5700}{659/\sqrt{36}} = 2.28$$

Step 3 Find the *P*-value. Using Table (here z table) find the corresponding area under the normal distribution for $z = 2.28$. It is 0.9887. Subtract this value for the area from 1.0000 to find the area in the right tail.

$1.0000 - 0.9887 = 0.0113$

Hence the *P*-value is 0.0113.

Step 4 Make the decision. Since the *P*-value is less than 0.05, the decision is to reject the null hypothesis. See Figure below.



Step 5 Summarize the results. There is enough evidence to support the claim that the tuition and fees at four-year public colleges are greater than \$5700.

Note: Had the researcher chosen $\alpha = 0.01$, the null hypothesis would not have been rejected since the *P*-value (0.0113) is greater than 0.01.

NC STATE UNIVERSITY

P-Value Decision Rule

- Normal curve is a useful visualization tool (z-value, p-value, critical region, etc.), however you do not need it to make decision.

Decision Rule When Using a *P*-Value

If *P*-value $\leq \alpha$, reject the null hypothesis.

If *P*-value $> \alpha$, do not reject the null hypothesis.

NC STATE UNIVERSITY

Important Points About P-Value

- First key difference: The α value is chosen by the researcher before the statistical test is conducted. The P-value is computed after the sample mean has been found.
- However, some researchers do not chose an α value, but report P-value and allow reader to decide whether the null hypothesis should be rejected. In that case, use the following guidelines

Guidelines for *P*-Values

If P-value ≤ 0.01, reject the null hypothesis. The difference is highly significant.

If P-value > 0.01 but P-value ≤ 0.05, reject the null hypothesis. The difference is significant.

If P-value > 0.05 but P-value ≤ 0.10, consider the consequences of type I error before rejecting the null hypothesis.

If P-value > 0.10, do not reject the null hypothesis. The difference is not significant.

NC STATE UNIVERSITY

Relationship With C.I.

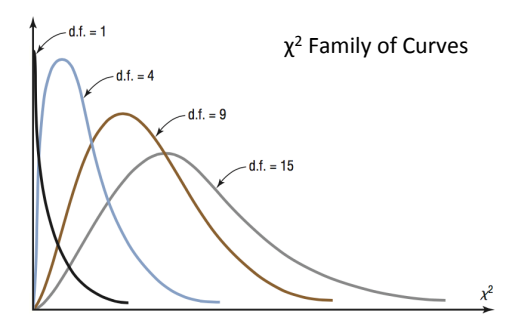
- When the null hypothesis is rejected in a hypothesis-testing situation, the confidence interval for the mean using the same level of significance *will not* contain the hypothesized mean. Likewise, when the null hypothesis is not rejected, the confidence interval computed using the same level of significance *will* contain the hypothesized mean.

NC STATE UNIVERSITY

What About Variance/SD

- Chi-square test for a Variance or Standard Deviation: χ^2
- χ^2 Distribution

The distribution is obtained from the values of $(n-1)s^2/\sigma^2$ when the samples are drawn from a normal distribution with variance σ^2

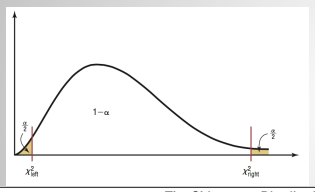


A chi-square variable cannot be negative, and the distributions are skewed to the right. At about 100 degrees of freedom, the chi-square distribution becomes somewhat symmetric. The area under each chi-square distribution is equal to 1.00, or 100%.

NC STATE UNIVERSITY

How to find χ^2 Values?

- Find the values for χ^2_{right} and χ^2_{left} for a 90% c.i. when n = 25?



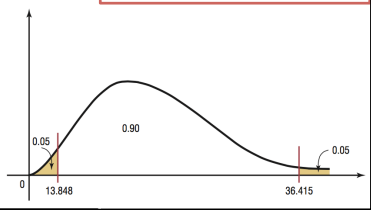
To find χ^2_{right}, subtract 1-0.90=0.10 and divide by 2 to get 0.05.
To find χ^2_{left}, subtract 1-0.05=0.95. Hence, use the 0.95 and 0.05 columns and the row corresponding to 24 d.f.

Two different values are used in the formula because the distribution is not symmetric

The Chi-square Distribution

Degrees of freedom	α 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1										
2										
⋮										
24							13.848			36.415

χ^2_{left} χ^2_{right}



NC STATE UNIVERSITY

C.I. for a Variance and S.D.

Formula for the Confidence Interval for a Variance

$$\frac{(n-1)s^2}{\chi^2_{\text{right}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{left}}}$$

$$\text{d.f.} = n - 1$$

Formula for the Confidence Interval for a Standard Deviation

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\text{right}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\text{left}}}}$$

$$\text{d.f.} = n - 1$$

NC STATE UNIVERSITY

Practice Question

- Find the 95% C.I. for the variance and standard deviation of the nicotine content of a cigarettes manufactured if a sample of 20 cigarettes has a standard deviation of 1.6 milligrams.

NC STATE UNIVERSITY

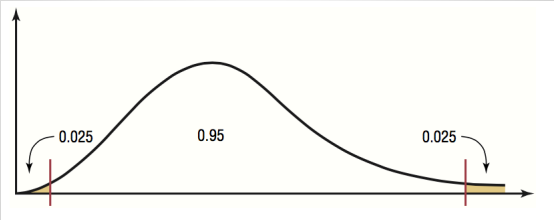
χ^2 test for a Variance and S.D.

- To find area under the chi-square distribution
  - If test is right-tailed, directly get the value at the intersection of α and d.f.
  - If test is left-tailed, then get the value at the intersection of (1-α) and d.f.
  - If the test is two-tailed, note that the area to the right is α/2 and area to the left is (1 – α/2), then find corresponding values from the table

NC STATE UNIVERSITY

Example

- Find the critical chi-square values for n=23 and α=0.05 when a two-tailed test is conducted



What are the α values for right and left?
Then look for corresponding values for d.f. = 22 in the table

NC STATE UNIVERSITY

χ^2 test for a Single Variance

Formula for the Chi-Square Test for a Single Variance

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

with degrees of freedom equal to $n - 1$ and where

n = sample size
s^2 = sample variance
σ^2 = population variance

Assumptions:
- The sample must be randomly selected from the population
- The population must be normally distributed for the variable under study.
- The observations must be independent of one another.

NC STATE UNIVERSITY

Hypothesis testing using χ^2

1. State the hypotheses and identify the claim.
2. Find the critical value(s).
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

NC STATE UNIVERSITY

Practice Examples

- An instructor wishes to see whether the variation in scores of the 23 students in her class is less than the variance of the population. The variance of the class is 198. Is there enough evidence to support the claim that the variation of the students is less than the population variance ($\sigma^2 = 225$) at $\alpha = 0.05$? Assume that the scores are normally distributed.
  - There is not enough evidence to support the claim that the variation in test scores of the instructor's students is less than the variation in scores of the population.

NC STATE UNIVERSITY

Other Tests

- Kolmogorov-Smirnov test
- Likelihood ratio test
- Bayesian test

NC STATE UNIVERSITY

Testing the Difference Between Two Means

- Suppose one wishes to determine the differences in average GRE scores of engineering graduate students admitted to NCSU and Duke. In other words, does the mean GRE scores of NCSU engineering differs from mean GRE scores of Duke?
- Here the hypotheses are:

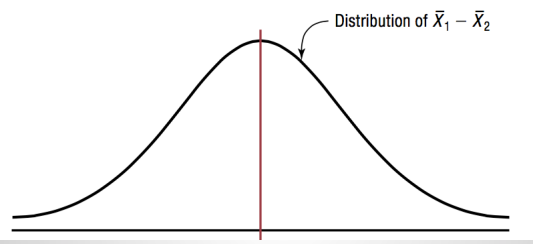
$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

Alternatively

$$H_0: \mu_1 - \mu_2 = 0 \quad H_1: \mu_1 - \mu_2 \neq 0$$

NC STATE UNIVERSITY

Distribution of $\bar{x}_1$ - $\bar{x}_2$



The variance of the difference is equal to the sum of the individual variances.

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}$$

where $\sigma^2_{\bar{X}_1} = \frac{\sigma_1^2}{n_1}$ and $\sigma^2_{\bar{X}_2} = \frac{\sigma_2^2}{n_2}$

So the standard deviation of $\bar{X}_1 - \bar{X}_2$ is

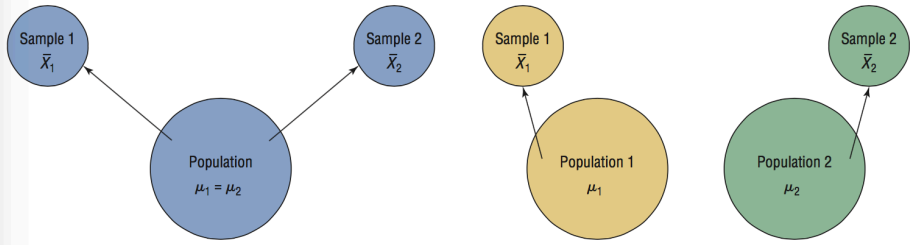
$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Formula for z Test

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

NC STATE UNIVERSITY

Hypothesis-Testing Situations



(a) Difference is not significant

Do not reject H_0: $\mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is not significant.

(b) Difference is significant

Reject H_0: $\mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is significant.

These tests can also be one-tailed, using the following hypotheses:

Right-tailed			Left-tailed		
H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$	or	H_0: $\mu_1 - \mu_2 = 0$ H_1: $\mu_1 - \mu_2 > 0$	H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 < \mu_2$	or	H_0: $\mu_1 - \mu_2 = 0$ H_1: $\mu_1 - \mu_2 < 0$

NC STATE UNIVERSITY

Basic Format of Hypothesis Testing

1. State the hypotheses and identify the claim.
2. Find the critical value(s).
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

NC STATE UNIVERSITY

Example

- A survey found that the average hotel room rate in Ashville is $88.42 and the average room rate in Raleigh is $80.61. Assume that the data were obtained from two samples of 50 hotels each and that the standard deviations of the populations are $5.62 and $4.83, respectively. At α=0.05, can it be concluded that there is a significant difference in the rates?

NC STATE UNIVERSITY

Solution

Step 1 State the hypotheses and identify the claim.

$H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$ (claim)

Step 2 Find the critical values. Since $\alpha = 0.05$, the critical values are $+1.96$ and -1.96.

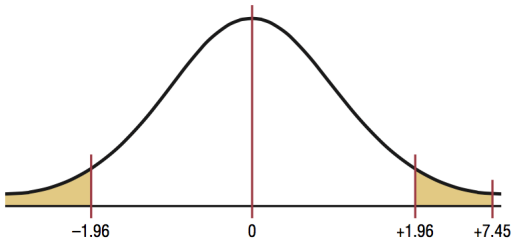
Step 3 Compute the test value.

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(88.42 - 80.61) - 0}{\sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}}} = 7.45$$

NC STATE UNIVERSITY

Solution

Step 4 Make the decision. Reject the null hypothesis at $\alpha = 0.05$, since $7.45 > 1.96$.



Step 5 Summarize the results. There is enough evidence to support the claim that the means are not equal. Hence, there is a significant difference in the rates.

NC STATE UNIVERSITY

Practice Example

- Determine the P-Value and make the decision
  - Given: value of two-tailed test is 1.40
- P-value can be determined using same procedure as described earlier

NC STATE UNIVERSITY

T Test to Determine the Difference

Variances are assumed to be unequal:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where the degrees of freedom are equal to the smaller of $n_1 - 1$ or $n_2 - 1$.

NC STATE UNIVERSITY

Practice Example

- The average size of a farm in Indiana County, Pennsylvania, is 191 acres. The average size of a farm in Greene County, Pennsylvania, is 199 acres. Assume the data were obtained from two samples with standard deviations of 38 and 12 acres, respectively, and sample sizes of 8 and 10, respectively. Can it be concluded at $\alpha = 0.05$ that the average size of the farms in the two counties is different? Assume the populations are normally distributed.

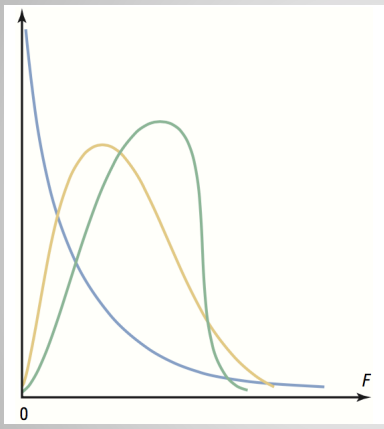
NC STATE UNIVERSITY

Testing the Difference Between Two Variances

- Is there a variation in the temperature for a given month between two different cities?
- Comparing the variance of the cholesterol of men with the variance of cholesterol of women
- For comparing two variances or standard deviations, an *F* test is used.

NC STATE UNIVERSITY

F Distribution



F distribution for different d.f.

If two independent samples are selected from two normally distributed populations in which the variances are equal ($\sigma_1^2 = \sigma_2^2$) and if the variances s^2_1 and s^2_2 are compared as s^2_1 / s^2_2, the sampling distribution of the variances is called the ***F* distribution.**

Characteristics of *F* distribution

- The values of *F* cannot be negative, because variances are always positive or zero.
- The distribution is positively skewed.
- The mean value of *F* is approximately equal to 1.
- The *F* distribution is a family of curves based on the degrees of freedom of the variance of the numerator and the degrees of freedom of the variance of the denominator.

NC STATE UNIVERSITY

Formulae for *F* Test

$$F = \frac{s_1^2}{s_2^2}$$

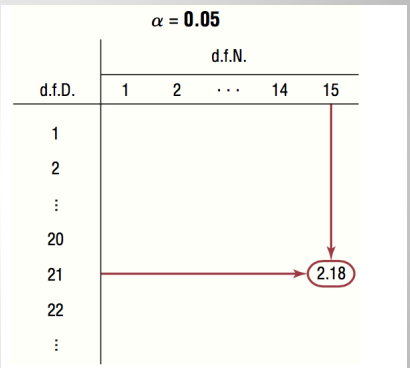
where the larger of the two variances is placed in the numerator regardless of the subscripts.

Note: When you are finding the *F* test value, *the larger of the variances is placed in the numerator of the F formula;* this is not necessarily the variance of the larger of the two sample sizes.

NC STATE UNIVERSITY

Finding C.V.

- Find the critical value for a right-tailed *F* test when α = 0.05, the degrees of freedom for the numerator (abbreviated d.f.N.) are 15, and the degrees of freedom for the denominator (d.f.D.) are 21.



NC STATE UNIVERSITY

Notes and Assumptions

Notes for the Use of the *F* Test

1. The larger variance should always be placed in the numerator of the formula regardless of the subscripts. (See note on page 518.)

$$F = \frac{s_1^2}{s_2^2}$$

2. For a two-tailed test, the α value must be divided by 2 and the critical value placed on the right side of the F curve.
3. If the standard deviations instead of the variances are given in the problem, they must be squared for the formula for the F test.
4. When the degrees of freedom cannot be found in Table H, the closest value on the smaller side should be used.

Assumptions for Testing the Difference Between Two Variances

1. The populations from which the samples were obtained must be normally distributed. (*Note:* The test should not be used when the distributions depart from normality.)
2. The samples must be independent of each other.

NC STATE UNIVERSITY

Example

- A medical researcher wishes to see whether the variance of the heart rates (in beats per minute) of smokers is different from the variance of heart rates of people who do not smoke. Two samples are selected, and the data are as shown. Using a 0.05, is there enough evidence to support the claim?

Smokers	Nonsmokers
$n_1 = 26$	$n_2 = 18$
$s_1^2 = 36$	$s_2^2 = 10$

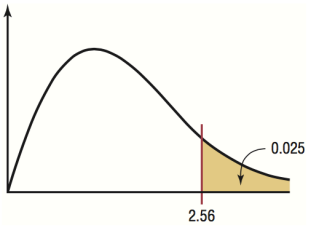
NC STATE UNIVERSITY

Solution

Step 1 State the hypotheses and identify the claim.

$H_0: \sigma_1^2 = \sigma_2^2$ and $H_1: \sigma_1^2 \neq \sigma_2^2$ (claim)

Step 2 Find the critical value. Use the 0.025 table in Table H since $\alpha = 0.05$ and this is a two-tailed test. Here, d.f.N. $= 26 - 1 = 25$, and d.f.D. $= 18 - 1 = 17$. The critical value is 2.56 (d.f.N. $= 24$ was used). See Figure below



Step 3 Compute the test value.

$$F = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3.6$$

Step 4 Make the decision. Reject the null hypothesis, since $3.6 > 2.56$.

Step 5 Summarize the results. There is enough evidence to support the claim that the variance of the heart rates of smokers and nonsmokers is different.

NC STATE UNIVERSITY

Acknowledgements

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