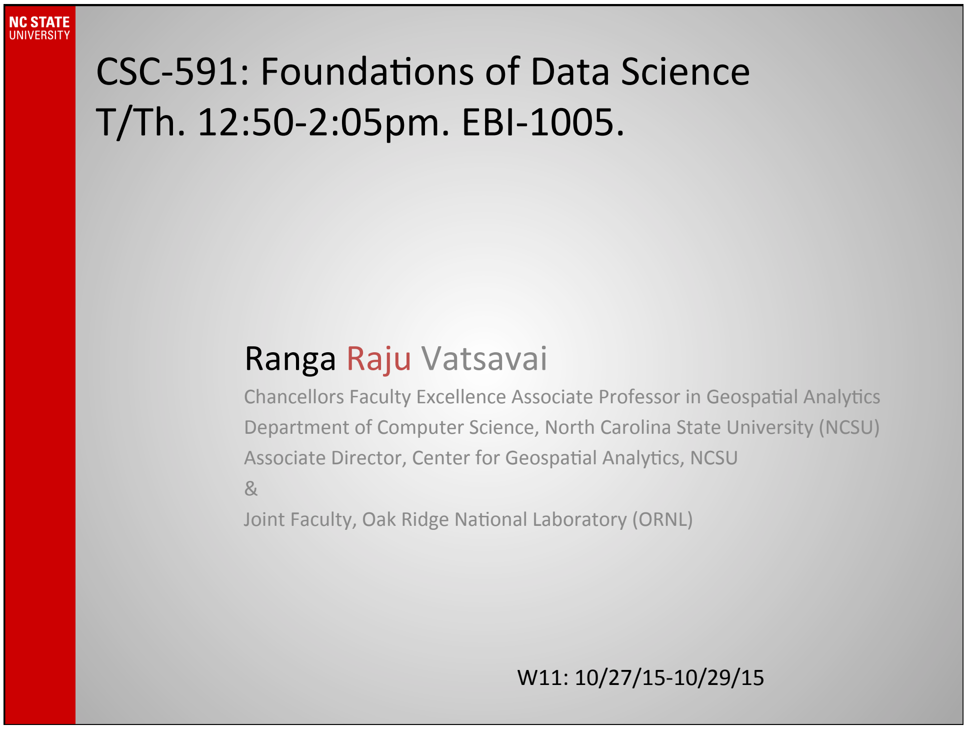


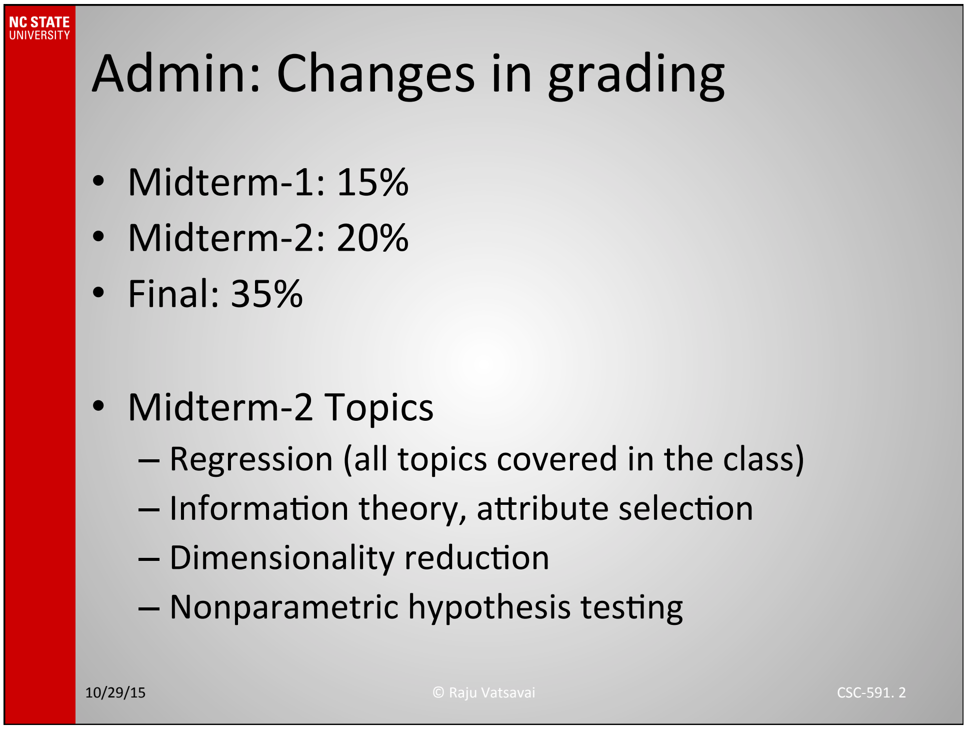
NC STATE UNIVERSITY

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)

W11: 10/27/15-10/29/15

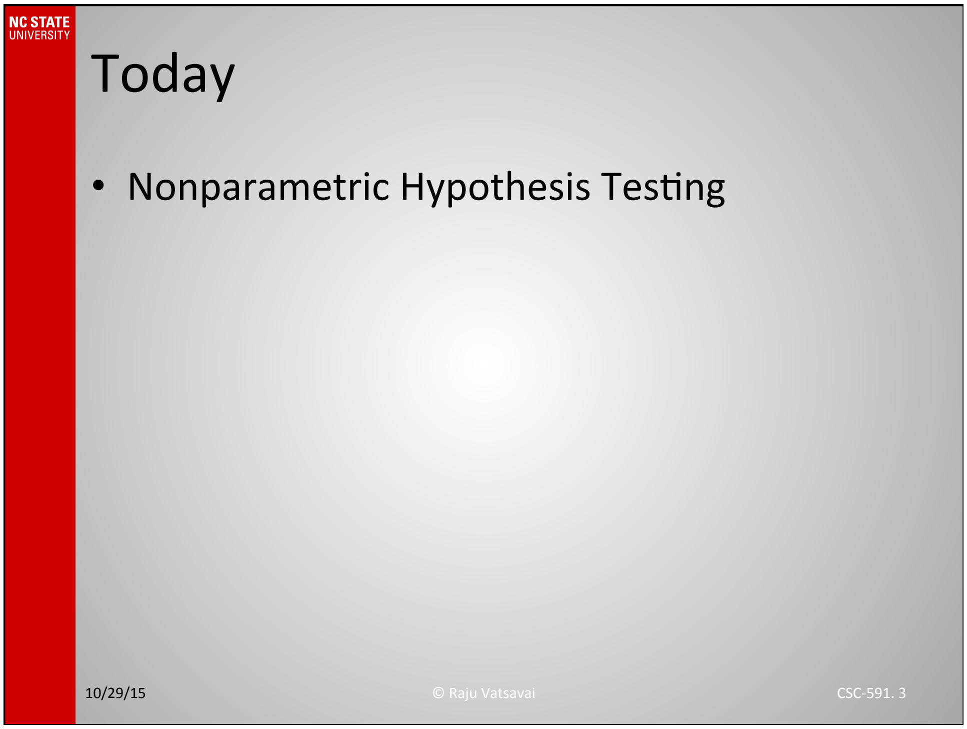


NC STATE UNIVERSITY

Admin: Changes in grading

- Midterm-1: 15%
- Midterm-2: 20%
- Final: 35%

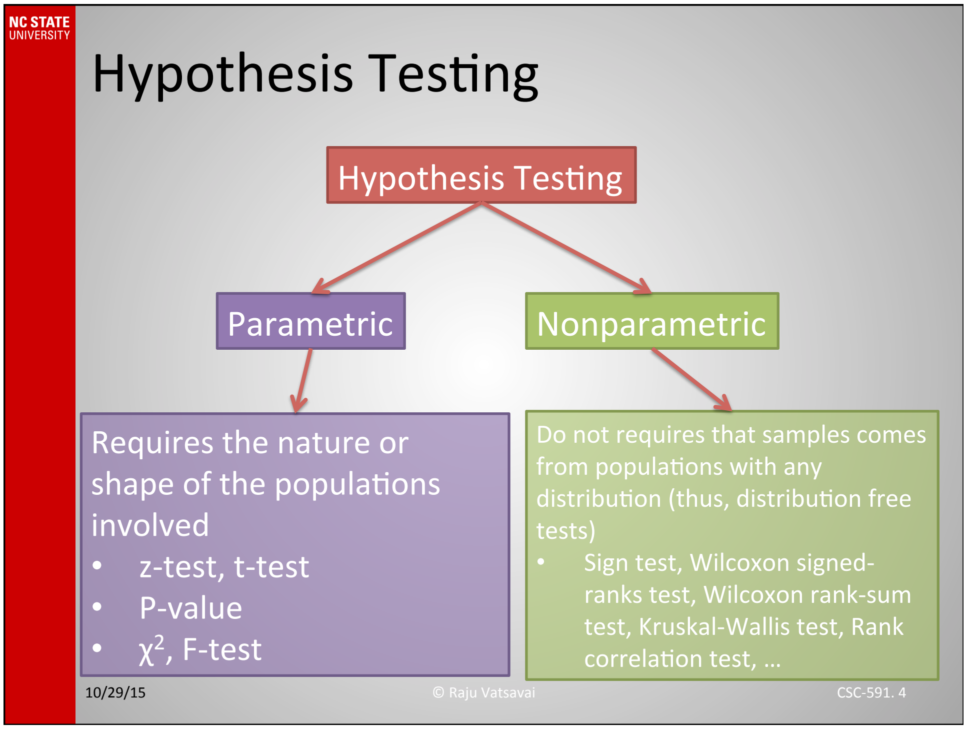
- Midterm-2 Topics
  – Regression (all topics covered in the class)
  – Information theory, attribute selection
  – Dimensionality reduction
  – Nonparametric hypothesis testing



NC STATE UNIVERSITY

Today

- Nonparametric Hypothesis Testing



NC STATE UNIVERSITY

Hypothesis Testing

Hypothesis Testing

Parametric

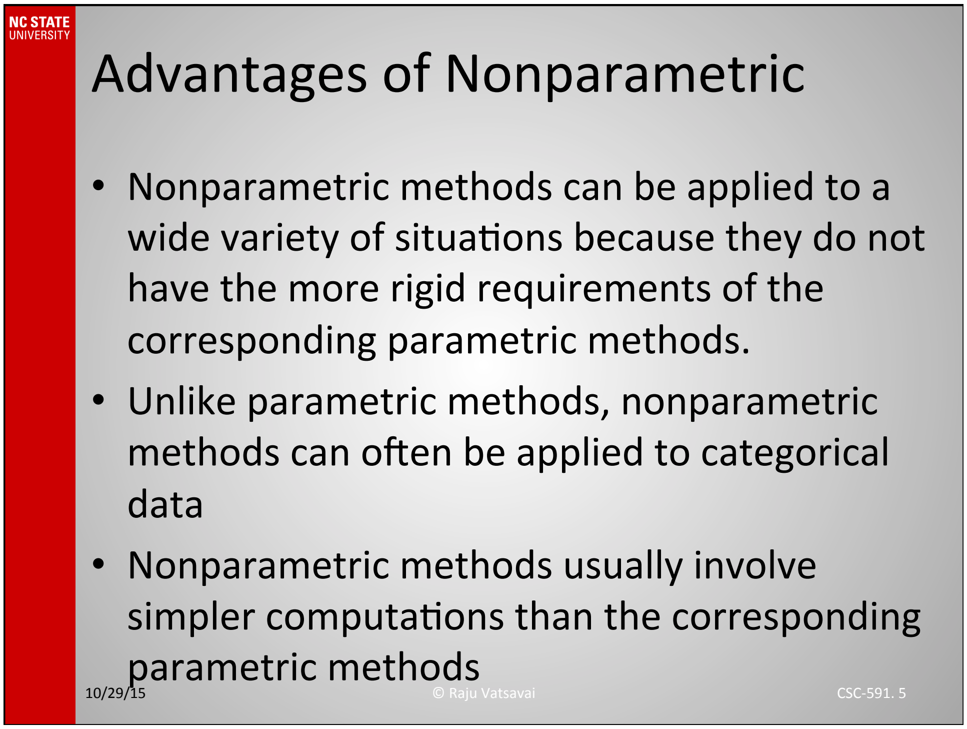
Nonparametric

Requires the nature or shape of the populations involved

- z-test, t-test
- P-value
- χ^2, F-test

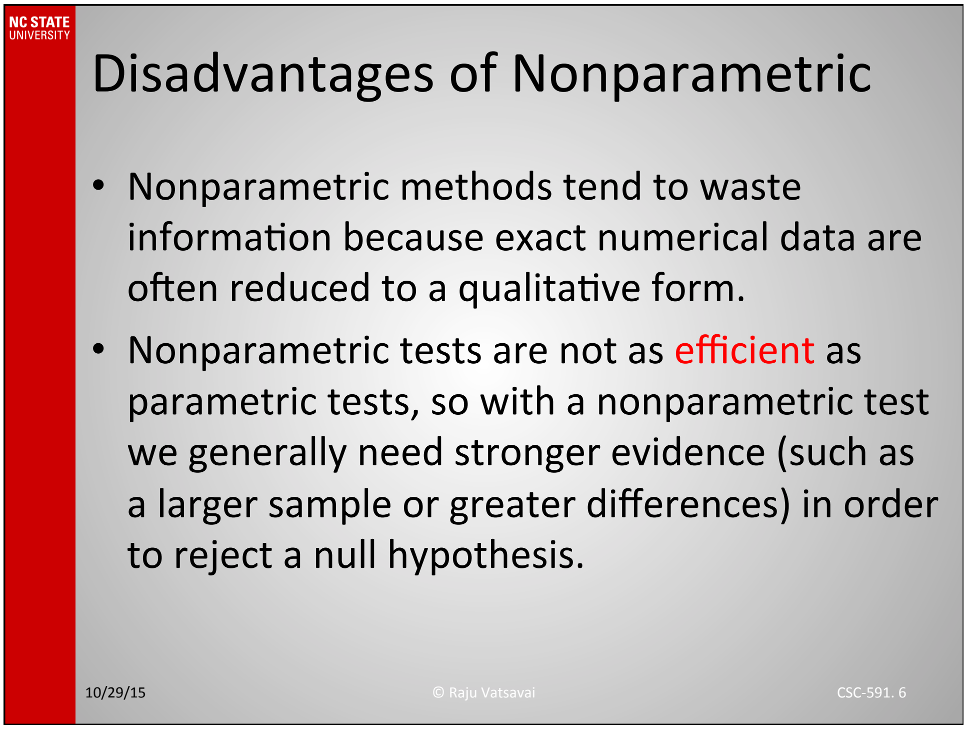
Do not requires that samples comes from populations with any distribution (thus, distribution free tests)

- Sign test, Wilcoxon signed-ranks test, Wilcoxon rank-sum test, Kruskal-Wallis test, Rank correlation test, …



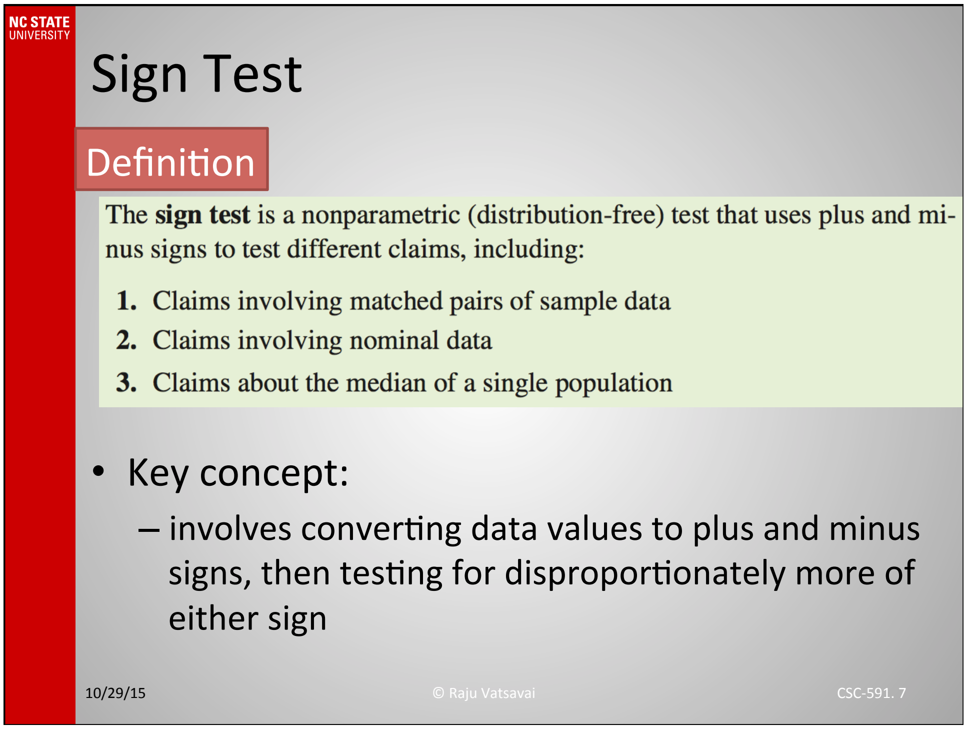
Advantages of Nonparametric

- Nonparametric methods can be applied to a wide variety of situations because they do not have the more rigid requirements of the corresponding parametric methods.
- Unlike parametric methods, nonparametric methods can often be applied to categorical data
- Nonparametric methods usually involve simpler computations than the corresponding parametric methods



Disadvantages of Nonparametric

- Nonparametric methods tend to waste information because exact numerical data are often reduced to a qualitative form.
- Nonparametric tests are not as efficient as parametric tests, so with a nonparametric test we generally need stronger evidence (such as a larger sample or greater differences) in order to reject a null hypothesis.



NC STATE UNIVERSITY

Sign Test

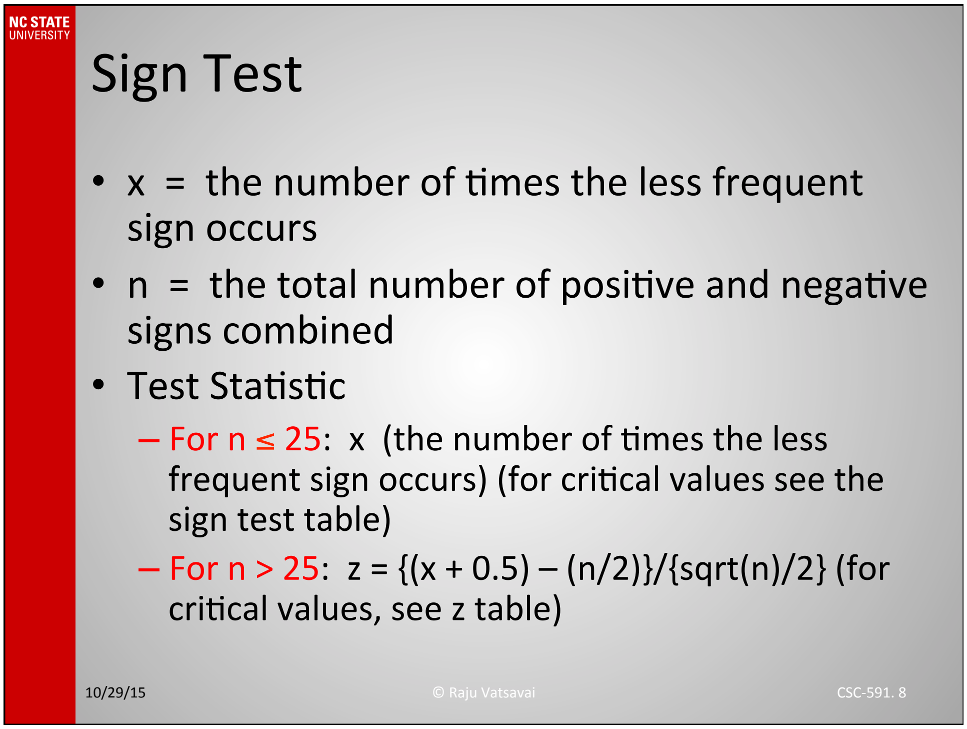
Definition

The **sign test** is a nonparametric (distribution-free) test that uses plus and minus signs to test different claims, including:

1. Claims involving matched pairs of sample data
2. Claims involving nominal data
3. Claims about the median of a single population

- Key concept:
  - involves converting data values to plus and minus signs, then testing for disproportionately more of either sign

10/29/15 © Raju Vatsavai CSC-591. 7

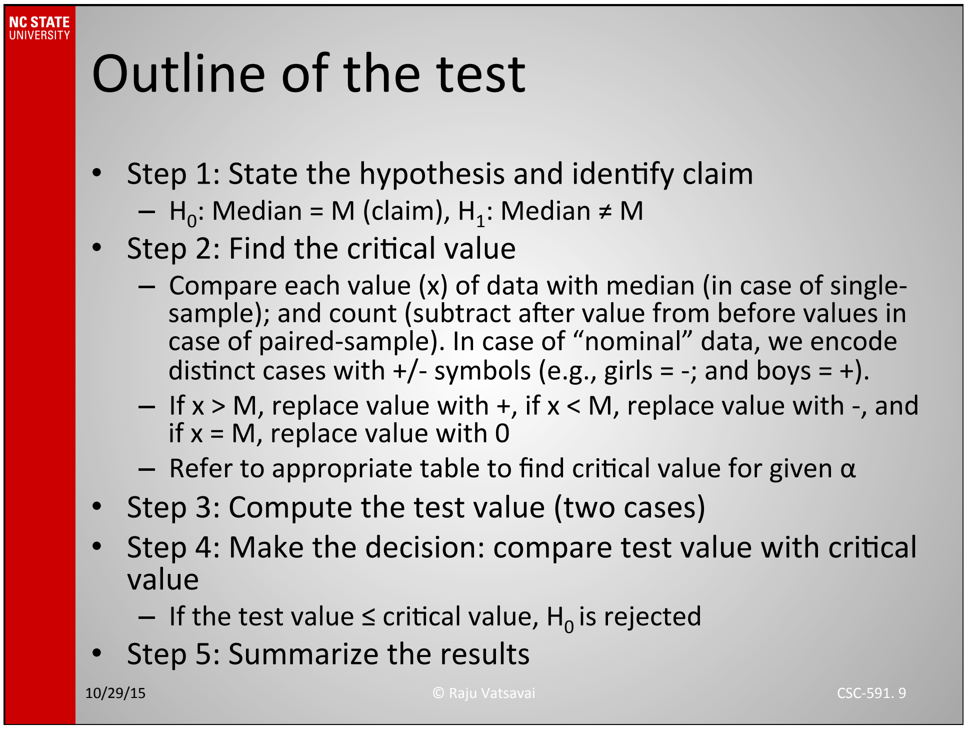


NC STATE UNIVERSITY

Sign Test

- x = the number of times the less frequent sign occurs
- n = the total number of positive and negative signs combined
- Test Statistic
  - For $n \leq 25$: x (the number of times the less frequent sign occurs) (for critical values see the sign test table)
  - For $n > 25$: z = {(x + 0.5) – (n/2)}/{sqrt(n)/2} (for critical values, see z table)

10/29/15 © Raju Vatsavai CSC-591. 8

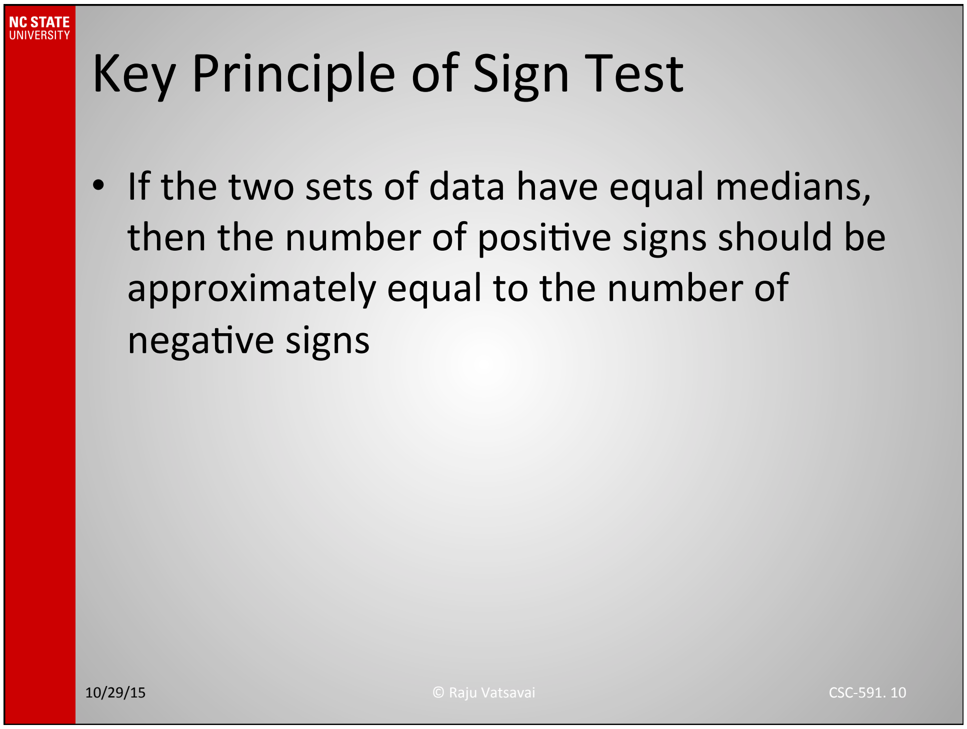


NC STATE UNIVERSITY

Outline of the test

- Step 1: State the hypothesis and identify claim
  - H_0: Median = M (claim), H_1: Median ≠ M
- Step 2: Find the critical value
  - Compare each value (x) of data with median (in case of single-sample); and count (subtract after value from before values in case of paired-sample). In case of “nominal” data, we encode distinct cases with +/- symbols (e.g., girls = -; and boys = +).
  - If x > M, replace value with +, if x < M, replace value with -, and if x = M, replace value with 0
  - Refer to appropriate table to find critical value for given α
- Step 3: Compute the test value (two cases)
- Step 4: Make the decision: compare test value with critical value
  - If the test value ≤ critical value, H_0 is rejected
- Step 5: Summarize the results

10/29/15 © Raju Vatsavai CSC-591. 9

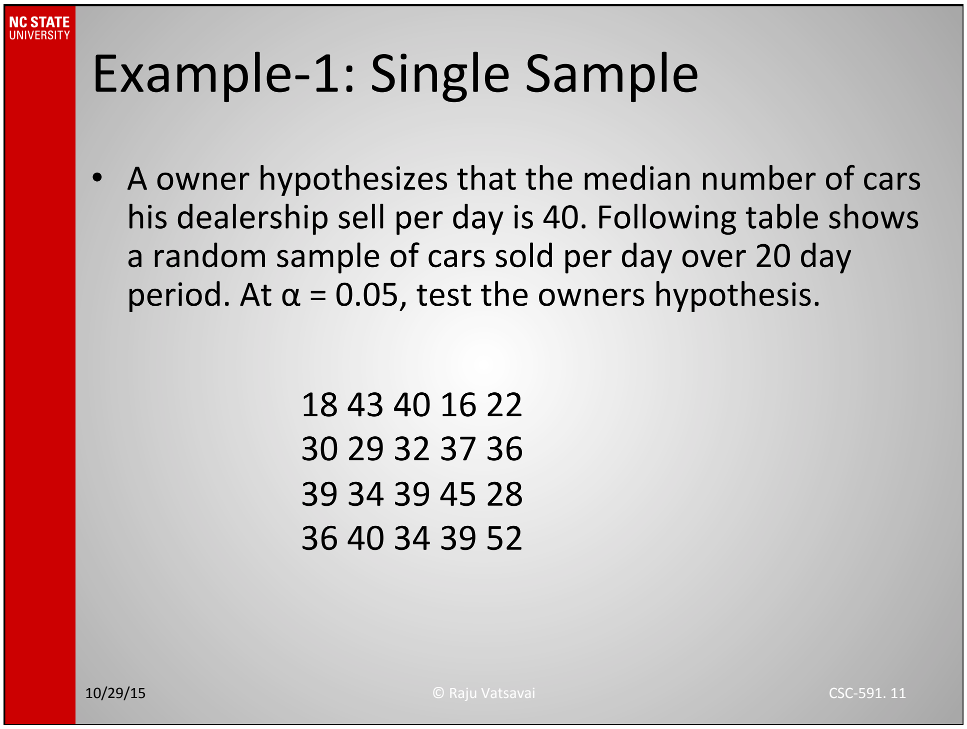


NC STATE UNIVERSITY

Key Principle of Sign Test

- If the two sets of data have equal medians, then the number of positive signs should be approximately equal to the number of negative signs

10/29/15 © Raju Vatsavai CSC-591. 10

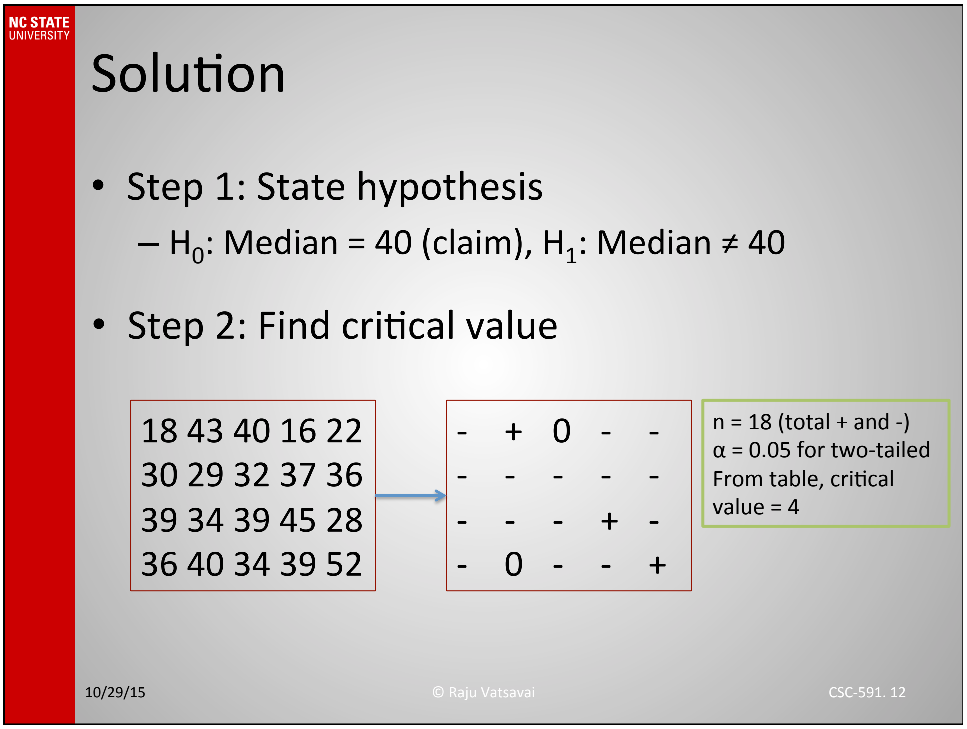


NC STATE UNIVERSITY

Example-1: Single Sample

- A owner hypothesizes that the median number of cars his dealership sell per day is 40. Following table shows a random sample of cars sold per day over 20 day period. At α = 0.05, test the owners hypothesis.

18 43 40 16 22
30 29 32 37 36
39 34 39 45 28
36 40 34 39 52



NC STATE UNIVERSITY

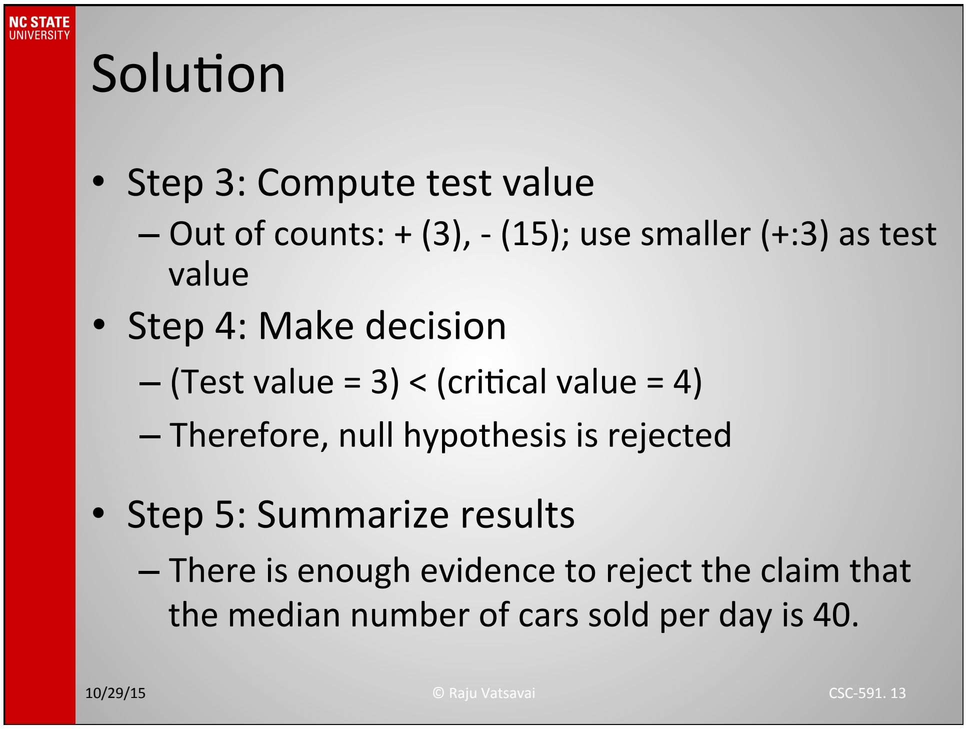
Solution

- Step 1: State hypothesis
  - H_0: Median = 40 (claim), H_1: Median ≠ 40
- Step 2: Find critical value

18	43	40	16	22
30	29	32	37	36
39	34	39	45	28
36	40	34	39	52

-	+	0	-	-
-	-	-	-	-
-	-	-	+	-
-	0	-	-	+

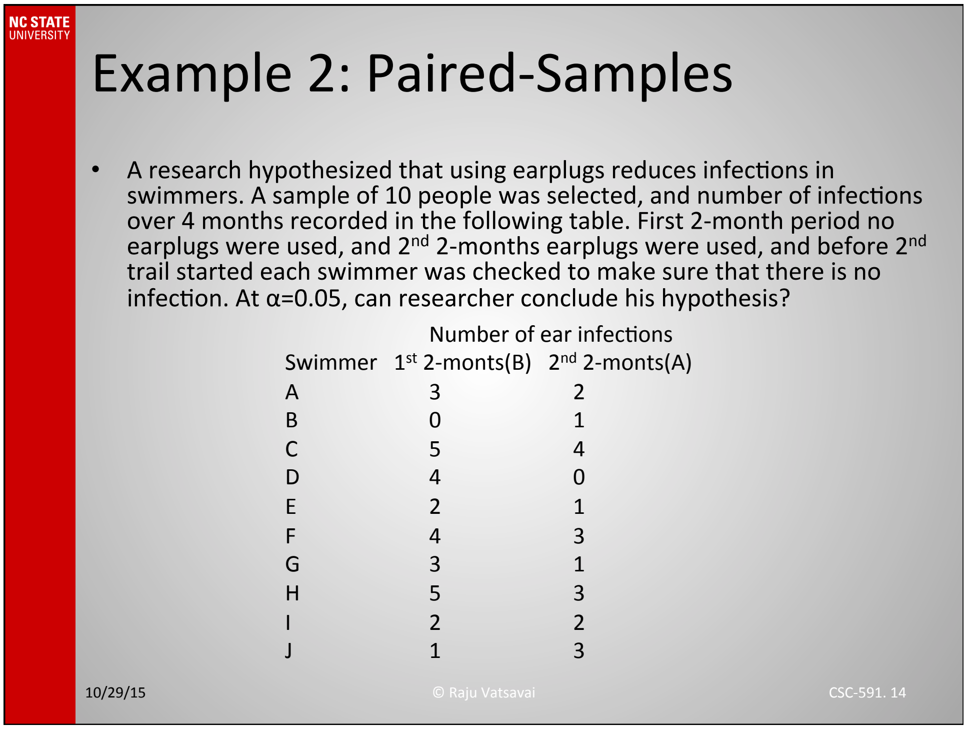
n = 18 (total + and -)
α = 0.05 for two-tailed
From table, critical value = 4



NC STATE UNIVERSITY

Solution

- Step 3: Compute test value
  - Out of counts: + (3), - (15); use smaller (+:3) as test value
- Step 4: Make decision
  - (Test value = 3) < (critical value = 4)
  - Therefore, null hypothesis is rejected
- Step 5: Summarize results
  - There is enough evidence to reject the claim that the median number of cars sold per day is 40.

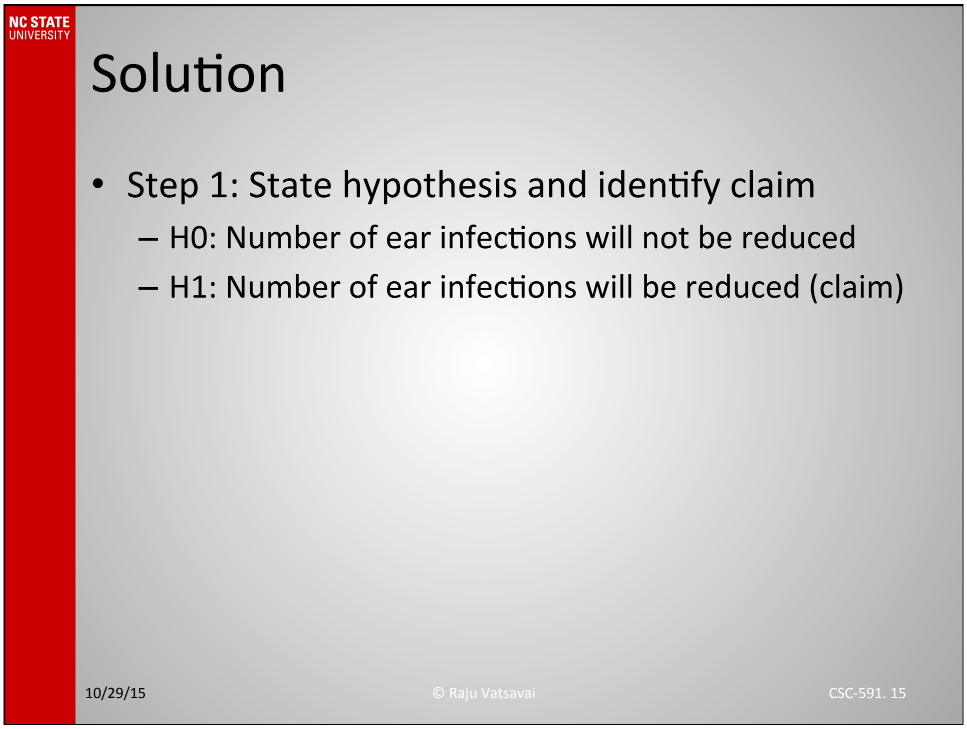


NC STATE UNIVERSITY

Example 2: Paired-Samples

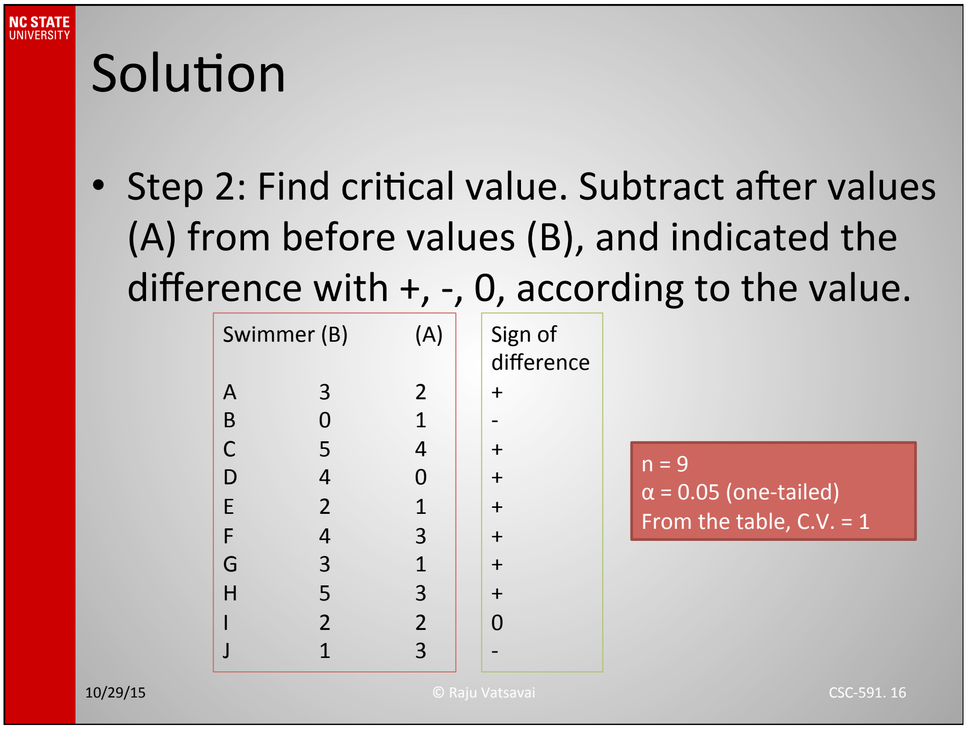
- A research hypothesized that using earplugs reduces infections in swimmers. A sample of 10 people was selected, and number of infections over 4 months recorded in the following table. First 2-month period no earplugs were used, and 2nd 2-months earplugs were used, and before 2nd trail started each swimmer was checked to make sure that there is no infection. At α=0.05, can researcher conclude his hypothesis?

	Number of ear infections	
Swimmer	1st 2-monts(B)	2nd 2-monts(A)
A	3	2
B	0	1
C	5	4
D	4	0
E	2	1
F	4	3
G	3	1
H	5	3
I	2	2
J	1	3



Solution

- Step 1: State hypothesis and identify claim
  - H0: Number of ear infections will not be reduced
  - H1: Number of ear infections will be reduced (claim)

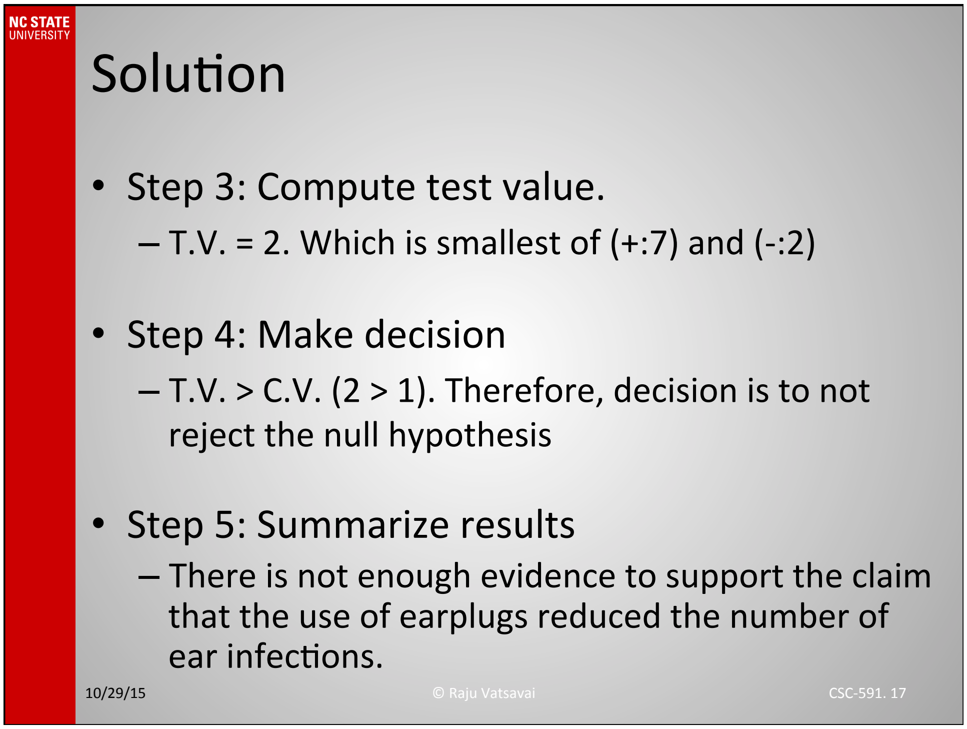


Solution

- Step 2: Find critical value. Subtract after values (A) from before values (B), and indicated the difference with +, -, 0, according to the value.

Swimmer	(B)	(A)	Sign of difference
A	3	2	+
B	0	1	-
C	5	4	+
D	4	0	+
E	2	1	+
F	4	3	+
G	3	1	+
H	5	3	+
I	2	2	0
J	1	3	-

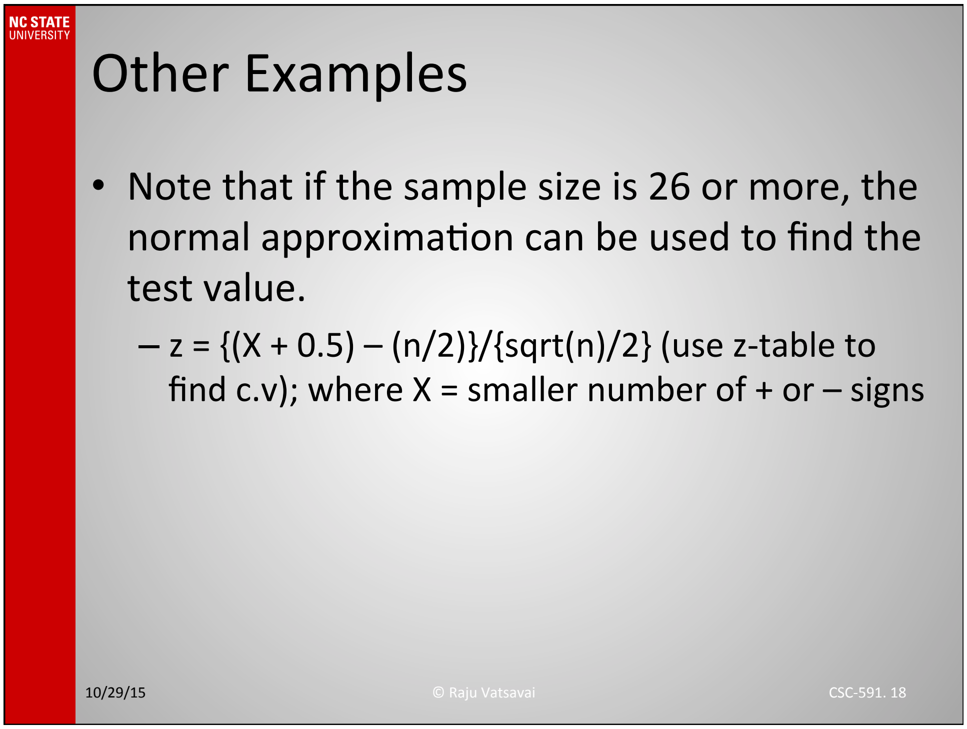
n = 9
$\alpha = 0.05$ (one-tailed)
From the table, C.V. = 1



NC STATE UNIVERSITY

Solution

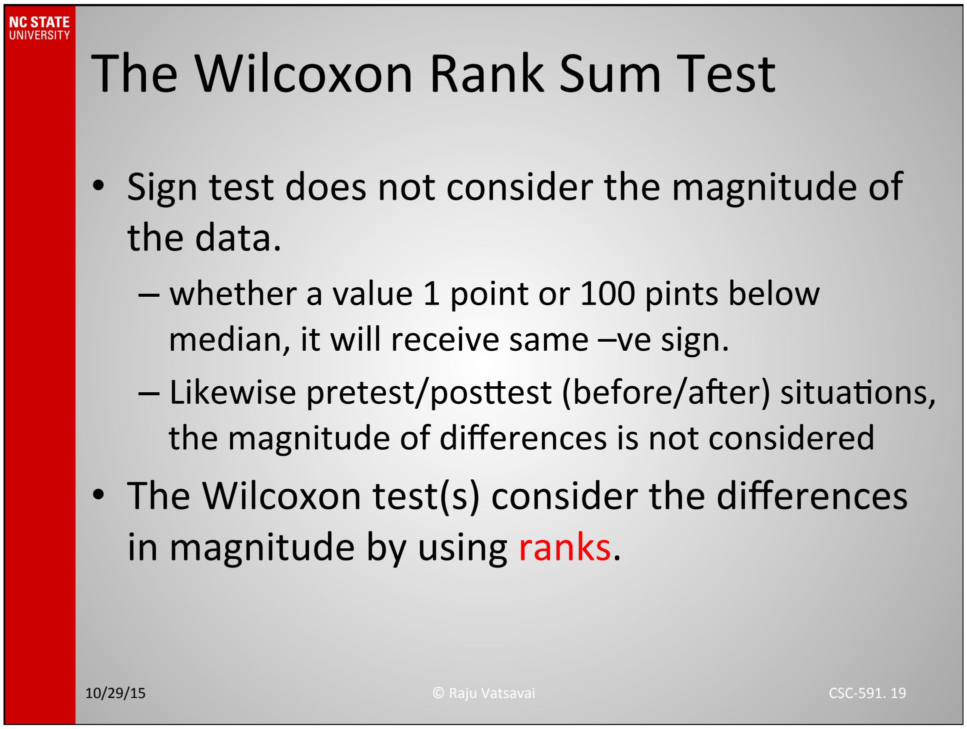
- Step 3: Compute test value.
  - T.V. = 2. Which is smallest of (+:7) and (-:2)
- Step 4: Make decision
  - T.V. > C.V. (2 > 1). Therefore, decision is to not reject the null hypothesis
- Step 5: Summarize results
  - There is not enough evidence to support the claim that the use of earplugs reduced the number of ear infections.



NC STATE UNIVERSITY

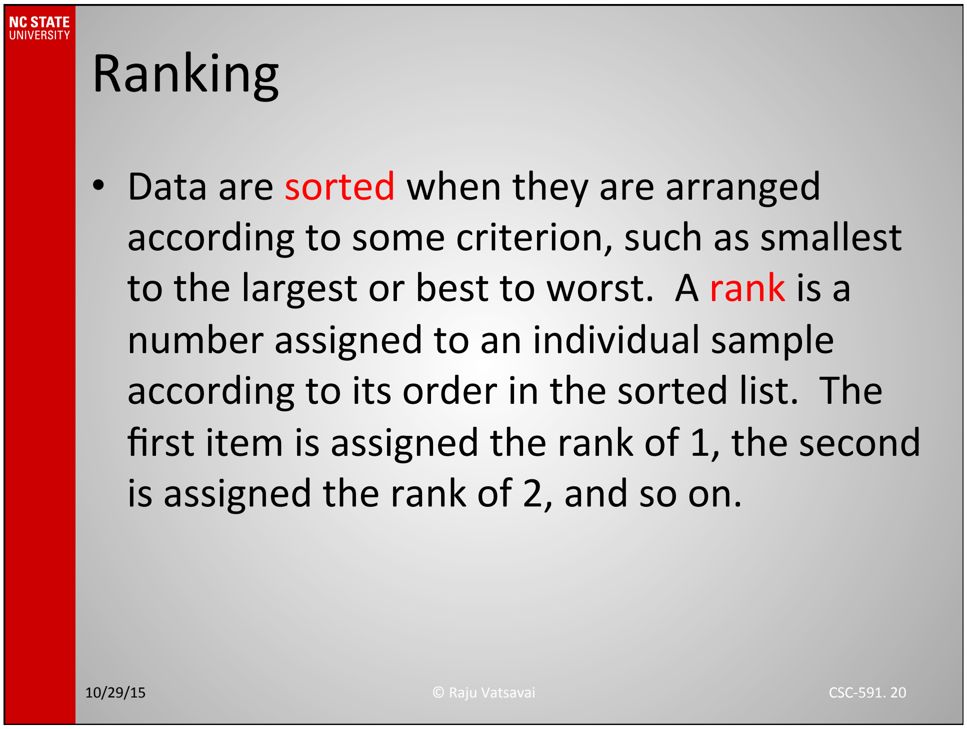
Other Examples

- Note that if the sample size is 26 or more, the normal approximation can be used to find the test value.
  - $z = \{(X + 0.5) – (n/2)\}/\{sqrt(n)/2\}$ (use z-table to find c.v); where X = smaller number of + or – signs



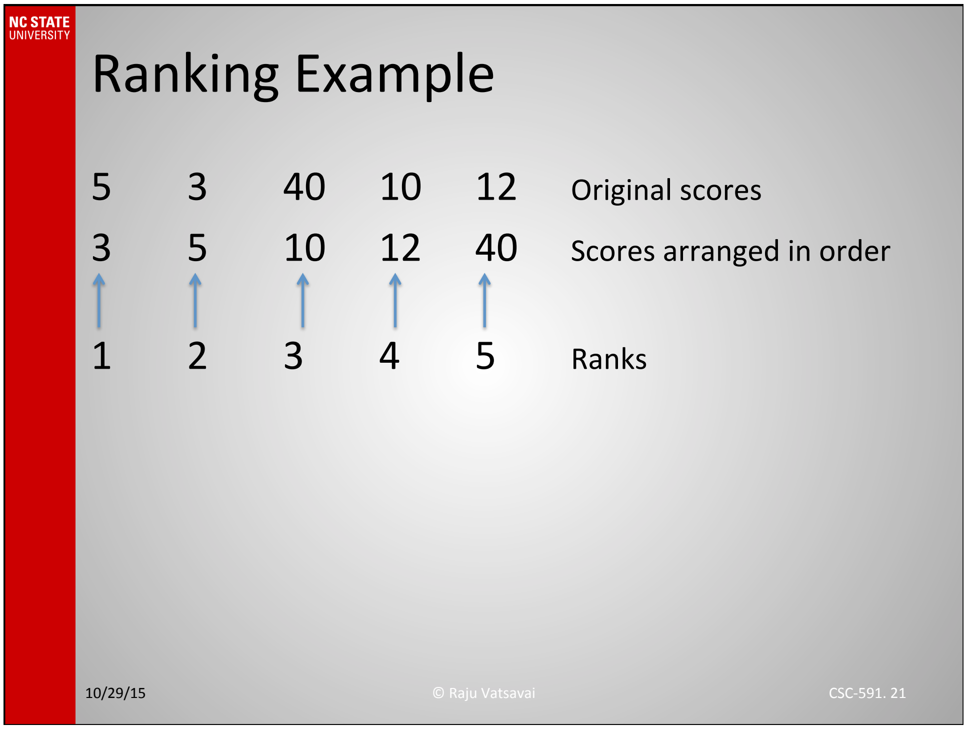
The Wilcoxon Rank Sum Test

- Sign test does not consider the magnitude of the data.
  - whether a value 1 point or 100 pints below median, it will receive same –ve sign.
  - Likewise pretest/posttest (before/after) situations, the magnitude of differences is not considered
- The Wilcoxon test(s) consider the differences in magnitude by using ranks.



Ranking

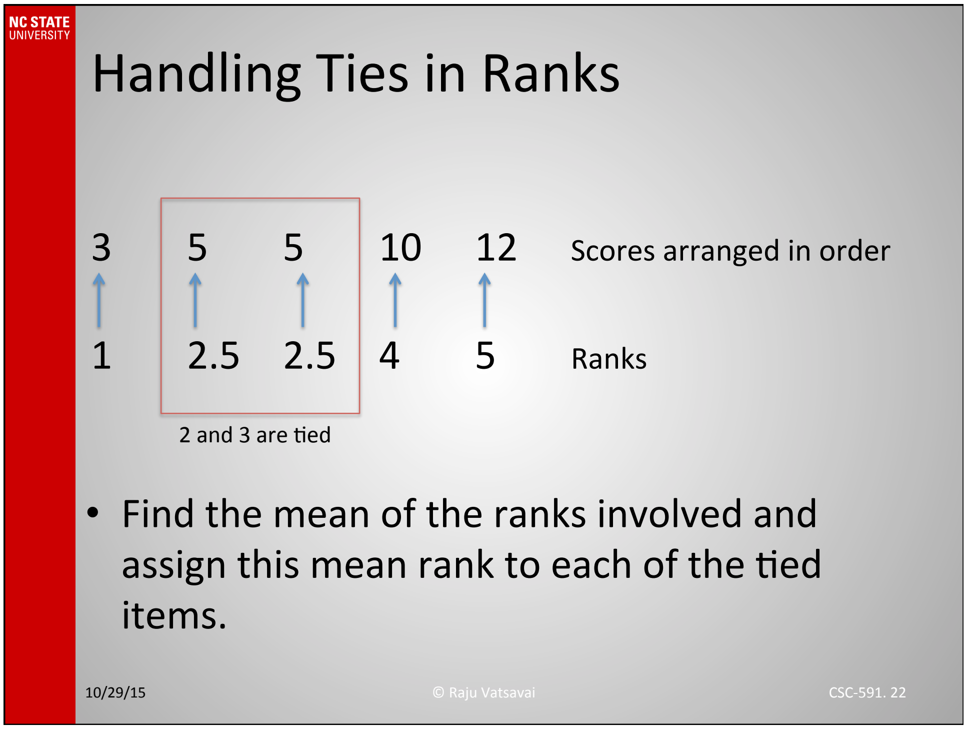
- Data are sorted when they are arranged according to some criterion, such as smallest to the largest or best to worst. A rank is a number assigned to an individual sample according to its order in the sorted list. The first item is assigned the rank of 1, the second is assigned the rank of 2, and so on.



NC STATE UNIVERSITY

Ranking Example

5	3	40	10	12	Original scores
3	5	10	12	40	Scores arranged in order
↑	↑	↑	↑	↑	
1	2	3	4	5	Ranks



NC STATE UNIVERSITY

Handling Ties in Ranks

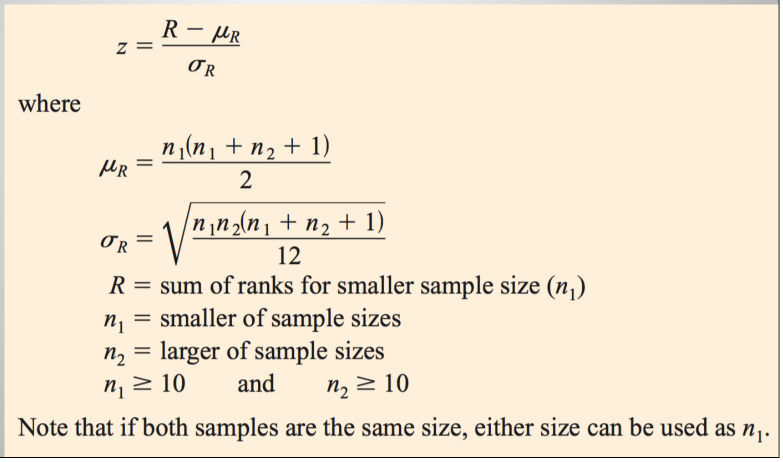
3	5	5	10	12	Scores arranged in order
↑	↑	↑	↑	↑	
1	2.5	2.5	4	5	Ranks

2 and 3 are tied

- Find the mean of the ranks involved and assign this mean rank to each of the tied items.

Wilcoxon Rank Sum Test

- Formula for the Wilcoxon Rank Sum test when samples are independent



$$z = \frac{R - \mu_R}{\sigma_R}$$

where

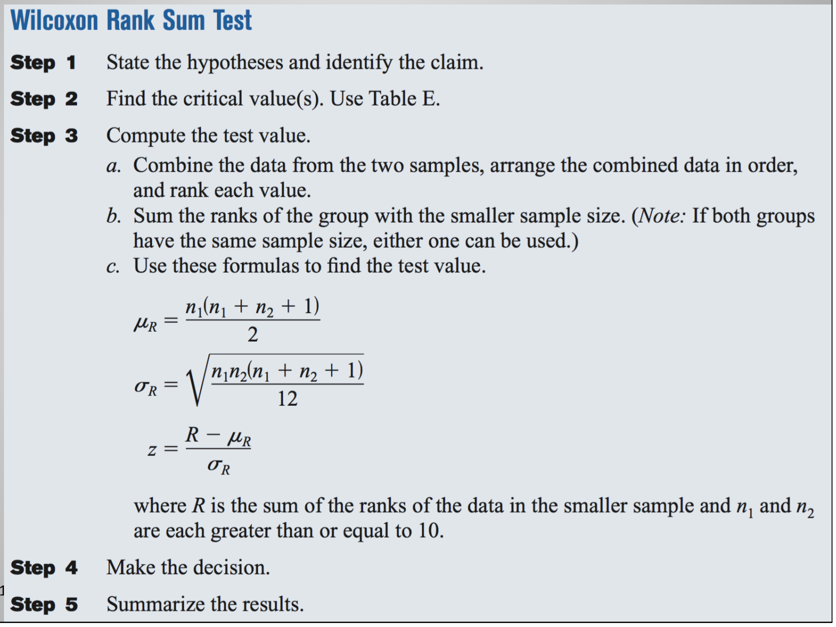
$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

R = sum of ranks for smaller sample size (n_1)
n_1 = smaller of sample sizes
n_2 = larger of sample sizes
$n_1 \geq 10$ and $n_2 \geq 10$

Note that if both samples are the same size, either size can be used as n_1.

Procedure



Wilcoxon Rank Sum Test

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value(s). Use Table E.

Step 3 Compute the test value.

a. Combine the data from the two samples, arrange the combined data in order, and rank each value.

b. Sum the ranks of the group with the smaller sample size. (*Note:* If both groups have the same sample size, either one can be used.)

c. Use these formulas to find the test value.

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

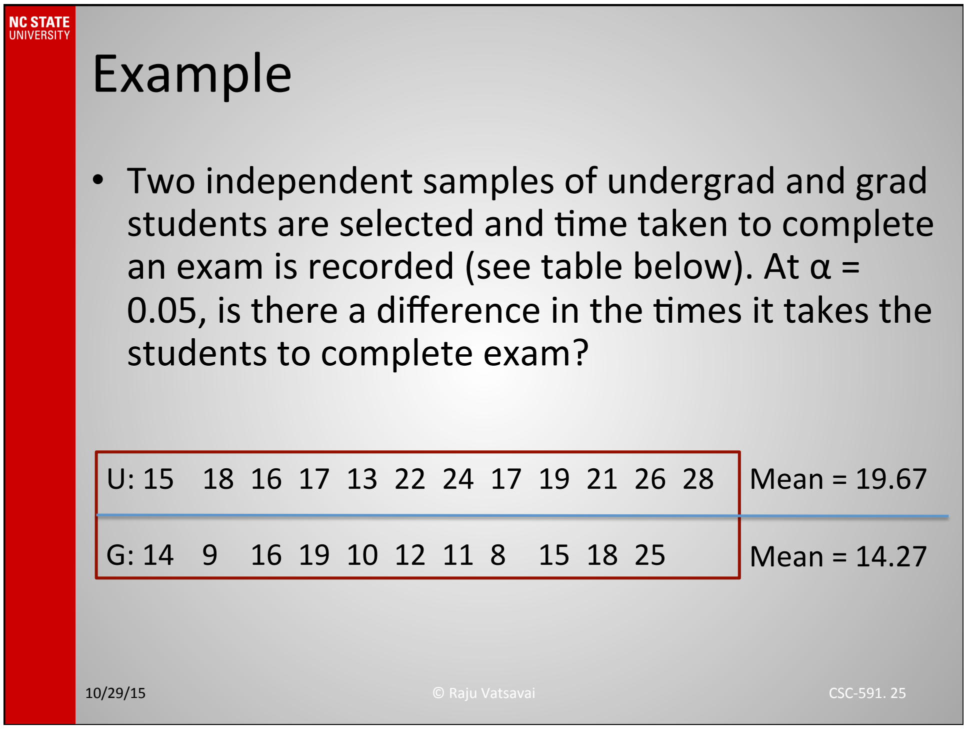
$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

$$z = \frac{R - \mu_R}{\sigma_R}$$

where R is the sum of the ranks of the data in the smaller sample and n_1 and n_2 are each greater than or equal to 10.

Step 4 Make the decision.

Step 5 Summarize the results.

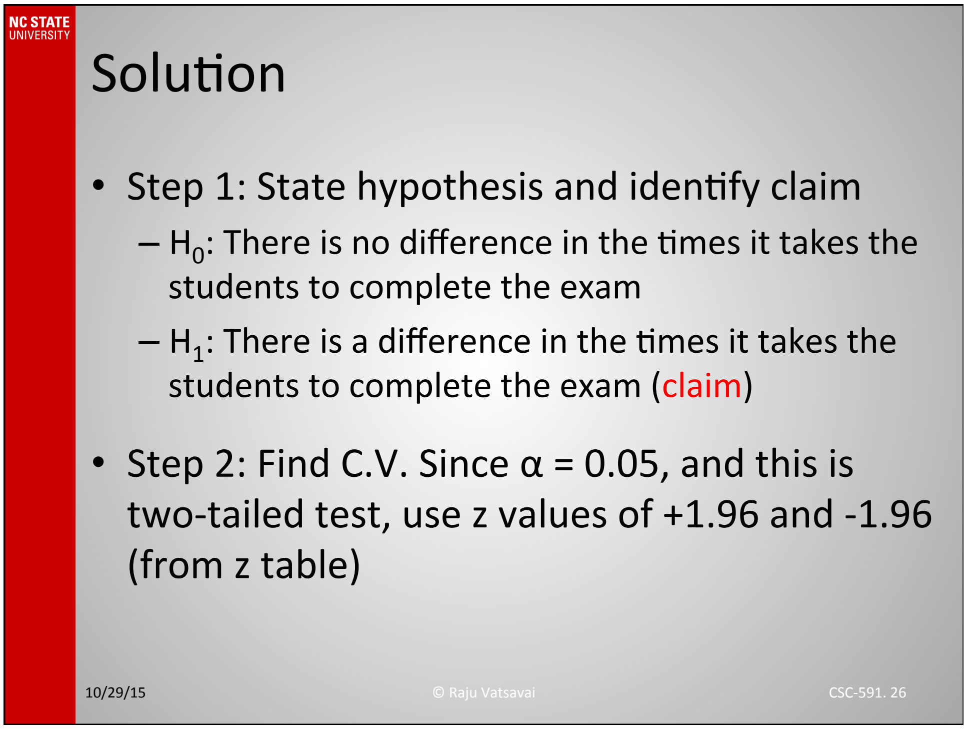


NC STATE UNIVERSITY

Example

- Two independent samples of undergrad and grad students are selected and time taken to complete an exam is recorded (see table below). At α = 0.05, is there a difference in the times it takes the students to complete exam?

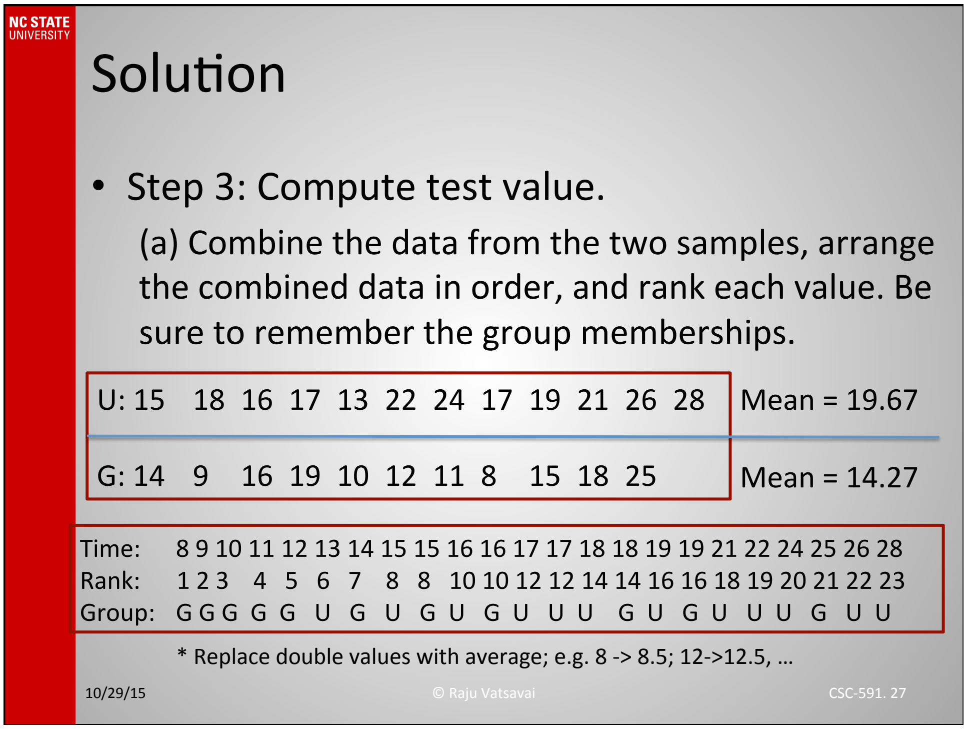
U: 15	18	16	17	13	22	24	17	19	21	26	28	Mean = 19.67
G: 14	9	16	19	10	12	11	8	15	18	25		Mean = 14.27



NC STATE UNIVERSITY

Solution

- Step 1: State hypothesis and identify claim
  - H_0: There is no difference in the times it takes the students to complete the exam
  - H_1: There is a difference in the times it takes the students to complete the exam (claim)
- Step 2: Find C.V. Since α = 0.05, and this is two-tailed test, use z values of +1.96 and -1.96 (from z table)



NC STATE UNIVERSITY

Solution

- Step 3: Compute test value.

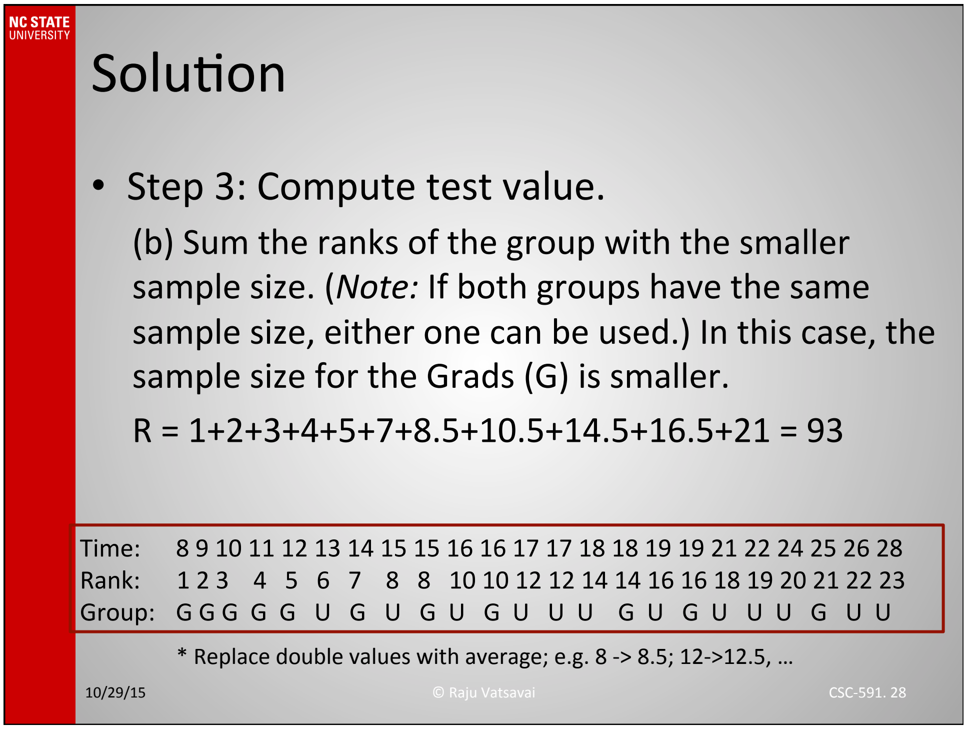
(a) Combine the data from the two samples, arrange the combined data in order, and rank each value. Be sure to remember the group memberships.

U: 15 18 16 17 13 22 24 17 19 21 26 28 Mean = 19.67

G: 14 9 16 19 10 12 11 8 15 18 25 Mean = 14.27

Time:	8	9	10	11	12	13	14	15	15	16	16	17	17	18	18	19	19	21	22	24	25	26	28
Rank:	1	2	3	4	5	6	7	8	8	10	10	12	12	14	14	16	16	18	19	20	21	22	23
Group:	G	G	G	G	G	U	G	U	G	U	G	U	U	U	G	U	G	U	U	U	G	U	U

* Replace double values with average; e.g. 8 -> 8.5; 12->12.5, …



NC STATE UNIVERSITY

Solution

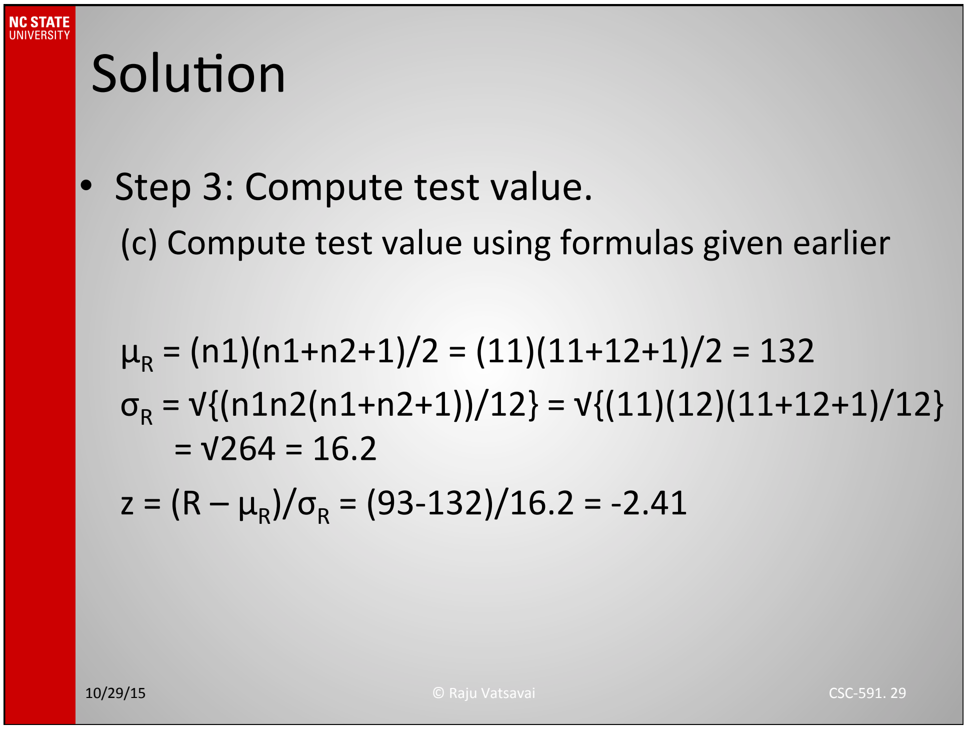
- Step 3: Compute test value.

(b) Sum the ranks of the group with the smaller sample size. (*Note:* If both groups have the same sample size, either one can be used.) In this case, the sample size for the Grads (G) is smaller.

R = 1+2+3+4+5+7+8.5+10.5+14.5+16.5+21 = 93

Time:	8	9	10	11	12	13	14	15	15	16	16	17	17	18	18	19	19	21	22	24	25	26	28
Rank:	1	2	3	4	5	6	7	8	8	10	10	12	12	14	14	16	16	18	19	20	21	22	23
Group:	G	G	G	G	G	U	G	U	G	U	G	U	U	U	G	U	G	U	U	U	G	U	U

* Replace double values with average; e.g. 8 -> 8.5; 12->12.5, …



NC STATE UNIVERSITY

Solution

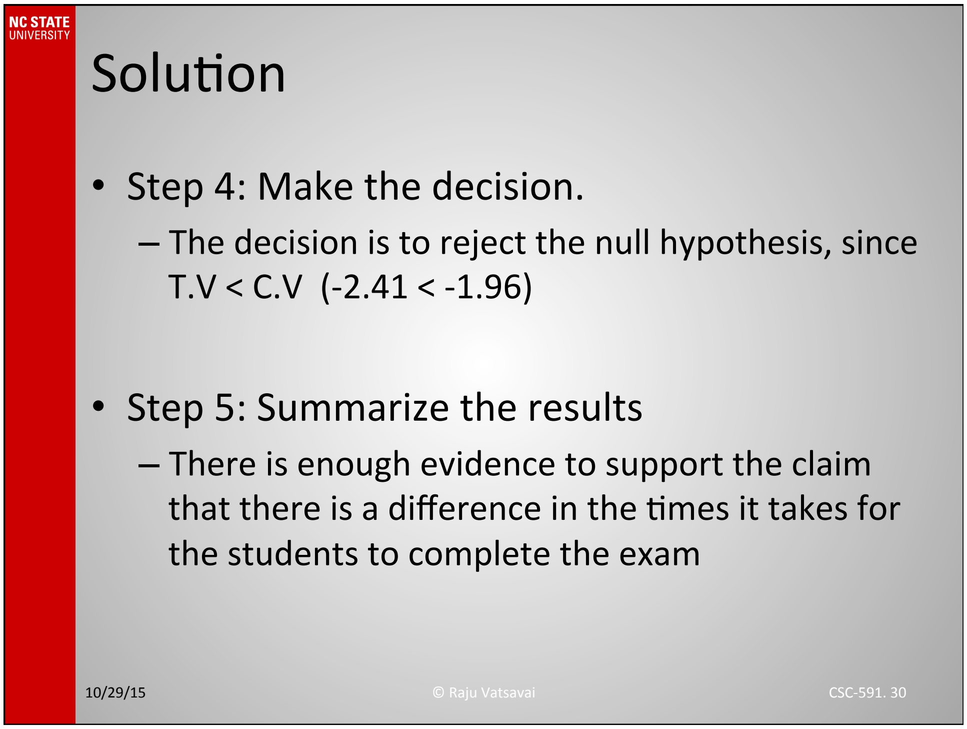
- Step 3: Compute test value.

  (c) Compute test value using formulas given earlier

  $\mu_R = (n1)(n1+n2+1)/2 = (11)(11+12+1)/2 = 132$

  $\sigma_R = \surd\{(n1n2(n1+n2+1))/12\} = \surd\{(11)(12)(11+12+1)/12\}$
  $= \surd 264 = 16.2$

  $z = (R - \mu_R)/\sigma_R = (93-132)/16.2 = -2.41$



NC STATE UNIVERSITY

Solution

- Step 4: Make the decision.
  - The decision is to reject the null hypothesis, since T.V < C.V (-2.41 < -1.96)

- Step 5: Summarize the results
  - There is enough evidence to support the claim that there is a difference in the times it takes for the students to complete the exam

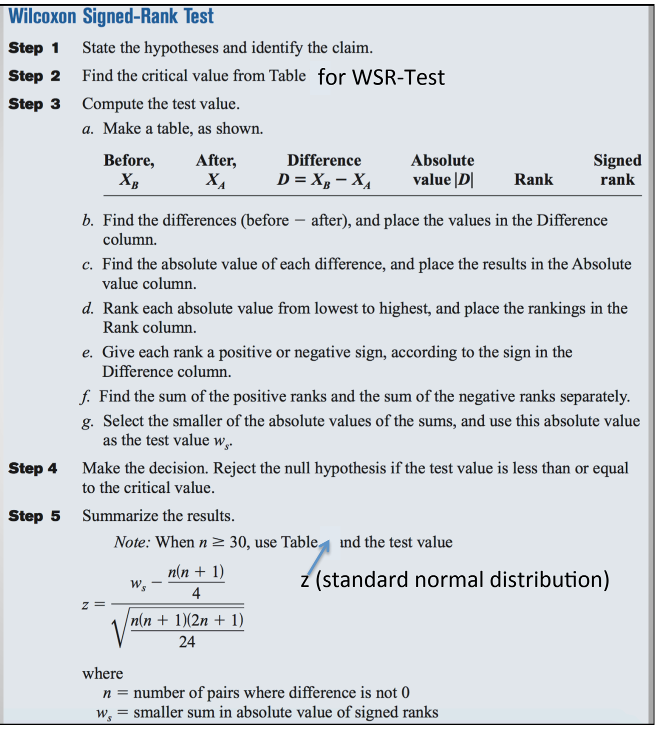
NC STATE UNIVERSITY

Wilcoxon Signed-Rank

- When the samples are dependent
  - Many times samples may be dependent, for example "before-and-after" experiments. We can use Wilcoxon Signed-Rank test in place of *t* test.

NC STATE UNIVERSITY

Procedure



Wilcoxon Signed-Rank Test

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value from Table for WSR-Test

Step 3 Compute the test value.

a. Make a table, as shown.

Before, X_B	After, X_A	Difference $D = X_B - X_A$	Absolute value $\|D\|$	Rank	Signed rank

b. Find the differences (before − after), and place the values in the Difference column.

c. Find the absolute value of each difference, and place the results in the Absolute value column.

d. Rank each absolute value from lowest to highest, and place the rankings in the Rank column.

e. Give each rank a positive or negative sign, according to the sign in the Difference column.

f. Find the sum of the positive ranks and the sum of the negative ranks separately.

g. Select the smaller of the absolute values of the sums, and use this absolute value as the test value w_s.

Step 4 Make the decision. Reject the null hypothesis if the test value is less than or equal to the critical value.

Step 5 Summarize the results.

Note: When $n \geq 30$, use Table and the test value

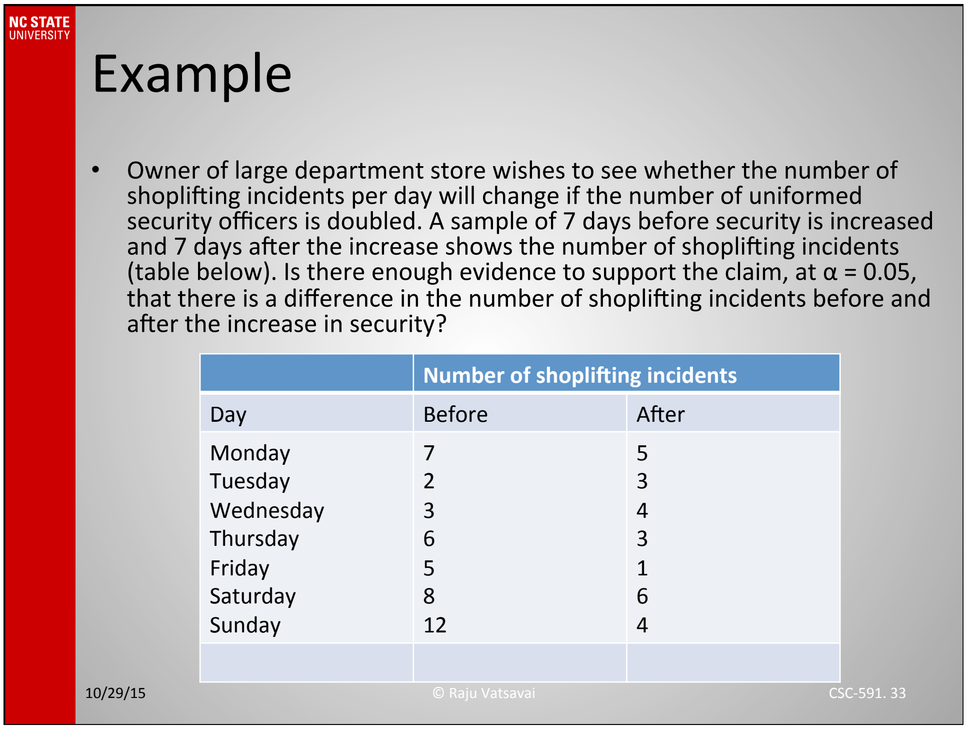
z (standard normal distribution)

$$z = \frac{w_s - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$$

where

n = number of pairs where difference is not 0

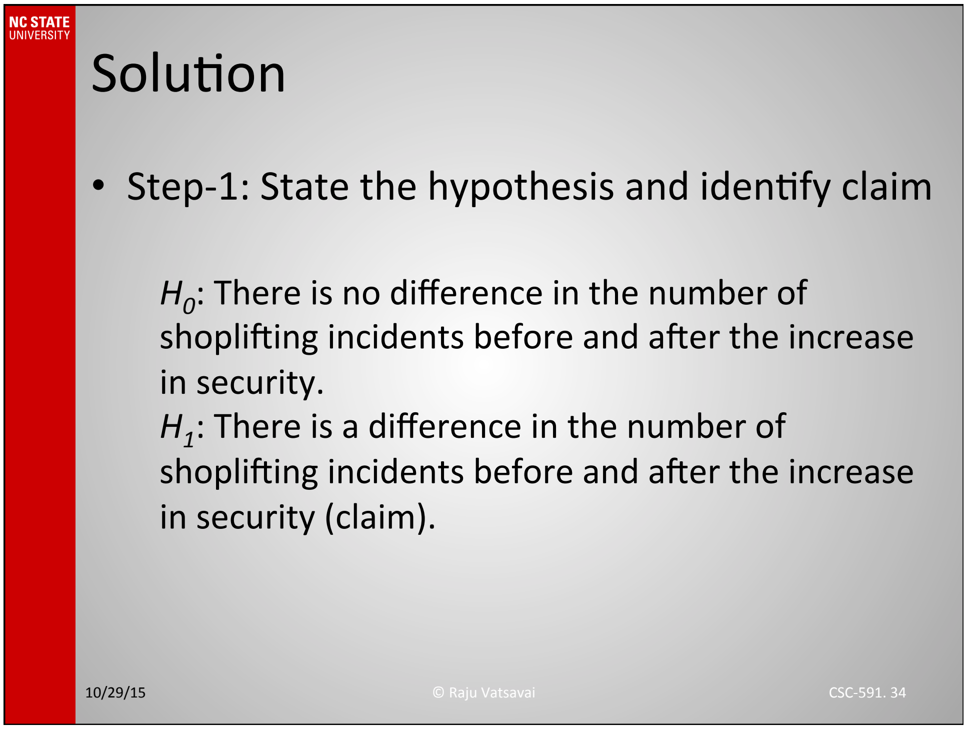
w_s = smaller sum in absolute value of signed ranks



Example

- Owner of large department store wishes to see whether the number of shoplifting incidents per day will change if the number of uniformed security officers is doubled. A sample of 7 days before security is increased and 7 days after the increase shows the number of shoplifting incidents (table below). Is there enough evidence to support the claim, at α = 0.05, that there is a difference in the number of shoplifting incidents before and after the increase in security?

	Number of shoplifting incidents	
Day	Before	After
Monday	7	5
Tuesday	2	3
Wednesday	3	4
Thursday	6	3
Friday	5	1
Saturday	8	6
Sunday	12	4



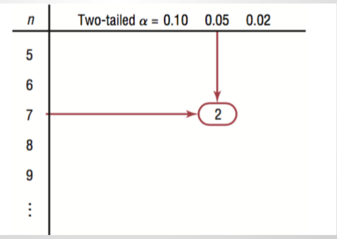
Solution

- Step-1: State the hypothesis and identify claim

H_0: There is no difference in the number of shoplifting incidents before and after the increase in security.
H_1: There is a difference in the number of shoplifting incidents before and after the increase in security (claim).

Solution

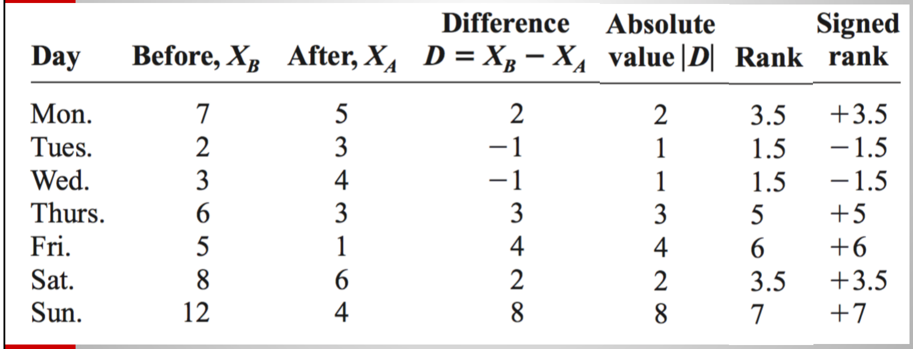
- Step-2: Find critical values from WSR table.
  - We have: n = 7, α = 0.05, two-tailed test
  - Therefore, c.v. = 2



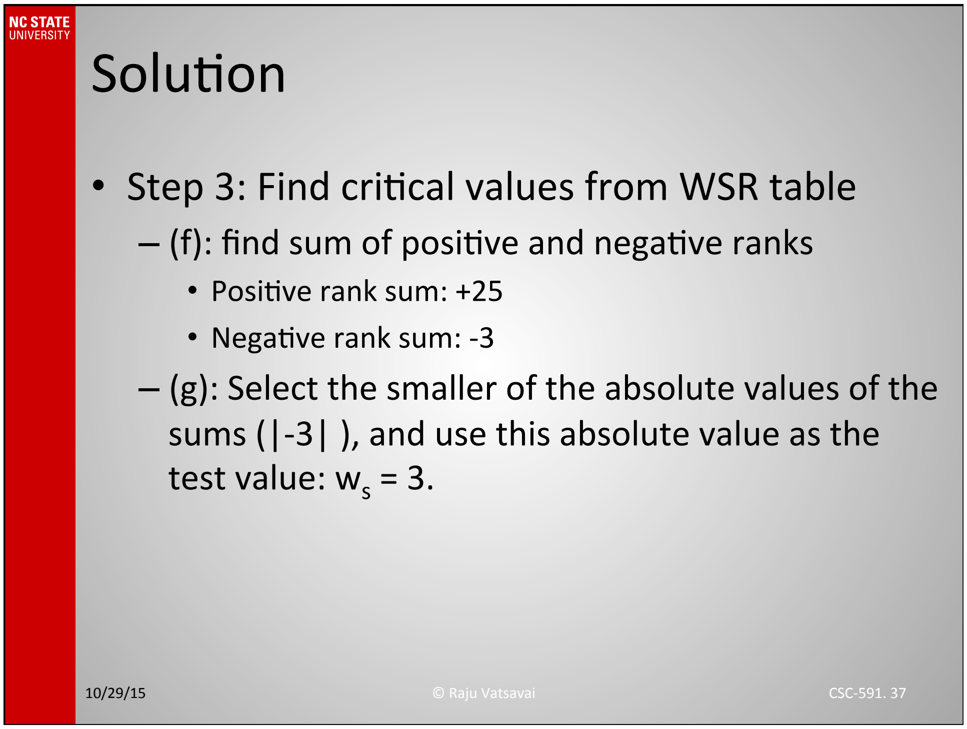
n	Two-tailed α = 0.10	0.05	0.02
5			
6			
7		2	
8			
9			
⋮			

Solution

- Step-3: Find test value
  - (a-e): Make table

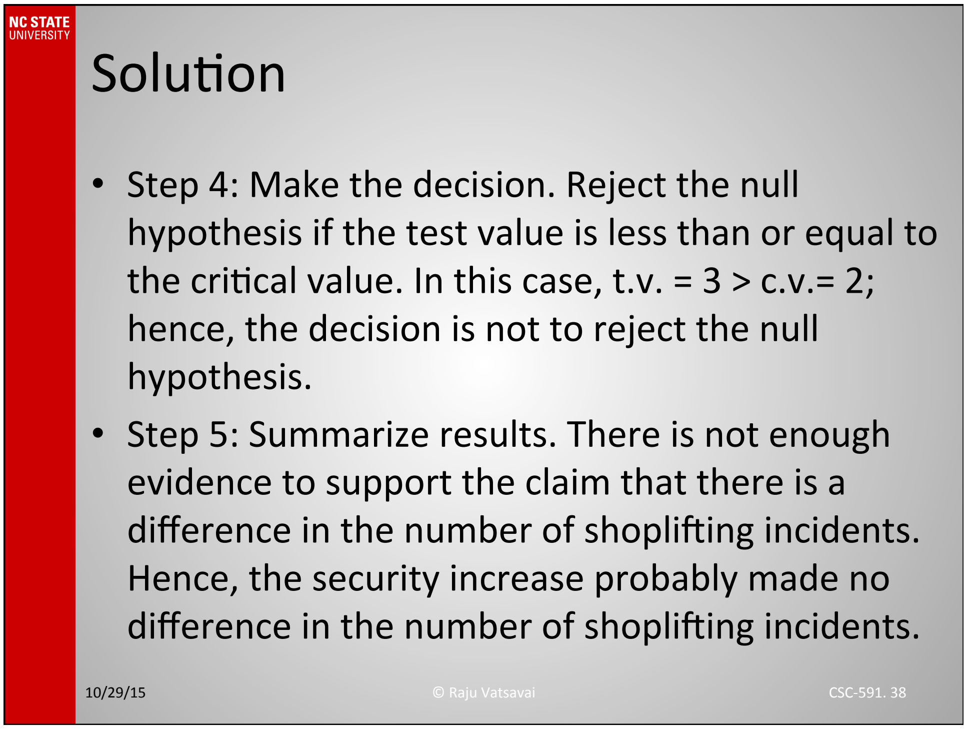


Day	Before, X_B	After, X_A	Difference $D = X_B - X_A$	Absolute value $\lvert D\rvert$	Rank	Signed rank
Mon.	7	5	2	2	3.5	+3.5
Tues.	2	3	−1	1	1.5	−1.5
Wed.	3	4	−1	1	1.5	−1.5
Thurs.	6	3	3	3	5	+5
Fri.	5	1	4	4	6	+6
Sat.	8	6	2	2	3.5	+3.5
Sun.	12	4	8	8	7	+7



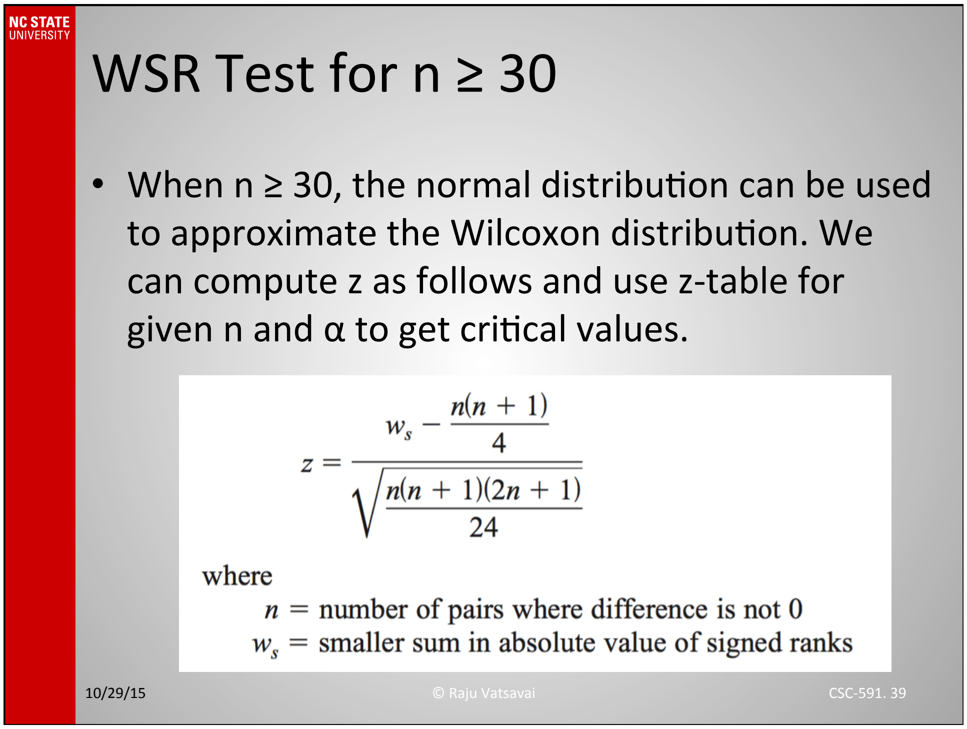
Solution

- Step 3: Find critical values from WSR table
  - (f): find sum of positive and negative ranks
    - Positive rank sum: +25
    - Negative rank sum: -3
  - (g): Select the smaller of the absolute values of the sums (|-3|), and use this absolute value as the test value: $w_s = 3$.



Solution

- Step 4: Make the decision. Reject the null hypothesis if the test value is less than or equal to the critical value. In this case, t.v. = 3 > c.v.= 2; hence, the decision is not to reject the null hypothesis.
- Step 5: Summarize results. There is not enough evidence to support the claim that there is a difference in the number of shoplifting incidents. Hence, the security increase probably made no difference in the number of shoplifting incidents.



NC STATE UNIVERSITY

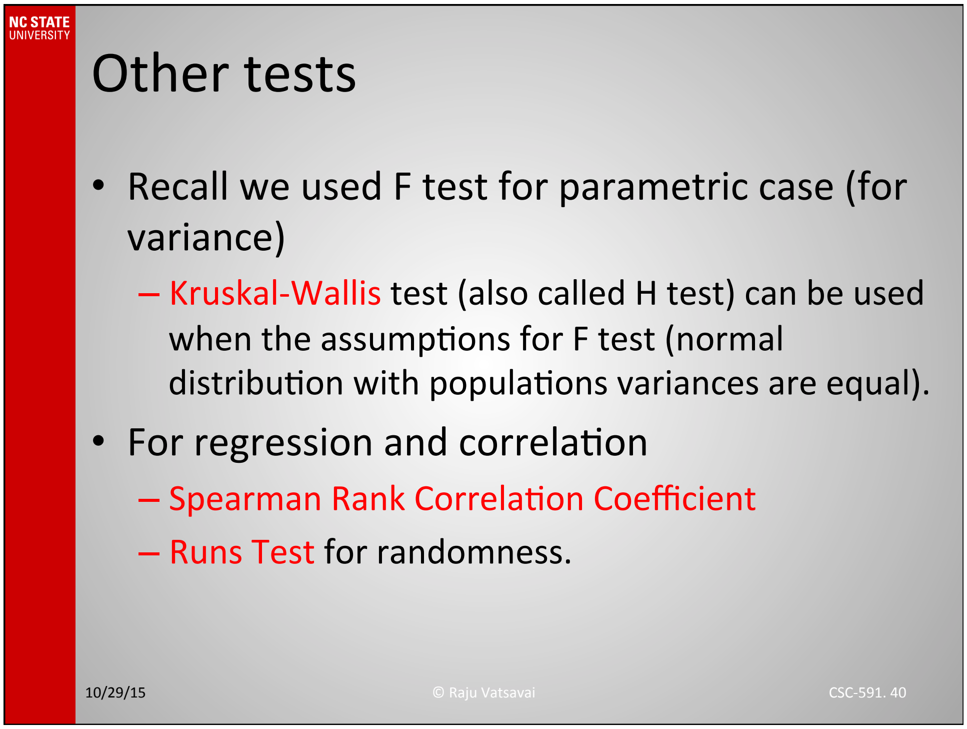
WSR Test for n ≥ 30

- When n ≥ 30, the normal distribution can be used to approximate the Wilcoxon distribution. We can compute z as follows and use z-table for given n and α to get critical values.

$$z = \frac{w_s - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

where

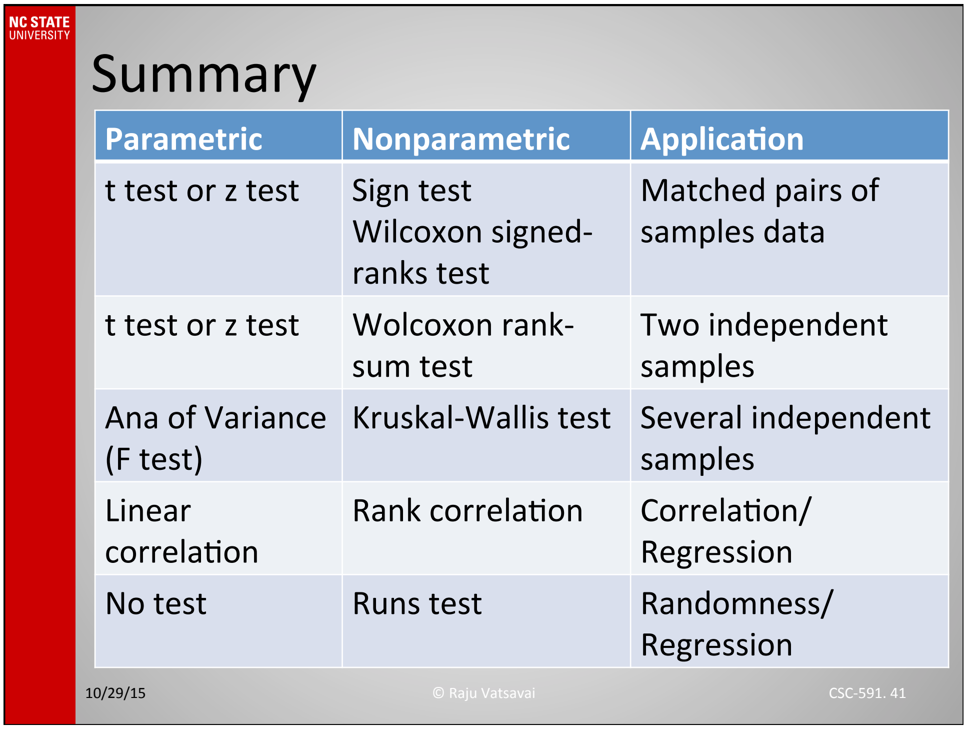
n = number of pairs where difference is not 0
w_s = smaller sum in absolute value of signed ranks



NC STATE UNIVERSITY

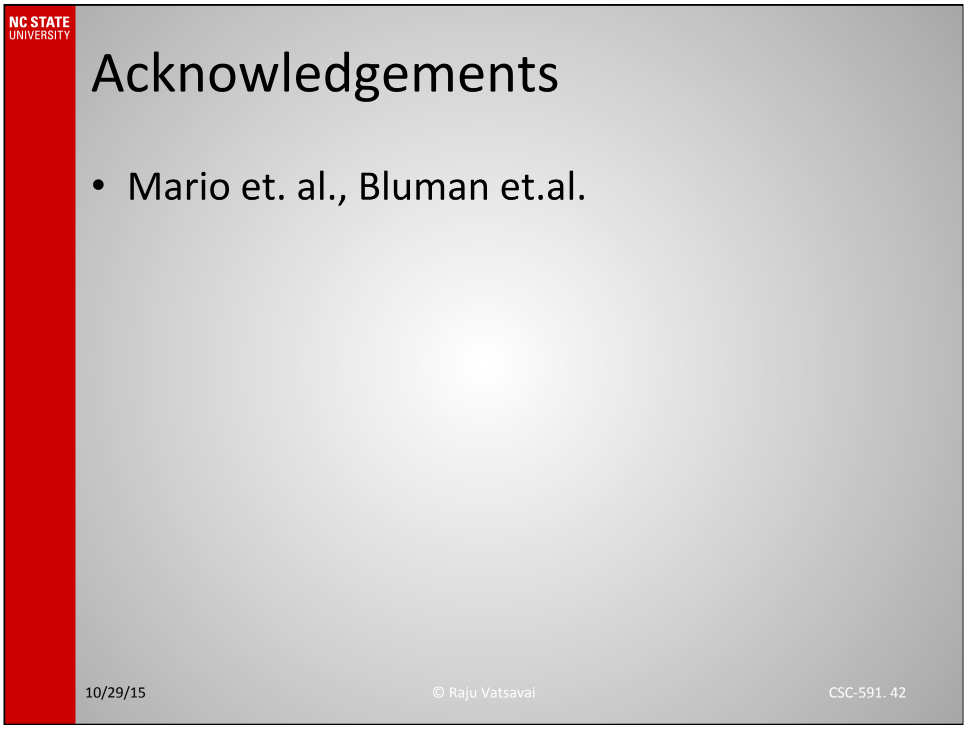
Other tests

- Recall we used F test for parametric case (for variance)
  - Kruskal-Wallis test (also called H test) can be used when the assumptions for F test (normal distribution with populations variances are equal).
- For regression and correlation
  - Spearman Rank Correlation Coefficient
  - Runs Test for randomness.



Summary

Parametric	Nonparametric	Application
t test or z test	Sign test Wilcoxon signed-ranks test	Matched pairs of samples data
t test or z test	Wolcoxon rank-sum test	Two independent samples
Ana of Variance (F test)	Kruskal-Wallis test	Several independent samples
Linear correlation	Rank correlation	Correlation/ Regression
No test	Runs test	Randomness/ Regression



Acknowledgements

- Mario et. al., Bluman et.al.