DATA ANALYTICS

Class # 11

Non-parametric tests

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

QUOTE OF THE DAY..

Try not to become a person of success, but rather try to become a person of value.

ALBERT EINSTEIN, Theoretical physicist

INTRODUCTION

- All of the tests presented in hypothesis testing are called **parametric tests** and are based on certain assumptions.
- For example, when running tests of hypothesis for means of continuous outcomes, all parametric tests assume that the outcome is approximately normally distributed in the population. This does not mean that the data in the observed sample follows a normal distribution, but rather that the outcome follows a normal distribution in the full population which is not observed.
- Many statistical tests are **robust,** which means that they maintain their statistical properties even when assumptions are not entirely met. Tests are robust in the presence of violations of the normality assumption when the sample size is large based on the Central Limit Theorem.
- When the sample size is small and the distribution of the outcome is not known and cannot be assumed to be approximately normally distributed, then alternative tests called **nonparametric tests** are appropriate.

WHEN TO USE A NONPARAMETRIC TEST

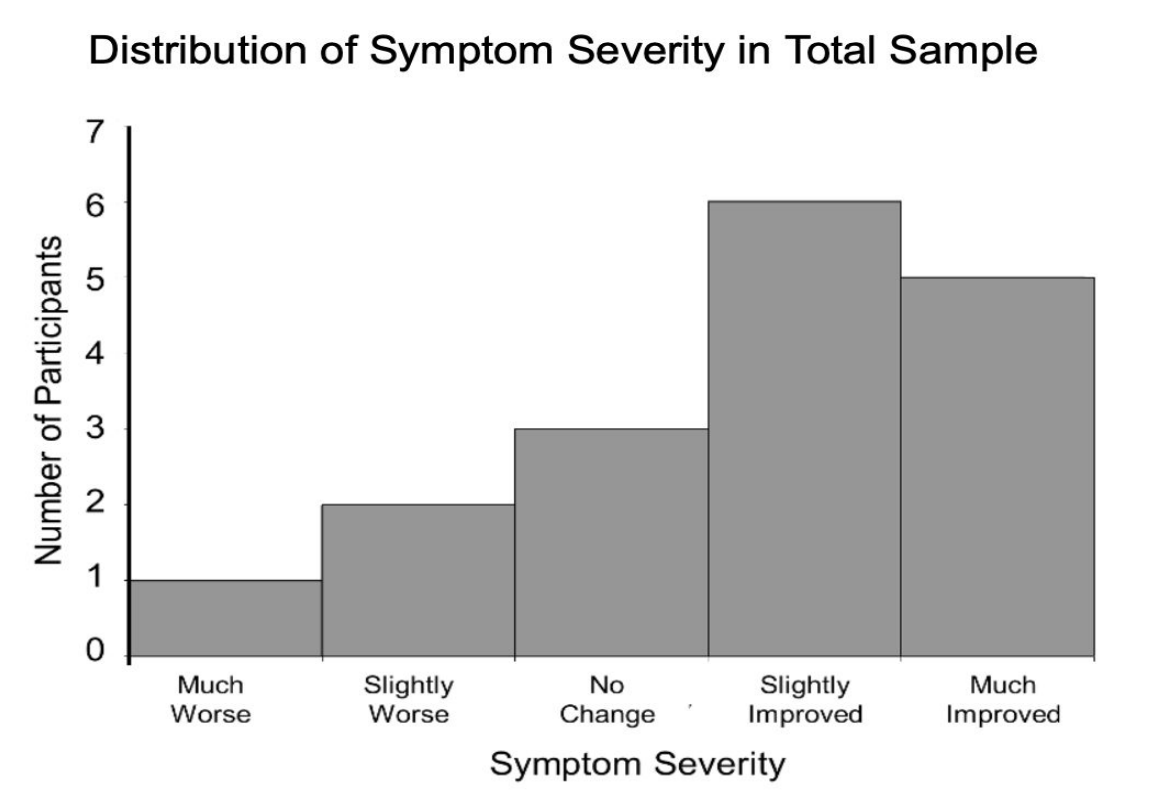
- Nonparametric tests are called **distribution-free tests** because they are based on fewer assumptions (e.g., they do not assume that the outcome is approximately normally distributed).
- Parametric tests involve specific probability distributions (e.g., the normal distribution) and the tests involve estimation of the key parameters of that distribution (e.g., the mean or difference in means) from the sample data. The cost of fewer assumptions is that nonparametric tests are generally less powerful than their parametric counterparts (i.e., when the alternative is true, they may be less likely to reject H_0).
- It can sometimes be difficult to assess whether a continuous outcome follows a normal distribution and, thus, whether a parametric or nonparametric test is appropriate.
- There are several statistical tests that can be used to assess whether data are likely from a normal distribution. The most popular are **the Anderson-Darling test, and the Shapiro-Wilk test**.

WHEN TO USE A NONPARAMETRIC TEST

- There are some situations when it is clear that the outcome does not follow a normal distribution. These include situations:
  - when the outcome is an **ordinal variable or a rank**,
  - when there are **definite outliers** or
  - when the outcome has **clear limits of detection**.

USING AN ORDINAL SCALE

Consider a clinical trial where study participants are asked to rate their symptom severity following 6 weeks on the assigned treatment. Symptom severity might be measured on a 5 point ordinal scale with response options: Symptoms got much worse, slightly worse, no change, slightly improved, or much improved. Suppose there are a total of n=20 participants in the trial, randomized to an experimental treatment or placebo, and the outcome data are distributed as shown in the figure below.



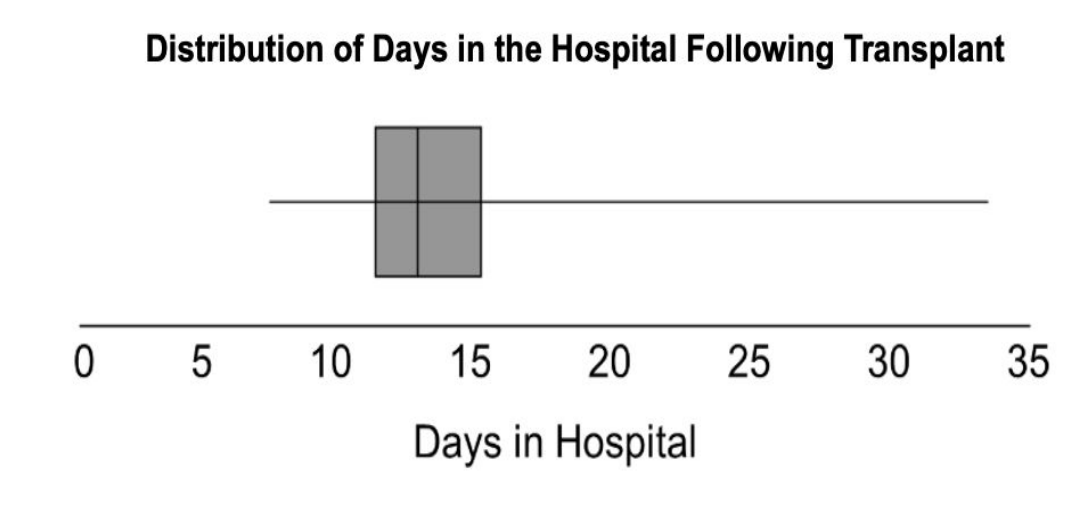
The distribution of the outcome (symptom severity) does not appear to be normal as more participants report improvement in symptoms as opposed to worsening of symptoms.

WHEN THE OUTCOME IS A RANK

In some studies, the outcome is a rank. For example, in new-born health studies an APGAR score is often used to assess the health of a new-born. The score, which ranges from 1-10, is the sum of five component scores based on the infant's condition at birth. APGAR scores generally do not follow a normal distribution, since most new-borns have scores of 7 or higher (normal range).

WHEN THERE ARE OUTLIERS

In some studies, the outcome is continuous but subject to outliers or extreme values. For example, days in the hospital following a particular surgical procedure is an outcome that is often subject to outliers. Suppose in an observational study investigators wish to assess whether there is a difference in the days patients spend in the hospital following liver transplant in for-profit versus nonprofit hospitals. Suppose we measure days in the hospital following transplant in n=100 participants, 50 from for-profit and 50 from non-profit hospitals. The number of days in the hospital are summarized by the box-whisker plot below.



- Q_1-1.5(Q_3-Q_1) as a lower limit and Q_3+1.5(Q_3-Q_1) as an upper limit to detect outliers.
- In the box-whisker plot above, Q_1=12 and Q_3=16, thus outliers are values below 12-1.5(16-12) = 6 or above 16+1.5(16-12) = 22.

Note that 75% of the participants stay at most 16 days in the hospital following transplant, while at least 1 stays 35 days which would be considered an outlier.

LIMITS OF DETECTION

In some studies, the outcome is a continuous variable that is measured with some imprecision (e.g., with clear limits of detection). For example, some instruments or evaluation cannot measure presence of specific quantities above or below certain limits.

HIV viral load is a measure of the amount of virus in the body and is measured as the amount of virus per a certain volume of blood. It can range from "not detected" or "below the limit of detection" to hundreds of millions of copies. Thus, in a sample some participants may have measures like 1,254,000 or 874,050 copies and others are measured as "not detected." If a substantial number of participants have undetectable levels, the distribution of viral load is not normally distributed.

ADVANTAGES OF NONPARAMETRIC TESTS

Nonparametric tests have some distinct advantages. With outcomes such as those described before, nonparametric tests may be the only way to analyse these data. Outcomes that are ordinal, ranked, subject to outliers or measured imprecisely are difficult to analyse with parametric methods without making major assumptions about their distributions as well as decisions about coding some values (e.g., "not detected"). Nonparametric tests can also be relatively simple to conduct.

Hypothesis Testing with Nonparametric Tests

In nonparametric tests, the hypotheses are not about population parameters (e.g., $\mu=50$ or $\mu_1=\mu_2$). Instead, the null hypothesis is more general. For example, when comparing two independent groups in terms of a continuous outcome, the null hypothesis in a parametric test is H_0: $\mu_1 = \mu_2$. In a nonparametric test the null hypothesis is that the two populations are equal, often this is interpreted as the two populations are **equal in terms of their central tendency**.

NONPARAMETRIC TESTING

Assigning Ranks

- The outcome variable (ordinal, interval or continuous) is ranked from lowest to highest and the analysis focuses on the ranks as opposed to the measured or raw values. For example, suppose we measure self-reported pain using a visual analog scale with anchors at 0 (no pain) and 10 (agonizing pain) and record the following in a sample of n=6 participants:

7 5 9 3 0 2

- The ranks, which are used to perform a nonparametric test, are assigned as follows: First, the data are ordered from smallest to largest. The lowest value is then assigned a rank of 1, the next lowest a rank of 2 and so on. The largest value is assigned a rank of n (in this example, n=6). The observed data and corresponding ranks are shown below:

Ordered Observed Data:	0	2	3	5	7	9
Ranks:	1	2	3	4	5	6

NONPARAMETRIC TESTING

- A complicating issue that arises when assigning ranks occurs when there are ties in the sample (i.e., the same values are measured in two or more participants). For example, suppose that the following data are observed in our sample of n=6:

Observed Data: 7 7 9 3 0 2

- The 4th and 5th ordered values are both equal to 7. When assigning ranks, the recommended procedure is to assign the mean rank of 4.5 to each (i.e. the mean of 4 and 5), as follows:

Ordered Observed Data:	0	2	3	7	7	9
Ranks:	1	2	3	4.5	4.5	6

- Suppose that there are three values of 7. In this case, we assign a rank of 5 (the mean of 4, 5 and 6) to the 4th, 5th and 6th values, as follows:

Ordered Observed Data:	0	2	3	7	7	7
Ranks:	1	2	3	5	5	5

NONPARAMETRIC TESTING

Note:

Using this approach of assigning the mean rank when there are ties ensures that the sum of the ranks is the same in each sample (for example, 1+2+3+4+5+6=21, 1+2+3+4.5+4.5+6=21 and 1+2+3+5+5+5=21). Using this approach, the sum of the ranks will always equal n(n+1)/2. When conducting nonparametric tests, it is useful to check the sum of the ranks before proceeding with the analysis.

NONPARAMETRIC TESTING

To conduct nonparametric tests, we again follow the five-step approach outlined in the hypothesis testing.

- Set up hypotheses and select the level of significance α. Analogous to parametric testing, the research hypothesis can be one- or two- sided (one- or two-tailed), depending on the research question of interest.
- Select the appropriate test statistic. The test statistic is a single number that summarizes the sample information. In nonparametric tests, the observed data is converted into ranks and then the ranks are summarized into a test statistic.
- Set up decision rule. The decision rule is a statement that tells under what circumstances to reject the null hypothesis. Note that in some nonparametric tests we reject H_0 if the test statistic is large, while in others we reject H_0 if the test statistic is small. We make the distinction as we describe the different tests.
- Compute the test statistic. Here we compute the test statistic by summarizing the ranks into the test statistic identified in Step 2.
- Conclusion. The final conclusion is made by comparing the test statistic (which is a summary of the information observed in the sample) to the decision rule. The final conclusion is either to reject the null hypothesis (because it is very unlikely to observe the sample data if the null hypothesis is true) or not to reject the null hypothesis (because the sample data are not very unlikely if the null hypothesis is true).

MANN WHITNEY U TEST (WILCOXON RANK SUM TEST)

- A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test.
- The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations.
- In contrast, the null and two-sided research hypotheses for the nonparametric test are stated as follows:

H0: The two populations are equal

H1: The two populations are not equal.

- This test is often performed as a two-sided test and, thus, the research hypothesis indicates that the populations are not equal as opposed to specifying directionality. A one-sided research hypothesis is used if interest lies in detecting a positive or negative shift in one population as compared to the other. The procedure for the test involves pooling the observations from the two samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to n1+n2, respectively.

MANN WHITNEY U TEST (WILCOXON RANK SUM TEST)

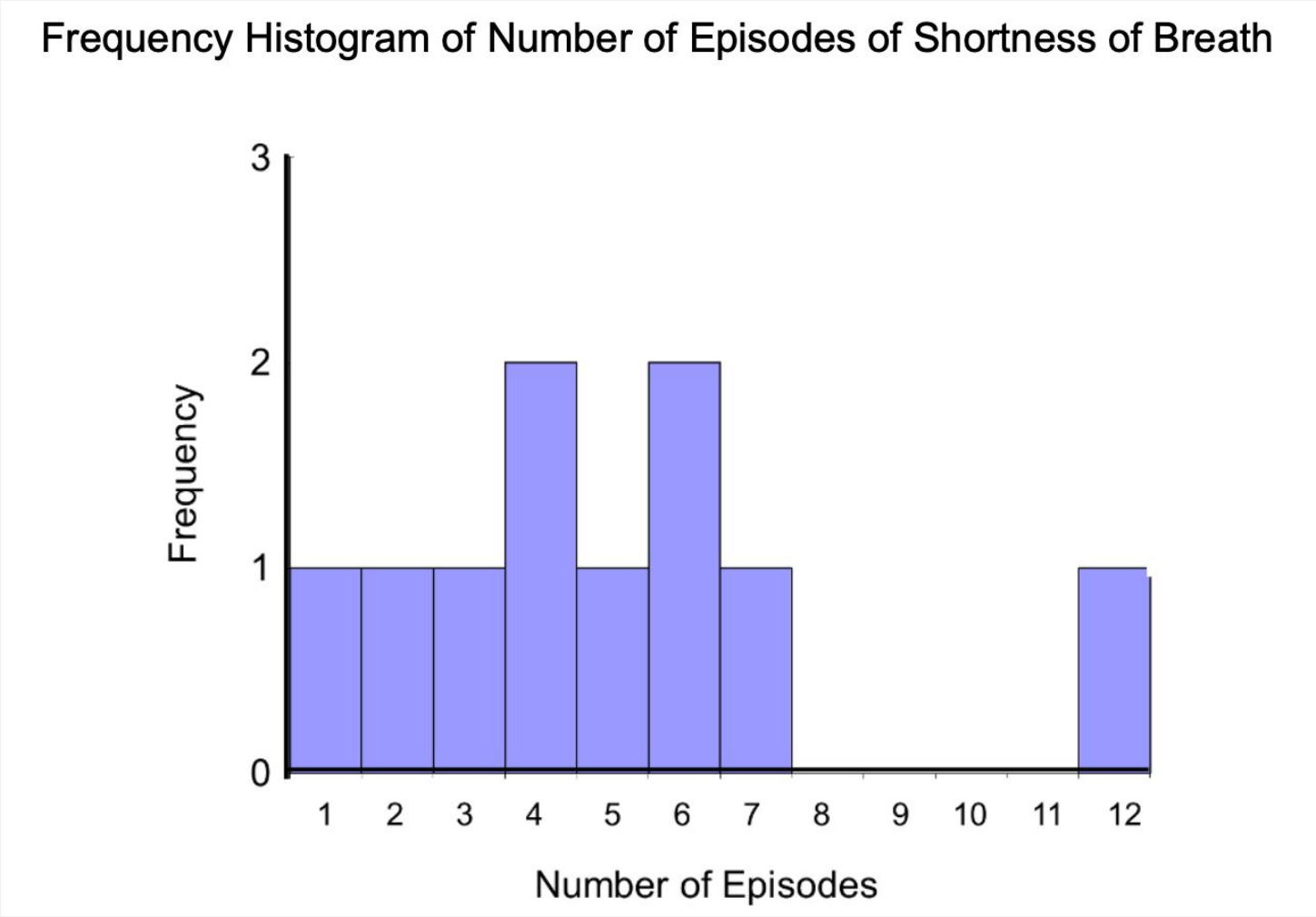
- **Example:**
- Consider a Phase II clinical trial designed to investigate the effectiveness of a new drug to reduce symptoms of asthma in children. A total of n=10 participants are randomized to receive either the new drug or a placebo. Participants are asked to record the number of episodes of shortness of breath over a 1 week period following receipt of the assigned treatment. The data are shown below.

Placebo	7	5	6	4	12
New Drug	3	6	4	2	1

Is there a difference in the number of episodes of shortness of breath over a 1 week period in participants receiving the new drug as compared to those receiving the placebo? By inspection, it appears that participants receiving the placebo have more episodes of shortness of breath, but is this statistically significant?

EXAMPLE 1

- **In this example, the outcome is a count and in this sample the data do not follow a normal distribution.**



EXAMPLE 1

- In addition, the sample size is small (n1=n2=5), so a nonparametric test is appropriate. The hypothesis is given below, and we run the test at the 5% level of significance (i.e., α=0.05).

H0: The two populations are equal

H1: The two populations are not equal.

- Note that if the null hypothesis is true (i.e., the two populations are equal), we expect to see similar numbers of episodes of shortness of breath in each of the two treatment groups, and we would expect to see some participants reporting few episodes and some reporting more episodes in each group. This does not appear to be the case with the observed data. A test of hypothesis is needed to determine whether the observed data is evidence of a statistically significant difference in populations.
- The first step is to assign ranks and to do so we order the data from smallest to largest. This is done on the combined or total sample (i.e., pooling the data from the two treatment groups (n=10)), and assigning ranks from 1 to 10, as follows. We also need to keep track of the group assignments in the total sample.

EXAMPLE 1

		Total Sample (Ordered Smallest to Largest)		Ranks	
Placebo	New Drug	Placebo	New Drug	Placebo	New Drug
7	3		1		1
5	6		2		2
6	4		3		3
4	2	4	4	4.5	4.5
12	1	5		6	
		6	6	7.5	7.5
		7		9	
		12		10	

The goal of the test is to determine whether the observed data support a difference in the populations of responses. First, we sum the ranks in each group. In the placebo group, the sum of the ranks is 37; in the new drug group, the sum of the ranks is 18. As a check on our assignment of ranks, we have $n(n+1)/2 = 10(11)/2=55$ which is equal to $37+18 = 55$.

For the test, we call the placebo group 1 and the new drug group 2. $R_1=37$ and $R_2=18$. If the null hypothesis is true (i.e., if the two populations are equal), we expect R_1 and R_2 to be similar. In this example, the lower values (lower ranks) are clustered in the new drug group (group 2), while the higher values (higher ranks) are clustered in the placebo group (group 1). This is suggestive, but is the observed difference in the sums of the ranks simply due to chance? To answer this we will compute a test statistic to summarize the sample information and look up the corresponding value in a probability distribution.

EXAMPLE 1

Test Statistic for the Mann Whitney U Test

The test statistic for the Mann Whitney U Test is denoted **U** and is the ***smaller*** of U_1 and U_2

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where R_1 = sum of the ranks for group 1 and R_2 = sum of the ranks for group 2.

$$U_1 = 5(5) + \frac{5(6)}{2} - 37 = 3 \qquad U_2 = 5(5) + \frac{5(6)}{2} - 18 = 22$$

U=3. Is this evidence in support of the null or research hypothesis?

Smaller values of U support the research hypothesis, and larger values of U support the null hypothesis.

Key Concept:

- For any Mann-Whitney U test, the theoretical range of U is from 0 (complete separation between groups, H_0 most likely false and H_1 most likely true) to $n_1 * n_2$ (little evidence in support of H_1).
- In every test, **$U_1 + U_2$ is always equal to $n_1 * n_2$.** In the example above, U can range from 0 to 25 and smaller values of U support the research hypothesis (i.e., we reject H_0 if U is small).

EXAMPLE 1

- In every test, we must determine whether the observed U supports the null or research hypothesis. Specifically, we determine a critical value of U such that **if the observed value of U is less than or equal to the critical value, we reject H_0 in favor of H_1 and if the observed value of U exceeds the critical value we do not reject H_0.**
- The critical value of U can be found from the table. To determine the appropriate critical value we need sample sizes (for Example: $n_1=n_2=5$) and our two-sided level of significance ($\alpha=0.05$). For the above Example, the critical value is 2, and the decision rule is to reject H_0 if $U \leq 2$. **We do not reject H_0 because 3 > 2.** We do not have statistically significant evidence at $\alpha = 0.05$, to show that the two populations of numbers of episodes of shortness of breath are not equal. However, in this example, the failure to reach statistical significance may be due to low power. **The sample data suggest a difference, but the sample sizes are too small to conclude that there is a statistically significant difference.**

EXAMPLE 2

- A new approach to prenatal care is proposed for pregnant women living in a rural community. The new program involves in-home visits during the course of pregnancy in addition to the usual or regularly scheduled visits. A pilot randomized trial with 15 pregnant women is designed to evaluate whether women who participate in the program deliver healthier babies than women receiving usual care. The outcome is the APGAR score measured 5 minutes after birth. APGAR scores range from 0 to 10 with scores of 7 or higher considered normal (healthy), 4-6 low and 0-3 critically low. The data are shown below.

Usual Care	8	7	6	2	5	8	7	3
New Program	9	9	7	8	10	9	6	

Is there statistical evidence of a difference in APGAR scores in women receiving the new and enhanced versus usual prenatal care?

EXAMPLE 2

- **Step 1.** Set up hypotheses and determine level of significance.

  H_0: The two populations are equal

  H_1: The two populations are not equal. $\alpha = 0.05$

- **Step 2.** Select the appropriate test statistic.
  - Because APGAR scores are not normally distributed and the samples are small ($n_1=8$ and $n_2=7$), we can use the Mann Whitney U test.
  - The test statistic is U, the smaller of U_1 and U_2

- **Step 3.** Set up decision rule.
- The appropriate critical value can be found from the table. To determine the appropriate critical value we need sample sizes ($n_1=8$ and $n_2=7$) and our two-sided level of significance ($\alpha=0.05$).
- The critical value for this test with $n_1=8$, $n_2=7$ and $\alpha = 0.05$ is 10 and the decision rule is as follows: **Reject H_0 if $U \leq 10$.**

EXAMPLE 2

- **Step 4.** Compute the test statistic.
- The first step is to assign ranks of 1 through 15 to the smallest through largest values in the total sample, as follows:

		Total Sample (Ordered Smallest to Largest)		Ranks	
Usual Care	New Program	Usual Care	New Program	Usual Care	New Program
8	9	2		1	
7	8	3		2	
6	7	5		3	
2	8	6	6	4.5	4.5
5	10	7	7	7	7
8	9	7		7	
7	6	8	8	10.5	10.5
3		8	8	10.5	10.5
			9		13.5
			9		13.5
			10		15
				R_1=45.5	R_2=74.5

EXAMPLE 2

- Next, we sum the ranks in each group. In the usual care group, the sum of the ranks is R_1=45.5 and in the new program group, the sum of the ranks is R_2=74.5. Recall that the sum of the ranks will always equal n(n+1)/2. As a check on our assignment of ranks, we have n(n+1)/2 = 15(16)/2=120 which is equal to 45.5+74.5 = 120.
- We now compute U_1 and U_2, as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 8(7) + \frac{8(9)}{2} - 45.5 = 46.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 8(7) + \frac{7(8)}{2} - 74.5 = 9.5$$

Thus, the test statistic is U=9.5.

- **Step 5.** Conclusion:

We reject H_0 because 9.5 $\leq$ 10. We have statistically significant evidence at α =0.05 to show that the populations of APGAR scores are not equal in women receiving usual prenatal care as compared to the new program of prenatal care.

Any question?

REFERENCE

- The detail material related to this lecture can be found in

- D'Agostino RB and Stevens MA. Goodness of Fit Techniques.
- Apgar, Virginia (1953). " A proposal for a new method of evaluation of the newborn infant ". Curr. Res. Anesth. Analg. 32 (4): 260-267.
- Conover WJ. Practical Nonparametric Statistics, 2nd edition, New York: John Wiley and Sons.
- Siegel and Castellan. (1988). "Nonparametric Statistics for the Behavioral Sciences," 2nd edition, New York: McGraw-Hill.