A psychiatric facility is considering a new, experimental, drug for the relief of panic. It is being investigated on just 5 patients suffering from panic disorder. The investigator chooses, *at random*, 3 to receive the new drug; the other 2 are given standard medication. This is a situation where the investigator doesn’t have sufficient sample size nor can assume normality. Thus it is not appropriate to perform a *two sample t-test*. However, it is possible to construct a valid *rank-based* test known as the *Wilcoxon Rank Sum Test*. pssst … You might reasonably argue at this point that a sample size of 5 is too small to do anything! But then I could not develop these course notes for you … so bear with me.

The rank-based analysis does not consider the actual relief scores. Instead, it considers their ranks (the highest relief score is given rank=1 and so on; the lowest relief score is given rank=5). Under the null hypothesis that the experimental drug is the same as the standard medication in alleviating panic, the 5 ranks of relief should be independent of the treatment received. Thus, under the null hypothesis assumed model, we can treat the 5 *ranks themselves* as being assigned at random to the experimental and standard care groups. The null hypothesis further says that all possible sets of 5 assignments are equally likely and that we can expect the average rank in the two groups to be the same (and equal to 3, which is the average of 1, 2, 3, 4, 5).

*Rank-based* analysis methods comprise a substantial (yes, “substantial” is correct; I’m not exaggerating) body of statistical techniques known as *nonparametrics*.
# Table of Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>3</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Some Virtues and Limitations of Nonparametrics</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Guidelines for Choosing the Correct Nonparametric Test</td>
<td>5</td>
</tr>
<tr>
<td>2. One Population – Single Sample</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Sign (Median) Test</td>
<td>6</td>
</tr>
<tr>
<td>2.2 R Illustration</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Stata Illustration</td>
<td>9</td>
</tr>
<tr>
<td>3. One Population – Paired Data</td>
<td>10</td>
</tr>
<tr>
<td>3.1 Sign Test</td>
<td>10</td>
</tr>
<tr>
<td>3.2 Wilcoxon Signed Rank Test</td>
<td>13</td>
</tr>
<tr>
<td>3.3 R Illustration</td>
<td>22</td>
</tr>
<tr>
<td>3.4 Stata Illustration</td>
<td>23</td>
</tr>
<tr>
<td>4. Two Independent Populations – Two Samples</td>
<td>24</td>
</tr>
<tr>
<td>4.1 Wilcoxon Rank Sum Test</td>
<td>24</td>
</tr>
<tr>
<td>4.2 Mann Whitney U Test (Mann Whitney Rank Sum, Rank Sum)</td>
<td>28</td>
</tr>
<tr>
<td>4.3 R Illustration</td>
<td>30</td>
</tr>
<tr>
<td>4.4 Stata Illustration</td>
<td>30</td>
</tr>
<tr>
<td>5. K Independent Populations - Analysis of Variance</td>
<td>31</td>
</tr>
<tr>
<td>5.1 Kruskal-Wallis One Way Analysis of Variance</td>
<td>31</td>
</tr>
<tr>
<td>5.2 Friedman Randomized Block Analysis of Variance</td>
<td>36</td>
</tr>
<tr>
<td>5.3 R Illustration</td>
<td>40</td>
</tr>
<tr>
<td>5.4 Stata Illustration</td>
<td>41</td>
</tr>
<tr>
<td>6. Correlation</td>
<td>42</td>
</tr>
<tr>
<td>6.1 Spearman Rank Correlation</td>
<td>42</td>
</tr>
<tr>
<td>6.2 R Illustration</td>
<td>47</td>
</tr>
<tr>
<td>6.3 Stata Illustration</td>
<td>47</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Introduction

Sometimes we can’t do normal theory based t-tests or analyses of variance, because we have insufficient sample size or because we cannot assume normality (oh and there’s more than one reason for why we might not be able to assume normality).

Example -

In the early testing of AZT for the treatment of AIDS, the drug might have been given to a very small number, \( n=7 \), of patients. In this pilot setting, the investigators might have wanted to test the null hypothesis of no effect against the one sided alternative of harm before proceeding any further with the development of this drug. In these preliminary analyses, the responses might have been coded using a simple 3-point likert scale: “deterioration”, “no change”, or “improvement”.

We cannot do a simple paired t-test to assess improvement in over time for a number of reasons:

1. The outcome (response to AZT) can assume at most a limited number (3) of values. Normality assumes that the range of possible values is infinite (-\( \infty \), +\( \infty \)).
2. The categories "deterioration", "no change", and "improvement" are discrete. Normality assumes that the possible values lie on a continuum.
3. The underlying distribution (especially with only 3 valid outcomes) is not even approximately normal. As well, it is entirely unknown.
4. With a small sample size, it is unreasonable to think that the central limit theorem applies which would let us assume that the distribution of the sample mean is approximately normal.

**Tip** – Consider using rank-based (nonparametric) methods if:

- Your sample size is small (<30); or
- The possible outcome values are limited to a small number of discrete possibilities; or
- Your data are clearly non-normal.
1.2 Some Virtues and Limitations of Nonparametric Tests

Virtues/Advantages:

1. **Nonparametric tests require only minimal assumptions.** Typically, these minimal assumptions are: independence, symmetry and constancy of variance.

2. **Nonparametric tests are intuitively straightforward.** With a little practice, you get the hang of it. Starting with the null, and assuming this model is true, you reason out “what is equally likely” under the null. You then consider the direction of the test statistic values when, instead, the alternative is true. The latter is used to define your p-value calculation, as the calculation is the null hypothesis probability of the “observed or more extreme (as in unfavorable to the null)”.

3. **Nonparametric test statistics are quick and easy** to calculate.

4. **Nonparametric tests are valid!** This is virtue “#1” again but highlights a related virtue. Consider that when the assumption of normality is not appropriate, a normal theory analysis may yield a wrong answer. A correct nonparametric analysis, provided its fewer assumptions are satisfied, will then yield a valid answer.

5. **With increasing sample size, nonparametric tests compete favorably even when it would have been appropriate to do normal theory tests. This is relative efficiency.** Suppose you are not sure if you can do normal theory tests. So to be on the safe side, you do a rank-based nonparametric test. The good news is that, with reasonable sample size, your caution has not cost you much. Much of the time (relative efficiency) you will come to the same conclusion as you would have reached had you performed the normal theory test:

<table>
<thead>
<tr>
<th>Non-Parametric Test</th>
<th>Normal Theory Test</th>
<th>Relative Efficiency (as n→infinity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sign Test</td>
<td>One Sample Paired t</td>
<td>.64</td>
</tr>
<tr>
<td>Wilcoxon Signed Rank</td>
<td>One Sample Paired t</td>
<td>.96</td>
</tr>
<tr>
<td>Wilcoxon Rank Sum</td>
<td>Two Sample t</td>
<td>.96</td>
</tr>
</tbody>
</table>

*Translation of relative efficiency = .96:
Under conditions that are appropriate for the normal theory one sample paired t-test, as the sample size increases substitution of the nonparametric counterpart, the Wilcoxon Signed Rank test, can be expected to yield the same conclusion 96% of the time. That’s pretty good!*

Limitations/Disadvantages:

1. Estimation and the construction of confidence intervals, not covered in these notes, is tedious.

2. The parameters estimated using ranks do not have straightforward interpretations.

3. The magnitudes of the observations are not used in the analysis.
## 1.3 Guidelines for Choosing the Correct Nonparametric Test

<table>
<thead>
<tr>
<th>Nature</th>
<th>Population/Sample</th>
<th>Observation/Data</th>
<th>Relationships/Modeling</th>
<th>Analysis/Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Population – Single Sample</td>
<td>Z-test, t-test</td>
<td>Sign (Median) test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Population – Paired Data</td>
<td>Paired t-test</td>
<td>Wilcoxon Signed Rank test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two Independent Populations – Two Samples</td>
<td>2 sample t-test</td>
<td>Wilcoxon Rank Sum test, Mann Whitney U test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K Independent Populations – Analysis of Variance</td>
<td>One way anova, Randomized block anova</td>
<td>Kruskal Wallis Test, Friedman randomized block anova</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>Pearson product moment</td>
<td>Spearman rank correlation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. One Population – Single Sample

2.1 Sign (Median) Test
Also called the Binomial test.

Example -
Consider again the pilot study of AZT introduced previously. Suppose the 7 patients were followed for a period of 24 hours with the following responses:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Response at 24 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>deterioration (+1)</td>
</tr>
<tr>
<td>2</td>
<td>no change (0)</td>
</tr>
<tr>
<td>3</td>
<td>improvement (-1)</td>
</tr>
<tr>
<td>4</td>
<td>deterioration (+1)</td>
</tr>
<tr>
<td>5</td>
<td>improvement (-1)</td>
</tr>
<tr>
<td>6</td>
<td>improvement (-1)</td>
</tr>
<tr>
<td>7</td>
<td>improvement (-1)</td>
</tr>
</tbody>
</table>

Patients #1 and #4 responses suggest that AZT is harmful. Patients #3, #5, #6 and #7 responses suggest AZT is beneficial. In the one sample sign (median) responses of “no change” are regarded as uninformative and dropped from the analysis. Thus, in this example the data for patient #2 response is regarded as uninformative and is dropped, yielding a final sample size for analysis n=6.

Research Question:
Do the responses for the n=6 patients whose data is informative suggest significant harm or benefit of AZT after 24 hours?

Assumptions:
(1) The responses of the individuals are independent. (how’s that for minimal assumptions!)

Null and Alternative Hypotheses:
H₀: AZT has no effect on patient status at 24 hours
Hₐ: AZT produces deterioration at 24 hours
Reason out *what are equally likely* when the null hypothesis is true:

- When the null is true ➔ The occurrence of improvement or deterioration does not depend in any way on AZT administration.

- If it is further assumed that the progression of the disease is negligible for the 24 hours of observation ➔ Then an individual is just as likely to report improvement at 24 hours as deterioration at 24 hours. ➔ This is equivalent to a "50-50" chance of deterioration at 24 hours:

  \[ \text{Probability} \{ \text{deterioration} \mid H_0 \text{ true} \} = \text{Probability} \{ \text{improvement} \mid H_0 \text{ true} \} = 0.5 \]

**Null and Alternative Hypotheses**

- \( H_0: \) Probability \{ "deterioration at 24 hours" \} = 0.5
- \( H_A: \) Probability \{ "deterioration at 24 hours" \} > 0.5

**Use what are equally likely when the null hypothesis is true to define the test statistic**

- When the null is true, Each individual patient response at 24 hours (“deterioration” v “not”) is a **Bernoulli** \((\pi = .5)\)

- ➔ When the null is true and provided independence holds, The number of patients with “event” (deterioration) out of 6 is a **Binomial** \((n=6, \pi = .5)\)
Definition Sign (Median) Test Statistic:

Let X = # individuals out of 6 who report deterioration at 24 hours.

- n = 6 is the “number of trials”
- π = 0.5 is the probability of event (“deterioration” in this example) when null is true
- X_{observed} = 2

Under the null hypothesis model assumption:

X is distributed Binomial(n=6, π=0.5) and therefore:

\[ E[X | \text{null true}] = n \pi = (6)(0.5) = 3 \]
\[ \text{VAR}[X | \text{null true}] = n \pi (1-\pi) = (6)(0.5)(0.5) = 1.5 \]

Rejection Rule:
The significance level is calculated using the Binomial(n=6, π=0.5). In this example, because H_A is one-sided, the p-value calculation is also one tailed. Specifically, it is in the direction of \( \pi > 0.5 \).

\[ p\text{-value } = \Pr[ X \geq 2 | \text{null is true}] = \Pr[ X \geq 2 | X \sim \text{Binomial}(6, \pi = 0.5) ] \]
\[ = 0.8926 \]

Online Calculator:  [www.artofstat.com > Web Apps > Binomial Distribution. At top, tab: Find Probability]

Interpretation:
No surprises here. We have only 5 observations. The p-value 0.8906 tells us that the assumption of the null hypothesis model and its application to the data have led to a very likely result (p-value = .89). Thus, we have no reason to reject the null hypothesis. These data provide no statistically significant evidence that AZT causes significant deterioration at 24 hours (Well, of course not! We have a teeny sample).
2.2 R Users

```r
# Single Sample - Sign Test (Binomial Test)
library(BSDA)

table1 = read.table(text="
patid  ydeteriorate
1.00  1.00
2.00  0.00
3.00  -1.00
4.00  1.00
5.00  -1.00
6.00  -1.00
7.00  -1.00", header=TRUE)
df1 <- as.data.frame.matrix(table1)

# HA is deterioration at 24 hrs -->
# p-value = Pr [observed or more + signs of deterioration]
# SIGN.test(dataframe$variable, md = 0, alternative = "greater", conf.level = 0.95)
# note: alternative can be either "two.sided", "greater" or "less"
SIGN.test(df1$ydeteriorate, md = 0, alternative = "greater", conf.level = 0.95)
```

One-sample Sign-Test
data: df1$ydeteriorate
s = 2, p-value = 0.8906
alternative hypothesis: true median is greater than 0

--- output omitted ---

2.3 Stata Users

```
. * SINGLE SAMPLE: Sign test ---*
. generate patid=.
. generate ydeteriorate=.
. *(2 variables, 7 observations pasted into data editor)
.
. * drop patid=2 because ydeteriorate=0
. drop if ydeteriorate==0
(1 observation deleted)
.
. * create 0/1 y for sign test
. generate y=ydeteriorate
. recode y (-1=0)
(y: 4 changes made)
.
. * bitest VARIABLE == nullp
. bitest y == 0.5

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Observed k</th>
<th>Expected k</th>
<th>Assumed p</th>
<th>Observed p</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>0.50000</td>
<td>0.33333</td>
</tr>
</tbody>
</table>
Pr(k >= 2) = 0.890625 (one-sided test)
Pr(k <= 2) = 0.343750 (one-sided test)
Pr(k <= 2 or k >= 4) = 0.687500 (two-sided test)
```
3. One Population – Paired Data

3.1 Sign Test
Suppose it is not appropriate to calculate a normal theory paired t-test.

Example -
Nine college students agree to participate in an experiment to test the hypothesis that marijuana use reduces short term memory retention. Each is asked to complete one quiz (pre-marijuana) prior to taking marijuana and a second quiz (post-marijuana) after smoking 5 joints. The two quizzes are identical in the number of questions asked and in degree of difficulty. Suppose the following are observed:

<table>
<thead>
<tr>
<th>Student</th>
<th>Pre-Marijuana # Correct</th>
<th>Post-Marijuana # Correct</th>
<th>(Post) – (Pre) Difference</th>
<th>Sign of Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>-1</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>3</td>
<td>-2</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>5</td>
<td>-1</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
<td>-2</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>-1</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>6</td>
<td>+1</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Students #1 - #6 exhibited memory loss, while student #7 exhibited improvement. Here, too, in the one sample sign test, changes “post-pre” = 0 are regarded as uninformative and dropped from the analysis. Thus, in this example the data for student #8 response is regarded as uninformative and is dropped, yielding a final sample size for analysis n=8.

Introduction to Signs
In this example, interest is in a limited question: “up” or down”. We are asking: does marijuana use result in an increase or decrease in short term memory? Thus, the focus is the direction only, positive or negative, of the change in test scores.

- A positive sign reflects an increase in retention
- A negative sign reflects a decrease in retention.
- The analysis is of the "signs" only.

Null and Alternative Hypotheses:
H₀: Marijuana use has no effect on quiz score
Hₐ: Marijuana use reduces memory retention, one sided (the “sign” of “post – pre” is negative)
Reason out what are equally likely when the null hypothesis is true:

- When the null is true ➔ the change “post – pre” does not depend in any way on marijuana use ➔
  - The null chances are “50-50” that the change “post-pre” is positive or negative ➔
  - Pr[ sign of “post-pre” is negative | null is true ] = 0.5

Use what are equally likely when the null hypothesis is true to define the test statistic

- When the null is true, For each student, the “sign” of change (“post-pre”) is a Bernoulli (π = .5)

- ➔ When the null is true and provided independence holds, The number of students with “sign” that is negative out of 8 is a Binomial(n=8, π = .5)

Assumptions: The individual responses, each defined as a "sign", are independent.

Null and Alternative Hypotheses:

\( H_0: \) marijuana use has no effect on memory retention
\( \pi = \text{probability \{negative sign\}} = 0.5 \)

\( H_A: \) marijuana use reduces memory retention
\( \pi = \text{probability \{negative sign\}} > 0.5 \)

Definition Sign Test Statistic:
Let \( X = \# \) individuals out of 8 for whom change “post-pre” is negative (sign is negative).

- \( n = 8 \) is the “number of trials”
- \( \pi = 0.5 \) is the probability of event (negative sign) when null is true
- \( X_{\text{observed}} = 7 \)

Under the null hypothesis model assumption:
\( X \) is distributed Binomial(n=8, \( \pi=0.5 \)) and therefore:
\[ E [X \mid \text{null true}] = n \pi = (8)(0.5) = 4 \]
\[ \text{VAR} [X \mid \text{null true}] = n \pi (1-\pi) = (8)(0.5)(0.5) = 2 \]
Rejection Rule:

The significance level is calculated using the Binomial(n=8, π=0.5). In this example, too, because $H_A$ is one-sided, the p-value calculation is also one tailed. Specifically, it is in the direction of $π > 0.5$.

\[
p\text{-value} = P[X \geq 7 \mid \text{null is true}] = P[X \geq 7 \mid X \sim \text{Binomial}(8, \pi = 0.5)]
\]

\[
= 0.0352
\]


![The Binomial Distribution](https://istats.shinyapps.io/BinomialDist/)

**Interpretation:**

The p-value = 0.0352 tells us that the assumption of the null hypothesis model and its application to the data have led to an unlikely result and thus a challenge to the null. The null hypothesis is rejected. In this example, the data provide modestly statistically significant evidence ("modestly" because here, too, the sample size is small) that marijuana use is associated with a reduction in memory retention.
### 3.2 Wilcoxon Signed Rank Test

Suppose it is not appropriate to calculate a normal theory paired t-test.

**Example**

A pilot study is being conducted to determine if hypnosis therapy results in increased sleep among insomniacs. Nine insomniacs agree to participate. Each is asked to keep a diary of the number of hours awake each night during the one month prior to and the one month following hypnosis therapy. For each patient, the pre- and post-therapy data are then averaged separately to obtain single “pre” and “post” scores of inability to sleep:

<table>
<thead>
<tr>
<th>Pt</th>
<th>Pre-Therapy Ave # hours awake</th>
<th>Post-therapy Ave # hours awake</th>
<th>Difference [ Pre – Post ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.83</td>
<td>0.878</td>
<td>+ 0.952</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.647</td>
<td>-0.147</td>
</tr>
<tr>
<td>3</td>
<td>1.62</td>
<td>0.598</td>
<td>+1.022</td>
</tr>
<tr>
<td>4</td>
<td>2.48</td>
<td>2.05</td>
<td>+0.430</td>
</tr>
<tr>
<td>5</td>
<td>1.68</td>
<td>1.06</td>
<td>+0.620</td>
</tr>
<tr>
<td>6</td>
<td>1.88</td>
<td>1.29</td>
<td>+0.590</td>
</tr>
<tr>
<td>7</td>
<td>1.55</td>
<td>1.06</td>
<td>+0.490</td>
</tr>
<tr>
<td>8</td>
<td>3.06</td>
<td>4.14</td>
<td>-1.080</td>
</tr>
<tr>
<td>9</td>
<td>1.30</td>
<td>1.29</td>
<td>+0.010</td>
</tr>
</tbody>
</table>

**Why not do the sign test? A case for a rank-based approach:**

- *A limitation of the sign test is that it considers merely “up” or down*. No use is made of the fact that some pre/post changes are big while others are small.
- *If it happens that the alternative is true, then we can reasonably expect to see some big pre/post changes; thus, we’d like our test statistic to be sensitive to these*. Under the alternative, depending on the direction of the alternative, large positive differences are more likely to be seen than are zero differences or large negative differences.
- In particular, large positive signs ought to provide stronger evidence in favor of the alternative than the signs alone.

Thus, rationale for extending the sign test is to incorporate "relative size" into the analysis. We do this using “ranks”. For analyzing a single sample of paired data, the appropriate rank based procedure is the Wilcoxon Signed Rank Test.
Introduction to Ranks

You’re already familiar with the idea - Imagine you are considering which of 3 houses to purchase. You might assign rank=1 to the least expensive, rank =2 to the mid priced house, and rank=3 to the most expensive house. All other things equal, you’re going to buy the rank=1 house!

A rank is a relative magnitude. The idea is to replace the actual data value (for example the price of the house) with its place in the ordered listing from smallest to largest values.

*Take care - the approach to assigning ranks depends on the setting and the question of interest.*

In constructing a Wilcoxon Signed Rank Test, we obtain signed ranks as follows.

**Step 1:**
For each “pre-post” value, obtain the absolute difference by dropping the sign.

**Step 2:**
Drop “pre-post” differences = 0 (they are not informative) from the analysis.
Reduce the sample size by the number of zero differences dropped.

**Step 3:**
Rank the absolute differences from smallest to largest.

*How to deal with ties -*
If two or more absolute differences have the same value, these are called "ties". Assign average ranks to each "tied" value in the group.

*Example –* Suppose that after listing the absolute differences in ascending order, the 5th, 6th, 7th, and 8th absolute differences are all equal.
It’s not possible to assign ranks 5, 6, 7 and 8
So instead we assign their average = \( \frac{5 + 6 + 7 + 8}{4} = 6.5 \) to all of these absolute differences

**Step 4:**
Put the “signs” back. For each ranked absolute difference, attach the “sign” of the observed difference “pre-post”. *The results are called signed ranks.*

*Example – continued*

<table>
<thead>
<tr>
<th>Pt</th>
<th>Signed Difference</th>
<th>Absolute Magnitude</th>
<th>Rank</th>
<th>Signed Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+0.952</td>
<td>0.952</td>
<td>7</td>
<td>+7</td>
</tr>
<tr>
<td>2</td>
<td>-0.147</td>
<td>0.147</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>+1.022</td>
<td>1.022</td>
<td>8</td>
<td>+8</td>
</tr>
<tr>
<td>4</td>
<td>+0.430</td>
<td>0.430</td>
<td>3</td>
<td>+3</td>
</tr>
<tr>
<td>5</td>
<td>+0.620</td>
<td>0.620</td>
<td>6</td>
<td>+6</td>
</tr>
<tr>
<td>6</td>
<td>+0.590</td>
<td>0.590</td>
<td>5</td>
<td>+5</td>
</tr>
<tr>
<td>7</td>
<td>+0.490</td>
<td>0.490</td>
<td>4</td>
<td>+4</td>
</tr>
<tr>
<td>8</td>
<td>-1.080</td>
<td>1.080</td>
<td>9</td>
<td>-9</td>
</tr>
<tr>
<td>9</td>
<td>+0.010</td>
<td>0.010</td>
<td>1</td>
<td>+1</td>
</tr>
</tbody>
</table>
Now we can see the advantages of using ranks over signs.

- The observed changes in numbers of hours awake include some large negative changes (the maximum negative difference has magnitude 1.08) and some large positive changes (the maximum positive difference has magnitude 1.02).

- We can make good use of their relative magnitudes via the use of ranks. For example, if a high percent of the study participants experience large drop (“pre-post” is large) in number of hours awake, this is evidence that hypnosis therapy improves sleep.

Null and Alternative Hypotheses:

H₀: Hypnosis therapy has no effect on average number of hours awake
Hₐ: Hypnosis therapy reduces average number of hours awake (pre-post is positive, yielding large +signed rank)

Assumptions:
The individuals are independent

Reason out what are equally likely when the null hypothesis is true:

- When the null is true ➔ hypnosis therapy has no effect on average number of hours awake. ➔
  - The “pre-post” changes in average hours awake should fluctuate evenly about zero.
  - We expect 50% of the “pre-post” changes to be negative and 50% to be positive.
  - MOREOVER! We also expect that the sizes of the “pre-post” changes to be evenly distributed around zero.

Here, it helps to reason out what we expect when the alternative is true:

- When the alternative is true ➔ hypnosis therapy tends to produce “pre-post” changes that are positive ➔
  - We expect to get “pre-post” changes that are negative infrequently and, when we do,
  - We expect the negative “pre-post” change to be small in size. ➔
  - We expect [ sum of negative signed ranks ] < [ sum of positive signed ranks ]

Introduction to Sum of Positive Ranks, Sum of Negative Ranks:

T⁻ = sum of negative signed ranks
T⁺ = sum of positive signed ranks

Example - continued

T⁻ = sum of negative signed ranks = 11.
T⁺ = sum of positive signed ranks = 34.
Use what are equally likely when the null hypothesis is true to define the test statistic

Step 1:
When the null is true, for each individual, there is a “50-50” chance that the sign of “pre-post” = “+”

\[
\text{Pr [ sign of the rank is positive | null true ]} = \text{Pr [ sign of the rank is negative | null true ]} = 1/2.
\]

Step 2:
Obtain the total # ways to assign “+” and “-” signs to all n of the individual ranks.

Answer:
For each individual, there are 2 possible ways to assign “+” or “-“
Thus, for the entire sample of n individual data values
Total # ways to assign “+” and “-” = (2)(2) … (2) = 2^n

Example -
In this example n=9
If there are 2 ways to assign “+/-” for each individual, then for 9 independent individuals,
Total # ways to assign “+” and “-” = (2)(2) … (2) = 2^9 = 512

Step 3:
Obtain the null model probability of each equally likely configuration of “+” and “-“

Answer:
When the null is true, all configurations of “+” and “-” are equally likely.
Thus, if total # ways to assign “+” and “-” = 2^n
Then, probability [ each configuration of “+” and “-” | null true ] = 1 / [ 2^n ]

\[
\text{Pr [ configuration of n "+" and ",-" signs | null true ]} = \left[ \frac{1}{2} \right]^n
\]

Example -
In this example, the observed signed ranks = \{+7, -2, +8, +3, +6, +5, +4, -9, +1\}
Thus, the observed configuration of signs, “+” and “-“, is \{ + - + + + + - +\}
Pr [ \{ + - + + + + - +\} | null true ] = 1/512
Step 3:
This is just like what we did in BIOSTATS 540 Unit 3, *Probability Basics*. See again pp 16-21. In particular, starting with equally likely configurations of signed ranks, we can solve for the null model probability of each sum of positive ranks $T^+$ and the sum of the negative ranks $T^-$. 

Answer:
Consider that, upon data collection, “+” and “-“ will be assigned to each of the 9 ranks 

$$\{+1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9\}$$ 

Total # ways to do this assignment is $(2 \text{ for rank}=1) \times (2 \text{ for rank}=2) \ldots \times (2 \text{ for rank}=9) = 2^9 \Rightarrow$ 

Pr [ each set of signed ranks | null true ] = $1 / [2^9]$ 

___ii) 
For each set of signed ranks, solve for $T^+$ and $T^-$ 

___iii) 
Now obtain null probability model for the possible outcomes of $T^+$ and $T^-$ 

**Example**
Consider a really small sample size, n=3. The ranks that are to be assigned are simply $\{+1 \ 2 \ 3\}$ 

The total # ways to attach signs to $\{+1 \ 2 \ 3\}$ = $(2)(2)(2) = 2^3 = 8$ 

When the null is true, each configuration occurs with probability = $1 / [8] = .125$ 

This yields the following null model distribution for the signed ranks and the associated null model distributions of $T^+$ and $T^-$ 

| Signed Ranks | Pr[ Signed Ranks | null true ] | $T^+$ | $T^-$ | Pr[$T^+$ | null true ] | $T^+$ | $T^-$ |
|--------------|-----------------|-------|-------|-----------------|-------|-------|
| 1            | +1 +2 +3        | 1/8 = .125 | 6     | 0               |      |       |
| 2            | -1 +2 +3        | 1/8 = .125 | 5     | 1               |      |       |
| 3            | +1 -2 +3        | 1/8 = .125 | 4     | 2               |      |       |
| 4            | +1 +2 -3        | 1/8 = .125 | 3     | 3               |      |       |
| 5            | +1 -2 -3        | 1/8 = .125 | 1     | 5               |      |       |
| 6            | -1 +2 -3        | 1/8 = .125 | 2     | 4               |      |       |
| 7            | -1 -2 +3        | 1/8 = .125 | 3     | 3               |      |       |
| 8            | -1 -2 -3        | 1/8 = .125 | 0     | 6               |      |       |
| **total**    |                 | **1.00**  |       |                 |      |       |

Tip! Notice that the sum of $T^+$ and $T^-$ is always $(1+2+3)=6$. This means that, once we know $T^-$ (or $T^+$), we can always get the other by subtraction.
Example, continued -
These data are from n=9 study participants.

- Observed signed ranks = {+7, -2, +8, +3, +6, +4, -9, +1} ⇒
  - $T^- = 11$
  - $T^+ = 34$

This is the basis for the calculation of significance levels for the Wilcoxon Signed Rank Statistic.

<table>
<thead>
<tr>
<th>Wilcoxon Signed Rank Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>o Sum of all the ranks = sum {1, 2, ..., n} = $\frac{n(n+1)}{2}$ = fixed total</td>
</tr>
<tr>
<td>o $T^- = \left[ \frac{n(n+1)}{2} \right] - T^+$</td>
</tr>
<tr>
<td>o $T^+ = \left[ \frac{n(n+1)}{2} \right] - T^-$</td>
</tr>
</tbody>
</table>

Example, continued

Assumptions:
The individual responses are independent.

Null and Alternative Hypotheses:

$H_0$: hypnosis therapy has no effect on hours awake (“pre – post” = 0)
- Sum of Positive Ranks = Sum of Negative Ranks
  - $T^+ = T^-$

$H_A$: hypnosis therapy reduces number of hours awake (“pre – post” > 0)
- Sum of Positive Ranks > Sum of Negative Ranks
  - $T^+ > T^-$
How to Get an Exact P-value

Because the total of $[1 + 2 + \ldots + n] = \text{fixed total} = (n)(n+1)/2$, we can calculate:

$$p\text{-value} = \Pr \left[ T^+ \geq 34 \right]$$
$$= \Pr \left[ T^- \leq 11 \right]$$

How to Calculate Exact Significance Levels Using Online Calculator

https://www.astatsa.com/WilcoxonTest/

**Step 1:** Choose Wilcoxon signed rank test for paired data, one sample

**Step 2:** Scroll down. At left, enter data as pairs. Choose one sided alternative greater. Choose exact p-value.

**Step 3:** At bottom, click: Proceed to calculate results. You should then see

**Results:**

Wilcoxon signed rank test

Test statistic $V : 34$
$p\text{-value} : 0.101562$
null hypothesis $\mu_0 = 0.0$
alternative hypothesis: greater, $\hat{\mu} > \mu_0$
95% confidence interval : -0.0685 -------- inf
sample estimate of pseudo-median $\hat{\mu} : 0.4375$
Step 4 (R Users, optional): Just below, you’ll get the associated R code (how nice is that!):

```
# Copy-paste these lines into the R command prompt.
# Lines that begin with the # character are taken as comment lines by R.
A <- c(1.83, 0.5, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.3)
B <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 4.14, 1.29)
wilcoxon.test(A, B, paired = TRUE, alternative = "greater", mu = 0.0,
  exact = TRUE, correct = TRUE, conf.int = TRUE, conf.level = 0.95)
```

**Z-Score Approximation**

It is also possible to get an approximate p-value using the Normal(0,1)

We can solve for a z-score test statistic approximation.

A large sample approximation requires knowing $E \{ T^+ \}$ and $\text{Var} \{ T^+ \}$ under the null hypothesis. With a little bit of algebra (I’ll spare you) it can be shown that:

$$E \{ T^+ \mid \text{null true} \} = \frac{n(n+1)}{4}$$

$$\text{VAR} \{ T^+ \mid \text{null true} \} = \frac{n(n+1)(2n+1)}{24}$$

If a continuity correction is incorporated, the Z-score is defined as follows:

$$Z\text{-score} = \left[ \frac{\left( T^+ - \frac{n(n+1)}{4} \right)^{-1/2}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \right]$$

Roughly, the large sample approximation method for the calculation of significance levels should only be used when the number of non-zero differences is 16 or larger.

**How to Handle Ties**

- **Rationale** - If there are ties, then $T^+$ will be less variable than if there are no ties.
- When computing an exact significance level associated with $T^+$, no adjustment is required. Simply use the average signed ranks in the calculation of $T^+$ and proceed as described.
- However, when computing an approximate significance level associated with $T^+$, adjustment is required.
• Specifically, in the solution for the z-score, the formula for the variance of $T^+$ should be made smaller by an amount that is related to the number and pattern of ties in the data.

• $g = \text{The number of groups of ties}$
  - Index the groups 1 to $g$ with $i=1, ..., g$
  - $t_i = \text{the number of tied absolute differences in group } "i"$

**Example, continued**
The number of nonzero differences is 9 and therefore the large sample approximation is not appropriate. However, for illustration:

$$Z = \frac{34 - \frac{9(10)}{4}}{\sqrt{\frac{9(10)(19)}{24}}} = 1.3032$$

• Approximate significance = Pr [ Normal (0,1) ≥ 1.3032 ] = 0.0968
• Compared to the exact p-value = 0.1016, the approximation is not very good.

**It is possible to get an approximate significance level using the same online calculator**
http://astatsa.com/WilcoxonTest/

**Step 5:** Under calculate p-value, click on *approximate, for large samples*.

**Results:**

*Wilcoxon signed rank test with continuity correction*

Test statistic $V$ : 34
p-value : 0.096259

None
null hypothesis $\mu_0 = 0.0$
alternative hypothesis: greater, $\mu > \mu_0$

**Interpretation:**
The approximate p-value is not appropriate to use here because the sample size is so small. The exact p-value = .1016 tells us that the assumption of the null hypothesis model and its application to the data have led to a likely result. The null hypothesis is NOT rejected. We conclude that these data do not provide statistically significant evidence that hypnosis reduces number of hours awake.
3.3 R Users

```r
# Single Sample - Paired Data. Sign Test
# Compute differences. Then do sign test on differences
library(BSDA)
table2 = read.table(text=
student pre post
1.00 4.00 3.00
2.00 5.00 3.00
3.00 6.00 5.00
4.00 2.00 1.00
5.00 3.00 1.00
6.00 5.00 4.00
7.00 5.00 6.00
8.00 3.00 3.00
9.00 2.00 1.00" , header=TRUE)
df2 <- as.data.frame.matrix(table2)
df2$ydiff <- df2$post - df2$pre

# HA is REDUCED memory -->
# p-value = Pr [observed or fewer + signs]
# SIGN.test(dataframe$variable, md = 0, alternative = "less", conf.level = 0.95)
# Alternative can be either "two.sided", "greater" or "less"
SIGN.test(df2$ydiff, md = 0, alternative = "less", conf.level = 0.95)

One-sample Sign-Test

data:  df2$ydiff
s = 1, p-value = 0.03516
alternative hypothesis: true median is less than 0
-- some output omitted --

# Single Sample - Wilcoxon Signed Rank Test
table3 = read.table(text=
patid pre post
1.00 1.83 0.88
2.00 0.50 0.65
3.00 1.62 0.60
4.00 2.48 2.05
5.00 1.68 1.06
6.00 1.88 1.29
7.00 1.55 1.06
8.00 3.06 4.14
9.00 1.30 1.29" , header=TRUE)
df3 <- as.data.frame.matrix(table3)

# wilcox.test( ) calculates difference = 1stvar - 2ndvar
# HA is hypnosis REDUCES hours awake --> (post-pre) is POSITIVE -->
# p-value = Pr [ sum positive ranks is observed or greater]
wilcox.test(df3$pre, df3$post, paired=TRUE,alternative="greater")

Wilcoxon signed rank test

data:  df3$pre and df3$post
V = 34, p-value = 0.1016
alternative hypothesis: true location shift is greater than 0
-- some output omitted --
```
3.4 Stata Users

* --- SINGLE SAMPLE PAIRED DATA: Sign Test ---*
. generate student=.
. generate pre=.
. generate post=.
. *(3 variables, 9 observations pasted into data editor)

. generate ydiff=post-pre
. drop if ydiff==0
(1 observation deleted)

. replace ydiff=0 if ydiff < 1
(7 real changes made)

. bitest ydiff==0.5

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Observed k</th>
<th>Expected k</th>
<th>Assumed p</th>
<th>Observed p</th>
</tr>
</thead>
<tbody>
<tr>
<td>ydiff</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>0.50000</td>
<td>0.12500</td>
</tr>
</tbody>
</table>

\[ Pr(k \geq 1) = 0.996094 \] (one-sided test)
\[ Pr(k \leq 1) = 0.035156 \] (one-sided test)
\[ Pr(k \leq 1 \text{ or } k \geq 7) = 0.070313 \] (two-sided test)

* --- SINGLE SAMPLE PAIRED DATA: Wilcoxon Signed Rank Test using command signrankex (exact) ---*
. * For approximate p-values, use command signrank
. generate patid=.
. generate pre=.
. generate post=.
. *(3 variables, 9 observations pasted into data editor)

. findit signrankex
. * Download st0297 and follow instructions to install. Then type help signrankex.
. signrankex pre=post

Wilcoxon signed-rank test

<table>
<thead>
<tr>
<th>sign</th>
<th>obs</th>
<th>sum ranks</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>7</td>
<td>34</td>
<td>22.5</td>
</tr>
<tr>
<td>negative</td>
<td>2</td>
<td>11</td>
<td>22.5</td>
</tr>
<tr>
<td>zero</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>9</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>

Ho: pre = post

\[ S = 11.500 \]
\[ \text{Prob} \geq |S| = 0.2031 \]
4. Two Independent Samples: Two Samples

4.1 Wilcoxon Rank Sum Test
Suppose it is not appropriate to calculate a normal theory two sample t-test.

Example -
The Department of Psychiatry at the University of Arizona is interested in testing the effectiveness of a new drug for the treatment of severe depression. Five "severely depressed" patients agree to participate in a randomized controlled clinical trial to address this question.

- **Group 1**: 2 receive placebo (old);
- **Group 2**: 3 receive the new drug (new).

After two weeks of therapy, the five patients are again evaluated by the Principal Investigator using the Hamilton Depression Scale (high values indicate more severe depression). The investigator is blind to the treatment assignment. If the three patients on the new drug are judged to be less severely depressed than the two patients who received the old drug, this suggests a relative benefit of the new treatment.

Low Hamilton Depression Scale scores are evidence of treatment benefit.

Assumptions:
(1) The individual responses to treatment are independent.
(2) The variability in the responses to treatment is the same for patients, regardless of drug administered, new versus old.

Null and Alternative Hypotheses:
H₀: Assuming comparability at baseline, the median depression level of patients at 2 weeks is the same for “new” and “old” drug patients.
Hₐ: The median depression assessment score is worse (higher) for patients receiving the old drug compared to those receiving the new drug (one sided).

In constructing a Wilcoxon Rank Sum test, the ranking procedure is now as follows.
**Pool the data.** Rank the pooled data.

**Example** – Suppose the following are obtained

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>old</td>
<td>old</td>
</tr>
<tr>
<td>Depression Score</td>
<td>46</td>
<td>41</td>
<td>35</td>
<td>53</td>
<td>40</td>
</tr>
<tr>
<td>Rank</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
Reason out what are equally likely when the null hypothesis is true:

Step 1: Obtain the total # ways to assign the ranks “1”, “2”, “3”, “4” and “5”

The total number of ways to assign 2 ranks to one group and 3 ranks to the other group is

\[
\binom{5}{2} = \binom{5}{3} = 10 \text{ because "5 choose 2" is the same as "5 leave 3 behind"}
\]

• And, in general, the number of ways to assign \(n_1\) rankings to group #1 and \(n_2\) rankings to group #2 is

\[
\binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2}
\]

Step 2: Each set of assignments of the ranks “1”, “2”, “3”, “4” and “5” is equally likely under the null.

• Example – In this example, when the null hypothesis is true, each of arrangements of rankings is observed with probability

\[
\frac{1}{\binom{5}{2}} = \frac{1}{10}
\]

• In general, when the null hypothesis is true, each of the arrangements of rankings is observed with probability

\[
\frac{1}{\binom{n_1+n_2}{n_1}} = \frac{1}{\binom{n_1+n_2}{n_2}}
\]

Step 3: Let \(S_1 = \text{[sum of ranks in group #1]}\) and \(S_2 = \text{[sum of ranks in group #2]}\).

We only need one (\(S_1\) or \(S_2\)), of course, because the sum of all the ranks is a fixed total.

Step 4: Obtain the complete null hypothesis distribution of \(S_1 = \text{[sum of ranks in group #1]}\)

Or \(S_2\), if you prefer.

• It is convenient to choose group #1 as the smaller sample size group.

• Following are all 10 “equally likely” arrangements of rankings together with their accompanying values of \(S_1\). To the right is the associated null distribution of \(S_1\).
### Wilcoxon Rank Sum Test

Useful tools regarding the two sums of ranks, $S_1$ and $S_2$

1. **Sum of all the ranks**: $\sum \{1, 2, \ldots, (n_1 + n_2)\} = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$ is the fixed total.

2. **$S_2$**:
   
   $$S_2 = \left[ \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} \right] - S_1$$

3. **$S_1$**:
   
   $$S_1 = \left[ \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} \right] - S_2$$

We have what we need to get the exact $p$-value

Quick look back.

$S_1 = \text{sum of ranks in group #1 (old)}$

### Null Hypothesis Distribution of $S_1$

<table>
<thead>
<tr>
<th>$S_1$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.10</td>
</tr>
<tr>
<td>4</td>
<td>.10</td>
</tr>
<tr>
<td>5</td>
<td>.20</td>
</tr>
<tr>
<td>6</td>
<td>.20</td>
</tr>
<tr>
<td>7</td>
<td>.20</td>
</tr>
<tr>
<td>8</td>
<td>.10</td>
</tr>
<tr>
<td>9</td>
<td>.10</td>
</tr>
<tr>
<td>Total=</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Key: Observed
Previously, we noted that evidence of a benefit of the “new” drug is LOWER depression scores. Thus, when the alternative is true, we expect $S_2 = \text{sum of ranks on “new”} = \text{low}$ and $S_1 = \text{sum of ranks on “old”} = \text{high}$. Thus the p-value calculation we want here is the following:

- p-value = Pr $[ S_1 \geq 7 | \text{null model}]$
  
  = Pr[S_1=7] + Pr[S_1=8] + Pr[S_1=9]$
  
  
  = .40

Note – I looked for an online calculator to do this p-value calculation, so that you would have it, but I was not able to find one that would let me do an analysis with sample sizes of 2 and 3 (okay; that’s fair).

**Interpretation:**
The exact p-value = .40. Thus, the assumption of the null hypothesis model and its application to the data have led to a likely result. The null hypothesis is NOT rejected. We conclude that these data do not provide statistically significant evidence that the “new” drug is effective in reducing depression.

**Z-Score Approximation**

It is also possible to get an approximate p-value using the Normal(0,1)

- Again, sparing you the slog, under the null hypothesis assumption model it can be shown that:

  $E \{ S_1 | \text{null true } \} = n_1(n_1 + n_2 + 1)/2$

  $\text{VAR } \{ S_1 | \text{null true } \} = n_1n_2(n_1 + n_2 + 1)/12$

**Z-Score Approximation to the Wilcoxon Rank Sum Statistic**

$Z$-score = $\left[ \frac{S_1 - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}} \right]$  

or  

$Z$-score = $\left[ \frac{S_2 - \frac{n_2(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}} \right]$

Notes:
(1) There are two choices of z-score, depending on which test statistic, $S_1$ or $S_2$, is being standardized.
(2) This approximation is not good if $n_1 \leq 10$ or $n_2 \leq 10$. 
4.2 Mann Whitney U Test

Also called Mann Whitney Rank Sum Test, or simply Rank Sum Test.

This section may be skipped.
I say “may be skipped” because the Mann Whitney U test is equivalent to the Wilcoxon Rank Sum test. In fact, it can be thought of as the application of some algebra to the Wilcoxon Rank Sum test.

The material that follows is for the interested reader who wants to understand the correspondence.

Notation

<table>
<thead>
<tr>
<th></th>
<th>Group #1</th>
<th>Group #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smaller sample size</td>
<td>$n_1$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Sum of ranks</td>
<td>$S_1$</td>
<td>$S_2$</td>
</tr>
</tbody>
</table>

The Mann Whitney U test focuses on a special kind of “pairing”

- **Definition of $T_1$**
  - Total # pairings
    How many ways can you pair one observation from group #1 with one observation from group #2?
    Answer = $n_1n_2$
  - $T_1 = \# \text{ pairings for which } [\text{ group } #1 \text{ observation } ] < [\text{ group } #2 \text{ observation } ]$
    This will be our test statistic $T_1$
  - Remember $S_1 = \text{ sum of ranks for observations in group } 1$?
    A nice bit of algebra reveals a very convenient equivalent formula that is much more usable:
    $$T_1 = S_1 - \left[ \frac{n_1(n_1 + 1)}{2} \right]$$

- **Definition of $T_2$**
  - Total # pairings
    How many ways can you pair one observation from group #2 with one observation from group #1?
    Answer = $n_1n_2$
  - $T_1 = \# \text{ pairings for which } [\text{ group } #2 \text{ observation } ] < [\text{ group } #1 \text{ observation } ]$
    This will be our test statistic $T_2$
  - Similarly, recall. $S_2 = \text{ sum of ranks for observations in group } 2$? Now we have
    $$T_2 = S_2 - \left[ \frac{n_2(n_2 + 1)}{2} \right]$$
Equivalence of Wilcoxon Rank Sum Test and Mann Whitney U Test p-value calculations

See again, pp 25-26. The smaller sample size group received the “old” drug and $n_1=2$. Previously, we noted that evidence of a benefit of the “new” drug is LOWER depression scores. Thus, when the alternative is true, we expect $S_2$ = sum of ranks on “new” = low and $S_1$ = sum of ranks on “old” = high. Thus the p-value calculation we want here is the following:

**Wilcoxon Rank Sum Test** p-value = $\Pr [ S_1 \geq 7 | \text{null model} ]$

$= \Pr [S_1=7] + \Pr [S_1=8] + \Pr [S_1=9]$

$= .20 + .10 + .10$

$= .40$

**Mann Whitney U Test** p-value is obtained as follows.

\[
p\text{-value calculation} = \Pr [ S_1 \geq 7 ] = \Pr ( S_1 - \left[ \frac{n_1(n_1+1)}{2} \right] \geq 7 - \left[ \frac{2(2+1)}{2} \right] )
\]

$= \Pr ( T_1 \geq 4 )$
### 4.3 R Users

```r
# Two Independent Samples - Wilcoxon Rank Sum/Mann Whitney U
table4 = read.table(text="
patid  group  ydepress
1.00  1.00  46.00
2.00  1.00  41.00
3.00  1.00  35.00
4.00  0.00  53.00
5.00  0.00  40.00", header=TRUE)
df4 <- as.data.frame.matrix(table4)

# wilcox.test( ) wants data in wide format
new <- subset(df4, group==1)
old <- subset(df4, group==0)

# wilcox.test( ) calculates sum of ranks in smaller sample size group
# HA is that drug reduces depression --> group=old will have higher scores
# p-value is based on smaller sample size group. This is group=old
# --> p-value = [sum of ranks in group=old is observed or greater]
wilcox.test(old$ydepress,new$ydepress, alternative="greater")

Wilcoxon rank sum test
data: old$ydepress and new$ydepress
W = 4, p-value = 0.4
alternative hypothesis: true location shift is greater than 0
```

### 4.4 Stata Users

```stata
. * ---- TWO INDEPENDENT SAMPLES: Wilcoxon Rank Sum/Mann Whitney using command ranksumex (exact) ---*
. * For approximate p-values, use command ranksum
. generate patid=. 
. generate group=. 
. generate ydepress=. 
. *(3 variables, 5 observations pasted into data editor)

. * ranksumex YVARIABLE, by(GROUPVARIABLE)
. ranksumex ydepress, by(group)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

<table>
<thead>
<tr>
<th>group</th>
<th>obs</th>
<th>rank sum</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>combined</td>
<td>5</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Exact statistics
Ho: ydepress(group==0) = ydepress(group==1)
Prob <= 5 = 0.4000
Prob >= 7 = 0.4000
Two-sided p-value = 0.8000
```

<table>
<thead>
<tr>
<th>Nature</th>
<th>Population/ Sample</th>
<th>Observation/ Data</th>
<th>Relationships/ Modeling</th>
<th>Analysis/ Synthesis</th>
</tr>
</thead>
</table>
5. K Independent Populations - Analysis of Variance

5.1 Kruskal Wallis One Way Analysis of Variance

Suppose it is not appropriate to calculate a normal theory one way analysis of variance.

Example -

A study is conducted to compare values of a particular urine determination among three groups of infants: (1) term (2) preterm and (3) preterm with acidosis at 1-3 weeks of age. There are N=15 babies in total, with 5 in each of the 3 groups.

Following are the urine determination values.

<table>
<thead>
<tr>
<th>Group 1 Term</th>
<th>Group 2 Pre-Term</th>
<th>Group 3 (K=3) Acidosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>3.2</td>
<td>7.3</td>
</tr>
<tr>
<td>3.9</td>
<td>4.6</td>
<td>8.4</td>
</tr>
<tr>
<td>5.0</td>
<td>5.1</td>
<td>6.9</td>
</tr>
<tr>
<td>4.8</td>
<td>4.9</td>
<td>8.2</td>
</tr>
<tr>
<td>4.1</td>
<td>4.3</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Assumptions:

(1) Independence - The individual responses are all mutually independent.
(2) Homogeneity of variance - The variability in the values of the urine determination is the same in all 3 groups

Null and Alternative Hypotheses:

H_0: Equality of Medians - The distributions of the urine determination values are identical in all 3 groups.
H_A: Not – In at least one group, the distribution of the urine determination values is different.

Ranking Procedure:

No surprise here. The approach is the same as what we do for the Wilcoxon Rank Sum Test comparison of 2 groups.

- Pool all the data values, across all K groups, into one series and rank.
- Then, for each group “i”, calculate the sum of the ranks for that group. Call this R_i

<table>
<thead>
<tr>
<th>Nature</th>
<th>Population/ Sample</th>
<th>Observation/ Data</th>
<th>Relationships/ Modeling</th>
<th>Analysis/ Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example, continued -

Ranks and values of $R_i$

<table>
<thead>
<tr>
<th>Group 1 Term</th>
<th>Group 2 Pre-Term</th>
<th>Group 3 Acidosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>11</td>
</tr>
</tbody>
</table>

$R_1 = 5+2+9+7+3 = 26$  
$R_2 = 1+6+10+8+4 = 29$  
$R_3 = 13+15+12+14+11 = 65$

Reason out what are equally likely when the null hypothesis is true:

When the null hypothesis is true, all possible assignments of 5 ranks to group 1, 5 ranks to group 2, and 5 ranks to group 3 are equally likely.

- The total number of ways to assign 5, 5, and 5 ranks to groups 1, 2, and 3 is

$$\binom{15}{5,5,5} = \binom{15}{5} \binom{10}{5} \binom{5}{5}$$

$$= \binom{15}{5} \binom{10}{5} [1] \text{ because } \binom{5}{5} = 1$$

- More generally, if there are N in total, K groups, and sample sizes $n_1, n_2, \ldots, n_K$ , the total number of ways to assign $n_1, n_2, \ldots, n_K$ ranks to groups 1, 2, ..., K is

$$\binom{N}{n_1, n_2, \ldots, n_K} = \binom{N}{n_1} \binom{N-n_1}{n_2} \binom{N-n_1-n_2}{n_3} \cdots \binom{N-n_1-\cdots-n_{K-1}}{n_K}$$

$$= \binom{N}{n_1} \binom{N-n_1}{n_2} \binom{N-n_1-n_2}{n_3} \cdots \binom{N-n_1-\cdots-n_{K-2}}{n_K}$$
Example, continued
When the null is true, each of the arrangements of 15 rankings, 5 per group, is observed with probability

\[
\Pr(\text{each assignment of 5, 5, and 5 ranks to groups 1, 2, and 3 | null true}) = 1/\left(\begin{array}{c}15 \\ 5 \ 5 \ 5 \end{array}\right)
\]

Use *what are equally likely when the null hypothesis is true* to define the test statistic

Recall that sum total of all the ranks 1, 2, …. N is (1 + 2 + … + N)

\[
\text{Sum of ranks (1,2,3,....N)} = \frac{N(N+1)}{2}
\]

So when the null is true, on average, *the rank of each individual* is expected to be equal to the average of (1, 2, … N):

\[
\text{Average of all ranks} = \bar{R} = \frac{1+2+...+N}{N} = \left(\frac{1}{N}\right) \frac{N(N+1)}{2} = \frac{(N+1)}{2}
\]

In turn, when the null is true, *the average rank in each group* is also expected to be equal to the average of (1, 2, … N).

\[
E [ \bar{R}_i | \text{null true}] = \bar{R}_-
\]

The *Kruskal Wallis test (K)* measures how close each group-specific \( \bar{R}_i \) is to the overall average \( \bar{R}_- \).

\[
K = \frac{12}{(N)(N+1)} \sum_{i=1}^{K} n_i (\bar{R}_i - \bar{R}_-)^2 = \frac{12}{(N)(N+1)} \sum_{i=1}^{K} n_i \left( \bar{R}_i - \frac{N+1}{2} \right)^2
\]
Alternative Formula
Here is another formula for the Kruskal Wallis test (K) that is easier to work with. It works with the sums of the ranks in each group, rather than the group-specific averages.

\[ K = \frac{12}{(N)(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(N+1) \]

Rejection of the null hypothesis occurs for LARGE values of K.

Example, continued -

\[ K = \frac{12}{(15)(16)} \left[ \frac{26^2}{5} + \frac{29^2}{5} + \frac{65^2}{5} \right] - 3(15+1) = 9.42 \]

We let the computer get the exact p-value for us.

How to Calculate Exact Significance Levels Using Online Calculator

http://vassarstats.net From home page > ordinal data > Kruskal Wallis test > For K=3. Enter raw data at right.

Dear reader – I suspect that because we have n=5 in each group, the online calculator did a chi square approximate p-value calculation. As you’ll see on the next page, the p-value matches my chi square approximate p-value.

Interpretation:
The exact p-value = .009. Assumption of the null hypothesis model and its application to the data have led to a very unlikely result. The null hypothesis is rejected. We conclude that these data provide statistically significant evidence that urine determination levels differ among the 3 groups: “term” versus “pre-term” versus “acidosis”. Note – at this point, further analyses would be performed to explore the nature of these group differences.
Chi Square Approximation
It is also possible to get an approximate p-value using the Chi Square distribution with \( df = (K - 1) \)

When there are too many groups and/or too many observations, obtain an approximate p-value using a Chi Square Distribution with \( df = K - 1 \)

Example, continued -
Observed Chi Square \( df=2 = 9.42 \)
p-value = .009


![Chi-Squared Distribution](https://istats.shinyapps.io/ChisqDist/)

**Adjustment for Ties**
There is an adjustment for ties but this is not discussed here. Unless there are many ties, the adjustment does not amount to much.
5.2 Friedman Randomized Complete Blocks Analysis of Variance
Suppose it is not appropriate to calculate a normal theory two way randomized complete blocks analysis of variance.

Recall the two way randomized complete blocks design.
This analysis of variance design is introduced in BIOSTATS 640 Unit 6 (Analysis of Variance). It is an extension of the paired t-test and is helpful when it is of interest to compare a set of treatments but the setting is one where there is a lot of variability among the individuals. The idea is to match/partition (as best as possible) the subjects into blocks, within which the individuals are homogeneous, and then to compare the treatments of interest within the blocks.

Example -
A study is being conducted to compare the effects of T=3 levels of dose of treatment on outcome Y in n=24 rabbits. Thus, the number of treatment conditions is T=3. The rabbits available for study, however, are extremely variable in their weight. To control for this, so as to improve the precision of the analysis of treatment effects, the 24 rabbits are grouped into 8 groups of 3 rabbits each. Within each group, the 3 rabbits are similar (“matched”/”homogeneous”) in weight. An anova “parlance”, we say the 24 rabbits are partitioned into N=8 “homogeneous” blocks of 3 animals each.

<table>
<thead>
<tr>
<th>Block</th>
<th>Treatment 1 (dose = 2.29)</th>
<th>Treatment 2 (dose = 3.63)</th>
<th>Treatment 3 (dose = 5.75)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>64</td>
<td>62</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>48</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>49</td>
<td>34</td>
<td>61</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>63</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>33</td>
<td>41</td>
<td>56</td>
</tr>
<tr>
<td>6</td>
<td>37</td>
<td>64</td>
<td>62</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
<td>34</td>
<td>57</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>64</td>
<td>72</td>
</tr>
</tbody>
</table>

Assumptions:
(1) Independence - The individual responses are all mutually independent.
(2) Homogeneity of variance - The variability in the outcomes is the same in all 8 blocks.

Null and Alternative Hypotheses:
\(H_0:\) Within each block, there is no effect of treatment (dose)
\(H_A:\) At least one treatment (dose) yields different outcomes than the others
Reason out **what are equally likely** when the null hypothesis is true:

Because this is an extension of pairing (in particular, instead of matching to obtain pairs we now match to obtain “homogenous” blocks), **ranking is done separately within each block.** Then, for each treatment (dose of drug) “i”, calculate the sum of the ranks for that treatment. Call this \( R_i \)

\[
R_i = \text{Sum of ranks of outcomes } Y \text{ for dose}=i, \text{ taken over all } N \text{ blocks}
\]

**Example, continued –**

**Ranks and values of \( R_i \)**

<table>
<thead>
<tr>
<th>Block</th>
<th>Treatment 1 (dose = 2.29)</th>
<th>Treatment 2 (dose = 3.63)</th>
<th>Treatment 3 (dose = 5.75)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

\[
R_1 = \frac{1+1+2+1+1+2+1}{6} = 10
\]

\[
R_2 = \frac{3+2+1+2+2+3+1+2}{8} = 16
\]

\[
R_3 = \frac{2+3+3+3+3+2+3+3}{8} = 22
\]

**Notation**

- \( T = \# \text{ treatments} = 3 \)
- \( N = \# \text{ blocks} = 8 \)
- \( R_{ij} = \text{Rank of } Y \text{ for rabbit in block “i” and treatment “j”} \)

**Use what are equally likely when the null hypothesis is true to define the test statistic**

Within each of the 8 blocks (i=1, 2, …, 8) the \( T=3 \) outcomes (j=1,2,3) \( Y_{ij} \) are ranked from 1 to 3. Under the null, all are equally likely.

**Example** -

In Block i=1:  
\[ \{ R_{11}, R_{12}, R_{13} \} = \text{rearrangement (permutation) of } [1, 2, 3] \]  \# ways = 3!

In Block i=2:  
\[ \{ R_{21}, R_{22}, R_{23} \} = \text{rearrangement (permutation) of } [1, 2, 3] \]  \# ways = 3!

In Block i=3:  
\[ \{ R_{31}, R_{32}, R_{33} \} = \text{rearrangement (permutation) of } [1, 2, 3] \]  \# ways = 3!

Total # arrangements of rankings = \((3!)(3!)(3!)=3!^3\)

\[
\Pr \{ \text{each arrangement } | \text{ null true } \} = 1 / \{3!^3\}
\]

**In general** -

In Block i=1:  
\[ \{ R_{11}, R_{12}, \ldots, R_{1T} \} = \text{rearrangement (permutation) of } [1, 2, \ldots, T] \]  \# ways = \(T!\)

In Block i=2:  
\[ \{ R_{21}, R_{22}, \ldots, R_{2T} \} = \text{rearrangement (permutation) of } [1, 2, \ldots, T] \]  \# ways = \(T!\)

\[ \ldots \]

In Block i=N:  
\[ \{ R_{N1}, R_{N2}, \ldots, R_{NT} \} = \text{rearrangement (permutation) of } [1, 2, \ldots, T] \]  \# ways = \(3!^3\)

Total # arrangements of rankings = \((T!)(T!)(T!)=(T!)^3\)

\[
\Pr \{ \text{each arrangement } | \text{ null true } \} = 1 / \{(T!)^3\}
\]
For each (“i-th”) block:

\[ \text{Sum of the T ranks in } i\text{th block} = \frac{T(T+1)}{2} \rightarrow \]

\[ \text{Average of the T ranks in } i\text{th block} = \left[ \frac{1}{T} \right] \frac{T(T+1)}{2} = \frac{(T+1)}{2} \]

\[ = \bar{R}_i \]

For each (“i-th”) block, reason out what we expect when the null hypothesis is true:

\[ \mathbb{E}[\bar{R}_i \mid \text{null true}] = \bar{R}_\cdot \]

The Friedman test (\(Q\)) measures how close each group-specific \(\bar{R}_i\) is to the overall average \(\bar{R}_\cdot\).

\[ Q = \frac{(12)(N)}{(T)(T+1)} \sum_{j=1}^{T} \left( \bar{R}_i - \frac{T+1}{2} \right)^2 \]

Here is an equivalent formula for \(Q\) that is easier to calculate if you are doing it by hand (probably you are not!)

\[ Q = \frac{12}{NT(T+1)} \sum_{j=1}^{T} R_i^2 - 3N(T+1) \]

Rejection of the null hypothesis occurs for LARGE values of \(Q\).

Example, continued -

\[ Q = \frac{12}{(8)(3)(4)} \sum_{j=1}^{T} [10^2 + 16^2 + 22^2] - 3(8)(4) = 9 \]
We let the computer get the exact p-value for us.

How to Calculate Exact Significance Levels Using Online Calculator

http://vassarstats.net From home page > ordinal data > Friedman test > For K=3. Enter raw data at right.

Dear reader – Here, the online calculator is telling me that “sufficiently large” begins at about N=7 blocks. So here, too, I suspect that the online calculator did a chi square approximate p-value calculation.

Interpretation:
The p-value = .01. Assumption of the null hypothesis model and its application to the data have led to an unlikely result. The null hypothesis is rejected. Conclude that these data provide statistically significant evidence that dose of treatment is associated with differences in outcome Y.

Chi Square Approximation
It is also possible to get an approximate p-value using the Chi Square distribution with df=(T-1)

Online Calculator www.artofstat.com > Web Apps > Chi Square Distribution. At top, tab: Find Probability

Example, continued -
Observed Chi Square \( df=2 = 9 \)

p-value = .01
## 5.3 R Users

**# K Independent Groups Kruskal Wallis Test**

**# Note - Can use this with K=2 for Wilcoxon Rank Sum/Mann Whitney**

```r
table5 = read.table(text="
infantid  group  yurine
1   0   4.5
2   0   3.9
3   0   5
4   0   4.8
5   0   4.1
6   1   3.2
7   1   4.6
8   1   5
9   1   4.9
10  1   4.3
11  2   7.3
12  2   8.4
13  2   6.9
14  2   8.2
15  2   6.2", header=TRUE)
df5 <- as.data.frame.matrix(table5)

#kruskal.test(Yvariable ~ GROUPvariable, data = dataframe)
kruskal.test(yurine ~ group, data = df5)

Kruskal-Wallis rank sum test

data:  yurine by group
Kruskal-Wallis chi-squared = 9.4118, df = 2, p-value = 0.009042
```

**# Friedman Randomized Completely Block Analysis of Variance**

```r
library(tidyr)  # to convert from WIDE to LONG

table6 = read.table(text="
block  dose229  dose363  dose575
1.00  17.00   64.00   62.00
2.00  21.00   48.00   72.00
3.00  49.00   34.00   61.00
4.00  54.00   63.00   91.00
5.00  33.00   41.00   56.00
6.00  37.00   64.00   62.00
7.00  40.00   34.00   57.00
8.00  16.00   64.00   72.00", header=TRUE)
df6 <- as.data.frame.matrix(table6)

# Convert WIDE to LONG: Command gather( ) in package=tidyr
# longdf <- widedf %>% gather(longpredictor, longyoutcome, 1STVAR:LASTVAR)
long_df6 <- df6 %>% gather(dose, yrabbit, dose229:dose575)

# Friedman Completely Randomized Block Analysis of Variance
# friedman.test(YVARIABLE ~ PREDICTOR|BLOCK, data=LONGDATAFRAME)
friedman.test(yrabbit ~ dose|block, data=long_df6)

Friedman rank sum test

data:  yrabbit and dose and block
Friedman chi-squared = 9, df = 2, p-value = 0.01111
```
5.4 Stata Users

. * ---- K Independent Samples: Kruskal Wallis One Way Anova ---*
  . generate infantid=.
  . generate group=.
  . generate yurine=.
  . * (3 variables, 15 observations pasted into data editor)
  . sort GROUPVARIABLE
  . * kwallis YVARIABLE, by(GROUPVARIABLE)
  . sort group
  . kwallis yurine, by(group)

Kruskal-Wallis equality-of-populations rank test

<table>
<thead>
<tr>
<th>group</th>
<th>Obs</th>
<th>Rank Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>26.50</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>28.50</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>65.00</td>
</tr>
</tbody>
</table>

chi-squared = 9.395 with 2 d.f.
probability = 0.0091

chi-squared with ties = 9.412 with 2 d.f.
probability = 0.0090

. * ---- Friedman Randomized Complete Blocks Anova ---*
  . generate block=.
  . generate dose229=.
  . generate dose363=.
  . generate dose575=.
  . * (4 variables, 8 observations pasted into data editor)
  . findit friedman
  . * Download snp_1 and follow instructions to install. Then type help friedman.

  . * drop the block variable
  . drop block

  . * transpose rows and columns so that rows = treatments and columns = blocks
  . transpose, clear

  . * friedman v1-vLAST where LAST = last block number
  . friedman v1-v8

Friedman = 9.0000
Kendall = 0.5625
P-value = 0.0111
6. Correlation

6.1 Spearman Rank Correlation
Suppose it is not appropriate to calculate a normal theory Pearson Product Moment correlation (r)

Recall the Pearson Product Moment Correlation (r)
See again BIOSTATS 540 Unit 12 (Simple Linear Regression and Correlation) pp 41-45 . The Pearson product moment correlation (r) is a measure of the strength of the linear relationship between two continuous variables, say X and Y. It is related to the associated simple linear regression.

Introduction to the Spearman Rank Correlation (r_s)
Provided that both variables, X and Y, are ordinal, it is possible to calculate a measure of the association between the two using the Spearman's Rank Correlation coefficient, r_s. The Spearman rank correlation (r_s) is a measure of the strength of the monotone increasing/decreasing relationship between two continuous or two ordinal variables, say X and Y.

The Spearman Rank Correlation (r_s) is the rank analogue of the Pearson Product Moment correlation. The calculations are the same as for the Pearson Product moment correlation but is based on the ranks in place of the values themselves.

Consider calculating the Spearman Rank Correlation (rs) if:

-- Your two variables X and Y are ordinal
-- The ranges are very limited
-- Interest is more general than that of a linear relationship
-- In particular, you are interested in monotone increasing/decreasing relationships

Example -
Is intelligence, as measured by IQ, associated with a personality score obtained from psychological testing? The psychological test is such that "Type A" personalities score high while "Type B" personalities score low. The following are observed for 8 individuals:

<table>
<thead>
<tr>
<th>Individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ=X</td>
<td>20</td>
<td>17</td>
<td>15</td>
<td>19</td>
<td>23</td>
<td>21</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>Personality Score=</td>
<td>90</td>
<td>94</td>
<td>100</td>
<td>103</td>
<td>113</td>
<td>114</td>
<td>118</td>
<td>119</td>
</tr>
</tbody>
</table>

Step 1: Rank the values of each variable separately

<table>
<thead>
<tr>
<th>Individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank(IQ)=R</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Rank(Personality Score)=S</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>
Step 2:
Use either of 2 (equivalent) formulae for calculating the Spearman’s Rank Correlation $r_S$:

**Formula #1**
Rank analogue of Pearson Product Moment Correlation, $r$:

\[
\begin{align*}
S_r &= \frac{\sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^{n} (S_i - \bar{S})^2}} = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sum_{i=1}^{n} (R_i - \bar{R})^2}
\end{align*}
\]

\[
\bar{R} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} R_i = (N+1)/2 \quad \text{and} \quad \bar{S} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} S_i = (N+1)/2
\]

\[
\text{This works because} \quad \sum_{i=1}^{n} (R_i - \bar{R})^2 = \sum_{i=1}^{n} (S_i - \bar{S})^2
\]

**Example, continued**
Since $\bar{R} = \bar{S} = (N+1)/2 = (8+1)/2 = 4.5$ we have

<table>
<thead>
<tr>
<th>Individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank(IQ)=R</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Rank(Personality Score)=S</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>$\left(R_i - \bar{R}\right)$</td>
<td>1.5</td>
<td>-0.5</td>
<td>-2.5</td>
<td>0.5</td>
<td>3.5</td>
<td>2.5</td>
<td>-1.5</td>
<td>-3.5</td>
<td>-</td>
</tr>
<tr>
<td>$\left(S_i - \bar{S}\right)$</td>
<td>-3.5</td>
<td>-2.5</td>
<td>-1.5</td>
<td>-0.5</td>
<td>0.5</td>
<td>1.5</td>
<td>2.5</td>
<td>3.5</td>
<td>-</td>
</tr>
<tr>
<td>$\left(R_i - \bar{R}\right) \left(S_i - \bar{S}\right)$</td>
<td>-5.25</td>
<td>1.25</td>
<td>3.75</td>
<td>-0.25</td>
<td>1.75</td>
<td>3.75</td>
<td>-3.75</td>
<td>-12.25</td>
<td>-11.0</td>
</tr>
<tr>
<td>$\left(R_i - \bar{R}\right)^2$</td>
<td>2.25</td>
<td>0.25</td>
<td>6.25</td>
<td>0.25</td>
<td>12.25</td>
<td>6.25</td>
<td>2.25</td>
<td>12.25</td>
<td>42.00</td>
</tr>
</tbody>
</table>

\[
r_S = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sum_{i=1}^{n} (R_i - \bar{R})^2} = \frac{-11.00}{42.00} = -.2619
\]
Formula #2
A Computationally simpler formula!

- Calculate \( D = \sum_{i=1}^{n} (S_i - R_i)^2 \) Then,

- \( r_s = 1 - \frac{6D}{N^3 - N} \)

Example, continued -

<table>
<thead>
<tr>
<th>Individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank(IQ)=R</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Rank(Personality Score)=S</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>((R_i - S_i)^2)</td>
<td>25</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>16</td>
<td>49</td>
<td>106</td>
</tr>
</tbody>
</table>

\[
r_s = 1 - \frac{6D}{N^3 - N} = 1 - \frac{6(106)}{512 - 8} = -0.2619 \text{ which matches.}
\]

Formula #3
Formula to use when there are ties (eeew – it’s ugly):

- \( g_X = \# \) groups of ties among the X values, indexed by "i"
- \( g_Y = \# \) groups of ties among the Y values, indexed by "j"
- \( t_i = \# \) ties in the \( i^{th} \) group of ties among the X values
- \( t_j = \# \) ties in the \( j^{th} \) group of ties among the Y values

\[
r_{S,TIES} = \frac{2\left[ N^3 - N \right] - \left[ \frac{\sum_{i=1}^{g_S} (t_i^3 - t_i)}{12} \right] - \left[ \frac{\sum_{j=1}^{g_T} (t_j^3 - t_j)}{12} \right]}{\sqrt{\left[ \frac{N^3 - N}{12} \right] - \left[ \frac{\sum_{i=1}^{g_S} (t_i^3 - t_i)}{12} \right] - \left[ \frac{\sum_{j=1}^{g_T} (t_j^3 - t_j)}{12} \right]}}
\]
Significance Tests for Zero Correlation Using $r_s$:

Null and Alternative Hypotheses:

$H_0$: Spearman’s Rank Correlation $r_s = 0$

$H_A$: Not.

**Solution 1 - # pairs is small (say 4-30)**

We let the computer get the exact p-value for us.

**How to Calculate Exact Significance Levels Using Online Calculator**

[http://vassarstats.net](http://vassarstats.net) From home page > ordinal data > Rank Order Correlation. Enter raw data at right.

![Image of Vassar Stats Online Calculator](image)

Interpretation:

The Vassar Stats online calculator is telling us that with $n=8$, if we want to know if our $r_s$ observed $= -0.26$ is significant at the .05 level, one sided, then the critical value is $r_s$ critical $= -0.62$. Since $r_s$ observed $= -0.26$ does not exceed $r_s$ critical $= -0.62$ in the negative direction, we conclude that this correlation is NOT statistically significantly different from zero.
Solution II - # pairs is > 30
We can use the t-test approximation. See again BIOSTATS 540 Unit 12 (Simple Linear Regression and Correlation) pp 41-45.

Example, continued

\[
t_{df=(n-2)} = r_S \sqrt{\frac{n-2}{1-r_S^2}}
\]

\[
t_{df=(n-2)} = r_S \sqrt{\frac{n-2}{1-r_S^2}} = \frac{(-0.2619)\sqrt{6}}{\sqrt{1-[-0.2619]^2}} = \frac{(-0.2619)(2.4495)}{0.9651} = -0.6647
\]

P-value (one sided) = \( \Pr [\text{Student's } t_{DF=6} \leq -0.6647 ] = 0.27 \)

Online Calculator www.artofstat.com > Web Apps > t Distribution. At top, tab: Find Probability

Interpretation:
The conclusion is the same. With an approximate p-value = .27 (note – we probably did not have sufficient sample size for this but I did the illustration anyway), we conclude that this correlation is NOT statistically significantly different from zero.
### 6.2 R Users

```
# Spearman Rank Correlation
table7 = read.table(text="
patid  xiq  yperson
1.00  20.00  90.00
2.00  17.00  94.00
3.00  15.00  100.00
4.00  19.00  103.00
5.00  23.00  113.00
6.00  21.00  114.00
7.00  16.00  118.00
8.00  12.00  119.00", header=TRUE)
df7 <- as.data.frame.matrix(table7)
cor(df7$xiq, df7$yperson, method="spearman")
   [1] -0.2619048
cor.test(df7$xiq, df7$yperson, method="spearman")
   data: df7$xiq and df7$yperson
   S = 106, p-value = 0.5364
   alternative hypothesis: true rho is not equal to 0
   sample estimates:
       rho
-0.2619048
```

### 6.3 Stata Users

```
. * ---- Spearman Rank Correlation ---*
. generate patid=.
. generate xiq=.
. generate yperson=.
   *(3 variables, 8 observations pasted into data editor)*

. spearman xiq yperson

    Number of obs =     8
   Spearman's rho  =   -0.2619

Test of Ho: xiq and yperson are independent
    Prob > |t|  =   0.5309
```