Unit 3.
Introduction to Nonparametrics

"Don't ask what it means, but rather how it is used"
- I. Wittgenstein

A psychiatric facility is considering a new, experimental, drug for the relief of panic. It is being investigated in a pilot study of just 5 patients suffering from panic disorder; understandably, pilot studies are often small. Here, the investigator chooses, ***at random***, 3 to receive the new drug; the other 2 are given standard medication.

As n=5 is a very small sample size, it is not possible to assume normality and, thus, 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***. Okay, you might reasonably argue that n=5 is still too small for meaningful inference. Bear with me here; I'm focused on the ideas only.

A rank-based analysis does not consider the outcomes themselves (here, the relief scores). It considers their ***ranks*** (the highest relief score is given rank=1 and so on; the lowest relief score is given rank=5).

So, what does the null hypothesis say here? 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 body of statistical techniques known as ***nonparametrics***.

Table of Contents

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 (note – there is 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 (null: zero improvement in over time) for several reasons:

1. ***There are too few possible outcomes***.
   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 outcomes are discrete***.
   The categories "deterioration", "no change", and "improvement" are discrete. Normality assumes that the possible values lie on a continuum.

3. ***Normality cannot be assumed***.
   The underlying distribution (especially with only 3 valid outcomes) is not even approximately normal. Also, it is entirely unknown.

4. ***The sample size is too small***.
   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 cannot be assumed to be distributed 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 test statistic values are likely to be when, instead, the alternative is true. Remember: the direction of the alternative is used to define your p-value calculation: p-value = 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 perform well (tend to get right answer) even when it would have been appropriate to do normal theory tests. This is the idea of 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 a 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:

Non-Parametric Test	Normal Theory Test	Relative Efficiency (as n→infinity)
Sign Test	One Sample Paired t	.64
Wilcoxon Signed Rank	One Sample Paired t	.96
Wilcoxon Rank Sum	Two Sample t	.96

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) The 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

	Parametric Test	Nonparametric Test
One Population – Single Sample	Z-test, t-test	Sign (Median) test
One Population – Paired Data	Paired t-test	Wilcoxon Signed Rank test
Two Independent Populations – Two Samples	2 sample t-test	Wilcoxon Rank Sum test, Mann Whitney U test
K Independent Populations – Analysis of Variance	One-way Anova, Randomized Block Anova	Kruskal Wallis Test, Friedman Randomized Block Anova
Correlation	Pearson product moment	Spearman rank correlation

2. One Population – Single Sample

2.1 Sign (Median) Test

Spoiler. You already know how to do this! This is just 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:

Patient	Response at 24 hours (score)
1	deterioration (+1)
2	no change (0)
3	improvement (-1)
4	deterioration (+1)
5	improvement (-1)
6	improvement (-1)
7	improvement (-1)

Patients #1 and #4 responses (deterioration scored as a +1) suggest that AZT is harmful. Patients #3, #5, #6 and #7 (improvement scored as a –1) 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 statistically 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:

H_O: AZT has no effect on patient status at 24 hours

H_A: AZT produces deterioration at 24 hours, *one sided*

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

- When the null hypothesis 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:

$$\text{Probability } \{\text{deterioration} \mid H_O \text{ true}\} = \text{Probability } \{\text{improvement} \mid H_O \text{ true}\} = 0.5$$

Null and Alternative Hypotheses
H_O: 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 Trial ($\pi = .5$)***

- When the null is true and provided independence holds,
  The count 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"
- $\pi = 0.5$ 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 ($n=6$, $\pi=0.5$) and therefore:

$E[X \mid \text{null true}] = n\,\pi = (6)\,(0.5) = 3$

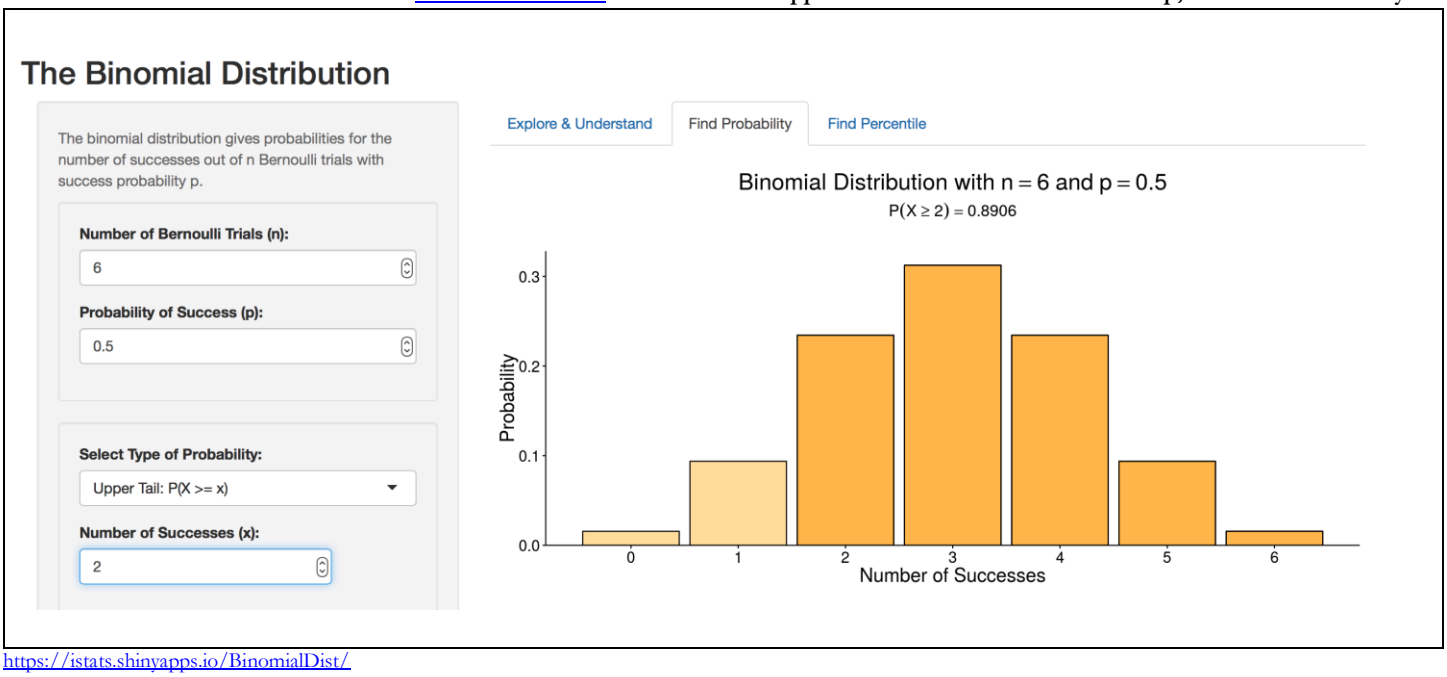
$VAR[X \mid \text{null true}] = n\,\pi\,(1-\pi) = (6)\,(0.5)\,(0.5) = 1.5$

Rejection Rule:

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

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

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



https://istats.shinyapps.io/BinomialDist/

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 Illustration

```
# Single Sample - Sign Test (Binomial Test)
# SIGN.test( ) in package {BSDA}.  Don't forget to install this package first.
library(BSDA)

table1 = read.table(text="                  # Note – This R code is a quick and easy way to create a small dataframe
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            Assumption of the null has NOT led to an unlikely result. Do NOT reject the null.
 alternative hypothesis: true median is greater than 0
  Upper Achieved CI     0.9922     -1    Inf

--- output omitted --
```

3. One Population – Paired Data

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

3.1 Sign Test

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

Example -

Nine college students (n=9) consent to participate in an experiment to test the hypothesis that marijuana use reduces short term memory retention. Participants 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:

Participant	Pre-Marijuana # Correct	Post-Marijuana # Correct	(Post) – (Pre) Difference	Sign of Difference negative or positive
1	4	3	-1	- "negative"
2	5	3	-2	-
3	6	5	-1	-
4	2	1	-1	-
5	3	1	-2	-
6	5	4	-1	-
7	5	6	+1	+ "positive"
8	3	3	0	0
9	2	1	-1	-

Participants #1 - #6 exhibited memory loss, while student #7 exhibited improvement. ***Here, 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:

H_O: Marijuana use has no effect on quiz score

H_A: 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
  - Probability [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 participant, the "sign" of change ("post-pre") is a ***Bernoulli ($\pi = .5$)***

- When the null is true and provided independence holds,
  The number of participants with a negative "sign" (out of 8) is a ***Binomial ($n=8, \pi = .5$)***

Assumptions:

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

Null and Alternative Hypotheses:

H_O: marijuana use has no effect on memory retention
π = probability {negative sign} = 0.5

H_A: marijuana use reduces memory retention
π = probability {negative sign} > 0.5

Definition Sign Test Statistic:

Let X = # participants (out of 8) for whom change "post-pre" is negative (sign is negative).

- n = 8 is the "number of trials"
- π = 0.5 is the probability of event (negative sign) when null is true
- $X_{observed} = 7$

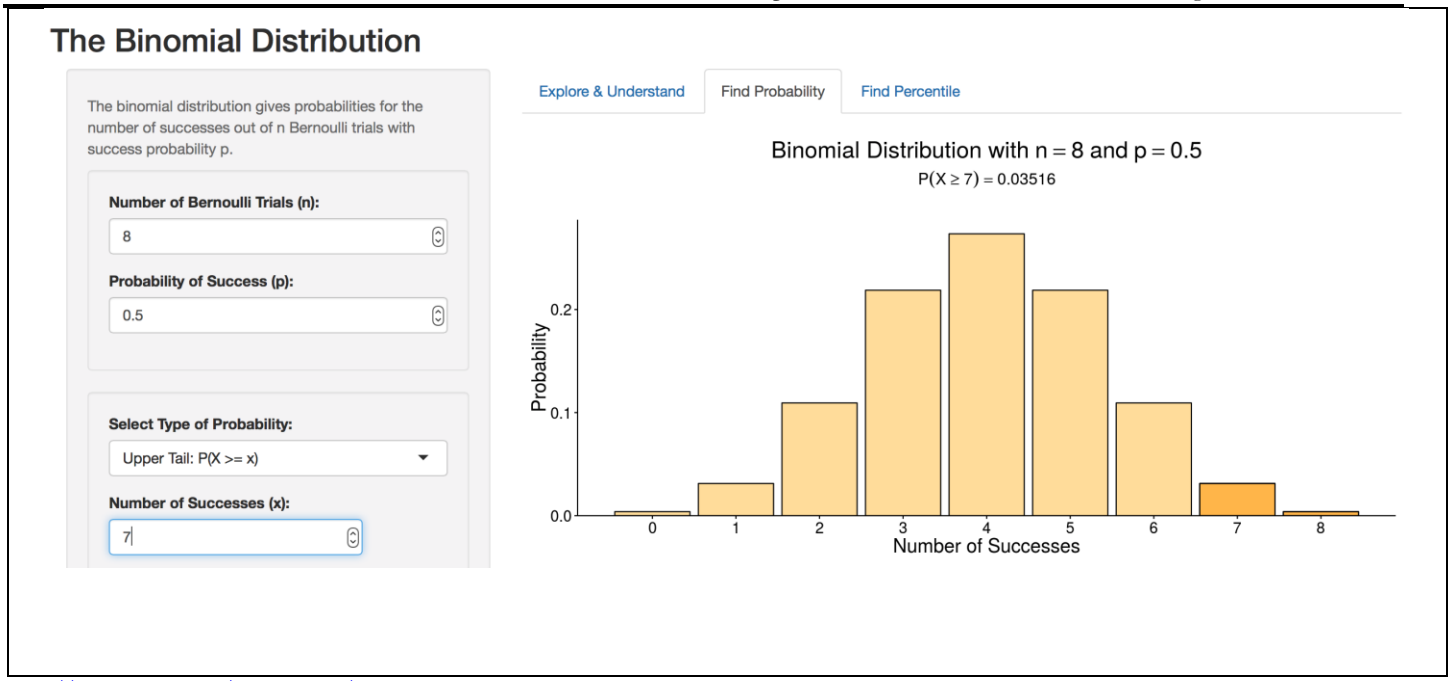
Under the null hypothesis model assumption:
X is distributed Binomial ($n=8, \pi=0.5$) and therefore:
E [X | null true] = $n \pi = (8)(0.5) = 4$
VAR [X | null true] = $n \pi (1-\pi) = (8)(0.5)(0.5) = 2$

Rejection Rule:

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

$$\text{p-value} = \Pr[X \geq 7 \mid \text{null is true}] = \Pr[X \geq 7 \mid X \sim \text{Binomial}(8, \pi = 0.5)]$$
$$= 0.0352$$

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



https://istats.shinyapps.io/BinomialDist/

Interpretation:
The p-value = .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 so 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 (Pt) consent 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:

Pt	Pre-Therapy Ave # hours awake	Post-therapy Ave # hours awake	Difference [Pre – Post]
1	1.83	0.878	+ 0.952

2	0.5	0.647	**-0.147**
3	1.62	0.598	**+1.022**
4	2.48	2.05	**+0.430**
5	1.68	1.06	**+0.620**
6	1.88	1.29	**+0.590**
7	1.55	1.06	**+0.490**
8	3.06	4.14	**-1.080**
9	1.30	1.29	**+0.010**

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 sizes of the pre/post changes. Some are big while others are small. It would be good to make use of this additional evidence!

- ***In patricular, if the alternative is true, then we can reasonably expect to see some big pre/post changes. 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 sized positive*** signs provide stronger evidence in favor of the alternative than the ***signs alone***.

Thus, the advantage of the Wilcoxon Signed Rank test over the sign test is that the former makes use of the relative magnitudes of the pre/post changes 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 actual outcomes (e.g., \$165K, \$210K, \$275K) are replaced by their position in an ordered line-up from smallest to largest, (e.g., RANK[\$165K] = 1, RANK[\$210K] = 2, and RANK[\$275K] = 3). ***Take care – as you will see, the approach to assigning ranks depends on the setting and the question of interest.***

How to obtain signed ranks for a Wilcoxon Signed Rank Test.

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". Take the ranks that would have been assigned, average these, and then assign the same average to each "tie".

Example – Suppose that after listing the absolute differences in ascending order, the 5th, 6th, 7th, and 8th absolute differences are all equal.
The ranks that would have been assigned are the following: 5, 6, 7 and 8
Average these: average = (5 + 6 + 7 + 8)/4 = 6.5 to all of these absolute differences.
Take care! The next rank will then be "9", as if no averaging had occurred and as if "5","6", "7" and "8" had been assigned.

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

Pt	Signed Difference [Pre – Post]	Absolute Magnitude [Pre – Post]	Rank	Signed Rank
1	+ 0.952	0.952	7	+ 7
2	-0.147	0.147	2	-2
3	+1.022	1.022	8	+8
4	+0.430	0.430	3	+3
5	+0.620	0.620	6	+6
6	+0.590	0.590	5	+5
7	+0.490	0.490	4	+4
8	-1.080	1.080	9	-9
9	+0.010	0.010	1	+1

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 proportion of the study participants have large "pre-post" changes in number of hours awake, this is evidence that hypnosis therapy improves sleep.

Null and Alternative Hypotheses:
H_O: Hypnosis therapy has no effect on average number of hours awake
H_A: 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 a person's 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.

Thinking ahead to p-value calculations, 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, each participant has a "50-50" chance of a positive (+) "pre-post" change

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

Step 2:
Obtain the total # ways to assign positive (+) and negative (-) signs to n ranks (n=number of participants).

Answer:
For each participant (hence, each rank), there are 2 possible ways to assign "+" or "-"
Thus, for the entire sample of n participants (hence, n ranks)
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 participant (hence, each rank),
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]

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

- **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:
Spoiler. For those of you that were in this class, this is just like what we did in BIOSTATS 540 Unit 3, ***Probability Basics***. See again pp 16-21 of the BIOSTATS 540 Unit 3 notes. First, make a list of all the equally likely configurations of signed ranks. Second, use this to solve for the null model probability of each sum of positive ranks $\mathbf{T^+}$ and the sum of the negative ranks $\mathbf{T^-}$.

Reasoning of the solution:

* Our data for analysis will be one (observed) configuration of positives (+) and negatives (-) assigned to the n=9 ranks **{ 1 2 3 4 5 6 7 8 9 }**

* How many configurations are possible?
Answer:
Total # ways to do this assignment is (2 for rank=1) x (2 for rank=2) …. x (2 for rank=9) = 2^9

* When the null hypothesis is true, all our equally likely. So, what is the null probability of each?
Answer:
Pr [each set of signed ranks | null true] = 1 / [2^9]

* For each configuration of signed ranks, solve for $\mathbf{T^+}$ and $\mathbf{T^-}$

* Now obtain null probability model for the possible outcomes of T^+ and $\mathbf{T^-}$

Example

For illustration, consider n=3. The ranks that are to be assigned are: **1 2 3**
The total # ways to attach signs to **1 2 3** = (2 ways) (2 ways) (2 ways) = $2^3 = 8$
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-**

Distribution of triplets of signed ranks				
	Signed Ranks	**Pr[Signed Ranks | 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		

Distribution T+ = sum of positive ranks	
T+	**Pr[T+ | null true] = Pr[T- | null true]**
0	1/8 = .125
1	1/8 = .125
2	1/8 = .125
3	1/8 + 1/8 = .250
4	1/8 = .125
5	1/8 = .125
6	1/8 = .125
	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; e.g., $T^+ = 6 - T^-$ and $T^- = 6 - T^+$

Example, continued -
In our example, n=9 and we observe the following.

- Observed signed ranks = {+7, -2, +8, +3, +6, +5, +4, -9, +1}
  T- = [2 + 9] = 11
  T+ = [7 + 8 + 3 + 6 + 5 + 4 + 1] = 34

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

Wilcoxon Signed Rank Test

- Sum of all the ranks = sum { 1, 2, …, n } = $\left[\frac{n(n+1)}{2}\right]$ = fixed total
- $T^- = \left[\frac{n(n+1)}{2}\right] - T^+$
- $T^+ = \left[\frac{n(n+1)}{2}\right] - T^-$

Example, continued

Assumptions:
The individual responses are independent.

Null and Alternative Hypotheses:

H_O: Hypnosis therapy has no effect on hours awake ("pre – post" = 0)
Sum of Positive Ranks = Sum of Negative Ranks
$\mathbf{T^+ = T^-}$

H_A: Hypnosis therapy reduces number of hours awake ("pre – post" > 0)
Sum of Positive Ranks > Sum of Negative Ranks
$\mathbf{T^+ > T^-}$

How to Get an Exact P-value

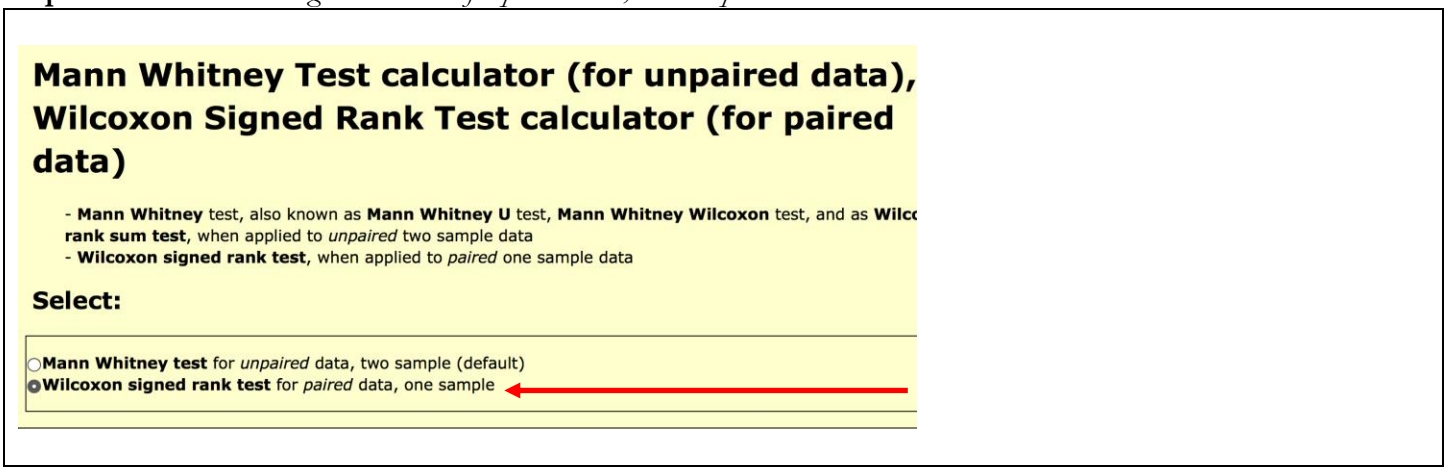
Because the total of [1 + 2 + …. + n] = fixed total = (n)(n+1)/2, we can calculate:

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

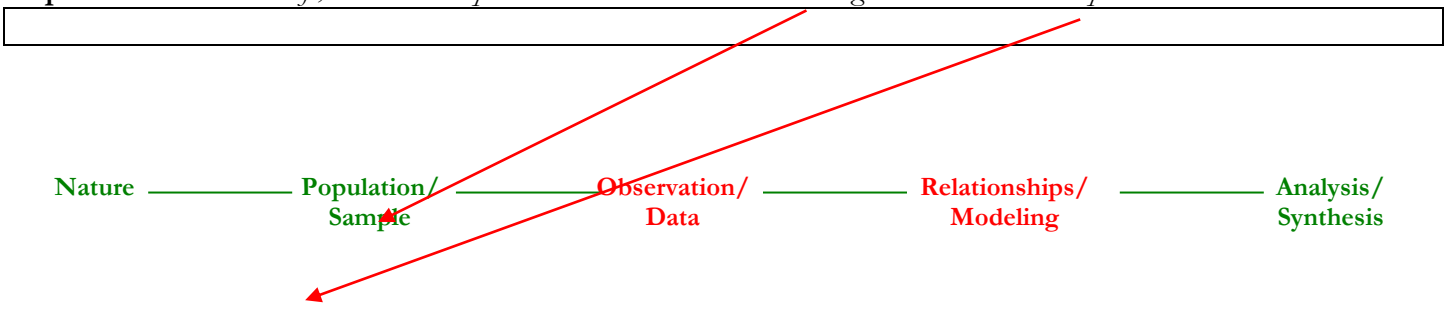
Online Calculator: Wilcoxon Signed Rank Test

There are several, obviously. Some are easier to use than others. You can choose. I found this one to be the most straightforward.
https://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.*



Enter your two columns of *paired* numerical data below:

Paired [A, B] across rows, comma or space separated

```
1.83  0.878
0.50  0.647
1.62  0.598
2.48  2.050
1.68  1.060
1.88  1.290
1.55  1.060
3.06  4.140
1.30  1.290
```

Clear

Additional parameters and options:

Alternative hypothesis	Null hypothesis	Confidence interval of $\hat{\mu}$
○ two sided, estimated location shift $\hat{\mu} \neq \mu_0$	location shift (mean) μ_0	● 95%
● greater, $\hat{\mu} > \mu_0$	0 Replace/ edit	○ 99%
○ lesser, $\hat{\mu} < \mu_0$	default $\mu_0 = 0$	
Calculate exact p-value	Continuity correction	
● exact	● True, Yes	
○ approximate, for large samples	○ False, No	

Proceed to calculate results ← Click here to view the results.

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

Results:

Wilcoxon signed rank test

Test statistic V : 34

p-value : 0.101562 ← Assumption of the null hypothesis has NOT led to an unlikely result (p-value = .10). Do NOT reject the null

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 (Optional – good to know?): *Just below, you'll get the associated* R *code (how nice is that!):*

R code to reproduce these results:

```
# 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)

wilcox.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 [$\mathbf{T^+}$] and Var [$\mathbf{T^+}$] under the null hypothesis. With a little bit of algebra (I'll spare you) it can be shown that:

$$E\{ \mathbf{T^+} \mid \text{null true} \} = n(n+1)/4$$
$$VAR\{ \mathbf{T^+} \mid \text{null true} \} = n(n+1)(2n+1)/24$$

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

Z-Score Approximation to the Wilcoxon Signed Rank Statistic

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

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. This makes sense, right?
- 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 = The number of groups of ties
  - Index the groups 1 to g with i=1, ..., g

- t_i = 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:

$$\text{Z-score} = \left[\frac{\left(T^{+} - \frac{n(n+1)}{4} \right) - 1/2}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \right] = \left[\frac{\left(34 - \frac{9(10)}{4} \right) - 1/2}{\sqrt{\frac{9(10)(19)}{24}}} \right] = 1.3032$$

- Approximate significance = Pr [Normal (0,1) $\geq$ 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 ← Not surprising. Again, assumption of the null hypothesis has NOT led to an unlikely result (p-value = .10). Do NOT reject the null.

None

null hypothesis $\mu_0 = 0.0$

alternative hypothesis: greater, $\hat{\mu} > \mu_0$

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 a person's average number of hours awake.

3.3 R Illustration

Sign Test

```
# Use function SIGN.test( ) in package {BSDA}
# In this setting a preliminary step is to compute differences.  This followed by a 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                    Assumption of the null HAS led to an unlikely result. REJECT the null.
 alternative hypothesis: true median is less than 0
 -- some output omitted --
```

Wilcoxon Signed Rank Test

```
# 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 --> (pre-post) is POSITVE -->
# 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        Assumption of the null has NOT led to an unlikely result. Do NOT reject the null.
 alternative hypothesis: true location shift is greater than 0
-- some output omitted --
```

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 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 (key: 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 might (okay the sample size is ridiculously small here) suggest 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:

H_O: Assuming comparability at baseline, the median depression level of patients at 2 weeks is the same for "new" and "old" drug patients.

H_A: 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; that is, ignore treatment group.

Example – The following are obtained. The best outcome is the lowest score; so it gets rank = 1, and so on.

Patient ID	1	2	3	4	5
Randomization	new	new	new	old	old
Depression Score	46	41	35	53	40
Rank	4	3	1	5	2

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

$$1/\binom{5}{2} = 1/10$$

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

$$1/\binom{n_1+n_2}{n_1} = 1/\binom{n_1+n_2}{n_2}$$

Step 3: Let S_1 = [sum of ranks in group #1] and S_2=[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 = [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.

Arrangement of Ranks				Null Hypothesis Distribution of S_1	
Group #1 (old)	**Group #2 (new)**	**Probability**	**Value of S_1**	**S_1**	**Probability**
1,2	3,4,5	1/10	3 = 1+2	3	.10
1,3	2,4,5	1/10	4	4	.10
1,4	2,3,5	1/10	5	5	.20
1,5	2,3,4	1/10	6	6	.20
2,3	1,4,5	1/10	5	7	.20
2,4	1,3,5	1/10	6	8	.10
2,5	1,3,4	1/10	7	9	.10
3,4	1,2,5	1/10	7	**Total=**	**1.00**
3,5	1,2,4	1/10	8		
4,5	1,2,3	1/10	9		

Key: This configuration is our observed data!

Wilcoxon Rank Sum Test

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

- **Sum of all the ranks = sum { 1, 2, …, $(n_1 + n_2)$ } =** $\left[\frac{(n_1+n_2)(n_1+n_2+1)}{2}\right]$ **= fixed total**
- **S_2 =** $\left[\frac{(n_1+n_2)(n_1+n_2+1)}{2}\right]$ **- 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

We use the null hypothesis distribution of S_1 that we just reasoned out! Here is it again.

S_1 = sum of ranks in group #1 (old)

Null Hypothesis Distribution of S_1	
S_1	Probability
3	.10
4	.10
5	.20
6	.20
7	.20
8	.10
9	.10
Total=	1.00

Key: Observed, more extreme relative to the null

A little juggling here. Sorry. To solve for the correct way to calculate the p-value we need to think in steps: When the alternative is true:

- Hamilton depression scores will be LOWER
- This means we expect S_2 = sum of ranks on "new" to be LOWER
- And therefore, this also means we expect ***S_1 = sum of ranks on "old" TO BE HIGHER***.
- Thus, the p-value calculation we want here is the following:

- p-value = $\Pr [S_1 \geq 7 \mid \text{null model}]$
  $= \Pr[S_1=7] + \Pr[S_1=8] + \Pr[S_1=9]$
  $= [.20] + [.10] + [.10]$
  $=$ **.40**

Sigh. I was not able to find an online calculator that would allow me to work with sample sizes of 2 and 3. That seems fair when you consider that you can't really learn much from sample sizes of 2 and 3.

Interpretation:

The exact p-value = .40 suggests that the assumption of the null hypothesis model and its application to the data have NOT led to an unlikely 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:

$$E \{ S_1 \mid \text{null true} \} = n_1 (n_1 + n_2 + 1)/2$$
$$VAR \{ S_1 \mid \text{null true} \} = n_1 n_2 (n_1 + n_2 + 1)/12$$

Z-Score Approximation to the Wilcoxon Rank Sum Statistic

$$\text{Z-score} = \left[\frac{\left(S_1 - \frac{n_1(n_1+n_2+1)}{2} \right)}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \right] \quad \textit{or} \quad \text{Z-score} = \left[\frac{\left(S_2 - \frac{n_2(n_1+n_2+1)}{2} \right)}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \right]$$

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.

It's okay if you want to skip this section ... maybe.
Why? 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.

Consider reading this section if you want to understand what makes the two tests equivalent.
Short answer: With some algebra, you can get from one test to the other.

Notation

	Group #1	Group #2
	Smaller sample size	Larger sample size
Sample size	n_1	n_2
Sum of ranks	S_1	S_2

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 = $n_1 n_2$
  - **T_1 = # pairings for which [group #1 observation] < [group #2 observation]**
    This will be our test statistic T_1
  - Remember S_1 = 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 = $n_1 n_2$

- **T_1 = # pairings for which [group #2 observation] < [group #1 observation]**
  This will be our test statistic T_2

- Similarly, recall. S_2 = 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 26-27. 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 = $\Pr[S_1 \geq 7 \mid \text{null model}]$
$= \Pr[S_1=7] + \Pr[S_1=8] + \Pr[S_1=9]$
$= [.20] + [.10] + [.10]$
$= .40$

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

$$\text{p-value} = \Pr[S_1 \geq 7]$$
$$= \Pr\left(S_1 - \left[\frac{n_1(n_1+1)}{2}\right] \geq 7 - \left[\frac{2(2+1)}{2}\right]\right)$$
$$= \Pr(T_1 \geq 4)$$

4.3 R Illustration

```
# Two Independent Samples - Wilcoxon Rank Sum/Mann Whitney U
# Use function wilcox.test( )
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)

# GOOD TO KNOW!!!  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.          Assumption of the null has NOT led to an unlikely result. Do NOT reject the null.
 alternative hypothesis: true location shift is greater than 0
```

5. K Independent Populations - Analysis of Variance

5.1 Kruskal Wallis One Way Analysis of Variance

Dear Class. This may seem a little bit out of order because it is an analysis of variance "like" setting. Analysis of variance is the focus of BIOSTATS 640 Unit 6., Analysis of Variance. However, because the Kruskal Wallis test is a non-parametric analysis, I am including it here, so that all the nonparametric inference related course notes are in one place. Apologies for any confusion – cb.

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.

The following data (urine determination values) are observed.

Group 1 ***Term***	**Group 2** ***Pre-Term***	**Group 3 (K=3)** ***Acidosis***
4.5	3.2	7.3
3.9	4.6	8.4
5.0	5.1	6.9
4.8	4.9	8.2
4.1	4.3	6.2

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_O: 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, ignoring group.
- Then, separately for each group "i", calculate the sum of the ranks for that group. Call this R_i

Example, continued -

Ranks and values of R_i

Group 1 ***Term***	**Group 2** ***Pre-Term***	**Group 3** ***Acidosis***
5	1	13
2	6	15
9	10	12
7	8	14
3	4	11
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

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

$$= \binom{15}{5}\binom{10}{5}[1]$$ because there is just 1 way to" choose all 5 from 5"

- 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

$$\binom{N}{n_1\ n_2\ \ldots\ n_K} = \binom{N}{n_1}\binom{N-n_1}{n_2}\cdots\binom{N-n_1-\ldots n_{K-2}}{n_{K-1}}\binom{n_K}{n_K}$$

$$= \binom{N}{n_1}\binom{N-n_1}{n_2}\cdots\binom{N-n_1-\ldots n_{K-2}}{n_{K-1}}$$

Example, continued

When the null is true, each of the arrangements of 15 rankings, 5 per group, is observed with probability

$$\Pr[\text{each assignment of 5, 5, and 5 ranks to groups1, 2, and 3} \mid \text{null true}] = 1/\binom{15}{5\ 5\ 5}$$

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

Recall that the sum of the ranks 1, 2, …. N is = 1 + 2 + … + N

$$\text{Sum of ranks } (1,2,3,\ldots N) = \frac{N(N+1)}{2}$$

When the null is true, ***each individual*** has expected rank equal to the average of the ranks (total/sample size):

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

Also, when the null is true, ***the average rank in each group*** is expected to be equal to the average of the ranks.

$$E[\bar{R}_i \mid \text{null true}] = \bar{R}_{..}$$

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

$$K = \frac{12}{(N)(N+1)}\sum_{i=1}^{K} n_i\left(\bar{R}_i - \bar{R}_{..}\right)^2 = \frac{12}{(N)(N+1)}\sum_{i=1}^{K} n_i\left(\bar{R}_i - \frac{N+1}{2}\right)^2$$

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!

Online Calculator: Kruskal Wallis One Way Analysis of Variance

http://vassarstats.net

From home page > ***ordinal data > Kruskal Wallis test > For K=3.*** Enter raw data at right.

Data Entry

	Ranks for Sample			Raw Data for Sample		
count	A	B	C	A	B	C
1	5	1	13	5	1	13
2	2	6	15	2	6	15
3	9	10	12	9	10	12
4	7	8	14	7	8	14
5	3	4	11	3	4	11
Reset	Calculate from Ranks			Calculate from Raw Data		

Mean Ranks for Sample		
A	B	C
5.2	5.8	13

H = 9.42
df = 2
P = 0.009 *

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.

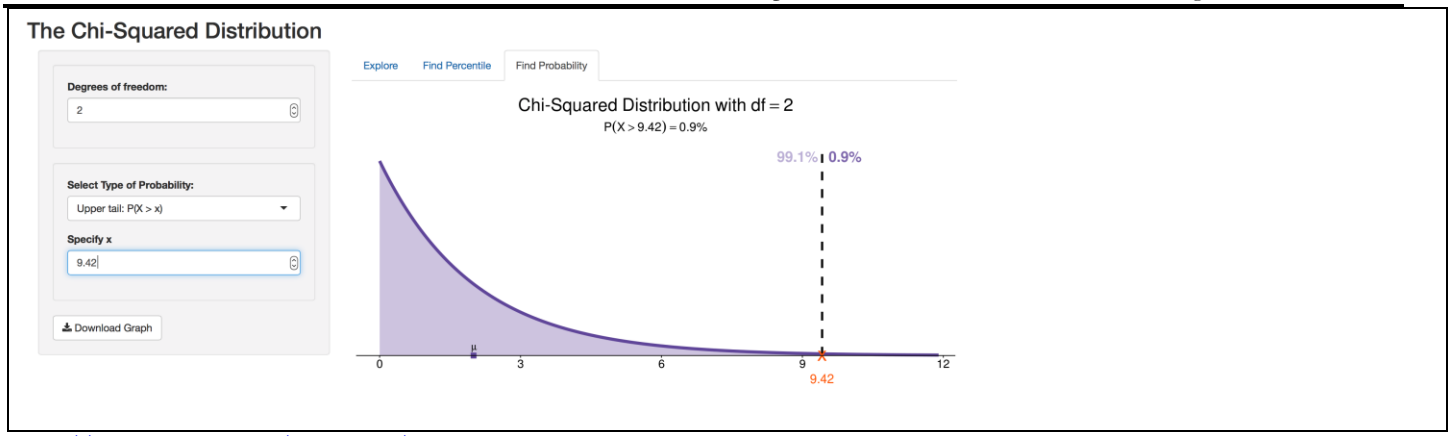
Interpretation:
The exact p-value = .009. Assumption of the null hypothesis model and its application to the data have led to a very unlikely result. The null hypothesis is rejected. We conclude that these data provide statistically significant evidence that urine determination levels are different in 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

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



https://istats.shinyapps.io/ChisqDist/

Adjustment for Ties

There is an adjustment for the presence of tied ranks but this is not discussed here. Unless there are many tied ranks, 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.

This is related to the two-way randomized complete blocks design.
This will be introduced in BIOSTATS 640 Unit 6 (***Analysis of Variance***). Briefly, the randomized complete blocks design is an extension of the paired test. Instead of 2 measurements on each participant (e.g., pre and post), there are now more than 2 measurements on each participant. The term "participant" or "unit" is replaced by the term "block". But the idea is the same. The goal is to compare treatments in an analysis that controls for "participant" effects. The manner in which we control for "participant" effects is to compare the treatments of interest within participants (now called "blocks"). This is analogous to what we do in performing a paired t-test.

Example -
A study is comparing the effects of T=3 levels of dose of treatment on outcome Y in 24 rabbits. 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 placed into N=8 carefully defined blocks (or groups) of 3 rabbits each. The N blocks are defined such that within each, 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.

Outcome, Y

Block	Treatment 1 *dose = 2.29*	Treatment 2 *dose = 3.63*	Treatment 3 *dose = 5.75*
1	17	64	62
2	21	48	72
3	49	34	61
4	54	63	91
5	33	41	56
6	37	64	62
7	40	34	57
8	16	64	72

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

Null and Alternative Hypotheses:
H_O: *Within each block*, there is no effect of treatment (dose)
H_A: At least one treatment (dose) yields different outcomes than the other two treatments

Reason out *what are equally likely* when the null hypothesis is true:
Because this is an extension of pairing (in particular, instead of within-pair homogeneity, we have within-block homogeneity), ***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 = Sum of ranks of outcomes Y for dose=i, taken over all N=8 blocks

Ranks and values of R_i = sum of ranks for each treatment

Block	Treatment 1 *dose = 2.29*	Treatment 2 *dose = 3.63*	Treatment 3 *dose = 5.75*
1	1	3	2
2	1	2	3
3	2	1	3
4	1	2	3
5	1	2	3
6	1	3	2
7	2	1	3
8	1	2	3
	R_1 = 1+1+2+1+1+1+2+1 = 10	R_2 = 3+2+1+2+2+3+1+2 = 16	R_3= 2+3+3+3+3+2+3+3 =22

Notation

T = # treatments = 3

N = # blocks = 8

R_{ij} = Rank of Y 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: $[R_{11}, R_{12}, R_{13}]$ = rearrangement (permutation) of [1, 2, 3] **# ways = 3!**

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

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

Total # arrangements of rankings = (3!)(3!)(3!) = $[3!]^3$

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

In general -

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

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

…

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

Total # arrangements of rankings = (T!)(T!) …. (T!) = $[T!]^N$

Pr [each arrangement | null true] = 1 / { $[T!]^N$ }

For each ("i^{th}") block:

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

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

$$= \bar{R}_{..}$$

For each ("i^{th}") block, reason out *what we expect* when the null hypothesis is true:

$$E[\bar{R}_i \mid \text{null true}] = \bar{R}_{..}$$

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

$$Q = \frac{(12)(N)}{(T)(T+1)} \sum_{j=1}^{T} \left(R_i - \frac{T+1}{2} \right)^2$$

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.

Online Calculator: Friedman Test for Randomized Block Design
http://vassarstats.net
From home page > ***ordinal data > Friedman test > For K=3.*** Enter raw data at right.

Data Entry:

count	Ranks within Rows A	B	C	Raw Data for Sample A	B	C
1	1	3	2	17	64	62
2	1	2	3	21	48	72
3	2	1	3	49	34	61
4	1	2	3	54	63	91
5	1	2	3	33	41	56
6	1	3	2	37	64	62
7	2	1	3	40	34	57
8	1	2	3	16	64	72
	Reset			Calculate		

Mean Ranks for Sample A	B	C
1.3	2	2.8

csq_r = 9
df = 2
P = 0.0111 *

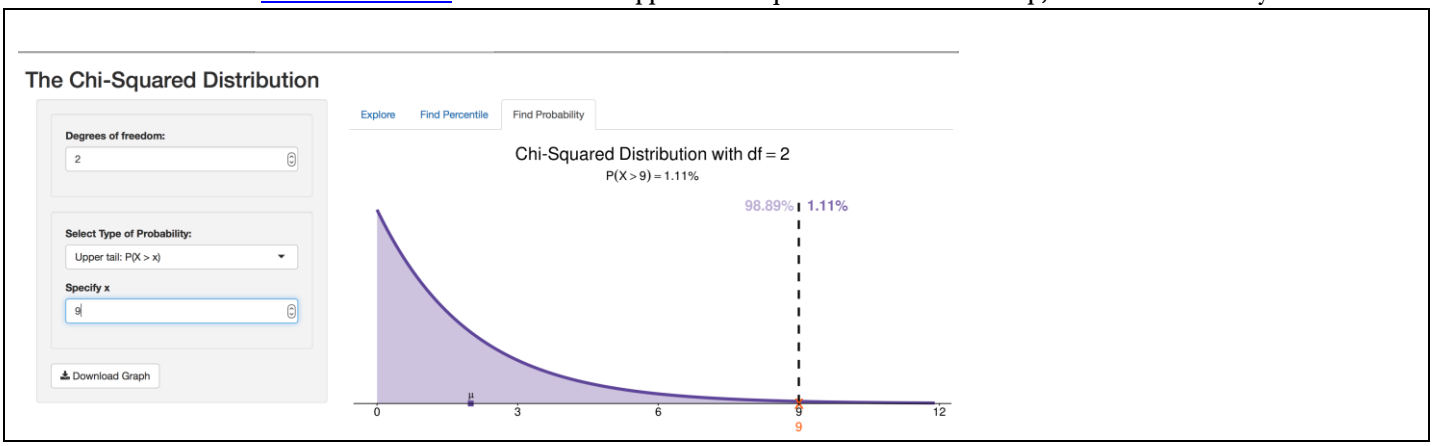
Dear reader – Here, the online calculator is telling me that "sufficiently large" begins at about N=7 blocks. So here, too, I suspect that the online calculator did a chi square approximate p-value calculation.

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. We 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 > **Online Web Apps > Chi Square Distribution. At top, tab: Find Probability**



https://istats.shinyapps.io/ChisqDist/

Example, continued -

Observed Chi Square $_{DF=2}$ = 9
p-value = .01

5.3 R Illustration

Kruskal Wallis Test: Nonparametric alternative to One Way Analysis of Variance

```
# K Independent Groups Kruskal Wallis Test
# Use function kruskal.test( )
# Note - Can use this with K=2 for Wilcoxon Rank Sum/Mann Whitney
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   Assumption of the null HAS to an unlikely result.
                                                                   REJECT the null.
```

Friedman Test: Nonparametric Alternative to Randomized Complete Blocks Analysis of Variance

```
# Friedman Randomized Completely Block Analysis of Variance
library(tidyr)         # we will need to use fucntion gather( ) in package {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          Assumption of the null HAS led to an unlikely result.
                                                              REJECT the null.
```

6. Correlation

6.1 Spearman Rank Correlation

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

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

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:

Individual	1	2	3	4	5	6	7	8
IQ=X	20	17	15	19	23	21	16	12
Personality Score=Y	90	94	100	103	113	114	118	119

Step 1: Rank the values of each variable separately

Individual	1	2	3	4	5	6	7	8
Rank (IQ)=R	6	4	2	5	8	7	3	1
Rank (Personality Score) =S	1	2	3	4	5	6	7	8

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:

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sum_{i=1}^{n}(R_i - \bar{R})^2}$$

- $\bar{R} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} R_i = (N+1)/2$ and $\bar{S} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} S_i = (N+1)/2$
- This works because $\sum_{i=1}^{n}(R_i - \bar{R})^2 = \sum_{i=1}^{n}(S_i - \bar{S})^2$

Example, continued -

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

Individual	1	2	3	4	5	6	7	8	Total
Rank (IQ)=R	6	4	2	5	8	7	3	1	-
Rank (Personality Score) =S	1	2	3	4	5	6	7	8	-
$(R_i - \bar{R})$	1.5	-0.5	-2.5	0.5	3.5	2.5	-1.5	-3.5	-
$(S_i - \bar{S})$	-3.5	-2.5	-1.5	-0.5	0.5	1.5	2.5	3.5	
$(R_i - \bar{R})(S_i - \bar{S})$	-5.25	1.25	3.75	-0.25	1.75	3.75	-3.75	-12.25	-11.0
$(R_i - \bar{R})^2$	2.25	0.25	6.25	0.25	12.25	6.25	2.25	12.25	42.00

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sum_{i=1}^{n}(R_i - \bar{R})^2} = \frac{-11.00}{42.00} = -.2619$$

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 -

Individual	1	2	3	4	5	6	7	8	Total
Rank (IQ)=R	6	4	2	5	8	7	3	1	-
Rank (Personality Score)=S	1	2	3	4	5	6	7	8	-
$(R_i - S_i)^2$	25	4	1	1	9	1	16	49	106

$$r_S = 1 - \frac{6D}{N^3 - N} = 1 - \frac{6(106)}{512 - 8} = -.2619$$ which matches.

Formula #3

Formula to use when there are ties (eew – 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

$$r_{S;\,TIES} = \frac{\left[\frac{N^3 - N}{6}\right] - D - \left[\frac{\sum_{i=1}^{g_X}(t_i^3 - t_i)}{12}\right] - \left[\frac{\sum_{j=1}^{g_Y}(t_j^3 - t_j)}{12}\right]}{2\sqrt{\left[\frac{N^3 - N}{12}\right] - \left[\frac{\sum_{i=1}^{g_X}(t_i^3 - t_i)}{12}\right]}\sqrt{\left[\frac{N^3 - N}{12}\right] - \left[\frac{\sum_{j=1}^{g_Y}(t_j^3 - t_j)}{12}\right]}}$$

Significance Tests for Zero Correlation Using r_s:

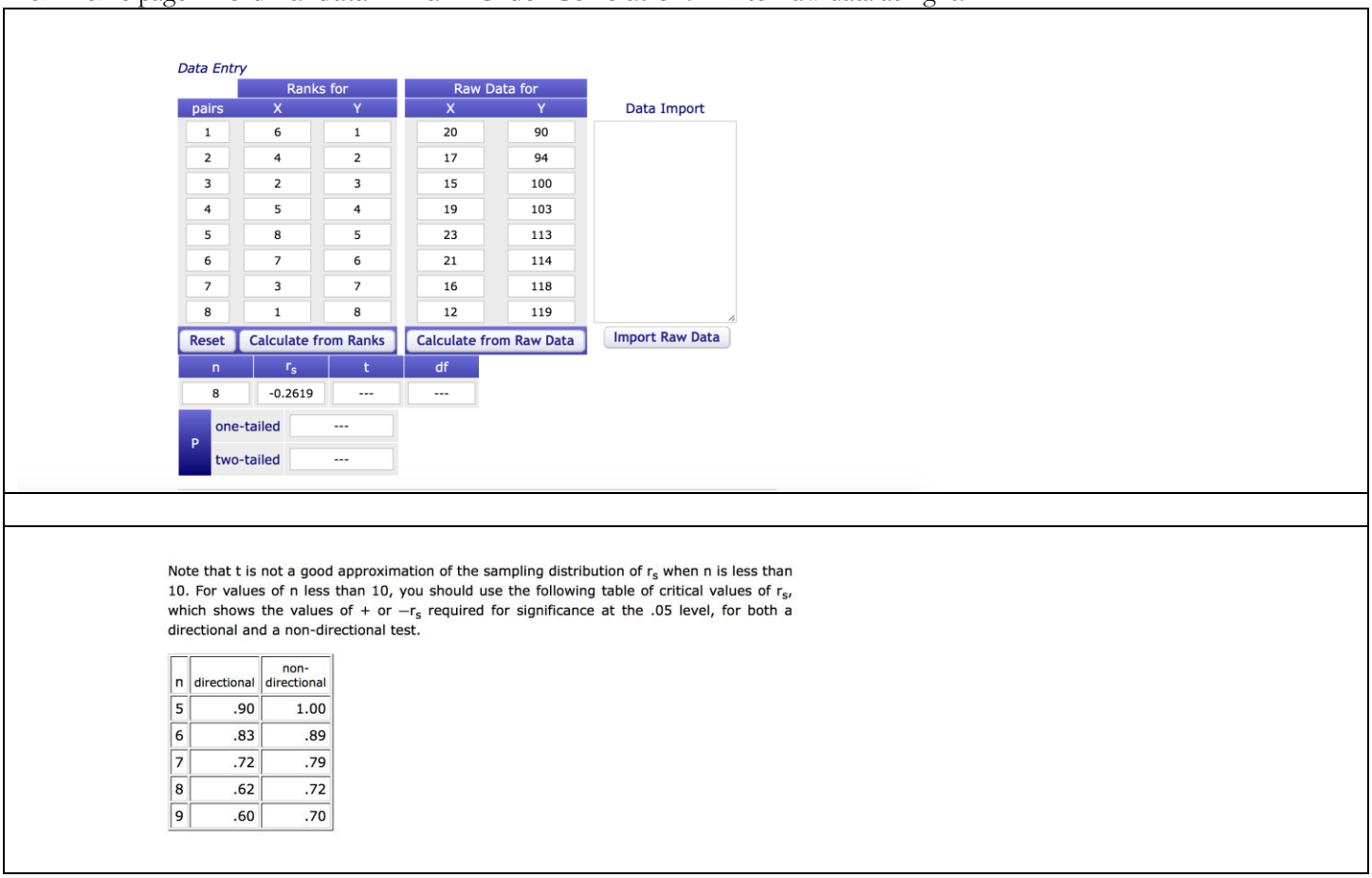
Null and Alternative Hypotheses:

H_O: Spearman's Rank Correlation $r_s = 0$
H_A: Not.

Solution I - the number of pairs is small (say 4-30)

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

Online Calculator: Spearman Correlation
http://vassarstats.net
From home page > ***ordinal data > Rank Order Correlation.*** Enter raw data at right.



Data Entry

pairs	Ranks for X	Ranks for Y	Raw Data for X	Raw Data for Y
1	6	1	20	90
2	4	2	17	94
3	2	3	15	100
4	5	4	19	103
5	8	5	23	113
6	7	6	21	114
7	3	7	16	118
8	1	8	12	119

Data Import

Reset | Calculate from Ranks | Calculate from Raw Data | Import Raw Data

n	r_s	t	df
8	-0.2619	---	---

P: one-tailed ---; two-tailed ---

Note that t is not a good approximation of the sampling distribution of r_s when n is less than 10. For values of n less than 10, you should use the following table of critical values of r_s, which shows the values of + or $-r_s$ required for significance at the .05 level, for both a directional and a non-directional test.

n	directional	non-directional
5	.90	1.00
6	.83	.89
7	.72	.79
8	.62	.72
9	.60	.70

Interpretation:
The Vassar Stats online calculator is telling us that with n=8, if we want to know if our $r_{s\ observed} = -.26$ is significant at the .05 level, *one sided*, then the critical value is $r_{s\ critical} = -.62$ Since $r_{s\ observed} = -.26$ does not exceed $r_{s\ critical} = -.62$ in the negative direction, we conclude that this correlation is NOT statistically significantly different from zero.

Solution II - the number of pairs is > 30

We can use the t-test approximation. See again BIOSTATS 540 Unit 12 (***Simple Linear Regression and Correlation)*** pp 41-45.

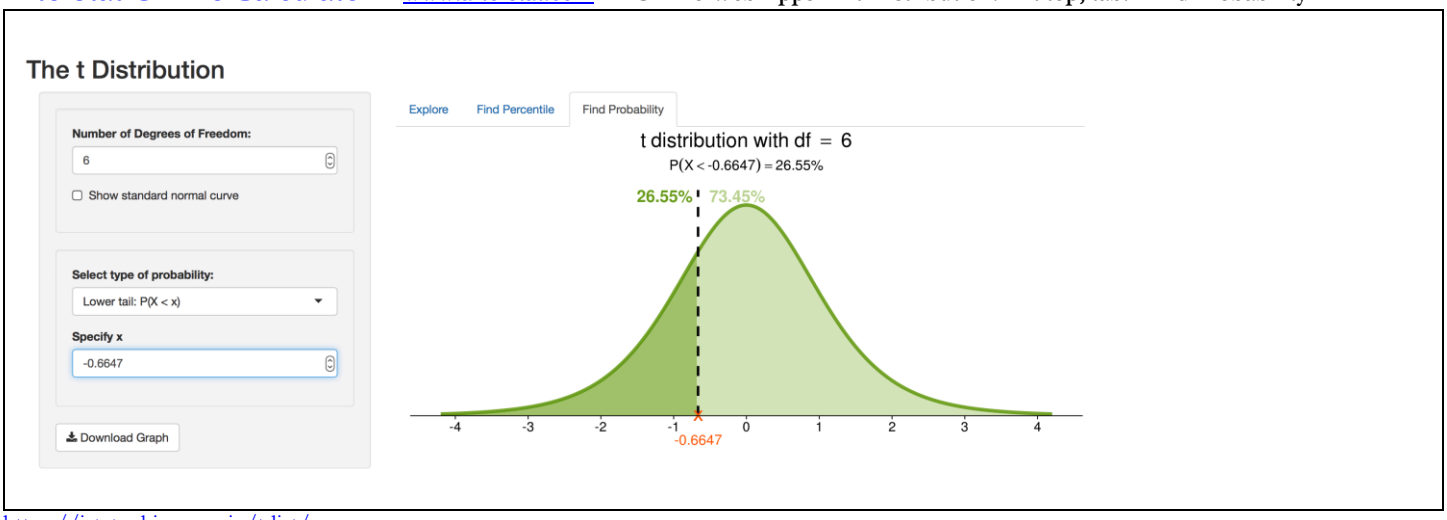
Example, continued

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

$$t_{df=(n-2)} = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} = \frac{(-.2619)\sqrt{6}}{\sqrt{1-[.2619]^2}} = \frac{(-.2619)(2.4495)}{.9651} = -.6647$$

P-value (one sided) = Pr [Student's $t_{DF=6} \leq -0.6647$] = **.27**

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



https://istats.shinyapps.io/tdist/

Interpretation:

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

6.2 R Illustration

```
# Spearman Rank Correlation
# Use function cor.test( ) with option method="spearman"

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")
  Spearman's rank correlation rho

 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
```