Randomization for Inference
(and some other nonparametric procedures)

STAT:5201

Week 4: Lecture 1

See OLRT Section 2.4
Randomization for Inference

- Classical hypothesis tests (e.g. a $t$-test for comparing means, a test for the regression slope, etc.) depend on the **model assumptions** for the tests being met which usually include...
  - Normality of errors
  - Constant variance
  - Independence

- Whether or not these assumptions are reasonably met will be considered AFTER the data are collected.

- A **randomization test** allows us to perform a hypothesis test with potentially fewer and less restrictive assumptions than the classical tests, while still providing a p-value for the test of interest.
We’ve already seen that we can draw a conclusion of statistical significance at the 0.05 level regarding the centers (or locations) of two distributions using Tukey’s quick test through just a visualization.

This test didn’t depend on the data being normally distributed, which is a common assumption in formal hypothesis testing, such as a two-sample $t$-test.
Randomization for Inference

- What if you want to compare two groups, but you do not have normality, nor approximate normality? What options do you have?
- We can try to use a nonparametric test (or some people prefer the term distribution-free tests).

Example (Randomization test in two-sample setting)

Consider the VERY small experiment with 2 treatments in a CRD...

\[ H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2 \]

Observed data below:

\[
\begin{array}{cc}
\text{TRT 1} & \text{TRT 2} \\
3 & 8 \\
7 & 10 \\
\end{array}
\]

\[
t_0 = \frac{5 - 9}{\sqrt{5(2/2)}} = -1.789
\]

Certainly, we can not assume \( t_0 \sim t_2 \) if \( H_0 \) is true.
In order to find a p-value, what we need is a distribution for \( t_0 \) under \( H_0 \) true (i.e. a null distribution) that does not depend on normality. We can use a randomization procedure to create our own null distribution!

We need a test statistic that has these properties:

1. Under the \( H_0 \) true, the test statistic is small (near zero)
2. Under the \( H_1 \) true, the test statistic is large (in magnitude)

Here, the classic \( t \)-test test-statistic has these properties.

\[
t_0 = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

We just need the actual null distribution for \( t_0 \) (we can’t assume it’s \( t_2 \)).

We will use randomization to establish a null distribution for \( t_0 \).
If \( H_0 \) is true, then the treatments are no different from each other.

If the treatments are no different, then any combination of the observed values is just as likely as any other combination.

If \( H_0 \) is true, then the treatment labels of the EUs are exchangeable.

Example (Randomization test in two-sample setting)

Under \( H_0 \) true, any combination of the 4 observed values is equally likely...

\[
\begin{array}{c|c}
\text{TRT1} & \text{TRT2} \\
3 & 7 \\
8 & 10 \\
\end{array}
\quad
\begin{array}{c|c}
\text{TRT1} & \text{TRT2} \\
7 & 3 \\
8 & 10 \\
\end{array}
\quad
\ldots \quad \text{etc.}
\]

\( t^* = -1.029 \quad t^* = 0.283 \)
Example (Randomization test in two-sample setting)

In how many different combinations could these EUs have been assigned to treatment groups? \( \binom{4}{2} = 4C_2 = 6 \)

All 6 combinations are equally likely to have been observed under \( H_0 \) true. Randomization shows us the ‘hypothetical’ data sets that could have occurred.

<table>
<thead>
<tr>
<th>Randomization</th>
<th>TRT 1</th>
<th>TRT 2</th>
<th>( t^* )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3,7</td>
<td>8,10</td>
<td>-1.789</td>
<td>1/6</td>
</tr>
<tr>
<td>2</td>
<td>3,8</td>
<td>7,10</td>
<td>-1.029</td>
<td>1/6</td>
</tr>
<tr>
<td>3</td>
<td>3,10</td>
<td>8,7</td>
<td>-0.283</td>
<td>1/6</td>
</tr>
<tr>
<td>4</td>
<td>8,7</td>
<td>3,10</td>
<td>0.283</td>
<td>1/6</td>
</tr>
<tr>
<td>5</td>
<td>7,10</td>
<td>3,8</td>
<td>1.029</td>
<td>1/6</td>
</tr>
<tr>
<td>6</td>
<td>8,10</td>
<td>3,7</td>
<td>1.789</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Our observed test statistic was \( t_0 = -1.789 \) (the most extreme value above).
Example (Randomization test in two-sample setting)

Under $H_0$ true, what is the probability of observing a test statistic this extreme or more extreme?

$$P(\lvert t^* \rvert \geq |-1.789|) = \frac{2}{6} = 0.33$$

This is the exact randomization test p-value in the two-sided test.

If you were instead interested in testing the one-sided test,

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 < \mu_2$$

then the p-value would be $\frac{1}{6} = 0.167$

Randomization for Inference

- The idea is that treatments are randomly assigned to EUs, and under $H_0$ true, the variability observed in the data is NOT due to treatment, but simply due to EU-to-EU variability.

- The key assumption in this nonparametric testing process is exchangeability of group membership under the null being true.

- The p-value is limited because of the discrete nature of the null distribution. And in small data sets, because there are only a small number of possible combinations of data, such as 6 below.

<table>
<thead>
<tr>
<th>Randomization</th>
<th>TRT 1</th>
<th>TRT 2</th>
<th>$t^*$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3,7</td>
<td>8,10</td>
<td>1.789</td>
<td>1/6</td>
</tr>
<tr>
<td>2</td>
<td>3,8</td>
<td>7,10</td>
<td>1.029</td>
<td>1/6</td>
</tr>
<tr>
<td>3</td>
<td>3,10</td>
<td>8,7</td>
<td>0.283</td>
<td>1/6</td>
</tr>
<tr>
<td>4</td>
<td>8,7</td>
<td>3,10</td>
<td>0.283</td>
<td>1/6</td>
</tr>
<tr>
<td>5</td>
<td>7,10</td>
<td>3,8</td>
<td>1.029</td>
<td>1/6</td>
</tr>
<tr>
<td>6</td>
<td>8,10</td>
<td>3,7</td>
<td>1.789</td>
<td>1/6</td>
</tr>
</tbody>
</table>
Randomization for Inference

- Because of the discreteness of the randomization distribution, it’s a good idea to calculate the randomization p-value as “≥” rather than just “>”.

This allows us to be conservative (in continuous distributions this doesn’t matter). If using just “>” and the observed value is the most extreme, then p-value=0 (not realistic).

Use “≥” rather than “>” when calculating a randomization p-value.

<table>
<thead>
<tr>
<th>Randomization</th>
<th>TRT 1</th>
<th>TRT 2</th>
<th>$t^*$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.7</td>
<td>8.10</td>
<td>-1.789</td>
<td>1/6</td>
</tr>
<tr>
<td>2</td>
<td>3.8</td>
<td>7.10</td>
<td>-1.029</td>
<td>1/6</td>
</tr>
<tr>
<td>3</td>
<td>3.10</td>
<td>8.17</td>
<td>0.283</td>
<td>1/6</td>
</tr>
<tr>
<td>4</td>
<td>8.17</td>
<td>3.10</td>
<td>0.283</td>
<td>1/6</td>
</tr>
<tr>
<td>5</td>
<td>7.10</td>
<td>3.8</td>
<td>1.029</td>
<td>1/6</td>
</tr>
<tr>
<td>6</td>
<td>8.10</td>
<td>3.7</td>
<td>1.789</td>
<td>1/6</td>
</tr>
</tbody>
</table>
You can use this procedure for any independent two-sample setting, but the number of combinations, or possible randomizations gets large quickly:

\[
\binom{10}{5} = 252 \quad \binom{20}{10} = 184,756
\]

One option is to only generate a subset of all possible randomizations and compute an approximate randomization test p-value by comparing the observed test statistic to the randomization distribution that you generated.

This is flexible, such that you can choose any appropriate test statistic and utilize the same randomization procedure.
Example (Randomization test in larger two-sample setting)

Here, we have a two treatments in a CRD with $n = 10$ in each group. There are 184,756 possible groupings, all equally likely under $H_0$ true, but we will use only a subset of those.

We’ll compute an approximate randomization test p-value using 1000 randomizations.

Original observed data

Group A mean = 10.23
Group B mean = 11.27
$t_0 = -2.23$
Example (Randomization test in larger two-sample setting)

One generated randomization:

Randomization data
Group A mean = 10.89
Group B mean = 10.61
\( t^* = 0.54 \)

This process is repeated 1000 times, capturing the \( t^* \) each time. The distribution of \( t^* \) values forms the null distribution for the hypothesis test.
Example (Randomization test in larger two-sample setting)

After generating 1000 randomizations, we can compare our observed test statistic to the randomization distribution, which is the null distribution to calculate our approximate randomization p-value.

There were 48 randomizations with a test statistic as extreme or more extreme in magnitude than the observed test statistic.

\[ P(|t^*| \geq 2.23) = \frac{48}{1000} = 0.048 \]

The is our approximate randomization p-value.
Example (Randomization test in larger two-sample setting)

In R:
# Original data has ‘y’ outcome and factor ‘trt’.
> orig<-t.test(y~trt)

> hold<-rep(NA,1000)
> for (i in 1:1000){
    # randomize the labels
    trt2<-sample(trt,size=20,replace=FALSE)
    # capture the null test statistic
    t.out<-t.test(y~trt2)
    hold[i]<-t.out$statistic
    
}

> mean(abs(hold)>=abs(orig$statistic))
[1] 0.048
Wilcoxon Rank-Sum Test (or Mann-Whitney Test)

- Another well-known nonparametric alternative to the two-sample t-test is the **Wilcoxon Rank-Sum Test**. If a client mentions the **Mann-Whitney U-Test**, this is essentially the same test (just a different test statistic).

- Under $H_0$ true, $N = n_A + n_B$ observations are being drawn from only one distribution (not two). Based on this idea, a test statistic was developed.

- This nonparametric test is based on the order in which the observations from the two samples fall.
We compare two types of flooring in a hospital setting: carpeted and bare. We have 8 rooms of each kind of flooring. The quantity of bacteria present on the given floor was recorded in colonies/ft$^3$.

<table>
<thead>
<tr>
<th>condition</th>
<th>response</th>
<th>Colonies/ft$^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bare</td>
<td>1</td>
<td>5.4</td>
</tr>
<tr>
<td>bare</td>
<td>2</td>
<td>12.1</td>
</tr>
<tr>
<td>bare</td>
<td>3</td>
<td>8.3</td>
</tr>
<tr>
<td>bare</td>
<td>4</td>
<td>3.8</td>
</tr>
<tr>
<td>bare</td>
<td>5</td>
<td>7.2</td>
</tr>
<tr>
<td>bare</td>
<td>6</td>
<td>12.0</td>
</tr>
<tr>
<td>bare</td>
<td>7</td>
<td>11.2</td>
</tr>
<tr>
<td>bare</td>
<td>8</td>
<td>10.1</td>
</tr>
<tr>
<td>carpeted</td>
<td>9</td>
<td>7.1</td>
</tr>
<tr>
<td>carpeted</td>
<td>10</td>
<td>12.9</td>
</tr>
<tr>
<td>carpeted</td>
<td>11</td>
<td>9.2</td>
</tr>
<tr>
<td>carpeted</td>
<td>12</td>
<td>14.1</td>
</tr>
<tr>
<td>carpeted</td>
<td>13</td>
<td>11.9</td>
</tr>
<tr>
<td>carpeted</td>
<td>14</td>
<td>11.3</td>
</tr>
<tr>
<td>carpeted</td>
<td>15</td>
<td>15.7</td>
</tr>
<tr>
<td>carpeted</td>
<td>16</td>
<td>15.1</td>
</tr>
</tbody>
</table>
Wilcoxon Rank-Sum Test (Exact test)

Example (Wilcoxon rank-sum test)

We rank all \( N = 16 \) observations from lowest to highest:

<table>
<thead>
<tr>
<th>Rank</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>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.8</td>
<td>5.4</td>
<td>7.1</td>
<td>7.2</td>
<td>8.3</td>
<td>9.2</td>
<td>10.1</td>
<td>11.2</td>
<td>11.3</td>
<td>11.9</td>
<td>12.0</td>
<td>12.1</td>
<td>12.9</td>
<td>14.1</td>
<td>15.1</td>
<td>15.7</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Under \( H_0 \) true, any ordering of the treatments is equally likely, and we wouldn’t expect to see a long string of one treatment at one end.

One group getting lots of small ranks or lots of big ranks is evidence against the null of equal distributions.
Wilcoxon Rank-Sum Test (Exact test)

**Definition (Wilcoxon rank-sum test statistic)**

Test statistic $W^*$:

$W^*$ = Sum of the ranks of one group
(use the smaller group if $n_A \neq n_B$, to match common tables).

Under $H_0$ true, $W^*$ has a known distribution. Critical values can be found in tables for small samples.

A very small sum or a very large sum suggests $H_0$ should be rejected.

* Or some transformation of the sum like $W^{**} = W^*_A - \frac{n_A(n_A+1)}{2}$. 
Example (Wilcoxon rank-sum test)

In R exact test:
> wilcox.test(response~condition, exact=TRUE)

Wilcoxon rank sum test

data:  response by condition
W = 14, p-value = 0.06496
alternative hypothesis: true location shift is not equal to 0

NOTE: This was an observational study. What may be some lurking variables causing the differences in bacteria colonies besides type of flooring? (What could be another “segregating factor” between these 2 groups of rooms? Potential confounding factor?)
Wilcoxon Rank-Sum Test (Exact test)

Example (Wilcoxon rank-sum test)

In SAS exact test:

```sas
proc npar1way wilcoxon data=dec;
    class condition;
    var response;
    exact wilcoxon;
run;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable response
  Classified by Variable condition

<table>
<thead>
<tr>
<th>condition</th>
<th>N</th>
<th>Sum of Scores</th>
<th>Expected Under H0</th>
<th>Std Dev Under H0</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>bare</td>
<td>8</td>
<td>50.0</td>
<td>68.0</td>
<td>9.521905</td>
<td>6.250</td>
</tr>
<tr>
<td>carpeted</td>
<td>8</td>
<td>86.0</td>
<td>68.0</td>
<td>9.521905</td>
<td>10.750</td>
</tr>
</tbody>
</table>

Exact Test
One-Sided Pr <= S 0.0325
Two-Sided Pr >= |S - Mean| 0.0650
Example (Wilcoxon rank-sum test)

In SAS exact test:

The NPAR1WAY Procedure

<table>
<thead>
<tr>
<th>condition</th>
<th>N</th>
<th>Scores</th>
<th>Expected</th>
<th>Std Dev</th>
<th>Mean</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>bare</td>
<td>8</td>
<td>50.0</td>
<td>68.0</td>
<td>9.521905</td>
<td>6.250</td>
<td></td>
</tr>
<tr>
<td>carpeted</td>
<td>8</td>
<td>86.0</td>
<td>68.0</td>
<td>9.521905</td>
<td>10.750</td>
<td></td>
</tr>
</tbody>
</table>

Exact Test
- One-Sided Pr <= S = 0.0325
- Two-Sided Pr >= |S - Mean| = 0.0650

NOTE: The total sum of the ranks = \( \frac{N(N+1)}{2} = 136 \).
Under \( H_0 \) true, each group sum is expected to be 68.
Wilcoxon Rank-Sum Test with Ties (Exact test)

- How to handle ties.

- If you use the base R function of `wilcox.test()` to perform the exact test and you have ties, you will get a warning message:
  
  ```
  Warning message:
  In wilcox.test.default(x = c(5.4, 12.1, 8.3, :
      cannot compute exact p-value with ties
  ```

- But you can use the `coin` package to compute an exact p-value with ties using the `wilcox_test()` function from that package.

  ```r
  > wilcox_test(y~trt, distribution="exact")
  ```

  **Exact Wilcoxon-Mann-Whitney Test**

  data:  y by condition (A, B)
  Z = -1.8918, p-value = 0.06263
  alternative hypothesis: true mu is not equal to 0
When you have ties in ranked analysis, in general, the tied observations are given the average rank.

Here is an example from Penn State with $n_A = 5$, $n_B = 5$:

Using the usual $W^*$ tables is OK if the number of ties is small, but otherwise, we may need a different table (or null distribution).
Wilcoxon Rank-Sum Test (Approximate test)

- When both groups have 10 or greater observations, we can use a normal approximation for the test statistic.

- $W_A$ is the sum of the ranks for treatment A. Under $H_0$, we have $W_A \sim N(\mu_A, \sigma^2_A)$ with...

  \[
  \mu_A = \frac{n_A(n_A+n_B+1)}{2} \quad \text{and} \quad \sigma_A = \sqrt{\frac{n_A n_B (n_A+n_B+1)}{12}}
  \]

**Example (Approximate Wilcoxon Rank-sum test)**

Our earlier example is a bit small for this approximation with $n = 8$, but I’ll show it as an example here... Get the observed $z$ test statistic:

\[
Z = \frac{W_A - \mu_A}{\sigma_A} = \frac{50 - 68}{\sqrt{8 \cdot 8 \cdot (8+8+1)/12}} = -1.89
\]

And then compare to standard normal distribution for a p-value.
Example (Approximate Wilcoxon Rank-sum test)

# from base function wilcoxon.test()
> wilcox_test(response~condition,distribution="approximate")

Approximative Wilcoxon-Mann-Whitney Test

data:  response by condition (bare, carpeted)
Z = -1.8904, p-value = 0.0672
alternative hypothesis: true mu is not equal to 0

# from coin package
> wilcox.test(response~condition,exact=FALSE)

Wilcoxon rank sum test with continuity correction

data:  response by condition
W = 14, p-value = 0.06608
alternative hypothesis: true location shift is not equal to 0
The Wilcoxon test does not require normality. It does require that the distributions from both populations are equal.

Replacing observed values by their ranks is a common trick in nonparametric tests.

If the original data had strong outliers, this will be suppressed by replacing them with their ranks, making the test much less sensitive to outliers.

If the assumptions of a traditional $t$-test are met, use the $t$-test as it will have more power than a nonparametric test.
Example (Red-Green paired samples on one chip)

Gene expression technology uses dyes (red and green) to label mRNA samples. The sample is “excited” and the level of fluorescence from a sample is quantified as the level of gene expression.

If one sample is split into 2 parts, each part has the same expression. But is there a dye effect? Does one dye naturally fluoresce higher than the other?
A Randomization Test for Paired Data

Example (Red-Green paired samples on one chip)

From each sample, we have a red and green measurement. There may be sample-to-sample variability (or chip-to-chip variability), but it is not of primary interest. The sample serves as a block. We are interested in knowing if there is a difference between the red and green. This is paired data.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Fluorescence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Red</td>
<td>Green</td>
<td>diff (Red-Green)</td>
</tr>
<tr>
<td>1</td>
<td>2.4</td>
<td>1.8</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>1.2</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>-0.4</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>4</td>
<td>3.0</td>
<td>2.5</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>-0.25</td>
<td>0.3</td>
</tr>
<tr>
<td>6</td>
<td>1.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

- $H_0 : \mu_{\text{red}} = \mu_{\text{green}}$ vs. $H_1 : \mu_{\text{red}} \neq \mu_{\text{green}}$
- but because paired... $H_0 : \mu_{\text{diff}} = 0$ vs. $H_1 : \mu_{\text{diff}} \neq 0$
A Randomization Test for Paired Data

- One option we have already seen would be to align the data for sample (or chip) and do Visual comparison test.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Aligned for Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Red</td>
</tr>
<tr>
<td>1</td>
<td>1.40</td>
</tr>
<tr>
<td>2</td>
<td>1.15</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>1.35</td>
</tr>
<tr>
<td>5</td>
<td>1.25</td>
</tr>
<tr>
<td>6</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>Green</td>
</tr>
<tr>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
</tr>
</tbody>
</table>

- There are at least 3 stragglers on both ends. They are statistically significantly different at the 0.05 level.

- But another option would be to perform a randomization test that takes the blocked structure into consideration.
A Randomization Test for Paired Data

Example (Red-Green paired samples on one chip)

Under $H_0$ true, the data still has the sample-to-sample variability, but the red-green differences are randomly interchangeable (no treatment effect).

The sign of the difference that we observe is equally likely to be “+” and it is to be “−”.

For example, we observed red=2.4 & green=1.8 (diff= +0.6), but red=1.8 & green=2.4 (diff= −0.6) was just as likely to have been seen if $H_0$ true. This is just an exchange of labels.

We want to choose a test statistic that will be relatively large in magnitude when $H_0$ is not true. Here, we’ll choose the ‘sum of the differences’ and label it as $S$.

What is the null distribution for $S$?
A Randomization Test for Paired Data

Example (Red-Green paired samples on one chip)

How many possible randomizations are there that coincides with switching labels within a block?

How many equally likely sets of differences? (we will keep the magnitude of the differences here)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Fluorescence Red</th>
<th>Fluorescence Green</th>
<th>diff (Red-Green)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.4</td>
<td>1.8</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>1.2</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>-0.4</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>4</td>
<td>3.0</td>
<td>2.5</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>-0.25</td>
<td>0.3</td>
</tr>
<tr>
<td>6</td>
<td>1.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>
A Randomization Test for Paired Data

Example (Red-Green paired samples on one chip)

We do not have to compute the sum of the differences for all $2^6 = 64$ randomizations. We just need to find the randomization p-value...

How many randomizations would give a sum that was as extreme or more extreme as the $S$ observed? $S_o = 2.0$ (only 1 negative difference)

Randomization       | Six computed differences
---------------------|---------------------------------------------------
1                    | 0.6 0.1 0.2 0.5 0.3 0.7 "all positive"
2                    | 0.6 -0.1 0.2 0.5 0.3 0.7
3                    | 0.6 0.1 -0.2 0.5 0.3 0.7 "observed"
4                    | -0.6 -0.1 -0.2 -0.5 -0.3 -0.7 "all negative"
5                    | -0.6 -0.1 -0.2 -0.5 -0.3 -0.7
6                    | -0.6 -0.1 -0.2 -0.5 -0.3 -0.7

There are 3 configurations giving an $S$ as extreme or more extreme on the ‘positive’ side, and 3 on the ‘negative’ side.

The 2-sided p-value = $\frac{6}{64} = 0.0938$
The sign test

- The sign test works similarly to the randomization test in the previous example, but we now disregard the magnitude of the differences, and just make our inference based on the count of “+” or “−” differences.

- If all differences are “+” or all differences are “−”, then this gives some evidence for reject $H_0$.

Example (Sign test for paired data)

Continuing example... We found that 1 of the 6 differences was negative.

Let $T =$ # of negative differences.
Under $H_0$ true, $T \sim \text{binomial}(6,0.5)$.

$T_o = 1.$

$$P(T \leq 1) = P(T = 0) + P(T = 1) = 0.1094$$ is the 1-sided p-value (not very powerful)
The Wilcoxon Signed-Rank Test (for paired data)

- Clearly, this will somehow utilize the ranks, and it is the rank of the absolute value of the paired differences that are used in the analysis.

Example (Wilcoxon Signed-Rank Test for paired data)

Continuing example with 6 paired differences...

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff</td>
<td>0.1</td>
<td>-0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Let $T$ be the sum of the ranks of the “+” differences\(^1\).

Here, $T$ will be a positive value as the ranks are positive.

Under $H_0$ false, $T$ will either be close to 0 (no differences were “+”) or very large (all differences were “+”). Under $H_0$ true, you’ll be in the middle of the range of the possible values for $T$.

---

\(^1\) Some use $T=$ sum of all signed ranks, which would have different critical values.
The Wilcoxon Signed-Rank Test (for paired data)

- As with the Wilcoxon Test for the two-sample setting, there is a distribution for $T$ under $H_0$ true that can be found in table. This can provide an exact test p-value.

- There is also a normal approximation for large sample sizes with

  \[ T \sim N\left( \frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24} \right) \]

Example (Approximate Wilcoxon Signed-Rank Test)

Continuing example with 6 paired differences...

Our observed $T_o = 19$.
Under $H_0$, we have the approximation $T \sim N(10.5, 22.75)$.

\[
P(T \geq 19) = P\left(Z \geq \frac{19-10.5}{\sqrt{22.75}}\right) = p(Z \geq 1.78) = 0.0375 \text{ as 1-sided p-value.}
\]
The Wilcoxon Signed-Rank Test (for paired data)

Example (Approximate Wilcoxon Signed-Rank Test)

Continuing example with 6 paired differences...

In R:
> red<-c(2.4,1.3,-0.4,3.0,0.05,1.2)
> green<-c(1.8,1.2,-0.2,2.5,-0.25,0.5)
> wilcox.test(red,green,paired=TRUE,
            alternative="greater",exact=FALSE,correct=FALSE)

Wilcoxon signed rank test
data:  red and green
V = 19, p-value = 0.03737
alternative hypothesis: true location shift is greater than 0

NOTE: The standard procedure for Wilcoxon Signed-Rank Test is to throw-out zeros, which can be a concern.
In general, if your assumptions are met, parametric tests such as a $t$-test or $F$-test are more powerful than the nonparametric tests.

The main assumption for these nonparametric tests is that the labels are exchangeable under the null being true, or “exchangeability” under $H_0$.

So, they can be applied under non-normality but they don’t fix the non-constant variance situation, but maybe a transformation can help that issue.

A nonparametric test for a CRD 1-way ANOVA (i.e. single factor, more than 2 levels) is called the Kruskal-Wallis test (an extension of the Wilcoxon rank-sum test). This test again analyzes the ranks instead of the raw data in the 1-way ANOVA setting.