Previous Lecture

Categorical (nominal data) analysis

- Contingency Tables
- Chi-squared Test
- Fisher’s Exact Test
Today’s Lecture

Parametric vs. Nonparametric testing

Paired nonparametric tests
  Sign Test
  Wilcoxon Signed-Rank Test

Comparing two samples
  Wilcoxon Rank-Sum Test
  Permutation test

Comparing many samples
Table of Contents

Parametric vs. Nonparametric testing

Paired nonparametric tests
  Sign Test
  Wilcoxon Signed-Rank Test

Comparing two samples
  Wilcoxon Rank-Sum Test
  Permutation test

Comparing many samples
## Data types

<table>
<thead>
<tr>
<th>Data type</th>
<th>What is it?</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinal</td>
<td>Meaningful to compute distance between measurements</td>
<td>Body weight, Temperature, Blood pressure</td>
</tr>
<tr>
<td>Nominal</td>
<td>Categorical data; measurements can be sorted into categories with no ordering</td>
<td>Drug type, Ethnicity</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Measurements can be ordered, but are not cardinal</td>
<td>Visual acuity, Pain scale</td>
</tr>
</tbody>
</table>

Cardinal data can be *ratio scale*, meaning a fixed zero point is defined, or *interval scale* where the zero point is arbitrary. Only with ratio scale data does comparing ratios (*this is twice as large as that*) make sense.
Hypothesis Testing

Use data from samples to ask questions regarding the underlying population.

Compute sample statistics to estimate population parameters (means, proportions, etc.). This means we often make assumptions about the populations parametric form.
Parametric vs nonparametric tests

Parametric Methods

▶ Parameters which determine the form of the population distribution are assumed known.
  ▶ Usually this means assuming data is normal (or approximately so under the Central Limit Theorem) and the mean, standard deviation, etc. are known.

Nonparametric Methods

▶ Few or no assumptions about the shape and/or parameters of the population distribution are made.
▶ Use when data is not normal (e.g., heavily skewed), data is not cardinal, or when the sample size is too small to assume normality under CLT.
Nonparametric equivalents for familiar parametric methods

<table>
<thead>
<tr>
<th>Goal</th>
<th>Parametric method</th>
<th>Nonparametric method</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compare two measurements from same individual</td>
<td>Paired t-test</td>
<td>Sign test or Wilcoxon signed-rank test</td>
<td>Change in blood pressure readings before and after treatment in treatment group?</td>
</tr>
<tr>
<td>Compare means between two groups</td>
<td>Two sample t-test</td>
<td>Wilcoxon rank-sum test or permutation test</td>
<td>Difference in blood pressure readings for patients in placebo and treatment groups?</td>
</tr>
<tr>
<td>Compare means between three or more groups</td>
<td>ANOVA</td>
<td>Kruskal-Wallis test</td>
<td>Difference in blood pressure readings for patients in either control, treatment 1, or treatment 2 groups?</td>
</tr>
</tbody>
</table>
Why don’t we always use nonparametric methods?

Nonparametric methods are less powerful than parametric methods when the data is approximately normal.

- Given fixed sample size, $p$-values from nonparametric methods will be greater than those from parametric tests.

Many tests use ranks of data instead of actually data and intuition regarding the results may suffer.

- Knowing that mean ranks differ by $X$ is less informative than knowing means differ by $X$. 
Table of Contents

Parametric vs. Nonparametric testing

**Paired nonparametric tests**
- Sign Test
- Wilcoxon Signed-Rank Test

Comparing two samples
- Wilcoxon Rank-Sum Test
- Permutation test

Comparing many samples
Setup

Say we ask beach goers to apply one type of sunblock to their left arm and another type to their right arm so that we can gauge the effectiveness of two new sunblock formulas.

- $x_i =$ amount of sunburn on left arm for $i$th person
- $y_i =$ amount of sunburn on right arm for $i$th person
- $d_i = x_i - y_i ;$ difference in sunburn for the $i$th person

We want to test whether these two sunblocks provide different levels of protection from sunburn, i.e. test the hypothesis:

\[ H_0 : \Delta = 0 \]
\[ H_1 : \Delta \neq 0 \]

where $\Delta =$ population median of the $d_i$s.
In the most restrictive case, we can only determine

- $d_i = 0$, both arms are equally sunburned
- $d_i > 0$, left arm is more sunburned than the right arm
- $d_i < 0$, right arm is more sunburned than the left arm

for each individual. Can’t actually observe magnitude of $d_i$.

From a sample of $N = 45$ people, we find

- 5 people are equally sunburned on both arms
- $C = 18$ people are more sunburned on the left arm
- $D = 22$ people are more sunburned on the right arm

In this case, we must use the **Sign Test**.
The Sign Test – Assuming normality

We have $C = \text{number of } d_i > 0 = 18$ and $n = \text{number of } d_i \neq 0 = C + D = 40$.

Under $H_0$, we expect that $\Pr(n) = 1/2$. This reduces the sign test to a special case of the one-sample binomial test ($H_0 : p = 1/2$ vs $H_1 : p \neq 1/2$). The normal approximation to the binomial is valid if:

$$n \left( \frac{1}{2} \right) \left( \frac{1}{2} \right) \geq 5 \implies n \geq 20$$

If this holds, we accept the null hypothesis if

$$\frac{n}{2} + \frac{1}{2} + z_{1-\alpha/2} \sqrt{n/4} < C < \frac{n}{2} - \frac{1}{2} - z_{1-\alpha/2} \sqrt{n/4}$$
We have $C =$ number of $d_i > 0 = 18$, $D =$ number of $d_i < 0 = 22$ and $n =$ number of $d_i \neq 0 = C + D = 40$.

$$p = \begin{cases} 1.0 & \text{if } C = D \\ 2 \times \left[ 1 - \Phi \left( \frac{|C-D|-1}{\sqrt{n}} \right) \right] & \text{if } C \neq D \end{cases}$$

**Figure 9.2 Computation of the p-value for the sign test**

- If $C < n/2$, then $p = 2 \times \text{area to the left of } \left( C - \frac{n}{2} + \frac{1}{2} \right) \sqrt{\frac{n}{4}} \text{ under an } N(0, 1) \text{ distribution}$.
- If $C > n/2$, then $p = 2 \times \text{area to the right of } \left( C - \frac{n}{2} - \frac{1}{2} \right) \sqrt{\frac{n}{4}} \text{ under an } N(0, 1) \text{ distribution}$.
If $n < 20$, then the normal approximation to the binomial distribution doesn’t hold, and the exact binomial probabilities are needed to compute a $p$-value.

$$p = \begin{cases} 
2 \times \sum_{k=C}^{n} \binom{n}{k} \left( \frac{1}{2} \right)^n & \text{if } C > n/2 \\
2 \times \sum_{k=0}^{C} \binom{n}{k} \left( \frac{1}{2} \right)^n & \text{if } C < n/2 \\
1.0 & \text{if } C = n/2
\end{cases}$$

This is a special case of the small-sample, one-sample binomial test.
Using the normal approximation, we have

\[ C = 18 \]
\[ D = 22 \]
\[ n = 40 \]
\[ p = 2 \times \left[ 1 - \Phi \left( \frac{|18 - 20| - 1}{\sqrt{40}} \right) \right] = 2 \times \Phi(-0.474) = 0.635 \]

We except \( H_0 \), meaning both sunblocks are equally effective at preventing sunburn. Alternatively, we can compute the \( p \)-value using the exact binomial methods. Let’s compute this in R:

```r
n <- 40
k <- 0:18
2 * sum(choose(n, k) * (0.5^n))
0.625828```
Comparing differences in paired observations – with magnitudes

We go back to the beach with a spectrophotometer which allows us to quantitatively assign a sunburn to a 1-10 scale, and collect the data to the right, consistent with the previous data set.

We wish to test the same hypothesis $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ with $\Delta = \text{population median of } d_i$.

Because our data now has magnitudes in addition to signs, we use the **Wilcoxon Signed-Rank Test** to test the hypothesis.

<table>
<thead>
<tr>
<th>$d_i$</th>
<th>$f_i$</th>
<th>$d_i$</th>
<th>$f_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

$22 \quad 18$
### Assign ranks to the data

| $|d_i|$ | $d_i$ | $f_i$ | $d_i$ | $f_i$ | Total | Range of ranks | Average Ranks |
|------|------|------|------|------|-------|----------------|----------------|
| 1    | -1   | 4    | 1    | 10   | 14    | 1–14          | 7.5            |
| 2    | -2   | 4    | 2    | 6    | 10    | 15–24         | 19.5           |
| 3    | -3   | 5    | 3    | 2    | 7     | 25–31         | 28.0           |
| 4    | -4   | 1    | 4    | 0    | 1     | 32            | 32.0           |
| 5    | -5   | 2    | 5    | 0    | 2     | 33–34         | 33.5           |
| 6    | -6   | 2    | 6    | 0    | 2     | 35–36         | 35.5           |
| 7    | -7   | 3    | 7    | 0    | 3     | 37–39         | 38.0           |
| 8    | -8   | 1    | 8    | 0    | 1     | 40            | 40.0           |
| 9    | -9   | 0    | 9    | 0    | 0     | –             | –              |
| 10   | -10  | 0    | 10   | 0    | –     | –             | –              |
|      |      | 22   | 18   |      |       |                |                |
Computing ranked sums

\[ R = \sum \langle \text{rank} \rangle \langle \text{observations with this rank} \rangle \]

By \( H_0 \), we expect the rank sum for positive \( d_i \) to be \( R_1 = n(n + 1)/4 \) with variance \( Var(R_1) = n(n + 1)(2n + 1)/24 \) (when no ties are present). A larger rank sum indicates better performance for the right arm sunblock while smaller indicate worse performance.
Wilcoxon Signed-Rank Test

1. Rank the differences and compute the average ranks.
2. Choose one group (positive or negative) and compute the ranked sum $R_1$ of that group. Choose either by interest or by which group’s ranked sum is easier to compute.
3. Compute the test statistic

$$T = \begin{cases} \left[ R_1 - \frac{n(n+1)}{4} \right] - 1 \right] / \sqrt{\text{Var}(T)} & \text{if } R_1 \neq \frac{n(n+1)}{4} \\ 0 & \text{otherwise} \end{cases}$$

using variance

$$\text{Var}(T) = \begin{cases} \sum_{j=1}^{n} r_j^2 / 4 & \text{if ties exists} \\ n(n+1)(2n+1)/24 & \text{if no ties (easier to compute)} \end{cases}$$
This test should be used only if the number of nonzero differences is $\geq 16$ and if the difference scores appear symmetric. For data with $n < 16$, small-sample tables must be used (Table 11 in the back of your book).

1. Reject $H_0$ if $T > z_{1-\alpha/2}$, otherwise accept $H_0$.
2. $p = 2 \times [1 - \Phi(T)]$

**Figure 9.5** Computation of the $p$-value for the Wilcoxon signed-rank test
Revisiting our data with Wilcoxon T

1. \( R_1 = 10(7.5) + 6(19.5) + 2(28.0) = 248 < 40(41)/4 = 410 \)
2. \( \text{Var}(T) = \left[ 14(7.5)^2 + 10(19.5)^2 + \ldots + 3(38)^2 + (40)^2 \right] = 5449.75 \)
3. \( T = (|248 - 410| - 0.5)/\sqrt(5449.75) = 2.19 \)
4. \( p = 2 \left[ 1 - \Phi(2.19) \right] = 0.029 \)

Reject \( H_0 \), the two sunblocks are not equally effective. \( R_1 < E(R_1) \) indicates the right arm more often sunburned than the left arm, meaning the sunblock on the left arm has the stronger effect.

This test can be used in lieu of the paired \( t \)-test in cases where sample size is small or you’re not sure if normality applies to your data.
Parametric vs. Nonparametric testing

Paired nonparametric tests
  - Sign Test
  - Wilcoxon Signed-Rank Test

Comparing two samples
  - Wilcoxon Rank-Sum Test
  - Permutation test

Comparing many samples
Setup

Say we have the data below, where people with different types of a genetic retinal disease have had their visual acuity measured. Can we test whether one genetic type deteriorates sight more than the other?

<table>
<thead>
<tr>
<th>Table 9.3</th>
<th>Comparison of visual acuity in people ages 10–19 with dominant and sex-linked RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual acuity</td>
<td>Dominant</td>
</tr>
<tr>
<td>20–20</td>
<td>5</td>
</tr>
<tr>
<td>20–25</td>
<td>9</td>
</tr>
<tr>
<td>20–30</td>
<td>6</td>
</tr>
<tr>
<td>20–40</td>
<td>3</td>
</tr>
<tr>
<td>20–50</td>
<td>2</td>
</tr>
<tr>
<td>20–60</td>
<td>0</td>
</tr>
<tr>
<td>20–70</td>
<td>0</td>
</tr>
<tr>
<td>20–80</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
</tr>
</tbody>
</table>
Wilcoxon Rank-Sum Test

We’d normally test the hypothesis

\[ H_0 : F_D = F_{SL} \]
\[ H_1 : F_D(x) = F_{SL}(x - \Delta) \quad \text{where} \quad \Delta \neq 0 \]

with a two-sample $t$-test, where

- $F_D =$ c.d.f of visual acuity for dominant group
- $F_{SL} =$ c.d.f of visual acuity for sex-linked group
- $\Delta =$ shift in the $F_{SL}$ relative to $F_D$.
  - $\Delta = 0$ indicates visual acuity of both groups are similar
  - $\Delta > 0$ indicates visual acuity of dominant group is worse
  - $\Delta < 0$ indicates visual acuity of sex-linked group is worse

However, because visual acuity (20-X vision) is ordinal, we need to use the nonparametric version of the two-sample $t$-test: the **Wilcoxon rank-sum test**.
Wilcoxon Rank-Sum Test – what to do

1. Combine samples into a single sample and order the values from best to worst (20-20 to 20-80).

2. Assign ranks to each measurement.

3. Measurements with the same value (ties) should be assigned an average rank (as was done for the signed-rank test).

4. Choose one group (arbitrarily) and compute the ranked sum $R_1$ of that group.

\[
T = \begin{cases} 
\left[ R_1 - \frac{n_1(n_1+n_2+1)}{2} \right] - 1/2 \bigg/ \sqrt{Var(T)} & \text{if } R_1 \neq \frac{n_1(n_1+n_2+1)}{2} \\
0 & \text{otherwise}
\end{cases}
\]

\[
Var(T) = \frac{n_1 n_2}{12} \left( n_1 + n_2 + 1 - \frac{\sum_{i=1}^{g} t_i(t_i^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)} \right)
\]

with $g$ tied groups.
This test should be used only if the \( \min(n_1, n_2) \geq 10 \). Also know that this test goes by the name **Mann-Whitney U test**. While the computation is different, the test statistic and \( p \)-value computed for the same data are identical.

1. Reject \( H_0 \) if \( T > z_{1-\alpha/2} \), otherwise accept \( H_0 \).
2. \( p = 2 \times [1 - \Phi(T)] \)

**Figure 9.6** Computation of the \( p \)-value for the Wilcoxon rank-sum test
Which genetic retinal disease type deteriorates more quickly?

1. \( R_1 = 5(3.5) + 9(13.5) + \ldots + 2(42.5) = 479 < 25(56)/2 = 700 \)

\[
\text{Var}(R_1) = \left[ \frac{25(30)}{12} \right] \left\{ 56 - \left[ 6(6^2 - 1) + 14(14^2 - 1) + \ldots + 1(1^2 - 1) \right] / 55(54) \right\} \\
= 62.5(56 - 5382/2970) = 3386.74
\]

2. \( T = \left( |470 - 700| - 0.5 \right) / \sqrt{3386.74} = 3.79 \)

3. \( p = 2 \left[ 1 - \Phi(3.79) \right] = 0.00015 \)

Therefore we reject \( H_0 \) as the visual acuities between the two groups are significantly different (sex-linked vision is much worse than dominant group).
Idea behind the permutation test

We are often testing the null hypothesis that two samples are drawn from the same distribution, i.e. the only reason we see any differences between samples is due to randomization and finite sample size.

The permutation test is a way to explicitly test this idea. If two samples $A, B$ consist of observations from the same underlying distribution, then a test statistic computed on a random sample from observations of $A, B$ shouldn’t give a significantly different result than that statistic computed on the original data.

The permutation test provides a way to estimate the $p$-value for many tests without having to rely on normal approximations at the cost of increased computational power.
How to perform a permutation test

For example, if we want to perform the Wilcoxon rank-sum test on data from samples $A, B$ with sizes $m, n$, we follow the following steps:

1. Compute the rank sum $R_1$ for sample $A$ with ranks determined from the combined sample of $N = m + n$ observations.
2. Enumerate all $\binom{N}{m}$ ways to scramble the $N$ observations into sample $A$, called permutations.
3. Recompute the rank sum $R_{1,\text{perm}}$ for sample $A$ under each of these permutations.
4. Compute $p$:

$$p = 2 \times \min [Pr(R_{1,\text{perm}} \leq R_1), Pr(R_{1,\text{perm}} \geq R_1), 0.5].$$
Permutation test caveats

If the sample sizes are large, enumerating all permutations of the samples can become problematic. For example, $n = m = 20$, $(\frac{40}{20}) > 137 \times 10^9$. For these cases, it is typical that some large number $R$ of permutations are randomly computed; 1,000, 5,000, or 10,000 permutations are common. The error in your estimate of the $p$-value

$$\sqrt{\frac{p_{true}(1 - p_{true})}{R}}$$

is reduced by increasing $R$. If computing a permutation test this way, make sure to mention this value when describing your methods.
Table of Contents

Parametric vs. Nonparametric testing

Paired nonparametric tests
   Sign Test
   Wilcoxon Signed-Rank Test

Comparing two samples
   Wilcoxon Rank-Sum Test
   Permutation test

Comparing many samples
Comparing means among more than two samples

Given normally distributed samples and cardinal data, this problem can be solved using one-way ANOVA. However, with data far from normality or with ordinal data, we need a nonparametric alternative: the Kruskal-Wallis Test.

The most intuitive way to understand the Krusal-Wallis test is to think of it as a generalization of the Wilcoxon rank-sum test in which data from two samples were pooled together. With the Krusal-Wallis test, ranks are computed from pooled data from all samples.
Setup

Take the following comparing anti-inflammatory effects of different drugs in rabbits.

\( H_0 \): treatments are equally effective vs \( H_1 \): some treatments are better than others.

### Ocular anti-inflammatory effects of four drugs on lid closure after administration of arachidonic acid

<table>
<thead>
<tr>
<th>Rabbit Number</th>
<th>Indomethacin</th>
<th>Aspirin</th>
<th>Piroxicam</th>
<th>BW755C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>Rank</td>
<td>Score</td>
<td>Rank</td>
</tr>
<tr>
<td>1</td>
<td>+2</td>
<td>13.5</td>
<td>+1</td>
<td>9.0</td>
</tr>
<tr>
<td>2</td>
<td>+3</td>
<td>20.0</td>
<td>+3</td>
<td>20.0</td>
</tr>
<tr>
<td>3</td>
<td>+3</td>
<td>20.0</td>
<td>+1</td>
<td>9.0</td>
</tr>
<tr>
<td>4</td>
<td>+3</td>
<td>20.0</td>
<td>+2</td>
<td>13.5</td>
</tr>
<tr>
<td>5</td>
<td>+3</td>
<td>20.0</td>
<td>+2</td>
<td>13.5</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>4.0</td>
<td>+3</td>
<td>20.0</td>
</tr>
</tbody>
</table>

\(^a\)(Lid-closure score at baseline - lid-closure score at 15 minutes)\textsubscript{drug eye} - (lid-closure score at baseline - lid-closure score at 15 minutes)\textsubscript{saline eye}
Pooled ranks

1. Create combined sample of size $N = \sum n_i$.
2. Assign ranks to measurements, using average ranks for ties.
3. Compute rank sums $R_i$ for each of the $k$ original samples.
Kruskal-Wallis test statistic

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

or, if there are ties,

\[ H = H^*/ \left[ 1 - \frac{\sum_{j=1}^{g} t_j^3 - t_j}{N^3 - N} \right] \]

This test should be used only if each sample \( n_i \geq 5 \).
$p$-value for Kruskal-Wallis test

- Accept $H_0$ if $H \leq \chi^2_{k-1,1-\alpha}$, otherwise reject $H_0$.
- $p = Pr(\chi^2_{k-1} > H)$. 

![Graph showing the $\chi^2$ distribution with a shaded area representing the $p$-value.](image)
Worked Example

\[ R_1 = 13.5 + 20.0 + \ldots + 4.0 = 97.5 \]

\[ R_2 = 9.0 + 20.0 + \ldots + 20.0 = 85.0 \]

\[ R_3 = 20.0 + 09.0 + \ldots + 20.0 = 91.5 \]

\[ R_4 = 9.0 + 4.0 + \ldots + 1.0 = 26.0 \]

\[ H = \frac{12}{24(25)} \left( 97.5^2 + 85^2 + 91.5^2 + 26.0^2 \right) / 6 - 3(25) \]

\[ = \frac{0.020(4296.583) - 75}{1 - \frac{1020}{13,800}} = 11.804 \]

We have 4 \(- 1 = 3 \) df, so we compare to \( \chi^2_3 \), which yields 0.005 \(< p < 0.01 \), so we should expect a significant difference in the anti-inflammatory effect of these drugs.