1) Introduction

- Most of our previous inferences have been based on the assumption of an underlying normal distribution for the test statistic. For example, it was pointed out previously that \( \bar{X} \) has a normal distribution through the central limit theorem for large sample sizes. To be more specific, one can use the 15/40 Rule. This states that \( \bar{X} \) is sufficiently close to normal so that one can safely use \( t \)-statistics as follows:
  - for \( n < 15 \), only if data are close to normal
  - for \( 15 < n < 40 \), if data does not have outliers or strong skewness
  - for \( n = 40 \), generally ok.

- In cases where the original data are skewed, we have often transformed the data to achieve normality, as e.g. in the rank transformation for correlation and the log transformation for the odds ratio. There is a body of methods which do not assume a normal distribution. These are called nonparametric, or distribution free, methods.

- There are nonparametric procedures for ANOVA and regression. Our coverage will be limited to comparing two means. Our procedures assuming the normal distribution were covered in Chapter 11. There were two cases, one for independent samples and the other for paired samples. We have nonparametric methods for each of these cases, also. We'll start with independent samples, and then cover paired samples. The classification of methodology is as follows.

<table>
<thead>
<tr>
<th>Normality Assumption</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Independent</td>
</tr>
<tr>
<td>Yes</td>
<td>two-sample ( t ) test</td>
</tr>
<tr>
<td>No</td>
<td>rank-sum test</td>
</tr>
</tbody>
</table>

2) Wilcoxon Rank Sum Test

- This is also called the Mann-Whitney \( U \) test.

- Suppose one has two independent samples. If both sample sizes are large (e.g. \( n_1, n_2 > 40 \)), the sample means are approximately normal under most circumstances by virtue of the central limit theorem. Suppose however that the sample sizes are both less than 40, and that we are unwilling to assume that the populations follow normal distributions. Under these conditions, we desire some alternative test.

- Most nonparametric tests are based on using ranks in some way. This test uses the rank sum. It assumes that the population distributions have the same general shape,
which may of course be nonnormal. It tests whether the population medians are equal, i.e.

\[ H_0 : M_1 = M_2 \]

Because the populations are assumed to have the same general shape, another equivalent way to state \( H_0 \) is

\[ H_0 : \text{the populations are identical} \]

vs.

\[ H_A : \text{one is shifted to the right (or left).} \]

• As an example, consider the problem in the book. Two populations of children with PKU (phenylketonuria) were examined, the first has low exposure to (unmetabolized) phenylalanine (a protein) and the other has high exposure. We desire to test whether the level of exposure is related to the child's mental age score, \( X \). The experimenters are not willing to assume that \( X \) has a normal distribution for each population. There were \( n_1 = 21 \) children with low exposure and \( n_2 = 18 \) with high. The data are displayed in Fig. 13.2.

Do the populations of \( X \) appear to be nonnormally distributed?

• The Wilcoxon rank sum test procedure is as follows:

1) Rank (from lowest to highest) the data from the combined samples. Tied observations receive the average rank. (Note this is done in the table).

2) Find the sum of ranks for both samples. Let \( W \) denote the smaller sum, with \( n_s \) the sample size corresponding to the smaller sum and \( n_L \) the other sample size.

3) If \( n_s \) and \( n_L < 10 \), the distribution of \( W \) under \( H_0 \) is given in Table A.7. For larger samples, one may assume \( W \) has a normal distribution. Its mean is

\[ \mu_W = n_s (n_s + n_L + 1) / 2 \]

with variance

\[ \sigma^2_W = n_s n_L (n_s + n_L + 1) / 12. \]

Hence

\[ z = (W - \mu_W) / \sigma_W \]

is a standard normal random variable, from which we can test \( H_0 \).

• As an illustration, we consider the previous problem.

A subset of the data with corresponding ranks is given below:
### Low Exposure

<table>
<thead>
<tr>
<th>x</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.5</td>
<td>2.0</td>
</tr>
<tr>
<td>37.5</td>
<td>6.0</td>
</tr>
<tr>
<td>39.5</td>
<td>7.0</td>
</tr>
<tr>
<td>40.0</td>
<td>8.0</td>
</tr>
<tr>
<td>45.5</td>
<td></td>
</tr>
</tbody>
</table>

### High Exposure

<table>
<thead>
<tr>
<th>x</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.0</td>
<td>1.0</td>
</tr>
<tr>
<td>35.0</td>
<td>3.0</td>
</tr>
<tr>
<td>37.0</td>
<td>4.5</td>
</tr>
<tr>
<td>37.0</td>
<td>4.5</td>
</tr>
<tr>
<td>43.5</td>
<td></td>
</tr>
</tbody>
</table>

- To complete the hypothesis test, one has:
  1) \( H_0 : M_1 = M_2 \) (equal medians)
  2) \( H_A : M_1 \neq M_2 \) (unequal medians)
  3) \( TS \): Wilcoxon rank sum \( W \)
  4) \( RR \): for \( \alpha = .05 \), because \( n_1 \) and \( n_2 > 10 \) reject if \( |z| > 1.96 \)
  5) Calculations:

\[
W = 313 \\
\mu_w = 18(18 + 21 + 1)/2 = 360 \\
\sigma_w^2 = (18)(21)(18 + 21 + 1)/12 = 1260 \\
z = (313 - 360)/\sqrt{1260} = -1.32
\]

Therefore, do not reject \( H_0 \), in fact, \( p = 2(0.093) = 0.186 \).

- If \( n_1 \) and \( n_2 < 10 \), the exact distribution of \( W \) is tabulated. For example, suppose \( n_1 = 5 \) and \( n_2 = 4 \). To use the table, we re-order, so that \( n_1 \) is the smaller, hence let \( n_1 = 4 \) and \( n_2 = 5 \). If \( W = 13 \), one finds on page A-17 that

\[
\text{Prob}(W \leq 13) = 0.0556,
\]

which is the p-value for a one-sided test. The p-value for a two-sided test is

\[
p = 2(0.0556) = .111.
\]

**3) Wilcoxon Signed-Rank Test**

- The Wilcoxon signed-rank test may be used for paired samples, when the differences are not assumed to follow a normal distribution as required for the paired \( t \)-test. The null hypothesis is that the median difference is 0, i.e.
$H_0: M_0 = 0.$

This implies that a difference is equally likely to be positive or negative. Why is this so?

- The alternative hypothesis could be one-sided or two sided.

- There is another test, namely the sign test (Section 13.1) which could also be used. It is not as powerful, nor used often in practice, hence we'll skip it.

- Consider the following example. The objective is to determine whether a drug is successful in reducing the forced vital capacity (FVC) of patients with cystic fibrosis. Because the inherent FVC varies so much among patients, a paired design was used where the FVC was measured for a given patient with the drug and with a placebo. The data for 14 patients follow.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Placebo</th>
<th>Drug</th>
<th>Difference</th>
<th>Rank</th>
<th>Signed Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>224</td>
<td>213</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>95</td>
<td>-15</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>33</td>
<td>42</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>541</td>
<td>440</td>
<td>101</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>74</td>
<td>-32</td>
<td>106</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>-28</td>
<td>113</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>293</td>
<td>448</td>
<td>-152</td>
<td>7</td>
<td>-7</td>
</tr>
<tr>
<td>8</td>
<td>-23</td>
<td>-178</td>
<td>155</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>525</td>
<td>367</td>
<td>158</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>-38</td>
<td>140</td>
<td>-178</td>
<td>10</td>
<td>-10</td>
</tr>
<tr>
<td>11</td>
<td>508</td>
<td>323</td>
<td>185</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>255</td>
<td>10</td>
<td>245</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>525</td>
<td>65</td>
<td>450</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>1023</td>
<td>343</td>
<td>680</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

- Because this is a paired design, the experimenter proceeds to find the difference in FVC reduction for each patient. A graph is given in Fig. 13.1. The experimenter is not willing to assume normality.

- The Wilcoxon signed-rank test statistic is found as follows.

  1) Find the differences, $d_i$

  2) Rank the absolute values of the non-zero differences. (A difference of 0 is ignored except for reducing the sample size by 1).

  3) Attach the signs of the differences to the ranks

<table>
<thead>
<tr>
<th>Signed Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
</tr>
<tr>
<td>-19</td>
</tr>
</tbody>
</table>
4) Find $T_+$, the sum of the positive ranks
5) Find $T_-$, the sum of the negative ranks
6) Let $T$ be the smaller of $T_+$ and $T_-$ in absolute value,

- If $n \leq 12$, $T$ is not assumed to follow a normal, and specialized tables are used. For $n > 12$, $T$ is approximately normal with mean
  \[ \mu_T = n(n + 1)/4 \]
  and variance
  \[ \sigma_T^2 = n(n + 1)(2n + 1)/24. \]
  Hence, we transform to
  \[ z = (T - \mu_T) / \sigma_T \]
  and use test procedures based on the standard normal.

- Some care must be exercised using this method the book outlines. $T$ is the smaller of $T_+$ and $T_-$, hence the calculated $z$ cannot be positive. One would always reject in the lower tail, with appropriate $\alpha$ or $\alpha/2$ critical value.

- To illustrate in our problem, note that the data were ordered to give consecutive ranks in column 5. Tied observations would be given the average rank. The signs of the differences are attached to the ranks in columns 6 and 7. The sum of positive and negative ranks are
  \[ T_+ = 86, \quad T_- = -19 \]
  whereupon
  \[ T = 19. \]

- Our hypothesis test is
  1) $H_0 : M_d = 0$
  2) $H_A : M_d > 0$, i.e. the drug diminishes the reduction in FVC.
  3) $T$ Wilcoxon $T$
  4) $RR$ Reject if $z < -1.645$.
    Note we always expect to find a negative $z$, because $T$ is the smaller sum.
  5) Calculations:

    $T = 19$
    \[ \mu_T = 14(15)/4 = 52.5 \]
    \[ \sigma_T^2 = 14(15)(29)/24 = 253.75 \]
    \[ z = (19 - 52.5) / \sqrt{253.75} = -2.10. \]

  Hence we reject $H_0$. We have statistical evidence to conclude that the drug tends to diminish the FVC reduction. ($p = 0.018$).
• If $n \leq 12$, the exact distribution of $T$ is tabulated in Table A.6. For example, with $n = 8$ and $T = 5$, one has under $H_0$

$$\text{Prob}[T \leq 5] = 0.0391.$$  

Hence the $p$-value for a one-sided test is $p = 0.039$.

4) Advantages and Disadvantages of Nonparametric Methods

• The chief advantage of nonparametric tests is that they do not require the assumption of normality for the underlying population(s). They can also be easier by avoiding, for example, transformations to normality (hence the description of 'quick and dirty' methods).

• The disadvantage occurs primarily when the normal distribution assumption is satisfied and one fails to take advantage of it. If one has normality and uses a Wilcoxon test, the power is approximately 0.95 of the corresponding $t$-tests. This may be interpreted to mean that whatever power a $t$-test achieves with $n = 19$ observations would require $n = 20$ observations with a Wilcoxon test. Another minor technical disadvantage is that the Wilcoxon test assumes that one doesn't have a large proportion of ties (which is unlikely to occur in most applications anyway).