Comparing More Than Two Treatments

Suppose we randomize patients to $K > 2$ treatments and are interested in testing whether there are treatment differences in response. We proceed by positing the null hypothesis that there are no treatment differences and then test this null hypothesis against the alternative hypothesis that there exist some difference in response among the treatments.

Let us first consider the case where the response is dichotomous; i.e. the patient either responds or doesn’t respond to treatment. Let $\pi_1, \ldots, \pi_K$ denote the population response rates for the $K$ treatments under study. The null hypothesis of no treatment differences is given by

$$H_0 : \pi_1 = \pi_2 = \ldots = \pi_K.$$  

The alternative hypothesis $H_A$ : states that the $K$ population response rates “the $\pi$’s” are not all equal.

To decide between $H_0$ and $H_A$, we conduct an experiment where we allocate $n_1, \ldots, n_K$ individuals on each of the $K$ treatments respectively and count the number that respond to each of the treatments. Therefore, the data from such a clinical trial can be viewed as realizations of the independent random variables $X_1, \ldots, X_K$, where

$$X_i \sim b(n_i, \pi_i), \; i = 1, \ldots, K.$$ 

Each of the population response rates $\pi_i$ are estimated using the sample proportions $p_i = X_i/n_i, \; i = 1, \ldots, K$.

The strategy for decision making is to combine the data from such an experiment into a test statistic for which larger values would provide increasing evidence against the null hypothesis. In addition, the distribution of this test statistic, under the null hypothesis, is necessary in order to gauge how extreme (evidence against $H_0$) the observed data are compared to what may have occurred by chance if the null hypothesis were true.

We begin by by first giving some general results regarding estimators that are normally distributed which will be used later for constructing test statistics.
7.1 Testing equality using independent normally distributed estimators

For \( i = 1, \ldots, K \), let the estimator \( \hat{\theta}_i \) be an unbiased estimator for the parameter \( \theta_i \) which has a normal distribution; i.e.

\[
\hat{\theta}_i \sim N(\theta_i, \sigma_i^2).
\]

Assume \( \sigma_i^2 \) is either known or can be estimated well (consistently) using the data.

Suppose we have \( K \) independent estimators \( \hat{\theta}_1, \ldots, \hat{\theta}_K \) of \( \theta_1, \ldots, \theta_K \) respectively such that

\[
\hat{\theta}_i \sim N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, K.
\]

If we assume that \( \theta_1 = \ldots = \theta_K \) (i.e. similar to a K-sample null hypothesis), then the weighted estimator

\[
\hat{\theta} = \frac{\sum_{i=1}^{K} w_i \hat{\theta}_i}{\sum_{i=1}^{K} w_i}, \quad w_i = \frac{1}{\sigma_i^2}
\]

is the best unbiased estimator for the common value \( \theta \).

**Remark:** By best, we mean that among all functions of \( \hat{\theta}_1, \ldots, \hat{\theta}_K \) that are unbiased estimators of \( \theta \), \( \hat{\theta} \) (defined above) has the smallest variance.

In addition, when \( \theta_1 = \ldots = \theta_K \), then the random variable

\[
\sum_{i=1}^{K} w_i (\hat{\theta}_i - \hat{\theta})^2 \tag{7.1}
\]

is distributed as a central chi-square distribution with \( K - 1 \) degrees of freedom. If the \( \theta_i, i = 1, \ldots, K \) are not all equal, then

\[
\sum_{i=1}^{K} w_i (\hat{\theta}_i - \hat{\theta})^2 \tag{7.2}
\]

is distributed as a non-central chi-square distribution with \( K - 1 \) degrees of freedom and non-centrality parameter equal to

\[
\sum_{i=1}^{K} w_i (\theta_i - \bar{\theta})^2, \tag{7.3}
\]

where

\[
\bar{\theta} = \frac{\sum_{i=1}^{K} w_i \theta_i}{\sum_{i=1}^{K} w_i}.
\]

**Note:** The non-centrality parameter is based on the true population parameters.
7.2 Testing equality of dichotomous response rates

How do the results above help us in formulating test statistics? Returning to the problem of dichotomous response, the null hypothesis is

\[ H_0 : \pi_1 = \ldots = \pi_K. \]

This is equivalent to

\[ H_0 : \sin^{-1}\sqrt{\pi_1} = \ldots = \sin^{-1}\sqrt{\pi_K}. \]

We showed in chapter 6 that if \( p_i = X_i/n_i \) is the sample proportion responding to treatment \( i \), then

\[ \sin^{-1}\sqrt{p_i} \sim N\left( \sin^{-1}\sqrt{\pi_i}, \frac{1}{4n_i} \right). \]

Letting

- \( \sin^{-1}\sqrt{p_i} \) take the role of \( \hat{\theta}_i \)
- \( \sin^{-1}\sqrt{\pi_i} \) take the role of \( \theta_i \)
- \( \frac{1}{4n_i} \) be \( \sigma_i^2 \); hence \( w_i = 4n_i \)

then by (7.1), the test statistic

\[ T_n = \sum_{i=1}^{K} 4n_i (\sin^{-1}\sqrt{p_i} - \bar{A}_p)^2, \quad (7.4) \]

where

\[ \bar{A}_p = \frac{\sum_{i=1}^{K} n_i \sin^{-1}\sqrt{p_i}}{\sum_{i=1}^{K} n_i}, \]

is distributed as a central chi-square distribution with \( K - 1 \) degrees of freedom under the null hypothesis \( H_0 \).

This test statistic is a reasonable measure for assessing the strength of evidence of the alternative hypothesis. If the population response rates \( \pi_i, i = 1, \ldots, K \) are all the same, then we would expect the sample proportions \( p_i \) to also be close to each other; in which case the test statistic \( T_n \) would be near zero. The more different the \( p_i \)'s are from each other the greater \( T_n \) would be. Consequently, a reasonable strategy would be to reject the null hypothesis when the statistic \( T_n \)
is sufficiently large. The degree of evidence is obtained by gauging the observed value of the test statistic to the chi-square distribution with $K - 1$ degrees of freedom. If we denote by $\chi^2_{\alpha; K-1}$, the $(1 - \alpha)$-th quantile of the central chi-square distribution with $K - 1$ degrees of freedom, then a test at the $\alpha$-th level of significance can be obtained by rejecting $H_0$ when

$$T_n \geq \chi^2_{\alpha; K-1}.$$

**Power and sample size calculations**

In order to assess the power of the test, we need to know the distribution of the test statistic under the alternative hypothesis. Suppose we entertain a specific alternative hypothesis that we believe is clinically important to detect (more about this later). Say,

$$H_A : \pi_1 = \pi_{1A}, \ldots, \pi_K = \pi_{KA},$$

where the $\pi_{iA}$’s are not all equal. Using (7.2), the distribution of $T_n$ is a non-central chi-square with $K - 1$ degrees of freedom and non-centrality parameter given by (7.3), which in this case equals

$$\phi^2 = \sum_{i=1}^{K} 4n_i(sin^{-1}\sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2,$$

where

$$\bar{A}_{\pi A} = \frac{\sum_{i=1}^{K} n_i sin^{-1}\sqrt{\pi_{iA}}}{\sum_{i=1}^{K} n_i}.$$

The level-$\alpha$ test rejects $H_0$ when $T_n \geq \chi^2_{\alpha; K-1}$. Therefore, the power of the test to detect $H_A$ is the probability that a non-central chi-square distribution, with $K - 1$ degrees of freedom and non-centrality parameter $\phi^2$, given by (7.5), exceed the value $\chi^2_{\alpha; K-1}$. See illustration below.

There are tables and/or computer packages available to carry out these probability calculations.

**Sample size calculations**

Suppose we allocate patients to each of the $K$ treatments equally. Thus, on average, $n_1 = \ldots = n_K = n/K$, where $n$ denotes the total sample size. In that case, the non-centrality parameter (7.5) equals

$$\phi^2 = \frac{4n}{K} \left\{ \sum_{i=1}^{K} (sin^{-1}\sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2 \right\},$$

(7.6)
where

$$\bar{A}_{\pi A} = K^{-1} \sum_{i=1}^{K} \sin^{-1} \sqrt{\pi_{iA}}.$$

**Note:** $\bar{A}_{\pi A}$ is just a simple averages of the $\sin^{-1} \sqrt{\pi_{iA}}$s for the alternative of interest.

Let us define by

$$\phi^2(\alpha, \beta, K - 1),$$

the value of the non-centrality parameter necessary so that a non-central chi-square distributed random variable with $K - 1$ degrees of freedom and non-centrality parameter $\phi^2(\alpha, \beta, K - 1)$ will exceed the value $\chi^2_{\alpha:K-1}$ with probability $(1 - \beta)$. Values for $\phi^2(\alpha, \beta, K - 1)$ are given in tables or derived from software packages for different values of $\alpha, \beta$ and degrees of freedom $K - 1$. For illustration, we have copied a subset of such tables that we will use later in examples.

Thus, if we want a level-$\alpha$ test of the null hypothesis $H_0 : \pi_1 = \ldots = \pi_K$ to have power $(1 - \beta)$ to detect the alternative hypothesis $H_A : \pi_1 = \pi_{1A}, \ldots, \pi_K = \pi_{KA}$ (not all equal), then we need the non-centrality parameter

$$\frac{4n}{K} \left\{ \sum_{i=1}^{K} (\sin^{-1} \sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2 \right\} = \phi^2(\alpha, \beta, K - 1),$$
or

\[ n = \frac{K \phi^2(\alpha, \beta, K - 1)}{4 \left\{ \sum_{i=1}^{K} (\sin^{-1} \sqrt{\pi_i A} - \bar{A}_\pi A)^2 \right\}}. \]  

(7.7)

Choosing clinically important alternatives

Specifying alternatives that may be of clinical importance for comparing \( K \) treatments may not be that straightforward. (It is not that easy even for two-treatment comparisons). One conservative strategy is to find the sample size necessary to have sufficient power if any of the \( K \) treatments differ from each other by \( \Delta A \) or more.

Remark: Keep in mind that by using the arcsin square-root transformation, all treatment differences are measured on this scale.

To do this we must find the least favorable configuration of population response probabilities subject to the constraint that at least two treatment response rates on the arcsin square-root scale differ by \( \Delta A \) or more. By the least favorable configuration, we mean the configuration which would lead to the smallest non-centrality parameter. If we can find such a least favorable configuration, then this would imply that any other configuration has a larger non-centrality parameter, thus, any other configuration would have power at least as large as the least favorable one.

For a fixed sample size we showed in (7.6) that the non-centrality parameter is proportional to \( \{ \sum_{i=1}^{K} (\sin^{-1} \sqrt{\pi_i A} - \bar{A}_\pi A)^2 \} \), although we will not prove this formally, it is intuitively clear that the least favorable configuration is obtained when two treatments, the one with the largest and smallest response, differ by \( \Delta A \) (in this case on the arcsin square root scale) and the other treatments have response half way in between.

It is clear from the picture above that

\[ \{ \sum_{i=1}^{K} (\sin^{-1} \sqrt{\pi_i A} - \bar{A}_\pi A)^2 \} = \left( \frac{\Delta A}{2} \right)^2 + 0 + \ldots + 0 + \left( \frac{\Delta A}{2} \right)^2 = \frac{\Delta A^2}{2}. \]

Substituting the result for the least favorable configuration in equation (7.7) yields

\[ n = \frac{K \phi^2(\alpha, \beta, K - 1)}{2 \Delta A^2}. \]  

(7.8)

Example: Suppose a standard treatment has a response rate of about .30. Another three
treatments have been developed and it is decided to compare all of them in a head to head randomized clinical trial. Equal allocation of patients to the four treatments is used so that approximately \( n_1 = n_2 = n_3 = n_4 = n/4 \). We want the power of a test, at the .05 level of significance, to be at least 90% if any of the other treatments has a response rate greater than or equal to .40. What should we choose as the sample size?

For this example:

- \( \alpha = .05 \)
- \( 1 - \beta = .90 \), or \( \beta = .10 \)
- \( K=4 \)
- \( \Delta_A = \sin^{-1}\sqrt{.40} - \sin^{-1}\sqrt{.30} = .1051 \)
- \( \phi^2(.05, .10, 3) = 14.171 \) (derived from the tables provided)

Therefore by (7.8), we get

\[
n = \frac{4 \times 14.171}{2(.1051)^2} = 2567,
\]

or about 2567/4 = 642 patients per treatment arm.

### 7.3 Multiple comparisons

If we had done a two-treatment comparison for the previous example, and wanted 90% power to detect a difference from .30 to .40 in response rate at the .05 (two-sided) level of significance, then the sample size necessary is

\[
n = \frac{2 \times 10.507}{2(.1051)^2} = 952,
\]

or 476 patients per treatment arm.

For a four-treatment comparison, we needed 642 patients per treatment arm under similar parameter and sensitivity specifications. Why is there a difference and what are the implications of this difference?
We first note that when we test four treatments simultaneously, we can make six different pairwise comparisons; i.e.

\[ 1 \text{ vs } 2, \ 1 \text{ vs } 3, \ 1 \text{ vs } 4, \ 2 \text{ vs } 3, \ 2 \text{ vs } 4, \ 3 \text{ vs } 4. \]

If each of these comparisons were made using separate studies, each at the same level and power to detect the same alternative, then six studies would be conducted, each with 952 patients, or 952 \( \times 6 = 5,712 \) total patients. This is in contrast to the 2,567 patients used in the four-treatment comparison.

This brings up the controversial issue regarding multiple comparisons. When conducting a clinical trial with \( K \) treatments, \( K > 2 \), there are \( \binom{K}{2} \) possible pairwise comparisons that can be made. If each comparison was made at the \( \alpha \) level of significance, then even if the global null hypothesis were true; (i.e. in truth, all treatments had exactly the same response rate), the probability of finding at least one significant difference will be greater than \( \alpha \). That is, there is an inflated study-wide type I error that is incurred due to the multiple comparisons.

Let’s be more specific. Denote by \( T_{nij} \) the test statistic used to compare treatment \( i \) to treatment \( j \), \( i < j \), \( i, j = 1, \ldots, K \). There are \( \binom{K}{2} \) such pairwise treatment comparisons, each using the test statistic

\[ T_{nij} = \frac{2(\sin^{-1}\sqrt{\pi_i} - \sin^{-1}\sqrt{\pi_j})}{(\frac{1}{n_i} + \frac{1}{n_j})^{1/2}}. \]

For a two-sample comparison, the level \( \alpha \) test (two-sided) would reject the hypothesis \( H_{0ij} : \pi_i = \pi_j \) if

\[ |T_{nij}| \geq Z_{\alpha/2}. \]

That is,

\[ P_{H_{0ij}}(|T_{nij}| \geq Z_{\alpha/2}) = \alpha. \]

However, if we made all possible pairwise treatment comparisons then the event of finding at least one significant result is

\[ \bigcup_{i<j, \ i, j = 1, \ldots, K} (|T_{nij}| \geq Z_{\alpha/2}). \]
It is clear that
\[ P_{H_0}\left\{ \bigcup_{i<j, i,j=1,\ldots,K} (|T_{nij}| \geq Z_{\alpha/2}) \right\} > P_{H_{0_{ij}}}(|T_{nij}| \geq Z_{\alpha/2}) = \alpha. \]

If we are concerned about this issue and want to control the overall study-wide type I error rate, then we would require greater evidence of treatment difference before declaring a pairwise-treatment comparison to be significant. One simple, albeit conservative, strategy is to use Bonferroni correction. This follows by noting that
\[ P_{H_0}\left\{ \bigcup_{i<j}(|T_{nij}| \geq c) \right\} \leq \sum_{i<j} P_{H_{0_{ij}}}(|T_{nij}| \geq c). \]

There are \( \binom{K}{2} = K(K-1)/2 \) elements in the sum on the right hand side of the formula above. Therefore, if we choose the constant \( c \) so that each of the probabilities in the sum is equal to
\[ \alpha/\left( \binom{K}{2} \right), \]
then we are ensured that the probability on the left will be less than or equal to \( \alpha \). Since under the null hypothesis \( T_{nij} \) follows a standard normal distribution, we choose
\[ c = Z_{\alpha/\{K(K-1)\}}. \]

Hence, with a Bonferroni correction, we would declare a pairwise treatment comparison (say treatment \( i \) versus treatment \( j \)) significant if
\[ |T_{nij}| \geq Z_{\alpha/\{K(K-1)\}}. \]

Using this as the convention, the probability of declaring any treatment pairwise comparison significant under the null hypothesis is
\[ P_{H_0}\left\{ \bigcup_{i<j}(|T_{nij}| \geq Z_{\alpha/\{K(K-1)\}}) \right\} \leq \sum_{i<j} P_{H_{0_{ij}}}(|T_{nij}| \geq Z_{\alpha/\{K(K-1)\}}) = \binom{K}{2} \times 2\alpha/\{K(K-1)\} = \alpha. \]

**Example:** If we were comparing four treatments, there would be six possible treatment comparisons. Using the Bonferroni method we would declare any such pairwise comparison significant.
at a study-wide global significance level of $\alpha$ if the nominal pairwise p-value was less than $\alpha/6$. For two sided tests, this would imply that the test statistic

$$|T_{nij}| \geq Z_{\alpha/12}.$$  

If we choose $\alpha = .05$, then a simple two-sample comparison requires the two-sample test to exceed 1.96 in order to declare significance. However, as part of a four-treatment comparison, if we use a Bonferroni correction, we need the two-sample test to exceed $Z_{.05/12} \approx 2.635$ to declare significance.

Typically, the way one proceeds when testing for treatment differences with $K > 2$ treatments is to first conduct a global test of the null hypothesis at level $\alpha$ using (7.4). If this test fails to reject the null hypothesis, then we conclude that there are no treatment differences. If the null hypothesis is rejected, then we may want to consider the different pairwise treatment comparisons to determine where the treatment differences occur. A pairwise comparison would be declared significant based on the Bonferroni correction described above. Such a strategy conservatively protects all type I errors that can be committed to be less than $\alpha$.

There are other more sophisticated methods for multiple comparisons which will not be discussed here. There are some clinical trials statisticians that object to the whole notion of correcting for multiple comparisons and believe that each pairwise comparison should be considered as a separate experiment.

**Quote:** Why should we be penalized for having the insight to test more than two treatments simultaneously which allows for $\binom{K}{2}$ treatment comparisons as opposed to conducting $\binom{K}{2}$ separate studies each of which individually would not be penalized. (By penalize we mean increased evidence of treatment difference before significance can be declared.)

The FDA’s position is that they will consider such arguments but it must be agreed to up front with good rationale.

**Chi-square tests for comparing response rates in $K$ treatments**

It was convenient, in developing the theory and deriving sample size calculations, to use the arcsin square-root transformation of sample proportions to test the equality of $K$ treatment
response rates. In general, however, the standard $K$-sample test used to test this null hypothesis is the chi-square test; namely,

$$
\sum_{\text{over } 2 \times K \text{ cells}} \frac{(O_j - E_j)^2}{E_j},
$$

where $O_j$ denotes the observed count in the $j$-th cell of a $2 \times K$ contingency table of response-non-response by the $K$ treatments, and $E_j$ denotes the expected count under the null hypothesis. For comparison, we will construct both the chi-square test and the test based on the arcsin square-root transformation given by (7.4) on the same set of data.

Table 7.1: Observed Counts

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td></td>
<td>206</td>
<td>273</td>
<td>224</td>
<td>275</td>
<td>978</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>437</td>
<td>377</td>
<td>416</td>
<td>364</td>
<td>1594</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>643</td>
<td>650</td>
<td>640</td>
<td>639</td>
<td>2572</td>
</tr>
</tbody>
</table>

For each cell in the $2 \times K$ table, the expected count is obtained by multiplying the corresponding marginal totals and dividing by the grand total. For example, the expected number responding for treatment 1, under $H_0$, is $\frac{978 \times 643}{2572} = 244.5$.

Table 7.2: Expected Counts

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td></td>
<td>244.5</td>
<td>247.2</td>
<td>243.4</td>
<td>243</td>
<td>978</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>398.5</td>
<td>402.8</td>
<td>396.6</td>
<td>396</td>
<td>1594</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>643</td>
<td>650</td>
<td>640</td>
<td>639</td>
<td>2572</td>
</tr>
</tbody>
</table>

The chi-square test is equal to

$$
\frac{(206 - 244.5)^2}{244.5} + \frac{(437 - 398.5)^2}{398.5} + \ldots + \frac{(364 - 396)^2}{396} = 23.43.
$$

Gauging this value against a chi-square distribution with 3 degrees of freedom we get a p-value $< .005$. 

Using the same set of data we now construct the K-sample test (7.4) using the arcsin square-root transformation.

<table>
<thead>
<tr>
<th>Treatment $i$</th>
<th>$p_i$</th>
<th>$\sin^{-1}\sqrt{p_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$206/643$ = .32</td>
<td>.601</td>
</tr>
<tr>
<td>2</td>
<td>$273/650$ = .42</td>
<td>.705</td>
</tr>
<tr>
<td>3</td>
<td>$224/640$ = .35</td>
<td>.633</td>
</tr>
<tr>
<td>4</td>
<td>$275/639$ = .43</td>
<td>.715</td>
</tr>
</tbody>
</table>

$$\bar{A}_p = \frac{643 \times .601 + 650 \times .705 + 640 \times .633 + 639 \times .715}{2572} = .664,$$

and

$$T_n = 4\{643(.601 - .664)^2 + 650(.705 - .664)^2 + 640(.633 - .664)^2 + 639(.715 - .664)^2\} = 23.65.$$

**Note:** This is in good agreement with the chi-square test which, for the same set of data, gave a value of 23.43. This agreement will occur when sample sizes are large as often is the case in phase III clinical trials.

**Example of pairwise comparisons**

Suppose the data above were the results of a clinical trial. Now that we’ve established that it is unlikely that the global null hypothesis $H_0$ is true (p-value < .005), i.e. we reject $H_0$, we may want to look more carefully at the individual pairwise treatment differences.

Additionally, you are told that treatment 1 was the standard control and that treatments 2, 3 and 4 are new promising therapies. Say, that, up front in the protocol, it was stated that the major objectives of this clinical trial were to compare each of the new treatments individually to the standard control. That is, to compare treatments 1 vs 2, 1 vs 3, and 1 vs 4. If we want to control the overall experimental-wide error rate to be .05 or less, then one strategy is to declare any of the above three pairwise comparisons significant if the two-sided p-value were less than .05/3=.0167. With a two-sided test, we would declare treatment $j$ significantly different than treatment 1, for $j = 2, 3, 4$ if the two-sample test

$$|T_{n1j}| \geq Z_{.0167/2} = 2.385, \ j = 2, 3, 4.$$
Remark: The exact same considerations would be made using a one-sided test at the .025 level of significance. The only difference is that we would reject when $T_{nij} \geq 2.385$. We must be careful in defining $T_{n1j}$ that the sign is such that large values are evidence against the one-sided null hypothesis of interest.

With that in mind, we consider the test based on the arcsin square-root transformation

$$T_{n1j} = 2 \left( \frac{n_1 n_j}{n_1 + n_j} \right)^{1/2} \frac{sin^{-1} \sqrt{p_j} - sin^{-1} \sqrt{p_1}}{\sqrt{p_j - p_1}}.$$ 

Substituting the data above we get

$$T_{n12} = 2 \left( \frac{643 \times 650}{643 + 650} \right)^{1/2} (.705 - .601) = 3.73*$$

$$T_{n13} = 2 \left( \frac{643 \times 640}{643 + 640} \right)^{1/2} (.633 - .601) = 1.14$$

$$T_{n14} = 2 \left( \frac{643 \times 639}{643 + 639} \right)^{1/2} (.715 - .601) = 4.08*$$

Thus we conclude that treatments 2 and 4 are significant better than treatment 1 (the standard control).

Suppose all four treatments were experimental, in which case, all six pairwise comparisons are of interest. To account for multiple comparisons, we would declare a pairwise comparison significant if the two-sided p-value were less than $.05/6=.0083$. Thus we would reject the null hypothesis $H_{0ij}: \pi_i = \pi_j$ when

$$|T_{nij}| \geq Z_{.0083/2} = 2.635.$$ 

Note: the comparisons of treatments 1 vs 2 and 1 vs 4 would still be significant with this more stringent criterion.

Rounding out the remaining pairwise comparisons we get

$$|T_{n23}| = |2 \left( \frac{650 \times 640}{650 + 640} \right)^{1/2} (.705 - .633)| = 2.59$$

$$|T_{n24}| = |2 \left( \frac{650 \times 639}{650 + 639} \right)^{1/2} (.705 - .715)| = 0.36$$
\[ |T_{n34}| = |2 \left( \frac{640 \times 639}{640 + 639} \right)^{1/2} \left( .633 - .715 \right)| = 2.94 * . \]

Clearly, treatments 2 and 4 are the better treatments, certainly better than control. The only controversial comparison is treatment 2 versus treatment 3, where, we may not be able to conclude that treatment 2 is significantly better than treatment 3 because of the conservativeness of the Bonferroni correction.

### 7.4 K-sample tests for continuous response

For a clinical trial where we randomize patients to one of \( K > 2 \) treatments and the primary outcome is a continuous measurement, then our primary interest may be to test for differences in the mean response among the \( K \) treatments. Data from such a clinical trial may be summarized as realizations of the iid random vectors

\[(Y_i, A_i), i = 1, \ldots, n,\]

where \( Y_i \) denotes the response (continuously distributed) for the \( i \)-th individual and \( A_i \) denotes the treatment \((1, 2, \ldots, K)\) that the \( i \)-th individual was assigned. Let us denote the treatment-specific mean and variance of response by

\[E(Y_i|A_i = j) = \mu_j, j = 1, \ldots, K\]

and

\[\text{var}(Y_i|A_i = j) = \sigma^2_{Yj}, j = 1, \ldots, K.\]

**Note:**

1. Often, we make the assumption that the treatment-specific variances are equal; i.e. \( \sigma^2_{Y1} = \ldots = \sigma^2_{YK} = \sigma^2_Y \), but this assumption is not necessary for the subsequent development.

2. Moreover, it is also often assumed that the treatment-specific distribution of response is normally distributed with equal variances; i.e.

\[(Y_i|A_i = j) \sim N(\mu_j, \sigma^2_Y), j = 1, \ldots, K\]

Again, this assumption is not necessary for the subsequent development.
Our primary focus will be on testing the null hypothesis

\[ H_0 : \mu_1 = \ldots = \mu_K. \]

Let us redefine our data so that \((Y_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, K)\) denotes the response for the \(i\)-th individual within treatment \(j\), and \(n_j\) denotes the number of individuals in our sample assigned to treatment \(j\) \((n = \sum_{j=1}^{K} n_j)\). From standard theory we know that the treatment-specific sample mean

\[ \bar{Y}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j} \]

is an unbiased estimator for \(\mu_j\) and that asymptotically

\[ \bar{Y}_j \sim N(\mu_j, \frac{\sigma^2_{Y_j}}{n_j}), j = 1, \ldots, K. \]

**Remark:** If the \(Y\)'s are normally distributed, then the above result is exact. However, with the large sample sizes that are usually realized in phase III clinical trials, the asymptotic approximation is generally very good.

Also, we know that the treatment-specific sample variance

\[ s^2_{Yj} = \frac{\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{n_j - 1} \]

is an unbiased estimator for \(\sigma^2_{Y,j}\), and that asymptotically

\[ \bar{Y}_j \sim N(\mu_j, \frac{s^2_{Yj}}{n_j}), j = 1, \ldots, K. \]

**Remark:** If the treatment specific variances are all equal, then the common variance is often estimated using the pooled estimator

\[ s^2_Y = \frac{\sum_{j=1}^{K} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{n - K}. \]

Returning to the general results of section 7.1 of the notes, we let

- \(\bar{Y}_j\) take the role of \(\hat{\theta}_j\)
- \(\mu_j\) take the role of \(\theta_j\)
• \( \frac{s^2_j}{n_j} \) take the role of \( \sigma^2_j \); hence \( w_j = \frac{n_j}{s^2_j} \).

Using (7.1), we construct the test statistic

\[
T_n = \sum_{j=1}^{K} w_j (\bar{Y}_j - \bar{Y})^2,
\]

where

\[
\bar{Y} = \frac{\sum_{j=1}^{K} w_j \bar{Y}_j}{\sum_{j=1}^{K} w_j},
\]

and \( w_j = \frac{n_j}{s^2_j} \).

For the case where we are willing to assume equal variances, we get that

\[
T_n = \frac{\sum_{j=1}^{K} n_j (\bar{Y}_j - \bar{Y})^2}{s^2_Y},
\]

where

\[
\bar{Y} = \frac{\sum_{j=1}^{K} n_j \bar{Y}_j}{n}.
\]

Under the null hypothesis \( H_0 \), \( T_n \) is approximately distributed as a chi-square distribution with \( K - 1 \) degrees of freedom. Under the alternative hypothesis

\[
H_A : \mu_1 = \mu_{1A}, \ldots, \mu_K = \mu_{KA}
\]

where the \( \mu_{jA} \)’s are not all equal, \( T_n \) is approximately distributed as a non-central chi-square distribution with \( K - 1 \) degrees of freedom and non-centrality parameter

\[
\phi^2 = \sum_{j=1}^{K} w_j (\mu_{jA} - \bar{\mu}_A)^2,
\]

where

\[
\bar{\mu}_A = \frac{\sum_{j=1}^{K} w_j \mu_{jA}}{\sum_{j=1}^{K} w_j},
\]

and \( w_j = \frac{n_j}{s^2_j} \).

Assuming equal variances, we get the simplification

\[
\phi^2 = \frac{\sum_{j=1}^{K} n_j (\mu_{jA} - \bar{\mu}_A)^2}{\sigma_Y^2},
\]

where

\[
\bar{\mu}_A = \frac{\sum_{j=1}^{K} n_j \mu_{jA}}{n}.
\]
Remark: In the special case where the response data are exactly normally distributed with equal treatment-specific variances, the test statistic $T_n$ given by (7.10), under the null hypothesis, has a distribution which is exactly equal to $K - 1$ times a central $F$ distribution with $K - 1$ numerator degrees of freedom and $n - K$ denominator degrees of freedom. That is

$$T_n/(K - 1) = \frac{\sum_{j=1}^{K} n_j (\bar{Y}_j - \bar{\bar{Y}})^2 / s_{Y}^2}{(K - 1)}$$

has an $F$ distribution under the null hypothesis. This is exactly the test statistic that is used for testing the equality of means in a one-way ANOVA model.

However, when the sample size $n$ is large, we can use the test statistic given by (7.9), based on an asymptotic chi-square distribution, to test $H_0$ without making either the normality assumption or the assumption of equal variances.

### 7.5 Sample size computations for continuous response

Let us consider the case where patients are allocated equally to the $K$ treatments so that

$$n_1 = \ldots = n_K = n/K,$$

and, for design purposes, we assume that the treatment specific variances are all equal which we posit to be the value $\sigma_Y^2$. The question is how do we compute the sample size that is necessary to have power $(1 - \beta)$ to detect an alternative where any two treatments population mean responses may differ by $\Delta_A$ or more? Using considerations almost identical to those used for testing equality of response rates for dichotomous outcomes, we can find the least favorable configuration which after substituting into (7.12), yields the non-centrality parameter

$$\frac{n\Delta_A^2}{2K\sigma_Y^2}.$$  \hspace{1cm} (7.13)

Hence, to obtain the desired power of $(1 - \beta)$, we need

$$\frac{n\Delta_A^2}{2K\sigma_Y^2} = \phi^2(\alpha, \beta, K - 1),$$

or

$$n = \frac{2K\sigma_Y^2 \phi^2(\alpha, \beta, K - 1)}{\Delta_A^2}. \hspace{1cm} (7.14)$$

Example:
We expand on the example used for two-sample comparisons given on page 84 of the notes, but now we consider $K = 4$ treatments. What is the sample size necessary to detect a significant difference with 90% power or greater if any pairwise difference in mean treatment response is at least 20 units using the K-sample test above at the .05 level of significance? We posit that the standard deviation of response, assumed equal for all treatments, is $\sigma_Y = 60$ units. Substituting into formula (7.14), we get that

$$n = \frac{2 \times 4 \times (60)^2 \times 14.171}{(20)^2} \approx 1020,$$

or about $1021/4 = 255$ patients per treatment arm.

**Remark:** The 255 patients per arm represents an increase of 35% over the 189 patients per arm necessary in a two-sample comparison (see page 84 of notes). This percentage increase is the same as when we compare response rates for a dichotomous outcome with 4 treatments versus 2 treatments. This is not a coincidence, but rather, has to do with the relative ratio of the non-centrality parameters for a test with 3 degrees of freedom versus a test with 1 degree of freedom.

### 7.6 Equivalency Trials

The point of view we have taken thus far in the course is that of proving the superiority of one treatment over another. It may also be the case that there already exists treatments that have been shown to have benefit and work well. For example, a treatment may have been proven to be significantly better than placebo in a clinical trial and has been approved by the FDA and is currently on the market. However, there still may be room for other treatments to be developed that may be equally effective. This may be the case because the current treatment or treatments may have some undesirable side-effects, at least for some segment of the population, who would like to have an alternative. Or perhaps, the cost of the current treatments are high and some new treatments may be cheaper. In such cases, the company developing such a drug would like to demonstrate that their new product is equally effective to those already on the market or, at least, has beneficial effect compared to a placebo. The best way to prove that the new product has biological effect is to conduct a placebo-controlled trial and demonstrate superiority over the placebo using methods we have discussed. However, in the presence of
established treatments that have already been proven effective, such a clinical trial would be un-ethical. Consequently, the new treatment has to be compared to one that is already known to be effective. The comparison treatment is referred to as an active or positive control.

The purpose of such a clinical trial would not necessarily be to prove that the new drug is better than the positive control but, rather, that it is equivalent in some sense. Because treatment comparisons are based on estimates obtained from a sample of data and thus subject to variation, we can never be certain that two products are identically equivalent in their efficacy. Consequently, a new drug is deemed equivalent to a positive control if it can be proved with high probability that it has response at least within some tolerable limit of the positive control. Of course the tricky issue is to determine what might be considered a tolerable limit for purposes of equivalency. If the positive control was shown to have some increase in mean response compared to placebo, say $\Delta^*$, then one might declare a new drug equivalent to the positive control if it can be proved that the mean response of the new drug is within $\Delta^*/2$ of the mean response of the positive control or better with high probability. Conservatively, $\Delta^*$ may be chosen as the lower confidence limit derived from the clinical trial data that compared the positive control to placebo. Let us assume that the tolerable limit has been defined, usually, by some convention, or in negotiations of a company with the regulatory agency. Let us denote the tolerable limit by $\Delta_A$.

Remark: In superiority trials we denoted by $\Delta_A$, the clinically important difference that we wanted to detect with desired power. For equivalency trials, $\Delta_A$ refers to the tolerable limit.

Let us consider the problem where the primary response is a dichotomous outcome. (Identical arguments for continuous response outcomes can be derived analogously). Let $\pi_2$ denote the population response rate for the positive control, and $\pi_1$ be the population response rate for the new treatment.

Evaluating equivalency is generally stated as a one-sided hypothesis testing problem; namely,

$$H_0 : \pi_1 \leq \pi_2 - \Delta_A \text{ versus } H_A : \pi_1 > \pi_2 - \Delta_A.$$  

If we denote by the parameter $\Delta$ the treatment difference $\pi_1 - \pi_2$, then the null and alternative hypotheses are

$$H_0 : \Delta \leq -\Delta_A \text{ versus } H_A : \Delta > -\Delta_A.$$  

The null hypothesis corresponds to the new treatment being inferior to the positive control. This
is tested against the alternative hypothesis that the new treatment is at least equivalent to the positive control. As always, we need to construct a test statistic, $T_n$, which, when large, would provide evidence against the null hypothesis and whose distribution at the border between the null and alternative hypotheses (i.e. when $\pi_1 = \pi_2 - \Delta_A$) is known. Letting $p_1$ and $p_2$ denote the sample proportion that respond on treatments 1 and 2 respectively, an obvious test statistic to test $H_0$ versus $H_A$ is

$$T_n = \frac{p_1 - p_2 + \Delta_A}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}.$$ 

where $n_1$ and $n_2$ denote the number of patients allocated to treatments 1 and 2 respectively.

This test statistic was constructed so that at the border of the null and alternative hypotheses; i.e. when $\pi_1 = \pi_2 - \Delta_A$, the distribution of $T_n$ will be approximately a standard normal; that is

$$T_n (\pi_1 = \pi_2 - \Delta_A) \sim N(0, 1).$$

Clearly, larger values of $T_n$ give increasing evidence that the null hypothesis is not true in favor of the alternative hypothesis. Thus, for a level $\alpha$ test, we reject when

$$T_n \geq Z_\alpha.$$ 

With this strategy, one is guaranteed with high probability ($\geq 1 - \alpha$) that the drug will not be approved if, in truth, it is not at least equivalent to the positive control.

**Remark:** Notice that we didn’t use the arcsin square-root transformation for this problem. This is because the arcsin square-root is a non-linear transformation; thus, a fixed difference of $\Delta_A$ in response probabilities between two treatments (hypothesis of interest) does not correspond to a fixed difference on the arcsin square-root scale.

**Sample size calculations for equivalency trials**

In computing sample sizes for equivalency trials, one usually considers the power, i.e the probability of declaring equivalency, if, in truth, $\pi_1 = \pi_2$. That is, if, in truth, the new treatment has a response rate that is as good or better than the positive control, then we want to declare equivalency with high probability, say $(1 - \beta)$. To evaluate the power of this test to detect the alternative ($\pi_1 = \pi_2$), we need to know the distribution of $T_n$ when $\pi_1 = \pi_2$. 
Because

\[ T_n = \frac{(p_1 - p_2 + \Delta_A)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx \frac{p_1 - p_2 + \Delta_A}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}, \]

straightforward calculations can be used to show that

\[ E(T_n) \approx \frac{\Delta_A}{\sqrt{\pi(1 - \pi)} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, \]

and

\[ \text{var}(T_n) \approx 1. \]

Hence

\[ T_n \sim N \left( \frac{\Delta_A}{\sqrt{\pi(1 - \pi)} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, 1 \right), \]

and the non-centrality parameter equals

\[ \phi(\cdot) = \frac{\Delta_A}{\sqrt{\pi(1 - \pi)} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}. \]

If \( n_1 = n_2 = n/2 \), then

\[ \phi(\cdot) = \frac{\Delta_A}{\sqrt{\pi(1 - \pi)} \left( \frac{4}{n} \right)}. \]

To get the desired power, we solve

\[ \frac{\Delta_A}{\sqrt{\pi(1 - \pi)} \left( \frac{4}{n} \right)} = Z_\alpha + Z_\beta \]

or

\[ n = \frac{(Z_\alpha + Z_\beta)^2 \times 4\pi(1 - \pi)}{\Delta_A^2}. \] (7.15)

Generally, it requires larger sample sizes to establish equivalency because the tolerable limit \( \Delta_A \) that the regulatory agency will agree to is small. For example, a pharmaceutical company has developed a new drug that they believe has similar effects to drugs already approved and decides to conduct an equivalency trial to get approval from the FDA to market the new drug. Suppose the clinical trial that was used to demonstrate that the positive control was significantly better than placebo had a 95% confidence interval for \( \Delta \) (treatment difference) that ranged from .10-.25. Conservatively, one can only be relatively confident that this new treatment has a response rate that exceeds the response rate of placebo by .10. Therefore the FDA will only allow a new
treatment to be declared equivalent to the positive control if the company can show that their new drug has a response rate that is no worse than the response rate of the positive control minus .05. Thus they require a randomized two arm equivalency trial to compare the new drug to the positive control with a type I error of $\alpha = .05$. The response rate of the positive control is about .30. (This estimate will be used for planning purposes). The company believes their drug is similar but probably not much better than the positive control. Thus, they want to have good power, say 90%, that they will be successful (i.e. be able to declare equivalency by rejecting $H_0$) if, indeed, their drug was equally efficacious. Thus they use formula (7.15) to derive the sample size

$$n = \frac{(1.64 + 1.28)^2 \times 4 \times .3 \times .7}{(.05)^2} = 2864,$$

or 1432 patients per treatment arm.