Multiple unpaired numerical samples

Suppose we have multiple unpaired samples:

\[
\begin{array}{cccc}
\text{group 1} & \text{group 2} & \cdots & \text{group J} \\
Y_{1,1} & Y_{1,2} & \cdots & Y_{1,J} \\
Y_{2,1} & Y_{2,2} & \cdots & Y_{2,J} \\
\vdots & \vdots & \cdots & \vdots \\
Y_{n_1,1} & Y_{n_2,2} & \cdots & Y_{n_J,J}
\end{array}
\]

The \(i\)th observation from group \(j\) is denoted \(Y_{ij}\).

There are \(n_j\) observations in group \(j\), so that \(N = n_1 + \cdots + n_J\) is the total sample size.

An experiment where the number of observations per group is the same is said to have a balanced design.

Summary statistics: summarize each sample independently of the others.


Example: Smoking and Heart Rate

We consider the smokers data (Larsen & Marx, case study 12.2.1).

<table>
<thead>
<tr>
<th>non</th>
<th>light</th>
<th>moderate</th>
<th>heavy</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>55</td>
<td>66</td>
<td>91</td>
</tr>
<tr>
<td>52</td>
<td>60</td>
<td>81</td>
<td>72</td>
</tr>
<tr>
<td>71</td>
<td>78</td>
<td>70</td>
<td>81</td>
</tr>
<tr>
<td>58</td>
<td>58</td>
<td>77</td>
<td>67</td>
</tr>
<tr>
<td>59</td>
<td>62</td>
<td>57</td>
<td>95</td>
</tr>
<tr>
<td>65</td>
<td>66</td>
<td>79</td>
<td>84</td>
</tr>
</tbody>
</table>

Non-smokers, light smokers, moderate smokers and heavy smokers (six in each group) undertook sustained physical exercise. Their heart rates were measured after resting for three minutes.

(The observations are not paired.)

Main question: Is smoking associated with a decrease in fitness level?
Testing for Equality of Means

Let $\mu_j$ be the population mean for group $j$. Consider testing

$$H_0 : \mu_1 = \cdots = \mu_J$$

versus

$$H_1 : \text{not all } \mu_j \text{'s are equal}$$

One-Way Analysis of Variance (ANOVA)

Let $\bar{Y}_{.j}$ and $S_j^2$ denote the $j$th sample mean and variance:

$$\bar{Y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}, \quad S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2$$

Let $\bar{Y}_{..}$ denote the (all-groups-combined) grand mean:

$$\bar{Y}_{..} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij} = \frac{\sum_{j=1}^{J} n_j \bar{Y}_{.j}}{N}$$

One-Way Analysis of Variance (ANOVA)

Compute the between groups sum of squares (aka treatment sum of squares)

$$SS_T = \sum_{j=1}^{J} n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

Compute the within groups sum of squares (aka error sum of squares)

$$SSE = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 = \sum_{j=1}^{J} (n_j - 1)S_j^2$$

The $F$-test rejects for large values of

$$F = \frac{SS_T/(J - 1)}{SSE/(N - J)}$$

Theory. If the observations are iid normal, then $F$ has the so-called F-distribution with $J - 1$ and $N - J$ degrees of freedom.
### One-way ANOVA table

Traditionally, the computations above are gathered in an ANOVA table:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>SS</th>
<th>MS = SS/Df</th>
<th>F-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>J - 1</td>
<td>SSₜ</td>
<td>SSₜ/J</td>
<td>F = MSₜ/MSE</td>
<td>P(F₍J₋₁,N₋J₎ &gt; F)</td>
</tr>
<tr>
<td>Residuals</td>
<td>N - J</td>
<td>SSE</td>
<td>SSE/(N - J)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sometimes, an ANOVA table also includes the total sum of squares:

\[ SS_{\text{Tot}} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 \]

Note that \( SS_{\text{Tot}} = SS_{\text{T}} + SSE \).

### One-Way ANOVA model assumptions

The p-value relies on the assumption that, at least approximately,

\[ Y_{ij} \sim N(\mu_j, \sigma^2) \]

and that these are independent.

This is often expressed as

\[ Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \]

where

- \( \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \) are the measurement errors.
- \( \mu = \frac{1}{N} \sum_{j=1}^{J} n_j \mu_j \) is the grand mean.
- \( \alpha_j = \mu_j - \mu \) is the \( j \)th treatment effect. Note that \( \sum_{j=1}^{J} n_j \alpha_j = 0 \).
- Testing \( \mu_1 = \cdots = \mu_J \) is equivalent to testing \( \alpha_1 = \cdots = \alpha_J = 0 \).
Checking Assumptions

□ We need to check that the $\varepsilon_{ij}$’s satisfy the following properties:

1. They have the same variance (homoscedasticity).
2. They are approximately normally distributed (important if the sample sizes are small).
3. They are independent of each other.

□ The errors are not available to us. As a substitute we use the residuals:

$$e_{ij} = Y_{ij} - \overline{Y}_j$$

1. We use a residual plot to check for homoscedasticity.
2. We use a Q-Q plot of the residuals to check for (approximate) normality.
3. Independence is usually assessed based on how the data was collected.

While the assumption of normality is "cured" in the large-sample limit (because of the CLT), the assumption of homoscedasticity is not. However there is a variant (the analog of the Welch’s t-test) that does not require homoscedasticity.

Tukey’s Honest Significant Differences

□ Although the $F$-test is significant, in the boxplots it appears that there is not much difference between heart rates for non smokers and light smokers

We might want to know which groups have different population means.

□ Tukey’s honest significant differences provides a method to perform all (or some selected tests) with an overall level of 5%.

The level is exact under the same assumptions as in ANOVA.

□ In our example, only the differences between heavy and light (or non) smokers are statistically significant.

This may seem paradoxical, as we accept

$$\mu_{\text{light}} = \mu_{\text{moderate}}, \quad \text{and} \quad \mu_{\text{moderate}} = \mu_{\text{heavy}}$$

while rejecting

$$\mu_{\text{light}} = \mu_{\text{heavy}}$$
Calibration by bootstrap

- The one-way ANOVA F-test and Tukey’s HSD both assume (approximate) normality and equal variances. (Although the Welch ANOVA F-test does not assume equal variances.) However, just as in the two-sample situation, a bootstrap calibration is possible. For the F-test, it works as follows:

1. Start by centering each group by removing its sample mean.
   - This is to place ourselves under the null where the means are all equal.

2. Let $B$ be a large integer. For $b = 1, \ldots, B$ do the following:
   - (a) For each (centered) group $j = 1, \ldots, J$, sample $n_j$ observations with replacement.
   - (b) Compute the corresponding F-ratio $F_b$.

3. The p-value is
   $$\frac{\#\{b : F_b \geq F_{\text{obs}}\} + 1}{B + 1}$$
   (roughly) the fraction of bootstrapped statistics ($F_b : b = 1, \ldots, B$) that are at least as large as the observed statistic $F_{\text{obs}}$.

Permutation test

- Just as in the two-sample situation, a calibration by permutation is appropriate if we are interested in comparing the group distributions (goodness-of-fit testing) instead of merely comparing the means. This leads to testing

$$H_0 : \text{all } J \text{ samples have the same distribution, meaning, } (Y_{ij} : i = 1, \ldots, n_j; j = 1, \ldots, J) \text{ are iid}$$

versus

$$H_1 : \text{this is not the case}$$

- Let $Z_1, \ldots, Z_n$ be the concatenated sample $(Y_{ij} : i = 1, \ldots, n_j; j = 1, \ldots, J)$. Under the null, the $Z_\ell$’s are exchangeable because the $Y_{ij}$’s are iid.

- A permutation test based on the treatment sum of squares works as follows:

1. For each permutation $\pi$ of $\{1, \ldots, N\}$, let
   $$Y_{ij}^\pi = Z_{\pi(i)} \text{ for } i = n_{j-1} + 1, \ldots, n_j \text{ (where } n_0 = 0)$$
   and compute
   $$SS_T^\pi = \sum_{j=1}^J n_j (\bar{Y}_{(j)}^\pi - \bar{Y}_\cdot)^2$$

2. The (exact) $p$-value is the fraction of $SS_T^\pi$ that are greater or equal to the observed $SS_T^{\text{obs}}$. 
Even for small \( n \), computing the exact p-value as done above may not be feasible. Indeed, assuming all the \( Y_{ij} \)'s are distinct, there are
\[
(n_1, \ldots, n_J)! = \frac{(n_1 + \cdots + n_J)!}{n_1! \cdots n_J!}
\]
non-equivalent permutations. This number is very large even for small sample sizes.

When this is the case, we estimate the p-value by Monte Carlo sampling. We sample \( B \) permutations (\( B \) is a large integer) uniformly at random.

**Kruskal-Wallis Test**

The *Kruskal-Wallis test* is a rank-based method, and as such, is a form of permutation test. The null hypothesis is the same as with any permutation test, that the \( J \) samples come from the same distribution (i.e., same population).

Let \( R_{ij} \) denote the overall rank of \( Y_{ij} \) and define the rank-sum for group \( j \)
\[
R_j = \sum_{i=1}^{n_j} R_{ij}
\]

The test rejects for large values of
\[
D = \frac{12}{N(N+1)} \sum_{j=1}^{J} \frac{R_j^2}{n_j} - 3(N+1)
\]

If the design is balanced, meaning \( n_1 = \cdots = n_J \), this is equivalent to rejecting for large values of
\[
\sum_{j=1}^{J} R_j^2
\]
To compare it with one-way ANOVA, note that

\[ D = (N - 1) \frac{\sum_{j=1}^{J} n_j \left( \frac{R_{..j}}{n_j} - \frac{R_{..}}{N} \right)^2}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \left( \frac{R_{ij}}{n_j} - \frac{R_{..}}{N} \right)^2} \]

So the total sum of squares is used in the denominator as opposed to the error sum of squares as in regular one-way ANOVA.

As with any rank-based method, the KW test is distribution-free (important for tabulation in the pre-personal computer age) and invariant with respect to strictly increasing transformations for the data.

Theory. Under the null, \( D \) has asymptotically (all the sample sizes tend to infinity in proportion to each other) a chi-squared distribution with \( J - 1 \) degrees of freedom.

Repeated measures: multiple paired numerical samples

Suppose now that we have multiple paired samples. This is often referred to as repeated measures, and presented as follows:

<table>
<thead>
<tr>
<th>subject 1</th>
<th>treatment 1</th>
<th>treatment 2</th>
<th>\ldots</th>
<th>treatment J</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject 2</td>
<td>( Y_{1,1} )</td>
<td>( Y_{1,2} )</td>
<td>\ldots</td>
<td>( Y_{1,J} )</td>
</tr>
<tr>
<td>subject 2</td>
<td>( Y_{2,1} )</td>
<td>( Y_{2,2} )</td>
<td>\ldots</td>
<td>( Y_{2,J} )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
</tr>
<tr>
<td>subject 1</td>
<td>( Y_{I,1} )</td>
<td>( Y_{I,2} )</td>
<td>\ldots</td>
<td>( Y_{I,J} )</td>
</tr>
</tbody>
</table>

There are \( I \) subjects (or experimental units) in the study and each subject receives all \( J \) treatments. This often occurs over a period of time.

This typically leads to a two-way analysis of variance, where the subjects are the blocks. (This is covered later in the slides.)
Friedman test

- The **Friedman test** is a rank-based procedure for repeated measures designs, to test

  \[ H_0 : \text{for all } i = 1, \ldots, I, \ (Y_{i,1}, \ldots, Y_{i,J}) \text{ are exchangeable} \]

  with the alternative \( H_1 \) being the negation of \( H_0 \).

- Let \( R_{ij} \) be the rank of \( Y_{i,j} \) among \( (Y_{i,1}, \ldots, Y_{i,J}) \).

  The treatment responses are compared within each subject.

- Define the sum of the ranks for treatment \( j \):

  \[ R_j = \sum_{i=1}^{I} R_{ij} \]

- The test rejects for large values of

  \[ G = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} R_j^2 - 3I(J+1) \]

- **Theory.** Under the null, \( G \) has asymptotically (as \( I \to \infty \)) the chi-squared distribution with \( J - 1 \) degrees of freedom.

Two-way designs

- Suppose we have a two-way design with two factors:

  - **Treatment:** \( I \) levels, indexed by \( i \).
  - **Blocking:** \( J \) blocks, indexed by \( j \).

  These define \( I \times J \) groups, also called cells, indexed by pairs \( (i, j) : i = 1, \ldots, I; j = 1, \ldots, J \).

- Let \( Y_{i,j,k} \) denote the \( k \)th observation in cell \( (i, j) \).

- For simplicity, we assume the design is balanced, with the same number \( n \) of observations per group.

- The observations in each cell, \( (Y_{i,j,k} : k = 1, \ldots, n) \), are assumed iid.

- **Summary statistics:** for each group independently.

- **Graphics:** side-by-side boxplots and interactions plot.
Pictorially, the data “looks” as follows: it is divided into groups indexed by the treatment and the blocking variables.

<table>
<thead>
<tr>
<th>Block 1</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>…</th>
<th>Treatment k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y_{1,1,1}, \ldots, Y_{1,1,n}$</td>
<td>$Y_{2,1,1}, \ldots, Y_{2,1,n}$</td>
<td>…</td>
<td>$Y_{k,1,1}, \ldots, Y_{k,1,n}$</td>
</tr>
<tr>
<td>Block 2</td>
<td>$Y_{1,2,1}, \ldots, Y_{1,2,n}$</td>
<td>$Y_{2,2,1}, \ldots, Y_{2,2,n}$</td>
<td>…</td>
<td>$Y_{k,2,1}, \ldots, Y_{k,2,n}$</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Block b</td>
<td>$Y_{1,b,1}, \ldots, Y_{1,b,n}$</td>
<td>$Y_{2,b,1}, \ldots, Y_{2,b,n}$</td>
<td>…</td>
<td>$Y_{k,b,1}, \ldots, Y_{k,b,n}$</td>
</tr>
</tbody>
</table>

### Example: orange juice vs ascorbic acid in tooth growth

- We consider the ToothGrowth data, already loaded in R.
  
  - The response is the length of odontoblasts (continuous).
  - Supplement type: $VC =$ ascorbic acid or $OJ =$ orange juice (categorical).
  - Dose is either 0.5, 1 or 2 mg (discrete).

- **Questions:**
  1. Does the delivery method have any effect on tooth growth when *controlling* for dosage?
  2. Is there any *interaction* between delivery method and dosage in their effect on tooth growth?

### Population means

- Let $\mu_{ij}$ denote the population mean of group $(i,j)$.

- Define

  $$\mu_. = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}, \quad \mu_i. = \frac{1}{J} \sum_{j=1}^{J} \mu_{ij}, \quad \mu.j = \frac{1}{I} \sum_{i=1}^{I} \mu_{ij},$$

- We have the decomposition:

  $$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

  - **Grand mean**: $\mu$ (same as $\mu_.$).
  - **Treatment effects**: $\alpha_i = \mu_i. - \mu$
  - **Blocking effects**: $\beta_j = \mu.j - \mu$
  - **Interactions terms**: $\gamma_{ij} = \mu_{ij} - \mu_i. - \mu.j + \mu$
Note that:
\[
\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \sum_{i=1}^{I} \gamma_{ij} = 0, \ \forall j, \quad \sum_{j=1}^{J} \gamma_{ij} = 0, \ \forall i
\]

Hypotheses about the means

- Testing for a treatment effect
  \[H_T : \alpha_i = \mu_i - \mu_\cdot = 0, \ \forall i\]
  This is often the main testing problem considered.

- Testing for a blocking effect
  \[H_B : \beta_j = \mu_j - \mu_\cdot = 0, \ \forall j\]
  This is often less important since the blocking variable is usually known to affect the response variable.

- Testing for interactions
  \[H_{TB} : \gamma_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu_\cdot = 0, \ \forall i, j\]

Sample means

- Define
  \[
  \bar{Y}_{ij} = \frac{1}{n} \sum_{k=1}^{n} Y_{ijk}, \quad \bar{Y}_{i..} = \frac{1}{nJ} \sum_{j=1}^{J} \sum_{k=1}^{n} Y_{ijk}, \quad \bar{Y}_{..j} = \frac{1}{nI} \sum_{i=1}^{I} \sum_{k=1}^{n} Y_{ijk}
  \]
  \[
  \bar{Y}_{...} = \frac{1}{nJI} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} Y_{ijk}
  \]
  - \(\bar{Y}_{ij}\) is the method of moments estimate (MME) for \(\mu_{ij}\)
  - \(\bar{Y}_{i..}\) is the MME estimate for \(\mu_i\)
  - \(\bar{Y}_{..j}\) is the MME estimate for \(\mu_j\)
  - \(\bar{Y}_{...}\) is the MME estimate for \(\mu_\cdot\)
### Sums of squares

**The treatment sum of squares**

\[ SS_T = nJ \sum_{i=1}^{I} (Y_{i..} - \bar{Y}_{...})^2 \]

**The blocks sum of squares**

\[ SS_B = nI \sum_{j=1}^{J} (Y_{.j.} - \bar{Y}_{...})^2 \]

**The interactions sum of squares**

\[ SS_{TB} = n \sum_{j=1}^{J} \sum_{i=1}^{I} (Y_{ij.} - \bar{Y}_{.j.} - \bar{Y}_{i..} + \bar{Y}_{...})^2 \]

**The error sum of squares**

\[ SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (Y_{ijk} - \bar{Y}_{ij.})^2 \]

**Assume equal variance**

\[ \text{Var}(Y_{ijk}) = \sigma^2, \quad \forall i, j, k \]

**We have:**

<table>
<thead>
<tr>
<th>SS</th>
<th>df</th>
<th>( \mathbb{E}(SS/df) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( SS_T )</td>
<td>( I - 1 )</td>
<td>( \sigma^2 + \frac{nJ}{I-1} \sum_i \alpha_i^2 )</td>
</tr>
<tr>
<td>( SS_B )</td>
<td>( J - 1 )</td>
<td>( \sigma^2 + \frac{nI}{J-1} \sum_j \beta_j^2 )</td>
</tr>
<tr>
<td>( SS_{TB} )</td>
<td>( (I - 1)(J - 1) )</td>
<td>( \sigma^2 + \frac{n}{(I-1)(J-1)} \sum_{ij} \gamma_{ij}^2 )</td>
</tr>
<tr>
<td>SSE</td>
<td>( (n - 1)JI )</td>
<td>( \sigma^2 )</td>
</tr>
</tbody>
</table>
**F-tests**

- For $H_T$, reject for large
  
  \[
  F_T = \frac{SS_T/(I-1)}{SSE/((n-1)JI)}
  \]

- For $H_B$, reject for large
  
  \[
  F_B = \frac{SS_B/(J-1)}{SSE/((n-1)JI)}
  \]

- For $H_{TB}$, reject for large
  
  \[
  F_{TB} = \frac{SS_{TB}/(I-1)(J-1)}{SSE/((n-1)JI)}
  \]

---

**Theory**. Assume the $Y_{ijk}$'s are independent normal with equal variance.

- Under $H_T$, $F_T$ has the F-distribution with $I-1$ and $(n-1)JI$ degrees of freedom.
- Under $H_B$, $F_B$ has the F-distribution with $J-1$ and $(n-1)JI$ degrees of freedom.
- Under $H_{TB}$, $F_{TB}$ has the F-distribution with $(I-1)(J-1)$ and $(n-1)JI$ degrees of freedom.

---

**Two-way ANOVA table**

- Traditionally, the computations above are gathered in an ANOVA table:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>SS</th>
<th>MS = SS/df</th>
<th>F-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>$I-1$</td>
<td>SS$_T$</td>
<td>$\frac{SS_T}{I-1}$</td>
<td>$F_T = \frac{MS_T}{MS_E}$</td>
<td>$P(F_{df_T,df_E} &gt; F_T)$</td>
</tr>
<tr>
<td>Blocking</td>
<td>$J-1$</td>
<td>SS$_B$</td>
<td>$\frac{SS_B}{J-1}$</td>
<td>$F_B = \frac{MS_B}{MS_E}$</td>
<td>$P(F_{df_B,df_E} &gt; F_B)$</td>
</tr>
<tr>
<td>Interactions</td>
<td>$(I-1)(J-1)$</td>
<td>SS$_{TB}$</td>
<td>$\frac{SS_{TB}}{(I-1)(J-1)}$</td>
<td>$F_{TB} = \frac{MS_{TB}}{MS_E}$</td>
<td>$P(F_{df_{TB},df_E} &gt; F_{TB})$</td>
</tr>
<tr>
<td>Residuals</td>
<td>$(n-1)JI$</td>
<td>SSE</td>
<td>$\frac{SSE}{(n-1)JI}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Sometimes, an ANOVA table also includes the total sum of squares:

\[
SS_{Tot} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (Y_{ijk} - \bar{Y})^2
\]

Note that

\[
SS_{Tot} = SS_T + SS_B + SS_{TB} + SSE
\]

- NOTE: sometimes the table is different and the tests for treatment and blocking are performed as if it were a one-way ANOVA.
Two-Way ANOVA model assumptions and diagnostics

- The p-value relies on the assumption that $Y_{ijk} \iid \mathcal{N}(\mu_{ij}, \sigma^2)$. This is equivalent to

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where the $\varepsilon_{ijk}$'s are the measurement errors.

- Checking the assumptions on the $\varepsilon_{ij}$'s is done as before based on the residuals:

$$e_{ijk} = Y_{ijk} - \bar{Y}_{ij}.$$ 

- If we are fitting a model without interactions, i.e.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

then we need to check that the model is appropriate. This can be done via an interactions plot, or a residual plot.

Permutation test

- A calibration by permutation is also possible for the null where the groups within each block come from the same population.

  This allows for the blocking variable to have an effect on the response, which it typically does. But it does not allow for interactions.

- Therefore, consider testing

$$H_0 : \text{the observations within each block are iid}$$

meaning, for each $j = 1, \ldots, J$, $(Y_{ijk} : i = 1, \ldots, I; k = 1, \ldots, n)$ are iid

versus

$$H_1 : \text{this is not the case}$$

- Choose a test statistic, for example, consider the treatment sum of squares. Permute the observations within each block, independently of the other blocks. This is done many times (say, $B = 2,000$ times) and the statistic is computed every time. The p-value is computed as usual.