

## Multiple comparison

Scheffé's test.....	1
Dunnett's .....	6
Bonferroni procedure .....	13
Overall comment .....	16

## Scheffé's test

Q: what is Scheffé's test

A:

Scheffé's test is a statistical method used to perform **post-hoc analysis**(students may not familiar with this term, which can be discourage to read when they see terms unknown in the beginning of the text) following analysis of variance (ANOVA). It is specifically designed to conduct pairwise comparisons between group means or to make more general contrasts, particularly when the number of comparisons is large(How large? This is not specified in the following context) and when the researcher wants to control the overall Type I error rate (the probability of incorrectly rejecting the null hypothesis).

Here's a summary of key features of Scheffé's test:

1. **Flexibility:** It is more flexible than other post-hoc tests, such as Tukey's or Bonferroni's, because it allows for not only pairwise comparisons but also complex contrasts (comparisons of **linear combinations of group means**).
2. **Conservative:** Scheffé's test is considered conservative, meaning it has a lower probability of committing Type I errors, but this can lead to a higher likelihood of Type II errors (failing to detect a true effect).
3. **Calculation:** The test calculates a statistic based on the ratio of between-group variance to within-group variance. This ratio is then compared to a critical value determined by the number of groups and the desired significance level.

4. **Post-hoc Use:** Scheffé's test is used after a significant ANOVA result to explore which specific means are different from each other or to evaluate more complex hypotheses about the group means.

To conduct Scheffé's test, the researcher first runs an ANOVA to determine if there are any significant differences among the group means. If the ANOVA indicates significance, Scheffé's test is applied to identify which specific groups differ and to examine any more general contrasts between them.

**Q: how to do a Scheffé's test**

**A:** To perform a **Scheffé's test** after conducting an **ANOVA** (Analysis of Variance), you follow a series of steps. Here's a clear guide on how to do it:

### 1. Conduct an ANOVA Test

- **Null Hypothesis ( $H_0$ ):** All group means are equal.
- **Alternative Hypothesis ( $H_1$ ):** At least one group mean is different from the others.

First, you need to run an ANOVA (usually one-way ANOVA) to determine whether there are significant differences among the group means. If the ANOVA result is significant (i.e.,  $p\text{-value} < \alpha$ , typically ("α is", two terms are mentioned, it may be confusing if not clearly stated) 0.05), you can proceed with the Scheffé's test.

### 2. Formulate Comparisons

- Scheffé's test allows for comparing all pairwise group differences, but also for more general contrasts (combinations of different groups, not necessarily just comparing two at a time).
- Define which comparisons or contrasts you are interested in. For example:
  - Pairwise comparisons of means.
  - More complex contrasts like comparing the average of several groups against a single group, or comparing the differences between several specific group pairs.

(these two bullet points seems redundant)

### 3. Calculate the Scheffé's Test Statistic

For each contrast or comparison, the test statistic is calculated using the formula:

$$F = \frac{\sum \left( \frac{(M_{\text{contrast}} - M_{\text{overall}})^2}{s_{\text{pooled}}^2} \right)}{k}$$

Where:

- $M_{\text{contrast}}$  is the difference between the group means or the contrast of interest.
- $M_{\text{overall}}$  is the overall mean of all groups.
- $s_{\text{pooled}}^2$  is the pooled estimate of variance.
- $k$  is the number of groups.

In practice, this calculation requires knowing the variance within each group and the number of observations per group.

- (1. It did not provide a formula for pooled variances.
- (2. Did not specify F is the F statistic, or why is these formula would follow f distribution.
- (3. This is a general formula, which could be confusing when students first see this. It may be broke down formula for Pairwise comparisons or More complex contrasts.)

#### 4. Determine the Critical Value

- Once you have the F-statistic(did not mention what is the F statistic previously) for each comparison or contrast, determine the critical value for the F-distribution.
- The critical value is based on the number of groups g(is it should be k which was referred previously in the formula, k-1), the total number of observations(not listed in the formula but should be the N for pooled variance, N-k), and the desired significance level (usually  $\alpha=0.05$ ).  
(based on how? Not clear on how to find the F critical value)
- You can find the critical F-value from statistical tables or using software (e.g., R, SPSS, or Python).

#### 5. Compare the Test Statistic to the Critical Value

- If the calculated F-statistic for a comparison or contrast is **greater than** the critical value from the F-distribution, you **reject the null hypothesis** for that comparison (indicating a significant difference).
- If the F-statistic is **less than** the critical value, you **fail to reject** the null hypothesis for that comparison (indicating no significant difference).

## 6. Interpret the Results

- Identify which comparisons or contrasts showed significant differences. For example, you may find that the mean of group A differs from group B, or that the combination of group C and D differs from group E.
- Scheffé's test will tell you whether any specific differences between groups (or combinations of groups) are statistically significant while controlling for Type I error.

### Example:

Let's say you have 3 groups: A, B, and C. After running ANOVA, you find a significant result ( $p\text{-value} < (\text{"alpha=" make sure student understand where does 0.05 come from})0.05$ ), and you wish to perform pairwise comparisons.

- **Group A vs. Group B:** Calculate the difference in their means, apply the Scheffé formula to get the test statistic, and compare to the critical value.
- **Group A vs. Group C:** Similarly, calculate the F-statistic and compare it to the critical value.
- **Group B vs. Group C:** Do the same for the remaining comparison.

If any of the F-statistics exceed the critical value, those group comparisons are significantly different.

## 7. Software Implementation

Performing Scheffé's test by hand can be tedious, but statistical software packages can do it quickly:

- **In R:** You can use the `aov()` function for ANOVA and the `TukeyHSD()` function for post-hoc tests. While `TukeyHSD()` is commonly used for pairwise comparisons, you can use the `glht()` function from the `multcomp` package to conduct Scheffé's test specifically.
- **In SPSS:** After running ANOVA, choose "Post-hoc tests" and select "Scheffé" from the options for post-hoc comparisons.

- **In Python (using statsmodels):** You can use statsmodels for ANOVA and then apply the Scheffé test using custom contrast calculations, though Python doesn't have a built-in function for Scheffé's test.

**Key Points:**

- Scheffé's test is used after ANOVA to perform post-hoc comparisons.
- It controls for Type I error across multiple comparisons.
- It's more flexible than some other post-hoc tests but can be conservative.

(Did not address the important question on "How Scheffe's test control the significance level?" to answer the potential question that "how is Scheff's test any different from other test")

## Dunnett's

Q: what is Dunnett's test

A:

Dunnett's test is a statistical method used to compare multiple treatment groups to a single control group, while controlling for the Type I error rate (false positives). It is commonly applied in experimental research when there are several experimental conditions (e.g., different doses or treatments) and a single control group.

### Key Features of Dunnett's Test:

- **Purpose:** To test whether each treatment group is significantly different from the control group, without comparing the treatment groups to each other.
- **Type of Analysis:** It is a **multiple comparison** procedure that controls the family-wise error rate (FWER). This is important because conducting multiple pairwise comparisons increases the risk of false positives.
- **Control Group:** The primary focus of Dunnett's test is the comparison between each treatment group and a reference or control group (usually the baseline or no-treatment group). Treatment groups are not directly compared to each other.
- **Statistical Assumptions:** Assumes that the data are normally distributed and that the variances across groups are roughly equal.

### How it Works:

1. **Null Hypothesis:** The null hypothesis for each comparison is that the mean of the treatment group is equal to the mean of the control group.
2. **Adjustments:** The test adjusts for the fact that multiple comparisons are being made, reducing the likelihood of finding a significant result just by chance.
3. **Results:** If the p-value for a given treatment group is less than the significance level (usually 0.05), the null hypothesis is rejected, indicating that the treatment has a significant effect compared to the control.

### Why Use Dunnett's Test?

- When you have many treatment groups and a single control, **Dunnett's test** is preferred over multiple t-tests, because it minimizes the risk of Type I errors.
- It is particularly useful in **pharmacology, biology, and clinical research**, where the goal is often to assess the effect of various treatments relative to a standard or control.

### Example:

If you are testing the effect of four different drugs (Drug A, Drug B, Drug C, Drug D) on blood pressure and you have a control group (no drug), Dunnett's test would allow you to compare each of the drugs individually against the control group. You would not compare Drug A directly to Drug B, for example, as the test focuses only on control vs treatment group comparisons.

Overall, Dunnett's test is a powerful tool for analyzing multiple treatment conditions in relation to a single baseline or control group while maintaining proper statistical rigor.

**Q: how to do a Dunnett's test**

**A:**

To perform **Dunnett's test**, follow these steps. This involves conducting pairwise comparisons between multiple treatment groups and a single control group, with adjustments to control for Type I error rate. Here's how to carry out Dunnett's test in a step-by-step manner, both manually (conceptually) and using software like R or SPSS.

### Step-by-Step Guide

#### 1. Set up your data

Ensure you have:

- A **control group** (reference group).
- Several **treatment groups** to compare against the control.
- A **continuous outcome variable** (e.g., measurements of weight, blood pressure, etc.).

For example:

- Control: No treatment
- Treatment 1: Drug A
- Treatment 2: Drug B
- Treatment 3: Drug C

## 2. Formulate hypotheses (Specify there are many sets of hypothesis)

- **Null hypothesis ( $H_0$ ):** Each treatment group mean is equal to the control group mean.
  - $H_0: \mu_{\text{control}} = \mu_{\text{treatment}}$
- **Alternative hypothesis ( $H_1$ ):** Each treatment group mean is different from the control group mean.
  - $H_1: \mu_{\text{control}} \neq \mu_{\text{treatment}}$

## 3. Conduct an ANOVA

Before performing Dunnett's test, you should first conduct an **Analysis of Variance (ANOVA)** to check if there is any significant overall difference between the groups (control and treatments). This tests the null hypothesis that all groups have the same mean.

- If the **ANOVA p-value** is **significant**, proceed with Dunnett's test.
- If the ANOVA p-value is **not significant**, there is no need for Dunnett's test since you would not reject the null hypothesis for the group means.

## 4. Choose a significance level ( $\alpha$ )

Typically, you use a significance level of **0.05** (5%).

## 5. Perform Dunnett's test

Dunnett's test is designed to compare each treatment group to the control group, adjusting the p-values for multiple comparisons.

Here's how you perform the test in different software.

~~~~~

(1. There is no theory of how to calculate the statistic by hand, so I asked another question.

2. Assumptions?)



Q: how to do step 5 in theory

A:

### Step 5 in Theory: Performing Dunnett's Test

In theory, **Dunnett's test** is a method used to compare several treatment groups against a **single control group** while adjusting for the fact that multiple comparisons are being made. It's a correction of the **t-test** for multiple comparisons, ensuring that the overall **Type I error rate** (the probability of incorrectly rejecting a true null hypothesis) is controlled.

Here's how you would perform **Dunnett's test** in theory, broken down into key steps:

---

#### 1. Assumptions (Preliminary Step)

Before conducting Dunnett's test, there are certain assumptions that need to be met:

- **Normality:** The data within each group should come from a normal distribution.
- **Homogeneity of variance:** The variance (spread of data) within each group should be approximately equal.

If these assumptions hold, you can proceed with the analysis.

(Are there any requirements for sample sizes? Equal/Unequal size.)

---

#### 2. Null Hypothesis ( $H_0$ ) and Alternative Hypothesis ( $H_1$ )

For each treatment group, you compare it to the control group. The hypotheses for each treatment group  $i$  would be:

- **Null hypothesis ( $H_0$ ):** The mean of the treatment group is equal to the mean of the control group:

$H_0: \mu_{\text{control}} = \mu_{\text{treatment}}$  for each treatment group

- **Alternative hypothesis ( $H_1$ ):** The mean of the treatment group is not equal to the mean of the control group:

$H_1: \mu_{\text{control}} \neq \mu_{\text{treatment}} \text{ for each treatment group}$

These hypotheses are tested for each treatment group relative to the control.

### 3. Conducting an Initial One-Way ANOVA

Before performing Dunnett's test, you need to determine if there are **any overall differences** among the groups (control + treatments). This is done with a **one-way ANOVA**.

The null hypothesis for the ANOVA is:

$H_0: \mu_{\text{control}} = \mu_{\text{treatment 1}} = \mu_{\text{treatment 2}} = \dots = \mu_{\text{treatment k}}$

If the p-value from the ANOVA is **significant ( $p < 0.05$ )**, this means that at least one group mean is different from the others, and you can proceed with Dunnett's test.

If the ANOVA p-value is **not significant**, then there is no evidence to suggest that the treatment groups differ from the control, and there is no need for Dunnett's test.

---

### 4. Performing Dunnett's Test (The Core Step)

Dunnett's test involves comparing each **treatment group** to the **control group** and adjusting the p-values to account for multiple comparisons. The steps in this procedure are as follows:

1. **Compute the difference between the means** of each treatment group and the control group:

$D_i = \mu_{\text{treatment } i} - \mu_{\text{control}}$  for each treatment group  $i$

2. **Calculate the standard error (SE)** of the difference between the means:

(Dunnett's test assumes the equal variance across groups. Use a pooled variance instead of separate group variances or MSE might be helpful)

$$SE_i = \sqrt{\frac{\text{Var}_{\text{control}}}{n_{\text{control}}} + \frac{\text{Var}_{\text{treatment}_i}}{n_{\text{treatment}_i}}}$$

Where:

- $\text{Var}_{\text{control}}$  is the variance of the control group.
- $\text{Var}_{\text{treatment}_i}$  is the variance of the treatment group.
- $n_{\text{control}}$  is the sample size of the control group.
- $n_{\text{treatment}_i}$  is the sample size of the treatment group.

3. Compute the t-statistic for each comparison:

$$t_i = \frac{D_i}{SE_i}$$

This t-statistic tests whether the difference between the treatment group and the control group is significantly different from zero.

3. **Adjust for multiple comparisons:** Since multiple comparisons increase the chance of a Type I error, Dunnett's test adjusts the significance level (p-value) for each test using a method that controls the family-wise error rate (FWER). (Define what is family-wise error rate may helps students to understand why we need to control it)

The **adjustment** is typically based on a correction factor that takes into account the number of comparisons you're making. In practice, this adjustment is done by comparing the computed t-statistic to a critical value from the **t-distribution**, using a correction for multiple comparisons.

(Need more details on how to find the critical value using Dunnett-critical value table, and point out that Dunnett-critical value table is different from t-distribution table )

The corrected p-values are then used to decide whether to reject the null hypothesis for each treatment group.

- If a treatment group's p-value is **less than the significance level (e.g., 0.05)**, you **reject the null hypothesis**, indicating a significant difference between that treatment and the control group.
  - If the p-value is **greater than the significance level**, you **fail to reject** the null hypothesis, indicating no significant difference between the treatment and the control group.
-

## 5. Conclusion

After performing Dunnett's test, you can conclude whether each treatment group differs significantly from the control group. Only the comparisons between the control and treatment groups are made, and the test adjusts the p-values to avoid inflating the error rate from making multiple comparisons.

For each treatment group, one of the following will occur:

- **Significant result:** The treatment significantly differs from the control (i.e., reject the null hypothesis for that group).
  - **Non-significant result:** The treatment does not significantly differ from the control (i.e., fail to reject the null hypothesis for that group).
- 

### Summary of the Theoretical Steps:

1. **Conduct an ANOVA** to check if there are overall differences among the groups.
  2. For each treatment group:
    - Compute the difference between its mean and the control group mean.
    - Compute the standard error of the difference.
    - Compute the t-statistic.
    - Compare the t-statistic to the critical value from the t-distribution, adjusting for multiple comparisons.
  3. **Interpret the results:** Compare the p-values to the significance level to decide if the treatment differs significantly from the control.
- 

This theoretical process ensures that you are controlling the **family-wise error rate** while comparing each treatment group to a single control, without inflating the probability of Type I errors due to multiple comparisons.

(1. What happens if the variance are not equal?)

(2. Mention the influence of if the sample size is equal.)

## Bonferroni procedure

Q: what is Bonferroni procedure

A:

The **Bonferroni procedure** is a statistical method used to adjust **p-values** when conducting multiple hypothesis tests to control the **family-wise error rate** (FWER). The family-wise error rate is the probability of making **at least one Type I error** (i.e., rejecting a true null hypothesis) when performing multiple comparisons.

### Why Use the Bonferroni Procedure?

When you perform multiple statistical tests, the chance of incorrectly rejecting a null hypothesis (i.e., finding a false positive) increases with the number of comparisons. The Bonferroni correction helps to reduce the risk of these false positives by adjusting the significance level to account for the number of comparisons being made.

### Basic Concept:

The Bonferroni procedure is based on the idea of adjusting the **alpha level** (significance level) to make it more stringent when there are multiple comparisons. In essence, it reduces the **probability of making Type I errors** by lowering the threshold for significance.

### How it Works:

1. **Divide the original alpha level ( $\alpha$ )** by the number of comparisons (tests) you are conducting.
  - If you were conducting  $m$  tests, and your original significance level is  $\alpha=0.05$ , the adjusted significance level for each test becomes:  
 $\alpha_{\text{adjusted}}=\alpha/m$
  - For example, if you are performing 5 tests and your desired significance level is 0.05, the Bonferroni-adjusted significance level for each test would be:  $\alpha_{\text{adjusted}}=0.05/5=0.01$

- In this case, you would only reject the null hypothesis for any test where the p-value is **less than 0.01**.

## 2. Perform the individual hypothesis tests.

- For each test, calculate the p-value and compare it to the adjusted significance level  $\alpha_{\text{adjusted}}$

## 3. Compare p-values to the adjusted alpha level:

- If the **p-value** for a test is **less than or equal to**  $\alpha_{\text{adjusted}}$ , reject the null hypothesis for that test.
- If the **p-value** is **greater than**  $\alpha_{\text{adjusted}}$ , fail to reject the null hypothesis for that test.

### Example:

Suppose you are testing 5 hypotheses (e.g., comparing 5 different treatments to a control group), and you want to control the family-wise error rate at  $\alpha=0.05$ .

- **Number of comparisons (m) = 5.**
- **Original significance level  $\alpha=0.05$**   $\alpha = 0.05$ .

Using the Bonferroni correction:

$\alpha_{\text{adjusted}} = 0.05/5 = 0.01$

Now, for each of your 5 tests, you will compare the **p-value** to 0.01:

- If the **p-value** for any test is **less than 0.01**, you reject the null hypothesis.
- If the **p-value** is **greater than 0.01**, you fail to reject the null hypothesis.

### Advantages of the Bonferroni Procedure:

1. **Simple and intuitive:** It is easy to understand and implement.
2. **Conservative approach:** It effectively controls the family-wise error rate, ensuring that the probability of making one or more Type I errors is kept under the desired threshold (usually 0.05).

### Disadvantages of the Bonferroni Procedure:

1. **Conservativeness:** The Bonferroni correction can be too conservative, especially when there are many tests (**what number will be considered too many to use this method**). This means it increases the risk of Type II errors

(failing to reject a false null hypothesis), as it requires a very stringent threshold for significance.

2. **Power loss:** By adjusting the alpha level for multiple comparisons, you reduce the likelihood of detecting true effects, especially when the number of tests is large. This can lead to a **loss of statistical power** (i.e., an increased chance of failing to detect a true effect).

### **Alternative Adjustments:**

Because of the conservativeness of the Bonferroni procedure, there are several alternatives designed to be less strict and preserve more statistical power:

- **Holm-Bonferroni method:** A step-down approach that adjusts the alpha level in a more flexible manner and is less conservative than the standard Bonferroni correction.
- **Benjamini-Hochberg procedure:** A method that controls the **false discovery rate** (FDR), which is the expected proportion of false positives among the rejected hypotheses, rather than controlling the family-wise error rate.

### **Conclusion:**

The **Bonferroni procedure** is a widely used method for controlling the family-wise error rate when performing multiple hypothesis tests. It works by adjusting the alpha level based on the number of tests you are conducting, making it a conservative method that reduces the chances of Type I errors. However, it can be too stringent and lead to loss of statistical power, especially when a large number of comparisons are involved.

## Overall comment

Overall, the output follows good structure to answer the question, but it had difficulty to ensure the

1. Consistency (ask the same question, you may get different ways of explanation).
2. Also, the output often tells you steps without telling why (Bonferroni's trade-off between Type I/II errors).
3. Some output could be misleading, or incomplete for a complicated concept (pooled variance in Scheffé's test without providing a detailed formula; assumptions was not mentioned for the data used).
4. As AI often provided a limited answer, it is not good for college students to build a comprehensive understanding between similar methods and make comparison.