A4. Multiplicity Adjustments

Understanding and controlling error rates in multiple comparisons

Dr. Paul Schmidt

To install and load all the packages used in this chapter, run the following code:

```
for (pkg in c("tidyverse", "emmeans", "multcomp", "multcompView")) {
  if (!require(pkg, character.only = TRUE)) install.packages(pkg)
}

library(tidyverse)
library(emmeans)
library(multcomp)
library(multcompView)
```

After fitting a linear model and running an ANOVA, the typical next step is to ask **which groups actually differ from each other**. This leads to multiple pairwise comparisons, and as soon as more than a single comparison is made, the probability of finding a "significant" difference by chance alone starts to grow. This chapter explains why that is a problem, what adjustment methods exist, and how to use them via the `emmeans` package.

The Motivating Problem

Whenever several hypotheses are tested simultaneously, the chance of at least one false positive grows with the number of tests. If a single test is conducted at the 5% level and the null hypothesis is true, the probability of a false rejection is 5%. If 10 independent tests at the 5% level are conducted under true null hypotheses, the probability of at least one false rejection is already about 40%. For 20 tests it exceeds 64%. This is the **multiple comparisons problem**.

> **i** Brief reminder: Type I and Type II errors
>
> A Type I error (α) occurs when a true null hypothesis is incorrectly rejected (a false positive). A Type II error (β) occurs when a false null hypothesis is not rejected (a false negative). Saying a test is "performed at the 5% level" means the Type I error rate is controlled at 5% for that one test. Multiplicity adjustments extend this idea of error control from a single comparison to a whole family of comparisons.

The figure above translates these definitions into a familiar setting: a weather forecast acting as a "test" for whether it will rain. The null hypothesis is the default state (no rain, no need for an umbrella), and rejecting it means taking action. A **Type I error** corresponds to bringing an umbrella on a sunny day - mildly annoying but harmless. A **Type II error** corresponds to leaving the umbrella at home and getting soaked - a much costlier mistake. This asymmetry is a useful intuition for later sections, where the cost of false positives versus false negatives drives the choice of adjustment method.

Two different notions of error control appear throughout the literature:

- **Family-Wise Error Rate (FWER):** the probability of making *at least one* Type I error among all comparisons in the family. Methods like Tukey, Dunnett, Bonferroni, Holm, Sidak and Scheffé control the FWER.
- **False Discovery Rate (FDR):** the expected *proportion* of false positives among all rejected hypotheses. This is a less strict criterion introduced by Y. Benjamini and Y. Hochberg [1] and is especially popular when many comparisons are made (e.g. in genomics).

Which one to control depends on the cost of a false positive. In a confirmatory field trial with a handful of treatments, FWER control via Tukey or Dunnett is the standard. In a screening setting with hundreds of comparisons, FDR is often more appropriate because strict FWER control would leave almost no power.

A Running Example

Throughout this chapter we use the built-in `PlantGrowth` dataset, which records dried plant weight under three conditions: a control group (`ctrl`) and two treatments (`trt1`, `trt2`).

```
mod <- lm(weight ~ group, data = PlantGrowth)
anova(mod)
```

```
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value  Pr(>F)
group      2  3.7663  1.8832  4.8461 0.01591 *
Residuals 27 10.4921  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA shows a significant effect of `group`, so it is reasonable to ask *which* groups differ. All pairwise comparisons can be obtained with `emmeans`:

```
emm <- emmeans(mod, specs = ~ group)
emm
```

```
 group emmean    SE df lower.CL upper.CL
 ctrl    5.03 0.197 27     4.63     5.44
 trt1    4.66 0.197 27     4.26     5.07
 trt2    5.53 0.197 27     5.12     5.93

Confidence level used: 0.95
```

```
pairs(emm, adjust = "none")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.1944
 ctrl - trt2   -0.494 0.279 27  -1.772  0.0877
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0045
```

With `adjust = "none"`, `emmeans` reports unadjusted p-values. These are essentially multiple t-tests and do **not** account for the fact that three comparisons are being made at once. This is sometimes called the Fisher LSD (Least Significant Difference) approach and is generally discouraged unless an omnibus F-test has already established overall significance and only very few comparisons are planned.

Methods at a Glance

The table below summarises the most common adjustment methods, when to use them, what they control, and how to request them in `emmeans`. Every method listed (except "none") is available via the `adjust` argument of `pairs()`, `contrast()`, or `summary()` on an `emmGrid` object.

Method	When to use	Controls	Conservativeness	`emmeans` call
None (Fisher LSD)	Only if justified; few planned comparisons after a significant F-test	Nothing (per-comparison error only)	Most liberal	`adjust = "none"`
Tukey HSD	All pairwise comparisons among group means	FWER	Moderate, optimal for all-pairs	`adjust = "tukey"`
Dunnett	Comparing several treatments against a single control	FWER	Less conservative than Tukey for this case	`adjust = "dunnett"` (with `trt.vs.ctrl`)
Bonferroni	Any set of pre-specified comparisons, simple to report	FWER	Very conservative, especially for many tests	`adjust = "bonferroni"`
Holm	Same situations as Bonferroni, strictly more powerful	FWER	Less conservative than Bonferroni	`adjust = "holm"`
Sidak	Independent tests, assumes independence	FWER	Slightly less conservative than Bonferroni	`adjust = "sidak"`
Scheffé	Any (including unplanned, data-driven) linear contrasts	FWER	Most conservative; very general	`adjust = "scheffe"`
Benjamini-Hochberg (FDR)	Large numbers of comparisons, screening context	FDR	Much less conservative than FWER methods	`adjust = "fdr"` (or `"BH"`)

Two structural distinctions are worth keeping in mind:

- **Simultaneous vs. sequential (stepwise):** Bonferroni and Sidak apply the same adjustment to every p-value at once (simultaneous). Holm, on the other hand, sorts the p-values from smallest to largest and applies progressively weaker adjustments in sequence. Holm uniformly dominates Bonferroni, which means it is always at least as powerful while controlling the same FWER.
- **Targeted vs. general:** Tukey is tailored to all-pairs comparisons and Dunnett to many-vs-one comparisons. Both exploit the correlation structure among the contrasts and are therefore more powerful than generic methods like Bonferroni when used in their intended setting.

Tukey versus Dunnett

Tukey HSD and Dunnett's test are both well-calibrated FWER methods, but they answer different questions. **Using the wrong one wastes power.**

- **Tukey** is the right choice when the research question is "which groups differ from which?" and all pairwise contrasts are of interest. For k groups this is $k(k-1)/2$ contrasts.
- **Dunnett** is the right choice when the research question is "which treatments differ from the control?" and only the $k-1$ treatment-vs-control contrasts are of interest. Dunnett explicitly exploits the fact that all contrasts share the control as a reference, making it more powerful than Tukey when only many-vs-one comparisons are needed.

Dunnett's test is underused in practice. Many analysts default to Tukey even when only control comparisons matter, losing statistical power unnecessarily.

```
# All pairwise comparisons with Tukey adjustment
pairs(emm, adjust = "tukey")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.3909
 ctrl - trt2   -0.494 0.279 27  -1.772  0.1980
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0120

P value adjustment: tukey method for comparing a family of 3 estimates
```

```
# Treatments vs. control (ctrl as reference) with Dunnett adjustment
contrast(emm, method = "trt.vs.ctrl", ref = "ctrl", adjust = "dunnett")
```

```
 contrast    estimate    SE df t.ratio p.value
 trt1 - ctrl   -0.371 0.279 27  -1.331  0.3296
 trt2 - ctrl    0.494 0.279 27   1.772  0.1582

P value adjustment: dunnettx method for 2 tests
```

Note how `method = "trt.vs.ctrl"` restricts the set of contrasts to "each treatment against the control", which is exactly the family Dunnett's test is designed for.

Bonferroni, Holm, Sidak

These three methods make no assumptions about the structure of the comparisons and can therefore be applied to any set of contrasts - planned or unplanned.

```
pairs(emm, adjust = "bonferroni")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.5832
```

```
 ctrl - trt2   -0.494 0.279 27  -1.772  0.2630
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0134

P value adjustment: bonferroni method for 3 tests
```

```
pairs(emm, adjust = "holm")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.1944
 ctrl - trt2   -0.494 0.279 27  -1.772  0.1754
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0134

P value adjustment: holm method for 3 tests
```

```
pairs(emm, adjust = "sidak")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.4771
 ctrl - trt2   -0.494 0.279 27  -1.772  0.2407
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0133

P value adjustment: sidak method for 3 tests
```

For this small example the three methods give similar results, but for larger comparison families the differences become more pronounced. Holm should generally be preferred over Bonferroni because it controls the FWER at the same level while being more powerful. Sidak is slightly less conservative than Bonferroni but assumes independence of the test statistics, which is rarely exactly true in ANOVA-type analyses.

False Discovery Rate

When many comparisons are made - for instance in genomic screens, high-throughput phenotyping, or large multi-environment trials - strict FWER control becomes so conservative that genuine effects can no longer be detected. The FDR approach of Y. Benjamini and Y. Hochberg [1] relaxes the criterion: instead of controlling the probability of *any* false rejection, it controls the expected *proportion* of false rejections among the rejected hypotheses.

```
pairs(emm, adjust = "fdr")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.1944
 ctrl - trt2   -0.494 0.279 27  -1.772  0.1315
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0134

P value adjustment: fdr method for 3 tests
```

The FDR adjustment is less conservative than any FWER method and is particularly well-suited to exploratory or screening analyses. It should not be used as a loophole to obtain more “significant” results in a confirmatory study with a small, pre-specified comparison family; there, an FWER method is the correct choice.

Scheffé

Scheffé’s method is the most general - and consequently the most conservative - of the FWER procedures. It protects against *any* possible linear contrast among the group means, including contrasts that were only formulated after looking at the data. For a set of pre-specified pairwise comparisons it is almost always beaten by Tukey or Dunnett.

```r
pairs(emm, adjust = "scheffe")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.4241
 ctrl - trt2   -0.494 0.279 27  -1.772  0.2265
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0163

P value adjustment: scheffe method with rank 2
```

Scheffé is most useful when post-hoc, data-driven contrasts are to be tested (e.g. “the average of groups A and B versus C”) without any prior specification.

Choosing a Method Before Looking at the Data

A critical methodological point: **the adjustment method should be chosen before inspecting the results, not after.** Trying several `adjust =` options and reporting whichever makes the most comparisons “significant” is a form of p-value hacking that inflates the effective Type I error rate well beyond the nominal 5%. The appropriate method follows from the research question and comparison family:

- All pairwise comparisons planned -> Tukey.
- Comparisons against a control -> Dunnett.
- Small pre-specified set of arbitrary contrasts -> Bonferroni or Holm.
- Very many comparisons, screening setting -> FDR.
- Truly post-hoc, unplanned contrasts -> Scheffé.

Transparent reporting is equally important: the method used, the comparison family it was applied to, and whether the analysis was pre-specified or exploratory should all be stated clearly.

Cross-reference: Compact Letter Display

A common way to visualise the results of multiple comparisons is the **Compact Letter Display (CLD)**, in which groups sharing a letter are not significantly different. A dedicated chapter covers the construction and interpretation of CLDs as well as their well-documented pitfalls - see the upcoming Appendix A5 on Compact Letter Displays.

Additional Resources

- F. Bretz, T. Hothorn, and P. Westfall [2] “Multiple Comparisons Using R” - the standard reference, covering both theory and implementation via the `multcomp` package.
- T. Hothorn, F. Bretz, and P. Westfall [3] “Simultaneous Inference in General Parametric Models” - the methodological paper behind `multcomp::glht()`.
- emmeans vignette “Comparisons and contrasts” - a practical walkthrough of all adjustment options available in `emmeans`.
- Y. Benjamini and Y. Hochberg [1] “Controlling the False Discovery Rate” - the original FDR paper.
- Lee S, Lee DK. What is the proper way to apply the multiple comparison test? Korean J Anesthesiol. 2018 Oct;71(5):353-360. doi: 10.4097/kja.d.18.00242

ℹ Key Takeaways

1. **Multiple comparisons inflate the Type I error rate.** Adjustments are needed whenever more than a handful of comparisons are made.
2. **FWER vs. FDR** is the first decision. FWER is standard for confirmatory analyses; FDR is appropriate for screening.
3. **Tukey for all-pairs, Dunnett for vs-control.** Using Dunnett when only control comparisons matter saves substantial power.
4. **Holm dominates Bonferroni** and should be preferred for arbitrary pre-specified comparison families.
5. **Scheffé is only for truly unplanned contrasts.** It is too conservative for standard pairwise comparisons.
6. **Choose the method before inspecting the results.** Picking the method post-hoc to maximise "significant" findings is p-value hacking.
7. **Report transparently** which method was used and to which comparison family it was applied.

Bibliography

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[2] F. Bretz, T. Hothorn, and P. Westfall, *Multiple Comparisons Using R*. Boca Raton: Chapman, Hall/CRC, 2011. doi: 10.1201/9781420010909.

[3] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008, doi: 10.1002/bimj.200810425.