Author: Jenny Freeman
Reviewer: Phillips Obasohan

Multiple testing including post hoc tests

Date created: 23/07/2020

Some useful definitions:

Type 1 error (also known as an alpha error, or a false positive error)

A type 1 error is the error that you commit when you incorrectly reject the null hypothesis. At its simplest, the null hypothesis states that there is no effect. The alternative hypothesis states that there is an effect. Thus a type 1 error is the error that is committed when you decide that there is an effect, when in fact there is no effect in the population. It can be thought of as a false positive error.

Type 2 error (also known as a beta error, or a false negative error)

A type 2 error is the error that you commit when you fail to reject the null hypothesis when it is false. It is the error that is committed when there is an effect, but you do not reject the null hypothesis. It can be thought of as a false negative error.

False discovery rate

This is the expected proportion of type 1 errors committed when you conduct multiple hypothesis tests. As mentioned above a type one error is when you incorrectly reject the null: you get a false positive result. The false discovery rate is the rate at which significant results are actually not significant. For example a false discovery rate of 5% would mean that of all the results declared significant, in 5% of the cases we have incorrectly rejected the null hypothesis. In standard significance testing we set alpha to be 0.05 to control the false discovery rate:

$FDR = E(V/R \mid R > 0)\, P(R>0)$
Where V=number of type 1 errors (i.e. number of false positives)
R= Number of rejected hypotheses

You can think of this as being the expected number significant results that are actually false positives. For example a false discovery rate of 5% means that 5% of significant results are actually false

Familywise error rate/experiment-wise error rate

This is the probability of committing at least one type 1 error. It is most easily thought of in terms of the opposite case. For a single significance test the probability of committing a type 1 error is 0.05, thus the probability of not committing a type one error is 1-0.05, that is 0.95. If you conduct two significance tests, the probability of committing at least one type one error is 1-probability of not committing any. Assuming that the tests are independent, the probability of committing no type 1 errors is 0.95x0.95. Thus the probability of committing at least one is 1-(0.95x0.95)

For n tests the familywise error rate is $1-(0.95)^n$

P-hacking

This refers to the practice of selectively reporting the most positive results. It occurs when an investigator tests several hypotheses on a dataset with a variety of subgroups, multiple secondary outcome measures, inclusion/exclusion criteria and selectively report only those that yield significant results. It leads to the mis-reporting of true effect sizes and can lead to bias in the understanding of outcomes. One way to prevent the practice is to specify in advance of a study all analyses that will be conducted, to ensure that there are no unauthorised post-hoc analyses carried out in the search for a significant result.

Multiple testing: what is the issue?

The significance level for a single significance test is usually set at 0.05, or 5%. This means that the probability of committing a type 1 error is set at 0.05. A type 1 error can be thought of as a false positive error. However, it's rarely the case that we do a single significance test on a given set of data and when we repeatedly test hypotheses we increase our risk of committing a type 1 error. To illustrate this, let's assume that we want to conduct 5 significance tests. For the sake of simplicity, let's also assume that these 5 tests are independent. For a single test, if the significance level is set at 0.05, then the probability of not committing a type 1 error is 1-0.05 = 0.95. Given that the tests are independent of each other, the probability that we do not commit a type 1 error in all 5 tests is 0.95 x 0.95 x 0.95 x 0.95 x 0.95 i.e. 0.95^5. Thus the probability of committing at least one type 1 error is $1- 0.95^5 = 0.23$. The probability of a type 1 error has increased from 0.05 to 0.23. The name for this, the error rate across tests conducted on the same data is known as the family-wise error rate (see below for more information).

In order to guard against this and control the overall type 1 error rate, various adjustments have been proposed. Particularly when conducting an analysis of variance, if the overall analysis of variance is significant, interest will focus on conducting post hoc analyses to understand where the differences lie i.e. which particular groups are different. In this case, it would be inappropriate to conduct separate t-tests because of the increased risk of a type 1 error, and it is recommended that post hoc analyses adjust for multiple testing in order to control the overall type 1 error rate. Which test is chosen depends on a number of factors.

It is worth noting that with multiple testing there is a balance to be struck between preserving the overall type 1 error rate whilst ensuring that the risk of committing a type 2 error does not increase to an unacceptable level. A type 2 error is the error that is committed when a real effect is not detected. This can be thought of as a false negative error. The power of a study is its ability to detect a genuine effect. As the type 1 error rate decreases the type 2 error rate increases – the two are inversely related. As the number of false positives goes down, the number of false negatives goes up

Things to consider when choosing a post hoc test:

- How conservative/liberal you want to be i.e. how sure do you want to be that your overall type 1 error rate is controlled. Tests that are referred to as conservative have much tighter control of the overall type 1 error rate.
- How powerful you want to be i.e. how sure you want to be that you can detect a difference if one exists (not commit a type 2 error). Tests with greater power are better able to detect a difference if one exists.
- Sample sizes of the groups.
- Variability of the groups.

Some common tests

Test	Description	Recommended usage
Bonferroni	Most conservative but simple to apply. Sets the significance cut-off at alpha/n. If conduct 5 tests, divide 0.05 by 5, and then only declare a result significant at the 5% level if an individual test p-value is < 0.01. Not very powerful, in part because it overcorrects for type 1 errors	Use when you want guaranteed control of type 1 error rate, but most conservative of the methods available, so can suffer from lack of power – poor at detecting a difference if one exists
Tukey (honest significant difference)	Similar to Bonferroni in controlling the type 1 error rate. Very conservative (lack power to detect a difference if one exists). More power over Bonferroni when testing a large number of differences. Also, generally has greater power than both Scheffé and Dunn	Use when have equal variances and equal sample sizes. Also better than Bonferroni when testing a large number of differences
REGWQ (Ryan, Einot, Gabriel and Welsch Q)	Good power and control of type 1 error rate. Recommended when want to test all pairs of means. Should not be used when group sizes are different	Use when have equal variances and equal sample sizes
Scheffe	More conservative than Tukey for reducing the risk of type 1 error, but at the cost of a lack of power – you will be less likely to detect an effect. Useful when interested in general comparisons across groups rather than individual pairwise comparisons	Use instead of Tukey when the group sizes are different
Dunn	Conservative test, lacking in power, especially when you have a large number of comparisons, in which case use Tukey. Also used for multiple comparisons for the Kruskall-Wallis test in preference to Bonferroni	Only use when making a small number of comparisons
Benjamin-Hochberg (BH)	Good when conducting a large number of tests as it accounts for the false discovery rate	Use when conducting a large number of post hoc tests
Gabriel's	Designed to be used when group sizes are different. Generally more powerful but can be too liberal when sample sizes are very different	Use when sample sizes are slightly different
Hochberg's GT2	Designed to be used when group sizes are different. But very unreliable when populations variances are different	Use when sample sizes are very different but not if the variances are different

Dunnett	Designed for the situation when one group is a ***control group*** and you want to compare all treatment groups to this control group	Use when there is a control group that you want to compare the other groups to.

The following have been developed for use when the population variances differ

Test	Description	Recommended use
Games-Howell	Most powerful of the 4 but can be liberal when sample sizes are small. But good when sample sizes are unequal	Recommended method
Tamhane's T2	Conservative – tight control of the type 1 error rate	
Dunnett's T3	Conservative – tight control of the type 1 error rate	
Dunnett's C	Conservative – tight control of the type 1 error rate	

Some common tests that are generally not recommended

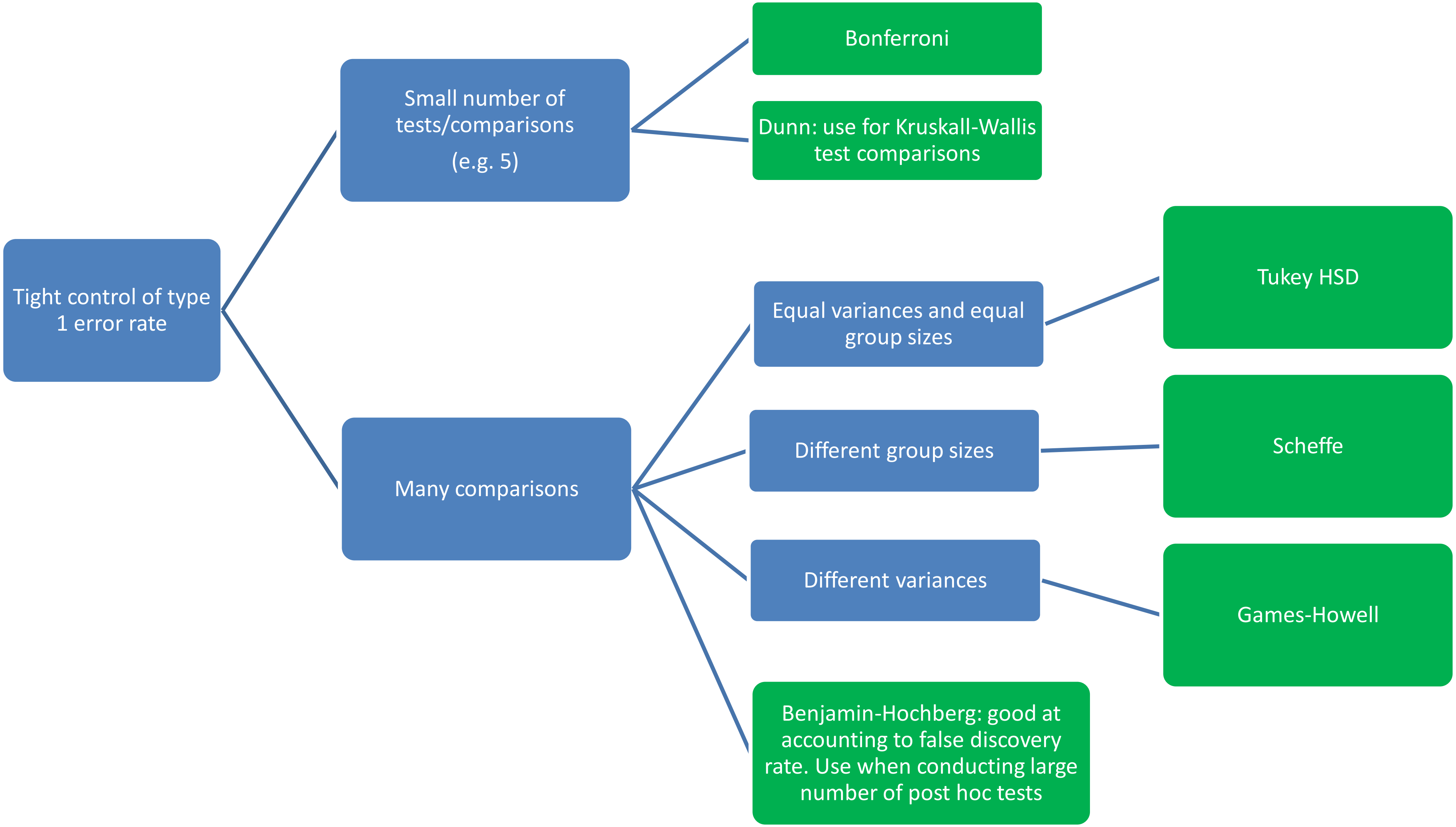
Duncan	Overall ANOVA result will tell you whether there are any differences between means, but not which means differ. Duncan's test identifies pairs of means that differ – don't think it makes any attempt to control the type 1 error rate, so similar to LSD in that respect	Not recommended as doesn't control the type 1 error rate
Studentized Newman-Keuls (SNK)	Liberal test in that it does not control the family-wise error rate very well. Good for detecting differences if they exist.	Not recommended as doesn't control the type 1 error rate very well
LSD (Least-significant difference)	No attempt to control the type 1 error rate; equivalent to conducting unadjusted multiple tests. Requires that the overall ANOVA is significant.	Not recommended as doesn't control the type 1 error rate

Multiple Comparison Procedures. Larry Toothaker. Sage. 1993

Author: Jenny Freeman
Reviewer: Phillips Obasohan

Multiple testing including post hoc tests

Date created: 23/07/2020



Tight control of type 1 error rate

Small number of tests/comparisons (e.g. 5)

Bonferroni

Dunn: use for Kruskall-Wallis test comparisons

Many comparisons

Equal variances and equal group sizes

Tukey HSD

Different group sizes

Scheffe

Different variances

Games-Howell

Benjamin-Hochberg: good at accounting to false discovery rate. Use when conducting large number of post hoc tests

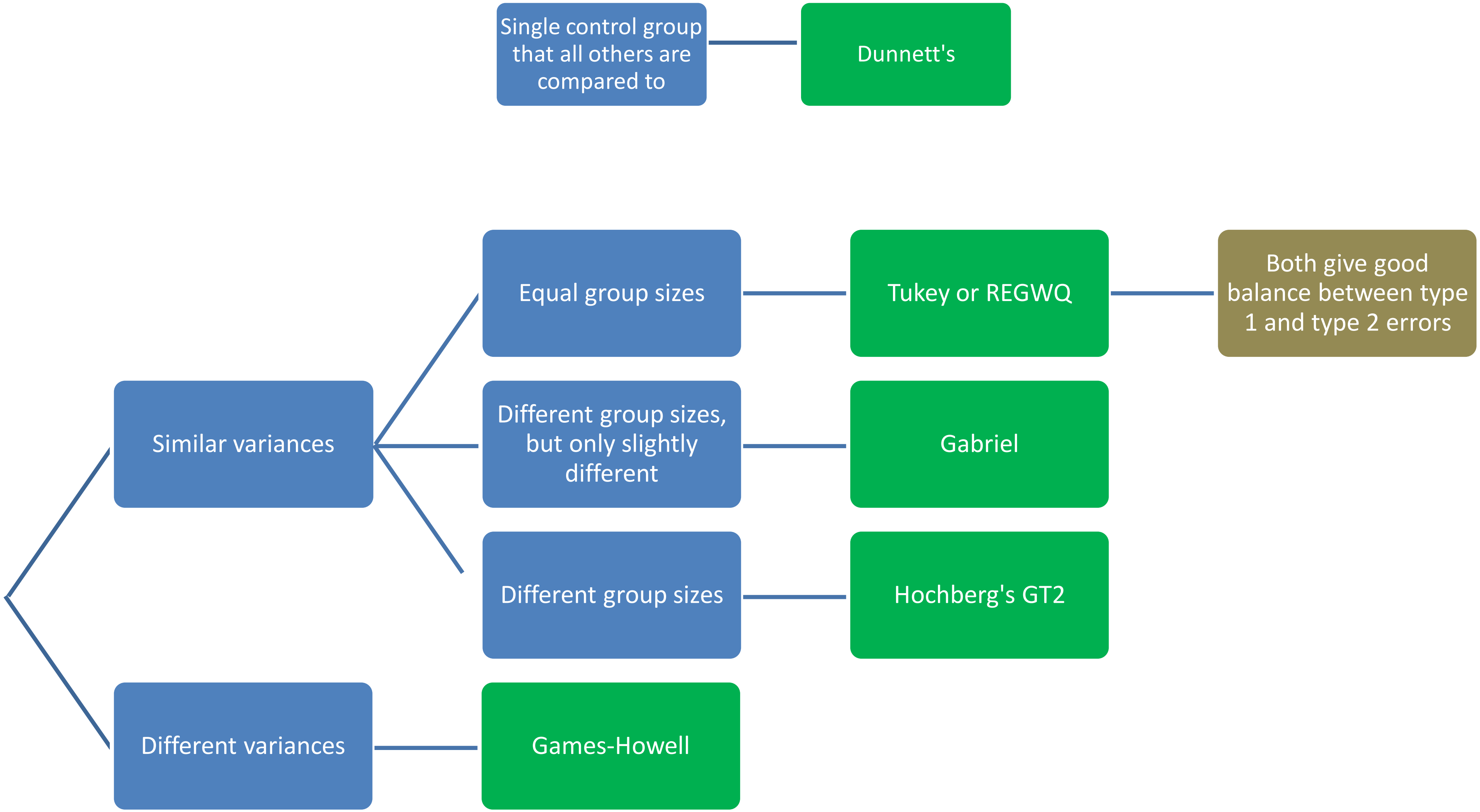
The University Of Sheffield.

301

Author: Jenny Freeman
Reviewer: Phillips Obasohan

Multiple testing including post hoc tests

Date created: 23/07/2020



Single control group that all others are compared to — Dunnett's

Similar variances

- Equal group sizes — Tukey or REGWQ — Both give good balance between type 1 and type 2 errors
- Different group sizes, but only slightly different — Gabriel
- Different group sizes — Hochberg's GT2

Different variances — Games-Howell