Chapter 18
Bonferroni Adjustments

1 General Purpose

The unpaired t-test can be used to test the hypothesis that the means of two parallel-group are not different (Chap. 7). When the experimental design involves multiple groups, and, thus, multiple tests, it will increase the chance of finding differences. This is, simply, due to the play of chance, rather than a real effect. Multiple testing without any adjustment for this increased chance is called data dredging, and is the source of multiple type I errors (chances of finding a difference where there is none). The Bonferroni adjusted t-test (and many other methods) are appropriate for adjusting the increased risk of type I errors. This chapter will assess how it works.

2 Schematic Overview of Type of Data File

Group (1,2,3,. . .)	Outcome
.	
.	
.	
.	
.	
.	
.	
.	
.	

T.J. Cleophas, A.H. Zwinderman, *Clinical Data Analysis on a Pocket Calculator*,
DOI 10.1007/978-3-319-27104-0_18

3 Primary Scientific Question

In a parallel-group study of three or more treatments, does Bonferroni t-test adequately adjust the increased risk of type I errors?

4 Bonferroni T-Test, Data Example

The underneath example studies three groups of patients treated with different hemoglobin improving compounds. The mean increases of hemoglobin are given.

	Sample size	Mean hemoglobin mmol/l	Standard deviation mmol/l
Group1	16	8.725	0.8445
Group 2	10	10.6300	1.2841
Group 3	15	12.3000	0.9419

An overall analysis of variance test produced a p-value of < 0.01. The conclusion is that we have a significant difference in the data, but we will need additional testing to find out, exactly where the difference is:

between group 1 and 2,
between group 1 and 3, or
between group 2 and 3.

The easiest way is to perform a t-test for each comparison. It produces a highly significant difference at $p < 0.01$ between group 1 versus 3 with no significant differences between the other comparisons. This highly significant result is, however, unadjusted for multiple comparisons. If one analyzes a set of data with three t-tests, each using a 5 % critical value for concluding that there is a significant difference, then there is about $3 \times 5 = 15$ % chance of finding a significant difference at least once. This mechanism is called the Bonferroni inequality. Bonferroni recommended a solution for the inequality, and proposed to follow in case of three t-tests to use a smaller critical level for concluding that there is a significant difference:

With 1t-test: critical level = 5 %
With 2t-tests: critical level = (5 %) /2 = 2,5 %
With 3t-tests: critical level = (5 %) /3 = 1.67 %.

a more general version of the equation is given underneath:

In case of k comparisons and an overall critical level (= null-hypothesis rejection level) of α the rejection p-value will become

$$\alpha \times 2/(k(k-1))$$

E.g. with k = 3, and α = 0.05 (5 %)

$$0.05 \times \frac{2}{3(3-1)} = 0.0166.$$

In the given example a p-value of 0.0166 is still larger than 0.01, and, so, the difference observed remained statistically significant, but using a cut-off p-value of 0.0166, instead of 0.05, means that the difference is not *highly* significant anymore.

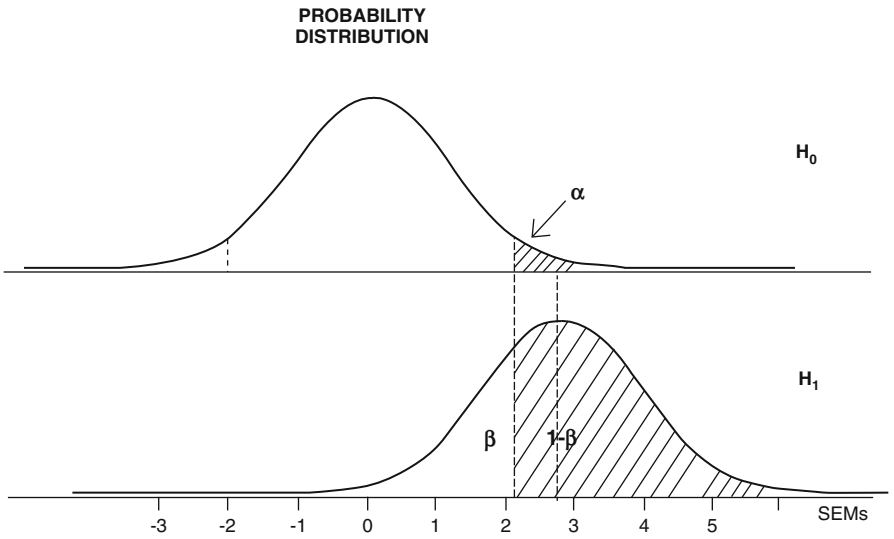
5 Bonferroni T-Test Over-conservative

With k-values not too large, this method performs well. However, if k is large (k > 5), then this Bonferroni correction will rapidly lead to very small rejection p-values, and is, therefore, called over-conservative. E.g., with 10 comparisons the rejection p-value will be only $0.05 \times 2/(10 \times 9) = 0.001$, which is a value hard to be obtained in a small study. Actually, a more realistic rejection p-value may be larger, because in most multiple test studies a positive correlation between multiple treatment comparisons exists. It means that multiple tests in one study are much more at risk of similar results than multiple tests in different studies (see also Chap. 10). The chance of twice a p-value of 0.05 may, therefore, not be 0.025, but, rather, something in between, like 0.035.

6 Conclusion

Bonferroni adjustment is adequate for adjusting the increased type I error of multiple testing, and can easily be performed with the help of a pocket calculator. Alternative methods include Tukey's honestly significant difference (HSD) method, Student-Newman-Keuls method, the method of Dunnett, and many more. They are, however, computationally more laborious, and require specific statistical tables. Statistical software programs like SPSS or SAS are helpful.

Bonferroni adjustments increase the risk of type II errors (β-values) of not finding a difference which does exist. This is illustrated in the underneath figure: with α = 0.05 (left vertical interrupted line), β is about 30 % of the area under the curve. With α = 0.167 (adjusted for three tests as demonstrated in the above Sect. 4) (right vertical interrupted line), β has risen to about 50 %. This rise caused loss of power from about 70 % to about only 50 % ((1−β)-values), (see also Chap. 11 for additional explanation of power assessments). H0 = null-hypothesis, H1 = alternative hypothesis, SEM = standard error of the mean.



In the current chapter only continuous outcome data are adjusted for multiple testing. However, binary data can equally be assessed using the Bonferroni equation.

7 Note

More background, theoretical and mathematical information of multiple comparisons and false positive studies are given in Statistics applied to clinical studies 5th edition, Chaps. 8 and 9, Springer Heidelberg Germany, from the same authors.