12.1: The need for ANOVA

Introduction

Moving from an experiment with two groups to multiple groups is deceptively simple: we move from one comparison to **multiple comparisons**. Consider an experiment in which we have randomly assigned patients to receive one of three doses of a statin drug (lower cholesterol), including a placebo (e.g., Tobert and Newman 2015). Thus, we have three groups or **levels** of a single treatment **factor** and we'll want to test the null hypothesis that the group (level) means are all equal as opposed to the alternate hypothesis in which one or more of the group means, e.g., group A, group B, group C, are different.

$$H_0 : \bar{X}_A = \bar{X}_B = \bar{X}_C$$

The correct procedure is to analyze multiple levels of a single treatment with a one-way analysis of variance followed by a suitable **post-hoc** ("after this") test. Two common post-hoc tests are **Tukey's range test** (aka Tukey's HSD [honestly significant difference] test), which is for all **pairwise comparisons**, or the **Dunnett's test**, which compares groups against the control group. Post-hoc tests are discussed in Chapter 12.6.

Thus, the **family-wise** (aka **experiment-wise**) error rate for multiple comparisons is kept at 5%, and each **individual-wise comparison** is compared against a more strict (i.e., smaller **Type I error rate**). Put another way, the family-wise error rate is the chance of a number of false positives: making a mistake when we consider many tests simultaneously. The simplest correction for the individual-wise error rate is the **Bonferroni correction**: test each individual comparison at Type I error equal to α/C, where C is the number of comparisons.

> **Note:**
>
> To get the number of "pairwise" comparisons (C), let
>
> $$C = \frac{k(k-1)}{2}$$

For our three group experiment, how many pairwise comparisons can be tested? Therefore, $k = 3$, and we have

$$C = \frac{3(3-1)}{2} = 3$$

Thus, for our three groups, A, B, and C, there are three possible pairwise comparisons.

$$H_1 : \bar{X}_A = \bar{X}_B$$
$$H_2 : \bar{X}_A = \bar{X}_C$$
$$H_3 : \bar{X}_B = \bar{X}_C$$

How many pairwise comparisons for a four-group experiment? Check your work, you should get $C = 6$.

The multiple comparison problem

Let's say that we're stubborn. We could do many single two-sample t-tests — certainly, your statistical software won't stop you — but this is a situation that calls for statistical reasoning. Here's why we should not: we will increase the probability of rejecting a null hypothesis when the null hypothesis is true (e.g., discussion in Jafari and Ansari-Pour 2019). That is, the chance we will commit a Type I error increases if we do not account for the lack of independence in these sets of pairwise tests evaluated by t-tests. This is **multiplicity** or the **multiple comparison problem**.

Review: when we perform a two-sample t-test we are willing to reject a true null hypothesis 5% of the time. This is what is meant by setting the critical probability value (alpha) = 0.05. By "willing" we mean that we know that our conclusions could be wrong because we are working with samples, not the entire population. (Of course at the time, we have no way of actually knowing WHEN we are wrong, but we do want to know how likely we could be wrong!) However, if we compare three population means we have three separate null hypotheses.

H_O: one or more of the means are different.

But if we conduct these as separate independent sample t-tests, then we are implicitly making the following null hypothesis statements:

$$H_1 : \bar{X}_O = \bar{X}_B$$
$$H_2 : \bar{X}_O = \bar{X}_C$$
$$H_3 : \bar{X}_O = \bar{X}_C$$

Thus, we have a 5% chance of being wrong for the first hypothesis and/or a 5% chance of being wrong for the second hypothesis and/or a 5% chance of being wrong for the third hypotheses. The chance that we will be wrong for at least one of these hypotheses must now be greater than 5%.

For three separate hypotheses there is a 14% chance of being wrong when we have the probability value for each individual t-test set at $\alpha = 0.05$. How did we get this result? The point is that these tests are not independent, they are done on the same data set; therefore, you can't simply apply the multiplication probability rule.

Here's how to figure this: for the set of three hypotheses, the probability of incorrectly rejecting at least one of the null hypotheses is $1-(1-\alpha)^3 = 1-0.957 = 0.143$

So, for three t-tests on the same experiment, the Type I error for the overall tests (experiment-wise) is actually 14%, not 5%. It gets worse as the number of combinations (groups and therefore hypotheses) increases. For four groups, Type I error is actually $1-(1-\alpha)^4 = 1-0.815 = 0.185$.

That's 18.5%, not 5%.

If we have just five populations means to compare, the probability of rejecting a null hypothesis when it is true climbs to 60%! How did this happen? The probability of correctly rejecting all of them is now $(1-\alpha)^5 = 0.774$

So, the probability of incorrectly rejecting one test (Type I error) is now $1-(1-\alpha)^5 = 1-0.774 = 0.226 = 22.6\%$ instead of the 5% we think we are testing.

This is the key argument for why you must use ANOVA to analyze multiple samples instead of a combination of t-tests!! ANOVA guarantees that the overall error rate is the specified 5%.

Why is the Type I error not 5% for each test? Because we conducted ONE experiment, we can conduct only ONE test (we could be right, we could be wrong 5% of the time). If we conduct the experiment over again, on new subjects, each time resulting in new and therefore independent data sets, then Type I error = 5% for each of these independent experiments.

Now, I hope I have introduced you to the issue of Type I error at the level of a single comparison and the idea of an experiment, holding Type I error-rates at 5% across all hypotheses to be evaluated in an experiment. You may wonder why anyone would make this mistake now. Actually, people make this "mistake" all the time and in some fields like evaluating gene expression for microarray data, this error was the norm, not the exception (see, for example, discussion of this in Jeanmougin et al 2010).

To conclude, if one does multiple tests on the same experiment, whether it is t-tests or some other test for that matter, then our subsequent tests are related. This is what we mean by independence in statistics — and there are many ways that nonindependence may occur in experimental research. For example, we introduced the concept of **pseudoreplication**, when observations are treated as if they are independent, but they are not (see Chapter 5.2). The "multiple comparison problem" specifically refers to the lack of independence when all the data set from a single experiment is parsed into lots of separate tests. Philosophically, there must be a logical penalty — and that is reflected in the increase in Type I error.

Clearly something must be done about this!

ANOVA is a solution

One possible solution for getting the correct experiment-wise error rate: adjust for differences in probability for multiple comparisons with the t-test. We used post-hoc tests presented above: you could evaluate the tests after accounting for the change in Type I error. This is what is done in many cases. For example, in genomics. In the early days of gene expression profiling by microarray, it was common to see researchers conduct t-tests for each gene. Since microarrays can have thousands of genes represented on the chip, then these researchers were conducting thousands of t-tests, arranging the t-tests by P-value and counting the number of p-values less than 5% and declaring that the differences were statistically significant.

This error didn't stand long, and there are now many options available to researchers to handle the "multiple comparisons" problem (some probably better than others, research on this very much an ongoing endeavor in biostatistics). The Bonferroni correction was

an available solution, largely replaced by the **Holm** (aka **Holm-Bonferroni**) **method** (Holm 1979). Recall that the Bonferroni correction judged individual p-values statistically significant only if they were less than α/C, where, again, α is the family-wise error rate (e.g., $\alpha = 0.05$) and C was the number of comparisons. The Holm method orders the C p-values from lowest to highest rank. The method then evaluates lowest p-value, if less than α/C, then reject hypothesis for that comparison. Proceed to next p-value, if less than α/C, then reject hypothesis for that comparison, and so on until no more comparisons are less than α/C.

A MUCH better alternative is to perform a single analysis that takes the multiple-comparisons problem into account: single-factor ANOVA, also called the one-way ANOVA, plus the **post-hoc tests** with error correction. We introduce one-way ANOVA in the next section. Post-hoc tests are discussed in Chapter 12.6.

Questions

1. You should be able to define and distinguish how Bonferoni correction, Dunnett's test, and Tukey's test methods protect against inflation of Type I error.
2. What will be the experiment-wise error rate for an experiment in which there are only two treatment groups?
3. Experiment-wise error rate may also be called _____ error rate.
4. List and compare the three described posthoc approaches to correct for multiple comparison problem,
5. Glycophosphate-tolerant soy bean is the number one GMO (genetically modified organism) crop plant worldwide. Glycophosphate is the chief active ingredient in Roundup, the most widely used herbicide. A recent paper examined "food quality" of the nutrient and elemental composition of plants drawn from fields which grow soy by organic methods (no herbicides or pesticides) and GMO plants subject to herbicides and pesticides. A total of 28 individual t-tests were used to compare the treatment groups for different levels of nutrients and elements (e.g., vitamins, amino acids, etc.,); the authors concluded that 10 of these t-tests were statistically significant at Type I error rate of 5%. Discuss the approach to statistical inference by the authors of this report; include correct use of the terms experiment-wise and individual-wise in your response and suggest an alternative testing approach if it is appropriate in your view.
6. In a comparative study about resting metabolic rate for eleven species of mammals, how many pairwise species comparisons can the study test?
7. In Chapter 4.2 we introduced a data set from an experiment. The experiment looked at DNA damage quantified by measuring qualities in a Comet Assay including the Tail length, the percent of DNA in the tail, and olive moment. The data set is copied to end of this page. In the next chapter I'll ask you to conduct the ANOVA on this experiment. For now, answer the following questions.
   a. What is the response variable?
   b. Explain why there is only one response variable.
   c. How many treatment variables are there?
   d. Why is this an ANOVA problem? Include as part of your explanation a statement of the null hypothesis.

Data used in this page

comet assay dataset

Data set, comet assay

Table 12.1.1. Comet assay data.

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	10	9.7732	2.1501
Copper-Hazel	6	4.8381	0.9676
Copper-Hazel	6	3.981	0.836
Copper-Hazel	16	12.0911	2.9019
Copper-Hazel	20	15.3543	3.9921
Copper-Hazel	33	33.5207	10.7266
Copper-Hazel	13	13.0936	2.8806
Copper-Hazel	17	26.8697	4.5679

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	30	53.8844	10.238
Copper-Hazel	19	14.983	3.7458
Copper	11	10.5293	2.1059
Copper	13	12.5298	2.506
Copper	27	38.7357	6.9724
Copper	10	10.0238	1.9045
Copper	12	12.8428	2.5686
Copper	22	32.9746	5.2759
Copper	14	13.7666	2.6157
Copper	15	18.2663	3.8359
Copper	7	10.2393	1.9455
Copper	29	22.6612	7.9314
Hazel	8	5.6897	1.3086
Hazel	15	23.3931	2.8072
Hazel	5	2.7021	0.5674
Hazel	16	22.519	3.1527
Hazel	3	1.9354	0.271
Hazel	10	5.6947	1.3098
Hazel	2	1.4199	0.2272
Hazel	20	29.9353	4.4903
Hazel	6	3.357	0.6714
Hazel	3	1.2528	0.2506

Rat lung cells treated with Hazel tea extract and exposed to copper metal. Tail refers to length of the comet tail, TailPercent is percent DNA damage in tail, and Olive moment refers to Olive's (1990), defined as the fraction of DNA in the tail multiplied by tail length.