Section 7: ANOVA II - Advanced Concepts

TF: Reagan Rose (rrose@g.harvard.edu)

October 20, 2015

Review of Multiple Comparisons

Consider an experiment to determine differences between three or more treatment groups. This is a generalization of a t-test, which compares two groups. How should we proceed? One way would be to perform all possible t-tests. But this raises the problem of **multiple comparisons**, in which when we're comparing k groups there are $k(k-1)/2$ pairwise comparisons to make, and therefore the chance that we see at least one significant result is much greater than α, our nominal Type I error rate.

Statisticians have developed many multiple comparisons procedures, and we discuss a few of them below. Just be aware that the reason we are talking about any of these methods is because **we want to use the most powerful, but least error-prone tests** to analyze our data. If we didn't, we could just use the t-test for everything!

Tukey HSD

Purpose

The purpose of Tukey's HSD test is to determine which groups in the sample differ. In particular, while an ANOVA can tell *whether* different groups in a sample differ, Tukey's HSD can help clarify *which* groups in the sample specifically have significant differences.

Procedure

Tukey's HSD test works through defining a value known as the **Honest Significant Difference** (HSD). The value of the HSD represents the minimum distance between two group means that must exist before the difference between the two groups is considered statistically significant.

Test Statistic:

$$Q_{obs} = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ (General)}, \quad Q_{obs} = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{S_p\sqrt{\frac{2}{n}}} \text{ (Equal sample sizes)}$$

Reference Distribution:

The reference distribution we are using to determine whether our observed test statistic, Q_{obs}, is feasible under the null hypothesis is called the **Studentized Range Distribution**.

$$Q^* = q(1 - \alpha, I, n - I).$$

Strength

Like other post-hoc tests, the Tukey HSD test is weak. That means that if a test of the difference between two specific means were designed *prior to collecting data*, that test would have more power (i.e. it would more likely yield significant results than Tukey's HSD test). However, Tukey's method is less conservative than the Bonferroni correction.

Multi-way ANOVA

Purpose

Recall that in the one-way ANOVA, we are comparing **multiple groups** on **one variable**. For example, we want to know whether the mean height of Harvard students is different between freshmen, sophomores, juniors, seniors, etc. But what if we want to compare **multiple groups** on **multiple variables**? Perhaps we also want to know whether height is different based on major. To do this, we use **Multi-way ANOVA**, which is just an extension of the regular ANOVA that allows us to compare groups on multiple variables all at once.

Why don't we just run separate one-way ANOVAs? Or a bunch of separate t-tests? Because a Multi-Way ANOVA will have more **power**, and can measure **interactions**. Intuitively, the less tests we do, the less chance we have of picking up a **spurious effect.**

Main Effects vs. Interactions

Main Effects

Main effects are what we've been working with so far in this class. A independent variable X has a **main effect** on a dependent (response) variable Y if X is significantly associated with Y without considering any other variables.

Interactions

However, sometimes variables only have an effect on the outcome **after accounting for other variables**. For example, diet and exercise are known to be predictive of weight loss. It is known (in general), that diet has a **moderate main effect**, exercise has a **small main effect**, and the combination of diet and exercise has a **large interaction effect**. So if I just diet or

just exercise, I may lose some weight. But if I diet and exercise I will probably lose a lot more than doing either on their own. We account for this phenomenon in statistics by including **Interaction terms** in our analyses.

Procedure

To incorporate multiple main effects and some interaction effects in an ANOVA, we simply **add the terms we think have main effects**, and **multiply the terms we think have interaction effects.** In R, this looks like

model1 = aov(height classyear + major, data=data) (main effects only)

model2 = aov(height classyear*major, data=data) (full factorial)

Kruskal-Wallis Test

Purpose

Recall in our t-testing framework, we have other testing options to use when our assumptions of normality and independence are strongly violated (i.e. permutation tests, rank-sum, signed-rank, etc.). In the ANOVA framework, if our assumptions of normality and equal variances are strongly violated, we can use the **Kruskal-Wallis Test**.

Procedure

1. Rank all the combined data, ignoring groups. Average the ranks of any ties.
2. Compute the observed test statistic

$$K_{obs} = (N-1)\frac{\sum_{i=1}^{I} n_i(\bar{R}_i - \bar{R})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2}$$

3. Compare your observed statistic K_{obs} against a known critical value, from the χ^2_{I-1} distribution.

$$\text{(two-sided) p-value} = P(|\chi^2_{1-\alpha/2,I-1}| \geq K_{obs}).$$

Exercises

Question 1. Getting familiar with Tukey

Research shows that some students learn most effectively with constant background noise, as opposed to unpredictable sound or no sound at all. We sample 15 Harvard students and divide them into 3 groups of $n = 5$. Group 1 is asked to study in a room with background noise, group 2 study in a room with noise that changes periodically, and group 3 study in silence. After studying, all students take a test over the material. A table with summary statistics and an ANOVA table for this data are shown below:

Group	n	Mean	Std. Dev
No sound	5	3.5	1.12
Sporadic sound	5	1.8	1.09
Constant sound	5	1	0.71

```
                    Df Sum Sq Mean Sq F value  Pr(>F)
factor(data$groups)  2   16.3   8.150   8.288 0.00548 **
Residuals           12   11.8   0.983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Summarize the results of the ANOVA and discuss whether the assumptions are met.

(b) Compare the 3 groups pairwise using contrast tests. Make sure to use an appropriate adjustment method to correct for the multiple comparisons.

(c) Now compare the groups using Tukey's method.

(d) Do the different tests give you different results? Explain why. Which test do you believe?

Question 2. A big huge messy simulation study

To study the performance and power of different tests, let's do a simulation. Simulate three groups of $n = 5$ based on the following procedures. Then run 10,000 simulations comparing the three groups, each time recording the pvalue from (i) Bonferroni corrected pairwise t-tests, (ii) Bonferroni corrected contrast tests, and (iii) the Tukey HSD test.

1. Simulate three groups each with $n = 5$ observations from a $N(1, 1)$ distribution and calculate the p-values from 1,000 simulations. This simulation should show you how the tests perform **under the null**.

2. Now simulate 2 groups from $N(1, 1)$ and one group from $N(5, 1)$. This should show you the power of these tests when there is a **large effect** for one group.

3. Now simulate 2 gruops from $N(1, 1)$ and one group from $N(2, 1)$. This should show you the power of the tests when there is a **small effect** for one group.