Chapter 27 - One-Way ANOVA: Comparing Several Means

Last Time:

- Review Activity

Today:

- Chapter 27 - ANOVA
- Quiz #7 - Inference for μ

Next Time:

- Completing Chapter 27
- Homework assigned

FiveThirtyEight

Politics Sports Science & Health **Economics** Culture Politics Podcast: Wh

SEP. 12, 2014 AT 7:37 AM

The Economic Guide To Picking A College Major

By Ben Casselman

Filed under Higher Education

Get the data on GitHub

Students walk across the campus of UCLA in Los Angeles. KEVORK DJANSEZIAN / GETTY IMAGES

The millions of American college students heading back to campus this month face a grim reality: A college degree is no guarantee of economic success. But through their choice of major, they can take at least some steps toward boosting their odds.

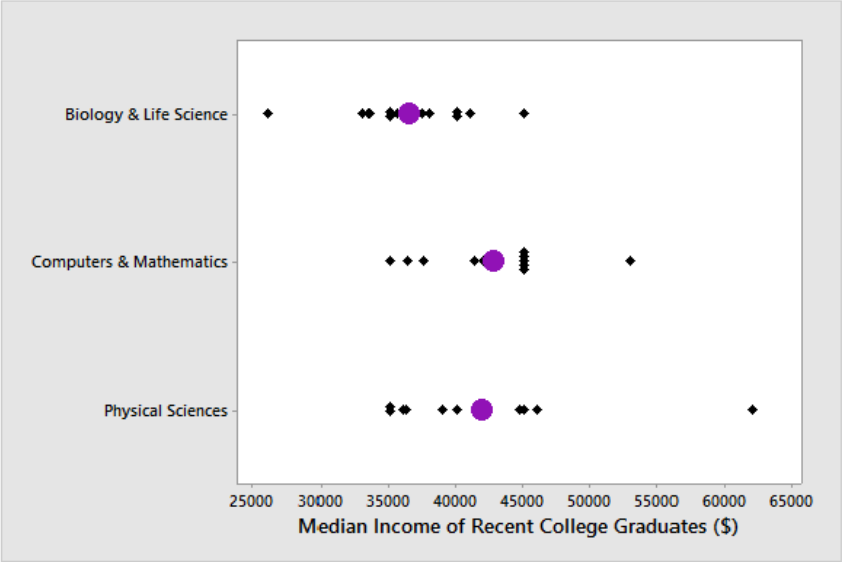
The link between education and earnings is notoriously fraught, with cause and effect often difficult to disentangle. But a look at detailed data on college graduates by major reveals some clear messages: Don't be pre-med if you aren't planning to go to medical school; don't assume that all "STEM" — science, technology, engineering and math — majors are the same; and if you study drama, be prepared to wait tables.

https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/

Salaries by Major

"Politicians love to tout the importance of science, technology, engineering and math majors. But when it comes to earnings, the "S" majors don't really belong with the "TEM" ones. Engineering majors are nearly all high-paying. So are most computer and math majors, and math-heavy sciences like astrophysics. But many sciences, particularly the life sciences, pay below the overall median for recent college graduates."

We are going to test the claim that *not all STEM majors are the same.*



We have data from the 2014 American Community Survey. Each dot represents the median salary of recent graduates (up to 5 years after graduation), from a particular major classified within a major category.

Thus the sample consists of graduates sampled from different majors and assigned a particular category, thus providing us with three independent "treatment" groups.

1. Biology & Life Science
2. Computers & Mathematics
3. Physical Sciences

Given this data, what type of question or analysis do you think we should be pursuing?

Major Category	N	Mean	StDev	Minimum	Maximum
Biology & Life Science	14	50821	6219	35000	60000
Computers & Mathematics	11	66273	11791	50000	92000
Physical Sciences	10	62400	2409	55000	80000

ANOVA: ANalysis Of VAriance

In Chapter 9, we defined a **factor** as a variable that can take one of several **levels** used to differentiate one group from another (e.g. treatment group).

We are interested in experiments or studies with **one-way** or **completely randomized designs**. This is when individuals are (randomly) assigned into the different levels of the factor. In other words, there is only *one-way* to group the variable.

When the variable studied within each factor level is **quantitative** in nature, we use Analysis of Variance (ANOVA) to determine whether the population means of each factor level are equal.

Specifically, we are studying the One-Way ANOVA, but to learn about more advanced designs, take MATH 217!

Salaries by Major - ANOVA Hypotheses

1. Define parameters:
   - μ_1 = true mean of the median salary of recent graduates who majored in Biology & Life Sciences.
   - μ_2 = true mean of the median salary of recent graduates who majored in Computers & Mathematics.
   - μ_3 = true mean of the median salary of recent graduates who majored in Physical Sciences.
2. The **null hypothesis** is that the group population means are all the same. That is:
3. The **alternative hypothesis** is

ANOVA Hypotheses

We are comparing multiple population means.

Note: these hypotheses imply that we have a **multiple comparisons test**.

Steps for Multiple Comparisons Tests

ANOVA is an example of a **multiple comparisons test**, since the hypotheses are "many-sided".

Statistical methods for multiple comparisons usually have two steps:

1. An **overall test** to see if there is good evidence of any differences among the parameters that we want to compare.
2. If there are differences, then a detailed **follow-up analysis** to decide which of the parameter(s) differ and to estimate how large the differences are.

Conditions for ANOVA

There are *two sets of conditions* for an analysis of variance. One set needs to be checked prior to conducting the analysis in order to make sure the procedure is valid and one set needs to be checked after the analysis to make sure the model fit the data appropriately.

Before the analysis, verify the following:

- Condition #1: We have I independent randomized samples, one from each of the I populations.

  Important: this condition is met if either a randomized sample was used to recruit subjects into each factor level, or subjects were randomly assigned to one of the factor levels.

- Condition #2: Each distinct population has a Normal distribution with unknown mean μ_i.

  ANOVA is robust against departures from Normality. This means that even if we have 4 or 5 observations in a group, then provided there is no *extreme* skewness in each sample, it is okay to proceed, but you **must check each sample distributions.**

- Condition #3: Each population has the same standard deviation, σ.

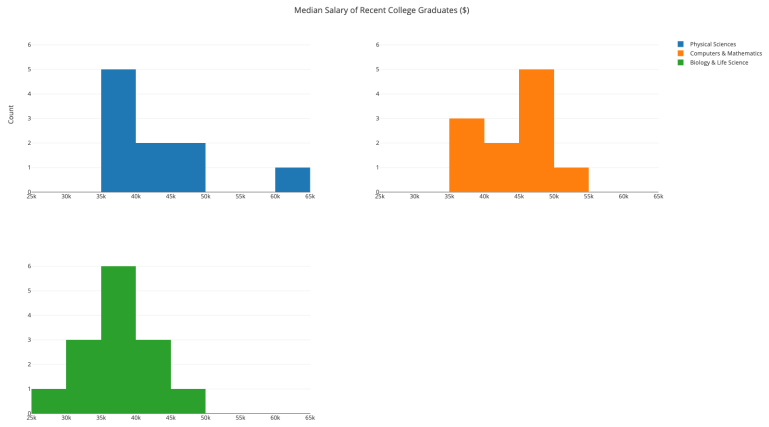
  Need to check by comparing the sample standard deviations.

After the analysis, verify the following:

- The *residuals* are Normally distributed with mean zero.
- The variance of the *residuals* is constant and they show no pattern.

Salaries by Major - Conditions

- Condition #1: We have I independent SRS's, one from each of the I populations.
  - Population 1: Students with majors in Biology & Life Science.
  - Population 2: Students with majors in Computers & Mathematics.
  - Population 3: Students with majors in Physical Science.
- Condition #2: Each distinct population has a Normal distribution with unknown mean μ_i.

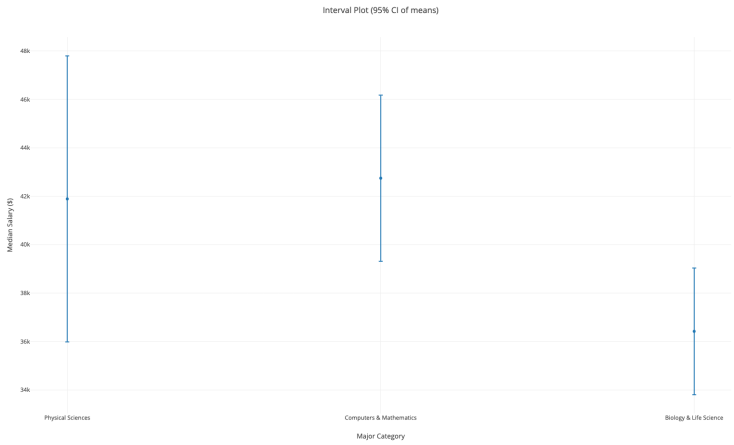


Check if the observations in each independent sample are approximately symmetric.

There does not appear to be extreme skewness in any of the three major categories/samples, therefore the condition is met.

- Condition #3: Each population has the same standard deviation, σ.

  → As usual the population standard deviation is unknown to us, so we look to the samples.



  → To determine if the sample standard deviations are approximately the same, we use the following rule of thumb:

  For the salaries by major:

Major Category	N	Mean	StDev
Biology & Life Science	14	36421	4529
Computers & Mathematics	11	42745	5109
Physical Sciences	10	41890	8252

ANOVA Test Statistic

The sample means estimate the respective population means, so we ask:

> Are the differences among the sample means due to **true differences** among the population means (alternative hypothesis), or merely due to sampling variability (null hypothesis)?

To answer this, we:

> Consider the variation *between* sample means relative to the variability *within* each sample.

The test statistic for comparing multiple independent population means is called the F-statistic.

THE ANOVA *F* STATISTIC

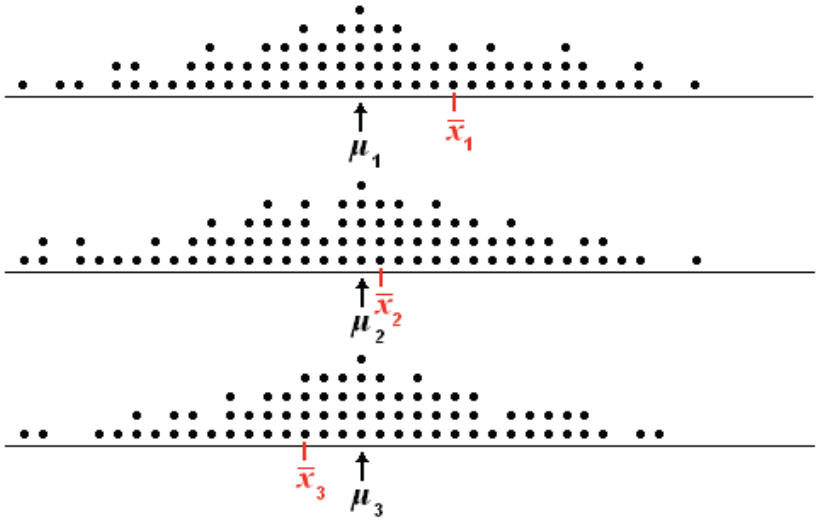
The **analysis of variance *F* statistic** for testing the equality of several means has this form:

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

Understanding the F-statistic

Each dotplot below represents the sampling distribution for the sample mean of group i. Thus, each one is approximately Normal and centered around the population mean of group i. We need to examine is the *variability* within each group (common $\sigma/\sqrt{n}$) compared to the *variability* across the groups.

SCENARIO # 1:

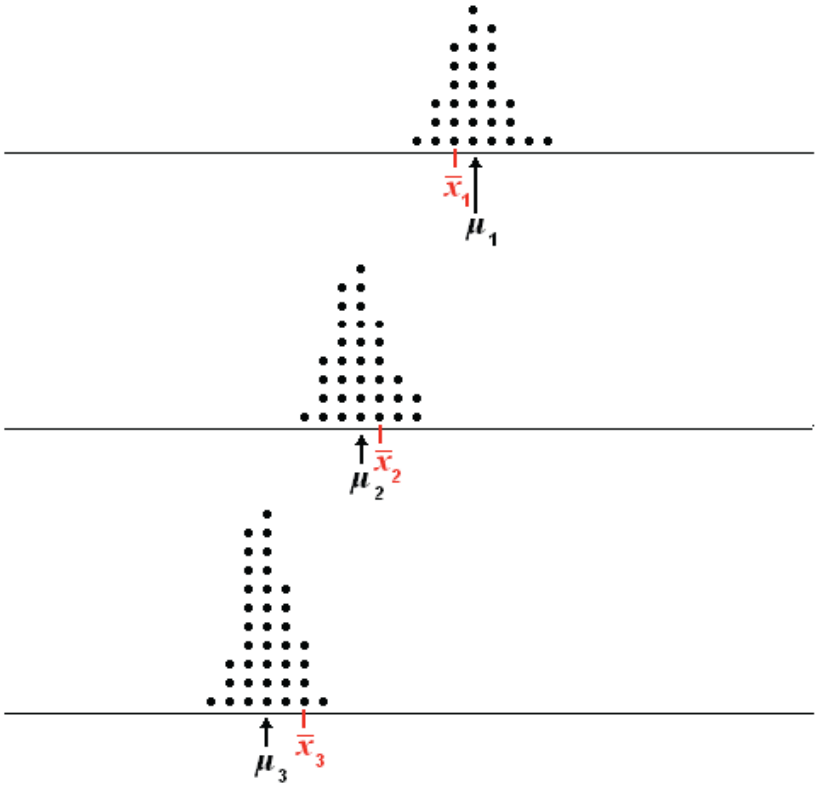


High sampling variability within each group, relative to the variation across the $\bar{x}$'s.

i.e. the sample means are pretty similar relative to the large variability in the sampling distributions.

Plausible that the $\bar{x}$'s could have come from sampling distributions with the same μ's.

SCENARIO # 2:



Low sampling variability within each group, relative to the variation across the $\bar{x}$'s.

i.e. the sample means are pretty dissimilar relative to the small variability in the sampling distributions.

Not plausible for the $\bar{x}$'s to have come from sampling distributions with the same μ's.

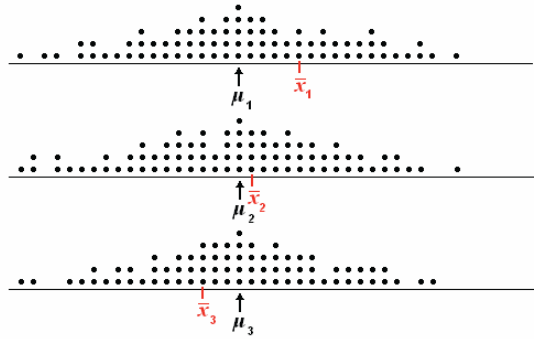
THE ANOVA F STATISTIC

The **analysis of variance F statistic** for testing the equality of several means has this form:

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

SCENARIO # 1:

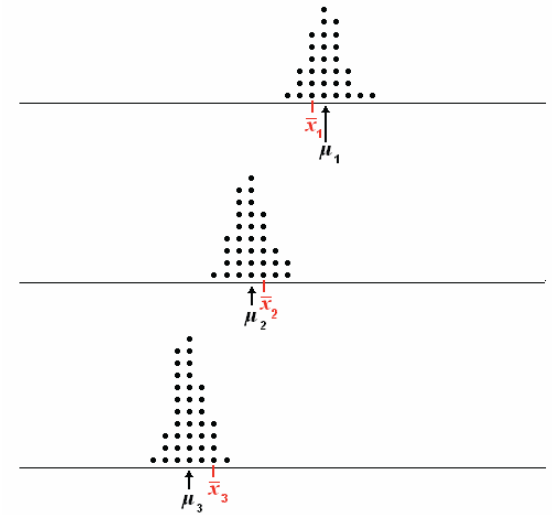
$$F = \frac{\text{Variation between the group } \bar{x}\text{'s}}{\text{Variation within the groups}}$$



Small variability across the groups
Large variability within the groups
$\Rightarrow$ Small F-statistic $\Rightarrow$ Evidence supporting H_0.

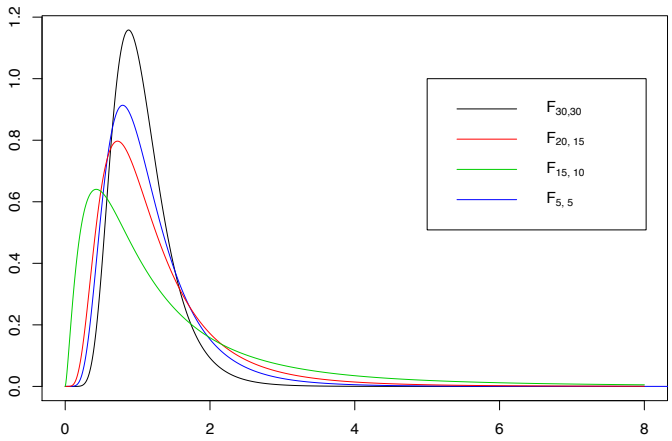
SCENARIO # 2:

$$\mathbf{F} = \frac{\text{Variation between the group } \bar{x}\text{'s}}{\text{Variation within the groups}}$$



Large variability across the groups
Small variability within the groups
$\Rightarrow$ Large F-statistic $\Rightarrow$ Evidence against H_0.

The F-statistic follows an F-distribution which is a positive, right skewed distribution, parameterized by 2 different degrees of freedom (numerator and denominator).



To determine the degrees of freedom, we must first note that

1. There are _____ levels of the factor (number of groups/means defined).
2. There are _____ total individuals in the study.

   $\rightarrow n_1$ in group 1, n_2 in group 2, $\ldots$, n_I in group I.

   $\rightarrow$ Such that, $n_1 + n_2 + \cdots + n_I = N$

Then, under H_0, the F-test statistic follows the F-distribution with _____ numerator degrees of freedom and _____ denominator degrees of freedom.

ANOVA Table

	df	SS	MS	F Score	P Value
Treatment					
Errors					
Total					

- Sum of Squares (SS): The sum of the squared differences for the treatment or error.
- Mean Squares Treatment (MS_T): The variability between the factor means.
- Mean Squares Errors (MS_E): Estimate of the pooled variability within each sample.

Salaries by Major - ANOVA Results & Conclusion

Recall

We are trying to determine if all STEM majors are equal. Or in other words if the true mean of the median salaries is the same for each major category.

	df	SS	MS	F Score	P Value
Treatment		297891584			
Errors					
Total		1438331429			

TABLE D F distribution critical values (*continued*)

		Degrees of freedom in the numerator					
	p	1	2	3	4	5	6
20	.100	2.97	2.59	2.38	2.25	2.16	2.09
	.050	4.35	3.49	3.10	2.87	2.71	2.60
	.025	5.87	4.46	3.86	3.51	3.29	3.13
	.010	8.10	5.85	4.94	4.43	4.10	3.87
	.001	14.82	9.95	8.10	7.10	6.46	6.02
25	.100	2.92	2.53	2.32	2.18	2.09	2.02
	.050	4.24	3.39	2.99	2.76	2.60	2.49
	.025	5.69	4.29	3.69	3.35	3.13	2.97
	.010	7.77	5.57	4.68	4.18	3.85	3.63
	.001	13.88	9.22	7.45	6.49	5.89	5.46
50	.100	2.81	2.41	2.20	2.06	1.97	1.90
	.050	4.03	3.18	2.79	2.56	2.40	2.29
	.025	5.34	3.97	3.39	3.05	2.83	2.67
	.010	7.17	5.06	4.20	3.72	3.41	3.19
	.001	12.22	7.96	6.34	5.46	4.90	4.51
100	.100	2.76	2.36	2.14	2.00	1.91	1.83
	.050	3.94	3.09	2.70	2.46	2.31	2.19
	.025	5.18	3.83	3.25	2.92	2.70	2.54
	.010	6.90	4.82	3.98	3.51	3.21	2.99
	.001	11.50	7.41	5.86	5.02	4.48	4.11

Conclusion:

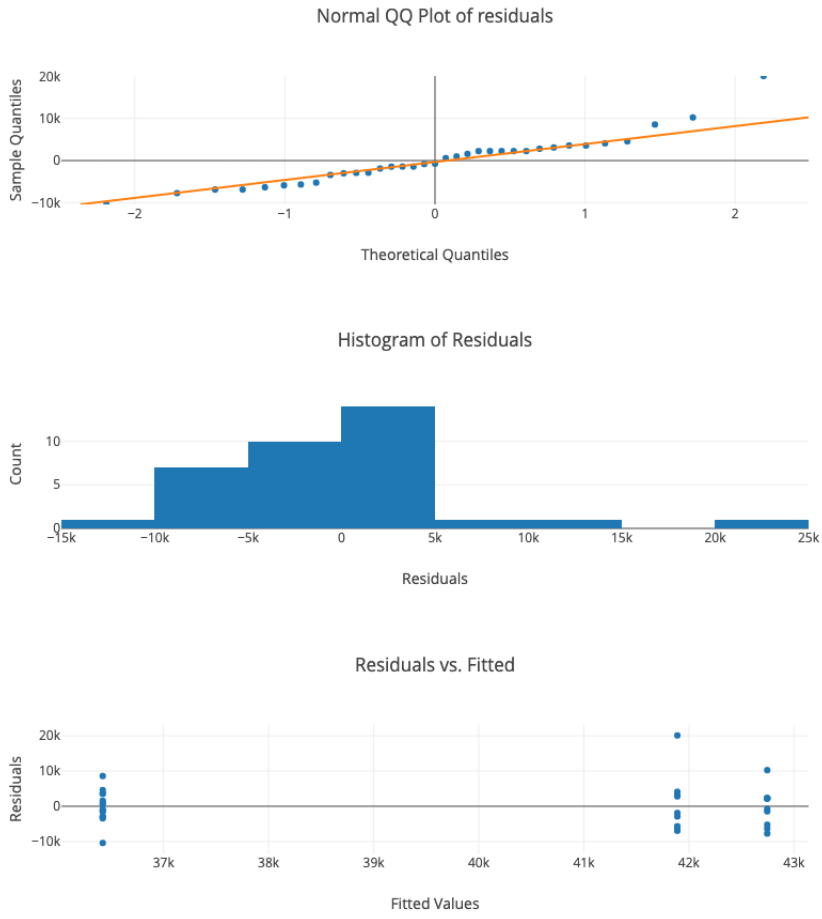
After Conditions

Once you have fit the model, you need to check the **residuals** to make sure they are "behaving." If the conditions surrounding the residuals are not met, then the conclusion to your test is not meaningful.

Recall from Chapter 5

Residuals vs. Fitted	Normal QQ Plot
We want a random scatter of the points above and below zero. This indicates that fitting an ANOVA is appropriate, and the residuals have constant variance.	We want the points to be close to the line. This indicates that the residuals are normally distributed with a mean of 0.

Salaries by Major - After Conditions



Recall:

An ANOVA is a **multiple comparisons test**. Thus, we must complete a follow-up analysis once we reject H_0, to determine *which mean(s) are actually different.*

Multiple Comparisons - Follow up Analysis

The ANOVA hypotheses state

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$
$$H_a : \text{at least one mean differs}$$

- This is not the same as saying $H_a : \mu_1 \neq \mu_2 \neq \ldots \neq \mu_I$.
- The alternative hypotheses is saying that *at least two means* differ.

Salary by Major - Follow Up

The ANOVA hypotheses state

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : \text{at least one mean differs}$$

That is, any of the following could support the alternative hypothesis

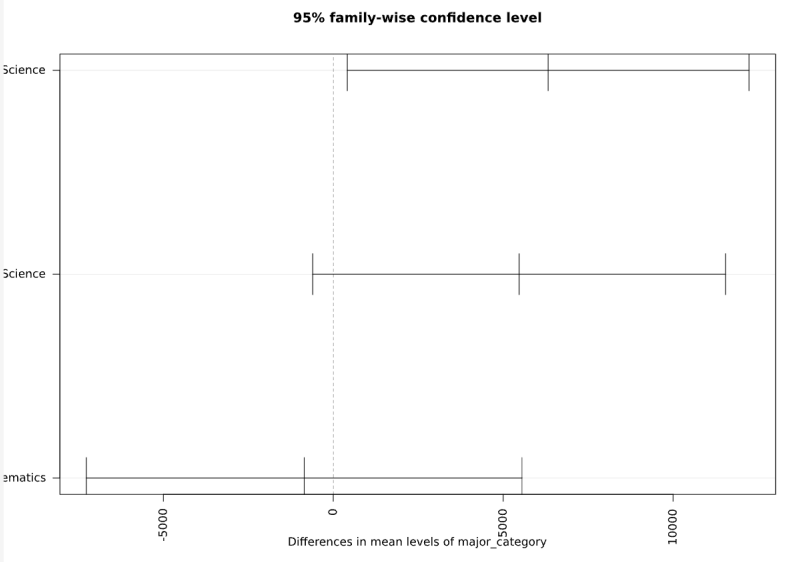
- $\mu_1 = \mu_2 \neq \mu_3$
- $\mu_1 \neq \mu_2 = \mu_3$
- $\mu_1 \neq \mu_2 \neq \mu_3$

So which is it?

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = median ~ major_category)

$major_category
                                                diff        lwr       upr
Computers & Mathematics-Biology & Life Science 6324.0260   413.2883 12234.764
Physical Sciences-Biology & Life Science       5468.5714  -605.4126 11542.555
Physical Sciences-Computers & Mathematics      -855.4545 -7265.2673  5554.358
                                                   p adj
Computers & Mathematics-Biology & Life Science 0.0339495
Physical Sciences-Biology & Life Science       0.0842998
Physical Sciences-Computers & Mathematics      0.9425406
```



ANOVA Procedure

1. Define population means and state hypotheses.
2. Verify (before) conditions.
3. Complete ANOVA table.
4. Make a conclusion in the context of the problem.
5. Verify (after) conditions.
6. If H_0 was rejected, complete follow-up analysis.

Executive Summary