1 Analysis of Variance

1.1 Review

The following is a partial list of statistical methods that we have discussed:
1. mean
2. median
3. mode
4. standard deviation
5. z-score
6. percentile
7. coefficient of variation
8. scatter plot
9. histogram
10. boxplot
11. normal-quantile plot
12. confidence interval for mean
13. confidence interval for difference in means
14. confidence interval for proportion
15. confidence interval for difference in proportions
16. one sample mean test
17. two independent sample mean test
18. match pair test
19. one sample proportion test
20. two sample proportion test
21. test of homogeneity
22. test of independence
23. linear correlation coefficient & test
24. regression

For each situation below, which method is most applicable?

• If it’s a hypothesis test, what are the null and alternative (state in words and mathematically).
• If it’s a graphical method, describe what you would be looking for.

Question 1. A college admission board wants to determine if student A with a GPA of 3.4 at a tough public school with an average student GPA of 2.2 and standard deviation of 0.5 is better than student B with a GPA of 3.9 from an easy private school with an average student GPA of 3.3 and standard deviation of 0.4.

Question 2. A scientist wants to determine if the length of penguin beaks are normally distributed.

Question 3. A bad comedian who can’t find work decides to do a study to determine if the style of shoe a person wears depends on hair color.

Question 4. A student is writing a report on Tucson, they need to find a number.
to describe how much the maximum daily temperature varies throughout the year.

**Question 5.** A professor wants to compare the distribution of Exam II scores for two sections of the same statistics class.

### 1.2 Introduction

**Example 1.** A clinical psychologist wishes to test three methods (A, B, C) for reducing hostility levels in university students to see if there is any real difference between the methods. A certain psychological test (HLT) was used

---

**ANOVA: ANALYSIS OF VALUE**

**Is Your Research Worth Anything?**

Developed in 1912 by geneticist R.A. Fisher, the Analysis of Variance is a powerful statistical tool designed to test the significance of one’s work.

Significance is determined by comparing one’s research with the Dull Hypothesis:

$$H_0 : \mu_1 = \mu_2$$

where,

- $H_0$ : the Dull Hypothesis
- $\mu_1$ : significance of your research
- $\mu_2$ : significance of a monkey typing randomly on a typewriter in a forest where no one hears it.

The test involves computation of the $F$ ratio:

$$F' = \frac{\sum (people \ who \ care \ about \ your \ research)}{world \ population}$$

This ratio is compared to the $F$ distribution with $l-1, N_0$ degrees of freedom to determine a $p$ value. A low $p$ value means you’re on to something good (though statistically improbable).

**Type I/II Errors**

- **Type I:** You incorrectly believe your research is not Dull. Good luck graduating.
- **Type II:** No conclusions can be made. Good luck graduating.

Of course, this test assumes both Independence and Normality on your part. Neither of which is likely true, which means it’s not your problem.

---

Figure 1: ANOVA Comic (Credit: Jorge Cham [http://phdcomics.com](http://phdcomics.com))
to measure the degree of hostility (higher scores indicate greater hostility). Eleven students participated in the experiment and the results are shown in the table below:

<table>
<thead>
<tr>
<th>score</th>
<th>method</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.00</td>
<td>A</td>
<td>MALE</td>
</tr>
<tr>
<td>83.00</td>
<td>A</td>
<td>FEMALE</td>
</tr>
<tr>
<td>78.00</td>
<td>A</td>
<td>MALE</td>
</tr>
<tr>
<td>68.00</td>
<td>A</td>
<td>FEMALE</td>
</tr>
<tr>
<td>83.00</td>
<td>A</td>
<td>MALE</td>
</tr>
<tr>
<td>54.00</td>
<td>B</td>
<td>FEMALE</td>
</tr>
<tr>
<td>78.00</td>
<td>B</td>
<td>MALE</td>
</tr>
<tr>
<td>71.00</td>
<td>B</td>
<td>FEMALE</td>
</tr>
<tr>
<td>82.00</td>
<td>C</td>
<td>MALE</td>
</tr>
<tr>
<td>95.00</td>
<td>C</td>
<td>FEMALE</td>
</tr>
<tr>
<td>88.00</td>
<td>C</td>
<td>MALE</td>
</tr>
</tbody>
</table>

**Question 6.** What is the null and alternative hypothesis?

**Question 7.** How could we test this using methods we know?

**The problem with multiple tests**

When we run a hypothesis test, there is a chance that we can make an error in our decision (Type I or Type II error). If we use multiple tests, we won’t know the overall Type I or Type II error.

We need a single test to answer our question and determine the overall Type I and Type II errors.

**Definitions**

Analysis of variance is often employed in statistical experiments.

**Definition 1.1**

**Experimental unit.**

Anthony Tanbakuchi  
MAT167
Analysis of Variance

An experimental unit is the object on which a measurement (or measurements) is taken.
Example: The student.

**FACTOR.**
A factor is an **independent variable** whose levels are controlled and varied by the experimenter. It’s the explanatory or predictor variable.
Example: Method for reducing hostility.

**LEVEL.**
A level is the intensity setting of a factor.
Example: Hostility methods A, B, C

**TREATMENT.**
A treatment is a specific combination of factor levels that an experimental unit receives.
Example: Student 1 gets method A, Student 2 gets method B, ...

**RESPONSE VARIABLE.**
The response variable is the dependent variable being measured by the experimenter.
Example: Student’s HLT test score for hostility.

*Example 2.* A group of students in a statistics class are randomly divided into 3 groups. Each group is fed a different breakfast and then given the same final exam for the class. Group 1 gets pickles for breakfast, group 2 gets ketchup for breakfast, and group 3 gets brown sugar for breakfast. A researcher wished to determine if final exam scores are effected by the type of breakfast a student eats.

**Question 8.** What is the experimental unit in this study?

**Question 9.** What is the factor that is varied in this study?

**Question 10.** What are the levels for the factor

**Question 11.** What does a treatment consist of?

**Question 12.** What is the response variable?

**Methods for testing relationships between variables**

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Dependent Variable</th>
<th>Quantitative</th>
<th>Categorical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation / Regression</td>
<td>ANOVA</td>
<td>Homogeneity</td>
<td></td>
</tr>
</tbody>
</table>

Anthony Tanbakuchi

MAT167
1.3 One-Way Analysis of Variance

**USE**

Often used to help answer:
1. Are means the same across groups of data?
2. Is at least one mean different across groups of data?

**COMPUTATION**

**Definition 1.6**

One-Way Analysis of Variance ANOVA.

Analysis of variance is a generalization of the two sample \( t \) test for many samples. For more than two samples, a one-way ANOVA analyzes sample variance to test the null hypothesis that all the sample means are equal. The alternative hypothesis is that at least one mean is different.

The response (dependent) variable must be quantitative. The predictor (independent) variable is generally categorical\(^1\).

**Versatility of ANOVA**
- Can study the relationship between a response variable and one or more explanatory (predictor) variables.
- Does not require any assumptions about the nature of the relationship (doesn’t have to be linear). It’s much more general than regression.

**Definition 1.7**

One-Way ANOVA Hypothesis Test.

Equality of means test for more than two samples:

**Requirements**
- (1) simple random samples,
- (2) independent samples,
- (3) only one factor,
- (4) populations are normal and have equal variance (loose).\(^2\)

**Null Hypothesis** \( H_0 : \mu_1 = \mu_2 = \cdots = \mu_n \) all treatments have the same mean (all treatments the same).

**Alternative Hypothesis** \( H_a \) at least one treatment mean is different.

**CONCEPTUAL DESCRIPTION**

Conceptual Test statistic

Compare the variation between treatment means to variation within treatments.

\[
F = \frac{\text{variation between factor level means}}{\text{pooled variation within factor levels}}
\]  
(1)

\(^1\)ANOVA is also frequently used when the predictor variable quantitative but the nature of the statistical relation is unknown. In these cases, the quantitative variable is broken up into categories and analysis of variance is used to detect if a relationship exists.

\(^2\)To check the assumption of equal variance: Levene Test for Equality of Variances or Bartlett’s Test. Bartlett’s give better performance if the data is normal.
Conceptually the value of $F$ can be interpreted as follows:
- If $F$ is small ($F \leq 1$) the sample means don’t vary significantly. Therefore treatment has no effect on the means.
- If $F$ is large ($F \gg 1$) the sample means vary with statistical significance. Therefore the treatment does effect the mean.

**MATHEMATICAL DESCRIPTION**

**Sum of Squares**

For one factor with $k$ levels, $n_j$ measurements in a level, $\bar{x}_j$ and $s^2_j$ factor level mean and variance, $\bar{x}$ overall mean, $N$ total measurements:

Overall variation of the data:

$$SS(\text{total}) = \sum_{i=1}^{N} (x_i - \bar{x})^2$$  \hspace{1cm} (2)

Variation between the $k$ sample means:

$$SS(\text{treatment}) = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$$  \hspace{1cm} (3)

Pooled variation of the $k$ samples:

$$SS(\text{error}) = \sum_{j=1}^{k} (n_j - 1) s^2_j$$  \hspace{1cm} (4)

Algebraically:

$$SS(\text{total}) = SS(\text{treatment}) + SS(\text{error})$$  \hspace{1cm} (5)

The $SS(\text{error})$ is assumed variation common to all the populations under study.

**Average variation**

By dividing the sum of squares by their corresponding number of degrees of freedom we can find the **mean squares**.

Mean overall variation of the data:

$$MS(\text{total}) = \frac{SS(\text{total})}{N - 1}$$  \hspace{1cm} (6)

Mean variation between the $k$ sample means:

$$MS(\text{treatment}) = \frac{SS(\text{treatment})}{k - 1}$$  \hspace{1cm} (7)

Mean pooled variation of the $k$ samples:

$$MS(\text{error}) = \frac{SS(\text{error})}{N - k}$$  \hspace{1cm} (8)
Test statistic

To test the hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$, the test statistic is:

$$F = \frac{MS(\text{treatment})}{MS(\text{error})}$$

having an $F$ distribution with two degrees of freedom:

$$df_1 = k - 1$$

$$df_2 = N - k$$

Like the $\chi^2$ distribution, we always find the area to the RIGHT of the test statistic.

R Command

<table>
<thead>
<tr>
<th>$F$ distribution CDF:</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>pf(F', df1, df2)</code></td>
</tr>
<tr>
<td>Finds the area to the left of $F'$ on the $F$ density, $p = P(F &lt; F') = F(F')$, given the numerator and denominator degrees of freedom $df_1$ and $df_2$.</td>
</tr>
</tbody>
</table>

The $F$ distribution

A WORKED OUT EXAMPLE: “MANUAL”

Example 3. A clinical psychologist wishes to test three methods (A, B, C) for reducing hostility levels in university students to see if there is any real difference between the methods. A certain psychological test (HLT) was used to measure the degree of hostility (higher scores indicate greater hostility). Eleven students participated in the experiment and the results are shown in the table below:
Analysis of Variance

<table>
<thead>
<tr>
<th>score</th>
<th>method</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75.00</td>
<td>MALE</td>
</tr>
<tr>
<td>2</td>
<td>83.00</td>
<td>FEMALE</td>
</tr>
<tr>
<td>3</td>
<td>78.00</td>
<td>MALE</td>
</tr>
<tr>
<td>4</td>
<td>68.00</td>
<td>FEMALE</td>
</tr>
<tr>
<td>5</td>
<td>83.00</td>
<td>MALE</td>
</tr>
<tr>
<td>6</td>
<td>54.00</td>
<td>FEMALE</td>
</tr>
<tr>
<td>7</td>
<td>78.00</td>
<td>MALE</td>
</tr>
<tr>
<td>8</td>
<td>71.00</td>
<td>FEMALE</td>
</tr>
<tr>
<td>9</td>
<td>82.00</td>
<td>MALE</td>
</tr>
<tr>
<td>10</td>
<td>95.00</td>
<td>FEMALE</td>
</tr>
<tr>
<td>11</td>
<td>88.00</td>
<td>MALE</td>
</tr>
</tbody>
</table>

Step 0: known info : Enter data:
R: A = c(75, 83, 78, 68, 83)
R: B = c(54, 78, 71)
R: C = c(82, 95, 88)

Enter number of levels k, level sample sizes n_j, and level means \( \bar{x}_j \):
R: k = 3
R: n = c(length(A), length(B), length(C))
R: n
[1] 5 3 3
R: x.bar = c(mean(A), mean(B), mean(C))
R: x.bar
[1] 77.400 67.667 88.333

Make one more vector \( x \) with all data and find total sample size \( N \) and overall mean \( \bar{x} \):
R: x = c(A, B, C)
R: N = length(x)
R: N
[1] 11
R: x.bar.bar = mean(x)
R: x.bar.bar
[1] 77.727

Step 1: Test ANOVA 1-Way
Step 2: Requirements (1) simple random samples, (2) independent samples, (3) only one factor, (4) populations are normal and have equal variance (loose).

Step 3: Hypothesis \( H_0 : \mu_A = \mu_B = \mu_C \) all means equal, \( H_a : \) at least one mean is different.

Step 4: Significance \( \alpha = 0.05 \)
Step 5: p-value manually
Find \( MS(\text{total}) \) using equations 2 and 6:
R: SStotal = sum((x - x.bar.bar)^2)
R: MSTotal = SStotal/(N - 1)
1.3 One-Way Analysis of Variance

R: MSTotal
   [1] 118.82

Find \( MS(\text{treatment}) \) using equations 3 and 7:
R: SStreatment = sum(n * (x.bar - x.bar.bar)^2)
R: MStreatment = SStreatment / (k - 1)
R: MStreatment
   [1] 320.82

Easiest way to find \( MS(\text{error}) \) is to solve equation 5 for \( SS(\text{error}) \) (rather than use equation 4) then use equation 8:
R: SSerror = SStotal - SStreatment
R: MSerror = SSerror / (N - k)
R: MSerror
   [1] 68.317

Step 5: (cont...) Now find the test statistic:
R: F = MStreatment/MSerror
R: F
   [1] 4.6961

Finally, find the \( p \)-value.
R: p.val = 1 - pf(F, df1 = k - 1, df2 = N - k)
R: p.val
   [1] 0.044765

Step 6: Decision  Reject \( H_0 \) since \( p \)-value \( \leq \alpha \).

Step 7: Conclusion Our data supports the conclusion at least one of the methods effect the mean hostility level.

The probability of observing our sample data assuming the null hypothesis that \( \mu_A = \mu_B = \mu_C \) is true (indicating the different methods have no significant effect on the mean hostility level) is only 0.0448. Since this is unlikely, our sample data supports the alternative hypothesis that at least one treatment has a different mean.

ANOVA IN R

1-WAY ANOVA IN R:
results=aov(depVarColName~indepVarColName, data=tableName)
summary(results)
boxplot(depVarColName~indepVarColName, data=tableName)

**R Command**

**Format of data for analysis**
R (and most other statistics packages) need the data in a table with each row representing a treatment listing each factor level and the response measurement.
Best to use Excel or some other spreadsheet application to enter the data. Then **export the sheet as a CSV file and load it into R**.

**R Command**

```
LOADING CSV DATA:
MyTable=read.csv(file.choose())
```

When you run this command, a file browser will open. Choose the CSV file with your data and then R will load it.

*MyTable* the name of the table to create in R that will contain your data.

A WORKED OUT EXAMPLE: R

*Example 4.* A clinical psychologist wishes to test three methods (A, B, C) for reducing hostility levels in university students to see if there is any real difference between the methods. A certain psychological test (HLT) was used to measure the degree of hostility (higher scores indicate greater hostility). Eleven students participated in the experiment and the results are shown in the table below:

<table>
<thead>
<tr>
<th>subject</th>
<th>hostility score</th>
<th>method</th>
<th>subject gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>A</td>
<td>male</td>
</tr>
<tr>
<td>2</td>
<td>83</td>
<td>A</td>
<td>female</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
<td>A</td>
<td>male</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>A</td>
<td>female</td>
</tr>
<tr>
<td>5</td>
<td>83</td>
<td>A</td>
<td>female</td>
</tr>
<tr>
<td>6</td>
<td>54</td>
<td>B</td>
<td>male</td>
</tr>
<tr>
<td>7</td>
<td>74</td>
<td>B</td>
<td>female</td>
</tr>
<tr>
<td>8</td>
<td>71</td>
<td>B</td>
<td>male</td>
</tr>
<tr>
<td>9</td>
<td>79</td>
<td>C</td>
<td>female</td>
</tr>
<tr>
<td>10</td>
<td>95</td>
<td>C</td>
<td>male</td>
</tr>
<tr>
<td>11</td>
<td>87</td>
<td>C</td>
<td>female</td>
</tr>
</tbody>
</table>

**Step 0: known info** : Enter data: I have already loaded the data into the table called *hostility*.

**Step 1: Test ANOVA 1-Way**
1.4 Summary

One-Way ANOVA Hypothesis Test
Equality of means test for more than two samples:

Requirements (1) simple random samples, (2) independent samples, (3) only one factor, (4) populations are normal and have equal variance (loose).
Null Hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_n$ all treatments have the same mean (all treatments the same).

Alternative Hypothesis $H_a$ at least one treatment mean is different.

**Test** in R

```r
results = aov(depVarColName ~ indepVarColName, data=tableName)
summary(results)
boxplot(depVarColName ~ indepVarColName, data=tableName)
```

Be sure to specify the actual table name, factor column name, and response variable column name.

### 1.5 Additional Examples

*Example 5.* A clinical psychologist wishes to determine if gender effects hostility levels in university students. A certain psychological test (HLT) was used to measure the degree of hostility (higher scores indicate greater hostility).

Load the ANOVA data file provided with this week’s HW. The hostility table has the data for this problem.

**Question 13.** What is the null and alternative hypothesis?

**Question 14.** What is the response variable & the factor (independent variable)?

**Question 15.** Run the ANOVA, what is the $p$-value and formal decision? (Check: $p$-val=0.353)

**Question 16.** What is the final conclusion?
Make a boxplot to visualize the data!