Introductory Statistics Lectures

Analysis of Variance

1-Way ANOVA: Many sample test of means

ANTHONY TANBAKUCHI
DEPARTMENT OF MATHEMATICS
PIMA COMMUNITY COLLEGE

(Compile date: Tue May 19 14:51:36 2009)

Contents

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 a 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 wring 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

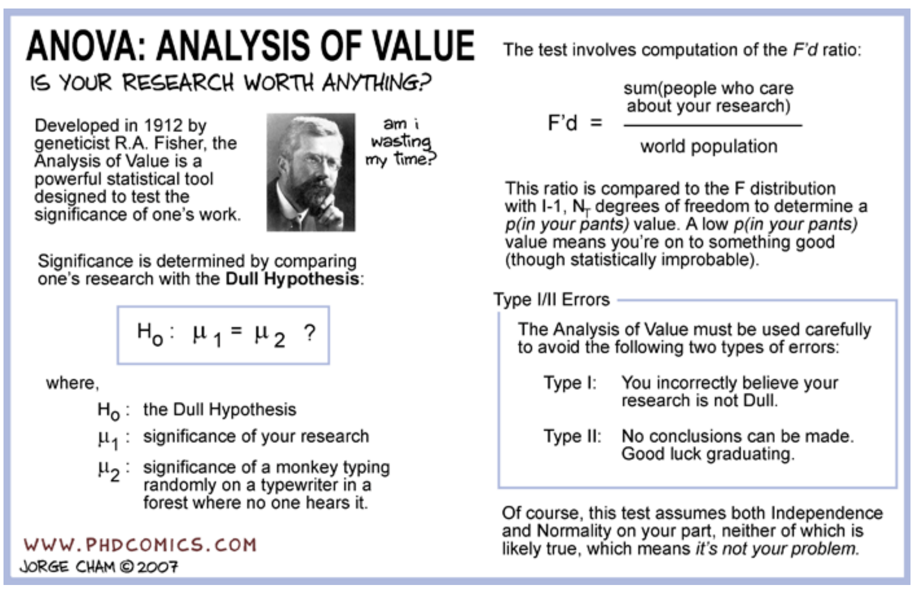


Figure 1: ANOVA Comic (Credit: Jorge Cham http://phdcomics.com)

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 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:

	score	method	gender
1	75.00	A	MALE
2	83.00	A	FEMALE
3	78.00	A	MALE
4	68.00	A	FEMALE
5	83.00	A	MALE
6	54.00	B	FEMALE
7	78.00	B	MALE
8	71.00	B	FEMALE
9	82.00	C	MALE
10	95.00	C	FEMALE
11	88.00	C	MALE

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.

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

Example: The student.

FACTOR. Definition 1.2

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. Definition 1.3

A level is the intensity setting of a factor.

Example: Hostility methods A, B, C

TREATMENT. Definition 1.4

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. Definition 1.5

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

		INDEPENDENT VARIABLE	
		quantitative	categorical
DEPENDENT	quantitative	correlation / regression	**ANOVA**
VARIABLE	categorical	—	homogeneity

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.

$$\boxed{F = \frac{\textbf{variation between}\text{ factor }\textbf{level means}}{\text{pooled }\textbf{variation within}\text{ factor }\textbf{levels}}} \tag{1}$$

[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_j^2 factor level mean and variance, $\bar{\bar{x}}$ overall mean, N total measurements:

Overall variation of the data:

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

Variation between the k sample means:

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

Pooled variation of the k samples:

$$SS(\text{error}) = \sum_{j=1}^{k} (n_j - 1) s_j^2 \tag{4}$$

Algebraically:

$$SS(\text{total}) = SS(\text{treatment}) + SS(\text{error}) \tag{5}$$

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:

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

Mean variation between the k sample means:

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

Mean pooled variation of the k samples:

$$MS(\text{error}) = \frac{SS(\text{error})}{N - k} \tag{8}$$

Test statistic

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

$$\boxed{F = \frac{MS(\text{treatment})}{MS(\text{error})}} \tag{9}$$

having an F distribution with two degrees of freedom:

$$(\text{numerator}) \qquad df_1 = k - 1 \tag{10}$$

$$(\text{denominator}) \qquad df_2 = N - k \tag{11}$$

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

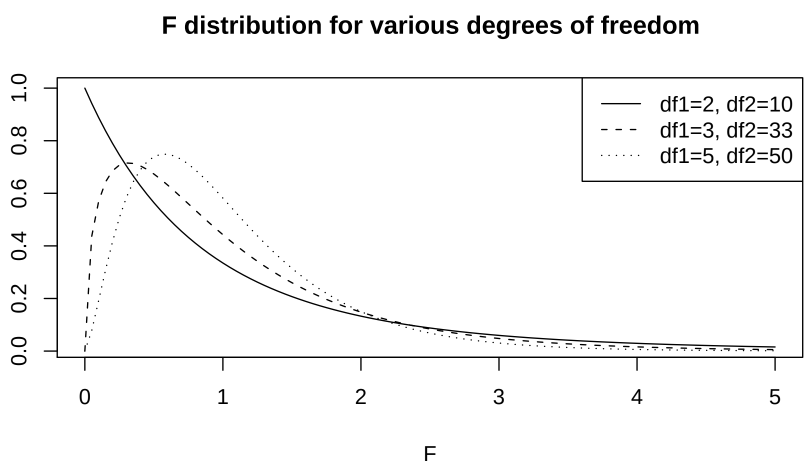
R Command

F distribution CDF:

`pf(F', df1, df2)`

Finds the area to the left of F' on the F density, $p = P(F < F') = F(F')$, given the numerator and denominator degrees of freedom `df1` and `df2` .

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:

	score	method	gender
1	75.00	A	MALE
2	83.00	A	FEMALE
3	78.00	A	MALE
4	68.00	A	FEMALE
5	83.00	A	MALE
6	54.00	B	FEMALE
7	78.00	B	MALE
8	71.00	B	FEMALE
9	82.00	C	MALE
10	95.00	C	FEMALE
11	88.00	C	MALE

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_i with all data and find total sample size N and overall mean $\bar{\bar{x}}$

```
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(total) using equations 2 and 6:

```
R: SStotal = sum((x - x.bar.bar)^2)
R: MSTotal = SStotal/(N - 1)
```

```
R: MSTotal
[1] 118.82
```

Find MS(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(error) is to solve equation 5 for SS(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

R COMMAND

1-WAY ANOVA IN R:

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

tableName name of table that contains your data.
depVarColName column name that is your response variable (dependent variable).
indepVarColName column name that is your factor (independent variable).

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.

subject	hostility score	method	subject gender
1	73	A	female
2	83	A	male
3	76	A	female
4	68	A	male
5	80	A	female
6	54	B	male
7	74	B	female
8	71	B	male
9	79	C	female
10	95	C	male
11	87	C	female

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**.

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.

R COMMAND

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:

```
R:  hostility
    score  method  gender
1      75       A    MALE
2      83       A  FEMALE
3      78       A    MALE
4      68       A  FEMALE
5      83       A    MALE
6      54       B  FEMALE
7      78       B    MALE
8      71       B  FEMALE
9      82       C    MALE
10     95       C  FEMALE
11     88       C    MALE
```

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

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 using R

Run ANOVA and view results: Data is in the table `hostility` , factor (dependent variable) is in column `method` , response variable is in column `scores` :

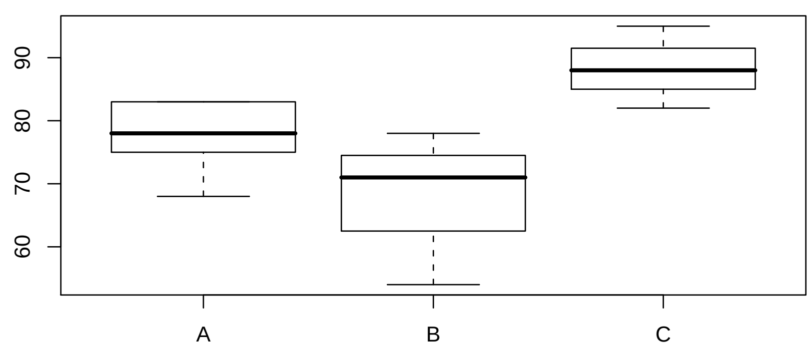
```
R: res = aov(score ~ method, data = hostility)
R: summary(res)
            Df Sum Sq Mean Sq F value Pr(>F)
method       2    642     321     4.7  0.045 *
Residuals    8    547      68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

We can visualize the effect of the different methods using **box plots**:

```
R: boxplot(score ~ method, data = hostility)
```



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

```
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!