Kerala School of Mathematics
Course in Statistics for Scientists
Analysis of Variance

T.Krishnan
Strand Life Sciences, Bangalore

6 March 2010

Examples and Purpose

Examples:

- Differential effects of say three fertilizers on the yield of a crop.
- Has a new traffic regulation reduced the pollution level?
- Is this new drug for hypertension better than the current ones?
  - Is the differential effect, if any is in the same direction and of the same magnitude for different life styles?

Purpose:

- to examine the differences between means of three or more groups
- to examine the combined effects of different factors (say, drugs, life styles)
- do factors interact? (say, new drug reduces blood pressure in active subjects but not in sedentary ones)

A Drug Trial

- Two established drugs A and B and a new experimental drug C for hypertension are to be compared in a clinical trial.
- Fifteen subjects are available for the trial.
- Reduction in systolic blood pressure is to be measured after the administration of the drug for six months.
- The fifteen subjects are divided into three groups **at random** each of size 5.
- Each group is administered one of the drugs.
- One subject in group B dropped out during the six months.
- Can we simply ignore this subject and carry on with the analysis?
- Since the dropping out was for reasons not connected with the trial, it was ignored and the data on the remaining 14 subjects were taken up for analysis.

Data for ANOVA

The following data relate to a medical study where A and B are established Ayurvedic and allopathic drugs respectively, and C is a new experimental drug for hypertension. The numbers are reduction in systolic blood pressure after six months of drug use.

Subject Sr no.	Drug	Reduction systolic BP
1	A	5.2
2	A	4.7
3	A	5.1
4	A	4.9
5	A	5.3
6	B	5.2
7	B	4.3
8	B	4.7
9	B	4.4
10	C	6.7
11	C	7.3
12	C	7.0
13	C	7.5
14	C	6.4

Some Questions

- Let us look at the means of the three groups:

Drug	No. of subjects	Mean
A	5	5.04
B	4	4.65
C	5	6.98

- From this can we conclude that drug C is better than drugs A and B?
- and drug A is better than drug B?
- Can differences of this magnitude arise even if the three drugs are not different?
- These differences arise from a combination of subject-to-subject differences and the drug effects.
- Notice that there is a difference even among subjects administered the same drug.
- The difference between two subjects from different groups is a combination of drug effect and the subject-to-subject difference.
- ANOVA asks the question: How much is the difference between the three means the result of the drugs and how much due to subject-to-subject difference?
- Is the difference between the means of the same order as the subject-to-subject difference and so cannot be attributed to the different drugs?

Basis of ANOVA

- In this case, ANOVA decomposes the total variation into two components.
- One component is due to subject-to-subject variation.
- Another component is due to the differential effects of drugs.
- Then it compares the two components and asks if they are of the same magnitude taking into account chance variation.
- If they are, then the conclusion is that the drug effects are not different; if they are not, then the effects are different.
- More precisely, it computes the chance that the relative value of the two components is as much as the observed value if there was no difference between the drug effects.
- This chance is the p-value you are familiar with.
- If the p-value is small then the hypothesis of no differential effect is rejected; else accepted.
- Appropriate calculations leading to the p value is where statistical calculations come in and the assumptions under which they are made.

ANOVA Calculations

- Let us rearrange the data as follows:

Drug A	Drug B	Drug C
5.2	5.2	6.7
4.7	4.3	7.3
5.1	4.7	7.0
4.9	4.4	7.5
5.3		6.4
5.04	4.65	6.98

- The last row gives the means of the three drugs.

Subject-to-Subject Variation

- The five subjects in Drug A column have been treated under identical conditions.
- So the variation in these five numbers is measured by the average squared deviations from their mean:

$$\frac{(5.2-5.04)^2+(4.7-5.04)^2+(5.1-5.04)^2+(4.9-5.04)^2+(5.3-5.04)^2}{4}=\frac{S_A^2}{4}.$$

- You know from previous discussions why we use a denominator of 4 and not 5 (n-1 and not n).
- Similar computations for the other two drugs.
- Each of these quantities measures the subject-to-subject variation in each drug.
- One of the assumptions made in the standard ANOVA computation is that the subject-to-subject variation is the same for all the drugs. This is often denoted by σ^2.
- This variation is often called (Experimental) Error or Residual.

Residual Mean Square

- Now it is necessary to pool these three subject-to-subject variations into a single number.
- Since they are based on (slightly) different numbers of observations (5, 4, 5), we need to combine them with these weights.
- That is given by

$$\hat{\sigma}^2 = (S_A^2 + S_B^2 + S_C^2)/(4+3+4)$$

- $S_A^2 + S_B^2 + S_C^2$ is called the Residual Sum of Squares (RSS or ESS) and $(S_A^2 + S_B^2 + S_C^2)/(4+3+4)$ the Residual Mean Square (RMS or EMS).
- The number (4+3+4) = 11 is called the **degrees of freedom** of this Sum of Squares.
- Thus

$$\text{Mean Square} = \frac{\text{Sum of Squares}}{\text{degrees of freedom}}$$

Between Drug Mean Square

- We have thus captured the variation **within** each column (drug).
- Now the remaining variation in the data is between the drugs, that is, between the three means 5.04, 4.65, 6.98.
- From your previous knowledge, you know that the variance of the mean based on n observations is $\frac{\sigma^2}{n}$.
- Thus the variances of the drug (column) means are $\frac{\sigma^2}{5}$, $\frac{\sigma^2}{4}$, and $\frac{\sigma^2}{5}$ respectively.
- If the three group means are the same (hypothesis H_0) then the squared differences between these means and the overall mean of 5.62 are $(5.04-5.62)^2, (4.65-5.62)^2, (6.98-5.62)^2$ combined suitably provide an estimate of the same σ^2.
- The appropriate way of combining them is $\frac{5(5.04-5.62)^2+4(4.65-5.62)^2+5(6.98-5.62)^2}{2} = \frac{\text{BSS}}{2} = BMS$, because they estimate $\frac{\sigma^2}{5}$, $\frac{\sigma^2}{4}$, and $\frac{\sigma^2}{5}$.
- The denominator of 2 (not 3) is for the same reason as '$(n-1)$' and not n in estimating variance. respectively.

Hypothesis Test

- A hypothesis test is conducted by comparing the variability between group means to variability within group observations.
- Under the hypothesis H_0 of equality of means, the ratio $F = \frac{\text{BSS}}{\text{RSS}}$ should be close to 1.
- Thus the hypothesis H_0 is to be rejected if $F >> 1$ (much greater than 1).
- What the threshold value for rejection is decided on the basis of the distribution of the statistic F under the hypothesis.
- This distribution has been worked out as the F distribution with degrees of freedom (2, 11) (for this case).
- If there are k groups with a total of n observations in all groups together, then the degrees of freedom of this F distribution is $(k-1, n-k)$.

Assumptions of ANOVA

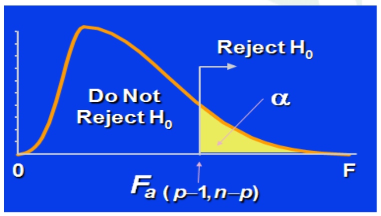
- The derivation of this distribution is made under certain assumptions, which are:
  - **independence**: the subjects form independent random samples from the corresponding populations.
  - **normality**: the characteristic observed is distributed normally in each population.
  - **homogeneity of variance**: the variance in each population is the same.
- This test can also be carried out for equality of two means for which you carried out the two-sample t-test under the same assumptions.
- In that case the F distribution will have (1, n-2) degrees of freedom and $F(1, n-2) \equiv t(n-2)^2$ and so it is the same test.
- However, the t-test cannot be carried out to test equality of > 2 means.

ANOVA Table

- The Analysis of Variance computations are set out neatly in an ANOVA table.
- Here is the SYSTAT ANOVA for our drugs data.
- The shaded area in the picture is the rejection region, say with $\alpha = 0.05$.
- If the observed F-value falls in that region, reject the hypothesis.

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-Ratio	p-Value
DRUG$	14.694	2	7.347	53.520	0.000
Error	1.510	11	0.137		



Multiple Comparisons

- The table shows that the three means are very different much more than what could be attributed to chance variations.
- In this situation, we would be interested in finding out where the differences lie.
- Is it due to the new drug C being different from established drugs A and B?
- Is it only due to A and B being different and C not different from either A or B?
- It will be useful to test A vs B, A vs C, B vs C, A+B vs C or anything else that seems interesting and useful to figure out where the differences lie?

Multiple Comparison Tests

- Also called **post hoc tests**.
- If you conduct several tests simultaneously, you reject some by chance even if there is no difference between the means.
- Multiple comparison tests make adjustments to guard against this situation.
- There are many procedures available.
- We illustrate one of them from SYSTAT.
- Look at the p-values and the confidence intervals.
- If the confidence interval includes 0, then it is not significant; else, it is.
- This shows that the new drug C is different from A and B, but A and B are not different.

Tukey's Honestly-Significant-Difference Test

DRUG$(i)	DRUG$(j)	Difference	p-Value	95% Confidence Interval	
				Lower	Upper
A	B	0.390	0.299	-0.281	1.061
A	C	-1.940	0.000	-2.573	-1.307
B	C	-2.330	0.000	-3.001	-1.659

A Second Factor

- Suppose it is envisaged that the effect of the drug might depend upon the lifestyle of the subject—active (B) or sedentary (S).
- This introduces another factor into the trial.
- It is possible that the differential effects of the drugs may be different—in direction and/or magnitude—in B and S subjects.
- If this happens we say that there is **interaction** between drugs and life style.
- Let us look at a table of means in a two-way classified table.

DRUG$(rows) by LIFESTYLE$(columns)

	B	S
A	5.100	5.000
B	4.750	4.550
C	6.967	7.000

Interaction

- Before you look at the difference between drugs, you should examine if there is interaction.
- If interaction is present then it is meaningless to look at the drug effect.
- For, suppose the difference between drugs are in different directions and of the same magnitude in the two life styles, then they will cancel out and the drug effect will be insignificant.
- But that may not be the true state of affairs—there may be drug effect but different drugs may be better for different life styles.
- It is important that the analysis brings out this aspect of the data.
- This analysis is done by a two-way analysis of variance, where the interaction effects and the effects of drug and life style are presented, based on computations based on the same logic.
- It is also helpful to graphically examine the presence of interaction by an interaction plot.
- From the plot and the ANOVA table, it appears that there is no interaction and the drug differences are significant and the life style differences are not.

Interaction Plot

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-Ratio	p-Value
DRUG$	14.244	2	7.122	39.115	0.000
LIFESTYLE$	0.027	1	0.027	0.146	0.712
DRUG$*LIFESTYLE$	0.030	2	0.015	0.083	0.921
Error	1.457	8	0.182