

Data Analysis and Visualisation

Lecture week 5

Analysis of Variance (ANOVA)

Introduction

- Topics covered
 - Computational steps
 - Strength of Association
 - Post hoc Analyses
 - The Scheffé Method of Post hoc Analysis
- The ANOVA test allows us to compare the means of multiple samples

Introduction

- Earlier, t Test determines if the means of two samples differ to a statistically significant degree
 - Comparing more than two samples?
 - multiple t Tests between each pair of samples not possible
 - We would have to do $K(K-1)/2$ tests where K is the number of samples
 - multiple t Tests increases likelihood of Type I errors
 - When, level of significance = 0.05
- Type I error = 5% (5 times out of 100)

Type I errors and Multiple *t* Tests

- With multiple *t* Tests, probability (*p*) of a Type I error is

$$p = 1 - (1 - \alpha)^c$$

- Where α is the alpha level and *c* is the number of comparisons
- For three samples $c = 3(3-1)/2 = 3$ If α is 0.05 then

$$p = 1 - (1 - 0.05)^3 = 1 - 0.93^3 = 0.14 = 14\%$$

- chance of making a Type I error increases to 14%
- To avoid this, *ANalysis Of VAriance* (**ANOVA**) is used
 - test statistical significance in the differences of means between more than two samples
 - compares variance between groups to variance within each group

Example

Group 1	Group 2	Group 3
22	6	8
16	10	6
17	13	4
18	13	5
	8	2
	4	

- Suppose we have three groups
 - Groups 1 and 2 have been given a treatment of some sort
 - Group 3 is the control group
- We want to know ‘has the treatment had any effect?’

Degrees of Freedom

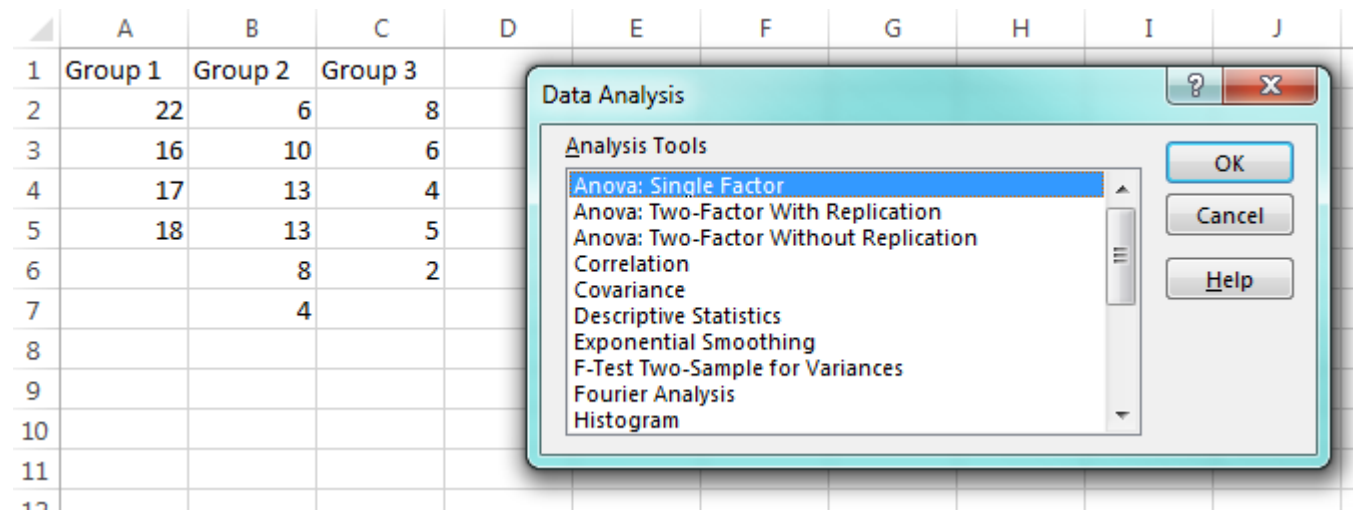
- The variance of a population is calculated from

$$SD_x^2 = \frac{(X - \bar{X})^2}{N - 1}$$

- Where $N-1$ is known as the *degrees of freedom* in the sample
 - $N-1$ is used rather than N to avoid bias in the sample
- This is the number of things that are allowed to vary in your sample
 - Suppose we have ten scores that add to 100
 - Nine of those scores can add up to anything at all, but the tenth score is fixed by the total of those nine
 - If the first nine scores add up to 89 then the tenth must be 11
 - The group of ten scores has only 9 degrees of freedom, so $df = 9$

ANOVA in Excel

- Calculating ANOVAs by hand is long and prone to mistakes so we're going to let Excel do the heavy lifting 😊
- ANOVA is part of the Data Analysis Toolpak and has three options
 - We want the Anova – Single Factor tool



ANOVA in Excel

- Input the data range for analysis, input a location for the data to be output, and tick the labels box if appropriate

- Excel to

The screenshot shows an Excel spreadsheet with data for an ANOVA test. The data is organized into three columns: Group 1, Group 2, and Group 3. The values for Group 1 are 22, 16, 17, and 18. The values for Group 2 are 6, 10, 13, and 13. The values for Group 3 are 8, 6, 4, and 2. The dialog box 'Anova: Single Factor' is open, showing the input range as '\$A\$1:\$C\$7', grouped by 'Columns', with 'Labels in first row' checked, and an alpha level of 0.05. The output range is set to '\$E\$1'.

	A	B	C
1	Group 1	Group 2	Group 3
2	22	6	8
3	16	10	6
4	17	13	4
5	18	13	5
6		8	2
7		4	

Anova: Single Factor

Input
Input Range:
Grouped By: ☒ Columns ☐ Rows
☒ Labels in first row
Alpha:

Output options
☒ Output Range:
☐ New Worksheet Ply:
☐ New Workbook

ANOVA in Excel

	A	B	C	D	E	F	G	H	I	J	K
1	Group 1	Group 2	Group 3		Anova: Single Factor						
2	22	6	8								
3	16	10	6		SUMMARY						
4	17	13	4		<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
5	18	13	5		Group 1	4	73	18.25	6.916667		
6		8	2		Group 2	6	54	9	13.6		
7		4			Group 3	5	25	5	5		
8											
9											
10					ANOVA						
11					<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
12					Between Groups	402.9833	2	201.4917	22.23356	9.21E-05	3.885294
13					Within Groups	108.75	12	9.0625			
14											
15					Total	511.7333	14				
16											

- The first part of the results shows the descriptive statistics for each of the group samples
 - Count, sum, average (mean) and variance (standard deviation squared)

ANOVA in Excel

	A	B	C	D	E	F	G	H	I	J	K
1	Group 1	Group 2	Group 3		Anova: Single Factor						
2	22	6	8								
3	16	10	6		SUMMARY						
4	17	13	4		<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
5	18	13	5		Group 1	4	73	18.25	6.916667		
6		8	2		Group 2	6	54	9	13.6		
7		4			Group 3	5	25	5	5		
8											
9											
10					ANOVA						
11					<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
12					Between Groups	402.9833	2	201.4917	22.23356	9.21E-05	3.885294
13					Within Groups	108.75	12	9.0625			
14											
15					Total	511.7333	14				
16											

- The second part of the results shows the values for
 - *SS* – the sum of squares between groups and within groups
 - *df* – the degrees of freedom between groups and within groups
 - *MS* – the mean square value between groups and within groups
 - *F*, *F crit* – the equivalent of t_{obt} and t_{crit} in the t Test
 - *P-value* – the probability that the variation is due to chance

ANOVA in Excel

	A	B	C	D	E	F	G	H	I	J	K
1	Group 1	Group 2	Group 3		Anova: Single Factor						
2	22	6	8								
3	16	10	6		SUMMARY						
4	17	13	4		<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
5	18	13	5		Group 1	4	73	18.25	6.916667		
6		8	2		Group 2	6	54	9	13.6		
7		4			Group 3	5	25	5	5		
8											
9											
10					ANOVA						
11					<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
12					Between Groups	402.9833	2	201.4917	22.23356	9.21E-05	3.885294
13					Within Groups	108.75	12	9.0625			
14											
15					Total	511.7333	14				
16											

- This shows that the variation between groups is much larger than the variation within groups
- It not clear if significant differences between all groups or just between two groups
- Nor does it tell us how strong the effect is – how closely correlated the effects are

Exercise

- A researcher is interested in differences between blondes, brunettes and redheads in terms of introversion/extroversion. She selects random samples from a college campus, gives each subject a test of social introversion and comes up with the results given

Blondes	Brunettes	Redheads
5	3	2
10	5	1
6	2	7
2	4	2
5	3	2
3	5	3

1. State your hypotheses
2. Select α , your significance level
3. Calculate F
4. Find F_{crit}
5. Decide whether or not to reject the null hypothesis

Strength of association

- With large samples - finding differences in means that are statistically significant but not important in terms of practical significance
- F test does not,inform about the strength of association for the treatment effect.
- One way to do that - Find the proportion of total variability that can be accounted for by the treatment effects with omega squared

Strength of Association

$$\hat{\omega}^2 = \frac{SS_B - (K - 1)MS_W}{SS_T + MS_W}$$

Where

- SS_B = Sum of squares between groups
- SS_T = Sum of squares for the total
- MS_W = Mean square within groups
- K = number of groups
- In our example

$$\hat{\omega}^2 = \frac{SS_B - (K - 1)MS_W}{SS_T + MS_W}$$

Strength of Association

- SS_B = Sum of squares between groups
- SS_T = Sum of squares for the total
- MS_W = Mean square within groups
- K = number of groups
- In our example

$$\hat{\omega}^2 = \frac{SS_B - (K - 1)MS_W}{SS_T + MS_W} = \frac{402.98 - (3 - 1)9.06}{511.73 + 9.06}$$

$$\hat{\omega}^2 = \frac{402.98 - 18.12}{520.79} = \frac{384.86}{520.79} = 0.74 = 74\%$$

- 74% – of the variance is due to the difference between the groups

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	402.9833	2	201.4917	22.23356	9.21E-05	3.885294
Within Groups	108.75	12	9.0625			
Total	511.7333	14				

Exercise

- A company is thinking of replacing their current flavour of drink with two new flavours. They do a taste test with each of the new flavours and their current flavour and get their testers to rate the tastes out of ten. These are their results

New flavour 1	New flavour 2	Current flavour
8	2	7
9	3	6
10	4	5
10	2	8
6	1	10
8	3	6
9	2	7

1. State your hypotheses
2. Select α , your significance level
3. Calculate F
4. Find F_{crit}
5. Decide whether or not to reject the null hypothesis
6. If you reject the null hypothesis, calculate the strength of association

Post Hoc Analysis

- In the last problem we discovered that there was a significant difference between the means
 - Current flavour: 7 New flavour 1: 8.57 New flavour 2: 2.42
- In Anova method which pairs are statistically different is not clear
- *Post Hoc Analysis* determines pairs which show significant differences
 - Scheffé method (*C*)
 - Groups of equal or unequal N's

The Scheffé method of Post Hoc Analysis

- The Scheffé method requires the calculation of C_{obt} which is then compared against C_{crit} (which is calculated, not looked up)

$$C_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_W \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}$$

Where

- $\bar{X}_1, \bar{X}_2$ are the means of the two groups being compared
- N_1, N_2 are the N s of the two groups
- MS_W is the within-group mean square from your ANOVA

$$C_{crit} = \sqrt{(K - 1)(F_{crit})}$$

- Where, K is the number of groups
- F_{crit} is the critical value of the F test from your ANOVA

The Scheffé method of Post Hoc Analysis

- Current flavour v. flavour 1

$$C_{obt} = \frac{\bar{X}_C - \bar{X}_1}{\sqrt{MS_W \left(\frac{1}{N_C} + \frac{1}{N_1} \right)}} = \frac{7 - 8.57}{\sqrt{1.86 \left(\frac{1}{7} + \frac{1}{7} \right)}} = \frac{-1.57}{\sqrt{1.86 \left(\frac{2}{7} \right)}}$$

$$C_{obt} = \frac{-1.57}{\sqrt{0.53}} = \frac{-1.57}{0.73} = -2.15$$

- The minus sign doesn't matter in this case as we could have calculated $\bar{X}_C - \bar{X}_1$ or $\bar{X}_1 - \bar{X}_C$, giving us

$$C_{obt} = 2.15$$

- Next, compared with the critical value

$$C_{crit} = \sqrt{(K - 1)(F_{crit})} = \sqrt{(3 - 1)(3.55)} = \sqrt{7.1}$$

$$C_{crit} = 2.66$$

$$C_{obt} < C_{crit}$$

- Conclusion:

No significant difference between the current flavour and new flavour 1

The Scheffé method of Post Hoc Analysis

- Current flavour v. flavour 2

$$C_{obt} = \frac{\bar{X}_C - \bar{X}_2}{\sqrt{MS_W \left(\frac{1}{N_C} + \frac{1}{N_2} \right)}} = \frac{7 - 2.43}{\sqrt{1.86 \left(\frac{1}{7} + \frac{1}{7} \right)}} = \frac{4.57}{\sqrt{1.86 \left(\frac{2}{7} \right)}}$$
$$C_{obt} = \frac{4.57}{\sqrt{0.53}} = \frac{4.57}{0.73} = 6.26$$

- This now has to be compared with the critical value

$$C_{crit} = \sqrt{(K - 1)(F_{crit})} = \sqrt{(3 - 1)(3.55)} = \sqrt{7.1} = 2.66$$

$$C_{obt} > C_{crit}$$

- Conclusion: Difference between the current flavour and new flavour 2

Interpreting the results

	New flavour 1	New flavour 2
Current flavour	$\bar{X}_c < \bar{X}_1$ $C_{obt} < C_{crit}$	$\bar{X}_c > \bar{X}_2$ $C_{obt} > C_{crit}$

- The rating for flavour 1 is higher but not significantly higher than for the current flavour
- The rating for flavour 2 is significantly lower than for the current flavour

Note: When conducting Scheffé post-hoc tests there is no increase in the likelihood of making Type I errors with repeated tests

Exercise

Group 1	Group 2	Group 3
22	6	8
16	10	6
17	13	4
18	13	5
	8	2
	4	

- Suppose we have three groups
 - Groups 1 and 2 have been given a treatment of some sort
 - Group 3 is the control group
- Conduct an ANOVA test to identify if there are significant differences between groups
 - If such differences exist, carry out *post hoc* tests to identify which treatments made a significant difference

Summary

- Topics covered
 - Computational steps
 - Strength of Association
 - Post hoc Analyses
 - The Scheffé Method of Post hoc Analysis
- Non-parametric tests allow us to handle data that does not follow a normal distribution

Coming soon

- **Next session:** Non-parametric tests
- **Homework:** Read chapter 12 of *Statistics for the Terrified*.