Experimenting: statistical analysis 2

Per Ola Kristensson

Research Methods
M.Phil. Advanced Computer Science
University of Cambridge

Michaelmas Term, 2009

Example

- Method A (baseline)
- Method B

- Between-subjects experiment
- One session
- Three participants in each condition
- Six participants in total

Raw data

	Group A	Group B
Observation 1	$X_{A1} = 2$	$X_{B1} = 6$
Observation 2	$X_{A2} = 3$	$X_{B2} = 7$
Observation 3	$X_{A3} = 1$	$X_{B3} = 5$
Sample mean	$\overline{X}_A = 2$	$\overline{X}_B = 6$

Two different variance estimates

- Within
  - Error
- Between
  - Effect
  - (Error)

Sum of squares

- Remember the residuals from last lecture:

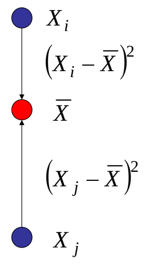
$$X_i - \overline{X}$$

- Sum of squares (SS) is simply the sum of the squared residuals:

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Why sum of squares?

- Intuitive explanation:
  - We have individual observations
  - We try to fit these observations to an expected value (the mean)
  - We do not know the true population mean
  - However, we do know an estimate – the sample mean
  - Sum of squares gives us a similarity measure or “goodness of fit”
  - Similar to linear regression



Group A (within)

Group A	Raw value	Residuals (squared)
Observation 1	$X_{A1} = 2$	$\left(X_{A1} - \overline{X}_A\right)^2 = (2-2)^2 = 0$
Observation 2	$X_{A2} = 3$	$\left(X_{A2} - \overline{X}_A\right)^2 = (3-2)^2 = 1$
Observation 3	$X_{A3} = 1$	$\left(X_{A3} - \overline{X}_A\right)^2 = (1-2)^2 = 1$
Sum of squares		$\sum_{i=1}^{3}\left(X_{Ai} - \overline{X}_A\right)^2 = 2$

Group B (within)

Group A	Raw value	Residuals (squared)
Observation 1	$X_{B1} = 6$	$\left(X_{B1} - \overline{X}_B\right)^2 = (6-6)^2 = 0$
Observation 2	$X_{B2} = 7$	$\left(X_{B2} - \overline{X}_B\right)^2 = (7-6)^2 = 1$
Observation 3	$X_{B3} = 5$	$\left(X_{B3} - \overline{X}_B\right)^2 = (5-6)^2 = 1$
Sum of squares		$\sum_{i=1}^{3}\left(X_{Bi} - \overline{X}_B\right)^2 = 2$

Group A+B (total)

Overall mean	$\overline{X}_{AB} = 4$
Squared residuals 1A+1B	$\left(X_{A1} - \overline{X}_{AB}\right)^2 + \left(X_{B1} - \overline{X}_{AB}\right)^2 = 8$
Squared residuals 2A+2B	$\left(X_{A2} - \overline{X}_{AB}\right)^2 + \left(X_{B2} - \overline{X}_{AB}\right)^2 = 10$
Squared residuals 3A+3B	$\left(X_{A3} - \overline{X}_{AB}\right)^2 + \left(X_{B3} - \overline{X}_{AB}\right)^2 = 10$
Sum of squares	$\sum_{i=1}^{3}\left(X_{Ai} - \overline{X}_{AB}\right)^2 + \sum_{i=1}^{3}\left(X_{Bi} - \overline{X}_{AB}\right)^2 = 28$

Intermediary summary

- Group A (within):
  - Mean: 2
  - Variance: 2
- Group B (within):
  - Mean: 6
  - Variance: 2
- Groups A+B (total):
  - Mean: 4
  - Variance: 28
- Question:
  - Is there a difference between the means of groups A and B that is due to effect rather than error?

Remember the logic of ANOVA

- Within-groups estimate (error)
- Between-groups estimate (effect of independent variable and error)
- H0: $\mu_1 = \mu_2$
- Given H0, the variance estimates should be equal
- This is because H0 assumes the effect of the independent variable does not exist
- Then both variance estimates reflect error and their ratio is 1
- A ratio larger than 1 suggests an effect of the independent variable

Remember the assumptions behind ANOVA

- The population has a mean
  - Remember, not all distributions have a mean
- The population is assumed to be normal
  - Each observation sampled from a Gaussian distribution
- Each observation is assumed to be independent

Chi-squared distribution

- A sum of squared independent normal random variables have a chi-squared distribution

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2, n > 0$$

F-distribution

- The F-distribution arises as the ratio of two independent chi-square estimates
- In our case:

$$\frac{SS_{between} / df_{between}}{SS_{within} / df_{within}}$$

- where *df* are the degrees of freedom for each chi-square estimate

Back to our example

- SS_{error} = 2 + 2 = 4 (for A and B)
- SS_{total} = 28 (for A + B)
- $SS_{effect} = SS_{total} - S_{error}$ = 28 – 4 = 24
- df_{error} = *N subjects* – *N* groups = 6 – 2 = 4
- df_{effect} = *N* groups – 1 = 2 – 1 = 1
- $MS_{error} = SS_{error} / df_{error}$ = 4 / 4 = 1
- $MS_{effect} = SS_{effect} / df_{effect}$ = 24 / 1 = 24
- $F = MS_{effect} / MS_{error}$ = 24 / 1 = 24
- p (for df[1,4] and F = 24) ≈ 0.08
- $F_{1,4}$ = 24.000, p = 0.08

Summary

- Between-subjects experiment
- Six participants in total
- Two groups
  - Mean[A] = 2, Mean[B] = 6
  - Variance[A] = Variance[B] = 2
  - Mean[Total] = 4, Variance[Total] = 28
  - Variance[Effect] = Variance[Total] – (Variance[A] + Variance[B]) = 28 – (2+ 2) = 24
- Degrees of freedom:
  - Effect = 2 groups – 1 = 1 df
  - Error = 6 participants – 2 groups = 4 df
- MS[error] = 4 / 4 = 1
- MS[effect] = 24 / 1 = 24
- F = MS[effect] / MS[error] = 24 / 1 = 24

Always remember the assumptions of ANOVA

- Independence
- Normality
  - Residuals are normal
- Homogeneity of variances
  - The groups should have equal variance

Typical real usage of ANOVA

- Participants are exposed to three different methods
- ANOVA is used to compute if there is a statistically significant difference between the means (omnibus test)
- ANOVA only tells that there is a difference, it does not tell us which means differ
- Now post-hoc analysis is carried out to compute pair-wise differences between the means
- ANOVA is powerful in this scenario because it protects us against over-testing the data without being too restrictive (unlike t-tests)

When not to use ANOVA

- Data that is inherently non-normal
  - Rank data (e.g. user ratings)
  - Data that cannot be reasonably transformed so that the residuals are approximately normal

Things to watch out for

- First, ANOVA is relatively robust
  - Against non-normal data
  - Against unequal variances (as long as the number of participants in both conditions are equal)
- Outliers can cause misleading results
  - Outliers that violate ANOVA's assumption of homogeneity of variances is particularly troublesome
- Again, rank/ratings data require a different test

Summary

- We want to find out if means (or sometimes medians) differ among different methods
- We identify the levels of our independent variable (which methods we are going to test)
- Need to find a suitable baseline
- We identify which dependent variables to measure
- We decide on an experimental design
- Get ethical approval (if necessary)
- We carry out a pilot study
- We recruit participants and carry out the experiment

Summary, continued

- To find out if our independent variable could account for the difference in observed means (or medians) we need to conduct a significance test
  - (The difference could be due to chance)
- A significance test tells us if we can reject the null hypothesis at a preset significance level
- Failing to reject the null hypothesis does not mean that the means are equal
  - It just means you failed to reject the null hypothesis (and nothing else)
- If we reject the null hypothesis we conclude that the difference in observed means (or medians) are due to our manipulation of the level of the independent variable

Summary, continued

- Analysis of variance is a popular method for testing for significant differences
- The idea is to partition the variance into variance related to error (sampling error) and effect (due to our manipulation of the independent variable)
- These variances are chi-squared estimates
- The F-distribution tells us the probability that the ratio of two chi-square estimates is the same
- We obtain a probability that we can use to reject the null hypothesis under a preset confidence level

Conclusion

- Choose the right baseline
- Think carefully about experimental design
- Carry out a pilot study
- Plot the data before analysing it further
- Observe all assumptions behind statistical tests
- Explain what you did so others can repeat it
- Motivate any out-of-the-ordinary modifications to experimental procedure or analysis you have carried out