Experimenting: statistical analysis 1

Per Ola Kristensson

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

Michaelmas Term, 2009

Running example in this lecture

- Spatiotemporal visualization

A. 2D
B. Space time cube

Informally: significant difference

- We sampled two groups from a population (students)
- We exposed each group to a different method (independent variable)
- We collected measures (dependent variable) from each group using said method
- We now believe a right way to compare these methods is to investigate if the means of the measures differ between the two groups
- The null hypothesis HO says for some predetermined confidence level there is no actual difference between the means and any measured difference is due to sampling error
- If we reject the null hypothesis H0 then we have a significant result at said confidence level

Type I and type II errors

- Type I
  - Rejecting H0 even though it is true
- Type II
  - Failing to reject H0 even though it is false

Different tests and assumptions

- Nature of data
- Experimental design
- Many, many pitfalls
- Also varies according to specific research fields

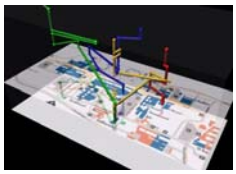
Back to our example

- Two conditions (independent variable)
  - 2D (baseline)
  - Space time cube
- Dependent variables
  - Error
  - Response time
- Between-subjects design

Our example, two conditions

A. 2D (baseline)

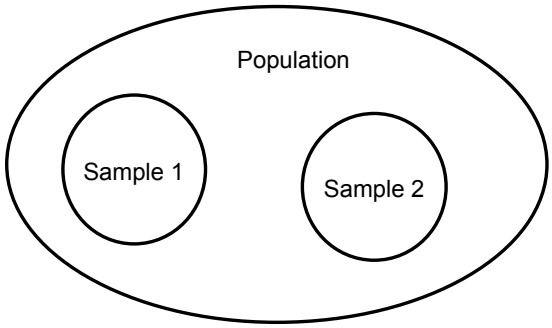
B. Space time cube



Analysis of variance (ANOVA)

- A versatile significance testing method which is very popular
- Many variations exist
- Typical usage:
  - Between-subjects experiment with more than two levels of the independent variable
  - Within-subjects designs
  - Mixed designs
  - As an omnibus test that is followed up by post-hoc tests *if* a significant difference between the means is detected

Sampling from the population



Error

- The amount an observation differs from the population mean
- Typically the population mean is unobservable

Residual

- The amount an observation differs from the sample mean
- Unlike the population mean, the sample mean is observable

A bit more formally

- Sample from a normal distribution:

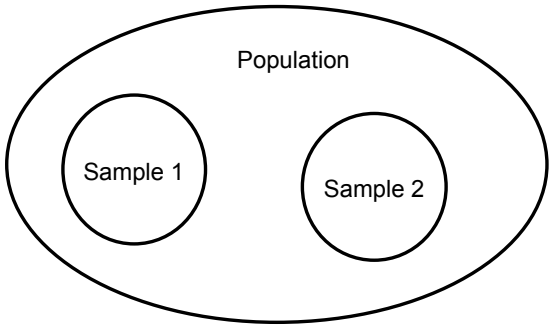
$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

- Sample mean:

$$\overline{X} = X_1 + X_2 + \ldots + X_n / n$$

- Error: $X_i - \mu$
- Residual: $X_i - \overline{X}$

Sampling again



Why would there be a difference between the means?

1. Because of group membership
   - Effect of independent variable on dependent variable
2. *Not* because of group membership
   - Sampling error

The logic of ANOVA

- There are two independent estimates of the population variance that can be obtained
  - Between-groups estimate (effect of independent variable and error)
  - Within-groups estimate (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

F-distribution

- The ratio of the between-groups estimate and the within-groups estimate is an F-distribution when H0 is true
- F varies as a function of a pair of degrees of freedom (one for each estimate of the variance)
- $F \geq 0$

Sums of squares

- Remember the residuals:

$$X_i - \overline{X}$$

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

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

Computing the F-score

- The SS are partitioned
  - $SS_{total} = SS_{within} + SS_{between}$
- $F = MS_{between} / MS_{within}$

Rejecting the null hypothesis

- Using the F-distribution we can compute the probability that the result was due to chance
- If the probability is less than a *preset* significance level α we reject H0
- Reject H0 if $p < \alpha$

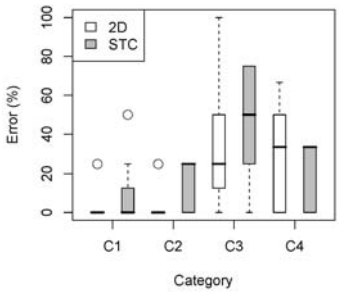
Assumptions of ANOVA

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

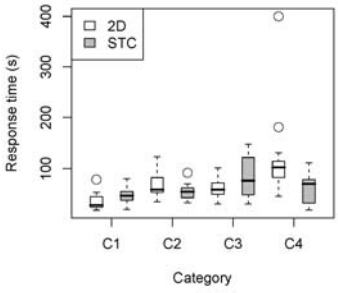
Measures, our example

- Error – number of errors in a test
- Response time – number of seconds it took for participants to answer questions

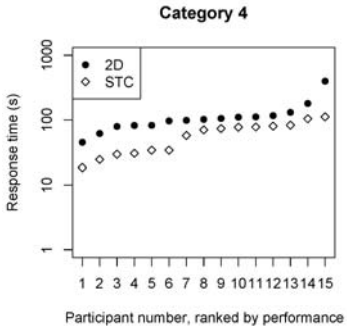
Plotting the data, error



Plotting the data, response time



Plotting the data, category 4



ANOVA results

Cat.	2D Mean	2D sd	STC Mean	STC sd	F	p
1	1.67	6.46	10.00	18.41	2.7344	0.1094
2	1.67	6.46	13.33	12.91	9.800	0.0041
3	33.33	29.38	48.33	24.03	2.343	0.1371
4	31.11	26.63	20.00	16.90	1.8622	0.1832

Cat.	2D Mean	2D sd	STC Mean	STC sd	F	p
1	36.57	16.65	45.77	16.37	2.326	0.1384
2	67.74	24.05	55.30	15.07	2.8835	0.1006
3	60.02	17.51	82.73	39.80	3.8563	0.0599
4	120.64	83.32	60.34	29.95	6.9571	0.0135

Reporting 1

All statistical tests in this paper were carried out using analysis of variance (ANOVA) at a significance level of = 0.05. Assumptions underlying the analysis of variance procedure were taken into account before performing any significance testing. We did not apply any transformations to the data (such as logarithmic transformations) when testing for significance.

Reporting 1

All statistical tests in this paper were carried out using **analysis of variance** (ANOVA) at a **significance level** of $\alpha = \mathbf{0.05}$. **Assumptions** underlying the analysis of variance procedure were **taken into account** before performing any significance testing. We did **not apply any transformations** to the data (such as logarithmic transformations) when testing for significance.

Reporting 2

We found a high-magnitude statistically significant difference in question category 4 where space time cube representation halved the average response time from 121 s in the baseline 2D system down to 60 s ($F_{1,28} = 6.957$, $p = 0.0135$). This result supports the hypothesis that space time cube representation is efficient in supporting users' understanding of complex spatiotemporal patterns in datasets.

Reporting 2

We found a high-magnitude **statistically significant difference** in question category 4 where space time cube representation halved the average response time from **121** s in the baseline 2D system down to **60** s ($\mathbf{F_{1,28}}$ = 6.957, p = 0.0135). This result **supports the hypothesis** that space time cube representation is efficient in supporting users' understanding of complex spatiotemporal patterns in datasets.

Next lecture

- A walk-through of ANOVA
- More about limitations and implications of statistical testing