

Power and pitfalls in statistical data analysis

Valentina Moskvina
(e-mail: MoskvinaV1@cardiff.ac.uk)

Prague
February 2007

DATA

- **Nominal** - categories without order (eye colour, marital status, ...)
- **Ordinal** - ordered categories (stage of a disease)
- **Binary** (yes/no)
- **Integer** (number of counts)
- **Ratio** (value independent of units)
- **Interval** (Distances between units are known (hours spent studying))

Data analysis

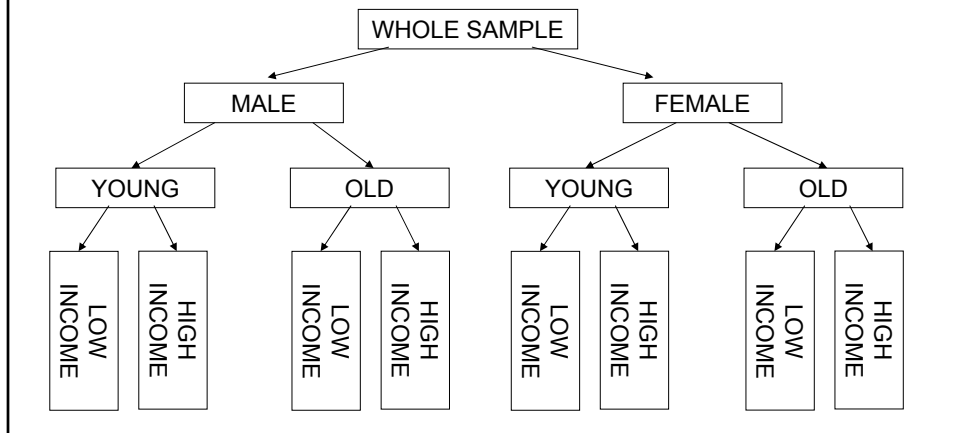
1. Data exploration and choice of statistical method
2. Analyses:
 - Parametric / Non-parametric
 - Univariate / Multivariate
 - Data reduction (reducing multiple testing)
 - Machine learning with cross-validations
3. Data interpretation:
 - Null Hypothesis - is that any difference seen occurred by chance.
 - Test is designed to disprove the Null Hypothesis
 - If the result is statistically significant ($p < 0.05$), the conclusion is that the null hypothesis is wrong (and therefore rejected)

Data interpretation

- Biological scientists usually accept a 5% probability that the difference detected is due to chance ($p=0.05$)
- When multiple tests were performed this probability is increased
- *Bonferroni* correction states that if one is testing n independent hypotheses, one should use a significance level of $0.05/n$
- Note that, since tests are rarely independent, this is a very conservative correction

Subgroup analysis

- Standard procedure adopted is to split data by categories:



Simulation

1000 simulations

At each simulation:

- 500 random Normally distributed values
- Split randomly into 8 groups of equal sizes
- Performed:
 - t-test compare all pair combinations
 - ANOVA with 8 categories
- Results:
 - t-tests: ~47% at least one pair was significant
 - ANOVA: ~5% sig. results

Rationale

- Subgroup analyses required if there is a reason to expect different effect in different groups (M/F, age groups,...)
- Do not consider too many groups
- Conclusions should NOT be based on p-values in subgroups:
It is incorrect to conclude that effect differs between two groups if p-value is sig. in one and not in another

Parametric vs non-parametric

Always plot data to determine need for parametric or non-parametric analyses

Parametric:

- assumes the data was from a defined distribution
- more powerful
- defined interpretation

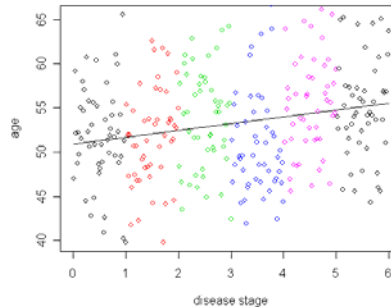
Non-parametric:

- no limitations on types of data
- more likely to give Type II errors (false negatives)
- less likely to be significant
- cannot relate back to any parametric properties of the data

Parametric vs non-parametric

300 individuals

Parametric analysis
(linear regression):



Null hypothesis:

age does not depend of the stage of the disease

Independent variable: disease stage

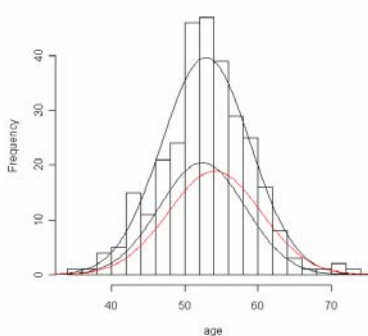
Dependent variable: age

Result: coef = 0.78, p=0.00015

Interpretation: age = 50.9 + stage*0.78
(12 months*0.78 ≈ 9 months)

Parametric vs non-parametric

Histogram of age



Split into 2 groups:

Cases (BLACK): mean=52.4, SD=5.85

Controls (RED): mean=54.1, SD=6.34

Non-parametric analysis: Chi-squared test

Categories: 0 - if age<53; 1 - otherwise

0 - early stages (0-2); 1 - late stages (3-5)

Result: **OR=1.45; p=0.11**

Odds Ratios (OR)

OR – ratio of odds for being young in the case and controls group

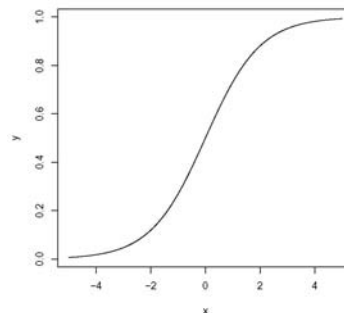
- In cases: the number of cases with age below 53 divided by the number of cases with age above 53
- in controls: the number of controls with age below 53 divided by the number of controls with age above 53

	age<53	age≥53	total
cases	a	b	a+b
controls	c	d	c+d
total	a+c	b+d	N

$$OR = \frac{a}{b} / \frac{c}{d} = \frac{ad}{bc}$$

Binary Outcomes

- Usually, binary data result from a non-linear relationship between probability of outcome (case/control) variable and the exposure (age group) x .
- A fixed change in x often has less impact when the probability is close to 0 or 1 than when it is close to $\frac{1}{2}$.
- In practise these relationships are monotonic and typically look like s-shaped curves.



Logistic regression

- Logistic regression – a method for the analysis of binary outcomes, used to
 - compare binary outcome variable between two exposure groups
 - compare more than two exposure groups
 - examine the effect of an ordered or continuous exposure variable

Logistic regression

Simple Logistic Regression model:

$$OR = \frac{\text{odds in cases}}{\text{odds in controls}}$$

$$\text{odds cases} = \text{odds in controls} \cdot OR = \text{baseline} \cdot OR$$

$$\ln \frac{p}{1-p} (\ln \text{ odds in cases}) = A + Bx_1$$

Results of the simple Logistic Regression:

	<i>B</i>	SE	p	exp(<i>B</i>)
Age group	0.37	0.232	0.110	1.45

General Logistic Regression Model:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$$

Confounders (false positive finding)

- Sub-population 1: 200 individuals

	cat ₁	cat ₂
case	160	160
control	40	40

OR=1, p=1

- Sub-population 2: 200 individuals

	cat ₁	cat ₂
case	160	40
control	160	40

OR=1, p=1

- Together: 400 individuals

	cat ₁	cat ₂
case	320	200
control	200	80

OR=0.64, p = 0.005

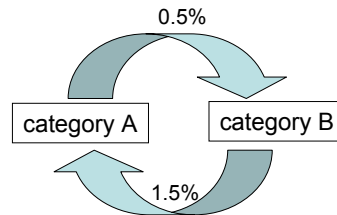
Rationale

- Use matched studies (appropriate statistic techniques):
 - Paired t-test, McNemar test
 - Conditional regression models
- Index of membership
 - self-reported ethnicity, geographical origin, possibly religion, social class
 - not always accurate / effects may be subtle
 - infer from an individual's background
 - *look for signatures of stratification* (latent class analysis)
 - *correct tests for inferred substructure*

Misclassification error

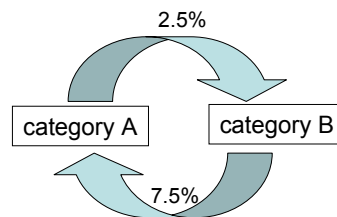
- Cases**

(average misclass. error 1%)



- Controls**

(average misclass. error 5%)



Type I error

- $N_{cas}=200$, $N_{con}=200$; $N_{sim}=5000$, χ^2 test

Null hypothesis:	Misclassification Error			
	in cases:1% in controls: 1%	in cases:1% in controls: 5%	in cases (1%) A→B: 0.5% B→A: 1.5%	in cases (1%) A→B: 0.5% B→A: 1.5%
Equal frequencies in cases and controls			in controls (1%) A→B: 0.5% B→A: 1.5%	in controls (5%) A→B: 2.5% B→A: 7.5%
0.5	0.052	0.050	0.050	0.086
0.4	0.051	0.050	0.046	0.125
0.3	0.048	0.070	0.053	0.181
0.2	0.044	0.126	0.051	0.284
0.1	0.053	0.263	0.047	0.536

Power

- $N_{\text{cas}}=100$, $N_{\text{con}}=100$; $N_{\text{sim}}=5000$, χ^2 test

Alternative hypothesis		Misclassification Error				
Freq in cases	Freq in controls	No error	in cases:1% in controls: 1%	in cases:1% in controls: 5%	in cases (1%) A→B: 0.5% B→A: 1.5%	in cases (1%) A→B: 0.5% B→A: 1.5%
0.5	0.3	0.986	0.918	0.909	0.912	0.642
0.5	0.35	0.856	0.693	0.685	0.698	0.327
0.5	0.4	0.528	0.367	0.364	0.371	0.111
0.5	0.45	0.179	0.126	0.128	0.129	0.052

Summary

- Always plot the data before the analysis
- Use parametric analysis if possible
- Think about possible confounders
- Pay attention to data quality
- Be aware of multiple testing problem