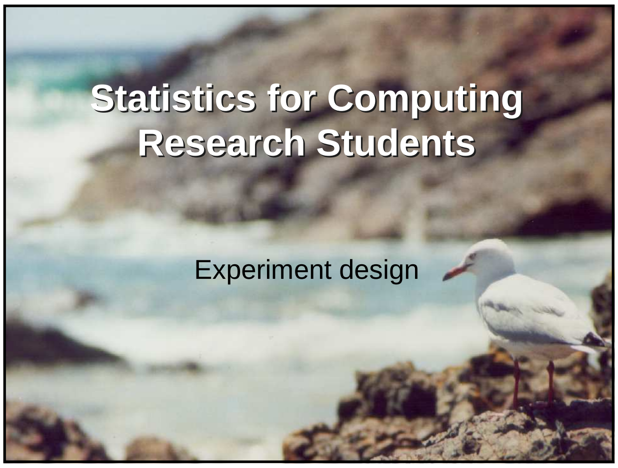


Example 1: Algorithm speed

- Aim: To compare speed of 3 sorting algorithms A, B, C.
- Method: Run each algorithm once on the same data.
- Results: A: 7.3s B: 6.5s C: 11.8s
- What can we conclude?

Variables

- Explanatory/covariate (independent)
  - Variable that is controlled/known in the experiment
- Response/outcome (dependent)
  - Variable that is measured outcome
- Variables in example 1?

Confounding Variables

- Related to both independent and dependent
  - As time passes, child grows taller and country's GDP increases.
    - Could falsely conclude: child's growth impacts GDP.
  - Study of gender risk of cancer
    - Smoking confounds
  - See also "Simpson's paradox"

Dealing with confounding variables

- Control
  - Remove all smokers from cancer study
    - Conclusions are limited to non-smokers
    - May bias results if choice to smoke is related to other cancer-causing factors (e.g. suburb)
- Measure and model
  - Include smoking as an independent variable in model
  - Estimate risk due to smoking **and** gender

Randomness

- Identical circumstances can produce different outcomes
- Real-world measurements are subject to measurement error
  - Response variable and/or covariate
- Individual cases are subject to unknown factors and real-world randomness
- Modelled as random noise in response variable; noise in covariate

Example 1: Algorithm speed

- Aim: To compare speed of 3 sorting algorithms A, B, C.
- Method: Run each algorithm once on the same data.
- Results: A: 7.3s B: 6.5s C: 11.8s
- What are:
  - Independent and dependent
  - Possible confounding variables
  - Sources of randomness?

A 'valid' conclusion

- B is fastest
  - On that data set (independent)
  - Using that code (independent)
  - In that programming language (ind)
  - With that compiler (ind)
  - On that machine (ind)
  - Running that OS (ind)
- Provided there were no other programs running during the tests!

Let's design an experiment

- Many data sets – easy
- Different sizes of data set – not so easy
- Different machines – not much choice
- Different languages – difficult
- Different programmers
- Different OS – not much choice
- Control/measure background activity

Example 1: Experiment

- A variety of data set sizes: 10,20,50,100,200,500,1000,…
- N random data sets of each size
- Run each algorithm on each data set
- Control other computer activity as much as possible
- Use different machines, compilers, OS

Allocation strategy

- Complete
  - All combinations of explanatory variables
    - Same data sets for each algorithm
    - Test all algorithms on all machines, OS, etc
- Randomised
  - Randomised blocks – balanced random selections
- Experiment design
  - Selects subset of cases to study

Example 2: Human factors

- Question: Do users find it easier to use web sites that have drop-down menus or ones that use on-screen menus?
- Ethics approval for human experiments

Variables

- Explanatory
  - Drop-down vs on-screen menus
- Response (what we measure)
  - User preference statements (informal)
  - Likert scale 1-5
    - "Site B was easier to use than site A"
    - SD, D, N, A, SA
  - Time taken to complete a task
- Frame the research question clearly…

Other explanatory variables

- Prior experience of the user
- Gender, age, ethnicity
- Physical ability (to control mouse etc)
- Input device (mouse, touch pad, etc)
- Site colours, appearance, fonts
- Colour blindness

Confounding variables

- Site complexity
  - Larger sites more likely to use drop-down menus, but may be more difficult to navigate because they are larger
- Site designer ability
  - More experienced designers may be more likely to use drop-down menus and also produce better site organisation + ease of use

Experiment design

- Control confounding variables?
  - Custom-built web sites for tests
  - Same content and design
  - Differ only in menu technology
  - But:
    - Are the test sites representative?
    - Design, structure, placement of menus – comparable?
    - Test site designer is an independent variable!

Learning

- Doing a task changes a person – they learn
- Using one test web site affects performance on paired test web site
- Cross-over design
- For limitations and alternatives:
  - http://www.uq.edu.au/~hmrburge/stats/twotrials.html

Sample size

- How much data do I need to have a strong chance of seeing the effect I am looking for if it is there?
  - An experiment that could never show the desired outcome is worse than useless.

Experiment validity

- Internal validity:
  - Is the experiment conducted properly?
  - Are there confounding variables etc not considered?
- External validity:
  - Do the results generalise?
- Test, re-test
  - Repeat the whole expt and analysis

Algorithm adaptation

- During algorithm development, we may **test** and then **improve** the algorithm iteratively.
- This can 'adapt' the algorithm to perform well on test data but it may not perform well on other data.

A (silly) example

- What is the fastest algorithm to sort the following data (assume it is in an array)?

  1 7 3 8 11 9 2 6 16 5 0 4

Real examples

- Choose the best statistical model (or Artificial Neural Network/Decision Tree/other learning system) for your data.
- Just about any program to extract information from data can be adapted.
  - Solving CAPTCHAs
  - Parsing English queries

Ways **not** to avoid adaptation

- My algorithm is based on fundamental principles
  - Only OK if truly established a priori
- I have a large data set that I use for testing
- All parameters are set from the data in my final algorithm

Avoiding adaptation

- Reserve a portion of data set for final testing
  - Once-off run of final tests, report those results whatever they are
- If your algorithm sets parameters from data, (e.g. learning or fitting a statistical model), use **cross-validation** for final testing

Cross validation: motivation

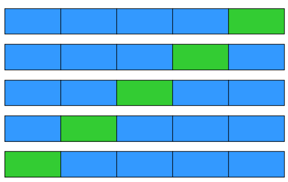
- Algorithm where:
  - Run A on training data to set parameters P
  - Run A(P) on new data to analyse it
- E.g.
  - ANNs and statistical models
  - Decision trees
  - Person (face/gait/voice/etc) recognition

Problem

- If we use data to set the parameters and then test performance on the same data, results are biased (‘adapted’)
- Idea:
  - Set parameters (train) on N/2
  - Test on remaining N/2
- Problem:
  - Limited training data (N/2) and test data (N/2)

Cross validation

- Train on, say 80% and test on 20%.
- Do that 5 times.

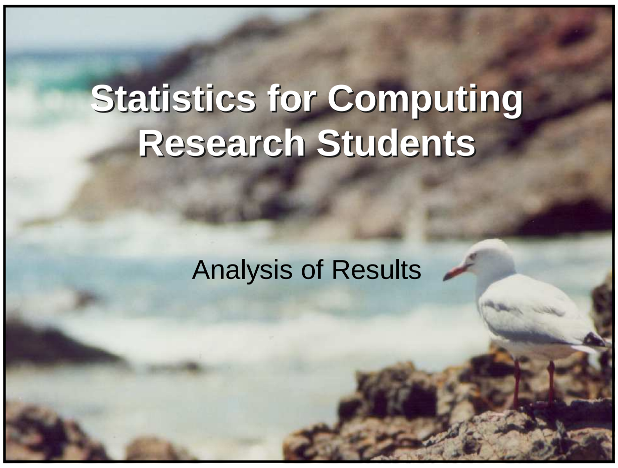


Take-home messages

- Think response, explanatory, confounding.
- Other variables – are they having a random effect or held constant?
- Formulate the research question clearly in advance.
- Understand what result is expected.
- Human experiments are more difficult.

Take-home messages

- Develop algorithms using a subset of your data.
- Test algorithms on data not previously used.
- Use cross-validation for algorithms that involve training.
- Design for analysis: next time



- You've done your experiment

  now what ?
- Depends on the model you are testing…

Models

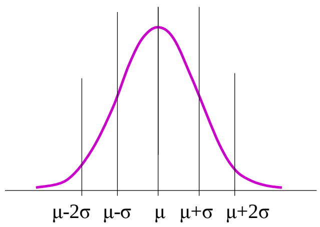
- Understand model before experiment…
- Mean + noise
  - The data items have a mean value plus noise
- Mean time to sort 10000 items
  - Algorithm A: 7.3s B: 6.5s C: 11.8s
  - Is B really better than A?
  - What about C?

Distributions and randomness

- Actual measurements:
  - A: 7.8 7.4 6.9 7.0 7.3 7.2 6.8 7.7 …
  - B: 6.3 6.7 6.2 6.5 6.6 6.8 6.4 6.9 …
  - C: 12.9 11.2 10.7 12.3 10.1 11.9 11.5 …
- Central Limit Theorem
  - Mean is Gaussian (Normal) with std dev $s/\sqrt{N}$ where s is sample std dev

Gauss distribution

- Well-known bell curve



Estimation

- A statistic T is an estimate of a true parameter θ
  - Average $\overline{X}$ is an estimate of mean μ
  - Std devn s is an estimate of σ
- The question is:
  how accurate is the estimate?

Confidence interval

- B: $\overline{X}$ = 6.5 s = 1.2 N = 100
  - We are 95% sure that an individual sample X will lie between μ – 2σ and μ + 2σ
    - i.e. 4.1 and 8.9 (approximating μ by $\overline{X}$, σ by s)
  - We are 95% sure that $\overline{X}$ lies between μ – 2σ/10 and μ + 2σ/10
    - i.e. $\overline{X}$ is within 2σ/10 of μ.
  - Therefore, we can say that 95% likely that the true mean μ lies within 2σ/10 of $\overline{X}$.
    - i.e. μ is likely between 6.2 and 6.8 (approximating σ by s)
    - This does not say whether B is better than A or not

Confidence interval

- Range of values with probability 1-α that the true parameter θ lies in the range
- e.g. Under normality, 95% (α = 0.05) CI for the mean is $\overline{X} \pm 1.96\sigma/\sqrt{N}$
  - (If s is used instead of σ, the CI changes somewhat – see Student's t-distribution)

Comparing two means

- T-test
  - Take difference between means
  - Test whether it is zero
- If data are paired (same test data for sorting in each pair), use paired t-test
- Assumes equal variance

Testing a hypothesis

- Hypothesis: "Algorithm B is better than algorithm A"
- More formally: "The mean execution time for algorithm B is less than A"
  - Median may be more appropriate?

The null hypothesis

- What would be the case if our hypothesis of signifiance is **not** true?
- "The mean execution time for algorithms A and B are the same"
- H_0: μA = μB
- H_0: μA – μB = 0

Hypothesis testing

- We say that the null hypothesis is rejected (and that there is a statistically significant effect) if
  - the **probability of**
  - **results at least as extreme** as the results we obtained
  - **occurring by chance**
  - is sufficiently **small** (<5%).

Hypothesis testing example

- Flip a coin N times and we happen to get heads every time.
- Is the coin 'fair' or is it a double-headed coin?
- N=2 25% chance of HH with fair coin
- N=4 6.25% chance of HHHH with fair
- N=10 0.1% chance of HHHHHHHHHH!

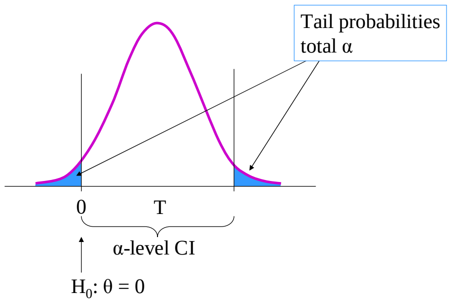
Hypothesis testing

- We are interested in the **probability** that a **result at least as extreme as our result** could happen **by chance if the null hypothesis is true** (i.e. if there is nothing significant happening).
- P-value: This probability.
- If p-value < 0.05, we say it is significant.
- Reporting p-values is sensible: p-value of 0.001 is much more significant!

Relating CI and Hypothesis testing

- If the null hypothesis lies inside an α-level confidence interval, then the null hypothesis is accepted.
- The α-level that puts the null hypothesis at the edge of the confidence interval is the p-value of the hypothesis test.

Relating CI and Hypothesis testing



Subtleties

- Single-sided test versus double-sided test
- Different kinds of confidence intervals

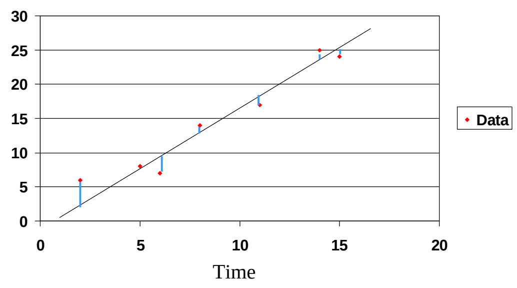
Aside

- How much data do I need?

Regression and residuals

- A linear model
  - Y = **X θ** + c + noise; Y is response variable
  - **θ**: vector of fitted parameters
  - **X**: row vector of carrier (indep.) variables
- Fitting model yields
  - $\hat{\mathbf{y}} = \mathbf{X}\,\hat{\boldsymbol{\theta}} + \hat{c}$; **X** is matrix of carrier data
- Residuals: error in model fit
  - $\mathbf{r} = \hat{\mathbf{y}} - \mathbf{y}$

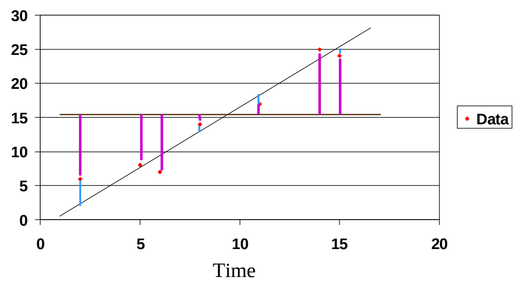
Regression and residuals



ANOVA

- ANalysis Of VAriance
- Determine whether linear model parameters are significant
- Assumes normality (Gaussian distribution) of residuals

Regression and residuals



ANOVA

- SS = **r**.**r** = $\Sigma \mathbf{r}_i^2$
- Compare mean model with line model
  - SS_{line} = **r**.**r**
  - SS_{mean} **=** $\Sigma(y-\bar{y})^2$
- ANOVA says:
  - $SS_{mean} = SS_{line} + SS_{slope}$
  - If SS_{slope} is large compared to SS_{line} then slope is significant

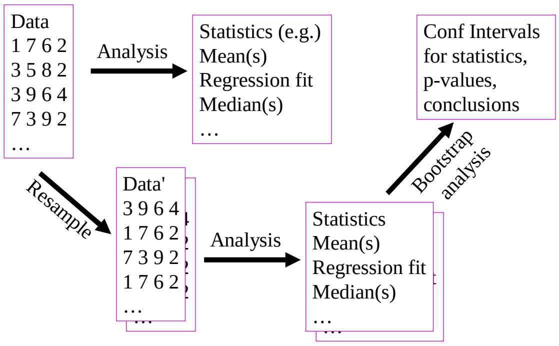
ANOVA

- SS_{slope} / (SS_{line} / N-2) is F(1,N-2) [a known distribution] – F-test
- Assumes residuals are:
  - Normal (Gauss) distribution
  - Zero mean
  - Equal variance
- MANOVA: Multivariate

More models

- GLM (Generalised Linear Models)
  - **y** = f(A**x** + **b**)
  - f is monotonic
- Feed-forward ANNs
  - **y** = $f(\Sigma_i f(A_i\mathbf{x} + \mathbf{b}) + A_0\mathbf{x} + \mathbf{b})$
  - More levels are possible but two levels gives a universal approximator

Bootstrap



Types of bootstrap

- Resampling method
  - Random with replacement
  - Blocks (time series, or other correlated)
- CI estimation methods
  - Percentiles (1st order accurate)
  - BC, studentized (2nd order accurate)

Comparison of bootstrap

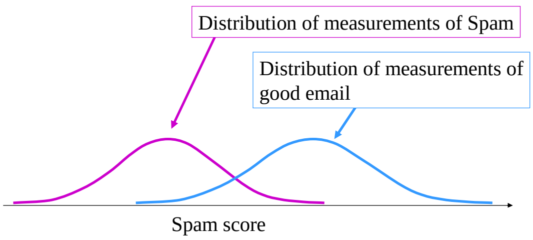
- Bootstrap: Non-parametric: distribution (of statistics) need not be known
  - Flexible: can provide confidence intervals for statistics that are not well understood (i.e. not means/variances under normality)
- ANOVA/t-test, etc: Parametric: based on analysis of distribution
  - More powerful to draw conclusions

Classification/recognition tasks

- Gait/face recognition
- Spam email classifier
- “Recognise photos of the Pope”

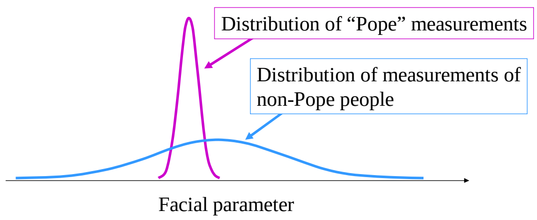
- Calculate some measurement(s)
- Classify as A/B (good/bad) etc.
  - Linear/non-linear classifier

Spam classifier



•Where to put the decision threshold?

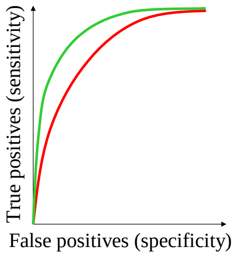
“Pope” recognition



•Where to put the decision thresholds?

Classification: ROC

- Receiver Operating Characteristic (ROC) curve
- Plot
  - True positives (= 1-false negatives)
  - False positives
- As threshold is varied



ROC

- If the system is made more sensitive to true positive cases, it is more likely to produce false positives as well.
- Depending upon cost/benefit ratio of false positive and false negative, can choose optimal operating threshold.

Classification, Testing and CI

- Spam classifier is a one-sided test
- Pope recogniser is a two-sided test
  - H_0: Photo is Pope
  - H_1: Photo is not Pope
  - Threshold range is a CI for H_0
    - Level (α) is the false negative rate
  - Better separation using more measurements (higher dimensionality)

Design for analysis

- Consider the formal hypothesis and null hypothesis
- Understand the planned analysis before conducting the experiments
- Ensure the data will enable the analysis

Take-home messages

- 'Formal' hypothesis
- Null hypothesis
- Model
- Statistical testing
  - Parametric
  - Non-parametric
- Confidence interval

Take-home messages

- Experimental work **requires** statistical analysis
- Plan for analysis before experiment
- Get help with statistics
  - Only certain techniques will be relevant to your particular questions and experiments.

References

- Bootstrap
  - http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore14.pdf
  - http://bcs.whfreeman.com/pbs/cat_140/chap18.pdf
  - http://www.wiley.com/legacy/wileychi/eoenv/pdf/Vab028-.pdf
- ANOVA
  - http://bcs.whfreeman.com/pbs/cat_140/chap14.pdf

Resources

- http://www.causascientia.org/math_stat/ProportionCI.html
  - Testing and CI calculator for proportions

Case Study: Solving CAPTCHAs

- Questions:
  - What CAPTCHA techniques are most difficult to solve automatically?
  - How do humans and computers compare at solving CAPTCHAs?

Questions

- What does it mean to say a CAPTCHA has been solved by computer?
- What is the role of the experimenter/programmer in developing solution algorithms?
- How to measure difficulty of solving by computer?

Experiments

- How might you measure human performance at solving CAPTCHAs?
- How might you measure computer performance?
- What makes it difficult to measure?

Analysis

- How valid would it be to extend results of study to the wider field of CAPTCHAs?
- What useful conclusions might we be able to draw?
- Should experiments be different to enable analysis to lead to useful conclusions?