Scientific Method

CSW

Susan Stepney

Department of Computer Science
THE UNIVERSITY *of York*

Requirements

was the project a success?

- Engineering project
  - **requirements**, design, build, test
  - *evaluate:* does it satisfy the **requirements**?

- Experimental project
  - **hypothesis**, design and perform experiment, results
  - *evaluate:* results confirm/reject the **hypothesis**?

- you have to be able to determine whether/how well you have answered the question
- so … you have to know *what the **question** is* !

have a clear question

- clear statement of what you are trying to build/discover
  - implement “user friendly telepathic application interface”
  - investigate “all birds are white”

- posed in such a way that you can tell if you have answered it
  - requirement: “user friendly”
    - how will you know if yours is?
  - all the birds you investigated were white swans
    - what has your investigation shown?

quantify requirements (1)

- all requirements must be *testable*
  - basis of acceptance tests in real life
  - basis of your project evaluation

- so *quantify* the requirement
- "high availability" ✘
- "availability of 99.9%" ✓
    - (although still not perfect)
  - this quantification *must not* be arbitrary shorthand for "lots" or "very high"
  - it must be derived and justified

quantify requirements (2)

- thinking about testability helps elicit the real requirement

- how would you test “easy to use”?
  - “the product shall be easy to use”
  - “the product shall be easy to use by the general public”
  - quantified: “85% of a representative sample group shall be able to successfully discover their bank balance on first use”

Design

controlled experiment (1)

- you want to test if a new coding style improves readability
  - measuring readability of new style alone tells you *nothing* !
    - "82% of schizophrenics hear voices"
    - "98% of schizophrenics brush their teeth"
  - have to measure readability using both new *and* original style, to have something to *compare*, and to measure the effect

controlled experiment (2)

- *the **control*** : the experiment in the original circumstances, or those with no change

- design of controls is *difficult*
  - "placebo effect" in medicine
    - patients given only dummy drug also get better!
  - "observer effect" in performance trials
    - subjects do better just because they are being watched

controlled experiment (3)

- make sure there isn't a "hidden" change
  - example: effect of vaccine on mice
    - control 0 : no injection
    - control 1 : inject with water
      - effect of *being injected*
    - inject with vaccine
      - effect of *being injected with vaccine*
  - example: effect on search algorithm of using information about its own performance
    - control 0 : no change
    - control 1 : performance measuring code present, data gathered but not used
      - effect of *presence of extra code and data*
    - performance data used
      - effect of *using the data*

double blind experiment

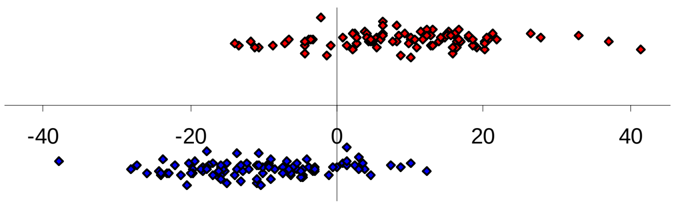
- quantification does not always remove all subjectivity
- testing a subjective effect, like “readability”
  - control group : original style
  - experimental group : new style
  - **subjective** : knowledge of which style can influence answers
- so, perform “double blind”
  - *neither* experimenter *nor* subject know who is in the control and who in the experimental group
    - randomised allocation
    - look at membership after results have been *collected*
      - are the two groups significantly different?
    - look at membership after results have been *analysed*
      - are there two populations that correspond to the two groups?
  - difficult to combine with “control 0” style

the null hypothesis (1)

- a way of casting the question
- usually, H_0 is a statement of the status quo
  - the change has no effect
  - the new design process is no different from the old
  - the new parameter values give the same results as before
- "universal" statements
  - impossible to prove true
    - no effect found, maybe because I haven't looked hard enough
  - but can be ***rejected***
    - exhibit a counter-example
    - exhibit ***statistically significant*** evidence against
      - even then, reject only at a given ***confidence level***
- must be possible, at least *in principle*, to reject H_0

the null hypothesis (2)

- H_0 is assumed true unless data indicate otherwise
    - **we are measuring $p(obs|H_0)$**
- a small value of p means ***reject*** the null hypothesis
    - H_0 = "all swans are white"
        - I observe a black swan
        - p("this swan is black" | "all swans are white") = 0
        - H_0 rejected!
    - H_0 = "no effect on readability"
        - that is, "the readabilities have the same distribution"
        - I observe an improvement
            - that is, I observe a *statistically significantly* different distbn
        - p("I see this sample" | "it follows H_0 distribution") $< \alpha$ (eg, 5%)
        - H_0 rejected, at the $1 - \alpha$ (here 95%) ***confidence level***



the null hypothesis (3)

- *lack of rejection* of H_0 **DOES NOT IMPLY** *proof* of H_0
  - **because we are measuring p(obs|H_0), *not* p(H_0|obs)**
  - H_0 = "all swans are white"
    - I observe a white swan
    - ***p***("this swan is white" | "all swans are white") = 1
    - ***p***("all swans are white" | "this swan is white") = ???
  - H_0 = "no effect on readability"
    - I don't observe an effect
      - I *don't* observe a *statistically significantly* different distbn
    - ***p***("I see this sample" | "H_0 holds") > α' (eg, 90%)
    - ***p***("H_0 holds" | "I see this sample") = ???
      - need to use Bayes' Theorem (and further information, ***p***("H_0 holds"), that you probably don't have) to work this out
- a kind of (probabilistic) "proof by contradiction"

"Does new style improve readability?"

- Null hypothesis H_0 = "*no effect* on readability"

- experimental design
  - trial new and old styles
    - controlled experiment
    - consider whether double-blind experiment is needed
  - gather statistics
    - this needs careful design!
  - analyse results
    - no *statistically significantly* effect on readability : H_0 not rejected
    - improvement in readability : H_0 rejected
    - *decrease* in readability : H_0 also rejected!

"Can reindeer fly?"

- Null hypothesis H_0 = "Reindeer cannot fly"

- experimental design
  - throw several reindeer off a roof
  - result : they all go *splat* on the ground
  - H_0 not rejected
    - this does not *prove* reindeer cannot fly
    - all you have shown is that:

      from this roof, on this day, under these weather conditions, these reindeer either could not, or chose not to, fly

      — [Christmas Guardian, 1980s]

- it is possible, in principle, to reject this H_0
  - by exhibiting a flying reindeer!

sampling bias (1)

- usually experiment on a sample of entire population
  - look at colour of *some* birds, test new alg on *some* data, …
  - need to pick this sample carefully
- are nearby stars more or less luminous than average?
  - H_0 = "distance has no effect on luminosity"
- gather data sample, from a standard star catalogue
  - plot luminosity against distance
  - see a *statistically significant* dependence on distance
    - less luminous stars tend to be closer, more luminous further away
- not a real result : ***sample bias***
  - can't *see* far away low luminosity stars, so not in catalogue
  - high luminosity stars are rare, so aren't many nearby
    - more volume of space further away than nearby ($4\pi r^2 \mathrm{d}r$)

sampling bias (2)

- sample not **representative** of the whole population
  - telephone polls : include only people with telephone (and no job…)
  - most psychology studies done on
      - mentally ill people
      - psychology undergraduates
    - very few done on "ordinary" people!
  - most medical experiment control groups are white males
    - ethnic groups and women introduce "too much variability" !
  - split your data into two samples, to train and to evaluate
    - is the split representative?
      - don't just take the first part to train, last part to evaluate, in case data is in some kind of order
- need unbiased control / test / evaluation samples

design to detect a difference

- problem too easy
  - control and your algorithm can solve it easily
    - everyone does equally well – everyone gets an “A”
  - difficult to detect a difference
- problem too hard
  - neither the control nor your algorithm can solve it at all
    - everyone does equally badly – everyone gets an “F”
  - difficult to detect a difference
- problem just right
    - results spread across the whole measurement scale
  - designed to detect a difference

compare like with like

- “my algorithm is better than X’s”
    - cast as H_0 = “*no different from X’s*”
  - my highly optimised algorithm is better than X’s prototype implementation
  - mine worked better than X’s the one time I ran it
    - I didn’t dare try again, in case it didn’t happen again
  - mine worked better than X’s on this artificial problem
    - which is highly unrepresentative of the real world use
      - over-simplified, inputs too small, unrepresentative synthetic data, …
  - mine worked better on *this* problem than X’s worked on *that* problem
    - and I had to search hard to find this problem

Evaluation and results

independent evaluation data (1)

- hypothesis-generating data
  - data used to help suggest the hypothesis H_0
    - ***I tried a new coding style, and it seemed to improve readability***
    - suggests H_0 = "style has *no effect* on readability"
  - **do not** use *this* data to reject H_0
    - it was used to suggest H_0 in the first place, so ***of course*** it will reject H_0!
- training data
  - data used to train your algorithm
  - **cannot** use this data to evaluate how well your algorithm works *in general*
    - because it might work on *only* its training data !
    - biased training set; "overfitting" the data; …

independent evaluation data (2)

- further ***independent*** test data
  - used to test H_0
  - used to evaluate the algorithm

- three sets of data
  - original data, partitioned into two sets
    - one used to train
    - second used to evaluate
    - "leave one out" cross validation
  - then, third *independent* set, used to evaluate rigorously
    - ***recognising tanks***

http://neil.fraser.name/writing/tank/

results

- you do a well designed, controlled experiment, comparing like with like, properly evaluated

- you get *statistically significant* repeatable results
  - you successfully reject H_0 ☺
    - "my algorithm *really is* better than X's!"
    - you hope you're going to get a good project mark
      - provided you write it up well…
- you *don't* get *statistically significant* repeatable results
  - you haven't rejected H_0 ☹
    - "I've *no evidence* that my algorithm is better than X's"
    - you fear you are going to get a dismal project mark …

understand your results

- actually, whichever result you get, it's not really that interesting, unless you can say ***why***
  - asking "why?" leads to ***new*** questions and hypotheses
    - "it's *statistically significantly* better, ***because*** …"
    - "it's not *statistically significantly* better, ***because*** …"
      - remember : *lack of rejection* of H_0 **DOES NOT IMPLY** *proof* of H_0
    - **"negative" results can be *more* interesting!**
      - presumably, you *expected* to reject H_0
      - so now you've *learned* something
    - if you write-up the "negative" result *well*, you'll get good marks
- need new experiments, get new results
  - good answers to good questions should lead on to even more good questions : "further work"
    - science is an **iterative** "open" process

repeatability (1)

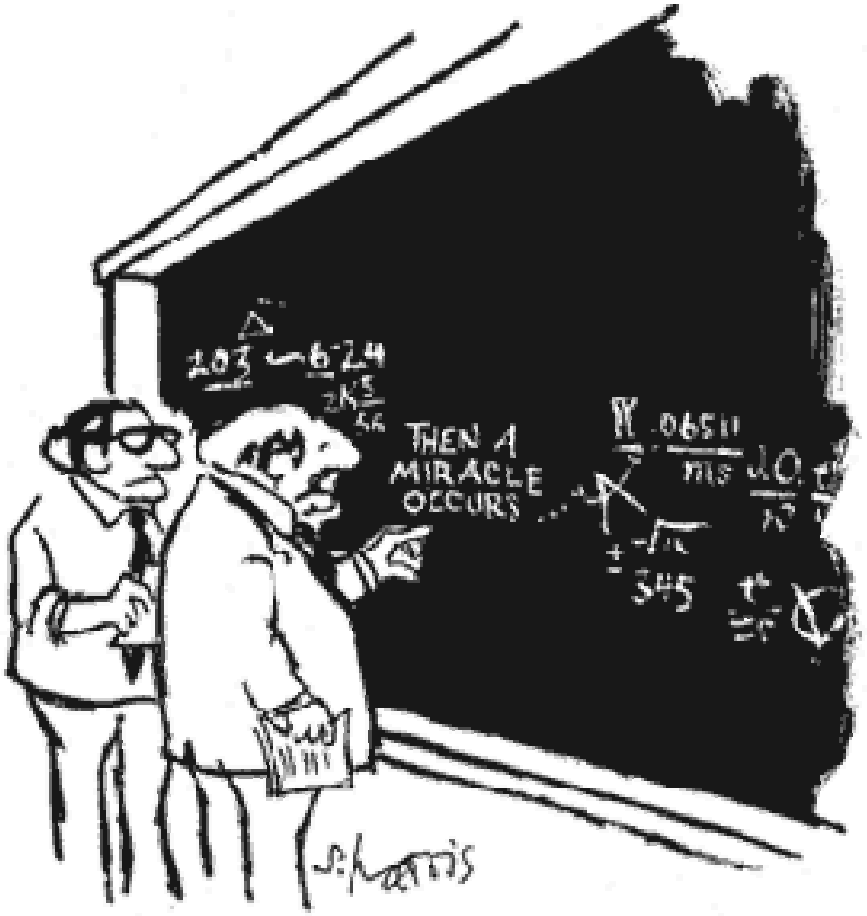
- your results suggest new experiments
- you want to use your original results as a control
  - ***but you can't repeat them !***
    - dependent on precise version of code you used
      - but you were gradually modifying it as time progressed
    - dependent on specific sample of data
      - but you've no record of which data sample you used
    - dependent on the random selection of parameter values
      - but you've no record of which *random seed* you used
  - (but if your results are *too* sensitive to particular parameters, you *might* need to think again)
- use **version control**
  - code, data sample, random seed, …

repeatability (2)

- someone else is running experiments
- they want to use your results as a control
  - ***but they can't repeat them !***

- think about making your code and data available to others
  - the Web makes this easier than it used to be

be explicit



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Scientific Method : summary

- a clear question
  - a quantified testable hypothesis
- a well-designed experiment
  - control
  - null hypothesis
- an evaluated result
  - *statistical significance* [next lecture]
  - "why", as well as "what"
  - repeatable