

Evidence and Experiments

Justin Zobel

University of Melbourne, Australia

Semester 2, 2018

Research revisited

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

The 'doing' of research involves:

- ▶ Forming a definite *question*, the answer to which will satisfy the aim of the research.
- ▶ Gathering of *evidence* that relates to the question.
- ▶ Connecting the question and evidence with an *argument* (chain of reasoning)

That is, experiments are used to gather the evidence that will support, or disprove, a hypothesis.

(Note: in this lecture, technical experiments are used as examples of mechanisms for gathering evidence – that is, data – but many forms of research use other mechanisms.)

Understanding research ...

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

Research concerns *phenomena*: things we can 'see' and measure.

A great deal of research concerns building of better 'eyes'; and verification of whether these 'eyes' are seeing real events.

In this perspective, critical components of research include:

- ▶ Modelling or describing some phenomena.
- ▶ Verifying that the phenomena do indeed occur, that is, taking of *measurements*.
- ▶ Verifying that the observed behaviours are not due to some other cause.

In applied research, the phenomena are created rather than natural.

Evidence

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

In planning to answer a research question, you need to consider:

- ▶ What data may be available – created by you, or sourced from elsewhere.
- ▶ What specific mechanisms will be used to gather and standardize the data.
- ▶ Whether the data will be sufficient in volume or quality to give a robust answer to the question.
- ▶ What domain knowledge may be required to properly interpret the data.
- ▶ What the limits, biases, flaws, and properties of the data are likely to be.
- ▶ How problems in the data will be addressed or contained.
- ▶ What the method of analysis of the data will be.
- ▶ What the results of analysis will show if the data supports the hypothesis; or if the hypothesis is false.

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

A *microarray* can be used to find the genetic profile of a human, by reporting the presence or absence of, say, 10^6 of the commonest genetic mutations.

Typical question: which mutations are linked to extreme height?

- ▶ *What data:* Profile some individuals directly, or (much cheaper!) obtain data from a public genomic database.
- ▶ *What mechanisms:* Data from different microarray platforms needs to be normalized.
- ▶ *Domain knowledge:* Whether the results are biologically plausible.

Evidence example ...

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

- ▶ *Limits*: Microarrays are noisy. Public samples may be biased towards (e.g.) people with disease. Laboratory conditions are variable. 'Extreme height' may be rare. (And etc.)
- ▶ *Addressing problems*: Careful manual data processing, following written guidelines.
- ▶ *Analysis*: Particular statistical or machine-learning methods designed for data where the number of features is absurdly greater than the number of samples.
- ▶ *Results*: Expected p -values; or, perhaps, the accuracy prediction of height from hold-out data.

Purpose of experiments

Experiments are used to learn about a system.

That is, we use experiments to gather data from a system, which we expect to illustrate some of the system's properties.

- ▶ Where are the bottlenecks in query evaluation?
- ▶ Which data structure seems to use the least memory?
- ▶ What is effective or poor about this particular web crawler?
- ▶ What are the in-practice obstacles to the use of this online mechanism for recording patient outcomes?

A system is anything under rigorous examination – it can involve the natural world, or people, as well as technical materials.

Purpose of experiments ...

Experiments are needed for verification of answers to research questions.

- ▶ Which compression algorithm is most effective for transmission of large volumes of data over the internet?
- ▶ Which compression algorithm is fastest for web pages?
- ▶ How much can English text be compressed?
- ▶ How much can a high-resolution image be compressed?

Note how each of these embodies a hidden assumption: the task, or scope.

Purpose of experiments ...

Use analysis, argument, observations to make interesting predictions about a (physical, virtual, ...) system.

Use experiments (that is, investigative data gathering) to test whether the predictions are correct.

- ▶ Must have something to measure, observe, count.
 - What aspects of the system's behaviour are of interest?
 - If it can't be measured, it isn't science. It's just speculation.
- ▶ Must have expectations of the outcomes.
 - What constitutes success? What constitutes failure?

Experiments *confirm* – that is, add weight to our belief in – a hypothesis, or disprove a hypothesis.

Hypotheses & models

Evidence

**Hypotheses &
models**

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

Experiments are typically designed to explore the behaviour of a system or test whether a belief or expectation is correct. A statement of belief is a *hypothesis*.

A hypothesis and a system are related, implicitly or explicitly, by a model (mathematical, algorithmic, informal, even intuitive) that describes likely behaviour. Detailed models yield precise predictions that, if confirmed, are highly persuasive.

Example

- ▶ 'Doubling of database size leads to only a slight increase in query evaluation time' is a testable prediction that follows from a model (an understanding) of how certain kinds of queries are evaluated.
- ▶ 'Social networking can be used to reduce recovery time from illness' is a testable prediction that follows from a model (an understanding) of how people in specific situations react to the influence of peers.

A hypothesis in which there is high confidence is a *theory*.

- Hence the theories of evolution and relativity, which are scientific statements about properties of the physical world in which there is high confidence and for which there is no significant contrary evidence.

Theories can't be conclusively proved, and often can't be conclusively disproved.

In common usage people often say *theory* where a scientist would say *hypothesis*.

(Some researchers say *theorize* when they mean *guess* or *speculate* – this use of 'theorize' is wrong!)

There is only limited use of this form of theory in computing (but plenty of theorems in the mathematical sense); in socially based research, the strict sense of 'theory' does not apply.

Empirical research – examples

Improvements to a database querying algorithm. Might consider:

- ▶ Index size.
- ▶ Index construction time.
- ▶ Query evaluation time (CPU or elapsed).
- ▶ Throughput of queries.
- ▶ Temporary space requirements.
- ▶ Network traffic, disk traffic.
- ▶ Lock requests, other measures of bottlenecks.
- ▶ The number of concurrent users.

That is, an investigation of an algorithm or system should be based on clear aims regarding the problem to be solved.

(And should be based on a clear scope! What queries? What data? How much? And so on.)

Empirical research – examples ...

Independent room navigation algorithm for a robot. Might consider:

- ▶ Time to process each room observation.
- ▶ Proportion of goals successfully achieved.
- ▶ Point-to-point travelling time.
- ▶ Number of collisions.
- ▶ Number of collisions successfully avoided.
- ▶ Number of near-collisions with other moving objects.
- ▶ Realism of plans – whether the robot tends to find the routes that an animal would take.

Specificity versus validity

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

Example research question: can a particular approach to traffic-light scheduling be used to reduce drive times.

- ▶ For an initial investigation, live testing is impractical.
- ▶ A model or simulation must make assumptions about:
 - Typical driver behaviours
 - Road properties (lanes, distances, shape)
 - Total traffic load
 - Road network topology
 - Which roads in the network are subject to peak load
 - Peak versus median load
 - Etc.

(This is an example of the study-versus-case-study issue.)

Research method?

The big question (maybe):

- ▶ One purpose of investigations is to establish facts (observations) on which a hypothesis might be based.
- ▶ Another purpose of investigations is to establish the validity of, or to confirm, a hypothesis.
- ▶ What allows us to use some finite set of investigations to assert that a theory is established
 - that is, to assert that something can for all practical purposes be regarded as true?
- ▶ What allows us to claim that a certain experimental design is good, or sufficient, or superior to the alternatives?

What is science?

It could be argued that science should be defined in terms such as 'a system for building theory on facts' but arguments of this kind are difficult to sustain.

- ▶ For example, the pre-Copernican, heliocentric model of the world was – prior to the invention of the telescope – less consistent with the observed facts than was the geocentric model.

There is no consistent, agreed description of the process of science. Many contenders illuminate or tell part of the story: empiricism, inductivism, logical positivism, falsification, Bayesian reasoning.

Some contenders may be initially appealing but difficult to sustain: relativism, Kuhnian paradigms – not falsifiable?!

(What are these concepts: see Wikipedia.)

Characteristics of science

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

Theory and hypothesis dictating a search for evidence (facts?), facts shaping theory.

- ▶ We learn from facts and use them to develop theories (induction).
- ▶ But we need initial theories to help us search for facts.

Recognising and learning from mistakes.

Use of hypotheses and falsifiability.

- ▶ It must be possible for an experiment to fail.
- ▶ Use of theories with a high information content, that is, are precise.
- ▶ Theories that make precise predictions are easy to falsify, and so unsuccessful attempts at falsification give us confidence.

Characteristics of science ...

Willingness to abandon theories in the face of contradictions.

Theories that are not ad hoc; an ad hoc theory can be used to explain observations, but doesn't suggest discriminating tests.

Flexibility in the face of failure; exploration of whether contradictions are due to an incorrect hypothesis, faulty experimental apparatus, or poor measurement of the experimental outcomes.

Confirmation and confirmability of theories.

- ▶ Seeking of plausible alternative explanations.
- ▶ Identification of experiments that discriminate between explanations and can shift weight of belief in one direction or another.
- ▶ Using prior knowledge in a rigorous way to define 'plausible'.

A variable is an aspect of a system that can influence the behaviour being measured.

Examples

- ▶ Consumption of cholesterol is a variable in heart-disease, but so are age and life-style.
- ▶ The number of processor cycles consumed is a factor in CPU time, but so is overall system architecture (caches, bus, prefetch strategy).

Variables ...

Variables can be unknown – that is, outcomes can be affected by factors other than those known to the researcher.

Example

- ▶ Studies on the effect of caffeine found relationships to lung and bowel cancer. The unknown factor was that caffeine consumption is correlated with smoking. (That is, coffee drinkers have higher than average likelihood of being smokers.)

Experiments need to *control* for variables so that only the factors under study can be responsible for the final result.

Unknowns can be controlled by use of statistical methods.

Explain outcomes:

- ▶ What variables are present in the experiments?
- ▶ What variables can influence the results?
- ▶ What variables can account for the results?
- ▶ How do the experiments distinguish between the effects of the different variables?

Computer systems exhibit complex behaviour. Many factors *might* lead to the same measurements.

Example: Measurement of the CPU time required to complete a task, using old algorithm A and new algorithm B. The implementations have uniform interfaces – only their internals differ. The algorithms take identical input, manipulate it, and produce identical output.

A simple case?

Variables and measurement

Variables that can affect the outcome:

- ▶ A is not as well implemented as is B.
- ▶ A uses more buffer space than B, leading to more paging.
- ▶ A uses floating-point operations that are not supported on the hardware of the test machine.
- ▶ On the test machine A is less able than B to make use of prefetch.
- ▶ Inaccuracies in the timing mechanism randomly favoured B.
- ▶ B was run after A and thus did not generate page faults.
- ▶ Properties of the input. (For example, if A and B are sorting algorithms and the input is sorted, their relative speeds in this case are unlikely to reflect relative speed on unsorted data.)
- ▶ A is intrinsically slower than B.

Variables and measurement ...

Evidence

Hypotheses &
models

Persuasiveness

Philosophy?

Variables

Humans

Statistics

Visualization

Analysis

In other cases:

- ▶ Users come and go, changing system load unpredictably.
- ▶ Initial conditions vary.
(The robot may not always start in the same state or from the same place.)
- ▶ Data varies.
- ▶ The code and platform vary.
(Measurements may be taken over a long period.)

Variables and measurement ...

Distinguish between *random* and *systematic* effects: random affects can average out in the long run, systematic effects introduce consistent bias.

- ▶ Random: Data properties can affect the performance of hash-based algorithms in unpredictable ways.
- ▶ Systematic: For a given small database, data and index are resident on the same disk. For a given large database, they are on separate disks, allowing speed improvements through concurrency.

The human factor

Many kinds of experiment require human assessment – a judgement, say, as to whether or not certain behaviour was observed.

If the assessment is made by the researcher, bias can be a factor in the results.

Example

- ▶ Did the robot traverse the room efficiently?
Is 'efficiently' consistently defined? (Time, resources, displacement, collisions ...?) Is it meaningful at all?

Also, successes are often more memorable than failures. To the researcher there often appear to be good reasons for discarding a negative outcome.

Define what is to be assessed before starting the assessment.

The human factor ...

Intuition is unreliable.

Example:

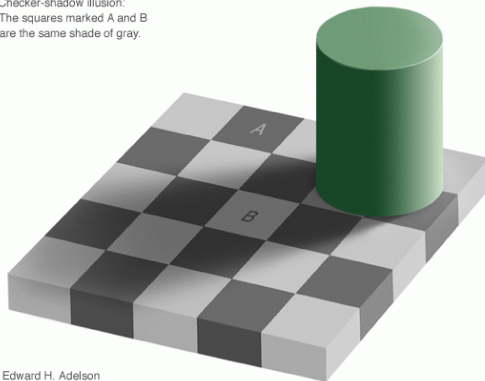
- ▶ To win a game in tennis, a two-point advantage is required.
- ▶ To win a set in tennis, a two-game advantage is required.
- ▶ To win a tennis match, a two-set advantage is required.

Suppose a particular player has a random chance of 40% of winning each point.

What is the probability that this player wins the match?

The human factor . . .

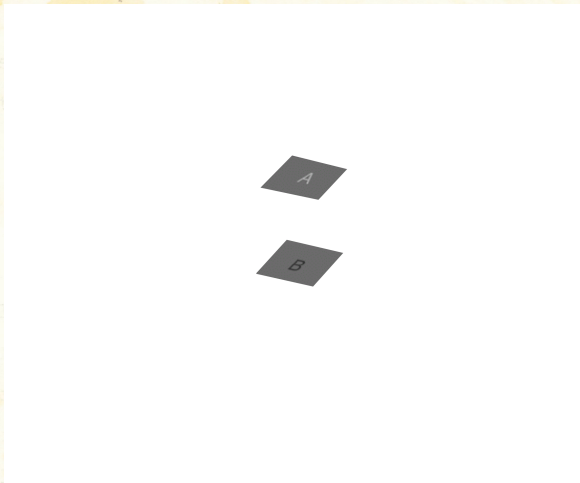
Checker-shadow illusion:
The squares marked A and B
are the same shade of gray.




Edward H. Adelson

(Copyright 1995, used with permission.)

The human factor . . .



(Copyright 1995, used with permission.)



A Test of
Tone Re-Creation

GIVEN BY
MR. GLEN ELLISON

AND
THE NEW EDISON
DIAMOND DISC
PHONOGRAPH

WE believe that every music lover has looked forward eagerly to the day when just such a demonstration of tone re-creation as we are offering today could be made.

Those who hear this test will realize fully for the first time how literally true it is that Mr. Edison has made possible the re-creation of the artist's voice.

No more exciting test could be made to demonstrate that the New Edison actually does re-create the voice of the artist then to play it side by side with the artist who made the records. This is the final proof.

Close your eyes. See if you can distinguish the voice of the New Edison from that of the artist. Did you ever believe it possible to re-create a voice?

Note that the voice of the artist and the voice of the Edison are indistinguishable. This would be impossible were it not for the fact that Mr. Edison has eliminated all mechanical timber. The wonderful diamond styles brings out all those delicate shades and fine distinctions which characterize the human voice.

This re-creation tone test will be particularly interesting to the music student. It shows what great possibilities lie in the New Edison as an aid to their study. It gives a perfect accompaniment, played by great artists and full orchestras—more complete than is at the disposal of the average student. And in the study of any voice or instrument, the student may have as a guide and inspiration the actual interpretations of the world's greatest artists—reproduced with the same tone quality as if they were personally present.

Glen Ellison records exclusively for the New Edison and the Edison Diamond Amberola. His selections have attained universal popularity.

NEW EDISON RE-CREATIONS

By GLEN ELLISON

Price \$1.00 Each

Face to Face With the Girl of My Dreams, <i>Howard</i>	50294
Held Your Hand Out, Naughty Boy, <i>Murphy-David</i>	50256
I Like Your Town, <i>Mellor-Gifford</i>	50277
I Love a Lizzie, <i>H. Lander-Groffon</i>	50361
Make Your Mind Up, Maggie MacKenzie, <i>Mills</i>	50256
My Big Little Soldier Boy, <i>Church</i>	50204
My Bonnie, Bonnie Jean, <i>H. Lander</i>	50352
Nanny (I Have Never Loved Another Girl But You), <i>H. Lander</i>	50352
She Is Ma Daisy, <i>H. Lander</i>	50361
Wee Little House That You Live In, <i>Mellor-Gifford-Godfrey</i>	50277
When I Leave the World Behind, <i>Berlin</i>	50301

EDISON BLUE AMBEROL RECORDS

By GLEN ELLISON

Price 50c Each

Held Your Hand Out, Naughty Boy, <i>Murphy-David</i>	2689
I Like Your Town, <i>Mellor-Gifford</i>	2696
Make Your Mind Up, Maggie MacKenzie, <i>Mills</i>	2667
My Bonnie, Bonnie Jean, <i>H. Lander</i>	2727
My Big Little Soldier Boy, <i>Church</i>	2546
Nanny (I Have Never Loved Another Girl But You), <i>H. Lander</i>	2905
She Is My Rosie, <i>H. Lander</i>	2871
Sing Us a Song of Bonnie Scotland, <i>Fayne</i>	2824
Wee Little House That You Live In, <i>Mellor-Gifford-Godfrey</i>	2721
When I Leave the World Behind, <i>Berlin</i>	2749

Statistics in experiments

It is unreasonable to expect to reliably observe behaviour on a single run:

- ▶ To do so would require controlling for all variables other than the one under study.
- ▶ Most studies involve unknown variables with unpredictable effects.

It is unreasonable to expect to manually discover pattern in large volumes of data.

Statistical techniques can be used as tools for exploring data and for giving confidence that effects are real, not the product of unknown variables.

Populations

An underlying assumption is that there is a *population* of data sets, from which a particular test set has been selected.

Populations are typically either extremely large or infinite.

It is not in general possible to examine every member of a population (though it is sometimes tempting to try ... :-)

However, it is important to understand what the population is, so that 'member' is clearly defined.

- ▶ What is a 'typical' obstacle course for a robot?

Is it not always clear what 'typical member' might mean. For example, should the definition of a 'typical' word in English allow for the fact that some words are more frequently used than others?

Measurement of a random member of a population is just that – random.

Indeed, a measurement of a single member is expected to be untypical!
There is surprisingly little relationship between individual behaviour and group behaviour.

- ▶ The number of babies born each month in Melbourne is almost constant.

Statistically, performance measured on a single member of a population is meaningless.

To gain *confidence* in a measurement it needs to be based on a *sample* of the population.

A sample is a set of members of a population, drawn at random.

The method of selection of members for a sample can control for known variables. For example, two samples of adult males might be selected so that both had the same proportion of smokers.

A sample is assumed to be representative of the population. There may be systematic and random biases due to selection error and 'luck'.

Randomised sampling can control for unknown variables.

Samples in computing experiments:

- ▶ A series of runs over the same data, giving typical speeds.
- ▶ Different network configurations.
- ▶ Different data.
- ▶ Different implementations over the same data.
- ▶ Different users.

(These are only a sample ...)

It is usually necessary to explain the sampling procedure so that biases are not concealed.

Statistical analysis

Given some measurements over each member of a sample, statistical tools can be used to analyse likely performance.

The simplest analytical tool is averaging.

- ▶ Even an average can be misused. Some distributions do not have a meaningful average.

Given an average value for a sample of a population, further tools can be used to state the likelihood that measurement of a random member of the population will be close to the average.

These tools use information such as the distribution of the values for the individuals in the sample, size of the sample.

Larger samples give greater confidence in the reported values, but increase the likelihood of outliers being observed.

Hypothesis testing

Samples can be compared to determine the likelihood that they represent different populations.

That is, suppose that we have two samples with different means.

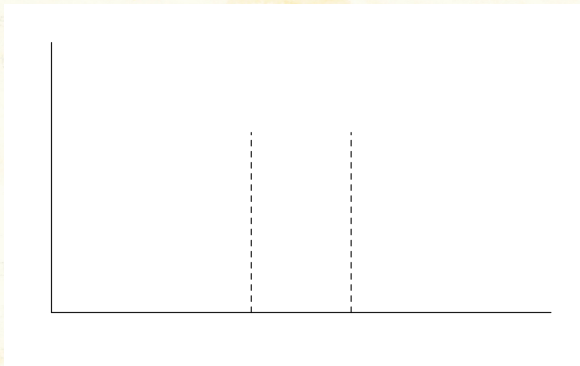
Statistical hypothesis tests can be used to determine the likelihood that the means of the underlying populations are indeed different.

- ▶ Comparing two image-processing algorithms on a collection of JPG files, the first correctly identifies objects more often than does the second. Is the first algorithm better?

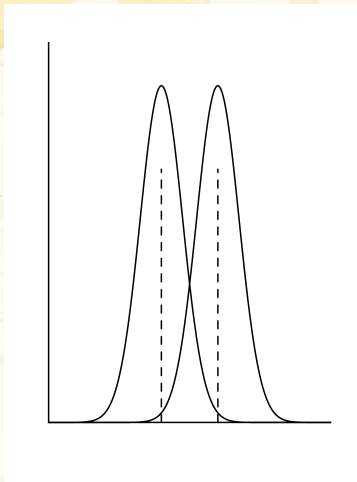
The *power* of a test depends on the data volume – it determines the size of the effect that can be observed with high confidence.

Power is used to ensure that an experiment is sufficiently sensitive.

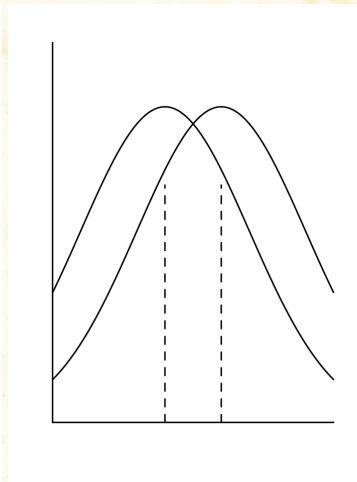
Hypothesis testing ...



Are these means drawn from the same population?



Probably not



Possibly

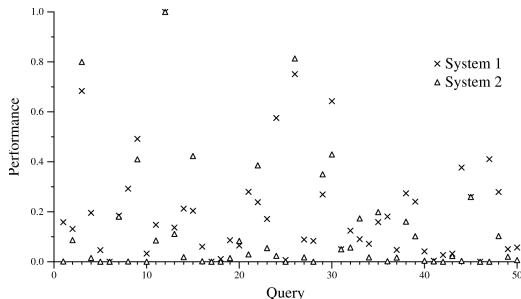
Use graphs and scatterplots to look at distributions and correlations.

Think of different ways to aggregate and visualise the data to draw out possible properties.

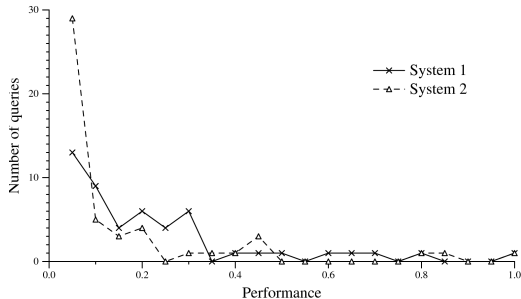
Visualisation is a valid way of presenting evidence.

However, visualisation is not a substitute for evidence. Graphs and figures need careful interpretation.

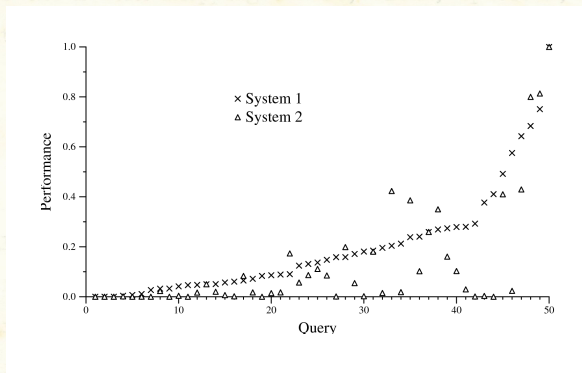
Consider the two systems represented in this graph.



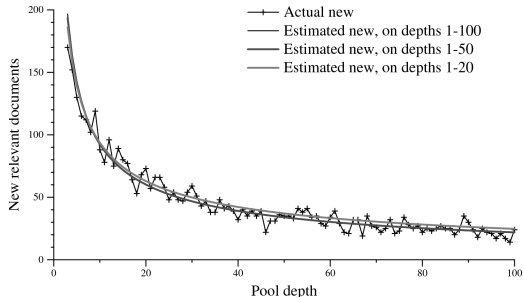
System 1 average is 0.20, system 2 average is 0.13. Which is superior?



This is the same data as shown earlier, with similar values aggregated.



The same data as before, sorted by performance on System 1.



A visual demonstration quality of fit of a curve to some data.

Analysing evidence

Experiments are noisy – data points can be flukes, freaks, outliers.

Discard outliers that are demonstrably in error.

Identify outliers that cannot be explained.

Consider all outcomes, not just favourable outcomes.

(Classification of outcomes can be a good starting point for further research but does not necessarily add weight to evidence.)

Undertake ‘failure analysis’ to explain why each outcome occurred. That is, look inside each experiment.

Explain why the outcomes are evidence for the hypothesis.

Use confidence intervals, hypothesis tests, test theory.

How likely is a coincidence? Very likely.

What distinguishes members of one half of a population from members of the other half? Something – by random luck. Only if the distinction has been predicted is it of interest.

Why do stone-age artifacts appear to have been influenced by aliens?
Because of sampling bias.

Why do the 10 poorest students on exam A do better on exam B?
Individuals that are outliers on one property are not necessarily outliers on another, even if the properties are correlated.