CS2309 CS Research Methodology

Science

Lee Wee Sun
School of Computing
National University of Singapore
leews@comp.nus.edu.sg

Semester 1, 2010/11

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Science

Scientists spend their lives trying to understand the world, unraveling the secrets of nature.

- Science is based on observation.
  - A scientist rejects authority as an ultimate basis for truth.
  - In practice, we need to use facts established by others, but ...
    - Scientist reserve the right to judge for themselves whether the methods used are appropriate and whether the facts are credible.
    - A scientist would repeat and test the work of others whenever he or she feels it is desirable.
    - A scientific publication is required to contain enough details to allow someone else to replicate the experiments.
- The collective judgement of scientists, when there is substantial agreement, constitute the body of science.

Outline

- Science
- Inductive Principle
  - Deductive and Inductive Reasoning
  - Bayes' Rule
  - Reproducibility
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Deductive and Inductive Reasoning

Scientists use both deductive and inductive reasoning.

- In deductive reasoning, we start from a set of assumptions and proceed to deduce the consequences.
- In inductive reasoning, we use observations in order to discover scientific "laws" and regularities in nature.

Deductive Reasoning

Here we look at simple deductive reasoning.

- Given a collection of premises and a conclusion, an argument is an assertion that the conjunction of the premises implies the conclusion
- In deductive reasoning, an argument is valid if the assertion is always true i.e. it is a tautology.

Example of a valid argument

- Premise 1: All men are mortal
- Premise 2: Socrates is a man
- Conclusion: Therefore, Socrates is mortal

The pattern of reasoning in this example is called a syllogism and is descended from the time of Aristotle.

The validity of the argument can be checked with a truth table

- Premise 1: p implies q
- Premise 2: p is true
- Conclusion: q is true

p	q	$p \rightarrow q$	$(p \rightarrow q) \wedge p$	$[(p \rightarrow q) \wedge p] \rightarrow q$
F	F	T	F	T
F	T	T	F	T
T	F	F	F	T
T	T	T	T	T

Another valid argument:

- Premise 1: p implies q
- Premise 2: q is false
- Conclusion: p is false

p	q	$p \rightarrow q$	$(p \rightarrow q) \wedge (\neg q)$	$[(p \rightarrow q) \wedge (\neg q)] \rightarrow (\neg p)$
F	F	T	T	T
F	T	T	F	T
T	F	F	F	T
T	T	T	F	T

The following is not a valid argument[1].

- Premise 1: p implies q
- Premise 2: q is true
- Conclusion: p is true

p	q	$p \to q$	$(p \to q) \wedge q$	$[(p \to q) \wedge q] \to p$
F	F	T	F	T
F	T	T	T	F
T	F	F	F	T
T	T	T	T	T

[1] However this pattern is often used by human reasoning in a weaker form. More about this later.

A Historical Example

Ptolemiac System

- From antiquity, people thought that earth was the center of the universe; sun, planets and stars revolve around it.
- Ptolemy's model (150 A.D.)
- Epicycles, where planets rotate around a small axis while rotating around a larger axis around earth, are used to explain the planets' retrograde motions.

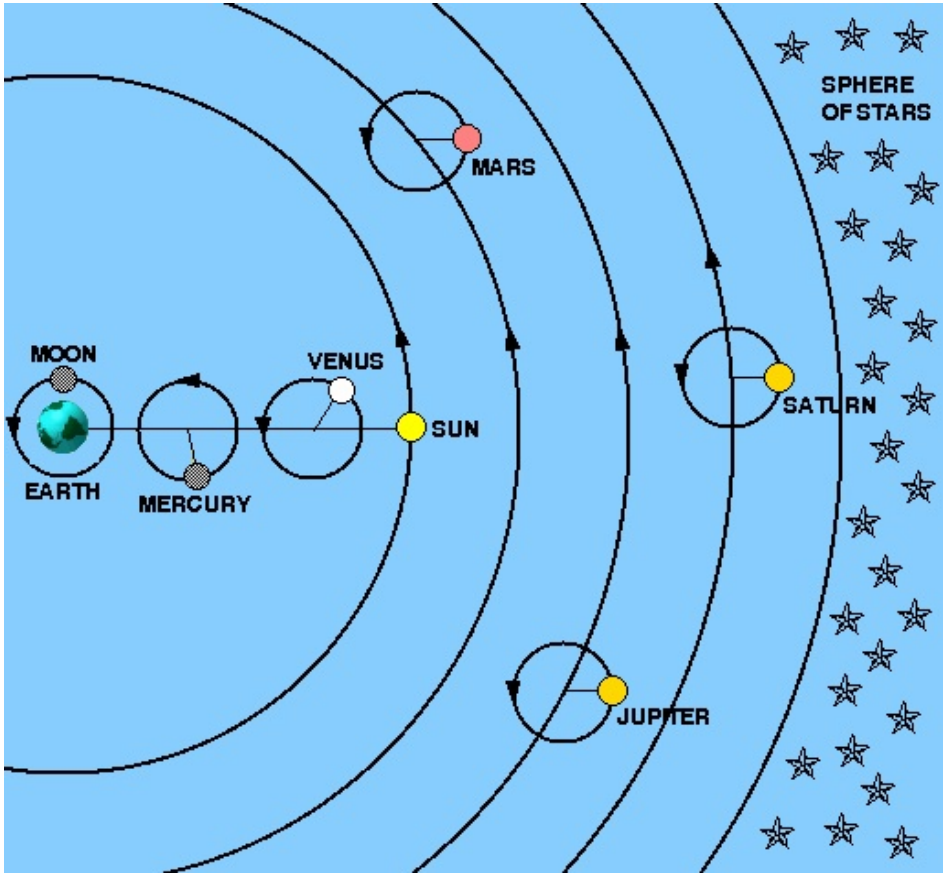


Image from http://www.shef.ac.uk/physics/people/vdhillon/teaching/phy105/phy105_ptolemy.html

Copernican System

- Copernicus (1473-1543) proposed a heliocentric system, where the sun is the center; earth and the planets revolve around the sun.
- We are no longer the center of the universe.
- At that time, observations were insufficient to distinguish which model is correct.

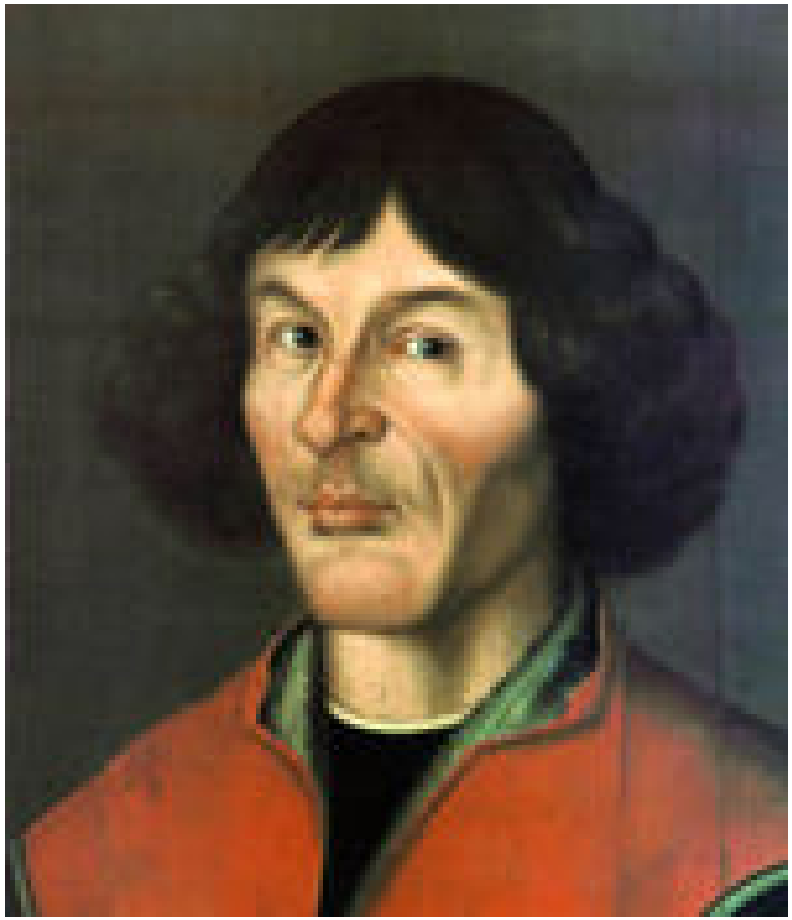


Image from `http://en.wikipedia.org/wiki/Image:Nikolaus_Kopernikus.jpg`

Galileo's telescope

- Along came Galileo (1564-1642) who invented the telescope, making observations that distinguish the models possible.
- Copernicus' theory implies that all phases of the Venus will be observed.
- Ptolemy's theory implies that not all phases can be observed.

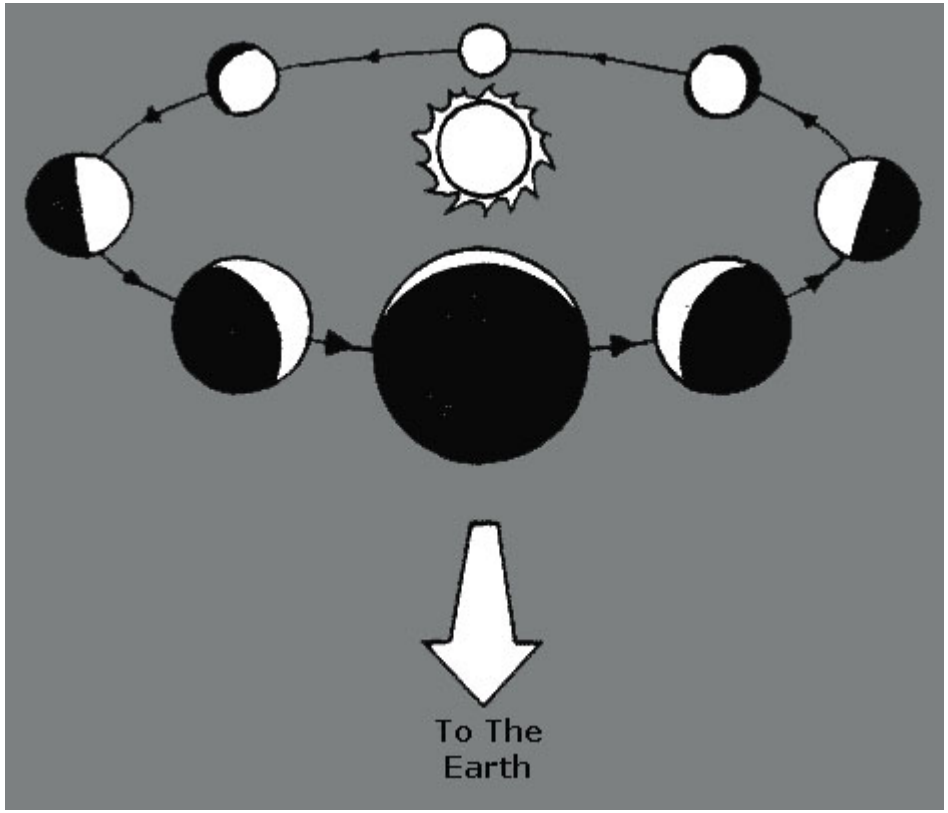


Image from http://www.souledout.org/nightsky/venusphases/venusphases.html

Watch video from http://phys23p.sl.psu.edu/phys_anim/astro/ptolemy_v_phases_avi.html and http://phys23p.sl.psu.edu/phys_anim/astro/copernicus_v_phases_avi.html.

Which of the following is a valid argument?

Copernicus is correct

- Premise 1: Copernicus correct implies all phases are observed
- Premise 2: All phases are observed
- Conclusion: Therefore, Copernicus is correct

Ptolemy is wrong

- Premise 1: Ptolemy correct implies not all phases can be observed
- Premise 2: All phases are observed
- Conclusion: Therefore, Ptolemy is wrong

Induction

- We cannot logically prove that Copernicus is correct. What can we do?
- Most scientists work by induction
  - All the swans that we have seen are white.
  - Therefore all swans are white.
- Is this actually a reasonable way to work??

Outline

- Science
- Inductive Principle
  - Deductive and Inductive Reasoning
  - Bayes' Rule
  - Reproducibility
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Bayes' Rule

- Probability theory provides some support for inductive reasoning.
- From Bayes' rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Converts prior probability $P(A)$ into posterior probability $P(A|B)$ of A having observed event B.
- If we expand $P(B)$ into

$$P(B) = P(B \wedge A) + P(B \wedge \neg A) = P(B|A)P(A) + P(B|\neg A)P(\neg A)$$

we get

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

- If a theory predicts an observation, then given
  - T is the event that the theory is true
  - O is the event that the observation confirms the prediction

  this is the same as saying $P(O|T) = 1$
- In this case, Bayes' rule gives

$$P(T|O) = \frac{P(O|T)P(T)}{P(O)} = \frac{P(T)}{P(O)}$$

- Since $P(O) \leq 1$, $P(T|O) \geq P(T)$.
- If an observation agrees with the theory's prediction, the probability that the theory is true is increased.
- Agrees with intuition.
- Copernicus has higher probability of being correct after the observation of all phases of Venus!

- Since $P(O|T) = 1$, we have $P(\neg O|T) = 0$.
- Bayes gives us

$$P(T|\neg O) = \frac{P(\neg O|T)P(T)}{P(\neg O)} = 0.$$

- This agrees with deductive logic. If a theory predicts some consequence and the consequence is found to be false, then the theory is false.
- In practice, there are always observation errors, so $P(O|T)$ is only approximately 1, but the qualitative behaviour is the same.
- Repeated testing with different predictions of the theory increases our confidence in the theory as long as it is not discredited.

Outline

- Science
- Inductive Principle
  - Deductive and Inductive Reasoning
  - Bayes' Rule
  - Reproducibility
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Reproducibility and Engineering

- After seeing one million white swan, there is still a small probability that we will see a black swan in future.
- Cannot get absolute truth.
- However, for most practical engineering purposes, reproducability is sufficient - we don't need absolute truth.
- As long as we can predict the outcome with high reliability, we can design our system sufficiently well
  - We still use Newton's theory of gravitation instead of Einstein's general theory of relativity for computation when we know it is reliable.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Popper

- Popper (1902-1994) was not satisfied with induction as the basis of science.
- He noted that a counter-example is enough to discredit a theory and proposed falsification as the basis of doing science.
    - According to Popper, scientists should use their creativity to produce bold theories such as those of Kepler and Newton.
    - The theories are tested as severely as possible, and discarded if found to be false.
- Attempts to prove has been replaced with attempts to disprove!

- Popper propose falsifiability as the criterion for demarcating science from non-science.
  - If a theory does not have to make predictions that can potentially be falsified through observation, then it is not science.

Falsifiable?

Which of these statements are falsifiable?

- My computer program can think.
- My computer can fool you into believing that you are communicating with another person instead of to a computer.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

The Scientific Method

The exact procedure varies from one area to another, but the scientific method consist roughly of:

- From prior observations, construct a hypothesis that explains the observations.
- Design experimental studies to test the hypothesis.
- Gather data.
- Use data to test hypothesis.
- Modify the hypothesis in light of the data and iterate.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
  - Correlation and Causation
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Observational and Manipulation Experiment

- In manipulation experiment, the scientist directly manipulates the variables and measure the changes in the outcome.
  - Properly designed manipulation experiment can give causal link between variables i.e. it tells you that if you can control the outcome by changing a particular variable.
  - Necessary if causal link needed as correlation does not imply causation.
- In observational research, the scientist observes what goes on without manipulating any of the variables.
  - Sometimes not possible to do manipulation research for various reasons, e.g. ethical reasons.
  - Experimental research sometimes cannot be done in the natural setting - the outcome may be biased by the setup necessary to do the experiment.

Smoking Causes Cancer

- To check whether smoking causes cancer, we can perform observational research by checking the rate of cancer among smokers and non-smokers.
- Assume we find that smokers have a higher rate of cancer.
- The tobacco industry can claim that that it is just a coincident that the smokers have a higher rate of cancer - there is a gene that tends to make a person likely to enjoy smoking and also likely to have cancer.
- We can rule out this argument by performing an manipulation research.
  - Randomly select two groups of people; force one group to smoke.
  - Check the cancer rate of the two groups after a sufficiently long time.
- This would give a causal link between smoking and cancer but it would also be unethical!

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
  - Control
  - Randomization
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Control

- The ideal experiment is one where all the variables are held constant except the one under study.
- One device for achieving this is the introduction of controls.
  - These are similar test specimens that are subjected to as nearly as possible the same treatment as the objects of the experiment, except for the change in the variable under study.
- A confounding variable is a variable that affects the dependent variable but has not been considered or controlled for.
  - The tobacco industry would claim that the gene that causes someone to both like smoking and be likely to develop cancer is a confounder.
- An experiment lacks validity if it fails to take care of confounding variables appropriately.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
  - Control
  - Randomization
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Randomization

- The standard way of preventing biases and unknown factors from affecting the validity of experiments is to use randomization.
- The allocation of subjects to the treatment and control group is randomized, e.g. by tossing a coin.
- By randomization, the unknown factors should be equally present in both groups and hence their effect should statistically cancel out leaving the variable under study as the only variable that changes.
- Randomization also provides the basis for calculating the probability of error.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
  - Sampling Bias
  - Order and Location Effect
  - Change of Equipment
  - Effect of Measuring
  - Effect of the Experimenter
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means

Sampling Bias

Intellectual Inferiority of Women

- In the 19th century, French neurologist, Paul Broca, inspired by evolution theory, investigated the development of human intelligence by carefully measuring brain weight.
- He found that
  - Caucasian men had larger brains than caucasian women.
  - Caucasian women in turn had larger brains than negroes.
  - Modern brains were heavier than mediaval brains.
  - French brains were heavier than German brains.
- French middle-class men were the pinnacle of evolution!

Le Bon (1879), one of the founders of social psychology

In the most intelligent races, as among the Parisians, there are a large number of women whose brains are closer in size to those of gorillas than to the most developed male brains. This inferiority is so obvious that no one can contest it for a moment: only its degree in discussion. All psychologists who have studied the intelligence of women ... recognize today that they represent the most inferior forms of human evolution and that they are closer to children and savages than to an adult, civilized man.

- Hold on. Before we relegate the women in this class to making coffee ... no relationship between brain weight and intelligence within a species has been found.
  - Size is not a valid measure of intelligence.
- Putting aside the validity, what gave rise to the differences in measurements that Broca found?
- According to Gould (1981)
  - His female brains mainly come from elderly women while male brains from younger males who died in accidents - brains shrink with age.
  - Brain size is related to body size - when normalized for body size, no difference was found.
- There is systematic biases in sampling.
  - How can we remove these biases?

The sampling distribution can be particularly important in surveys.

Biases in Surveys

Are there any problems with the following survey methods?

- Doing an internet survey.
- Performing survey at an MRT station.
- Go door to door during the day. How about evening?

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
  - Sampling Bias
  - Order and Location Effect
  - Change of Equipment
  - Effect of Measuring
  - Effect of the Experimenter
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means

Order and Location Effect

Location, time, and order of presention can often have an effect on the outcome.

Industrial lab experiment

- In an industrial lab, experiments were performed to determine the effect of the length of time of pressing in the mold on the strength of a plastic part.
- Hot plastic was introduced in the mold, pressed for 10 seconds, removed
- Another batch was pressed for 20 seconds, and so on with time increasing with each batch.
- It was found that the strength increased with the duration.
- It was suggested that the mold gets warmer as time goes on, so the increased strength may just be because of that.
- Suggest an alternative design.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
  - Sampling Bias
  - Order and Location Effect
  - Change of Equipment
  - Effect of Measuring
  - Effect of the Experimenter
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means

Change of Equipment

Care needs to be taken that different equipment does not make comparisons invalid.

Comparing Algorithms

You wish to compare the running time of your new sorting algorithm, fancy-sort, against other algorithms. You obtain the results of other algorithms from the original research papers. You then run your algorithm on the same datasets. In comparison, your algorithm runs much faster than the published result.

- Is this a good way to do the comparison?
- Suggest a better way.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
  - Sampling Bias
  - Order and Location Effect
  - Change of Equipment
  - Effect of Measuring
  - Effect of the Experimenter
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means

Effect of Measuring

Doing the measurement itself can often affect the outcome.

Placebo Effect

The placebo effect refers to the effect produced by any treatment when the subject believes that he or she has been given an effective treatment.

- Suggest a design in light of the placebo effect.

Surveys

Survey methods have to be carefully designed...

- Asking questions related to prestige or embarassment in face to face survey.
- Question: *NUS is a very well respected university internationally. What ranking would you give NUS among universities worldwide?*

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
  - Sampling Bias
  - Order and Location Effect
  - Change of Equipment
  - Effect of Measuring
  - Effect of the Experimenter
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means

Effect of the Experimenter

Often the experimenter make (unconscious) decisions that affect the outcome of the experiment.

Lanarkshire Milk Experiment

- In spring 1930, an experiment was carried out in Lanarkshire, England to determine the effect of providing free milk on the height and weight of school children.
- Initially, assigned at random, but teachers allowed to use their judgment in switching children between treatment and control to obtain a better balance between undernourished and well-nourished individuals in each group.
- Later, the controls were found to have been distinctly heavier and taller than the subjects before the trials began.
- Teachers could (perhaps unconsciously) have adjusted initial randomization because of sympathy, obscuring the result.

Double Blind Experiments

- To avoid the placebo effect and the effect of the experimenter, medical experiments are often double-blind.
- Neither the doctor nor the patient knows whether the patient is in the treatment or control group - assigned randomly and not known to the doctor.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Example: MYCIN

MYCIN, developed at Stanford in the 1970s is one of the best known successes of AI.

- Recommend therapy for blood and meningitis infection.
- Does as well as human experts.

How is MYCIN shown to do as well as human experts?

- Randomized blind experiment (control for human bias): judges do not know whether human or MYCIN is making a particular recommendation. Eight humans, each solve 10 problems. Another 10 from MYCIN and 10 more from the original attending physician for 100 recommendation in total.

- Of the 8 humans only 5 are experts. One is a research fellow, one a resident doctor and one a medical student. Control for problem being too easy (ceiling/floor effect) such that everyone gets it right.
  - If MYCIN does as well as experts, it would do better than normal doctors and medical students - including them allows good to be distinguised from average.
  - The setup also tests that having more knowledge gives better performance - experts have more knowledge compared to normal doctor. MYCIN would also allow manipulation experiments to be done by subtracting rules from database.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Constructing Hypothesis

- How do you know what hypothesis to test? What to look out for in prior observations?
- According to Mills (1843)
- The method of agreement states that
  - if the circumstances leading up to a given event have in all cases had one factor in common, that factor may be the cause sought.
- The method of difference states that
  - if two sets of circumstances differ in only one factor and the one containing the factor leads to the event and the other does not, this factor can be considered a potential cause of the event.

- The method of concomitant variation states that,
  - if there is a factor which varies together with the effect we are interested in, e.g. the strength of the effect increases when the strength of the factor increases, then the factor is a potential cause of the effect.
- Preliminary exploratory experiments may be necessary.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Hypothesis Testing

- Experiments are often carried out to test hypotheses
- It is important to know whether the effects observed are real or merely the result of chance.
- For example, if we do an experiment to test whether a new drug helps to cure a certain disease
  - One possibility is that the result of the experiment is purely due to chance e.g. the average recovery rate is the same whether the drug is used or not. This is commonly used as a null hypothesis H_0.
  - Another possibility is that the drug has some positive or negative effect e.g. the average recovery rate is different. We may use this as an alternative hypothesis H_A.
- Note that hypothesis testing only tells you whether the effect is likely to be due to chance, not whether the effect is important. Need domain knowledge to judge that.

The Lady Tasting Tea

At a tea party attended by academics and their wives in Cambridge in the 1920s, one of the ladies claimed that tea taste different depending on whether

- tea is poured into milk, or
- milk is poured into tea.

Ronald Fisher proposed an experiment to test her claim:

- Prepare 8 cups of tea, four prepared in each way.
- The lady, who did not see the tea being prepared, has to figure out which is which by tasting.

The lady apparently got all of them right!

- What is the probability that the lady obtained the result purely by chance?
- Before answering that, note the clever use of randomization.
  - We are able to quantify the probability because we know the process that generates the randomness - in this case all possible outcomes are equally likely.
  - The use of randomness eliminated other factors that may have confounded the result.
- There are $\binom{8}{4} = 70$ possible outcomes out of which only one is the all correct outcome.
- The probability of being correct by chance is $1/70 = 0.014$.
- Fisher would usually reject the null hypothesis if the probability of it being correct is less than 5% - so this is unlikely to be a chance event.

Type I Error

Since the test is statistical, two types of errors are possible

Type I error

- The null hypothesis H_0 is rejected when it is true.
- In our example, this means rejecting the hypothesis that the observations are due to chance when they actually are due to chance.
- The probability of this happening is normally denoted by α and is called the significance level of the test.
- Fisher would be upset if the experiment indicate that the lady can tell the difference in the tea when she could not!
  So he would try to design his experiment to have a small α value.
- The set of outcomes for which the null hypothesis is rejected is called the critical region: if we set the critical region to the event that the lady is totally correct, then $\alpha = 1/70$.

Type II Error

Type II Error

- The null hypothesis H_0 may be accepted when it is false.
- In our example, this means accepting that the observations are due to chance when they are actually not.
- The probability of this happening is normally denoted β.
- The probability that H_0 is rejected when it is false, $1 - \beta$, is called the power of the test.
- The lady would be upset if the experiment fails to detect that she can tell the difference in the tea
  She would demand a more powerful test!

There is a trade-off between type I and type II errors - accepting the lady's claim even when she makes more errors would decrease type II error but increase type I. Increasing the sample size would improve both the power and level of significance of a test.

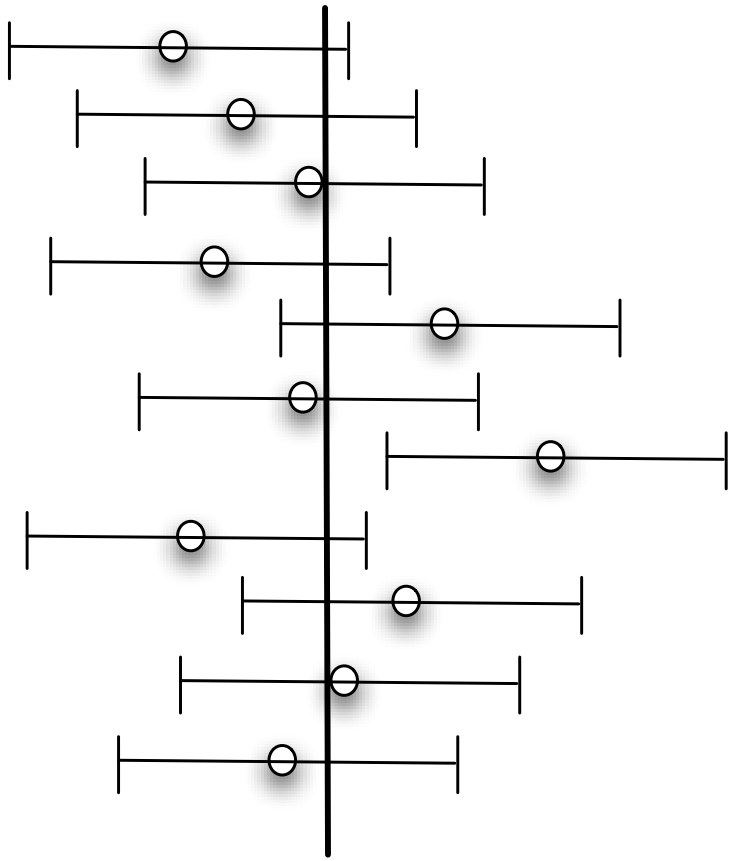
Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Confidence Interval

- Statistical inference is about infering parameter values from statistics.
- Statistics are function on samples while parameters are functions on populations.
- Hypothesis testing answers yes-no questions about the population parameters e.g. the mean value. It gives probabilities that the answer is wrong.
- We are often interested in the range of values that the parameter is likely to take.
- The confidence interval gives the region where the parameter is likely to fall into, given the statistic.

- For example, if we have a Gaussian random variable, the interval $m \pm 1.96\sigma$, where m is the value of the observation and σ is the standard deviation, would be a 95% confidence interval.
- Assume that we are given a 95% confidence interval for the parameter. What does this mean?
  - The correct interpretation is that if we repeatedly do the experiments, 95% of the time the true parameter would be within the interval we specify.



Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Gaussian Distribution

Gaussian Distribution

- Also called normal distribution or bell curve.
- Most commonly used continuous distribution with density function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(x-\mu)^2/2\sigma^2}$ where μ is the mean and σ is the standard deviation.
- Models many natural measurements well - one reason is the Central Limit Theorem.

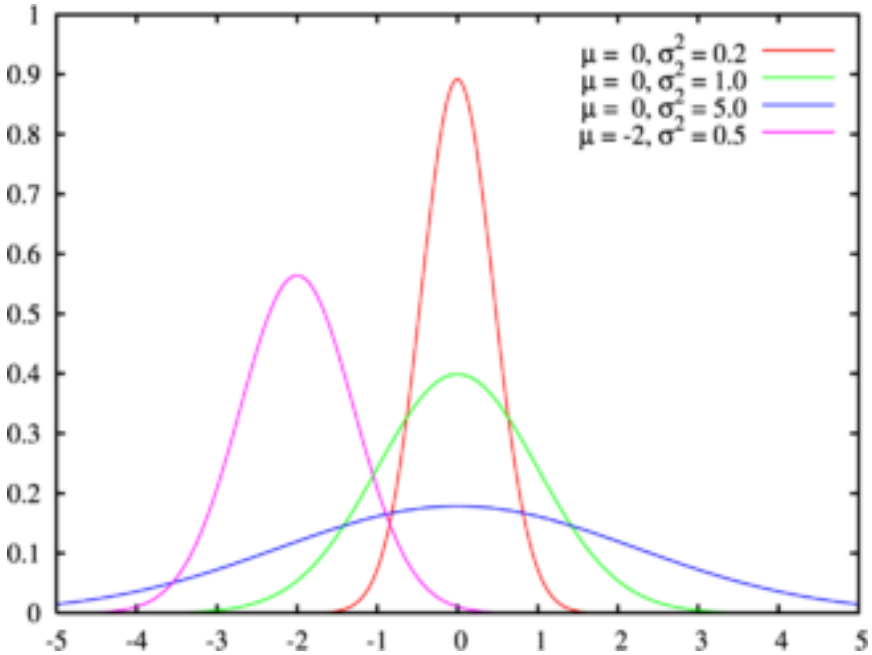


Image from http://en.wikipedia.org/wiki/Image:Normal_distribution_pdf.png

Central Limit Theorem

- The sampling distribution of means approaches the normal distribution as the sample size increases.
- More specifically, for samples drawn from a distribution with mean μ and standard deviation σ, the distribution of means approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{N}$ where N is the sample size.
- This holds regardless of the original distribution, even if it is highly skewed.
- The standard deviation of the sampling distribution of the mean is often called the standard error and decreases as N increases.
- The approximation is usually regarded as reasonable for $N \geq 30$.
- We can reasonably assume that the means of the variable are normally distributed for our hypothesis testing, even when we do not know the distribution of the original variable!

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
  - Z test
  - t test
  - Two sample t test
  - Paired t test
- Error Bars

Z Test on the Mean

Comparing Algorithms

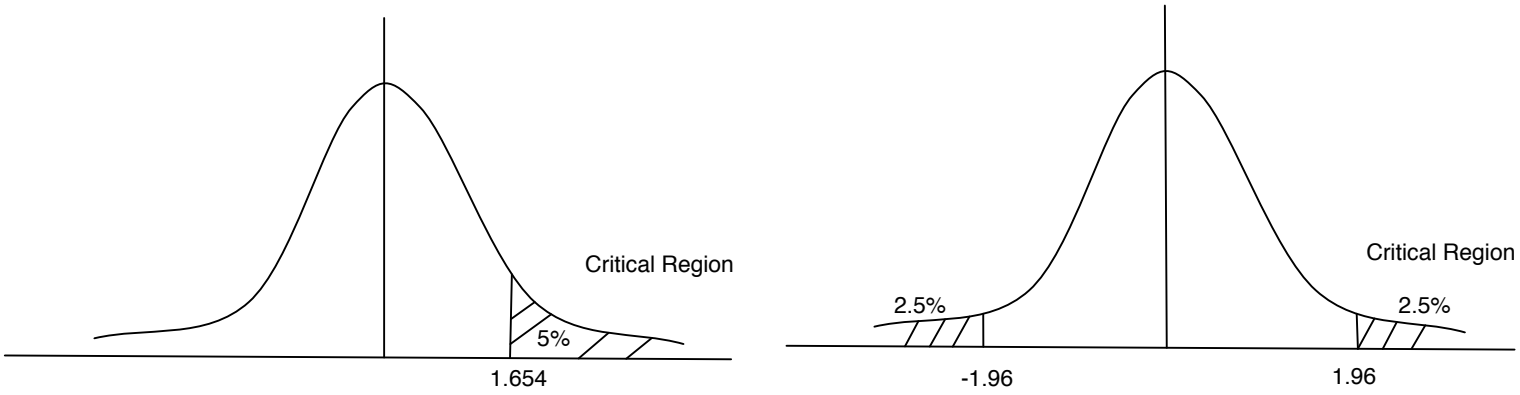
Assume that we are comparing our new algorithm with an old one. Assume that we have been running the old algorithm for a long time and we know that its average performance (running time, accuracy, etc.) is μ_d and the standard deviation is σ.

The Z test is appropriate for this. In the example, we may form the following null and alternative hypothesis

- H_0: $\mu_d = \mu$
- H_1: $\mu_d > \mu$

If all we know about the population is that the mean is μ and std dev is σ, the central limit theorem indicates that assuming that the sampling distribution of the mean is normal with mean μ and standard deviation $\sigma/\sqrt{N}$ is reasonable when N is large.

- Let $\bar{x}$ be the measured sample mean. If we center it by subtracting μ and normalize the result by dividing by $\sigma_{\bar{x}} = \sigma/\sqrt{N}$, the resulting random variable $Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ is a Gaussian random variable with zero mean and standard deviation of 1.
- The null hypothesis is that the sample mean has the standard Gaussian distribution (zero mean, std dev 1).
- From the Gaussian distribution, we can calculate (look up tables, or use computer program) that the probability that $Z > 1.645$ is no more that 0.05.
- Hence, if we reject the null hypothesis when the value of Z is more than 1.645, we will be wrong at most 5% of the time, giving a level of significance (p value) of 0.05.



- The previous example is a one-tailed test as the rejection or critical region is only on one side, corresponding to the additional knowledge provided by the alternative hypothesis that the sample mean is larger than the hypothesized mean.
- If we do not know if it is larger or smaller, we may want to do a two-tailed test where the alternative is H_1: $\mu_d \neq \mu$.
- In this case we form rejection regions on both sides of the mean.
- For a standard normal distribution, we can reject the null hypothesis if $|Z| > 1.96$ to obtain a p value of 0.05.

- When the population standard deviation is not known, it is common to estimate it from data by the sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{N}(x-\bar{x})^2}{N-1}}$.
- In this case, the Z test is often run with $Z = \frac{\bar{x}-\mu}{s/\sqrt{N}}$.
- If $N > 30$ the approximation used in the Z test is usually reasonable. But when N is small, the t test is usually used.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
  - Z test
  - t test
  - Two sample t test
  - Paired t test
- Error Bars

The t-Test

- When the distribution from which the sample is drawn is normal, the resulting sample mean has the t-distribution.
- Useful when sample size N is small. Quite robust, even when the distribution is not normal.
- Looks like normal distribution but has heavier tails - Z test on small samples risk rejecting null hypothesis incorrectly.
- Run the test just like the Z test. Use $t = \frac{\bar{x}-\mu}{s/\sqrt{N}}$.
- Instead of comparing a Z score to the normal distribution, compare the t score to the t distribution.
- Actually, there is a family of t distributions, and you need to compare it to the distribution with the appropriate number of degrees of freedom, $N-1$.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
  - Z test
  - t test
  - Two sample t test
  - Paired t test

Two Sample Independent t-Test

Comparing Algorithms

- Assume that we are comparing our new algorithm with an old one.
- We randomly select n datasets to run with the old algorithm and m separate datasets to run with the new algorithm.
- We can do a one-sided test with the null hypothesis that there is no difference in the average running time and the alternative that our new algorithm has better average running time.

- Assumes we have a sample $X_1, \ldots, X_n$ drawn from a normal distribution with mean μ_X and an independent sample $Y_1, \ldots, Y_m$ drawn from a normal distribution with mean μ_Y.
- Assume unknown but common variance for the two distributions.
- The null and alternative hypotheses are
  H_0: $\mu_1 - \mu_2 = d_0$
  H_1: $\mu_1 - \mu_2 \neq d_0$ or
  H_1: $\mu_1 - \mu_2 < d_0$ or
  H_1: $\mu_1 - \mu_2 > d_0$ depending on what we are interested in.

- The test statistic is $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{1/N_1 + 1/N_2}}$ where $s_p^2 = \frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}$ is the estimated pooled variance, computed as a weighted average of the two sample variances s_1^2 and s_2^2.
- The number of degrees of freedom is $N_1 + N_2 - 2$.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
  - Z test
  - t test
  - Two sample t test
  - Paired t test
- Error Bars

Paired t-Test

Comparing Algorithms

- Consider the previous example.
- This time we randomly select n datasets, run both algorithms and measure the difference in the running time for each dataset.
- Both algorithms run on same dataset - should be correlated.
- Better design.

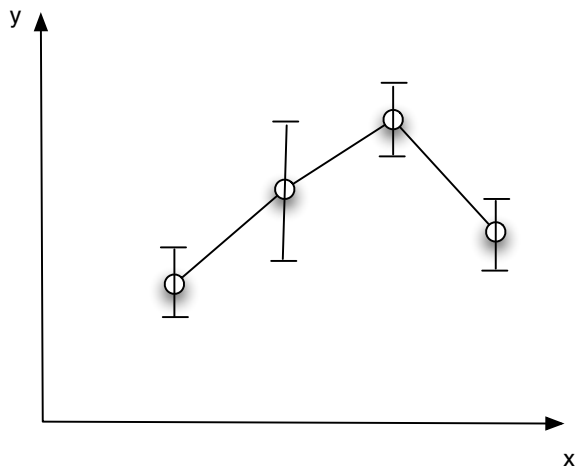
- In many experiments, the samples are paired, e.g. by age or weight.
- One member of the pair is randomly assigned to the treatment group while the other to the control group.
- Each observation is the difference of each pair.
- If the pairs are positively correlated (tend to be larger or smaller than their respective means at the same time), the variance of the estimate is smaller - more powerful test.

- If the differences are normally distributed with unknown variance, paired t-test is appropriate.
- The null and alternative hypothesis:
  H_0: $\mu_d = d_0$
  H_1: $\mu_d < d_0$ or
  H_1: $\mu_d > d_0$ or
  H_1: $\mu_d \neq d_0$ depending on what we are interested in, where μ_d is the difference in the means.
- Use $t = \frac{\bar{x}_d - d_0}{s_d/\sqrt{N_d}}$ where $\bar{x}_d$ is the sample average of the paired differences, s_d is the sample standard deviation of the paired differences and N_d is the number of pairs.
- Number of degrees of freedom is $N_d - 1$.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Error Bars



- Confidence intervals can be used to make plots more informative as shown above.
- If the variance is known, we can use the $\bar{x} = \mu \pm 1.96\sigma/\sqrt{N}$ to construct 95% confidence intervals for error bars.
- If the variance is not known, use $\bar{x} = \mu \pm t_{.025}s/\sqrt{N}$, where $t_{.025}$ depends on the degree of freedom, to construct 95% confidence intervals.

Outline

- Science
- Inductive Principle
- Falsification
- Scientific Method
- Design Principles
- Validity Traps
- Example: MYCIN
- Constructing hypothesis
- Hypothesis testing
- Confidence Interval
- Gaussian Distribution and Central Limit Theorem
- Tests on Means
- Error Bars
- Occam's Razor

Bonferroni Method

- We may sometimes need to do many hypothesis tests.
- For example, if we have multiple samples instead of just two, we may want to know if the mean of one of the samples is higher than the rest - can do t-tests for each pair of samples to know which means are different.
- However, we cannot maintain the same level of statistical significance if we do multiple tests.

Stockbroker's Predictions

You receive an email from a stockbroker claiming that he can predict whether the stock market will rise or fall each day. For the next ten days, he emails you his prediction at the beginning of the day, and you check his prediction at the end of the day. After 10 days, you find that he was correct for each of the 10 days. Should you become his customers?

- The Bonferroni method is a simple way to deal with this: if we do n tests, set the significance level of each test to α/n to obtain a significance level of α for the entire process.

Occam's Razor

- You have some observations and you want to formulate a theory or explanation for the observations.
- There are many potential explanation for the observations
- What explanation or theory should you choose?
- A commonly used scientific principle is to choose the simplest explanation This principle is usually known as Occams razor, named after William of Occam, who said

Entities shall not be multiplied unnecessarily.

- There are many ways to formulate complex explanations and fewer ways to formulate simple explanations.
- If we try many times to fit the data, we are likely to fit it just by chance.
- Using simple explanation forces us to fit the data using fewer possible ways - if we succeed it is less likely that it is by chance and more likely to be reproducible.
- By analogy to the Bonferroni method, if we test fewer hypothesis, we can be more confident of the outcome of the test.
- In searching for useful hypotheses, try to keep it simple Otherwise you will be wasting a lot of time investigating hypotheses that are unlikely to be useful.

References

- Andy Field and Graham Hole, *How to Design and Report Experiments*. SAGE Publications, 2003.
- E. Bright Wilson, Jr., *An Introduction to Scientific Research*. Dover Publications, 1990.
- Paul R. Cohen, *Empirical Methods in Artificial Intelligence*. MIT Press, 1995.
- John A. Rice, *Mathematical Statistics and Data Analysis*. Duxbury press, 1995.
- Research Methods Knowledge Base, Web Center for Social Research Methods, `http://www.socialresearchmethods.net/kb/`