Science

Scientists aim to understand the world: to unravel the secrets of nature. Ultimately, science is based on observations. Mathematical deduction is used to form theories and hypotheses but the hypotheses must be validated using experiments.

A scientist rejects authority as the basis for truth and reserves the right to judge for themselves whether the methods used are appropriate and whether the facts are credible. Scientists should be able to replicate works of others to confirm results for themselves. A scientific publication is required to contain enough details to allow others to replicate the experiments. When there is substantial experimental evidence and agreement in the judgements of scientists, a theory may be considered to as established; even then, a theory is never considered to be absolutely true as there is always a possibility that it may need to be refined to take into account new observations.

1 Inductive Principle

Deductive reasoning is commonly used in mathematics. Given a collection of premises and a conclusion, an argument is an assertion that the conjunction of premises implies the conclusion. An argument is *valid* if the assertion is always true. The example below is an example of a valid argument called a syllogism, descended from the time of Aristotle:

Premise 1: All men are mortal.

Premise 2: Socrates is a man.

Conclusion: Therefore, Socrates is mortal.

Another commonly used valid argument is to use the contrapositive statement.

Premise 1: p implies q

Premise 2: q is false

Conclusion: p is false.

This argument pattern is commonly used in science, as we will describe later.

The following is a pattern of reasoning commonly used by people. However, it not a valid argument.

Premise 1: p implies q

Premise 2: q is true

Conclusion: p is true

We can confirm that it is not a valid argument from the truth table as $((p \rightarrow q) \wedge q) \rightarrow p$ is not a tautology.

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

1.1 Ptolemiac vs Copernican System

We will consider a historical example to illustrate these patterns of reasoning.

From antiquity, people have thought that the earth was the center of the universe with the sun, planets and stars revolving around it. This model explains why the celestial objects move against the background stars. However, there are some unusual movements that are not well explained. For example, planets generally move from west to east, but sometimes change direction for a while. This is called the retrograde motion of the planets. Ptolemy's model (from 150A.D.) is constructed to explain these motions. In Ptolemy's model, epicycles, where planets rotate around a small axis while rotating around a larger axis around the earth, are used to explain retrograde motion (see Figure 1[1]).

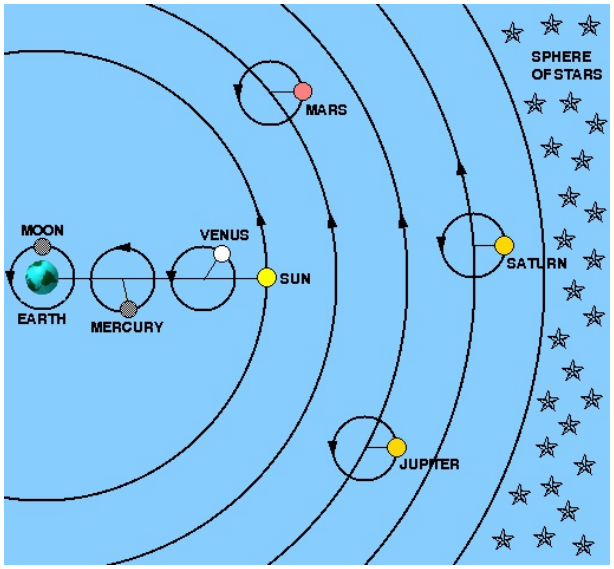

Figure 1: Epicycles in Ptolemy's model.

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

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

Then came Galileo (1564-1642), who invented the telescope, making observations that distinguish the models possible. The Copernican and Ptolemy theories make different predictions about the phases of the moon. Copernicus' theory implies that all phases of the Venus will be observed (see Figure 2[2]). Ptolemy's theory implies that not all phases can be observed.

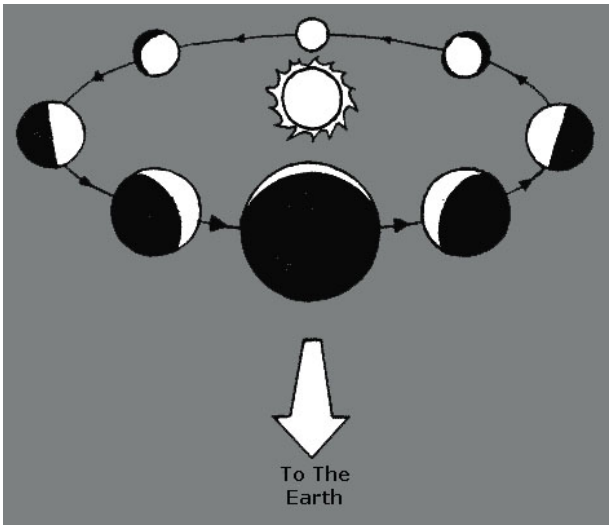

Figure 2: All phases of Venus can be observed in the Copernican model.

With the telescope, it is possible to prove that Ptolemy is wrong. This follows from the following valid argument.

Premise 1: Ptolemy correct implies not all phases can be observed

Premise 2: All phases are observed

Conclusion: Therefore, Ptolemy is wrong

However, it is not possible to show that Copernicus is correct, as the following in **not** a valid argument.

Premise 1: Copernicus correct implies all phases are observed

Premise 2: All phases are observed

Conclusion: Therefore, Copernicus is correct

[2]Image from `http://www.souledout.org/nightsky/venusphases/venusphases.html`

1.2 Induction

Despite not being able to logically prove that Copernicus is correct, we tend to feel more confident in his theory following the observation that it made a correct prediction. This is called inductive reasoning.

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

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

Bayes' rule 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. Copernicus has higher probability of being correct after the observation of all phases of Venus!

On the other hand, since $P(O|T) = 1$, we have $P(\neg O|T) = 0$. Bayes' rule 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.

So, overall Bayes' rule shows that repeated testing with different correct predictions of the theory increases our confidence in the theory.

1.3 Truth and Reproducibility

We have argued that it is not possible to show that a theory is true (or even false) with certainty. However, for most practical purposes, truth is not necessary. If we can show that the outcome is reproducible with high probability, this is often sufficient. For example, we still use Newton's theory of gravitation instead of Einstein's general relativity for most purposes because we know that it is reliable under most conditions. Only under conditions when it is necessary do we use Einstein's theory.

1.4 Popper's Falsifiability Principle

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. Under Popper's philosophy, 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.

2 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.

How do you know what hypothesis to test? What should you 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 before you can consider your first hypothesis.

2.1 Observational and Manipulation Experiment

In a manipulation experiment, the scientist directly manipulates the variables and measure the changes in the outcome. A properly designed manipulation experiment can give causal link between variables i.e. it tells you if you can control the outcome by changing a particular variable. Manipulation is necessary if a causal link is needed as *correlation does not imply causation*.

In observational research, the scientist observes what goes on without manipulating any of the variables. Often it is not possible to do manipulation research for various reasons, e.g. ethical reasons, or the outcome may be biased by the setup necessary to do the manipulation.

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 the following 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!

3 Design Principles

3.1 Use of 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. For example, 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.

3.2 Use of 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 can often also be used as the basis for calculating the probability of error (more on hypothesis testing later).

4 Validity Threats

4.1 Order and Location Effects

The order in which experiments are done or the physical location of the experimental subject often has an effect that has to be considered.

Strength of plastic: 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, and then 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 thought that the length of time was the factor determining the strength. However, someone pointed out that the mold gets warmer as the experiment proceeded, so the strength may be caused by that, invalidating the results of the experiment.

Once, the confounding variable has been identified, it may be possible to control it. However, it is difficult to think of all possible confounding variables. One way to potentially remove some unknown confounder is to randomize. Order and location often have unknown confounders associated with them, so randomizing them is often a good idea. In the example above, the effect of the temperature bias would have been reduced by randomizing the order of pressing the pieces so that some of the pieces that were pressed longer were also pressed earlier.

4.2 Change of Equipment

Always make sure that the equipment used (e.g. the computer) is the same when comparing two experiments. Otherwise, the different outcomes of the experiment may be caused by the difference in equipment and not the difference in the control and treatment groups.

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 the result valid?

In the example, the comparison is not valid. You should at least use the same machines (and programming language and compiler) to do the experiments. Even then, the implementation of an

algorithm often affects its efficiency, so care is required in interpreting the results of such experiments.

4.3 Effect of Measuring

In medicine, 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. Subjects often get better simply because they believe they are being treated, even though they are actually given placebo (some harmless alternative to the medication). The usual way to control for the placebo effect is to give the control group a placebo (something that has no effect but looks indistinguishable from the treatment to the subject). Similar care needs to be taken with experiments involving human in the computing areas, e.g. in human-computer interaction, to avoid problems with effects similar to the placebo effect.

The effects of measurements is also often seen in surveys. Asking questions related to prestige or embarrassment in a face-to-face survey is likely to result in biased answers. Similarly, leading questions, such as "NUS is a very well respected university internationally. What ranking would you give NUS among universities worldwide?" is likely to produce biased outcome. Care needs to be taken in designing the questions. Having several different questions measure the same effect and perhaps administered to different groups is often useful in detecting problems.

4.4 Effect of 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, the children were assigned at random, but the teachers were 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 making the results questionable. It was thought that the teachers could (perhaps unconsciously) have adjusted initial randomization because of sympathy.

The method to overcome the effect of the experimenter is to take the decision away from them by using randomization. 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.

4.5 Distribution Bias

In manipulation studies, we are usually interested in discovering causes. A causal effect is usually less sensitive to the distribution of the objects under study.

However, in observation studies, the effect we are studying may not be causal and may be seriously effected by the distribution. For example, in doing a survey, we need to ensure that the sample we select is representative of the population we are interested in. If we do an internet survey, the results would probably not apply to young children or the aged. If we perform a survey at an MRT station during peak hour, we are likely to get mostly working adults.

5 Example: MYCIN

MYCIN, developed at Stanford in the 1970s, is one of the best known successes of AI. It's role is to recommend therapy for blood and meningitis infection. MYCIN was shown to do as well as human experts in it's recommendation. We describe how that was done.

A randomized blind experiment was done to control for human bias: the judges did not know whether a human or MYCIN was making a particular recommendation. Eight humans were asked to each solve 10 problems. MYCIN solved another 10 problems and 10 more were taken from the original attending physician for 100 recommendation in total.

Of the 8 humans *only* 5 are experts. One is a research fellow, one is a resident doctor and one is a medical student. The presence of non-experts allows for control of the *ceiling/floor* effect: the problem being too easy 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.
- Manipulation experiments can also be done with MYCIN, by subtracting rules from database and observing the effect on the prediction quality.

6 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.

The alternative hypothesis can be *two-sided* as seen above or *one-sided*, e.g. the average recovery rate is better with the new drug, depending on the aim of the study.

Note that hypothesis testing only tells you whether the effect is likely to be due to chance, not whether the effect is important. Domain knowledge is needed to judge importance.

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? Note that the clever use of randomization means that the probability distribution is known. In this case all possible outcomes are equally likely. The use of randomness has also 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 $1/70 = 0.014$. Statisticians, such as 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.

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

In 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*; in this case, the critical region is the set of outcomes where the lady identified all cups correctly. If he sets the critical region to be the outcomes where at most one of the four cups prepared the first way being wrongly identified, then the level of significance is 17/70 as there are 16 ways to wrongly identify one of the four cups prepared the first way.

In 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!

Is not possible to compute the power of a test unless we specify a specific probability distribution as the alternative hypothesis. Assume that, in the alternative hypothesis, the probability of correctly identifying all cups is 0.6 and the probability of wrongly identifying 1, 2, 3 and 4 cups prepared the first way are 0.2, 0.1, 0.05 and 0.05 respectively. The the test that requires all cups to be correctly identified would have power of 0.6 while the test that allows at most one cup poured the first way to be wrongly identified will have power of 0.8.

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.

7 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 (see Figure 3).

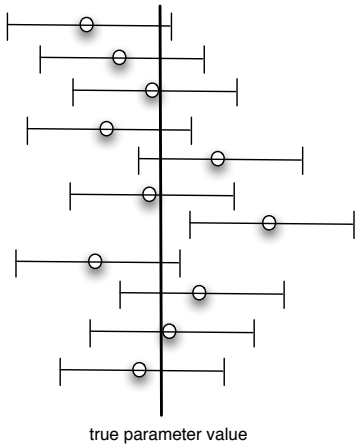

Figure 3: Illustration of confidence interval. In this figure, the true parameter is in all the confidence intervals.

8 Hypothesis Testing on Means and Confidence Intervals

We will only look at tests on the means (average values). This is probably the most commonly done type of test, e.g. we may want to know whether the average performance of one algorithm is better than the average performance of another algorithm. Other types of tests are possible; you should consult a statistics text when you need to perform other tests.

We first look at the Gaussian distribution, which allows good approximation for hypothesis testing and confidence intervals when the sample size is sufficiently large.

Gaussian Distribution: This is also called normal distribution or bell curve. It is the 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. Gausian distribution models many natural measurements

well - one reason is the *Central Limit Theorem* which we will describe next. Examples of some Gaussian distributions are shown in Figure 4[3].

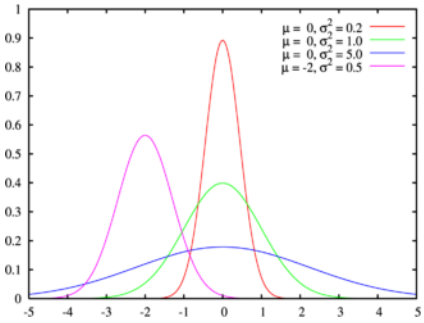

Figure 4: Gaussian distributions with different means and variances..

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!

8.1 The 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 and the standard deviation is σ. You now measure the performance of the new algorithm over a sample of size n. Assume that the population mean of the new algorithm is μ_d. We want to know if the mean performance of the new algorithm is different from the mean performance of the old algorithm.

In the example, we may form the following null and alternative hypothesis

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

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 not too small.

[3]Image from `http://en.wikipedia.org/wiki/Image:Normal_distribution_pdf.png`

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 $Z > 1.96$ is no more that 0.05. To do the test we compute the measured $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ and reject the null hypothesis when the value of $|z|$ is more than 1.96: we will be wrong at most 5% of the time, giving a level of significance (p value) of 0.05.

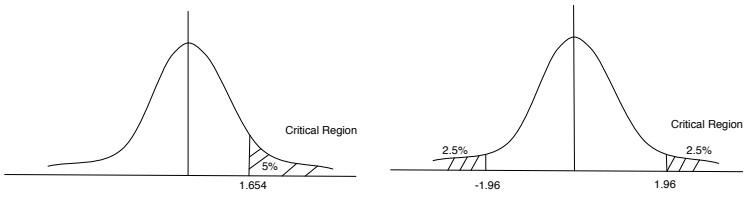

Figure 5: Critical region for one and two tailed tests.

The example shows a *two-tailed test* where the rejection or critical region is on both sides, corresponding to our lack of knowledge on whether the sample mean is larger or smaller than the hypothesized mean. If we do know if it is larger (or smaller), we may want to do a *one-tailed test* where the alternative is H_1: $\mu_d > \mu$. In this case we form rejection regions on only one sides of the mean (see Fig. 5). For a standard normal distribution, we can reject the null hypothesis if $|z| > 1.645$ 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 $n \geq 30$ this approximation is usually reasonable.

However, when n is small (e.g. $n < 30$), we may need to be more careful so that the results are still valid. When the sample distribution is known to be Gaussian (or nearly Gaussian) but the variance is unknown, the t test, to be described next, is usually used. When the distribution is unknown or known to be badly non-Gaussian, and $n < 30$ (so that the Z-test approximation is poor), we can often used a non-parametric test (not described here, please consult a statistics textbook).

8.2 The t-test

When the distribution from which the sample is drawn is normal, we can derive the distribution of the sample mean: the sample mean has the t-distribution. The t-distribution looks like the normal distribution but has heavier tails. By using a Z test (using standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}}$ instead of the true std dev σ) on small samples (e.g. $n < 30$), we risk rejecting null hypothesis incorrectly as the probability that the sample mean deviates significantly from the true mean is

higher in the t-distribution. The t-test is quite robust, even when the distribution is not exactly normal, hence it is commonly used when sample size n is small.

The t-test is run just like the Z test but we use the statistic $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ instead. Instead of comparing a z score to the normal distribution, we compare the t score to the t distribution. For example, we would use $|t| > t_{.025}$, instead of $|z| > 1.96$ as the critical region to reject the null hypothesis at $p = 0.05$ for a two-sided test. 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*, $n - 1$. The critical regions for t-tests are available from lookup tables or computer programs.

8.3 Confidence Intervals

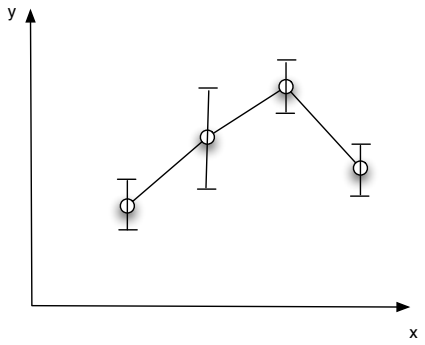

Figure 6: Error bars using confidence intervals.

Confidence intervals can be used to make plots more informative as shown in Figure 6. 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.

8.4 Two Sample Tests

Comparing Algorithms: Assume that we are comparing our new algorithm with an old one. Instead of assuming that the mean and standard deviation of the old algorithm as known, we randomly select n_1 datasets to run with the old algorithm and n_2 separate datasets to run with the new algorithm.

Two Sample Z-test: In this case, we can do a two sample test. Assume that the first distribution has mean μ_1 and standard deviation σ_1 and that the second distribution has mean μ_2 and standard deviation σ_2. Then it is known that the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

has a standard normal distribution.

Generally, the null hypothesis can be written as

$$H_0 : \mu_1 - \mu_2 = d_0.$$

You can set d_0 to zero to test if the two means are the same. The alternative can be two-sided or one sided.
H_1: $\mu_1 - \mu_2 \neq d_0$ or
H_1: $\mu_1 - \mu_2 < d_0$ or
H_1: $\mu_1 - \mu_2 > d_0$

The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

and the same table or computer program as before can be used to find the critical region for a particular level of significance.

Two Sample t-test: When the variances of the two distributions are unknown the two sample t test can be used. The two sample t-test assumes that 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. It further assume unknown but common variance for the two distributions (the case of different variances is slightly more complicated). The null and alternative hypotheses are the same as in the two sample Z-test.

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 $n_1 + n_2 - 2$.

8.5 Paired Tests

Comparing Algorithms: Consider the previous example. This time we randomly select n datasets, run both algorithms and measure the difference in the performance for each dataset.

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. This results in a more powerful test.

If the differences are normally distributed with unknown variance, paired t-test is appropriate. The null and alternative hypotheses are:
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.

To run the paired t-test, 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. The number of degrees of freedom is $n_d - 1$. If the sample size is sufficiently large, the paired Z test can be used in a similar manner.

9 Occam's Razor

9.1 Bonferroni Method

Stockbroker's Prediction: 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. Your hypothesis test indicates that it is unlikely that his correct predictions are due to chance. What went wrong?

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.

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.

9.2 Induction Revisited

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. Finally, bear in mind that simplicity can be relative -

simple explanations still has to agree with all known facts; the explanation may look complicated because the known facts are complicated but is actually simple because it differs only a little from the known facts.

10 Exercise

1. Einsteins theory of relativity predicted three consequences. For each of the following, answer true or false. Explain

   (a) If all three were to be confirmed the probability that the theory is correct would be very high

   (b) If two consequences were to be confirmed and the third found to be false, the probability that the theory is correct would be increased only a little

   (c) If any one of the consequences were found to be false the theory must be false.

2. Your friend Ahmad handed you a pack of cards and told you, I can read minds, and furthermore I can show by probabilistic reasoning that it is almost certainly true that I can read minds. He then asked you to shuffle the cards in any way you like. Take the top card, look at it but dont show me what it is. I will read your mind and tell you what the card is, Ahmad said. You repeated the process 10 times with him and astonishingly, Ahmad was correct in all 10 trials. Ahmad then said, Lets say that your prior belief $P(M)$ that I can read minds is very low at 10^{-10}.

   Let D be the proposition that I am correct in all 10 trials. Then $P(D|M) = 1$ since being able to read minds implies that I will be correct.

   If I cannot read mind, then my guess for each trial is no better than random and $P(D|\neg M)$ equals $(1/52)^{10} = 6.92 \times 10^{-18}$ since there are 52 cards in the pack. Also $P(\neg M) = 1 - P(M)$ is approximately 1.

   Now $P(D) = P(D|M)P(M) + P(D|\neg M)P(\neg M)$ approximately equals 10^{-10}. By Bayes rule, $P(M|D) = P(D|M)P(M)/P(D)$ is approximately 1. So you have to agree that it is almost certain that I can read minds. Should you believe Ahmad?

3. Ann Landers once asked readers of her advice column. If you had it to do over again, would you have children? She received nearly 10,000 responses, almost 70% saying NO! Can it be true that 70% of parents regret having children?

4. Your company buy several crates of oranges each week. To make sure that the oranges are good, the company examine a few oranges from the top of each crate. Criticize the procedure.

5. In October, 1976, a nationwide vaccination program was started against swine flu in the US. The first shots were given to the group most at risk the elderly. During the first week of the

program, 24,000 persons aged 65 and over were given shots, and three of these persons died. As a result, eight states suspended the vaccination program. Were the actions of the eight states appropriate?

6. If we look at median professors salaries and sales of alcoholic beverages each year, we find a strong positive association. Can we conclude that teachers are spending their extra pay on booze, thus driving sales up?

7. A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds) and the median number of days that patients remain in hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital?

8. People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

9. Many studies have found an association between cigarette smoking and heart disease. Recently, the Framingham study found an association between coffee drinking and heart disease. Should you conclude that coffee drinking causes heart disease? Or can you explain the association between coffee drinking and heart disease in some other way?

10. A 2-year study in North Carolina found that 75% of all industrial accidents in the state happened to workers who had skipped breakfast. Comment.

11. A 15-year study of more than 45,000 Swedish soldiers revealed that heavy users of marijuana were six times more likely than nonusers to develop schizophrenia. Comment.

12. Which of the following questions does a test of significance help answer? Explain

    (a) Is the sample or experiment properly designed?

    (b) Is the observed effect due to chance?

    (c) Is the observed effect important?

13. From the article So, now we know what Americans do in bed. So? in The New York Times, 9 Oct 1994, by Tamar Lewin, we find the following: ... Unlike previous sex studies, whose subjects were self selected, this study was carried out by in-person interviews with a random sample of 3,432 men and women aged 18 to 59. The results will be published in a book "Sex in American" (Little, Brown and Co. with Gina Kolata as a co-author with the investigators.) The survey showed a marked contrast between the sex that most people have compared to T.V. and movie images of sexual behavior. American women typically have two sexual partners during their life and men about six. 78% of the men and 86% of women say they have been faithful to their spouses while married.

    How is it possible that men have a lot more sexual partners than women?

14. In 1977, the Viking spacecraft orbiting Mars took a photograph of a human looking stone face about a mile wide. You have a theory that the face was constructed by extra terrestrial beings that visited the solar system millions of years ago. You hypothesize that these extra terrestrial beings must have left similar constructions all over the solar system. You send a space craft to the moon to look for a formation that looks like a human face from above. After an exhaustive search, you find a formation that looks similar to a human face.

True or false: your theory has been strongly supported by experiment. Explain.

15. We are interested to know whether consumption of alcohol reduces motion sickness. To test the hypothesis, we recruited 100 first year SoC students for a field trip to Tioman Island. We bought lots of alcohol and wait for a storm during the monsoon season to do the experiment. Describe how you would do the experiment. What would you measure and what statistical test would you do?

16. The lecturer is interested to know whether doing CS2309 is correlated with success in UROP. He looked into the records to get the students who did UROP in the last three years and found their UROP marks as well as whether they did CS2309. What type of statistical test should he do? What conclusion can he draw?

17. You developed a new optimizing compiler. You wish to show that it outperforms the current best compiler. You randomly select 30 commonly used programs, compile them with both compilers and measure the running of the resulting programs.

    (a) What statistical test should you use?

    (b) Your new compiler has 50 different parameters that can be set. Your test with the default configuration did not show that your new compiler is better than the old one. You are not ready to give up, so you exhaustively searched through the parameter settings and found a best configuration for the set of 30 programs and use that as the new default configuration. You now claim that your new compiler is better based on statistical significance test on the same 30 programs. Is your claim valid? Why?

References

[1] P.R. Cohen. *Empirical methods for artificial intelligence*. MIT press, 1995.

[2] A. Field and G. Hole. *How to design and report experiments*. Sage Thousand Oaks, CA:, 2003.

[3] E.B. Wilson. *An introduction to scientific research*. Dover Pubns, 1991.