Applied Data Analysis

We have seen how simple hypothesis testing works in principle but there are many more issues to discuss for real experiments.

How do you choose a? — in fact probability that we make Type I error — that we think we have seen an effect when there isn't one.

That is, if we use $a = 0.05$, and do lots of independent experiments, then if all the null hypotheses are true, on 1 experiment in 20 we get $P < a$ and decide that there is an effect that is not there.

So if we don't mind making this kind of error one time in 20, we choose $a = 0.05$; if we want a stricter criterion, choose $a = 0.01$.

This approach means that we never have to calculate the exact value of P for a given experiment — we just look up value of the statistic we're using that would give $P = a$.

Then if, when we do the experiment, the statistic is more extreme than this critical value, we reject H_0; otherwise accept H_0.

Errors and Distributions

Picture regions of different probability by drawing a graph of the distribution of a statistic — suppose statistic has continuous values.

For commonly used statistics, the distribution will have a hump in the middle for the likely values and tail off for extreme values.

For one-tailed test, split area under curve into two parts: part under one tail occupying 5% of the area and part under the rest occupying 95% of area — 5% region represents rejection region for $a = 0.05$.

For two-tailed test, 5% has to be split across two tails of distribution.

Knowing probability of a Type I error is useful, but probability of a Type II error (not seeing an effect that is really there) is useful too

Intuitively, more data means lower chance of a Type II error (for a given a), and indeed this is the case for any reasonable test.

If the null hypothesis is accepted because P is large, there is nothing simple to say about what the chance of a Type II error is.

There might be real effect which is not shown up, either because there is not enough data, or because the statistic used is not a good one for detecting the particular kind of difference that has occurred.

Full Example

There are significance tests to cover many different situations — need to spend time analysing the nature of the measurements and the experimental design to find the correct one — textbooks.

Another test is widely applicable — used when we want to know if two independent sets of data obtained under different conditions differ significantly.

For each condition some outcome is observed in a number of trials; the outcome must be measured with a number — we want to know whether the outcome is significantly different in the two conditions.

Since we have number of trials in each condition, we have indication of spread of likely values of outcome measure.

It seems reasonable to suppose that information about whether the conditions differ significantly is available without making further assumptions.

One test that will handle this situation is *Kolmogorov-Smirnov* (K-S) two sample test — statistic that this uses is the maximum difference in the cumulative distributions of the two outcome measures.

This is easiest to explain with an example.

Data Sets

Suppose we conduct a series of trials — say 10 — in one condition — say using one kind of crossover operator in a genetic algorithm.

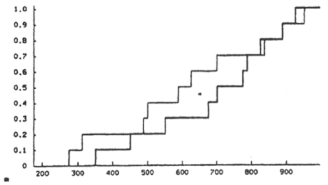
In each trial we measure number of generations to reach a particular state of the population, and get the following results:

630 890 700 270 500 480 320 950 836 585

We then do the same thing in an independent set of trials (no pairing with the first set) using a different operator — suppose this gives:

784 456 893 555 678 699 350 821 921 772

Is there a significant difference between these two sets of numbers? — find statistic for K-S test by making a cumulative frequency graph for the data sets by counting number less than any given value:



We can see that the graphs show how many values are less than any given value we observe in each condition.

Testing Data Sets

The statistic needed for K-S test is largest vertical difference between the two graphs, which we call K.

We can see by inspecting them that largest such difference is $K = 0.3$, near the asterisk, between values of the outcome from 630 to 678.

In practice, we would calculate this statistic by ordering the two data sets independently, then comparing the ordering between them and finding the point in the sequence with biggest difference:

270	
320	
	350
	456
480	
500	
	555
585	
630	
	678
	699
700	
	772
	784
	821
836	
890	
	893
	921
950	

The statistic $K = 0.3$ can be looked up in tables for the test — turns out not significant for 10 trials in each dataset even at $a = 0.05$.

Parametric and Non-parametric Statistics

The distribution of statistic K under H_0, ie both sets of data come from same underlying distribution, is known independently of what the distribution of the data actually is.

There is no assumption that data come from Gaussian distribution, or indeed any other distribution — test with this property is known as a *non-parametric* test.

Tests that make assumptions about distributions are known as *parametric* tests — typically they assume distribution is Gaussian.

A parametric test that could be applied to these data, if we make the necessary assumption, is the *unrelated t-test*, which uses statistical difference in means of the two data sets, normalised by estimate of standard deviation of the data.

In general, parametric tests are more powerful and involve simpler calculations but if assumption of Gaussian distribution of the data is incorrect then they can give misleading results.

There are numerous other significance tests that can be useful — important one is *chi-square test*, which is useful when some data need to be compared with expected frequencies.

Combining Significance Levels

Sometimes hypothesis is tested in two experiments which yield independent P values.

The best way to combine results is to find a way of treating two experiments as one, and finding an overall statistic that can be used in a test of significance.

When this is not possible, it can be useful to know how to combine more than one significance level in a sensible way.

In particular, correct way to combine is not to take their product, or their maximum or minimum — though all sometimes suggested.

If two significance levels are P1 and P2, the significance level of two experiments taken together is in fact:

$$P = P_1 P_2 (1 - \log(P_1 P_2))$$

where the logarithm is to base e (a natural logarithm).

The generalisation of this formula to N experiments is:

$P = g \sum_{r=0}^{N-1} \frac{(-\log g)^r}{r!}$ where g is the the product of the N separate significance levels, and r! is the factorial function of r.

Problems with Hypothesis Testing

Hypothesis testing is a respectable and sometimes valuable way to assess results of experiments — however, it has difficulties.

An important problem is that if the methodology were taken literally, hypotheses about, say, effectiveness of a new drug would be accepted or rejected when it was known that there was a definite probability that an error was being made.

This problem is exacerbated by the asymmetry in treatment of null and alternative hypotheses, which means that probabilities of Type I errors are accurately controlled but probabilities of Type II errors have to be largely guessed.

In practice, approach is not followed literally and common sense prevails — rather than setting a in advance and acting accordingly, most researchers tend to treat the P value obtained for their data as a kind of standardised descriptive statistic.

They report these P values, then let others draw their conclusions; such conclusions will often be that further experiments are needed.

The problem then is that there is no standard approach to arriving at a final conclusion and everything remains tentative — perhaps this is how it should be.

It means that statistical tests are a component in a slightly ill-defined mechanism for accumulating evidence, rather than tidy cut-and-dried way that their inventors were trying to establish.