

Introduction to Statistics for Computer Science Projects

Peter Coxhead

Introduction

Whole modules are devoted to statistics and related topics in many degree programmes, so in this short session all I aim to do is to introduce some key ideas and concepts. You will need to follow up on any that are relevant to your project.

Variability

Statistics is primarily a collection of techniques to deal with variability. Variability arises in two main ways: measurement errors and intrinsic variation.

- Suppose we want to know the value of some fixed quantity which can be determined precisely. An example might be the number of lines of code in a particular program. Provided that we can count every line exactly, statistical analysis is irrelevant.
- More commonly, there will be measurement errors. These are more usual in measuring physical quantities. For example, we might want to measure the speed of transmission of given set of data under fixed conditions where the rate will not vary. Although we can count the data size exactly, there will always be errors, however small, in measuring the time.
- Usually, there will be both measurement errors and intrinsic variability. For example, we may want to know the run time of an algorithm solving a given problem. The time cannot be measured exactly, but more importantly, the run time will vary because of varying loads on the processor (e.g. from the operating system) and for some languages (e.g. Java) because of the need for periodic garbage collecting if memory runs short.

A **variable** or **variate** is a quantity whose value varies for whatever reason. Statistical techniques are designed to allow us both to *describe this variation* and to *draw some conclusions in spite of it*.

Descriptive Statistics

Given a set of values for some variable, we want to be able to describe these to other people, in some more meaningful way than just listing the raw data.

For example, I converted a Java library for matrix manipulation into JavaScript, and was interested in the time behaviour of some of the functions. In one test, I generated 100 random matrices of size 70 x 60 and used the JavaScript Date object to time the calculation of the pseudo-inverse of each matrix in milliseconds.

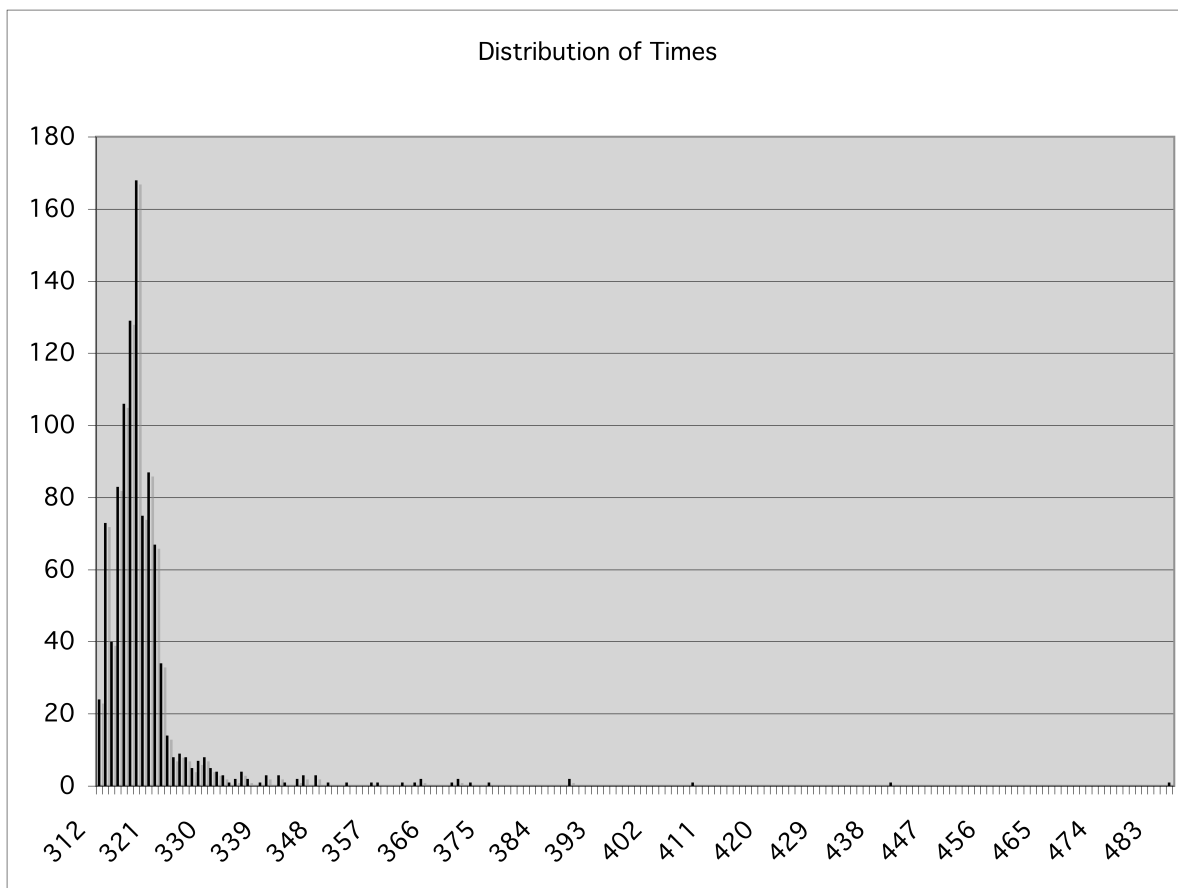
How should I describe this data?

I could just list the 100 values: 318, 314, 315, 315, 313, 314, 315, 314, 314, 315, 313, 313, 315, 313, 314, 315, 314, 314, 315, 316, 315, 315, 314, 314, 314, 314, 314, 315, 314, 314, 316, 315, 314, 314, 315, 315, 316, 315, 313, 314, 313, 314, 314, 313, 313, 313, 315, 313, 312, 312, 313, 316, 313, 315, 315, 315, 313, 313, 312, 314, 314, 313, 313, 315, 314, 314, 315, 314, 314, 315, 313, 313, 314, 312, 312, 316, 314, 315, 315, 315, 315, 315, 314, 314, 313, 314, 314, 315, 313, 315, 316, 314, 315, 314, 323, 314, 314, 315, 314, 310.

More useful is to look at the **distribution**. This shows the relative number of occurrences of each value. (Microsoft Excel can be used to do a lot of simple statistical processing; in this case I used a pivot table.)

Time in ms	No. of occurrences
310	1
312	5
313	20
314	36
315	30
316	6
318	1
323	1

The first point you should notice is that the distribution is not symmetrical. Although most of the values are clustered into the 310-318 range, there is an odd **outlier** at 323. To understand a distribution like this, we really need more data. I repeated the test with 1000 random matrices. A **histogram** of the resulting data is shown below.



This shape is classic for the run times of complex algorithms on random data. There is a non-zero lower bound (in this case 312 ms, although one run took 310 ms in the earlier test). The ‘worst case’ performance will be many times the value of the lower bound, but may not be found in random trials. Often small sample sizes will seriously underestimate the possible worst case performances.

Always ensure that you understand the distribution of the variables you are dealing with.

Summary Measures

The distribution contains the maximum information about a variable. However, we usually want to reduce it to some more easily grasped numbers.

‘**Measure of central tendency**’ is the statistician’s term for what is often loosely called the ‘average’. There are three commonly used measures, whose values will differ for a skewed distribution like the one shown above.

The **mode** is the most common value. Its value is 318 in the test of 1000 runs. Because it uses only one piece of information – the value which has the highest peak in the distribution – it tends to be relatively variable from one set of data to another.

The **median** is the value which has half the values at or below it and half at or above it. Here its value is the same as the mode (which is not always the case): 318 ms for the 1000-run test. It uses more information than the mode, but takes no account of the actual size of the values, only their ordering.

The **mean** is the ordinary arithmetic average: the sum of the values divided by the number of values. Here the value is 319.583 ms (to three decimal places) for the 1000-run test. The mean will be larger than the median when the distribution is **positively skewed** (with a tail to the high end) and smaller than the median when the distribution is **negatively skewed** (with a tail to the low end). It makes the most use of the information, but the impact of skew must be remembered. (The difference between median earnings and mean earnings offers a good example.)

The most common measure of ‘spread’ – the **deviation** from the centre or the **variance** about the centre – is based on the mean (although there are measures based on the median). The **variance** is defined as the average of all the squared differences between each value and the mean. The **standard deviation** is the square root of the variance. Squares are used to remove negative values (and because they make formal analysis much easier).

For the 1000-run test, the standard deviation is 10.487 (to three decimal places). The standard deviation is not easy to interpret (because of the squaring and square rooting).

- Very roughly, it’s the average difference between a value and the mean (but the squaring weights it more towards higher differences). So we expect that on average values will be about 10 away from the mean of 320 (rounding to whole numbers).
- For a particular kind of symmetrical distribution, called the normal distribution, about 95% of the values should lie in the range (mean ± 2 * the standard deviation). (Strictly the 2 should be 1.96, but the difference rarely matters.) Here (rounding to 1 decimal place), the range would be 319.6 ± 21.0 , i.e. 298.6 to 340.6. Since the data consists of integers, the actual range calculated in this way is 299 to 340. It’s clear that this range is misleading because of the skew, since the lowest value observed was 312. In fact, 97% of the values lie in this range: the small ones that would be outside the range if the distribution was actually normal are missing.

Standard Error and Significant Figures in Reports

Earlier, I gave the mean as 319.583 ms. You should have thought that this was a silly thing to do! Why? Because it’s spuriously accurate. If I repeated the test with another 1000 random matrices, I wouldn’t expect to get the same mean.

What would I expect? If the distribution of the values were normal (a particular shape of symmetrical distribution), then I can make an estimate of the *variability of the mean* between successive sets of data. Note that this is *not* the same as the *variability of the data values* about the mean.

The **standard error** of the mean is the expected standard deviation of the mean between repeats of the whole data sampling exercise. It’s calculated as the standard deviation of the data values divided by the square root of the sample size.

Here the standard error is $10.487/\sqrt{1000} = 0.3316$. What this means is that, only approximately because the distribution is skewed and hence clearly not normal, if we repeated the test with 1000 random matrices with all other conditions being the same, we would expect roughly 95% of the means to be in the range $319.583 \pm 2 * 0.3316 = 318.9$ to 320.2. This is called the **95% confidence range** for the mean.

What this makes clear is that it’s silly here to quote the mean more accurately than to 1 decimal place (dp), given that the approximate 95% confidence range is ± 0.6 .

Always take the standard error into account when quoting summary statistics. Quoting a silly number of decimal places/significant figures is a sure indicator of statistical illiteracy.

Exercise Suppose we had observed the same mean, 319.583 ms, and the same standard deviation, 10.487 ms, but with only a sample size of 100 random matrices. How should we report the mean?

Evidence shows that most people find numbers with more than 3 digits hard to grasp, so even if the standard error is small, because of a large sample size, think carefully about how to report your data. For example, it's almost never worth reporting percentages to more than whole numbers: is there any meaningful difference between 56% and 56.4%?

Significance Testing

Suppose I want to compare two algorithms for finding the pseudo-inverse of random matrices.

In the simplest approach, I could collect two sets of data, say 100 random matrices run with the first algorithm and then a different 100 random matrices run with the second algorithm.

Here are some results I obtained, with and without a minor 'tweak' to a standard algorithm.

	N	Mean	S.d.
Algorithm 1	100	312.5	3.517
Algorithm 2	100	311.2	4.628

Is Algorithm 2 'really' better than Algorithm 1? To answer this we use an apparently round-about logic.

- Assume that the algorithms would not produce different averages if each were applied to the entire population, i.e. all possible random matrices of the relevant size.
- Under this assumption, calculate the probability of a difference as large as or larger than that observed having occurred by chance. This calculation will require various other assumptions to be made about the nature of the data, its distribution, etc. The full set of assumptions forms the 'null hypothesis'.
- If the probability of the observations given the null hypothesis is lower than some threshold figure, decide that it's too improbable to have occurred by chance, so reject the null hypothesis. Such a result is said to be **statistically significant**. It's important not to confuse this specialized use of the word 'significant' with its 'normal' use: see below.

There are literally thousands of different **statistical tests**, i.e. ways of calculating the probability of a set of observations assuming some null hypothesis is true. Even for a simple set of data, such as that used here, there are many possible ways of trying to calculate the required probability.

If in doubt, always consult an expert as to which statistical test to use (but expect different answers from different experts!).

One possibility here is to apply a t-test. An argument against is that it assumes normal distributions, but we know that they are actually skewed. On the other hand, the t-test is known to be robust, i.e. it's unlikely to wrongly declare a result to be significant even if the distributions are somewhat skewed.

Microsoft Excel will perform the calculation of the probability (I chose the options: 2-tailed,¹ two samples, equal variance). The value for the data above is $p = 0.021$, i.e. 2.1%.

Thus if (and it's a crucial if) all the assumptions made in the null hypothesis are true, the chances of observing a difference as large as or larger than the one seen here (311.2 ms vs. 312.5 ms) are only 2.1%.

¹ Unless you are a statistical expert, 2-tailed is always the right choice.

Conventionally, if the probability is less than 1% (the ‘1% significance level’), we reject the null hypothesis. If the probability is less than 5%, we may reject the null hypothesis, but should treat our conclusion as somewhat tentative. So it looks as if Algorithm 2 is indeed better than Algorithm 1, but not as certainly as we would like.

An alternative statistical test for a difference in two distributions (not specifically a difference in the means) which can be applied in this case is the Mann-Whitney U. This makes no assumptions about the form of the distribution. Excel doesn’t do this test, but can be used to prepare the necessary data for it; alternatively there are some online web pages which will do it all for you. For the data discussed above, I obtained a value of $p = 1.2 \times 10^{-5}$ (2-tailed), i.e. according to this test, if the distributions of run times is in fact the same for each algorithm when applied to the entire population of random matrices, the probability of the observed data being as or more different is about 0.0012%. This is so unlikely that we can conclude that the algorithms are indeed different.

Independent Samples vs Repeated Measures

There’s another way I could try to determine whether the two algorithms are different when applied to all possible random matrices (of the relevant size). I could generate say 100 random matrices and apply each algorithm to each of the 100 matrices. There are different ways of describing the two approaches (i.e. using a new random matrix for each run of either algorithm or using the same random matrix twice, once for each algorithm). Commonly the first is called ‘independent samples’ and the second ‘repeated measures’.

In principle, wherever repeated measures designs can be used, they should, since they remove one source of random differences.

I rewrote my program to use the two algorithms in turn on the same set of random matrices. The results were as below.

	N	Mean	S.d.
Algorithm 1	100	313.5	5.567
Algorithm 2		311.8	3.340

If we *incorrectly* apply the independent samples t-test, we obtain $p = 0.014$, i.e. the result is significant at the 5% level but not at the 1% level. If we correctly apply the repeated measures t-test, we obtain $p = 5.6 \times 10^{-7}$, i.e. the result is highly implausible, well beyond the 1% significance level.

Since the normal distribution assumption is violated, we might try the repeated measures version of the Mann-Whitney test, usually called the Wilcoxon sign test. This yields $p = 1.2 \times 10^{-10}$. So, either way we can be apparently be pretty sure that there is a real difference between the two algorithms.

Notice that if you incorrectly apply an independent samples test to repeated measures data, almost always the result will be that the calculated significance level will be too high, i.e. less likely to be statistically significant. The opposite is more complicated, since to apply a repeated measures test to independent samples data, you first have to artificially pair up the values, and the result depends on how this is done.

Always use the correct test – independent samples or repeated measures – for the design you have used.

Statistical Significance vs Real Significance (Size of Effect)

It’s important to understand that the statistical significance of a difference depends on several factors:

- the size of the difference
- the sample size(s)
- the ‘power’ of the test applied.

With a large enough sample, even a very small difference will turn out to be statistically significant, i.e. not due to chance. But this does *not* mean that it's a meaningful or worthwhile difference. In my example above, it's not clear that reducing the run time from around 313 ms to around 311 ms, a reduction of less than 1%, makes any practical difference at all.

Always report both an appropriate measure of the size of the effect you found as well as its statistical significance. Don't use a low value for the probability under the null hypothesis as a measure of the meaningfulness of the effect. It isn't.

What next?

I deliberately haven't attempted to describe lots of different statistical tests. Seek advice and information elsewhere.

- In my experience Wikipedia is reliable on statistics, including statistical tests, although usually not written for beginners.
- Richard Lowry, of Vassar College, NY, has an extremely useful website:
<http://faculty.vassar.edu/lowry/webtext.html> is an online textbook
<http://faculty.vassar.edu/lowry/VassarStats.html> contains a large number of web pages implementing JavaScript programs to perform statistical calculations.
- Consult your project supervisor.
- E-mail me; although I'm typically only in the university about one day a week, I read my e-mail regularly.
- Dr Allan White runs a free University-wide statistical advisory service. Contact him via e-mail: A.P.White@bham.ac.uk.

Some of the things you need to be clear about in relation to your data when seeking advice are:

1. Its **level of measurement**, since this determines which statistical tests can be applied.
 - Most physical data, like a set of run times, is normally at the **ratio** level of measurement: values can be added, subtracted, multiplied or divided and still be meaningful. There is a proper zero.
 - Some physical data, like temperature in °C, can be added and subtracted, but not multiplied or divided, since there is not a proper zero. This is the **interval** level of measurement.
 - Much psychological data can at best be ordered in size, but cannot meaningfully be added or subtracted. This is the **ordinal** level of measurement.
 - Occasionally you might need to work with classifications into distinct, un-orderable categories. This is the **nominal** level of measurement.
2. Its **distribution**. Data which has a more-or-less symmetrical, roughly bell-shaped distribution around some central value can be analysed using a wide variety of statistical tests. Heavily skewed data should either be scaled in some way to 'fix up' its distribution, or else analysed using statistical tests which are not dependent on assumptions of normality.
3. The **design** of any experiments you carried out. Was each observation (each data value) subject to quite independent random effects (independent samples)? Or was some randomly affected data used more than once (repeated measures)? Or was there some mixture of the two?
4. Can you define clearly the null hypothesis or hypotheses involved in your data analysis? In other words, what is it that you are trying to show is *not* the case?