Statistical significance for practical experiments

The primary reason for running significance tests is so that we do not draw conclusions and take actions (or allow others to do so) based on evidence which is too weak. Sometimes apparent effects are simply due to chance: most importantly for us, they may be due to natural variation in the test data. Significance testing when deciding whether or not a new system is better than a baseline is all about reducing the possibility we are fooling ourselves because of a chance effect. The **null hypothesis** for that situation is that the new system has performance which is equivalent to the baseline: we are trying to see whether we have good evidence that the null hypothesis is false.

Researchers are normally optimistic that their new system will work. In many cases, we really want our new system to work. That might enable us to write a paper, get more funding, persuade our boss that we are a good person or whatever. The most essential thing to remember about significance testing is that we need to adopt a quite different perspective and be sceptical. We have to start by assuming that our new system does not actually work any better than the baseline (that the null hypothesis is true) and then try and see whether the results can persuade our sceptical selves otherwise. In fact, this is part of the more general process of doing research: to see whether we can defend our approach against reasonable doubt we have to (temporarily) put ourselves in the position of a sceptic and see whether the results we find have another explanation than the one we want to claim. Significance testing is just one aspect of this.

There are standard significance levels which are used in published work. This is just a convention: we use these particular levels because that is what everyone else does. Which level to use depends on the cost of the decisions we might make. For instance, if we have a new system which is very easy to deploy, we need less evidence to use it than if we're considering a complex change. Obviously, the extent of the improvement also matters: we might have done lots of trials and so be very confident that a new system is better, but decide that the degree of improvement is insufficient to justify switching. (Most significance tests are sensitive to the degree of improvement, in that it will be easier to show significance if the improvement is large, but not all are.)

For the moment, let's just consider significance at the 1% level. What we are claiming when we say a result is significant at that level is that there is less than 1% chance that we have fooled ourselves into thinking that the null hypothesis is false. Suppose we try a test on 200 items of test data and obtain a result with our new system which is significantly different at the 1% level from the baseline. What we are claiming is equivalent to saying that if we tested the baseline system 100 times on different sets of 200 items, we would only once get as good performance as our new system.

In fact, if we had that much test data, we could run such an experiment. In some cases, simulating an experiment explicitly is actually a good way of doing significance testing. But we usually use standard significance tests which rely on assumptions about the way our data behaves. The particular test we will examine in this course is the **sign test**.

Sign test

The sign test is appropriate when we have paired data, so it is a natural fit for a situation where we have tried a baseline system on some test data and want to compare a new system with the baseline. We also make the assumption that the individual tests on data items are not linked. In the sentiment systems, we consider each document in isolation, so this condition holds. Our null hypothesis is that each system is equally good. This means that we would expect, on average, that they would obtain the same overall

accuracy on a particular test set. We would not expect them to get exactly the same results on each piece of data, but the cases where the new system is better than the baseline would be approximately balanced by the cases where it was worse.

To see how the sign test works, let's first consider a case where the ground truth data is numerical, and the systems being tested both produce fine-grained numerical values. Under such circumstances, ties would be vanishingly rare. For instance, we might have systems that estimate the weight of elephants and give results in milligrams. We can essentially always say whether or not the new system is closer to ground truth than the baseline, even if it is only very slightly closer. Suppose do **n** trials: we count the cases where the new system is better, and call that number **plus**, and call the cases where it is worse **minus**. If the null hypothesis is true, the baseline and the new system are actually equally good, but just happen to give slightly different results on particular elephants. The probability of getting **plus** on a particular trial is equal to the probability of getting **minus**. Thus the observed counts of **plus** and **minus** obey a binomial distribution with a mean of 0.5n under the null hypothesis. The sign test simply checks how likely it is that the actual observations (or more extreme observations than the actual observations) could arise in that situation: i.e., it is a special use of a binomial distribution.

The binomial distribution gives the following probability for an observation to be exactly **k** given **n** trials and a probability of success of **q**:

$$\binom{n}{k} q^k (1-q)^{n-k}$$

We want to include the probabilities of the more extreme observations. Assuming for the moment we are only interested in the lower half of the distribution (a one-tailed test, see below), we are interested in the probability that a trial is less than or equal to **k**, where **k** is the minimum of **plus** and **minus**:

$$\sum_{i}^{k} \binom{n}{i} q^i (1-q)^{n-i}$$

For the sign test, q is always 0.5:

$$\sum_{i}^{k} \binom{n}{i} 0.5^n$$

However, we are actually interested in the probability that the observed value is more extreme in either direction (our system might actually be worse than baseline). Because the binomial is symmetric, we can simply double this value to obtain the two-tailed sign test.

$$2 \sum_{i}^{k} \binom{n}{i} 0.5^n$$

(A note: the elephant estimation example is purely for expository purposes. The sign test ignores the magnitude of the difference between the ground truth and the estimate. Two systems could be equivalent according to the sign test, even if one occasionally made huge errors in the approximation. So it's unlikely we'd actually use the sign test in this situation.)

Ties

In testing our sentiment systems, we naturally get a large number of ties when both systems are correct or both incorrect. In order to apply the sign test, we distribute these ties evenly between plus and minus. Since all our numbers must be integers, if we get an odd number of ties, we pretend we did one more test which came out as a tie. Thus, if we call the actual number of ties **Null**:

$n = 2\lceil \frac{Null}{2} \rceil + Plus + Minus$

$k = \lceil \frac{Null}{2} \rceil + \min\{Plus, Minus\}$

Note that we have a slight problem with the formula we gave for the sign test in the situation where the value of **k** is actually equal to the mean of the binomial and **n** is an even number, because we can end up counting part of the probability twice. For instance, if **n** is 4, the terms of the binomial are 1, 4, 6, 4, 1. In this case, if k is 2, we will include the probability mass for k twice in:

$$2\sum_{i}^{k}\binom{n}{i}0.5^n$$

This effect shows up if we compare a system with itself: we can end up with a "probability" slightly greater than 1. In practice, this is unimportant, because we would never apply the sign test when the observed value was equal to the predicted mean, since it is obvious that the null hypothesis is not disproved.

Using the normal approximation to the binomial

Calculating the values for the sign test can be slightly tricky, because we are dealing with very small and very large numbers for usual values of **n**. While it is possible to do this, it is easier to simply use the normal approximation to the binomial. That is, we approximate the binomial distribution with a normal distribution with the same mean, and with variance equal to 0.25n. For instance, with 100 trials, the mean is 50 and the variance is 25. Standard deviation is the square root of the variance: i.e., 5 in this example. Because the normal distribution is continuous, we have to add 0.5 to k and test that value for significance.

This test is equivalent to checking whether or not k is inside the relevant confidence interval. For instance, the 95% confidence interval is about 1.96 x standard deviation, which gives us a very convenient rule of thumb calculation that we can test for significance by checking whether the observed value is more than two standard deviations from the mean.

One-tailed or two-tailed?

The decision to use a one-tailed or two-tailed test can be the cause of considerable controversy. Our recommendation to use the two-tailed test is simply because this is harder to beat and hence less open to criticism (in the usual situation where you are testing your own system against a baseline). You can use the one-tailed test if you are absolutely sure you know what you're doing, but you must never switch from a two-tailed test to a one-tailed test if the two-tailed test does not demonstrate significance!

Variance and error bars

Consider a system for estimating elephant weights again. Suppose we have done lots of trials that show us how close the system gets to the real weight of the elephant, and we have adjusted this system so that it is

equally likely to overestimate or underestimate the elephant weight. Perhaps we find that the system is within 100kg of the real elephant weight 95% of the time. Then, when we estimate the weight of an unseen elephant, we could quote a figure as +/- 100kg. For instance: 6,400 +/- 100 kg. These confidence limits are shown on a graph using error bars.

In many ways, variance is more useful than p-values (see https://www.youtube.com/watch?v=5OL1RqHrZQ8).

In this course, the ML methods we are using are trained deterministically, so for given training data and parameter settings, you should always get the same result on some given test data. But when you do cross-validation, the variance gives you some idea of how much the results will vary with different data. Some other ML techniques, including most neural network approaches, will give different results on the same data. When using such techniques, it is good practice to do at least three replications (more if there is a lot of variability) and to report the mean and variance of the results.

Soft indications of variance

Consider elephant weighing again. If we know that our system is only accurate to within 100kg, we should not quote the weight it estimates to the nearest 1kg. The number of significant digits we use gives a soft indication of accuracy even if we are not using error bars. An extremely frequent mistake is to quote ratios to an excessive number of decimal places. For instance, if we do 123 trials and obtain the correct result in 63 cases, we should give the accuracy as 0.51 rather than 0.5122. An easy way to think about this is just to think about the raw numbers we are considering, and compare with the case where we are doing a nice round number of trials (e.g., 100). It makes no sense to quote the accuracy out of 123 trials as .5122, just as it would make no sense to say that we got 51.22 results correct out of 100.

We might decide to use even fewer significant figures when we know there are errors in measurements. For instance, if we do 1230 trials and obtain 630 correct results, we could potentially give the accuracy as 0.512, but if we know that the 95% confidence limits on these results are around +/- 10, it would be better to quote the number as 0.51 instead.