

Hypothesis Testing

Scientific facts are established through experiment. The first time a fact was established, it typically pushed the limits of the instruments of the day. The finite precision of real instruments means that actual measurements are never perfect. And randomness can lie in the data itself, or effects of outside influences. e.g. Monte Carlo algorithms have running times that are random variables. Suppose Monte Carlo algorithm A has expected running time which is less than B. But on specific runs, B may occasionally run faster than A. If we have only observations of the two algorithms to go by, how can we decide which is faster? How confident can we be in that decision? If you see some unusual network traffic in a network trace, is it random variation or evidence of intrusion? These are the basic questions of analytical statistics. To make progress on any of these questions, we have to make a *Hypothesis* and *Test* it.

An Example

Suppose we have two types of dice: fair dice, which produce uniform random values from 1 to 6; and loaded dice, which always produce a 5 or 6 uniformly. We pick one of these dice, and throw it three times. It comes up 6, 5, 6. We have one of two hypotheses:

- (A) The die is fair.
- (B) The die is loaded in the manner described above.

While there can be many possible hypothesis, one is given a privileged status as the “Null hypothesis” while the others are called “alternate” hypotheses. The “Null hypothesis” is the only one whose distribution we have to analyze, so it should be as simple as possible. Paradoxically, the null hypothesis is usually the opposite of what the experimenter wants to show. i.e. to show that a drug has an effect, the null hypothesis is that it has no effect. To show that one algorithm is faster than another, the null hypothesis is that their speed is the same. For the dice example, the Null hypothesis is hypothesis A that the dice is fair. The alternate is hypothesis B. How do we decide whether to reject A, and what does that tell us about hypothesis B?

Bayesian Approach

We could take a Bayesian approach to making this decision. First we define some events:

Event A: The die is fair.

Event B: The die is loaded.

Event C_i : The i^{th} throw produces a 1, 2, 3 or 4

Event D_i : The i^{th} throw produces a 5 or 6

Then we can compare the probability $\Pr(A|D_1 \cap D_2 \cap D_3)$ with $\Pr(B|D_1 \cap D_2 \cap D_3)$, the probabilities of the two hypothesis given the data. If one of these is very small, we would reject that hypothesis and accept the other (whose probability would be close to 1). Bayes' rule gives us:

$$\Pr(A|D_1 \cap D_2 \cap D_3) = \frac{\Pr(D_1 \cap D_2 \cap D_3|A) \Pr(A)}{\Pr(D_1 \cap D_2 \cap D_3)}$$

where

$$\Pr(D_1 \cap D_2 \cap D_3) = \Pr(D_1 \cap D_2 \cap D_3|A)\Pr(A) + \Pr(D_1 \cap D_2 \cap D_3|B)\Pr(B)$$

We can compute most of these probabilities easily, except for $\Pr(A)$ and $\Pr(B)$. We have to choose those. There is no way around this. Bayesian inferences involves a-priori assumptions about the hypothesis. In this case we assume they are equally likely, i.e. their *prior* probabilities are:

$$\Pr(A) = \Pr(B) = 1/2$$

Since the tosses are independent, its easy to see that

$$\Pr(D_1 \cap D_2 \cap D_3|A) = \left(\frac{1}{3}\right)^3 = \frac{1}{27}$$

while

$$\Pr(D_1 \cap D_2 \cap D_3|B) = 1$$

and substituting we find that

$$\Pr(A|D_1 \cap D_2 \cap D_3) = \frac{\frac{1}{27} \frac{1}{2}}{\frac{1}{27} \frac{1}{2} + \frac{1}{2}} = \frac{1}{28}$$

This is quite a small probability and it seems safe to reject hypothesis (A). (B) is the complement of (A), so its probability conditioned on the observations must be 27/28 which we might as well accept.

Problems with Bayes

Surprisingly, traditional statistical testing does not work this way. Inevitably, only the “forward” probability $\Pr(D_1 \cap D_2 \cap D_3|A)$ of the observations given the null hypothesis is used. The reasons are not obvious, but they are compelling.

Determining Priors

First of all, $\Pr(A|D_1 \cap D_2 \cap D_3)$ is a posterior probability that requires use of Bayes' rule. And this in turn requires choice of prior probabilities $\Pr(A)$ and $\Pr(B)$. That is, experimenters would have to choose how much to “bet” on their favored hypothesis. One might require that equal probabilities be assigned, but in many cases this would be completely unrealistic. Also, if the experimenter picks unrealistic alternate hypotheses even with high priors, Bayes rule will tend to give high probability to the null hypothesis even though it is false. With all these choices, there is far to much scope for experimenters to influence the outcome of the experiment, intentionally or not.

Enumerating Alternates

A second difficulty is that for many tests, the Null hypothesis is simple while its complement is complex or intractable. For the dice example above we assumed we knew exactly how the biased dice were biased. But this is very unrealistic. In reality, the bias could be almost anything. Representing it and making inference would be difficult or impossible even for this simple example. On the other hand, the Null hypothesis that the dice is fair is extremely easy to work with. More generally, if we are comparing means of two samples to decide if the samples are different, the null hypothesis would be that they are from a common normal distribution. This is very easy to analyze. The complement of the null hypothesis would at least assume the populations are from distinct normal distributions with different means and variances, and may further allow non-normal distributions. Analyzing these alternate hypotheses is difficult and requires many assumptions about additional parameters which can be another source of error or bias.

Analytical Statistics

A mathematician who discovers an implication $A \Rightarrow B$, and then determines that B is false has no trouble asserting that A is false as well (modus tollens). Analytical statistics works in a similar way. The null hypothesis A generates a distribution of observations, some of which are less likely than others. When an observation B is sufficiently unlikely, we reject A . We can do this even though we haven't determined the probability of another hypothesis. But in order to come up with a good statistical test, we normally do look at the alternative hypotheses in order to pick a test that help us reject the null hypothesis assuming an alternative hypothesis is true.

Our goal is to come up with a test that gives a low $\Pr(\text{Observation}|\text{Null hypothesis})$ when some other hypothesis holds. But the space of possible observations will often be enormous and any specific observation will have very small probability. We are not quite “saying what we mean” by a low-probability observation given the hypothesis.

The solution is to choose a *test statistic*. That is real-valued function on the outcomes of the experiment. Under the Null hypothesis, we have a known probability distribution on the outcomes, and we can predict the distribution of the test statistic. Generally, we try to find a statistic that has a bell-shaped distribution. An observation is unlikely if its statistic is in the tails of this distribution. We assign the outcome a finite probability which is the total probability of all outcomes whose statistics have equal or greater distance from the mean.

Example

Consider our dice experiment from earlier. A die is thrown 3 times, and the outcomes are 6,5,6.

Null Hypothesis

The die is fair, the result of each toss is independent and uniformly distributed in $\{1, \dots, 6\}$.

Test Statistic

The statistic T is the sum of the values on the 3 die throws.

Under the Null hypothesis of fair die, the test statistic has a simple bell-shaped distribution shown below. There are 216 equally-likely outcomes. The observation 6, 5, 6 has a statistic of 17, which is near the right tip of the distribution.

One-sided test

We want to find the probability of observations “at least as extreme” as our experimental observations. There are only 4 outcomes which produce a sum of 17 or greater, out of $6^3 = 216$. So $\Pr(T \geq 17) = 4/216 = 0.0185$.

p-value

The probability 0.0185 is called the “p-value” for the experiment. It is the probability of observations T at least as extreme as the experiment outcome.

One subtlety of experiment design is the meaning of “extreme”. 17 is an unusual value both because it is large, but also because it is far from the mean. The value 4 is equally far from the mean, and therefore equally extreme under the second definition. Should we consider one-sided or two-sided tails when computing p-values? This is an experiment design question. If one is sure that alternate hypotheses move the test statistic in one direction, one need only check the tail area in that direction. But if one has no a-priori knowledge of the affect of alternates, the two-sided t-test is more appropriate. In this case:

Two-sided test

There are 8 outcomes that produce a T value at least as far from the mean (10.5) as the experiment value. So $\Pr(|T - \mu| \geq 17 - \mu) = 8/216 = 0.037$.

In this case, the two-sided p-value is exactly twice the one-sided value. This is true for any symmetric test statistic distribution. Most common test statistics are in fact symmetric, and for all of these the two-sided p-value is twice the one-sided value.

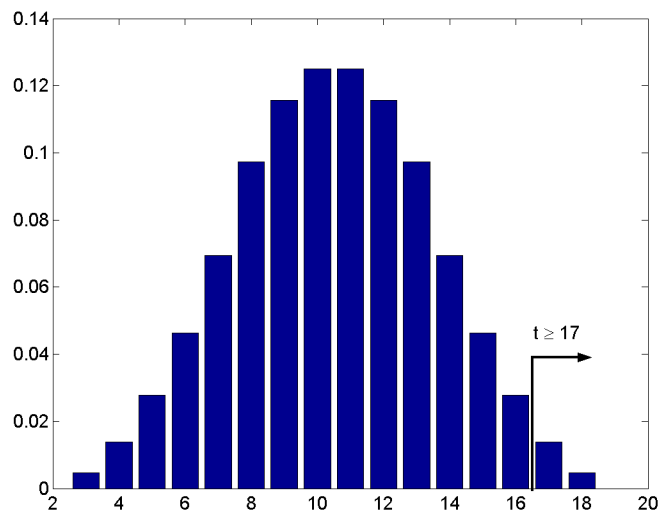


Figure 1: Probability distribution for the sum of 3 tosses of a fair die, and the test statistic value plotted at 17.

If we do reject the Null hypothesis, we cannot assert that hypothesis B is likely as we did in the Bayesian case. We never directly analyzed the probability of B. Rather, we reject the Null hypothesis and *assume* that some other hypothesis holds. It may be the hypothesis we used to derive the test statistic, but its very possible that something different is going on that we never accounted for.

Significance level

Finally, we need to make a decision. For this, we should specify how large a p-value we can live with. The maximum acceptable p-value is called the *significance* of the experiment. There is no hard and fast rule for picking this. Commonly used values include 0.05, 0.02, 0.01, 0.001. When the risk of an incorrect conclusion is small, larger values are acceptable. For our example, would could reject the Null hypothesis at the 0.05 level. But if the risks are high, e.g. if the null hypothesis is that a certain drug has serious side-effects, then the significance level should be orders of magnitude lower.

Significance values are nice round numbers because they are chosen *before* the experiment happens. The criteria for acceptable levels of error and risk have nothing to do with the experiment itself, and so should be set ahead of time.

Limitations of the Analytic Statistical Approach

Analytical statistics looks only at the forward probability of observation given the null hypothesis. That means a test using this approach will occasionally reject the Null hypothesis even though its alternates are even *less* likely. However, if the test is well-designed, the likelihood of this is low and one can argue that most of the time (i.e. for most repetitions of the experiment), when the null hypothesis has low probability, some alternate hypothesis will have high probability. But recognize that this is not true for any particular experiment - since we never actually look at the probability of the alternate hypotheses, we do not know. That is why statisticians are very careful how they report the result of an experiment, and experimenters should be too. You cannot “accept” an alternate hypothesis just because you rejected the Null hypothesis. You can say the experiment is consistent with some alternate, but that is about all.

Statistical Formalities

The example from the last section illustrates a lot of the subtleties of real statistical testing. Now that we are going to define some more tests, we need some definitions to clarify exactly what a test does, and how good it is. First we need to describe what can go wrong.

Type I and II errors

There are two ways a null hypothesis test can fail. Conventionally, the null hypothesis is labelled “Null” because it is the opposite of what we want to show. That is a “positive” outcome to the test is a rejection of the Null hypothesis. In our example, a positive outcome would be labelling the die as loaded. A negative outcome, or failure of the test, means that we were unable to reject the Null hypothesis and we still believe that the die might be fair.

Type I error is a false positive, or rejection of the Null hypothesis when it is in fact true. e.g. we might conclude that a drug has an effect on a test population when it does not.

Type II error is a false negative, or failure to reject the Null hypothesis when it is actually false. e.g. we conclude that a drug has no effect when it actually does.

There is a direct trade-off between these two errors types through choice of the significance level. Smaller significance thresholds reduce the rate of type I errors at the expense of an increase in type II error rate, and vice versa. The best trade-off involves analysis of the *risks* of the test. If you are just trying to find the faster of algorithm A and B when you compare them, the two types of error involve essentially the same risk. So you might as well adjust the test so they are equal. But if the null hypothesis is that a drug has no serious side effects, then type II errors, which mean failing to recognize the side effects, are much more serious. Consequently you would adjust the test so the probability of a type II error was extremely low. You would do this at the expense of type I errors, flagging innocuous drugs as potentially harmful.

Significance Level and Power

We defined the significance level earlier as the probability threshold at which we reject the null hypothesis. It was the probability of seeing the observations by chance if the null hypothesis were true. Therefore the significance level is exactly the probability of a type I error. By setting a threshold in the tail of the test statistic distribution as we did in our example, we fix the significance level and the probability of a type I error. The more extreme the value, the lower the significance level, and the lower the probability of a type I error. However, the probability of type II errors increases because the alternate hypothesis distribution may not produce values which are extremal enough to reach the significance threshold. Unlike type I errors, there is no simple graphical relationship between the null distribution and the probability of a type II error. Analyzing type II errors requires us to go beyond the null hypothesis, and consider the distribution induced by alternate hypothesis.

Power of a test relative to an alternate

The power of a test relative to an alternate is the probability of rejecting the null hypothesis if the alternate is true. Thus the power of a test is one minus the probability of a type II error for that alternate hypothesis. There will often be many alternate hypothesis for the same test, and the power will in general be different for each alternate hypothesis. e.g. suppose the null hypothesis is that the means of two samples are equal. One alternate would be that the means differ by 1, another that they differ by 2 etc. Almost any test will have greater power relative to the second hypothesis.

Effect Size and Sensitivity

To get a better handle on the power of a test, we need the concept of *effect size*. An effect size is a measurable quantity for an actual population of experimental subjects. e.g. it might be the measured frequencies of 5's and 6's for a particular die. Or it may be the actual difference in average running times for algorithms A and B. The larger the effect size, the more power a given statistical test will have. Furthermore, if the null hypothesis is rejected, the estimated effect size is a very useful reportable "outcome" of the experiment.

Effect size and significance level influence but do not fully determine the power of a test. The final attribute of the test is its sensitivity. Sensitivity is the size of the smallest effect needed to trigger rejection of the null hypothesis. Sensitivity can be increased by more careful controls on the experiment, or by increasing the sample size. e.g. toss the die more often, study a larger population of people, or do additional runs of algorithm A and B.

Example

For instance, suppose we have 8 observations of the running time for algorithm A, and 6 for B. Here they are:

Algorithm A	95.7	83.3	101.2	102.9	88.5	111.9	112.0	99.6
Algorithm B	112.9	96.6	117.1	126.3	103.1	118.6		

The mean time for algorithm A is 99.4 secs, while it is 112.4 for algorithm B. But what does that really tell us? How likely is this to happen by chance if they are really the same, or if B is actually faster than A? Lets formulate the null hypothesis. To make this tractable, we will assume we already know the variance of both running times (say 100).

Null hypothesis 1 Assume that running times for algorithm A and B come from the same distribution, and are independent, identically-distributed normal random variables with the same mean and variance 100.

Before we conduct the test, we should decide on the significance level we can live with. There is very little risk associated with a type I error, and we would just like to know if there is a speed winner and which algorithm it is. So 0.05 seems appropriate. This is an upper bound on the risk of a type I error.

Test Statistic We want to distinguish the alternate hypothesis (that the two algorithms have different running times) from the null hypothesis. A natural test statistic is the difference in means between algorithms A and B. Under the null hypothesis, this difference has expected value zero. Under the alternate hypothesis, it should be large.

Apart from doing a good job of distinguishing the hypotheses, this test statistic makes it very easy to analyze the null distribution. Although we dont know the actual mean of the null distribution, it doesnt affect the distribution of the test statistic because we take the difference between the two sample means. This makes life much easier (otherwise we would have to integrate over the posterior density of a continuous parameter which is actual null distribution mean).

The variance of the mean of 8 measurements of A is $100 \cdot 8 / 8^2 = 12.5$, and for B is $100 \cdot 6 / 6^2 = 16.67$. The variance of the difference between them is the sum of their variances which is normal with mean zero and variance 29.17. This distribution is plotted below.

We havent dealt with continuous probability distributions before, so the figure needs a little explanation. You cant assign finite probability to specific values on the x axis, but you can to a range of x values. The probability that x lies in a range $a \leq x \leq b$ is just the area under the curve between a and b. The total area under the curve is therefore 1. You can think of this as a very dense histogram.

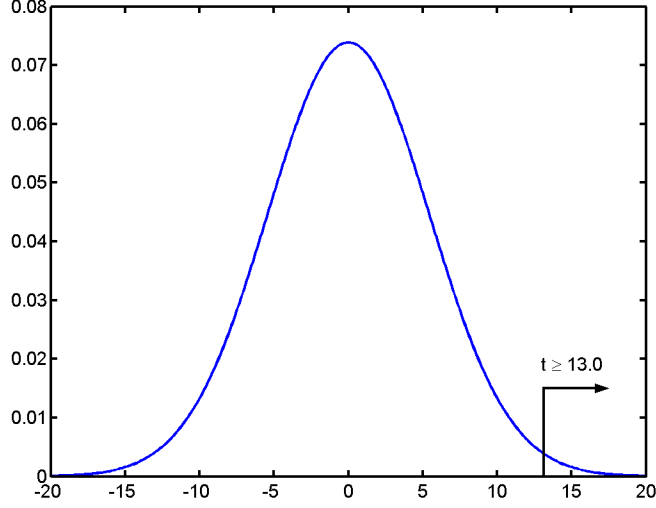


Figure 2: Fixed-variance normal distribution for the difference between the sample means, and the test statistic at 13

The difference between the two sample means (13.0) is also shown on this plot. To assign a finite probability to this outcome, we consider the area under the probability curve to the right of 13. This is the total probability of an outcome that is at least as extreme as what we observed. In this case the one-sided p-value 0.008 (two-sided is twice this). We would see such an event less than one time in a hundred by chance.

This test was rather unrealistic. If we don't know the mean running time for either algorithm, it's unlikely we know their variances. A more realistic null hypothesis would be:

Null hypothesis 2 Assume that running times for algorithm A and B come from the same distribution, and are independent, identically-distributed normal random variables with the same mean and variance.

We could try to use the same test statistic as before, which helps by eliminating the effect of the null distribution mean. Unfortunately, we still have to take into account the unknown variance. We haven't developed the tools to do this, but there is a standard method for doing so. It turns out if \bar{X}_A and \bar{X}_B are the means of the running time for A and B respectively, and if we define the sample variances as:

$$S_A = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} (X_{Ai} - \bar{X}_A)^2 \quad S_B = \frac{1}{n_B - 1} \sum_{i=1}^{n_B} (X_{Bi} - \bar{X}_B)^2$$

then the quantity

$$T = \frac{(\bar{X}_B - \bar{X}_A)}{S_{AB}}$$

no longer depends on the unknown mean or variance, where

$$S_{AB} = \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}$$

The resulting statistic T has a *t-distribution*¹. It depends only on the number of *degrees of freedom* of the combined statistic, which in this case is $n_A + n_B - 2 = 12$.

If we substitute the values from our sample, we find that $\bar{X}_A = 99.4$, $\bar{X}_B = 112.4$, $S_A = 102.54$, $S_B = 117.5$, and finally $S_{AB} = 5.63$. The T-statistic is

$$T = 2.31$$

which we show plotted on the graph below. The area in the right tail of the curve (one-sided p-value) is 0.0195. Since we have no a-priori reason to favor one algorithm over the other, the two-sided p-value (0.039) should be used here. We should still reject the null hypothesis according to our significance threshold of 0.05, but the p-value and probability of a type I error is increased compared to the test assuming known variance. This is to be expected since Hypothesis 2 makes fewer assumptions than 1. Hypothesis 2 admits more distributions to explain the data, and typically it will be harder to dismiss it.

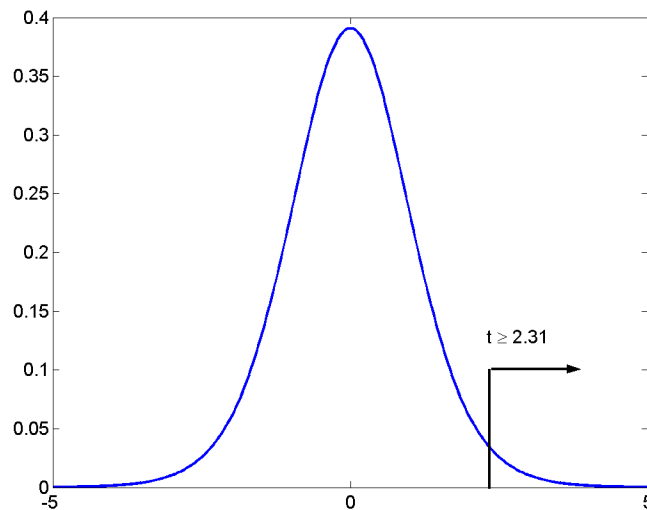


Figure 3: T statistic probability density and the test statistic for hypothesis 2

Finally, we may not know whether the variances of our two samples are the same or not, so it may be more appropriate to consider a hypothesis like this:

¹The full name is Student's T-distribution. "Student" was actually William Sealy Gosset, a statistician working at the Guinness Brewery in Dublin in the early 1900's. Gosset was under a confidentiality agreement with his employer not to publish any of his research, hence he published his seminal work on T-tests under the pseudonym "Student". There are several theories about Guinness' secretiveness, one of which is that they felt that use of statistical methods gave them a competitive advantage - a secret weapon for brewers!

Null hypothesis 3 Assume that running times for algorithm A and B are independent, normally-distributed random variables with the same mean but possibly different variances.

A variant of the t-statistic still applies to this case. The only change is to the definition of S_{AB} which becomes:

$$S_{AB} = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$$

Since this is a more inclusive hypothesis, we again expect the ease of rejection to decrease, and therefore for the p-value to increase. The new t-value becomes 2.29, only slightly smaller than before, and the new two-sided p-value is 0.0408.