Statistical Tests - 1

In a way, we could wrap up this module in a few lines. Statistical tests can be viewed at first glance, as a "reverse read" of interval estimation. Indeed, in their simplest form, they are.

A simple approach to tests

Consider the following problem: we would like to know if the "true mean" of a random variable is equal to a certain value μ. For example, we would like to determine if the batteries our factory is producing truly produce a voltage of mean 1.5 V. We take a sample, and, using our expertise in interval estimation, produce an interval estimate for the mean, with a confidence level we feel comfortable with. If 1.5 falls inside this interval, we can say that the test was passed, if it doesn't, we have to say that the test was failed.

Of course, we can repeat the pattern for any other estimation problem we have studied, or will learn about in the future. And, in a way, that's almost all there is to statistical tests. However, this would be a bit naive, as proved by the enormous amount of space this question takes in any statistics course. Part of this space is not far from fluff, as is the case in other areas of statistics, but, in fact, there is more to the problem of testing than the short paragraph above. However, maybe not surprisingly, the "more" concerns mostly the interpretation of test results, rather than the technique, which does not go much beyond what we described above.

A more sophisticated approach: *Hypotheses*

A small disclaimer

The history of statistical tests is somewhat complex and surprisingly acrimonious. Also, its theory dates back to the early 20th Century, is full of personalty conflicts, and this might explain the somewhat rigid, and sometimes confusing terminology that it comes with. In fact, as we will see momentarily, some of its formulations are outright misleading.

Additionally, this methodology comes, so to speak, in two parts. The first, and best known, was formulated and strongly pushed by Fisher. The second, possibly just as important, if not more, was formulated slightly later by Neyman and Pearson, and was bitterly opposed by Fisher. The confusion that this semi-religious debate created did not help to make it as clear and simple as it really is.

As an aside, we do not address in this course the interesting, but very different, approach to statistics known as *Bayesian Statistics*, which has had a significant upswing in recent decades. Just so you know, though, the theory we discuss here is completely meaningless from a Bayesian point of view. Since, however, the reasonable attitude to the existence of diverging methodologies is that they fit different problems, classical testing has a huge role in all our lives (after all, what we eat, in terms of FDA approval, the medical treatments we take, in the same terms, the whole warranty system for what we buy, and on, and on, relies on applications of this theory), and we should definitely learn what it is, what it can do for us, and, even more importantly, what it **cannot do** for us.

Null and Alternate Hypothesis

As it often happens, the terminology used in statistical testing is somewhat peculiar, but we can make sense of it if we stop and think for a moment. This first part of the discussion is essentially due to Fisher, who put testing on a precise basis.

Traditionally, the statement tested is called "the Null Hypothesis", and labeled H_0. This is a statement assumed to be true, until proved false. For example, we may test the hypothesis that a certain distribution has expected value equal to a specific number μ_0. The "Alternate Hypothesis" should be the negation of the Null Hypothesis, hence we would, in this case, set it as "H_1: $\mu \neq \mu_0$", where μ is the "true" expectation of our random variable.

The logic of a statistical test is the following: we observe a sample with the distribution we are studying, and look how it turns out. We then choose a function of the sample that should behave in a well defined way if the Null Hypothesis was true, and verify if it indeed behaved that way. We will not go into the methodology of choosing such a function in this introductory course, but, in our case, it can be shown that the smart choice is, in most cases—and certainly in the case when we can rely on the Central Limit Theorem—the sample mean is the best choice. If the distribution was assumed to be normal, or, at least, if the CLT can be relied to be in force, the sample mean will be normally distributed around the true mean, and we can use our knowledge of the normal distribution to verify that it did not stray too far from our assumed expectation μ_0. In simple words, then, if the sample mean turns out to be not too far from μ_0, we may feel comforted in assuming that that is indeed the correct expectation. If, instead, the sample mean is far off, we have strong evidence that μ_0 may not be the true expectation. Statistical tests simply formalize this argument in a standard protocol.

For simplicity, we'll discuss here in some detail tests on the mean of normal random variable, when the variance is known. Other situations are similar: you would use different "statistics", and different distributions (for example, when testing for the mean with unknown variance, you would use the expression discussed in the estimation module for estimating the mean in this case, and use the t distribution with the appropriate number of degrees of freedom, while to test for the variance we would use S^2, or s^2—look back at the corresponding sections in the "Estimation" module—depending on whether the "true" mean was known or not).

A Simple Example: Testing the Mean When the Variance is Known

Let's look at the simplest case: assume we somehow know that our normally distributed random variable has a known variance σ^2. Then, if the expected value is indeed μ_0, $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ is a standard normal variable, and we know its likely values. In fact, we may say that with 95% probability, Z should fall between -1.98, and 1.98. The quantitative step we still have to make is to decide what is our definition of "straying too far", that is, when do we feel that it is an unlikely chance event if $\bar{X}$ (and hence Z) is too far off. Typical (that is, traditional) thresholds for "unlikely" are values that fall in the "tails" of the distribution, as in the left- and right-most 5% (that is we say the value is *not* "far off" if it falls in the central 90% of the distribution), or in the left- and right-most 2.5% (the "reasonable" area is now the central 95%—this is often the most common choice), or in the left- and right-most 0.05% (accepting the central 99% of the distribution). If the test fails, say so: "the hypothesis has to be rejected" is the standard term.

If the test is passed, we don't say "the hypothesis is accepted", because, if we look carefully, the expected value could easily have been different from μ_0, and the mean could have still fallen outside the so-called "rejection zone" with reasonable probability—we say "the hypothesis cannot be rejected". This caution was prominently pushed by Fisher, and it fits well with the first half of the 20th Century scientific philosophy, where positive results are welcomed, but never considered final, while the real breakthroughs come when the experiment "falsifies" the original assumption. In Fisher's, now standard, terminology, a failed test is "significant", and a passed test is not.

It is clear that our test may force us to reject the Null Hypothesis, even if it was true, and we just hit an unexpected large "fluctuation" around the mean. Expanding the "acceptance region" from 90% to 95%, to 99%, reduces our risk of falling into this error—called "Error of Type I". Unfortunately, the wider our acceptance region, the smaller the "rejection region", so that we may more easily fail to reject a hypothesis that is false, simply because the true mean was different from μ_0, but $\bar{X}$ still could reasonably fall in the "acceptance region". This second possible error is called "Error of Type II", and it is clear that we cannot keep both errors down simultaneously, since they push in opposite directions.

We will see how to manage this conundrum shortly, but, whatever we may do, note that no result of a statistical test asserts an absolute answer: accepting or rejecting a hypothesis is done on the basis of *plausibility*, as in, for example, "if the true value had really been μ_0, our observation would correspond to an unusual swing away from the center, hence we feel it more reasonable to reject the assumption that the true mean was μ_0".

Let's look at a couple of concrete examples to show how this works in practice. Suppose we are testing our batteries, to check whether they provide indeed 1.5 V of electricity, and are using an instrument that yields measurements with a known standard deviation of 0.05 V. We sample 16 batteries from our line and consider the resulting average reading. What should we conclude?

- First we can set our Type I error—this is called the *significance level* of our test. The most common choices are the usual ones: 90%, 95%, 99%. We can also wait and not commit ourselves yet.

- Next we compute the value of $Z = \frac{\bar{X} - 1.5}{\sigma/\sqrt{n}}$, to change our average into an approximate standard normal variable, assuming that the batteries do indeed produce 1.5 V.

There are two ways to proceed. The simplest way is to fix the significance level and see whether Z falls within the *acceptance region*, that is within the interval around 0 that has probability equal to the level we have chosen:

- For a 90% level test, that's $(-1.64, 1.64)$

- For a 95% level test, that's $(-1.96, 1.96)$

- For a 99% level test, that's $(-2.58, 2.58)$

If Z falls within the interval of our choice, we *cannot reject the Null Hypothesis*, if it doesn't we reject it.

In the more cautious approach, we don't set out with a set significance level, but rather compute the so-called *p-value* for our Z: that's the highest significance level that would allow us not to reject the Null Hypothesis. Thus, for example, if we ended up with $Z = -1.64$, the *p-value* would be 90% (or, rather, 10%, if you decide to use the complementary probability—usage differs among practitioners, but the second is more common). Since a decision has to be made, at this point it would be up to you to decide whether the p-value warrants rejecting or not. This approach has the merit of showing whether yours was a borderline decision, or fairly clear-cut, and is the preferred method: you should report the p-value of your test.

Suppose we ended up with an average reading of 1.3 V. Then, find

$$4 \cdot \frac{1.3 - 1.5}{0.05} = -80 \cdot 0.2 = -16$$

The p-value (in the second sense) of this result is practically 0, so that there is little doubt that we have to reject the hypothesis that the batteries are good. What if we had found an average of 1.4?

$$4 \cdot \frac{-0.1}{0.05} = -8$$

The p-value is a little larger, but is still practically zero. Let's try 1.45

$$4\frac{-0.05}{0.05} = -4$$

This has a p-value of about 0.001. We would still reject the hypothesis, of course (we would not reject it, if we decided to set our significance at 99.9%—in other words, we would need observations that occur once in a thousand or less to falsely reject the hypothesis, and this is so conservative that it amounts to cheating). The point here is that our sample is not large, but our standard deviation is very small—we have a very sensitive instrument. What reading would cause us to at least start considering the possibility that the batteries are within specifications? Well, if we found $Z = -1.95$, we would probably not reject the batch, but we would be at the border. That would correspond to

$$\bar{X} = 1.5 - 1.95 \cdot \frac{\sigma}{\sqrt{n}} = 1.5 - 1.95 \cdot \frac{0.05}{4} \approx 1.48$$

Two-tailed and One-tailed Tests

The previous example is a "two-tailed" test, in that we would reject the Null Hypothesis both if Z fell into the left, or the right tail of the distribution. Thus, $Z = -4$ is just as bad as $Z = 4$.

Sometimes, we are only worried that our quantity may be either too large or too small, but not both. For example we may have a target amount of a toxic substance in a product (that is an amount that is still safe), but will be fine if it is less. On the other hand, we might instead test for an amount that is too large (and if it was even larger, that certainly would not change our conclusion). Which choice we make makes a huge difference—a point we need to have very clear.

In any case, a test like the first one we mentioned will often be presented as

H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

The other option would be similar, with the opposite inequality.

This way of writing is a little odd, because, in fact, the first null hypothesis is, logically speaking, H_0: $\mu \leqslant \mu_0$. Indeed, if our sample mean turned out to be abnormally low, we would be very happy, even though it is clear that μ_0 is not a likely value for the "true" mean. To put it differently. to figure out how a test works, *check the alternate hypothesis*: that's the one that really defines the test. In fact, one could say (even if this is not common) that *the Null Hypothesis is the negation of the Alternate Hypothesis*. Since the core of a test is in the rejection, not in the acceptance, this is perfectly logical, but habits are hard to change, so expect to see this "asymmetric" test definitions (including in our On Line Stat book).

To see how this type of test works out, let's look at a hypothetical test set as

$$H_0: \mu = 10$$
$$H_1: \mu > 10$$

(again, it would be more rational to write H_0: $\mu \leqslant 10$) and suppose the variance is know to be equal to 1, while the observation of a sample of 9 results in a sample mean of 10.5.

Now, since we are not worried about small values, if we are looking at, say, a significance level of 95%, the 5% rejection region is all to the right, and not split between the two tails, as in the previous problem. Let's compute our new Z:

$$Z = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} = 3 \cdot \frac{10.5 - 10}{1} = 1.5$$

To determine the p-value, we look at the probability of a standard normal variable to take values larger than 1.5, which is, approximately, 0.144—-definitely not something that would suggest rejecting the hypothesis (it would be accepted at a significance level of even less than 90%). Note that "you are testing if you are at 10", and this is the assumption that you cannot reject.

Now, suppose you were making the opposite test (with the same data), instead, this time testing whether the dangerous level of 11 was reached:

$$H_0: \mu = 11$$
$$H_1: \mu < 11$$

(again, you may rather think of the Null Hypothesis to be $H_0: \mu \geqslant 11$, that is, you are verifying whether the contaminant is above its safe level, rather than whether it is within its safe amount). The rest being equal, $Z = -1.5$, and the argument is the same (everything is in the opposite direction, but the numbers turn out the same): we have the same p-value, and hence we will definitely not reject the Null Hypothesis. However, in this case, the Null Hypothesis implies the exact opposite of the Null Hypothesis in the previous case. So, the same data will lead to two opposite conclusions, depending on the question asked: if we ask "are we safe", the answer is "yes"; if we ask "are we unsafe", the answer is "yes" as well!

Did you notice what happened? Fact is a statistical test is "stacked" in favor of the Null Hypothesis. Tests are designed to reject it only when there is really strong evidence against it. Clearly, we should not make too much of a passed test, even though the p-value would give us a better understanding of what the data seems to suggest.

However, there is a more detailed analysis that can be performed, giving a much more refined picture of what the test result is. From the previous example we may notice that the data actually cannot really allow us to tell whether μ is 10 or 11. The two values are simply too close to be distinguished. The next module will set out a method to turn this reflection into a quantitative method to understand what we can say in a situation like this.