

# Statistics and Hacking: An Introduction to Hypothesis Testing

[Originalartikel](#)

[Backup](#)

In the early 20<sup>th</sup> century, Guinness breweries in Dublin had a policy of hiring the best graduates from Oxford and Cambridge to improve their industrial processes. At the time, it was considered a trade secret that they were using statistical methods to improve their process and product.

One problem they were having was that the z-test (a commonly used test at the time) required large sample sizes, and sufficient data was often unavailable. By studying the properties of small sample sizes, [William Sealy Gosset](http://en.wikipedia.org/wiki/William_Sealy_Gosset) developed a statistical test that required fewer samples to produce a reasonable result. As the story goes though, chemists at Guinness were forbidden from publishing their findings.

So he did what many of us would do: realizing the finding was important to disseminate, he adopted a pseudonym (Student) and published it. Even though we now know who developed the test, it's still called Student's t-test; and it remains widely used across scientific disciplines.

It's a cute little story of math, anonymity, and beer; but what can we do with it? As it turns out, it's something we could probably all be using more often, given the number of Internet-connected sensors we've been playing with. Today our goal is to cover hypothesis testing and the basic z-test, as these are fundamental to understanding how the t-test works. We'll return to the t-test soon; with real data.

I recently purchased two of the popular DHT11 temperature-humidity sensors. The [datasheet](http://akizukidenshi.com/download/ds/aosong/DHT11.pdf) (PDF warning) says that they are accurate to +/- 2 degrees C and 5% relative humidity within a certain range. That's fine and good, but does that mean the two specific sensors I've purchased will produce significantly different results under the same conditions? Different enough to affect how I would use them? Before we discuss how to quantify that, we'll have to go over some basic statistical theory. If you've never studied statistics before, it can be less than intuitive, so we'll go over a more basic test before getting into the details of Student's t-test.

It's worth starting by mentioning that there are two major schools of statistics; Bayesian and Frequentist (and there's a bit of a holy war between them). A detailed discussion of each does not belong here, although if you want to know more [this article](http://scienceblogs.com/goodmath/2008/04/07/schools-of-thought-in-probabil/) provides a reasonable summary. Or if you prefer a comic, [this one](http://xkcd.com/1132/) should do. What's important to remember is that while our test will rely upon the frequentist interpretation of statistics, there are other correct ways of approaching the problem.

For our example, imagine for a moment you are working quality control in a factory that makes 100k resistors. The machinery is never perfect, so while the average value of the resistors produced is 100k, individual resistors have slightly different values. A measure of the spread of the individual values around the 100k average is the standard deviation (σ). If your machine is working correctly, you would probably also notice that there are fewer resistors with very high deviations from 100k, and more resistors closer to 100k. If you were to graph the number of resistors produced of each value, you would probably get something that looks like this:



data-permalink=<https://hackaday.com/2017/12/01/statistics-and-hacking-an-introduction-to-hypothesis-te>

sting/empirical\_ruleyl/“

data-orig-file=„[https://hackadaycom.files.wordpress.com/2017/11/empirical\\_ruleyl.png?w=800](https://hackadaycom.files.wordpress.com/2017/11/empirical_ruleyl.png?w=800)“

data-orig-size=„769,558“ data-comments-opened=„1“ data-image-

meta=„{&quot;aperture&quot;:&quot;0&quot;,&quot;credit&quot;:&quot;&quot;,&quot;camera&quot;:&quot;&quot;,&quot;caption&quot;:&quot;&quot;,&quot;created\_timestamp&quot;:&quot;0&quot;,&quot;copyright&quot;:&quot;&quot;,&quot;focal\_length&quot;:&quot;0&quot;,&quot;iso&quot;:&quot;0&quot;,&quot;shutter\_speed&quot;:&quot;0&quot;,&quot;title&quot;:&quot;&quot;,&quot;orientation&quot;:&quot;0&quot;}“ data-image-title=„Empirical\_Ruleyl“ data-image-description=„“ data-medium-file=„[https://hackadaycom.files.wordpress.com/2017/11/empirical\\_ruleyl.png?w=800?w=400](https://hackadaycom.files.wordpress.com/2017/11/empirical_ruleyl.png?w=800?w=400)“

data-

large-file=„[https://hackadaycom.files.wordpress.com/2017/11/empirical\\_ruleyl.png?w=800?w=769](https://hackadaycom.files.wordpress.com/2017/11/empirical_ruleyl.png?w=800?w=769)“

class=„aligncenter wp-image-282553 size-full“

src=„[https://hackadaycom.files.wordpress.com/2017/11/empirical\\_ruleyl.png?w=800](https://hackadaycom.files.wordpress.com/2017/11/empirical_ruleyl.png?w=800)“ alt=„“

srcset=„[https://hackadaycom.files.wordpress.com/2017/11/empirical\\_ruleyl.png](https://hackadaycom.files.wordpress.com/2017/11/empirical_ruleyl.png) 769w,

[https://hackadaycom.files.wordpress.com/2017/11/empirical\\_ruleyl.png?w=250](https://hackadaycom.files.wordpress.com/2017/11/empirical_ruleyl.png?w=250) 250w,

[https://hackadaycom.files.wordpress.com/2017/11/empirical\\_ruleyl.png?w=400](https://hackadaycom.files.wordpress.com/2017/11/empirical_ruleyl.png?w=400) 400w“ sizes=„(max-

width: 769px) 100vw, 769px“/></p><p>This is a bell curve, also called a normal or Gaussian distribution, which you have probably seen before. If you were very astute, you might also notice that 95% of your resistor values are within two standard deviations of our average value of 100&#160;&#937;. If you were particularly determined, you could even make a table for later reference defining what proportion of resistors would be produced within different standard deviations from the mean. Luckily for us, <a

href=„<http://sixsigmastudyguide.com/wp-content/uploads/2014/04/z-table.jpg>“ target=„\_blank“>such tables already exist</a> for normally distributed data, and are used for the most basic of hypothesis tests: the z-test.</p><p>Let&#8217;s say you then bought a machine that produces

100&#160;&#937; resistors &#8212; you quit your job in QC and have your own factory now. The vendor seemed a bit shady though, and you suspect the machine might actually be defective and produce resistors centered on a slightly different value. To work this out, there are four steps: develop a set of hypotheses, sample data, check if the sampled data meets the assumptions of your test, then run the test.</p><h2>Developing Hypotheses</h2><p>There are only two possibilities in our case: the machine either produces resistors that are significantly different from 100&#160;&#937;, or it doesn&#8217;t. More formally you have the following hypotheses:</p><p>H0: The machine does not produce resistors that are significantly different from 100&#160;&#937;</p><p>HA: The machine produces resistors that are significantly different from 100&#160;&#937;</p><p>H0 is called our null hypothesis. In classical statistics, it&#8217;s the baseline, or the hypothesis to which you&#8217;d like to give the benefit of the doubt. Here, it&#8217;s the hypothesis that we don&#8217;t find a difference between the two machines. We don&#8217;t want to go complaining to the manufacturer unless we have clear evidence that the machine isn&#8217;t making good resistors.</p><p>What we will do is use a <a

href=„<http://sixsigmastudyguide.com/wp-content/uploads/2014/04/z-table.jpg>“ target=„\_blank“>z-score table</a> to determine the probability that some sample we take is consistent with H0. If the probability is too low, we will decide that H0 is unlikely to be true. Since the only alternative hypothesis is HA, we then decide to accept HA as true.</p><p>As part of developing your hypotheses, you will need to decide how certain you want to be of your result. A common value is 95% certainty (also written as &#945;=0.05), but higher or lower certainty is perfectly valid. Since in our situation we&#8217;re accusing someone of selling us shoddy goods, let&#8217;s try to be quite certain first and be 99% sure (&#945;=0.01). You should decide this in advance and stick to it &#8211; although no one can really check that you did. You&#8217;d only be lying to yourself though, it&#8217;s up to your readers to decide whether your result is strong enough to be

convincing.

## Sampling and Checking Assumptions

Next you take a random sample of your data. Lets say you measure the resistance of 400 resistors with your very accurate multimeter, and find that the average resistance is  $100.5 \pm 1$ , with a standard deviation of  $1 \pm 1$ .

The first step is to check if your data is approximately shaped like a bell curve. Unless you've purchased a statistical software package, the easiest way I've found to do this is using the scipy stats package in Python:


```
import scipy.stats as stats
list_containing_data=[]
result = stats.normaltest(list_containing_data)
print result
```

As a very general rule, if the result (output as the  $p$ -value) is more than 0.05, you're fine to continue.

Otherwise, you'll need to either choose a test that doesn't assume a particular data distribution or apply a transformation to your data. We'll discuss both in a few days.

As a side note, testing for normality is sometimes ignored when required, and the results published anyway. So if your friend forgot to do this, be nice and help them out. No one wants this pointed out for the first time publicly (e.g. a thesis defense or after a paper is published).

## Performing the Test



data-

permalink=„<https://hackaday.com/2017/12/01/statistics-and-hacking-an-introduction-to-hypothesis-testing/stderryel/>“ data-orig-file=„<https://hackadaycom.files.wordpress.com/2017/11/stderryel.png>“

data-orig-size=„800,419“ data-comments-opened=„1“ data-image-

meta=„{&quot;aperture&quot;:&quot;0&quot;,&quot;credit&quot;:&quot;&quot;,&quot;camera&quot;:&quot;&quot;,&quot;caption&quot;:&quot;&quot;,&quot;created\_timestamp&quot;:&quot;0&quot;,&quot;copyright&quot;:&quot;&quot;,&quot;focal\_length&quot;:&quot;0&quot;,&quot;iso&quot;:&quot;0&quot;,&quot;shutter\_speed&quot;:&quot;0&quot;,&quot;title&quot;:&quot;&quot;,&quot;orientation&quot;:&quot;0&quot;}“ data-image-title=„stderryel“ data-image-description=„“ data-medium-

file=„<https://hackadaycom.files.wordpress.com/2017/11/stderryel.png?w=400>“

data-large-file=„<https://hackadaycom.files.wordpress.com/2017/11/stderryel.png?w=800>“ class=„wp-image-282550 size-thumbnail“


src=„<https://hackadaycom.files.wordpress.com/2017/11/stderryel.png?w=250&h=131>“ alt=„“ width=„250“ height=„131“

srcset=„<https://hackadaycom.files.wordpress.com/2017/11/stderryel.png?w=250&h=131> 250w, <https://hackadaycom.files.wordpress.com/2017/11/stderryel.png?w=500&h=262> 500w, <https://hackadaycom.files.wordpress.com/2017/11/stderryel.png?w=400&h=210> 400w“

sizes=„(max-width: 250px) 100vw, 250px“/><figcaption class=„wp-caption-text“>#963;: standard deviation, n: number of samples</figcaption></figure>

Now that the hard part is over, we can do the rest by hand. To run the test, we determine how many standard errors away from  $100 \pm 1$  the our sample average is. The standard error is the standard deviation divided by the square root of the sample size. This is why bigger sample sizes let you be more certain of your results. Everything else being equal, as sample size increases your standard error decreases.

In our case the standard error is  $0.05 \pm 1$ .



data-

permalink=„<https://hackaday.com/2017/12/01/statistics-and-hacking-an-introduction-to-hypothesis-testing/ztestyel/>“ data-orig-file=„<https://hackadaycom.files.wordpress.com/2017/11/ztestyel.png>“ data-

orig-size=„800,358“ data-comments-opened=„1“ data-image-

meta=„{&quot;aperture&quot;:&quot;0&quot;,&quot;credit&quot;:&quot;&quot;,&quot;camera&quot;:&quot;&quot;,&quot;caption&quot;:&quot;&quot;,&quot;created\_timestamp&quot;:&quot;0&quot;,&quot;copyright&quot;:&quot;&quot;,&quot;focal\_length&quot;:&quot;0&quot;,&quot;iso&quot;:&quot;0&quot;,&quot;shutter\_speed&quot;:&quot;0&quot;,&quot;title&quot;:&quot;&quot;,&quot;orientation&quot;:&quot;0&quot;}“ data-image-title=„ztestyel“ data-image-description=„“ data-medium-

file=„<https://hackadaycom.files.wordpress.com/2017/11/ztestyel.png?w=400>“

data-large-file=„<https://hackadaycom.files.wordpress.com/2017/11/ztestyel.png?w=800>“ class=„wp-image-282547 size-thumbnail“  
src=„<https://hackadaycom.files.wordpress.com/2017/11/ztestyel.png?w=250&h=112>“ alt=„“  
width=„250“ height=„112“  
srcset=„<https://hackadaycom.files.wordpress.com/2017/11/ztestyel.png?w=250&h=112> 250w,  
<https://hackadaycom.files.wordpress.com/2017/11/ztestyel.png?w=500&h=224> 500w,  
<https://hackadaycom.files.wordpress.com/2017/11/ztestyel.png?w=400&h=179> 400w“  
sizes=„(max-width: 250px) 100vw, 250px“/><figcaption class=„wp-caption-text“>M: a value we’re testing against, 181;: the sample average</figcaption></figure><p>Next we calculate the test statistic, z. This is the difference between the sample mean of 100.5% and the value we’re testing against of 100%, divided by the standard error. That gives us a z value of 10, which is rather large as z-statistic tables typically only go up to 3.49. This means the probability (p) that our null hypothesis is correct is less than 0.001 (or less than 0.1% if you prefer). We would normally report this as  $p < 0.001$ , as no one really cares what the precise value of p is when it’s that small.</p><h2>What Does it Mean?</h2><p>Since our calculated p is lower than our threshold value of 0.01 we reject the null hypothesis that the average value of resistors produced by the machine is 100%; there’s definitely an offset, but do we call our vendor?</p><p>In real life, statistical significance is only part of the equation. The rest is effect size. So yes, our machine is significantly off specification; but with a standard deviation of 1%, it wasn’t supposed to be good enough to produce 1% tolerance resistors anyway. Even though we’ve shown that the true average value is higher than 100%, it’s still close enough that the resistors could easily be sold as 5% tolerance. So while the result is significant, the (fictional) economic reality is that it probably isn’t relevant.</p><p>This is all well and good for our fictional example, but in real life data tends to be expensive and time-consuming to collect. As hackers, we often have very limited resources. This is an important limitation to the z-test we’ve covered today, which requires a relatively large sample size. While Internet-connected sensors and data logging are inexpensive these days, a test that puts more knowledge within the reach of our budget would be great.</p><p>We’ll return in a short while to cover exactly how you can achieve that using a t-test, with examples in Python using a real data set from IoT sensors.</p></html>

From:  
<https://schnipsl.qgelm.de/> - Qgelm

Permanent link:  
[https://schnipsl.qgelm.de/doku.php?id=wallabag:statistics-and-hacking\\_-an-introduction-to-hypothesis-testing](https://schnipsl.qgelm.de/doku.php?id=wallabag:statistics-and-hacking_-an-introduction-to-hypothesis-testing)

Last update: 2021/12/06 15:24

