5

Power, Effect Size, *P*-Values, and Estimating Required Sample Size Using Python

CHAPTER OBJECTIVES

- Understand the nature of *p*-values and their proper use in scientific inference.
- Distinguish between a *p*-value and an effect size and understand why effect sizes are often more important than *p*-values.
- Understand the nature of statistical power and how degree of power relates one-to-one with *p*-values.
- Estimate statistical power and required sample size in *t*-tests to better appreciate how *p*-values, effect size, and sample size relate to one another.

In this chapter, we survey the concepts of **statistical power**, **effect size**, and ***p*-values**. Though some of this information we have already studied earlier in the book in one way or another, the current chapter seeks to unify and explain how these concepts "go together" and influence one another. For example, a researcher interpreting a *p*-value but not understanding statistical power is a researcher who should not be interpreting *p*-values! Yes, understanding how these elements relate to one another is that important! Likewise, the interpretation of effect size, without understanding the role of *p*-values, makes such interpretation incomplete and potentially misguided or misleading. For both the producer and consumer of research, it is essential that these concepts be clearly understood if one is to attempt to produce or interpret any scientific evidence at all. Hence, this chapter is mandatory reading for newcomers to statistics and data analysis, but also to experienced researchers who may not be familiar with how these elements relate to one another or who have never truly understood it. This chapter is extremely important for both such audiences.

5.1 What Determines the Size of a *P*-Value?

In order to understand statistical power and sample size, one must have an excellent understanding of what makes a *p*-value large or small. Otherwise, one will be inclined to potentially draw conclusions from "$p < 0.05$" that are not warranted and attribute

Applied Univariate, Bivariate, and Multivariate Statistics Using Python: A Beginner's Guide to Advanced Data Analysis, First Edition. Daniel J. Denis.

scientific worth to the result of an experiment where there may be none. The good thing about understanding how one inferential statistical test works is that you then have a good grounding in virtually how they all work, and hence, come to know why a *p*-value may be large in one case and small in another.

The easiest and most straightforward demonstration of the mechanics of *p*-values is with the simple univariate *t*-test for means. The principle can be demonstrated just as easily with a *z*-test for means, but since *t*-tests are most often used in research (i.e. we rarely know population variances and have to estimate them), we will use the *t*-test as a prototype for our example. In unpacking the components of the *t*-test, we learn what makes the size of *t*, and hence the size of the *p*-value, large or small. Recall the one-sample *t*-test for a mean:

$$t = \frac{\bar{y} - E(\bar{y})}{\hat{\sigma}_M} = \frac{\bar{y} - E(\bar{y})}{s/\sqrt{n}}$$

Recall that the numerator of the *t*-test, $\bar{y} - E(\bar{y})$, features a difference in means, that of the obtained sample mean minus the expectation of the mean under the null hypothesis. This expectation is given by $E(\bar{y})$, where E denotes the expectation operator. It can be easily shown that the expectation of the sample mean is equal to the population mean, that is, $E(\bar{y}) = \mu$:

$$\bar{y} = \frac{(y_1 + y_2 + \cdots + y_n)}{n}$$

$$E(\bar{y}) = E\left(\frac{(y_1 + y_2 + \cdots + y_n)}{n}\right)$$

It is well known in mathematical statistics that **the expectation of a sum of random variables is equal to the sum of individual expectations**. Given this, along with the above results, we can now write the expectation of the sample mean $\bar{y}$ as

$$E(\bar{y}) = \frac{E(y_1 + y_2 + \cdots + y_n)}{n}$$

$$= \frac{[E(y_1) + E(y_2) + \cdots + E(y_n)]}{n}$$

What is more, since the expectation of each of the individual values y_1 through y_n is $E(y_1) = \mu$, $E(y_2) = \mu$,..., $E(y_n) = \mu$, we can replace each y_1 through y_n with μ. Hence, we end up with

$$E(\bar{y}) = \frac{[\mu + \mu + \cdots + \mu]}{n}$$

$$E(\bar{y}) = \frac{n\mu}{n}$$

$$E(\bar{y}) = \mu$$

Since the n values cross out, we end up with simply μ. Hence, we have shown that $E(\bar{y}) = \mu$. So, what this means for our *t*-test is that the numerator $\bar{y} - E(\bar{y})$ can just as well be written as $\bar{y} - \mu$, yielding for the *t*-test

$$t = \frac{\bar{y} - E(\bar{y})}{\hat{\sigma}_M} = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

The numerator thus represents a difference between means, that of the sample from the population. When you think about it, this should be the reason why we are conducting the *t*-test, because we want to know whether our obtained sample mean differs from the population mean. Is this not the goal of your study or experiment? Now, of course, we already know it will differ to some degree. What are the odds that it doesn't? **Next to nil.** To understand this, consider obtaining the mean IQ of Californians on a sample size of 25. When we compare this to a population mean IQ of 100, we already know there will be at least **some** difference between the sample mean $\bar{y}$ and the population parameter μ. Hence, there being a difference is hardly in question. The true question, as you will recall from our discussion earlier in the book, is whether the difference in the sample is large enough to conclude a true population difference. To evaluate this, we need to compare it to the denominator of the test, which is the estimated standard error of the mean, $s/\sqrt{n}$. Notice what the standard error of the mean is made up of. It has two components. The first component is s, the sample standard deviation, while the second component is a function of sample size, specifically the square root of the sample size, $\sqrt{n}$. Hence, we see that the size of the resulting *t*-statistic will be determined by three things:

- The actual difference in means observed, that of $\bar{y} - \mu$.
- The size of the standard deviation s.
- The size of the sample taken for the test, given as the square root $\sqrt{n}$.

The reason you did the experiment or study is for the **first of these factors**, not the other two. That is, you surely did not conduct a statistical test to demonstrate that you can obtain a small standard deviation or obtain a large sample size. Though in some cases collecting data to demonstrate a low standard deviation or variance may be a worthwhile research pursuit, it is usually not the reason why scientists conduct studies. You presumably did the test because you were interested in the mean difference in the numerator, but it is clear that **the significance test is not simply a function of this difference**. This fact is extremely important! It is a function of the size of s and $\sqrt{n}$. And, as the size of t gets larger, it becomes an increasingly **unlikely statistic**, such that it lies off into the tail of the corresponding sampling distribution. Hence, it stands that statistical significance, the infamous "$p < 0.05$" can be achieved even for a constant mean difference $\bar{y} - \mu$ by simply increasing the sample size and/or finding a way to ensure s is small.

In most cases, decreasing s can be challenging, but it can be done. For example, if we were to test a treatment on those suffering from COVID-19 vs. those not suffering, we could minimize the variance in our samples by screening beforehand and selecting those from a given age or health background. That way, we are reducing inherent variability in our design and will be able to more easily detect a treatment effect if there is one to be found. In a moment, we will call this ability to detect a treatment difference that of statistical power. **Statistical power** will be the probability of detecting that difference or effect if it truly does exist in the population. Getting a handle on s can be quite difficult, and hence we are usually and most conveniently limited to increasing sample size to increase statistical power. We survey that possibility now.

5.2 How *P*-Values Are a Function of Sample Size

P-values can at times largely be a function of sample size. Be sure this is clear. It means that when you achieve "$p < 0.05$," it may, for all purposes, simply imply that you have collected a large sample. Hence, if you are in the habit of making scientific conclusions and implementing policy or clinical decisions based on the presence of *p*-values alone, you should be very concerned by this, as doing so essentially constitutes **a serious misunderstanding of what the *p*-value communicates**. The influence of sample size has been observed and known since the inception of null hypothesis significance testing, at least since Berkson (1938). The following is what he had to say on the subject. His comments were in relation to the chi-squared test, but they apply equally well to virtually any significance test:

> I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small. Having observed this, and on reflection, I make the following dogmatic statement, referring for illustration to the normal curve: "If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large – for instance, on the order of 200,000 – the chi-square P will be small beyond any usual limit of significance.
>
> (Berkson, 1938, p. 526)

With Berkson's words in mind, let us look a bit more closely at how this works in the *t*-test. For any difference however small in the numerator of the *t*-test, $\bar{y} - \mu$, increasing sample size will necessarily lead to a larger test statistic. This is not a hypothetical result, it is an arithmetic one (and based on good statistical theory to back it up). That is, the statistic **has** to increase in value for a constant difference $\bar{y} - \mu$ but an increasing sample size. For example, suppose the mean difference in IQ is equal to 1, with a standard deviation of 2. For a sample size of 4, our *t* computation comes out to be

$$t = \frac{\bar{y} - E(\bar{y})}{\hat{\sigma}_M} = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{101 - 100}{2/\sqrt{4}} = \frac{1}{1} = 1$$

Now, suppose we increase sample size to 16, leaving the rest of our *t* computation unchanged:

$$t = \frac{\bar{y} - E(\bar{y})}{\hat{\sigma}_M} = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{101 - 100}{2/\sqrt{16}} = \frac{1}{0.5} = 2$$

Notice that by a simple change in sample size, *t* has increased from 1 to a value of 2. For a sample size of 100, we have

$$t = \frac{\bar{y} - E(\bar{y})}{\hat{\sigma}_M} = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{101 - 100}{2/\sqrt{100}} = \frac{1}{0.2} = 5$$

We can see that, clearly, the value for *t* is increasing as a function of sample size for a constant mean difference. That is, notice that the mean difference in the numerator

has not changed at all. It has remained at a value of 1. Meanwhile, the value for t of 5 on a sample size of 100 is easily statistically significant, while for a sample size of 4, the statistic is not nearly as large. If you automatically associate scientific importance with statistical significance, then you can easily and mistakenly conclude a scientific effect for a mean difference of a single IQ point! If based on our discussion so far you are thinking, "**Wow, *p*-values, from a scientific point of view at least, may be quite meaningless**," then you are thinking correctly!

As mentioned, noting how the *p*-value is a function of sample size is not new. In addition to Berkson (1938), Bakan (1966) and other methodologists (e.g. see Cohen, 1990, among a myriad of other methodologists, especially in psychology) have long pointed out the problems with null hypothesis significance testing and how *p*-values can be made small as a function of sample size. In addition, since the numerator of the *t*-test in this situation will in practicality never equal 0, that is, $\bar{y} - \mu$ will always be different from zero to at least some decimal place, so long as the denominator $s/\sqrt{n}$ can be made small, **statistical significance is essentially assured**. So, what to do about this? The solution is not to abandon the *p*-value as some have advocated. The *p*-value is necessary to secure inferential support for the test statistic. The solution to the problem is to accompany all significance tests with a measure of **effect size**. **Effect sizes, not *p*-values, are what science is all about**. How does the size of effect influence the degree of statistical power? We consider that issue next.

P-values can quite easily be made almost entirely a function of sample size. Hence, whether you obtain a large or small p-value could depend extensively on whether or not you collected a small or large sample size. That is, statements such as "p < 0.05" do not necessarily reflect a meaningful scientific effect. It may simply indicate that a large-enough sample size was collected that resulted in a small-enough p-value to reject the null hypothesis. To make any sense of any scientific finding that might be present, a measure of effect size should accompany the reporting of p-values. Never, ever, interpret a p-value without an accompanying effect size when needing to draw a firm conclusion from a scientific report.

5.3 What is Effect Size?

Effect size is a general term that takes on different computations depending on the given statistical model or context. It either typically reports a "variance explained" figure or a mean difference of some kind. In most cases, effect sizes in one context can be quite easily translated to effect sizes in another via algebraic manipulation. That is, a Cohen's *d*, for instance, can be quite easily translated into an R-squared statistic, and so on.

In the context of our *t*-test, we can easily demonstrate how effect size is determined, and how this determination is not the same as for that of the *p*-value. Let us start from scratch and recall our *t*-test:

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

Again, as we have emphasized, as a scientist, what interests you most in the formula? You presumably did the study or experiment because you wanted to observe a

difference in means. That is, you did the study because you were hoping to see a distance in the numerator, $\bar{y} - \mu$. To use the IQ example once more, if the population mean IQ were equal to $\mu = 100$ and you treated a sample of individuals with a "magic pill" to see if it would make them smarter, yielding a sample mean $\bar{y} = 130$, it is the distance between means 130–100 = 30 that is of most interest, not the fact that you may have obtained a small standard error. After all, we just discussed that the standard error is essentially a measure of how much variance you have in your sample, along with the size of sample used. Hardly of great scientific interest! You did the study because you wanted to see a **difference in means**; hence, any measure of effect size should be primarily about this observed difference, $\bar{y} - \mu$. **It is not the size of the computed statistic that should dominate your scientific interests.** From a statistical point of view, yes, small standard errors are interesting, but in terms of the science, they are not priority. The difference in means in the numerator is what should be guiding your scientific interests.

However, if we simply report the distance of 30 as a measure of effect, it is incomplete. But why? Why is the distance between means not sufficient to quantify the size of effect in this case? It is incomplete because it does not take into account the baseline amount of variation in the population. What this means is that the distance of 30 may or may not be "impressive" depending on how much variability there is in the population, or at least an estimate of it. For example, if IQ scores have lots of inherent variability, then a distance of 30 is not quite as impressive as if the variability is quite low. With very low variability, a distance of 30 becomes much more impressive. Hence, we need a measure that puts the distance in means in some kind of **statistical context**. This is exactly what **Cohen's *d*** accomplishes. Cohen's d (Cohen, 1988), a measure of **statistical distance**, is given by

$$d = \left| \frac{\bar{y} - \mu_0}{\sigma} \right|$$

where $\bar{y}$ is our observed sample mean, μ_0 is the population mean under the null, and σ is the population standard deviation. If σ is not known, then s can be used to estimate it. The bars around the fraction mean to take the absolute value of the distance. Now, as a scientist, you naturally have a hypothesis as to the direction of the effect, that is, you are hoping for the IQ case, for instance, that $\bar{y}$ is larger than μ_0 since such would denote the group receiving the magic pill is doing better IQ-wise than what we would expect under the null. However, Cohen's d itself does not care much about the direction. And if the distance were reversed such that $\bar{y}$ were equal to 70 instead of 130, you would probably be interested in that result as well, even if it did not align with your theoretical prediction. Hence, this is the reason why we take the **absolute value** and thus ignore the **sign** of the result. That is, Cohen's d really does not care about your scientific interests, it simply wants to measure the magnitude difference in means.

What is the influence of σ on the size of effect? We can easily demonstrate this numerically. Using our previous example with a distance of 30, suppose σ were equal to 1. Then

$$d = \left| \frac{\bar{y} - \mu_0}{\sigma} \right| = \left| \frac{130 - 100}{1} \right| = 30$$

Of course, this would be a whopping effect! However, if σ were equal to 30, then

$$d = \left| \frac{\bar{y} - \mu_0}{\sigma} \right| = \left| \frac{130 - 100}{30} \right| = 1$$

Notice carefully that, in both cases, the actual distance in means in the numerator did not change. Only the population standard deviation changed and had a drastic effect on the size of d. This is because under the first scenario, the population standard deviation is much smaller; hence, when we observe a difference in means, it is less likely to occur simply due to natural variation in the population. When variability in the population is rather large, such as in the second scenario, then seeing rather large distances of the type $\bar{y} - \mu_0$ would be expected to be more likely. Thus, this is the reason why Cohen's d will be smaller in such situations.

5.4 Understanding Population Variability in the Context of Experimental Design

The size of σ can be understood as a general concept in scientific experimentation and is not restricted to simply being used in statistical formulae. Rather, it is a central feature of how you should be thinking about research studies, and, even more, how you should plan such research studies. Experiments in physics, for example, usually have quite low values of σ in their investigations. Why is that? It is because **physicists usually have a lot of control over the objects they study**. Molecular movement, for example, is a very precise area of investigation; there is not much variability to begin with regardless of the "treatment" imposed. Medical sciences, in some cases, may also operate in areas that have relatively low population variances. Whether a medication increases or decreases heart rate is likely to be tested on a pool of subjects with a relatively low variance in heart rates, at least practically speaking even due to the natural range of the number of possible beats generated in a given minute. Or, a researcher can purposely screen for participants who have a particular range of heartbeats in the attempt to **factor out** as much of the inherent variability in the data as possible before imposing a treatment. For instance, unless your heartbeat is in the range of 72 to 80, maybe you are disqualified from the study from the start. Only accepting participants within that range will likely afford the experiment more **statistical power** and **sensitivity**, since there is not as much "noise" to filter out when trying to discern the presence or absence of an effect. Then, if the treatment is working or has an effect on heart rate, it can be much more easily detected than if the variance in the population is quite high.

In many other areas of science, however, populations have inherently more variability, and hence our "numerator" in our effect size above will have to be large enough to compensate for this inherent variability to be "noticeable." By analogy, if you have a lot of noise in the water, such that the wind is generating impressive white caps on the lake (i.e. when the waves are so high it shows a white splash), it will take a much larger rock (effect) to generate a splash that is noticeable. When scientists measure people's IQ, for example, the range and variance can be quite high. There is also a lot of

measurement error built into things. **Your job as a scientist is to make the waters as calm as possible before skipping the rock to observe its effect**. If you are okay with stormy waters, then do not expect to see a noticeable splash from your rock, even if the rock is indeed making a splash. In physics, for instance, the waters are usually quite calm. In economics and psychology, on the other hand, often the waters are naturally stormy. Physics may be able to detect effects with very small samples. Psychology and economics? Usually not.

Hence, an experimenter would prefer population variance to be as small as possible, allowing him or her to more easily detect any effect that may be present. All else equal, this will more easily generate effects that are noticeable, and hence Cohen's *d* (or similar computable effect size) will become larger in value. Other than using screening tools and controls, one could also choose a **blocking design** in which nuisance variables are blocked in order to remove variability in the ensuing *t*-test, *F*-test, or what have you. We briefly explored block designs in our previous chapter, and will discuss it further in the following chapter. By blocking, we reduce MS error, allowing MS between to better shine through. **Any method by which within variability is reduced will allow between variability to shine.** This is a principle not only of statistics but also of science in general. If you do experiments (and not simply correlational designs), this concept of between vs. within should always be on your mind! Calm the waters, then skip the rock. If the rock makes a splash, you're bound to see it! If you do not calm the waters first, by way of analogy, you will not see whether the treatment for COVID-19 is working! Calm the waters as much as you can first!

5.5 Where Does Power Fit into All of This?

Having surveyed what makes a *p*-value small, as well as a measure of effect size, we are now well equipped to understand statistical power and how it fits in with all of these components. Recall our definition of power was it being the probability of rejecting the null hypothesis given that it is false. More power means bigger test statistics. Hence, whatever makes the *t* or *F* or whatever test statistic you are using large, you have more power. Given our discussion of *p*-values and effect size, we can now list the determinants of statistical power as it concerns the *t*-test:

- Size of distance in means $\bar{y} - \mu$. All else being equal, the larger the distance, the greater the statistical power.
- Size of σ. All else equal, the smaller the population variance, the greater the statistical power.
- Sample size, n. All else equal, the larger the sample size, the greater the statistical power.

Hence, through the *t*-test, we have seen what makes a test statistic large vs. small, translating to what makes a *p*-value large or small, which translates into the degree of power for the given experiment or study. We can easily see now why increasing sample size is usually, even if typically expensive, the easiest way to boost power. An increase in sample size has the effect of usually decreasing the standard error of the statistic. In

the limit (which, if you are not familiar with calculus, crudely means "in the long run" in this case), as $n \to \infty$, the standard error goes to 0, which means, pragmatically, that we now have the entire population. That is, increasing n generates a smaller and smaller $s/\sqrt{n}$, which implies whatever is going on in the numerator of our statistic, $\bar{y} - \mu$, begins to look quite large relative to the diminishing standard error. We also see from our discussion that statistical power is no mystery. It is a function of quantifiable things in our test statistic, and if you understand the determinants of it, *p*-values will forever lose their mystery and you will stop automatically and necessarily associating small *p*-values with anything of scientific import. A small *p*-value may hint toward a meaningful scientific result, but on its own it fails to reveal that much about what is going on scientifically.

Now, we said that increasing sample size usually has the effect of decreasing the standard error and leading to a smaller *p*-value. The qualifier "usually" should not be ignored. When increasing sample size, we in reality do not know what will happen to the variance of the sample, nor do we know what will happen to the distance between means (whether a one-sample test or two) in the numerator. If these figures change, as they may very well due to incorporating new information (sample size) into our design, then it may very well be that the resulting *p*-value actually increases in size with a limited and defined increase in sample size. However, this is extremely rare. In most cases, an increase in sample size overpowers any possible or plausible change in effect size or variance, such that we are almost always guaranteed to observe a smaller and smaller *p*-value. But, if ever you notice a *p*-value increase as you collect more subjects or objects, it may simply be because you are tapping more into the population, and the population is much more variable than your sample data had originally suggested. But in the long run, an increase in sample size will be overpowering to the *p*-value and will eventually drive it to be a very small value. That is, as sample size increases, you will at some point reject the null. As was the case for the *Titanic*, it sinking was a **mathematical certainty**.

5.6 Can You Have Too Much Power? Can a Sample Be Too Large?

Having discussed the determinants of power, it may at first seem to the newcomer that an increase in sample size is a "negative" event, since it would appear to artificially lower *p*-values and make rejecting null hypotheses all too "easy." After all, if Berkson, Bakan and others are correct (hint: *they are*) and *p*-values can be made largely a function of sample size, then maybe we should advise researchers to gather small samples so that the *p*-value is not artificially made small – or at least not collect "too big" of a sample size to accommodate the potential drawbacks of the *p*-value. This idea, while at first glance appears to be plausible, is entirely and utterly **misguided**. We will explain why this is the case in a moment, but first, let us state a fundamental principle of research to get us started:

You cannot have too large of a sample size for your experiment, nor can you have too much power. Period.

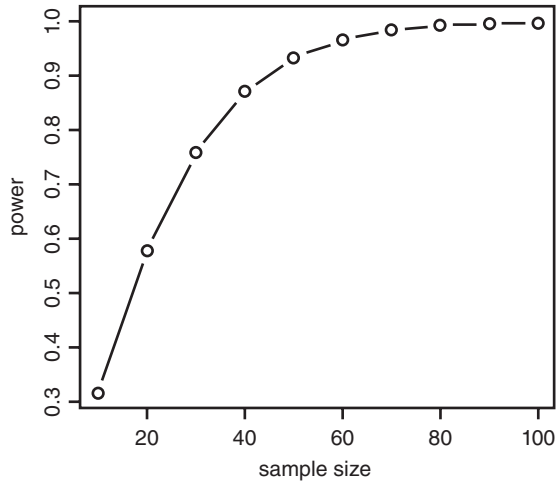
Take notice of the above! You cannot have "too big" a sample size! Recall that the goal of a research study is to learn of population parameters. If we had the population parameters, we would not be computing inferential statistics in the first place. That is, we would have no need to compute, for instance, sample means, and could just get on with computing population parameters and be done with it. No inference required! However, we know things usually do not work this way. Populations are usually quite large, if not infinite in size, such as in the case for the behavior of coins in which we can theoretically flip the coin an infinite number of times. Thus, if the goal is to learn of population parameters in the first place, then **this implies that collecting as large of a sample size as possible could never be "wrong," at least not in the sense of statistical inference.** However, what it does mean is that relying on the *p*-value as any indicator of scientific evidence is severely misguided, which is why coupling it with effect size measures is advised. But, to suggest we "appease" the *p*-value by collecting less than large sample sizes is nothing short of ridiculous.

Now, of course, as with all things, there are caveats. While you cannot theoretically go wrong with collecting an increasingly large sample size, **it may still not be worth your while**. **Enter pragmatics.** That is, thanks to inferential statistics, you may be able to learn just as much, or nearly as much, from a much smaller sample of information than collecting a much larger sample. There is a trade-off of resources you should be aware of, and there is a decreasing **rate of return** for collecting increasingly larger samples.

Let us unpack the above concept a bit with an example. Suppose again we are searching for a treatment for COVID-19 and choose to evaluate our new drug on a sample of 1,000 test subjects. Suppose our power for the experiment is quite high at 0.95. Power ranges from 0 to 1.0, and hence power equal to 0.95 is, in most cases, quite respectable. Now, we have said that you cannot have "too big" of a sample, right? While this may be true theoretically, **you can have too big of a sample size pragmatically**. In other words, while you can double your sample size to 2,000 subjects, such a doubling will probably not tell you that much more than your original 1,000 units. Even if power increased from 0.95 to, say, 0.99, the cost incurred, whether that cost be time, financial expense, or fatigue in collecting and using the larger sample size, may simply not be "worth it." Hence, sticking with the original 1,000 units is probably the most economical choice in giving you the most "bang for your buck" in terms of the scientific experiment. It would be "better" to have the 2,000 subjects, but simply not worth the investment given the rate of return. In other words, you can conclude just about as much from the 1,000 subjects as you can from the 2,000.

The above principle can be easily demonstrated by a survey of **power curves**. Below is an example of a power curve where power is given as a function of sample size. Notice that as sample size increases, the rate of return on power begins to plateau. The shape of the curve is **logarithmic**, which for this case implies that for a sample size of 60 or so, we have reached sufficient power. Collecting more of a sample size will not benefit us much in terms of the inferences we wish to make. To be blunt, collecting a sample size of 500, for example, would be a waste of resources. We can do just as well in terms of our statistical inferences with far less of a sample size. Of course, how we collect the sample is extremely important (e.g.

usually random sampling is advisable), but the point is that the rate of return on collecting larger and larger sample sizes is usually quite minimal past a certain point.



You cannot have "too much" power or sample size. However, at some point, at a practical level, increasing power levels and sample size simply is not worth the investment. That is, the rate of return for seeking higher and higher levels of power is not usually worth the cost of collecting a larger sample, whether those costs be financial or other. Increasing sample size beyond power levels of approximately 0.95 usually affords little benefit in most cases and will incur more expense in the form of time, resources, etc., than is necessary. Aim for a healthy sample size and high power, but there is little need to waste time or resources collecting increasingly larger sample sizes.

5.7 Demonstrating Power Principles in Python: Estimating Power or Sample Size

We can easily demonstrate the above-mentioned principles in Python, showing how sample size and effect size help to determine the degree of statistical power. The question often arises as to whether you should estimate power or sample size. It really does not matter which you estimate, since, as we have seen, one is a function of the other. However, in planning an experiment, it is typical of the researcher to designate his or her degree of preferred power and then learn how much of a sample size will be required to achieve this. Hence, the researcher may set power at, say 0.90, estimate the given effect size, and then solve for sample size. In other cases, the researcher may be restricted by sample size and would like to learn of the degree of statistical power afforded by the fixed sample size. Either way will give you the same estimates. Sometimes researchers have fixed sample sizes and cannot collect more. In such cases, they may simply want to know their chances at rejecting the null and whether the

experiment or study is worth doing at all beyond a **pilot study** (which is a study done on a very small number of subjects, presumably to inform the decision to pursue or not pursue a larger-scale study).

As an example, for an independent samples *t*-test, we can use **statsmodels** to provide us with a sample size estimate. For this example, we set **effect size** at 0.8, **alpha** at 0.05, and desire a level of **power** equal to 0.8:

```
from statsmodels.stats.power import TTestIndPower
effect = 0.8
alpha = 0.05
power = 0.8
analysis = TTestIndPower()
result = analysis.solve_power(effect, power=power, nobs1=None,
ratio=1.0, alpha=alpha)
print('Sample Size: %.3f' % result)
```

```
Sample Size: 25.525
```

We see that the estimated sample size is equal to approximately **25 subjects per group**. To be on the safe side, it is always wise to round sample size estimates **upward** to ensure sufficient power. That is, the estimate is for 25.52, but it does not "hurt" to round up in this case. The worst thing that can happen is that we end up having slightly more sample size than we actually need. If we round down, however, we risk having insufficient sample size. Now, in this case, the fraction of 0.525 will make little difference either way, but in principle, **always round up**. The `ratio = 1.0` designates that we have equal numbers of participants in each sample. Though this may change over the course of data collection, it is usually wise to start off assuming equal sample sizes per group.

Now, suppose we adjusted the requested power to 0.9 instead of 0.8. If you are understanding power correctly, this should imply that for such an increase in power we require a greater sample size. Recall that since increasing sample size is one essentially sure way of boosting power, this is what we would expect. Indeed, this is what we find:

```
from statsmodels.stats.power import TTestIndPower
effect = 0.8
alpha = 0.05
power = 0.9
analysis = TTestIndPower()
result = analysis.solve_power(effect, power=power, nobs1=None,
ratio=1.0, alpha=alpha)
print('Sample Size: %.3f' % result)
```

```
Sample Size: 33.826
```

We can see that as our demand for power increases, this is associated with a likewise increase in sample size. In English, what this means is that if you want more of a chance to reject the null, that is, greater sensitivity, you are going to need to collect more subjects.

5.8 Demonstrating the Influence of Effect Size

We can also easily demonstrate the influence of effect size on sample size requirements. For instance, if we hypothesized a much smaller effect (0.2), this should entail requiring a much **larger sample size** to detect it. We keep power set at 0.8, as well as the significance level at 0.05. Let us see what happens to estimated sample size given these parameters:

```
from statsmodels.stats.power import TTestIndPower
effect = 0.2
alpha = 0.05
power = 0.8
analysis = TTestIndPower()
result = analysis.solve_power(effect, power=power, nobs1=None,
ratio=1.0, alpha=alpha)
print('Sample Size: %.3f' % result)
```

```
Sample Size: 393.406
```

We can see that as the effect size drops to 0.2, the required sample size increases to a whopping approximately 393! But why should this make sense? A drop in effect size from 0.8 to 0.2 implies that we are attempting to detect something that is much smaller than before. As an analogy, when the eye doctor asks you to read off the very small tiny letters at the optical exam, you will need a stronger lens to detect it:

A B C

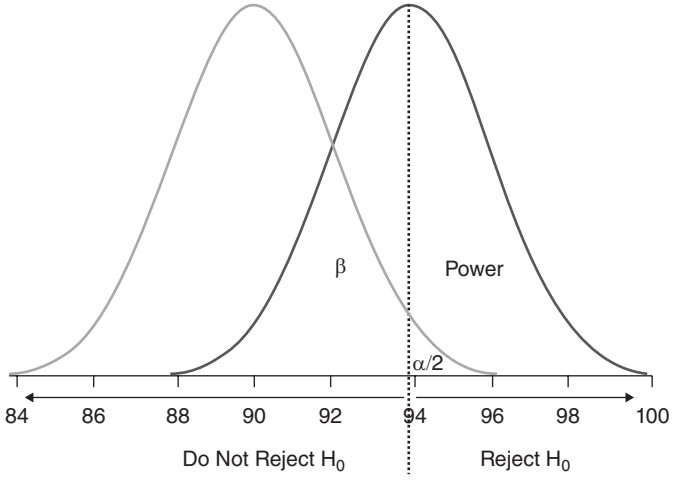
A B C

In the above, the smaller letters require greater sensitivity than the larger letters, since the effect is much smaller in the first case than in the second. That is, **we need a more powerful test to detect the smaller effect than the large**. The tiny letters represent a measure of effect, in which in this case the effect is very small. Hence, you need a stronger "prescription" in order to detect it. If the letters are extremely small, you might even need a "powerful" set of binoculars. Understand the analogy? If the letters were large (i.e. a huge effect), you could probably identify them with no prescription at all. If the Yellowstone bison is directly in front of you (well, run!), you do not need binoculars to see him. You have well sufficient power because the effect (size and proximity of the bison) is quite large! Hence, **the smaller the effect size, all else equal, the greater sensitivity of the test required**. That is, all else being equal, you require more subjects to see something that is hardly there in the first place.

5.9 The Influence of Significance Levels on Statistical Power

We have demonstrated how changing the sample size or altering the anticipated effect size can have a significant impact on power. Conversely, we have seen how adjusting the power level and effect necessitates differing numbers of subjects.

But what about the significance level? A change in significance level will also have a direct influence on statistical power since it necessarily changes the type I error rate. We can easily see why this must be true by studying power curves for distributions, as shown below:



Notice that if we were to move the criterion decision line (the vertical dotted line dividing β, the type II error rate, and power) over toward the left (imagine pushing the line leftward), thereby increasing the type I error rate (i.e. α / 2 in the curve), the type II error rate would decrease, and, consequently, since power is the balance of this error rate, power would increase. Hence, if you increase the type I error rate from, say, 0.05 to 0.10, you necessarily increase statistical power. We can easily demonstrate this by using our earlier example in Python. In what follows, we adjust the significance level from 0.05 to 0.20, while maintaining the effect size and power equal to 0.8, respectively:

```
from statsmodels.stats.power import TTestIndPower
effect = 0.8
alpha = 0.20
power = 0.8
analysis = TTestIndPower()
result = analysis.solve_power(effect, power=power, nobs1=None,
ratio=1.0, alpha=alpha)
print('Sample Size: %.3f' % result)
```

```
Sample Size: 14.515
```

We can see that keeping other parameters the same as in our earlier example (where we estimated the number of subjects to equal approximately 25 per group), but adjusting the significance level from 0.05 to 0.20, the required sample size is now only 14 or so, down from 25. But if we just said that this should give us more power, why is power still equal to 0.80? It is equal to 0.80 because we set it **a priori** as such, in other words

it was **fixed** at that level. Given that it was fixed there, and the effect at 0.8, there is only one other way to accommodate the new alpha level, and that is via the estimated sample size. This is why the sample size decreased to 14.5.

5.10 What About Power and Hypothesis Testing in the Age of "Big Data"?

With all of the hype around "big data" these days, it is a reasonable question to ask where statistical power fits into this discussion. Big data is a term that connotes a few different things, including the employment of **data engineering** on extremely large data sets. Hence, as the name suggests, when you are working with big data, it implies from a statistical inference perspective that you have sufficient power to make inferences toward populations. However, make no mistake, **inferences are still required**, even if they are not **formalized**. Big data does not somehow make statistical inference and hypothesis testing obsolete in any way, it simply means that we are working with such large data sets that we will usually have sufficient data to make very good estimates and guesses as to the nature of population parameters. Hence, if you have such large data, then calculating statistical power will usually be a waste of time since you will usually have enough power to reject virtually any null hypothesis you put up. The focus must then be on **effect size** and describing or quantifying the degree or extent of variance explained in the response variable or variables you are working with.

For example, when studying data in the COVID-19 pandemic, even if entire populations could not be studied because the data were still coming in day by day, most models had sufficient power or sensitivity since they were based on relatively large sample sizes. Hence, in studies where sample size is easy to come by, such as in big data or areas where samples are extremely large, power is still a concern **implicitly**, but it does not rear its head as much because sample sizes are usually well beyond sufficient. Where power is a much more pressing issue is in **well-calibrated designed experiments**, where the scientist is purposely planning a demonstration to show whether a treatment works or does not. The idea of power as part of null hypothesis significance testing originally and historically arose in the context of performing rather precise experiments and quality control designs, and so when you are building a study from the ground up, you should definitely be paying close attention to statistical power. Most samples in such designs will be potentially difficult and expensive to recruit and hence you definitely do not want to waste resources on collecting more subjects than you may need if indeed collecting subjects is an expensive endeavor. For a big data project in which COVID-19 cases are analyzed on world populations, statistical inference still occurs, but is more **implicit**. Hence, power concerns are likewise more implicit as well. They still exist, however, and inferences are still taking place, again, even if implicitly and we do not bring up *p*-values into the discussion.

What about hypothesis testing? Does hypothesis testing somehow vanish simply because one is using exceedingly large data sets? Absolutely not! Now, from a statistical point of view, yes, rejecting the null hypothesis when you have hundreds of thousands of observations is quite the meaningless result. With such large sample sizes, as we have discussed, null hypotheses will be jettisoned left and right. However, **hypothesis testing is not merely a statistical exercise. It is a scientific one**. Hence, when you reject the null, you are also simultaneously attempting to infer a suitable

alternative hypothesis. The principles of hypothesis testing still remain, the only difference being that little if any focus will be on *p*-values. Most of the focus, as it should be, will be on **effect sizes** or other measures of pragmatic description (such as **prevalence** or **incidence** of disease, etc.). But, rest assured, the principles of hypothesis testing still remain, even in the age of big data.

Statistical power and hypothesis testing are both just as relevant in the age of big data as they were before. The only difference is that since some data sets are so large, usually statistical power is nothing more than an afterthought since it is almost a virtual guarantee that enough participants or objects are being subjected to analysis. And while rejecting a null hypothesis in such situations may no longer be meaningful from a statistical perspective, the idea of detecting a good alternative hypothesis is just as relevant. Hence, big data changes things in terms of software engineering and other facets of research, such as data recruitment, but essential statistical principles are still at play. Even when algorithmic approaches dominate, essential principles of basic statistics and research remain, that of collecting a sample, studying it, then making inferences toward the population.

5.11 Concluding Comments on Power, Effect Size, and Significance Testing

This is as far as we take our discussion of statistical power, sample size, and how these relate to effect size and significance testing. The chapter was only meant as a brief overview and summary of these issues, with a few demonstrations in Python. Short as this chapter may be, it is essential that you clearly understand and appreciate the issues at play if you are at all to interpret *p*-values, significance tests, and effect sizes. Numerous times, the author has witnessed even experienced and well-published researchers and scientists demonstrate a complete and utter lack of understanding of how these concepts interplay. Such a misunderstanding leads one to make or draw substantive conclusions from research papers that simply do not exist. If you do not understand the contents of this chapter, you should not be interpreting statistical evidence. That is not an exaggeration or an idealistic remark. These fundamental issues are **that** important to grasp. The most common misunderstanding and errors of interpretation are usually the following, and are the ones you should be acutely aware of so that you do not make them:

- Assuming that since the experimental report reads "$p < 0.05$" (or $p < 0.01$, etc.), that somehow this translates into a meaningful scientific result. This is not the case. Without a meaningful effect size measure, "$p < 0.05$" may, on a scientific level, be quite meaningless.
- Not appreciating or understanding that statistical power is intimately related to the size of the *p*-value, and that insufficient power makes rejecting a null hypothesis virtually impossible, even if the effect size is quite large. If you are interpreting *p*-values, you need to be acutely aware of how they relate to statistical power. Otherwise, you will likely draw conclusions from your research that are not warranted.
- Failing to appreciate how *p*-values can largely be a function of sample size, but this is generally not the case for effect sizes. While increasing the sample size to large numbers virtually almost guarantees a rejection of the null hypothesis, the effect size may increase or decrease depending on what the new data have to say.

- Believing that an increase in sample size is always a good idea, in that if you double sample size, for instance, it will result in an equivalent doubling of statistical power or otherwise more sensitivity to detect the alternative hypothesis under consideration. This misunderstanding fails to understand and appreciate the diminishing returns of increasing sample size in relation to power. For a given effect size, at some point, increasing sample size will not afford that much more power, and hence if the sample size is not cheap, then doing so would be a waste of resources.

If any of the principles of this chapter are not clear, you are encouraged to consult good introductory textbooks on statistics targeted toward scientists that further elucidate and explore these topics. Hays (1994) is an excellent reference for all matters discussed in this chapter.

Review Exercises

1. Cite and discuss the determinants of the ***p*-value**. What are the factors that make a *p*-value large or small?
2. Why does the statement "$\boldsymbol{p < 0.05}$" not necessarily imply a scientifically meaningful result? Explain.
3. Why is it the case that for **increasing sample size**, this might not result in a rejection of the null hypothesis? Explain.
4. How does reducing **population variance** contribute to increasing statistical power? Try to explain by highlighting what it means to conduct a quality experiment.
5. Someone says to you that having **too much statistical power** for an experiment is a bad thing. Respond and explain, and correct if necessary.
6. Your boss says to you, "Let's collect an additional 2,000 participants for our study over and above the 2,000 we already have." What kinds of concerns might you have over this plan? Respond to your boss with information that he/she may want to consider further.
7. How does an **increase in sample size** virtually guarantee a rejection of the null hypothesis, yet not necessarily a small or large effect size?
8. Why is the **distance** between the sample mean and the population mean insufficient on its own as a measure of effect size? That is, why is it necessary to divide by the population standard deviation to better contextualize this distance? Explain.
9. Suppose a medical researcher says, "This COVID-19 vaccine has worked successfully on 10,000 individuals we tested it on." Explain why it may not necessarily work on the 10,001st individual using what you know about the principles of statistical inference.
10. Demonstrate using **power curves** for two distributions why decreasing the type II error rate, all else being equal, has the effect of increasing statistical power. Further, can you actively and deliberately decrease type II error in a real research context given a fixed significance level? Why or why not?
11. Using Python, for a one-sample *t*-test, demonstrate the computation for power for an effect equal to 0.0. Does this result surprise you? Why or why not?
12. Estimate sample size in Python for a two-sample *t*-test with both a very small effect size and then a rather large one. Note the difference you observed in required sample sizes. Does this surprise you at all? Why or why not?