Effect Sizes: Primer

Wow! That had a huge effect!

Dr. Gordon Wright

Power

When there is a true effect out there to measure, you want to make sure your design is sensitive enough to detect the effect, otherwise what's the point. We've already talked about the idea that an effect can have different sizes. The next idea is that your design can be more less sensitive in its ability to reliabily measure the effect. We have discussed this general idea many times already in the textbook, for example we know that we will be more likely to detect "significant" effects (when there are real differences) when we increase our sample-size. Here, we will talk about the idea of design sensitivity in terms of the concept of power. Interestingly, the concept of power is a somewhat limited concept, in that it only exists as a concept within some philosophies of statistics.

A digresssion about hypothesis testing

In particular, the concept of power falls out of the Neyman-Pearson concept of null vs. alternative hypothesis testing. Up to this point, we have largely avoided this terminology. This is perhaps a disservice in that the Neyman-Pearson ideas are by now the most common and widespread, and in the opinion of some of us, they are also the most widely misunderstood and abused idea, which is why we have avoided these ideas until now.

What we have been mainly doing is talking about hypothesis testing from the Fisherian (Sir Ronald Fisher, the ANOVA guy) perspective. This is a basic perspective that we think can't be easily ignored. It is also quite limited. The basic idea is this:

1. We know that chance can cause some differences when we measure something between experimental conditions.
2. We want to rule out the possibility that the difference that we observed can not be due to chance

3. We construct large N designs that permit us to do this when a real effect is observed, such that we can confidently say that big differences that we find are so big (well outside the chance window) that it is highly implausible that chance alone could have produced.
4. The final conclusion is that chance was extremely unlikely to have produced the differences. We then infer that something else, like the manipulation, must have caused the difference.
5. We don't say anything else about the something else.
6. We either reject the null distribution as an explanation (that chance couldn't have done it), or retain the null (admit that chance could have done it, and if it did we couldn't tell the difference between what we found and what chance could do)

Neyman and Pearson introduced one more idea to this mix, the idea of an alternative hypothesis. The alternative hypothesis is the idea that if there is a true effect, then the data sampled into each condition of the experiment must have come from two different distributions. Remember, when there is no effect we assume all of the data cam from the same distribution (which by definition can't produce true differences in the long run, because all of the numbers are coming from the same distribution). The graphs of effect-sizes from before show examples of these alternative distributions, with samples for condition A coming from one distribution, and samples from condition B coming from a shifted distribution with a different mean.

So, under the Neyman-Pearson tradition, when a researcher find a signifcant effect they do more than one things. First, they reject the null-hypothesis of no differences, and they accept the alternative hypothesis that there was differences. This seems like a sensible thing to do. And, because the researcher is actually interested in the properties of the real effect, they might be interested in learning more about the actual alternative hypothesis, that is they might want to know if their data come from two different distributions that were separated by some amount...in other words, they would want to know the size of the effect that they were measuring.

Back to power

We have now discussed enough ideas to formalize the concept of statistical power. For this concept to exist we need to do a couple things.

1. Agree to set an alpha criterion. When the p-value for our test-statistic is below this value we will call our finding statistically significant, and agree to reject the null hypothesis and accept the "alternative" hypothesis (sidenote, usually it isn't very clear which specific alternative hypothesis was accepted)
2. In advance of conducting the study, figure out what kinds of effect-sizes our design is capable of detecting with particular probabilites.

The power of a study is determined by the relationship between

1. The sample-size of the study
2. The effect-size of the manipulation
3. The alpha value set by the researcher.

To see this in practice let's do a simulation. We will do a t-test on a between-groups design 10 subjects in each group. Group A will be a control group with scores sampled from a normal distribution with mean of 10, and standard deviation of 5. Group B will be a treatment group, we will say the treatment has an effect-size of Cohen's d = .5, that's a standard deviation shift of .5, so the scores with come from a normal distribution with mean =12.5 and standard deivation of 5. Remember 1 standard deviation here is 5, so half of a standard deviation is 2.5.

The following R script runs this simulated experiment 1000 times. We set the alpha criterion to .05, this means we will reject the null whenever the p-value is less than .05. With this specific design, how many times out of of 1000 do we reject the null, and accept the alternative hypothesis?

```
p<-length(1000)
for(i in 1:1000){
  A<-rnorm(10,10,5)
  B<-rnorm(10,12.5,5)
  p[i]<-t.test(A,B,var.equal = TRUE)$p.value
}

length(p[p<.05])
```

```
[1] 186
```

The answer is that we reject the null, and accept the alternative 186 times out of 1000. In other words our experiment succesfully accepts the alternative hypothesis 18.6 percent of the time, this is known as the power of the study. Power is the probability that a design will succesfully detect an effect of a specific size.

Importantly, power is completely abstract idea that is completely determined by many assumptions including N, effect-size, and alpha. As a result, it is best not to think of power as a single number, but instead as a family of numbers.

For example, power is different when we change N. If we increase N, our samples will more precisely estimate the true distributions that they came from. Increasing N reduces sampling error, and shrinks the range of differences that can be produced by chance. Lets' increase our N in this simulation from 10 to 20 in each group and see what happens.

```
p<-length(1000)
for(i in 1:1000){
  A<-rnorm(20,10,5)
  B<-rnorm(20,12.5,5)
  p[i]<-t.test(A,B,var.equal = TRUE)$p.value
}

length(p[p<.05])
```

```
[1] 371
```

Now the number of significant experiments i 371 out of 1000, or a power of 37.1 percent. That's roughly doubled from before. We have made the design more sensitive to the effect by increasing N.

We can change the power of the design by changing the alpha-value, which tells us how much evidence we need to reject the null. For example, if we set the alpha criterion to 0.01, then we will be more conservative, only rejecting the null when chance can produce the observed difference 1% of the time. In our example, this will have the effect of reducing power. Let's keep N at 20, but reduce the alpha to 0.01 and see what happens:

```
p<-length(1000)
for(i in 1:1000){
  A<-rnorm(20,10,5)
  B<-rnorm(20,12.5,5)
  p[i]<-t.test(A,B,var.equal = TRUE)$p.value
}

length(p[p<.01])
```

```
[1] 142
```

Now only 142 out of 1000 experiments are significant, that's 14.2 power.

Finally, the power of the design depends on the actual size of the effect caused by the manipulation. In our example, we hypothesized that the effect caused a shift of .5 standard deviations. What if the effect causes a bigger shift? Say, a shift of 2 standard deviations. Let's keep N= 20, and alpha < .01, but change the effect-size to two standard deviations. When the effect in the real-world is bigger, it should be easier to measure, so our power will increase.

```
p<-length(1000)
for(i in 1:1000){
  A<-rnorm(20,10,5)
  B<-rnorm(20,30,5)
  p[i]<-t.test(A,B,var.equal = TRUE)$p.value
}

length(p[p<.01])
```

```
[1] 1000
```

Neat, if the effect-size is actually huge (2 standard deviation shift), then we have power 100 percent to detect the true effect.

Power curves

We mentioned that it is best to think of power as a family of numbers, rather than as a single number. To elaborate on this consider the power curve below. This is the power curve for a specific design: a between groups experiments with two levels, that uses an independent samples t-test to test whether an observed difference is due to chance. Critically, N is set to 10 in each group, and alpha is set to .05

In Figure 1 power (as a proportion, not a percentage) is plotted on the y-axis, and effect-size (Cohen's d) in standard deviation units is plotted on the x-axis.

```
power<-c()
for(i in seq(0,2,.1)){
sd_AB <- 1
n<-10
C <- qnorm(0.975)
se <- sqrt( sd_AB/n + sd_AB/n )
delta<-i
power <- c(power,1-pnorm(C-delta/se) + pnorm(-C-delta/se))
}

plot_df<-data.frame(power,
                    effect_size = seq(0,2,.1))

ggplot(plot_df, aes(x=effect_size, y=power))+
  geom_line()+
  theme_classic()
```

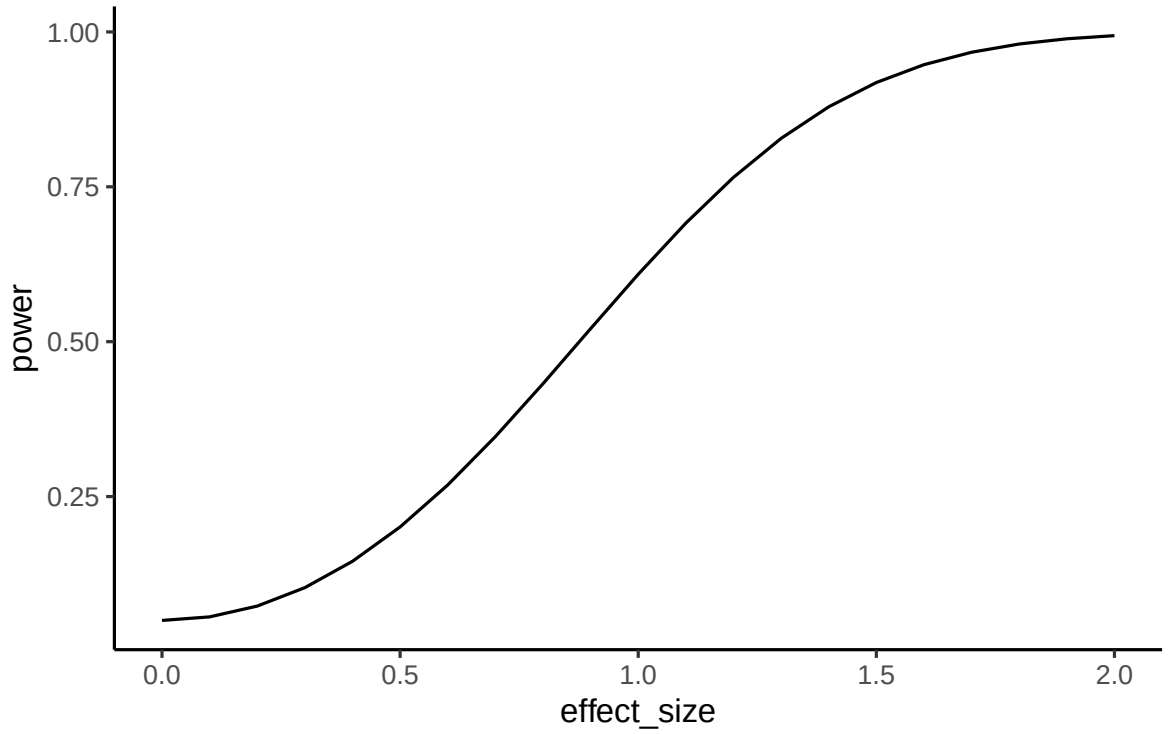

Figure 1: This figure shows power as a function of effect-size (Cohen's d) for a between-subjects independent samples t-test, with N=10, and alpha criterion 0.05.

A power curve like this one is very helpful to understand the sensitivity of a particular design. For example, we can see that a between subjects design with N=10 in both groups, will detect an effect of d=.5 (half a standard deviation shift) about 20% of the time, will detect an effect of d=.8 about 50% of the time, and will detect an effect of d=2 about 100% of the time. All of the percentages reflect the power of the design, which is the percentage of times the design would be expected to find a $p < 0.05$.

Let's imagine that based on prior research, the effect you are interested in measuring is fairly small, d=0.2. If you want to run an experiment that will detect an effect of this size a large percentage of the time, how many subjects do you need to have in each group? We know from the above graph that with N=10, power is very low to detect an effect of d=0.2. Let's make Figure 2 and vary the number of subjects rather than the size of the effect.

```
power<-c()
for(i in seq(10,800,10)){
sd_AB <- 1
n<-i
```

```
C <- qnorm(0.975)
se <- sqrt( sd_AB/n + sd_AB/n )
delta<-0.2
power <- c(power,1-pnorm(C-delta/se) + pnorm(-C-delta/se))
}

plot_df<-data.frame(power,
                    N = seq(10,800,10))

ggplot(plot_df, aes(x=N, y=power))+
  geom_line()+
  theme_classic()+
  geom_hline(yintercept=.8, color="green")
```

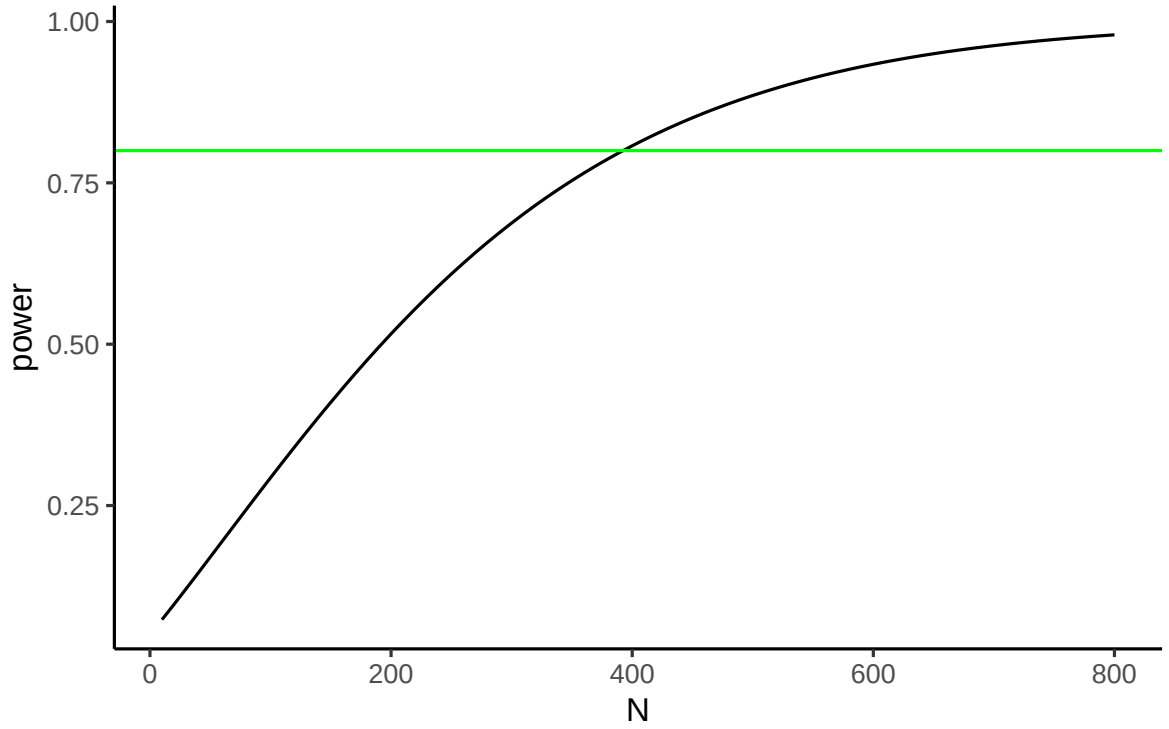

Figure 2: This figure shows power as a function of N for a between-subjects independent samples t-test, with d=0.2, and alpha criterion 0.05.

The figure plots power to detect an effect of d=0.2, as a function of N. The green line shows where power = .8, or 80%. It looks like we would nee about 380 subjects in each group to measure an effect of d=0.2, with power = .8. This means that 80% of our experiments would

succesfully show p < 0.05. Often times power of 80% is recommended as a reasonable level of power, however even when your design has power = 80%, your experiment will still fail to find an effect (associated with that level of power) 20% of the time!

Planning your design

Our discussion of effect size and power highlight the importance of the understanding the statistical limitations of an experimental design. In particular, we have seen the relationship between:

1. Sample-size
2. Effect-size
3. Alpha criterion
4. Power

As a general rule of thumb, small N designs can only reliably detect very large effects, whereas large N designs can reliably detect much smaller effects. As a researcher, it is your responsibility to plan your design accordingly so that it is capable of reliably detecting the kinds of effects it is intended to measure.

Some considerations

Low powered studies

Consider the following case. A researcher runs a study to detect an effect of interest. There is good reason, from prior research, to believe the effect-size is d=0.5. The researcher uses a design that has 30% power to detect the effect. They run the experiment and find a significant p-value, (p<.05). They conclude their manipulation worked, because it was unlikely that their result could have been caused by chance. How would you interpret the results of a study like this? Would you agree with thte researchers that the manipulation likely caused the difference? Would you be skeptical of the result?

The situation above requires thinking about two kinds of probabilities. On the one hand we know that the result observed by the researchers does not occur often by chance (p is less than 0.05). At the same time, we know that the design was underpowered, it only detects results of the expected size 30% of the time. We are face with wondering what kind of luck was driving the difference. The researchers could have gotten unlucky, and the difference really could be due to chance. In this case, they would be making a type I error (saying the result is real when it isn't). If the result was not due to chance, then they would also be lucky, as their design only detects this effect 30% of the time.

Perhaps another way to look at this situation is in terms of the replicability of the result. Replicability refers to whether or not the findings of the study would be the same if the experiment was repeated. Because we know that power is low here (only 30%), we would expect that most replications of this experiment would not find a significant effect. Instead, the experiment would be expected to replicate only 30% of the time.

Large N and small effects

Perhaps you have noticed that there is an intriguiing relationship between N (sample-size) and power and effect-size. As N increases, so does power to detect an effect of a particular size. Additionally, as N increases, a design is capable of detecting smaller and smaller effects with greater and greater power. For example, if N was large enough, we would have high power to detect very small effects, say d= 0.01, or even d=0.001. Let's think about what this means.

Imagine a drug company told you that they ran an experiment with 1 billion people to test whether their drug causes a significant change in headache pain. Let's say they found a significant effect (with power =100%), but the effect was very small, it turns out the drug reduces headache pain by less than 1%, let's say 0.01%. For our imaginary study we will also assume that this effect is very real, and not caused by chance.

Clearly the design had enough power to detect the effect, and the effect was there, so the design did detect the effect. However, the issue is that there is little practical value to this effect. Nobody is going to by a drug to reduce their headache pain by 0.01%, even if it was "scientifcally proven" to work. This example brings up two issues. First, increasing N to very large levels will allow designs to detect almost any effect (even very tiny ones) with very high power. Second, sometimes effects are meaningless when they are very small, especially in applied research such as drug studies.

These two issues can lead to interesting suggestions. For example, someone might claim that large N studies aren't very useful, because they can always detect really tiny effects that are practically meaningless. On the other hand, large N studies will also detect larger effects too, and they will give a better estimate of the "true" effect in the population (because we know that larger samples do a better job of estimating population parameters). Additionally, although really small effects are often not interesting in the context of applied research, they can be very important in theoretical research. For example, one theory might predict that manipulating X should have no effect, but another theory might predict that X does have an effect, even if it is a small one. So, detecting a small effect can have theoretical implication that can help rule out false theories. Generally speaking, researchers asking both theoretical and applied questions should think about and establish guidelines for "meaningful" effect-sizes so that they can run designs of appropriate size to detect effects of "meaningful size".

Small N and Large effects

All other things being equal would you trust the results from a study with small N or large N? This isn't a trick question, but sometimes people tie themselves into a knot trying to answer it. We already know that large sample-sizes provide better estimates of the distributions the samples come from. As a result, we can safely conclude that we should trust the data from large N studies more than small N studies.

At the same time, you might try to convince yourself otherwise. For example, you know that large N studies can detect very small effects that are practically and possibly even theoretically meaningless. You also know that that small N studies are only capable of reliably detecting very large effects. So, you might reason that a small N study is better than a large N study because if a small N study detects an effect, that effect must be big and meaningful; whereas, a large N study could easily detect an effect that is tiny and meaningless.

This line of thinking needs some improvement. First, just because a large N study can detect small effects, doesn't mean that it only detects small effects. If the effect is large, a large N study will easily detect it. Large N studies have the power to detect a much wider range of effects, from small to large. Second, just because a small N study detected an effect, does not mean that the effect is real, or that the effect is large. For example, small N studies have more variability, so the estimate of the effect size will have more error. Also, there is 5% (or alpha rate) chance that the effect was spurious. Interestingly, there is a pernicious relationship between effect-size and type I error rate

Type I errors are convincing when N is small

So what is this pernicious relationship between Type I errors and effect-size? Mainly, this relationship is pernicious for small N studies. For example, the following figure illustrates the results of 1000s of simulated experiments, all assuming the null distribution. In other words, for all of these simulations there is no true effect, as the numbers are all sampled from an identical distribution (normal distribution with mean =0, and standard deviation =1). The true effect-size is 0 in all cases.

We know that under the null, researchers will find p values that are less 5% about 5% of the time, remember that is the definition. So, if a researcher happened to be in this situation (where there manipulation did absolutely nothing), they would make a type I error 5% of the time, or if they conducted 100 experiments, they would expect to find a significant result for 5 of them.

Figure 3 reports the findings from only the type I errors, where the simulated study did produce $p < 0.05$. For each type I error, we calculated the exact p-value, as well as the effect-size (cohen's D) (mean difference divided by standard deviation). We already know that the true effect-size is zero, however take a look at this graph, and pay close attention to the smaller sample-sizes.

```r
all_df<-data.frame()
for(i in 1:1000){
  for(n in c(10,20,50,100,1000)){
    some_data<-rnorm(n,0,1)
    p_value<-t.test(some_data)$p.value
    effect_size<-mean(some_data)/sd(some_data)
    mean_scores<-mean(some_data)
    standard_error<-sd(some_data)/sqrt(length(some_data))
    t_df<-data.frame(sim=i,sample_size=n,p_value,effect_size,mean_scores,standard_e
    all_df<-rbind(all_df,t_df)
  }
}

type_I_error <-all_df[all_df$p_value<.05,]
type_I_error$sample_size<-as.factor(type_I_error$sample_size)

ggplot(type_I_error,aes(x=p_value,y=effect_size, group=sample_size,color=sample_siz
  geom_point()+
  theme_classic()+
  ggtitle("Effect sizes for type I errors")
```

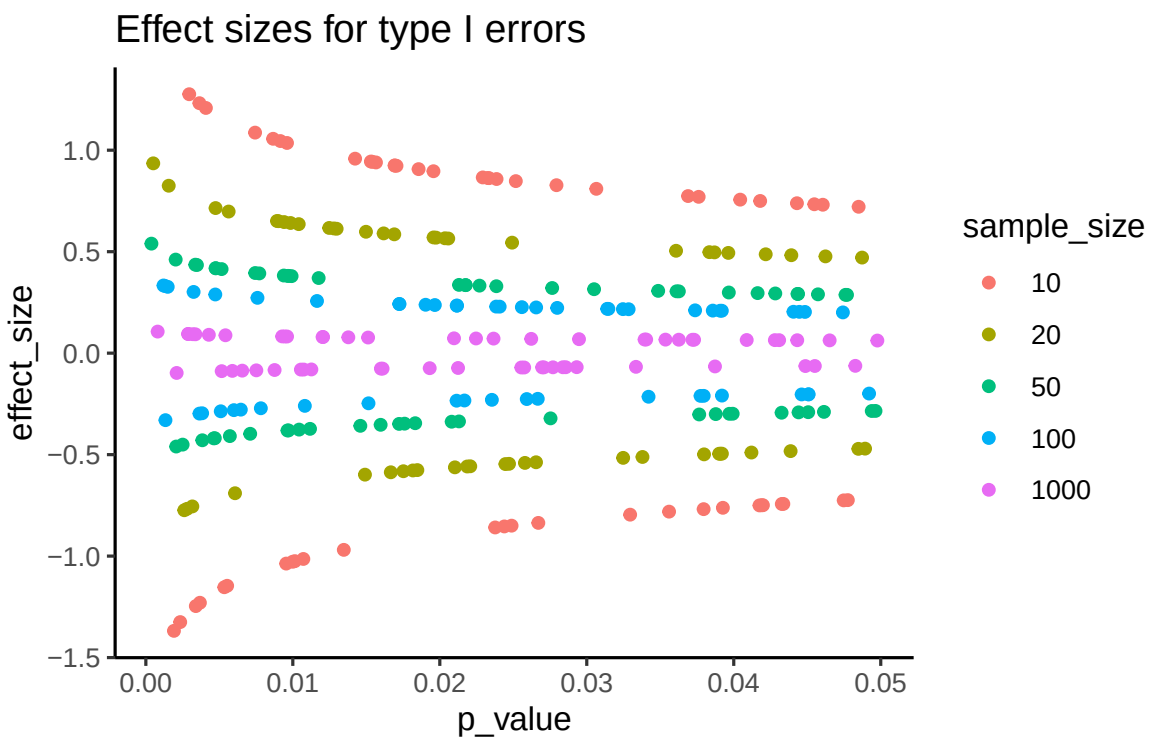

Figure 3: Effect size as a function of p-values for type 1 Errors under the null, for a paired samples t-test.

For example, look at the red dots, when sample size is 10. Here we see that the effect-sizes are quite large. When p is near 0.05 the effect-size is around .8, and it goes up and up as when p gets smaller and smaller. What does this mean? It means that when you get unlucky with a small N design, and your manipulation does not work, but you by chance find a "significant" effect, the effect-size measurement will show you a "big effect". This is the pernicious aspect. When you make a type I error for small N, your data will make you think there is no way it could be a type I error because the effect is just so big!. Notice that when N is very large, like 1000, the measure of effect-size approaches 0 (which is the true effect-size in the simulation shown in Figure 4).

```
all_df<-data.frame()
for(i in 1:10000){
  sample      <- rnorm(50,100,20)
  sample_mean <- mean(sample[1:25]-sample[26:50])
  sample_sem  <- sd(sample)/sqrt(length(sample))
  sample_t    <- t.test(sample, mu=100)$statistic
  sample_d    <- (mean(sample)-100)/sd(sample)
```

```
  t_df<-data.frame(i,sample_mean,
                   sample_sem,
                   sample_t,
                   sample_d)
  all_df<-rbind(all_df,t_df)
}

library(ggpubr)
a<-ggplot(all_df,aes(x=sample_mean))+
  geom_histogram(color="white")+
  theme_classic()
b<-ggplot(all_df,aes(x=sample_sem))+
  geom_histogram(color="white")+
  theme_classic()
c<-ggplot(all_df,aes(x=sample_t))+
  geom_histogram(color="white")+
  theme_classic()
d<-ggplot(all_df,aes(x=sample_d))+
  geom_histogram(color="white")+
  theme_classic()

ggarrange(a,b,c,d,
          ncol = 2, nrow = 2)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

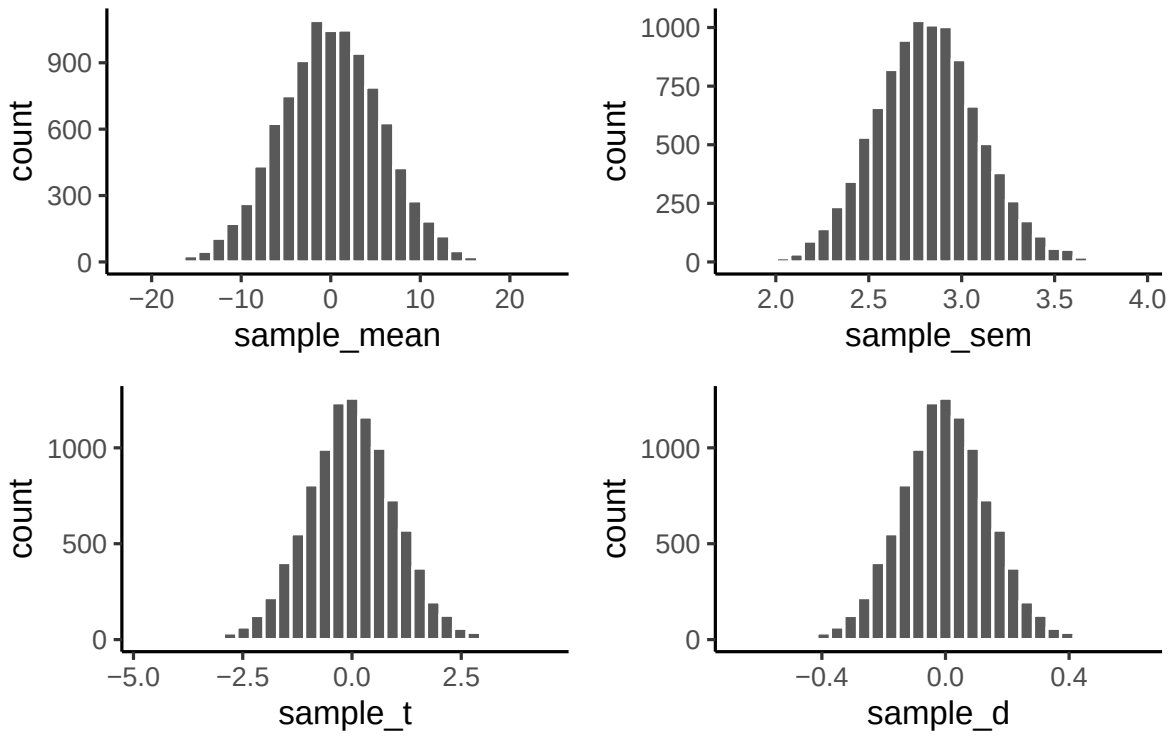

Figure 4: Each panel shows a histogram of a different sampling statistic.