12.4: 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.

The following graph 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.

```
library(ggplot2)
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)$p.value
    effect_size<-mean(some_data)/sd(some_data)
    mean_scores<-mean(some_data)
    standard_error<-sd(some_data)/sqrt(length(some_data))
    t_df<-data.frame(sim=i,sample_size=n,p_value,effect_size,mean_scores,standard_error)
    all_df<-rbind(all_df,t_df)
  }
}
type_I_error <-all_df[all_df$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_size))+
  geom_point()+
  theme_classic()+
  ggtitle("Effect sizes for type I errors")
```

run | restart | restart & run all

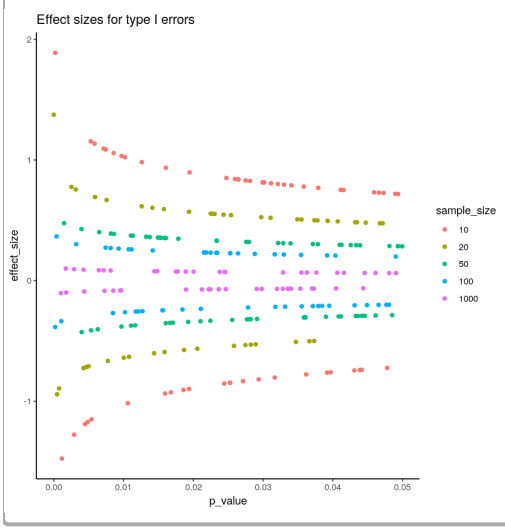

Figure \(\PageIndex{1}\): Effect size as a function of p-values for type 1 Errors under the null, for a paired samples t-test.

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).
