ABTestingConcepts

```
library(ggplot2)
```

In the world of hypothesis testing, rejecting the null hypothesis when it is actually true is called a type 1 error. Committing a type 1 error is a false positive because we end up recommending something that does not work.

Conversely, a type 2 error occurs when you accept the null hypothesis when it is actually false. This is a false negative because we end up sitting on our hands when we should have taken action. We need to consider both of these types of errors when choosing the sample size.

Visual Representation of the Power and the Significance Level

The concepts of power and significance level can seem somewhat convoluted at first glance. A good way to get a feel for the underlying mechanics is to plot the probability distribution of Z assuming that the null hypothesis is true. Then do the same assuming that the alternative hypothesis is true, and overlay the two plots.

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

p1 and p2 are the event rates for the control and test groups, respectively n1 and n2 are the sample sizes for the control and test groups, respectively p is the blended rate: (x1+x2)/(n1+n2) where x1 and x2 are event counts

Consider the following example:

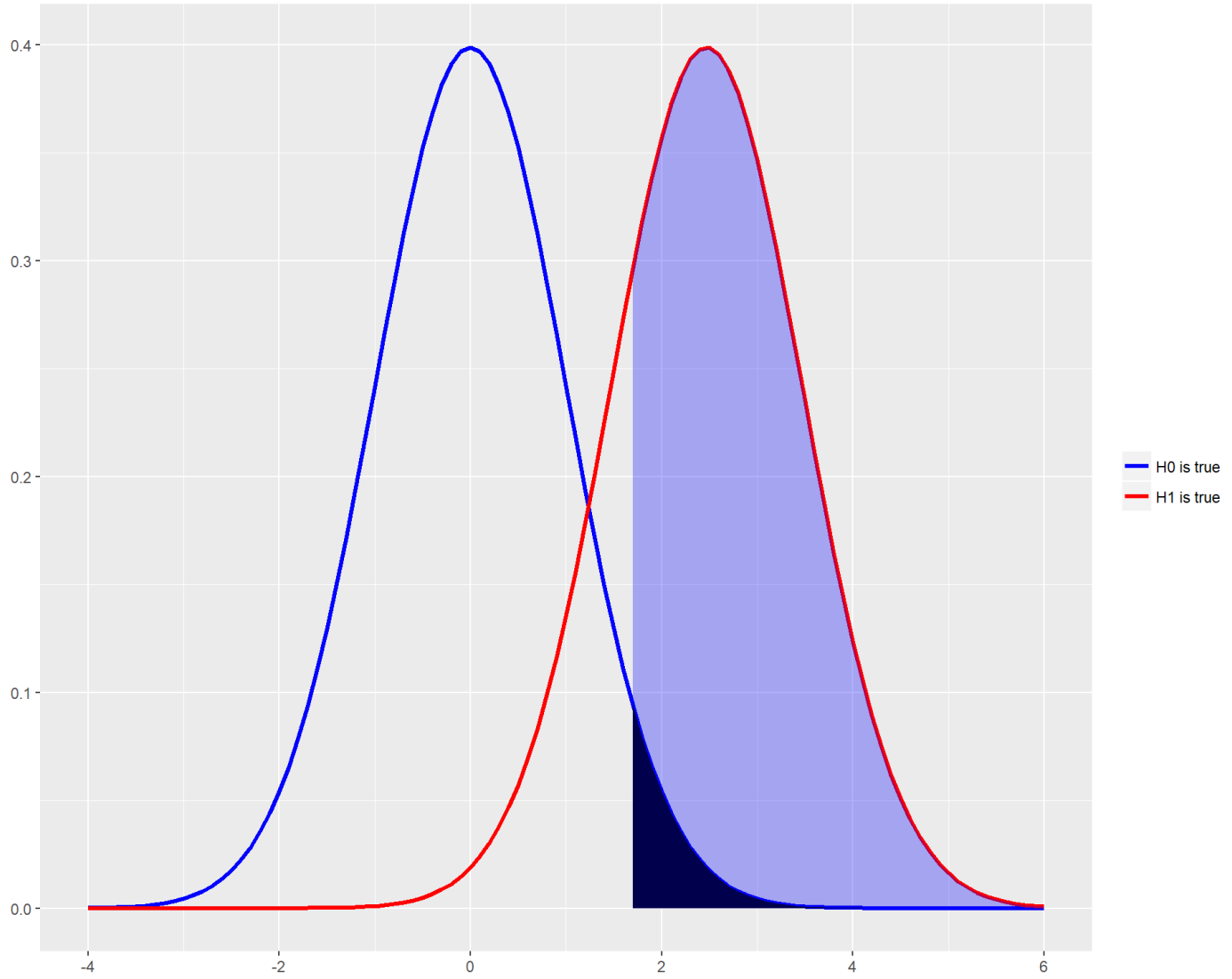
- H0: p1=p2, H1: p1>p2. A one-sided test was chosen here for charting-simplicity.
- Our chosen significance level is 5%. The corresponding decision rule is |Z|>1.65. The number (1.65) is the cutoff that corresponds to the upper 5% on the standard normal distribution.
- N=5,000 (2,500 in each cell).
- Say we decide that we need to observe a difference of 0.02 in order to be satisfied that the intervention worked (i.e., p1=0.10 and p2=0.08). We will discuss how to make this decision later in the post. The desired difference of 0.02 under the alternative hypothesis corresponds to Z=2.47 (using the formula for Z above). We are assuming that the variance is roughly the same under the null and alternative hypotheses (they're very close).

A typical requirement for the power is 80%. In other words, we will tolerate a 20 chance of a type 2 error (1 - power). As mentioned above, the typical requirement for the significance level is 5%.

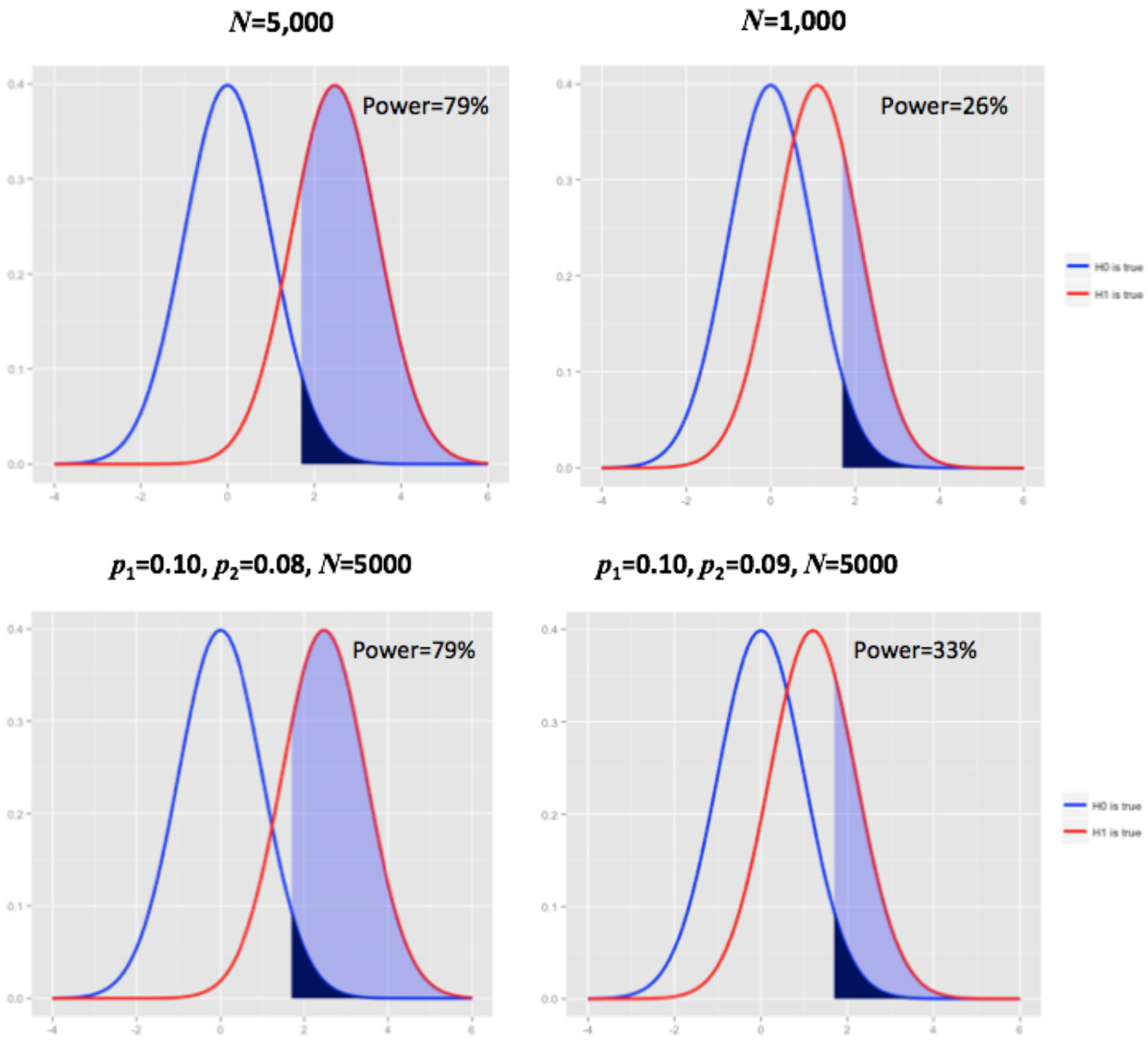
```
x <- seq(-4, 6, 0.1)
mean1 <- 0.00
mean2 <- 2.47
dat <- data.frame(x = x, y1 = dnorm(x, mean1, 1), y2 = dnorm(x, mean2, 1))
ggplot(dat, aes(x = x)) +
  geom_line(aes(y = y1, colour = 'H0 is true'), size = 1.2) +
  geom_line(aes(y = y2, colour = 'H1 is true'), size = 1.2) +
  geom_area(aes(y = y1, x = ifelse(x > 1.65, x, NA)), fill = 'black') +
  geom_area(aes(y = y2, x = ifelse(x > 1.65, x, NA)), fill = 'blue', alpha = 0.3) +
  xlab("") + ylab("") + theme(legend.title = element_blank()) +
  scale_colour_manual(breaks = c("H0 is true", "H1 is true"), values = c("blue", "red"))
```

```
## Warning: Removed 57 rows containing missing values (position_stack).

## Warning: Removed 57 rows containing missing values (position_stack).
```



The light blue area represents power. In this case, it is 79% The dark blue area is significance level of 5%. We reject the null hypothesis when z > 1.65.



Power Analysis

Let's say we require the significance level to be 5% and the power to be 80%. This means we have now specified two key components of a power analysis:

A decision rule of when to reject the null hypothesis. We reject the null when the p-value is less than 5%. Our tolerance for committing type 2 error (1???80%=20%). Our next task now is to find the sample size that meets these two criteria. On the surface this sounds easy, but it actually poses a challenge. In fact, this is akin to solving one equation (power=80%) with two unknowns: the sample size and the detectable difference. The detectable difference is the level of impact we want to be able to detect with our test. It is almost always the bottleneck in a power analysis and directly dictates the precision of the sample. We cannot solve for the sample size without first specifying the level of impact we want to be able to detect with our test.

In order to explain the dynamics behind this, let's go back to the definition of power: the power is the probability of rejecting the null hypothesis when it is false. Hence for us to calculate the power, we need to define what "false" means to us in the context of the study. In other words, how much impact, i.e., difference between test and control, do we need to observe in order to reject the null hypothesis and conclude that the action worked?

Let's consider two illustrative examples: if we think that an event rate reduction of, say, 10^{-10} is enough to reject the null hypothesis, then we need a very large sample size to get a power of 80%. This is pretty easy to deduce from the charts above: if the difference in event rates between test and control is a small number like 10^{-10}, the null and alternative probability distributions will be nearly indistinguishable. Hence we will need to increase the sample size in order to move the alternative distribution to the right and gain power. Conversely, if we only require a reduction of 0.02 in order to claim success, we can make do with a much smaller sample size. The smaller the detectable difference, the larger the required sample size.

Choosing the Detectable Difference

Unlike the significance level and the power, there are no plug-and-play values we can use for the detectable difference. However, if we put the detectable difference in the context of what we are trying to get out of the study, things become more clear. First, let's start with some guiding principles and then move on to specific suggestions:

Avoid wasteful sampling: Let's say it takes an absolute difference of 0.02 between test and control in order for the treatment to pay off. In this case, aiming for a 0.01 detectable difference would just lead to more precision than we really need. Why have the ability to detect 0.01 if we don't really care about a 0.01 difference? If there is no cost to sampling and/or you have an infinite pool of clients to choose from, this is a moot point. But in many cases, sampling for unnecessary precision can be costly, financially or in terms of over-contacting your clients.

Avoid missed opportunities: Conversely, if we are analyzing a sensitive metric where small changes can have a large impact, we have to aim for a small detectable difference. If we choose an insufficient sample size, we may end up sitting on our hands and missing an opportunity (type 2 error).

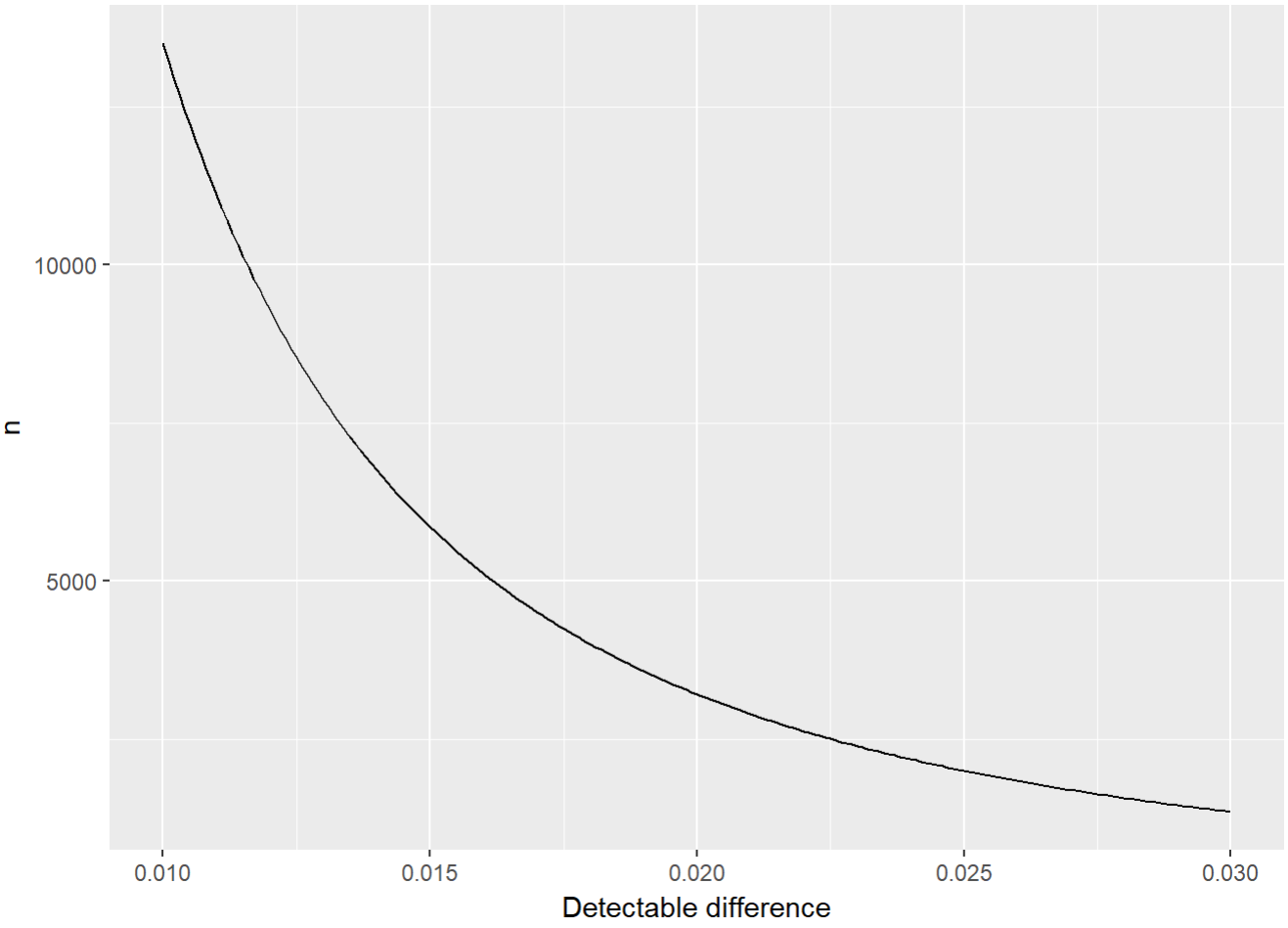
Clearly, the key is to define what "pay off" means for the study at hand, which depends on what the adverse event is a well as the cost of the action. Hence the answer should come from a cross-functional analysis/discussion between the data scientist and the business stakeholder.

Running Power Analyses in R

Once we have a viable range for the detectable difference, we can evaluate the sample size required for each option. This is easy to do in R using the code below. Here we are using the basic power functionalities in core R, but if you are running more complex analyses the pwr package may come in handy.

For example, let's say that p1=0.10 and we want the detectable difference to be between 0.01 and 0.03. Clearly, we'd rather be able to detect a difference of 0.01 but it may be too costly and hence we want to evaluate more conservative options as well.

```
library(scales)
library(ggplot2)
p1 <- 0.1 # baseline rate
b <- 0.8 # power
a <- 0.05 # significance level
dd <- seq(from = 0.01, to = 0.03, by = 0.0001) # detectable differences
result <- data.frame(matrix(nrow = length(dd), ncol = 2))
names(result) <- c("DD", "ni")
for (i in 1:length(dd)) {
  result[i, "DD"] <- dd[i]
  result[i, "ni"] <- power.prop.test(sig.level = a, p1 = p1, p2 = p1 - dd[i], alternative = 'tw
o.sided', power = b)
}
ggplot(data = result, aes(x = DD, y = ni)) +
  geom_line() + ylab("n") + xlab("Detectable difference") + scale_x_continuous(labels = comma)
```




```
#+
#  geom_point(data = result[ceiling(result\\(n / 10) * 10 == 5000, ],
#             aes(x = DD, y = ni), colour = "red", size = 5)
```

This graph shows that we need roughly 10x more observations to get a detectable difference of 0.01 compared to 0.03. This is because the power increases with $\sqrt{N}$ Hence, settling for a detectable difference around the middle of the range in terms of sample size requirement - e.g., 0.016 - is perhaps the most prudent choice. This leads to a decent power at a sample size of 10,000 (5,000 in each cell). Again, this is a made-up example and you can easily plug your own numbers into the code:

```
power.prop.test(sig.level=0.05, p1=0.1, p2=0.10-0.016, alternative='two.sided', n=5000)$power
```

```
## [1] 0.7905305
```

Power analysis for continuous variables

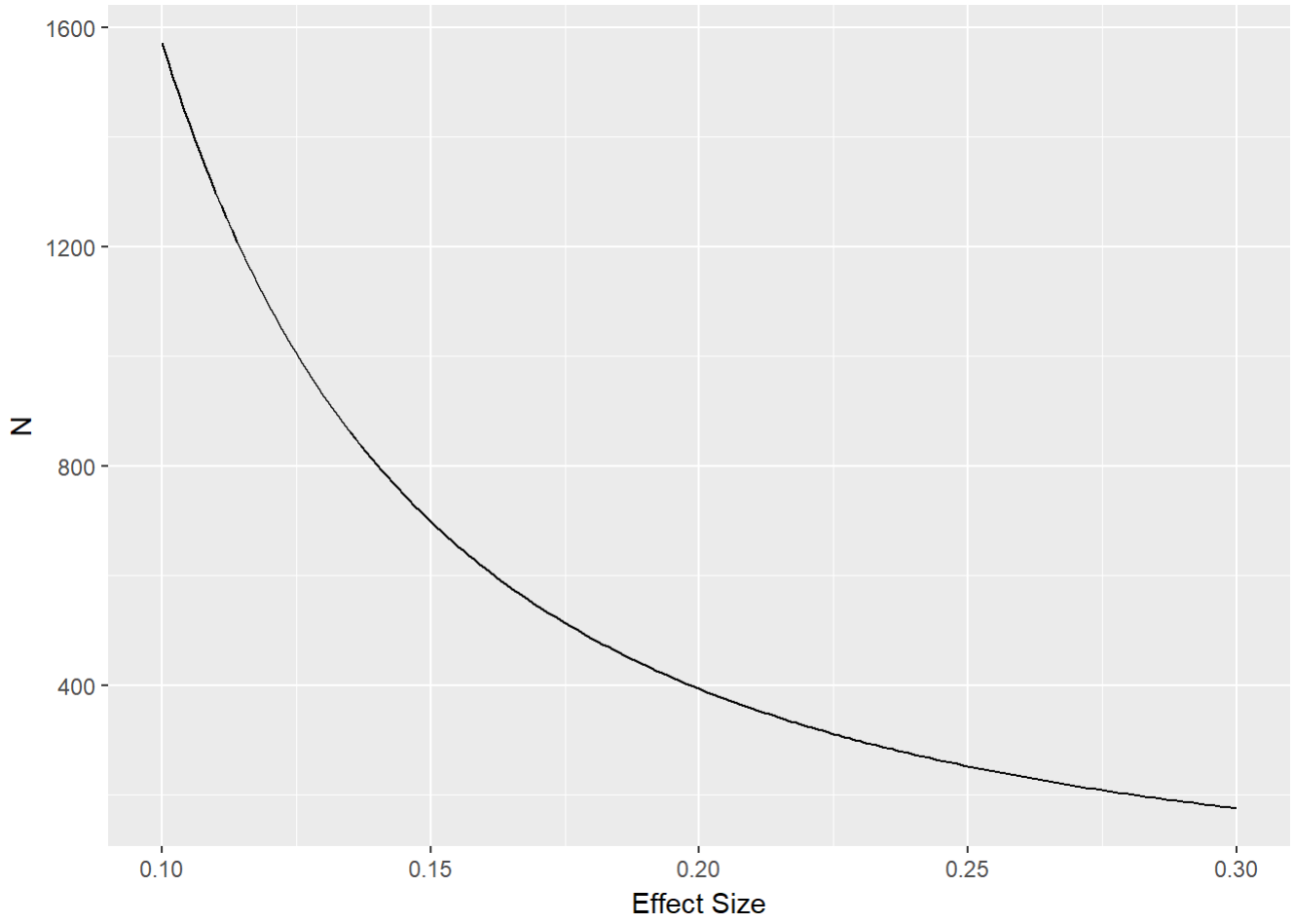
In the example above, we did a test of proportions since the underlying outcome variable is binary. When comparing averages of continuous variables across two samples (such as revenue), the simplest test statistic is the t-test. This test statistic is similar to the test we use for proportions - i.e., subtract the means and divide by the pooled standard error. However, the detectable differences are typically expressed in terms of the effect size given by:

$$D = \frac{\bar{X}_1 - \bar{X}_2}{S}$$

where S is the pooled standard deviation, X1 is the average of the control group, and X2 is the average of the test group. When choosing the D, we can follow the suggested framework above for the detectable difference. Another option is to use Cohen's guidelines for small, medium, and large effects (D=0.2,0.5,0.8). However, these values are tied to the historical variance of the data (the S in the denominator) as opposed to the business context.

In order to run these types of tests in R, simply replace power.prop.test with power.t.test.

```
library(ggplot2)
es <- seq(from = 0.1, to = 0.3, by = 0.001) # effect sizes
result <- data.frame(matrix(nrow = length(es), ncol = 2))
names(result) <- c("ES", "ni")
for (i in 1:length(es)){
  result[i, "ES"] <- es[i]
  result[i, "ni"] <- power.t.test(sig.level = a, d = es[i], sd = 1,
                                  alternative = 'two.sided', power = b)$n
}
ggplot(data = result, aes(x = ES, y = ni)) + geom_line() + xlab("Effect Size") + ylab("N") +
  scale_x_continuous(labels = comma)
```



Online calculator

https://www.abtasty.com/sample-size-calculator/ (https://www.abtasty.com/sample-size-calculator/)

https://www.evanmiller.org/ab-testing/ (https://www.evanmiller.org/ab-testing/)