Quantitative Understanding in Biology
1.5 Experimental Power and Design

Jason Banfelder

September 22, 2020

In our last two sessons, we've emphasized the importance of correctly controlling for Type I error rates (i.e., reporting a significant result when there is fact no real underlying effect) to ensure that we don't under-report the chances of reaching an irreproducible inference in our work. We're now going to shift gears and focus on Type II errors. In other words, we'll be addressing what we can do to control for the chances that we'll miss a genuine effect when there really is one.

We've seen over and over again that confidence intervals computed by statistical tests will be narrower in experiments that include more samples. In this section, we will use our knowledge of how CIs (and, equivalently, p-values) vary with n to plan experiments of an appropriate size.

Before we begin, it is important to recall that it is not valid to incrementally add samples to a study until you obtain a significant result. In other words, if you perform an experiment with a sample size of six and obtain a p-value of 0.07, you can't go back to the bench and add two more samples so you can rerun your statistics with $n = 8$, because you'd be understating your Type I error rate. But, of course, the motivation in adding more n is to reduce the Type II error rate. Since this probably isn't an argument you want to get into with your PI, it is a really good idea to carefully consider experimental design and sample size up front.

In this section, we'll define the precision of a CI as its half-width. In the case of a symmetric CI, the CI is written as

$$\text{CI} : \overline{x} \pm \text{precision} \tag{1}$$

Other texts may define precision differently, so if you look at other sources or use a computer program to run these calculations, make sure you know how this term is defined and adjust your interpretation accordingly. Some of the literature reasons in terms of "effect size".

This is usually denoted with the symbol d, and is defined by the relation...

$$d = \frac{\text{precision}}{\text{SD}} \tag{2}$$

You can think of the effect size as a normalized precision. The SD and the precision have the same units of measure as the quantity being measured, so d is a dimensionless quantity.

Statistical power calculations only give you estimates of the sample size that will allow you (with some likelihood) to conclusively observe a desired effect size, or one that is larger. Furthermore, when you use these methods, you'll need to estimate quantities like the standard deviation (SD) of the quantities that you'll measure. The bottom line is that most of what is presented in this section is approximate, so (1) we'll feel free to use approximations in our formulae, and (2) it is a good idea to be conservative when we provide our estimates.

To perform power calculations using R, you'll want to install the `pwr` package. Installing a package can be done via the GUI on Mac and Windows implementations of R, or at the command line...

```
install.packages(c('pwr'))
```

You only need to do this once to install the package on your computer. In each R session where you want to use functions from this package, you'll need to load the library with the command:

```
library(pwr)
```

1 Single Mean

From our previous lectures, we already know enough to estimate sample sizes for some special cases. Recall that the 95% CI for a univariate distribution with large n is given by...

$$95\%\ \text{CI: } \overline{x} \pm 1.96 \cdot \text{SEM} \tag{3}$$

$$95\%\ \text{CI: } \overline{x} \pm 1.96 \cdot \frac{\text{SD}}{\sqrt{n}} \tag{4}$$

If we rearrange and take $1.96 \cong 2$, then we can write...

$$n \cong 4 \cdot \left(\frac{\text{SD}}{\text{precision}} \right)^2 \tag{5}$$

This is a useful rule of thumb to have at your disposal for estimating how many measurements need to be taken to estimate the true mean to a desired precision.

Note that in order to use this formula, you'll need to estimate the SD of a population that you haven't taken samples from yet. You can usually get a rough idea of what this quantity will be by looking at previously obtained data. If you're not sure, be conservative and choose something on the high-side of what you expect.

The formula above only applies when n is sufficiently large to make the approximation that $t^* \cong 2$. So if you get $n = 4$ from the above formula, you should appreciate that you are likely to be underestimating n significantly.

This line of reasoning can be generalized by recalling that...

$$(1 - \alpha)\,\text{CI: } \overline{x} \pm t^* \cdot \text{SEM} \tag{6}$$

...which implies...

$$n \cong \left(t^* \cdot \frac{\text{SD}}{\text{precision}} \right)^2 \tag{7}$$

Of course, t^* is a function of n (and α), so this equation has to be solved iteratively. These sorts of calculations can be performed in R. We will show some examples in the next section.

The above equations do not guarantee that if you perform n measurements, you'll obtain a CI with the desired half-width. In fact, if all of the assumptions in the analysis hold, you'll have a 50% chance of obtaining such a CI, or narrower. Put in another way, **your power to obtain the desired precision will be 0.5**. The power of an experiment is an important quantity, and it is helpful to have an estimate of the power of an experiment before you perform it. Formally, the power of an experiment is one minus the probability of a type II error (assuming an effect of the specified size is actually present). Informally, power is the chance that you'll be able to measure an effect of a given size.

2 Difference Between Two Means

If you want to be able to determine the difference between the means of two groups of measurements to a certain desirable precision, the rule of thumb is...

$$n_{\text{each group}} \sim 8 \cdot \left(\frac{\text{SD}_{\text{each group}}}{\text{precision}} \right)^2 \tag{8}$$

There is an assumption that the SDs of the measurements from both groups are roughly the same. As before, the sample size given by this formula will give you a 50% chance

of realizing your desired precision. You'll need significantly more samples than calculated above for a 95% chance of hitting your target.

Example: In a series of knockdown experiments on MDCK cells, it was desired to confirm that preparations of the knockdown prevent the formation of functional tight junctions (TJs). This is assessed by measuring (among other things) transepithelial resistance (TER). Inspection of previous studies shows that the mean value of TER for wild-type cells that are known to form TJs is about $130\,\Omega\,\mathrm{cm}^2$, and the standard deviation of TER measurements is about $30\,\Omega\,\mathrm{cm}^2$. In this experiment, we are only interested in whether tight junctions form, not on the specific effects that a knockdown has on TER (perhaps via the regulation of TJs). We might say that variations of up to 35% in TER would still be indicative of TJ formation. The required precision for this experiment is therefore not particularly high: we just want a CI with a precision of roughly $\pm 45\,\Omega\,\mathrm{cm}^2$. According to our rule of thumb, we'll need...

$$n_{\text{each group}} \sim 8 \cdot \left(\frac{30\,\Omega\,\mathrm{cm}^2}{45\,\Omega\,\mathrm{cm}^2}\right)^2 = 3.5 \tag{9}$$

This tells us that we'll need at least four samples per group. However, since the resultant n is small, we suspect this is a significant underestimation. So we turn to R's `pwr` package to help us do a better job.

Working in R requires that we pose our question in terms of effect size instead of precision. In this case $d = 45/30 = 1.5$.

```
pwr.t.test(d = 45 / 30, power = 0.5, sig.level = 0.05,
           type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 4.566802
##               d = 1.5
##       sig.level = 0.05
##           power = 0.5
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

R is telling us that we need five samples per group. Note that a two-sample, two-sided t-test with a significance level of 0.05 is the default, so these parameters didn't need to be explicitly specified.

This computation is analogous to the rule of thumb, and tells us that if we use five samples, we'll have a 50/50 chance of obtaining a 95% CI with a precision equal to or better than

our desired precision. If we want to have an 80% chance, we'll need nine samples per group, and for a 95% assurance that our CI will be as narrow as we want, we'd need 13 samples per group.

```
pwr.t.test(d = 45 / 30, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 8.060321
##               d = 1.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr.t.test(d = 45 / 30, power = 0.95)
```

```
##
##      Two-sample t test power calculation
##
##               n = 12.59873
##               d = 1.5
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

In the above example, we were only hoping to detect relatively large effect sizes. If our experiment was looking not simply to determine if tight junctions were being formed, but rather to quantify potentially subtle effects of preparation methodology on TER, then we might say that we want to be able to resolve 10% changes in mean TER. Our precision would then be $13\,\Omega\,\mathrm{cm}^2$, and our effect size would be $\frac{13}{30} = 0.43$. Our rule of thumb then tells us $n = 8(\frac{30}{13})^2 = 42.6$, so we estimate that we'd need 43 samples in each group to have a 50/50 chance of obtaining such a narrow CI.

A more precise calculation in R...

```
pwr.t.test(d = 13 / 30, power = 0.5)
```

```
##
##      Two-sample t test power calculation
```



```
##
##               n = 41.88941
##               d = 0.4333333
##       sig.level = 0.05
##           power = 0.5
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr.t.test(d = 13 / 30, power = 0.95)
```

```
##
##      Two-sample t test power calculation
##
##               n = 139.373
##               d = 0.4333333
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

. . . indicates that we need 140 samples in each group for a 95% assurance. That's 280 samples in all, and assuming that you allow for some experimental problems, you likely need to plan (and budget) for 300 or so preparations.

If you thought that the above computed samples sizes were surprisingly high, you are not alone. Often when studies are planned (an all too rare event in the first place), the first power calculations along these lines can be quite depressing. Although we can use power calculations as above to compute a required n, budget, time and other constraints often put an upper bound on n. What is usually needed in practice is a more holistic view of the interplay and tradeoffs among power, effect size, and n that will aid in the selection of a pragmatic experimental plan. Preparation of plots can be very helpful in this regard.

The `pwr.t.test` function in R and its analogs for other tests are convenient in that they can compute any unknown given the other three. So if we want to know what effect size we can reasonably (power = 80%) expect to measure in a given experiment, with a total of 50 samples (25 in each group), we can compute. . .

```
pwr.t.test(power = 0.8, n = 25)
```

```
##
##      Two-sample t test power calculation
```

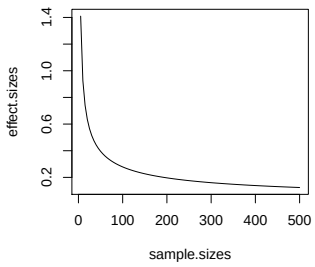


```
##
##               n = 25
##               d = 0.8087121
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

So in our example above, we could reasonably expect 95% CIs with a precision of 24 $\Omega\,cm^2$. Remember that the significance level defaults to 0.05, so if you want R to compute the significance level you must explicitly override the default value with the specification: `sig.level = NULL`.

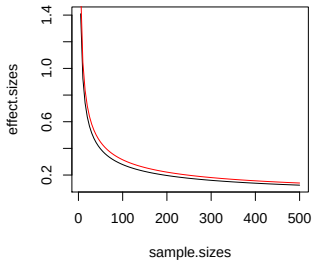
We can plot effect sizes as a function of sample size...

```
sample.sizes <- seq(5, 500, 5)
effect.sizes <- numeric(length(sample.sizes))
for (i in 1:length(sample.sizes)) {
  effect.sizes[i] = pwr.t.test(n = sample.sizes[i], power = 0.5)$d
}
plot(sample.sizes, effect.sizes, type = "l")
```



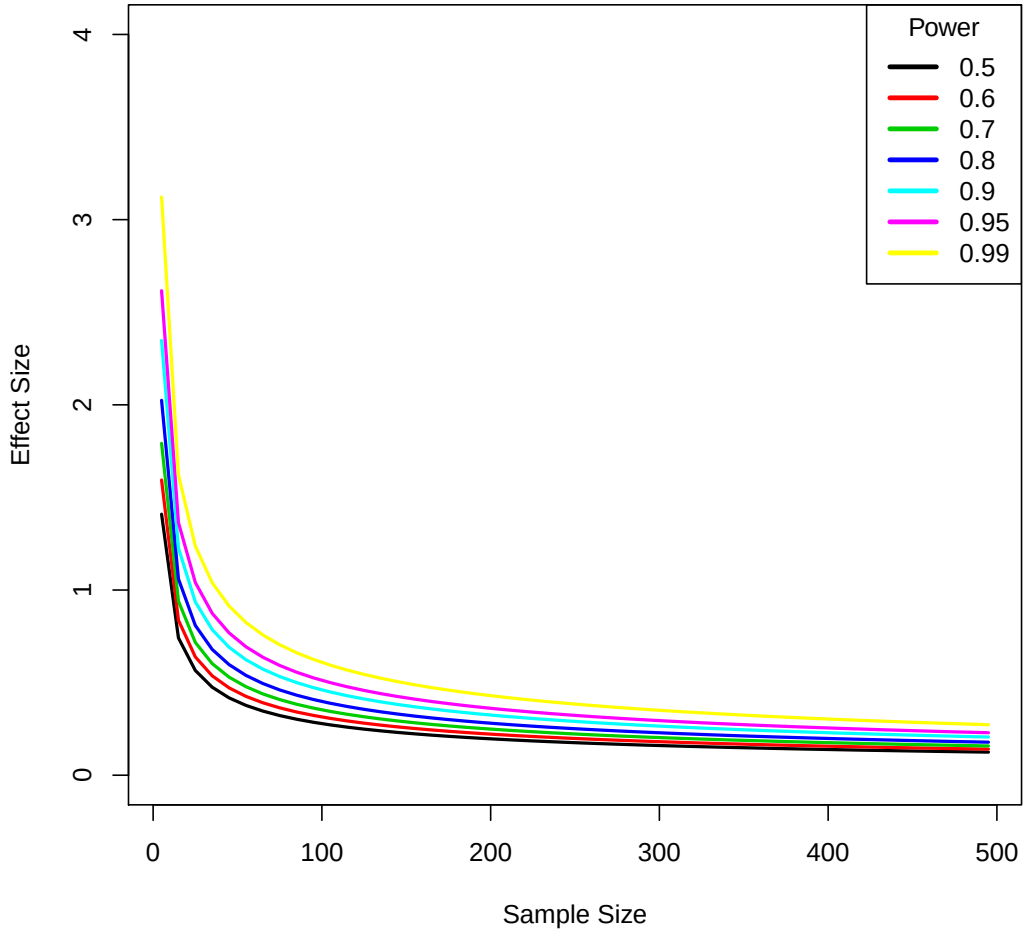
...and we can add a second line for a different level of power:

```
for (i in 1:length(sample.sizes)) {
  effect.sizes[i] = pwr.t.test(n = sample.sizes[i], power = 0.6)$d
}
lines(sample.sizes, effect.sizes, col='red')
```

Preparing a complete plot is not all that much harder if we use some loops.

```
power.plot.1 <- function() {
  sample.sizes <- seq(5, 500, 10)
  powers = c(0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99)
  effect.sizes <- array(numeric(), c(length(sample.sizes), length(powers)))
  for (i in 1:length(sample.sizes)) {
    for (j in 1:length(powers)) {
      effect.sizes[i, j] = pwr.t.test(n = sample.sizes[i], power = powers[j])$d
    }
  }
  xrange <- c(floor(min(sample.sizes)), ceiling(max(sample.sizes)))
  yrange <- c(floor(min(effect.sizes)), ceiling(max(effect.sizes)))
  plot(xrange, yrange, xlab = 'Sample Size', ylab = 'Effect Size', type = 'n')
  for (j in 1:length(powers)) {
    lines(sample.sizes, effect.sizes[ ,j], col = j, lwd = 2)
  }
  legend('topright', title = 'Power', as.character(powers), lwd = 3, col = 1:j)
}
power.plot.1()
```

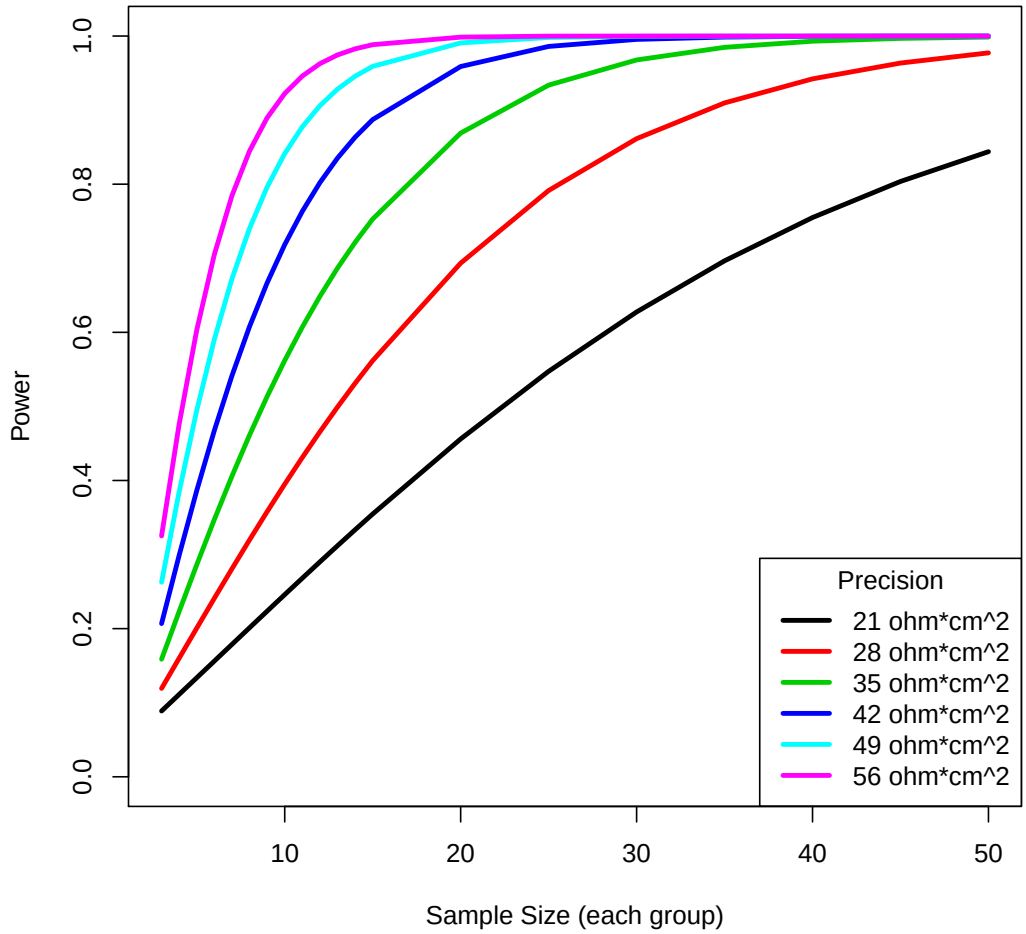



An alternative view can be obtained by plotting curves of power as a function of sample size for different effect sizes. Here we will also convert dimensionless effect sizes to precisions that are expressed in the units of measure for our TER example.

```
power.plot.2 <- function() {
  sample.sizes <- c(3:15, seq(20, 50, 5))
  effect.sizes <- seq(0.6, 1.6, 0.2)
  precisions <- effect.sizes * 35
  powers <- array(numeric(), c(length(sample.sizes), length(effect.sizes)))
  for (i in 1:length(sample.sizes)) {
    for (j in 1:length(effect.sizes)) {
      powers[i, j] = pwr.t.test(n = sample.sizes[i], d = effect.sizes[j])$power
    }
  }
```



```
  xrange <- c(floor(min(sample.sizes)), ceiling(max(sample.sizes)))
  yrange <- c(0, 1)
  plot(xrange, yrange, xlab = 'Sample Size (each group)', ylab = 'Power', type = 'n')
  for (j in 1:length(effect.sizes)) {
    lines(sample.sizes, powers[ ,j], col = j, lwd = 3)
  }
  legend('bottomright', title = 'Precision',
   paste(as.character(precisions), "ohm*cm^2"), lwd = 3, col = 1:j)
}
power.plot.2()
```



Plots like these typically are the most useful for planning experiments. All curves will have the same basic shapes because there is always zero power as the sample size approaches

zero, and power can be made arbitrarily high by increasing sample size to something very large.

If your power is very low, you may be better off not doing the experiment at all (this is always an option). Similarly, if your experimental plan puts you on the upper flat part of these curves, you might consider reducing your sample size a bit.

To review: In reality, the decision to include a certain number of samples in an experiment is driven not by a single power calculation, but by understanding the tradeoffs among power, sample size, and precision. The precision you need (or want) is something that should be guided by your scientific judgment and understanding of the underlying biology of your system.

All of the results hinge on having a reasonable estimate of the variation of your data (this appears as the SD in these analyses). Recall that in many biological studies variation can come from both measurement error and biological diversity. You can do something about measurement error by being more careful at the bench, or by switching to more precise methods, but realize that a good deal of intrinsic biological variation is typically unavoidable.

3 Non-Equal Sample Sizes

As mentioned above, the decision to include a certain number of samples in an experiment is usually driven in part by budget and time constraints. In some cases, the constraints on sample size may be hard limits if you only have access to a fixed number of consenting patients with a rare disease or a limited number of surgical tissue specimens. In many such cases, the hard constraint is imposed on the number of samples in one group only. You can still gain some statistical power by increasing the number of samples in the other group (typically the control group), but there are limits to this.

You'll always need the fewest total samples when sample sizes are equal, but you can use unequal sample sizes if you need to. For example, R shows us that we need roughly 64 samples in each group to have an 80% chance of measuring an effect that is half the size of the SD of the data we are collecting:

```
pwr.t.test(power = 0.80, d = 0.5)$n
```

```
## [1] 63.76561
```

If we have access to only 48 experimental samples, we can compute how many control samples we would need to achieve the same goals...


```
pwr.t2n.test(n1 = 48, power = 0.80, d = 0.5)$n2
```

```
## [1] 94.48827
```

There are limits to how far you can go. For example, if you only have access to 30 experimental samples, you simply cannot measure an effect of this size with a power of 80%.

```
pwr.t2n.test(n1 = 30, power = 0.80, d = 0.5)$n2
```

```
## Error in uniroot(function(n2) eval(p.body) - power, c(2 + 1e-10, 1e+09)):
f() values at end points not of opposite sign
```

As above, plots can be prepared to gain insight into the tradeoffs that are at work under these circumstances.

4 Experimental Design by Simulation

In this lecture, we've focused on one specific example where a two-sample t-test is applicable. The `pwr` package in R has a number of other tests for other scenarios (comparing proportions, dealing with contingency tables, etc.). The same principles apply, and you should be able to use these tests to prepare plots if you need to.

Another approach is to use simulation to compute power. Here is a function analogous to `pwr.t.test` that computes power by repeatedly performing numerical experiments:

```
pwr.t.test.sim <- function(d = 1,
                           n = 100,
                           sig.level = 0.05,
                           trial.count = 10000) {

  trials <- 1:trial.count
  p.values <- numeric(trial.count)

  for (i in trials) {

    x <- rnorm(n)
    y <- rnorm(n, mean = d)

    p.values[i] <- t.test(x, y)$p.value

  }
```



```
  hist(p.values, breaks = seq(0, 1, sig.level))
  power <- sum(p.values <= sig.level) / trial.count
  return(power)
}
```

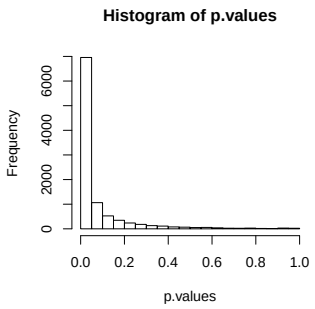
We can demonstrate that this works by performing the same computation using standard statistical methods, and by simulation:

```
pwr.t.test(n = 50, d = 0.5)
```

```
##
##      Two-sample t test power calculation
##
##               n = 50
##               d = 0.5
##       sig.level = 0.05
##           power = 0.6968934
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr.t.test.sim(n = 50, d = 0.5)
```

```
## [1] 0.6957
```



Don't underestimate the utility of simulation. We often need to turn to simulation to answer real-world questions for which analytical statistical methods are difficult to find and apply properly, or for which they don't exist at all. For example, all of the two sample t-tests that we've performed in this section assume that the SD is the same for both groups. If we have reason to believe that this is not the case, we could easily modify our simulation to probe how the power of our experiment would be affected:


```
pwr.t2sd.test.sim <- function(n = 100,
         mean1 = 0, sd1 = 1.0,
         mean2 = 1, sd2 = 1.0,
         sig.level = 0.05,
         trial.count = 10000) {

  trials <- 1:trial.count
  p.values <- numeric(trial.count)

  for (i in trials) {

    x <- rnorm(n, mean = mean1, sd = sd1)
    y <- rnorm(n, mean = mean2, sd = sd2)

    p.values[i] <- t.test(x, y)$p.value

  }
  hist(p.values, breaks=seq(0, 1, sig.level))
  power <- sum(p.values <= sig.level) / trial.count
  return(power)
}
```

It would not be a stretch to modify this to include differing group size, to test underlying distributions other than the normal distribution, etc. You would expect the basic principles of the tradeoffs among the effect size, sample size, and power to still hold, but quantifying complex situations confidently without simulation can be challenging.

5 Exercise

A casino is running a simple game where a coin is flipped, and the house wins when the coin comes up heads, and the patron wins when the coin comes up tails.

You're investigating a complaint from an irate gambler that the casino is cheating, and plan to go to the casino and inspect their coin by flipping it repeatedly, looking for a suspicious number of heads.

Show that you would need to plan on making around 39,000 coin flips to have an 80% chance of detecting a coin that comes up heads 51% of the time if you want to be very conservative about falsely accusing the casino of cheating by requiring that you're 99.9% sure that any accusation you make is correct.

