Power Analysis

Nicholas Kortessis

A tactic that is commonly used in hypothesis testing is called a **power analysis**. Power analysis answers the question: What is the probability that I can detect particular effect given my test?

This is really about figuring out how to modify the test to have the best reasonable chance to detect an effect of a particular size. Power analysis relies on 4 fundamental features:

1. **Effect Size** - effect size is the difference between some specified outcome and the null hypothesis. Effect size has two meanings. One is the effect of an alternative hypothesis to the null. The other is the effect size from a particular observation from data, which is the difference between the estimate from a sample and the null hypothesis. All else equal, tests have more power to large effects than small effects.
2. **Sample Size** - sample size is the number of statistical individuals in a single sample. All else equal, studies with larger sample sizes have greater power.
3. **Type I Error** - Type I error is the probability of falsely rejecting the null (or a false-positive error). It is typically depicted with α. Keeping Type I error probabilities low trades off against power. As such, increasing Type I error rates typically increases power.
4. **Population Variability** - Population variability is σ, and typically represents the difficulty in estimation. The larger σ, the more difficult it is to estimate anything about the population because the estimates can vary so much from sample to sample. As such, when σ is larger, all else equal, power is lower. Some statisticians refer to population variability as "problem difficulty".

Since power analysis answers the question, what is the chance I can detect a particular effect from my test?, we need to have a good understanding of effect sizes.

Effect sizes are just the difference of the populations or samples from the null. Before a study is done, you could presume a the population has a particular population parameter that is linked to your question. How much that parameter differs from the null hypothesis is the effect size. After a study is done, the data in hand have some difference with the null hypothesis (even if it is very small and the difference is not statistically significant). The difference between the estimated parameters from the data and the null hypothesis is the effect size.

A quick example of a power analysis: A binomial test

The setting

Imagine that we are sampling a population for males and females and we are interested in whether males and females equally common in the population. Male versus female is a binary, categorical characteristic of an individual (a single organism of the species under study). We can characterize the biology probabilistically with the following probability distribution

$$X_i = \begin{cases} \text{Male} & \text{with probability } p \\ \text{Female} & \text{with probability } 1-p \end{cases}$$

where X_i is the character of the ith individual in the population.

The sampling distribution

So long as we sample randomly, we can use the binomial as a sampling distribution. The binomial has the parameters p (probability of "male") and n (sample size). The binomial comes with an implicit assumption of random sampling, exemplified by the specific assumptions that the individual probabilities of success are identical and independent of one another (also known as **iid** for independent and identically distributed).

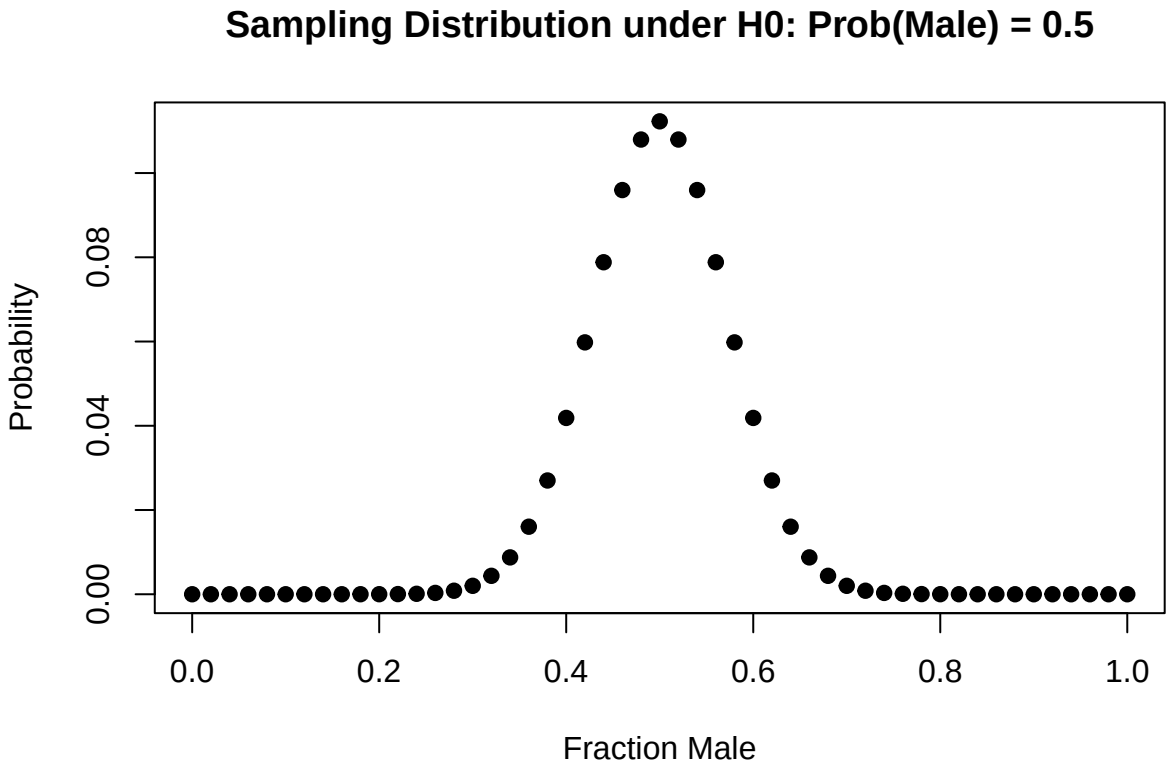
If we are interested in whether there is evidence that males and females are unequal in the population, we compare this males and females are equally likely, a reasonable null hypothesis is defined on the parameter p. such that Prob(Male) = 0.5. That is a statement that males and females are equally likely in the population. Here it is in model form.

$$H_0 : p = 0.5$$

$$H_A : p \neq 0.5.$$

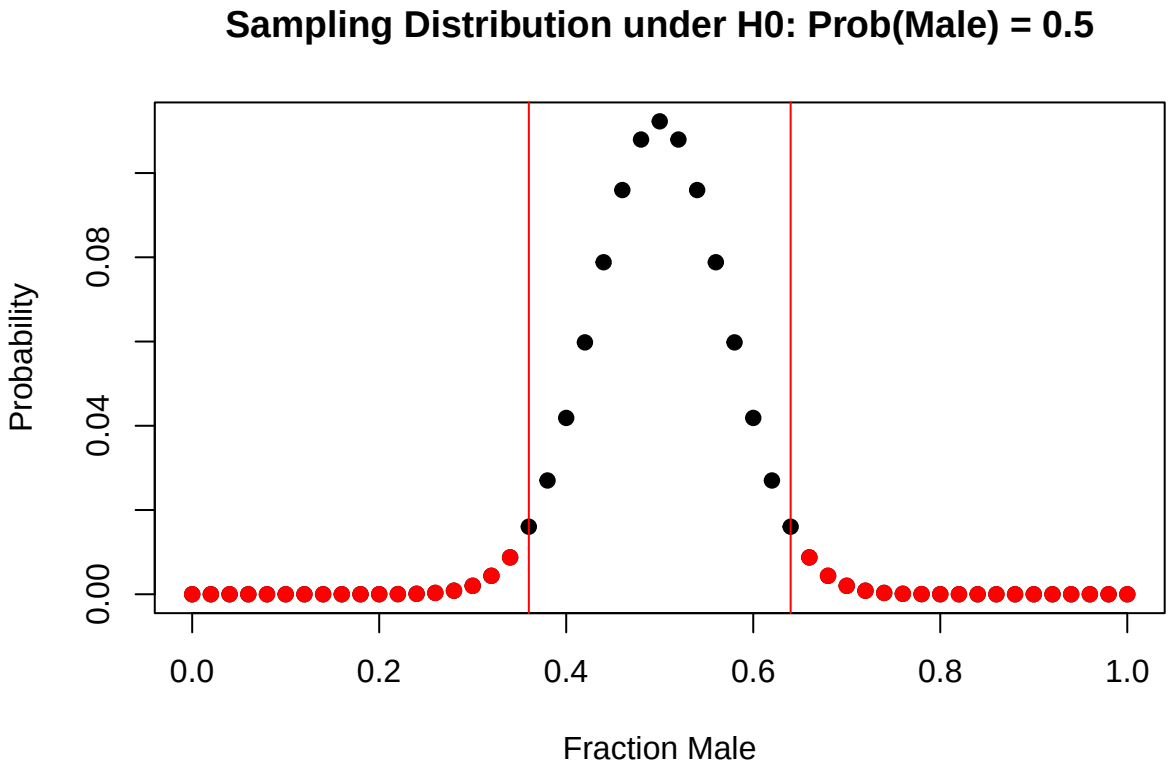
To build a test, we develop a sampling distribution of the number of males under the null hypothesis that $p = 0.5$. If S is the number of males, then $S \sim \text{Binomial}(n, p = 0.5)$ is the sampling distribution under the null hypothesis. Here is an example with a sample size of $n = 50$.

```
rm(list = ls())
null.p <- 0.5
sample.size <- 50

# Here is a visual of the sampling distribution under the null hypothesis.
# Possible outcomes
poss.num.males <- 0:sample.size
# Probability of each outcome
prob <- dbinom(poss.num.males, sample.size, null.p)
# Convert the outcomes to 'proportion male in sample'
poss.frac.males <- poss.num.males/sample.size
# Make plot
plot(poss.frac.males, prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = 'Sampling Distribution under H0: Prob(Male) = 0.5')
```



Now, before we collect any data, we need to decide what our type I error is. This is the probability of falsely rejecting the null hypothesis in the case where it is true. We constructed a sampling distribution under the assumption that the null hypothesis is true. So to practically use the Type I error, we just need to use the α values to find the rejection region. Let's set $\alpha = 0.5$.

```
# Now we pick a type I error level test.
alpha <- 0.05

# Find the critical values for the test.
crit.test.values <- qbinom(c(alpha/2, 1 - alpha/2),
                           size = sample.size,
                           prob = null.p)/sample.size
# Plot the critical values
plot(poss.frac.males, prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = 'Sampling Distribution under H0: Prob(Male) = 0.5')
abline(v = crit.test.values, col = 'red')
# Find the rejection region and label points in red.
rej.indx <- poss.frac.males < crit.test.values[1] | poss.frac.males > crit.test.values[2]
rej.region <- poss.frac.males[rej.indx]
# Label the rejection region points in red.
points(rej.region, prob[rej.indx], pch = 19, col = 'red')
```



At this point, our test is set up. We now sample 50 individuals, calculate the fraction that are male, and then evaluate whether it is in the rejection region. If it is, we reject the null hypothesis. If it is not, we fail to reject it.

But what if we have a suspicion that the actual fraction that is male is 55%, what is our chance of rejecting the null? This is the essence of power. We want to know the probability of correctly rejecting the null given the actual fraction male is something different than 0.5. To do this, we need to pick a possible alternative to the null. Here, we will pick prob(male) = 0.55.

What we are doing now is evaluating the power to detect a particular "effect size". An effect size is the difference between the null and what you aim to detect. Let's calculate it.

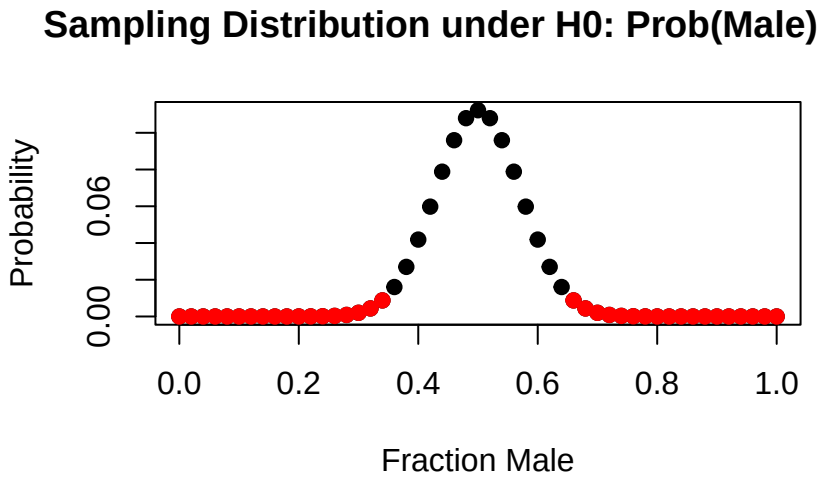
```
hyp.real.p <- 0.55 # Presumed real p value
(effect.size <- hyp.real.p - null.p)
```

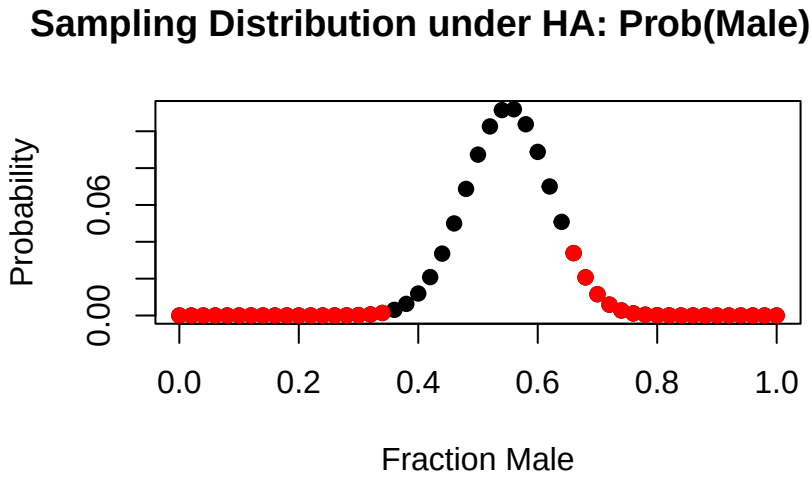
```
## [1] 0.05
```

Now we create what the sampling distribution would be if the male fraction is actually 55%. We do that again with a binomial distribution, but we need to now use a different probability, prob(Male) = 0.55.

```
alt.prob <- dbinom(poss.num.males, sample.size, hyp.real.p)

# Let's plot it on a paired plot.
par(mfcol = c(2,1))
# First, the sampling distribution under the null.
plot(poss.frac.males, prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = 'Sampling Distribution under H0: Prob(Male) = 0.5')
points(rej.region, prob[rej.indx], pch = 19, col = 'red')

# Now, the sampling distribution under the alternative of a particular effect
# size.
```

```r
plot(poss.frac.males, alt.prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = paste('Sampling Distribution under HA: Prob(Male) = ',hyp.real.p,
                  sep = ''))
# Now let's plot the rejection region of the null hypothesis test on this
# sampling distribution. By doing so, we are getting a picture of how often we
# would reject the null IF THE ALTERNATIVE OF A SPECIFIC EFFECT WERE TRUE.
# This is the essence of a power analysis.
points(rej.region, alt.prob[rej.indx], pch = 19, col = 'red')
```






```r
# The power is the probability of rejecting the null under the alternative.
# To calculate that probability, we just sum of the probability of the red
# values of the sampling distribution under the alternative.

(power <- sum(alt.prob[rej.indx]))
```

```
## [1] 0.0787655
```

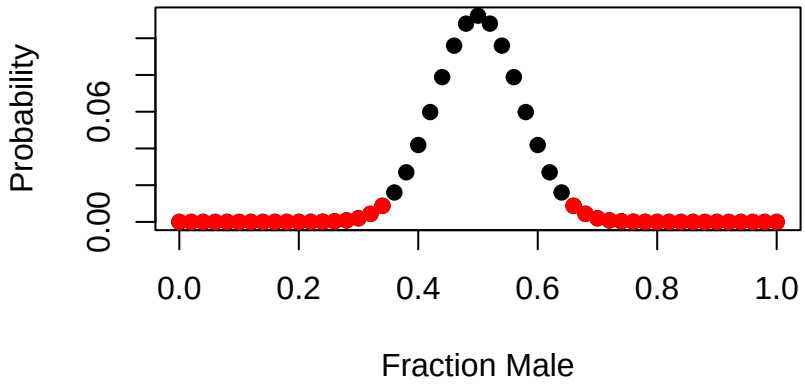
Increasing Power

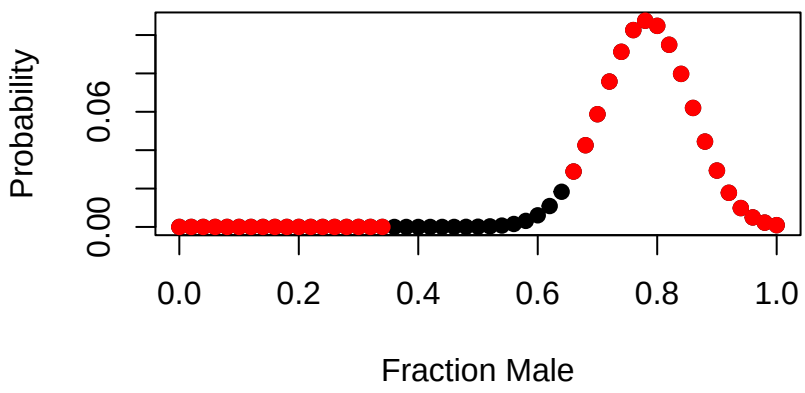
In this case, there are three ways to increase power. Really all we have here is a statement about what the world is like (the effect size), our sampling scheme (the sample size), and our choice about Type I error. Changing any one alters the power.

1. Power is larger when trying to detect larger effect sizes

For the same α value and the same sample size, power is higher when detecting larger effects.

```
sample.size <- 60
effect.size <- 0.15
alpha = 0.05
hyp.real.p <- null.p + effect.size
alt.prob <- dbinom(poss.num.males, sample.size, hyp.real.p)
par(mfcol = c(2,1))
plot(poss.frac.males, prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = 'Sampling Distribution under H0: Prob(Male) = 0.5')
points(rej.region, prob[rej.indx], pch = 19, col = 'red')
plot(poss.frac.males, alt.prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = paste('Sampling Distribution under HA: Prob(Male) = ',hyp.real.p,
                  sep = ''))
points(rej.region, alt.prob[rej.indx], pch = 19, col = 'red')
```

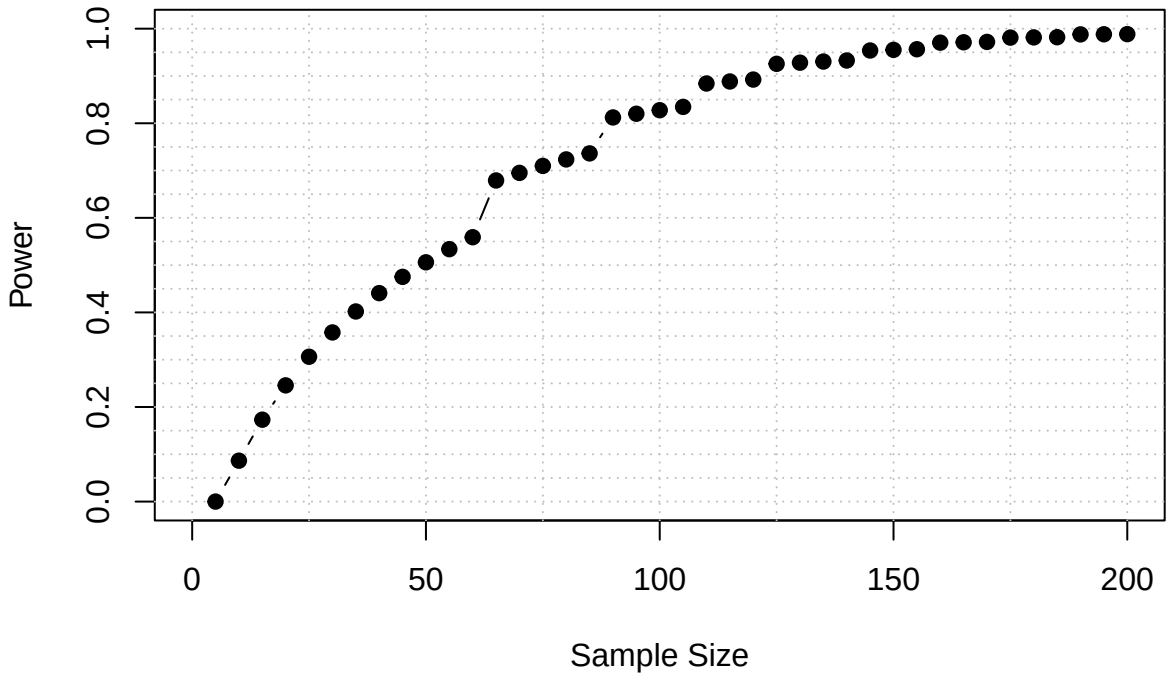





```
(power <- sum(alt.prob[rej.indx]))
```

```
## [1] 0.958361
```

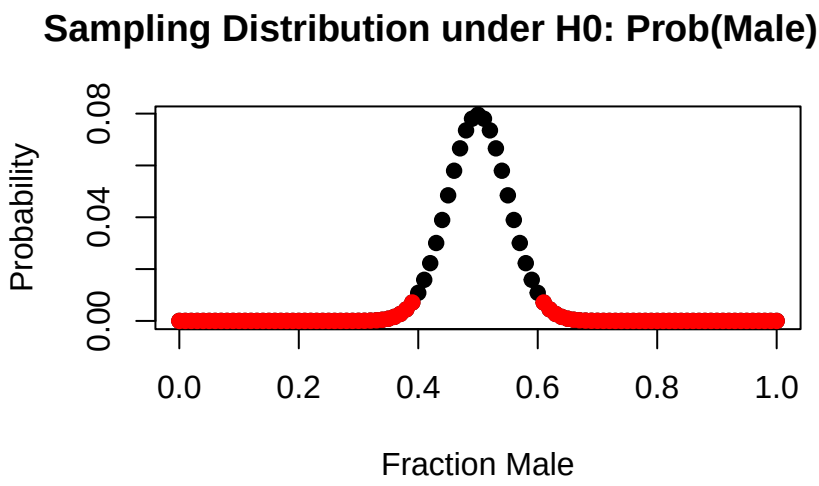
Here is a summary of this example with many different sample sizes.

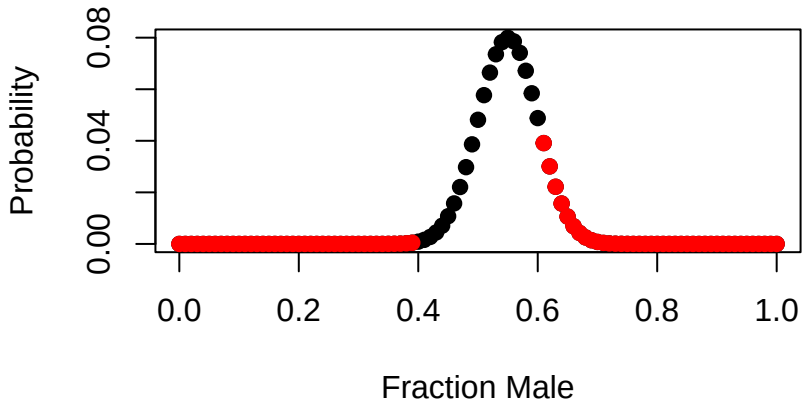


2. Power is larger with larger samples

For the same alpha value and the same effect size, power is higher when you have a larger sample size.

```
effect.size <- 0.05
sample.size <- 100
alpha <- 0.05
hyp.real.p <- null.p + effect.size
poss.num.males <- 0:sample.size; poss.frac.males <- poss.num.males/sample.size
null.prob <- dbinom(poss.num.males, sample.size, null.p)
alt.prob <- dbinom(poss.num.males, sample.size, hyp.real.p)
par(mfcol = c(2,1))
plot(poss.frac.males, null.prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = 'Sampling Distribution under H0: Prob(Male) = 0.5')
crit.test.values <- qbinom(c(alpha/2, 1 - alpha/2),
                           size = sample.size,
                           prob = null.p)/sample.size
rej.indx <- poss.frac.males < crit.test.values[1] | poss.frac.males > crit.test.values[2]
rej.region <- poss.frac.males[rej.indx]
points(rej.region, null.prob[rej.indx], pch = 19, col = 'red')
plot(poss.frac.males, alt.prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = paste('Sampling Distribution under HA: Prob(Male) = ',hyp.real.p,
                  sep = ''))
points(rej.region, alt.prob[rej.indx], pch = 19, col = 'red')
```



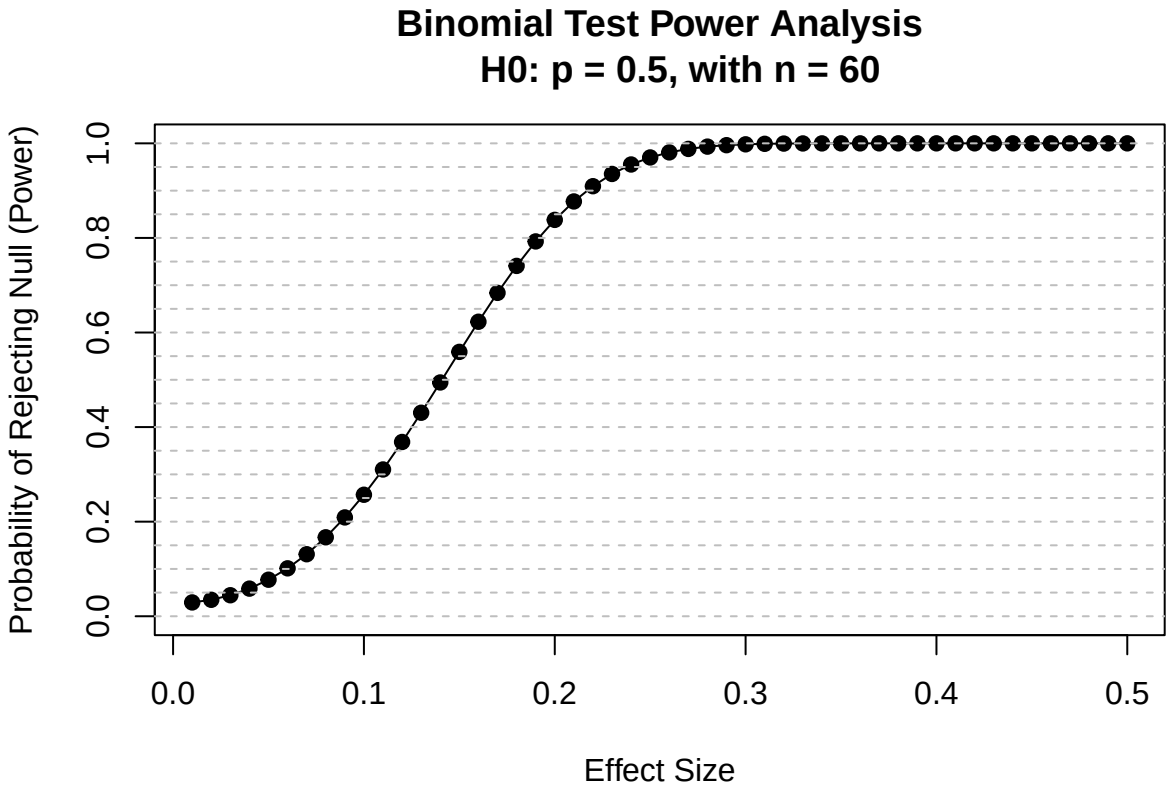



```
(power <- sum(alt.prob[rej.indx]))
```

```
## [1] 0.1351923
```

```
# Increasing the sample size here almost doubled the power.
```

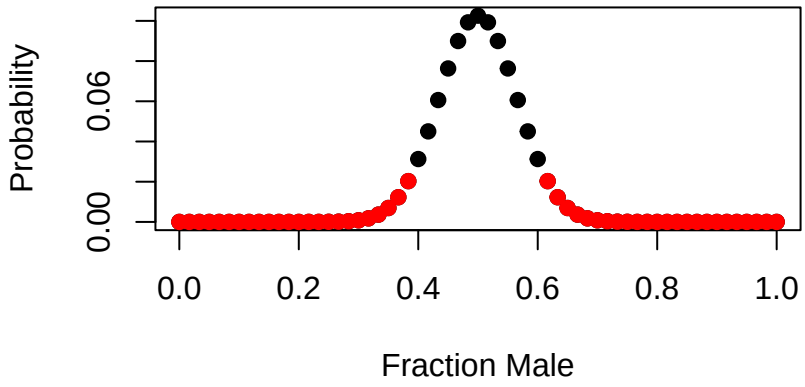
Here is a summary figure of how effect sizes influence power in this example of a binomial test.

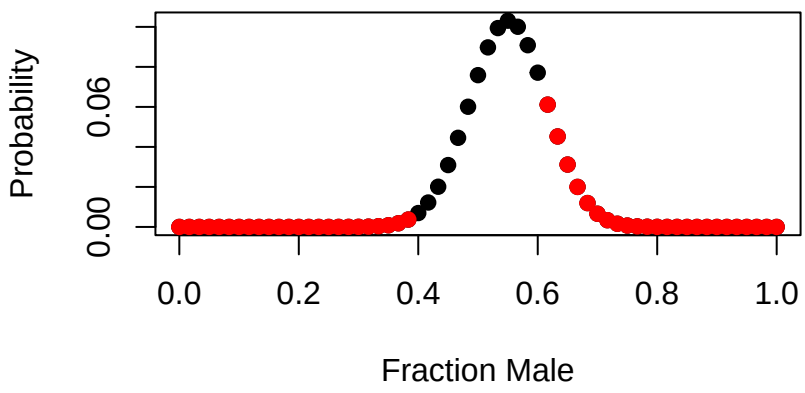


3. Power is larger with larger α values (Type I errors)

Last, increasing the type I error with the same sample size can also increase the power. But, of course, this trades off with Type I error. Let's increase Type I error to 0.1

```
effect.size <- 0.05
sample.size <- 60
alpha <- 0.1
hyp.real.p <- null.p + effect.size
poss.num.males <- 0:sample.size; poss.frac.males <- poss.num.males/sample.size
null.prob <- dbinom(poss.num.males, sample.size, null.p)
alt.prob <- dbinom(poss.num.males, sample.size, hyp.real.p)
par(mfcol = c(2,1))
plot(poss.frac.males, null.prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = 'Sampling Distribution under H0: Prob(Male) = 0.5')
crit.test.values <- qbinom(c(alpha/2, 1 - alpha/2),
                           size = sample.size,
                           prob = null.p)/sample.size
rej.indx <- poss.frac.males < crit.test.values[1] | poss.frac.males > crit.test.values[2]
rej.region <- poss.frac.males[rej.indx]
points(rej.region, null.prob[rej.indx], pch = 19, col = 'red')
plot(poss.frac.males, alt.prob, pch = 19,
     xlab = 'Fraction Male', ylab = 'Probability',
     main = paste('Sampling Distribution under HA: Prob(Male) = ',hyp.real.p,
                  sep = ''))
points(rej.region, alt.prob[rej.indx], pch = 19, col = 'red')
```



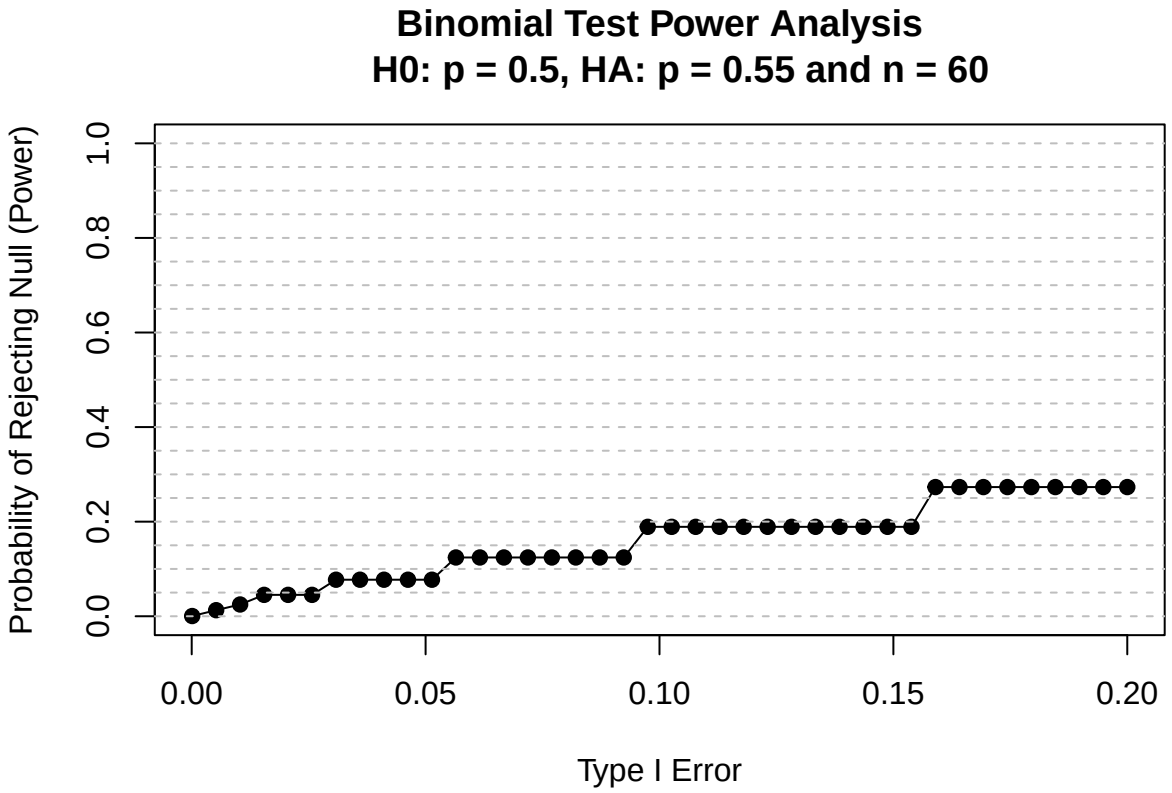



```
(power <- sum(alt.prob[rej.indx]))
```

```
## [1] 0.1890749
```

```
# Power in this case is about 19%.
```

Here is a summary of how changing the α value changes the power in this test.



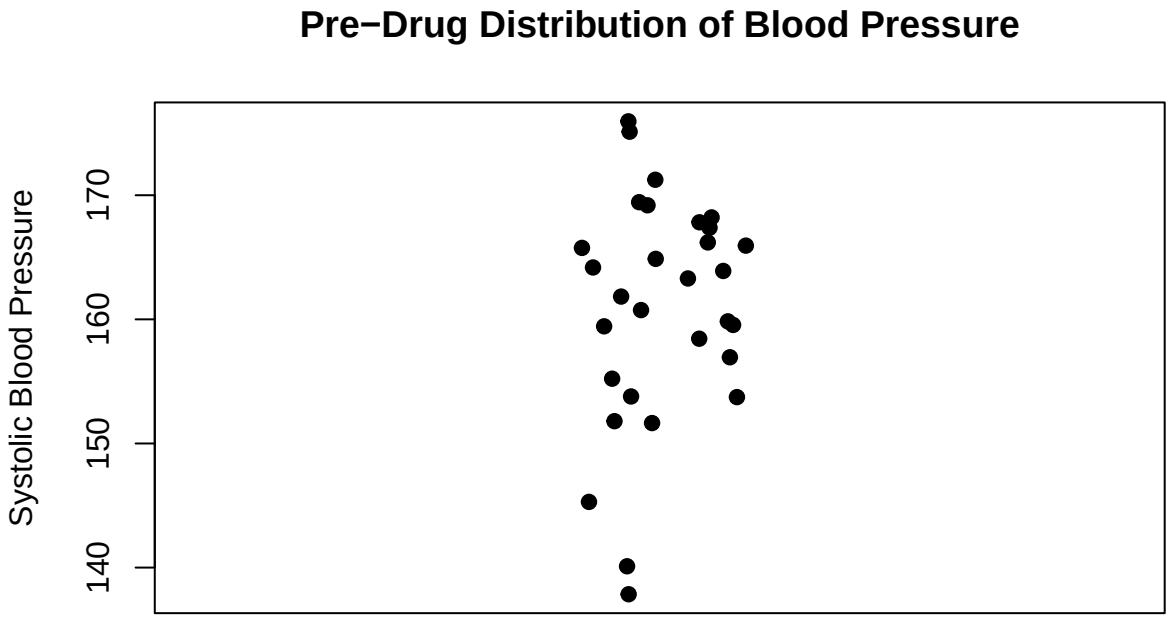
A last note about power: Power is larger in less variable populations (σ^2).

More variable populations made it more difficult to be certain of any particular population property. As such, there is generally lower power to estimate any particular effect. This is difficult to show in the case of the binomial distribution because the proportion of successes directly relates to the variability in successes. To take the sex ratio as an example, if the fraction of the population that is male is p, then the variability in sex in the population is $p(1-p)$.

But in other cases, we can imagine scenarios where we have the same sample size, same Type I error, and the same effect size, but one population is more variable than another. In that case, there is greater power in the population that is less variable.

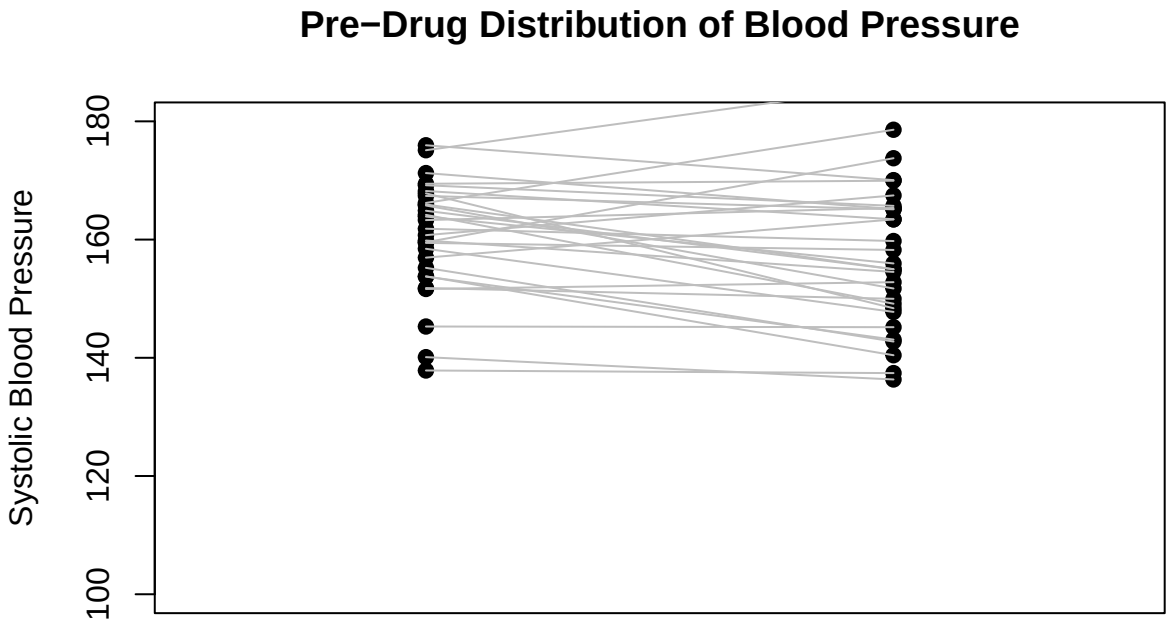
To see this, let's consider a paired t-test where we give a blood pressure drug to patients and measure their blood pressure before and after they receive the drug. Let's do this with, say, $n = 30$ people in our experimental trial. All have high blood pressure before receiving the drug. We will use systolic blood pressure as our measure. Here is the population.

```
n <- 30
set.seed(1)
pre.drug.bp <- rnorm(n, 160, 10)
stripchart(pre.drug.bp, method = 'jitter', vertical = T,
         pch = 19, ylab = 'Systolic Blood Pressure',
         main = 'Pre-Drug Distribution of Blood Pressure')
```



Now let's assume that some drug lowers blood pressure by an average of 5 points, but with some variability in effectiveness between individuals. We will call this `drug.effect.sd`. This means our effect size is 5. After taking the drug, the population has the following blood pressure.

```
effect.size <- -5
drug.effect.sd <- 8
post.drug.bp <- pre.drug.bp + rnorm(n, effect.size, drug.effect.sd)

stripchart(pre.drug.bp, vertical = T,
         pch = 19, ylab = 'Systolic Blood Pressure',
         main = 'Pre-Drug Distribution of Blood Pressure',
         at = 1, xlim = c(0.5,2.5), ylim = c(100, 180))
stripchart(post.drug.bp,vertical = T,
           pch = 19, add = T, at = 2)
for (i in 1:n){
  lines(c(1,2), c(pre.drug.bp[i], post.drug.bp[i]), typ = 'l',
        col = 'gray')
}
```



If we apply a paired t-test, we get

```
t.test(pre.drug.bp, post.drug.bp, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  pre.drug.bp and post.drug.bp
## t = 2.3674, df = 29, p-value = 0.0248
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.4869837 6.6705588
## sample estimates:
## mean difference
##        3.578771
```

To calculate the post-hoc power, we just bootstrap the drug effect, which mimics the idea of re-running the trial over and over again and evaluates our ability to detect an effect of a particular size under a particular test given sampling variability.

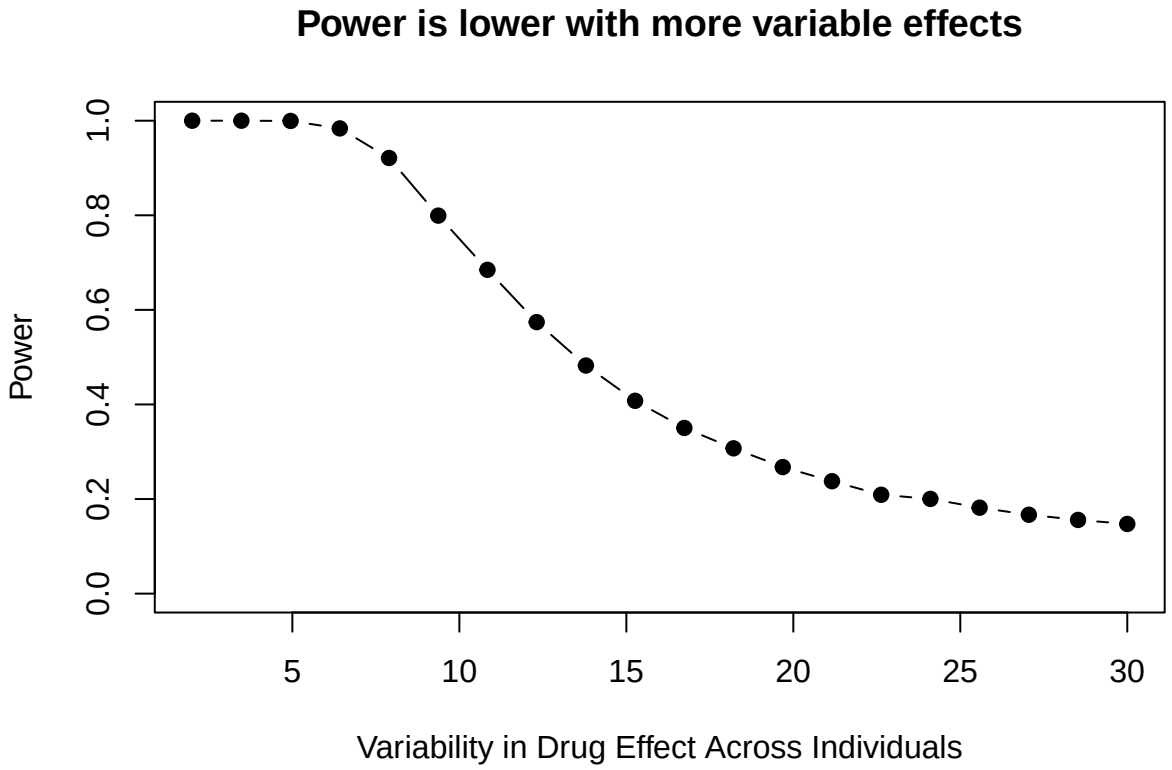
```
boot.samples <- 10000
test.p.value <- rep(NA, boot.samples)
for (i in 1:boot.samples){
  post.drug.bp <- pre.drug.bp + rnorm(n, effect.size, drug.effect.sd)
  boot.test <- t.test(pre.drug.bp, post.drug.bp, paired = TRUE)
  test.p.value[i] <- boot.test$p.value
}
power <- sum(test.p.value < 0.05)/boot.samples
power
```

```
## [1] 0.9103
```

Power in this case is pretty high.

To see how power is lower in more variable populations, we simply need to change how variable individuals are in their response to the drug, which we control with `drug.effect.sd`. Let's look at a lot of values and plot the power for each one.

```r
drug.effect.sd <- seq(from = 2, to = 30, length = 20)
power <- rep(NA, length(drug.effect.sd))
boot.samples <- 10000

for (j in 1:length(drug.effect.sd)){
  test.p.value <- rep(NA, boot.samples)
  for (i in 1:boot.samples){
    post.drug.bp <- pre.drug.bp + rnorm(n, effect.size, drug.effect.sd[j])
    boot.test <- t.test(pre.drug.bp, post.drug.bp, paired = TRUE)
    test.p.value[i] <- boot.test$p.value
  }
  power[j] <- sum(test.p.value < 0.05)/boot.samples
}

plot(drug.effect.sd, power, typ = 'b', pch = 19,
     xlab = 'Variability in Drug Effect Across Individuals',
     ylab = 'Power',
     main = 'Power is lower with more variable effects',
     ylim = c(0,1))
```



What this graph shows is that drugs with the same **average effect** are much more difficult to detect **if their effects are inconsistent across individuals**. This is important justification for doing preliminary data collection. It helps estimate the variability in the effects under study with smaller samples, and this can help determine what the sample size should be to have a good chance of detecting an effect.

Summary

- **Power** is the probability to reject a null hypothesis under a give effect size and sampling design

- **Effect size** is the difference between the null hypothesis and a particular value of the alternative
- Power can be determined by simulating a sampling distribution under a given effect size and then applying the test to the entire sampling distribution. Power is then the sum of the probabilities of all outcomes of the sampling distribution that leads to rejection of the null hypothesis.
- Power can also be evaluated by simulation. Repeatedly randomly sample from populations with a particular effect size and apply to the null hypothesis test to each random sample. The fraction of samples that reject the null is power.
- Power is larger for studies with larger sample sizes.
- Power is larger for studies detecting larger effect sizes.
- Power is larger for tests with larger Type I error probabilities (α).
- Power is larger for studies where populations are less variable.
- Power can be determined **prospectively**, that is, before a study is done. In such cases, the goal is to ask how many individuals should be sampled to detect an effect of a particular size. This is helpful in experimental design.
- Power can also be determined **retroactively**, that is, after a study has already been done. A particular study has an observed effect size that was not statistically significant, and the researcher wants to know whether they had low power to detect such an effect. This helps to answer the question of whether there really is no effect or whether their study design had little ability to detect a potentially subtle effect. If power is low for a study, there was little ability to detect the effect in the first place.