Section 7: some notes on inference and hypothesis testing

Advanced Quantitative Methods (PLSC 504)

Fall 2017

Inference and Hypothesis Testing (Review)

This should already be familiar to you from previous courses. We review some of the core concepts here mostly to illustrate how everything comes together, and because there appears to still be some confusion in the homework submissions.

Evaluating estimators

Let θ be our quantity of interest or estimand and $\hat{\theta}$ be our estimator. In statistics, we often talk about the distribution of $\hat{\theta}$. The distribution of $\hat{\theta}$ is called the **sampling distribution**. Why? Because in frequentist statistics we assume that our target parameter θ is fixed and that $\hat{\theta}$ is a random variable. Why is $\hat{\theta}$ a random variable? Because it depends on the data.

If $\mathbb{E}[\hat{\theta} - \theta] = \epsilon$ for $|\epsilon| > 0$ then our estimator is biased. **Bias** is a concept that does not depend on what happens as the sample size increases. The relevant concept here is **consistency**. We say that an estimator is consistent for some target parameter if it converges to the true value of the target parameter as we collect more and more data. Formally, $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{P} \theta$, where the $\xrightarrow{P}$ means "converges in probability".

What does this mean? It just means that the difference between $\hat{\theta}$ and θ approaches zero as we collect more data. Formally, $\Pr\left(|\hat{\theta} - \theta| > \epsilon\right) \to 0$ as $n \to \infty$. Here ϵ is a placeholder for the concept "very small". That is, the difference gets very small as we collect more data. Another way to think about it is that the sampling distribution of a consistent estimator becomes every more tightly concentrated around the truth as the sample size increases. Some estimators converge faster than others. In general, we would like to have an estimator that converges very fast.

One criteria that we have used to evaluate estimators is **mean squared error** (MSE). If $\hat{\theta}$ is our estimator then

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})$$

Let $\hat{\theta}_1, \hat{\theta}_2, \cdots$ be a sequence of point estimators for θ. A useful fact is that if $\text{MSE}(\hat{\theta_n}) \to 0$ as $n \to \infty$ then $\hat{\theta_n}$ is a consistent estimator. Why?

$$\begin{aligned}
\Pr\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) &= \Pr\left(|\hat{\theta}_n - \theta|^2 \geq \epsilon^2\right) \\
&\leq \frac{\mathbb{E}[\hat{\theta_n - \theta}]^2}{\epsilon^2} \\
&= \frac{\text{MSE}(\hat{\theta}_n)}{\epsilon^2}
\end{aligned}$$

The second line is true because of something called "Markov's Inequality" (see Blitzstein and Hwang, p. 429). So if $\text{MSE}(\hat{\theta}_n) \to 0$ as $n \to \infty$ then the estimator is consistent.

Let's bring in a simple example to illustrate some of these concepts: flipping a biased coin. Formally, let $X_1, \ldots, X_n \sim \text{Bern}(p)$ denote iid draws from the biased coin. Suppose that $p = 0.3$. I don't know this but I think I can nevertheless learn about it. Suppose there is a machine that will flip the coin n times for us.

Person 1 has a clever idea for an unbiased estimator: just record the first value that the flipping machine reports. Formally

$$\hat{p}_1 = X_1$$

This is, after all, an unbiased estimator. They even prove it,

$$\mathbb{E}[\hat{p}_1 - p] = \mathbb{E}[X_1] - \mathbb{E}[p] = p - p = 0$$

Person 2 proposes another clever estimator: "The sample mean plus $34/n$". They claim this is a better estimator because even though it its biased, it's consistent. They claim consistency is the most important criteria for an estimator. Person 3 proposes another estimator: "just take the sample mean, you idiots." How should we evaluate these estimators?

Perhaps using the MSE. For an unbiased estimator, the MSE is just the variance. What's the MSE of $\hat{p}_1$?

$$\text{MSE}(\hat{p}_1) = \text{Var}(X_1) = p(1-p)$$

$$\begin{aligned}
\text{MSE}(\hat{p}_2) &= \text{Var}\left(\bar{X} + \frac{37}{n}\right) + \left(\frac{37}{n}\right)^2 \\
&= \frac{p(1-p)}{n} + \left(\frac{37}{n}\right)^2
\end{aligned}$$

$$\begin{aligned}
\text{MSE}(\hat{p}_3) &= \text{Var}\left(\bar{X}\right) \\
&= \text{Var}\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2}\sum_i \text{Var}(X_i) \\
&= \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}
\end{aligned}$$

So, $\text{MSE}(\hat{p}_3) > \text{MSE}(\hat{p}_2) > \text{MSE}(\hat{p}_1)$. Which estimator is the "best"? It depends on our criteria. If we *just* cared about bias there would be no answer, since both $\hat{p}_1$ and $\hat{p}_3$ are unbiased. If we *just* cared about consistency there would also be no answer, since both $\hat{p}_2$ and $\hat{p}_3$ are both consistent. If we aim to minimize MSE, then we should pick $\hat{p}_3$.

The standard error of this estimator is $\sqrt{Var(\hat{p}_3)} = \sqrt{p(1-p)/n}$. We often don't even know the distribution that generates the data we see, so we don't know the standard error. If we knew what p was then we would know the standard error of the estimator here. But we don't know p. Instead we estimate it: $\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$.

```
# Our parameter. It is fixed.
p <- 0.3

# Simulate from our DGP. Suppose n = 100.
n <- 100
X <- rbinom(n = n, size = 1, prob = p)

# What is our estimate of p?
p_hat <- mean(X)
p_hat
```

```
## [1] 0.29
```

```r
# What is our *estimated* standard error.
se_hat <- sqrt(p_hat*(1-p_hat)/n)
se_hat
```

```
## [1] 0.04537621
```

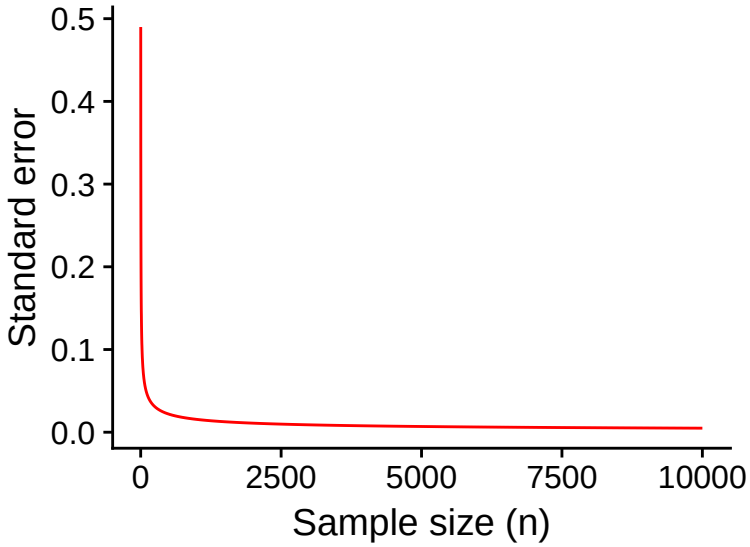
```r
# But what is the true standard error of our *estimator*, p_hat?
se <- sqrt(p*(1-p)/n)
se
```

```
## [1] 0.04582576
```

```r
# And what is the true variance of X, our Bernoulli RV?
p*(1-p)
```

```
## [1] 0.21
```

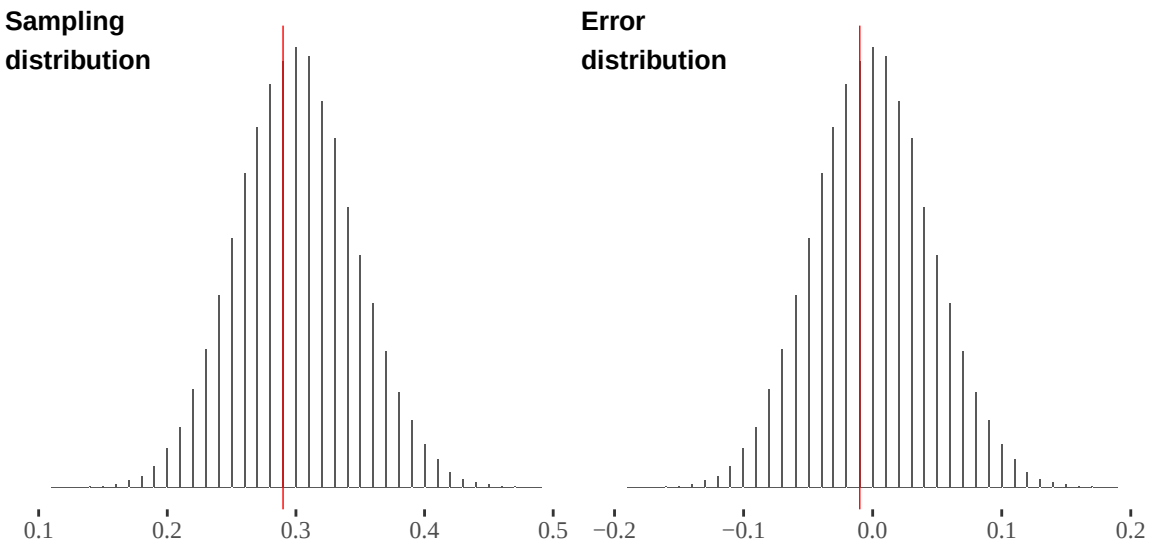
Does the Bernoulli distribution have a "standard error"? No. It's not an estimator. The standard error is the square root of the variance of an *estimator*. Does the sample mean of X have a standard error? Yes. Why? Because it's an estimator (of p). It has a variance. Is this a consistent estimator for p? Yes. Why? Because the bias (in this case it's just zero) and the se vanish as n approaches infinity. We can see this analytically, but here is a picture:



What is the sampling distribution of our estimator for p? We don't get to know this from taking one draw. Typically all we get to see is one estimate, but we can always do a Monte Carlo simulation:

```r
n <- 100
p <- 0.3
nsims <- 10^5
p_hat_sim <- rep(NA, nsims)
for(i in 1:nsims) {
  X <- rbinom(n = n, size = 1, prob = p)
  p_hat_sim[i] <- mean(X)
}
```

Now we have simulated the sampling distribution of our estimator. We can also look at the **error distribution**, $\hat{p}_3 - p$. Let's make a picture. Here the red line in the left panel denotes our observed value of $\hat{p}_3$ (0.29), the estimate we actually saw. The red line in the right panel denotes the observed difference between p and this single estimate.



So the estimate we actually observed was pretty close to the center of the sampling distribution (and error distribution), as expected. We note that the pictures above have a familiar shape.

This is to be expected because our estimator is **asymptotically Normal**. In general we say that an estimator $\hat{\theta}$ is *asymptotically Normal* if $\hat{\theta} \approx N(\theta, \hat{se}^2)$. In our example, this means that our estimator has a limiting distribution with mean p and variance, $\left(\sqrt{\hat{p}(1-\hat{p})/n}\right)^2$.

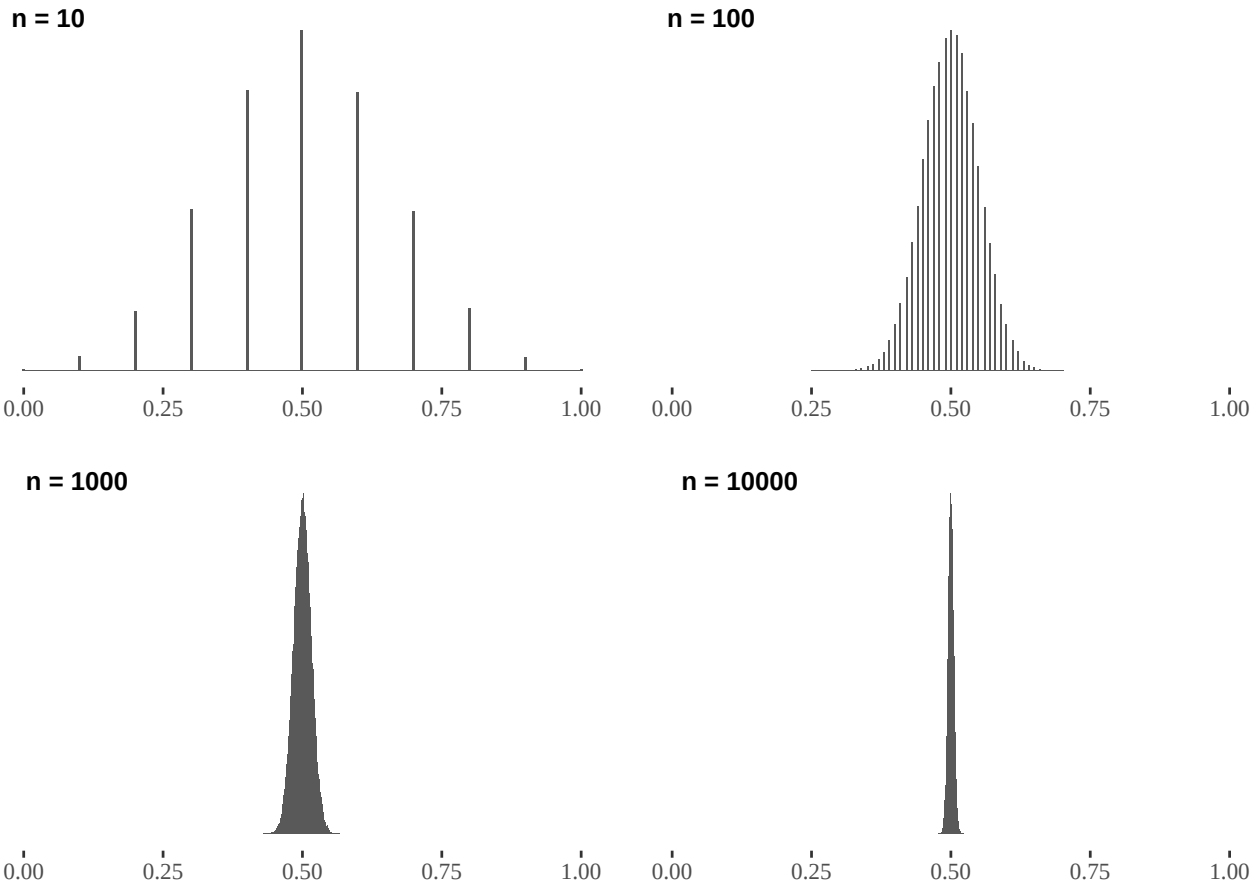
This follows from the Central Limit Theorem. Why is this useful? Because we can construct approximate confidence intervals for p. Why are they "approximate"? Because we appeal to the Normal distribution. These interval only have correct coverage in large samples. What do we need to construct an 89% confidence interval for p? Just a single draw. In this case, we can do it with the 100 coin flips the machine gave us,

```
alpha <- 0.11
ub <- p_hat + qnorm(1-alpha/2)*sqrt(p_hat*(1-p_hat)/n)
lb <- p_hat - qnorm(1-alpha/2)*sqrt(p_hat*(1-p_hat)/n)

c(lb, ub)
```

```
## [1] 0.2174801 0.3625199
```

Wider confidence intervals imply uncertainty. An estimator with a wide sampling distribution (e.g. $\hat{se}$ is large) will have wide confidence intervals. The width of the sampling distribution for p is a function of two things: 1) the sample size n, and 2) the variance $p(1-p)$. Note that the variance is maximized when $p = 1/2$.[1] So when $p = 1/2$ and n is small, the width of the sampling distribution will be largest. The plot below illustrates this concept. I fix $p = 1/2$ for each, but increase n. The estimator is the sample mean, which we know is unbiased and consistent for p.

[1] To see this, take the first derivative and solve for p, then apply the second derivative test.


n = 10
n = 100
n = 1000
n = 10000
0.00
0.25
0.50
0.75
1.00

Hypothesis testing

Null Hypothesis Significance Testing (NHST) is the canonical paradigm in political science. You may be aware of the ongoing discussion about abandoning the NHST paradigm. We won't cover other paradigms in this course (like Bayesian inference for example). These discussions have been going on for many many years and little has changed. NHST is almost certainly the framework you will be operating under during your research life as a political scientist. Indeed reviewers of empirical research will expect you to perform this ritual over and over again.[2] It's important to understand what's going on in order to be a critical consumer of quantitative research. The NHST recipe:

1. Decide on a test statistic, e.g. $T(Y_i, W_i)$ if our estimator is the difference in means or Horvitz-Thompson in the case of an experiment. This is some function of the observed data. This function can be as complicated as you want it to be. Our $T(\cdot)$ has been the difference in means in most of the homeworks.
2. Derive the distribution of $T(\cdot)$ under the null hypothesis. How do we do this? **We assume the null is true**. In practice, the null hypothesis is almost always that some difference is zero.
3. Calculate the p-value: the probability of seeing a test statistic at least as extreme as what we in fact see, **assuming the null is true**.
4. If $p < 0.05$ reject the null. (Optional: put a * in your regression table next to the estimate you obtained.)

Is $p < 0.05$ an arbitrary threshold? Yes, completely. Who decided that we should use $p < 0.05$? A guy called Ronald Fisher, technically "Sir Ronald Fisher":

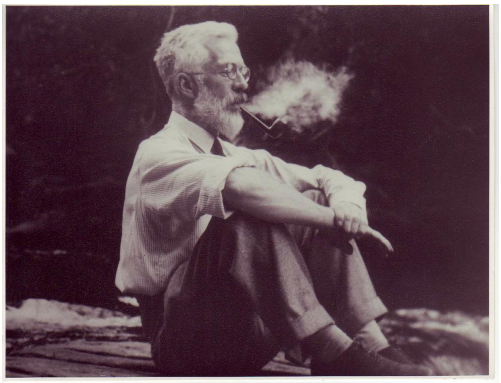


Figure 1: Ronald Fisher smoking a pipe.

> "The value for which P=.05, or 1 in 20, is 1.96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."
> -*Statistical Methods for Research Workers*, 1925.

For trivia night: Fisher smoked tobacco from a pipe. He was also famous for publishing his doubts about the link between smoking and lung cancer, arguing that "correlation isn't causation"[3]. He was also a consultant for tobacco companies, and a prominent supporter of eugenics. He died (from complications related to a colon cancer surgery) in Adelaide, Australia. You can visit his remains, buried under the floor in St Peter's Cathedral, if you wish. We pay a little tribute to his ghost every time we declare $p < 0.05$.

Let's consider a concrete example of how hypothesis testing is applied in political science. A professor might download the 2016 American National Election Study (ANES) survey, type "reg y x" into a software package and conclude from the output that the "effect" of "being a Republican" on "support for welfare" is negative because the estimated coefficient on our predictor is negative and the relationship is "statistically significant".

What has happened here? He has used an estimator, call it $\hat{\beta}$, to obtain a single estimate, call this $\hat{\beta}_1$, about the association between y and x. He has assumed a null hypothesis. Namely, that $\beta = 0$. But the data

[2] Some might argue that this is "Cargo Cult Science" (http://calteches.library.caltech.edu/51/2/CargoCult.htm)
[3] for example, https://www.nature.com/nature/journal/v182/n4635/abs/182596a0.html

suggest, on the contrary, that $\beta \neq 0$. That is, the probability of observing $\hat{\beta}_1$ under the null is very unlikely. How do we make sense of this? Because $\hat{\beta}$ has a sampling distribution: $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_2, \cdots$. Some values are more likely to present themselves to us than others.

The logic of hypothesis testing is like proof by contradiction. We want to show that $\beta \neq 0$, so we assume that it is in fact true and then ask – what's the likelihood of seeing the data we observe if this null hypothesis is indeed true. In symbols, we have some null hypothesis H (e.g. Republicans are indifferent about welfare), and then we look at some data D (e.g. relationship between support for welfare and party identification in a survey) to evaluate $\Pr(D|H)$. That is, what is the probability of seeing this data given our hypothesis is true? Note that we don't evaluate the likelihood that a null hypothesis is true. That is, we don't "accept the null". We simply "fail to reject it".

Let's return to the coin. We want to test whether the coin is fair (e.g. $p = 0.5$). So we need a null hypothesis (H_o) – the thing we want to refute. In this case, the null is: "the coin is fair" (e.g. $H_o : p = 0.5$). What's the alternative? The coin is not fair (e.g. $H_A : p \neq 0.5$).

Now comes the harder part. What test statistic should we choose? Remember that the test statistic is a function of the observed data. In this setting, the observed data are 100 flips from the machine. If the null is true, we would expect about 50/100 of these flips to land heads. Why? Because $p = 0.5$ under the null, but the coin flipping process is not deterministic.

In setting up our test, we **assume the null is true**. A reasonable test statistic is the number of heads in 100 flips. Next, we derive the distribution of our test statistic under the null. In this case, we know the distribution of n Bernoulli trials is a Binomial. So the null distribution is $f(x \mid p = 0.5) \sim \text{Bin}(n = 100, p = 0.5)$. How many heads do we expect to see if the null is true? About 50. So if we see 49 or 51 heads, can we conclude that the coin is not fair? We could, but that would be pretty unreasonable. Why? Because the likelihood of observing 49 or 51 heads in 100 flips is about the same as the likelihood of observing 50 heads.

Instead, we should just consider extreme deviations from 50, like 80 out of 100, to be evidence against the null. But what about 60 out of 100? Does that also count as "extreme"? Well, that's subjective, but we can nonetheless be precise about it. In order to do this, we need to define a **rejection region** such that if we observe a test statistic in this region we will reject the null. In symbolcs, we reject if $T(X) \in R$. Suppose we decide on the following rejection region: $\{0, 1, \ldots, 39, 61, 62, \ldots, 100\}$. That is, if we observe less than 40 heads or more than 60 heads in 100 tosses we will reject the null hypothesis that the coin is fair. Is this a reasonable test?

In order to answer this question, there are four more things I might want to know. First, what's the likelihood that I claim the coin is unfair when it really is fair? Second, what's the likelihood that I claim the coin is unfair when it really is unfair? Third, what's the likelihood that I claim the coin is fair when it actually is? Fourth, what's the likelihood that I claim the coin is fair when it is not? Table 1 provides a summary.

Table 1: Important probabilities for a test

$\Pr(\text{reject } H_o \mid H_o)$	$\Pr(\text{reject } H_o \mid H_a)$
$\Pr(\text{retain } H_o \mid H_o)$	$\Pr(\text{retain } H_o \mid H_a)$

The diagonal quantities in Table 1 correspond to error probabilities. Ideally, we would like these to be small, whereas we want to off diagonals to be large. $\Pr(\text{reject } H_o|H_o)$ is the "significance level" of our test. Convention is to name this α. The smaller our significance level the less likely it is that our test throws a "false positive" (e.g. we reject the null when we should not). We call $\Pr(\text{reject } H_o|H_a)$ the "power" of a test. Convention is to denote this $1 - \beta$, where $\beta = \Pr(\text{retain } H_o|H_a)$. The larger this value is the less likely it is that our test throws a "false negative" (e.g. we don't reject the null when we should). The first error is called a "Type I" error and the second is called a "Type II" error.

I think the easiest way to remember the distinction is to think about a hypothetical jury trial. In the United States a defendant is (in theory) "innocent until proven guilty". So the null hypothesis is that the defendent is innocent. The alternative is that the defendant is guilty. α is the probability of convicting the innocent, β

is the probability of letting the guilty go free, and so $1-\beta$ is the probability of convicting the guilty. The prosecutor cares more about a small β than a small α. The defense attorney, on the other hand, cares more about having a small α than a small β. A benevolent dictator might make them both small.

Now let's go back to our test of the coin. What is α here?

$$\begin{aligned}\Pr(\text{claim coin unfair} \mid \text{coin is fair}) &= \Pr(\text{reject } H_o | H_o) \\ &= \Pr(T(X) \in R |\ p = 0.5) \\ &= \Pr(\bar{X} < 0.4 \mid p = 0.5) + \Pr(\bar{X} > 0.6 \mid p = 0.5) \\ &\approx 0.046\end{aligned}$$

We know the distribution so we can calculate this exactly. It's very easy to do this in R,

```
# Using the CDF:
pbinom(40, size=100, prob  = 0.5) + (1-pbinom(60, size=100, prob  = 0.5))
```

```
## [1] 0.04604407
```

```
# Alternatively, using the PDF:
sum(dbinom(0:40, size=100, prob  = 0.5)) +
  sum(dbinom(61:100, size=100, prob  = 0.5))
```

```
## [1] 0.04604407
```

Now, what is $1-\beta$, the power of this test?

$$\begin{aligned}\Pr(\text{claim coin unfair} \mid \text{coin is unfair}) &= \Pr(\text{reject } H_o | H_a) \\ &= \Pr(T(X) \in R |\ p \neq 0.5) \\ &= \Pr(\bar{X} < 0.4 \mid p \neq 0.5) + \Pr(\bar{X} > 0.6 \mid p \neq 0.5)\end{aligned}$$

This is somewhat of a harder problem. Why? Because we also need to speculate about various values for p. For example, if $p = 0.6$ then the power of this test is,

```
pbinom(40, size=100, prob  = 0.60) + (1-pbinom(60, size=100, prob  = 0.60))
```

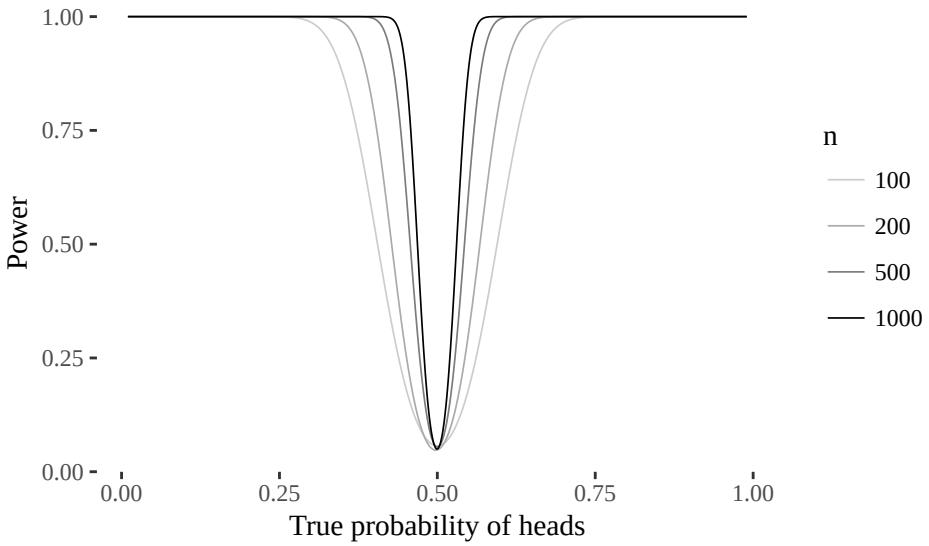
```
## [1] 0.4621178
```

So this test will reject the null when the null is false less than half the time if $p = 0.6$. What if $p = 0.7$?

```
pbinom(40, size=100, prob  = 0.70) + (1-pbinom(60, size=100, prob  = 0.70))
```

```
## [1] 0.9790114
```

This is a huge improvement. But we see that the power of the test crucially depends on how biased the coin is. If the bias is small, it's really hard to distinguish the unfair coin from a fair one. To visualize this, let's fix α at the conventional 0.05 and experiment with different values of n.



The moral of the story is that if you're looking for a small effect, you had better collect a heap of data.

The Permutation Test

We usually do hypothesis tests and compute p-values using an approximate approach based on the Normal distribution (see Aronow and Miller, Chapter 3). In our coin flipping example we know the parametric distribution, so we don't need to use the large sample theory approximation approach. That is, we could calculate an "exact" p-value here. However, this is still a parametric test, because we are using the binomial distribution. A permutation test is another method for calculating an exact p-value that does not rely on any particular distribution. In this sense, it is true to say that a permutation test is both exact and "non-parametric".

Who came up with this approach to hypothesis testing? It was, again, Ronald Fisher. Allegedly Fisher came up with the idea after meeting a "tea lady" (Dr. Muriel Bristol) who claimed that tea tasted different if milk was poured into the cup before or after the tea. Of course, Fisher didn't believe her. Thus we have the infamous "Lady Tasting Tea" problem:

> "A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. [...] Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject in random order. [...] Her task is to divide the cups into two sets of 4, agreeing, if possible, with the treatments received. [...] The element in the experimental procedure which contains the essential safeguard is that the two modifications of the test beverage are to be prepared "in random order". This is in fact the only point in the experimental procedure in which the laws of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced. [...] it may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged."
> - *The Design of Experiments* (1935)

So we have $n = 8$ cups. We randomly choose 4 cups to pour in the tea first, and the remaining 4 cups get the tea poured in second. There are $\binom{8}{4} = 70$ ways to do this. What's the null hypothesis here? Of course, it's that the lady cannot actually tell which of these 8 cups should be labeled T (tea first), and which should be labeled $\neg T$ (milk first). A reasonable test statistic is the number of cups correctly classified. Unlike in the previous examples, there is no need to specify an alternative hypothesis here.

Under the null, all 70 permutations of labels are equally likely to occur. Suppose our permutation of cups in the experiment is $\{\neg T, \neg T, T, T, \neg T, T, \neg T, T\}$. This occurs with probability 1/70. In the real example,

Muriel classified all 8 of the cups correctly, e.g. her labeling was $\{\neg T, \neg T, T, T, \neg T, T, \neg T, T\}$. What's the probability that Muriel came up with this by chance? It's $1/70 \approx 0.01$. This is the p-value.

The Sharp Null

In causal inference, the sharp null hypothesis states that the treatment effect is zero **for all units**. This uses the permutation test logic, and it is very easy to test because it can always be specified. Formally,

$$H_o : \tau_i = Y_i(1) - Y_i(0) = 0 \quad \forall i$$

Note that the null hypothesis of no average treatment effect ($\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = 0$) **does not imply the sharp null**. Why?

```
Y1 <- c(1:20, 25, 24)
Y0 <- c(1:20, 24, 25)

tau <- Y1-Y0
mean(tau)
```

```
## [1] 0
```

```
tau
```

```
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1 -1
```

Let's consider a simple example to fix the concept. Suppose that we observe $Y_i = (3, 5, 7, 9)$ and $W_i = (0, 0, 1, 1)$, where W_i is allocated using complete random assignment. The sharp null means that $Y_i(1) = Y_i(0) = (3, 5, 7, 9)$. What is the null distribution here? Well, there are $\binom{4}{2} = 6$ possible ways to assign treatment. Suppose our test statistic is the difference in means, $T(Y_i, W_i) = 1/2 \sum_i Y_i W_i - 1/2 \sum_i Y_i(1 - W_i)$. The distribution of our test statistic under the null is displayed in Table 2.

Table 2: Null distribution

Permutation	$T(Y_i, W_i)$	Probability
(1,1,0,0)	-4	1/6
(1,0,1,0)	-2	1/6
(0,0,1,1)	4	1/6
(0,1,0,1)	2	1/6
(0,1,1,0)	0	1/6

What's the observed value of our test statistic?

```
t_ate <- function(y, w) {
  N <- length(w)
  m <- sum(w)
  sum(w*y)/m - sum((1-w)*y)/(N-m)
}

Y <- c(3, 5, 7, 9)
W <- c(0, 0, 1 , 1)
t_ate(Y, W)
```

```
## [1] 4
```

The (two-sided) p-value is the proportion of test statistics as extreme as the one we observed under the null of no treatment effect for any unit: $\Pr(|T| \geq 4|H_o) = 2/6 = 1/3$. Typically it is not feasible to write down

all possible permutations under the null even when we know the distribution. For example, I didn't even do this with the lady tasting tea example, which is relatively simple. But it gets ugly very fast. Imagine that we have $N = 20$ units and want to put 10 in treatment and 10 in control. This yields 184,756 potential randomizations.

When we are in a situation like this, we just simulate the null distribution because it's 2017 and computation is cheap. The p-value is then the fraction of times that the simulated test statistic is more extreme than our observed test statistic. Let's try an example with 20 units, using the same test statistic as before.

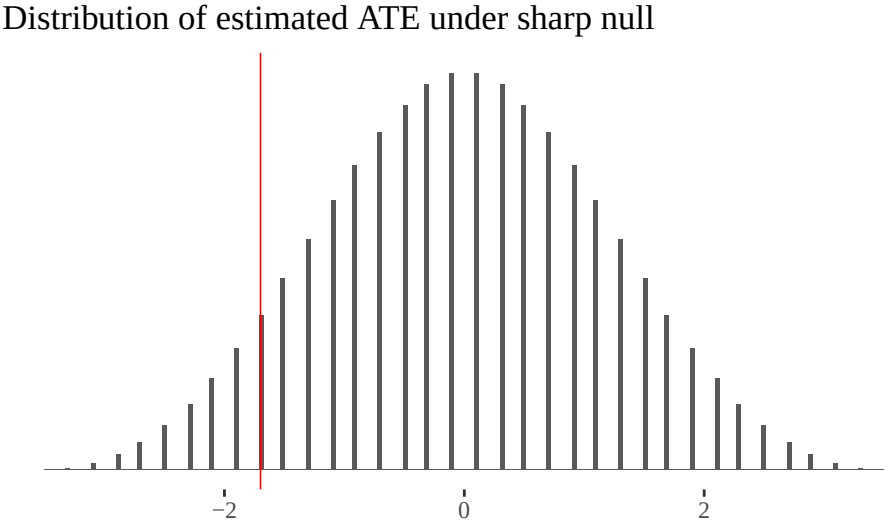
```
Y <-  c(8, 3, 1, 0, 4, 1, 6, 1, 1, 1, 2, 2, 1, 0, 0, 2, 2, 2, 9, 5)
W <- c(0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0)
t_obs <- t_ate(Y, W)
t_obs
```

```
## [1] -1.7
```

```
# Generate matrix of all possible permutations of treatment assignment
permuter <- function(n, m) {
  combn(1:n, m, function(x) replace(rep(0, n), x, 1))
}

# A 20 x 184,756 matrix of potential randomizations
theperms <- permuter(n = 20, m = 10)

# Distribution of the ATE under the sharp null. This is *every* possible value
#   of the test statistic that could be obtained under the sharp null.
null_dist <- apply(theperms, 2, function(x) t_ate(Y, x))

# The exact p-value
sum(abs(null_dist) >= abs(t_obs))/ncol(theperms)
```

```
## [1] 0.1757994
```

In this setting we would fail to reject the sharp null hypothesis if our standard was $p < 0.05$. The plot below illustrates this logic graphically. The red line here is the observed value of our test statistic.



Although we still don't need to invoke any distributional assumptions in this situation, there might nevertheless be too many permutations to use an exact method. The method still works pretty good even if we just take, say 2,500 random draws from the null distribution with the `sample()` function. Note that in the best case scenario where each of the 2500 random assignments was unique we would only see about 1% (2500/184756)

of the null distribution. This tiny fraction is, of course, the upper bound of what we will see using this inefficient method of simulating the null since we cannot even rule out sampling the same permutation of the treatment assignment vector more than once. Nonetheless, the p-value we get, even though an approximation, is pretty close to the "exact" p-value.

```
nsims <- 2500
t_sims <- rep(NA, nsims)
for(i in 1:nsims){
  w <- sample(W)
  t_sims[i] <- t_ate(Y, w)
}
# Approximate p-value
sum(abs(t_sims) >= abs(t_obs))/nsims
```

```
## [1] 0.1784
```