Gov 2000 - 5. Univariate Inference I

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

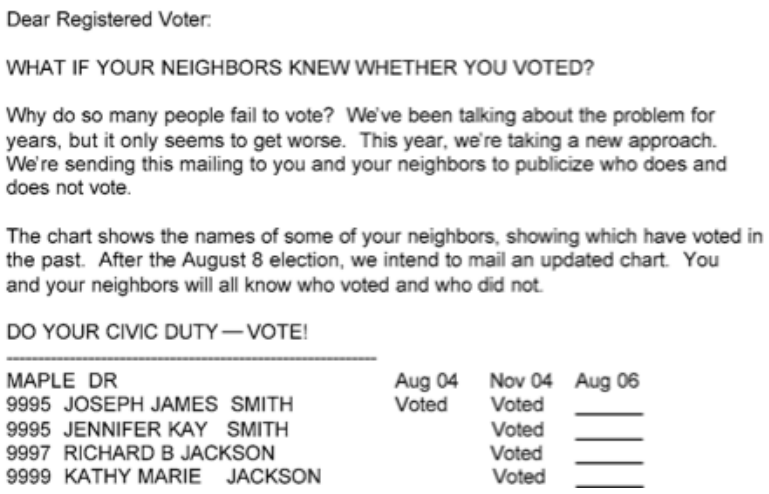
Where are we? Where are we going?

- For the last few weeks we have talked about probability—that is, if we knew how the world worked, we are describing what kind of data we should expect.
- Now we want to move the other way. If we have a set of data, can we estimate the various parts of the probability distributions that we have talked about. Can we estimate the mean, the variance, the covariance, etc?
- Moving forward this is going to be very important. Why? Because we are going to want to estimate the population conditional expectation in the linear regression model.

INTRODUCTION

Motivating example

- Gerber, Green, and Larimer (APSR, 2008) studied an experiment where they random assigned some households to get the following mailer:



Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____

- They wanted to know if social pressure got people to turn out more than a mailer that just emphasized a person's civic duty to vote. Let's just load up their data and see what they find:

```
load("gerber_green_larimer.RData")
## turn turnout variable into a numeric
social$voted <- 1 * (social$voted == "Yes")
neigh.mean <- mean(social$voted[social$treatment == "Neighbors"])
neigh.mean
```

```
## [1] 0.3779482
```

```
contr.mean <- mean(social$voted[social$treatment == "Civic Duty"])
contr.mean
```

```
## [1] 0.3145377
```

```
neigh.mean - contr.mean
```

```
## [1] 0.06341057
```

- We've estimated a difference of roughly 6.3 percentage points between the Neighbors group and the Civic Duty group. Is this a big effect?
- But this estimate depends on the sample that we took. What if we had taken a different sample from the population? Would the difference in means be? How can we tell if there is a "real" effect here or its just due to random chance?

Goal 1: Inference

- Inference: given that we observed this difference in means, what is our best guess about where it will be in other samples.
- In order to think about this, we need to think about the distribution of estimates across samples will look like. This can get fairly confusing, but we'll try to be as clear as possible.
- Last week we did repeated sampling to show how estimates differ from sample to sample. But obviously then we had access to the population (Fulton County, remember). Usually we will not have that—we don't know what the true difference in means is for this experiment. So why do we still want to know about how the estimator acts across samples?
- Remember back to the Lady Tasting Tea example in the first lecture. There we did a statistical though experiment. What would the world look like if the Lady was guessing at random? Here we are going to do the same thing. What would the world look like *if* there was no difference in means? Would our particular estimate of the difference in means be fairly typical or unusual?
- Basically, the sampling distribution will help us answer the following question: would an estimate like our observed estimate be unusual in some hypothetical world?

Goal 2: Compare estimators

- Above we use the simple difference in sample means ($\bar{Y} - \bar{X}$) to estimate the population difference in means. But there are other, alternative estimators for that difference. One is the post-stratification estimator, where we estimate the estimate the difference among two subsets of the data (male and female, for instance) and then take the weighted average of the two:

$$\hat{\theta}_{ps} = (\bar{Y}_f - \bar{X}_f)\bar{Z} + (\bar{Y}_m - \bar{X}_m)(1 - \bar{Z})$$

- Should we use the simple estimator or the post-stratification estimator? It's probably unclear from first glance (it turns out that in this case, the answers is that it depends on the population distribution). But, in general, we need to be able to compare different estimators to see which is the most appropriate in a given situation.
- We have to decide between estimators all the time: what is the best way to estimate the conditional expectation function, $\mathbb{E}[Y|X]$? What's the best way to

deal with missing data? How best to design an experiment? How to best to measure some concept like ideal points? For each of these empirical problems, you're going to need to understand how well an estimator performs at that task.

POPULATIONS AND SAMPLES

- In the last few weeks, we have talked about distributions and their characteristics—means, variances, covariances. I said before that we're going to want to estimate those properties, but we've been vague about what that means.
- **Probability** was about saying what types of data/outcomes we should expect given some distribution. **Inference** is about learning about the distribution given some data/outcomes.

Populations

- Typically, we want to learn about the distribution of random variable (or set of random variables) for a **population** of interest. As an example, we might want to know the distribution of votes for Hillary Clinton in the population of registered voters in the United States. This is an example of a finite population.
- Sometimes the population will be more abstract, such as the population of all possible television ads. This is an example of an infinite population.
- With either a finite or infinite population our main goal in inference is to learn about the **population distribution** or particular aspects of that distribution, like the mean or variance, which we call a **population parameter** (or just parameter).
- We sometimes call the population distribution the **data generating process** and represent it with a pmf or pdf, $f(x;\theta)$. Ideally we would place no restrictions on f and learn everything we can about it from the data. This **nonparametric** approach is difficult due to the fact that the space of possible distributions is vast! Instead, we will often make a **parametric** assumption and assume that the formula for f is known up to some unknown parameters.
- Thus, f has two parts: the known part which is the formula for the pmf/pdf (sometimes called the parametric model and comes from the distributional assumptions) and the unknown part, which are the parameters, θ. For instance, suppose we have a binary r.v. such as intending to vote for Hillary Clinton

($X = 1$). Then we might assume that the population distribution is Bernoulli with unknown probability of $X = 1$, θ. The we would have:

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}$$

Our goal is to learn about the probability of someone voting for Hillary Clinton from a sample of draws from this distribution.

- Probability tells us what types of samples we should expect for different values of θ.
- For some problems, such as estimating the mean of a distribution, we actually won't need to specify a parametric model for the distribution as we'll see.

Sampling

- With inference, we have a collection of data that (we assume) come from a common probability distribution—the population distribution. In this case, we will say that $X_1, X_2, \ldots, X_n$ are an independent and identically distributed set of random variables. We often shorten this to say that they are i.i.d. with pmf/pdf $f(x; \theta)$.
- What does this mean, exactly? Well, independent just means that each is independent of the other. Identical just means they all come from the same distribution. This could be Bernoulli with p or Normal with mean μ and σ^2.
- We call iid samples from a distribution **random samples**.

Example: finite populations and Fulton County

- With finite populations, we can imagine taking a random sample of the population values. Then the population probability distribution is just the distribution of the population data. For instance, suppose our population of interest was all registered voters in Fulton County, GA and we were interested in whether they turned out to vote ($X = 1$) or not ($X = 0$).
- In this case, we want to learn about this specific population, not the general infinite population of potential voters in Fulton County. Conveniently, we have access to data on the entire population of registered voters in Fulton County:

```
## load file of all registered voters in
load("fulton.RData")

## size of the population
nrow(fulton)
```

```
## [1] 339186
```

- The population distribution of turnout can be completely summarized by the expected value, which we can calculate from this data:

```
## calculate the population mean/proportion of people turning out
## this is a little pedantic because we are using
## the defintion of expected value
pop.mean <- 0 * sum(fulton$turnout == 0)/nrow(fulton) + 1 * sum(fulton$turnout == 1)/nrow(fulton)
pop.mean
```

```
## [1] 0.4393607
```

- This expected value/mean would represent the (usually unknown) parameter that we want to estimate.
- Let's take a sample of size $n = 10$ from this population, using sampling with replacement (as if we drew names from a hat, putting the names back in after each draw):

```
## set the seed so we can replicate everything!
set.seed(02143)

## quick reminder on how to select certain rows from a data frame
## with a matrix like this, we select the rows before the comma, columns after
first.five <- fulton[c(1,2,3,4,5), "turnout"]
first.five
```

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

```
## or more succintly because 1:5 is equivalent to c(1,2,3,4,5)
first.five.alt <- fulton[1:5, "turnout"]
first.five.alt
```

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

```
## leaving either before or after the comma blank gives us all the rows/columns
first.five.allcols <- fulton[1:5, ]
first.five.allcols
```

```
##   turnout black sex age dem rep urban percblk lvbdist school firest church
## 1       0     0   1  19   0   0     0  0.0523  3.4836      0      0      1
## 2       0     0   0  35   0   0     0  0.0288  3.2913      1      0      0
## 3       0     1   0  36   0   0     1  0.9924  2.8767      1      0      0
## 4       1     0   0  27   0   0     1  0.1112  2.5618      0      0      0
## 5       1     1   1  79   1   0     1  0.9923  2.7935      1      0      0
```

```
## first draw the indices of rows that we want to draw:
## to get a sample of the rows of the data, we sample the row indices
## remember that 1:nrow(fulton) is just a list of numbers from 1 to the
## number of rows of the data
samp.rows <- sample(1:nrow(fulton), size = 10, replace = TRUE)
samp.rows
```

```
##  [1] 209608 199760 194250 151559 322897   9128  46681  81429 113986 126544
```

```
## now we want to create a new object that is just those sampled rows:
f.sample <- fulton[samp.rows,]
f.sample$turnout
```

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

- Here we can see that we've sampled 10 observations from the population. Implicitly this also means that we have drawn a random sample of size 10 from the population distribution of turnout.

Showing how the mean varies across samples

- Remember that the sample mean is also a random variable, so it will vary from sample to sample as well.
- **Definition**: The **sample mean**, denoted $\bar{X}$ or $\hat{\mu}$, of a set of r.v.s, $\{X_1, \ldots, X_n\}$ is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- The sample mean is just the simple average of all of the values in the sample. Let's calculate this in our Fulton County sample of size 10:

```
## using R functions:
mean(f.sample$turnout)
```

```
## [1] 0.6
```

- But here is something to consider: we got a particular sample from the population. What would happen if had gotten a different sample:

```
## first draw the indices of rows that we want to draw:
samp.rows2 <- sample(1:nrow(fulton), size = 10, replace = TRUE)
f.sample2 <- fulton[samp.rows2,]
mean(f.sample2$turnout)
```

```
## [1] 0.5
```

- These two values are different! Somewhat obviously, this is because the sample mean is a function of the iid random variables. Because these change from sample to sample, then the estimate of the mean (the sample mean) will change from sample to sample.
- Basically, because $\bar{X}$ is a function of random variables, it too is a random variable. So it varies between different random samples of the data and it has its own probability distribution with its own center and spread.
- The distribution of the estimates across different random samples is called the **sampling distribution** and this is a crucial idea for us in this class and in all of statistics. In the coming weeks we will talk a lot about these sampling distributions. Be warned, though: sampling distributions are incredibly tricky and yet they are extremely fundamental. You need to make sure that you understand them.
- In the coming weeks, the sampling distribution is going to help us choose among rivals estimators and help us quantify our uncertainty about our estimates.
- **Definition**: The **sample variance**, denoted S^2 or $\hat{\sigma}^2$, of a set of r.v.s, $\{X_1, \ldots, X_n\}$ is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

```
## using R function
var(f.sample$turnout)
```

```
## [1] 0.2666667
```

- We'll see that the sample variance is going to be important in characterizing the sampling distribution of the mean. But you should remember that the sample variance has its own sampling distribution. It varies from sample to sample as well:

```
## using R function
var(f.sample2$turnout)
```

```
## [1] 0.2777778
```

Example: simulating the sampling distribution

- Can we tell what the average of the sampling distribution of the mean is? Well, here we can do this by simulation:

1. Take a random sample of size 10 from the population
2. Calculate the sample mean from the sample
3. Repeat 1 and 2 a lot of times (> 10,000) and store the results
4. Take the mean of the resulting distribution of estimates

```
## set up some stuff like a holder for the results
nsims <- 10000
my.means <- rep(NA, times = nsims)

## we're going to do this a bunch of times
for (i in 1:nsims) {
  samp.rows <- sample(1:nrow(fulton), size = 10, replace = TRUE)

  ## access the turnout variable for only the sampled rows
  ## and store the results
  my.means[i] <- mean(fulton[samp.rows, "turnout"])
}

summary(my.means)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3000  0.4000  0.4396  0.5000  1.0000
```

- Here we can see some feature of this sampling distribution. Its mean is actually fairly close to the overall population mean. But sometimes the sample mean was 0 (so there were no voters in the sample) or it was 1 (the sample was all voters). Next week we'll prove this formally and talk more about this distribution.

POINT ESTIMATION

Parameters

- Trying to estimate **parameters** of population distributions:
  - $\mu = E[Y]$: the mean
  - $\sigma^2 = \text{Var}[X]$: the variance
  - σ: the standard deviation
  - $\mu_1 - \mu_0 = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0]$: the difference in means between two groups
  - $\mathbb{E}[Y|X] = \alpha + \beta X$: intercept ($\alpha$) and slope ($\beta$) of the regression line
- We'll generically refer to the parameter we're trying to estimate as θ.
- These are the things we want to learn about.

Estimators

- Remember that we have a random sample $Y_1, \ldots, Y_n$ from some population distribution $f(y; \theta)$.
- **Definition**: An **estimator** , $\hat{\theta}$ of some parameter θ, is a function of the sample: $\hat{\theta} = h(Y_1, \ldots, Y_n)$.
- **Question** Why is the following statement wrong: "My estimate was the sample mean and my estimator was 0.377"?
- Remember that these estimators are like machines that take in a sample and output a number—we call this number the **estimate** of θ. But the estimator is the function or the rule for calculating estimates.
- This can get confusing because the names of the estimators are sometimes used as shorthand for the estimates themselves: "The sample mean was 0.377." What this really means is that the sample mean produced an estimate of 0.377 in this sample. The sample mean would have produced a different estimate with a different sample.

Examples of Estimators

- For the population mean, μ, we have many different possible estimators:
  - $\hat{\theta} = \bar{Y}$ the sample mean
  - $\hat{\theta} = Y_1$ just use the first observation
  - $\hat{\theta} = \max(Y_1, \ldots, Y_n)$
  - $\hat{\theta} = 3$ always guess 3

```
# mean
mean(social$voted[social$treatment == "Neighbors"])
```

```
## [1] 0.3779482
```

```
# first observation
social$voted[social$treatment == "Neighbors"][1]
```

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

```
# maximum
max(social$voted[social$treatment == "Neighbors"])
```

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

```
# always choose 3
3
```

```
## [1] 3
```

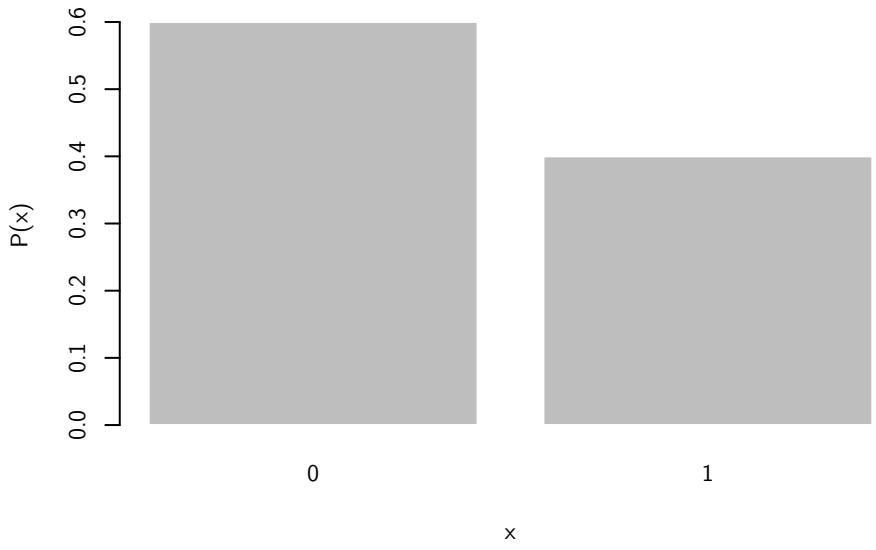
The Three Distributions

- There are three important distributions to keep straight. Let's review them using the frame of estimating the mean voter turnout of the "Neighbors" group to be our goal.
- Population Distribution: the data-generating process (Bernoulli in the case of the social pressure/voter turnout example)
- Sample distribution: $Y_1, \ldots, Y_n$ (series of 1s and 0s in the above example)

- Sampling distribution: distribution of the estimator over repeated samples from the population distribution (the .377 sample mean in the "Neighbors" group is one draw from this distribution)
- **Question**: If $Y_1, \ldots, Y_n$ is a random sample from a (population) Bernoulli distribution with mean/probability μ, will sampling distribution of the sample mean ($\bar{Y}$) be Bernoulli as well?

Sampling Distributions

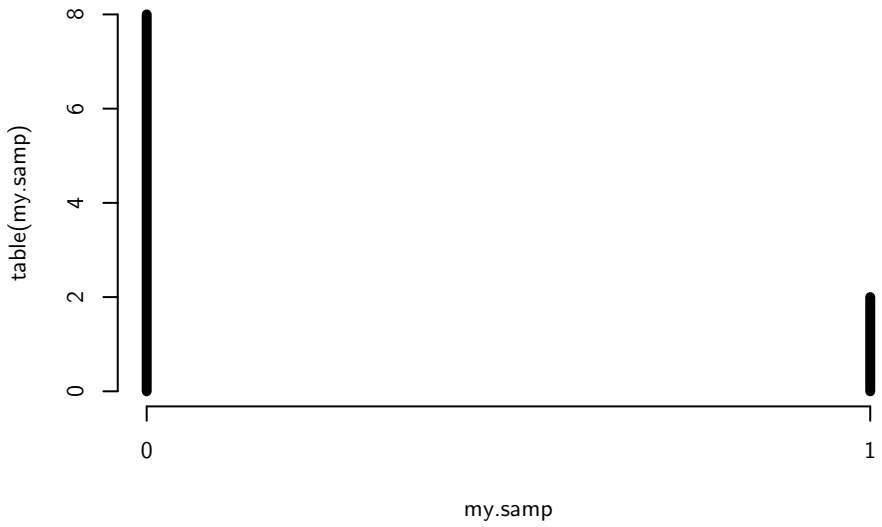
```
## population distribution is Bernoulli
## with probability of success 0.4
barplot(dbinom(0:1,size = 1, prob = 0.4), names.arg = c(0,1),
        border = "white", ylab = "P(x)", xlab = "x")
```




```
## we're going to take **one** sample of size 10
my.samp <- rbinom(n = 10, size = 1, prob = 0.4)
table(my.samp)
```

```
## my.samp
## 0 1
## 8 2
```

```
plot(table(my.samp), type = "h", lwd = 5, bty = "n")
```




```
## now we take the mean of the this sample, which is one draw from the
## **sampling distribution**
mean(my.samp)
```

```
## [1] 0.2
```

```
## let's take another draw from the population Distribution
my.samp2 <- rbinom(n = 10, size = 1, prob = 0.4)
table(my.samp2)
```

```
## my.samp2
## 0 1
## 7 3
```

```
## Let's feed this sample to the sample mean estimator and get another
## estimate, which is another draw from the sampling distribution
mean(my.samp2)
```

```
## [1] 0.3
```

- Remember that we are going to get to see one draw from the sampling distribution. Ideally we would want our estimator to have a sampling distribution that

puts all of the probability mass on one value: the true value of θ. But this isn't possible, so we want to make sure that we are using an estimator that has other good properties.

- **Question** The sampling distribution refers to the distribution of θ, true or false. Explain your answer.

FINITE-SAMPLE PROPERTIES OF ESTIMATORS

- Is our estimator good? It is better than some other estimators? How do we evaluation these little machines that take in samples and output estimates?
- Why did we use the sample difference in means to estimate the population difference in means for the above experiment? Why not the difference in sample medians or the difference in sample modes?
- There are two ways we evaluate estimators: the properties of its sampling distribution for a fixed sample size n (finite-sample) and the properties of the sampling distribution as we let $n \to \infty$. On the homework, you'll do both of these.
- As a shorthand, we will often indicate that an estimator is operating on a given sample size by giving it an n subscript:

$$\hat{\theta}_n = \hat{\theta}(Y_1, \ldots, Y_n)$$

- There are a few things we might want an estimator to do. We probably want it to get the right answer on average and we would ideally like it bounce around very little from sample to sample.

Unbiasedness

- Estimators can either be biased or unbiased. What does unbiased mean? It means that if we were to take many many samples from the population distribution, apply the estimator to each sample, and look at the distribution of those estimates, the average of that distribution would be equal to the true parameter value. More succinctly: **on average, we get the right answer**.
- **Definition**: The **bias** of an estimator $\hat{\theta}$ for population parameter θ is

$$\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

- **Definition**: An estimator $\hat{\theta}$ of θ is **unbiased** if

$$\mathbb{E}[\hat{\theta}] = \theta$$

- **Example** It's fairly straightforward to prove the sample mean is unbiased for the population mean:

$$\begin{aligned}\mathbb{E}[\bar{Y}_n] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} Y_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \frac{1}{n} n\mu \\ &= \mu\end{aligned}$$

- Sometimes we can prove unbiasedness **analytically**, where we just apply the expectation operator to the estimator (remember it's a random variable).
- **Problem** Let $Y_1, \ldots, Y_n$ be an iid sample from a population distribution with mean μ. Show that $\hat{\theta} = Y_1$ is an unbiased estimator for the mean.

- We can also check the bias of an estimator using simulation:

```
nsims <- 10000
mean.holder <- rep(NA, times = nsims)
first.holder <- rep(NA, times = nsims)
for (i in 1:nsims) {
    my.samp <- rbinom(n = 100, size = 1, prob = 0.4)
    mean.holder[i] <- mean(my.samp) ## sample mean
    first.holder[i] <- my.samp[1] ## first obs
}
mean(mean.holder) - 0.4
```

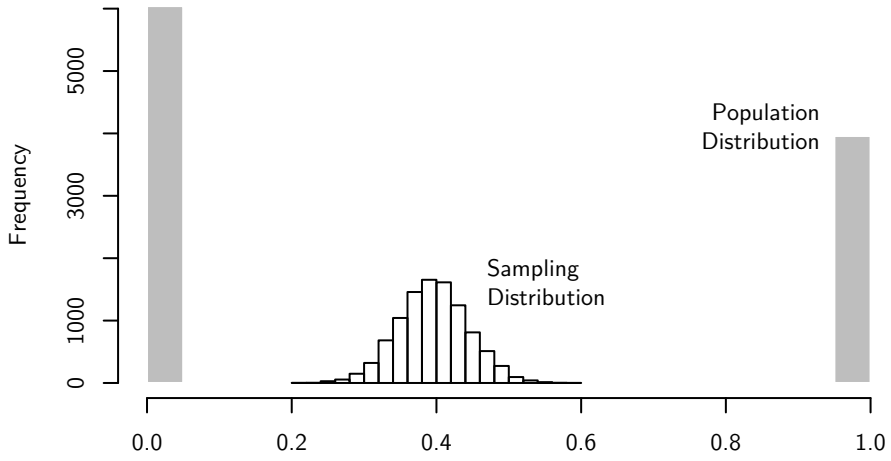
```
## [1] 0.000388
```

```
mean(first.holder) - 0.4
```

```
## [1] -0.0041
```

- Both are pretty close to 0!
- What does the sampling distribution look like relative to the population distribution? To see this, we'll use the "first observation" estimator since this is just a series of draws from the population distribution

```
hh <- hist(first.holder, plot = FALSE)
hh2 <- hist(mean.holder, plot = FALSE)
plot(hh, main = "", xlim = c(0,1), col = "grey", border = "white", xlab = "")
plot(hh2, add = TRUE)
text(x = 0.95, y = 10000*0.4, "Population\nDistribution", pos = 2)
text(x = 0.45, y = 1500, "Sampling\nDistribution", pos = 4)
```



- The population distribution can only take on the values 0 or 1, whereas the sampling distribution of the mean takes values that in between. Furthermore, you can see that
- **Problem** Calculate the bias of the estimator $\hat{\theta} = 3$ in terms of the population mean, μ. When will this estimator be unbiased?

- **Problem** Suppose that $Y_1, \ldots, Y_n$ is an iid sample from a population with mean μ_y and $X_1, \ldots X_n$ is an iid sample from a population with mean μ_x. Show that the sample difference in means $\bar{Y} - \bar{X}$ is an unbiased estimator for population difference in means, $\mu_y - \mu_x$. (For a bigger challenge, show that the post-stratification estimator is also unbiased for the same difference in population means. Protip: use the law of iterated expectations.)

Sampling Variance

- Intuitively, we would like to say that if an estimator is unbiased, then the one estimate that we get in our one sample should be "close" to the true value of the parameter. But unbiasedness is only about central tendency so the estimate might be very far from the truth. Thus, we need to think about the spread of the sampling distribution.
- The first observation, Y_1 and the sample mean $\bar{Y}_n$ are both unbiased, but obviously the second feels like a better estimator. We obviously need more criteria to evaluate estimators.
- **Question** If we have an unbiased estimator, do we want the sampling distribution to have higher or lower variance?
- **Definition**: The **sampling variance** of an estimator is simply its variance over repeated samples, $\text{Var}[\hat{\theta}]$.
- **Definition**: The **standard error** of an estimator is the standard deviation of the sampling distribution, $SE[\hat{\theta}] = \sqrt{\text{Var}[\hat{\theta}]}$
- These quantities are incredibly important because they tell us how uncertain our estimates are. If the sampling variance is very high, then our estimator will produce very different estimates from sample to sample. If it is very low, then the estimates will be very similar from sample to sample.
- Again, we can calculate the sampling variance analytically in some cases.

- **Example** Let's calculate the sampling variance of the sample mean of an iid sample, $Y_1, \ldots, Y_n$, with mean μ and variance σ^2:

$$\begin{aligned}
\text{Var}[\bar{Y}_n] &= \text{Var}[\frac{1}{n}\sum_{i=1}^{n} Y_i] && \text{(definition of variance)} \\
&= \frac{1}{n^2}\text{Var}[\sum_{i=1}^{n} Y_i] && \text{(properties of variance)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}[Y_i] && \text{(independence of observations)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 && \text{(definition)} \\
&= \frac{1}{n^2} n\sigma^2 && \text{(sum of a constant)} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

- The standard error of the sample mean, then, is just $\sigma/\sqrt{n}$.
- We can also investigate the sampling variance by simulation. Take the above code and let's look at the variance of the two sampling distributions:

```
var(mean.holder)
```

```
## [1] 0.002342664
```

```
var(first.holder)
```

```
## [1] 0.2391871
```

- Obviously, the sample mean has a much lower variance than just using the first observation. And since they are both unbiased, this means that our estimates will be closer to the truth on average.
- **Problem** Suppose $Y_1, \ldots, Y_{n_y}$ is an iid sample of size n_y from a population with mean μ_y and variance σ_y^2 and $X_1, \ldots, X_{n_x}$ is an iid sample of size n_x from a population with mean μ_x and variance σ_x^2. Calculate the sampling variance and the standard error of the difference in sample means estimator, $\bar{Y} - \bar{X}$.

Estimating the Sampling Variance/Standard Error

- Above we saw that the sampling variance (the variance of the sampling distribution) of the sample mean is σ^2/n when we apply the mean to an iid sample of size n from a population distribution with variance σ^2.
- Obviously we know n for our sample, so we just need to know σ^2 in order to know how much variability there will be in the sample mean between samples.
- In order to get an estimate of the sampling variance for the sample mean, we just need an estimator for the population variance. Turns out, know one of those, the sample variance:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

- Thus, our estimate of the sampling variance of the sample mean is S_n^2/n and the standard error of the sample mean is the $S_n/\sqrt{n}$.
- These are estimates of how uncertain our estimates are.

Efficiency

- **Definition**: If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of θ, then $\hat{\theta}_1$ is **efficient** relative to $\hat{\theta}_2$ when $\text{Var}[\hat{\theta}_1] \leq \text{Var}[\hat{\theta}_2]$ for any possible value of θ with strict inequality for at least one value of θ.
- **Problem** Show that the sample mean, $\bar{Y}$ is efficient relative to the "first observation" estimator, $\hat{\theta} = Y_1$ when $n > 1$.

Bias-Variance Tradeoff

- In many situations, there is tradeoff between bias and variance. In the extreme, think about comparing the sample mean $\hat{\theta} = \bar{Y}$ to the dumb, always selection 3 estimator: $\hat{\theta}_3 = 3$. The always choose 3 estimator has bias that can be quite high when the population mean is very far from 3, but it has 0 variance.

Mean Squared Error

- Obviously, if both of the estimators are unbiased, then all we want is lower sampling variance. But what about comparing biased and unbiased estimators or two biased estimators. How do we figure out which is better.

- One way to compare them is to see their squared estimation errors. For example, we might prefer a slightly biased estimator if it reduces the sampling variance.
- **Definition**: The **mean squared error (MSE)** of an estimator $\hat{\theta}$ for θ is $\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$. We can write this as:

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

- **Problem** What is the mean squared error of an unbiased estimator?

LARGE-SAMPLE PROPERTIES OF ESTIMATORS

- Remember that we have been investigating the properties of some estimator for a fixed number of observations, $\hat{\theta}_n$. And this is useful because we usually only have one sample that has a fixed sample size.
- Another way to compare estimators is to see how perform as we increase the sample size, n.
- Intuitively, we might want to know that the estimator "converges" to true parameter in some way.
- In addition, as n gets large, it turns out that many estimators have an approximately Normal distribution. This is really useful, since it allows us to figure out (approximately) the probability of seeing estimates that are larger or smaller than a particular value: $\mathbb{P}(\bar{Y}_n > 0)$. This will be very important next week when we want to form confidence intervals and think about hypothesis tests.
- Some of the math here gets a little complicated, but the intuitions are fairly simple.

Convergence in Probability

- A series of numbers, x_n converges to some value x if for every $\varepsilon > 0$, there exists an N such that for $n > N$ we have $|x_n - x| < \varepsilon$. Basically, as n grows, x_n gets arbitrarily close to x (though it may never reach it). What does it mean for a an estimator, or more generally a random variable, to "converge to some value"?
- **Question**: Suppose we have an iid sample, $Y_1, \ldots, Y_n$ from a population distribution with mean μ. What is $\mathbb{P}(\bar{Y}_n = \mu)$ as $n \to \infty$?

- **Definition**: A sequence of random variables, $X_1, X_2, \ldots$, is said to **converge in probability** to a value c if for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n - c| > \varepsilon) \to 0,$$

as $n \to \infty$. We write this $X_n \xrightarrow{p} c$.

- **Intuition** What is this saying intuitively? It says that the probability of X_n being more than a small value (ε) away from c goes to 0 as n gets large. In other words, the distribution of X_n becomes concentrated around c as n gets large.

- Wooldridge writes $\text{plim}(X_n) = c$ if $X_n \xrightarrow{p} c$.

- Properties of the convergence in probability:

1. if $X_n \xrightarrow{p} c$, then $g(X_n) \xrightarrow{p} g(c)$ for any continuous function g.
2. if $X_n \xrightarrow{p} a$ and $Z_n \xrightarrow{p} b$, then
    - $X_n + Z_n \xrightarrow{p} a + b$
    - $X_n Z_n \xrightarrow{p} ab$
    - $X_n / Z_n \xrightarrow{p} a/b$ if $b > 0$

Consistency

- **Definition** An estimator $\hat{\theta}_n$ is **consistent** for θ if $\hat{\theta}_n \xrightarrow{p} \theta$.

- **Definition**: Let $\hat{\theta}_n$ be an estimate of θ from a sample $Y_1, \ldots, Y_n$ of size n. Then, $\hat{\theta}_n$ estimator is **consistent** for θ for every $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty$$

- An intuitive way to understand consistency is that the sampling distribution of $\hat{\theta}_n$ is collapsing around the true value. As n increases, the entire distribution of estimates is getting closer to the truth.

- If an estimator is unbiased, then it is easy to figure out if it is consistent: an unbiased estimator is consistent if the sampling variance goes to 0 as $n \to \infty$ or $\lim_{n\to\infty} \text{Var}[\hat{\theta}_n] = 0$.

- An estimator is inconsistent if $\hat{\theta}_n \not\xrightarrow{p} \theta$. This might be because it converges to another value or because it never converges at all. Either of these are bad properties—imagine giving your estimator more data and your estimates either staying the same or getting worse! These estimators violate the first golden rule of statistics: more data is better.

- **Example** The "first observation" estimator, Y_1 is unbiased as we said, but it is inconsistent. This is because the sampling distribution never collapses to any value. As we add more observations, we simply drop those added observations and only use the first.

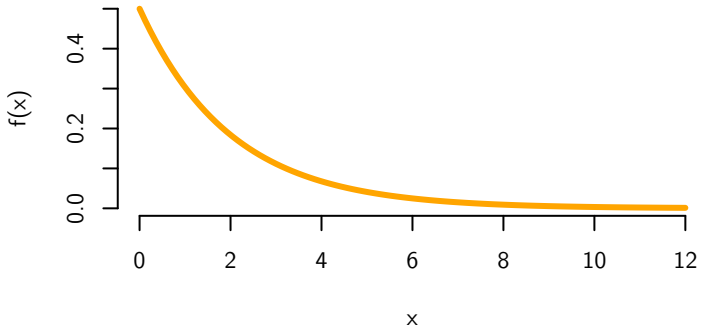
Law of Large Numbers

- One consistency result has a special name and that is the idea that the sample mean is consistent for the population mean.
- **Theorem** (Weak Law of Large Numbers) Let $Y_1, \ldots, Y_n$ be a an iid draws from a distribution with mean μ and let $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then, $\bar{Y}_n \xrightarrow{p} \mu$.
- Note that this is very easy to prove if the population distribution has a finite variance, σ^2 given that the sample mean is unbiased and the sampling variance of the sample mean is σ^2/n. Thus, as n gets large, the variance around the true values goes to 0.

Consistency example

- Let's compare two estimators that we might experience out in the real world. In both, we'll take iid samples from the population distribution which is Exponential with rate parameter 0.5. It's not vital that we know much about the Exponential, just that the mean of this population distribution will be 2. We want to estimate this mean.

```
curve(dexp(x, rate = 0.5), from = 0, to = 12, col = "orange", lwd = 3, bty = "n",
      ylab = "f(x)", xlab = "x")
```



- Suppose that we are interested in how long a government lasts in parliamentary democracies. So what we do is start today in 2014 record when each government dissolves and how long it took from when we started counting (9/30/2014).

Call this X. Now, suppose that the distribution of X is Exponential with rate 0.5.

- Now imagine two data collection schemes: one in which we wait until we've collected n observations from the process (this is just an iid sample) and one in which we stop collecting after 3 years, which we call a **censored sample**.
- What are the properties of these two approaches? We can investigate these by thinking about the two implied estimators: first, the usual sample mean for the entire sample of size n; and second, the censored sample mean, which is just the usual sample mean applied to any of the n observations that are less than 3. What is going to happen to these estimators as we increase n?

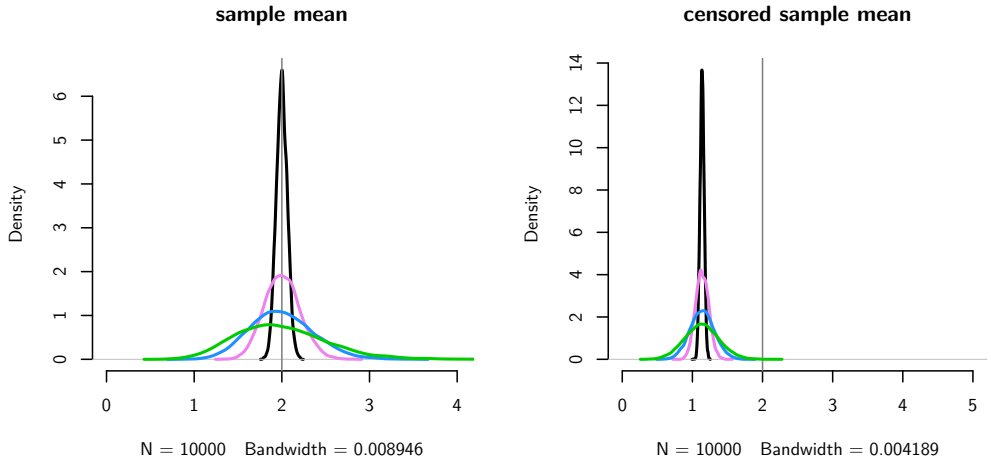
```
nsims <- 10000
holder <- matrix(NA, nrow = nsims, ncol = 6)
bad.holder <- matrix(NA, nrow = nsims, ncol = 6)
for (i in 1:nsims) {
  s5 <- rexp(n = 5, rate = 0.5)
  s15 <- rexp(n = 15, rate = 0.5)
  s30 <- rexp(n = 30, rate = 0.5)
  s100 <- rexp(n = 100, rate = 0.5)
  s1000 <- rexp(n = 1000, rate = 0.5)
  s10000 <- rexp(n = 10000, rate = 0.5)

  holder[i,1] <- mean(s5)
  holder[i,2] <- mean(s15)
  holder[i,3] <- mean(s30)
  holder[i,4] <- mean(s100)
  holder[i,5] <- mean(s1000)
  holder[i,6] <- mean(s10000)

  bad.holder[i,1] <- mean(s5[s5 < 3])
  bad.holder[i,2] <- mean(s15[s15 < 3])
  bad.holder[i,3] <- mean(s30[s30 < 3])
  bad.holder[i,4] <- mean(s100[s100 < 3])
  bad.holder[i,5] <- mean(s1000[s1000 < 3])
  bad.holder[i,6] <- mean(s10000[s10000 < 3])

}
```

```
par(mfrow = c(1,2))
plot(density(holder[,5]), xlim = c(0,4), bty = "n", main = "sample mean", lwd = 2)
abline(v=2, col = "grey50")
lines(density(holder[,4]), col = "violet", lwd = 2)
lines(density(holder[,3]), col = "dodgerblue", lwd = 2)
lines(density(holder[,2]), col = "green3", lwd = 2)
plot(density(bad.holder[,5]), xlim = c(0,5), bty = "n", main = "censored sample mean", lwd = 2)
abline(v=2, col = "grey50")
lines(density(bad.holder[,4]), col = "violet", lwd = 2)
lines(density(bad.holder[,3]), col = "dodgerblue", lwd = 2)
lines(density(bad.holder[,2]), col = "green3", lwd = 2)
```



- As you can see, as we increase the sample size, the usual sample mean converges to the true mean, 2 (what theorem do we have to thank for that?). But the sampling distribution of the censored sample mean starts to concentrates all of its mass very far away from the truth.
- Why is this important? The censored sample mean here will *almost never* be close to the true mean in large samples. Thus, inconsistent estimators are really misleading.

Convergence in Distribution

- One nice property of many estimators is that their sampling distribution is approximately Normal in large samples.

- **Definition**: A sequence of random variables, $X_1, X_2, \ldots$, is said to **converge in distribution** to Z if

$$\lim_{n \to \infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(Z \leq x),$$

which we write as $X_n \xrightarrow{d} Z$.

Asymptotic Normality

- **Definition**: An estimator is said to be **asymptotically Normal** if

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\text{Var}(\hat{\theta}_n)}} = \frac{\hat{\theta}_n - \theta}{SE(\hat{\theta})} \xrightarrow{d} N(0,1).$$

Central Limit Theorem

- **Theorem** (Central Limit Theorem) Let $Y_1, \ldots, Y_n$ be a an iid draws from a distribution with mean μ and variance σ^2. Then

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

- **Intuition** This is saying that for large samples (usually greater than 30 with an iid sample), the sampling distribution of the standardized sample mean is just the standard Normal distribution.
- Note that the central limit theorem also applies to sums and differences in means.
- Why do we standardize the sample mean here? Because we already know that the sample mean converges in probability to the population mean, so the sampling distribution of the sample mean is just converging to a single point. Standardizing ensures that this random variable is centered at 0 and has a variance that remains constant (at 1).

Empirical Rule for the Normal Distribution

- If $Z \sim N(0,1)$, then the following are roughly true:
  - Roughly 68% of the distribution of Z is between -1 and 1.
  - Roughly 95% of the distribution of Z is between -2 and 2.
  - Roughly 99.7% of the distribution of Z is between -3 and 3.
- You can use the `pnorm()` function in R to figure out any probability questions about the Normal distribution.

Asymptotic Normality Example

- Going back to the Gerber, Green, and Larimer paper. Remember that we saw that there was roughly a 0.063 difference between the Neighbors group and Civic Duty group. That's our estimate in this example.
- **Problem** Show that the Central Limit Theorem also applies to the difference in sample means when the group sizes are the same ($n = n_y = n_x$). Obviously the result applies to other cases, but this is easiest to show using what we have learned so far.
- Suppose that in this case we have estimated the standard error of the difference in means to be $\hat{SE}(\hat{\theta}) = 0.02$.
- Let's do one of the Lady Tasting Tea statistical thought experiments. What if there was no difference in means in the population ($\mu_y - \mu_x = 0$)? What would the sampling distribution of the difference in sample means look like in large samples?
- Well, we know by asymptotic Normality and the thought experiment assumption that

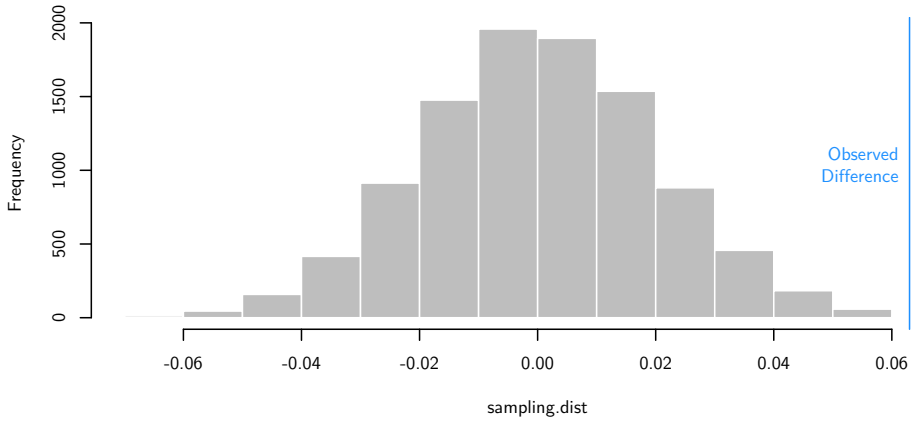
  $$(\hat{\theta} - 0)/SE(\hat{\theta}) \sim N(0, 1)$$

  by the properties of Normals, we know that this implies that

  $$\hat{\theta} \sim N(0, SE(\hat{\theta}))$$

- What will this distribution look like? We know it's going to be centered around 0, but what about the spread? Here we can plug in the estimated SE to get our best guess as to what the distribution will look like. Let's do that and plot the results along with the observed difference in means:

```
nsims <- 10000
sampling.dist <- rnorm(n = nsims, mean = 0, sd = 0.02)
hist(sampling.dist, xlim = c(-0.07, 0.07), main = "", col = "grey",
     border = "white")
abline(v = 0.063, col = "dodgerblue")
text(x = 0.063, y = 1000, "Observed\nDifference", pos = 2, col = "dodgerblue")
```



- **Question** Does the observed difference in means seem plausible if there really were no difference between the two groups in the population?