

# Gov 2000 - 5. Statistical Inference and Estimation

Matthew Blackwell

*Harvard University*

[mblackwell@gov.harvard.edu](mailto: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:
- 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:

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!

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

Figure 1: Mailer

```
load("../data/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 thought 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 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 answer 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,  $\theta$ . 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  $X = 1$ ,  $\theta$ . Then 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  $\theta$ .
- 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  $X_1, X_2, \dots, X_n$  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  $\mu$  and  $\sigma^2$ .
- We call iid samples from a distribution **random samples**.

## POINT ESTIMATION

### *Parameters*

- Trying to estimate **parameters** of population distributions:
  - $\mu = E[Y]$ : the mean
  - $\sigma^2 = \text{Var}[X]$ : the variance
  - $\sigma$ : 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  $\theta$ .
- These are the things we want to learn about.

### Estimators

- Remember that we have a random sample  $Y_1, \dots, Y_n$  from some population distribution  $f(y; \theta)$ .
- **Definition:** An estimator,  $\hat{\theta}$  of some parameter  $\theta$ , is a function of the sample:  $\hat{\theta} = h(Y_1, \dots, Y_n)$ .
- **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  $\theta$ . 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,  $\mu$ , we have many different possible estimators:
  - $\hat{\theta} = \bar{Y}$  the sample mean
  - $\hat{\theta} = Y_1$  just use the first observation
  - $\hat{\theta} = \max(Y_1, \dots, Y_n)$
  - $\hat{\theta} = 3$  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
```

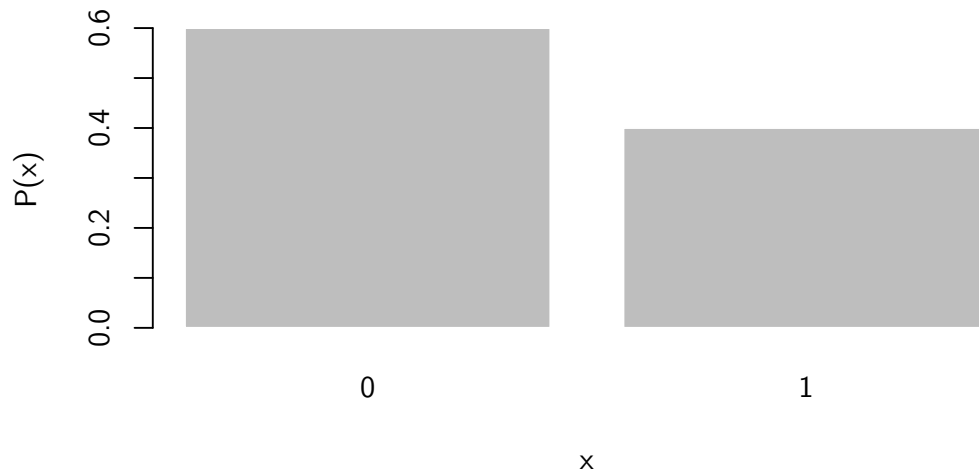
```
## [1] 3
```

### *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)
- Empirical distribution:  $Y_1, \dots, Y_n$  (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  $Y_1, \dots, Y_n$  is a random sample from a (population) Bernoulli distribution with mean/probability  $\mu$ , 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)
```

```
## here is the empirical distribution:  
table(my.samp)
```

```
## my.samp  
## 0 1  
## 6 4
```

```
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] 0.4
```

```
## 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
## 6 4
```

```
## 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] 0.4
```

- 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  $\theta$ . 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  $\theta$ , 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 evaluate 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  $n \rightarrow \infty$ . 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:

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

- 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  $\theta$  is

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

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

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

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

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

- Sometimes we can prove unbiasedness **analytically**, where we just apply the expectation operator to the estimator (remember it's a random variable).

- **Problem** Let  $Y_1, \dots, Y_n$  be an iid sample from a population distribution with mean  $\mu$ . 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
```

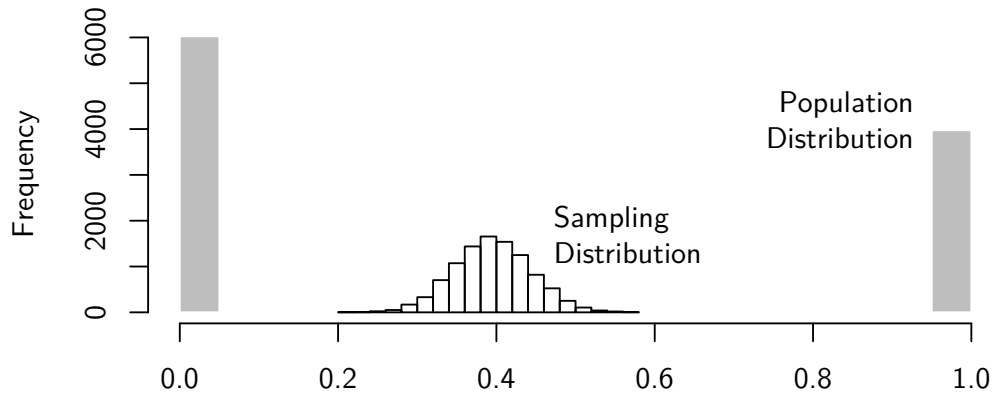
```
## [1] 0.000175
```

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

```
## [1] -0.0025
```

- 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,  $\mu$ . When will this estimator be unbiased?
- **Problem** Suppose that  $Y_1, \dots, Y_n$  is an iid sample from a population with mean  $\mu_y$  and  $X_1, \dots, X_n$  is an iid sample from a population with mean  $\mu_x$ . 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,  $Y_1, \dots, Y_n$ , with mean  $\mu$  and variance  $\sigma^2$ :

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

- 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] 0.002381598
```

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

```
## [1] 0.2395177
```

- 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  $Y_1, \dots, Y_{n_y}$  is an iid sample of size  $n_y$  from a population with mean  $\mu_y$  and variance  $\sigma_y^2$  and  $X_1, \dots, X_{n_x}$  is an iid sample of size  $n_x$  from a population with mean  $\mu_x$  and variance  $\sigma_x^2$ . 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  $\sigma^2/n$  when we apply the mean to an iid sample of size  $n$  from a population distribution with variance  $\sigma^2$ .
- Obviously we know  $n$  for our sample, so we just need to know  $\sigma^2$  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  $\theta$ , 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  $\theta$  with strict inequality for at least one value of  $\theta$ .
- **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  $\theta$  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 an estimator, or more generally a random variable, to “converge to some value”?
- **Question:** Suppose we have an iid sample,  $Y_1, \dots, Y_n$  from a population distribution with mean  $\mu$ . What is  $\mathbb{P}(\bar{Y}_n = \mu)$  as  $n \rightarrow \infty$ ?
- **Definition:** A sequence of random variables,  $X_1, X_2, \dots$ , is said to **converge in probability** to a value  $c$  if for every  $\varepsilon > 0$ ,

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

as  $n \rightarrow \infty$ . 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 ( $\varepsilon$ ) 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  $\theta$  if  $\hat{\theta}_n \xrightarrow{P} \theta$ .
- **Definition:** Let  $\hat{\theta}_n$  be an estimate of  $\theta$  from a sample  $Y_1, \dots, Y_n$  of size  $n$ . Then,  $\hat{\theta}_n$  estimator is **consistent** for  $\theta$  for every  $\varepsilon > 0$ ,

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

- 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  $n \rightarrow \infty$  or  $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}_n] = 0$ .
- 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.

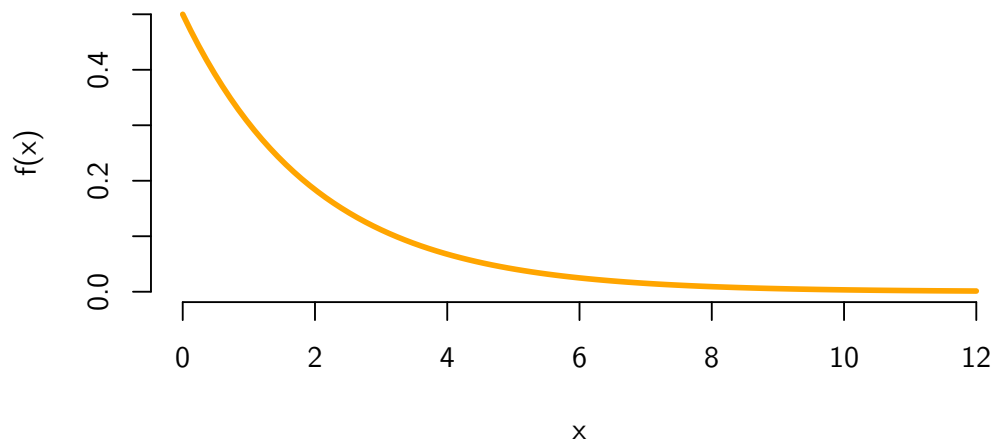
### Law of Large Numbers

- 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  $Y_1, \dots, Y_n$  be a an iid draws from a distribution with mean  $\mu$  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,  $\sigma^2$  given that the sample mean is unbiased and the sampling variance of the sample mean is  $\sigma^2/n$ . 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")
```

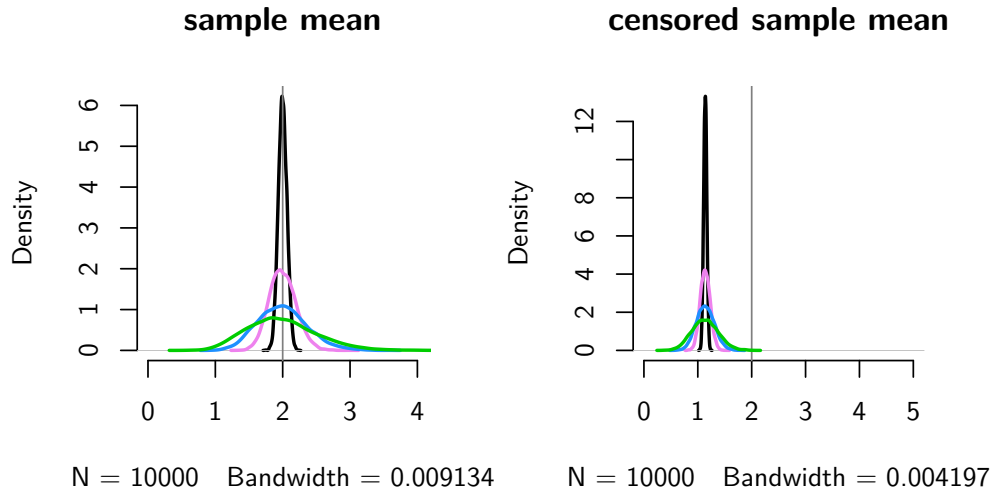


- Suppose that we are interested in how long a government lasts in parliamentary democracies. So what we do is start today in 2014 record when each government dissolves and how long it took from when we started counting (9/30/2014). Call this  $X$ . Now, suppose that the distribution of  $X$  is Exponential 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 concentrate 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,  $X_1, X_2, \dots$ , is said to **converge in distribution** to  $Z$  if

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

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  $Y_1, \dots, Y_n$  be a an iid draws from a distribution with mean  $\mu$  and variance  $\sigma^2$ . 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  $\widehat{SE}(\hat{\theta}) = 0.02$ .
- 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

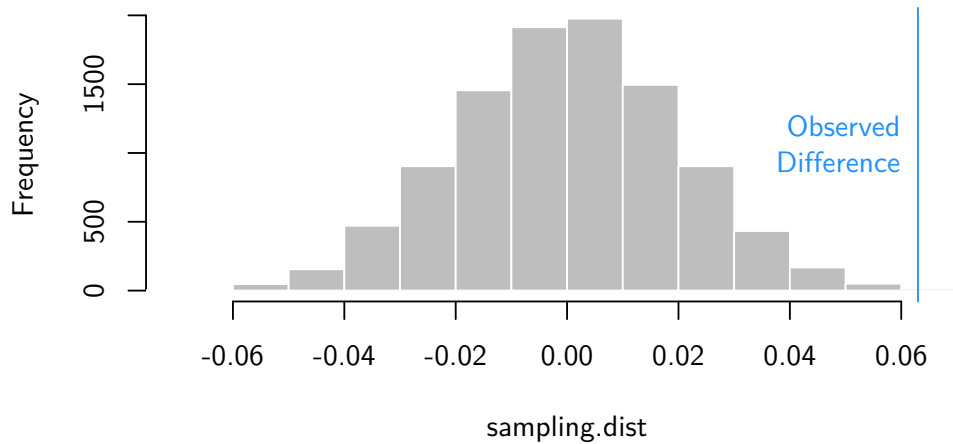
$$(\hat{\theta} - 0)/SE(\hat{\theta}) \sim N(0, 1)$$

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?

## INTERVAL ESTIMATION

### *Interval estimation - what and why?*

- Last week we talked about producing a single estimate of some population parameter. For example, we estimated the population difference in means using the sample difference in means. But that gave us one number. That might be our “best guess” about the true value, but what is the probability that our estimate will be equal to the true value? Zero!
- A different (complementary) approach to estimation is to produce a range of values that will contain the truth with some fixed probability.
- An **interval estimate** of the population mean,  $\mu$ , consists of two bounds within which we expect  $\mu$  to reside:

$$LB \leq \mu \leq UB$$

- A confidence interval is a kind of interval estimator that guarantees that the true value of the parameters will be in interval in some fixed proportion of random samples.
- How can we possibly figure out such an interval? We’ll rely on the distributional properties of estimators that we learned last week. Today we’ll focus on developing these types of interval estimators for the sample mean, but the same ideas apply to all estimators. Later in the course, we’ll see how they apply to regression coefficients.

### *Estimating the standard error of the sample mean*

- We’re going to make heavy use of the sampling distribution of the sample mean to derive interval estimates, called confidence intervals, for the sample mean.
- By the CLT, we know that in large samples, we know that the sample mean will be approximately Normally distributed:

$$\bar{Y} \sim N(\mu, SE[\bar{Y}]^2)$$

- By the properties of probability limits from last week, if we have a consistent estimator,  $\widehat{SE}[\bar{Y}] \xrightarrow{p} SE[\bar{Y}]$ , then when  $n$  is large, we know can replace the true  $SE$  with the estimate in the approximating distribution:

$$\bar{Y} \sim N(\mu, \widehat{SE}[\bar{Y}]^2)$$

- For the sample mean, remember that the sampling variance is  $\frac{\sigma^2}{n}$ . Remember that we know  $n$  and we can estimate  $\sigma^2$  with the sample variance:

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

- To get an estimate of the population standard deviation, we just use the sample standard deviation:  $S_n = \sqrt{S_n^2}$
- We can plug this into our formula for the standard error the sample mean:

$$\widehat{SE}[\bar{Y}] = \frac{S}{\sqrt{n}}$$

*Derivation of the Confidence Interval for the Population Mean*

- **Definition** A  $100(1 - \alpha)\%$  **confidence interval** with confidence level  $1 - \alpha$  is an interval estimator for a population parameter  $\theta$  that will contain/cover the true value,  $\theta$ ,  $100(1 - \alpha)\%$  of the time.
- We going to use the above distributional results to come up with a rule for constructing a confidence interval. This rule will be an estimator just like the sample mean or the sample variance, but it will produce two values instead of one: the upper and lower values of the intervals.
- If  $Y_1, \dots, Y_n$  are iid and  $n$  is large, then we have

$$\begin{aligned}\bar{Y} &\sim N(\mu, \widehat{SE}[\bar{Y}]^2) \\ \bar{Y} - \mu &\sim N(0, \widehat{SE}[\bar{Y}]^2) \\ \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} &\sim N(0, 1)\end{aligned}$$

- One fact that you should memorize is that that 95% of the probability mass of a standard Normal falls between -1.96 and 1.96, so we know that:

$$\mathbb{P}(-1.96 \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq 1.96) = 0.95$$

- Let's work backwards to derive the confidence interval:



$$\begin{aligned} \mathbb{P}\left(-1.96 \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq 1.96\right) &= 0.95 \\ \mathbb{P}\left(-1.96 \times \widehat{SE}[\bar{Y}] \leq \bar{Y} - \mu \leq 1.96 \times \widehat{SE}[\bar{Y}]\right) &= 0.95 \\ \mathbb{P}\left(-1.96 \times \widehat{SE}[\bar{Y}] - \bar{Y} \leq -\mu \leq 1.96 \times \widehat{SE}[\bar{Y}] - \bar{Y}\right) &= 0.95 \\ \mathbb{P}\left(\bar{Y} - 1.96 \times \widehat{SE}[\bar{Y}] \leq \mu \leq \bar{Y} + 1.96 \times \widehat{SE}[\bar{Y}]\right) &= 0.95 \end{aligned}$$

*Different confidence intervals*

- Confidence intervals take the following form:  $\bar{Y} \pm 1.96 \times \widehat{SE}[\bar{Y}]$ , which is shorthand for  $[LB, UB]$  where these are:

$$\begin{aligned} LB &= \bar{Y} - 1.96 \times \widehat{SE}[\bar{Y}] \\ UB &= \bar{Y} + 1.96 \times \widehat{SE}[\bar{Y}] \end{aligned}$$

- Where did the 1.96 come from?

$$\mathbb{P}\left(-1.96 \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq 1.96\right) = 0.95$$

- What about confidence intervals other than 95%? Say we want an  $(1 - \alpha)\%$  confidence interval:

$$\begin{aligned} \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq z_{\alpha/2}\right) &= (1 - \alpha) \\ \mathbb{P}\left(\bar{Y} - z_{\alpha/2} \times \widehat{SE}[\bar{Y}] \leq \mu \leq \bar{Y} + z_{\alpha/2} \times \widehat{SE}[\bar{Y}]\right) &= (1 - \alpha) \end{aligned}$$

- General formula for a  $(1 - \alpha)\%$  confidence interval:

$$\bar{Y} \pm z_{\alpha/2} \times \widehat{SE}[\bar{Y}]$$

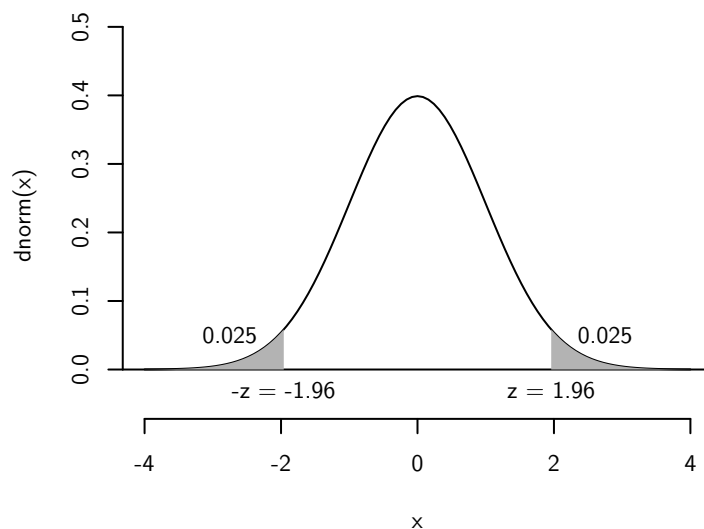
- Here we call the  $z_{\alpha/2}$  values the z-values for the particular confidence intervals.

### Finding the $z$ values

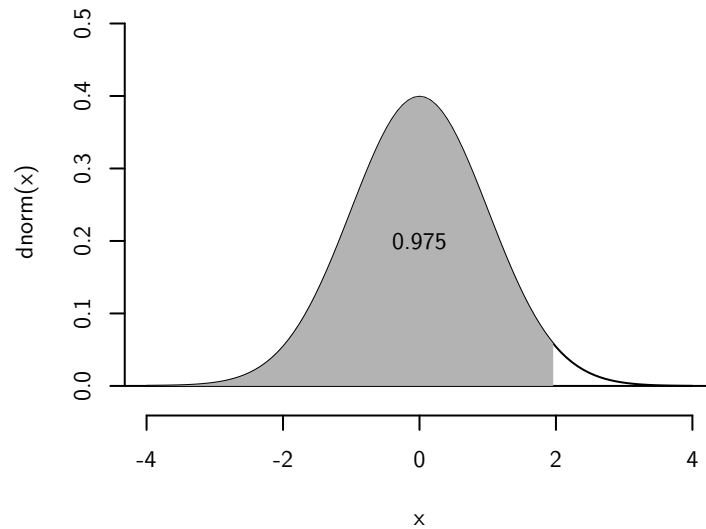
- How do we figure out what  $z_{\alpha/2}$  will be? We will need to find the values such that the following holds. Given a standard Normal variable,  $Z \sim N(0, 1)$ :

$$\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

- Thus, we want the values that puts  $\alpha/2$  in each of the tails.
- For example, with  $\alpha = 0.05$  for a 95% confidence interval, we want the  $z$  values that put 0.025 (2.5%) in each of the tails:



- But how did we find the 1.96 in this case? For these types of two-sided confidence intervals, we will simply find the  $z$ -value that puts  $\alpha/2$  of the probability mass above it or, equivalently,  $1 - \alpha/2$  below it. Thus, we want to know the value  $z$ , such that  $\mathbb{P}(Z \leq z) = 1 - \alpha/2$ :



- This is just the CDF of the standard Normal evaluated at  $1 - \alpha/2$ ! We can find this using the `qnorm()` function in R.
- Procedure:
  1. Choose a value  $\alpha$  (0.1 for example) for a  $100(1 - \alpha)\%$  confidence interval (90% in this case)
  2. Convert this to  $1 - \alpha/2$  (0.95 in this case)
  3. Plug this value into `qnorm()` to find  $z_{\alpha/2}$ :

```
qnorm(0.95)
```

```
## [1] 1.644854
```

- **Question** What happens to the size of the confidence interval when we increase our confidence, from say 95% to 99%? Do confidence intervals get wider or shorter?

*Confidence intervals you should probably memorize*

- For a 90% CI, we have  $\alpha = 0.1$  and  $z_{0.05} = 1.64$
- For a 95% CI, we have  $\alpha = 0.05$  and  $z_{0.025} = 1.96$
- For a 99% CI, we have  $\alpha = 0.01$  and  $z_{0.005} = 2.58$

### *Interpreting the confidence interval*

- **Caution** An often recited, but **incorrect** interpretation of a confidence interval is the following: “I calculated a 95% confidence interval for the sample mean of [0.05,0.13], which means that there is a 95% chance that the true population mean in is that interval.” This is WRONG.
- The true value of the population mean,  $\mu$ , is **fixed**. It is either in the interval or it isn't—there's no room for probability at all. The randomness is in the interval:  $\bar{Y} \pm 1.96 \times \hat{SE}[\bar{Y}]$ . This is what varies from sample to sample.
- Correct interpretation: across 95% of random samples, the constructed confidence interval will contain the true value.

### *Confidence interval simulation*

- Let's do a simulation to show how this works. We're going to draw samples of size 500 from a  $N(1, 10)$  and then calculate the confidence interval for population mean and then calculate the proportion of the intervals that contain the true value of the population mean.

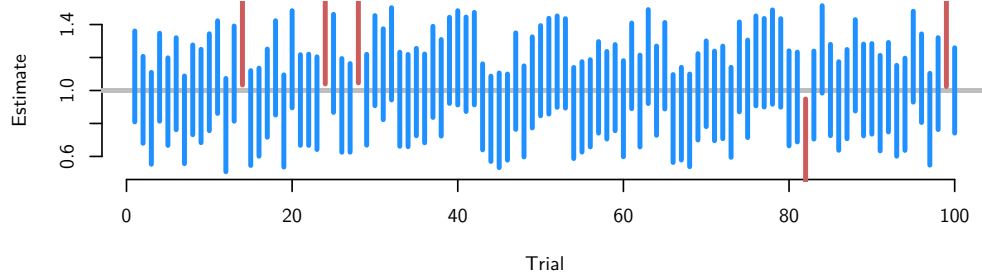
```
set.seed(02143)
sims<- 10000
cover <- rep(NA, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for(i in 1:sims){
  draws <- rnorm(500, mean=1, sd=sqrt(10))
  low.bound[i] <- mean(draws) - sd(draws)/sqrt(500) * 1.96
  up.bound[i] <- mean(draws) + sd(draws)/sqrt(500) * 1.96
  if (low.bound[i] < 1 & up.bound[i] > 1) {
    cover[i] <- 1
  } else {
    cover[i] <- 0
  }
}
mean(cover)
```

```
## [1] 0.9498
```

- We can see that roughly 95% of the confidence intervals (across the samples) contain the true value. Let's plot the first 100 of these interval estimates from

the first 100 samples and then color code the intervals by whether they cover the true value:

```
cols <- ifelse(cover == 1, "dodgerblue", "indianred")
plot(x = NULL, y = NULL, xlim = c(1,100), ylim = c(0.5, 1.5), bty = "n",
     xlab = "Trial", ylab = "Estimate")
abline(h = 1, col = "grey", lwd = 3)
segments(x0 = 1:100, y0 = low.bound[1:100], y1 = up.bound[1:100],
        col = cols[1:100], lwd = 3)
```



- You can see that in these 100 samples, exactly 95 of the calculated confidence intervals contains the true value.
- **Question** What happens to the intervals when we increase  $n$ , the sample size? Do the intervals get shorter or longer? Do more or fewer of them contain the true value?

#### *Interval estimation of the difference in means*

- Suppose we have two iid samples (also independent from each other) from possibly different distributions:
  - $Y_{11}, Y_{12}, \dots, Y_{1n_1}$  with population mean  $\mu_1$  and population variance  $\sigma_1^2$
  - $Y_{21}, Y_{22}, \dots, Y_{2n_2}$  with population mean  $\mu_2$  and population variance  $\sigma_2^2$
- Last week we saw that the sample difference in means is unbiased so we know that  $\mathbb{E}[\bar{Y}_1 - \bar{Y}_2] = \mu_1 - \mu_2$ .

- Let's calculate the sampling variance of the difference in means:

$$\begin{aligned}
 \text{Var}[\bar{Y}_1 - \bar{Y}_2] &= \text{Var}[\bar{Y}_1] + \text{Var}[\bar{Y}_2] \quad (\text{independence + variance properties}) \\
 &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \text{Var}[Y_{1i}] + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \text{Var}[Y_{2i}] \\
 &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sigma_1^2 + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sigma_2^2 \\
 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
 \end{aligned}$$

- Putting all of this together, we know that in large samples, we have the following:

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

- Using the logic from the last section, we can replace the true population variances with sample estimates:

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})$$

- Combining this with the arguments above, we can construct the following confidence interval for the difference in means in large samples:

$$(\bar{Y}_1 - \bar{Y}_2) \pm z_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

### *Interval estimation of the population proportion*

- Let's say that we have a sample of iid Bernoulli random variables,  $Y_1, \dots, Y_n$ , where each takes  $Y_i = 1$  with probability  $\pi$ . Note that this is also the **population proportion** of ones. We have shown in previous weeks that the expectation of one of these variable is just the probability of seeing a 1:  $\mathbb{E}[Y_i] = \pi$ .
- The **variance of a Bernoulli random variable** is a simple function of its mean:  $\text{Var}(Y_i) = \pi(1 - \pi)$ .
- **Problem** Show that the sample proportion,  $\hat{\pi} = \frac{1}{n} \sum_{i=1}^n Y_i$ , of the above iid Bernoulli sample, is unbiased for the true population proportion,  $\pi$ , and that the sampling variance is equal to  $\frac{\pi(1-\pi)}{n}$ .

- Note that if we have an estimate of the population proportion,  $\hat{\pi}$ , then we also have an estimate of the sampling variance:  $\frac{\hat{\pi}(1-\hat{\pi})}{n}$ .
- Given the facts from the previous problem, we just apply the same logic from the population mean to show that the following confidence interval:

$$P\left(\hat{\pi} - z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = (1 - \alpha)$$

*Gerber, Green, and Larimer experiment*

- Let's go back to the Gerber, Green, and Larimer experiment from last week. Here are the results of their experiment:

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Figure 2: Experimental results

- Let's use what we have learned up until now and the information in the table to calculate a 95% confidence interval for the difference in proportions voting between the Neighbors group and the Civic Duty group.
- You may assume that the samples within each group are iid and the two samples are independent.
- First, we know that the sample proportion for the Civic Duty  $\hat{\pi}_C$  group will be approximately Normal around the population mean  $\pi_C$  with sampling variance  $(\pi_C(1-\pi_C))/n$ . Same goes for the sample proportion in the Neighbors group,  $\hat{\pi}_N$ .
- Given that these are just sample means (of the underlying Bernoulli voter turnout variables), we can apply the difference in means results from above to show that the difference in proportions has the following distribution:

$$\hat{\pi}_N - \hat{\pi}_C \sim N\left(\pi_N - \pi_C, \frac{\pi_N(1-\pi_N)}{n_N} + \frac{\pi_C(1-\pi_C)}{n_C}\right)$$

- We can replace the variances with our estimates from the sample proportions:

$$\hat{\pi}_N - \hat{\pi}_C \sim N \left( \pi_N - \pi_C, \frac{\hat{\pi}_N(1 - \hat{\pi}_N)}{n_N} + \frac{\hat{\pi}_C(1 - \hat{\pi}_C)}{n_C} \right)$$

- Now, we can use this in the same way as every other example to calculate the 95% confidence interval:

$$(\hat{\pi}_N - \hat{\pi}_C) \pm 1.96 \times \sqrt{\frac{\hat{\pi}_N(1 - \hat{\pi}_N)}{n_N} + \frac{\hat{\pi}_C(1 - \hat{\pi}_C)}{n_C}}$$

- Let's plug in the values into R and go!

```
n.n <- 38201
samp.var.n <- (0.378 * (1 - 0.378))/n.n
n.c <- 38218
samp.var.c <- (0.315 * (1 - 0.315))/n.c
se.diff <- sqrt(samp.var.n + samp.var.c)
se.diff
```

```
## [1] 0.003435201
```

```
## lower bound
lb <- (0.378 - 0.315) - 1.96 * se.diff
lb
```

```
## [1] 0.05626701
```

```
## upper bound
ub <- (0.378 - 0.315) + 1.96 * se.diff
ub
```

```
## [1] 0.06973299
```

- Thus, the confidence interval for the effect is [0.056267, 0.069733].

### *Interval estimation of the population variance*

- **Challenge Problem** Suppose that you have an iid sample,  $Y_1, \dots, Y_n$ , from a population with variance,  $\sigma^2$ . Derive a  $(1 - \alpha)\%$  confidence interval for the sample variance,  $S_n^2$ . (Hint: you may have to look in Wooldridge and learn about a new distribution.)