Social Science Inquiry II

Week 5: Uncertainty and inference, part II

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2024

Loading packages for this class

```
> library(ggplot2)
> set.seed(60637)
```

Continuing inference

- Last class, we assumed we had a finite population that we observe all of, and the source of randomness in what we observed was due to random assignment of treatment. The inference we used there is called *randomization inference.*

Continuing inference

- Last class, we assumed we had a finite population that we observe all of, and the source of randomness in what we observed was due to random assignment of treatment. The inference we used there is called *randomization inference.*
- Now, we'll assume that our data is produced from a random generative process, where we're sampling from some (potentially infinite) population distribution that is not fully observed. The inference we will use in this setting is the type of inference we use for survey sampling.

Continuing inference

- Last class, we assumed we had a finite population that we observe all of, and the source of randomness in what we observed was due to random assignment of treatment. The inference we used there is called *randomization inference.*
- Now, we'll assume that our data is produced from a random generative process, where we're sampling from some (potentially infinite) population distribution that is not fully observed. The inference we will use in this setting is the type of inference we use for survey sampling.
- It's important to consider *what the source of randomness is* and *what population we're making inferences about.*

Back to estimation

- Returning to our example where we flip a coin twice, let X be the number of heads we observe. Our coin is *not* fair, and the probability of getting a heads is 0.75.

Back to estimation

- Returning to our example where we flip a coin twice, let X be the number of heads we observe. Our coin is *not* fair, and the probability of getting a heads is 0.75.
- The random variable's probability distribution is then:

$$f(x) = \begin{cases} 1/16 & x = 0 \\ 3/8 & x = 1 \\ 9/16 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Back to estimation

- We use the coin flip example, but we can compare this to any "random" event where we can code the outcome in terms of a one or zero.

Back to estimation

- We use the coin flip example, but we can compare this to any "random" event where we can code the outcome in terms of a one or zero.
- For example, with respect to the Pager (2003) data, we can use a process like this to model the probability that an employer will hire a white applicant without a criminal record.

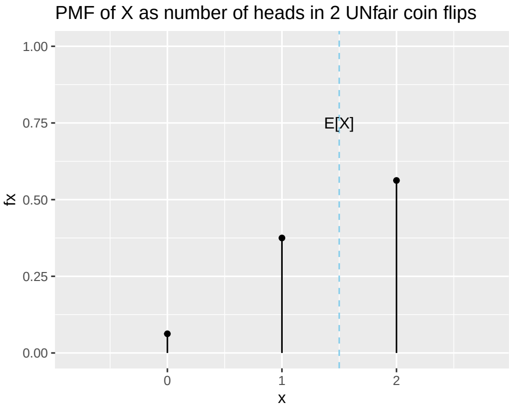
Back to estimation

- We use the coin flip example, but we can compare this to any "random" event where we can code the outcome in terms of a one or zero.
- For example, with respect to the Pager (2003) data, we can use a process like this to model the probability that an employer will hire a white applicant without a criminal record.
- We might say that there are different random processes, with different probabilities of success, for whites with and without criminal records, and blacks with and without criminal records.

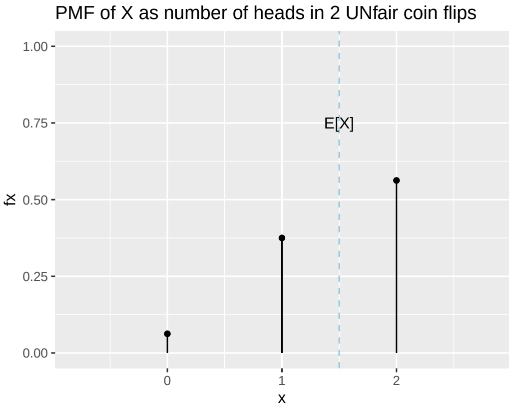
Back to estimation

- We use the coin flip example, but we can compare this to any "random" event where we can code the outcome in terms of a one or zero.
- For example, with respect to the Pager (2003) data, we can use a process like this to model the probability that an employer will hire a white applicant without a criminal record.
- We might say that there are different random processes, with different probabilities of success, for whites with and without criminal records, and blacks with and without criminal records.
- Here, where we have multiple coin flips, we can compare that to the probability distribution of hires for two people with the same profile.

Let's take a look at the mean.

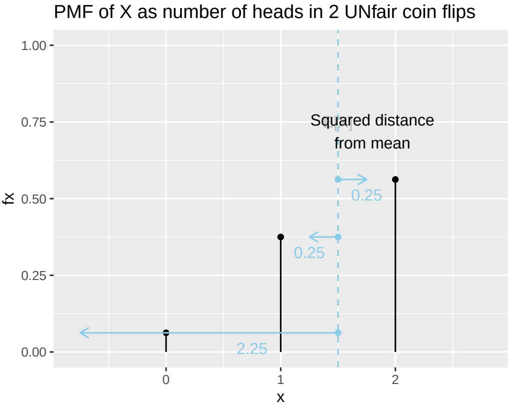


Let's take a look at the mean.



$$\mathrm{E}[X] = \sum_x x f x$$
$$= 0 \times \frac{1}{16} + 1 \times \frac{3}{8} + 2 \times \frac{9}{16} = \frac{24}{16}$$
$$= 1.5$$

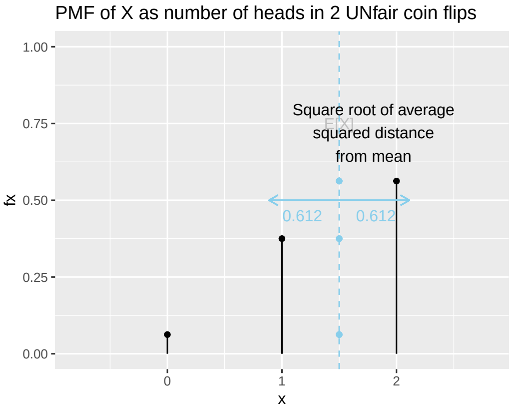
And the spread.



Variance = average squared distance from the mean

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= 2.25 \times \frac{1}{16} + 0.25 \times \frac{3}{8} + 0.25 \times \frac{9}{16} \\ &= 0.375\end{aligned}$$

And the spread.



SD = square root of variance

$$= \sqrt{0.375} = 0.612$$

- We can check our calculations of the expectation and spread in R.

- We can check our calculations of the expectation and spread in R.
- First, we'll want to simulate the random process.

- We can check our calculations of the expectation and spread in R.
- First, we'll want to simulate the random process.

```
> n <- 1000
> X <- c(0, 1, 2)
> probs <- c(1/16, 3/8, 9/16)
> x_observed <- sample(X, prob = probs,
+                      replace = TRUE,
+                      size = n)
> head(x_observed)
```

```
[1] 1 0 1 1 2 2
```

- We can check our calculations of the expectation and spread in R.
- First, we'll want to simulate the random process.

```
> n <- 1000
> X <- c(0, 1, 2)
> probs <- c(1/16, 3/8, 9/16)
> x_observed <- sample(X, prob = probs,
+                      replace = TRUE,
+                      size = n)
> head(x_observed)
```

```
[1] 1 0 1 1 2 2
```

```
> mean(x_observed)
```

```
[1] 1.514
```

- We can check our calculations of the expectation and spread in R.
- First, we'll want to simulate the random process.

```
> n <- 1000
> X <- c(0, 1, 2)
> probs <- c(1/16, 3/8, 9/16)
> x_observed <- sample(X, prob = probs,
+                      replace = TRUE,
+                      size = n)
> head(x_observed)

[1] 1 0 1 1 2 2

> mean(x_observed)

[1] 1.514

> var(x_observed)

[1] 0.3661702

> sd(x_observed)

[1] 0.60512
```

- The process that we just did – sampling and estimation based on observed data – is a very common process in empirical research.

- The process that we just did – sampling and estimation based on observed data – is a very common process in empirical research.
- But we may notice that the mean, variance, and standard deviation are not exactly what we calculated analytically.

Let's try it again.

Let's try it again.

```
> x_observed <- sample(X,
+                      prob = probs,
+                      replace = TRUE,
+                      size = n)
> mean(x_observed)
[1] 1.496
> var(x_observed)
[1] 0.3823664
> sd(x_observed)
[1] 0.6183578
```

- The values that we get are close, but not identical.

- The values that we get are close, but not identical.
- This is because what we are observing in practice is a *sample* from the data.

Sampling and statistics

- Very often, we only observe a limited number of observations, which are drawn from a larger population.

Sampling and statistics

- Very often, we only observe a limited number of observations, which are drawn from a larger population.
- Review:
  - We can summarize the data we observe with *statistics*. Statistics are functions of the data we observe.

Sampling and statistics

- Very often, we only observe a limited number of observations, which are drawn from a larger population.
- Review:
  - We can summarize the data we observe with *statistics*. Statistics are functions of the data we observe.

$$T_n = h(X_1, \ldots, X_n)$$

Sampling and statistics

- Very often, we only observe a limited number of observations, which are drawn from a larger population.
- Review:
  - We can summarize the data we observe with *statistics*. Statistics are functions of the data we observe.

$$T_n = h(X_1, \ldots, X_n)$$

  - (Estimators are a class of statistics that we use to approximate specific estimands. Test statistics are the specific statistics we use to test hypotheses.)

Sampling and statistics

- Because our sampling process is a random process, these statistics themselves are random variables, with their own distributions.

Sampling and statistics

- Because our sampling process is a random process, these statistics themselves are random variables, with their own distributions.
- We can describe our sample, but we might also like to *make inferences* about the larger population–i.e., to summarize what we know about that population based on the data we observe.

Sampling and statistics

- Because our sampling process is a random process, these statistics themselves are random variables, with their own distributions.
- We can describe our sample, but we might also like to *make inferences* about the larger population–i.e., to summarize what we know about that population based on the data we observe.
- This is what we use statistics for, and why we talk about probability AND statistics.

Sampling and statistics

- Because our sampling process is a random process, these statistics themselves are random variables, with their own distributions.
- We can describe our sample, but we might also like to *make inferences* about the larger population–i.e., to summarize what we know about that population based on the data we observe.
- This is what we use statistics for, and why we talk about probability AND statistics.
- Probability gives us a model of the world.

Sampling and statistics

- Because our sampling process is a random process, these statistics themselves are random variables, with their own distributions.
- We can describe our sample, but we might also like to *make inferences* about the larger population–i.e., to summarize what we know about that population based on the data we observe.
- This is what we use statistics for, and why we talk about probability AND statistics.
- Probability gives us a model of the world.
- Statistics give us a way to relate the data that we see to the model.

- In our two coin flip example, suppose we don't know whether the coin is fair or not. We can observe the results of a large number of coin flips, and make an educated guess about the underlying population value.

- In our two coin flip example, suppose we don't know whether the coin is fair or not. We can observe the results of a large number of coin flips, and make an educated guess about the underlying population value.
- Formally, that educated guess is called *estimation*.

Sample mean

Let's repeat our random sampling from the double coin flip, but we'll consider a smaller sample, of size $n = 100$.

```
> n <- 100
> x_observed <- sample(X,
+                      prob = probs,
+                      replace = TRUE,
+                      size = n)
> head(x_observed)
[1] 1 1 1 2 2 1
```

Sample mean

- Our *sample mean* is the mean we observe in our data.

Sample mean

- Our *sample mean* is the mean we observe in our data.
- This is one of the most commonly used sample statistics. It's called a plug-in estimator, because we just "plug in" the sample analog of the population quantity that we're interested in.

Sample mean

- Our *sample mean* is the mean we observe in our data.
- This is one of the most commonly used sample statistics. It's called a plug-in estimator, because we just "plug in" the sample analog of the population quantity that we're interested in.

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample mean

- Our *sample mean* is the mean we observe in our data.
- This is one of the most commonly used sample statistics. It's called a plug-in estimator, because we just "plug in" the sample analog of the population quantity that we're interested in.

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

```
> mean(x_observed)
[1] 1.44
```

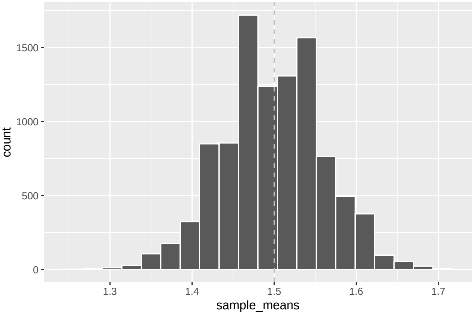
- We differentiate the *sample mean* from the *population mean* because the sample mean will vary with every new sample we draw.

- We differentiate the *sample mean* from the *population mean* because the sample mean will vary with every new sample we draw.
- We'll use a simulation with `replicate()` to see what would happen if we took a sample of size $n = 100$ from the population distribution many times.

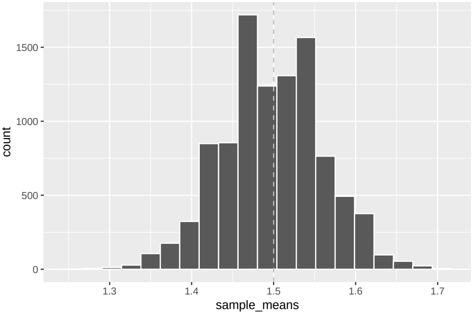
- We differentiate the *sample mean* from the *population mean* because the sample mean will vary with every new sample we draw.
- We'll use a simulation with `replicate()` to see what would happen if we took a sample of size $n = 100$ from the population distribution many times.

```
> n_iter <- 10000
> x_mat <- replicate(n_iter, sample(X,
+                                   prob = probs,
+                                   replace = TRUE,
+                                   size = n))
> dim(x_mat)
[1]   100 10000
> head(x_mat[,1])
[1] 0 2 2 2 2 2
> head(x_mat[,2])
[1] 2 2 1 2 1 0
```

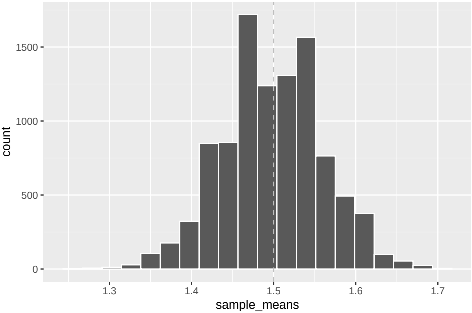
```
> sample_means <- apply(x_mat, 2, mean)
> length(sample_means)
[1] 10000
> head(sample_means)
[1] 1.59 1.51 1.54 1.47 1.43 1.46
```



count
1500
1000
500
0
1.3
1.4
1.5
1.6
1.7
sample_means



We see the sample means are roughly distributed around the mean of the underlying population, Ex.



We see the sample means are roughly distributed around the mean of the underlying population, `Ex`.

The expected value of the sample mean is the population mean.

Sample variance

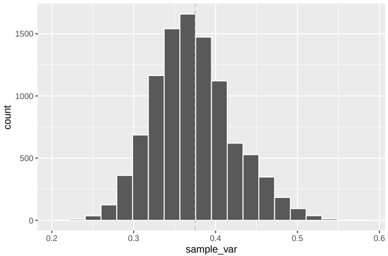
We can estimate the mean of the population using the sample mean.
What about the sample variance?

Sample variance

We'll do the same process with our simulations.

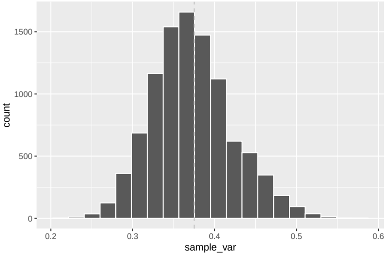
Sample variance

We'll do the same process with our simulations.



Sample variance

We'll do the same process with our simulations.



We see the sample variances are roughly distributed around the variance of the underlying population, $\texttt{sdx}^2$.

The formula for the unbiased sample variance is:

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

The formula for the unbiased sample variance is:

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

This looks a little bit different from a straightforward sample analog to the population variance formula,

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}\left[X\right])^2\right]$$

The formula for the unbiased sample variance is:

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

This looks a little bit different from a straightforward sample analog to the population variance formula,

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$$

Why do we divide by $n-1$, instead of n?

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

- The sample mean, $\bar{X}_n$, has an expected value of $\mathrm{E}[X]$.

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

- The sample mean, $\bar{X}_n$, has an expected value of $\mathrm{E}[X]$.
- However, because it is made up of the $1, \ldots, n$ X_i that we actually observe, the expected difference between $(X_i - \bar{X}_n)$ is a little bit smaller than the expected difference between $(X_i - \mathrm{E}[X])$.

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

- The sample mean, $\bar{X}_n$, has an expected value of $\mathrm{E}[X]$.
- However, because it is made up of the $1, \ldots, n$ X_i that we actually observe, the expected difference between $(X_i - \bar{X}_n)$ is a little bit smaller than the expected difference between $(X_i - \mathrm{E}[X])$.
- To account for this, we divide by $n - 1$, instead of n.

We can check to see how R calculates in the `var()` function.

We can check to see how R calculates in the `var()` function.

```
> head(x_observed)
[1] 1 1 1 2 2 1
```

We can check to see how R calculates in the `var()` function.

```
> head(x_observed)
[1] 1 1 1 2 2 1
> var(x_observed)
[1] 0.329697
```

We can check to see how R calculates in the `var()` function.

```
> head(x_observed)
[1] 1 1 1 2 2 1
> var(x_observed)
[1] 0.329697
> sum( (x_observed - mean(x_observed) )^2)/(n-1)
[1] 0.329697
>
```

We can check to see how R calculates in the `var()` function.

```
> head(x_observed)
[1] 1 1 1 2 2 1
> var(x_observed)
[1] 0.329697
> sum( (x_observed - mean(x_observed) )^2)/(n-1)
[1] 0.329697
>
```

R uses the formula for the unbiased sample variance.

Standard error of the estimator

- The sample mean is itself a random variable, and so it has its own mean and variance. The mean of the sample mean is the population mean. The variance of the sample mean is:

$$\mathrm{Var}[\bar{X}_n] = \frac{\mathrm{Var}[X]}{n}$$

Standard error of the estimator

- The sample mean is itself a random variable, and so it has its own mean and variance. The mean of the sample mean is the population mean. The variance of the sample mean is:

$$\mathrm{Var}[\bar{X}_n] = \frac{\mathrm{Var}[X]}{n}$$

- And the standard deviation of the sample mean is:

$$\sqrt{\mathrm{Var}[\bar{X}_n]} = \sqrt{\frac{\mathrm{Var}[X]}{n}}$$

Standard error of the estimator

- We often refer to the standard deviation of an estimator as the *standard error*.

Standard error of the estimator

- We often refer to the standard deviation of an estimator as the *standard error*.
  - The *standard error* describes the sampling variation of an **estimator**; i.e., how much our estimates will vary based on the random sample that we draw.

Standard error of the estimator

- We often refer to the standard deviation of an estimator as the *standard error*.
    - The *standard error* describes the sampling variation of an **estimator**; i.e., how much our estimates will vary based on the random sample that we draw.
    - *standard deviation* describes the underlying variation in the **population distribution**.

Let's check this in our simulation. We saw that mathematically, $\mathrm{Var}[X]$ was 0.375. So

$$\frac{\mathrm{Var}[X]}{n} = \frac{0.375}{100} = 0.00375$$

Let's check this in our simulation. We saw that mathematically, $\mathrm{Var}[X]$ was 0.375. So

$$\frac{\mathrm{Var}[X]}{n} = \frac{0.375}{100} = 0.00375$$

```
> var(sample_means)
[1] 0.00376872
```

Let's check this in our simulation. We saw that mathematically, $\mathrm{Var}[X]$ was 0.375. So

$$\frac{\mathrm{Var}[X]}{n} = \frac{0.375}{100} = 0.00375$$

```
> var(sample_means)
[1] 0.00376872
```

It's not exactly what we calculated mathematically.

- In fact, from our 10000 separate samples, we calculate 10000 separate sample means. From the variation in these sample means, we again *estimate* the variance of the sample mean.

- In fact, from our 10000 separate samples, we calculate 10000 separate sample means. From the variation in these sample means, we again *estimate* the variance of the sample mean.
- But *this estimate is itself a random variable*, with, again, its own sampling distribution. We will get slightly different estimates of the sampling variance of the sample mean each time we take our 10000 separate samples.

- In fact, from our 10000 separate samples, we calculate 10000 separate sample means. From the variation in these sample means, we again *estimate* the variance of the sample mean.
- But *this estimate is itself a random variable*, with, again, its own sampling distribution. We will get slightly different estimates of the sampling variance of the sample mean each time we take our 10000 separate samples.
- In practice, we will estimate the standard error of the sample mean by plugging our unbiased sample variance formula into the standard error formula:

$$\hat{\text{se}} = \sqrt{S_n^2/n}$$

Weak Law of Large Numbers

- As our sample size n grows, we are increasingly likely to observe a sample mean $\bar{X}_n$ that is close to the mean of the distribution, $\mathrm{E}[X]$.

Weak Law of Large Numbers

- As our sample size n grows, we are increasingly likely to observe a sample mean $\bar{X}_n$ that is close to the mean of the distribution, $\mathrm{E}[X]$.
- Formally,
  If $X_1, \ldots, X_n$ are i.i.d. random variables, then

$$\bar{X}_n \xrightarrow{p} \mathrm{E}[X].$$

Weak Law of Large Numbers

- As our sample size n grows, we are increasingly likely to observe a sample mean $\bar{X}_n$ that is close to the mean of the distribution, $\mathrm{E}[X]$.
- Formally,
  If $X_1, \ldots, X_n$ are i.i.d. random variables, then

$$\bar{X}_n \xrightarrow{p} \mathrm{E}[X].$$

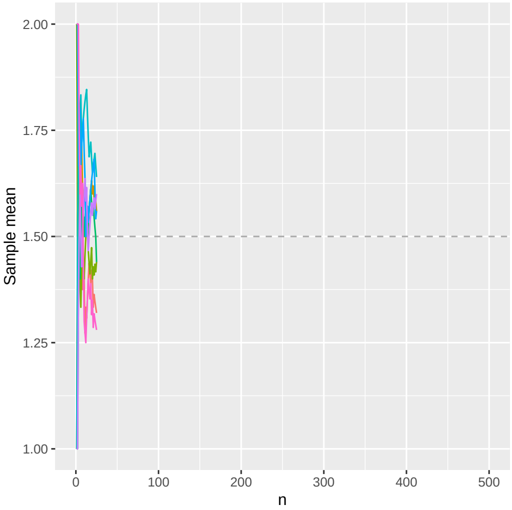
- Convergence in probability, $\xrightarrow{p}$, here means that the probability that we measure a value of $\bar{X}_n$ that is any arbitrary distance away from $\mathrm{E}[X]$ is decreasing with our sample size.

Weak Law of Large Numbers

To see the WLLN in action, we'll try simulating our coinflip process, but with an increasing number of samples used to calculate the sample mean.

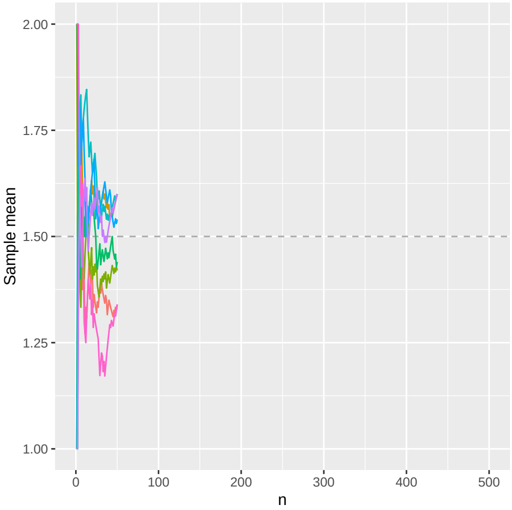
Weak Law of Large Numbers

To see the WLLN in action, we'll try simulating our coinflip process, but with an increasing number of samples used to calculate the sample mean.



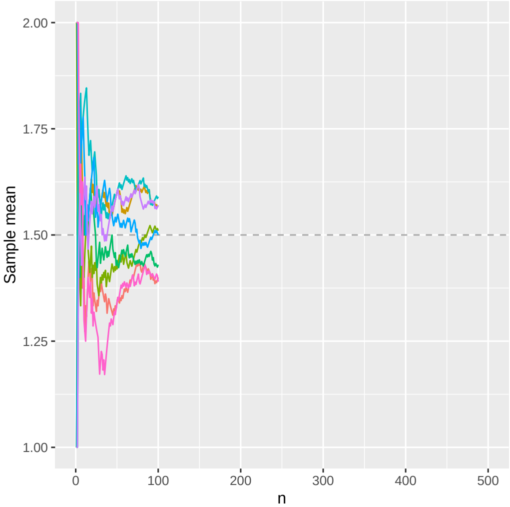
Weak Law of Large Numbers

To see the WLLN in action, we'll try simulating our coinflip process, but with an increasing number of samples used to calculate the sample mean.



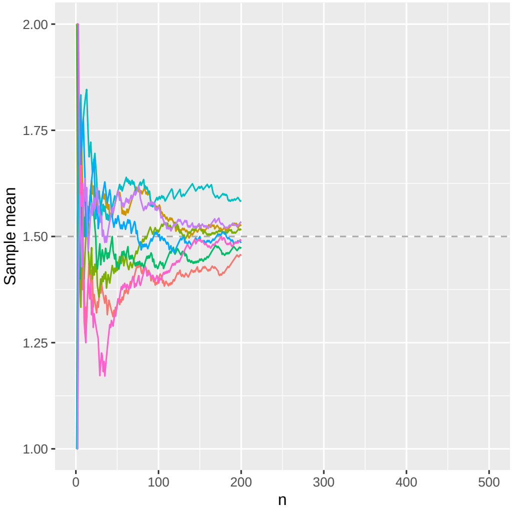
Weak Law of Large Numbers

To see the WLLN in action, we'll try simulating our coinflip process, but with an increasing number of samples used to calculate the sample mean.



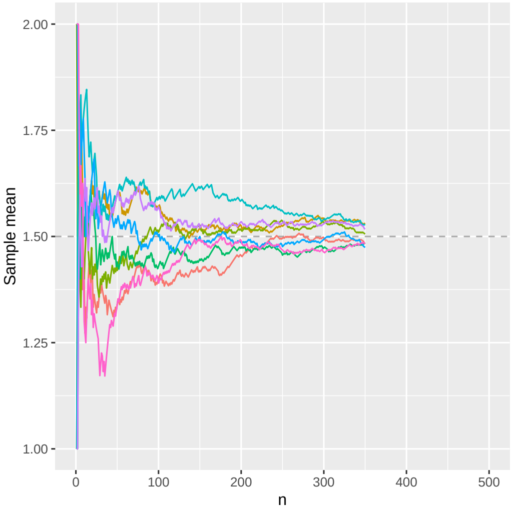
Weak Law of Large Numbers

To see the WLLN in action, we'll try simulating our coinflip process, but with an increasing number of samples used to calculate the sample mean.



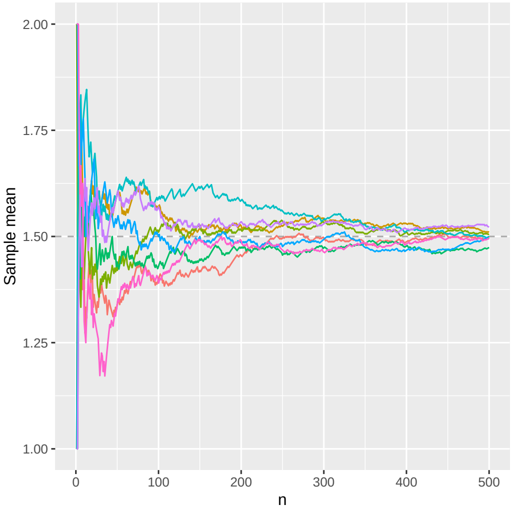
Weak Law of Large Numbers

To see the WLLN in action, we'll try simulating our coinflip process, but with an increasing number of samples used to calculate the sample mean.



Weak Law of Large Numbers

To see the WLLN in action, we'll try simulating our coinflip process, but with an increasing number of samples used to calculate the sample mean.



Weak Law of Large Numbers

- Why is the WLLN so helpful to us?

Weak Law of Large Numbers

- Why is the WLLN so helpful to us?
- Given a sufficient sample from a population, we can estimate features of a random variable to arbitrary precision

Weak Law of Large Numbers

- Why is the WLLN so helpful to us?
- Given a sufficient sample from a population, we can estimate features of a random variable to arbitrary precision
- This is why we can use sample analogs of population features, like the sample mean, as plug-in estimators to estimate the population quantities.

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question?

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:
  - Data (Where does it come from/how is it generated? What is the sample population? What is being measured?)

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:
  - Data (Where does it come from/how is it generated? What is the sample population? What is being measured?)
  - Research design/strategy

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:
  - Data (Where does it come from/how is it generated? What is the sample population? What is being measured?)
  - Research design/strategy
  - Statistical tools

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:
  - Data (Where does it come from/how is it generated? What is the sample population? What is being measured?)
  - Research design/strategy
  - Statistical tools
- What is the answer that the authors get to the main question?

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:
  - Data (Where does it come from/how is it generated? What is the sample population? What is being measured?)
  - Research design/strategy
  - Statistical tools
- What is the answer that the authors get to the main question?

How would you answer these questions with the Pager (2003) paper?

References I

Pager, D. (2003). The mark of a criminal record. American journal of sociology, 108(5):937–975.