Lecture 3: Estimators and their distributions

STA238: Probability, Statistics, and Data Analysis II

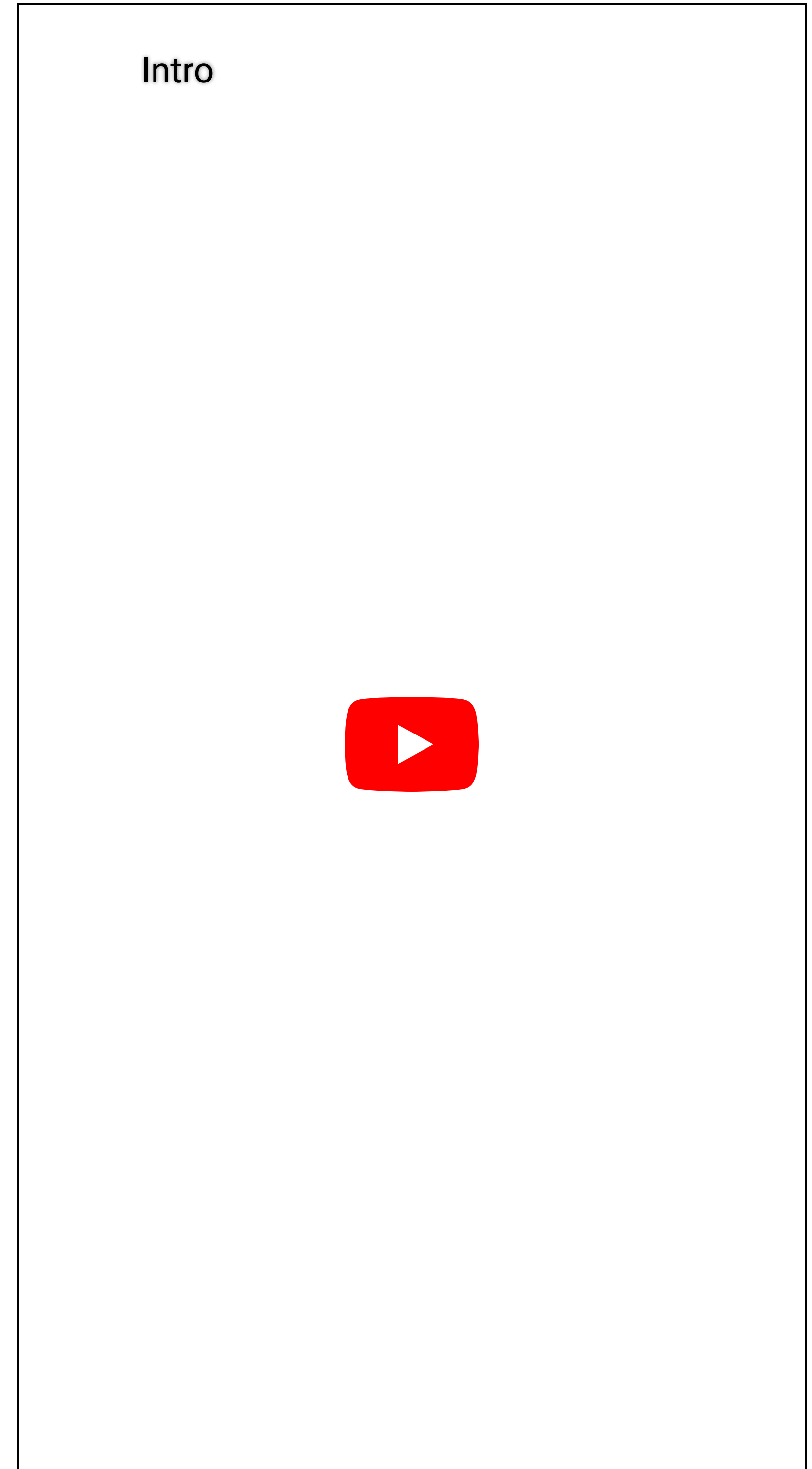
Fred Song

Week of July 8th, 2024 Lec 3

Statistic, Estimator, and Estimates

Example: Stealing bases

The video and data used in the example is from *Stolen Bases* module, SCORE Network, https://isle.stat.cmu.edu/SCORE/stolen-bases-module/ by Andrew Lee and Jacob Hurtubise, 2023.



Example: Pitch delivery and catcher throw times

Questions

- What is the mean time it takes from the start of the pitching until the ball is caught at the second base?
- What is probability of getting to the base before the ball is caught at the second base if you can run in 3.5 seconds?

```r
stealingbases
```

```
# A tibble: 166 × 2
   pitcher_delivery catcher_throw
              <dbl>         <dbl>
 1             1.38          1.82
 2             1.53          1.89
 3             1.15          1.89
 4             1.35          1.89
 5             1.54          1.9
 6             1.22          1.91
 7             1.22          1.92
 8             1.34          1.92
 9             1.25          1.92
10             1.53          1.93
11             1.3           1.93
12             1.2           1.93
13             1.23          1.93
14             1.4           1.93
15             1.12          1.93
16             1.31          1.93
```

Example: Pitch delivery and catcher throw times

```r
stealingbases
```

```
# A tibble: 166 × 2
   pitcher_delivery catcher_throw
              <dbl>         <dbl>
 1             1.38          1.82
 2             1.53          1.89
 3             1.15          1.89
 4             1.35          1.89
 5             1.54          1.9 
 6             1.22          1.91
 7             1.22          1.92
 8             1.34          1.92
 9             1.25          1.92
10             1.53          1.93
11             1.3           1.93
12             1.2           1.93
13             1.23          1.93
14             1.4           1.93
15             1.12          1.93
16             1.31          1.93
```

Questions

- What is the mean time it takes from the start of the pitching until the ball is caught at the second base?
- What is probability of getting to the base before the ball is caught at the second base if you can run in 3.5 seconds?

Model

- `pitcher_delivery` is a random sample of $X \sim \mathcal{N}(\mu_p, \sigma_p^2)$.
- `catcher_throw` is a random sample of $Y \sim \mathcal{N}(\mu_c, \sigma_c^2)$.

Example: Pitch delivery and catcher throw times

Questions

- What is the mean time it takes from the start of the pitching until the ball is caught at the second base?
- What is probability of getting to the base before the ball is caught at the second base if you can run in 3.5 seconds?

Model

- `pitcher_delivery` is a random sample of $X \sim \mathcal{N}(\mu_p, \sigma_p^2)$.
- `catcher_throw` is a random sample of $Y \sim \mathcal{N}(\mu_c, \sigma_c^2)$.

Parameters of interest

Let $W = X + Y$.

- $\mathbb{E}(W)$
- $P(W > 3.5)$

Population

Recall...

> In studying data, we call the collection of objects being studied the **population** of interest and the quantity of interest from the population a **parameter**.

Population

Recall...

> In studying data, we call the collection of objects being studied the **population** of interest and the quantity of interest from the population a **parameter**.

- The population is pitch delivery times by all pitchers and catcher throw times by all catchers...
- ...in all games, including future games...
- It is *impossible* to observe the parameter, a population quantity.

Sample

Recall...

> The subset of the objects collected in the data is the **sample** and an **estimator** is a rule using sample that estimates a parameter. The resulting value is an **estimate** of the parameter.

Sample

Recall…

> The subset of the objects collected in the data is the **sample** and an **estimator** is a rule using sample that estimates a parameter. The resulting value is an **estimate** of the parameter.

- We observe samples.
- We can use sample statistics as estimators based on modelling assumptions.
- e.g., sample mean as an estimator for the population mean assuming the data is a random sample.

Estimator and estimate

An **estimator** T of a parameter θ is a random variable defined as

$$T = h\left(X_1, X_2, \ldots, X_n\right)$$

to approximate the unknown parameter θ based on the sample $\{X_1, X_2, \ldots, X_n\}$.

An **estimate** t is the realized value of an estimator. That is,

$$t = h\left(x_1, x_2, \ldots, x_n\right).$$

t only depends on the observed data set, $\{x_1, x_2, \ldots, x_n\}$.

Estimator and estimate

- An *estimator* is the method or device for estimation.
- An *estimate* is the specific value computed using an estimator.
- An *estimator* is a sample statistic.

An **estimator T** of a parameter θ is a random variable defined as

$$T = h\left(X_1, X_2, \ldots, X_n\right)$$

to approximate the unknown parameter θ based on the sample $\{X_1, X_2, \ldots, X_n\}$.

An **estimate** t is the realized value of an estimator. That is,

$$t = h\left(x_1, x_2, \ldots, x_n\right).$$

t only depends on the observed data set, $\{x_1, x_2, \ldots, x_n\}$.

Example: Pitch delivery and catcher throw times

Model

- `pitcher_delivery` is a random sample of $X \sim \mathcal{N}(\mu_p, \sigma_p^2)$.
- `catcher_throw` is a random sample of $Y \sim \mathcal{N}(\mu_c, \sigma_c^2)$.
- $W = X + Y$ is the time from pitcher to the second baseman.

Parameters of interest

- $\mathbb{E}(W)$

Example: Pitch delivery and catcher throw times

Model

- `pitcher_delivery` is a random sample of $X \sim \mathcal{N}(\mu_p, \sigma_p^2)$.
- `catcher_throw` is a random sample of $Y \sim \mathcal{N}(\mu_c, \sigma_c^2)$.
- $W = X + Y$ is the time from pitcher to the second baseman.

Parameters of interest

- $\mathbb{E}(W)$

Estimator

- The sample mean, $\bar{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i$.

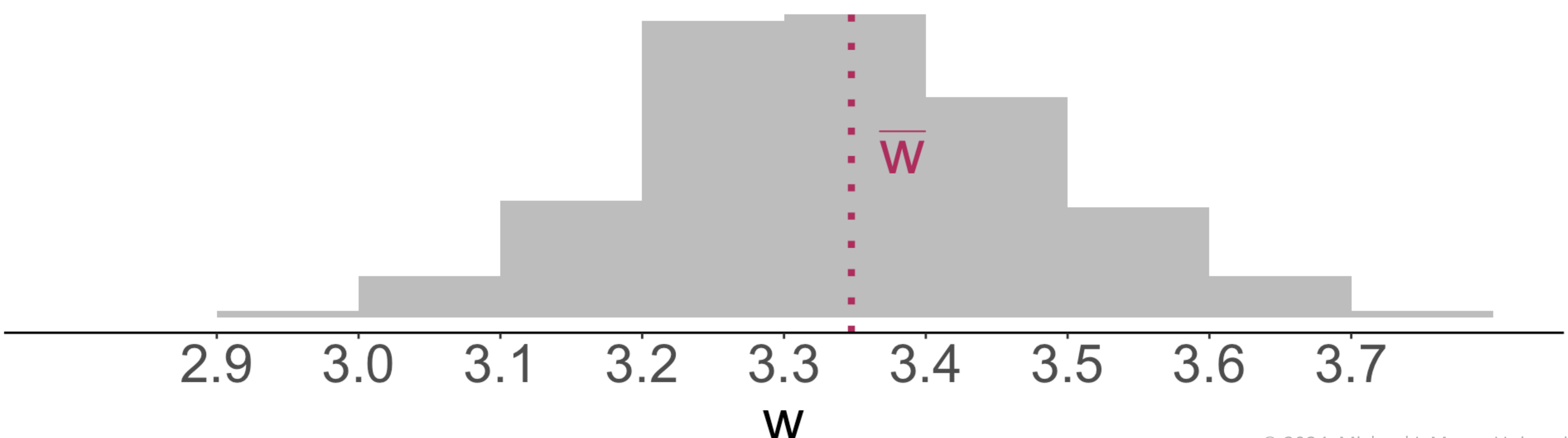
Estimate

```
stealingbases |>
  mutate(w = pitcher_delivery + catcher_
  summarise(wbar = mean(w))
```

```
# A tibble: 1 × 1
   wbar
  <dbl>
1  3.35
```

$\widehat{\mu}_W = \bar{w}_n$ = 3.35

> We often use the hat notation $\hat{\theta}$ to denote the estimator or estimate of the parameter θ.



Example: Pitch delivery and catcher throw times

Model

- `pitcher_delivery` is a random sample of $X \sim \mathcal{N}(\mu_p, \sigma_p^2)$.
- `catcher_throw` is a random sample of $Y \sim \mathcal{N}(\mu_c, \sigma_c^2)$.
- $W = X + Y$ is the time from pitcher to the second baseman.

Parameters of interest

- $P(W > 3.5)$ — let p denote the parameter.

Estimator

- The relative frequency of $W_i > 3.5$,

$$\hat{p}_{(1)} = \frac{\sum_{i=1}^{n} \mathbb{I}(W_i > 3.5)}{n}.$$

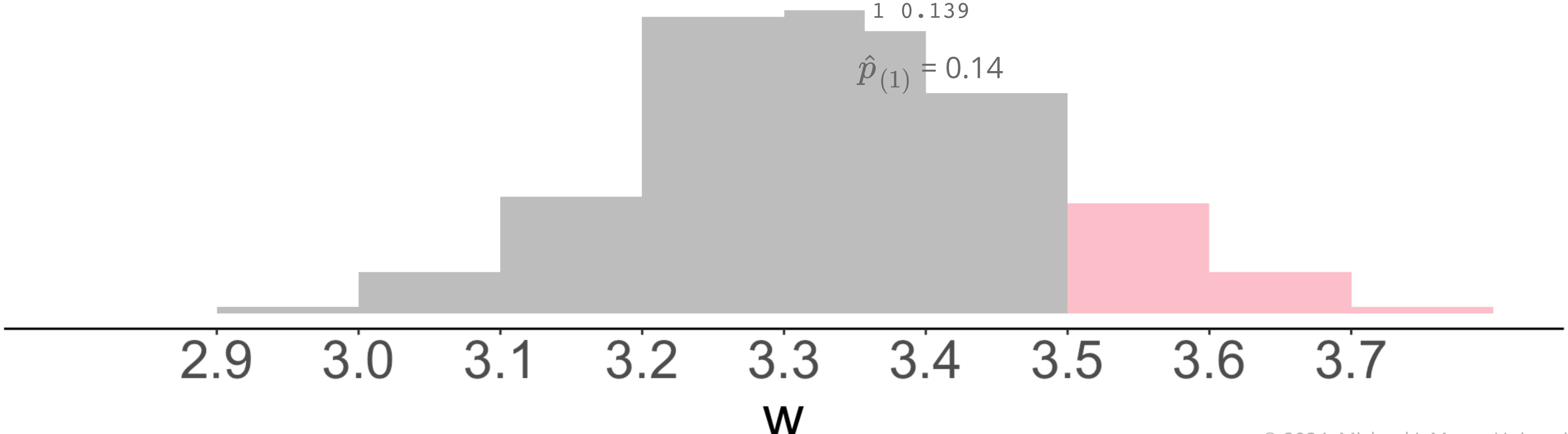
> $\mathbb{I}(W_i > 3.5) = 1$ when $W_i > 3.5$ and 0 otherwise.

Estimate

```r
stealingbases |>
  mutate(w = pitcher_delivery + catcher_
  summarise(phat1 = mean(w > 3.5))
```

```
# A tibble: 1 × 1
  phat1
  <dbl>
1 0.139
```

$\hat{p}_{(1)}$ = 0.14



Example: Pitch delivery and catcher throw times

Model

- `pitcher_delivery` is a random sample of $X \sim \mathcal{N}(\mu_p, \sigma_p^2)$.
- `catcher_throw` is a random sample of $Y \sim \mathcal{N}(\mu_c, \sigma_c^2)$.
- $W = X + Y$ is the time from pitcher to the second baseman.

Parameters of interest

- $P(W > 3.5)$ — let p denote the parameter.

Estimate

- $\hat{p}_{(1)} = \frac{\sum_{i=1}^{n} \mathbb{I}(W_i > 3.5)}{n} =$ 0.14

Estimator

- Probability based on the estimated model $\mathcal{N}(\widehat{\mu}_W, \sigma_W^2)$.

What is the distribution of W?

> For simplicity, assume σ_P^2=0.0181 and σ_C^2=0.0029 are known.

Example: Pitch delivery and catcher throw times

Model

- `pitcher_delivery` is a random sample of $X \sim \mathcal{N}(\mu_p, \sigma_p^2)$.
- `catcher_throw` is a random sample of $Y \sim \mathcal{N}(\mu_c, \sigma_c^2)$.
- $W = X + Y$ is the time from pitcher to the second baseman.

Parameters of interest

- $P(W > 3.5)$ — let p denote the parameter.

Estimate

- $\hat{p}_{(1)} = \frac{\sum_{i=1}^{n} \mathbb{I}(W_i > 3.5)}{n} = 0.14$

Estimator

- Probability based on the estimated model $\mathcal{N}(\widehat{\mu}_W, \sigma_W^2)$.

Assuming the performance of pitchers and catchers are independent,

$$W \sim \mathcal{N}(\mu_W, \sigma_p^2 + \sigma_c^2)$$

We can use $\widehat{\mu}_W = \bar{w}_n$ to estimate the distribution, $\widehat{F}$ and $\hat{p}_{(2)} = 1 - \widehat{F}(3.5)$.

Estimate

```r
wbar <- stealingbases |>
  mutate(w = pitcher_delivery + catcher_
  summarise(wbar = mean(w)) |>
  pull(wbar)
# use the wbar estimate
1 - pnorm(3.5, wbar, sqrt(0.0181 + 0.002
```

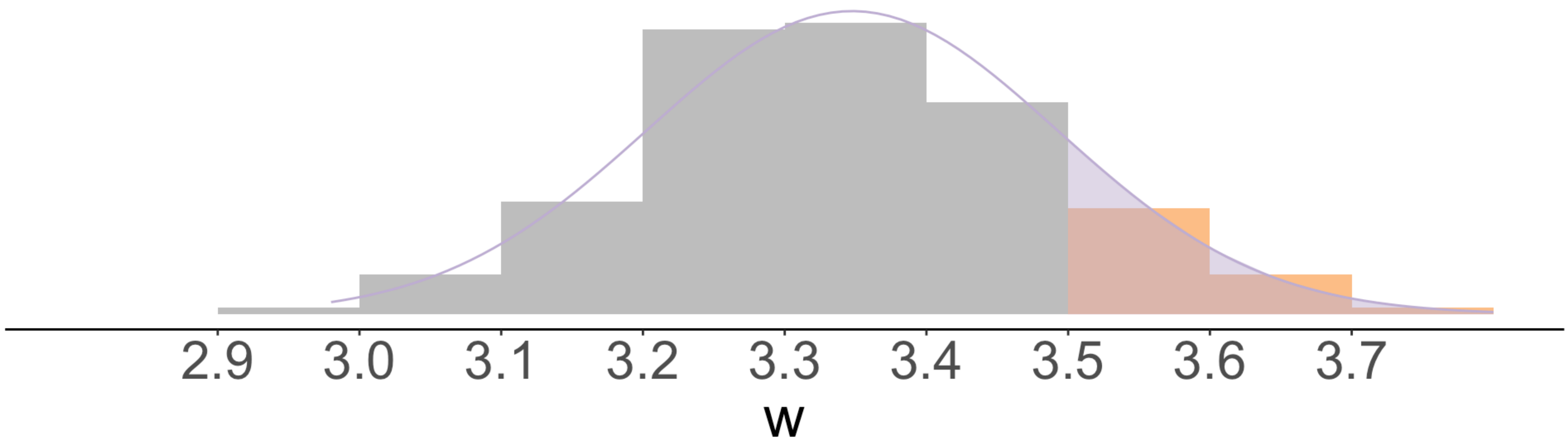
```
[1] 0.1463672
```

$\hat{p}_{(2)}$ = 0.15

Which estimator is better?

$$\hat{p}_{(1)} = \frac{\sum_{i=1}^{n} \mathbb{I}(W_i > 3.5)}{n} = 0.14$$

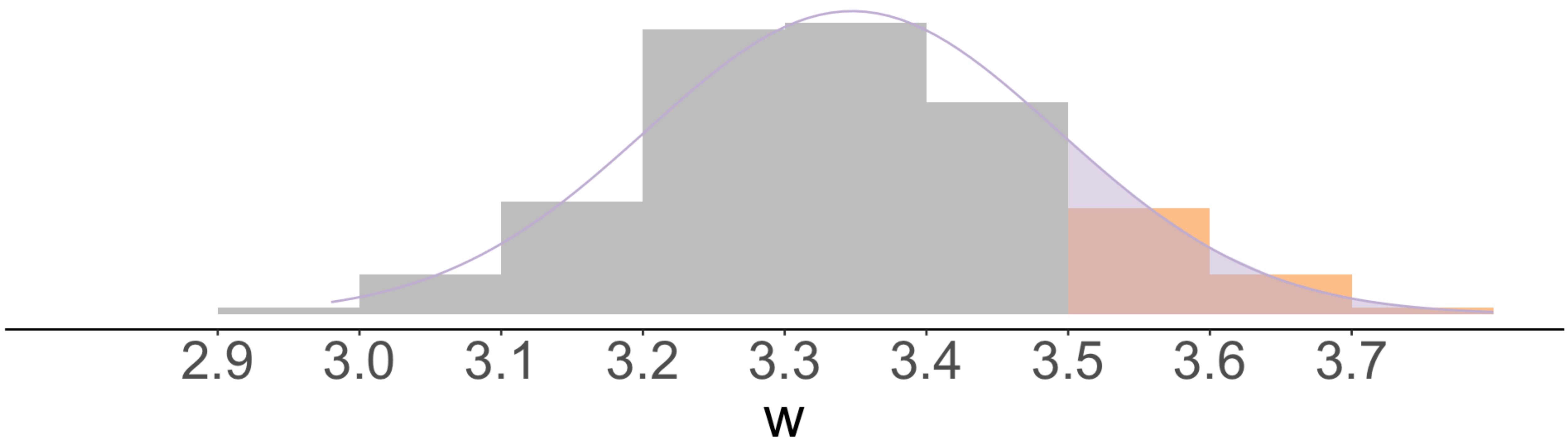
$$\hat{p}_{(2)} = 1 - \widehat{F}(3.5) = 0.15$$



Which estimator is better?

$$\hat{p}_{(1)} = \frac{\sum_{i=1}^{n} \mathbb{I}(W_i > 3.5)}{n} = 0.14$$

$$\hat{p}_{(2)} = 1 - \widehat{F}(3.5) = 0.15$$



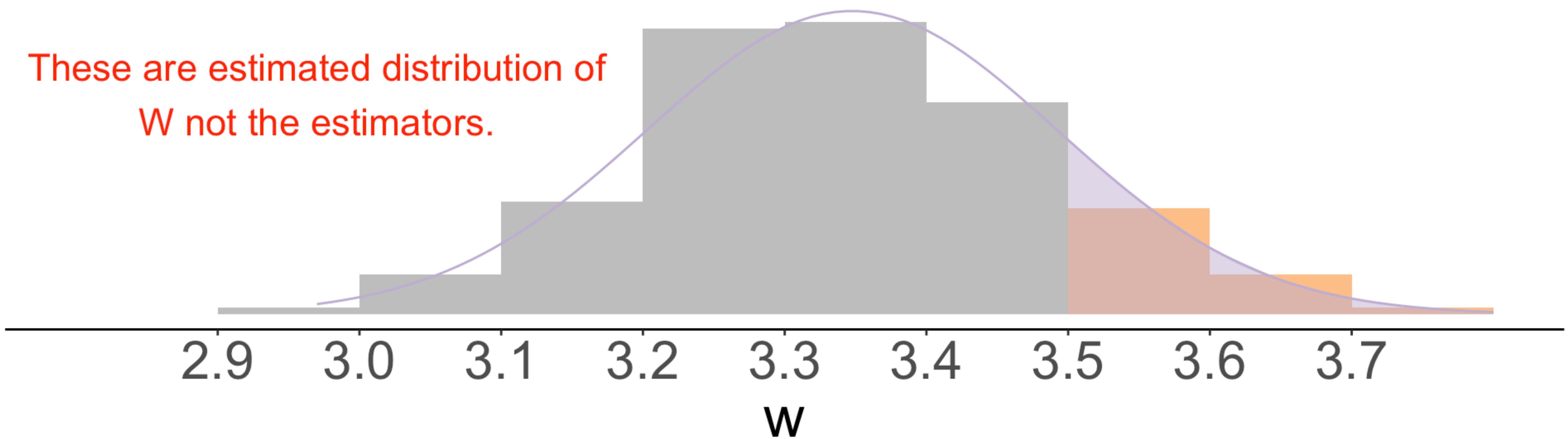
We can study the distributions of the estimators.

Which estimator is better?

$$\hat{p}_{(1)} = \frac{\sum_{i=1}^{n} \mathbb{I}(W_i > 3.5)}{n} = 0.14$$

$$\hat{p}_{(2)} = 1 - \widehat{F}(3.5) = 0.15$$

These are estimated distribution of
W not the estimators.



We can study the distributions of the estimators.

Sampling distribution

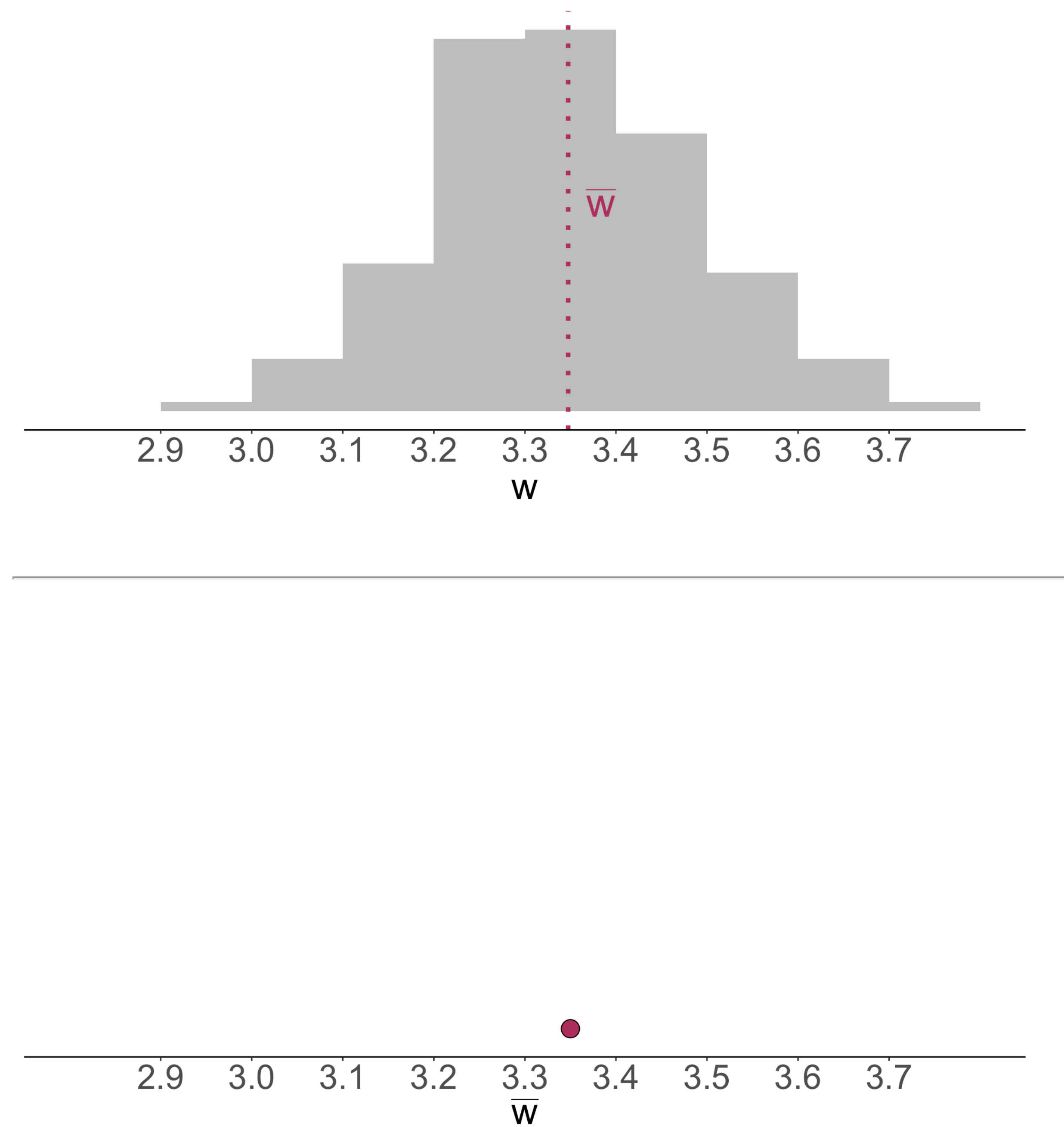
Sampling distribution

- An estimator, like any sample statistic, is a random variable with a probability distribution associated with it.

Let $T = h(X_1, X_2, \ldots, X_n)$ be an estimator based on a random sample X_1, X_2, ..., X_n. The probability distribution of T is called the **sampling distribution** of T.

Example: Sample mean of W

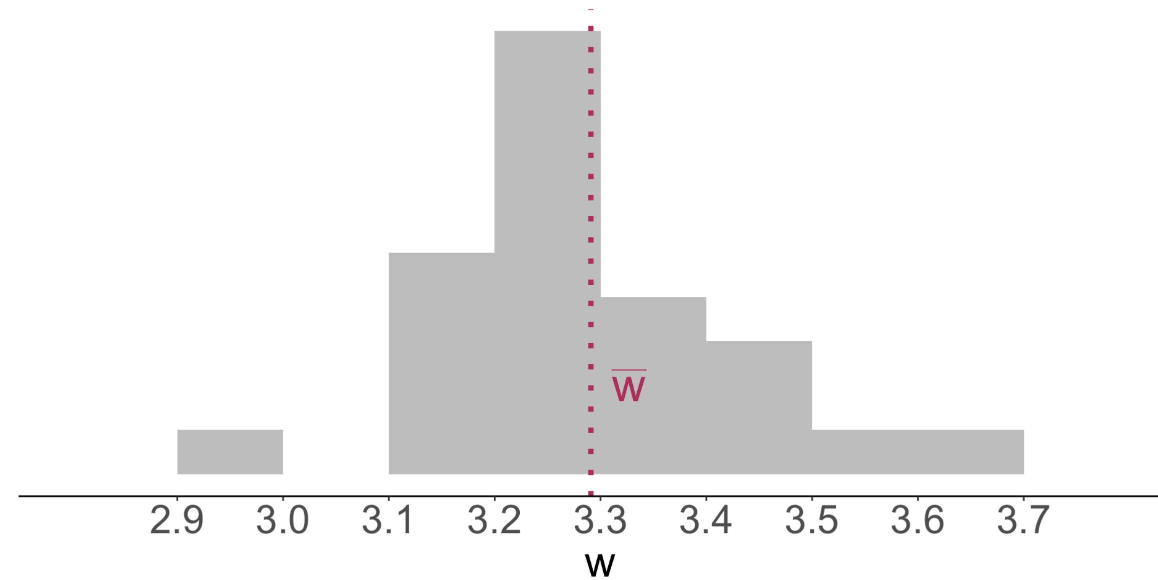
Suppose $W \sim \mathcal{N}(3.28, 0.021)$. This implies the observed data is a random sample of the distribution.

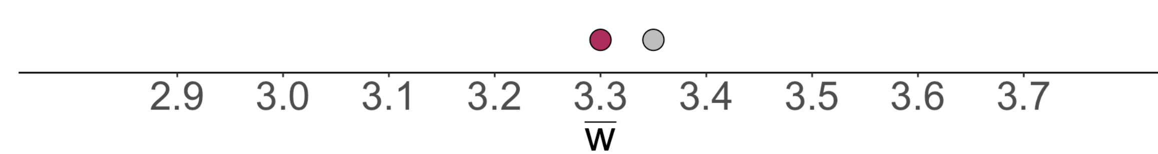


Example: Sample mean of W

Suppose $W \sim \mathcal{N}(3.28, 0.021)$. This implies the observed data is a random sample of the distribution.

If we observed another set of data, we may get a different estimate.



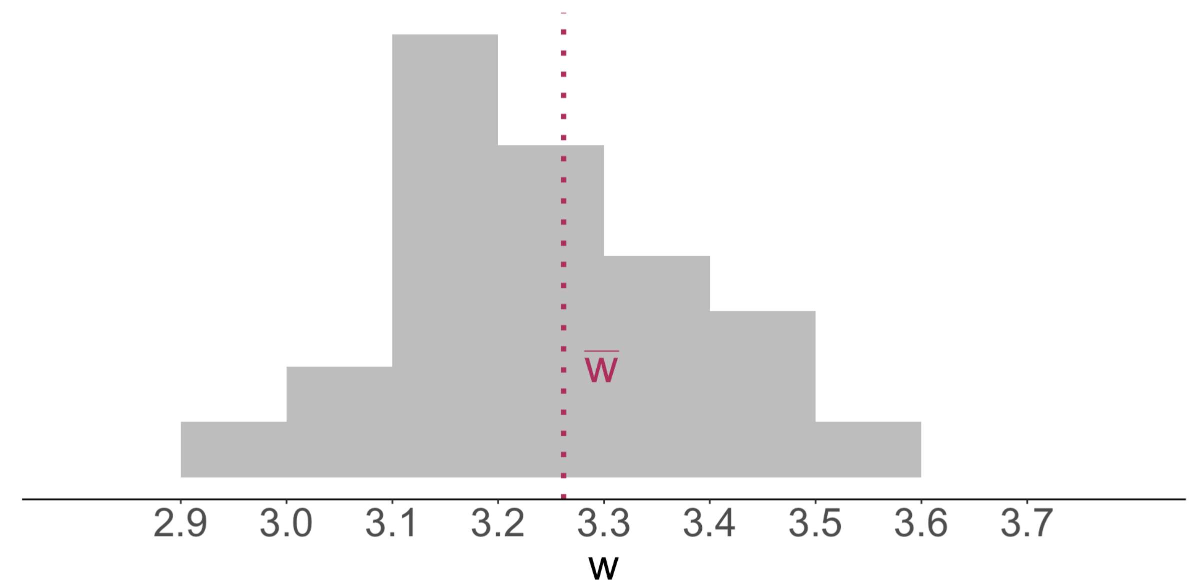


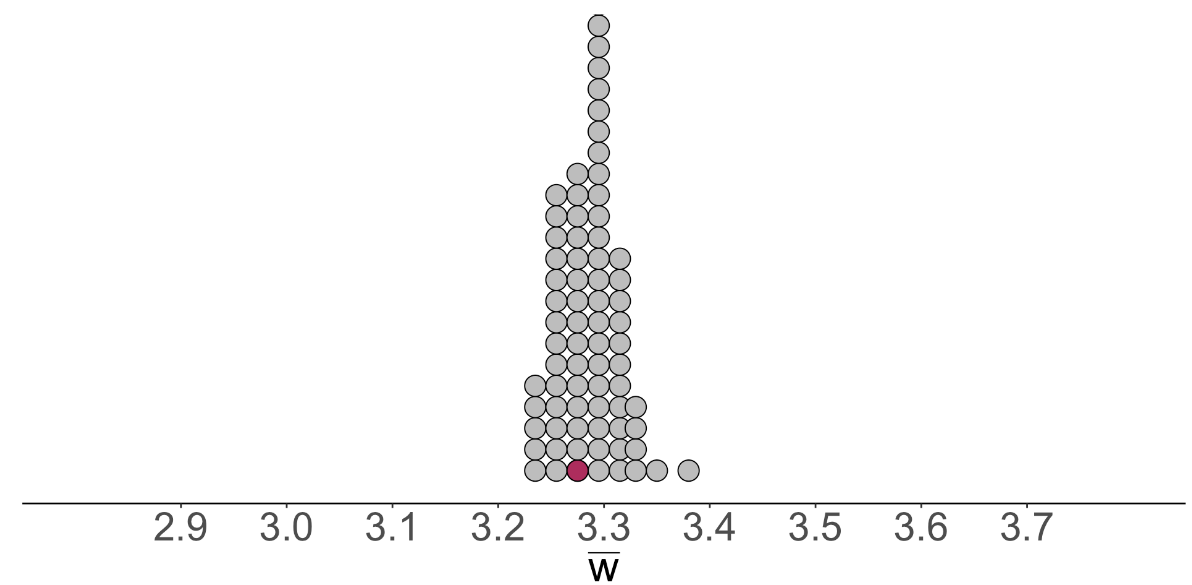
Example: Sample mean of W

Suppose $W \sim \mathcal{N}(3.28, 0.021)$. This implies the observed data is a random sample of the distribution.

If we observed another set of data, we may get a different estimate.

If we could repeat the procedure many times, we would be able to estimate the sampling distribution.





Example: Sample mean of W

Suppose $W \sim \mathcal{N}(3.28, 0.021)$. This implies the observed data is a random sample of the distribution.

If we observed another set of data, we may get a different estimate.

If we could repeat the procedure many times, we would be able to estimate the sampling distribution.

Based on the model, what is the distribution of $\bar{W}_n$?

Example: Sample mean of W

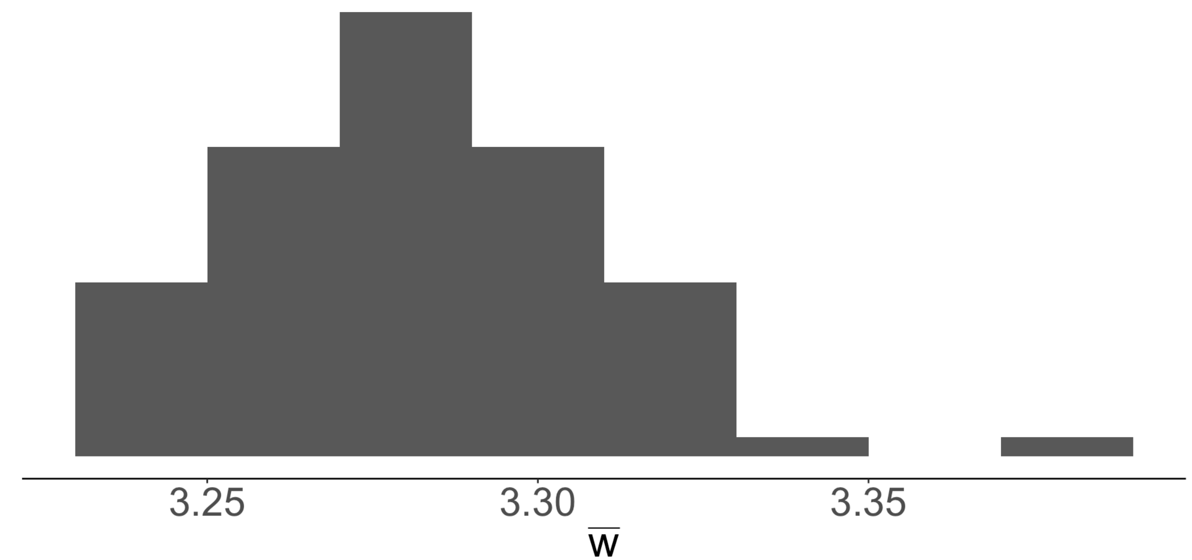
Suppose $W \sim \mathcal{N}(3.28, 0.021)$. This implies the observed data is a random sample of the distribution.

If we observed another set of data, we may get a different estimate.

If we could repeat the procedure many times, we would be able to estimate the sampling distribution.

Based on the model, what is the distribution of $\bar{W}_n$?

$$\bar{W}_n \sim \mathcal{N}\left(3.28, \frac{0.021}{n}\right)$$



Sampling distribution of a normal mean

In general, the sample mean of a random sample from $\mathcal{N}(\mu, \sigma^2)$ follows the following sampling distribution:

$$\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

What does the sampling distribution tell us about the quality of the estimator?

Let's begin with the expected value.

We want the expected value to be

________________________ .

What does the sampling distribution tell us about the quality of the estimator?

Let's begin with the expected value.

$\bar{W}_n$ is an example of a ________________ estimator of $\mathbb{E}(W)$.

An estimator T is called an **unbiased estimator** for the parameter θ if

$$\mathbb{E}(T) = \theta$$

irrespective of the value of θ.

The difference $\mathbb{E}(T) - \theta$ is called the **bias** of T; if the difference is nonzero, then T is called **biased**.

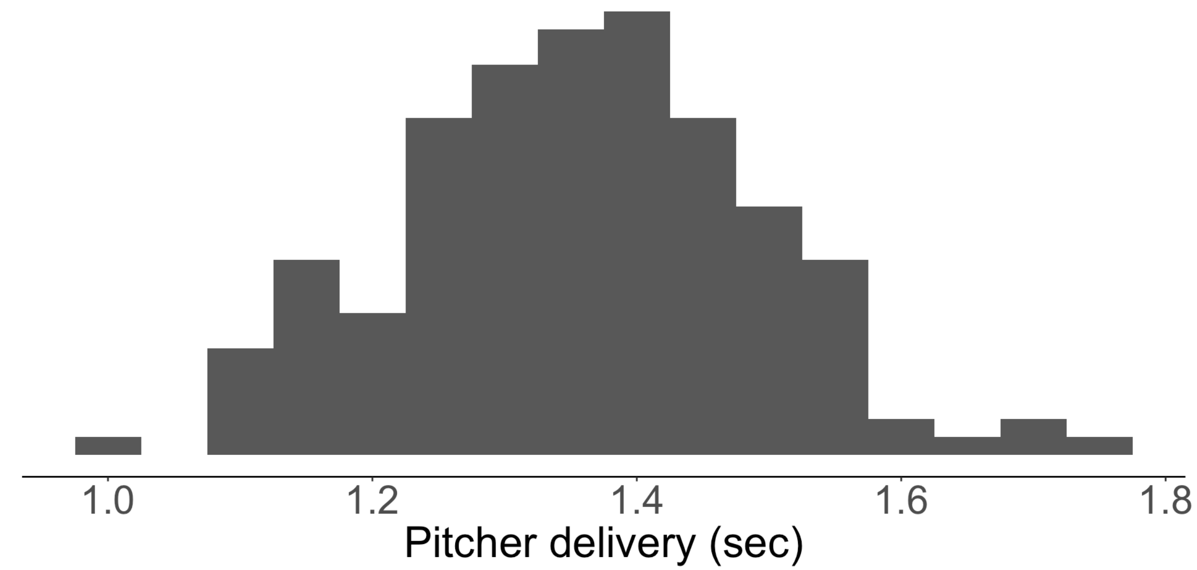
Example: Unknown variance

Suppose the population variance σ_p^2 for pitcher delivery time was unknown.

Is $\widetilde{S}_n^2$ defined as

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

where $\bar{X}_n = \sum_{i=1}^{n} X_i \big/ n$ a *good* estimator of σ_p^2?



Example: Unknown variance

Suppose the population variance σ_p^2 for pitcher delivery time was unknown.

Is $\widetilde{S}_n^2$ defined as

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

where $\bar{X}_n = \sum_{i=1}^{n} X_i \big/ n$ a *good* estimator of σ_p^2?

We can show that

$$\widetilde{S}_n^2 \xrightarrow{p} \sigma_p^2.$$

We won't prove this in this class.

> When an estimator converges in probability to the parameter of interest, we say the estimator is a **consistent** estimator.

Is it unbiased?

Example: Unknown variance

Suppose the population variance σ_p^2 for pitcher delivery time was unknown.

Is $\widetilde{S}_n^2$ defined as

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

where $\bar{X}_n = \sum_{i=1}^{n} X_i \big/ n$ a *good* estimator of σ_p^2?

Example: Unknown variance

Suppose the population variance σ_p^2 for pitcher delivery time was unknown.

Is $\widetilde{S}_n^2$ defined as

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

where $\bar{X}_n = \sum_{i=1}^{n} X_i \big/ n$ a *good* estimator of σ_p^2?

1. $\widetilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \bar{X}_n^2$
2. $\mathrm{Var}(\bar{X}_n) = \mathbb{E}(\bar{X}_n^2) - \left[\mathbb{E}\left(X_1\right)\right]^2$
3. $\mathrm{Var}(X_1) = \mathbb{E}(X_1^2) - \left[\mathbb{E}\left(X_1\right)\right]^2$

Together, they imply that

$$\mathbb{E}(\widetilde{S}_n^2) = \frac{n-1}{n}\sigma_p^2.$$

> $\widetilde{S}_n^2$ is a biased estimator for σ_p^2. It has a negative bias.

On the other hand,

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

is an unbiased estimator of the variance.

Unbiased estimators for expectation and variance

> This solves the mystery of $1/(n-1)$ in sample variance.

Suppose X_1, X_2, ..., X_n is a random sample from a distribution with finite expectation μ and finite variance σ^2.

Then the sample mean

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

is an unbiased estimator for μ and the sample variance

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

is an unbiased estimator for σ^2.

Example: Coffee shop

Suppose the daily count of coffees sold at a coffee shop follows a Poisson distribution. The daily counts from the past week are shown below. Assume they form a random sample.

2 4 2 5 6 0 3

How can we estimate the following quantities using the data?

1. Mean number of coffees sold per day.
2. The probability of selling zero coffee in a day.

Example: Coffee shop

Suppose the daily count of coffees sold at a coffee shop follows a Poisson distribution. The daily counts from the past week are shown below. Assume they form a random sample.

2 4 2 5 6 0 3

How can we estimate the following quantities using the data?

1. Mean number of coffees sold per day.
2. The probability of selling zero coffee in a day.

Let $Y \sim \mathrm{Pois}(\lambda)$ represent the true daily coffee sales count.

Example: Coffee shop

Suppose the daily count of coffees sold at a coffee shop follows a Poisson distribution. The daily counts from the past week are shown below. Assume they form a random sample.

2 4 2 5 6 0 3

How can we estimate the following quantities using the data?

1. Mean number of coffees sold per day.
2. The probability of selling zero coffee in a day.

Evaluate whether the estimators are unbiased.

Let $Y \sim \text{Pois}(\lambda)$ represent the true daiily coffee sales count.

1. $\bar{y}_7$=3.143.
   - $\bar{Y}_7$ is an unbiased estimator.
2. Let $p_0 = P(Y = 0)$.
   - $\hat{p}_0 = 1/7$ using relative frequency. This is an unbiased estimator.
   - $\hat{p}_0' = e^{-3.143}$ using the probability mass function with $\widehat{\lambda} = \bar{y}_7$. This is a biased estimator.

```
1/7
```

```
[1] 0.1428571
```

```
exp(-3.143)
```

```
[1] 0.04315314
```

> See the example discussed in Section 19.1 from Dekking *et al.*

Summary

- An **estimator** for a parameter is a sample statistic devised to provide *estimates* of the parameter.
- While we can get *consistent* estimators based on the law of large numbers, we can study the **sampling distribution** to learn about the quality of an estimator
- **Unbiased** estimators have a mean value that equals to the parameter.

More on sampling distributions

Example: Stealing bases

```r
stealingbases <- stealingbases |>
  mutate(w = pitcher_delivery + catcher_tl
stealingbases
```

```
# A tibble: 166 × 2
   pitcher_delivery catcher_throw
              <dbl>         <dbl>
 1             1.38          1.82
 2             1.53          1.89
 3             1.15          1.89
 4             1.35          1.89
 5             1.54          1.9 
 6             1.22          1.91
 7             1.22          1.92
 8             1.34          1.92
 9             1.25          1.92
10             1.53          1.93
11             1.3           1.93
12             1.2           1.93
13             1.23          1.93
14             1.4           1.93
15             1.12          1.93
16             1.31          1.93
```

Question

- What is probability of getting to the base before the ball is caught at the second base if you can run in 3.5 seconds?

Model

- `pitcher_delivery` is a random sample of $X \sim F_1$.
- `catcher_throw` is a random sample of $Y \sim F_2$.
- Let $W = X + Y$ be the time from the start of pitching until the ball is caught at the second base.

Parameter of interest

- $p = P(W > 3.5)$

Estimator

- $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(W_i > 3.5)$

> We won't assume the specifics of the distributions.

Example: Sampling distribution of the proportion

- Let $I_i = \mathbb{I}(W_i > 3.5)$ for $i = 1, 2, \ldots n$.
- We already know that $\mathbb{E}(\hat{p}_n) = \mathbb{E}(\bar{I}_n) = \mathbb{E}(I) = P(W > 3.5) = p$.
- Can we fully specify the sampling distribution in terms of the parameter?

Example: Sampling distribution of the proportion

- Let $I_i = \mathbb{I}(W_i > 3.5)$ for $i = 1, 2, \ldots n$.
- We already know that $\mathbb{E}(\hat{p}_n) = \mathbb{E}(\bar{I}_n) = \mathbb{E}(I) = P(W > 3.5) = p$.
- Can we fully specify the sampling distribution in terms of the parameter?

$$P(I_1 = x) \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\implies I_1 \sim \mathrm{Ber}(p)$$

with

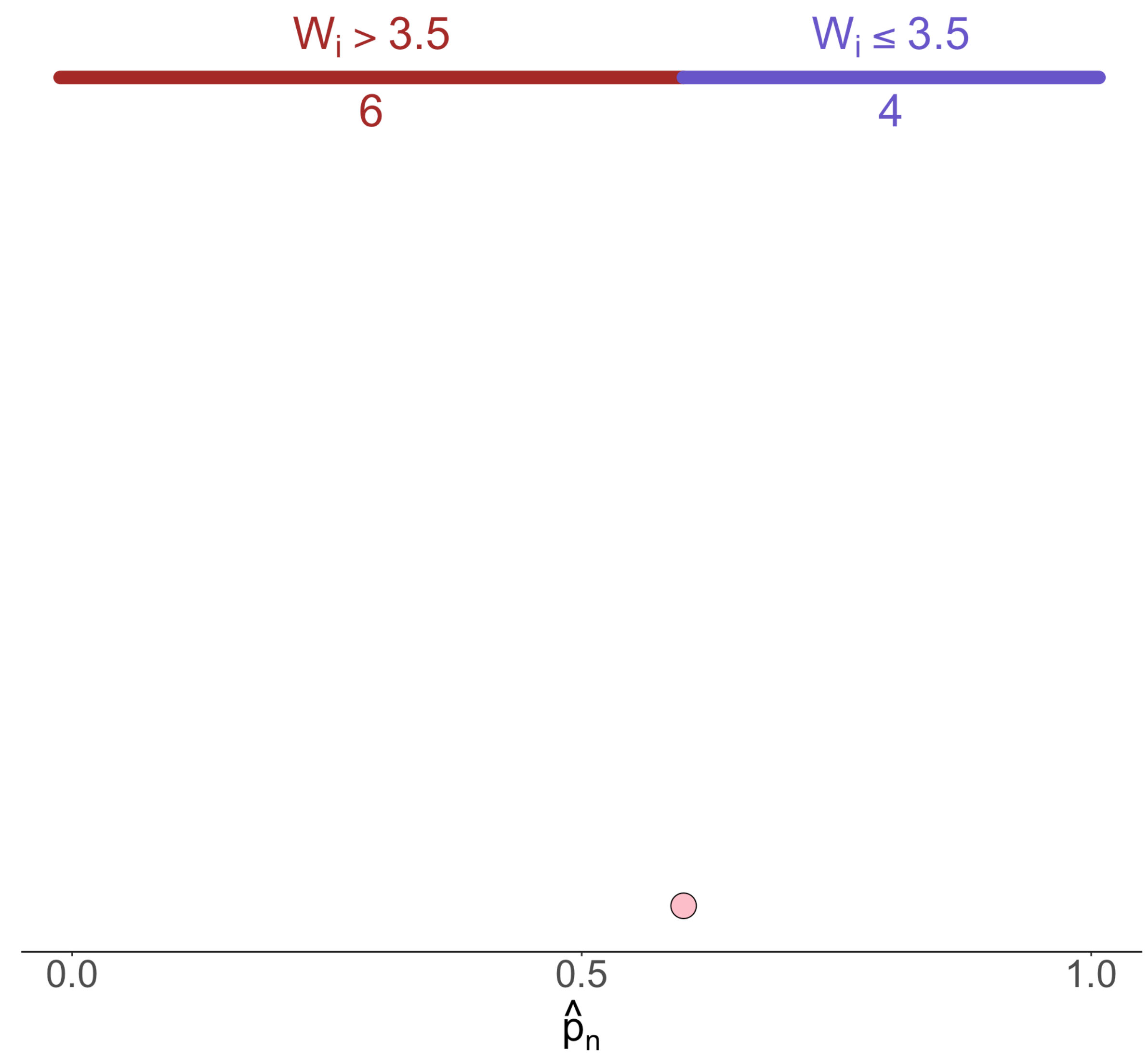
$$\mathbb{E}(I_1) = p \quad \text{and} \quad \mathrm{Var}(I_i) = p(1-p).$$

$$\implies \mathbb{E}(\hat{p}_n) = p \quad \text{and} \quad \mathrm{Var}(\hat{p}_n) = \frac{p(1-p)}{n}$$

How about the shape of the distribution?

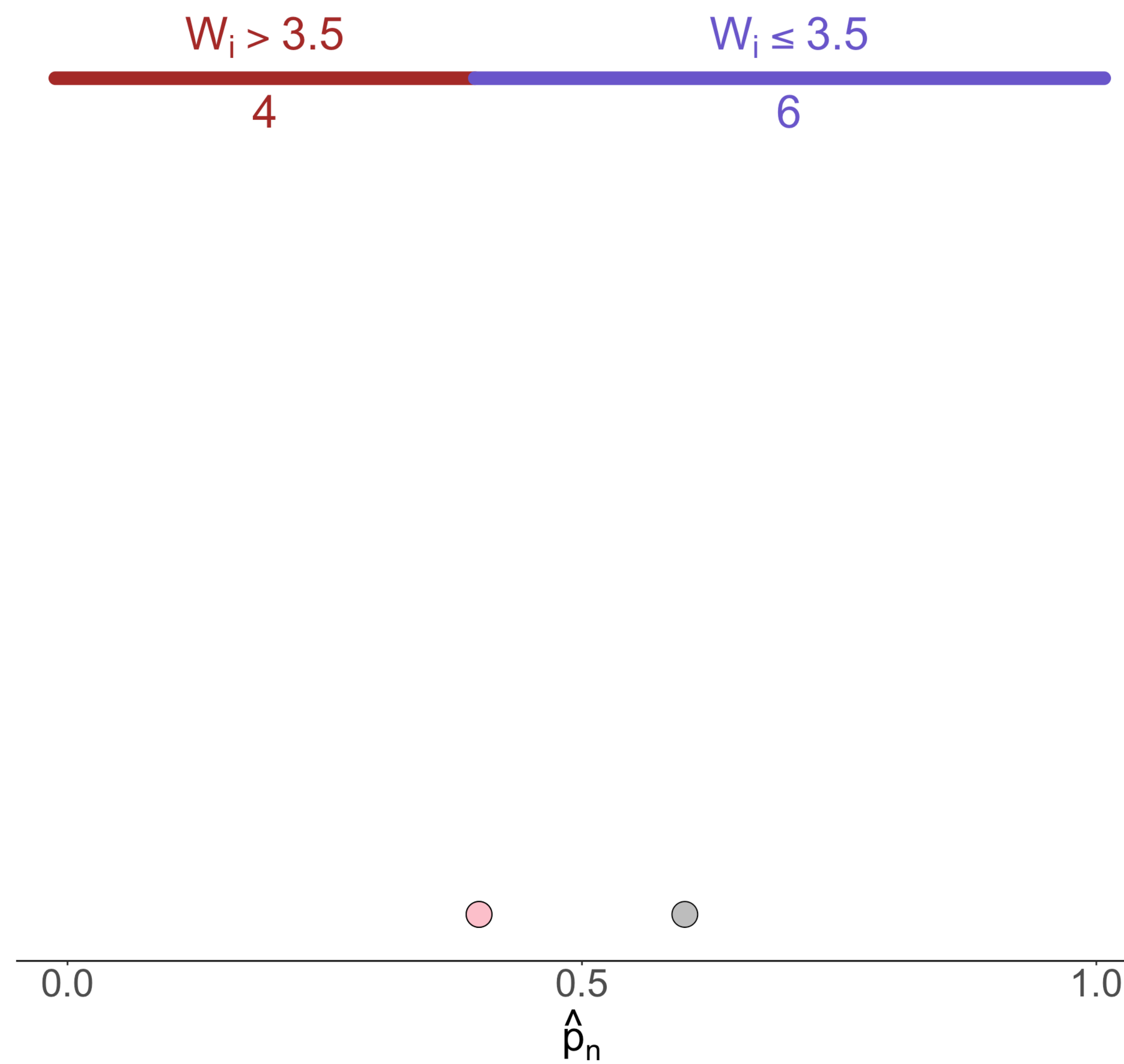
Example: Sampling distribution of the proportion

If we could repeatedly conduct experiments with $n = 10$...



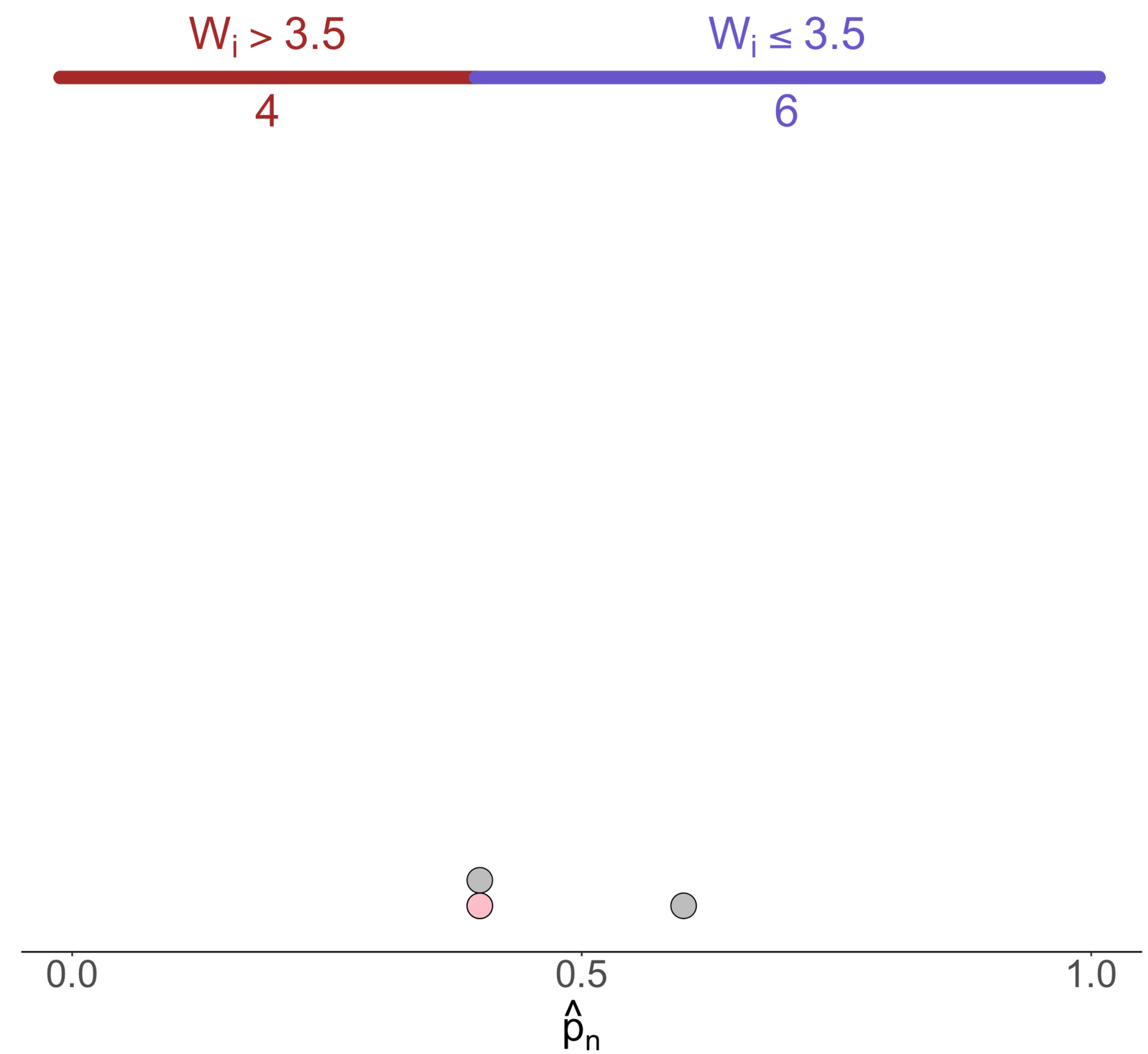
Example: Sampling distribution of the proportion

If we could repeatedly conduct experiments with $n = 10$...



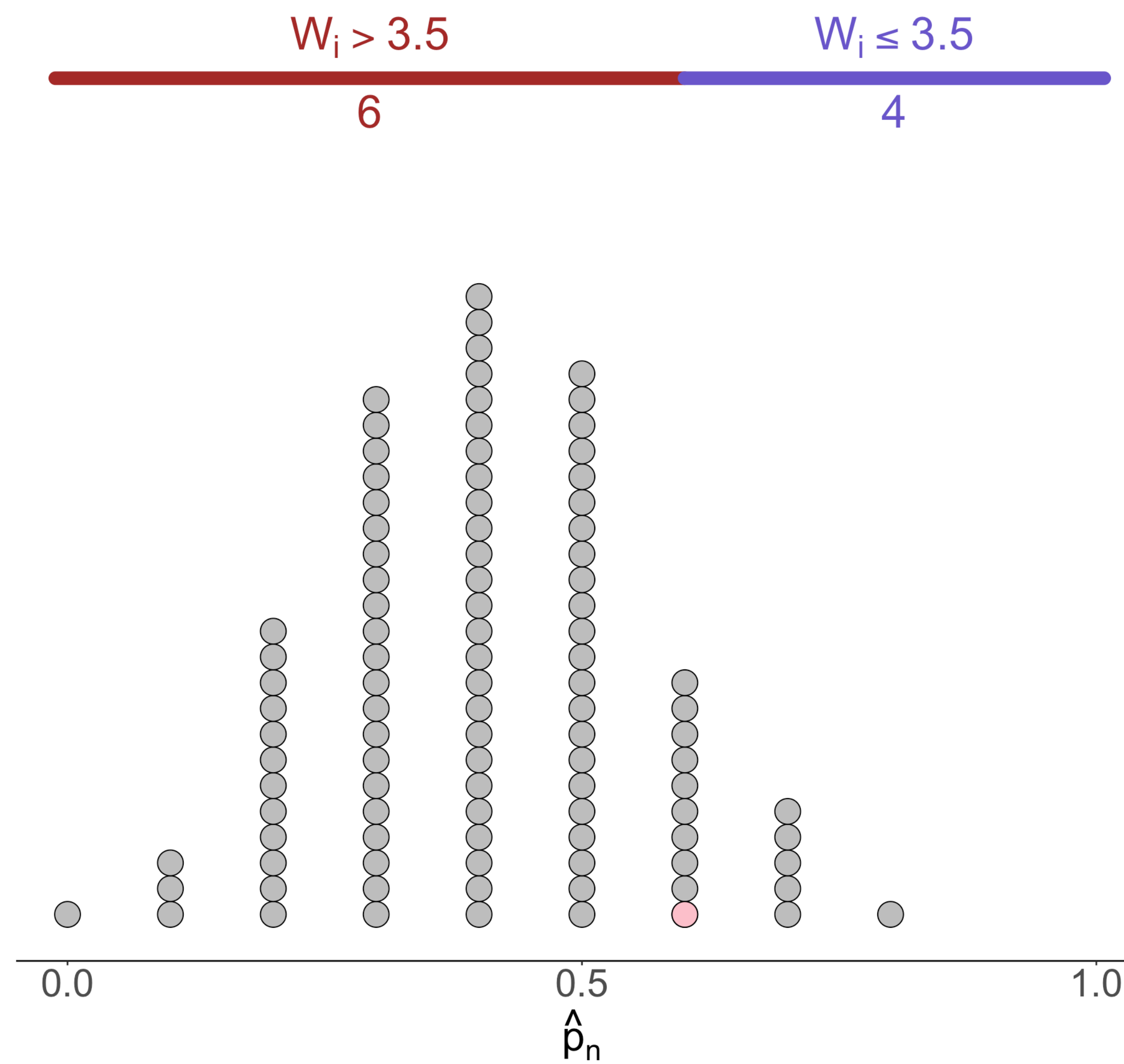
Example: Sampling distribution of the proportion

If we could repeatedly conduct experiments with $n = 10$...



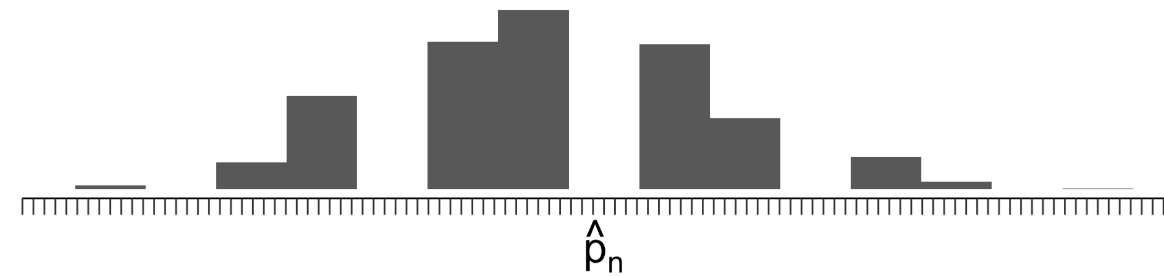
Example: Sampling distribution of the proportion

If we could repeatedly conduct experiments with $n = 10$...

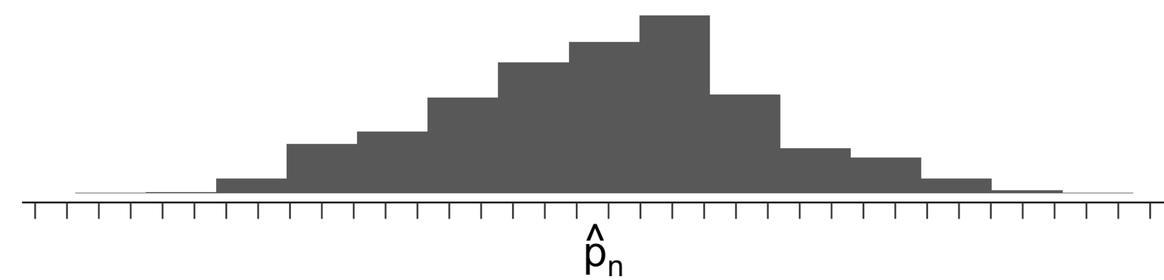


Example: Sampling distribution of the proportion

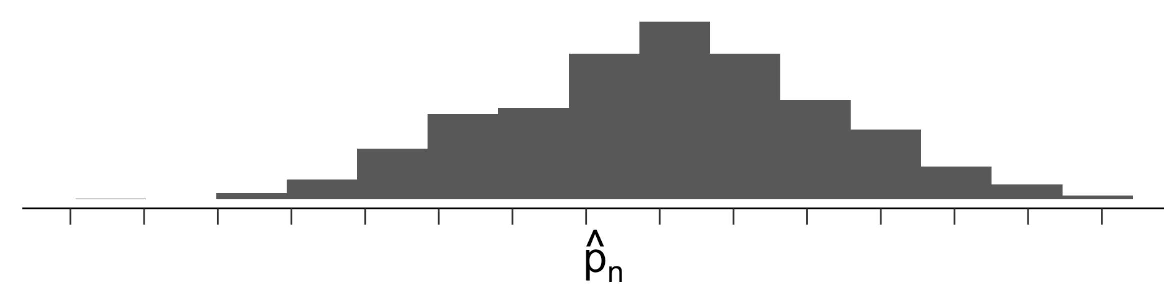
If we could repeatedly conduct experiments with $n = 10$...



If we could repeatedly conduct experiments with $n = 100$...



If we could repeatedly conduct experiments with $n = 500$...



Recall: The Central Limit Theorem

- We say, Z_n converges in distribution to $Z \sim \mathcal{N}(0,1)$. That is,

$$Z_n \xrightarrow{d} Z \quad \text{where} \quad Z \sim \mathcal{N}(0,1).$$

Let $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} F$ for any distribution F with a finite mean μ and a finite variance $\sigma^2 > 0$. Then

$$\lim_{n\to\infty} P(Z_n \le a) = \Phi(a)$$

where $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ and $\Phi(a)$ is the cumulative distribution function of the standard normal distribution.

Equivalently, we have

$$\lim_{n\to\infty} P(\bar{X}_n \le) = P(Y \le a)$$

where $Y \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Example: Result of CLT for $\hat{p}_n$

- $\hat{p}_n$ is a sample mean, $\bar{I}_n = \frac{1}{n}\sum_{i=1}^{n} I_i$.

- As you increase n, $\hat{p}_n$ converges in distribution to a normal random variable.

$$\hat{p}_n \xrightarrow{d} Y$$

where $Y \sim \mathcal{N}(\_\_\_\_\_\_, \_\_\_\_\_\_)$.

- Equivalently,

$$\_\_\_\_\_\_\_\_\_\_\_\_ \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Standardization of random variables

For any random variable X with mean μ and variance σ^2,

$$\frac{X - \mu}{\sigma}$$

results in a mean of _________ and variance of _________.

For any sample mean $\bar{X}_n$ from population mean μ and population variance σ^2,

$$\_\_\_\_\_\ \frac{\bar{X}_n - \mu}{\sigma}$$

results in a mean of _________ and variance of _________.

Sampling distribution of the sample proportion

- $n\hat{p}_n$ follows a binomial distribution.
- In practice, the normal approximation is often used; it's considered a reasonable approximation when np and $n(1-p)$ exceeds a certain threshold — often 5 or 10.

Consider an event A in the sample space of a random experiment with $p = P(A)$. Let Y be the number of times A occurs when the experiment is repeated n independent times.

$$\hat{p}_n = \frac{Y}{n}$$

is an estimator for p with

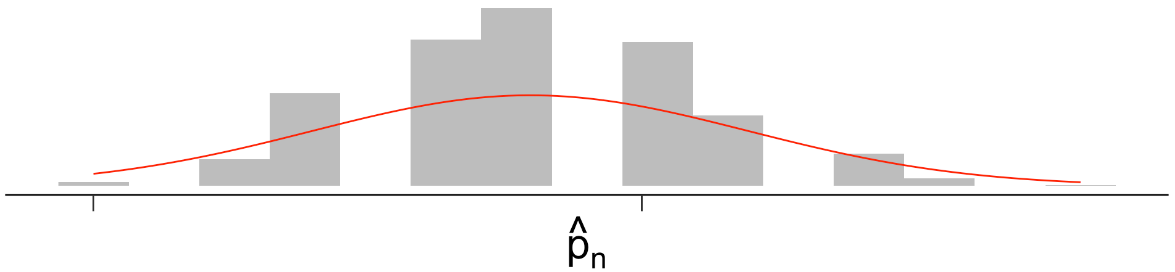
a. $\mathbb{E}(\hat{p}_n) = p$;

b. $\mathrm{Var}(\hat{p}_n) = \frac{p(1-p)}{n}$; and

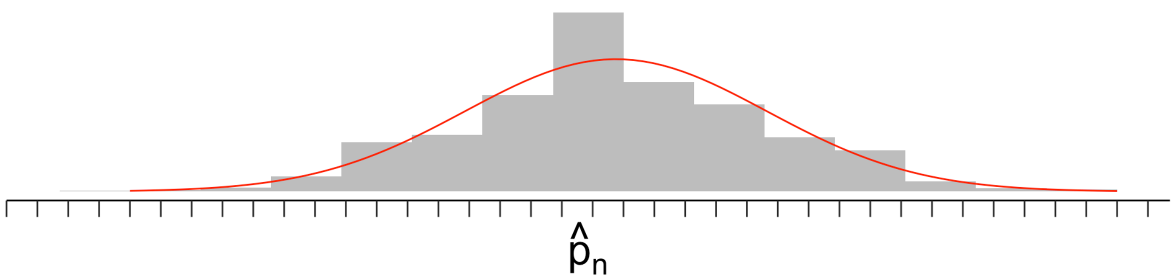
c. $\hat{p}_n \xrightarrow{d} Y$, where $Y \sim \mathcal{N}\left(p, p(1-p)/\,n\right)$.

Example: Normal approximation of $\hat{p}_n$

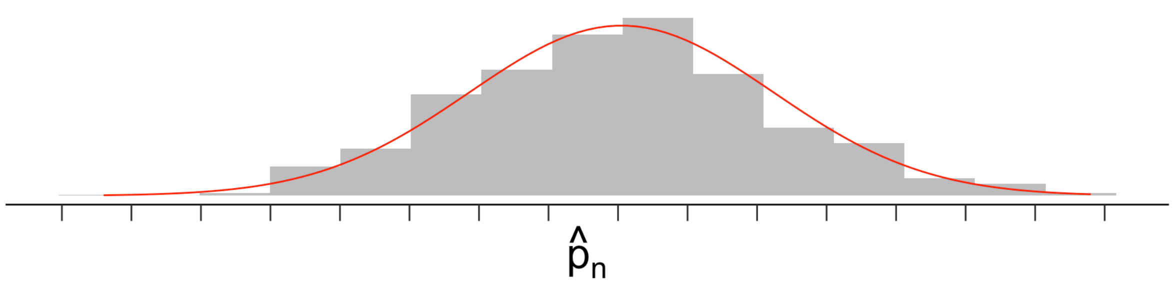
$n = 10$



$n = 100$



$n = 500$



Exercises

Suppose each customer at the coffee shop leaves a Google review with a probability of $p = 0.3$. Assuming the events of customer leaving Google reviews are independent events, what is the probability that more than 40 customers out of the next 100 will write Google reviews?

Exercises

Suppose each customer at the coffee shop leaves a Google review with a probability of $p = 0.3$. Assuming the events of customer leaving Google reviews are independent events, what is the probability that more than 40 customers out of the next 100 will write Google reviews?

```r
1 - pnorm(.4, mean = .3, sd = sqrt(.3 * .7 / 100))
```

```
[1] 0.01454817
```

```r
1 - pbinom(40, 100, .3)
```

```
[1] 0.01249841
```

Exercises

Suppose you simulate a random sample of size $n = 100$ for a complex experiment. You don't know the full distribution of these quantities but you know the population mean is 5 and the population variance is 2. What is the probability that the sample mean is within 0.1 distance from 5?

Exercises

Suppose you simulate a random sample of size $n = 100$ for a complex experiment. You don't know the full distribution of these quantities but you know the population mean is 5 and the population variance is 2. What is the probability that the sample mean is within 0.1 distance from 5?

```r
pnorm(5.1, mean = 5, sd = sqrt(2 / 100)) - pnorm(4.9, mean = 5, sd = sqrt(2 / 100))
```

```
[1] 0.5204999
```

Distributions related to the normal distribution

The $\chi^2(n)$ distribution

- "Chi-squared" distribution with n degrees of freedom.
- $\mathbb{E}(X) = n$ for $X \sim \chi^2(n)$.

The χ^2 distribution with n degrees of freedom is the distribution of the random variable

$$X = \sum_{i=1}^{n} Z_i^2$$

where Z_1, Z_2, ..., Z_n are independent standard normal random variables.

The $\chi^2(n)$ distribution

- It allows us to study the sampling distribution of the sample variance of a normal random distribution.
- We can show this by investigating the squared sum of standardized Y_i's.

Let $Y_1, Y_2, \ldots, Y_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then

$$\frac{(n-1)S_n^2}{\sigma^2}$$

where S_n^2 is the sample variance has the $\chi^2(n-1)$.

The $t(n)$ distribution

- As $n \to \infty$, the $t(n)$ distribution converges in distribution to the standard normal distribution.

The t distribution with n degrees of freedom is the distribution of the random variable

$$Y = \frac{Z}{\sqrt{X_n / n}}$$

where $Z \sim \mathcal{N}(0,1)$, $X_n \sim \chi^2(n)$, and Z and X_n are independent.

The $t(n)$ distribution

- The standardized sample mean of a normal distribution follows the t distribution.
- It doesn't follow the standard normal distribution because the sample variance is also a random variable.

Let $Y_1, Y_2, \ldots, Y_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then

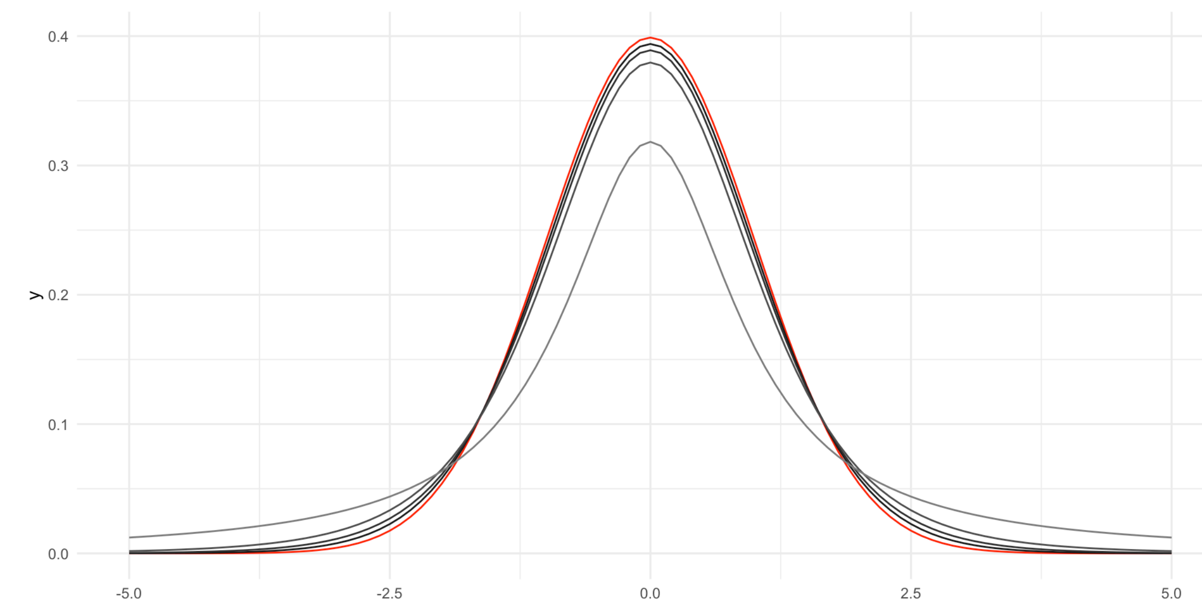
$$\frac{\bar{Y}_n - \mu}{\sqrt{S_n^2 / n}}$$

where $\bar{Y}_n$ is the sample mean and S_n^2 is the sample variance has the $t(n-1)$ distribution.

The $t(n)$ distribution

Probability density functions for the standard normal distribution and the t distributions with 1, 5, 10, and 20 degrees of freedom respectively.

```r
ggplot() +
  theme_minimal() +
  xlim(c(-5, 5)) +
  geom_function(fun = dnorm, colour = "red") +
  geom_function(fun = dt, args = list(df = 20),
  geom_function(fun = dt, args = list(df = 10),
  geom_function(fun = dt, args = list(df = 5),
  geom_function(fun = dt, args = list(df = 1),
```



$F(m, n)$ distribution

- This is useful when you want to compare variance of two random variables.

The F distribution with m and n degrees of freedom is the distribution of the random variable

$$W = \frac{X_m/m}{Y_n/n}$$

where $X_m \sim \chi^2(m)$, $Y_n \sim \chi^2(n)$, and X_m and Y_n are independent.

Summary

- Binary data can be modelled as a realization of a binomial random variable or a random sample of a Bernoulli random variable.
- With a large sample size, sampling distributions can often be estimated using a normal distribution based on the central limit theorem.
- Known sampling distributions related to the random sample of a normal distribution makes it convenient to work with models based on normal distributions.