Data Science Inference and Modeling

The textbook for the Data Science course series is freely available online.

This course corresponds to the textbook chapters Statistical Inference and Statistical Models.

Learning Objectives

- The concepts necessary to define estimates and margins of errors of populations, parameters, estimates, and standard errors in order to make predictions about data
- How to use models to aggregate data from different sources
- The very basics of Bayesian statistics and predictive modeling

Course Overview

Section 1: Parameters and Estimates

You will learn how to estimate population parameters.

Section 2: The Central Limit Theorem in Practice

You will apply the central limit theorem to assess how close a sample estimate is to the population parameter of interest.

Section 3: Confidence Intervals and p-Values

You will learn how to calculate confidence intervals and learn about the relationship between confidence intervals and p-values.

Section 4: Statistical Models

You will learn about statistical models in the context of election forecasting.

Section 5: Bayesian Statistics

You will learn about Bayesian statistics through looking at examples from rare disease diagnosis and baseball.

Section 6: Election Forecasting

You will learn about election forecasting, building on what you've learned in the previous sections about statistical modeling and Bayesian statistics.

Section 7: Association Tests

You will learn how to use association and chi-squared tests to perform inference for binary, categorical, and ordinal data through an example looking at research funding rates.

Introduction to Inference

The textbook for this section is available here

In this course, we will learn:

- *statistical inference*, the process of deducing characteristics of a population using data from a random sample
- the statistical concepts necessary to define *estimates* and *margins of errors*
- how to *forecast future results* and estimate the precision of our forecast
- how to calculate and interpret *confidence intervals and p-values*

Key points

- Information gathered from a small random sample can be used to infer characteristics of the entire population.
- Opinion polls are useful when asking everyone in the population is impossible.
- A common use for opinion polls is determining voter preferences in political elections for the purposes of forecasting election results.
- The *spread* of a poll is the estimated difference between support two candidates or options.

Section 1 Overview

Section 1 introduces you to parameters and estimates.

After completing Section 1, you will be able to:

- Understand how to use a sampling model to perform a poll.
- Explain the terms **population**, **parameter**, and **sample** as they relate to statistical inference.
- Use a sample to estimate the population proportion from the sample average.
- Calculate the expected value and standard error of the sample average.

Sampling Model Parameters and Estimates

The textbook for this section is available here and here; first part

Key points

- The task of statistical inference is to estimate an unknown population parameter using observed data from a sample.
- In a sampling model, the collection of elements in the urn is called the *population*.
- A *parameter* is a number that summarizes data for an entire population.
- A *sample* is observed data from a subset of the population.
- An *estimate* is a summary of the observed data about a parameter that we believe is informative. It is a data-driven guess of the population parameter.
- We want to predict the proportion of the blue beads in the urn, the parameter p . The proportion of red beads in the urn is $1 - p$ and the *spread* is $2p - 1$.

- The sample proportion is a random variable. Sampling gives random results drawn from the population distribution.

Code: Function for taking a random draw from a specific urn

The **dslabs** package includes a function for taking a random draw of size n from the urn:

```
if(!require(tidyverse)) install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages -----------------------------------------------------------------------------
```

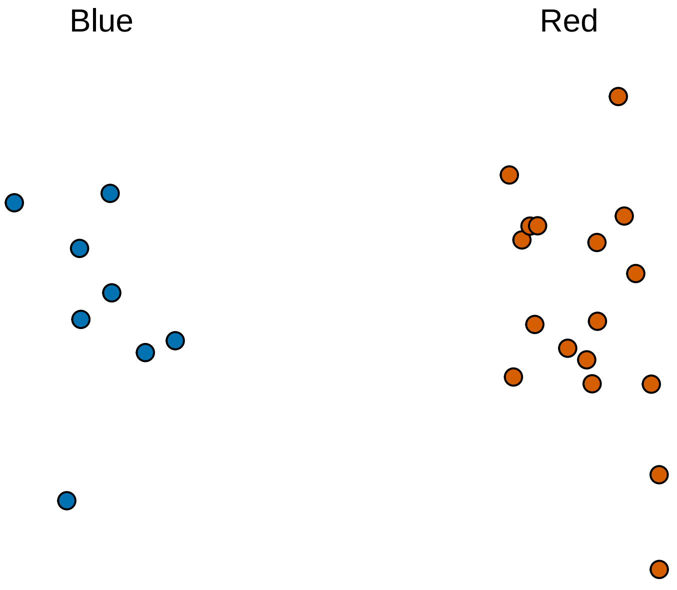
```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require(dslabs)) install.packages("dslabs")
```

```
## Loading required package: dslabs
```

```
library(tidyverse)
library(dslabs)
take_poll(25)    # draw 25 beads
```



The Sample Average

The textbook for this section is available here and here

Key points

- Many common data science tasks can be framed as estimating a parameter from a sample.
- We illustrate statistical inference by walking through the process to estimate p. From the estimate of p, we can easily calculate an estimate of the spread, $2p - 1$.
- Consider the random variable X that is 1 if a blue bead is chosen and 0 if a red bead is chosen. The proportion of blue beads in N draws is the average of the draws $X_1, ..., X_N$.
- $\overline{X}$ is the *sample average*. In statistics, a bar on top of a symbol denotes the average. $\overline{X}$ is a random variable because it is the average of random draws - each time we take a sample, $\overline{X}$ is different.

$\overline{X} = \frac{X_1+X_2+...+X_N}{N}$

- The number of blue beads drawn in N draws, $N\overline{X}$, is N times the proportion of values in the urn. However, we do not know the true proportion: we are trying to estimate this parameter p.

Polling versus Forecasting

The textbook for this section is available here

Key points

- A poll taken in advance of an election estimates p for that moment, not for election day.
- In order to predict election results, forecasters try to use early estimates of p to predict p on election day. We discuss some approaches in later sections.

Properties of Our Estimate

The textbook for this section is available here

Key points

- When interpreting values of $\overline{X}$, it is important to remember that $\overline{X}$ is a random variable with an expected value and standard error that represents the sample proportion of positive events.
- The expected value of $\overline{X}$ is the parameter of interest p. This follows from the fact that $\overline{X}$ is the sum of independent draws of a random variable times a constant $1/N$.

$E(\overline{X}) = p$

- As the number of draws N increases, the standard error of our estimate $\overline{X}$ decreases. The standard error of the average of $\overline{X}$ over N draws is:

$SE(\overline{X}) = \sqrt{p(1-p)/N}$

- In theory, we can get more accurate estimates of p by increasing N. In practice, there are limits on the size of N due to costs, as well as other factors we discuss later.
- We can also use other random variable equations to determine the expected value of the sum of draws $E(S)$ and standard error of the sum of draws $SE(S)$.

$E(S) = Np$

$SE(S) = \sqrt{Np(1-p)}$

Assessment - Parameters and Estimates

1. Suppose you poll a population in which a proportion p of voters are Democrats and $1-p$ are Republicans.

Your sample size is $N = 25$. Consider the random variable S, which is the **total** number of Democrats in your sample.

What is the expected value of this random variable S?

☐ A. $E(S) = 25(1-p)$
☒ B. $E(S) = 25p$
☐ C. $E(S) = \sqrt{25p(1-p)}$
☐ D. $E(S) = p$

2. Again, consider the random variable S, which is the **total** number of Democrats in your sample of 25 voters.

The variable p describes the proportion of Democrats in the sample, whereas $1-p$ describes the proportion of Republicans.

What is the standard error of S?

☐ A. $SE(S) = 25p(1-p)$
☐ B. $SE(S) = \sqrt{25p}$
☐ C. $SE(S) = 25(1-p)$
☒ D. $SE(S) = \sqrt{25p(1-p)}$

3. Consider the random variable S/N, which is equivalent to the sample average that we have been denoting as $\overline{X}$.

The variable N represents the sample size and p is the proportion of Democrats in the population.

What is the expected value of $\overline{X}$?

☒ A. $E(\overline{X}) = p$
☐ B. $E(\overline{X}) = Np$
☐ C. $E(\overline{X}) = N(1-p)$
☐ D. $E(\overline{X}) = 1-p$

4. What is the standard error of the sample average, $\overline{X}$?

The variable N represents the sample size and p is the proportion of Democrats in the population.

☐ A. $SE(\overline{X}) = \sqrt{Np(1-p)}$
☒ B. $SE(\overline{X}) = \sqrt{p(1-p)/N}$
☐ C. $SE(\overline{X}) = \sqrt{p(1-p)}$
☐ D. $SE(\overline{X}) = \sqrt{N}$

5. Write a line of code that calculates the standard error `se` of a sample average when you poll 25 people in the population.

Generate a sequence of 100 proportions of Democrats `p` that vary from 0 (no Democrats) to 1 (all Democrats).

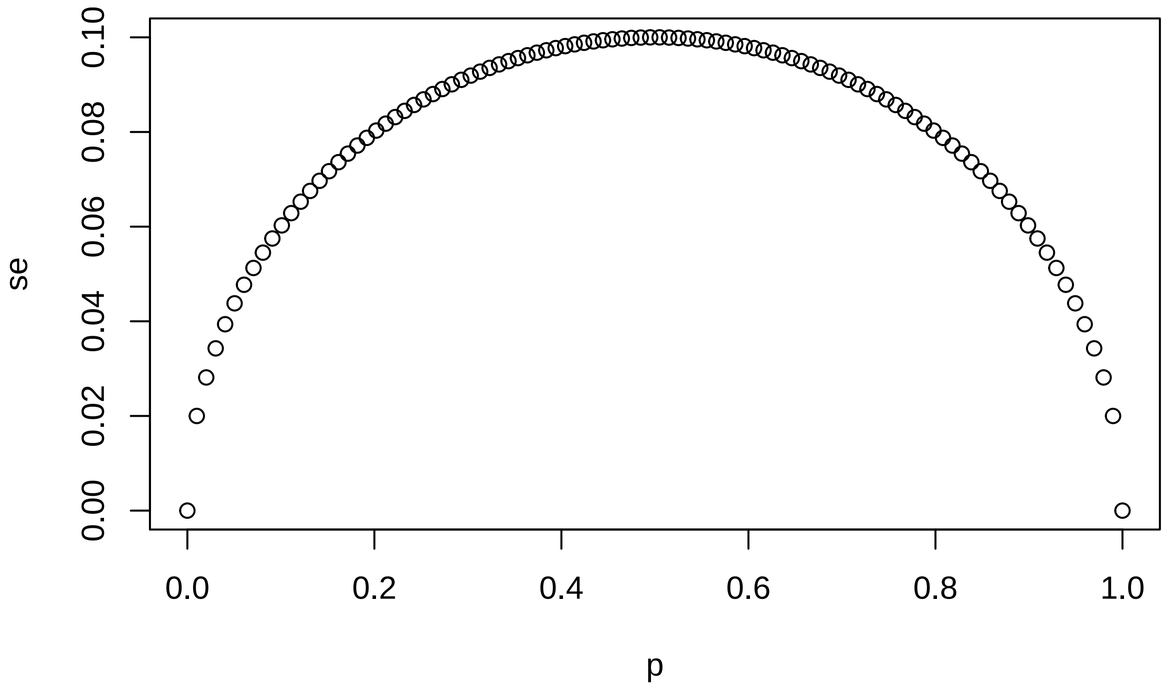
Plot `se` versus `p` for the 100 different proportions.

```r
# `N` represents the number of people polled
N <- 25

# Create a variable `p` that contains 100 proportions ranging from 0 to 1 using the `seq` function
p <- seq(0, 1, length.out = 100)

# Create a variable `se` that contains the standard error of each sample average
se <- sqrt(p * (1 - p)/N)

# Plot `p` on the x-axis and `se` on the y-axis
plot(p,se)
```

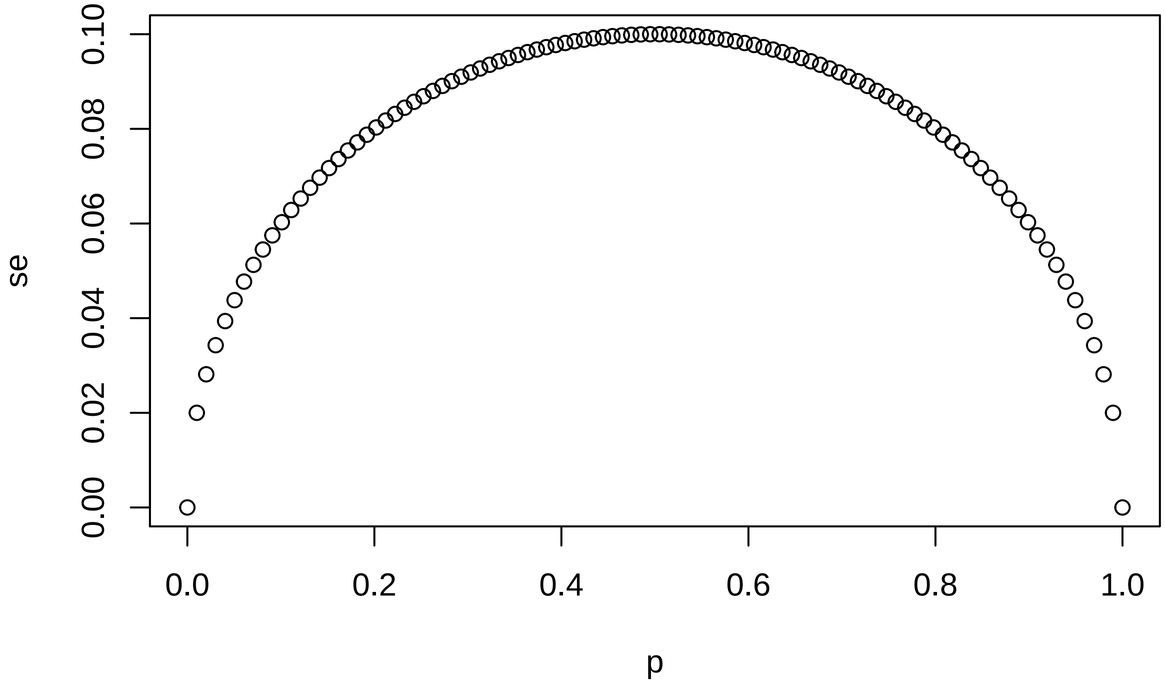


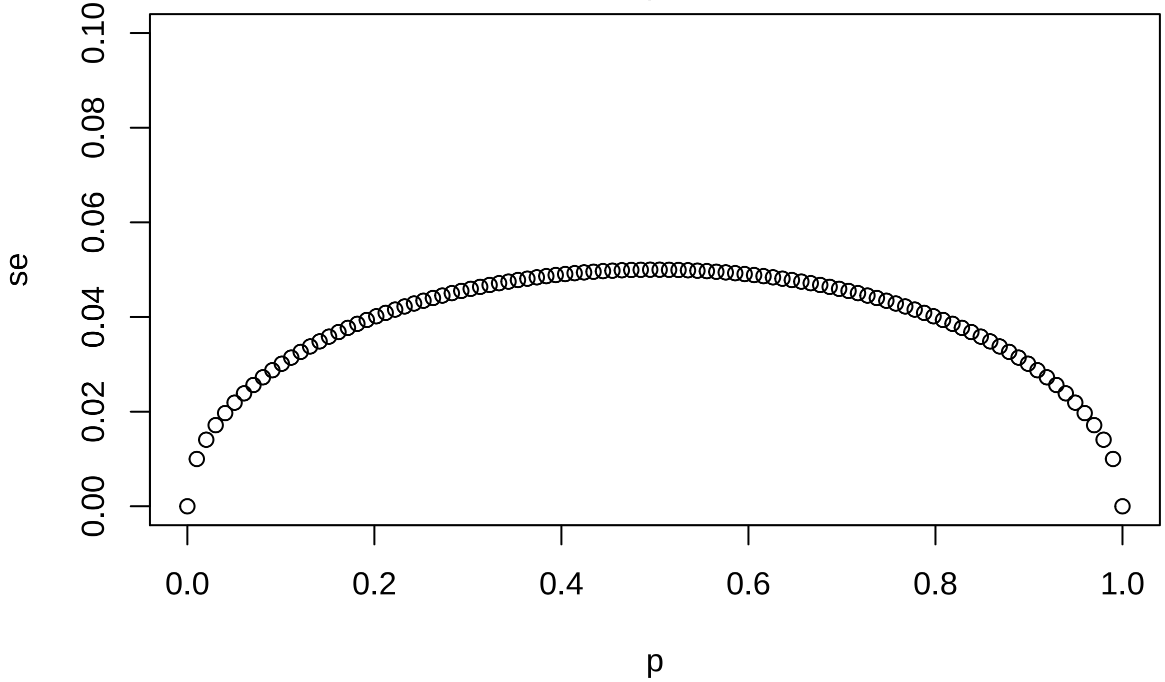
6. Using the same code as in the previous exercise, create a for-loop that generates three plots of `p` versus `se` when the sample sizes equal $N = 25$, $N = 100$, and $N = 1000$.

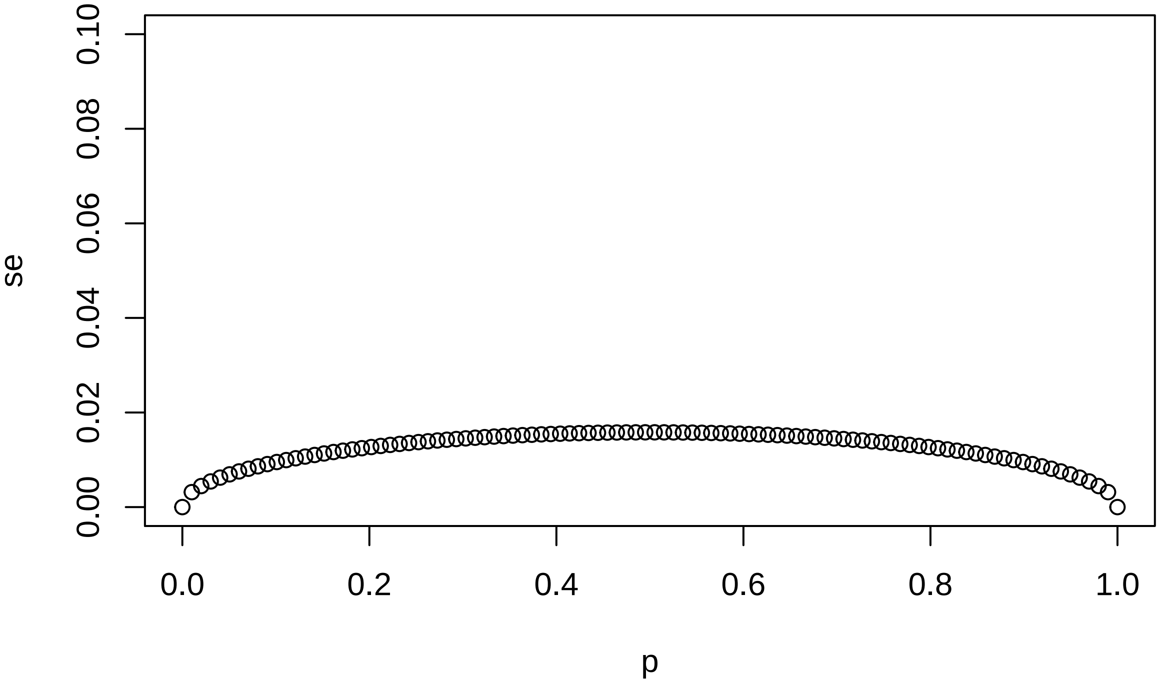
```r
# The vector `p` contains 100 proportions of Democrats ranging from 0 to 1 using the `seq` function
p <- seq(0, 1, length = 100)

# The vector `sample_sizes` contains the three sample sizes
sample_sizes <- c(25, 100, 1000)

# Write a for-loop that calculates the standard error `se` for every value of `p` for each of the three
for (N in sample_sizes)
{
se <- sqrt(p * (1 - p)/N)
plot(p,se,ylim = c(0,0.1))
}
```


se
0.00
0.02
0.04
0.06
0.08
0.10
0.0
0.2
0.4
0.6
0.8
1.0
p


se
0.00
0.02
0.04
0.06
0.08
0.10
0.0
0.2
0.4
0.6
0.8
1.0
p



7. Our estimate for the difference in proportions of Democrats and Republicans is $d = \overline{X} - (1 - \overline{X})$.

Which derivation correctly uses the rules we learned about sums of random variables and scaled random variables to derive the expected value of d

☐ A. $E\left[\overline{X} - (1 - \overline{X})\right] = E\left[2\overline{X} - 1\right] = 2E\left[\overline{X}\right] - 1 = N(2p - 1) = Np - N(1 - p)$
☐ B. $E\left[\overline{X} - (1 - \overline{X})\right] = E\left[\overline{X} - 1\right] = E\left[\overline{X}\right] - 1 = p - 1$
☐ C. $E\left[\overline{X} - (1 - \overline{X})\right] = E\left[2\overline{X} - 1\right] = 2E\left[\overline{X}\right] - 1 = 2\sqrt{p(1 - p)} - 1 = p - (1 - p)$
☒ D. $E\left[\overline{X} - (1 - \overline{X})\right] = E\left[2\overline{X} - 1\right] = 2E\left[\overline{X}\right] - 1 = 2p - 1 = p - (1 - p)$

8. Our estimate for the difference in proportions of Democrats and Republicans is $d = \overline{X} - (1 - \overline{X})$.

Which derivation correctly uses the rules we learned about sums of random variables and scaled random variables to derive the standard error of d?

☐ A. $SE\left[\overline{X} - (1 - \overline{X})\right] = SE\left[2\overline{X} - 1\right] = 2SE\left[\overline{X}\right] = 2\sqrt{p/N}$
☐ B. $SE\left[\overline{X} - (1 - \overline{X})\right] = SE\left[2\overline{X} - 1\right] = 2SE\left[\overline{X} - 1\right] = 2\sqrt{p(1 - p)/N} - 1$
☒ C. $SE\left[\overline{X} - (1 - \overline{X})\right] = SE\left[2\overline{X} - 1\right] = 2SE\left[\overline{X}\right] = 2\sqrt{p(1 - p)/N}$
☐ D. $SE\left[\overline{X} - (1 - \overline{X})\right] = SE\left[\overline{X} - 1\right] = SE\left[\overline{X}\right] = \sqrt{p(1 - p)/N}$

9. Say the actual proportion of Democratic voters is $p = 0.45$.

In this case, the Republican party is winning by a relatively large margin of $d = -0.1$, or a 10% margin of victory. What is the standard error of the spread $2\overline{X} - 1$ in this case?

```
# `N` represents the number of people polled
N <- 25
```

```
# `p` represents the proportion of Democratic voters
p <- 0.45

# Calculate the standard error of the spread. Print this value to the console.
2*sqrt((p*(1-p)/N))
```

```
## [1] 0.1989975
```

10. So far we have said that the difference between the proportion of Democratic voters and Republican voters is about 10% and that the standard error of this spread is about 0.2 when $N = 25$.

Select the statement that explains why this sample size is sufficient or not.

- ☐ A. This sample size is sufficient because the expected value of our estimate $2\overline{X} - 1$ is d so our prediction will be right on.
- ☒ B. This sample size is too small because the standard error is larger than the spread.
- ☐ C. This sample size is sufficient because the standard error of about 0.2 is much smaller than the spread of 10%.
- ☐ D. Without knowing `p`, we have no way of knowing that increasing our sample size would actually improve our standard error.

Section 2 Overview

In Section 2, you will look at the Central Limit Theorem in practice.

After completing Section 2, you will be able to: - Use the Central Limit Theorem to calculate the probability that a sample estimate X̄ is close to the population proportion p. - Run a Monte Carlo simulation to corroborate theoretical results built using probability theory. - Estimate the spread based on estimates of X̄ and ŜE(X̄). - Understand why bias can mean that larger sample sizes aren't necessarily better.

The textbook for this section is available here

Assessment 2.1: Introduction to Inference

1. Sample average

Write a function called take_sample that takes the proportion of Democrats p and the sample size N as arguments and returns the sample average of Democrats (1) and Republicans (0).

Calculate the sample average if the proportion of Democrats equals 0.45 and the sample size is 100.

Instructions - Define a function called take_sample that takes p and N as arguments. - Use the sample function as the first statement in your function to sample N elements from a vector of options where Democrats are assigned the value '1' and Republicans are assigned the value '0'. - Use the mean function as the second statement in your function to find the average value of the random sample.

```
# Write a function called `take_sample` that takes `p` and `N` as arguements and returns the average va
take_sample <- function(p, N){
    X <- sample(c(0,1), size = N, replace = TRUE, prob = c(1 - p, p))
    mean(X)
}
```

```
# Use the `set.seed` function to make sure your answer matches the expected result after random samplin
set.seed(1)

# Define `p` as the proportion of Democrats in the population being polled
p <- 0.45

# Define `N` as the number of people polled
N <- 100

# Call the `take_sample` function to determine the sample average of `N` randomly selected people from
take_sample(p,N)

## [1] 0.46
```

2. Distribution of errors - 1

Assume the proportion of Democrats in the population p equals 0.45 and that your sample size N is 100 polled voters. The take_sample function you defined previously generates our estimate, $\bar{X}$.

Replicate the random sampling 10,000 times and calculate p—$\bar{X}$ for each random sample. Save these differences as a vector called errors. Find the average of errors and plot a histogram of the distribution.

Instructions - The function take_sample that you defined in the previous exercise has already been run for you. - Use the replicate function to replicate subtracting the result of take_sample from the value of p 10,000 times. - Use the mean function to calculate the average of the differences between the sample average and actual value of p.

```
# Define `p` as the proportion of Democrats in the population being polled
p <- 0.45

# Define `N` as the number of people polled
N <- 100

# The variable `B` specifies the number of times we want the sample to be replicated
B <- 10000

# Use the `set.seed` function to make sure your answer matches the expected result after random samplin
set.seed(1)

# Create an objected called `errors` that replicates subtracting the result of the `take_sample` functi
errors <- replicate(B, p - take_sample(p, N))

# Calculate the mean of the errors. Print this value to the console.
mean(errors)

## [1] -4.9e-05
```

3. Distribution of errors - 2

In the last exercise, you made a vector of differences between the actual value for p and an estimate, $\bar{X}$. We called these differences between the actual and estimated values errors.

The errors object has already been loaded for you. Use the hist function to plot a histogram of the values contained in the vector errors. Which statement best describes the distribution of the errors?

Possible Answers - [ ] A. The errors are all about 0.05. - [ ] B. The error are all about -0.05. - [X] C. The errors are symmetrically distributed around 0. - [ ] D. The errors range from -1 to 1.

4. Average size of error

The error $p-\bar{X}$ is a random variable. In practice, the error is not observed because we do not know the actual proportion of Democratic voters, p. However, we can describe the size of the error by constructing a simulation.

What is the average size of the error if we define the size by taking the absolute value $p-\bar{X}$?

Instructions - Use the sample code to generate errors, a vector of $p-\bar{X}$. - Calculate the absolute value of errors using the abs function. - Calculate the average of these values using the mean function.

```
# Define `p` as the proportion of Democrats in the population being polled
p <- 0.45

# Define `N` as the number of people polled
N <- 100

# The variable `B` specifies the number of times we want the sample to be replicated
B <- 10000

# Use the `set.seed` function to make sure your answer matches the expected result after random sampling
set.seed(1)

# We generated `errors` by subtracting the estimate from the actual proportion of Democratic voters
errors <- replicate(B, p - take_sample(p, N))

# Calculate the mean of the absolute value of each simulated error. Print this value to the console.
mean(abs(errors))
```

```
## [1] 0.039267
```

5. Standard deviation of the spread

The standard error is related to the typical size of the error we make when predicting. We say size because, as we just saw, the errors are centered around 0. In that sense, the typical error is 0. For mathematical reasons related to the central limit theorem, we actually use the standard deviation of errors rather than the average of the absolute values.

As we have discussed, the standard error is the square root of the average squared distance $(\bar{X}-p)2$. The standard deviation is defined as the square root of the distance squared.

Calculate the standard deviation of the spread.

Instructions - Use the sample code to generate errors, a vector of $p-\bar{X}$. - Use ˆ2 to square the distances. - Calculate the average squared distance using the mean function. - Calculate the square root of these values using the sqrt function.

```
# Define `p` as the proportion of Democrats in the population being polled
p <- 0.45

# Define `N` as the number of people polled
N <- 100

# The variable `B` specifies the number of times we want the sample to be replicated
B <- 10000
```

```
# Use the `set.seed` function to make sure your answer matches the expected result after random sampling
set.seed(1)

# We generated `errors` by subtracting the estimate from the actual proportion of Democratic voters
errors <- replicate(B, p - take_sample(p, N))

# Calculate the standard deviation of `errors`
sqrt(mean(errors^2))
```

```
## [1] 0.04949939
```

6. Estimating the standard error

The theory we just learned tells us what this standard deviation is going to be because it is the standard error of $\bar{X}$.

Estimate the standard error given an expected value of 0.45 and a sample size of 100.

Instructions - Calculate the standard error using the sqrt function

```
# Define `p` as the expected value equal to 0.45
p <- 0.45

# Define `N` as the sample size
N <- 100

# Calculate the standard error
sqrt(p*(1-p)/N)
```

```
## [1] 0.04974937
```

7. Standard error of the estimate

In practice, we don't know p, so we construct an estimate of the theoretical prediction based by plugging in $\bar{X}$ for p.

Calculate the standard error of the estimate:

$\mathbf{\hat{SE}(\bar{X})}$

Instructions - Simulate a poll X using the sample function. - When using the sample function, create a vector using c() that contains all possible polling options where '1' indicates a Democratic voter and '0' indicates a Republican voter. - When using the sample function, use replace = TRUE within the sample function to indicate that sampling from the vector should occur with replacement. - When using the sample function, use prob = within the sample function to indicate the probabilities of selecting either element (0 or 1) within the vector of possibilities. - Use the mean function to calculate the average of the simulated poll, X_bar. - Calculate the standard error of the X_bar using the sqrt function and print the result.

```
# Define `p` as a proportion of Democratic voters to simulate
p <- 0.45

# Define `N` as the sample size
N <- 100
```

```
# Use the `set.seed` function to make sure your answer matches the expected result after random samplin
set.seed(1)

# Define `X` as a random sample of `N` voters with a probability of picking a Democrat ('1') equal to `
X <- sample(0:1, N, replace=T, p=c(1-p,p))

# Define `X_bar` as the average sampled proportion
X_bar <- mean(X)

# Calculate the standard error of the estimate. Print the result to the console.
sqrt(X_bar*(1-X_bar)/N)
```

```
## [1] 0.04983974
```

8. Plotting the standard error

The standard error estimates obtained from the Monte Carlo simulation, the theoretical prediction, and the estimate of the theoretical prediction are all very close, which tells us that the theory is working. This gives us a practical approach to knowing the typical error we will make if we predict p with X̄. The theoretical result gives us an idea of how large a sample size is required to obtain the precision we need. Earlier we learned that the largest standard errors occur for p=0.5.

Create a plot of the largest standard error for N ranging from 100 to 5,000. Based on this plot, how large does the sample size have to be to have a standard error of about 1%?

```
N <- seq(100, 5000, len = 100)
p <- 0.5
se <- sqrt(p*(1-p)/N)
```

Possible Answers - [ ] A. 100 - [ ] B. 500 - [X] C. 2,500 - [ ] D. 4,000

9. Distribution of X-hat

For N=100, the central limit theorem tells us that the distribution of X̄ is…

Possible Answers - [ ] A. practically equal to p. - [X] B. approximately normal with expected value p and standard error √(p(1−p)/N). - [ ] C. approximately normal with expected value X̄ and standard error √(X̄(1−X̄)/N). - [ ] D. not a random variable.

10. Distribution of the errors

We calculated a vector errors that contained, for each simulated sample, the difference between the actual value p and our estimate X̄.

The errors X̄−p are:

Possible Answers - [ ] A. practically equal to 0. - [X] B. approximately normal with expected value 0 and standard error √(p(1−p)/N). - [ ] C. approximately normal with expected value p and standard error √(p(1−p)/N). - [ ] D. not a random variable.

11. Plotting the errors