Brexit poll analysis - Part 1

Directions

There are 12 multi-part problems in this comprehensive assessment that review concepts from the entire course. The problems are split over 3 pages. Make sure you read the instructions carefully and run all pre-exercise code.

For numeric entry problems, you have 10 attempts to input the correct answer. For true/false problems, you have 2 attempts.

If you have questions, visit the "Brexit poll analysis" discussion forum that follows the assessment.

IMPORTANT: Some of these exercises use **dslabs** datasets that were added in a July 2019 update. Make sure your package is up to date with the command `update.packages("dslabs")`. You can also update all packages on your system by running `update.packages()` with no arguments, and you should consider doing this routinely.

Overview

In June 2016, the United Kingdom (UK) held a referendum to determine whether the country would "Remain" in the European Union (EU) or "Leave" the EU. This referendum is commonly known as Brexit. Although the media and others interpreted poll results as forecasting "Remain" ($p > 0.5$), the actual proportion that voted "Remain" was only 48.1% ($p = 0.481$) and the UK thus voted to leave the EU. Pollsters in the UK were criticized for overestimating support for "Remain".

In this project, you will analyze real Brexit polling data to develop polling models to forecast Brexit results. You will write your own code in R and enter the answers on the edX platform.

Important definitions

Data Import

Import the `brexit_polls` polling data from the **dslabs** package and set options for the analysis:

```
# suggested libraries and options
library(tidyverse)
options(digits = 3)
```

```
# load brexit_polls object
library(dslabs)
data(brexit_polls)
```

Final Brexit parameters

Define $p = 0.481$ as the actual percent voting "Remain" on the Brexit referendum and $d = 2p - 1 = -0.038$ as the actual spread of the Brexit referendum with "Remain" defined as the positive outcome:

```
p <- 0.481    # official proportion voting "Remain"
d <- 2*p-1    # official spread
```

Question 1: Expected value and standard error of a poll

6.0/6.0 points (graded)

The final proportion of voters choosing "Remain" was $p = 0.481$. Consider a poll with a sample of $N = 1500$ voters.

What is the expected total number of voters in the sample choosing "Remain"?

722 ✔ **Answer:** 722

722

Explanation

You can calculate the expected number of "Remain" voters with the following code:

```
p <- 0.481
N <- 1500
N*p
```

What is the standard error of the total number of voters in the sample choosing "Remain"?

19.4 ✔ **Answer:** 19.4

19.4

Explanation

You can calculate the standard error of the expected number of "Remain" voters with the following code:

```
sqrt(N*p*(1-p))
```

What is the expected value of $\hat{X}$, the proportion of "Remain" voters?

0.481

✔ **Answer:** 0.481

0.481

Explanation

The expected value of $\hat{X}$ is $p = 0.481$.

What is the standard error of $\hat{X}$, the proportion of "Remain" voters?

0.0129

✔ **Answer:** 0.0129

0.0129

Explanation

You can calculate the standard error $\hat{X}$ with the following code:

```
sqrt(p*(1-p)/N)
```

What is the expected value of d, the spread between the proportion of "Remain" voters and "Leave" voters?

-0.038

✔ **Answer:** -0.038

-0.038

Explanation

Given the proportion p, the expected value of the spread is $2p - 1$:

```
2*p-1
```

What is the standard error of d, the spread between the proportion of "Remain" voters and "Leave" voters?

0.0258

✔ **Answer:** 0.0258

0.0258

Explanation

The standard error of the spread is twice the standard error of $\hat{X}$:

```
2*sqrt(p*(1-p)/N)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

Question 2: Actual Brexit poll estimates

4.0/4.0 points (graded)

Load and inspect the `brexit_polls` dataset from **dslabs**, which contains actual polling data for the 6 months before the Brexit vote. Raw proportions of voters preferring "Remain", "Leave", and "Undecided" are available (`remain` , `leave` , `undecided`) The spread is also available (`spread`), which is the difference in the raw proportion of voters choosing "Remain" and the raw proportion choosing "Leave".

Calculate `x_hat` for each poll, the estimate of the proportion of voters choosing "Remain" on the referendum day ($p = 0.481$), given the observed `spread` and the relationship $\hat{d} = 2\hat{X} - 1$. Use `mutate` to add a variable `x_hat` to the `brexit_polls` object by filling in the skeleton code below:

```
brexit_polls <- brexit_polls %>%
        mutate(x_hat = ___________)
```

What is the average of the observed spreads (`spread`)?

0.0201

✔ **Answer:** 0.0201

0.0201

Answer code

```
brexit_polls <- brexit_polls %>%
mutate(x_hat = (spread + 1)/2)
mean(brexit_polls$spread)
```

What is the standard deviation of the observed spreads?

0.0588 ✔ **Answer:** 0.0588

0.0588

Answer code

```
sd(brexit_polls$spread)
```

What is the average of `x_hat`, the estimates of the parameter p?

0.51 ✔ **Answer:** 0.51

0.51

Answer code

```
mean(brexit_polls$x_hat)
```

What is the standard deviation of `x_hat`?

0.0294 ✔ **Answer:** 0.0294

0.0294

Answer code

```
sd(brexit_polls$x_hat)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 3: Confidence interval of a Brexit poll

3/3 points (graded)

Consider the first poll in `brexit_polls`, a YouGov poll run on the same day as the Brexit referendum:

```
brexit_polls[1,]
```

Use `qnorm` to compute the 95% confidence interval for $\hat{X}$.

What is the lower bound of the 95% confidence interval?

0.506 ✔ **Answer:** 0.506

0.506

Answer code

```
x_hat <- 0.52
N <- 4772
se_hat <- sqrt(x_hat*(1-x_hat)/N)
x_hat - qnorm(.975)*se_hat
```

What is the upper bound of the 95% confidence interval?

0.534 ✔ **Answer:** 0.534

0.534

Answer code

```
x_hat + qnorm(.975)*se_hat
```

Does the 95% confidence interval predict a winner (does not cover $p = 0.5$)? Does the 95% confidence interval cover the true value of p observed during the referendum?

- The interval predicts a winner and covers the true value of p.
- **The interval predicts a winner but does not cover the true value of p.** ✔
- The interval does not predict a winner but does cover the true value of p.
- The interval does not predict a winner and does not cover the true value of p.

Answer code

```
!between(0.5, x_hat - qnorm(.975)*se_hat, x_hat + qnorm(.975)*se_hat)    # predicts winner
between(0.481, x_hat - qnorm(.975)*se_hat, x_hat + qnorm(.975)*se_hat)    # does not cover p
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Continue the comprehensive assessment on the next page.