Inference problems

Group 9

M5 ICA4

Introduction

Quantities $\bar{x}$ and $\hat{p}$ are good point estimates for the population mean and population proportion, respectively. They are used in estimation and hypothesis testing. These point estimates vary from one sample to another. We will use package `infer` to get a deeper understanding of bothsimulation-based and theoritical inference. To get started, load packages `tidyverse` and `infer`.

```
library(tidyverse)
library(infer)
```

Below is a basic custom theme. Feel free to try it out when you use `ggplot()`. Simply add it as layer to your plot. Rather than using `theme_bw()` you can use `theme_custom()`. This custom theme increases the font point size on axes and their labels.

```
theme_custom <- function() {
  theme_bw() +
  theme(axis.title = element_text(size = 16),
        title = element_text(size = 20),
        axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        plot.caption = element_text(size = 10))
}
```

Inference on population proportion

The American Automobile Association claims that 54% of fatal car/truck accidents are caused by driver error. A researcher studies 35 randomly selected accidents and finds that 14 were caused by driver error. Test the claim at significance level of $\alpha = 0.05$.

Hypotheses

State the null and alternative hypotheses given the problem above.

Null Hyp: The proportion of fatal car accidents is 0.54 from the American Automobile Association are due to driver error.

Alternative Hyp: The proportion of fatal car accidents is not equal to 0.54 from driver error.

Simulation-based inference

Below we create a data set to use with package `infer`. We will create a data frame `driver` that has "yes" or "no" outcomes with regards to the question: Was the accident caused due to driver error?

```
# A tibble: 2 x 2
  value proportion
  <chr>      <dbl>
1 no           0.6
2 yes          0.4
```

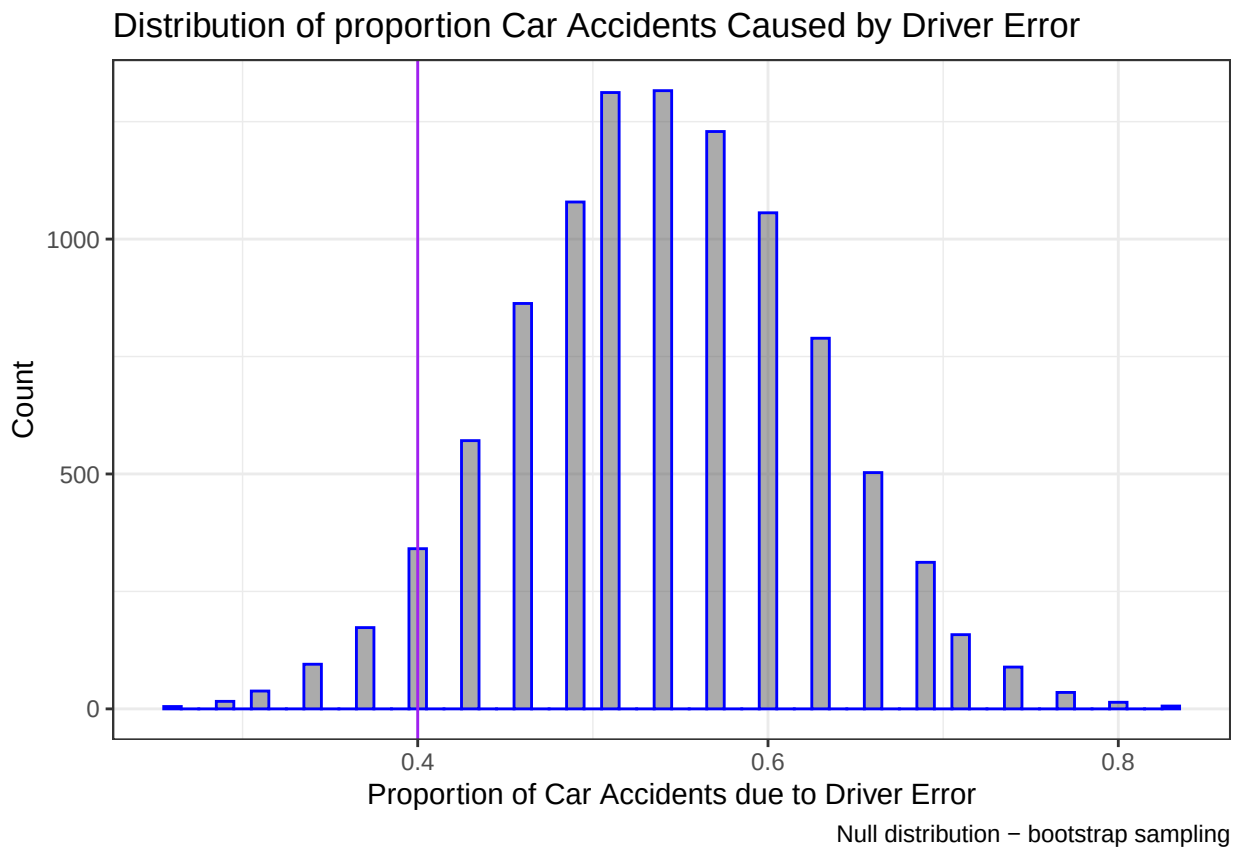
Simulated null distribution

Simulate the null distribution using a sequence of functions from package `infer`. Follow the template in the notes, previous assignments, andtake a look at the help for each function. Plot a histogram of the null distribution and place a vertical line at the value of the observed sample proportion of 0.4.

```
null.dist.prop<-driver%>%
  specify(response = value, success = "yes") %>%
  hypothesize(null = "point", p = 0.54) %>%
  generate(reps = 10000, type="simulate") %>%
  calculate(stat = "prop")
head(null.dist.prop)
```

```
Response: value (factor)
Null Hypothesis: point
# A tibble: 6 x 2
  replicate  stat
  <fct>     <dbl>
1 1         0.629
2 2         0.457
3 3         0.6
4 4         0.6
5 5         0.571
6 6         0.6
```

```
null.dist.prop%>%
  ggplot(mapping = aes(x = stat)) +
  geom_histogram(binwidth = .01, color = "blue", alpha = .5) +
  labs(x = "Proportion of Car Accidents due to Driver Error", y = "Count",
       title = "Distribution of proportion Car Accidents Caused by Driver Error",
       caption = "Null distribution - bootstrap sampling") +
  theme_bw()+
  geom_vline(xintercept = 0.4, color="purple")
```



Compute the p-value

Use the simulated null distribution to compute the p-value. Recall that the p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, given that the null hypothesis is true.

```
null.dist.prop%>%
  get_p_value(obs_stat = 0.4, direction = "two-sided")
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.134
```

Conclusion from the hypothesis testing

State your conclusion to "Does this provide convincing evidence that 54% of fatal car/truck accidents are caused by driver error", at the 5% significance level?

Will your conclusion change if the testing was done at the 1% significance level?"

Yes, we fail to reject the null hypothesis because there is sufficient evidence stating that 54% of car/truck accidents are caused by driver error.

If tested at the 1% significance level, our conclusion would not chnage becuase our P-value is still greater.

Explore the plausible values: Confidence interval

Compute a 95% confidence interval for the proportion of all accidents caused due to driver error. You may assume all the necessary assumptions are satisfied.

```
boot.means <- driver %>%
  specify(response = value, success = "yes") %>%
  generate(reps = 10000, type="bootstrap") %>%
  calculate(stat = "prop")

boot.means %>%
  summarise(lower95 = quantile(stat, probs = .025),
            upper95 = quantile(stat, probs = .975),
            lower99 = quantile(stat, probs = 0.005),
            upper99 = quantile(stat, probs = .995))
```

```
# A tibble: 1 x 4
  lower95 upper95 lower99 upper99
    <dbl>   <dbl>   <dbl>   <dbl>
1   0.257   0.571     0.2   0.600
```

Compare the 95% confidence interval with the 99% confidence interval.

The lower 95 is greater than the lower 99 while the upper 95 is less than the upper 99.The 99 has a larger width which indicates larger confidence.

Hypothesis testing and Confidence interval

Does the confidence interval constructed in the previous example reflect the results obtained from the hypothesis testing? Justify your answer in 2-5 sentences.

The p-value obtained from the hypothesis testing does not fall in the 95 percent confidence interval of 0.22 and 0.57. Therefore, we cannot confidently reject the null hypothesis. This reflects what we have obtained in the testing above.

Inference on population mean

A certain chemical pollutant in the Genesee River has been constant for several years with mean = 34 ppm (parts per million) and standard deviation = 8 ppm. A group of factory representatives whose companies discharge liquids into the river is now claiming that they have lowered the average with improved filtration devices. A group of environmentalists will test to see if this is true at the *4%* level of significance. Assume that their sample of size 50 gives a mean of 32.5 ppm. Perform a hypothesis test at the 4% level of significance and state your conclusion.

Hypotheses

State the null and alternative hypotheses given the problem above.

Null Hypothesis: The chemical pollutant in the Genesee River has a mean = 34ppm

Alt Hypothesis: The chemical pollutant in the Genesee River has a mean less than 34 ppm.

Simulated null distribution

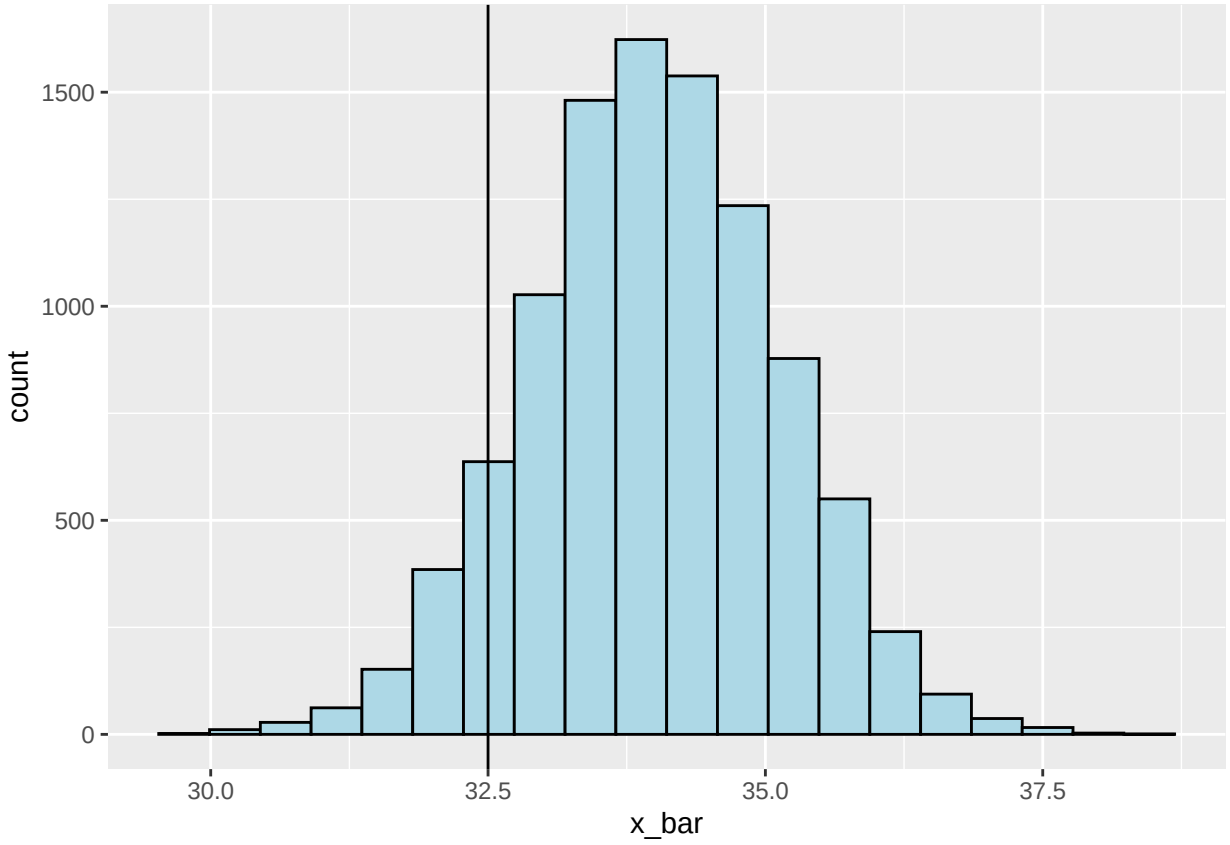
Plot a histogram of the simulated null distribution and place a vertical line at the value of the observed sample mean.

```
population <- tibble(value = rnorm(n = 1000000, mean = 34, sd = 8))

sample.means <- population %>%
  infer::rep_sample_n(size = 50, replace = TRUE, reps = 10000)

x_bar_reps<-sample.means%>%
  group_by(replicate) %>%
  summarise(x_bar = mean(value))

# Graphs sampling distribution
x_bar_reps %>%
  ggplot(mapping = aes(x = x_bar)) +
  geom_histogram(bins = 20, fill = "lightblue", color = "black") +
  geom_vline(xintercept = 32.5, color = "black")
```



Compute the p-value

Use the simulated null distribution to compute the p-value. Recall that the p-value is the probability of observing data at least as favorable to the alternative hypothesis as the current data set, given that the null

hypothesis is true. NOTE: For this section you will first have to generate the population and then randomly generate samples (maybe 1000) of size 50, calcualte the means and then plot the sampling distribution.

```
x_bar_reps %>%
  filter(x_bar <= 32.5) %>%
  summarize(p.value = n()/nrow(x_bar_reps))
```

```
# A tibble: 1 x 1
  p.value
    <dbl>
1  0.0917
```

Conclusion from the hypothesis testing

Perform a hypothesis test to determine if there convincing evidence to go with the factory representatives claim. Use a 4% significance level?

Will your conclusion change if the testing was done at the 1% significance level?"

We fail to reject the null hypothesis indicating that there was not sufficient evidence to prove workers were not to lower the chemical pollutant in the river. The river still has a mean of 34 ppm.

This conclusion would not change at the 1% signicance level as the p-value is still larger than 0.01%.

Explore the plausible values: Confidence interval

Confidence interval

Compute a 96% and 99% confidence interval for the average chemical pollutant in the Genesee River. You may assume all the necessary assumptions are satisfied.

```
sample.means <- population %>%
  infer::rep_sample_n(size = 50, replace = TRUE, reps = 1)
sample.means
```

```
# A tibble: 50 x 2
# Groups:   replicate [1]
   replicate value
       <int> <dbl>
 1         1  21.0
 2         1  50.0
 3         1  40.4
 4         1  33.9
 5         1  38.8
 6         1  26.7
 7         1  20.7
 8         1  27.7
 9         1  39.1
10         1  33.0
# ... with 40 more rows
```

```r
boot.means <- sample.means %>%
  specify(response = value ) %>%
  hypothesize(null = "point", mu = 32.5) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

boot.means %>%
  summarise(lower96 = quantile(stat, probs = .02),
            upper96 = quantile(stat, probs = .98),
            lower99 = quantile(stat, probs = 0.005),
            upper99 = quantile(stat, probs = .995))
```

```
# A tibble: 1 x 4
  lower96 upper96 lower99 upper99
    <dbl>   <dbl>   <dbl>   <dbl>
1    30.0    34.9    29.5    35.5
```

References

1. https://cran.r-project.org/web/packages/infer/vignettes/flights_examples.html
2. http://math.oxford.emory.edu/site/math117/probSetHypothesisTestsOneProportion2/