IST772 Summary Template: Chapter 4 – Introducing the Logic of Inference Using Confidence Intervals

Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Hendi Kushta
Date: 02/04/2023

****Important** Copying and/or pasting anything from the textbook will not be acceptable for your chapter notes submissions. You must write your notes in your own words and generate your own code, results, and graphs in R. This is what forces your brain to process the material that you read.**

The practice of using sample data to draw conclusions or forecasts about a population is known as inference in statistics. It entails using sample data to infer generalizations about population characteristics (such mean, variance, etc.). An example might be every person that tries lemon, tend to do a sour-face.
Induction is another type of inference. It is the process of predicting future events based on past data. It includes drawing conclusions about a population by using patterns and trends from a sample. Examples might be stock price prediction, or rainfalls etc.
The results that can be taken from inference or samples can not prove anything. Making estimates and/or inferences about a population with a certain level of certainty or confidence is the aim of statistical inference.
The author of the book uses an already prebuilt dataset in R. The mtcar dataset, which is a dataset that compares the fuel consumption and performance between cars.
I will take in consideration hp or horsepower attribute. In a car, the power is measured in horsepower. It shows the engine's capacity for work and how rapidly it can transform fuel into kinetic energy to propel the car. In general, an automobile with greater horsepower may accelerate more quickly and travel at higher speeds and most of the times the higher the HP, the more fuel the car uses.

```
> nrow(mtcars[mtcars$am == 1,])# Automatic transmissions
[1] 13
> nrow(mtcars[mtcars$am == 0,])# Manual transmissions
[1] 19
```

There are 13 cars with automatic transmission and 19 cars with manual transmission.
Below I am showing the mean and the standard deviation for the HP for each of the groups of cars automatic or manual.

IST772 Summary Template: Chapter 4 – Introducing the Logic of Inference Using Confidence Intervals

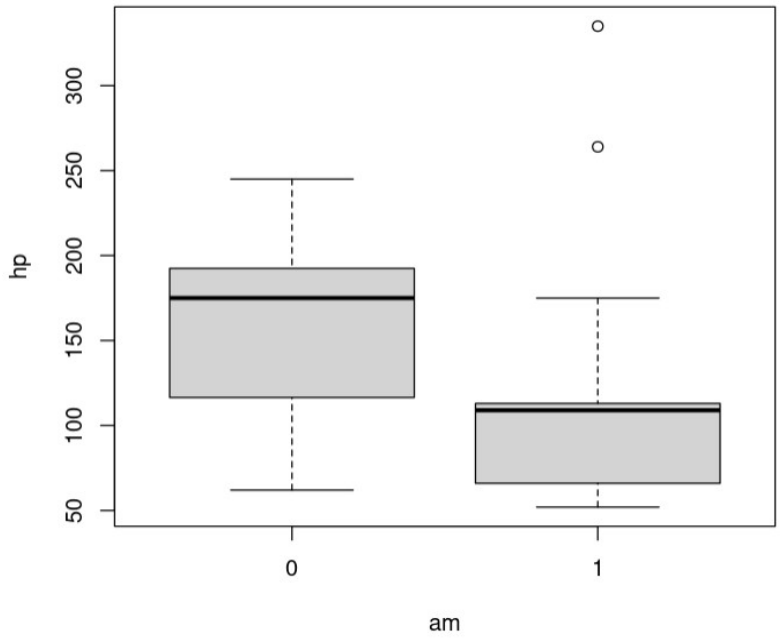
Originality Assertion: By submitting this file you affirm that this writing is your own.

```
> mean(mtcars$hp[ mtcars$am == 0 ])# Automatic transmissions
[1] 160.2632
> mean(mtcars$hp[ mtcars$am == 1 ])# Manual transmissions
[1] 126.8462
>
> sd(mtcars$hp[ mtcars$am == 0 ])# Automatic transmissions
[1] 53.9082
> sd(mtcars$hp[ mtcars$am == 1 ])# Manual transmissions
[1] 84.06232
```

In this case we can not say nothing. The mean of horse power for automatic transmission is higher meaning that automatic cars use more fuel, but according to standard deviation, the manual transmission cars use more fuel. To get a better comparison between these results, we might plot a box plot chart.

```
boxplot(hp ~ am, data=mtcars)          # Boxplot of hp, grouped by am
```

As we can see from the box plots, the manual transitions cars are mostly spread between 60 to 110, so the mean that high might be because of the outliers too. The manual transmission cars are less variable than the automatic ones.



IST772 Summary Template: Chapter 4 – Introducing the Logic of Inference Using Confidence Intervals

Originality Assertion: By submitting this file you affirm that this writing is your own.

EXPLORING THE VARIABILITY OF SAMPLE MEANS WITH REPETITIOUS SAMPLING

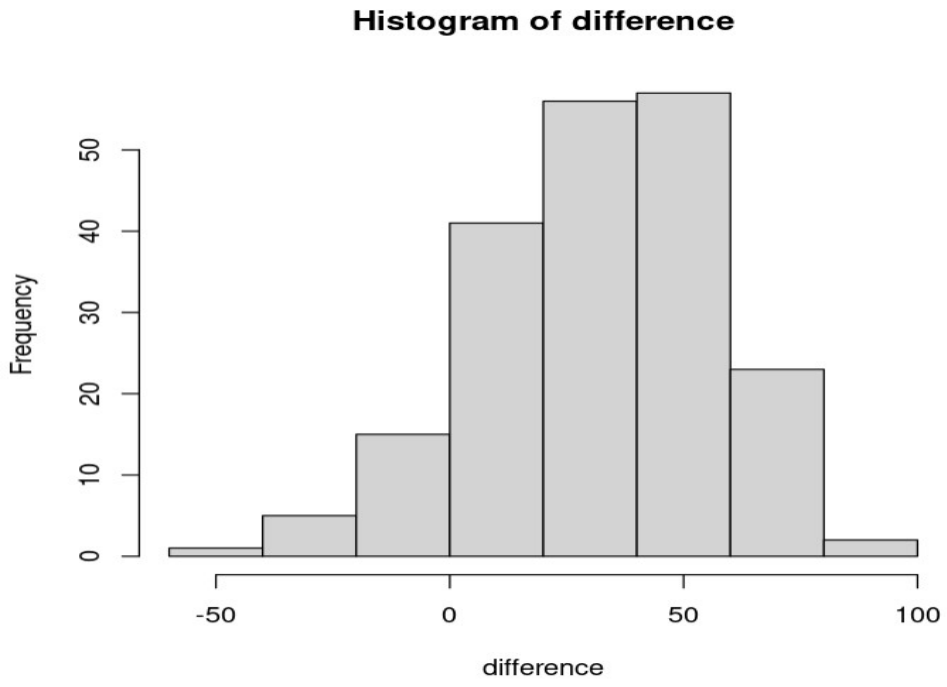
There will be comparison between automatic and manual transmission sample means with replacement for the horse powers in this section.

```
> mean( sample(mtcars$hp[ mtcars$am == 0 ],size=19,replace=TRUE) )
[1] 171.9474
> mean( sample(mtcars$hp[ mtcars$am == 1 ],size=13,replace=TRUE) )
[1] 118.2308
```

There is a mean of 172 horsepower for automatic and 118 for manual transmission, but every time I will run this lines of code, the results will be different.

Now lets just get the difference between these means and replicate the results 200 times to find the most convincing mean and add it to a variable so we can then draw the histogram to see the difference in means distribution.

```
> difference <- replicate(200,mean( sample(mtcars$hp[ mtcars$am == 0 ],size=19,replace=TRUE) ) -
+   mean( sample(mtcars$hp[ mtcars$am == 1 ],size=13,replace=TRUE) ))
> hist(difference)
```



IST772 Summary Template: Chapter 4 – Introducing the Logic of Inference Using Confidence Intervals

Originality Assertion: By submitting this file you affirm that this writing is your own.

OUR FIRST INFERENTIAL TEST: THE CONFIDENCE INTERVAL

An inferential test is a procedure used in statistical inference to evaluate a population-level hypothesis using data from a sample. Inferential testing seeks to draw generalizations about a population from data from a sample and use those generalizations to inform decisions or forecasts about the population. Regression analysis, ANOVA, and t-tests are examples of common inferential tests.
In the script below, we will see t-test. T-test or Student t-test is used to evaluate whether there is a significant difference between the means of two groups that are not related with one another.

```
> t.test(mtcars$hp[mtcars$am==0] ,mtcars$mpg[mtcars$am==1])

        Welch Two Sample t-test

data:  mtcars$hp[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
t = 10.883, df = 18.685, p-value = 1.593e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 109.7094 162.0323
sample estimates:
mean of x mean of y
160.26316  24.39231
```

These two samples were used in the t-test to determine a confidence interval with a mean difference of 109 to 162 horsepower. The values from 109 up to 162 are also known as interval estimate of population value.
It's crucial to consider the width of the confidence interval. A wide interval would imply that the population mean difference may range over a sizable area, in which case there would be significant doubt regarding the population value. It would be clear from a narrow interval that there is little uncertainty regarding the location of the population mean difference.
The confidence interval increases the weight of the evidence supporting our ideas, which is in line with how we should be thinking about inferential reasoning.

IST772 Summary Template: Chapter 4 – Introducing the Logic of Inference Using Confidence Intervals

Originality Assertion: By submitting this file you affirm that this writing is your own.

BOX ON P.61: FORMULAS FOR THE CONFIDENCE INTERVAL

Formulas for the Confidence Interval

A confidence interval is a range of values that is likely to contain the true value of a population parameter with a certain level of confidence.

Example: A 95% confidence ~~parameter~~ interval means that if the same sample was drawn multiple times and confidence level was interpreted every time, approximately 95% of the intervals would contain the true population parameter.

The confidence interval falls within Lower bound (left side) and Upper bound (right side)

$$\text{Lower bound} = \underbrace{(\bar{x}_1 - \bar{x}_2)}_{\text{is the difference between 2 sample means, so in the example from the horse power for automatic and manual transmissions,}} - \underbrace{t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}_{\text{calculates the width of the interval of confidence by substracting it from the means difference so we can find the lower bound.}}$$

$\bar{x}_1$ → sample mean for automatic

$\bar{x}_2$ → sample mean for manual

$\bar{x}_1 - \bar{x}_2 = 160,2632 - 126,8462$

$= 33,417$

t^* ⇒ is a critical value which is found in t-distribution table.

$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ ⇒ Standard error → represents the average difference between the sample estimate and the population

s^2 are variances for each samples

n are the number of observations for each sample.

* ⇒ To find the Upper bound its the same thing, but now we add the second part.

$$\text{Upper bound} = (\bar{x}_1 - \bar{x}_2) + t^* \sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}$$