Introduction to Logic of Inferences Using Confidence Intervals

T-Test

Introduction

- **Inference**- It is defined as a method in which we justify an information with reasons and that leads to conclusion.
- We are never certain about the conclusions that we have reached.
- We cannot make conclusion on the whole population as a result we draw samples from it and make our conclusion on that which ideally becomes the conclusion of whole population.

Boxplot Visualization:

- Boxplot is also referred as box and whiskers plot.
- I created a box plot on no. of drivers that got killed in Great Britain between 1969-84 due to Seatbelt law that came into effect in 1983.
- I calculated the mean of drivers that got killed when the law was not in affect v/s when the law came into effect.
- Code is as follows:
  - mean(data$DriversKilled[data$law==0])
  - mean(data$DriversKilled[data$law==1])
  - Here 0 is for no law present and 1 is for law being present.

Boxplot Code:

- **boxplot(DriversKilled~law,data = data)**
- Here **DriversKilled~law** indicates that Drivers killed is the dependent variable with ~ operator.
- Anything on the right of ~ is an independent variable.
- The dark band line in the middle indicates the median value i.e., 50^{th} percentile.
- Below and above that we have the 25^{th} and 75^{th} percentile.
- The mean for drivers killed during no law i.e., 0 is 125.86
- The mean for drivers killed during law i.e., 1 is 100.26

Exploring the Variability of Sample Means with Repetitious Sampling

- I'll create samples from our population of particular size with replacement using the sample () and get the average of no. of drivers that got killed.
- Code is as follows:
  - Law=0 mean(sample(data$DriversKilled[data$law==0],size=15, replace = TRUE))
  - Law=1 mean(sample(data$DriversKilled[data$law==0],size=15, replace = TRUE))

- We will get different mean from both the samples and they will change every time we run the above code.
- We can plot histogram of the difference between 2 values and replicate 100 times to get the normal distribution.

T

Code is as follows:

- diff<-replicate(100,mean(sample(data$DriversKilled[data$law==0],size=15,replace=TRUE))-mean(sample(data$DriversKilled[data$law==1],size=12, replace = TRUE)))
- hist(diff)
- We can also find the quantiles using the quantile for our diff variable.

Code:

- quantile(diff,c(0.025,0.975))
- This gives me the value at 2.5% lower bound region and 97.5% upper bound region.

Our First Inferential Test: The Confidence Interval

- In most of the cases we would not know the population standard deviation so we would not be able to calculate the standard error and perform the hypothesis test.
- Our 1st choice in Hypothesis testing is the Z-test

Critical value method

P-value method

- This test won't be useful when we don't have any data about our population for analysis. We would not have an idea about true population mean and population standard deviation.
- When we don't have an idea about population standard deviation we use the t-test.

T-Test

- It is also a bell shaped curve just like normal distribution but it is shorter and flatter.
- There are multiple t-distribution and are distinguished by the degrees of freedom.
- As the sample size increases degrees of freedom also increases.
- At sample size 30 t-distribution comes alike as normal distribution. i.e., a t-test becomes a z-test if we take a sample size is greater than 30.
- Null Hypothesis: The difference b/w the two mean value is 0 $\mu_1-\mu_2=0$
- Alternate Hypothesis: The difference b/w the two mean value is not 0 i.e., $\mu_1-\mu_2$ 0

R-Code

- t.test(data$DriversKilled[data$law==0],data$DriversKilled[data$law==1])
- In the output I get the following things:
  - t=5.12 which is the t-statistic
  - df=29.6 degrees of freedom
  - p-value=0.00001693
- alternate hypothesis-Since, p-value<0.05 we reject the null hypothesis.
- 95 percent confidence interval-[15.39,35.81]
- Sample estimates- This gives us the sample mean of each group i.e., For 0 (Law not present)=124.86 and for 1(Law present)=100.26
- Without replicating the values if I subtract the 2 mean values, I get 21.08 as result.
- The above value lies in the confidence interval of 95%.
- If we have more data we can conduct a t-test to make better conclusion

T

R Code Fragment and Explanation

Dataset:

Using Seatbelts data set which shows the Road Casualties in Great Britain from 1969-84.

````r
```{r}
# I am using the Seatbelts dataset.
# It is about the Road Casualties in Great Britain 1969-84. Compulsary wearing of Seatbelts
View(Seatbelts)
```

```{r}
# Reading the Dataset
data<-data.frame(Seatbelts)
head(data)
```
````

	DriversKilled <dbl>	drivers <dbl>	front <dbl>	rear <dbl>	kms <dbl>	PetrolPrice <dbl>	VanKilled <dbl>	law <dbl>
1	107	1687	867	269	9059	0.1029718	12	0
2	97	1508	825	265	7685	0.1023630	6	0
3	102	1507	806	319	9963	0.1020625	12	0
4	87	1385	814	407	10955	0.1008733	8	0
5	119	1632	991	454	11823	0.1010197	10	0
6	106	1511	945	427	12391	0.1005812	13	0

6 rows

Boxplot

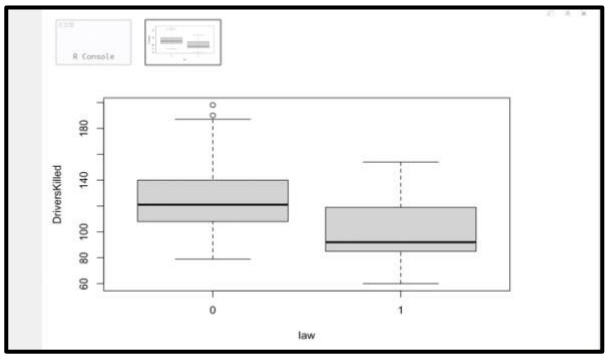
- Boxplot here shows us the whiskers of our two groups.
- On the left we have the group when the law for seat belts was not present.
- On the right we have the group when the law came into action.
- For 0: The median value is quite close to 1st and 3rd quartile.
- For 1: The median value is closer to the 1st quartile rather than the 3rd quartile
- Median of Law-1 is less than the median of Law-0.

````r
```{r}
# Boxplot
# Avg. no. drivers killed when the seatbelt law was not in effect i.e,0
mean(data$DriversKilled[data$law==0])

# Avg. no. drivers killed when the seatbelt law was in effect i.e, 1
mean(data$DriversKilled[data$law==1])

#Boxplot
boxplot(DriversKilled~law,data = data)
```

R Console
[1] 125.8698
[1] 100.2609
````



Sample Mean with Repetitious Sampling

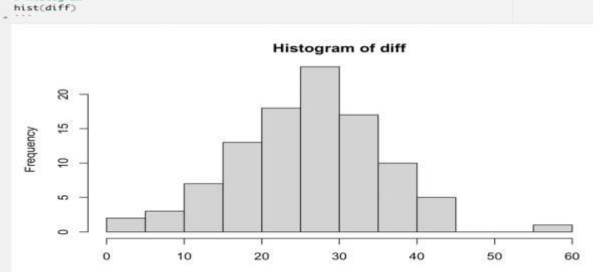
- We can find the mean of both the samples using the mean()
- Both the samples clearly have a different mean
- We can plot the histogram of difference in the two values over 1000 repetitions.
- Avg. no. of Drivers that were killed when there was no law is 124.933 and during law it is 104.41

````r
```{r}
# Sampling
# Law=0
mean(sample(data$DriversKilled[data$law==0],size=15, replace = TRUE))

# Law=1
mean(sample(data$DriversKilled[data$law==1],size=12, replace = TRUE))
```

[1] 124.9333
[1] 104.4167
````

````r
```{r}
# Histogram
hist(diff)
```
````



T

Quantiles

We can get the quantile values using the quantile() on diff at 2.5% and 97.5%

- It comes out to be 6.28 and 42.29.

```
```{r}
# Quantiles
quantile(diff,c(0.025,0.975))
```

     2.5%   97.5%
 6.287083 42.293333
```

T-Test

- We can run the t-test as per the below code in R:
- We just have to define the two parameters on which it will take place.
- Here Law=0 means that there was no law for seatbelts and Law=1 indicates that the law was introduced.
- The output generated will actually show us whether the diff value lies in the confidence interval of 95% or not.
- We get to see the mean of 2 groups i.e., for Law-0 it is 125.86 and for Law-1 it is 100.26

```
```{r}
# T-Test
t.test(data$DriversKilled[data$law==0],data$DriversKilled[data$law==1])
```

	Welch Two Sample t-test

data: data$DriversKilled[data$law == 0] and data$DriversKilled[data$law == 1]
t = 5.1253, df = 29.609, p-value = 1.693e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 15.39892 35.81899
sample estimates:
mean of x mean of y
 125.8698 100.2609
```