LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: **James Connolly**

Assessment Title: **Hypothesis Testing**

Work to be submitted to: **James Connolly**

Date for submission of work: **29-08-2018**

Place and time for submitting work: **3:00 pm**

To be completed by the Student

Student's Name: **PRATEEK PARASHER**

Class: **Msc in Big Data**

Subject/Module: **Data Science**

Word Count (where applicable): _____

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: **PRATEEK PARASHER** Date: **29/8/2018**

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations.]

Plagiarism: Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).

Cheating: The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination.

ABSTRACT: - Using statistical analysis I am interested to examine my dataset, but I am not thinking there must be some correlation between weather red warnings & road accidents. But I can't make that decision on my hypothesis or my assumption I need to perform statistical hypothesis analysis testing. I applied power test, t-test , correlation test. Correlation test is used to find the sample size to perform the power analysis with effect size of 0.5 with 80% certainty and no more than a 5% chance of inaccuracy. Then the sample of 29 records is used to do the power analysis. And the output of p-value is significantly high from 0.05. Therefore, this proves that the null hypothesis is true

DATA DESCRIPTION

```
> head(data)
  Years Months accidents year month warning_colors days warning_element
1  2018    mar        25 2018     3            red    6        snow_ice
2  2018    feb        27 2018     2            red    4        snow_ice
3  2018    jan        20 2018     1            red    1        snow_ice
4  2017    dec         9 2017    12     no-warning   30          normal
5  2017    nov        10 2017    11     no-warning   30          normal
6  2017    oct        23 2017    10            red    6            wind
```

```
> str(data)
'data.frame':   29 obs. of  8 variables:
 $ Years          : int  2018 2018 2018 2017 2017 2017 2017 2017 2017 2015 ...
 $ Months         : Factor w/ 12 levels "april","aug",..: 8 4 5 3 10 11 12 2 6 3 ...
 $ accidents      : int  25 27 20 9 10 23 7 4 9 19 ...
 $ year           : int  2018 2018 2018 2017 2017 2017 2017 2017 2017 2015 ...
 $ month          : int  3 2 1 12 11 10 9 8 7 12 ...
 $ warning_colors : Factor w/ 2 levels "no-warning","red": 2 2 2 1 1 2 1 1 1 2 ...
 $ days           : int  6 4 1 30 30 6 30 30 30 2 ...
 $ warning_element: Factor w/ 5 levels "normal","normal
                                  "| __truncated__,..: 4 4 4 1 1 5 2 2 2 3 ...
```

Dataset containing year 2012-2018 data of road accident and red weather warning data with different -2 warning elements.

Type of data -> continuous
No. of sample -> two – sample
Hypothesis testing -> correlation

HYPOTHESIS TESTING

H0 = The road accident is not related to the red weather warnings
H1 = The road accident is related to the red weather warnings

probability that IF the null hypothesis were true, sampling variation would produce an estimate that is further away from the hypothesised value than my data estimate

There are many statistical methods out of them I am using correlation method

predetermined cutoff (0.05) is called the significance level of the test that's why I am using 0.05 value for my calculation

```
> effect_size <- cohen.ES(test= "r", size= "large")
> effect_size

     Conventional effect size from Cohen (1982)

           test = r
           size = large
    effect.size = 0.5
```

```
> sample_size <- pwr.r.test(r = effect_size$effect.size, sig.level = 0.05, power = 0.8)
> sample_size

     approximate correlation power calculation (arctangh transformation)

              n = 28.24841
              r = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

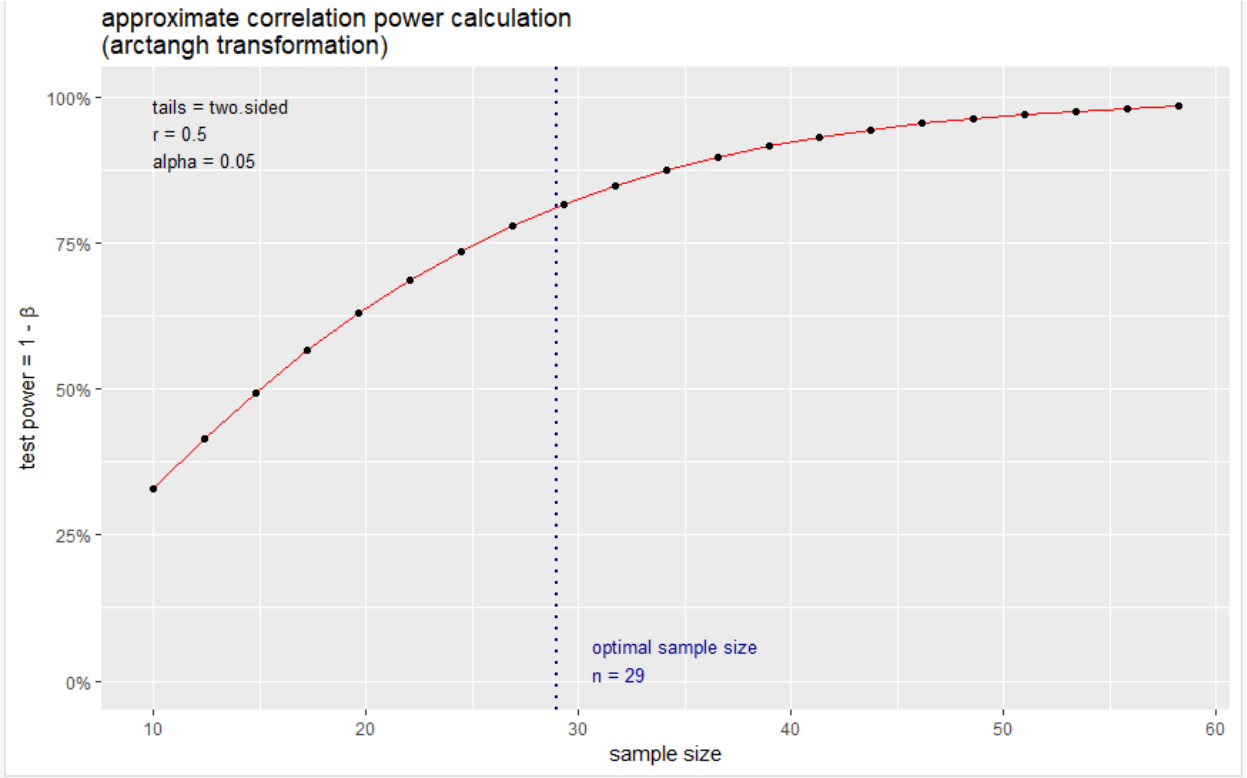

Fig 1 :- The results suggest that we need 29 records to detect an effect size of 0.5 with 80% certainty and no more than a 5% chance of inaccuracy.

In above experiment I choose **size = large** and **power = 0.8**

Effect size is the magnitude of the effect under the alternate hypothesis **which is 0.3 with size = medium**

```
           test = r
           size = medium
    effect.size = 0.3

approximate correlation power calculation (arctangh transformation)

              n = 111.8068
              r = 0.3
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

when I choose the size medium instead of large effect size is 0.3 and power is directly proportional to N(sample size)

s.no	test	size	effect size	N	power
1	r	med	0.3	111.8	0.9
2	r	med	0.3	84	0.8
3	r	med	0.3	66.5	0.7
4	r	large	0.5	37	0.9
5	r	large	0.5	28.2	0.8
6	r	large	0.5	22.7	0.7
7	r	small	0.1	1045.8	0.9
8	r	small	0.1	781.7	0.8
9	r	small	0.1	615.1	0.7

From above table we can see that how conventional size having impact on effect size(r) N(sample size) inversely proportional to conventional size.

I have data with 30 rows so I can perform my test with N 28.2 & 22.7 these are minimum sample size required to perform correlation test with N 28.2 I already did.

```
> effect_size <- cohen.ES(test= "r", size= "large")
> effect_size

     Conventional effect size from Cohen (1982)

           test = r
           size = large
    effect.size = 0.5

approximate correlation power calculation (arctangh transformation)

              n = 22.71375
              r = 0.5
      sig.level = 0.05
          power = 0.7
    alternative = two.sided
```

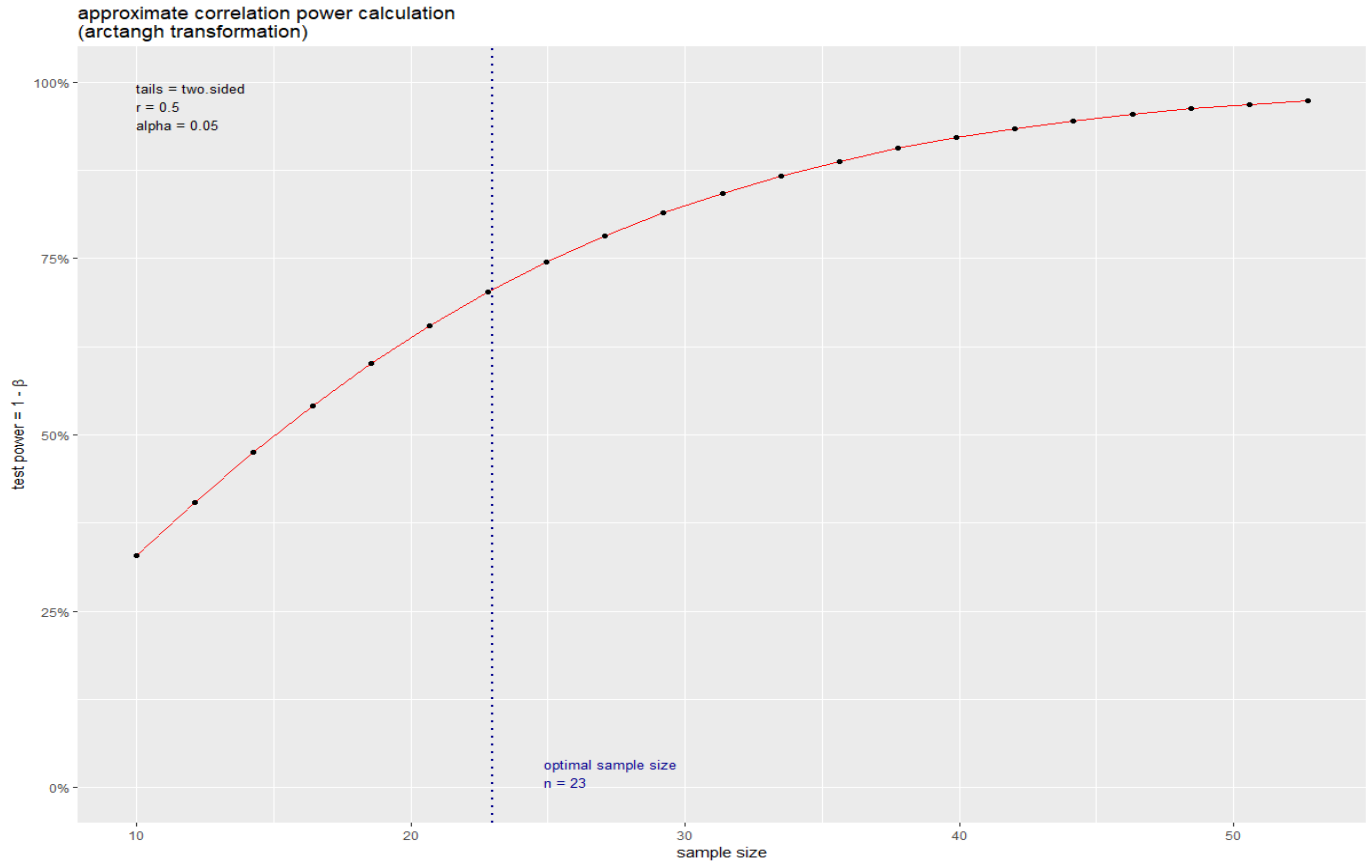

Fig 2 :- The results suggest that we need 23 records to detect an effect size of 0.5 with 70% certainty and no more than a 5% chance of inaccuracy.

```
cor.test( sample_data$accident, sample_data$month)

        Pearson's product-moment correlation

data:  sample_data$accident and sample_data$month
t = -2.0368, df = 21, p-value = 0.05447
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.700998665  0.007261029
sample estimates:
       cor
-0.4061568
```

In this experiment we change our power 0.7 with N = 23 In this result **P value is 0.0544**

p-value is really close to 0.05, the results should be considered marginally significant - the decision could go either way.

RESULTS

```
> cor.test( sample_data$accident, sample_data$month)

        Pearson's product-moment correlation

data:  sample_data$accident and sample_data$month
t = -1.3234, df = 27, p-value = 0.1968
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5624347  0.1316059
sample estimates:
       cor
-0.2468042
```

- P value is 0.1968 in that case we have less strict cut-offs, such as 0.10, requiring less evidence

The results are considered non-significant - fail to reject Ho.

result it is seen that p-value is significantly high from 0.05. Therefore, this proves that the null hypothesis is true which means that the alternate hypothesis is false. That is, there is a relation between road accidents and red weather warnings.

Conclusion :- From this hypothesis test, I am concluding that the red weather warnings is having relation with road accident. By this result my initial null hypothesis false and I will continue my prediction and further findings with alternative hypothesis.

GitHub Link :- https://github.com/prateekparasher/web_scrap