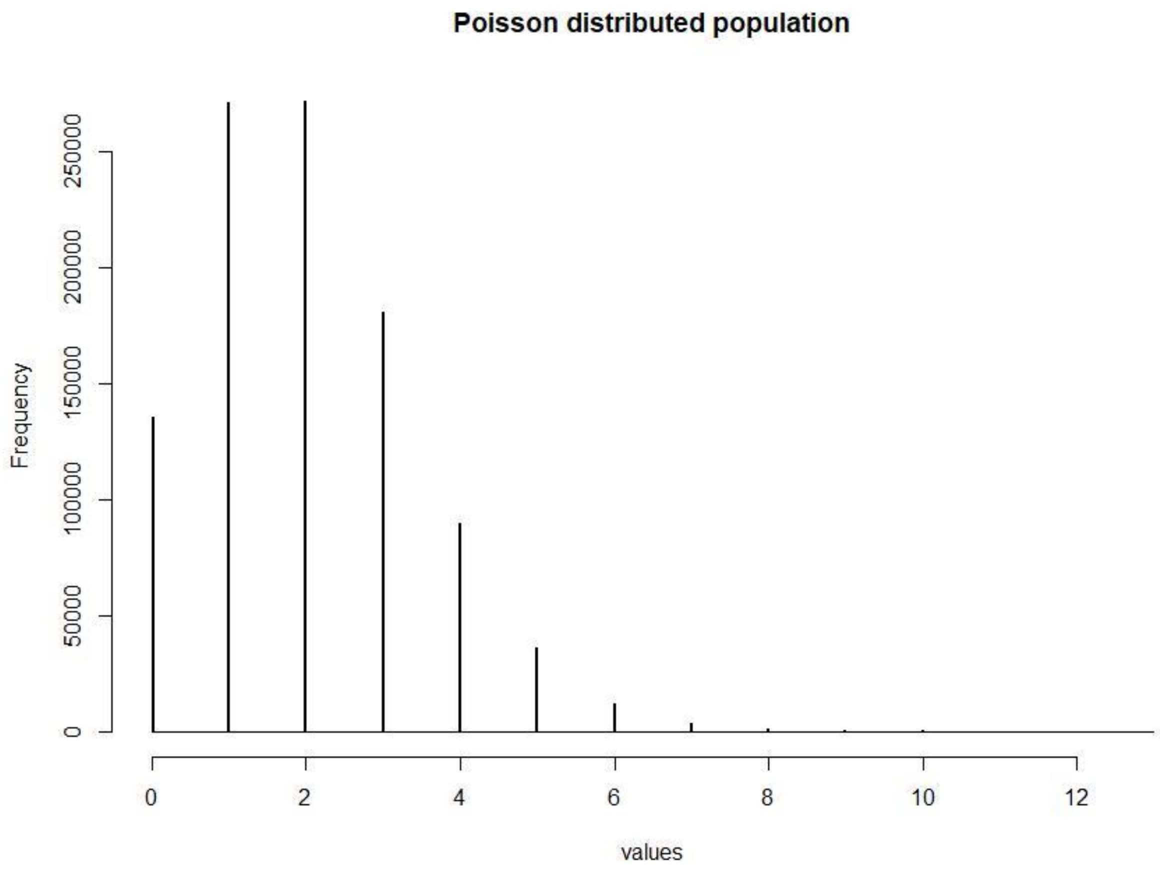
BE303 LAB ASSIGNMENT : REPORT

Ayush Nigam B20005

2.1
Subpart (a)_____________



Population distribution (lambda = 2)

Subpart (b)_____________

We know that samples are withdrawn randomly from poisson distribution randomly so they must be similar and any dissimilarity found between two samples is on account of random chance. **[1]**

HYPOTHESIS TESTING

H0 : two samples drawn from population are similar
H1 : two samples drawn from population will be dissimilar

To test similarity between two samples we use t-test.

From [1] Null hypothesis will not be rejected most of the times (assuming large sample size ; an assumption of t-test which is being used here)

Type 1 error is when we reject the null hypothesis when it’s actually true.

It is unlikely to commit a type 1 error here if our tests work properly because our null hypothesis is true most of the times [1]

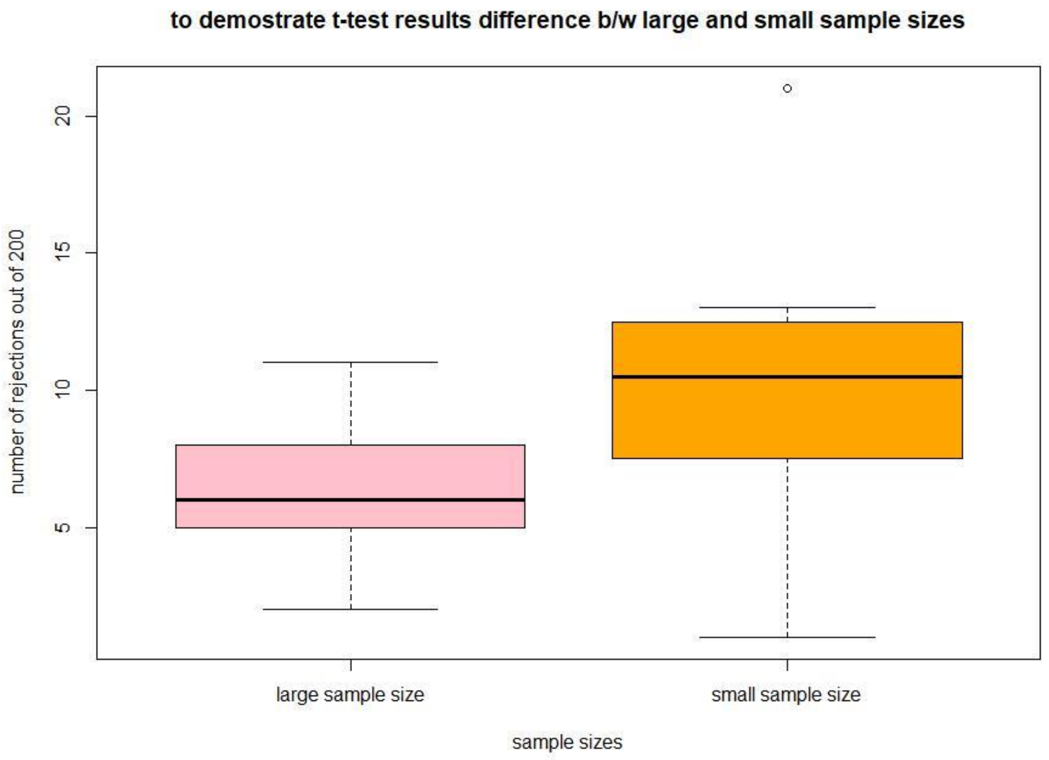
Out of 20 experiments where 200 iterations of t-tests are done on random sample population collected (sample size = 10000) are observed at a time following are the number of times null hypothesis is rejected out of those 200

```
> large_sample_size
 [1]  8425585661038741158576
```

Here we can see that it is very hard to reject the null hypothesis.

The alpha value taken is low (5%) which causes the above observation. **Thus alpha value represents type 1 error.**

Subpart (c)__________



Above is a box plot representing an experiment described as follows -

We define two sample sizes -
Large = 10000
Small = 7

We draw 200 sample populations at a time from the total population with large sample size. We evaluate the acceptance or rejection of null hypothesis for these 200 sample populations based on the methodology explained in previous subparts.
This Process is done 20 times and the number of rejections of null hypothesis out of 200 sample populations in each iteration is the variable of interest which is being used for each category of boxplot

Rejecting Null hypothesis (out of 200 sample populations) :

```
> large_sample_size
 [1]84255856610387411158576

> small_sample_size
 [1] 121313 611 8 710 6 8 612 112 91313102112
```

Results from 20 iterations for both sample sizes are stated above

We see from the boxplot that when sample size is reduced, the null hypothesis gets rejected more frequently which is not desirable because it will imply that samples taken are not similar (which is not the case because they are being randomly sampled from the main population).

This is a direct consequence of the failure of t-test for small sample size, because t-test used an assumption -

> **The fourth assumption is a reasonably large sample size is used. A larger sample size means the distribution of results should approach a normal bell-shaped curve.**
>
> **Source - [2]**

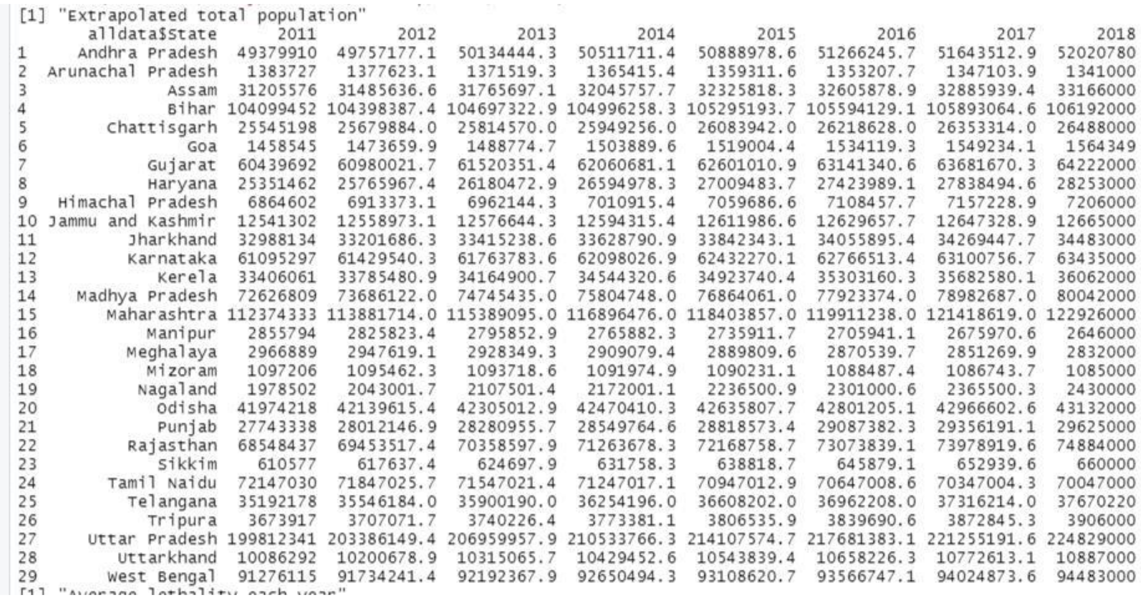
2.2

Extrapolating Total population assuming a linear regression :

Since only two data points namely population in 2011 and 2018 have been given to extrapolate, we don’t need to fit a regression model, the population values in 2012, 2013 and so on will lie exactly on the line connection population in 2011 and 2018 if they are plotted on 2D space.

We find the difference b/w these extreme values and divide it by number of years in b/w i.e 7 and repeatedly add it to the lower value i.e population 2011 to find the population in subsequent years

Extrapolated data

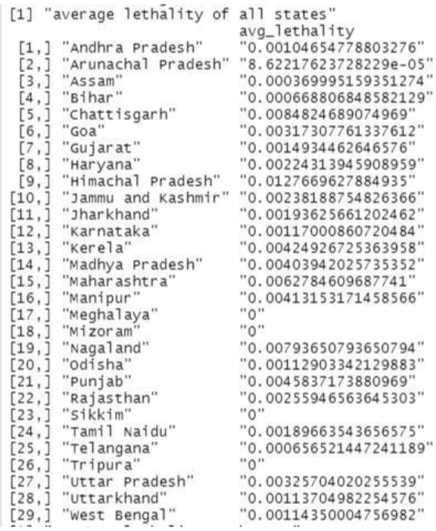


```
[1] "Extrapolated total population"
       alldata$State      2011        2012        2013        2014        2015        2016        2017      2018
1     Andhra Pradesh  49379910  49757177.1  50134444.3  50511711.4  50888978.6  51266245.7  51643512.9  52020780
2  Arunachal Pradesh   1383727   1377623.1   1371519.3   1365415.4   1359311.6   1353207.7   1347103.9   1341000
3              Assam  31205576  31485636.6  31765697.1  32045757.7  32325818.3  32605878.9  32885939.4  33166000
4              Bihar 104099452 104398387.4 104697322.9 104996258.3 105295193.7 105594129.1 105893064.6 106192000
5        Chattisgarh  25545198  25679884.0  25814570.0  25949256.0  26083942.0  26218628.0  26353314.0  26488000
6                Goa   1458545   1473659.9   1488774.7   1503889.6   1519004.4   1534119.3   1549234.1   1564349
7            Gujarat  60439692  60980021.7  61520351.4  62060681.1  62601010.9  63141340.6  63681670.3  64222000
8            Haryana  25351462  25765967.4  26180472.9  26594978.3  27009483.7  27423989.1  27838494.6  28253000
9   Himachal Pradesh   6864602   6913373.1   6962144.3   7010915.4   7059686.6   7108457.7   7157228.9   7206000
10 Jammu and Kashmir  12541302  12558973.1  12576644.3  12594315.4  12611986.6  12629657.7  12647328.9  12665000
11         Jharkhand  32988134  33201686.3  33415238.6  33628790.9  33842343.1  34055895.4  34269447.7  34483000
12         Karnataka  61095297  61429540.3  61763783.6  62098026.9  62432270.1  62766513.4  63100756.7  63435000
13            Kerela  33406061  33785480.9  34164900.7  34544320.6  34923740.4  35303160.3  35682580.1  36062000
14    Madhya Pradesh  72626809  73686122.0  74745435.0  75804748.0  76864061.0  77923374.0  78982687.0  80042000
15       Maharashtra 112374333 113881714.0 115389095.0 116896476.0 118403857.0 119911238.0 121418619.0 122926000
16           Manipur   2855794   2825823.4   2795852.9   2765882.3   2735911.7   2705941.1   2675970.6   2646000
17         Meghalaya   2966889   2947619.1   2928349.3   2909079.4   2889809.6   2870539.7   2851269.9   2832000
18           Mizoram   1097206   1095462.3   1093718.6   1091974.9   1090231.1   1088487.4   1086743.7   1085000
19          Nagaland   1978502   2043001.7   2107501.4   2172001.1   2236500.9   2301000.6   2365500.3   2430000
20            Odisha  41974218  42139615.4  42305012.9  42470410.3  42635807.7  42801205.1  42966602.6  43132000
21            Punjab  27743338  28012146.9  28280955.7  28549764.6  28818573.4  29087382.3  29356191.1  29625000
22         Rajasthan  68548437  69453517.4  70358597.9  71263678.3  72168758.7  73073839.1  73978919.6  74884000
23            Sikkim    610577    617637.4    624697.9    631758.3    638818.7    645879.1    652939.6    660000
24       Tamil Naidu  72147030  71847025.7  71547021.4  71247017.1  70947012.9  70647008.6  70347004.3  70047000
25         Telangana  35192178  35546184.0  35900190.0  36254196.0  36608202.0  36962208.0  37316214.0  37670220
26           Tripura   3673917   3707071.7   3740226.4   3773381.1   3806535.9   3839690.6   3872845.3   3906000
27     Uttar Pradesh 199812341 203386149.4 206959957.9 210533766.3 214107574.7 217681383.1 221255191.6 224829000
28        Uttarkhand  10086292  10200678.9  10315065.7  10429452.6  10543839.4  10658226.3  10772613.1  10887000
29       West Bengal  91276115  91734241.4  92192367.9  92650494.3  93108620.7  93566747.1  94024873.6  94483000
[1] "Average lethality each year"
```

Total population

Disease - **dengue**

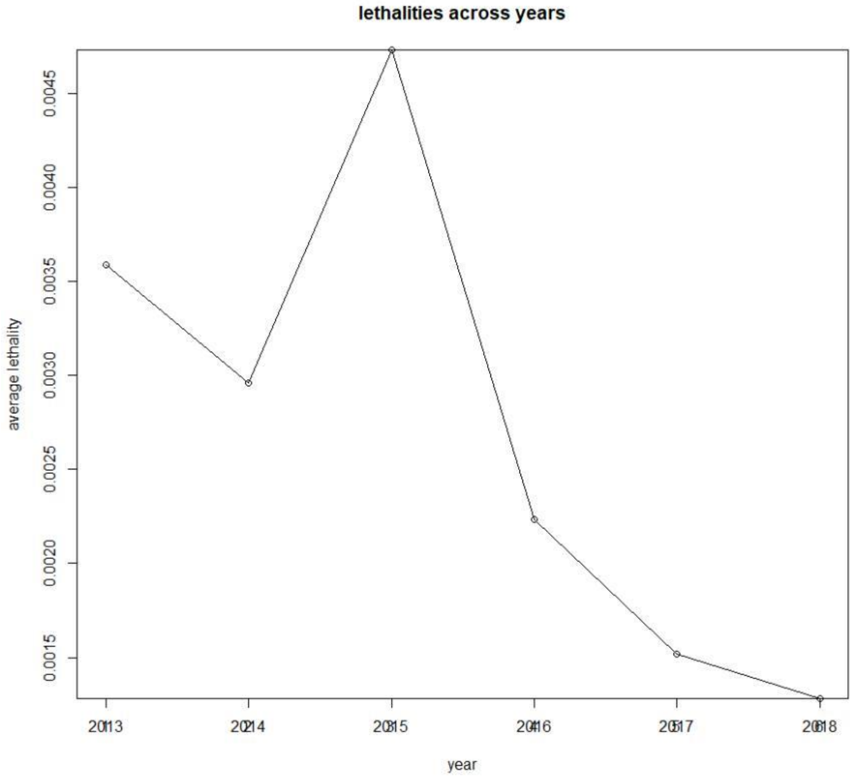
Average lethality of all states : across years



```
[1] "average lethality of all states"
                                avg_lethality
 [1,] "Andhra Pradesh"    "0.00104654778803276"
 [2,] "Arunachal Pradesh" "8.62217623728229e-05"
 [3,] "Assam"             "0.000369995159351274"
 [4,] "Bihar"             "0.000668806848582129"
 [5,] "Chattisgarh"       "0.0084824689074969"
 [6,] "Goa"               "0.00317307761337612"
 [7,] "Gujarat"           "0.0014934462646576"
 [8,] "Haryana"           "0.00224313945908959"
 [9,] "Himachal Pradesh"  "0.0127669627884935"
[10,] "Jammu and Kashmir" "0.00238188754826366"
[11,] "Jharkhand"         "0.00193625661202462"
[12,] "Karnataka"         "0.00117000860720484"
[13,] "Kerela"            "0.00424926725363958"
[14,] "Madhya Pradesh"    "0.00403942025735352"
[15,] "Maharashtra"       "0.0062784609687741"
[16,] "Manipur"           "0.00413153171458566"
[17,] "Meghalaya"         "0"
[18,] "Mizoram"           "0"
[19,] "Nagaland"          "0.00793650793650794"
[20,] "Odisha"            "0.00112903342129883"
[21,] "Punjab"            "0.0045837173880969"
[22,] "Rajasthan"         "0.00255946563645303"
[23,] "Sikkim"            "0"
[24,] "Tamil Naidu"       "0.00189663543656575"
[25,] "Telangana"         "0.000656521447241189"
[26,] "Tripura"           "0"
[27,] "Uttar Pradesh"     "0.00325704020255539"
[28,] "Uttarkhand"        "0.00113704982254576"
[29,] "West Bengal"       "0.00114350004756982"
```

Average lethality across years :

```
lethality 2013 lethality 2014 lethality 2015 lethality 2016 lethality 2017 lethality 2018(P)
0.003587346 0.002957288 0.004728888 0.002232982 0.001516794 0.001283662
```



Confidence interval for each year

```
[1] "Confidence Intervals for each year : "
[1] "lethality 2013"
[1] 0.002431015 0.004743676
[1] "lethality 2014"
[1] 0.001994208 0.003920368
[1] "lethality 2015"
[1] 0.002329102 0.007128673
[1] "lethality 2016"
[1] 0.001445079 0.003020884
[1] "lethality 2017"
[1] 0.001028418 0.002005170
[1] "lethality 2018(P)"
[1] 0.0009227569 0.0016445679
```

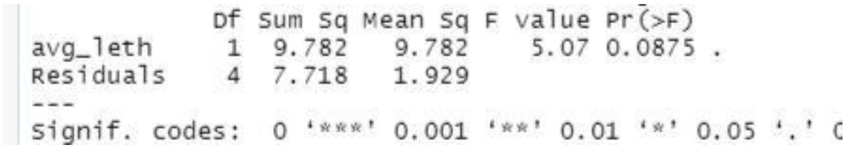
It is observed that average values are contained in the confidence interval.

A one way Anova test can be computed to assess the similarity between lethality across the years.

HYPOTHESIS

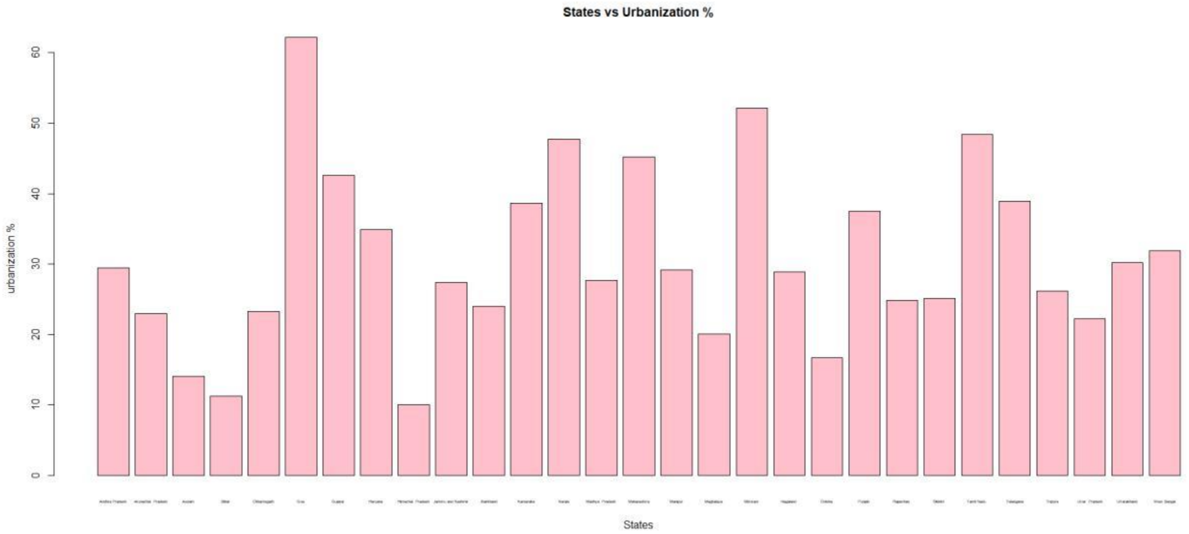
H_0 : mean lethality across the years is same
H_1 : mean lethality across the years is not same

RESULTS

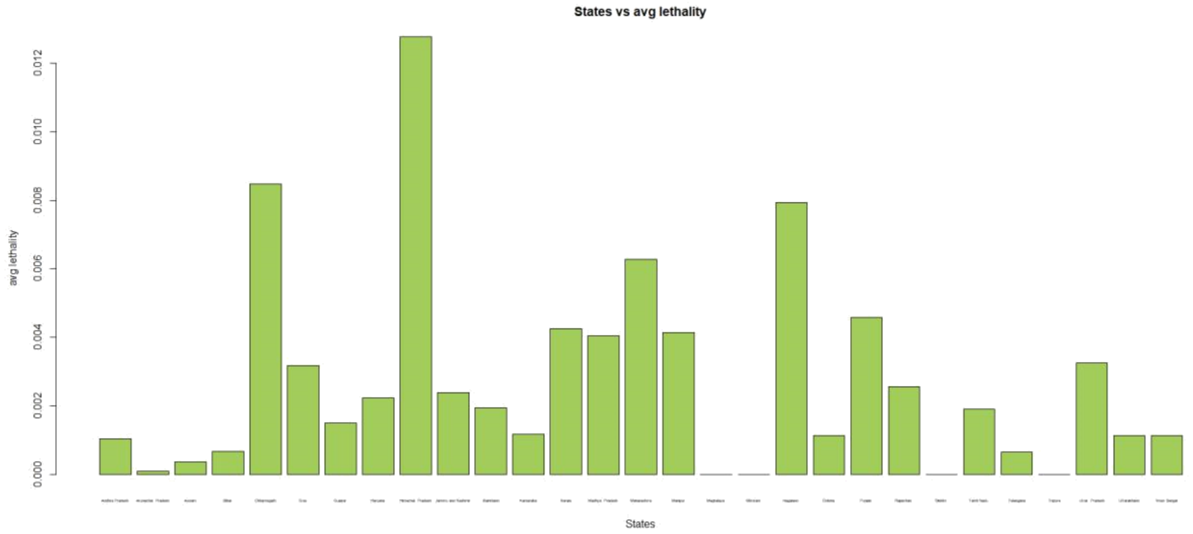


```
                Df Sum Sq Mean Sq F value Pr(>F)
avg_leth         1  9.782   9.782    5.07 0.0875 .
Residuals        4  7.718   1.929
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ (
```

We can see that the **F value** for the test is quite large (5.07) which portrays that there is a **significant difference among average lethalities across the years.**



States vs urbanization %



States vs average lethality across years

Correlation between these variable -
average lethality and urbanization % of states

Pearson correlation coefficient = -0.09495053

Which is not a significant correlation which can also be seen from the graphs which have very different distributions.

Thus lethality does not correlate with urbanization levels for *dengue.*

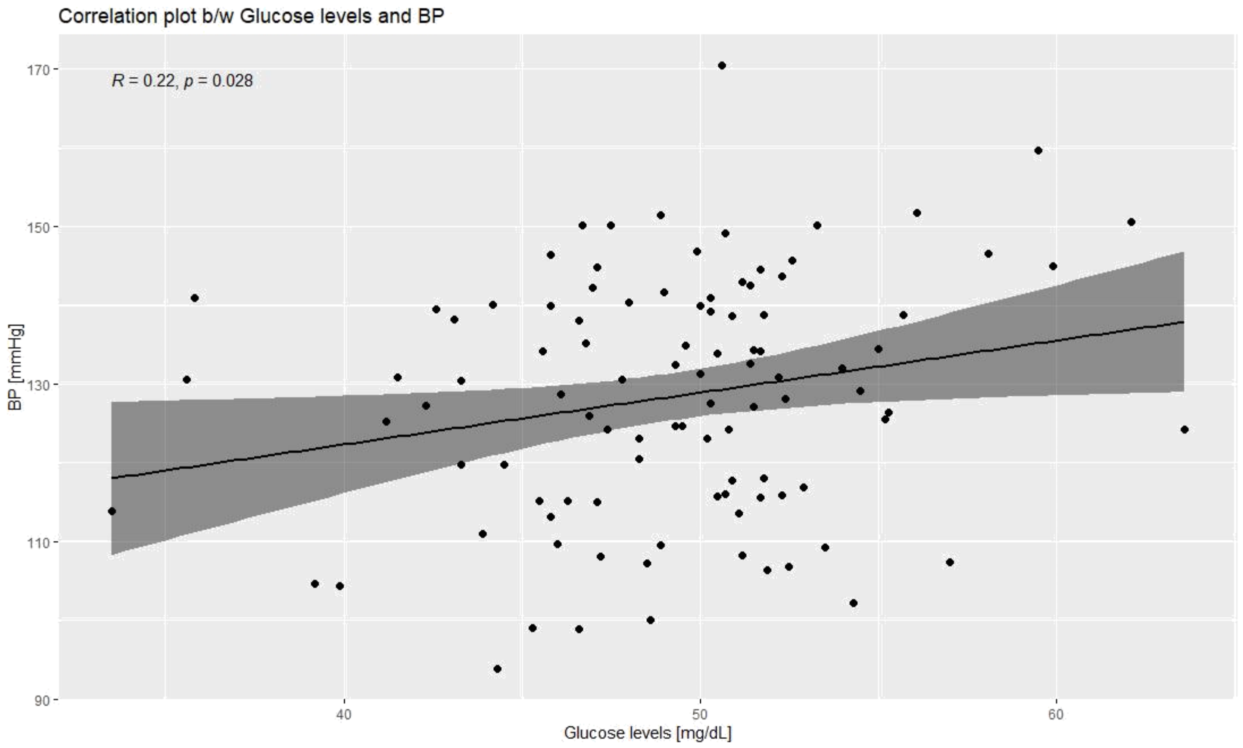
2.3

Research Question - Is there a relation between glucose levels and high blood pressure ?

H0 : There is no relation between glucose levels and high blood pressure.
H1 : There is a relation between glucose levels and high blood pressure.

The given dataset has two variables Glucose levels (mg/dL) and BP (mmHg).

The strength of dependency of the two variables can be measured with **pearson correlation**.

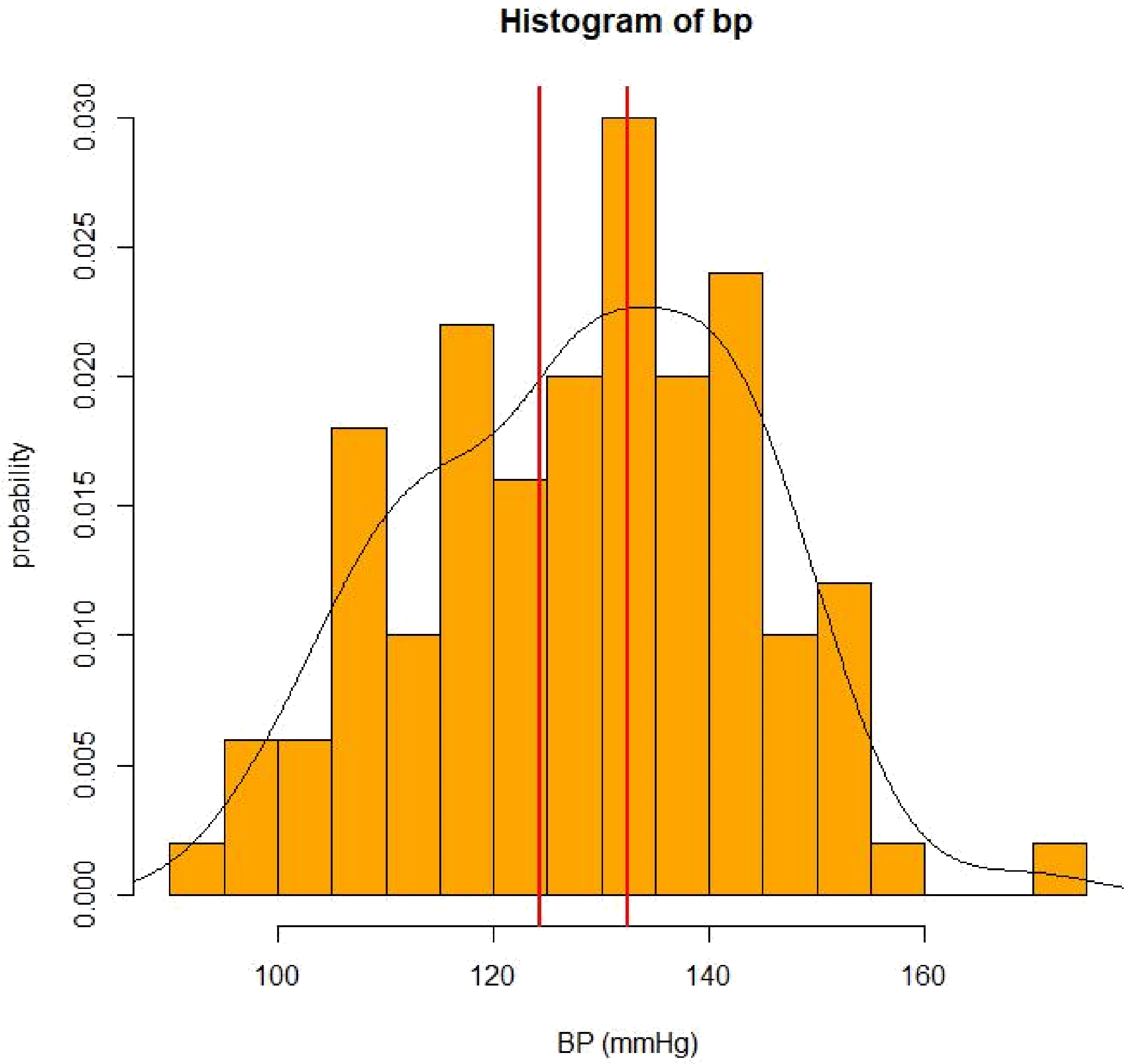


The Pearson correlation coefficient for the given variables is **0.2202548**.

This is not a high value of correlation so ***this model cannot (should not) be used to extrapolate data beyond the given data points.***

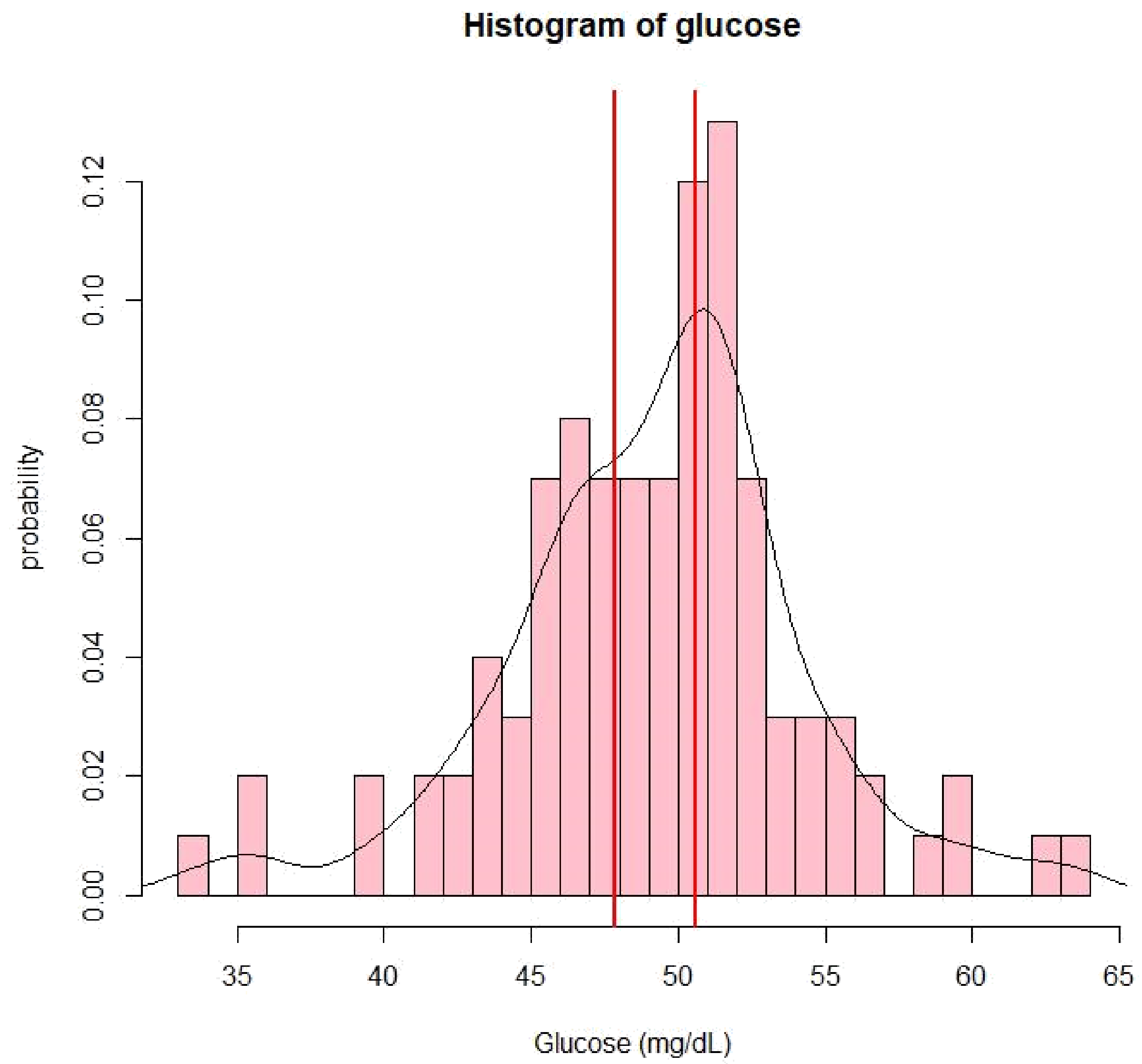
From observing the above graph as well, the data points are dispersed scattered which is not a good indicator of an underlying pattern.

Remarks and Conclusion -
The data given is very small (100) points, this should be considered before arriving at a judgment.
The current dataset gives a correlation of 0.2 which is positive (thus there is a dependency) but is weak and likely unimportant. However with a large dataset we stand a better chance of inferring the correct relationship.



Histogram of BP with confidence intervals in RED

lower bound of CI for BP = 124.367726853539 mmHg
upper bound of CI for BP = 132.442273146461 mmHg

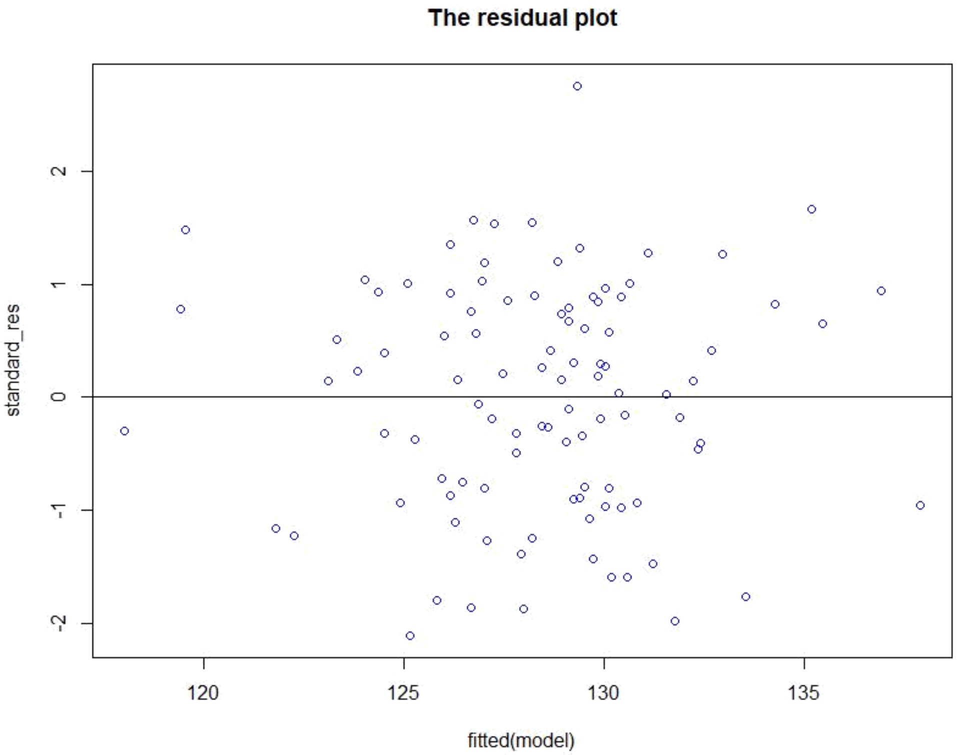


Histogram of Glucose levels with confidence intervals in RED

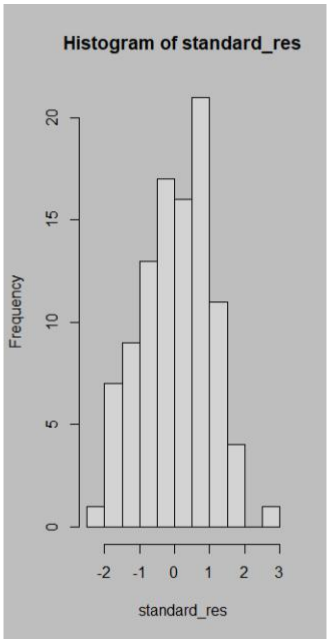
lower bound of CI for glucose levels is 47.8755965422861 mg/dL
upper bound of CI for glucose levels is 50.5664034577139 mg/dL

Residuals :

Min	1Q	Median	3Q	Max
-31.353	-12.268	2.186	12.334	41.184



The residual plot has similar density all over but is symmetrically distributed around origin, so the regression is not biased towards a side - positive or negative.



The standard residuals (hence errors) are normally distributed which is a good characteristic to check the unbiasedness of a model.

APPENDIX -

[2] https://www.investopedia.com/ask/answers/073115/what-assumptions-are-made-when-conducting-ttest.asp#:~:text=The%20common%20assumptions%20made%20when,of%20variance%20in%20standard%20deviation.