PAPER PRESENTATION

Electoral Statistics

DUSHYANT DIXIT AND CHAITANYA KUMARIA
[Email address]

Abstract

We live in the World's largest democracy and a vital cog in this democracy is the conduction of free and fair election. We all in our lives have seen many elections and what goes hand in hand with elections is News outlets publishing something called "Exit Poles", the day final voting ends. So, do they just randomly go and ask voters on some booths and decide their numbers, or is there some Mathematical knowledge involved? How does a news channel in his prime time show with confidence call a seat?

I am sure these questions would have eluded you as well.

So here, we are presenting to you the Mathematics(more specifically statistics) behind calling a seat.

Here, we have taken 2 Lok Sabha Seats (Lucknow and Kanpur). Using data election data since the 1st recorded election (Lok Sabha and Vidhan Sabha). We are looking at the following objectives:

1. To understand and see if elections in India follow a theoretical probability distribution.

2. Can we find a range wherein we can call elections?

3. Confidence interval in prediction.

Before proceeding, here is a list of important Statical concepts we have used.

Descriptive statistics:

A descriptive statistic (in the count noun sense) is a summary statistic that quantitatively describes or summarizes features from a collection of information, while descriptive statistics (in the mass noun sense) is the process of using and analyzing those statistics.

Inferential Statistics:

Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn.

Hypothesis testing :

Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory that applies to a population. Hypothesis testing uses sample data to evaluate a hypothesis about a population.

Chi-square Test

The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population.

Hypothesis testing has been likened to a criminal trial, in which a jury must use evidence to decide which of 2 possible truths, innocence (H0) or guilt (HA), is to be believed. Just as a jury is instructed to assume that the defendant is innocent unless proven otherwise, the investigator should assume there is no association unless there is strong evidence to the contrary. A jury's verdict must be either guilty or not guilty, in which case a not-guilty verdict does not equal innocence. Rather, it indicates that the burden of proof has not been met. Similarly, an investigator can only reject H0 or fail to reject it; failure to reject does not prove that the null H0 is true.

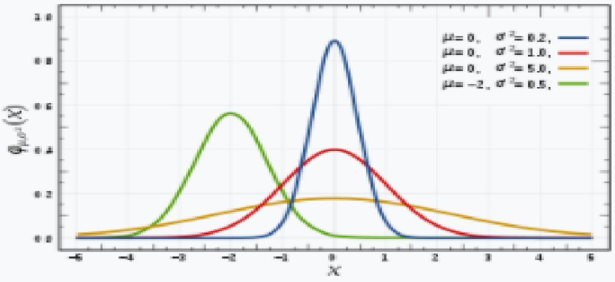
Major Probability Distributions used

Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of zero and a standard deviation of 1. The standard normal distribution is centered at zero and the degree to which a given measurement deviates from the mean is given by the standard deviation.

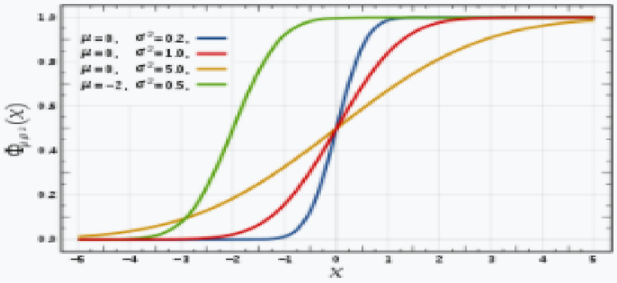
Normal distribution

Probability density function



The red curve is the *standard normal distribution*

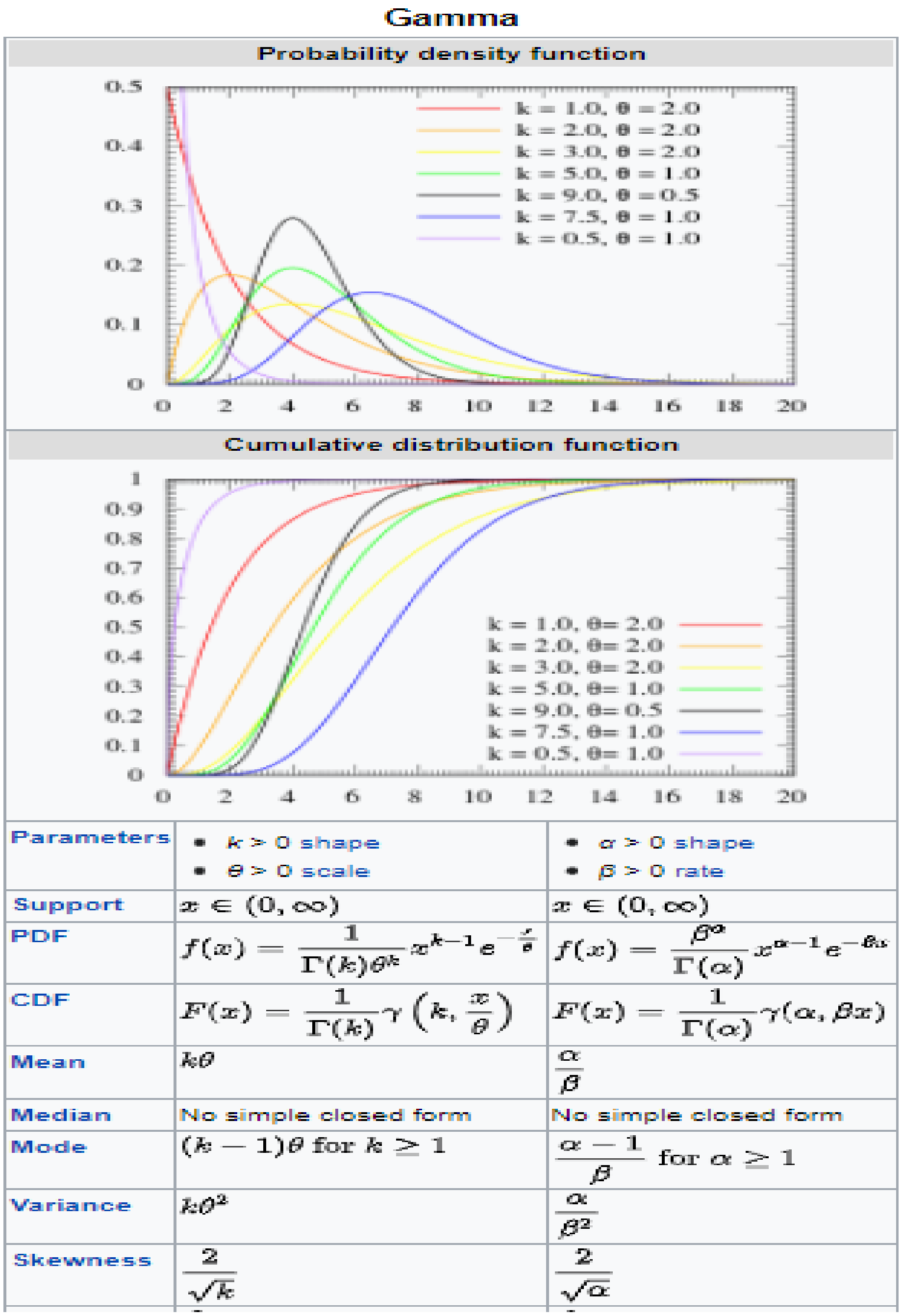
Cumulative distribution function



Notation	$\mathcal{N}(\mu,\sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 \in \mathbb{R}_{>0}$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2p-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
MAD	$\sigma\sqrt{2/\pi}$
Skewness	0

Gamma distribution

Gamma Distribution is one of the distributions widely used in the field of Business, Science, and Engineering to model the continuous variable that should have a positive and skewed distribution. Gamma distribution is a kind of statistical distribution that is related to the beta distribution. This distribution arises naturally in which the waiting time between Poisson distributed events is relevant to each other.



Parameters	• $k > 0$ shape • $\theta > 0$ scale	• $\alpha > 0$ shape • $\beta > 0$ rate
Support	$x \in (0, \infty)$	$x \in (0, \infty)$
PDF	$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
CDF	$F(x) = \frac{1}{\Gamma(k)} \gamma\left(k, \frac{x}{\theta}\right)$	$F(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta x)$
Mean	$k\theta$	$\frac{\alpha}{\beta}$
Median	No simple closed form	No simple closed form
Mode	$(k-1)\theta \text{ for } k \geq 1$	$\frac{\alpha - 1}{\beta} \text{ for } \alpha \geq 1$
Variance	$k\theta^2$	$\frac{\alpha}{\beta^2}$
Skewness	$\frac{2}{\sqrt{k}}$	$\frac{2}{\sqrt{\alpha}}$

Computer Software Used:

- **Python Programming Language:** We have used Python Programming language version 3.8 for our calculations and plots. We have used certain python libraries as well these include
  - **Numpy: Numerical Python**

  - **Pandas: used to import data in .csv, or .xlxs or .txt format in python paradigm**
  - **Matplotlib: This is used for plotting curves, and scatter plot**
  - **Statsmodel.api:** stats models is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.
  - **Scipy.stats: Used for theoretical probability distributions**
- **Microsoft Excel:** We have used excel notebooks to collect, clean, and sanitize data. We have also used these excel sheets in the python framework for Calculations.

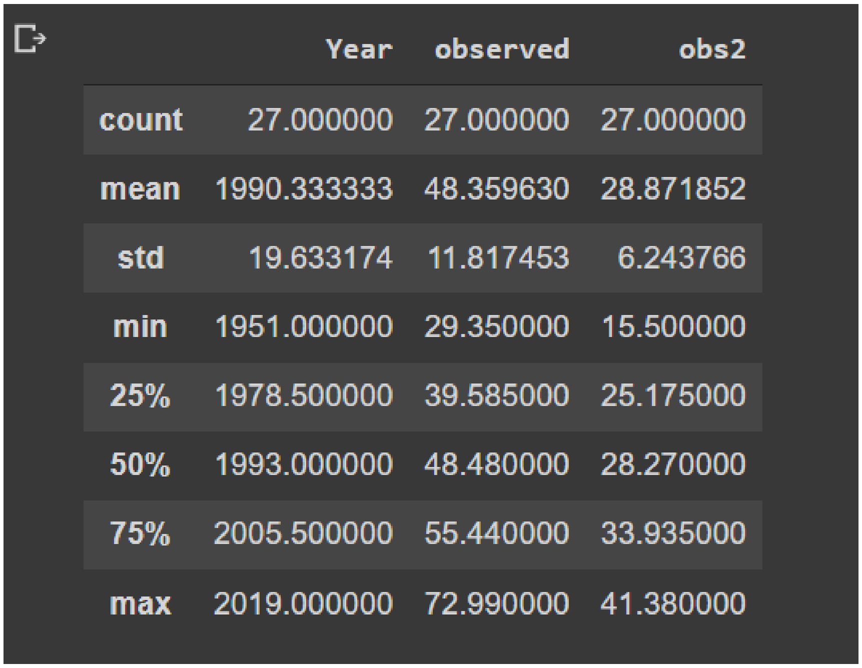
Assumptions:

Like any Statistical Analysis, here also we are assuming a few things:

1. Sample collected will always be homogeneous that is sample will be a true representation of the population
2. Voting pattern is homogeneous across Vidhan Sabha and Lok Sabha elections. It doesn't mean that the same party will be voted, it just means that a similar percentage of votes would decide the winner.

Procedure Followed

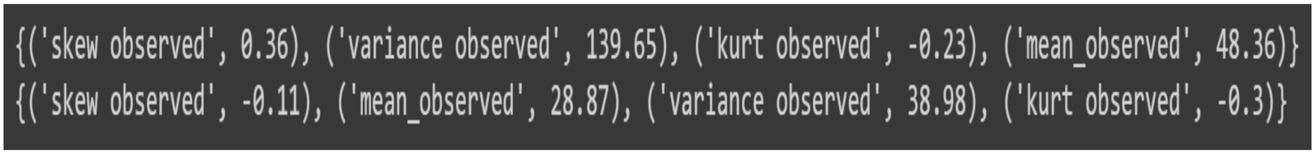
1. We are analyzing two seats Lucknow, and Kanpur. We have data for all the elections from 1952. We have taken the vote share of the winner and the Runners-up.
2. Following were our workings for Lucknow.
   We have,



	Year	observed	obs2
count	27.000000	27.000000	27.000000
mean	1990.333333	48.359630	28.871852
std	19.633174	11.817453	6.243766
min	1951.000000	29.350000	15.500000
25%	1978.500000	39.585000	25.175000
50%	1993.000000	48.480000	28.270000
75%	2005.500000	55.440000	33.935000
max	2019.000000	72.990000	41.380000

observed: Winner
obs2 : Runners up

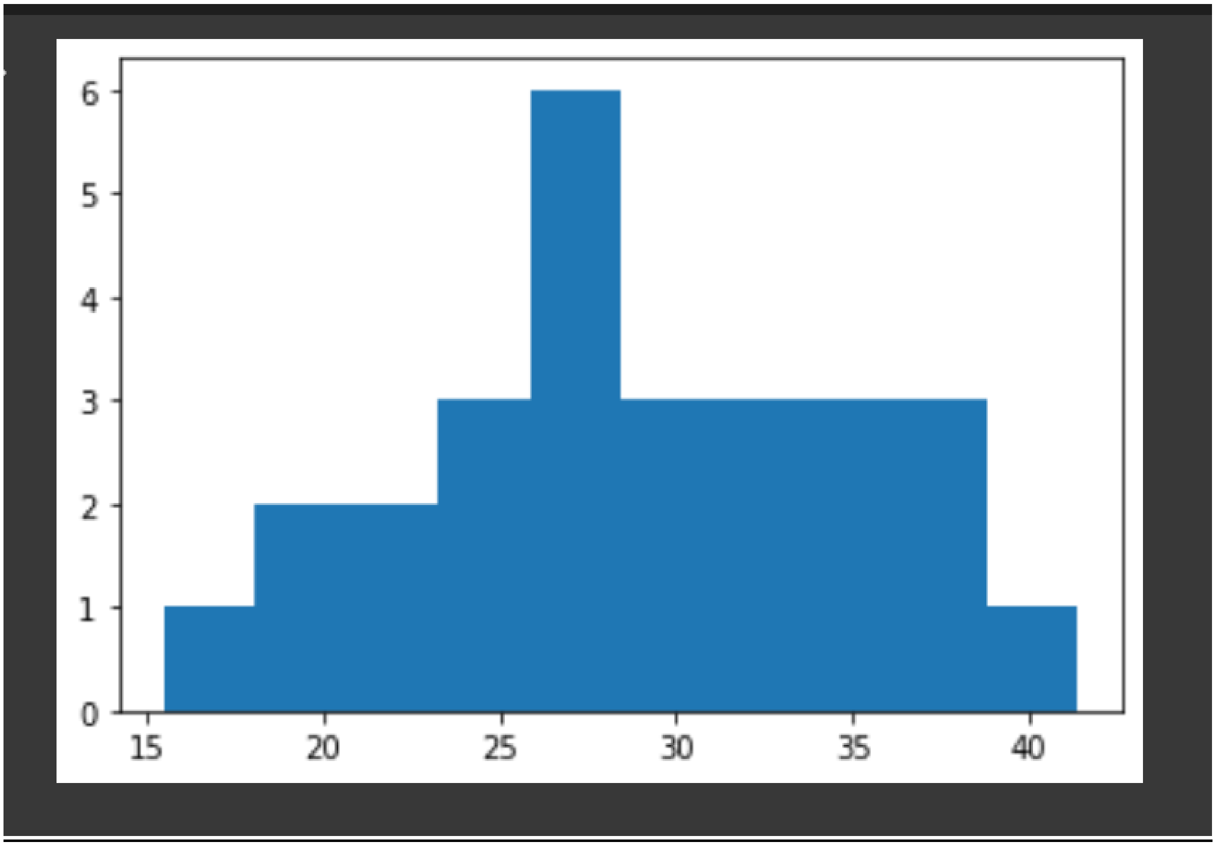
Following were other descriptive statistics of observed and obs2 respectively



```
{('skew observed', 0.36), ('variance observed', 139.65), ('kurt observed', -0.23), ('mean_observed', 48.36)}
{('skew observed', -0.11), ('mean_observed', 28.87), ('variance observed', 38.98), ('kurt observed', -0.3)}
```

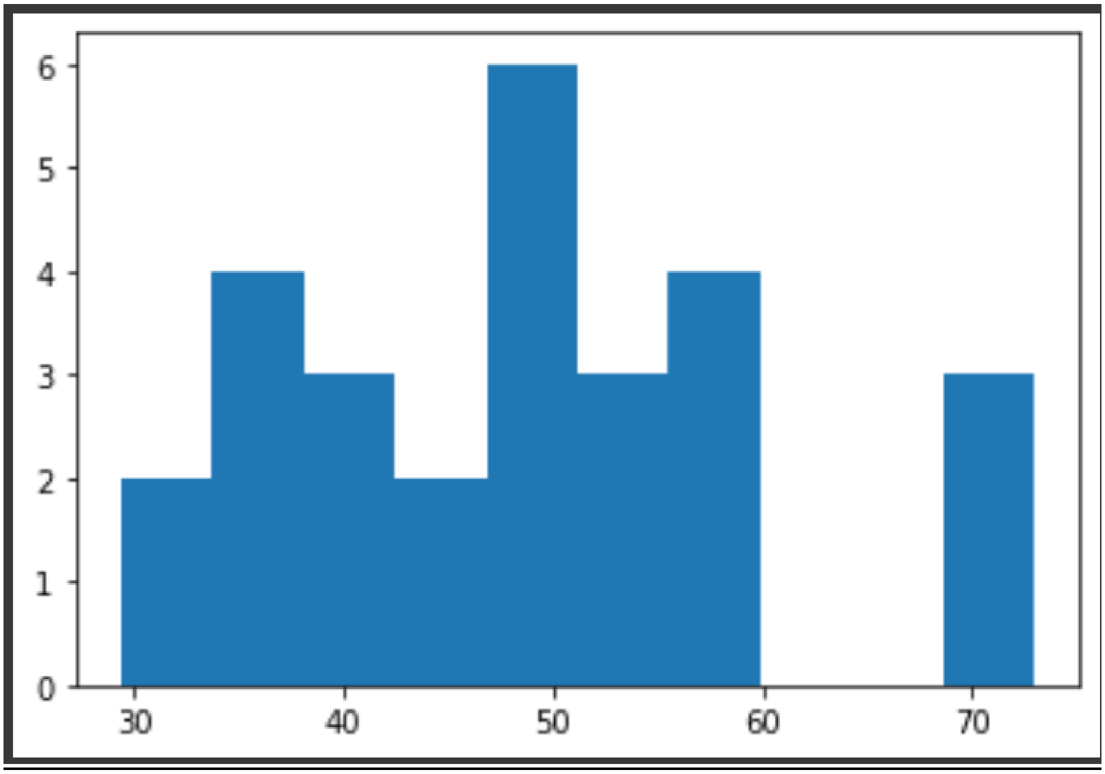
This is the histogram of runners-up data. Here, the x-axis represents vote share received

While Y-axis represents the s.no of election



This is the histogram of Winner’s data. Here, the x-axis represents vote share received

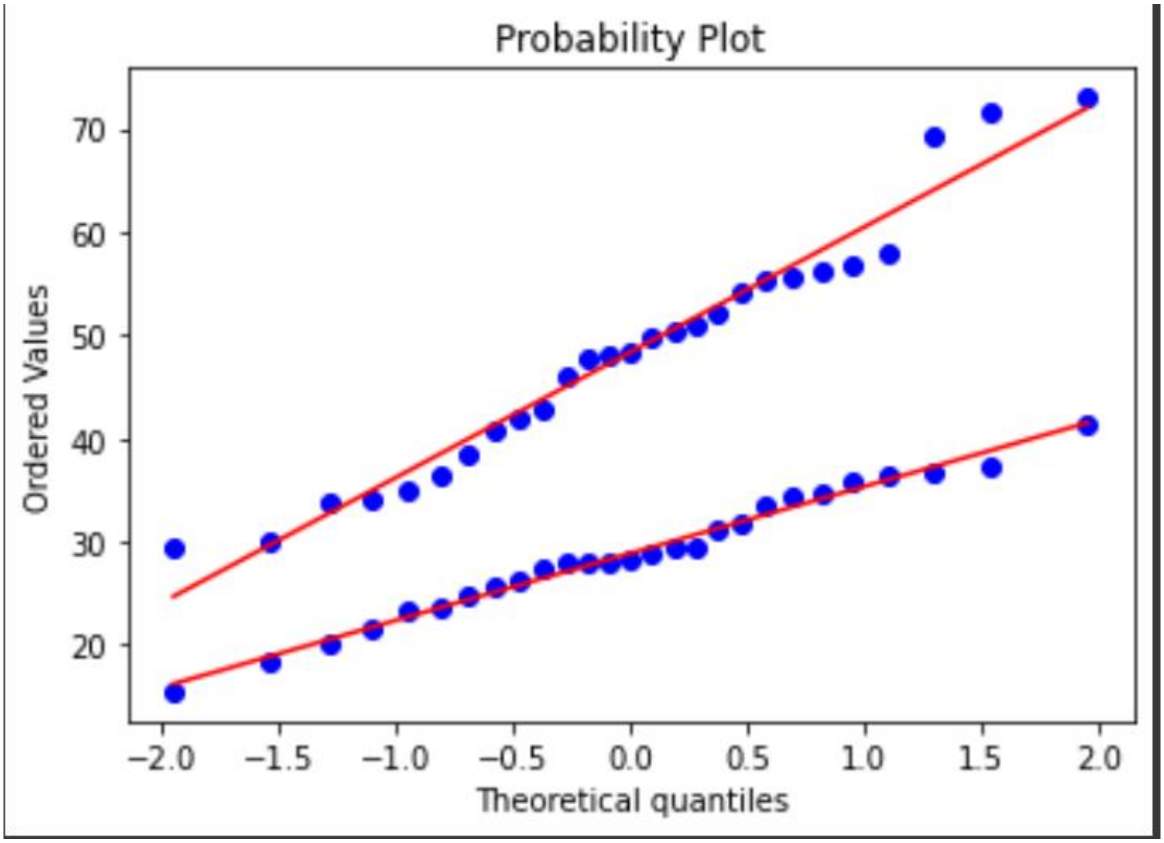
While Y-axis represents the no. of values in that range



Seeing the histogram and near 0 skewness and kurtosis along with the mean being close to median gave us the sense that the data is following a “normal distribution”

So, we created a Null Hypothesis that the data follows Normal distribution against the alternate hypothesis that the data does not follow Normal distribution

We began our testing by using a Probability Plot, (A plot used to check if the body of the sample comes from a given distribution). This gave us the following result.

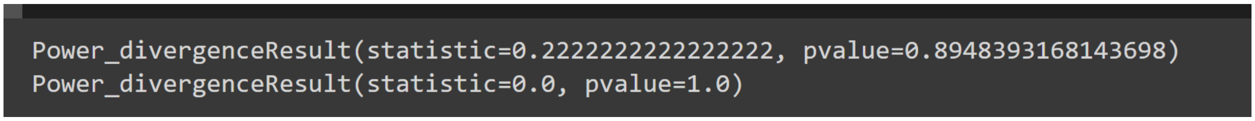


Note: Lower curve is for runners up, the higher one is for winners

Clearly, this shows that the data derived is very closely related to Normal Distribution.

Now we decide to do the chi-square test to check our null Hypothesis.

The test gave us the following result



```
Power_divergenceResult(statistic=0.2222222222222222, pvalue=0.8948393168143698)
Power_divergenceResult(statistic=0.0, pvalue=1.0)
```

p-value shows at till what % can we accept $H_0(Null\ Hypothesis)$.So we get that there is nearly a 90% confidence that winners data comes from Normal Distribution. Whereas our runners-up sample shows a 100% confidence that the data comes from Normal Distribution.

So, here are our conclusions for this seat

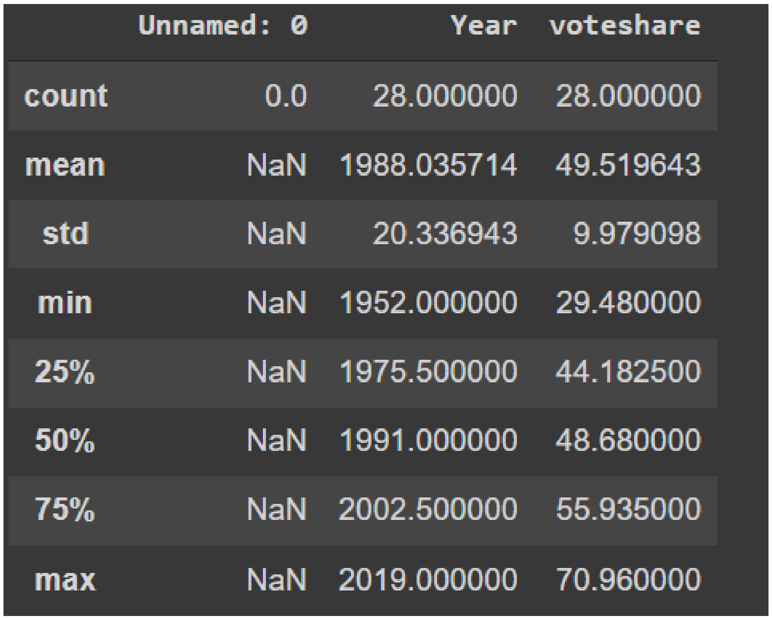
Conclusion: We can conclude from the above that winners in Lucknow (Lucknow west Vidhan Sabha seat and Lucknow Lok Sabha seat) that there is a 99.9% confidence that the runner up's voting percentage is between A =(15.5,41.3) there is 89.48% confidence that winner's vote share will be between B = (29.35, 72)

If when collecting data we have some number between A ∩ B, then we can categorize it to be as a swing seat, and if we get a % less than 15.5 or greater than 72% we need more data to conclude

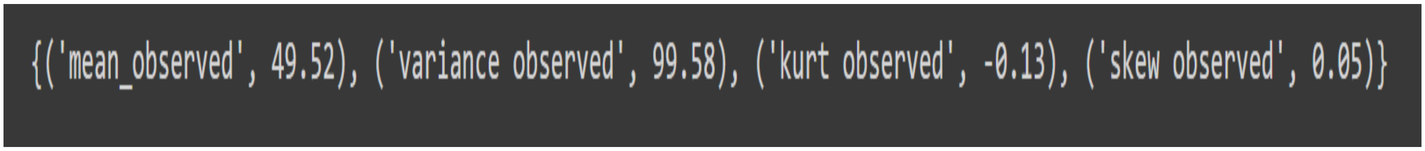
Doing similar Exercise for Kanpur, we have

For Winner,

Descriptive Statistics

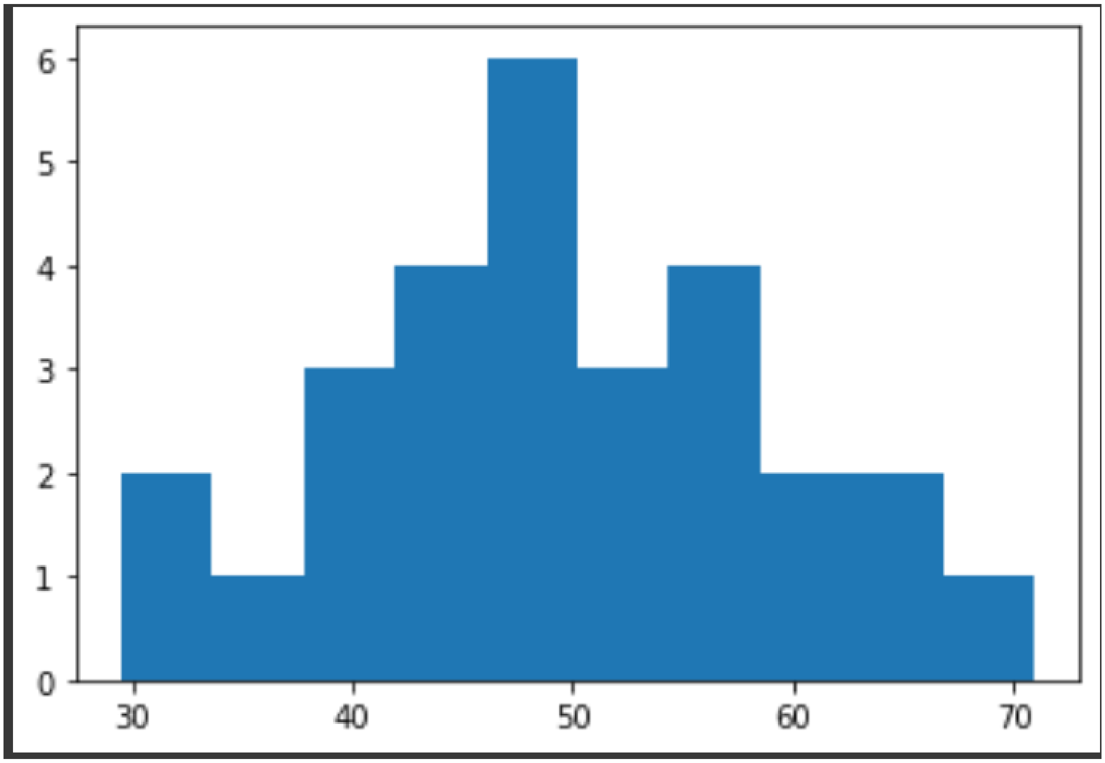


	Unnamed: 0	Year	voteshare
count	0.0	28.000000	28.000000
mean	NaN	1988.035714	49.519643
std	NaN	20.336943	9.979098
min	NaN	1952.000000	29.480000
25%	NaN	1975.500000	44.182500
50%	NaN	1991.000000	48.680000
75%	NaN	2002.500000	55.935000
max	NaN	2019.000000	70.960000



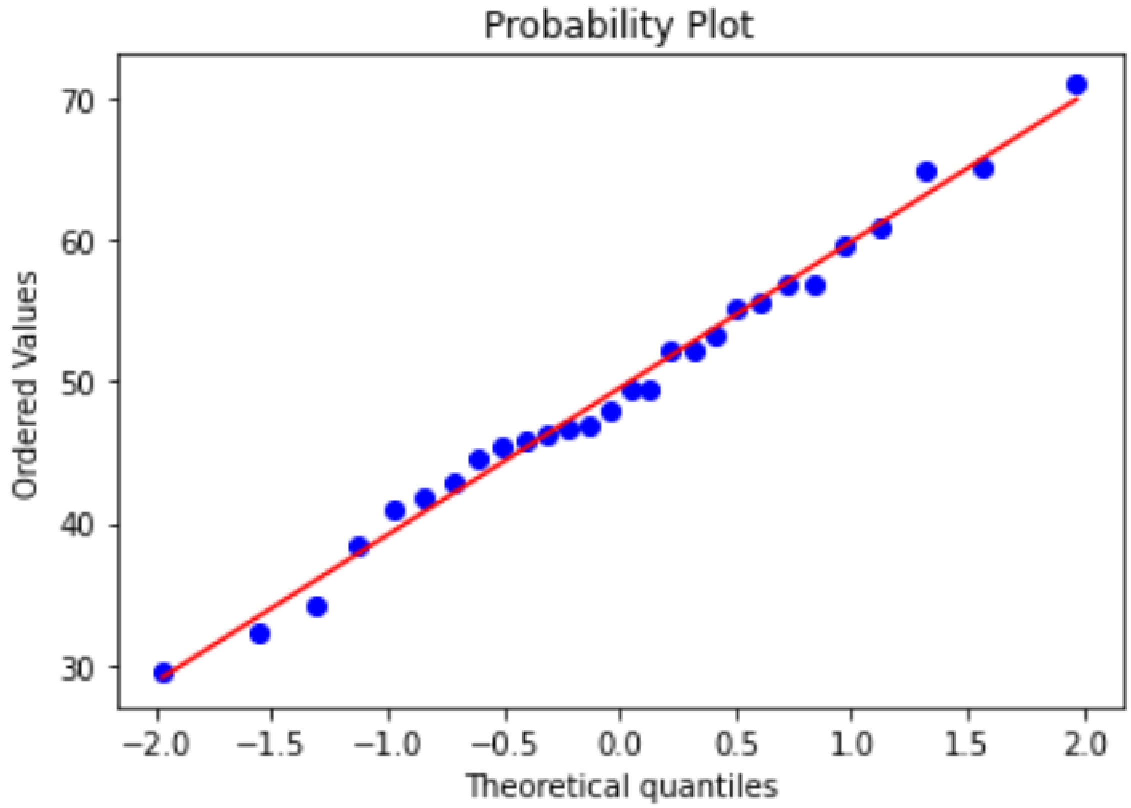
```
{('mean_observed', 49.52), ('variance observed', 99.58), ('kurt observed', -0.13), ('skew observed', 0.05)}
```

<u>Histogram</u>



Again, for this data, we have A peak in the centre and near-symmetrical buildings on sides, along with Kurtosis and Skewness being very close to 0. Therefore, again intuitively we try and fit a Normal Distribution.

Probability Plot



<u>This shows a body of the data is closely related to Normal Distribution</u>

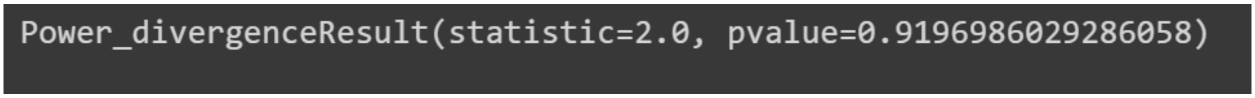
Let’s lookat the Hypothesis test

Here,

H_0: $The\ data\ comes\ from\ Normal\ Distribution$

H_1: $The\ data\ does\ not\ come\ from\ Normal\ distribution$

We have following result

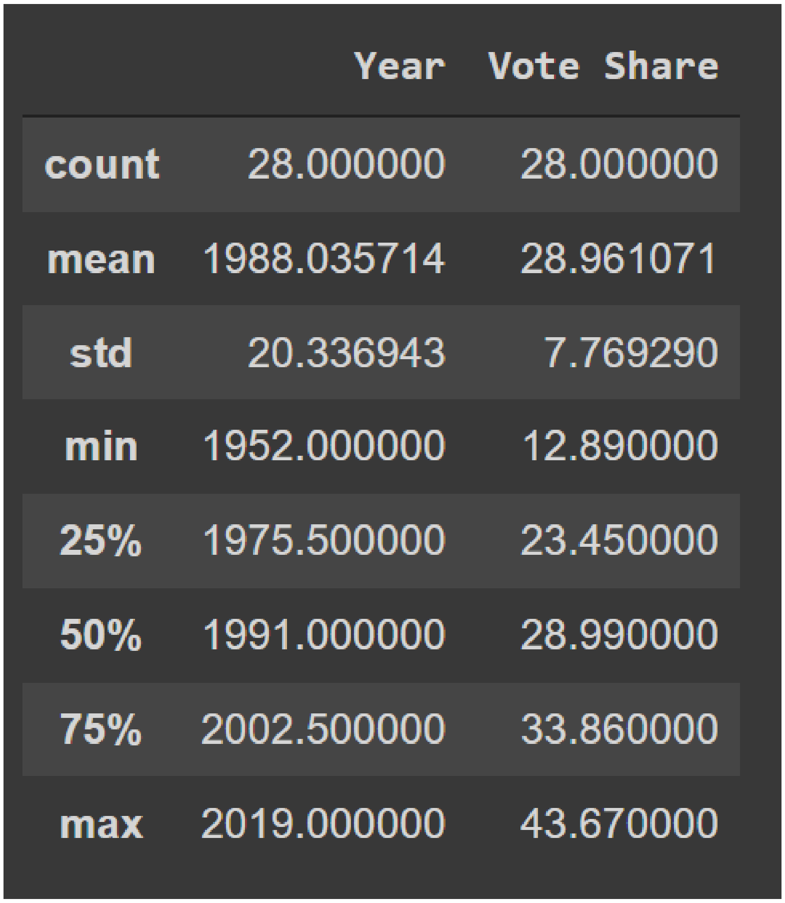


```
Power_divergenceResult(statistic=2.0, pvalue=0.9196986029286058)
```

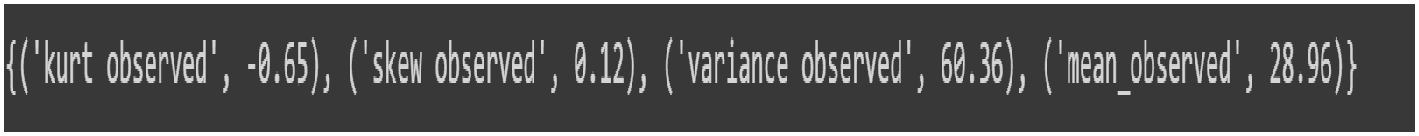
We can Conclude that winners on this seat follow Normal distribution. and there is almost 92% confidence that a winner in Kanpur will get between (29.4,70.9)

<u>Now lets look at the Runners Up</u>

<u>Descriptive Statistics</u>

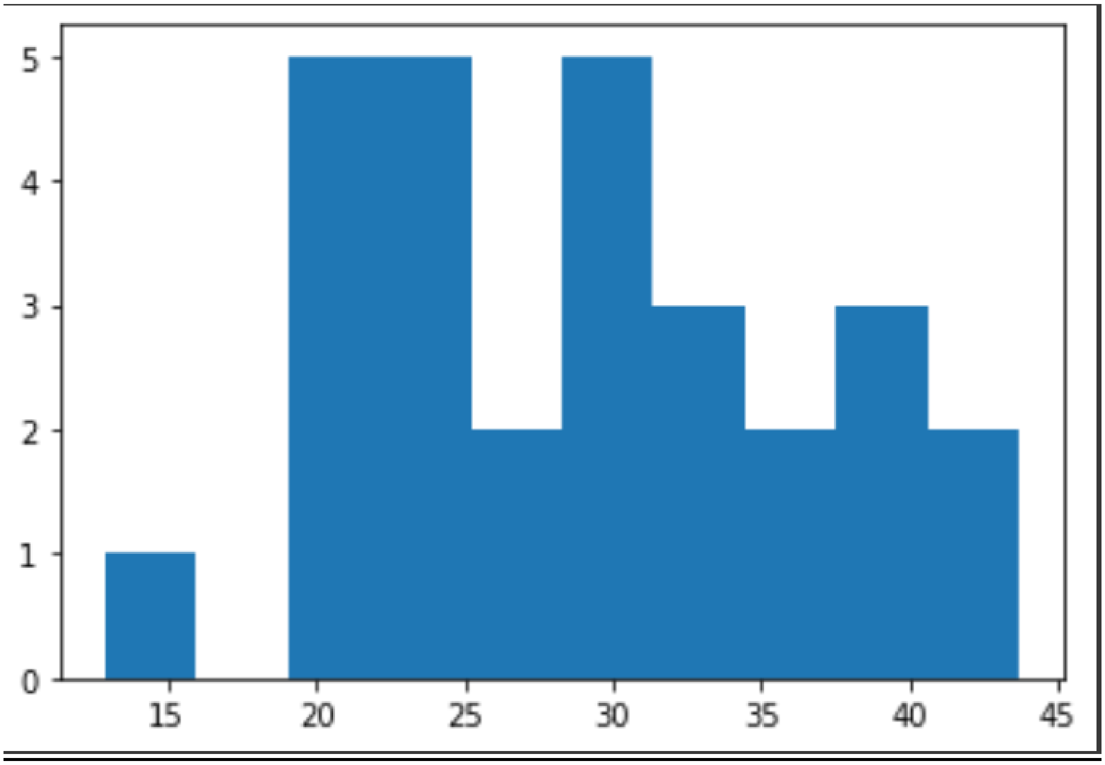


	Year	Vote Share
count	28.000000	28.000000
mean	1988.035714	28.961071
std	20.336943	7.769290
min	1952.000000	12.890000
25%	1975.500000	23.450000
50%	1991.000000	28.990000
75%	2002.500000	33.860000
max	2019.000000	43.670000



```
{('kurt observed', -0.65), ('skew observed', 0.12), ('variance observed', 60.36), ('mean_observed', 28.96)}
```

HISTOGRAM



Now, here Though the skewness is close to 0 Kurtosis is closer to -1, meaning a right-tailed distribution, also the histogram verifies this claim.

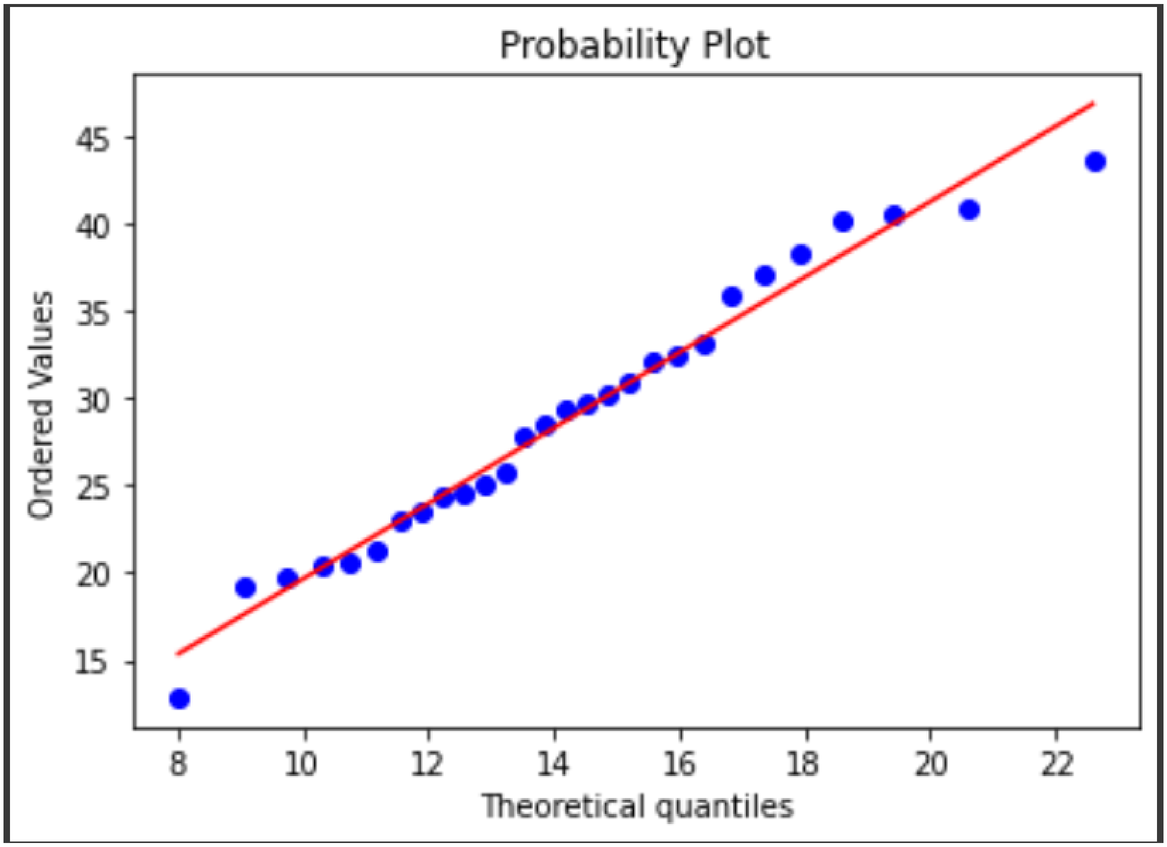
So, intuitively this sample doesn't follow Normal. So which distribution can it follow?

Lets look at Gamma.

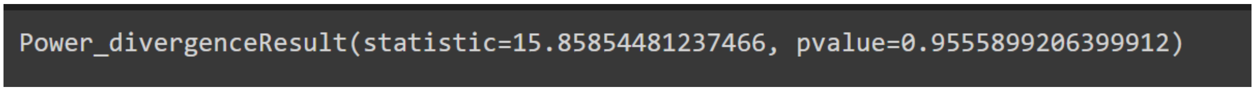
Creating

$H_0: The\ data\ comes\ from\ Gamma\ \ Distribution$

$H_1: The\ data\ does\ not\ come\ from\ Gamma\ \ distribution$



The Probability plot Gives us some solace that the data might come from Gamma distribution



```
Power_divergenceResult(statistic=15.85854481237466, pvalue=0.9555899206399912)
```

Since p val = 0.955 we fail to reject the null hypothesis at a 95% confidence interval. Therefore this data corresponds to gamma. That is in the future we have a 95% CI that runner up will have values b/w (12.89,43.67)

CONCLUSION

Given our assumptions, we can fit a theoretical distribution to election data, (at least on given data)

There is no 1 fit for all here, as no 1 distribution fits all the data here.

The usefulness of this data to the psephologist is that when she/he sees data of a particular seat, she/he can conclude what are the chances of winner or runner-up getting this amount of vote

Bibliography

https://eci.gov.in/

https://en.wikipedia.org/wiki/Kanpur_(Lok_Sabha_constituency)

https://en.wikipedia.org/wiki/Lucknow_(Lok_Sabha_constituency)

https://towardsdatascience.com/sql-pandas-or-both-analysing-the-uk-electoral-system-24fa01d33d05

Link to the Python Codes

Kanpur Winner

https://colab.research.google.com/drive/16gMe_dN0atD1Jnu3gSxkj7M9C8cpf4ub

Kanpur Runners up (Gamma Distribution)

https://colab.research.google.com/drive/1sjB7Jgc-9PS0Gn-jjO9frP-VnAh9OtpG

Lucknow Results

https://colab.research.google.com/drive/1A47od5BYk_v0NPtc1HJtYJ6TzKMFq-P-

Note Python files cannot be run as there is no excel file attached here