Method1: Trace driven simulation:
Let us say we have a monthly sales volume for the last three years which essentially means that I have 36 values in my data-set. So, instead of first fitting a distribution to the 36 values and then using the distribution in my further analysis I can directly use these 36 values in my analysis. So, if I want to simulate, I will simulate directly using these 36 values, this is generally called trace driven simulation.

Method2: Theoretical distributions:
What do you mean by theoretical distribution, theoretical distribution are normal distribution, uniform distribution, binomial distribution for discrete, Poisson distribution for discrete, exponential distribution for continuous these are all theoretical distributions.

Method 3: Empirical distribution: building a distribution from our collected data. Here, we are not fitting our collected data to a distribution, but building a distribution from our collected data.

Essential building blocks:

- define the density/distribution functions.
- Estimate the parameters (mean, median, std etc)

Empirical distribution for ungrouped data:

For ungrouped data:

Let $X_{(i)}$ denote the ith smallest of the X_j's so that: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

$$F(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)\left(X_{(i+1)} - X_{(i)}\right)} & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for } i = 1,2,\cdots,n-1 \\ 1 & \text{if } X_{(n)} \le x \end{cases}$$

Empirical distribution for grouped data:

For grouped data:

- Suppose that n X_j's are grouped in k adjacent intervals $[a_0,a_1)$, $[a_1,a_2)$,...$[a_{k-1},a_k)$ so that jth interval contains n_j observations. $n_1 + n_2 + \ldots n_k = n$.
- Let a piecewise linear function G be such that $G(a_0) = 0$, $G(a_j) = (n_1 + n_2 + \ldots n_j)/n$, then:

$$G(x) = \begin{cases} 0 & \text{if } x < a_0 \\ G\left(a_{j-1}\right) + \dfrac{x - a_{j-1}}{a_j - a_{j-1}}\left[G\left(a_j\right) - G\left(a_{j-1}\right)\right] & \text{if } a_{j-1} \le x < a_j, j = 1,2,\cdots k \\ 1 & \text{if } a_k \le x. \end{cases}$$

How to guess the distribution for method 2:

Clues for summary statistics:

- For the symmetric distributions mean and median should match. In the sample data, if these values are sufficiently close to each other, we can think of a symmetric distribution (e.g. normal).
- Coefficient of variation (*cv*): (ratio of std dev and the mean) for continuous distributions. The *cv = 1* for exponential dist. If the histogram looks like a slightly right-skewed curve with *cv >1*, then lognormal could be better approximation of the distribution.

Note: For many distributions *cv* may not even be properly defined. When? Examples?

CV = std / mean for standard normal distribution mean = 0 in which case we cannot use cv for analysis

- Lexis ratio: same as *cv* for discrete distributions.
- Skewness (*v*): measure of symmetry of a distribution. For normal dist. $v = 0$. For $v>0$, the distribution is skewed towards right (exponential dist, $v = 2$). And for $v<0$, the distribution is skewed towards left.

Next is parameter estimation: after going through the summary stats of our dataset, representing the data with boxplot, bar chart etc we conclude our dataset to a distribution and then try and find the distributions parameter for our dataset.

- Once distribution is guessed, the next step is estimating the parameters of the distribution.

- Each distribution has a set of parameters.
- ✓ Normal distribution has mean and standard deviation
- ✓ Exponential distribution has a "λ".

- Most common method of parameter estimation: MLE (What is this?)

Next is how to know the fitted distribution is good enough?

- It can be checked by several methods:
  1. Frequency comparison (a bit technical)
  2. Probability plots (visual tool)
  3. Goodness-of-fit tests (statistical test of goodness. Very widely used).

Probability plots: PP Plot:

- assuming we have two distributions (f and g) and a point of evaluation z (any value), the point on the plot indicates what percentage of data lies at or below z in both f and g (as per definition of the CDF).

- If the points deviate from the 45-degree line then the distributions deviate

- P-P plots are well suited to compare regions of high probability density (centre of distribution) because in these regions the empirical and theoretical CDFs change more rapidly than in regions of low probability density.

- P-P plots can be used to visually evaluate the skewness of a distribution.

- P-P plots are most useful when comparing probability distributions that have a nearby or equal location.

QQ Plot:

Similarly, to P-P plots, Q-Q (quantile-quantile) plots allow us to compare distributions by plotting their quantiles against each other.

Some key information on Q-Q plots:

- a point on the chart corresponds to a certain quantile coming from both distributions (again in most cases empirical and theoretical).

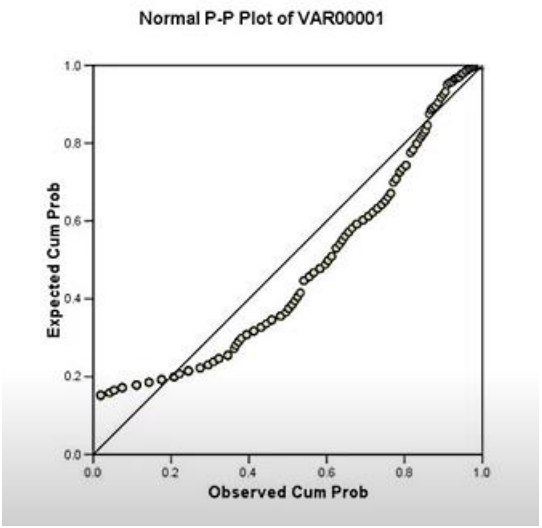
- If the points deviate from the 45-degree line then the distributions deviate

- Q-Q plot gets very good resolution at the tails of the distribution but worse in the centre (where probability density is high)

- Q-Q plots can be used to visually evaluate the similarity of location, scale, and skewness of the two distributions.

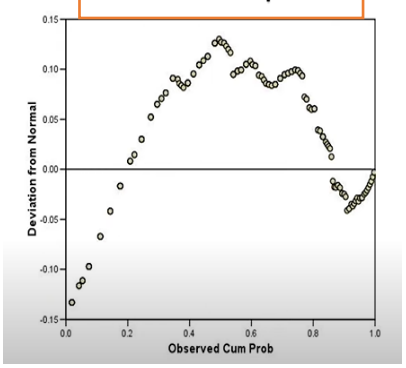
- The *Q-Q* plot will amplify the differences between the tails of the model distribution and the sample distribution.
- Whereas, the *P-P* plot will amplify the differences at the middle portion of the model and sample distribution.

Example:



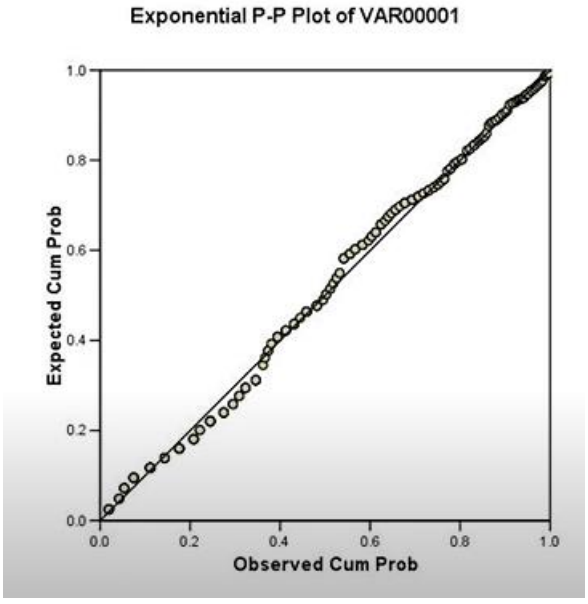
this pp plot diagram shows that the normal distn is not really a good fit for the dataset

Residual plot

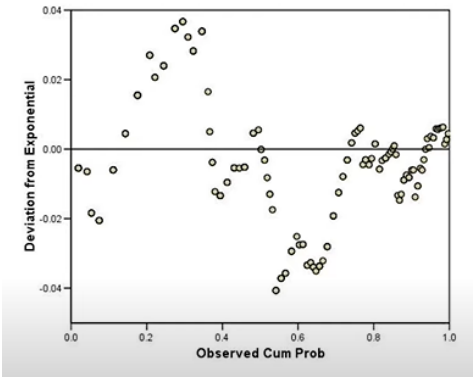


also, the points seem to deviate in the middle portion and the left portion from the normal. The y axis runs from −0.15 to 0.15

deviation from the normal? Means that the data points that differ from the central tendency (mean, median, mode)

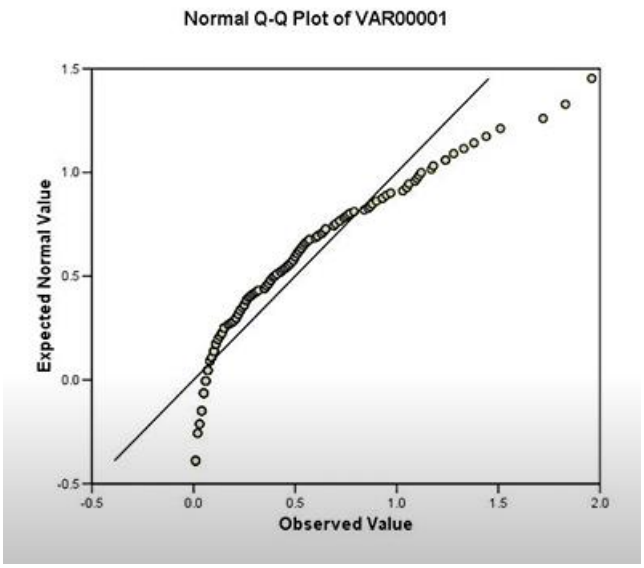


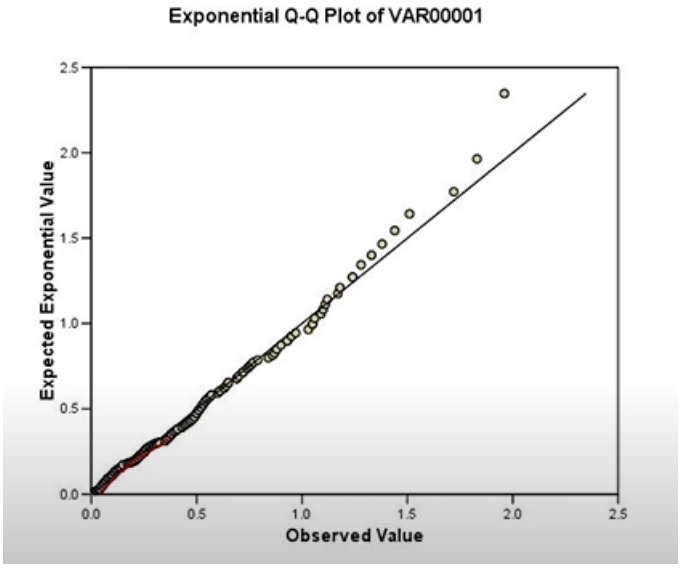
this pp plot diagram shows that the exponential distn is really a good fit for the dataset



here the deviation from exponential may look large but the y axis runs from −0.04 to 0.04 which is very low.

Example:





clearly exponential is a better fit according to qq plot too.

Goodness-of-fit test:

- A goodness-of-fit test is a statistical hypothesis test that is used to assess formally whether the observations $X_1, X_2, X_3 \ldots X_n$ are an independent sample from a particular distribution with function $F^{\wedge}$.

H_0: The X_i's are IID random variables with distribution function $F^{\wedge}$.

- Two famous tests:

1. Chi-square test
2. Kolmogorov - Smirnov test

Right skewed:

Mean>median>mode.

Left skewed:

Mode<median<mean.

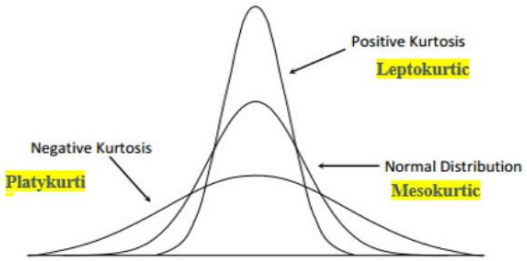
If the skewness is between −0.5 & 0.5, the data are nearly symmetrical.

If the skewness is between −1 & −0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.

If the skewness is lower than −1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.

Kurtosis:

- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called mesokurtic.
- A distribution with kurtosis <3 (excess kurtosis <0) is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- A distribution with kurtosis >3 (excess kurtosis >0) is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.



Chi-square tests: helps us to find out if two variables are related to each other or not

Chi-square tests are based on a null/alternative hypothesis.

Null hypothesis: the given data follows xyz distribution.

Alternative hypothesis: the given data does not follow xyz distribution.

null hypothesis is true when:

The tabulated chi-squared statistic (scipy.stats.chi2.ppf) > calculated value (stats.chisquare(obs_freq, expected_freq))

Also,

as the calculated value approaches zero, there is more evidence that the null hypothesis is true.

Also,

In terms of p-value, if p-value is greater than (1 − confidence level = alpha), we see that the null hypothesis is true

In the stats library of scipy, we can call for two chi-square test command. one is chi2_contingency and another is chisquare.But which one to use when?

We use chi-square to when we want to find any relation between two categorical groups. For example, there are is a gender variable (with males and female) and a

mode of travel variable (Public transport and own vehicle). Now, there might be a relation among them say, men travel more in public transport than their own vehicle or female prefer public transport than their own vehicle. To check that statistically, we use `chi2_contingency`.The null hypothesis is: two groups have no significant difference. We get the chi-square statistic, p-value, degrees of freedom, and the table of expected observation. Now if the p-value is greater than our specified alpha (or threshold value) we accept the null, concluding there is no statistically significant difference between two groups.

On the other hand, `chisquare` is used when we want to see if a set of discreet random variable is distributed evenly or not. If the data is spread evenly or not is checked by computing a expected frequency of the selected discreet random variable and then it is tallied with the original frequency distribution. How well the data is fitted (Goodness of Fit), is checked. Here the null hypothesis is: the discreet random variable is not significantly different from the expected distribution. Unlike `chi2_contingency`, `chisquare`returns chi-square statistic and the p-value. To look at the expected distribution (by the algorithm) call ‘expected_freq’ from ‘scipy.stats’ and pass the observed value table in it. If a data frame contains 5 (categorical) columns is passed, then five chi-square statistic and five corresponding p-values will be generated, in the order, the columns are in the data frame