STATISTICS FOR DATA SCIENCE

Normal Probability Plots

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Normal Probability Plots

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

STATISTICS FOR DATA SCIENCE

Topics to be covered...

- ✓The Normal Probability Plot.
- ✓Understanding Q-Q Plot.
- ✓Interpreting the Probability Plots.

What are Probability Plots basically mean?

- So far, we have always worked with randomly selected samples from some population.
- We have used an appropriate probability distribution to fit in the data accordingly.
- The probability plot is one way of accessing it through graphical representation.
- By visualizing the data, we can achieve tremendous amount of information.
- For instance our data may be skewed, or be bi-modal, and typically determines the distribution from which population it has come from.

Normal Probability Plots

- This addresses the following queries,

  a) Can we conclude that the data is normally distributed?

  b) It also shows whether the data is skewed or having longer or shorter tails.

- The data that is been plotted in the theoretical normal distribution should form a straight line. This denotes the normality of the data.

- A straight diagonal line depicts that the data is normally distributed.

- Identifies whether the data is skewed to left or right which does not fit the normal distribution.

How can I claim that my data is normally distributed?

For larger samples,

- Histogram will have a bell shaped curve which we call as symmetric and there will not be any outliers.

- The mean, median and mode will be similar and lie at the same point.

- In the similar way, 68% of observations lies within one standard deviation of the mean. 95% within two and 99.7% with three standard deviations.

For small samples,

- Histogram does not provides good visual presence, hence to conform its normality we can use Probability Plots.

Construction of a Probability Plot

1) Sort the data.
2) Assign evenly spaced values to the data between 0 and 1.
3) For each x_i in the data set,

$$\frac{(i-0.5)}{n}$$

Where,
i is the position of the data item
n is the size of the data set.

4) Find theoretical quantiles - Q_i.
5) Plot every point (x_i , Q_i).
6) Plot (x_i , x_i)
7) Look into the observation whether it forms approximately straight line. This helps us to understand the type of distribution.

Different ways of dividing data equally between 0 and 1.

Methods	Plotting Position Method
Blom	$(i-0.375)/(n+0.25)$
Benard	$(i-0.3)/(n+0.4)$
Hazen	$(i-0.5)/n$
Van der Waerden	$i/(n+1)$
Kaplan-Meier	i/n

Problem – Normal Probability Plot

Problem:

Construct a normal probability plot for the following data. Do these data appear to come from an approximately normal distribution?

3.01, 3.35, 4.79, 5.96, 7.89.

Solution:

Sort the values

3.01, 3.35, 4.79, 5.96, 7.89 / n = 5

Solution - Find the Plotting Position using Hazen Method

i	X_i	$\frac{(i-0.5)}{5}$
1	3.01	0.1
2	3.35	0.3
3	4.79	0.5
4	5.96	0.7
5	7.89	0.9

The value $(i-0.5)/n$ is chosen to reflect the position of X_i in the ordered sample.

There are $i-1$ values less than X_i , and i values less than or equal to X_i .

The quantity $(i-0.5)/n$ is a compromise between the proportions $i-1/n$ and i/n.

The distribution that the sample come from is $N(5,2^2)$

Solution - Understanding behind Normal Probability Plot

- From the plot we can infer that (X_1, 0.1) intersects at the point (Q_1, 0.1). We understand that Q_1 is at the 10^{th} percentile of the $N(5,2^2)$ distribution.

- Applying similar reasoning to the remaining points, we would expect each Q_i to be close to its corresponding X_i by 20^{th}, 30^{th}, 40^{th} and so on.

- The **probability plot consists of the points (X_i, Q_i).**

- **Since the distribution that** generated the Q_i was a normal distribution, this is called a **normal probability plot.**

Solution - Understanding behind Normal Probability Plot

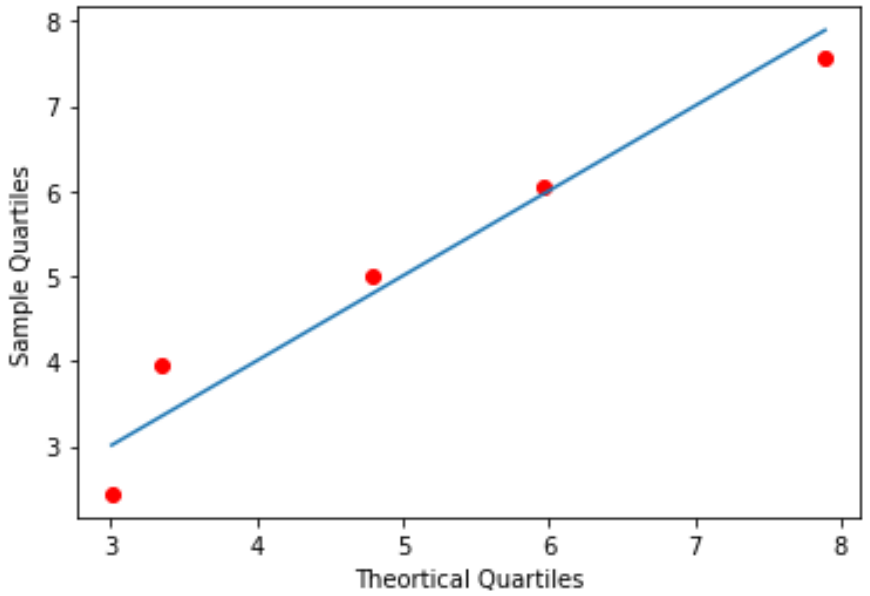
- If $X_1, \ldots, X_n$ do in fact come from the distribution that generated the Qi , the points should lie close to a straight line.
- To construct the plot, we must compute the Q_i.
- These are the 100(i − 0.5)/n percentiles of the distribution that is suspected of generating the sample.
- In this example the Qi are the 10th, 30th, 50th, 70th, and 90th percentiles of the $N(5, 2^2)$ distribution.
- We could approximate these values by looking up the z-scores corresponding to these percentiles, and then converting to raw scores.

Solution - Find the Theoretical Quartiles Q_i

i	X_i	$\frac{(i-0.5)}{5}$	Closest Area in z - Table	Z-score	(Q_i) X = z * σ + μ
1	3.01	0.1	0.1003	-1.28	-1.28 * 2 + 5 = 2.44
2	3.35	0.3	0.3015	-0.52	-0.52 * 2 + 5 = 3.95
3	4.79	0.5	0.5000	0.00	0.00 * 2 + 5 = 5.00
4	5.96	0.7	0.6985	0.52	0.52 * 2 + 5 = 6.05
5	7.89	0.9	0.8997	1.28	1.28 * 2 + 5 = 7.56

Solution - Normal Probability Plot

- The figure shows a normal probability plot for the sample $X_1, \ldots, X_5$.
- A straight line is superimposed on the plot, to make it easier to judge whether the points lie close to a straight line or not.
- The sample points are close to the line, so it is quite plausible that the sample came from a normal distribution.
- The sample points $X_1, \ldots, X_n$ are called empirical quantiles.



Q-Q Plot

- The points $Q_1, \ldots, Q_n$ are **called quantiles (divides distribution into equal sized areas)** of the distribution.
- These are the points in the data below which a certain proportion of the data falls.
- The probability plot is sometimes called a quantile–quantile plot, or QQ plot.
- We can use this Q-Q plot to check the assumption of Normality of the data.
- Determines whether if two set of quantiles come from the populations of same distribution. If, yes roughly forms a straight line.

How to use Probability plots for different sample sizes?

- Probability plots work better with larger samples. A good rule of thumb is to require at least 30 points before relying on a probability plot.

- Probability plots can still be used for smaller samples, but they will detect only fairly large departures from normality.

STATISTICS FOR DATA SCIENCE

Interpreting Probability Plots

- It's best not to use hard-and-fast rules when interpreting a probability plot. Judge the straightness of the plot by eye.

- When deciding whether the points on a probability plot lie close to a straight line or not, do not pay too much attention to the points at the very ends (high or low) of the sample, unless they are quite far from the line.

- It is common for a few points at either end to stray from the line somewhat.

- However, a point that is very far from the line when most other points are close is an outlier, and deserves attention.

THANK YOU

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering