



PROBABILITY PLOTS

D. Uma

Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Normal Probability Plot

D. Uma

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Topics to be covered...

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




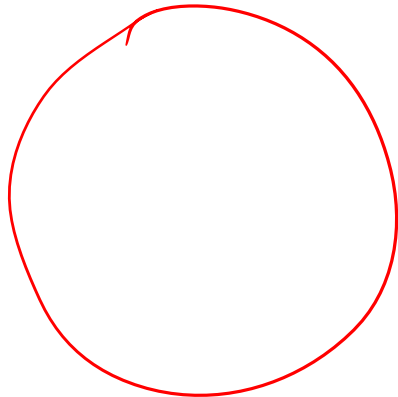
How do you know data is normally distributed?

Mean =
Median =
Mode =

} \Rightarrow Data values follow Normal Distribution

$N > 30$
 $N = 10$

sample 

Population 

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.

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 / $n = 5$

STATISTICS FOR DATA SCIENCE

Problem – Normal Probability Plot

Do these data appear to come from an approximately normal distribution?

$$\bar{x} = 5, s = 2$$

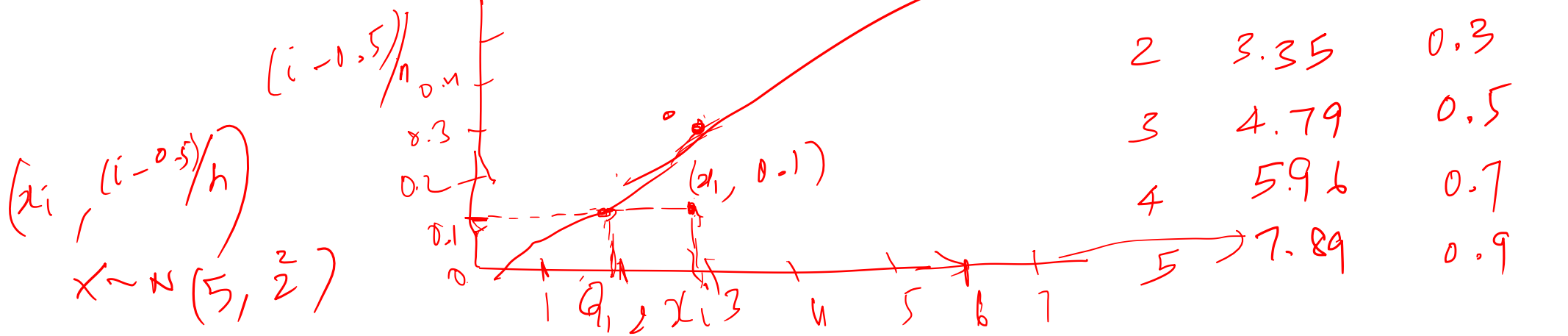
1 2 3 4 5
3.01, 3.35, 4.79, 5.96, 7.89.
 x_1, x_2, x_3, x_4, x_5

$$n = 5, x \sim N$$

Sample
 x_1, x_2, x_3, x_4, x_5

1. Sort the values

2. Assign ordered rank for each x_i



STATISTICS FOR DATA SCIENCE

Solution - Find the Theoretical Quartiles Q_i

$$\bar{x} = 5$$
$$s = 2$$

i	X_i	$\frac{(i - 0.5)}{5}$	Closest Area in z - Table	Z-score	$(Q_i) \checkmark$ $X = z * \sigma + \mu$
1	3.01	0.1	0.1003	-1.28	$-1.28 * 2 + 5 =$
2	3.35	0.3			
3	4.79	0.5			
4	5.96	0.7			
5	7.89	0.9			

Q_1

STATISTICS FOR DATA SCIENCE

Solution - Find the Theoretical Quartiles Q_i

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

x_i

Q_1

Q_2

Q_3

Q_4

Q_5

(x_1, Q_1)
 (x_2, Q_2) - - - (x_5, Q_5)

(Q_i, x_i)
 (x_i, x_i)

- 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,

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

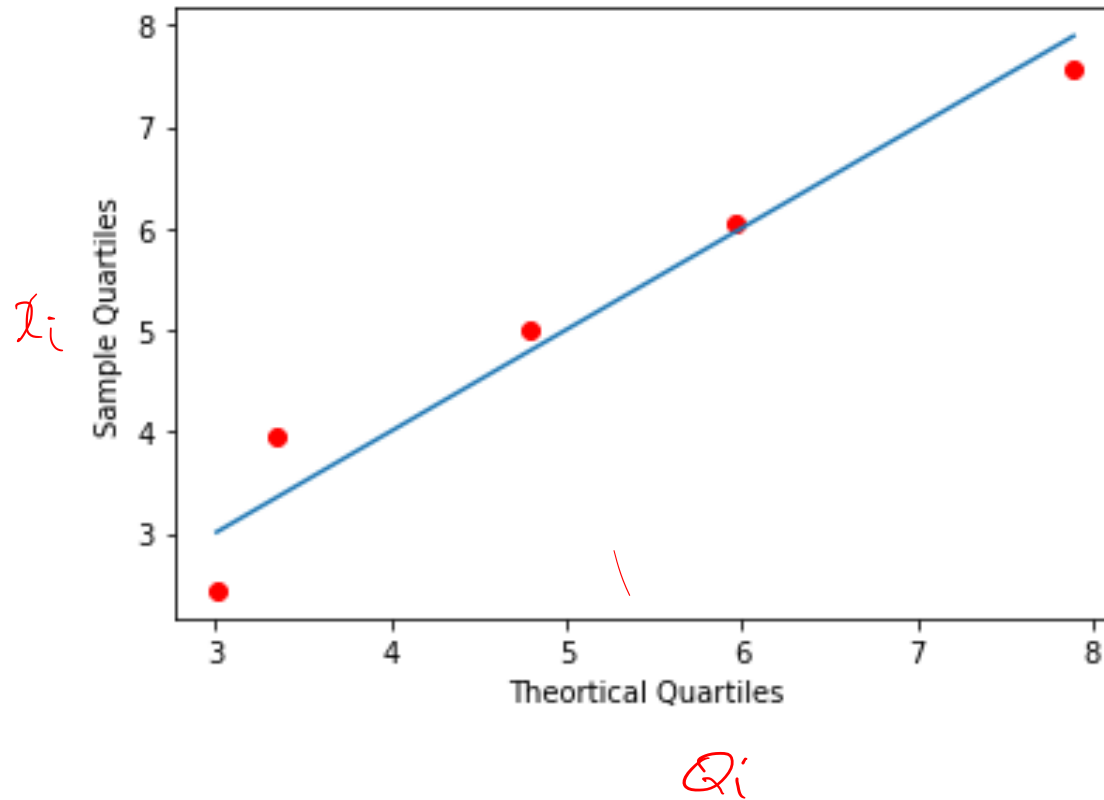
Where,

i is the position of the data item

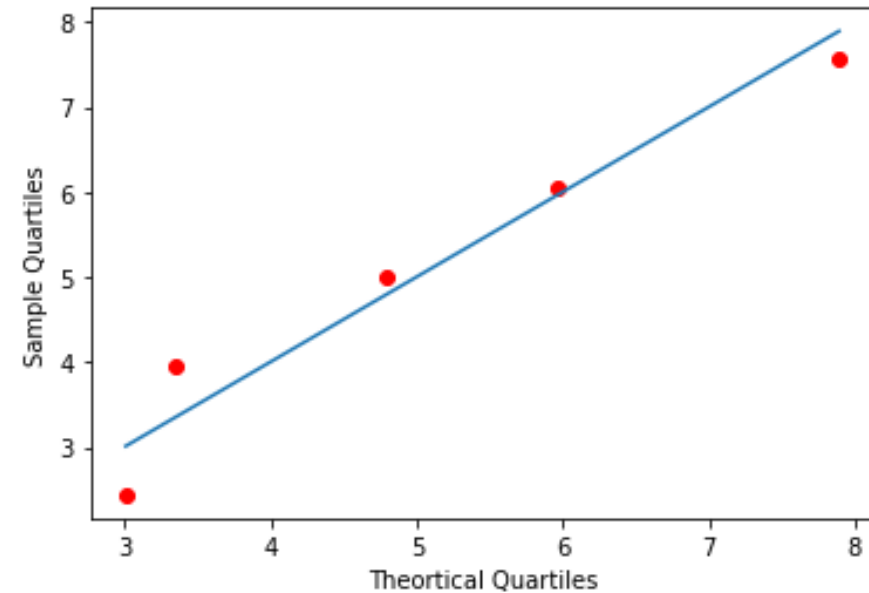
n is the size of the data set.

- 4) Find theoretical quantiles - Q_i .
- 5) Plot every point (x_i, Q_i) or (Q_i, x_i)

Note: Look into the observation whether it forms approximately straight line. This helps us to understand the type of distribution.



- The figure shows a normal probability plot for the sample X_1, \dots, 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, \dots, X_n are called empirical quantiles.



Q-Q Plot



- The points Q_1, \dots, 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.

$$X_i : \frac{i-1}{n} \text{ or } \frac{i}{n}$$

$$Q_1 = 0.1 \rightarrow 10^{\text{th}} \text{ percentile}$$

$$Q_3 = 0.3 \rightarrow 30^{\text{th}} \\ N(5, 2^2)$$

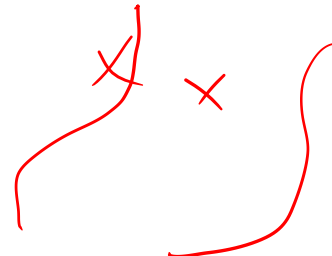
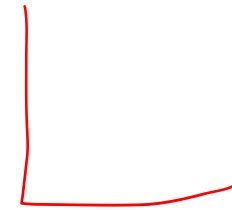
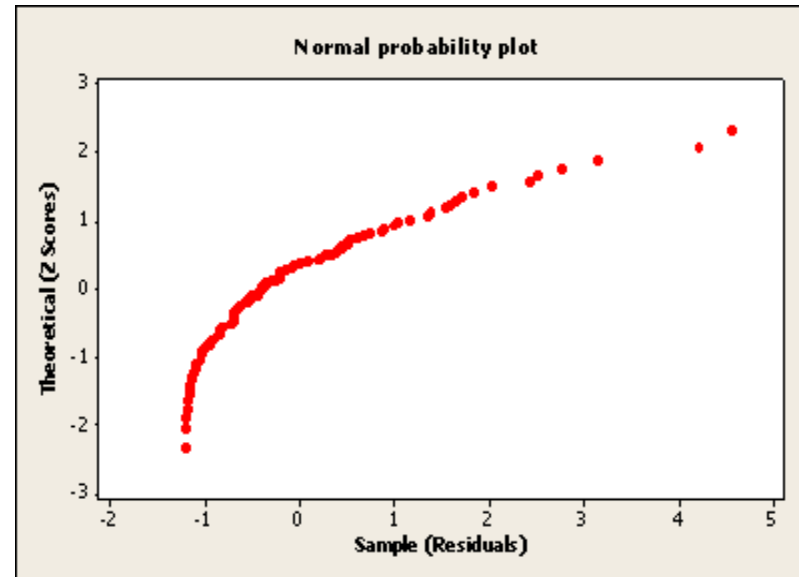
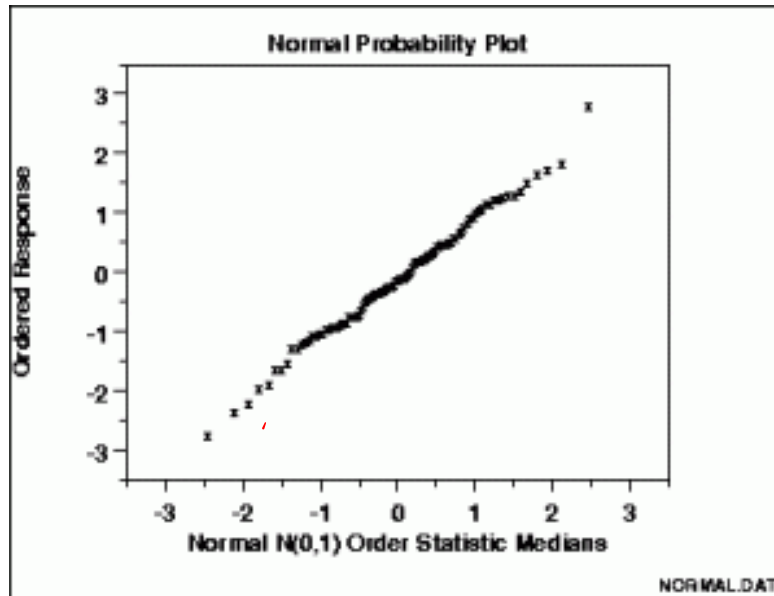
$$Q_5 = 50^{\text{th}} -$$

STATISTICS FOR DATA SCIENCE

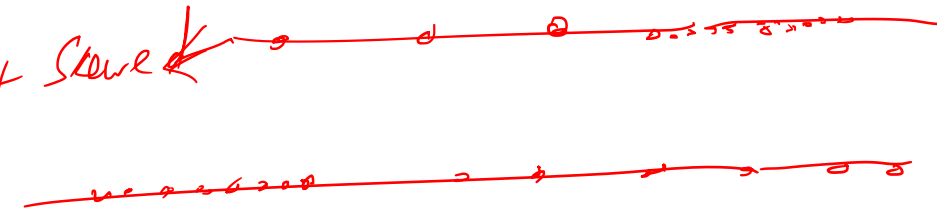
Interpreting Probability Plots



PES
UNIVERSITY
ONLINE



left skewed



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

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 10th 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 20th, 30th, 40th 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, \dots, X_n do in fact come from the distribution that generated the Q_i , 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 Q_i 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.

- 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.

- 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.

Why Probability Plots?

- 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.

- 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.



THANK YOU

D. Uma

Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335