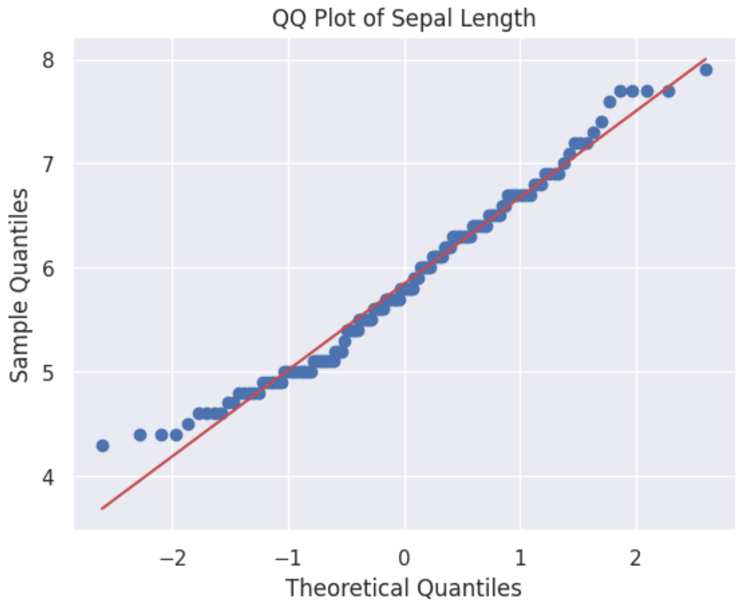
# How to interpret QQ plot for Univariate analysis



## A. Interpretation of the QQ Plot Components:

- **Horizontal Axis (X-axis):** Represents the **theoretical quantiles** of a chosen theoretical distribution, which is typically a standard normal distribution (mean=0, standard deviation=1). These are the expected values if your data followed that theoretical distribution perfectly.
- **Vertical Axis (Y-axis):** Represents the **sample quantiles** of your actual "Sepal Length" data. These are the ordered values from your dataset plotted at their corresponding quantile positions.
- **The Red Line:** This is a **reference line** that passes through the first and third quartiles of both the theoretical and sample distributions. If your data perfectly follows the chosen theoretical distribution, all the blue points should fall directly along this line.
- **The Blue Points:** Each blue point represents a data point from your "Sepal Length" data. Its x-coordinate is the theoretical quantile corresponding to its rank in the ordered data, and its y-coordinate is the actual value of that data point.

### B. Interpreting the QQ Plot for "Sepal Length":

By examining how the blue points deviate from the red line, we can assess how well the "Sepal Length" data fits a normal distribution (since the x-axis represents theoretical normal quantiles):

- **Overall Linearity:** If the blue points form a roughly straight line that closely follows the red reference line, it suggests that the sample data is approximately normally distributed.
- **Deviations from Linearity:** Deviations from the straight line indicate departures from normality:
    - **S-shaped curve:** Suggests that the tails of the sample distribution are heavier or lighter than the tails of a normal distribution. An upward curve at both ends indicates heavier tails (more extreme values), while a downward curve at both ends indicates lighter tails (fewer extreme values).
    - **Curvature at one end:** Suggests skewness. A curve that is convex upwards indicates negative skew (tail extending to the left), and a curve that is concave upwards indicates positive skew (tail extending to the right).
    - **Steps or plateaus:** Can indicate that the data is discrete or has many repeated values.

### Diagram analysis:

- The blue points generally follow the red line, suggesting that the "Sepal Length" data is **approximately normally distributed**.
- However, there are some noticeable deviations, particularly at the tails:
    - At the lower tail (left side of the plot), the points seem to fall slightly below the line, which might suggest a slightly lighter left tail than a perfect normal distribution.
    - At the upper tail (right side of the plot), the points also deviate slightly from the line, indicating some departure from perfect normality in the higher sepal length values. The curve seems to

bend slightly upwards, hinting at a potentially slightly heavier right tail.

- The central portion of the data follows the line more closely, indicating that the middle values are quite consistent with a normal distribution.

### QQ plots are the best choice when you want to:

- **Assess if a dataset follows a specific theoretical distribution:** Primarily used to check for normality, but can also be used to compare data against other distributions (e.g., exponential, Weibull) by changing the theoretical quantiles on the x-axis.
- **Visually identify departures from a theoretical distribution:** QQ plots make it easier to spot patterns of deviation (skewness, heavy/light tails, non-normality) than simply looking at histograms or summary statistics.
- **Compare the shapes of two distributions:** You can create a QQ plot comparing the quantiles of two different datasets to see if they have similar distributions, even if their means or variances differ.
- **Diagnose the assumptions of statistical models:** Many statistical tests and models assume that the residuals (the differences between the observed and predicted values) are normally distributed. QQ plots of the residuals are a common way to check this assumption.
- **Provide a more sensitive test for distributional assumptions than histograms:** While histograms give a general sense of the distribution's shape, QQ plots can reveal subtle deviations from the assumed distribution that might not be apparent in a histogram.

### In contrast to histograms, box plots, and density plots:

- **Histograms:** Show the frequency distribution but don't directly compare the data's quantiles to a theoretical distribution.
- **Box Plots:** Summarize the distribution through quartiles and outliers but don't provide a detailed assessment of how well the data fits a specific theoretical distribution across all quantiles.
- **Density Plots:** Show the estimated probability density function but don't directly compare quantiles to a theoretical distribution in the same way as a QQ plot.

In summary, QQ plots are the best choice when the primary goal is to **formally or informally assess how well a dataset conforms to a specific theoretical distribution**, especially normality, and to **diagnose the nature of any deviations**. They provide a quantile-based comparison that can reveal subtle distributional characteristics.