Motivation

In the field of statistics, the assumption of normality is important because many statistical techniques perform calculations assuming that the data is normally distributed. The techniques that assume that the data is Gaussian or Gaussian-like distributed includes, but not limited to:

Techniques That Assumes Normality

1. Hypothesis testing through t-test and z-test
2. Analysis of variance (ANOVA)
3. Sequential Gaussian simulation in spatial analysis
4. Control limits in control chart

Unfortunately, many real-life data are not normal. Permeability distribution of rock samples is lognormal. Time required to repair a malfunctioning component follows exponential distribution, and reliability analysis for machine performance with respect to time follows Weibull distribution. What should you do if our data fails a normality test, or is not Gaussian-like?

You have two options:

1. Use it as it is or fit non-normal distribution
2. Try non-parametric method
3. Transform the data into normal distribution

\*\*Use it as it is\*\*

Your data is inherently normally distributed, but it may not look normal when plotted because there are too few samples data. For example, test results of college students follow normal distribution. If you know for certain that your data is normally distributed by nature, then according to the Central Tendency Theorem, your data will eventually become Gaussian when you obtain a greater number of sample data. This means that you can still use the famous standard deviation method to assign letter grades to student, even if your students’ test result does not look Gaussian at all. If you have more students to take your exam, the test result will become Gaussian

<img test distribution image>

On the other hand, if you have plenty enough samples that are representative of the true population, you can fit different types of distributions to better describe your data. Different methods exist for different distributions and maybe you will be able to achieve your goal without using techniques that strictly requires Gaussian distribution. The code snippet below fits three different distributions on the sample data: lognormal, normal, and Weibull distributions. Through visual inspection, it can be observed that the sample data is the best represented by a lognormal distribution. With this conclusion, we can move forward by employing techniques that assume lognormal distribution.

Note

Visual inspection through plotting various distribution fits is one option to assess the performance of the fitted distributions. The other option would be to use hypothesis testing to numerically assess the performance. For example, if you want to numerically assess how well your data matches Gaussian distribution, you can test your hypothesis through D'Agostino-Pearson normality test, Anderson-Darling Test, or Shapiro-Wilk Test.

Let’s say that you have college students’ test result data. Test result of college students is normally distributed. However when you plot the data, it does not look like

The following table shows the non-parametric alternatives to parametric method

.

.

.

Note:

Parametric methods are the type of methods that assumes a certain shape of a distribution. For example, the following equation is used to calculate the confidence interval of a mean of a distribution :

CI of mean = stats of interest ± (distribution score × Standard Error )

The variable in the equation, distribution score, depends on the type of the distribution. If you do not know the distribution shape of your data, it is very difficult to obtain the value of the distribution score.

On the other hand, non-parametric methods do not assume anything about a distribution. In this example, a non-parametric alternative is to use Bootstrapping to calculate the confidence interval of a mean

If Gaussian-like, no need. – if you know the distribution is usually normal, according to the central limit theorem it will be gaussian once you obtain more data, if outliers, robust method