A BRIEF REVIEW OF GRAPHICAL TESTS FOR NORMALITY

Hoang Thanh Hai[1]
[1] *Thai Nguyen University of Economics and Business Administration, Viet Nam*

ABSTRACT

In statistics, it is quite common to assume that the data are normal. A variety of statistical methods are based on this assumption. If this assumption is violated the inferential results would be inaccurate and unreliable. For this reason, it is essential to check this assumption before such statistical analysis. Tests for normality are divided into two main branches. They consist of graphical methods and analytical test procedures. In this paper, we present a brief review of the most commonly used graphical tools for checking the normality assumption.

Keyword: - *tests for normality , graphical methods, histogram, stem-and-leaf plot, normal probability plot*

1. INTRODUCTION

The normal distribution is the most important one in all of probability and statistics. Many numerical populations have distributions that can be fit very closely by an appropriate normal curve. Examples include heights, weights, and other physical characteristics, measurement errors in scientific experiments, reaction times in psychological experiments, scores on various tests, and numerous economic measures and indicators [1]. In statistics, the most commonly used methods are estimation and hypotheses testing. All of them are based on an important assumption: the population from where data are collected is normally distributed. The violation of the normality assumption can lead to invalid inferential statements and inaccurate predictions. Hence, testing the normality assumption is compulsory to ensure the validity of conclusions.

Tests for normality could be divided into two main classes. The first one is graphical methods, and the second one is hypothesis tests. Graphical tools provide visual assessments about data distribution, skewness, or symmetry. Meanwhile, hypothesis testing gives statistically significant conclusions. The prime objective of this paper is to give a brief review of the most common graphical methods for accessing normality.

A continuous random variable X is said to have a normal distribution with parameters μ and σ^2 where $-\infty < \mu < +\infty, \sigma > 0$), if the probability density function of X is

$$f(x;\ \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \qquad -\infty < x < +\infty.$$

The statement that X is normally distributed with parameters μ and σ^2 is often abbreviated $X \sim N(\mu, \sigma^2)$.
The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called standard normal distribution.

If X is normally distributed with parameters μ and σ^2 then its density function graph is bell-shaped and symmetric about μ. Figure 1 presents graphs of $f(x;\ \mu, \sigma^2)$ for several different (μ, σ^2) pairs.

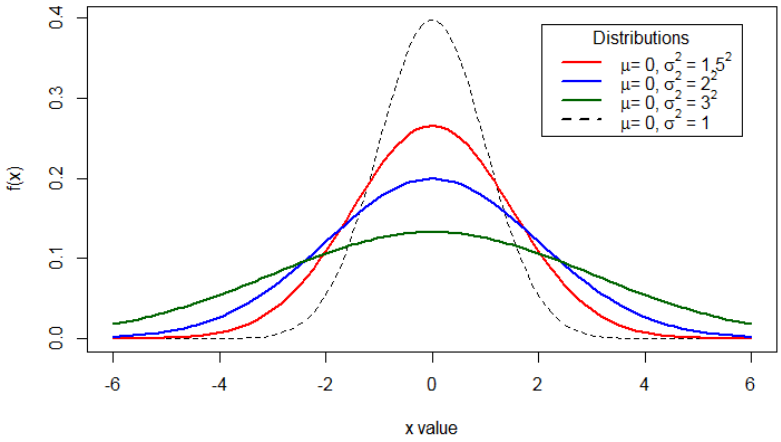

Fig - 1: Graphs of normal density functions for several different (μ, σ^2) pairs

2. GRAPHICAL METHODS FOR CHECKING NORMALITY

Before doing certain statistical procedures, to examine the assumption that the data are normally distributed, we normally start with using graphical tools to reveal data distribution. These judgments can be used to support the conclusion we receive when performing hypothesis tests. Some of the most commonly used graphical methods for checking normality are histograms, stem-and-leaf plots, normal probability plots, and empirical cumulative distribution function plots.

2.1 Histogram

The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). For each bin, the number of points from the data set that falls into the bin is counted. This number is called the frequency of the bin. The histogram graphically shows the following: center of the data, spread of the data, skewness of the data, presence of outliers, and presence of multiple modes in the data. These features provide strong indications of the data distribution.

Data that can be represented by a bell-shaped, symmetric histogram with most of the frequency counts bunched in the middle and with the counts gradually decreasing in the tails are called to have a normal distribution. Figure 2 is the histogram of the height of 8239 students [5]. The histogram is roughly bell-shaped, so we can conclude that the student height data is normally distributed.

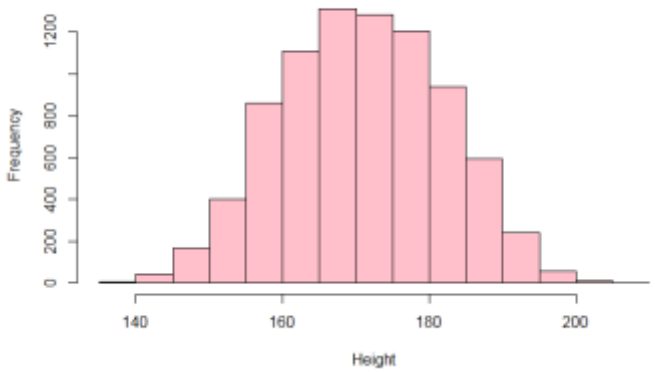

Fig - 2: Histogram of student height data

The histogram of the weight of these students is shown in figure 3. The distribution is skewed right. Its right tail is considerably longer than the other tail. Hence, the weight of student is non-normal.

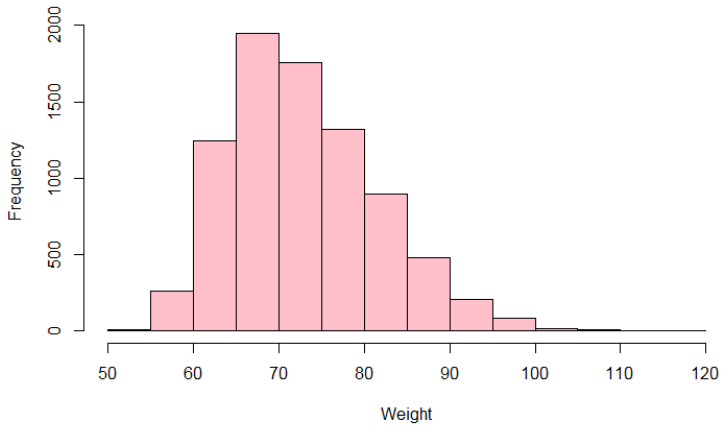

Fig - 3: Histogram of student weight data

2.2 Stem-and-leaf plot

In order to construct the stem-and-leaf plot, the data is classified into class intervals, as in the histogram. The classes are of equal length. Consider $n = 100$ values of $Y = \ln X$ where X is the yarn-strength [lb./22 yarns] of woolen fibers [3, p. 24]. The data is shown in Table 1.

Table - 1: A sample of 100 log (yarn strength)

2.4016	1.1514	4.0017	2.1381	2.5364
2.5813	3.6152	2.5800	2.7243	2.4064
2.1232	2.5654	1.3436	4.3215	2.5264
3.0164	3.7043	2.2671	1.1535	2.3483
4.4382	1.4328	3.4603	3.6162	2.4822
3.3077	2.0968	2.5724	3.4217	4.4563
3.0693	2.6537	2.5000	3.1860	3.5017
1.5219	2.6745	2.3459	4.3389	4.5234
5.0904	2.5326	2.4240	4.8444	1.7837
3.0027	3.7071	3.1412	1.7902	1.5305
2.9908	2.3018	3.4002	1.6787	2.1771
3.1166	1.4570	4.0022	1.5059	3.9821
3.7782	3.3770	2.6266	3.6398	2.2762
1.8952	2.9394	2.8243	2.9382	5.7978
2.5238	1.7261	1.6438	2.2872	4.6426
3.4866	3.4743	3.5272	2.7317	3.6561
4.6315	2.5453	2.2364	3.6394	3.5886
1.8926	3.186	3.2217	2.8418	4.1251
3.8849	2.1306	2.2163	3.2108	3.2177
2.0813	3.0722	4.0126	2.8732	2.4190

The 100 values in Table 1 are classified into 10 classes. The stem-and-leaf diagram presents only the first two digits to the left, with rounding. The first class consists of 4 values, which are represented as 1.2, 1.2, 1.3, 1.4. The ones

digits will be the stem values, and the tenths will be the leaves. Similarly, all other classes are represented. Figure 4 is the stem-and-leaf display of the log yarn-strength data. The histogram indicates a symmetric moderate tailed distribution, so the normal distribution is a good model for the data. In general, stem-and-leaf displays are useful for displaying the shape of the data, giving the reader a quick overview of the distribution. They retain most of the raw numerical data. They are also useful for determining outliers and the mode. However, stem-and-leaf displays are only useful for moderately sized data sets (around 15 – 50 observations). With large data sets, a stem-and-leaf plot will become very cluttered, since each data point must be represented numerically [4].

```
The decimal point is at the |

1 | 2234
1 | 55556778899
2 | 111112223333334444
2 | 5555555666666777788999
3 | 0001111222223444
3 | 555556666677789
4 | 00001334
4 | 55668
5 | 1
5 | 8
```

Fig - 4: Stem-and-leaf diagram of log yarn-strength data

2.3 Normal probability plot

The details involved in constructing probability plots differ a bit from source to source. The basis for the construction is a comparison between percentiles of the sample data and the corresponding percentiles of the distribution under consideration. The (100p)th percentiles of the continuous distribution with cumulative distribution function $F(x)$ is the number x_p that satisfies $F(x_p) = P(X \leq x_p) = p$. Consider a numerical sample of n observations. Order the n observations from smallest to largest. The ith smallest observation in the list is called to be the $[100(i-0.5)/n]$ th sample percentiles.

Consider the (population percentile, sample percentile) pairs – that is, the pairs

$$\left(\left[\frac{100(i-0.5)}{n}\right]\text{th percentile of the distribution},\ \ i\text{th smallest sample observation}\right)$$

for $i = 1,2,\ldots,n$. Each such pair can be plotted as a point on a two – dimensional coordinate system. If the sample percentiles are close to the corresponding population distribution percentiles, the first number in each pair will be roughly equal to the second one. The plotted points will then fall close to a 45^0 line. Substantial deviations of the plotted points from a 45^0 line suggest that the assumed distribution might be wrong.

An investigator is typically not interested in knowing whether a specified probability function, such as the standard normal distribution, is a plausible model for the population distribution from which the sample was selected. Instead, the investigator will want to know whether some member of family of probability distributions, like the family of normal distributions, specifies a plausible model.

The key to construct a normal probability plot is the relationship between standard normal z percentiles and those for any other normal distribution

$$\text{percentile for a normal } (\mu, \sigma^2) \text{ distribution} = \mu + \sigma.(\text{corresponding } z \text{ percentile}).$$

A plot of n pairs

$$([100(i-0.5)/n]\text{th } z \text{ percentile}, i\text{th smallest observation})$$

on a two-dimensional coordinate system is called a normal probability plot. If the sample observations are in fact drawn from a normal distribution with mean value μ and standard deviation σ, the points would fall close to a straight line with slope σ and intercept μ. Thus a plot for which the points fall close to some straight line suggest that the assumption of a normal population distribution is reasonable [1, p.210 – 214].

Figure 5 shows the normal probability plot of log yarn-strength data. The pattern in the plot is quite straight, indicating that it is plausible that the population distribution is normal.

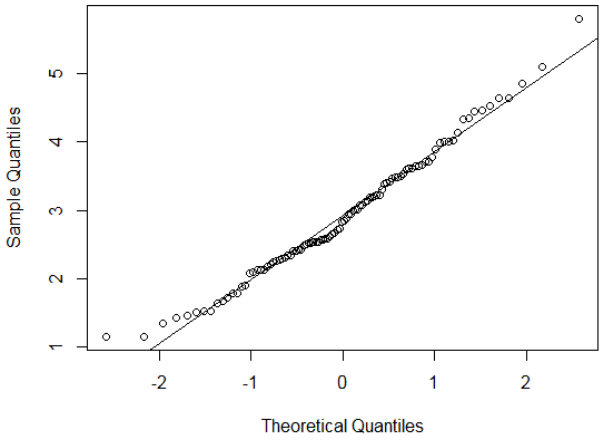

Fig - 5: Normal probability plot of log yarn-strength data

4. CONCLUSIONS

Many formal procedures from statistical inference are based on the assumption that the population distribution is of normal distribution. The use of such a procedure is inappropriate if the actual underlying distribution differs greatly from the assumed type. Normality can be assessed through two methods: graphical methods and statistical tests. Graphical tools help examine normality visually. However, they are somehow subjective. Hence, doing some normal tests after using graphical methods is highly recommended.

5. REFERENCES

[1]. Jay L. Devore, Kenneth N. Berk, 2012, *Modern mathematical Statistics with applications*, 2nd Edition, Springer.
[2]. Keya Rani Das, A. H. M. Rahmatullah Imon, 2016, *A brief Review of Tests for Normality*, American Journal of Theoretical and Applied Statistics, 5(1): 5-12
[3]. Ron S. Kenett, Shelemyahu Zacks, 2014, *Modern Industrial statistics with applications in R, Minitab and JMP*, John Wiley & Sons Ltd.
[4]. https://en.wikipedia.org/wiki/Stem-and-leaf_display
[5].https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Continous-Random-Variables/The-Standard-Normal-Distribution/The-Standard-Normal-Distribution-An-Example/index.html