Probability and Statistics

Topic 16 - Assessing Normality

Aamir Alaud Din, PhD

November 15, 2024

TABLE OF CONTENTS

TABLE OF CONTENTS

RECAP

- Area under the normal curve represents the probability of occurrence of a continuous random variable x given the mean and standard deviation.
- In order to consult Table V, we must convert the value of the continuous random variable x into the corresponding z-score.
- Given the probability $P(x)$of the continuous random variable x, the value of the x can be obtained given the mean and standard deviation using formula

$$x = \mu + z\sigma$$

- Different software packages compute the probability (area under the curve) using different methods/models/simulations *e.g.*, Monte Carlo Simulation.

TABLE OF CONTENTS

OBJECTIVES

After learning this topic and studying, you should be able to:

1. Use Normal Probability Plots to Assess Normality

TABLE OF CONTENTS

THE WHY SECTION

- Up to this point, we have said that a random variable X is normally distributed, or at least approximately normal, provided the histogram of the data is symmetric and bell-shaped.

- This works well for large data sets, but the shape of a histogram drawn from a small sample of observations does not always accurately represent the shape of the population.

- For this reason, we need additional methods for assessing the normality of a random variable X when we are looking at a small set of sample data.

- Our aim is to know and apply methods to assess normality of a given data set.

TABLE OF CONTENTS

USE NORMAL PROBABILITY PLOTS TO ASSESS NORMALITY

- A normal probability plot is a graph that plots observed data versus normal scores.
- A normal score is the expected z-score of the data value, assuming that the distribution of the random variable is normal.
- The expected z-score of an observed value depends on the number of observations in the data set.
- Drawing a normal probability plot requires the following steps:

Drawing a Normal Probability Plot

1. Arrange the data in ascending order.

(Continued)

USE NORMAL PROBABILITY PLOTS TO ASSESS NORMALITY

Drawing a Normal Probability Plot

2. Compute $f_i = \frac{i-0.375}{n+0.25}$, where i is the index (the position of the data value in the ordered list) and n is the number of observations. The expected proportion of observations less than or equal to the i^{th} data value is f_i.

3. Find the z-score corresponding to f_i from Table V.

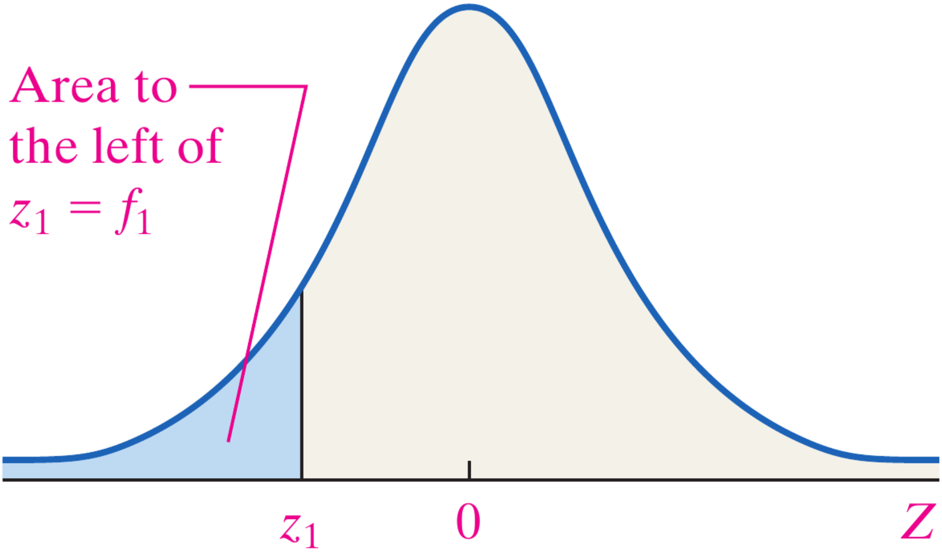
4. Plot the observed values on the horizontal axis and the corresponding expected z-scores on the vertical axis.

- The idea behind finding the expected z-score is that, if the data come from a normally distributed population, we could predict the area to the left of each data value.

USE NORMAL PROBABILITY PLOTS TO ASSESS NORMALITY

- The value of f_i represents the expected area to the left of the i^{th} observation when the data come from a population that is normally distributed.
- For example, f_1 is the expected area to the left of the smallest data value, f_2 is the expected area to the left of the second-smallest data value, and so on.
- See figure below.
- Once we determine each f_i, we find the z-scores corresponding to f_1, f_2, and so on.
- The smallest observation in the data set will be the smallest expected z-score, and the largest observation in the data set will be the largest expected z-score.
- Also, because of the symmetry of the normal curve, the expected z-scores are always paired as positive and negative values.

USE NORMAL PROBABILITY PLOTS TO ASSESS NORMALITY



USE NORMAL PROBABILITY PLOTS TO ASSESS NORMALITY

- Values of normal random variables and their z-scores are linearly related ($x = \mu + z\sigma$), so a plot of observations of normal variables against their expected z-scores will be linear.

- We conclude the following:

  "If sample data are taken from a population that is normally distributed, a normal probability plot of the observed values versus the expected z-scores will be approximately linear."

- It is difficult to determine whether a normal probability plot is "linear enough."

- However, we can use a procedure based on the research of S. W. Looney and T. R. Gulledge in their paper "Use of the Correlation Coefficient with Normal Probability Plots," published in the American Statistician.

USE NORMAL PROBABILITY PLOTS TO ASSESS NORMALITY

- Basically, if the linear correlation coefficient between the observed values and expected z-scores is greater than the critical value found in Table VI in Appendix A, then it is reasonable to conclude that the data could come from a population that is normally distributed.
- Normal probability plots are typically drawn using graphing calculators or statistical software.
- However, it is worthwhile to go through an example that demonstrates the procedure to better understand the results supplied by technology.

EXAMPLE 1

The data in Table 4 represent the finishing time (in seconds) for six randomly selected races of a greyhound named Barbies Bomber in the $\frac{5}{16}$-mile race at Greyhound Park in Dubuque, Iowa. Is there evidence to support the belief that the variable "finishing time" is normally distributed?

USE NORMAL PROBABILITY PLOTS TO ASSESS NORMALITY

Table 4	
31.35	32.52
32.06	31.26
31.91	32.37

Source: Greyhound Park, Dubuque, IA

EXAMPLE 2

Draw a normal probability plot of the data in Table 4 using technology. Is there evidence to support the belief that the variable "finishing time" is normally distributed?

TABLE OF CONTENTS

SUMMARY

- The normality of a data set is assessed using normal probability plot.
- For normal probability plot, the data must be arranged in ascending order.
- Normal probability plot is the plot of expected z-scores against the corresponding data points.
- If a strong correlation is established between the z-scores against the corresponding data points, we say that the data is normally distributed.


Thank You!