

Virtual Laboratories > 9. Hypothesis Testing > 1 2 3 4 5 6 7

7. Probability Plots

A **probability plot** is a goodness of fit test, like the [chi-square goodness of fit test](#), but graphical and informal.

Derivation of the Test

Suppose that we observe real-valued data (x_1, x_2, \dots, x_n) from a [random sample](#) of size n . We are interested in the question of whether the data could reasonably have come from a [continuous distribution](#) with [distribution function](#) F . First, we order that data from smallest to largest; this gives us the sequence of observed values of the [order statistics](#): $(x_{n,1}, x_{n,2}, \dots, x_{n,n})$

1. Show that $x_{n,i}$ is the *sample* quantile of order $\frac{i}{n+1}$.

Of course, by definition, the *distribution* quantile of order $\frac{i}{n+1}$ is

$$y_{n,i} = F^{-1}\left(\frac{i}{n+1}\right)$$

If the data really do come from the distribution, then we would expect the points

$$((x_{n,1}, y_{n,1}), (x_{n,2}, y_{n,2}), \dots, (x_{n,n}, y_{n,n}))$$

to be close to the diagonal line $y = x$; conversely, strong deviation from this line is evidence that the distribution did not produce the data. The plot of these points is referred to as a **probability plot**.

In the following exercises, we will explore probability plots for the [normal](#), [exponential](#), and [uniform](#) distributions.

2. In the [probability plot experiment](#), set the sampling distribution to the standard normal distribution and the sample size to $n = 20$. For each test distribution given below, run the experiment 50 times and note the geometry of the probability plot.

- Standard normal
- Uniform on the interval $[0, 1]$
- Exponential with parameter 1.

3. In the [probability plot experiment](#), set the sampling distribution to the uniform distribution on $[0, 1]$ distribution and the sample size to $n = 20$. For each test distribution given below, run the experiment 50 times and note the geometry of the probability plot.

- a. Standard normal
- b. Uniform on the interval $[0, 1]$
- c. Exponential with parameter 1.

4. In the **probability plot experiment**, set the sampling distribution to the exponential distribution with parameter 1, and the sample size to $n = 20$. For each test distribution given below, run the experiment 50 times and note the geometry of the probability plot.

- a. Standard normal
- b. Uniform on the interval $[0, 1]$
- c. Exponential with parameter 1.

Location-Scale Families

Usually, we are not trying to fit the data to a *particular* distribution, but rather to a parametric *family* of distributions (such as the normal, uniform, or exponential families). We are usually forced into this situation because we don't know the parameters; indeed the next step, after the goodness of fit, may be to estimate the parameters. Fortunately, the probability plot method has a simple extension for any **location-scale** family of distributions.

Suppose that G is a given distribution function. Recall that the location-scale family associated with G has distribution function

$$F(x) = G\left(\frac{x - a}{b}\right), \quad x \in \mathbb{R}$$

where $a \in \mathbb{R}$ is the location parameter and $b \in (0, \infty)$ is the scale parameter.

5. For $p \in (0, 1)$, let z_p denote the quantile of order p for G and y_p the quantile of order p for F . Show that

$$y_p = a + b z_p$$

From Exercise 5, it follows that if the probability plot constructed with distribution function F is nearly linear (and in particular, if it is close to the diagonal line), then the probability plot constructed with distribution function G will be nearly linear. Thus, we can use the distribution function G without having to know the location and scale parameters.

6. In the **probability plot experiment**, set the sampling distribution to normal distribution with mean 5 and standard deviation 2. Set the sample size to $n = 20$. For each of the following test distributions, run the experiment 50 times and note the geometry of the probability plot:

- a. Standard normal
- b. Uniform on the interval $[0, 1]$

c. Exponential with parameter 1.

7. In the **probability plot experiment**, set the sampling distribution to the uniform distribution on $[4, 10]$. Set the sample size to $n = 20$. For each of the following test distributions, run the experiment 50 times and note the geometry of the probability plot:

- Standard normal
- Uniform on the interval $[0, 1]$
- Exponential with parameter 1.

8. In the **probability plot experiment**, Set the sampling distribution to the exponential distribution with parameter 3. Set the sample size to $n = 20$. For each of the following test distributions, run the experiment 50 times and note the geometry of the probability plot:

- Standard normal
- Uniform on the interval $[0, 1]$
- Exponential with parameter 1.

Data Analysis Exercises

9. Draw the normal probability plot with **Michelson's velocity of light data**. Interpret the results.

10. Draw the normal probability plot with **Cavendish's density of the earth data**. Interpret the results.

11. Draw the normal probability plot with **Short's parallax of the sun data**. Interpret the results.

12. Draw the normal probability plot for the petal length variable in **Fisher's iris data**, using the following cases. Compare the results.

- All cases
- Setosa only
- Verginica only
- Versicolor only

Interpreting the Results

From your experiments, we hope that you have reached a few general conclusions. First, the probability plot method is of very little use with small samples. With just five points, for example, it is essentially impossible to judge the linearity of the probability plot. Even with large samples, the results can be rather subtle. Experience with a variety of distributions helps in making the fine judgments.