Estadística Española • Volumen 55, número 181 / 2013, pp. 177-186

Normal S-P Plots and Distribution Curves: a tool for studying the distributional behaviour of a data set

Sonia Castillo-Gutiérrez
Departamento de Estadística e Investigación Operativa. Universidad de Jaén

María Dolores Estudillo-Martínez
Departamento de Estadística e Investigación Operativa. Universidad de Jaén

Emilio Damián Lozano-Aguilera
Departamento de Estadística e Investigación Operativa. Universidad de Jaén

Abstract

The distribution curves are useful on P-P Plots to identify viable alternative probability models for the sample data when the hypothetical distribution is rejected. In this paper, an extension of the distribution curves on Normal S-P Plots is provided. Likewise, the key features of the main probability plots, Q-Q Plots, P-P Plots and S-P Plots, used as a visual tool for assessing the fit of a given probability model to some data, are reviewed. Moreover, an R script to construct some distribution curves on a Normal S-P Plot, when the normal distribution is rejected for the sample observations, is developed.

Keywords: Normal S-P Plots; Distribution Curves; Plotting Positions; R Script; Q-Q Plots; P-P Plots

AMS Classification: 62-09; 62-07; 62G10; 62G30

Resumen

Las curvas de distribución se usan en los P-P Plots para identificar posibles modelos de probabilidad alternativos para una muestra cuando se rechaza la hipótesis distribucional de partida. En este artículo, se realiza una extensión de las curvas de distribución para los S-P Plots Normales. Así mismo, se analizan las características fundamentales de los principales gráficos de probabilidad: Q-Q Plots, P-P Plots y S-P Plots, que son usados como una herramienta para determinar de forma gráfica el ajuste de un modelo de probabilidad dado a unos datos concretos. Además, se desarrolla un programa o script en R para construir algunas curvas de distribución en un S-P Plot Normal, cuando la distribución normal es rechazada para las observaciones muestrales.

Palabras clave: S-P Plots Normales; Curvas de distribución; Puntos de Posición Gráfica; R Script; Q-Q Plots; P-P Plots

Clasificación AMS: 62-09; 62-07; 62G10; 62G30

1. Introducción

Graphical techniques are used as a tool for studying the distributional behaviour of a set of observations as an alternative to analytical techniques since they are more intuitive and easily interpretable.

The probability plots are those that represent theoretical quantiles or probabilities against empirical quantiles or probabilities, respectively. Theoretical quantiles or probabilities correspond to the hypothetical distribution that we want to assess if a set of observations have it. Empirical quantiles or probabilities are referred to the sample data studied.

The main probability plots are: Quantile-Quantile Plot or Q-Q Plot, Probability-Probability Plot or P-P Plot and Stabilized-Probability Plot or S-P Plot.

Let $\{x_1, x_2, \dots, x_n\}$ be a simple random sample of size n from a distribution $F(x)$ with unknown location and scale parameters denoted by μ and σ, respectively. Let $F_0(x)$ be the distribution function of the hypothesized probability distribution, i.e., that with which the distribution of observations is compared.

The Q-Q Plot is constructed representing the quantiles of empirical distribution (that of the set of observations studied) against the corresponding quantiles of the theoretical distribution, previously selected. Therefore, the Q-Q Plot will plot the empirical quantile, i.e., the ordered observations from lowest to highest, $x_{(i)}, \ \ i = 1, \dots, n$, against the theoretical quantiles, i.e., $F_0^{-1}(p_i), \ i = 1, \dots, n,$, where p_i is an appropriate plotting position. If the theoretical distribution is a good approximation of the data distribution, the plotted points would have a straight configuration or almost straight configuration.

In the P-P Plots are represented cumulative probabilities of the ordered observations against the expected cumulative probabilities of the hypothesized distribution. More specifically, the standardized P-P Plots are constructed plotting the cumulative probabilities $F_0\big((x_{(i)} - \mu)/\sigma\big), \ i = 1, \dots, n$ against plotting position, $p_i, \ i = 1, \dots, n$. In case of μ and σ are unknown, they will be replaced by their maximum likelihood estimators, $\hat{\mu}$ and $\hat{\sigma}$ and in case of normal distribution, the usual unbiased variance estimate will be used [6,15]. If the theoretical distribution is a good approximation of the data distribution, plotted points are arranged around the bisectrix of the first quadrant, i.e., about the line $y=x$ defined between 0 and 1.

The S-P Plot [10] or Stabilized-Probability Plot appears as a transformation of the P-P Plot to stabilize the variance of the plotted points, i.e., in this type of plot the above mentioned variances are approximately equal.

The origin of the S-P Plots is that in the Q-Q Plots and P-P Plots some points have higher variance than others. For example, when the theoretical distribution is the normal

distribution, in the Q-Q Plots the points nearest to the centre of the graph have a smaller variance than the points of the tails, while in the P-P Plots the opposite happens.

The values

$$r_i = \left(\frac{2}{\pi}\right) \arcsin\left(\sqrt{p_i}\right), \qquad i = 1, \ldots, n$$

against

$$s_i = \left(\frac{2}{\pi}\right) \arcsin\left(\sqrt{u_i}\right), \qquad i = 1, \ldots, n$$

where p_i are some appropriate plotting positions and

$$u_i = F_0\left(\frac{x_{(i)} - \mu}{\sigma}\right), \qquad i = 1, \ldots, n$$

are represented in the S-P Plots. As in the P-P Plots, if the theoretical distribution is a good approximation of the empirical distribution, plotted points are arranged near the line $y=x$ defined between 0 and 1.

In the construction of the three probability plots described above we can see that the plotting positions p_i involved are key elements. In the literature, there are many proposals for these values [2,4]. Most of them are derived from the expression

$$p_i = \frac{i - c}{n - 2c + 1} \qquad 0 \leq c \leq 1$$

giving different values to c.

Some different definitions arise when it is considered that the p_i must be determined from measurements of localization of order statistics [1, 5, 8, 9, 12, 13, 16].

In [2] we can find a comparative study of different plotting positions. This study shows that the most extreme proposals are those of Hazen (1930) [7] where c=0.5 and Weibull (1939) [14] where c=0.

In this paper, the definition of plotting positions proposed by Yu and Huang (2001) [16] where c=0.326, is used, i.e.:

$$p_i = \frac{i - 0.326}{n + 0.348} \qquad i = 1, \ldots, n$$

This formula of the plotting position is based on an approximation to the median of the order statistics.

2. Distribution Curves on Normal S-P Plots

As stated above, by using the probability plots we can assess whether the sample data come from of a certain distribution. If the plotted points do not follow a straight configuration, that indicates that the observations do not have the hypothetical distribution.

In this paper, we will focus on normal distribution, i.e., we can assess whether a set of observations has a normal distribution. Therefore, in this case, $F_0 = \Phi$ is the standard normal distribution function.

When the plotted points on a Normal S-P Plot do not have a rectilinear configuration, then, the sample data do not have a normal distribution. The question now is how to determine an alternative probability model for the data set.

With similar intention, Gan, Koehler and Thompson (1991) [6] constructed the so-called distribution curves applied to the P-P Plots. These curves allow us to propose an alternative distribution of probability for observations when the hypothetical distribution is rejected. The idea is to include in the graph different distribution curves and to study whether the plotted points are near to someone of them, so that, visually, we can choose an appropriate probability model. To check if really the observations follow the actually selected alternative distribution, it would build the corresponding probability plot for the distribution and check that the points on the graph have an approximately straight configuration.

In this contribution, an extension of the distribution curves on Normal S-P Plots is provided.

The procedure to construct a *G* distribution curve on a Normal S-P Plot for a location-scale family is the following [2, 3]:

1. Construct a Normal S-P Plot with the sample observations (i. e. $\{x_1,\ x_2, \dots, x_n\}$).
2. Select a number k and compute the k plotting positions p_i by the following expression:

$$p_i = \frac{i - 0.326}{k + 0.348} \qquad i = 1, \dots, k$$

3. Apply the transformation to stabilize the variance:

$$r_i^* = \left(\frac{2}{\pi}\right) \arcsin\left(\sqrt{p_i}\right) \quad i = 1, \dots, k$$

4. Obtain the values $G^{-1}(r_i^*)$ for each $i = 1, \dots, k$, where *G* is the distribution function of the probability model selected.
5. Use the values obtained in the previous step to calculate the estimates of location-scale parameters of normal distribution, $\hat{\mu}$ and $\hat{\sigma}$.

6. Calculate the values

$$y_i = \Phi\left(\frac{G^{-1}(r_i^*) - \hat{\mu}}{\hat{\sigma}}\right) \quad i = 1, \ldots, k$$

7. Apply arcsin transformation to stabilize the variance:

$$s_i^* = \left(\frac{2}{\pi}\right)\arcsin\left(\sqrt{y_i}\right) \quad i = 1, \ldots, k$$

8. Plot the pairs of points (r_i^*, s_i^*) $i = 1, \ldots, k$ on the Normal S-P Plot of step 1, joining them to get a smooth curve.

3. R script

In this section the basic content of an R script [11] to construct some distribution curves on a Normal S-P Plot is presented. When this R script is run, a Normal S-P Plot and four distribution curves corresponding to four probability models (uniform, exponential, Cauchy and Gumbel distributions) are plotted. A number of k=100000 has been selected. The R script proposed is the following:

```
sp.curvas <– function(x)
{
x <– sort(x)
n <– length(x)
pyuhuang <– c((1:n-0.326)/(n+0.348))  # Plotting Position of Yu and Huang
r <– (2*asin(sqrt(pyuhuang)))/pi
mu <– mean(x)
sd <– sd(x)
s <– (2*asin(sqrt(pnorm((x-mu)/sd))))/pi
plot(r, s, xlim=c(0,1), ylim=c(0,1), xlab="Transformed Plotting Positions",
ylab="Transformed Probabilities")
abline(a=0, b=1)
k <– 100000
p.curve <– c((1:k-0.326)/(k+0.348))
r.curve <– (2*asin(sqrt(p.curve)))/pi
library("evd")

    # Distribution curve: Gumbel
y.gumbel <– qgumbel(r.curve)
mu.gumbel <– mean(y.gumbel)
sd.gumbel <– sd(y.gumbel)
gumbel <– pnorm((y.gumbel-mu.gumbel)/sd.gumbel)
s.gumbel <– (2*asin(sqrt(gumbel)))/pi
points(r.curve, s.gumbel, type="l", lty=1)
```

```
    # Distribution curve: Cauchy
y.cauchy <– qcauchy(r.curve)
mu.cauchy <– mean(y.cauchy)
sd.cauchy <– sd(y.cauchy)
cauchy <– pnorm((y.cauchy-mu.cauchy)/sd.cauchy)
s.cauchy <– (2*asin(sqrt(cauchy)))/pi
points(r.curve, s.cauchy, type="l", lty=2)

    # Distribution curve: Exponential
y.exp <– qexp(r.curve)
mu.exp <– mean(y.exp)
sd.exp <– sd(y.exp)
exp <– pnorm((y.exp-mu.exp)/sd.exp)
s.exp <– (2*asin(sqrt(exp)))/pi
points(r.curve, s.exp, type="l", lty=3)

    # Distribution curve: Uniform
y.unif <– qunif(r.curve)
mu.unif <– mean(y.unif)
sd.unif <– sd(y.unif)
unif <– pnorm((y.unif-mu.unif)/sd.unif)
s.unif <– (2*asin(sqrt(unif)))/pi
points(r.curve, s.unif, type="l", lty=4)
}
```

4. Numerical example

Next, an application of the R script is developed. In Figure 1, a Normal S-P Plot of a simulated sample of 100 observations from a exponential distribution is plotted. Moreover, four distribution curves are also represented, these that correspond to the uniform, exponential, Cauchy and Gumbel distributions.

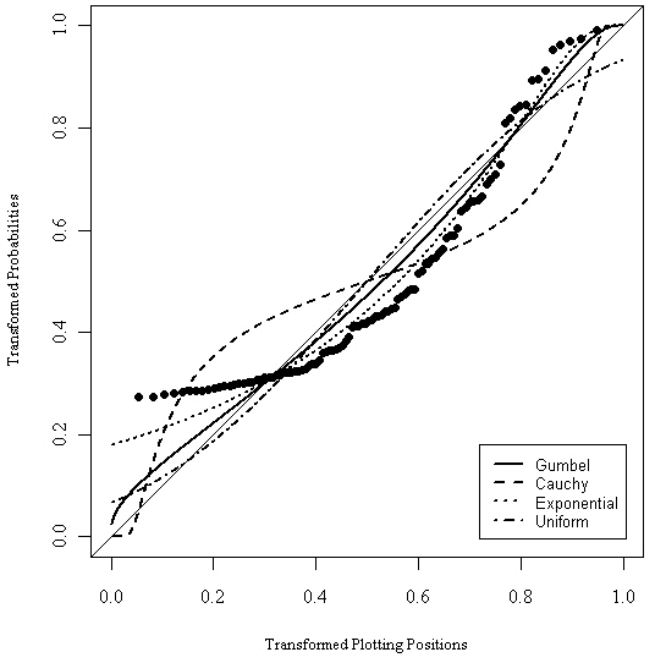
Figure 1 shows that the sample observations do not have a rectilinear configuration and this indicates that the sample data do not come from a normal distribution.

It can be interested in determining an alternative probability model to the normal distribution. For this, four distribution curves from uniform, exponencial, Cauchy and Gumbel distributions are also displayed. The configuration of sample data shows that a exponential distribution can be the appropriate alternative to the hypothesized probability distribution, the normal distribution.

To check if the sample data come from a exponential distribution, it has to construct an Exponential S-P Plot and to study if the plotted points have a rectilinear configuration.

Figura 1

Normal S-P Plot and distribution curves



In this example, the size of the sample data is 100. To determine how many sample observations are necessary to be able to select a distribution as a good candidate as alternative probability model to the normal distribution, let us see the following example.

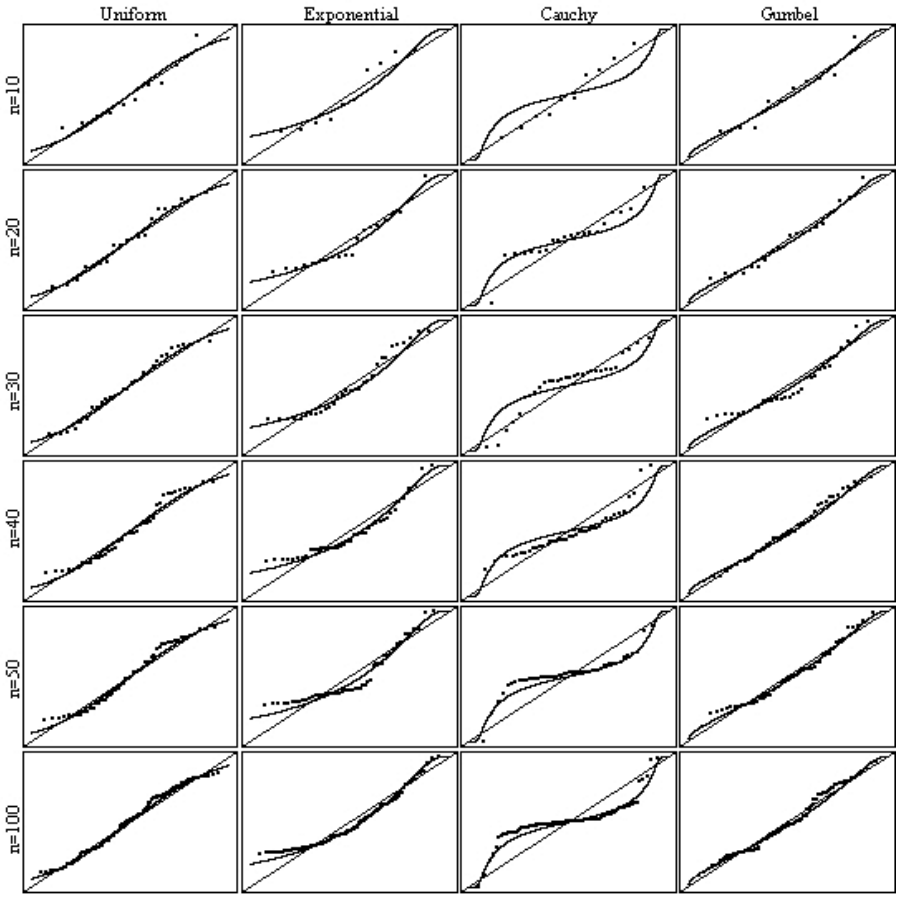
Six random samples of size 10, 20, 30, 40, 50 and 100 have been generated from uniform, exponential, Cauchy and Gumbel distributions. With each of these samples a Normal S-P plot is constructed. On the Normal S-P Plot of uniform samples a uniform distribution curve is displayed and the same with the other three distributions. Figure 2 shows the twenty-four plots.

In Figure 2, first, we can observe that the plotted points do not have a rectilinear configuration in the twenty-four plots. This indicates that the sample data do not come from a normal distribution. Then, we study if the distribution curves provide a viable alternative distribution of probability for different sample sizes.

We can observe that when we have small sample sizes some alternative probability model can be intuited. Sample size of 20 data provides weak information but points to possible distributions to choose. As the sample size increases the distribution curves clearly provide an alternative probability distribution to the normal distribution.

Figura 2

Normal S-P Plot and distribution curves for four distributions and different sample sizes (*n*=10,20,30,40,50,100)



5. Concluding remarks

In this paper, a review of the principal probability plots is presented. Q-Q Plots, P-P Plots and S-P Plots are a useful graphical tool for assessing if a sample data come from a determined probability distribution.

When the hypothetical distribution is rejected, the distribution curves can provide a viable alternative probability model. In this contribution, an extension of the distribution curves on Normal S-P Plots and an R script to construct some distribution curves on a Normal S-P Plot are provided.

Moreover, by an example, a study of the influence of sample size is presented. We can see that even with moderate sample sizes, the distribution curves displayed on the Normal S-P Plot, clearly points to a viable alternative probability model to the normal distribution for the sample data.

References

[1] BENARD, A. AND BOS-LEVENBACH, E. (1953) «Het uitzetten van waarnemingen op waarschijnlijkdeids-papier (The Plotting of Observations on Probability Paper)», *Stat Neerl*, 7 , pp. 163–173.

[2] CASTILLO-GUTIÉRREZ, S. (2011), «Gráficos de probabilidad: una herramienta para el análisis y el contraste de normalidad», *Editorial Académica Española.*

[3] CASTILLO-GUTIÉRREZ, S.; LOZANO-AGUILERA, E.D. AND ESTUDILLO-MARTÍNEZ, M.D. (2011), «Normal S-P Plots and Distribution Curves», *Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE*, Spain, pp. 315–318.

[4] CASTILLO-GUTIÉRREZ, S.; LOZANO-AGUILERA, E.D. AND ESTUDILLO-MARTÍNEZ, M.D. (2012), «Selection of a Plotting Positions for a Normal Q-Q Plot. R script», *Journal of Communication and Computer*, Vol. 9, Nº 3, pp. 243–250.

[5] FILLIBEN, J. (1975), «The Probability Plot Correlation Coefficient Test for Normality», *Technometrics*, 17, pp. 111–117.

[6] GAN, F.F.; KOEHLER, K.J. AND THOMPSON, J.C. (1991), «Probability Plots and Distribution Curves for Assessing the Fit of Probability Models», *Am Stat*, 45, Nº 1, pp. 14–21.

[7] HAZEN, A. (1930), «Flood Flows. A Study of Frecuencies and Magnitudes», Wiley, New York.

[8] LA BRECQUE, J. (1977), «Goodness-of-fit Tests based on Nonlinearity in Probability Plots», *Technometrics*, 19, 3, pp. 293–306.

[9] LOZANO-AGUILERA, E.D. (1995), «Aportaciones a las técnicas gráficas para el estudio de normalidad y las causas de su pérdida», *Ph.D. diss.*, University of Granada.

[10] MICHAEL, J.R. (1983), «The Stabilized Probability Plot», *Biometrika*, 70, Nº 1, pp. 11–17.

[11] R DEVELOPMENT CORE TEAM (2008), «R: A Language and Environment for Statistical Computing», Viena, Austria, software available at http://www.r-project.org/.

[12] SHAPIRO, S. AND FRANCIA, R. (1972), «An Approximate Analysis of Variance Test for Normality», *J. Am. Stat. Assoc.*, 67, pp. 215–216.

[13] SHAPIRO, S. AND WILK, M. (1965), «An Analysis of Variance Test for Normality (Complete Samples)», *Biometrika*, 52, 3, 4, pp. 591–611.

[14] WEIBULL, W. (1939), «The Phenomenon of Rupture in Solids», *Ing. Vet. Ak. Handl.*, 17.

[15] WILK, M.B. AND GNANADESIKAN, R. (1968), «Probability Plotting Methods for the Analysis of Data», *Biometrika,* 55, 1, pp. 1–17.

[16] YU, G.-H. AND HUANG, C.-C. (2001), «A distribution free plotting position», *Stoch. Environ. Res. Risk Assess.*, 15, pp. 462–476.