

Suppose we have a random sample of size  $n$ :  $x_1, x_2, \dots, x_n$  from an unknown distribution function  $F(x)$  and we wish to ascertain if the sample has originated from a specified family of distribution functions such as the normal, exponential, Weibull, or others. Note that each family of distributions in the field of statistics has a specific cdf (cumulative distribution function). For example, the cdf of a Gaussian  $N(\mu, \sigma^2)$  is given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \quad (\text{No closed-form inverse function exists because it}$$

is impossible to directly solve exactly for  $x$  in terms of  $F(x)$ .)

The cdf of the exponential density is given by  $F(x) = 1 - e^{-\lambda x}$  and its inverse function is obtained by first solving  $x$  in terms of  $F(x) = 1 - e^{-\lambda x}$  as follows.

$$e^{-\lambda x} = 1 - F(x) \rightarrow -\lambda x = \ln[1 - F(x)] \rightarrow x = \frac{1}{\lambda} \ln[1 - F(x)]^{-1} \rightarrow$$

$$F^{-1}(x) = \frac{1}{\lambda} \ln[(1 - x)^{-1}]. \quad \text{Note that } F[F^{-1}(x)] = F^{-1}[F(x)] = x.$$

Similarly, the cdf of a Weibull rv with minimum life  $\delta = 0$ , characteristic life  $\theta$ , and slope  $\beta$  is given by  $F(x) = 1 - e^{-(x/\theta)^\beta}$ . Directly inverting this cdf results in  $x = \theta \ln[(1 - F)^{-1/\beta}] \rightarrow F^{-1}(x) = \theta \ln[(1 - x)^{-1/\beta}]$ . Again note that  $F[F^{-1}(x)] = F^{-1}[F(x)] = x$ .

The random sample of size  $n$  can be ordered ascendingly as  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ , where the 1<sup>st</sup>-order statistic  $x_{(1)}$  is the sample minimum and  $x_{(n)}$  is the sample maximum. Then one estimate of the population cdf corrected for continuity (cfc) at the  $i^{\text{th}}$  order statistic is given by  $\hat{F}(x_{(i)}) = \frac{i - 0.5}{n}$ . Another estimate of  $F(x_{(i)})$  is

$$\text{given by } \hat{F}(x_{(i)}) = \frac{i - 0.30}{n + 0.40}, i = 1, 2, \dots, n; \text{ these latter } n \text{ estimates } \frac{i - 0.30}{n + 0.40} \text{ are}$$

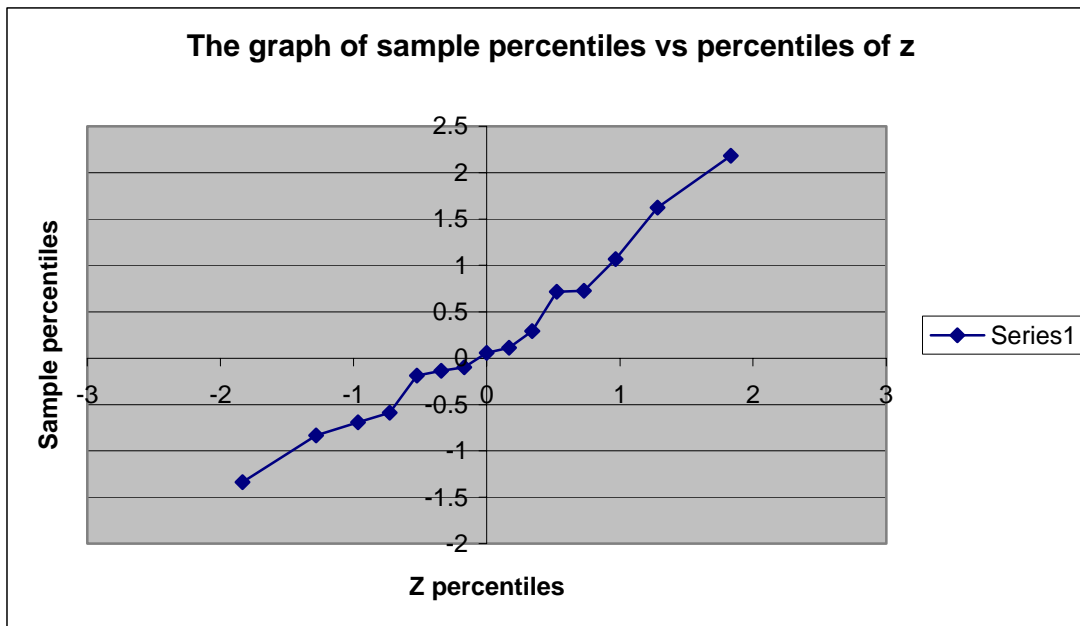
called median ranks and are commonly used in Reliability Engineering. For

completeness, one can also use the mean rank estimates  $\hat{F}(x_{(i)}) = \frac{i}{n+1}$ . For example, if our random sample size  $n = 15$ , then our three estimates of the population cdf at the 4<sup>th</sup>-order statistic are  $\hat{F}(x_{(4)}) = \frac{3.5}{15} = 0.23333$ ,  $\hat{F}(x_{(4)}) = \frac{4 - 0.30}{15 + 0.40} = 0.24026$  (median-rank estimate), and  $\hat{F}(x_{(4)}) = \frac{4}{16} = 0.25$  (the mean – rank estimate). Any one these three estimates is acceptable, and for  $n > 30$  the 1<sup>st</sup> two estimators are fairly close. The most commonly used estimates are  $\hat{F}(x_{(i)}) = \frac{i - 0.5}{n}$  and the median rank estimate  $\hat{F}(x_{(i)}) = \frac{i - 0.30}{n + 0.40}$ . When the sample size  $n = 15$ , then  $\hat{F}(x_{(4)}) = \frac{4 - 0.30}{15 + 0.40} = 0.24026$  implies we are 50% confident that 24.026% of the population lies below the 4<sup>th</sup>-order statistic  $x_{(4)}$ . Henceforth, we will employ the 1<sup>st</sup> estimator  $\hat{F}(x_{(i)}) = \frac{i - 0.5}{n}$ , whose value for the 4<sup>th</sup>-order statistic is  $\hat{F}(x_{(4)}) = \frac{3.5}{15} = 0.23333$ , although Minitab uses the Median-Rank estimates. This implies that  $x_{(4)}$  is an estimate of the 23.333 percentile of the population. Similarly, for  $n = 15$ ,  $\hat{F}(x_{(8)}) = \frac{7.5}{15} = 0.50$  implies that the 8<sup>th</sup>-order statistic  $x_{(8)}$  is an estimate of the 50<sup>th</sup> percentile (or the median) of the population. Thus, we have established that sample order statistics are estimates of population percentiles. Again for example, for  $n = 15$  and  $i = 2$  because  $\frac{2 - 0.5}{15} = 1.5/15 = 0.10$  then the 2<sup>nd</sup>-order statistic is an estimate of the 1<sup>st</sup> decile of the population, i.e.,  $x_{(2)} = \hat{x}_{0.10}$ .

**Definition.** A probability plot is the graph of the  $i^{\text{th}}$ -order statistic of the sample (or the sample percentile) versus the  $(\frac{i - 0.5}{n}) \times 100^{\text{th}}$  percentile of a specified distribution. Thus in pr plotting, the abscissa is the percentile of the specified

distribution (such as the normal, exponential, or the Weibull) while the ordinate, or the y-axis, represents the order statistics of the sample. If the data do not violate the assumption that the population cdf is  $F(x)$ , then the sample percentiles should almost equal to the corresponding population percentiles and as a result the graph of  $x_{(i)}$  versus the  $(\frac{i-0.5}{n}) \times 100^{\text{th}}$  percentile of the specified distribution should be a straight line thru the origin at roughly  $45^\circ$  degrees (recall that the graph of the function  $y = x$  is a straight line thru the origin with a slope of  $45^\circ$  degrees).

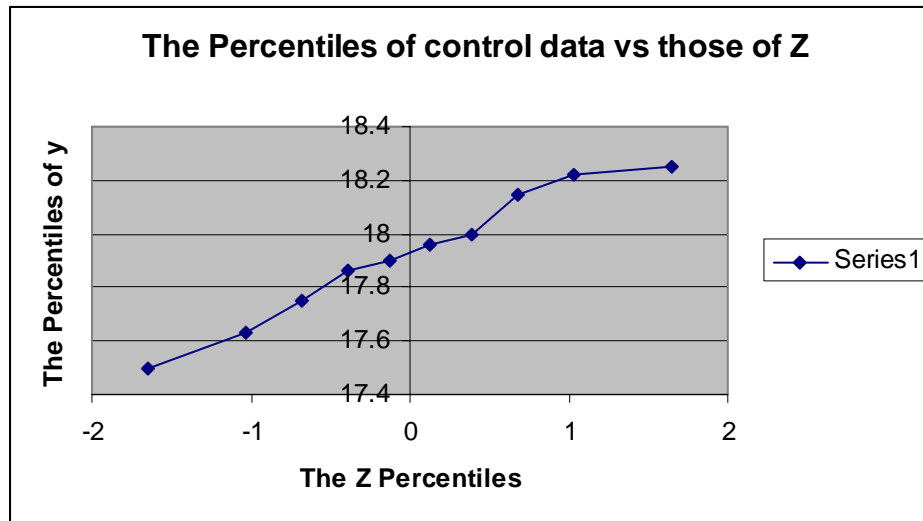
**Example.** I used the Matlab function: random ('Normal', 0, 1, 1, 15), to generate a random sample of size 15 from the  $N(0, 1)$ : -0.1867, 0.7258, -0.5883, 2.1832, -0.1364, 0.1139, 1.0668, 0.0593, -0.0956, -0.8323, 0.2944, -1.3362, 0.7143, 1.6236, -0.6918. The order statistics are -1.3362, -0.8323, -0.6918, -0.5883, -0.1867, -0.1364, -0.0956, 0.0593, 0.1139, 0.2944, 0.7143, 0.7258, 1.0668, 1.6236, 2.1832. Obviously this data set is normally distributed with  $\mu = 0$  and  $\sigma = 1$  because the sample values were generated from the  $N(0, 1)$ . However, suppose we did not have this information and we wish to test that the data originated from a normal distribution. How do go about doing this task? One way is to graph the sample percentiles  $0.5 \times 100/15 = 3.3333$  percentile,  $1.5 \times 100/15 = 10.00^{\text{th}}$  percentile,  $2.5 \times 100/15 = 16.6667$ , 23.3333, 30.0000, 36.6667, 43.3333, 50.0000, 56.6667, 63.3333, 70.0000, 76.6667, 83.3333, 90.0000, and  $14.5 \times 100/15 = 96.6667^{\text{th}}$  percentile versus the same percentiles of a normal population with zero mean and unit variance. The corresponding percentiles of  $z$  are as follows:  $z_{0.96667} = -1.8339$ ,  $z_{0.90} = -1.2816 =$  the  $10^{\text{th}}$  percentile, the  $16.6667^{\text{th}}$  percentile = -0.9674, -0.7279, -0.5244, -0.3407, -0.1679,  $z_{0.50} = 0$ , 0.1679, 0.3407, 0.5244, 0.7279, 0.9674, 1.2816,  $z_{0.0333} = 1.8339$ . The 15 ordered pairs (-1.8339, -1.3362), (-1.2816, -0.8323), (-0.9674, -0.6918), ..., (1.8339, 2.1832) are graphed atop the next page in Figure 1. Since the graph in Figure 1 roughly is a straight line thru the origin at  $45^\circ$  degrees, then the assumption of  $N(0, 1)$  distribution is indeed tenable.



**Figure 1.**

Next what happens if we graph the order statistics of a data set with nonzero mean and non-unit variance versus the z percentiles. Recall that if  $y$  is  $N(\mu, \sigma^2)$ , then its  $p \times 100^{\text{th}}$  percentile is given by  $y_p = \mu + z_{1-p} \times \sigma$ . For example, the  $90^{\text{th}}$  percentile of  $y$  is given by  $y_{0.90} = \mu + z_{0.10} \times \sigma = \mu + 1.2816 \times \sigma$ . Thus, the graph of  $y_p = \mu + \sigma \times z_{1-p}$  versus  $z$  will be a straight line with y-intercept at  $\mu$  and slope  $\sigma$ . As an example, let's graph the data of the control group  $y_{2j} : 17.50, 17.63, 17.75, 17.86, 17.90, 17.96, 18.00, 18.15, 18.22, 18.25$  given in Table 2.1 of Montgomery on page 22 versus the corresponding z percentiles  $-1.64485348, -1.03643347, -0.67448953, -0.3853206, -0.12566125, 0.125661246, 0.385320604, 0.674489526, 1.036433474, 1.644853476$ , which is provided in Figure 2.

Although the normal pr plot in Figure 2 is not exactly a straight line (the  $n^{\text{th}}$  order statistic 18.25 seems a bit too small), but one has to allow for sampling variation specially for such a small sample size  $n = 10$ . Note that the line crosses the y-axis near the mean of the data  $\bar{y} = 17.9220$  and its slope is close to the data standard deviation  $s = 0.2479158$ .

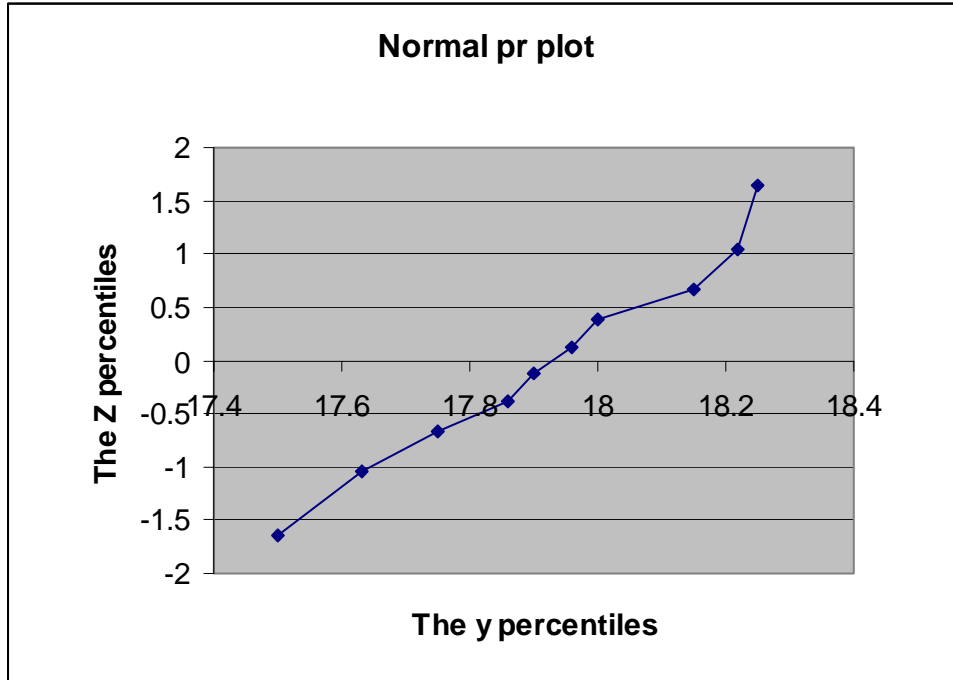


**Figure 2.**

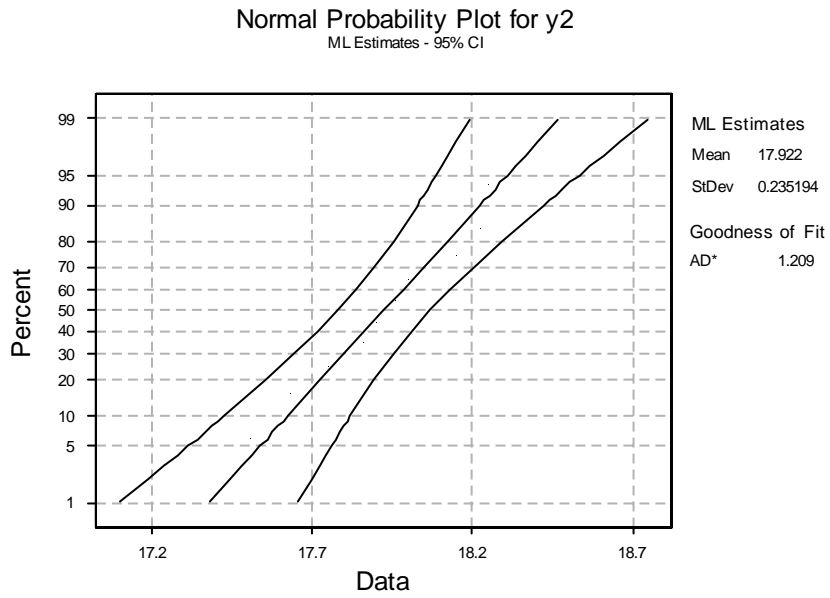
To give the reader more intuition, suppose we exchange the abscissa and ordinates in the above graph, then the graph will still stay roughly linear. We will then get the normal pr plot as do statistical softwares if we replace the z values on the ordinate with the corresponding cdf of the standardized normal. These two graphs are given in Figures 3 and 4, respectively.

From the above discussions it should be clear that in practice the sample percentiles are not graphed versus percentiles of a specified distribution. In fact, some statistical softwares (such Minitab) graph the cdf of the specified distribution (the ordinate) versus sample order statistics. For this reason, commercial pr (probability) plot papers (such Normal and Weibull) are available for pr plotting. SAS actually graphs the sample percentiles versus those of z.

The ordinate of a normal pr plot is scaled (or calibrated) such that the cumulative prs 0.00003169, 0.00135, 0.02275, 0.158655, 0.50, 0.841345, 0.97725, 0.99865, and 0.99996831 are placed at equidistance from each other on its ordinate



**Figure 3.**



**Figure 4.**

because these cumulative prs exactly correspond to the standardized normal deviates of -4, -3, -2, -1, 0, 1, 2, 3, and 4, respectively. That is to say, the standardized normal inverses of the cdfs 0.00003169, 0.00135, 0.02275, 0.158655, 0.50, 0.841345, 0.97725, 0.99865, and 0.99996831 are equal to -4, -3, -2, -1, 0, 1, 2, 3, and 4, respectively. The above calibration is demonstrated in Figure 3-4 on page 78 of Montgomery. In Figure 3-4 on page 78 of Montgomery the inverses of the cumulative prs on the left ordinates are  $z_{0.99} = -2.3263$ ,  $z_{0.95} = -1.6449$ ,  $z_{0.90} = -1.2816$ ,  $z_{0.80} = -0.8416$ ,  $z_{0.70} = -0.5244$ ,  $z_{0.50} = 0.0000$ ,  $z_{0.30} = 0.5244$ ,  $z_{0.20} = 0.8416$ ,  $z_{0.10} = 1.2816$ ,  $z_{0.05} = 1.6449$ ,  $z_{0.01} = 2.3263$ .