Journal of Empirical Research in Accounting & Auditing
ISSN (2384-4787)
J. Emp. Res. Acc. Aud. 3, No. 2 (October-2016)

On Uses of Dot Chart

Elsayed A. H. Elamir

Department of Statistics and Mathematics, faculty of Commerce, Benha University, Egypt and Management & Marketing Department, College of Business, University of Bahrain, P.O. Box 32038, Kingdom of Bahrain

Received: 15March 2016, Revised: 4 August 2016, Accepted: 5 Sept. 2016, Published: (Oct) 2016

Abstract: A simple and informative dot-chart is presented that gives clear picture and insight into the nature of the data by partitions the unity using cumulative distribution function into two parts below and above the mean to obtain measure of skweness that is zero for any symmetric distribution and to three parts based on 50% confidence interval for normal random variable to obtain a measure of data-density or data-concentration that is zero for normal distribution. This type of dot-chart has advantages over histogram and quantile-quantile plot where it does not required bin as histogram or plotting position as QQ-plot..

Keywords: Dot Chart, Histogram, Skewness, QQ-Plot.

Introduction

It is important for data analyst to give a description of a distribution and the natural step after treating location, spread, skewness is to characterize kurtosis. Kurtosis plays important roles in various applications, where it has been used in tests of normality, robustness, outliers, modified tests and estimation, large sample inferences, and other situations. The kurtosis parameter is embedded in many inference problems, for example, the asymptotic variance of process capability indices, coefficient of variation and effect size index depend on kurtosis parameter, among other parameters; see, Balanda and MacGillivray (1988), DeCarlo (1997), Schmid and Trede (2003), Kim and White (2004), Zenga (2006), Lihua and Ahmed (2008) and Fiori (2008). More recently, Fiori and Zenga (2009) gave a very good review for the original of kurtosis. The classical notion of kurtosis $\beta_2 = E\{(X-\mu)^4/\sigma^4\}$ or kurtosis excess $\gamma_2 = \beta_2 - 3$, is given by the standardized fourth central moment. The normal distribution has $\beta_2 = 3$ $(\gamma_2 = 0)$ and the uniform distribution has a flat top, with $\gamma_2 = -1.2$. Values of $\gamma_2 < -1.2$ may suggest that the distribution is bimodal Darlington (1970) but bimodal distributions can have high kurtosis if the modes are distant from the shoulders; see, for example, Hildebrand (1971), Groeneveld and Meeden (1984), Blanda and MacGillivary (1988) and Wang and Serfling (2005).

In this paper the unity of cumulative distribution function is divided to two parts using cumulative distribution function under and above the mean to obtain a measure of skewness that is zero for symmetric distributions and to three parts in terms of 50% confidence interval of normal random variable to obtain a measure of data-concentration that is zero for normal distribution. Based on these partitions an informative chart called dot-chart is presented that can provide more insight into the nature of the data.

In Section 2 the unity is divided to two and three parts based on mean and confidence interval of normal distribution and measures of skewness and data-concentration are introduced. The estimation of skewness and data-concentration measure is presented in Section 3. Section 4 is devoted to the conclusion.

Partitions of one and the shape measures

Let $X_1, X_2, \dots, X_n$ be a random sample from a continuous distribution with density function $f(x)$, quantile function $x(F) = F^{-1}(x) = Q(F)$, $0 < F < 1$, cumulative distribution function $F(x) = F$, mean $\mu = E(X)$, σ is the standard deviation of the distribution where $\int_{-\infty}^{\infty} f(x)dx = 1$, and $X_i \neq \mu, \mu - 0.675\,\sigma$ and $\mu + 0.675\,\sigma$ for $i = 1, \dots, n$. The unity can be partitions to two parts using CDF as

$$1 = F(x < \mu) + F(x > \mu) = F^- + F^+$$

E-mail: qm.group70@gmail.com

Also the unity can be partitions to three parts as

$$1 = F(x < \mu - c\sigma) + F(\mu - c\sigma < x < \mu + c\sigma) + F(x > \mu + c\sigma)$$

The skewness measure can be defined as

$$S = F(x < \mu) - F(X > \mu) = F^{-} - F^{+} = 2F(\mu) - 1$$

This measure is zero for any symmetric distribution and is bounded by -1 and 1. This measure is defined by Groeneveld and Meeden (1984).

The measure S can be shown graphically on the dot-chart that shows the pattern of the data-density or data-concentration on x-axis with respect to the mean by using the function dotchart(x) in R-software. Therefore, F^{-}represents the data-density or data-concentration for the values less than the mean and F^{+}represents the data-concentration for the values more than mean.

Figure 1 shows the dot-chart for 100 observations from uniform, normal, logistic, and Laplace distributions. The graph shows symmetric pattern of right side data-concentration with left side data-concentration from the mean for all distributions.

Moreover, the dot-chart shows short tail and equally distributed data-concentration for the uniform distribution, medium tails with more data-concentration near the mean than uniform for the normal distribution, long tails with more data-concentration near the mean than normal for the logistic and Laplace distributions.

Figure 2 shows the dot-chart for 100 observations from chi-square with 1 and 2 degree of freedom, beta with parameters (1,0.5) and (1,0.15) distributions. The graph shows a degree of asymmetric pattern of right side data-concentration and left side data-concentration from the mean for all distributions. Moreover, the dot-chart shows very strong left side data-concentration and very light right data-concentration for chi-square (1) distribution that indicates very strong right skewed and strong left side data-concentration with more data-concentration near the mean for chi-square (2) distribution that indicates strong right skewed while the dot-chart shows very strong right side data-density and very light left data-density for beta(1,0.15) that indicates very strong left skewed and medium right side data-density with more data-density near the mean for beta (1,0.5) distribution that indicates medium left skewed.

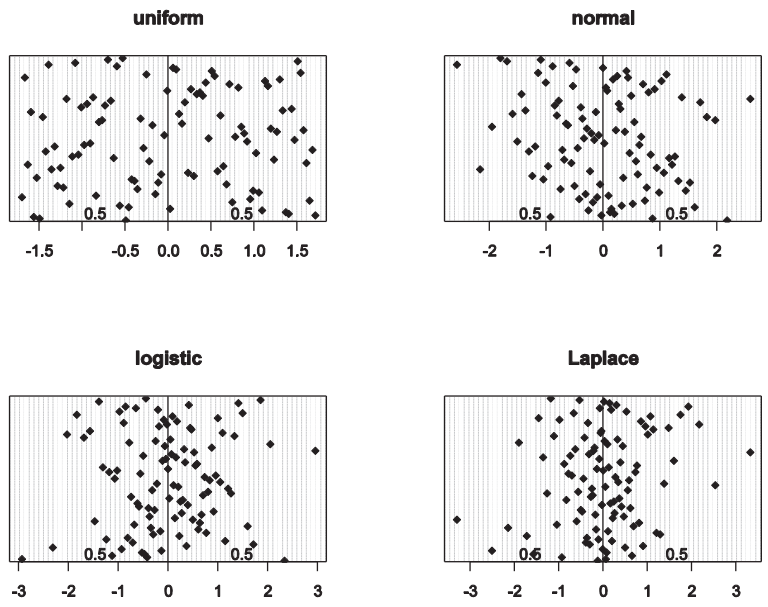

Figure 1. dot-chart with vertical line at mean for 100 standardized observations from uniform, normal , logistic, and Laplace distributions.

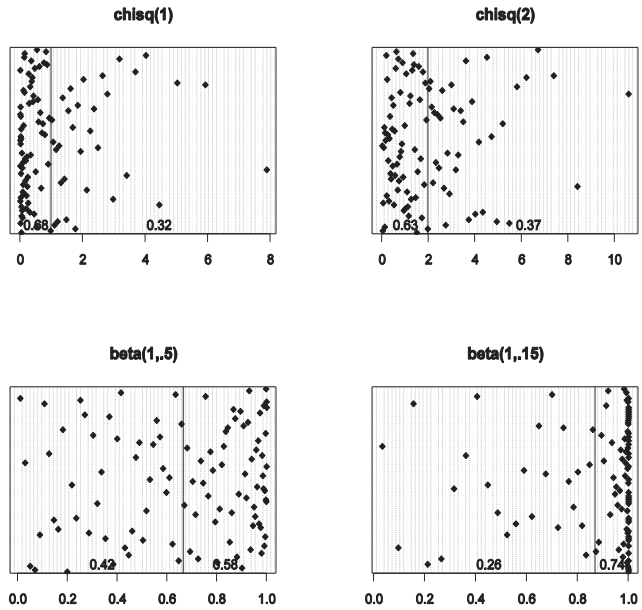

Figure 2. dot-chart with vertical line at mean for 100 observations drawn from chisq(1), chisq(2), beta(1,0.5) and beta (1,0.15) distributions.

The proposed measure for data-concentration is defined as

$$C = C_p - (C_r + C_l)$$

where

$$C_p = \int_{\mu-0.675\sigma}^{\mu+0.675\sigma} f(x)dx = F(\mu + 0.675\sigma) - F(\mu - 0.675\sigma),$$

$$C_r = \int_{\mu+0.675\sigma}^{\infty} f(x)dx = 1 - F(\mu + 0.675\sigma)$$

and

$$C_l = \int_{-\infty}^{\mu-0.675\sigma} f(x)dx = F(\mu - 0.675\sigma)$$

This measure is bounded by -1 and 1 for all distributions and the choice of $c = 0.675$ to obtain data-concentration equal to zero for the benchmark distribution (the normal distribution) or 50% confidence interval for a normal random variable $(\mu - 0.675\sigma,\ \mu + 0.675\sigma)$. Therefore, C_p represents the data-density that concentrated in the middle of the distribution (middle-data-density), C_r represents the data-density in the right side of the distribution (right side data-density), C_l represents the data-density in the left side of the distribution (left side data-density). Moreover the parameter $C_l + C_r$ can be interpreted as the data-density that concentrated in the sides or tails of the distribution (side-data-density) i.e. the C measure compares the side-data-density with middle-data-density in terms of the benchmark distribution, therefore, if $C = 0$, side-data-density equal to middle-data-density, $C > 0$ then middle-data-density is more than side-data-density or lighter sides-density and heavier middle-density than normal and $C < 0$ then middle-data-density is less than sides-data-density or heavier sides-density and lighter middle-density than normal. The measure C can be shown graphically on the dot-chart that shows the pattern of the scatter or density of the data in middle and sides of the distribution. This pattern gives indications about the shape of the distribution.

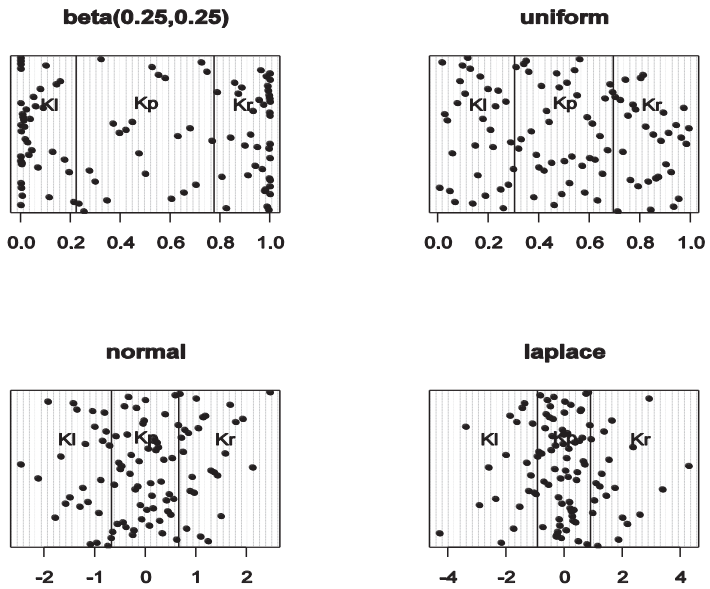

Figure 3. C_p, C_r and C_l on dot-chart for Beta(0.25,0.25) ($C_p = 0.22, C_r = 0.38, C_l = 0.38$ and $C = -0.54$), uniform ($C_p = 0.38, C_r = 0.30, C_l = 0.30$ and $C = -0.22$), normal ($C_p = 0.50, C_r = 0.25, C_l = 0.25$ and $C = 0$) and Laplace ($C_p = 0.61, C_r = 0.19, C_l = 0.19$ and $C = 0.23$) distributions using $n = 100$. Note that C is replaced by K on the graph.

Figure 3 shows the dot-chart for the beta, uniform, normal and Laplace distributions. The graph shows the data-density pattern for each distribution, for beta(0.25,0.25) distribution the sides-data-density is much more than the middle-data-density that indicates strong negative data-concentration (most likely not unimodal distribution), the uniform distribution shows randomly distributed data and the sides-data-density is more than middle-data-density that indicates negative data-concentration, the normal distribution shows more middle-data-density about mean than uniform and the sides-density is equal to middle-data-density that indicates zero data-concentration and the Laplace distribution shows much more middle-data-density about mean than normal with less sides-data-density with long scatter data (long tails) that indicates positive data-concentration.

Table 1. Values of S and C for some symmetric distributions

Set A	S	C	Set B	S	C
Beta(0.25,0.25)	0	-0.541	gl*(0,1,-0.85,-0.85)	0	0.985
Beta(0.5,0.5)	0	-0.366	gl(0,1,-0.75,-0.75)	0	0.969
Uniform	0	-0.220	gl(0,1,-0.5,-0.5)	0	0.762
Beta(1.5,1.5)	0	-0.157	gl(0,1,-0.25,-0.25)	0	0.345
Normal	0	0	gl(0,1,-0.15,-0.15)	0	0.224
Logistic	0	0.091	gl(0,1,-0.10,-0.10)	0	0.174
Laplace	0	0.230	gl(0,1,-0.05,-0.05)	0	0.130
t(5)	0	0.153	gl(0,1,-0.01,-0.01)		

*gl stands for generalized lambda distribution with four parameters;see, Ramberg et al. (1979)

DeCarlo (1997) and others have pointed out that the Laplace distribution is clearly more peaked than the t_5 distribution but $\beta_2 = 6$ for the Laplace and $\beta_2 = 9$ for the t_5. In contrast, $K = 0.230$ for the Laplace and $K = 0.153$ for the t_5 and thus K correctly classifies these distributions according to middle-density which is shown in Table 1 and Figure 4. According to Oja (1981), a valid measure of shape must be location and scale invariant and also must obey van Zwet ordering which rank orders the distributions in Set A of Table 1 from smallest to largest. Parameter C is location and scale invariant and rank the distributions in Set A from smallest to largest. Note that C exists in distributions where mean and variance exist while β_2 exists in distributions where fourth moment exists.

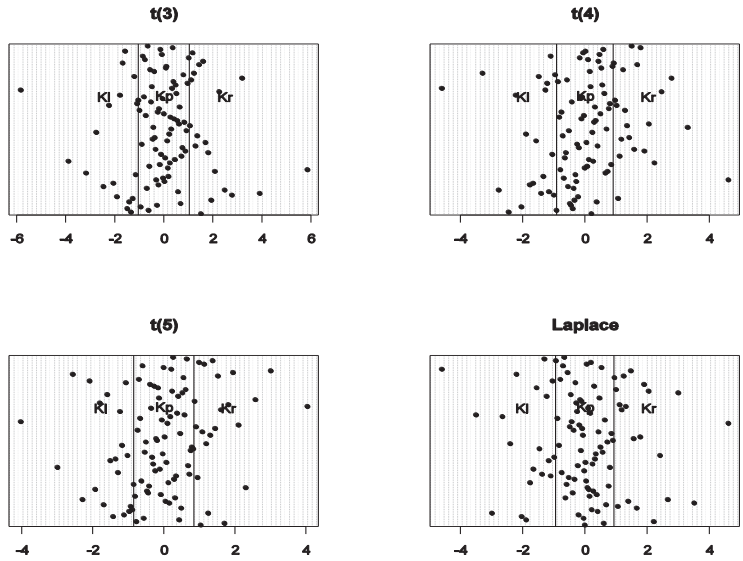

Figure 4. C_p, C_r and C_l on dot-chart for t(3) ($C_p = 0.66, C_r = 0.17, C_l = 0.17$ and $C = 0.32$), t(4) ($C_p = 0.60, C_r = 0.20, C_l = 0.20$ and $C = 0.20$), t(5) ($C_p = 0.57, C_r = 0.21, C_l = 0.21$ and $C = 0.15$) and Laplace ($C_p = 0.61, C_r = 0.19, C_l = 0.19$ and $C = 0.23$) distributions using $n = 100$. Note that C is replaced by K on the graph.

Estimation

We now consider estimators of the population S and C using a random sample of size n, $x_1, x_2, \ldots, x_n$ where $\bar{x}$ is sample mean and s is the sample standard deviation. The estimates are

$$c_p = \frac{1}{n}\sum_{i=1}^{n} I(\bar{x} - 0.675s < x_i < \bar{x} + 0.675s),$$

$$c_r = \frac{1}{n}\sum_{i=1}^{n} I(x_i > \bar{x} + 0.675s)$$

and

$$c_l = \frac{1}{n}\sum_{i=1}^{n} I(x_i < \bar{x} - 0.675s)$$

I is the indicator function.

Hence,

$$s = \frac{2}{n}\sum_{i=1}^{n} I(x_i < \bar{x}) - 1$$

and

$$c = c_p - (c_r + c_l)$$

The empirical mean and variances of these estimates from normal distribution are given in Table 2 using 10000 randomly generated normal samples for each sample size.

Table 2: empirical mean and variances of c from uniform and normal distributions using 10000 replications

	Uniform		Normal	
	c		c	
n	mean	Var.	Mean.	Var.
10	-0.171	0.0571	-0.015	0.05829
20	-0.194	0.0281	-0.0089	0.03020
30	-0.203	0.0189	-0.0006	0.02095
50	-0.211	0.0113	-0.0003	0.01233
75	-0.213	0.0077	-0.0001	0.00849
100	-0.216	0.0057	-0.0008	0.00640
200	-0.219	0.0028	-0.0005	0.00318
500	-0.219	0.0011	0.0005	0.00128
1000	-0.219	0.0006	0.0004	0.00064

From Table 2 and for uniform distribution

$$\hat{v}(k) = 0.6/n$$

For the normal distribution

$$\hat{v}(k) = 0.64/n$$

Figure 5 shows the dot-charts for faithful data from R-software.

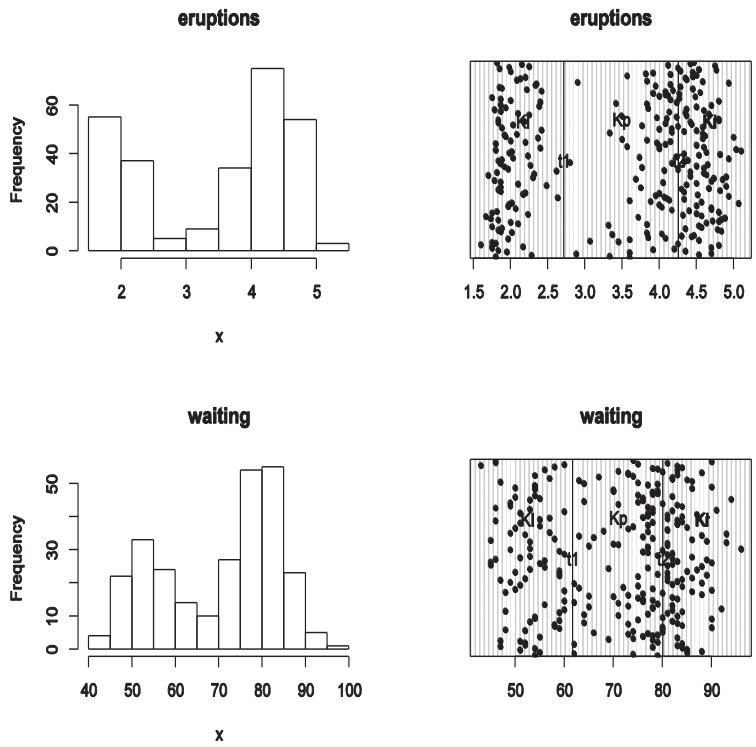

Figure 5. the histogram and dot-chart for faithful data in R-software. the eruption data has $c = -0.42$ ($c_p = 0.29$, $c_l = 0.34$ and $c_r = 0.37$) and waiting data has $c = -0.23$ ($c_p = 0.39, c_l = 0.31, c_r = 0.31$).

Conclusion

A measure of data-concentration is proposed: C_p which reflects information about the middle region and its complement $C_t = 1 - C_p$ which reflects information about the tails of the distribution. It is illustrated that large value of C_p could mean high or wide peak and small value of C_t could mean long or thick tail. Moreover, the proposed measure has flexibility to work on part of the distribution rather than on the entire distribution, for example, if we are interested in the right tail we could work with C_r.

On the other hand, there are limitations of the proposed measure. It can not be used when μ and σ are not finite, the Cauchy distribution provides an example, in that C_p is not defined. Also, what is the best choice of c which distinguish among distributions in terms of C_p. This may need sensitivity analysis.

References

Balanda, K.P. and MacGillivray, H.L. 1988. Kurtosis: A critical review. *American Statistician*. Vol. **42**, pp. 111–119.

Darlington, R.B. 1970. Is kurtosis really "peakedness"? *American Statistician*, Vol. **24**, pp. 19-22.

DeCarlo,L.T. 1997. On the meaning and use of Kurtosis. *Psychological Methods*,Vol. 2, pp. 292-307.

Fiori, A.M. and Zenga,M. 2005. The meaning of kurtosis, the influence function and an early intuition by L. Faleschini. *Statistica*,Vol. **65**, pp. 131–140.

Fiori, A.M. 2008. Measuring kurtosis by right and left inequality orders. *Communication in Statistics – Theory & Methods*, Vol. **37**, pp. 2665–2680.

Fiori, A.M. and Zenga,M. 2009. Karl Pearson and the origin of kurtosis. *International Statistical Review*, Vol. **77**, pp. 44-50.

Groeneveld, R.A. and Meeden, G., 1984. Measuring skewness and kurtosis. *The Statistician*, Vol. **33**, pp. 391–399.

Hildebrand, D.K. 1971. Kurtosis measures bimodality. *American Statistician*, Vol. **25**, pp. 42-43.

Kim, T. and White, H. 2004. On more robust estimation of skewness and kurtosis. *Finance Research Letters*, Vol. **1**, pp. 56-73.

Lihua, A. and Ahmed, S.E. 2008. Improving the performance of kurtosis estimator. *Computational Statistics & Data Analysis*, Vol. **52**, pp. 2669-2681.

Schmid, F. and Trede, M., 2003. Simple tests for peakedness, fat tails and leptokurtosis based on quantiles. *Computational Statistics & Data Analysis*, Vol. **43**, pp. 1–12.

Wang, J. and Serfling, R. 2005. Nonparametric multivariate kurtosis and tailweight measures. *Journal of Nonparametric Statistics*, Vol. **17**, pp. 441-456.

Zenga, M. 2006. Kurtosis. In *Encyclopedia of Statistical Sciences*, 2nd ed. Eds. S. Kotz, C.B. Read, N. Balakrishnan & B. Vidakovic. (Online edition) New York: John Wiley & Sons.