AN ORDER STATISTIC ESTIMATE OF THE SAMPLE STANDARD DEVIATION

Fred E. Cromer*

الخلاصة

ثم العينة من التوزيع الموحد والطبيعي والكاي المربع حصل عليها والعينة قورنت لهذه الاحصائية (العينة).
لقد استنتج أن الفروق بين هذه الاحصائيات (العينة و Standard Deviation) كانت على أصغرها
لعينات عددها من ١٥ الى ٣٠ من التوزيعات التي كانت تقريباً طبيعية .

ABSTRACT

A statistic was devised to estimate the sample standard deviation. Then samples from uniform, normal, and chi-square distributions were generated and the sample standard deviation was compared to this new statistic. It was found that the differences between these statistics were smallest for samples of size 15 to 30 from distributions that were approximately normal.

Quite early in a basic course in data analysis and descriptive statistics, measures of central tendency and dispersion are discussed. The student soon sees that even for very small data sets, the commonly used sample mean and sample variance are rather tedious to compute. For relatively small samples, the median is a very easily found measure of central tendency and can be used as an estimate of the sample mean. This gives one an initial conception of the center of the sampling distribution even before any arithmetic calculations (by hand or machine) are done.

The range of the sample is the most commonly discussed measure of dispersion which is based on order statistics, but it is well known that the range is not a good estimate of the sample standard deviation. Quasi-ranges or linear combinations of qusi-ranges, which may provide more efficient estimators, also increase the amount of computation required [1,3]. Since the sample variance is defined as the average of the squared deviations from the mean, a convenient estimate of the sample variance might be obtained by considering a "typical " score above and a "typical" score below the median and calculating their average squared deviation from the median.

Let y_1 be the minimum score, y_m be the median, and y_n be the maximum. If the midpoint between y_n and y_m is considered to be a "typical" score above the median and the midpoint between y_1 and y_m is considered to be a "typical" score below the median, the statistic c^2 is defined as follows:

$$c^2 = \frac{(y_n - y_m)^2 + (y_1 - y_m)^2}{8} \qquad (1)$$

The square root of c^2 can be used as an estimate of the sample standard deviation. Thus, after ordering a set of data, both a measure of central tendency and a measure of dispersion,which estimates the sample standard deviation, are available with a minimal amount of computation. It should be noted that for large data sets, ordering the sample is itself a major task so in that case little or no savings result in the calculation of such an estimate for the sample standard deviation.

Since c, the square root of c^2, is claimed to be an estimate of s, the sample standard deviation:

$$s = [\sum_{i=1}^{n} (x_i - \bar{x})^2/n]^{\frac{1}{2}}, \qquad (2)$$

it may be instructive to test whether this claim can be verified. A method for testing this claim which is quite instructive, even in the most elementary statistics classes, is an empirical study where data are collected and the statistics c and s are compared. This technique can be useful in further reinforcing many of the ideas covered in a basic descriptive statistics course. Such a Monte Carlo study was performed using randomly generated data sets from uniform, normal, and chi-square distributions.

The uniform distributions considered had ranges from 0 to 10, from 0 to 50, and from 0 to 100 [u(0,10), u(0,50), u(0,100)]. The normal distributions, denoted by $n(\mu,\sigma^2)$, were n(0,1), n(50,100), and n (100,225). The chi-square distributions had parameters 1, 5 and 20 [χ^2 (1), χ^2(5), χ^2 (20)] . In each of these nine cases, 50 samples each of size 5, 10, 15, 20, 25, 30, 40, 50, 75 and 100 were generated and the statistics c and s were compared.

* *College of Industrial Management, University of Petroleum and Minerals, Dhahran, Saudi Arabia*

The samples were generated using the IBM supplied subroutines RANDU and GAUSS [2]. The chi-square samples were obtained by using the fact that the sum of the squares of k independent standard normal variates is a chi-square variate with parameter k.

The results of this Monte Carlo investigation are reported in Tables 1 and 2. Table 1 indicates the behavior of the mean and standard deviation of the sampling distributions of c and s. Table 2 lists the relative errors of estimating s by c where relative error,

$$RE = \bar{d} / \bar{s}, \qquad (3)$$

d is the average of the absolute differences $|s_i - c_i|$ and s is the average value of s among the 50 samples of that given sample size.

Tables 1 and 2 indicate that changes of the parameters for the uniform and normal distributions do not influence RE, but changes in the parameter of the chi-square distribution does influence RE. This is not surprising since as the parameter increases, the chi-square distribution changes from extremely skewed to more symmetric and approaches the normal distribution in the limit. However, Tables 1 and 2 indicate that even for very small parameters of the chi-square, the relationship between RE and sample size is similar to that for the normal distribution.

As would be expected, c underestimates s for samples from a uniform distribution, since sample points are as likely to fall at the extremes as at the center of the distribution and s is sensitive to each of these values, while c uses only one value at each extreme. For both the normal and chi-square distributions c underestimates s for small samples and as sample size increases, c tends to approach s and then becomes an overestimator of s as the sample size gets large. Overall, it appears that for distributions which are approximately normal, even if they are still quite asymmetric, c is a fairly good estimator of s for sample sizes from 10 to 40.

It is also noted (Table 1) that for very small sample sizes and for the uniform distribution in general, it appears that the standard error (standard deviation of the sampling distribution) of c is smaller than the tasndard error of s.

Table 3 lists the correlations between c and s. Note that for very small sample sizes there is nearly a perfect linear relationship between c and s and for all distributions the correlations tend to decrease as sample size increases. It also appears that the linear relationship between c and s is highest for distributions which are skewed and peaked, such as $\chi^2(1)$. As the distributions become flatter and more symmetric the correlations tend to decrease so that the lowest correlations occurred for large samples from a uniform distribution.

For very large sets of data, c does not appear to be a very good estimator of s, but in these cases the ranking of the data is a tedious task in itself, so one might as well calculate the sample mean and variance directly rather than try to estimate them.

One very useful application of this statistic is for a quick analysis of classroom quiz results. Frequently, class size is in the range from 10 to 40 students, and by ranking this small set of scores, determining the median and calculating c, one quickly has an idea about the distribution of the quiz scores.

In summary, the development and discussion of the c-statistic is conceptually quite simple and is helpful in developing the statistical intuition of students when discussed as a measure of dispersion for a sample. Furthermore, it provides a very quick and useful initial estimate of the sample standard deviation which can be used along with the median to provide an idea of the shape of a sample distribution.

REFERENCES

[1] David, H.A., *Order Statistics*, New York: John Wiley & Sons, Inc., 1970.

[2] I.B.M. *System* 360 *Scientific Subroutine Package*, White Plains, New York: International Business Machines Corporation, 1968.

[3] Kendall, M.G. & Stuart, A. *The Advanced Theory of Statistics*, Vol. I, Third Edition, London: Charles Griffin & Company Limited, 1969.

Rcference Code for AJSE Information Retrieval QA 1276 CR 1. Paper Received May 20, 1975.

TABLE 1

Means and (Standard Deviation)* of S and C for Various Distributions

Distribution	Sample Size Population St. Dev.		5	10	15	20	25	30	40	50	75	100
u(0,10)	2.887	C	1.859(.530)	2.148(.312)	2.274(.237)	2.333(.155)	2.339(.141)	2.354(.106)	2.418(.083)	2.414(.089)	2.441(.061)	2.466(.030)
		S	2.526(.658)	2.687(.459)	2.780(.366)	2.825(.314)	2.852(.300)	2.812(.256)	2.880(.203)	2.865(.212)	2.881(.160)	2.900(.150)
u(0,50)	14.434	C	8.82(2.86)	10.73(1.42)	11.57(1.09)	11.46(.91)	11.90(.65)	12.09(.56)	11.99(.46)	12.21(.38)	12.19(.30)	12.36(.18)
		S	12.08(3.76)	13.46(1.99)	14.19(1.75)	14.00(1.62)	14.32(1.30)	14.40(1.16)	14.26(1.08)	14.38(.85)	14.28(.81)	14.46(.63)
u(0,100)	28.868	C	17.85(5.21)	21.36(2.81)	22.83(1.77)	23.05(1.65)	23.39(1.58)	23.97(1.26)	24.07(1.01)	24.13(.85)	24.39(.68)	24.60(.45)
		S	24.71(7.29)	26.33(4.53)	28.07(2.92)	27.88(2.71)	27.90(2.73)	28.90(2.65)	28.51(1.81)	28.24(1.93)	28.54(1.50)	28.78(1.58)
n(0,1)	1.000	C	.630(.199)	.819(.221)	.887(.178)	.957(.157)	.980(.171)	1.014(.167)	1.068(.205)	1.142(.145)	1.170(.117)	1.242(.181)
		S •	.864(.269)	.927(.221)	.965(.184)	.957(.133)	.978(.117)	.958(.123)	.991(.123)	.992(.095)	.982(.074)	.992(.068)
n(50,100)	10.000	C	6.97(2.03)	8.14(2.13)	8.72(1.74)	8.92(1.70)	10.35(1.91)	10.26(1.48)	10.68(1.24)	11.33(1.73)	11.78(1.55)	12.45(1.52)
		S	9.16(2.39)	9.35(1.91)	9.37(1.72)	9.38(1.63)	9.71(1.39)	9.90(1.30)	9.86(.97)	9.78(1.07)	9.91(.82)	9.94(.59)
n(100,225)	15.000	C	9.30(3.59)	11.77(2.72)	12.76(2.50)	14.07(2.90)	14.53(2.65)	15.34(2.90)	16.08(2.39)	16.89(2.45)	18.09(2.19)	19.34(2.33)
		S	12.56(4.61)	13.71(2.58)	13.66(2.43)	14.48(2.45)	14.04(1.90)	14.31(1.87)	14.64(1.80)	14.88(1.69)	15.28(1.32)	14.94(1.00)
χ^2 (1)	1.414	C	.730(.612)	1.129(.583)	1.321(.691)	1.345(.633)	1.607(.600)	1.721(.644)	1.815(.588)	2.090(.635)	2.255(.701)	2.339(.610)
		S	.923(.725)	1.188(.560)	1.173(.501)	1.134(.405)	1.274(.377)	1.253(.392)	1.244(.335)	1.347(.282)	1.332(.227)	1.339(.218)
χ^2 (5)	3.162	C	2.029(.848)	2.513(.866)	2.720(.876)	3.133(.966)	3.296(1.024)	3.290.(711)	3.464(.880)	3.870(.820)	3.953(.776)	4.533(.954)
		S	2.656(1.021)	2.818(.751)	2.775(.722)	2.973(.671)	3.044(.617)	2.951(.458)	2.968(.542)	3.065(.432)	2.966(.352)	3.146(.333)
χ^2 (20)	6.325	C	3.916(1.914)	5.052(1.661)	.738(1.735)	5.725(1.097)	6.324(1.554)	6.684(1.435)	6.874(1.301)	7.077(1.268)	7.531(1.489)	8.104(1.232)
		S	5.213(2.543)	5.783(1.625)	5.932(1.282)	5.856(.880)	5.985(1.090)	6.123(.971)	6.059(.749)	6.046(.654)	6.079(.651)	6.285(.556)

**For each sample size the mean of the statistic is listed followed by the standard deviation (in parentheses).*

TABLE 2

Relative Errors In Estimating s by c

Sample Size Distribution	5	10	15	20	25	30	40	50	75	100
u(0,10)	.264	.201	.182	.174	.180	.163	.161	.158	.153	.150
u(0,50)	.270	.203	.185	.181	.169	.161	159	.151	.146	.145
u(0,100)	.278	.189	.187	.173	.162	.171	.156	.145	.145	.145
n(0,1)	.271	.129	.104	.094	.091	.086	.118	.153	.192	.252
n(50,100)	.239	.141	.101	.110	.099	.073	.094	.169	.192	.252
n(100,225)	.259	.155	.102	.070	.087	.108	.112	.137	.184	.295
$\chi^2(1)$	.209	.116	.177	.213	.265	.377	.461	.552	.693	.747
$\chi^2(5)$	.236	.126	.111	.141	.134	.155	.187	.266	.333	.441
$\chi^2(20)$	.249	.138	.110	.095	.114	.118	.146	.177	.241	.290

TABLE 3

Correlations Between c and s

Sample Size Distribution	5	10	15	20	25	30	40	50	75	100
u(0,10)	.92	.81	.66	.60	.71	.62	.30	.68	.54	.36
u(0,50)	.97	.81	.66	.64	.62	.43	.42	.32	.59	.30
u(0,100)	.95	.76	.78	.65	.64	.56	.48	.34	.50	.40
n(0,1)	.97	.90	.87	.74	.73	.81	.72	.57	.67	.69
n(50,100)	.96	.88	.85	.75	.82	.83	.78	.76	.64	.63
n(100,225)	.97	.87	.83	.90	.78	.75	.72	.73	.64	.62
$\chi^2(1)$	.99	.96	.96	.93	.92	.91	.86	.84	.83	.77
$\chi^2(5)$	.97	.95	.91	.86	.88	.80	.82	.77	.73	.67
$\chi^2(20)$	.98	.94	.89	.78	.85	.78	.77	.76	.71	.55