The Technical Whitepaper Series

2

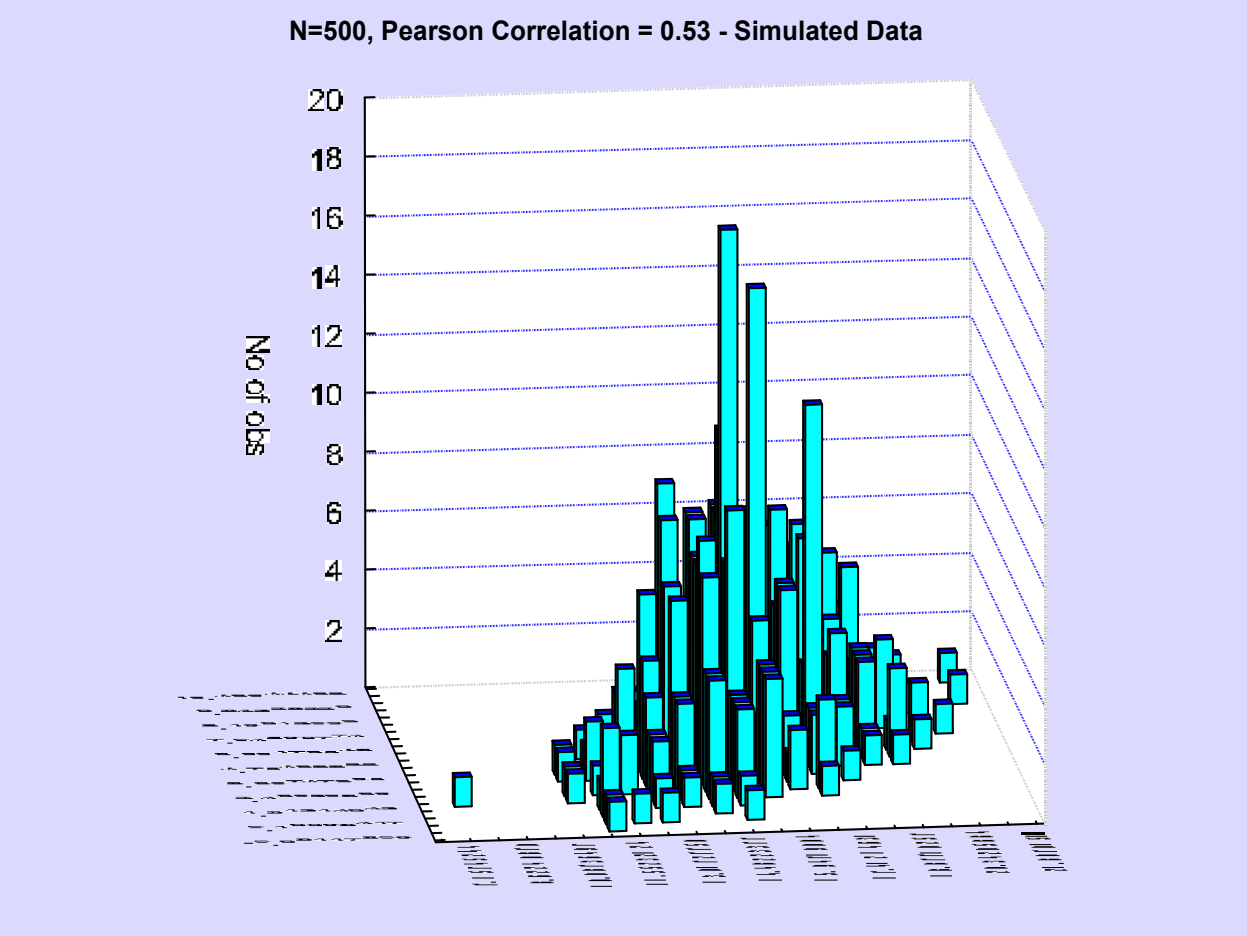
Skewness & Pearson Correlations

Attenuation of coefficient size as a function of skewed data

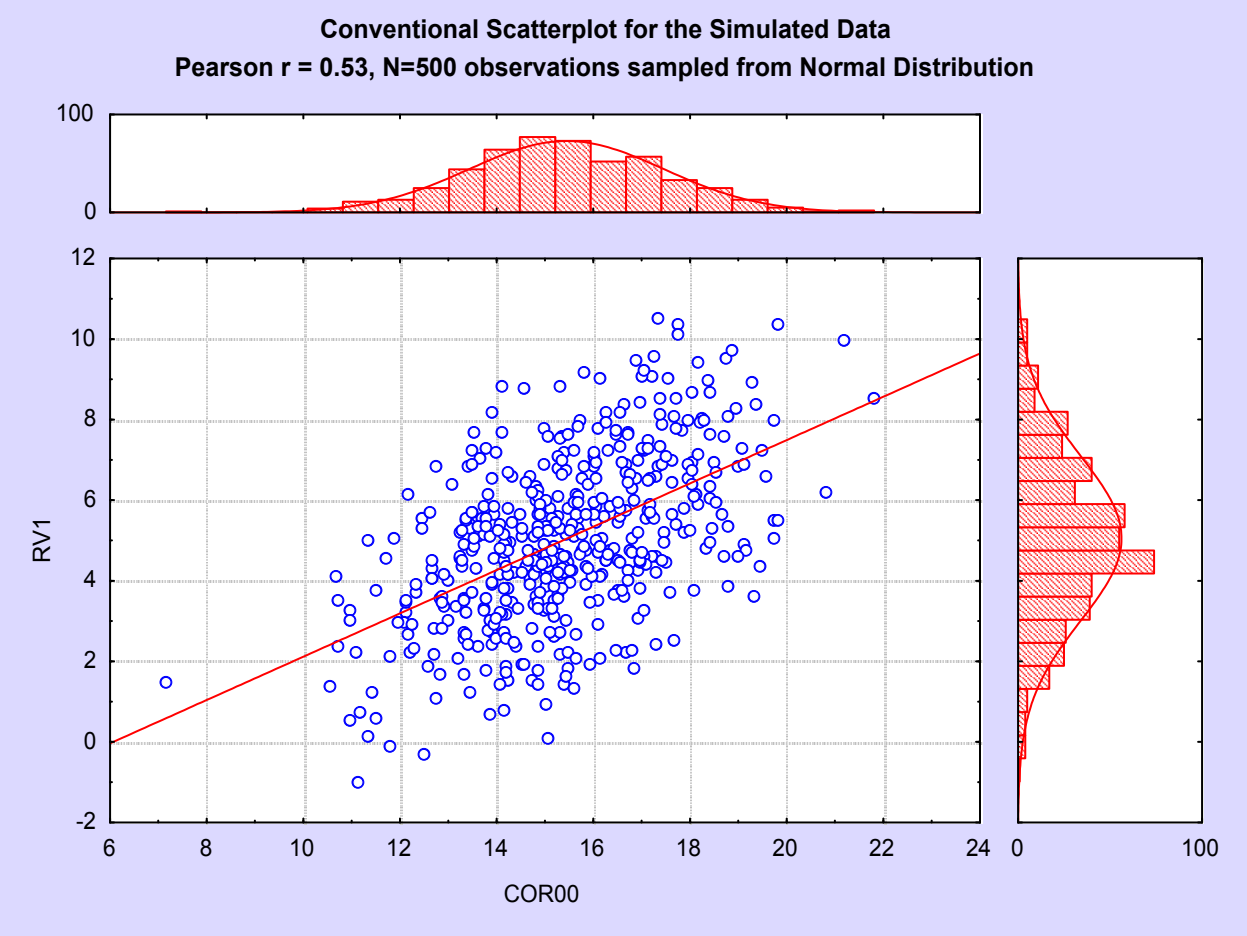
pbarrett.net
Advanced Projects R&D

July, 2001

Why is Skewness and distribution asymmetry of scale scores so important?
This is because as indicated in Kendall and Stuart (1958), using variables with skewness above |2.0| can cause problems with attenuation bias (a reduction in the "true" size) of product moment (Pearson) correlation coefficients. Skewed distributions can occur for many reasons, and also be associated with restriction of measurement range, outliers, small biased samples, and other sample irregularities. It is useful here to demonstrate the effect of increasing skew in one variable's observations, on a correlation computed between two normally distributed variables. I generated data (500 observations) for two variables, sampling from a perfect normal distribution in each case. I then transformed one of the variables such that it would correlate with the first variable at around 0.50 (in fact, 0.53 in this instant). The bivariate histogram of the data is



The conventional Scatterplot is presented over the page ..



COR00 is the random normal variable (RV2) suitably transformed to correlate at 0.53 with the random normal variable RV1.

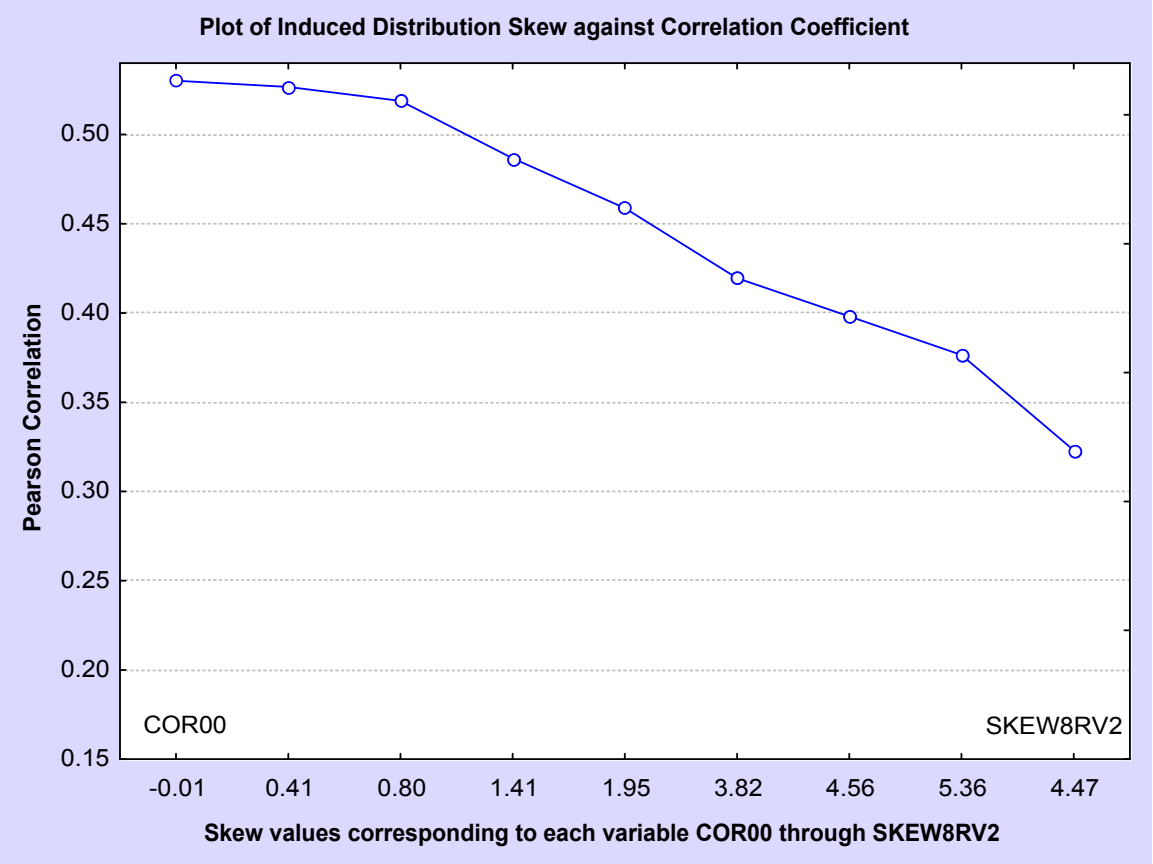
Now what we do is to gradually induce a positive (right directional) skew in the COR00 variable data – by expressing its values in increasing powers. Specifically, we computed powers of 2, 3, 6, 8, 9, 10, and 11, with a dual power scaling as the last series where we raised the 6^{th} power dataset to the 3^{rd} power. This latter operation was implemented to show that a more complex skew generator equation can yield a moderate skewness value yet have drastic attenuation effects on a correlation coefficient. Anyway, the table of statistics of the generated datasets is:

Variable	Descriptive Statistics (Skewness.sta)									
	Valid N	Mean	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Std.Dev.	Skewness	Kurtosis
RV1	500	5.02	4.92	-0.9812	10.5	3.700	6.41	2.03	0.068486	-0.08002
COR00	500	15.40	15.31	7.1575	21.8	14.084	16.81	2.00	-0.011222	0.28285
SKEW1RV2	500	241.14	234.27	51.2291	475.5	198.358	282.53	61.96	0.414209	0.26163
SKEW2RV2	500	3836.70	3585.69	366.6700	10370.2	2793.658	4748.94	1475.16	0.801509	0.91508
SKEW3RV2	500	166.94	128.57	1.3445	606.6	78.045	225.53	125.20	1.416908	1.83378
SKEW4RV2	500	4756.80	3012.04	6.8876	25000.0	1548.087	6371.79	4828.40	1.945630	3.99722
SKEW5RV2	500	84.03	46.10	0.0493	1115.2	21.803	107.10	109.28	3.824841	23.68850
SKEW6RV2	500	1470.53	705.63	0.3528	24319.9	307.075	1800.23	2205.49	4.559540	32.94284
SKEW7RV2	500	26034.86	10800.26	2.5255	530345.0	4324.827	30259.48	44751.03	5.361355	44.29127
SKEW8RV2	500	1575.51	212.54	0.0002	30000.0	47.538	1147.09	3913.70	4.467052	23.34413

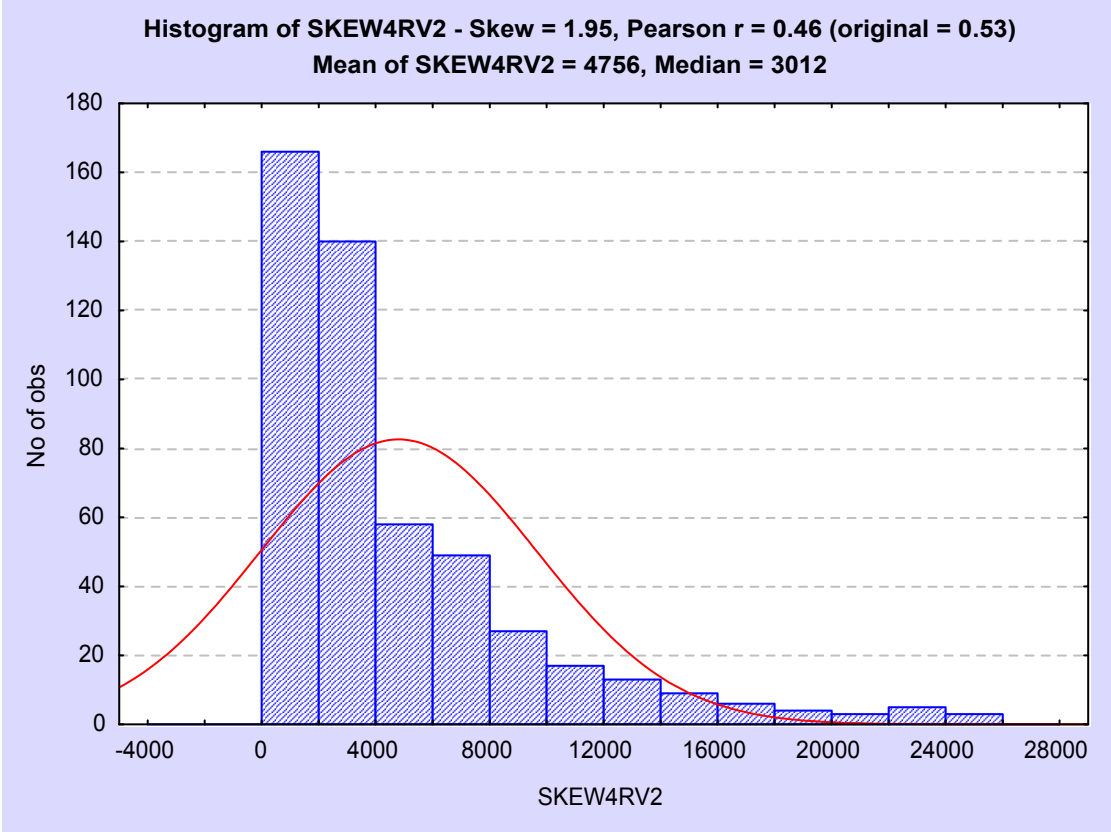
The variables above correspond to the power transformation (scaled to keep the numbers reasonable, with excessive outliers caused by the power transformation trimmed to keep the distributions reasonably tapered) ..

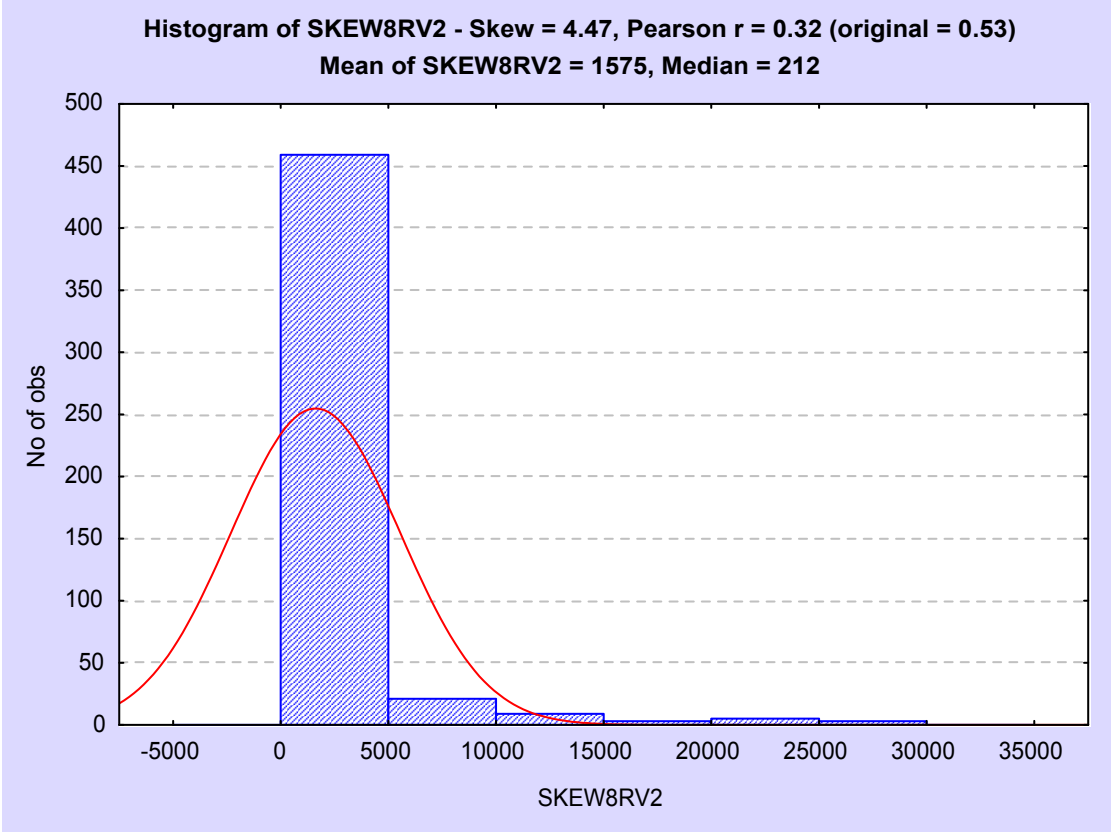
COR00 = no transformation
SKEW1RV2 = power of 2 of COR00 values
SKEW2RV2 = power of 3 of COR00 values
SKEW3RV2 = power of 6 of COR00 values
SKEW4RV2 = power of 8 of COR00 values
SKEW5RV2 = power of 9 of COR00 values
SKEW6RV2 = power of 10 of COR00 values
SKEW7RV2 = power of 11 of COR00 values
SKEW8RV2 = power of 3 of SKEW3RV2 values

Plotting the size of correlation between RV1 and the variables COR00 through SKEW8RV2



As can be seen from the above graph, skewness above about 4.0 causes up to 20% attenuation of the original correlation of 0.53. The histograms of a couple of these skewed variables are:


Histogram of SKEW4RV2 - Skew = 1.95, Pearson r = 0.46 (original = 0.53)
Mean of SKEW4RV2 = 4756, Median = 3012
No of obs
180
160
140
120
100
80
60
40
20
0
-4000
0
4000
8000
12000
16000
20000
24000
28000
SKEW4RV2


Histogram of SKEW8RV2 - Skew = 4.47, Pearson r = 0.32 (original = 0.53)
Mean of SKEW8RV2 = 1575, Median = 212
No of obs
500
450
400
350
300
250
200
150
100
50
0
-5000
0
5000
10000
15000
20000
25000
30000
35000
SKEW8RV2

As can be seen from the graphs above, as the skew increases, so invariably does the Mean – Median disparity (in a symmetrical distribution, these will be equal). So, when we compile our exception report, we report on both Mean-Median disparity and Skewness. Within the analyses below, we will use a hard criterion from our Skewness reporting. That is, we report values above |3.0| as exceptions. This is in addition to those cases where a median is 0 or near zero (which indicates that virtually no-one has scored greater than 0 on a particular scale score).

Reference

Kendall, M. G., & Stuart, A. 1958. *The advanced theory of statistics* New York: Hafner