2015 ■ vol. 11 ■ no. 1

DOI: 10.20982/tqmp.11.1.p032

The standard error of the Pearson skew

Bradley Harding [a], Christophe Tremblay [a], Denis Cousineau [a]

[a] École de psychologie, Université d'Ottawa

Abstract ■ The Pearson skew is a measure of asymmetry of a distribution, based on the difference between the mean and the median of a distribution. Here we show how to calculate the Pearson skew, estimate its standard error and the confidence interval. The derivation is based on a population following a normal distribution. Simulations explored the validity of this expression when the normality assumption is met in comparison to when the normality assumption is not met. The standard error of the Pearson skew revealed very robust in case of non-normal populations, compared to the Fisher Skew as presented in Harding, Tremblay & Cousineau (2014).

Keywords ■ Standard error; Pearson skew; descriptive statistics; shape of a distribution

denis.cousineau@uottawa.ca

Introduction

The Pearson skew (Sk_P) is an alternative measure for the skewness of a sample (noted as Sk_4 in Kim and White, 2003; also see Kendall and Stuart, 1983). It is a measure of the asymmetry in a data set based on the discrepancy between the mean and the median, standardized with a division by the standard deviation. Sk_P is therefore given by the following:

$$Sk_P = \frac{\overline{\mathbf{X}} - \tilde{\mathbf{X}}}{s_{\mathbf{X}}} \qquad (1)$$

where $\overline{\mathbf{X}}$ is the mean of the sample **X** of sample size *n*, $\tilde{\mathbf{X}}$ is the median, and $s_{\mathbf{X}}$ is the standard deviation of the sample.

The Pearson skew is an alternative to the Fisher skew; it is also more robust than the Fisher skew since it is less affected by the presence of outliers (see Daszykowski, Kaczmarek, Vander Heyden & Walczak, 2007). Table 1 provides the theoretical value of the Pearson skew for some commonly used distributions.

To be useful, a statistic must be accompanied by its standard error and confidence intervals (Harding, Tremblay & Cousineau, 2014). We hereby provide the expression for these assuming that the population is normally distributed and the sample size is large, as is common practice. We present the results first, followed by their derivations. Finally, we will present Monte Carlo experiments that confirm the results and their limits.

The standard error of the Pearson Skew is given by:

$$SE_{Sk_P} = \hat{\sigma}_{Sk_P} = \frac{1}{\sqrt{n}}\sqrt{\frac{\pi}{2} - 1} \qquad (2)$$

and the $1-\alpha$ confidence interval is obtained with:

$$\mathrm{CI}_{1-\alpha} = Sk_P + SE_{Sk_P} \times t_{n-1}(\alpha/2, 1-\alpha/2) \quad (3)$$

using the same notation as in Harding et al. (2014).

Demonstration

We know that the standard error of the mean is $\sigma_{\overline{\mathbf{X}}} = \sigma_{\mathbf{X}}/\sqrt{n}$ and the standard error of the median is $\sigma_{\tilde{\mathbf{X}}} = \sqrt{\pi/2}\,\sigma_{\mathbf{X}}/\sqrt{n}$ where $\sigma_{\mathbf{X}}$ is the population's standard deviation. We need the covariance between the mean and the median of a sample. From Ferguson (2003), we have that the covariance is asymptotically given by

$$\sigma^2_{\overline{\mathbf{X}},\tilde{\mathbf{X}}} = \frac{E\,|\mathbf{X} - \nu|}{2\,n\,f(\nu)}$$

where ν is the population true median, f is the density function of the population and $|x|$ is the absolute value of x. Assuming a normally distributed population with parameter $\mu = \nu$ and $\sigma_{\mathbf{X}}$, we find that $f(\nu) = 1/(\sqrt{2\,\pi}\,\sigma_{\mathbf{X}})$; as the distribution of $|\mathbf{X} - \nu|$ is

Table 1 ■ Theoretical values of the Pearson skew for some commonly used distributions.

Distribution	Pearson skew	Fisher skew
Normal	0	0
Student t*	0	0
Exponential	$1-\log(2)\approx 0.3069$	2
Weibull**	$\frac{\Gamma\left(1+\frac{1}{\gamma}\right)-\sqrt{\gamma}\log(2)}{\sqrt{\Gamma\left(\frac{\gamma+2}{\gamma}\right)-\Gamma\left(1+\frac{1}{\gamma}\right)^2}}$	$\frac{2\Gamma(1+\frac{1}{\gamma})-3\Gamma(1+\frac{1}{\gamma})\Gamma(1+\frac{2}{\gamma})+\Gamma(1+\frac{3}{\gamma})}{\left(\Gamma(1+\frac{2}{\gamma})-\Gamma(1+\frac{1}{\gamma})\right)^{3/2}}$
Lognormal	$\frac{e^{\mu+\frac{\sigma^2}{2}}-e^{\mu}}{\sqrt{\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}}}$	$\sqrt{e^{\sigma^2}-1}(2+e^{\sigma^2})$
Gumbel***	$-\frac{\sqrt{6}}{\pi}(\log(\log(2))+C)\approx -0.1643$	$\frac{-12\sqrt{6}\zeta(3)}{\pi^3}\approx -1.1396$

Note: Pearson skew tends to be approximately 6 times smaller than Fisher skew.
There is no closed-form expression for the median of the Wald and the Ex-Gaussian distributions.
*: d.f. must be greater than two for the Pearson skew and greater than 3 for the Fisher skew
**: γ is the shape parameter and Γ is the Gamma function.
***: C is Euler gamma ≈ 0.5772 and ζ is the Rieman zeta function, with $\zeta(3)\approx 1.2012$.

half-normal (Forbes, Evans, Hastings, & Peacock, 2010; Leemins, & McQuestion, 2008), the expected value of $|\mathbf{X}-\nu|$ is $\sqrt{2/\pi}\,\sigma_{\mathbf{X}}$. Hence,

$$\sigma^2_{\overline{\mathbf{X}},\tilde{\mathbf{X}}}=\frac{\sqrt{\frac{2}{\pi}}\sigma_{\mathbf{X}}}{2\,n\frac{1}{\sqrt{2\pi}\sigma_{\mathbf{X}}}}=\frac{\sigma^2_{\mathbf{X}}}{n}.$$

Using a Taylor series expansion (e. g., Ku, 1966), we have that

$$\begin{aligned}\sigma^2_{\overline{\mathbf{X}}-\tilde{\mathbf{X}}}&\approx\sigma^2_{\overline{\mathbf{X}}}+\sigma^2_{\tilde{\mathbf{X}}}-2\sigma^2_{\overline{\mathbf{X}},\tilde{\mathbf{X}}}\\&=\frac{\sigma^2_{\mathbf{X}}}{n}+\frac{\sigma^2_{\mathbf{X}}}{n}\frac{\pi}{2}-2\frac{\sigma^2_{\mathbf{X}}}{n}\\&=\frac{\sigma^2_{\mathbf{X}}}{n}\left(\frac{\pi}{2}-1\right)\end{aligned}$$

which is the squared standard error of the numerator of the Pearson skew.

As of the denominator, the standard error of $s_{\mathbf{X}}$ is $\sigma_{s_{\mathbf{X}}}=\sigma_{\mathbf{X}}/\sqrt{2(n-1)}$. Using again a Taylor series expansion, for a ratio this time, we have that

$$\begin{aligned}\sigma^2_{(\overline{\mathbf{X}}-\tilde{\mathbf{X}})/s_{\mathbf{X}}}\approx&\left(\frac{\overline{\mathbf{X}}-\tilde{\mathbf{X}}}{s_{\mathbf{X}}}\right)^2\times\\&\left(\frac{\sigma^2_{\overline{\mathbf{X}}-\tilde{\mathbf{X}}}}{(\overline{\mathbf{X}}-\tilde{\mathbf{X}})^2}+\frac{\sigma^2_{s_{\mathbf{X}}}}{s^2_{\mathbf{X}}}-2\frac{\sigma^2_{\overline{\mathbf{X}}-\tilde{\mathbf{X}},s_{\mathbf{X}}}}{(\overline{\mathbf{X}}-\tilde{\mathbf{X}})\times s_{\mathbf{X}}}\right)\end{aligned}$$

The last covariance term, $\sigma^2_{\overline{\mathbf{X}}-\tilde{\mathbf{X}},s_{\mathbf{X}}}$, is based on the correlation between the numerator and the denominator of the Pearson skew. Yet, assuming a normal distribution, these two terms are uncorrelated. Thus,

$$\begin{aligned}\sigma^2_{(\overline{\mathbf{X}}-\tilde{\mathbf{X}})/s_{\mathbf{X}}}&=\frac{\sigma^2_{\overline{\mathbf{X}}-\tilde{\mathbf{X}}}}{s^2_{\mathbf{X}}}+\frac{\sigma^2_{s_{\mathbf{X}}}}{s^2_{\mathbf{X}}}\times Sk_P^2-0\\&=\frac{\frac{\sigma^2_{\mathbf{X}}}{n}\left(\frac{\pi}{2}-1\right)}{s^2_{\mathbf{X}}}+\frac{\frac{\sigma^2_{\mathbf{X}}}{2(n-1)}}{s^2_{\mathbf{X}}}\times Sk_P^2\end{aligned}$$

An estimate is obtained by replacing $\sigma^2_{\mathbf{X}}$ by the observed variance $s^2_{\mathbf{X}}$. Further, note that the second term is expected to be close to zero (as $E(Sk_P)=0$) so that we can ignore it. Thus,

$$\hat{\sigma}^2_{Sk_P}=\frac{1}{n}\left(\frac{\pi}{2}-1\right)$$

and the standard error is found in Eq. 2.

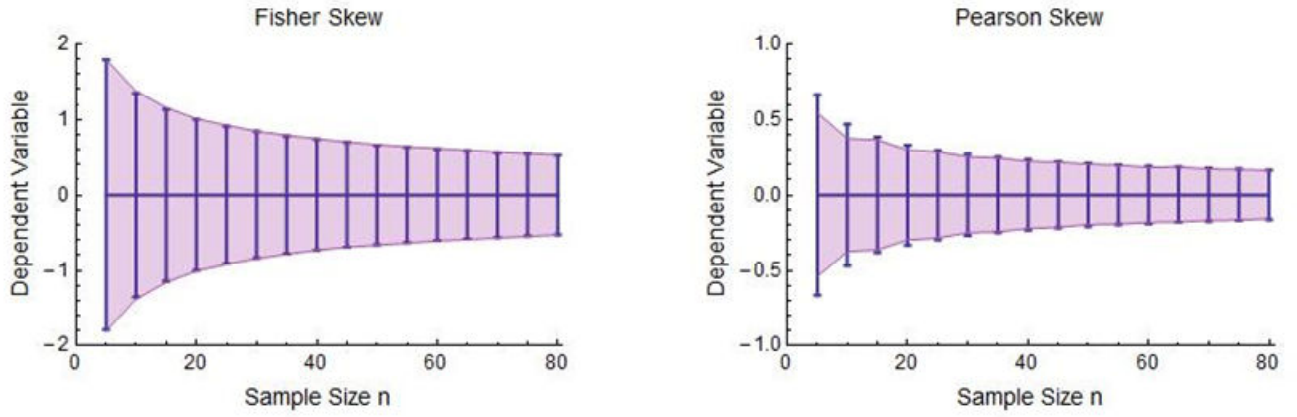
Regarding the distribution of Sk_P, knowing that the distribution of the sample mean of a normal population is also normal, that the distribution of the sample median is well approximated by a normal population as well, the numerator of Sk_P is normal with mean zero. The denominator is related to the χ^2 distribution and thus, the ratio has student t distribution with $n-1$

Table 2 ■ Summary of the standard error for the Pearson Skew in the same format presented in Harding et al. (2014)

Descriptive Statistics	Equation	Standard Error	Confidence interval
Pearson skewness*	$Sk_P=\frac{\overline{\mathbf{X}}-\tilde{\mathbf{X}}}{s_{\mathbf{X}}}$	$SE_{Sk_P}=\hat{\sigma}_{Sk_P}=\frac{1}{\sqrt{n}}\sqrt{\frac{\pi}{2}-1}$	$\mathrm{CI}_{1-\alpha}=Sk_P+SE_{Sk_P}\times t_{n-1}(\alpha/2,1-\alpha/2)$

Note: Estimates that assume a normally distributed population are marked by an asterisk (*);

Figure 1 ■ Mean estimated Fisher and Pearson skewness (horizontal blue line) as well as estimated (error bars) and actual (shaded area) 95% confidence intervals as a function of sample size. Each point is based on 50,000 data points sampled from a normal distribution with a true mean of 100 and a true standard deviation of 3 (replicating the simulations presented in Harding et al., 2014). The Fisher skew (as presented in Harding et al. , 2014) is found in the left panel whereas the Pearson skew, studied here, is found in the right panel.



degrees of freedom. Consequently, a confidence interval for the Pearson's skew is found using *t* critical values, as was indicated in Eq. 3.

Table 2 summarizes the relevant equations for the measure of the Pearson Skew following the same layout as Table 1 used in Harding et al. (2014).

Monte Carlo experiments

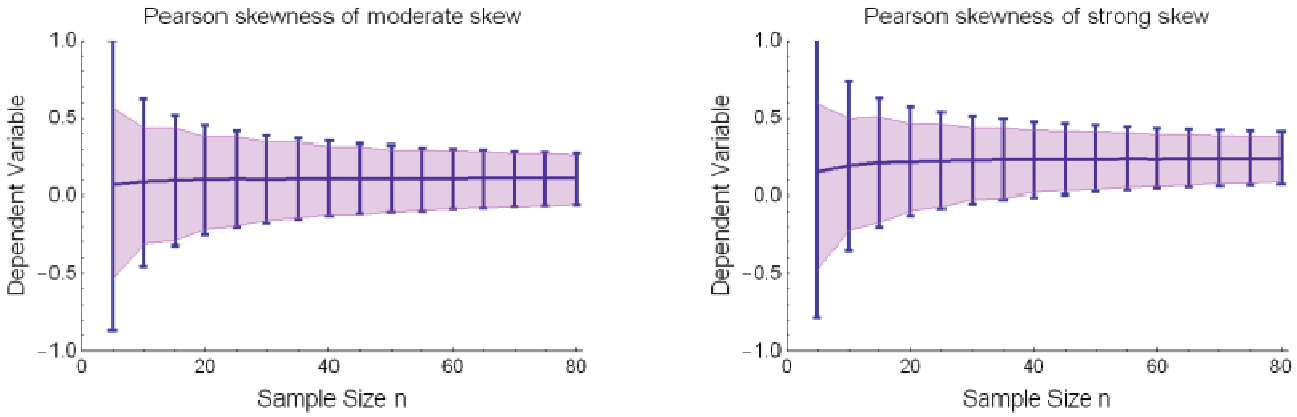
To verify the reliability of the Pearson skew standard error estimator we compared the actual standard error (the standard deviation of the Pearson skew over a large number of simulated samples) to the estimated standard error (measured using Eq. 2). We have followed the same methodology utilized in Harding et al. (2014) by simulating 50,000 random samples taken from a normal distribution with parameters $\mu = 100$, and $\sigma = 3$. We also varied the sample size by increasing *n* by increments of 5 from a very small sample ($n = 5$) to a fairly large sample ($n = 80$).

Figure 1 presents the results of the simulation for the actual and estimated standard error of the Pearson skew as compared to Fisher skew for increasing sample sizes. As seen, the results of the estimated standard error are very reliable when compared to the actual standard error. Note that for smaller sample sizes (below $n = 10$), the 95% confidence interval for Pearson skew (Eq. 2) slightly overestimates actual standard error values. However, larger sample sizes lead to reliable results. Note that although the Pearson skew and the Fisher skew are both interpreted by their relation to zero (a skewness of zero is given to a symmetrical distribution such as the normal distribution), the scales are different. The Pearson skew is measured with a smaller scale than the Fisher skew although they are interpreted the same. For comparison purposes, simulations on Fisher skew were added in Figure 1 (See Harding et al. for more information on the Fisher skew and its standard error). Unlike the Pearson skew, the confidence interval for Fisher skew is unaffected by smaller sample sizes.

Case example in which the normality of distribution assumption is violated

In the present section, we verify the reliability of the standard error estimator for the Pearson skew when the normality assumption is not met. The objective of this example is to see if a normally distributed population is required to use the standard error estimator of the Pearson skew as is the case for the Fisher skew. We sampled 50,000 data points from a Weibull population distribution with a scale parameter of $\beta = 60$, a shift parameter of $\alpha = 300$. We used a shape parameter of either $\gamma = 2$ or $\gamma = 1.25$ to observe the reliability when the normality assumption is not met and when it is violated outright. These simulation details replicate those used by Harding et al. (2014) in

Figure 2 ■ Mean estimated Pearson skew (horizontal blue line) as well as estimated (error bars) and actual (shaded area) 95% confidence intervals as a function of sample size. The distribution used is a Weibull distribution with a scale parameter of $\beta = 60$, a shift parameter of $\alpha = 300$ and a shape parameter of either $\gamma = 2$ or $\gamma = 1.25$ (simulation details are identical to those found in Appendix B of Harding et al., 2014). The Pearson skew of a moderately skewed distribution ($\gamma = 2$) is found in the left panel whereas the Pearson skew of a strongly skewed distribution ($\gamma = 1.25$) is found in the right panel.



Appendix B of their review of standard error estimators and their confidence intervals. Figure 2 presents the results of the simulation.

As seen, when the population's distribution is moderately skewed (left panel) the standard error estimator of the Pearson skew slightly overestimates skewness for sample sizes smaller than $n = 40$. For sample sizes larger than $n = 40$ the standard error estimator of the Pearson skew seems to estimate the actual standard error quite accurately. From this simulation alone we could warrant the use of the Pearson skew when the population's distribution is moderately skewed. When the population's distribution is strongly skewed (right panel), the Pearson skew standard error estimator consistently overestimates the actual standard error of positive skewness. However, negative skews seem to be consistently accurate as sample size grows (especially after $n = 30$). From these simulations we could advocate for the use of the Pearson skew when the population's distribution is not normal: the estimator is quite reliable and gains reliability as the sample size grows, the estimator slightly overestimates the actual result by a consistent margin, and the estimated values converge towards the actual standard error values.

We also repeated this simulation with the Fisher skew to verify how this measure of skewness estimates the standard error when the population's distribution is not normal. Figure 3 shows the results of this simulation (details of the simulation are identical as the one presented above for the Pearson skew). As is seen, when the distribution is moderately skewed, positive skews are slightly underestimated and negative skews are overestimated. When the distribution is strongly skewed (right panel of Figure 3), the estimated standard error underestimates greatly positive skews yet seems to estimate negative skews somewhat accurately (although it begins to underestimate negative skews as the sample size grows). What is important to note however is that although the standard error for the Fisher skew converges to a smaller interval, the actual standard error for the Fisher skew does not follow suit. Actual standard error for the Fisher skew does not follow the same trend as its estimator calculates as it requires a normally distributed population.

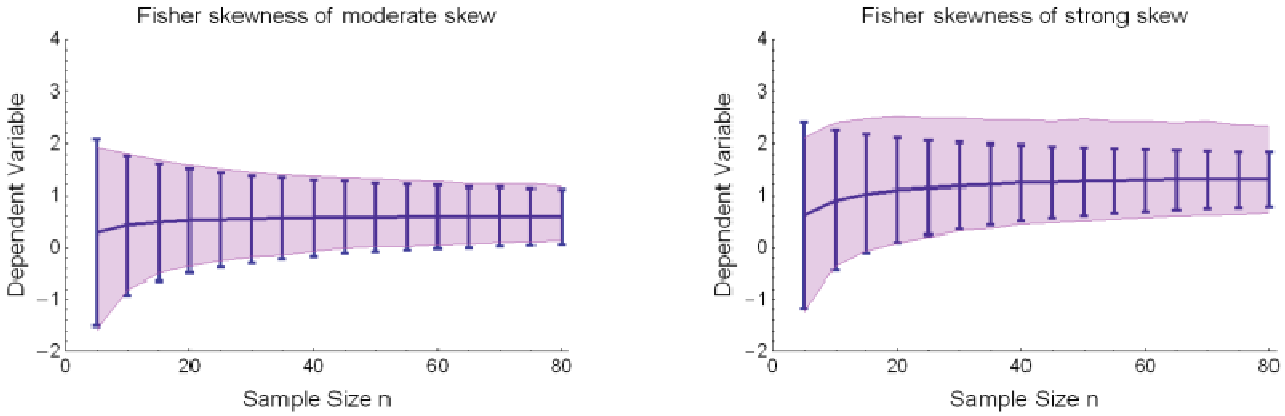
Based on these simulations we can conclude that the use of the Pearson skew in a situation where the population is not normally distributed is preferred over the use of the Fisher skew.

References

Daszykowksi, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis - A review: Basic concepts. *Chemometrics and Intelligent Laboratory Systems*, *85*, 203-219. doi: 10.1016/j.chemolab.2006.06.016

Ferguson, T. S. (2003). *Asymptotic Joint Distribution of Sample Mean and a Sample Quantile*, Internet resource found at http://www.math.ucla.edu

Figure 3 ■ Mean estimated Fisher skew (horizontal blue line) as well as estimated (error bars) and actual (shaded area) 95% confidence intervals as a function of sample size. Simulation details are identical to the ones presented in Figure 2. The Fisher skew of a moderately skewed distribution ($\gamma = 2$) is found in the left panel whereas the Fisher skew of a strongly skewed distribution ($\gamma = 1.25$) is found in the right panel.



/~tom/papers/unpublished/meanmed.pdf, last consulted 24 November 2014.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2010). *Statistical Distributions*. New York: Wiley.

Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations. *The Quantitative Methods for Psychology*, *10*, 107-123.

Kendall, M.G. & Stuart, A. (1983). *The advanced theory of statistics*. London: C. Griffin.

Kim, T.-H., & White, H. (2003). *On More Robust Estimation of Skewness and Kurtosis: Simulation and Application to the S&P500 Index*, Internet resource found at http://www.cirano.qc.ca/realisations/grandes_conferences/methodes_econometriques/white.pdf, last consulted 24 April 2012.

Ku, H. H. (1966). Notes on the use of propagation of error formulas. Journal of Research of the National Bureau of Standards - C. Engineering and instrumentation, 70C, 263-273.

Leemins, L. M., & McQuestion, J. T. (2008). Univariate distribution relationships. *The American Statistician*, *62*, 45-53. doi: 10.1198/000313008X270448

Citation

Harding B., Tremblay, C., & Cousineau, D. (2015). The standard error of the Pearson skew. *The Quantitative Methods for Psychology*, *11* (1), 32-36.

Received: 14/12/14