Entry from Lovric, M., Ed. (2010). *International Encyclopedia of Statistical Science*. New York: Springer.

Skewness

Paul von Hippel, Ohio State University, USA

Skewness is a measure of distributional asymmetry. Conceptually, skewness describes which side of a distribution has a longer tail. If the long tail is on the right, then the skewness is rightward or positive; if the long tail is on the left, then the skewness is leftward or negative. Right skewness is common when a variable is bounded on the left but unbounded on the right. For example, durations (response time, time to failure) typically have right skewness since they cannot take values less than zero; many financial variables (income, wealth, prices) typically have right skewness since they rarely take values less than zero; and adult body weight has right skewness since most people are closer to the lower limit than to the upper limit of viable body weight. Left skewness is less common in practice, but it can occur when a variable tends to be closer to its maximum than its minimum value. For example, scores on an easy exam are likely to have left skewness, with most scores close to 100% and lower scores tailing off to the left. Well-known right-skewed distributions include the Poisson, chi-square, exponential, lognormal, and gamma distributions. We are not aware of any widely used distributions that always have left skewness, but there are several distributions that can have either right or left skew depending on their parameters. Such ambidextrous distributions include the binomial and the beta.

Mathematically, skew is usually measured by the third standardized moment $E\left((X-\mu)/\sigma\right)^3)$, where X is a random variable with mean μ and standard deviation σ. The third standardized moment can take any positive or negative value, although in practical settings it rarely exceeds 2 or 3 in absolute value. Because it involves cubed values, the third standardized moment is sensitive to outliers (Kim & White 2004), and it can even be undefined for heavy-tailed distributions such as the Cauchy density or the Pareto density with a shape parameter of 3 . When the third standardized moment is finite, it is zero for symmetric distributions, although a value of zero does not necessarily mean that the distribution is symmetric (Ord 1968; Johnson and Kotz 1970, p. 253). To estimate the third standardized moment from a sample of n observations, a biased but simple estimator is the third sample moment $1/n\ \Sigma\ ((x-\bar{x})/s)^3$, where $\bar{x}$ is the sample mean and s is the sample standard deviation. An unbiased estimator is the third k statistic, which is obtained by taking the third sample moment and replacing $1/n$ with the quantity $n\ /\ ((n-1)(n-2))$ (Rose and Smith 2002).

Although the third standardized moment is far and away the most popular definition of skew, alternative definitions have been proposed (MacGillivray 1986). The leading alternatives are bounded by –1 and +1, and are zero for symmetric distributions, although again a value of zero does not guarantee symmetry. One alternative is Bowley's (1920) quartile formula for skew: $((q_3 - m) - (m - q_1))\ /\ (q_3 - q_1)$, or more simply $(q_1 + q_3 - 2m)\ /\ (q_3 - q_1)$, where m is the median and q_1 and q_3 are the first (or left) and third (or right) quartiles. Bowley's skew focuses on the part of the distribution that fits in between the quartiles: if the right quartile is further from the median than is the left quartile, then Bowley's skew is positive; if the left quartile is further from the median than the right quartile, then Bowley's skew is negative. Because it doesn't cube any values and doesn't use any values more extreme than the quartiles, Bowley's skew is more robust to outliers than is the conventional third-moment formula (Kim & White 2004). But the quantities in Bowley's formula are arbitrary: instead of the left and right quartiles—i.e., the 25^{th} and 75^{th} percentiles—Bowley could just as plausibly have used the 20^{th} and 80^{th} percentiles, the 10^{th} and 90^{th} percentiles, or more generally the $100p^{th}$ and $100(1-p)^{th}$ percentiles $F^{-1}(p)$ and

$F^{-1}(1 - p)$. Substituting these last two expressions into Bowley's formula, Hinkley (1975) proposed the generalized skew formula $(F^{-1}(1 - p) + F^{-1}(p) - 2\ m) / (F^{-1}(1 - p) - F^{-1}(p))$, which is a function of high and low percentiles defined by p. Since it is not clear what value of p is most appropriate, Groeneveld & Meeden (1984) averaged Hinkley's formula across all ps from 0 to 0.5. Groeneveld & Meeden's average was $(\mu - m) / E\ |X - m|$, which is close to an old skew formula that is attributed to Pearson: $(\mu - m) / \sigma$ (Yule 1911).

The Pearson and Groeneveld-Meeden formulas are consistent with a widely taught rule of thumb claiming that the skew determines the relative positions of the median and mean. According to this rule, in a distribution with positive skew the mean lies to the right of the median, and in a distribution with negative skew the mean lies to the left of the median. If we define skew using the Pearson or Groeneveld-Meeden formulas, this rule is self-evident: since the numerator of both formulas is simply the difference between the mean and the median, both will give positive skew when the mean is greater than the median, and negative skew when the situation is reversed. But if we define skew more conventionally, using the third standardized moment, the rule of thumb can fail. Violations of the rule are rare for continuous variables, but common for discrete variables (von Hippel 2005). A simple discrete violation is the binomial distribution with n=10 and p=0.09 (cf. Lesser 2005). In this distribution, the mean 0.9 is left of the median 1, but the skew as defined by the third standardized moment is positive, at 0.906, and the distribution, with its long right tail, looks like a textbook example of positive skew. Examples like this one argue against using the Pearson, Groeneveld-Meeden, or Bowley formulas, all of which yield a negative value for this clearly right-skewed distribution. Most versions of Hinkley's skew also contradict intuition here: Hinkley's skew is negative for $0.5>p>0.225$, zero for $0.225 \geq p>0.054$, and doesn't become positive until $p \leq 0.054$.

Since many statistical inferences assume that variables are symmetrically or even normally distributed, those inferences can be inaccurate if applied to a variable that is skewed. Inferences grow more accurate as the sample size grows, with the required sample size depending on the amount of skew and the desired level of accuracy. A reliable rule is that, if you are using the normal or t distribution to calculate a nominal 95% confidence interval for the mean of a skewed variable, the interval will have at least 94% coverage if the sample size is at least 25 times the absolute value of the (third-moment) skew (Cochran 1977, Boos & Hughes-Oliver 2000). For example, a sample of 50 observations should be plenty even if the skew is as large as 2 (or –2).

In order to use statistical techniques that assume symmetry, researchers sometimes transform a variable to reduce its skew (von Hippel 2003). The most common transformations for reducing positive skew are the logarithm and the square root, and a much broader family of skew-reducing transformations has been defined (Box an Cox 1964). But reducing skew has costs as well as benefits. A transformed variable can be hard to interpret, and conclusions about the transformed variable may not apply to the original variable before transformation (Levin, Liukkonen, & Levine, 1996). In addition, transformation can change the shape of relationships among variables; for example, if X is right-skewed and has a linear relationship with Y, then the square root of X, although less skewed, will have a curved relationship with Y (von Hippel 2010). In short, skew reduction is rarely by itself a sufficient reason to transform a variable. Skew should be treated as an important characteristic of the variable, not just a nuisance to be eliminated.

References

Box, G.E.P., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26** (2): 211–252.

Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley.

Boos, D.D., & Hughes-Oliver, J.M. (2000). How large does *n* have to be for *Z* and *t* intervals? *The American Statistician* **54**(2), 121-128.

Bowley, A.L. (1920). *Elements of Statistics*. New York: Scribner.

Groeneveld, R.A., and Meeden, G. (1984). Measuring skewness and kurtosis. *The Statistician* **33**, 391-399.

Groeneveld, R.A. (1986), Skewness for the Weibull Family, *Statistica Neerlandica*, 40, 135-140.

Hinkley, D.V. (1975). On power transformations to symmetry. *Biometrika* **62**, 101-111.

Johnson, N.L., & Kotz, S. (1970). *Continuous Univariate Distributions* **1**. Boston: Houghton Mifflin.

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

Lesser, L.M. (2005). Letter to the editor [comment on von Hippel (2005)]. *Journal of Statistics Education* 13(2). http://www.amstat.org/publications/jse/v13n3/lesser_letter.html

Levin, A., Liukkonen, J., & Levine, D. W. (1996). Equivalent inference using transformations. *Communications in Statistics, Theory and Methods, 25*(5), 1059–1072.

MacGillivray, H.L. (1986). Skewness and asymmetry: Measures and orderings. *Annals of Statistics*, **14**(3), 994-1011.

Ord, J.K. (1968). The discrete student's t distribution. *Annals of Mathematical Statistics* **39**, 1513-1516.

Rose, C., and Smith, M. (2002). *Mathematical Statistics with Mathematica*. New York: Springer.

Sato, M. (1997). Some remarks on the mean, median, mode and skewness. *Australian Journal of Statistics* **39**(2), 219-224.

von Hippel, P.T. (2003). Normalization. In *Encyclopedia of Social Science Research Methods* (Lewis-Beck, Bryman, and Liao, eds.). Thousand Oaks, CA: Sage.

von Hippel, P.T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education* 13(2). www.amstat.org/publications/jse/v13n2/vonhippel.html

von Hippel, P.T. (2010). How to impute skewed variables under a normal model. Unpublished manuscript, under review.

Yule, G.U. (1911). *Introduction to the Theory of Statistics*. London: Griffith.