

Estimating and Comparing Kurtosis and Skewness from an Arbitrary Population

Lihua An and S. Ejaz Ahmed

University of Windsor

MSUG meeting, Dearborn

June 8, 2006

- Introduction
- Preliminaries and background
 - population shape parameters
 - sample measures adapted by various statistical packages
- Comparison of skewness and kurtosis measures
 - adapted by software packages
 - two improved estimators
- A bias-corrected kurtosis measure for non-normal data
- Conclusions

- Skewness and Kurtosis are relevant to issues of
 - test of normality;
 - robustness;
 - outliers;
 - modified tests and estimation;
 - large sample inferences;
 - and etc.
- Some difficulties remain due to problems with estimating these parameters from an arbitrary population.

Let μ_r be the r^{th} population moment about the mean.

Population skewness:

$$\gamma_1 = \frac{E(X - \mu)^3}{(E(X - \mu)^2)^{3/2}} = \frac{\mu_3}{\sigma^3} \quad (1)$$

Population Kurtosis:

Pearson's $\beta_2 = \frac{E(X - \mu)^4}{(E(X - \mu)^2)^2} = \frac{\mu_4}{\sigma^4} \quad (2)$

Fisher's $\gamma_2 = \beta_2 - 3 \quad (3)$

Table: Skewness and Kurtosis Indices for Selected Distributions

Probability Distribution	Population skewness γ_1	Population Kurtosis γ_2
Normal	0	0
Student-t	0	$\frac{6}{df-4}$
Central χ^2	$\frac{8}{df}$	$\frac{12}{df}$
Uniform	0	-1.2
Binomial(n, p)	$\frac{1-2p}{\sqrt{npq}}$	$\frac{1-6pq}{npq}$
mixture of normals (p_j, μ_j, σ_j)	$\frac{1}{\sigma^3} \sum_{j=1}^k p_j (\mu_j - \mu) \cdot [3\sigma_j^3 + (\mu_j - \mu)^2]$	$\frac{1}{\sigma^4} \sum_{j=1}^k p_j [3\sigma_j^4 + 6(\mu_j - \mu)^2 \sigma_j^2 + (\mu_j - \mu)^4]$

- Let X_1, X_2, \dots, X_n be a random sample of size n from an unknown population. Let m_r be the r^{th} sample moment about the mean, defined by $m_r = \frac{\sum(x_i - \bar{x})^r}{n}$.
- Natural estimators of the population skewness and kurtosis are:

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{n^{1/2} \sum(x_i - \bar{x})^3}{(\sum(x_i - \bar{x})^2)^{3/2}}; \quad (4)$$

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{n \sum(x_i - \bar{x})^4}{(\sum(x_i - \bar{x})^2)^2} - 3, \quad (5)$$

respectively.

- However, (4) and (5) are biased.

- Unbiased estimators of the moments:

$$K_2 = \frac{n}{n-1} m_2,$$

$$K_3 = \frac{n^2}{(n-1)(n-2)} m_3,$$

$$K_4 = \frac{n^2}{(n-1)(n-2)(n-3)} (n+1)m_4 - 3(n-1)m_2^2.$$

- Measures based on the unbiased estimators of moments:

$$G_1 = \frac{K_3}{K_2^{3/2}} = \frac{\sqrt{n(n-1)}}{n-2} \cdot g_1 \quad (6)$$

$$G_2 = \frac{K_4}{K_2^2} = \frac{(n-1)}{(n-2)(n-3)} \{(n+1)(g_2) + 6\} \quad (7)$$

- G_1 , G_2 and g_1 , g_2 are both available in SAS:
 - PROC MEAN produces G_1 and G_2 ;
 - BiasKur option in CALIS procedure gives g_1 and g_2 .
- In contrast, the skewness and kurtosis measures adapted by MINITAB are:

$$b_1 = \frac{m_3}{s^3} = \left(\frac{n-1}{n}\right)^{3/2} g_1, \quad (8)$$

$$b_2 = \frac{m_4}{s^4} - 3 = \left(\frac{n-1}{n}\right)^2 g_2 - 3. \quad (9)$$

where, $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$.

In general,

$$\text{Var}(G_1) = \left\{ \frac{\sqrt{n(n-1)}}{(n-2)^2} \right\} \text{Var}(g_1) \simeq \left(1 + \frac{3}{n} \right) \text{Var}(g_1)$$

$$\text{Var}(b_1) = \left(\frac{n-1}{n} \right)^3 \text{Var}(g_1) \simeq \left(1 - \frac{3}{n} \right) \text{Var}(g_1)$$

$$\text{Var}(G_2) = \left\{ \frac{(n-1)(n+1)}{(n-2)(n-3)} \right\}^2 \text{Var}(g_2) \simeq \left(1 + \frac{10}{n} \right) \text{Var}(g_2)$$

$$\text{Var}(b_2) = \left(\frac{n-1}{n} \right)^4 \text{Var}(g_2) \simeq \left(1 - \frac{4}{n} \right) \text{Var}(g_2)$$

$$\text{Var}(b_1) < \text{Var}(g_1) < \text{Var}(G_1) \quad (10)$$

$$\text{Var}(b_2) < \text{Var}(g_2) < \text{Var}(G_2) \quad (11)$$

- Skewness measures g_1 , G_1 and b_1 are all unbiased for normal distribution.
- $mse(b_1) < mse(g_1) < mse(G_1)$.
- Kurtosis measures g_2 , G_2 and b_2 :

$$Bias(G_2) = 0,$$

$$Bias(g_2) = -\frac{6}{n+1},$$

$$Bias(b_2) = 3\frac{(n-1)^3}{n^2(n+1)} - 3 \simeq \frac{-12}{n+1}.$$

- $mse(g_2) < mse(b_2) < mse(G_2)$, for all $n \geq 2$.

Since $\text{Var}(b_2) < \text{Var}(g_2) < \text{Var}(G_2)$, define

$$ug_2 = g_2 + 6/(n + 1) \quad (12)$$

and

$$ub_2 = b_2 + 3 \frac{(n - 1)^3}{n^2(n + 1)} - 3 \quad (13)$$

It is easily seen that

$$mse(ub_2) < mse(ug_2) < mse(g_2) < mse(b_2) < mse(G_2), \quad \text{for all } n \geq 2.$$

or

$$mse(new1) < mse(new2) < mse(SAS2) < mse(MINITAB) < mse(SAS1)$$

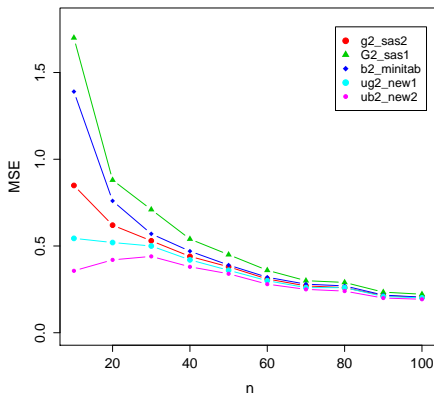


Figure: MSE of the kurtosis estimators for normal data

For a student t distribution with ν degrees of freedom,

$$\gamma_2 = \frac{6}{\nu - 4}, \quad \nu > 4 \quad (14)$$

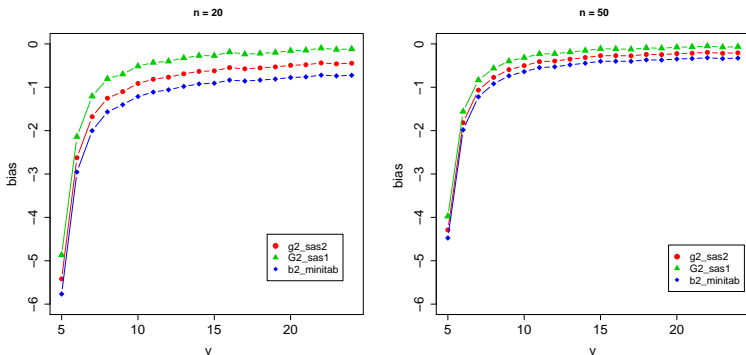


Figure: Bias of the kurtosis measures for t distribution

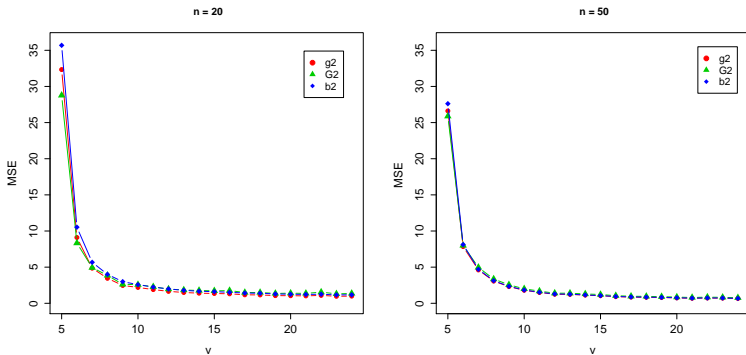


Figure: MSE of the kurtosis measures for t distribution

For a χ^2 distribution with ν degrees of freedom,

$$\gamma_2 = \frac{12}{\nu} \quad (15)$$

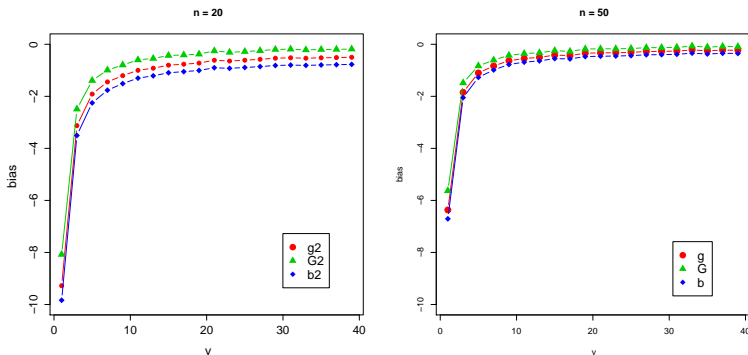


Figure: Bias of the kurtosis measures for χ^2 distribution

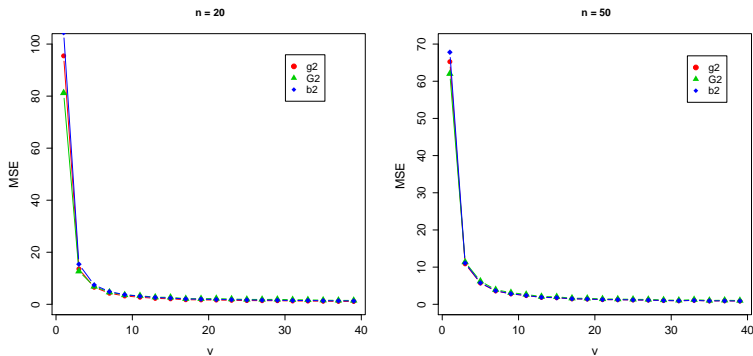


Figure: MSE of the kurtosis measures for χ^2 distribution

Theorem

If X is a mixture of k normals, then its mean, variance, skewness, and kurtosis are,

$$\mu = \sum_{j=1}^k p_j \mu_j,$$

$$\sigma^2 = \sum_{j=1}^k p_j (\sigma_j^2 + \mu_j^2) - \mu^2,$$

$$\gamma_1 = \frac{1}{\sigma^3} \sum_{j=1}^k p_j (\mu_j - \mu) [3\sigma_j^3 + (\mu_j - \mu)^2],$$

$$\beta_2 = \frac{1}{\sigma^4} \sum_{j=1}^k p_j [3\sigma_j^4 + 6(\mu_j - \mu)^2 \sigma_j^2 + (\mu_j - \mu)^4].$$

Corollary

If X is a mixture of two normals with $\mu_1 = \mu_2$, then its kurtosis is $\beta_2 = \frac{3(p_1\sigma_1^4 + p_2\sigma_2^4)}{(p_1\sigma_1^2 + p_2\sigma_2^2)^2}$. This kurtosis is maximized when $p_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$ and $p_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$, and the maximum value of β_2 is $\frac{3}{4}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + 2\right)$.

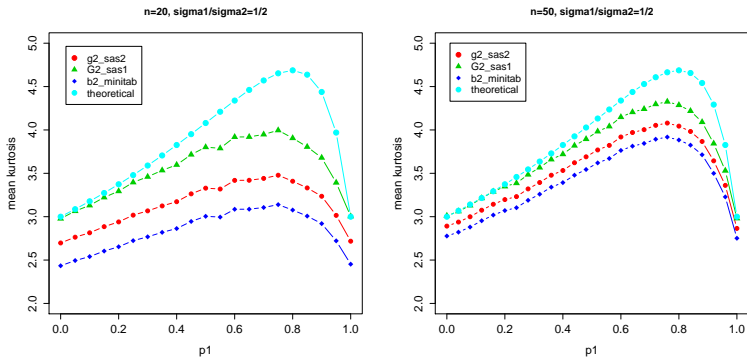


Figure: Bias of the kurtosis measures for mixture of normal distribution

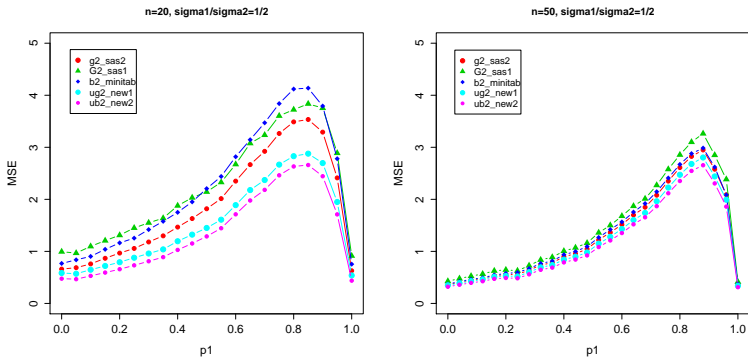


Figure: MSE of the kurtosis measures for mixture of normal distribution

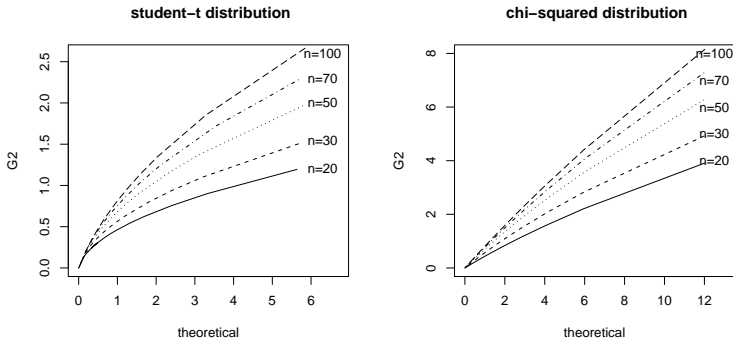


Figure: G_2 vs. γ_2 for t and χ^2 distributions

Table: An empirical bias-corrected kurtosis estimation for t and χ^2

n	t distribution	χ^2 distribution
20	$0.5255G_2 + 3.5094(G_2)^2$	$2.2064G_2 + 0.2213(G_2)^2$
30	$0.5556G_2 + 2.1388(G_2)^2$	$1.6662G_2 + 0.1564(G_2)^2$
50	$0.6813G_2 + 1.1531(G_2)^2$	$1.3632G_2 + 0.0873(G_2)^2$
70	$0.7743G_2 + 0.7654(G_2)^2$	$1.2583G_2 + 0.0535(G_2)^2$
100	$0.8067G_2 + 0.5213(G_2)^2$	$1.2100G_2 + 0.0324(G_2)^2$

Conclusions

- Skewness and kurtosis measures adapted by various software packages have been compared, with emphasis given to kurtosis measures; Measures adapted by SAS are favorable in terms of *MSE*;
- The two estimators proposed have lower *MSE* comparing to all currently used measures within the distributions studied;
- All above estimators have considerably large bias and therefore large *MSE* when underlying population distribution is not normal, therefore a bias-corrected estimator is essential for improving inferences established on population kurtosis;
- An empirical bias-corrected kurtosis estimator is provided for student-t (heavy-tailed) and χ^2 (skewed) data; however, extra variation is introduced.

Greeting

Thanks for Your Attention
and
Have a Nice Day!