IPRPD
International Journal of Business & Management Studies
ISSN 2694-1430 (Print), 2694-1449 (Online)
Volume 06; Issue no 04: April, 2025
DOI: 10.56734/ijbms.v6n4a9

ESTIMATING THE DISTRIBUTION OF NET WORTH BY AGE USING JOHNSON'S BIVARIATE SU DISTRIBUTION

Pilsun Choi[1], Insik Min[2]

[1]*Department of International Trade, Konkuk University, Republic of Korea*

[2]*Department of Economics, Kyung Hee University, Republic of Korea*

Abstract

This study proposes the use of Johnson's (1949) multivariate S_U distribution for estimating the joint distribution among economic well-being variables. The S_U distribution exhibits extremely high flexibility, enabling it to effectively capture the extreme skewness and kurtosis commonly seen in wealth and income variables. As this distribution imposes no constraints on variable ranges, it is well-suited for estimating the distributions of net worth or disposable income, which often include non-positive values. As an illustrative example, we employ the bivariate S_U model to estimate the joint distribution of net worth and age using the US survey data.

Keywords

Parametric Distribution, Johnson S_U, Multivariate, Correlation

JEL Classification: C13; C46; D31

I. Introduction

Historically, the focus on economic well-being and inequality centered primarily on income, but recent attention has shifted towards the relationship among various variables of economic well-being, such as income and wealth (Balestra and Oehler, 2023; Fisher et al., 2022; Keister and Lee, 2017; Kuhn et al., 2020; Piketty and Saez, 2003, among others).

This study proposes the use of Johnson's (1949b) multivariate S_U distribution for estimating the joint distribution among wealth, income, consumption, and other economic well-being variables. This distribution is constructed by transforming the normal distribution, making it remarkably straightforward to create a multivariate distribution. Its marginal distributions exhibit extremely high flexibility, capturing a wide range of skewness and kurtosis. Notably, as the S_U distribution imposes no constraints on variable ranges, it is well-suited for estimating the distributions of net worth and disposable income, which often include non-positive values. Due to these advantages, Choi and Min (2025) recently introduced the S_U distribution to net worth and disposable income, demonstrating its goodness of fit. Here, we aim to extend this approach to estimate the joint distribution of net worth and age.

We advocate for the use of a "normalized" correlation instead of the Pearson correlation coefficient as a means of measuring the association between two economic well-being variables such as income and wealth. This is because such variables are typically highly skewed and contain extreme outliers, which can distort the association between variables. The normalized correlation coefficient is derived by transforming each marginal distribution into a normal distribution and then measuring the Pearson correlation coefficient between them. In fact, it is estimated by one of the parameters defining the bivariate S_U distribution. As an illustrative example, we employ the bivariate S_U model to estimate the joint distribution of net worth and age using the US survey data.

II. Johnson's S_U Distribution

The S_U distribution first appeared in the pathbreaking article of Johnson (1949a). The S_U variable X is generated by the transformation to normality in the following manner.

$$\sinh^{-1}\left(\frac{X-m}{s}\right) = \lambda + \theta Z, \qquad -\infty < X < \infty, s > 0, \theta > 0 \tag{1}$$

where Z is a standard normal variable. The symbol S_U is for 'unbounded system' implying that the range of X is unbounded. The probability density function (PDF) of S_U is:

$$f(x) = \theta^{-1}[(x-m)^2 + s^2]^{-1/2}\phi(z), \tag{2}$$

where $z = \theta^{-1}\left[\sinh^{-1}\left(\frac{x-m}{s}\right) - \lambda\right]$, and $\phi(\cdot)$ is the PDF of a standard normal variable. The mean and variance of X are $m + s \cdot e^{\theta^2/2}\sinh(\lambda)$ and $\frac{1}{2}s^2\left(e^{\theta^2/2} - 1\right)\left(e^{\theta^2/2}\cosh(2\lambda) + 1\right)$, respectively.

Among the four parameters in (1), the sign of λ determines the direction of skewness. When $\lambda = 0$, the distribution is symmetric, while a positive (negative) λ indicates positive (negative) skewness. As λ and θ move further away from 0, the distribution deviates from the normal distribution, resulting in increased asymmetry and thicker tails. Johnson (1949a) shows that the S_U distribution is an extremely flexible distribution capable of capturing the wide range of combinations of skewness and excess kurtosis. Furthermore, unlike several parametric distributions traditionally used for income distribution estimation, the S_U distribution can be applied to variables with negative values, such as net worth or disposable income.

The S_U distribution can be easily extended to multivariate dimensions (Johnson, 1949b). For an $N \times 1$ random vector Z that follows a multivariate standard normal distribution, the joint PDF is expressed as:

$$\phi_R(z) = (2\pi)^{-N/2}|R|^{-1/2}\exp\left(-\frac{1}{2}z'R^{-1}z\right), \tag{3}$$

where R is the correlation coefficient matrix with an off-diagonal element r_{ij}, and $|R|$ is the determinant of R. A multivariate S_U random vector X can be obtained by the inverse hyperbolic sine transformation of each variable X_i to a normal variable, i.e., $\sinh^{-1}\left(\frac{X_i-m_i}{s_i}\right) = \lambda_i + \theta_i Z_i$ where $s_i > 0$ and $\theta_i > 0$. Hence, the joint PDF of X is:

$$f(x) = (2\pi)^{-N/2}|R|^{-1/2}\prod_{i=1}^{N}\theta_i^{-1}[(x_i - m_i)^2 + s_i^2]^{-1/2}\exp\left(-\frac{1}{2}z'R^{-1}z\right), \tag{4}$$

where $z_i = \theta_i^{-1}\left[\sinh^{-1}\left(\frac{x_i-m_i}{s_i}\right) - \lambda_i\right]$.

Due to the nonlinear transformation, the correlation of Z is not the same as the correlation of X. The Pearson's correlation coefficient ρ_{ij} between X_i and X_j is:

$$\rho_{ij} = \frac{e^{(\theta_i^2+\theta_j^2)/2}}{\sigma_i\sigma_j}\left[\frac{1}{2}e^{r_{ij}\theta_i\theta_j}\cosh(\lambda_i+\lambda_j) - \frac{1}{2}e^{-r_{ij}\theta_i\theta_j}\cosh(\lambda_i-\lambda_j) - \sinh(\lambda_i)\sinh(\lambda_j)\right], \tag{5}$$

where $\sigma_k = \left[\frac{1}{2}\left(e^{\theta_k^2} - 1\right)\left(e^{\theta_k^2}\cosh(2\lambda_k) + 1\right)\right]^{1/2}$, $k = i, j$. If $i = j$, then ρ_{ij} becomes 1. Conversely, if X follows multivariate S_U distribution with correlation matrix Σ whose off-diagonal element is ρ_{ij}, the correlation r_{ij} between Z_i and Z_j is:

$$r_{ij} = \frac{1}{\theta_i\theta_j}\ln\left(\frac{B_{ij} + \sqrt{B_{ij}^2 + \cosh(\lambda_i+\lambda_j)\cosh(\lambda_i-\lambda_j)}}{\cosh(\lambda_i+\lambda_j)}\right), \tag{6}$$

where $B_{ij} = \rho_{ij}\sigma_i\sigma_j\exp\left(-\frac{1}{2}\left(\theta_i^2 + \theta_j^2\right)\right) + \sinh(\lambda_i)\sinh(\lambda_j)$.

In the following section, we estimate the joint distribution of net worth and age using the 2022 Survey of Consumer Finances (SCF). Consider a bivariate S_U distribution with $\sinh^{-1}\left(\frac{X_1-m_1}{s_1}\right) = \lambda_1 + \theta_1 Z_1$, $\sinh^{-1}\left(\frac{X_2-m_2}{s_2}\right) = \lambda_2 + \theta_2 Z_2$, and r, the correlation coefficient between Z_1 and Z_2. From (4), the joint PDF is:

$$f(x_1, x_2) = \frac{(\theta_1\theta_2)^{-1}\left([(x_1-m_1)^2+s_1^2][(x_2-m_2)^2+s_2^2]\right)^{-1/2}}{2\pi\sqrt{1-r}}\exp\left(-\frac{1}{2(1-r^2)}\left(z_1^2 - 2r^2 z_1 z_2 + z_2^2\right)\right), \tag{7}$$

where $z_1 = \theta_1^{-1}\left[\sinh^{-1}\left(\frac{x_1-m_1}{s_1}\right)-\lambda_1\right]$, and $z_2 = \theta_2^{-1}\left[\sinh^{-1}\left(\frac{x_2-m_2}{s_2}\right)-\lambda_2\right]$.

The conditional distribution of X_1 given $X_2 = x_2$ is of the same S_U system as X_1, but with λ_1 and θ_1 replaced, respectively by $\lambda_1^* = \lambda_1 + r\theta_1\theta_2^{-1}\left(\sinh^{-1}\left(\frac{x_2-m_2}{s_2}\right)-\lambda_2\right)$ and $\theta_1^* = \theta_1\sqrt{1-r^2}$ (Kotz et al., 2000):

$$X_1|X_2 = x_2 \sim S_U(m_1, s_1, \lambda_1^*, \theta_1^*). \quad (8)$$

III. Estimation of Joint and Conditional Distributions for Net Worth and Age

We have selected two variables, net worth (in million dollars, nominal) and age from 2022 SCF. Estimation was performed using maximum likelihood estimation based on (7). In fact, it is possible to perform the estimation in a two-step manner, where individual marginal distributions are estimated first and then the correlation parameters among them are estimated. Such a two-step estimation may be used when dealing with a large number of variables. However, in our case with only two variables, we estimated all parameters in one step. Maximum likelihood estimation was performed using the Python SciPy "optimize" module with limited-memory BFGS (i.e., L-BFGS-B) algorithm. In our study, all statistics were calculated using the weights provided in the SCF.

Table 1 presents the estimated parameters of the bivariate S_U distribution. Based on these, we estimated the conditional distribution of net worth by age. Among the ages, we specifically chose the ages corresponding to the time of college graduation, namely 22, 23, and 24. In the 2022 SCF data, the number of observations for reference persons under 25 years old is quite limited. As shown in Table 2, even when combining the ages of 22 to 24, there are only 72 observations (1.58% of the total). Due to this small sample size, the calculated statistics of mean and standard deviation (S.D.) vary significantly by age. Unlike the empirical outcomes, the S_U model-based estimates for ages 22, 23, and 24 are very similar to each other. The S_U model demonstrates its advantage in obtaining robust estimates for conditional distributions when there are not many observations corresponding to the given conditions.

	m	s	λ	θ	r
Net worth (X_1)	-0.005	0.015	3.021	2.077	0.345
Age (X_2)	-2.301	123.780	0.426	0.129	

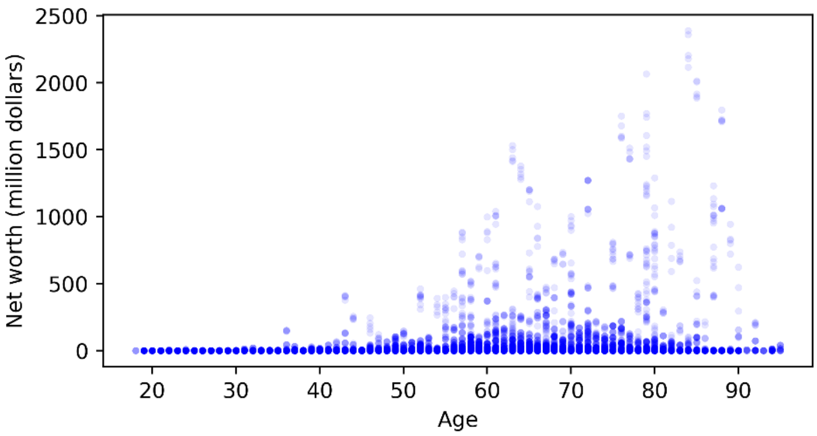
Table 1. Estimated parameters of the bivariate S_U distribution

Age	Descriptive statistics					Model-based estimates	
	Obs.	Min.	Max.	Mean	S.D.	Mean	S.D.
22	24	-0.009	5.413	0.122	0.472	0.282	1.953
23	27	-0.156	0.416	0.041	0.097	0.295	2.040
24	21	-0.046	9.576	0.386	1.638	0.309	2.131
22-24	72	-0.156	9.576	0.163	0.924	0.323	2.068

Table 2. Conditional distribution of net worth by age

Now, we analyze the correlation between net worth and age. The Pearson correlation coefficient estimate between the two variables is 0.071. It is positive as expected, but its magnitude is very small. The reason for this low Pearson correlation coefficient is not due to a lack of association between the variables, but because the distribution of the net worth variable has extreme skewness and excess kurtosis. Figure 1 demonstrates this well. The left panel of the figure uses the original variables of net worth and age to create a scatter plot, while the right panel uses "normalized" variables. Here, "normalization" means transforming the variable into a normally distributed variable with zero skewness and excess kurtosis. In our case, the normalization means $\sinh^{-1}\left(\frac{x_{1i}-\hat{m}_1}{\hat{s}_1}\right)$ and $\sinh^{-1}\left(\frac{x_{2i}-\hat{m}_2}{\hat{s}_2}\right)$, and the scatter plot of these two normalized variables is the right panel in Figure 1. The linear correlation in the right panel appears stronger than in the left panel. As shown in Table 1, the Pearson correlation coefficient estimate for these normalized variables is $\hat{r} = 0.345$, which is much bigger than 0.071, the correlation coefficient between x_{1i} and x_{2i}. Figure 1 demonstrates how the extreme skewness and excess kurtosis, commonly found in wealth or income variables, can distort the association between variables. And it shows the correlation between the normalized variables estimated by the S_U model serves as a better indicator for measuring the association.

(a) Original variables



(b) Normalized variables

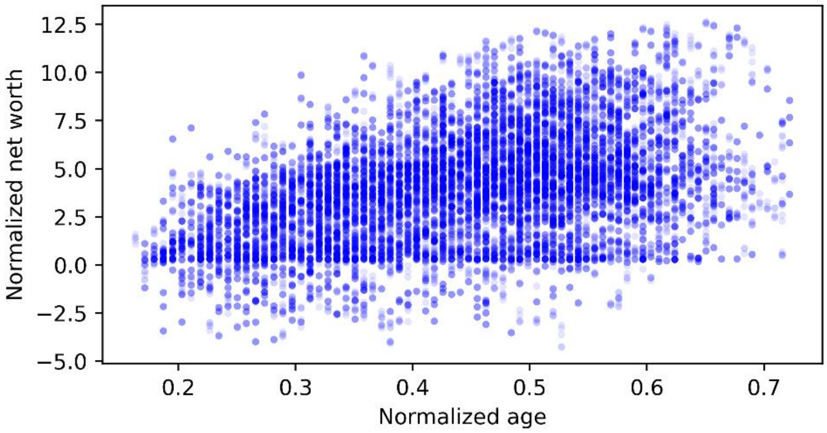


Figure 1. Scatter plot of net worth and age

IV. Concluding Remarks

In the literature, there is an approach that uses copula functions to construct the joint distribution of household income and wealth (Jäntti et al., 2015, among others). The multivariate S_U model discussed in this study can also be understood using the concept of copulas. That is, the multivariate S_U distribution can be considered as a distribution that combines each marginal S_U variable using Gaussian copula function. Since S_U variables are transformed from normal variables, it is quite simple to transform them back to normals and combine them using a Gaussian copula. Due to the fact that it is derived by transforming the normal distribution, the S_U distribution has several advantages. The joint PDF has relatively simple form, making maximum likelihood estimation relatively straightforward, even in one-step estimation. Furthermore, generating multivariate S_U random numbers is also straightforward, making it advantageous for simulation analyses in a multivariate dimension. In our example, we considered two variables: net worth and age. However, when dealing with more than two variables—for instance, when estimating the joint distribution of wealth, income, and consumption—the multivariate S_U model is likely to be an attractive option.

References

Balestra, C., & Oehler, F. (2023). Measuring the joint distribution of household income, consumption and wealth at the micro level, Working Paper, OECD.

Choi, P., & Min, I. (2025). Estimating the distribution of net worth and disposable income using Johnson's S_U distribution, Working Paper.

Fisher, J. D., Johnson, D. S., Smeeding, T. M., & Thompson, J. P. (2022). Inequality in 3-D: Income, Consumption, and Wealth. *Review of Income and Wealth*, 68(1), 16-42.

Jäntti, M., Sierminska, E. M., & Van Kerm, P. (2015). Modeling the joint distribution of income and wealth. In *Measurement of Poverty, Deprivation, and Economic Mobility* (Vol. 23, pp. 301-327). Emerald Group Publishing Limited.

Johnson, N. L. (1949a). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2), 149-176.

Johnson, N. L. (1949b). Bivariate distributions based on simple translation systems. *Biometrika*, 36(3/4), 297-304.

Keister, L. A., & Lee, H. Y. (2017). The double one percent: Identifying an elite and a super-elite using the joint distribution of income and net worth. *Research in Social Stratification and Mobility*, 50, 1-12.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions: Models and Applications*. John Wiley & Sons.

Kuhn, M., Schularick, M., & Steins, U. I. (2020). Income and wealth inequality in America, 1949–2016. *Journal of Political Economy*, 128(9), 3469-3519.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913–1998. *Quarterly Journal of Economics*, 118(1), 1-41.