Running Head: CLT AND SAMPLE SIZE

Central Limit Theorem and Sample Size

Zachary R. Smith

and

Craig S. Wells

University of Massachusetts Amherst

Paper presented at the annual meeting of the Northeastern Educational Research Association, Kerhonkson, New York, October 18-20, 2006.

Abstract

Many inferential statistics that compare means require that the scores within groups in the population be normally distributed. Unfortunately, most variables are not normally distributed, especially in the social sciences (Micceri, 1989). However, thanks to the central limit theorem (CLT), which states that as the sample size increases, the sample mean will be normally distributed for most underlying distributions, hypothesis tests are robust against the violation of normality. In this study, a simulation study was employed to generate sampling distributions of the mean from realistic non-normal parent distributions for a range of sample sizes in order to determine when the sample mean was approximately normally distributed. When data are rounded to the nearest integer, as is often practiced, larger samples are needed for the CLT to work. As the skewness and kurtosis of a distribution increase, the CLT stops working, even up to samples of 300. This study will benefit researchers and statisticians in that it will provide guidance in selecting appropriate sample sizes that will lead to robust conditions against the violation of normality.

Central Limit Theorem and Sample Size

Inferential statistics are a powerful technique used by researchers and practitioners for a wide array of purposes such as testing the falsehood of theories and identifying important factors that may influence a relevant outcome. Every hypothesis test requires that certain assumptions be satisfied in order for the inferences to be valid. A common assumption among popular statistical tests is normality; e.g., the two-sample *t*-test assumes the scores on the variable of interest are normally distributed in the population. Unfortunately, variables are rarely normally distributed, especially in the social sciences (Micceri, 1989). However, even if the scores depart severely from normality, the sample mean may be normally distributed for large-enough samples due to the central limit theorem. The purpose of the present study is to examine how large the sample size must be in order for the sample means to be normally distributed.

How common is normality?

Micceri (1989) analyzed 440 distributions from all different sources, such as achievement and psychometric variables. The sample size for the various distributions ranged from 190 to 10,893. With such a large number a variables, the results covered most types of distributions observed in educational and psychological research (Micceri, 1989).

When the Kolmogorov-Smirnov (KS) test, described in detail below, was used by Micceri, it was found that none of the distributions were normally distributed at the 0.01 alpha level. “No distributions among those investigated passed all tests of normality, and very few seem to be even reasonably close approximations to the Gaussian” (Micceri,

1989, p. 161). It seems that "normality is a myth; there never was, and never will be, a normal distribution" (Geary, 1947, as cited in Micceri). Still, normality is an assumption that is needed in many statistical tests. Determining the best way to approximate normality is the only option, since true normality does not seem to exist. This theorem, also described briefly below, only implies that the sampling means are approximately normally distributed when the sample size is large enough.

Micceri suggests conducting more research on the robustness of the normality assumption based on the fact that real world data are often contaminated. Furthermore, few studies have dealt with "lumpiness and multimodalities" in the distributions (Micceri, 1989).

Types of Non-normal Distributions

While there are many types of distributions that educational and psychological variables may follow besides Gaussian, many may be classified as either skewed (positively or negatively), heavy or thin tailed (kurtosis), and multimodal. A skewed distribution is one that has a majority of scores shifted to one end of the scale with a few trailing off on the other end of the scale. These distributions can be positively or negatively skewed, which depends upon which side the tail is located. A positively skewed distribution is one with the tail pointing towards the positive side of the scale. Positively skewed and negatively skewed distributions were examined in this study.

Kurtosis is a property of a distribution that describes the thickness of the tails. The thickness of the tail comes from the amount of scores falling at the extremes relative to the Gaussian distribution. Most distributions taper off to zero, but some distributions

have a lot of kurtosis; i.e., there are many scores located at the extremes, giving it a thick tail.

The distributions selected for this study were based on those commonly observed as reported by Micceri (1989).

Central Limit Theorem

The central limit theorem (CLT), one of the most important theorems in statistics, implies that under most distributions, normal or non-normal, the sampling distribution of the sample mean will approach normality as the sample size increases (Hays, 1994). Without the CLT, inferential statistics that rely on the assumption of normality (e.g., two-sample *t*-test, ANOVA) would be nearly useless, especially in the social sciences where most of the measures are not normally distributed (Micceri, 1989).

It is often suggested that a sample size of 30 will produce an approximately normal sampling distribution for the sample mean from a non-normal parent distribution. There is little to no documented evidence to support that a sample size of 30 is the magic number for non-normal distributions. Arsham (2005) claims that it is not even feasible to state when the central limit theorem works or what sample size is large enough for a good approximation, but the only thing most statisticians agree on is "that if the parent distribution is symmetric and relatively short-tailed, then the sample mean reaches approximate normality for smaller samples than if the parent population is skewed or long-tailed" (What is Central Limit Theorem? section, para. 3). Nevertheless, it is interesting to examine if the sample mean is normally distributed for variables that realistically depart normality. The Kolmogorov-Smirnov test may be used to examine the distribution of the sample means.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test is used to determine if a sample of data is consistent with a specific distribution, for example, the normal distribution. The typical approach to using this method starts with stating under the null hypothesis (H_0) that the true cumulative distribution function (CDF) follows that of a normal distribution while the alternative hypothesis (H_1) states that the true CDF does not follow the normal distribution. The KS-test essentially compares the empirical distribution with that of a specified theoretical distribution such as the Gaussian. The difference between the two distributions is summarized as follows:

$$T = \sup\left|F^*(x) - S(x)\right| \tag{1.1}$$

where $F^*(x)$ and $S(x)$ represent the empirical and comparison distribution, respectively.

The purpose of this study is to provide researchers a better understanding of what sample sizes are required in order for the sample mean to be normally distributed for variables that may depart from normality. Educational and psychological measures are not often normally distributed (Micceri, 1989). The results will be very useful in guiding researchers when selecting sample sizes a priori in order to be confident that the normality assumption is robust to violation, given the expected departure from normality and sample size. The CLT could be a much more helpful tool in psychometrics and statistics with an established sample size for specific non-normal distributions. The results of this study will be helpful for researchers and statisticians, providing guidance regarding appropriate sample size selection in designing studies.

Method

A Monte Carlo simulation study was used to sample from various realistic distributions to determine what sample size is needed to make the distribution of the sample means approximately normal. Using the computer program, S-PLUS, observations were randomly sampled from the following eight distributions, which were selected to represent realistic distributional characteristics of achievement and psychometric measures as described in Micceri (1989): Normal ($\mu = 50, \sigma = 10$), Uniform (Min=10, Max=30), Bimodal (see Figure 1 for the density), four heavily skewed distributions with large kurtosis, and from an empirical distribution of an actual data set. It is important to note that the observations for the Normal, Uniform, and Bimodal distributions were rounded to the nearest integer in order to represent common measurement practice (i.e., for practical purposes, even continuous variables are reported at the integer level).

Insert Figure 1 about here

Normal, Uniform, and Bimodal Distributions

From each of the three populations, 10,000 samples of size 5, 10, 15, 20, 25, and 30 were randomly drawn, rounding to the nearest integer, as is often observed in educational and psychological measures. The mean for each sample was then computed, using four decimal places. Twenty replications were performed for each condition (i.e., the process was repeated 20 times for each sample size).

After the sampling distribution for the mean was constructed for a replication, a one sample Kolmogorov-Smirnov (KS) test was used to test whether the sample mean followed a normal distribution. This test was used on each sample size taken from all distributions. The proportion of replications in which the distribution of the sample means were concluded to depart normality according to the KS-test was recorded.

Skewed Distributions

Fleishman (1978) provided an analytic method of producing skewed distributions by transforming observations drawn from a normal distribution. The following equation is the polynomial transformation he provided for simulating skewed distributions:

$$Y = a + bX + cX^2 + dX^3 \quad (1.2)$$

X represents the value drawn from the normal distribution while *a*, *b*, *c*, and *d* are constants (note that *a*=-*c*).

The four skewed distributions chosen for this study represent positively skewed distributions of various degrees (the values for the constants are reported in Table 1). The skewness levels were the four largest reported in Fleishman's work. They all had the highest kurtosis level as well. These distributions were chosen to determine if the central limit theorem would work even for such heavily skewed data that are often encountered in psychological data (Micceri, 1989). Figures 2, 3, 4, and 5 illustrate the amount of skewness produced by the constants.

Insert Table 1 about here

Insert Figures 2, 3, 4, & 5 about here

Samples of 5 through 300 at intervals of 5 were drawn with no rounding from each of the four skewed distributions over 10,000 repetitions. After the transformation, the sampling means were rounded to the fourth decimal place. Once the sampling distribution for the mean is complete, the one sample Kolmogorov-Smirnov (KS) test was applied.

Empirical Data

To make the results more generalizable to actual studies and test data, a distribution based on real data was also used to replace the simulated parent distributions. Data from a personality inventory was chosen to represent a negatively skewed distribution. Figure 6 provides the histogram representing the distribution for this psychological measure. The distribution of the forgiveness data had a skewness of -0.1155 and a kurtosis of 0.2544.

Insert Figure 6 about here

Samples of size 5 through 200 at intervals of 5 were drawn from this parent distribution for 10,000 samples. The one sample Kolmogorov-Smirnov (KS) test was used to determine at what point the samples would begin to follow a normal distribution.

For all uses of the KS-test, the *p*-values produced were compared to an alpha level of 0.05. When the *p*-value was less that 0.05, the null hypothesis was rejected and it was concluded that the sample did not follow a normal distribution. If the *p*-value was greater than 0.05, the null hypothesis could not be rejected and it is likely that the sampling distribution was approximately normal.

Results

Table 2 displays the proportion of data sets that exhibited non-normal distributions for the sample mean for the normal, uniform, and bimodal conditions.

Insert Table 2 about here

Interestingly, even when sampling from the normal distribution, several of the 20 repetitions produced non-normal sample means for sample sizes of 5 or 10. This is due to the fact that the scores sampled from the population were rounded to the nearest integer to simulate common testing practices (this effect disappeared when the observations sampled were rounded to four or more decimal places). When the parent distribution was uniform, nineteen out of twenty replications departed normality at the largest sample size of 30. With the bimodal parent distribution, a sample size of 30 may be sufficient enough to invoke the CLT.

The results yielded from the heavily skewed distributions provided very interesting results. At the most skewed level (skewness=1.75, kurtosis=3.75), the highest sample size of 300 did not produce sample means that followed the normal distribution. There was no consistency in the *p*-value as the sample size increased and a large majority

of the samples still departed normality. At the least skewed level (skewness=1.00, kurtosis=3.75), the percentage of samples that followed a normal distribution according to the KS Test increased.

For the simulation conducted on the empirical data from the personality inventory, the data began to consistently follow a normal distribution at a sample size of 175. Before the sample of 175, the data followed the normal distribution sporadically.

Discussion

For the normal, uniform, and bimodal data simulation, it was interesting to find the sample size needed was higher than one might expect solely because of rounding the data to the nearest whole number. For example, even with the normal distribution, a sample size of 15 was needed before the distribution of the sample means became approximately normal. With the bimodal distribution, a sample size of 30 seemed to be enough. Yet for the uniform distribution, a sample size of 30 was not even enough for the CLT to be invoked.

The use of 10,000 samples and 20 replications for this simulation work should provide enough power to make these results meaningful. It is an important finding for researchers and educators. Much of the time, test scores are rounded before calculations are done. Statisticians should be aware of the need for larger samples with rounded data, but many practitioners might not be so aware of this fact. The results here will give them a basis to begin their research whenever they have rounded data.

The data from the personality inventory provides a realistic view of when the sample means may be normally distributed for a realistic non-normal distribution. It took a sample size of 175 for the distribution of the sample means to become approximately

normal. A few samples before 175 were approximately normal, but at 175 is when this result became consistent up through a sample size of 200. Compared to the simulated normal, uniform, and bimodal distributions, this needed a much larger sample size. But when it's compared to the simulated skewed distributions, this was a much smaller sample size.

With the simulated skewed data so extreme, even a sample size of 300 was not large enough to make the results consistent with the normal distribution. Determining if this would happen over many replications is certainly possible. Unfortunately, the properties of the KS-test in S-PLUS limit the work. The *p*-values for the present study were all compiled by hand over the multiple replications. A program is needed to calculate the *p*-values and make a judgment on them compared to the alpha level chosen. This program would save time by compiling the results faster, as well as eliminating a good deal of human error that could occur over so many replications.

The chosen skewed distributions were also so acute that they did not represent most realistic skewed distributions. For example, they could represent data from a depression survey applied to a general population consisting of predominately non-depressed individuals. The floor effect in the positive direction with a large peak in the negative would show that the majority or all of the participants were not depressed. Usually, a depression scale is given to participants that have depression symptoms or feel that they may be depressed as a diagnosis. This can occur, but particularly the most extremely skewed distribution is atypical.

With real data harder to come by, simulation work should be conducted using Fleishman's power method weights for more typical skewed distributions to make the

results more generalizable to realistic research. Since the samples will not have to be so large, more replications can be simulated which should yield more reliable results.

Further simulation work should be completed to establish more specific sample sizes for the normal, uniform, and bimodal sampling distributions. With the samples already taken at intervals of 5, there is a solid basis for figuring out where to sample more specifically. For example, with the bimodal distribution, a sample size of 20 began following a normal distribution. The next step is to determine at the specific sample size between 15 and 20 that begins following the CLT.

Although we examined whether sample means were not normally distributed given a parent distribution and sample size, it is important to consider the implications on the Type I and II error rates for particular statistical tests. For instance, even though the sample means are not normally distributed for certain parent distributions and sample sizes, how much will the Type I and II error rates be influenced under such conditions? It is possible that the effect will be minimal even though the sample means do not follow the normal distribution exactly. This question must be examined in future research.

References

Arsham, H. (2005). *System Simulation: The Shortest Route to Applications*, Version 9. Retrieved 5/31/06 from http://home.ubalt.edu/ntsbarsh/simulation/sim.htm.

Fleishman, A. I. (1978). A Method For Simulating Non-Normal Distributions. *Psychometrika, 43* (4), 521-532.

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Holt, Rinehart and Winston.

Micceri, T. (1989). The Unicorn, The Normal Curve, and Other Improbable Creatures. *Psychological Bulletin, 105* (1), 156-166.

Table1. Fleishman's (1978) power method weights for simulating the skewed distributions.

Skew	Kurtosis	A	B	C	D
1.75	3.75	-0.399	0.930	0.399	-0.036
1.50	3.75	-0.221	0.866	0.221	0.027
1.25	3.75	-0.161	0.819	0.161	0.049
1.00	3.75	-0.119	0.789	0.119	0.062

Table2. Percentage of replications that departed normality based on the KS-test.

	Sample Size					
	5	10	15	20	25	30
Normal	100	95	70	65	60	35
Uniform	100	100	100	100	100	95
Bimodal	100	100	100	75	85	50

Figure 1. *Density of the bimodal distribution.*

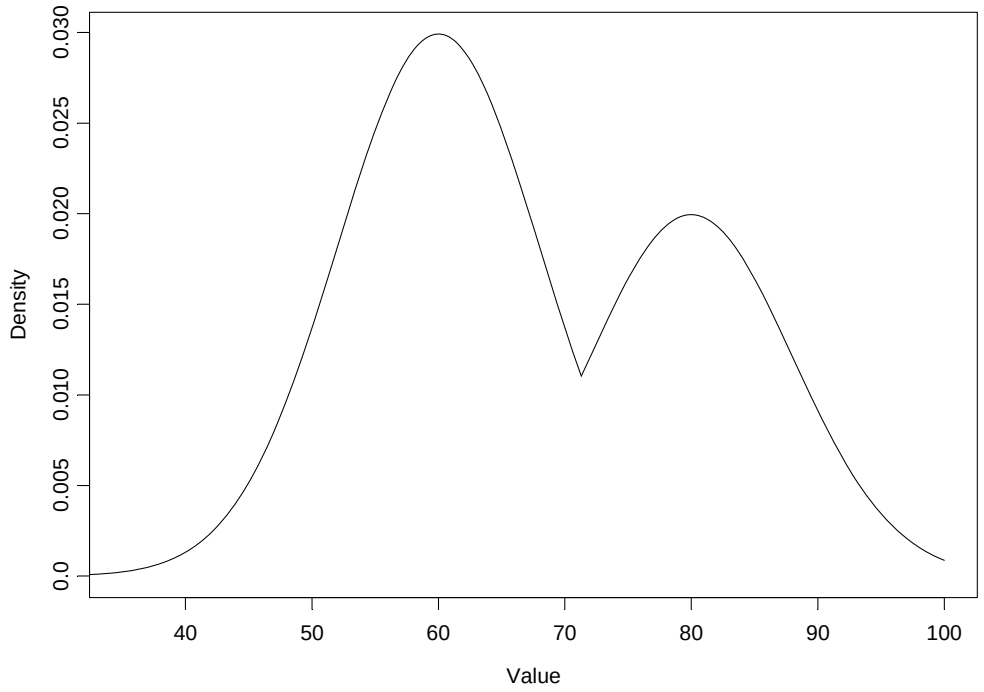

Figure 2. *Density of the most skewed distribution (skewness=1.75, kurtosis=3.75) based on Fleishman's power method weights (a = -0.40, b = 0.93, c = 0.40, d = -0.04).*

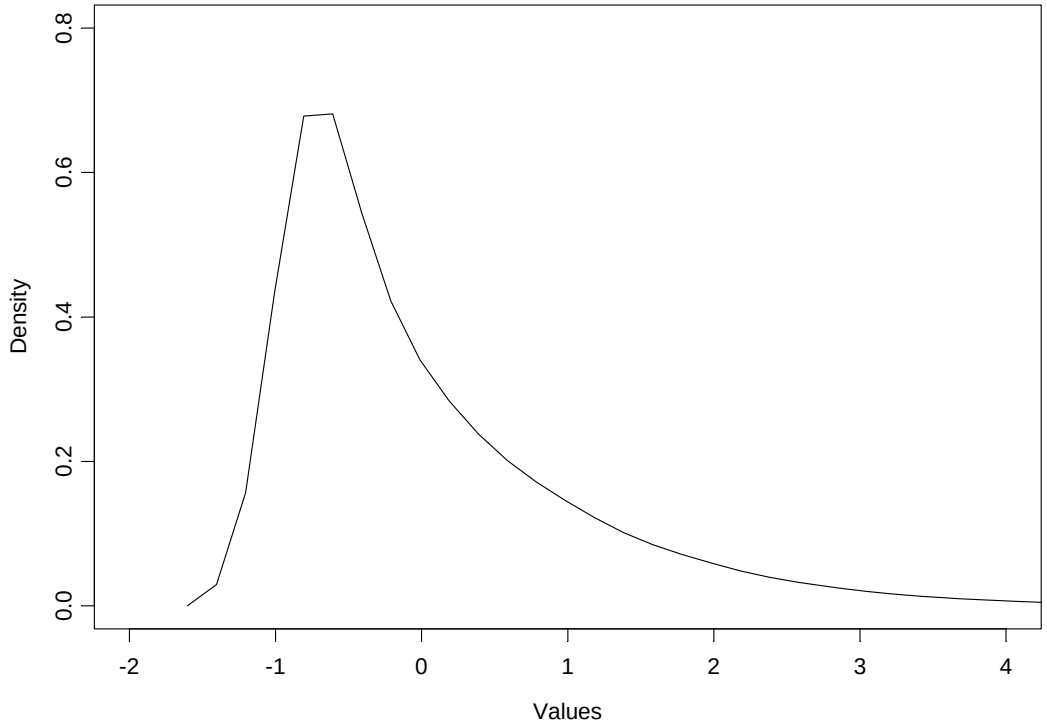

Figure 3. *Density of the second most skewed distribution (skewness=1.50, kurtosis=3.75) based on Fleishman's power method weights (a = -0.22, b = 0.87, c = 0.22, d = 0.03).*

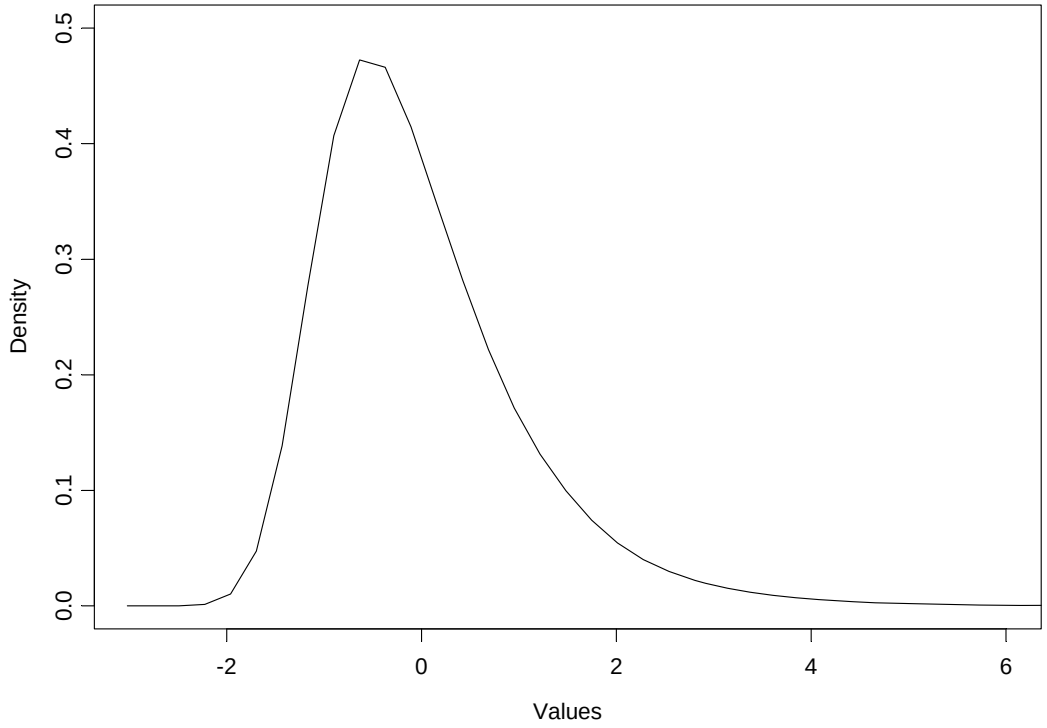

Figure 4. *Density of the third most skewed distribution (skewness=1.25, kurtosis=3.75) based on Fleishman's power method weights (a = -0.16, b = 0.82, c = 0.16, d = 0.05).*

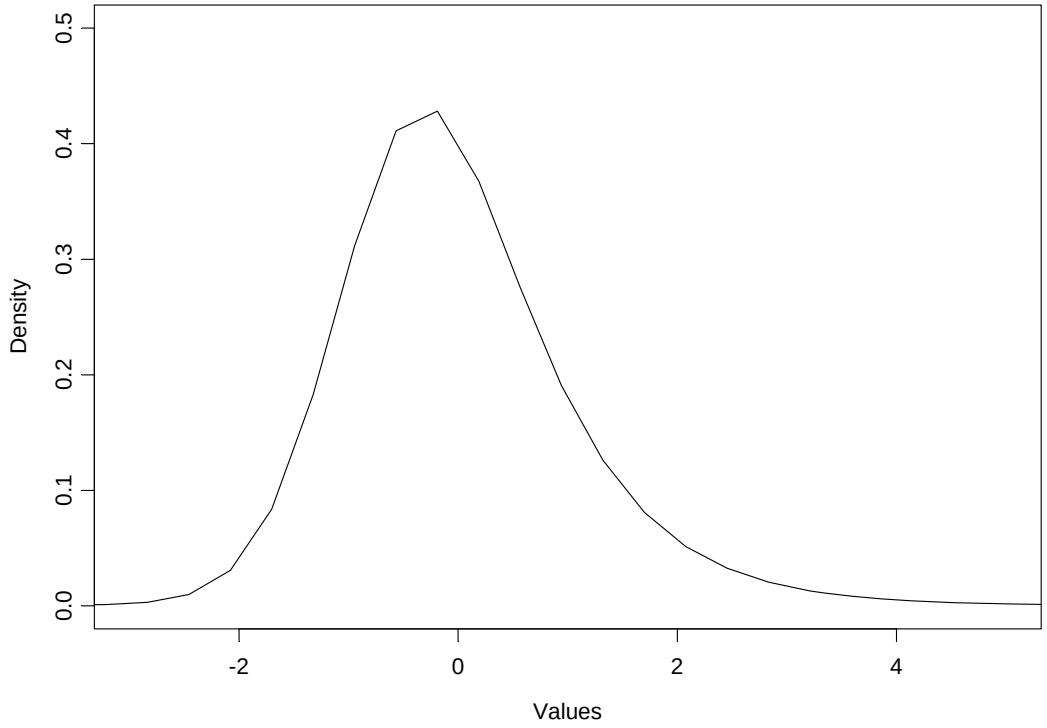

Figure 5. *Density of the least skewed distribution (skewness=1.00, kurtosis=3.75) based on Fleishman's power method weights (a = -0.12, b = 0.78, c = 0.11, d = 0.06).*

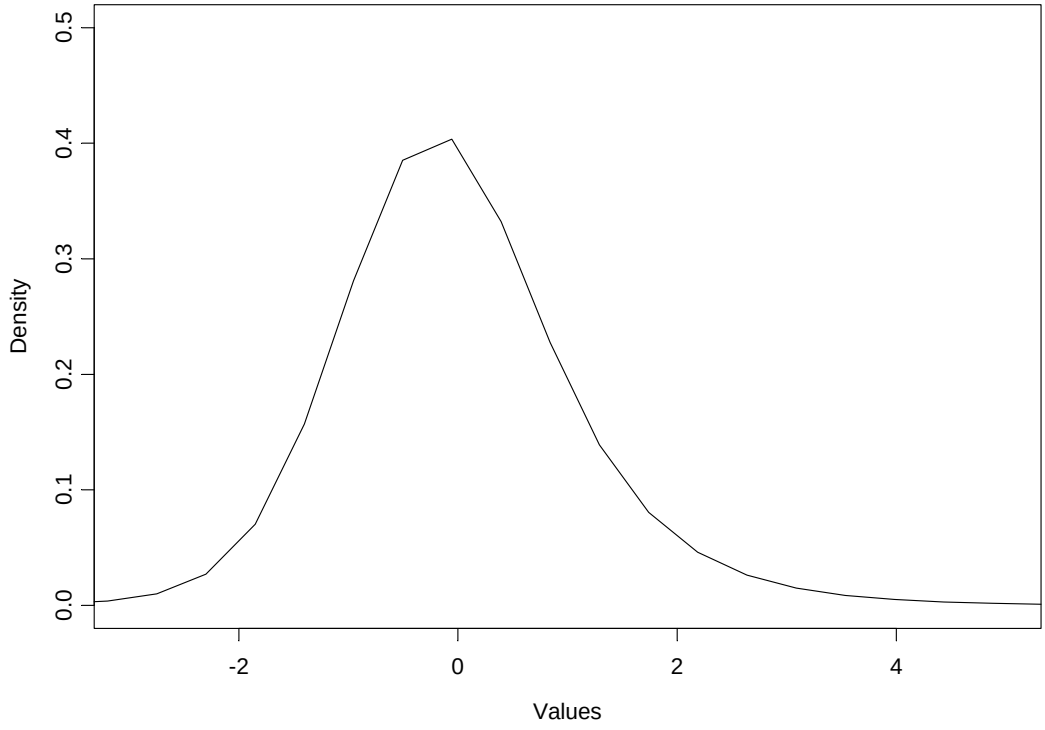

Figure 6. *Density of the distribution for the personality inventory (skewness=-0.116, kurtosis=0.254).*