September 25, 2012

Ms. Elizabeth M. Murphy
Secretary
U.S. Securities and Exchange Commission
100 F Street NE Washington, DC 20549-1090

Re: Suggested Sampling Procedure for Section 953(b) of the
Dodd-Frank Wall Street Reform and Consumer Protection Act

Dear Ms. Murphy,

I am writing as a law student at Stanford Law School and as a doctoral student in Stanford University's Department of Management Science and Engineering to offer a suggestion for how the Securities and Exchange Commission (SEC) can implement section 953(b) of the Dodd-Frank Act in a way that provides accurate information to investors, complies with the requirements of the statute, and minimizes implementation costs for companies.

Introduction

The Securities and Exchange Commission has received considerable public comment and discussion regarding how it should implement Section 953(b) of the Dodd-Frank Act, which calls upon companies to disclose the ratio of their CEO's compensation to their median worker compensation. In particular, many comment letters have expressed concern that it may be very difficult and expensive for companies, particularly large and complex companies, to determine their median worker compensation. This comment letter presents a simple procedure that will enable any company, through random sampling, to produce an estimate of its median compensation with a high degree of confidence. The degree of precision of this estimate can be made a function of the sample size, and the SEC, in consultation with businesses and the public can select the size that best balances the costs and benefits of this statutorily mandated reporting.

Overview of Procedure

Any company that wishes to estimate its median compensation can accomplish this by taking a random sample of its employees, calculating the compensation for each of those employees, and finding the median of that sample. A company can conduct such a sampling by assigning a unique identifying number to each of the company's employees and then using a computer to randomly select a given number of those identifiers. More complicated procedures, such as stratified sampling, will be unnecessary, regardless of the size of a company, how many countries it operates in, or how many subsidiaries it has.

The sample median generated through this process, for any odd-sized sample, will be a median-unbiased estimate of the company's true median compensation.[1] This means that the sample median will be no more likely to overestimate the true median than it is to underestimate it, and thus does not contain any "bias" to err in one direction more often than another.[2] Even a sample as small as 199 individuals will enable a company to achieve a 90% confidence level that the true median will be between the 89th and 111th entries in the sample, ranked by

[1] M. Mahamunulu Desu & R. H. Rodine, *Estimation of the Population Median*, SCANDINAVIAN ACTUARIAL JOURNAL 67-70 (1969:1-2).

[2] George Brown, *On Small-Sample Estimation,* THE ANNALS OF MATHEMATICAL STATISTICS: VOLUME 18 583 (1947).

compensation.[3] This will enable most companies to produce an estimate of their true median compensation that is, at a 90% confidence level, within $1300 of the actual median compensation. Larger samples would allow even greater levels of confidence and precision, as detailed below.

Procedural Details

Through the use of basic laws of probability, as outlined in Appendix A, it can be shown that if a company randomly selects a sample of N employees and computes the total compensation for each employee, then the company's level of confidence that the true median compensation will be between the kth largest and the kth smallest value in the sample is given by: $1-\sum_{j=0}^{k-1}\binom{N}{j}\left(\frac{1}{2}\right)^{N-1}$.[4] Using this formula, whose meaning and derivation are explained in Appendix A, the following table lists different sample sizes and corresponding levels of confidence and precision that they yield for an estimate of the true median. Thus, for example, if N = 199, and k = 89, a company could take a sample of 199 employees, rank them in ascending order of compensation, and thereby achieve 90% confidence that its true median compensation will be between the compensation of the 89th and the 111th individuals in the sample.

Table 1 – Exact k-Values for Confidence Intervals

Sample Size	**90% Confidence Interval**	**95% Confidence Interval**
	k - value	**k – value**
N = 99	k = 42	k = 40
N = 199	k = 89	k = 86
N = 399	k = 184	k = 180
N = 499	k = 232	k = 228
N = 999	k = 474	k = 469

These figures, however, only indicate that the median is likely to be between certain ranked values of the sample. They do not indicate how wide the range is likely to be in absolute dollar terms. Of course, any company that applies this methodology would immediately be able to determine the level of compensation of the kth lowest and kth highest paid employees in their sample, and so the dollar value that describes the width of the interval would be easy to report. Nevertheless, in determining the appropriate values to set for N and k, the SEC will likely want to prospectively consider how wide of intervals, in dollar terms, different values of N and k will produce. Clearly, this will be a matter for cost/benefit consideration that the SEC will be best poised to conduct with input from companies and the public.

The ideal would be for private companies or trade associations with access to employee compensation databases to work with the SEC as it implements this rule by using their databases to test the results from different values of N and k. Barring this, however, it is possible to

[3] *See generally* Morris H. DeGroot & Mark J. Schervish, Probability and Statistics 487, 491-93 (4th ed. 2010) (providing a technical discussion on the meaning and interpretation of confidence intervals).
[4] *See generally* John A. Rice, MATHEMATICAL STATISTICS AND DATA ANALYSIS 395-97 (3rd ed. 2007) (providing a general procedure for calculating confidence intervals for medians, upon which the methodology in this article is based).

generate simulated databases of employee compensation levels in order to estimate how wide in dollar terms the confidence levels produced by this procedure would be, were it applied to actual companies.

Using reasonable assumptions about the distribution of a company's compensation, as detailed in Appendix B, it can be shown that the confidence bands around the sample median would be of the size given in the table below:

Table 2 – Estimated Margins of Error for Simulated Company

Sample Size	Margin of Error 90% Confidence Level	Margin of Error 95% Confidence Level
N = 99	+/- $1905	+/- $2444
N = 199	+/- $1286	+/- $1668
N = 399	+/- $938	+/- $1178
N = 499	+/- $842	+/- $1035
N = 999	+/- $607	+/- $725

Thus, as can be seen, a relatively small sample can generate a narrow band that predicts the population median with a high degree of confidence.

The advantage of this procedure is that it is robust to different possible distributions of an employer's compensation. Although the dollar value that designates the size of the confidence intervals may vary from company to company, all companies will have the same level of confidence for the interval between the kth smallest and kth largest entry in a sample of a given size N. Nevertheless, it is certainly plausible that a particular company, with more detailed knowledge of its unique distribution of employee compensation, may be able to devise another sampling procedure that achieves predictions of comparable precision and accuracy to those specified here. Thus, the SEC could also specify in its rulemaking that if any company devises another procedure that it can demonstrate generates a median-unbiased estimator of the median within an X% confidence interval of +/- $Y of the estimated median, with the SEC selecting values for X and Y to best fit its cost/benefit analysis, then the company can substitute that estimator and procedure for this one specified here.

Conclusion

Section 953(b) does not specify how issuers must calculate the median of the annual total compensation of all employees. Because the median is a statistical term used to describe a set of observations, it is reasonable for the SEC to permit issuers to sample their employee populations to calculate the median. This approach will provide highly accurate information to investors with reduced compliance costs for issuers.

Please do not hesitate to contact me at ohlrogge@stanford.edu if you have any questions about this proposed methodology or if I can be of further assistance to the Commission in its efforts to implement the provisions of Section 953(b).

Yours truly,

Michael Ohlrogge,
J.D. Candidate, Stanford Law School,
Doctoral Student, Stanford Department of Management Science and Engineering.

APPENDIX A: Derivation of Formula for the Level of Confidence Between the kth Smallest and kth Largest Values in a Sample of Size N

The median is the value in a distribution that is greater than or equal to exactly half of the other values and less than or equal to exactly half of the other values. This special feature of the median lends itself very well to establishing a precise confidence interval for a company's true median compensation, based on taking just a small sample of employees and measuring their total compensation.

To start out with, suppose a company randomly and independently selects a sample of fifty employees and wants to know what the probability is that the true median compensation level is somewhere between the compensation level for the lowest paid employee and the highest paid employee. For every individual in the sample, there is a 50% chance that that individual's compensation is below the true median and a 50% chance that that individual's compensation is above the true median.[5] Therefore the chance that *all fifty employees* in the sample have compensation above the true median (in other words, the probability that the true median is less than the smallest value in the sample) is given by $\left(\frac{1}{2}\right)^{50}$. Likewise, the chance that *all fifty employees* in the sample have compensation below the true median (in other words, the probability that the true median is greater than the largest value in the sample) is also given by $\left(\frac{1}{2}\right)^{50}$. Furthermore these two events are mutually exclusive – it is impossible for all fifty employees in the sample to have compensation that is both above *and* below the actual company median. Therefore, the probability that either of these conditions holds is simply the sum of the probabilities that each of them individually holds, which is given by: $\left(\frac{1}{2}\right)^{50} + \left(\frac{1}{2}\right)^{50} = 2 \bullet \left(\frac{1}{2}\right)^{50} = \left(\frac{1}{2}\right)^{49}$.

In other words, if a company samples 50 employees and determines their individual compensations, then the probability that the company's true median compensation will not be somewhere between the lowest and highest paid individuals is given by $\left(\frac{1}{2}\right)^{49}$ = .000000000000002. Therefore, the company can achieve 99.9999999999998% confidence that the true median compensation will be somewhere between the lowest paid individual in the sample and the highest paid individual. By using the lowest and the highest compensations in the sample, the company will be able to achieve a 99.9999999999998% confidence interval for its true median compensation.

Clearly then, by forming an estimate based on the smallest and largest compensations in a sample, a company can achieve a very high confidence level. In practice, however, there is likely to be quite a wide range between the lowest and the highest compensated individuals, even in a 50-person sample. Thus, achieving 99.9999999999998% confidence that the true median is somewhere within such a wide range will not be tremendously informative. This procedure can be easily extended, however, to give confidence percentages for smaller intervals within a sample.

[5] Technical note: When sampling from an unknown continuous distribution, such as that of a company's employees, the value of any given observation is considered to be a continuous random variable. In statistical theory, the probability that a continuous random variable will assume any single value is considered for technical reasons to be zero. Thus, this section will refer to the probability that an individual employee has compensation that is *less than* the true median, rather than the probability that an employee has compensation that is *less than or equal to* the median. *See id.* at 47, 396.

To start with, suppose that a company wants to know the probability that the true median will be somewhere between the 2nd and the 49th observations in its sample, when those observations are ranked from smallest to largest. As with before, it is useful to break this down into pieces. First off, consider the probability that the company calculates compensations for a sample of fifty employees and that *forty-nine out of the fifty employees* in the sample have compensation above the true median (in other words, the probability that the true median is below the compensation of at least forty-nine out of the fifty employees). This equals the probability that the company would draw 50 observations, all 50 of which are greater than the population median (since if all 50 are greater than the true median, then it is also true that 49 are greater than the true median), plus the probability that the company would draw 50 observations, with exactly 49 of them greater than the true median. The first of these two probabilities was already calculated above as $\left(\frac{1}{2}\right)^{50}$.

For the second probability, there are 50 different combinations that could produce this because any of the fifty employees in the sample could be the lone employee whose compensation is less than the true median. For any given employee in the sample, the chance that their compensation is less than the true median is 50%, and the chance that all other 49 employees have compensations greater than the true median is $\left(\frac{1}{2}\right)^{49}$. Thus, the probability that that one particular employee's compensation is below the true median, *and* that all other employees have compensations above the true median is given by: $\left(\frac{1}{2}\right)\bullet\left(\frac{1}{2}\right)^{49}=\left(\frac{1}{2}\right)^{50}$. But, since there are fifty different employees for whom this could be true, the probability that *<u>any one</u>* employee has compensation below the true median, while all others have compensation above the true median, is given by: $50\bullet\left(\frac{1}{2}\right)^{50}$.

In situations such as this, where all of the different possible combinations that could satisfy a condition are added up, mathematicians use the binomial coefficient to represent the number of possibilities. In the case of calculating how many different combinations of one employee can be drawn out of a sample of 50 employees (i.e. how many different ways there could be exactly one employee whose compensation is below the sample median), the binomial coefficient is written as: $\binom{50}{1}$, which is read as "fifty choose one." Technically speaking, the binomial coefficient is defined as: $\binom{n}{j}=\frac{n!}{j!(n-j)!}$ for n ≥ j ≥ 0, where the factorial operator " ! " signifies $n!=n(n-1)(n-2)\cdots(1)$ and where by convention 0! = 1. Many standard calculators and computer programs also have built-in functions for the binomial coefficient, and in this case, it is easy to calculate that $\binom{50}{1}=50$. Thus, the probability that exactly one employee in a sample will have compensation below the true median is given by the number of possible combinations that could produce this (fifty), times the probability of any one of the possibilities ($\left(\frac{1}{2}\right)^{50}$). In other words, the probability that exactly one employee in a sample will have compensation below the true median is given by: $\binom{50}{1}\bullet\left(\frac{1}{2}\right)^{50}=50\bullet\left(\frac{1}{2}\right)^{50}$, exactly as before but now written using the binomial coefficient.

In fact, the binomial coefficient is implicitly present in the calculation of the probability that exactly fifty employees have compensation above the true median (in other words, the probability that exactly zero employees have compensation below the true median). In this case, the binomial coefficient is represented as $\binom{50}{0} = 1$ (note that there is only one way to choose exactly zero items out of a sample of fifty). Thus, the probability that all fifty employees will have compensation below the true median is given by $\binom{50}{0} \bullet \left(\tfrac{1}{2}\right)^{50} = 1 \bullet \left(\tfrac{1}{2}\right)^{50}$, again, exactly as before.

Putting all of these calculations together then, the probability that the true median will be less than the 2nd ranked sample compensation (i.e. the second lowest) is given by $\binom{50}{0}\left(\tfrac{1}{2}\right)^{50} + \binom{50}{1}\left(\tfrac{1}{2}\right)^{50} = \sum_{j=0}^{k-1}\binom{50}{j}\left(\frac{1}{2}\right)^{50}$, where here, k = 2 because the procedure is calculating the probability that the true median is less than the 2nd ranked sample.

As with above, the problem is symmetrical and therefore the probability that the true median will be greater than the 49th ranked employee in the sample (i.e. the second highest) is the same as the probability that the true median will be less than the 2nd ranked employee in the sample. Thus, the probability that the true median will be *between* the 2nd and the 49th ranked employees in the sample is given by $1 - 2 \bullet \sum_{j=0}^{k-1}\binom{50}{j}\left(\frac{1}{2}\right)^{50} = 1 - \sum_{j=0}^{k-1}\binom{50}{j}\left(\frac{1}{2}\right)^{49}$, which equals 0.99999999999991. Thus, a company would be able to report a 99.999999999991% confidence interval for their true median compensation by giving the interval bounded by the second lowest and the second highest compensations in a sample they took of 50 employees.

Finally then, this formula can be completely generalized. For a sample size of N, the confidence level that the true median is between the kth smallest and kth largest value in the sample is given by: $= 1 - \sum_{j=0}^{k-1}\binom{N}{j}\left(\frac{1}{2}\right)^{N-1}$.

APPENDIX B: Derivation of Sizes of Confidence Bands for Simulated Companies

In order to generate simulated employee compensation databases, it is necessary to make certain assumptions about the underlying distribution of employee compensation within companies. It is widely known that income within the population as a whole tends to be distributed in a log-normal form.[6] In such a distribution, there are a large number of employees who make low to medium salaries and a small number of employees who make very large salaries. Given the distribution of income within the population as a whole, it is a reasonable assumption that intra-company compensation distributions are also approximately log-normal.[7]

In formal mathematical terms, a log-normal distribution is created by taking a constant value and raising it to the power of each value found within a normal distribution. Thus, if Y is a normally distributed variable, with a mean of μ and a variance of σ^2, then $Z = e^Y$ will be a log-normally distributed variable, that is, $\ln(Z) \sim N(\mu, \sigma^2)$.[8] Thus, the only parameters whose values must be assumed in order to generate a simulated employee database with log-normally distributed compensation, are μ and σ^2.

The mean or expected value of a log-normally distributed population is given by $E(z) = e^{\mu + \sigma^2/2}$.[9] The median of a normal distribution is equal to its mean, and because the exponential function does not change the ordering of numbers it is applied to, the median of a log-normal distribution is simply e^{μ}.

Given these properties, an approximately 100,000 person company with a log-normal wage distribution that has a mean compensation of around $33,000 / year, a median compensation of around $5000 / year, and a CEO paid around $15,000,000 / year, would have approximate values $\mu \approx 8.5$ and $\sigma^2 \approx 3.8$. In order to generate the dollar estimates for the margins of error provided in Table 2, these values were inserted into a Monte Carlo simulation that created 10,000 simulated companies with log-normally distributed compensation according to values of μ = 8.5 and $\sigma^2 = 3.8$. Each of these 10,000 simulated companies was then randomly sampled, using sample sizes as indicated in the table. Finally, the difference between the kth smallest and the kth largest values in each of the samples was taken, and the median of the 10,000 differences from the simulated companies was calculated. The margins of error in the table are each one-half the size of these median figures.

[6] *See, e.g.*, Ronald Oaxaca & Michael Ransom, *Using Econometric Models for Intrafirm Equity Salary Adjustments*, JOURNAL OF ECONOMIC INEQUALITY 230 (Dec 2003; vol. 1, issue 3).

[7] *Id.*

[8] R. Carter Hill, William E. Griffiths & Guay C. Lim, PRINCIPLES OF ECONOMETRICS 103 (3rd ed. 2008).

[9] *Id.* at 104.