E C O N O M I C S B U L L E T I N

Standard error and confidence interval for QALY weights

Hiroshi Gunji
Japan Society of the Promotion of Science

Chie Hanaoka
Graduate School of Social Sciences, Hosei University

Abstract

There are some problems with the standard errors of QALY weights proposed by Groot (2000, Journal of Health Economics 19). The standard errors show smaller values than those of Groot when we recalculate using his method. Moreover, we correct the derivation of his approximation and derive corrected values. Because mean and variance do not exist for a distribution of QALY weights, using standard errors for statistical inference may lead to problems even when an approximation is used. In this paper, we verify the statistical properties of Groot's standard errors by simulation. We find that the corrected standard errors hold the same properties as a normal distribution under specific conditions. In general, however, it would be appropriate to use our simulation method to obtain critical values or p−value.

We would like to thank Seiritsu Ogura for his many helpful comments. Of course, all remaining errors are ours.
Citation: Gunji, Hiroshi and Chie Hanaoka, (2004) "Standard error and confidence interval for QALY weights." *Economics Bulletin*, Vol. 9, No. 1 pp. 1−12
Submitted: August 29, 2003. **Accepted:** March 26, 2004.
URL: http://www.economicsbulletin.com/2004/volume9/EB−03I10001A.pdf

1 Introduction

Measuring the level or quality of health is one of the most important issues in the economic evaluation of health care. When determining priorities in health care programs, this evaluation allows us to obtain information about which programs are appropriate for allocating resources efficiently. The health benefits that an individual derives from a particular health care intervention is defined according to enhanced quality and length of life. Quality-adjusted Life Years (QALYs) are widely used as an indicator of health benefits or health outcomes. The QALY measure incorporates both quality and quantity of life. It assigns utilities for health states, called QALY weights, on a scale from 0 to 1, where perfect health is defined as 1, and death as 0. This weight is then multiplied by the length of time spent in that health state. Put differently, QALYs are years of life in perfect health, used to value improvements in quality of life resulting from health care interventions.

QALY weights are measured using various methods. One method is to estimate ordered probit regressions using information on self-reported quality of health measures in a random sample (Cutler and Richardson, 1997). However, specific individual situations and characteristics may bias the answers to survey questions on subjective health state. For instance, people tend to take account of their age as they evaluate their health, so elders might consider their health quality higher than expected. This has been termed 'state dependent reporting errors' (Karkhofs and Lindeboom, 1995) or, 'scale of reference bias' (Groot, 2000).

In his seminal paper, Groot (2000) proposes a method to purge individual bias from self-reported health states, and to calculate QALY weights. People in the same health states may perceive their health state differently. To correct the individual bias, he uses an ordered probit model with varying bounds instead of the constant bounds assumed in the normal ordered probit model. Furthermore, he derives the standard errors for QALY weights and estimates them using the data of the British Household Panel Survey 1995. From the results, he concludes that since the weights corrected for individual bias are significantly smaller than those with the bias, his survey data has an upward individual bias.

However, there are some problems with the standard errors of QALY weights adopted by Groot. First, he makes two calculation errors. One is that he mistakes the standard deviation for the variance in computing the standard errors of the weights. The other is that he makes a calculation error in Taylor approximation. Consequently, we recalculate them and obtain the standard error that he originally attempts to calculate. Secondly, there is an issue with the approximation itself. Due to the fact that mean and variance do not exist for the Cauchy distribution, the estimates may be biased even if the approximation is used. Thus, we will attempt to determine how well Groot's approximation fits the data and which conditions lead to poor results. Moreover, we shall propose how to obtain critical values rather than the approximation.

The main results of the paper are follows. First, the standard errors recalculated with the Taylor approximation are far less than those of Groot. Therefore the parameters estimated by Groot are very reliable. Second, concerning statistical inference, the approximation method proposed by Groot may generally lead to bias although his analysis is fortunately the exception. Consequently, it is better to use critical values by simulation.

The rest of the paper is organized as follows. In Section 2, we derive a standard error using the Taylor approximation and recalculate them using the estimation results of Groot. In Section 3, we introduce a method of inference for confidence intervals by simulation and verify whether Groot's method is appropriate. Section 4 concludes the paper.

2 Groot's standard error

Groot uses the following procedure to estimate QALY weights. Suppose that the objective measure of health, H^0, linearly relates to the true quality of health, H^*,

$$H_n^* = \beta_0 + \beta_1 H_n^o + \beta_2 x_n + u_n, \tag{1}$$

where n $(n = 1, \ldots, N)$ is a single observation, x_n is a variable of individual characteristics, and u_n is a standard normal distributed error term. To simplify the notation, independent variable vectors are replaced with scalars. The true quality of health cannot be observed directly, but the subjective measure of health, H_n^s, can be observed. Because this variable is derived from the self-reported quality of health measure in survey data, we need to use a qualitative response model. Assume that

$$H_n^s = j \Leftrightarrow c_{j-1,n} < H_n^* \leq c_{j,n}, \quad j = 1, \ldots, J$$

The subjective health measure used by Groot has five ordered categories, because $j = 1, \ldots, 5$, that is, $J = 5$. When the standard ordered probit model is estimated, one can obtain δ_j which is the estimate of $c_{j,n}$. As noted in his corrigendum (Groot, 2001), he estimates δ_j using the ordered probit model with random bounds as proposed by Bolduc and Poole (1990).[1] The maximum likelihood estimation is attained from equation (1) and

$$c_{j,n} = \delta_j + \alpha_j H_n^o + \gamma_j x_n + \varepsilon_{j,n} \quad j = 2, \ldots, J-1$$

where $\varepsilon_{j,n}$ is a standard normal distributed error term and $c_{0,n} = -\infty$, $c_{1,n} = 0$, and $c_{J,n} = \infty$.[2] The QALY weight calculated by these estimators is

$$\text{QALY} = 1 - \frac{\beta}{\delta}, \tag{2}$$

and the QALY weight with biases is

$$\text{QALY}_b = 1 - \frac{\beta}{\delta + \alpha}, \tag{2'}$$

where $\beta \equiv \beta_1$, $\delta \equiv \delta_4$, and $\alpha \equiv \alpha_4$ are assumed to be independently distributed. Equation (2) depends on a Cauchy distribution because both β and δ are normally distributed. Recall that mean and variance do not exist for a Cauchy distribution, so we cannot carry out statistical inference. The problem arises from this type of QALY weight estimation, e.g., Catler and Richardson (1997). Consequently, Groot tries to estimate the standard errors using the Taylor approximation. Nevertheless, there are two problems with this method.

[1] Although Bolduc and Poole propose two models with random bounds, Groot does not note which model is used.

[2] Due to this assumption, equations (2) and (3) are derived.

2.1 The estimation of standard error

The problem is that he mistakes the standard deviation for the variance in calculating standard errors. If we want to determine the standard errors of equation (2), it follows that we should calculate the variance of β/δ. Hereafter, tables with *Arabic* numerals represent those in Groot's paper while tables with *Roman* numerals are ours. In his paper, Table 2 presents the coefficients of the standard ordered probit model. Table 3 presents the estimates of the ordered probit model with varying bounds. Table 4 shows QALY weights computed by the estimates from Tables 2 and 3. Let x_1 and x_2 be independent normal distributed random variables, $x_1 \sim N(\beta, \sigma_1)$ and $x_2 \sim N(\delta, \sigma_2)$, and $z = x_1/x_2$, following Groot (2000, p. 410). It is important to note that both σ_1 and σ_2 are *variances* rather than standard deviations. In his calculation, the first- and second-order moments are

$$\mathrm{E}(z) = \frac{\beta}{\delta} + \frac{\sigma_1}{\delta} + \frac{2\beta\sigma_2}{\delta^3} \tag{3}$$

$$\mathrm{E}(z^2) = \frac{\beta^2}{\delta^2} + \frac{2\sigma_1}{\delta^2} + \frac{6\beta\sigma_2}{\delta^4}. \tag{4}$$

From these results, we obtain the standard error,

$$\begin{aligned}\mathrm{SE}_g &= \sqrt{\mathrm{E}(z^2) - \mathrm{E}(z)^2} \\ &= \left[\frac{\sigma_1}{\delta^2}(2 - 2\beta - \sigma_1) + \frac{2\beta\sigma_2}{\delta^4}(3 - 2\beta - 2\sigma_1) - \frac{4\beta^2(\sigma_2)^2}{\delta^6}\right]^{1/2}.\end{aligned}$$

Using the estimates $\beta = 0.374$ and $\delta = 3.570$ of "Problems with arms, legs, etc." in his Table 2, and the t-values $t_\beta = 23.476$ and $t_\delta = 85.885$, for example, the moments are

$$\mathrm{E}(z) \simeq 0.1891, \quad \mathrm{E}(z^2) \simeq 0.0358.$$

Therefore, the standard error is

$$\mathrm{SE}_g \simeq 0.0081.$$

However, Groot's Table 4, which presents estimated QALY weights and their standard errors, shows that the standard error calculated above is 0.04. This value arises unfortunately from mistaking the standard deviation for the variance. The variances, $\sigma_1 = (0.674/23.476)^2$ and $\sigma_2 = (3.570/85.885)^2$, are needed for the calculation of each variance. Nevertheless, the standard errors, $\sigma_1 = 0.674/23.476$ and $\sigma_2 = 3.570/85.885$, are computed in his Table 4. In short, both σ_1 and σ_2 are variances by definition, but the standard errors are substituted. Therefore, we need to recalculate all the standard errors in his Table 4.

In our Table I, we present the standard errors of Groot's replication, SE_0, and that of his approximation, SE_g. The SE_g's are far smaller than the SE_0's. In other words, the QALY weights he estimates are quite reliable.

2.2 Taylor Approximation

Another problem is left: miscalculations in the Taylor approximation. Therefore, we need to recalculate it in the following way. First, we derive the first-order

moment. Defining $z = f(x_1, x_2) = x_1/x_2$, the partial derivatives of $f(x_1, x_2)$ with respect to x_1 and x_2 are

$$f_1(x_1, x_2) = \frac{1}{x_2}, \quad f_2(x_1, x_2) = -\frac{x_1}{(x_2)^2}$$

$$f_{11}(x_1, x_2) = 0, \quad f_{12}(x_1, x_2) = -\frac{1}{(x_2)^2}, \quad f_{22}(x_1, x_2) = \frac{2x_1}{(x_2)^3}.$$

Then a second-order Taylor-series approximation of $f(x_1, x_2)$ around (β, δ) yields

$$\begin{aligned} f(x_1, x_2) \simeq\ & f(\beta, \delta) + \frac{1}{1!}\left[f_1(\beta, \delta)(x_1 - \beta) + f_2(\beta, \delta)(x_2 - \delta)\right] \\ & + \frac{1}{2!}\left[f_{11}(\beta, \delta)(x_1 - \beta)^2 + 2f_{12}(\beta, \delta)(x_1 - \beta)(x_2 - \delta) + f_{22}(\beta, \delta)(x_2 - \delta)^2\right]. \end{aligned}$$

Therefore, we have

$$f(x_1, x_2) = \frac{\beta}{\delta} + \frac{1}{\delta}(x_1 - \beta) - \frac{\beta}{\delta^2}(x_2 - \delta) - \frac{1}{\delta^2}(x_1 - \beta)(x_2 - \delta) + \frac{\beta}{\delta^3}(x_2 - \delta)^2. \quad (5)$$

Taking the expectation of equation (5), we have

$$\mathrm{E}(z) = \frac{\beta}{\delta} + \frac{\beta\sigma_2}{\delta^3}. \quad (6)$$

This differs from equation (3).

Next, we will compute the second-order moment. Defining $z^2 = g(x_1, x_2) = (x_1)^2/(x_2)^2$, the partial derivatives of $g(x_1, x_2)$ with respect to x_1 and x_2 are

$$g_1(x_1, x_2) = \frac{2x_1}{(x_2)^2}, \quad g_2(x_1, x_2) = -\frac{2(x_1)^2}{(x_2)^3}$$

$$g_{11}(x_1, x_2) = \frac{2}{(x_2)^2}, \quad g_{12}(x_1, x_2) = -\frac{4x_1}{(x_2)^3}, \quad g_{22}(x_1, x_2) = \frac{6(x_1)^2}{(x_2)^4}.$$

Then a second-order Taylor-series approximation of $g(x_1, x_2)$ around (β, δ) yields

$$\begin{aligned} g(x_1, x_2) \simeq\ & \frac{\beta^2}{\delta^2} + \frac{2\beta}{\delta^2}(x_1 - \beta) - \frac{2\beta^2}{\delta^3}(x_2 - \delta) \\ & + \frac{1}{\delta^2}(x_1 - \beta)^2 - \frac{4\beta}{\delta^3}(x_1 - \beta)(x_2 - \delta) + \frac{3\beta^2}{\delta^4}(x_2 - \delta)^2. \end{aligned} \quad (7)$$

Taking the expectation of equation (7), we have

$$\mathrm{E}(z^2) = \frac{\beta^2}{\delta^2} + \frac{\sigma_1}{\delta^2} + \frac{3\beta^2\sigma_2}{\delta^4}. \quad (8)$$

This differs from equation (4).[3] Hence, the corrected standard error is

$$\mathrm{SE}_c = \left[\frac{\sigma_1}{\delta^2} + \frac{\beta^2\sigma_2}{\delta^4} - \frac{\beta^2(\sigma_2)^2}{\delta^6}\right]^{1/2}.$$

[3]Although we also calculate these moments for third- and fourth-order approximations, the results are different from equations (3) and (4).

The estimates of SE_c are presented in our Table I. For instance, the moments of the corrected approximation about "problems with arms, legs, etc." are

$$\mathrm{E}(z) \simeq 0.1888, \quad \mathrm{E}(z^2) \simeq 0.0357.$$

The corrected standard error is

$$\mathrm{SE}_c \simeq 0.0083.$$

In short, the standard errors of QALY weights are very small when SE_c is used. Therefore we need to reexamine the interpretation of the estimation results of Groot's QALY weights. For instance, he suggests that "the QALY weights for difficulties in hearing and seeing are not statistically different from 1." This suggestion may be right if we use SE_0 from the third column of the second row in his Table 2, i.e., $(0.92 - 1)/0.08 = -1.00$. Nevertheless, it is significantly different from 1 when we use SE_c, i.e., $(0.92 - 1)/0.0083 \simeq -9.64$.

3 Simulation

The second problem with Groot's approximation originates from idea in the approximation itself. Because mean and variance do not exist for Cauchy distributions, estimates may be biased even if approximations are used. Thus, we examine the distribution of QALY weights by simulation in order to investigate the properties of Groot's standard error. First, we obtain critical values in the following way:

Step 1: Generate two independent random numbers, $x_1 \sim N(\beta, \sigma_1)$ and $x_2 \sim N(\beta, \sigma_2)$, and retain QALY weights.

Step 2: Repeat Step 1. (say 100,000 times)

Step 3: Sort the simulated QALY weights in ascending order with each upper and lower quartile p (say, 0.05) being respectively assigned a critical value of $100 \times p$ and $100 \times (1 - p)$ percent.

It is straightforward to draw out p-values from this simulation. Yet, we are interested in the distribution of QALY weights, so we will show the percentile.

Table II presents a simulation of QALY weights using Groot's parameters. The replication is done 100,000 times. All critical values are symmetric (right-left mirror images) to QALY weights and the 0.50 percentile, i.e., the median, are equivalent to the estimates of QALY weights in Table I. Groot also verifies whether QALY weights are statistically different from 1. We find that the QALY weights are sufficiently different from 1 for all health conditions except alcohol, in which the value may be greater than 1.0.

We focus on an interesting result, the 68% confidence interval. This half-width of the confidence interval is identical with the corrected standard error, SE_c. For instance, "arms" in the first row in Table II is $(0.820-0.803)/2 = 0.008$. Recall that in a normal distribution, the half-width of the 68% confidence interval is equal to the standard error. This result leads us to the idea that QALY weights may be normally distributed. Actually, the values in Table II are very close to the normal distribution with mean $1 - \beta/\delta$ and variance $(\mathrm{SE}_c)^2$. That

is, the distribution of estimated QALY weights is extremely close to a normal distribution, and Groot's standard error, SE_c, can be considered a prominent approximation in the above examination. This fact, however, depends on whether the standard error of x_2 is very close to 0. If x_2 is a constant δ, that is, $\sigma_2 \to 0$, then x_1/x_2 depends only on the distribution of x_1. Because x_1 is normally distributed, then x_1/x_2 is normally distributed with the mean β/δ and the variance σ_1/δ^2. This distribution is considered to be almost the same distribution of the normal distribution, with the mean and the variance using Groot's approximation. Since the t-value of δ estimated by Groot is an enormous number, 85.885, its standard error is very small. Above all, Groot's approximation is fortunately adapted when the standard error of δ happens to be extremely small.

If so, what sizes of standard errors make Groot's approximation sufficient? Figure I represents the normal distribution with mean $1-\beta/\delta$ and variance $(\mathrm{SE}_c)^2$ and the histogram created by 100,000 random numbers of $1-x_1/x_2$ where $\beta = 0.674$, $\delta = 3.570$, and $t_\beta = 2$ in various t_δ's. In the case of $t_\delta = 2$, the distribution of QALY weights is skewed to the right even if δ is considered to be at a sufficient level. In $t_\delta = 5$, this appears to be very close to the normal distribution. In the case of $t_\delta = 10$ or $t_\delta = 30$, it is almost the same as the normal distribution. However, this bias may become serious as we usually face t-values such as 2 or 3. To confirm the difference between the distribution by the estimated QALY weights and the normal distribution, we perform the Jarque-Bera (JB) test for normality on a series of the simulated QALY weights. The null hypothesis of the test is that series is normally distributed, and the alternative hypothesis is that it is not. Table III shows the p-value of the JB test using the series in various t_δ's. Normality is not accepted until t_δ become extremely large. This is different from the impression from Figure I. For example, the null hypothesis is rejected until $t_\delta = 20$. The normality of the series is not accepted until the series of $t_\delta = 22$ under the 5% significance level. Moreover, normality is stably accepted when the series is greater than $t_\delta = 42$.

The most crucial thing for us in estimating QALY weights is to obtain the probability that we may reject the null hypothesis when it is, in fact, true using Groot's standard error. We test the sample, $1-x_1/x_2$, using Groot's parameters again and determine the probability of a Type I error carrying out the statistical test under a normal distribution with the mean $1-\beta/\delta$ and variance $(\mathrm{SE}_c)^2$. We use the statistic,

$$Z = \frac{\mathrm{QALY} - (1-\beta/\delta)}{\mathrm{SE}_c},$$

to test whether it depends on the standard normal distribution. Table IV shows the result from the 1,000,000 iterations if each nominal size is 0.01, 0.05, and 0.10. In this test, the actual size is nearly identical to the nominal size when t_δ is 20, compared with the JB test in which normality is not accepted even if t_δ is enormous. The size distortion is also serious when $t_\delta = 2$. For example, the actual size reaches 12% if the nominal size is 1%. Put differently, it is rejected almost 12% even if QALY weights follow the true distribution.

Consequently, Groot's approximation shows very good performance. It depends on the result of t_δ being enormous. Nevertheless, in general, statistical inference using Groot's approximation may have huge bias. Consequently, we recommend our method using the confidence interval of QALY weights.

4 Conclusion

Two calculation errors and the idea of approximation cause problems with the standard errors of QALY weights proposed by Groot. Concerning the first two erros, we recalculate them and find that his estimates are quite reliable. For the second problem, we demonstrate that his method of approximation is limited to a specific condition, where the value of t_δ is large. So, in general it is appropriate to use our simulation method to obtain confidence intervals or p-values for QALY weights.

References

[1] Bolduc, Denis and Erik Poole (1990), "Ordinal probit model with random bounds," *Economics Letters* 33: 239-244.

[2] Cutler, David M. and Elizabeth Richardson (1997), "Measuring the health of United States population," *Brookings Papers on Economic Activity, Microeconomics* 3: 217-271.

[3] Groot, Wim (2000), "Adaptation and scale of reference bias in self-assessments of quality of life," *Journal of Health Economics* 19: 403-420.

[4] Groot, Wim (2001), "Corrigendum to 'Adaptation and scale of reference bias in self-assessments of quality of life'," *Journal of Health Economics* 20: 145.

[5] Karkhofs, M. and M. Lindeboom (1995), "Objective health measures and state dependent reporting errors," *Health Economics* 4: 221-235.

Table I: QALY weights by health condition (revision)

	Using estimates from Table 2				Using estimates from Table 3				Using estimates from Table 3			
	$1-\frac{\beta}{\delta}$	SE_0	SE_g	SE_c	$1-\frac{\beta}{\delta}$	SE_0	SE_g	SE_c	$1-\frac{\beta}{\delta+\alpha}$	SE_0	SE_g	SE_c
Problems with arms, legs, etc.	0.811	0.044	0.008	0.008	0.837	0.072	0.030	0.017	0.841	0.075	0.030	0.017
Difficulty in seeing	0.923	0.078	0.018	0.015	0.937	0.105	0.037	0.025	0.936	0.112	0.037	0.025
Difficulty in hearing	0.970	0.077	0.016	0.012	0.937	0.094	0.031	0.019	0.942	0.088	0.031	0.019
Skin conditions, allergies	0.966	0.072	0.014	0.011	0.965	0.084	0.024	0.014	0.968	0.078	0.024	0.014
Chest, breathing problems	0.834	0.051	0.010	0.010	0.863	0.080	0.031	0.018	0.865	0.084	0.031	0.018
Heart, blood	0.841	0.054	0.011	0.011	0.828	0.076	0.032	0.022	0.844	0.071	0.032	0.022
Stomach, liver, kidney	0.798	0.049	0.011	0.013	0.766	0.058	0.032	0.034	0.782	0.060	0.032	0.034
Diabetes	0.856	0.076	0.021	0.021	0.853	0.101	0.046	0.040	0.860	0.102	0.046	0.040
Nerves, anxiety, depression	0.787	0.045	0.010	0.013	0.816	0.079	0.036	0.030	0.816	0.089	0.036	0.030
Alcohol, drugs	0.820	0.088	0.041	0.049	0.846	0.116	0.076	0.084	0.871	0.105	0.076	0.084
Epilepsy	0.867	0.098	0.037	0.036	0.850	0.111	0.060	0.060	0.861	0.111	0.060	0.060
Migraine, chronic headaches	0.937	0.073	0.015	0.012	0.927	0.089	0.030	0.018	0.929	0.092	0.030	0.018
Other	0.764	0.040	0.009	0.015	0.804	0.077	0.036	0.034	0.797	0.090	0.036	0.034

Table II: Critical values for QALY weights by health condition

	Percentile										
	0.01	0.025	0.05	0.10	0.16	0.50	0.84	0.90	0.95	0.975	0.99
Using estimates from Table 2											
Arms, legs	0.792	0.795	0.797	0.801	0.803	0.811	0.820	0.822	0.825	0.828	0.831
Seeing	0.889	0.895	0.899	0.905	0.909	0.923	0.938	0.942	0.947	0.952	0.957
Hearing	0.942	0.946	0.950	0.954	0.958	0.970	0.982	0.985	0.989	0.993	0.998
Skin	0.942	0.946	0.949	0.953	0.956	0.966	0.977	0.980	0.984	0.987	0.991
Chest	0.812	0.815	0.818	0.822	0.825	0.834	0.844	0.847	0.851	0.854	0.857
Heart	0.816	0.820	0.823	0.827	0.830	0.841	0.851	0.854	0.858	0.861	0.865
Stomach	0.767	0.772	0.776	0.781	0.785	0.798	0.811	0.815	0.820	0.824	0.829
Diabetes	0.805	0.813	0.820	0.828	0.834	0.856	0.877	0.883	0.891	0.898	0.906
Nerves	0.757	0.762	0.766	0.771	0.774	0.787	0.800	0.804	0.808	0.812	0.817
Alcohol	0.706	0.725	0.740	0.758	0.772	0.820	0.869	0.883	0.900	0.916	0.934
Epilepsy	0.784	0.797	0.808	0.821	0.831	0.867	0.903	0.913	0.926	0.937	0.951
Headaches	0.909	0.913	0.917	0.921	0.925	0.937	0.949	0.952	0.957	0.960	0.965
Other	0.730	0.735	0.740	0.745	0.749	0.764	0.778	0.783	0.788	0.793	0.798
Using estimates from Table 3											
Arms, legs	0.791	0.800	0.806	0.814	0.819	0.837	0.853	0.857	0.863	0.867	0.872
Seeing	0.875	0.885	0.894	0.904	0.911	0.937	0.962	0.969	0.978	0.986	0.995
Hearing	0.889	0.897	0.904	0.911	0.917	0.937	0.956	0.961	0.968	0.974	0.981
Skin	0.931	0.936	0.941	0.947	0.951	0.965	0.979	0.983	0.988	0.992	0.997
Chest	0.816	0.824	0.831	0.838	0.844	0.863	0.881	0.886	0.892	0.897	0.903
Heart	0.771	0.781	0.789	0.798	0.805	0.828	0.848	0.854	0.861	0.867	0.874
Stomach	0.679	0.694	0.707	0.720	0.731	0.766	0.799	0.808	0.819	0.828	0.840
Diabetes	0.755	0.771	0.785	0.801	0.813	0.853	0.893	0.904	0.918	0.930	0.943
Nerves	0.740	0.753	0.764	0.776	0.785	0.816	0.845	0.853	0.863	0.872	0.882
Alcohol	0.644	0.677	0.705	0.737	0.762	0.846	0.929	0.952	0.983	1.009	1.039
Epilepsy	0.704	0.727	0.748	0.771	0.789	0.849	0.908	0.925	0.946	0.965	0.987
Headaches	0.883	0.891	0.897	0.904	0.909	0.927	0.945	0.950	0.956	0.961	0.967
Other	0.718	0.733	0.745	0.759	0.770	0.804	0.837	0.846	0.857	0.867	0.878
Using estimates from Table 3											
Arms, legs	0.800	0.807	0.813	0.820	0.825	0.841	0.856	0.860	0.866	0.870	0.875
Seeing	0.872	0.883	0.892	0.901	0.909	0.936	0.961	0.969	0.978	0.986	0.995
Hearing	0.899	0.906	0.912	0.919	0.924	0.942	0.959	0.964	0.971	0.976	0.982
Skin	0.938	0.943	0.947	0.952	0.955	0.968	0.981	0.984	0.989	0.993	0.997
Chest	0.818	0.826	0.833	0.840	0.846	0.865	0.882	0.887	0.893	0.898	0.904
Heart	0.798	0.806	0.812	0.820	0.825	0.844	0.862	0.867	0.873	0.879	0.884
Stomach	0.705	0.718	0.729	0.741	0.750	0.782	0.812	0.820	0.830	0.839	0.849
Diabetes	0.767	0.782	0.796	0.810	0.822	0.860	0.897	0.908	0.922	0.933	0.947
Nerves	0.740	0.753	0.764	0.776	0.785	0.816	0.845	0.853	0.863	0.872	0.881
Alcohol	0.702	0.730	0.754	0.780	0.802	0.872	0.940	0.960	0.986	1.008	1.032
Epilepsy	0.727	0.749	0.768	0.789	0.805	0.861	0.915	0.931	0.951	0.968	0.988
Headaches	0.885	0.892	0.898	0.905	0.911	0.929	0.946	0.951	0.957	0.962	0.968
Other	0.706	0.721	0.734	0.749	0.760	0.797	0.831	0.840	0.852	0.862	0.874

Figure I: QALY weights and normal distribution

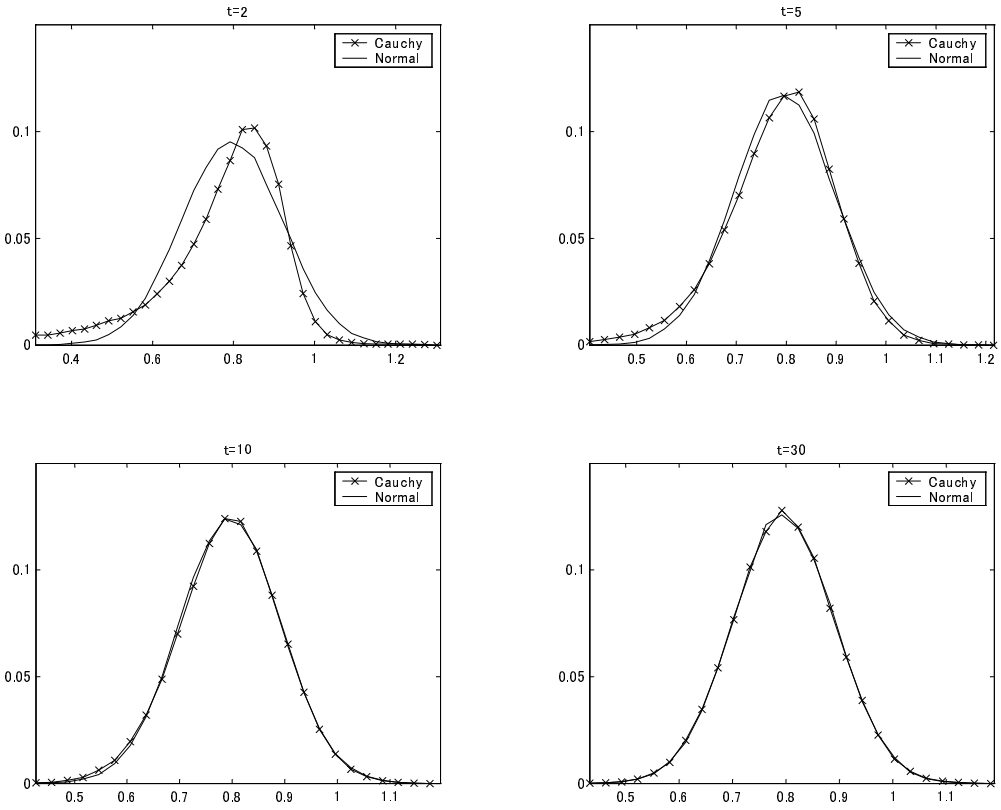

Table III: Jarque-Bera test for normality

t_δ	p-value	t_δ	p-value	t_δ	p-value	t_δ	p-value
1	0.000	21	0.001	41	0.008	61	0.207
2	0.000	22	0.051	42	0.967	62	0.405
3	0.000	23	0.002	43	0.440	63	0.741
4	0.000	24	0.007	44	0.973	64	0.697
5	0.000	25	0.001	45	0.805	65	0.975
6	0.000	26	0.243	46	0.667	66	0.441
7	0.000	27	0.096	47	0.262	67	0.891
8	0.000	28	0.082	48	0.734	68	0.026
9	0.000	29	0.778	49	0.174	69	0.085
10	0.000	30	0.025	50	0.910	70	0.602
11	0.000	31	0.025	51	0.724	71	0.405
12	0.000	32	0.193	52	0.654	72	0.803
13	0.000	33	0.447	53	0.433	73	0.473
14	0.000	34	0.085	54	0.014	74	0.980
15	0.000	35	0.346	55	0.363	75	0.815
16	0.000	36	0.053	56	0.496	76	0.467
17	0.000	37	0.058	57	0.368	77	0.728
18	0.000	38	0.511	58	0.288	78	0.134
19	0.000	39	0.115	59	0.649	79	0.236
20	0.000	40	0.766	60	0.796	80	0.215

Table IV: Type I error

t_δ	Nominal size		
	0.01	0.05	0.10
1	0.289	0.335	0.389
2	0.120	0.156	0.188
3	0.062	0.104	0.146
4	0.037	0.081	0.127
5	0.026	0.070	0.119
6	0.020	0.063	0.112
7	0.017	0.060	0.109
8	0.015	0.057	0.108
9	0.014	0.056	0.106
10	0.013	0.055	0.105
11	0.012	0.054	0.104
12	0.012	0.053	0.103
13	0.012	0.053	0.102
14	0.011	0.052	0.103
15	0.011	0.052	0.102
16	0.011	0.052	0.101
17	0.011	0.052	0.102
18	0.011	0.051	0.101
19	0.011	0.051	0.101
20	0.011	0.051	0.101