POINT AND INTERVAL ESTIMATES FROM SEQUENTIAL SAMPLING

by

Andrew P. Robinson

and

Thomas E. Burk

Staff Paper Series No. 122

DEPARTMENT OF FOREST RESOURCES

College of Natural Resources
and
Minnesota Agricultural Experiment Station
University of Minnesota
St. Paul, Minnesota

For more information about the Department of Forest Resources and its teaching, research and outreach programs contact the department head at:

Department of Forest Resources
University of Minnesota
115 Green Hall
1530 North Cleveland Avenue
St. Paul, MN 55108

Ph: 612-624-3400; Fax: 612-625-5212
Email: aek@forestry.umn.edu

POINT AND INTERVAL ESTIMATES FROM SEQUENTIAL SAMPLING [1]

by

Andrew P. Robinson

and

Thomas E. Burk

JULY 1997

Staff Paper Series No. 122

---

1 Research supported by the College of Natural Resources and the Minnesota Agricultural Experiment Station, University of Minnesota, St. Paul, the McIntire-Stennis Cooperative Forestry Research Program, and the USDA Small Business Innovation Research Program. Published as MAES Paper no. 974420122 of the Minnesota Agricultural Experiment Station.

The authors are Research assistant and Professor, Department of Forest Resources, Univ. of Minnesota, St. Paul.

ABSTRACT

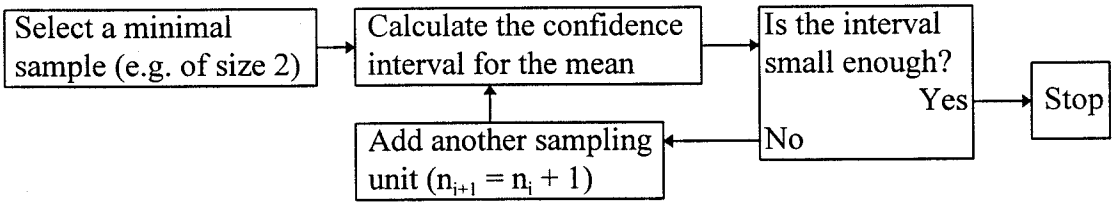
This paper reports a simulation-based exploration into the computation of point and interval estimates for data arising from sequential sampling. We conclude that the coverage probability of the standard frequentist confidence interval estimates is overstated. However, there are other interval estimates which do not overstate the coverage probability if the underlying population is Gaussian. Further, the effect of non-Gaussian behavior in the underlying population upon the properties of the interval estimate varies, depending upon the severity and the flavor, that is, whether it is skewness, kurtosis, etc.

INTRODUCTION

Sequential sampling has been a popular data collection technique within the biological sciences for the last 50 years (see for example Smith and Ker 1958, Mason 1969, Mukhopadhyay 1991). Whereas simple random sampling is effectively sampling in a batch, with a predetermined sample size, sequential sampling instead uses a sampling rule which permits the conclusion of sampling when a certain criterion, such as a sufficiently small confidence interval, has been reached. This is a simple variation of adaptive sampling presented by Thompson (1990) and Thompson and Seber (1996), in which the location, as well as the number, of the next sampling points depend on previously sampled values.

It is widely known among statisticians that one cannot safely apply the standard frequentist confidence intervals to data which have been collected by sequential sampling. This is less well-known within the broader community of researchers who apply such sampling techniques for data collection. In this paper, we express the problem in both statistical and heuristic terms. We then review the performance of a proposed solution and the assumptions which it requires.

We constrain our considerations to the single-level case where one applies the following algorithm:



The estimate of the interval is then calculated as:

$$(C_L, C_U) = \left(\bar{x} - t_{\alpha, n-1} \, s_x / \sqrt{n}, \bar{x} + t_{\alpha, n-1} \, s_x / \sqrt{n} \right) . \qquad (1)$$

This algorithm is easily expressed in Splus (Statistical Sciences, Inc. 1995), in which all subsequent computations have been performed.

At first examination it seems that calculating the confidence interval as though the data were collected within a simple random sample (Cochran 1977) seems quite reasonable. The key is that in frequentist theory, the sample size is considered fixed, whereas under the sequential sampling plan, the sample size becomes a random variable. Therefore the frequentist intervals are calculated conditional on the sample size, which means that the sample size is treated as though it were a constant. Under the sequential plan the sample size clearly is not constant, and it is not clear that it's reasonable to go ahead and pretend that it is.

The critical point to consider with estimates of frequentist intervals and point estimates is the sampling distribution from which the sample(s) are assumed to have been drawn. In performing simple random sampling, we are sampling a random interval from a population of intervals. In this sequential sampling however, we are only accepting the random interval if it is smaller than a predefined cutoff. Otherwise, we reject the sample and move to a different sample size. So, if we condition on the sample size, then any estimate of the interval arriving from a sample of that size actually comes from a censored distribution. Therefore the frequentist assumptions and large-sample theory will not be justifiable.

However, the intervals are only unacceptable from the viewpoint of frequentism. If the analyst prefers to take a Bayesian stance the objections become irrelevant. Such a decision should not be made lightly (see, e.g. Berger 1985). The analysis of this particular case is problematic, as the prior must be updated at each new realization, an expensive process.

The root of the problem is likely to be that data arising from such a design will not be exchangeable. Briefly, exchangeability requires that any permutation of the order of the units within the sample will be as likely to occur as any other permutation (Berger 1985). The lack of exchangeability is illustrated in the following example: consider a sequence of units sampled by the sequential sampling technique outlined above. Sampling concludes when the confidence interval estimated from the sample is sufficiently small. Therefore the probability of obtaining a sequence with its most extreme value last in the sequence is obviously much less likely than obtaining the same set of units but with the least extreme observation last. These two possibilities are therefore not exchangeable.

We emphasize that point estimates arising from sequential sampling are not subject to these concerns if the underlying distribution is Gaussian. However, there is a possibility that in an asymmetric distribution, especially one on the positive real number line such as arises for many biological variables, the estimate of the mean might not be independent of the sample size. To see this, recall that outliers will lead to large standard errors for estimates. In the sequential case, therefore, any outliers will have to be compensated for

by the collection of extra data. Since outliers can really only occur in one direction for positive, real-valued populations, they will also affect the mean. Therefore it is possible that there will be some dependence between the estimate of the mean and the sample size.

LITERATURE REVIEW

Some attention has been paid to this problem in the past. Its origin seems to be with Stein (1945) in what he called two-stage sampling, in which the first stage sample is used to estimate the variance to design for the second stage sample size. A version much more similar to the present problem was also explored by Anscombe (1953). More recently, Hall (1983) introduced a variation called accelerated sequential sampling, which involves taking a sample in a fixed-size batch, then sampling sequentially until a determined cutoff is reached, and then sampling another batch.

Edelman (1991) presented a quick introduction as well as a simple solution for Gaussian populations. Presenting theoretical and simulated results to justify it, he recommends simply removing 5 more degrees of freedom for obtaining the t-statistic used to calculate the interval before comparing with the target interval. In other words, when using the *t*-statistic, take n-6 degrees of freedom instead of n-1.

Most recently, Mukhopadhyay has taken an explicitly frequentist decision theoretic approach, see e.g. Mukhopadhyay (1991) and Mukhopadhyay and Datta (1996). This involves explicitly writing the objective function to include the cost of sampling (most often with a linear cost function) and then applying optimization to the resultant function.

SIMULATION RESULTS

We wrote code (Appendix 1) for Splus simulations following Edelman's (1991) recommendations. The results are summarized in Table 1 and the figures following the appendix.

Table 1. 5000 runs, population size 500 units, sampling with replacement.

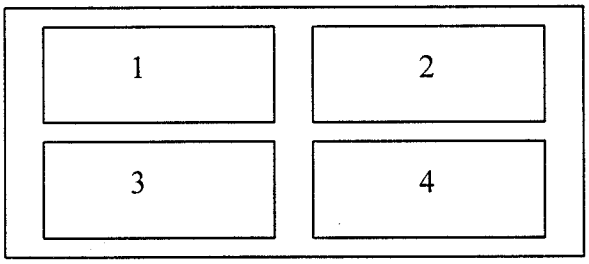
Distribution	Corr.	Target	Coverage	SE (Mean)	SBeSE M	SBeSE S	n (mean)	Figure
Gaussian	0	1	0.89	0.46	0.32	0.12	2.5	1
	5	1	0.99	0.32	0.31	0.06	9.8	2
	0	0.5	0.88	0.32	0.22	0.05	15.0	3
	5	0.5	0.95	0.24	0.22	0.02	18.0	4
	0	0.25	0.91	0.18	0.12	0.02	56.4	5
	5	0.25	0.95	0.13	0.12	0.001	62.2	6
Exp(1)	0	1	0.84	0.41	0.28	0.13	5.0	7
	5	1	0.97	0.31	0.29	0.09	9.8	8
T(2)	0	1	0.90	0.51	0.39	0.13	16.0	9
	5	1	0.98	0.43	0.41	0.08	21.8	10
T(5)	0	1.265	0.92	0.53	0.36	0.15	5.8	11
	5	1.265	0.99	0.40	0.37	0.09	9.8	12
N(5,1)+Exp(1)	0	1	0.82	0.95	0.42	0.12	18.9	14
	5	1	0.93	0.54	0.46	0.02	22.1	15

In Table 1, *Distribution* gives the distribution of the population, *Correction* is the number of extra degrees of freedom removed from the interval, *Target* is the desired interval width, *Coverage* is the actual coverage probability estimated through simulation (targeted at 0.95), *SE (Mean)* is the true standard error of the mean from the sampling run, *SBeSE M* is the quadratic mean of all of the sample-based estimates of the standard errors of the mean, *SBeSE S* is the standardized deviation from the quadratic mean, and *n (mean)* is the average sample size required to achieve the target interval. *Figure* refers to the corresponding diagnostic quartet of graphs in the appendix.

We chose sampling with replacement in order to simulate an infinite population size.

When reading the table and interpreting the figures, it is important to note that the distributions and summary statistics of the sample-based standard errors of the means cannot be interpreted in the usual frequentist way, that is, conditional on sample size. Although these are the long-run frequency accumulations of sample-based estimates of standard error, they do not arise from a single distribution as we would recognize it, and it is not clear that the summary statistics are meaningful in the usual way. We recommend using these merely as relative indicators of the performance of the underlying sampling process.

Each figure corresponds to one simulation, and consists of four plots as follows:



Plot 1 (titled *QQ-Plot for Normality*) gives the quantile-quantile plot for the set of sampled means against the Gaussian distribution.

Plot 2 (*Empirical Density*) gives the empirical frequency of the estimates of the means, with 5 lines superimposed. The central full line is the true mean of the population, the central broken line is the mean of the estimates. The interval delimited by full lines is the central 95% interval directly from the sampling distribution. The interval delimited by the broken lines is the quadratic mean of the sample-based intervals. These are for reference only and should not be interpreted as a comparison of coverage.

Plot 3 (also *Empirical Density*) gives the empirical frequency of the sample-based estimates of the standard errors of the means, and the broken line represents the standard deviation from the sampling distribution of the means.

Finally, plot 4 (*Estimates against Sample Sizes*) is exactly that, with a lowess smooth. Any trend in the smooth indicates a possible relationship between the sample size and the estimates of the means.

DISCUSSION

Starting with the simple case of Gaussian-distributed data, with a target interval width of 1 (interval units are in standard deviations) it is apparent from the comparison of plots 1 and 2 that the uncorrected estimate of SE(Mean) is quite unrealistically low, compared to the corrected estimate. The estimates arising from uncorrected sampling seem Gaussian apart from a few outliers, which all seem to correspond to samples of size 2. Finally, the empirical density of the means show that the uncorrected sample-based intervals are inside the true 95% quantiles, confirming that the coverage probability is overstated. There is no evidence of any trend in the relationship between sample size and the estimates of the mean.

One would expect the stipulation of a smaller target interval would improve the overall performance, and reduce the distinction between the corrected and the uncorrected estimates. This was the case in our simulation, although of the three intervals lengths compared the second was very close to the first in effect, presumably an artifact of sampling. The results are summarized in Figures 3, 4, 5 and 6. Again we see that for the corrected interval, the estimates are largely Gaussian, and the intervals are pretty much exact, but the sample-based standard errors of the estimates are quite some way under the standard deviation of the entire set of means.

There is little to choose now between the distributions of the sample-based estimates of the standard error of the mean of the corrected and the uncorrected versions, indicating that the effect of choosing a larger sample size might have washed out at this point. This is the subject of further inquiry. The empirical density of the uncorrected means again shows the poor coverage probability, and the QQ-plot implies that the density itself is not Gaussian. Again, there is no evidence of any trend in the relationship between sample size and the estimates of the mean.

As noted above, we suspected that introducing skewness into the population would alter the effectiveness of the estimator. The most striking aspect of plots 7 and 8 is that in both cases there is now a trend between the sample sizes and the estimates of the mean, and the trend seems more pronounced for the corrected interval. The empirical densities of the sample-based estimates of standard error seem fairly similar. As could be expected due to low sample sizes, the sample distributions of the means are not Gaussian in either case. The empirical densities of the mean estimates indicate that the estimate arising from the corrected interval is more likely to cover the true interval of interest.

The comparisons for long tailed distributions were with populations with the *t*-distribution on 2 and 5 degrees of freedom respectively. We consider the *t*-distribution with 2 degrees of freedom first. As the variance is infinite for this distribution, it's

impossible to present the target interval in units of standard deviation; we left it at 1. The coverage was better than we expected. The results, as summarized in graphs 9 and 10, are pretty much as expected except for the rather odd behavior in the sample sizes (see plot 4 in each figure). There seem to be clusters of points around sample sizes in both graphs. We suspect that this might be partly due to the infinite variance.

For the *t*-distribution with 5 degrees of freedom the results were only slightly more promising (Figures 11, 12). We set a target of 1.265, which is 1 s.d. for this distribution. The coverage probability for the corrected estimate was clearly high above the required value, but that of the uncorrected estimate wasn't too far below. We were encouraged to note that the average sample size for the corrected estimate with t-distributed data and target 1.265 was identical to that for the corrected estimate with Gaussian data and target 1. It seems that the proposed adjustment is adequate for these heavy-tailed distributions.

The final comparison was performed on a bimodal population, created using a mixture distribution which comprised a Gaussian distribution N(5,1) and an exponential distribution exp(1). A frequency distribution diagram is included in Figure 13. We chose this distribution because bimodality can occur in biological populations, for example the classic mixture of a Poisson(λ) with a spike at 0 (see e.g. Welsh at al. 1996). The results for both the corrected and the uncorrected sampling techniques were not even as good as those for the exponential population, although again the corrected results were considerably better than the uncorrected results.

CONCLUSIONS

The coverage of frequentist confidence intervals arising from the sequential sampling design is overstated. In the simple case we investigated the actual coverage probability for the 95% confidence intervals was about 82% - 89%, depending on the underlying distribution. This is due to the negatively biased sample-based estimate of the standard error of the mean.

As the average sample size increased, that is as the target interval decreased in length, the coverage improved. This seems reasonable; we would assume that as the target interval decreases, the expected sample size increases and the difference obtained by removing 5 degrees of freedom also decreases.

As the population from which the sample was drawn became less Gaussian (either skewed or more heavy tailed) the coverage probabilities got worse. This was most extreme in the bimodal and skewed distributions. There is also a suggestion of dependence between the mean and the sample size when sampling from skewed distributions.

We have discussed a fallacy held by some analysts of data collected from certain sequential sampling designs. Simply stated, it is sometimes assumed that standard frequentist confidence intervals can be used for data collected using sequential sampling

designs. Our overall conclusion confirms that this is not if one wishes to maintain a frequentist perspective. A simple and effective correction, due to Edelman (1991), is examined by simulation and found to be effective although occasionally conservative. Finally, although it is intended for use only for Gaussian populations, it performs well under certain alternatives.

ACKNOWLEDGEMENTS

We wish to thank both Professor Alan Ek of the University of Minnesota and Professor Gary Fowler of the University of Michigan for constructive suggestions which led to considerable improvements in this report.

LITERATURE CITED

Anscombe, F. J. 1953. Sequential estimation. *Journal of the Royal Statistical Society - Series B* 15: 1-21.

Berger, J. O. 1985. *Statistical decision theory and Bayesian analysis*. New York, Springer-Verlag.

Cochran, W. G. 1977. *Sampling techniques*, 3rd Edition. New York, John Wiley & Sons.

Edelman, D. 1991. The five-degree-of-freedom rule of thumb for fixed-width confidence intervals for a normal mean. *Biometrics* 47(2): 733-740.

Hall, P. 1983. Sequential estimation saving sampling operations. *Journal of the Royal Statistical Society - Series B* 45: 219-223.

Mason, R. R. 1969. *Sequential sampling of Douglas fir tussock moth populations*. Research Note PNW-102, Pacific North West Forest and Range Experiment Station. 11p.

Mukhopadhyay, N. 1991. Parametric sequential point estimation. *Handbook of sequential analysis*. Ed. B. K. Ghosh and P. K. Sen. New York, Marcel Dekker.

Mukhopadhyay, N. and S. Datta. 1996. On sequential fixed-width confidence intervals or the mean and second-order expansions of the associated coverage probabilities. *Ann. Inst. Statist. Math.* 48(3): 497-507.

Smith, J. H. G. and J. W. Ker. 1958. Sequential sampling in reproduction surveys. *Journal of Forestry* **56**: 107-109.

Statistical Sciences, Inc. 1995. *S-PLUS User's Manual, Version 3.3 for Windows*. Seattle, Statistical Sciences, Inc.

Stein, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* 18: 427-433.

Thompson, S. T. 1990. Adaptive cluster sampling. *Journal of the American Statistical Association* **85**(412): 1050-1058.

Thompson, S. K. and G. A. F. Seber. 1996. *Adaptive sampling*. New York, Wiley.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer. 1996. Modelling the abundance of rare species - statistical models for counts with extra zeros. *Ecological Modelling* **88**(1-3): 297-308.

APPENDIX 1: CODE

```
sequential.test <- function (size = 500, runs = 5000, target = 1, correction = 5)    {
# Since default variance is 1, target is expressed in units of standard deviation
# Select the sample from the population.  Populations varied.
       sequential.population <- rnorm (size)
# Declare the variables to be accreted.
       sequential.interval <- sequential.sbese <- sequential.sizes <-
              sequential.coverage <- sequential.means <- c()
       true.mean <- mean(sequential.population)
# Loop the sample selection
       for (i in 1:runs) {
# Take the first sample, of size 'correction + 2'
              sequential.sample <- sample (sequential.population, (correction + 2))
# Calculate the first sample-based confidence interval
              n <- correction + 2
              sequential.sbci <- (qt(p = 0.975, df = 1) *
                     sqrt (var (sequential.sample) / n))
# Enter while loop, which increments the sample until the sample-based
# confidence interval is sufficiently small
              while (sequential.sbci > target) {
                            sequential.sample <- c(sequential.sample,
                                   sample (sequential.population, 1))
                            n <- length (sequential.sample)
                            sequential.sbci <- (qt(p = 0.975, df =
                            (n - correction - 1)) * sqrt (var(sequential.sample) / n)) }
# Store mean in a vector
              sequential.means[i] <- mean (sequential.sample)
# Store sample-based estimate of standard error in a vector
              sequential.sbese[i] <-
sqrt(var(sequential.sample)/length(sequential.sample))
# Store the interval length
              sequential.interval[i] <- sequential.sbci
# Store the sample size in a vector
              sequential.sizes[i] <- n
# Store whether or not (T/F) the true mean is included in the sample-based interval
              sequential.coverage[i] <-
                     (((sequential.means[i] - sequential.sbci) < true.mean)
                     & ((sequential.means[i] + sequential.sbci) > true.mean))
              outbit <- list(Iteration = i,  Size = n)
              print (outbit)    }
# Data summaries: coverage probability, mean and standard error of the
# sample-based estimates of the standard errors, bootstrap estimate
# of the standard error of the mean, mean and sd of the necessary sample size,
       coverage <- mean (sequential.coverage)
       sbesemean <- mean (sequential.sbese)
       sbestdev <- sqrt(var(sequential.sbese))
       semean <- sqrt(var(sequential.means))
       mean.samples <- mean(sequential.sizes)
       interval <- sqrt(mean(sequential.interval ^ 2))
       stdev.samples <- sqrt(var(sequential.sizes))
# Graphical outputs
       par(mfrow = c(2,2))
       qqnorm(sequential.means, main = "QQ-Plot for Normality")
       plot(density(sequential.means), xlab="Estimate of Mean", ylab="",
              main="Empirical Density", type="l")
       addenda <- c(true.mean, mean(sequential.means), quantile(sequential.means, 0.025),
              quantile(sequential.means, 0.975), true.mean-interval, true.mean+interval)
       abline(v=addenda[1], lty=1); abline(v=addenda[2], lty=2)
       abline(v=addenda[3], lty=1); abline(v=addenda[4], lty=1)
       abline(v=addenda[5], lty=2); abline(v=addenda[6], lty=2)
       plot(density(sequential.sbese), type="l", xlab="Estimate of SE(Mean)",
              main="Empirical Density", ylab="",
              sub="Sample-Based Estimates of S.E. (Mean)")
       abline(v=semean, lty=2)
       plot(x=sequential.sizes, y=sequential.means, xlab="Sample sizes",
              ylab="Estimates of the Mean", main= "Estimates against Sample Sizes",
              sub="Beware any Trend!")
       lines(loess.smooth(sequential.sizes, sequential.means))
       list(Coverage = coverage, SE.mean = semean, SbeSE.mean = sbesemean,
              SbSE.stdev = sbestdev, Samplesizes.mean = mean.samples,
              Samplesizes.stdev = stdev.samples)
}
```

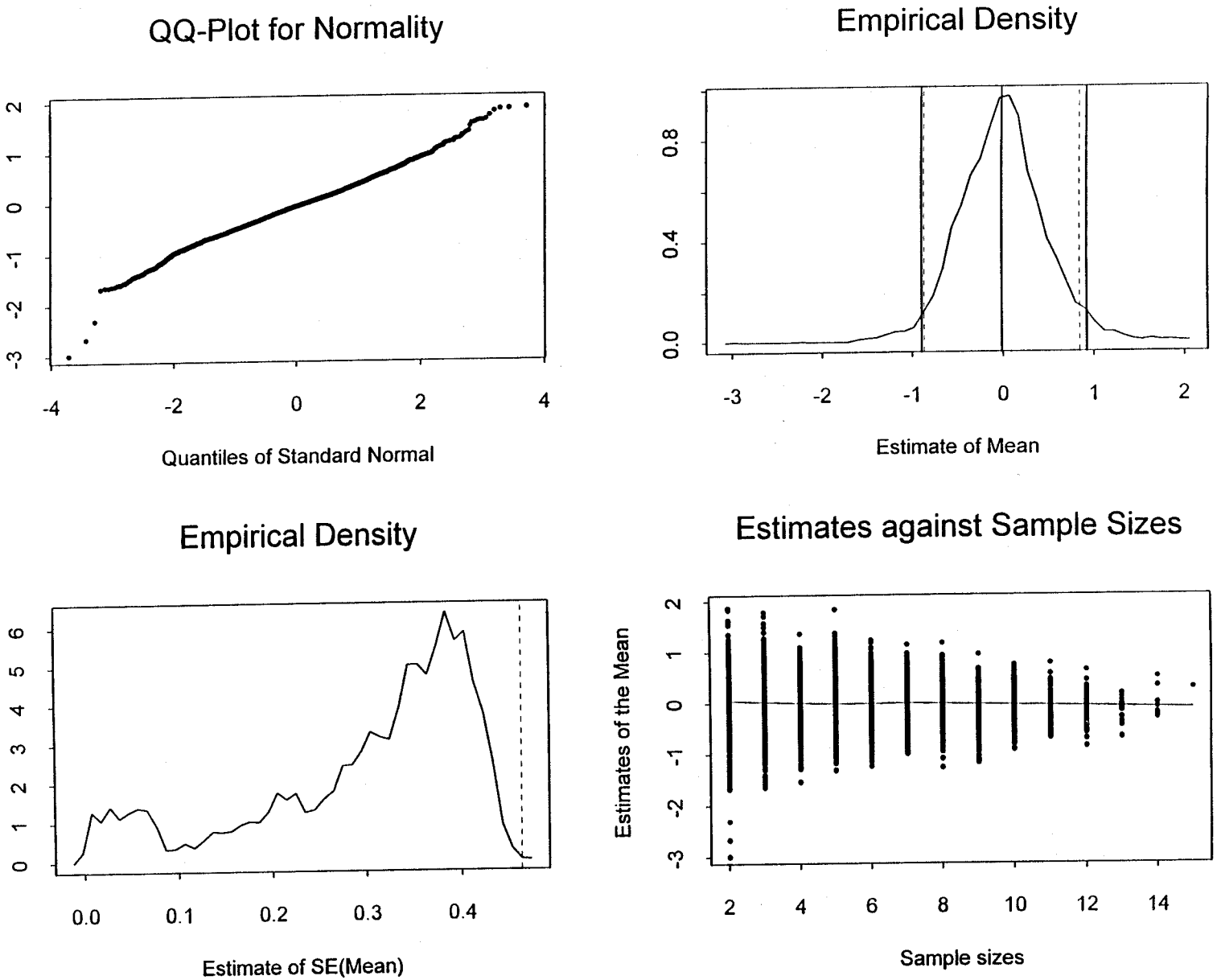


Figure 1. Standard Gaussian distribution, Target = 1, no correction.

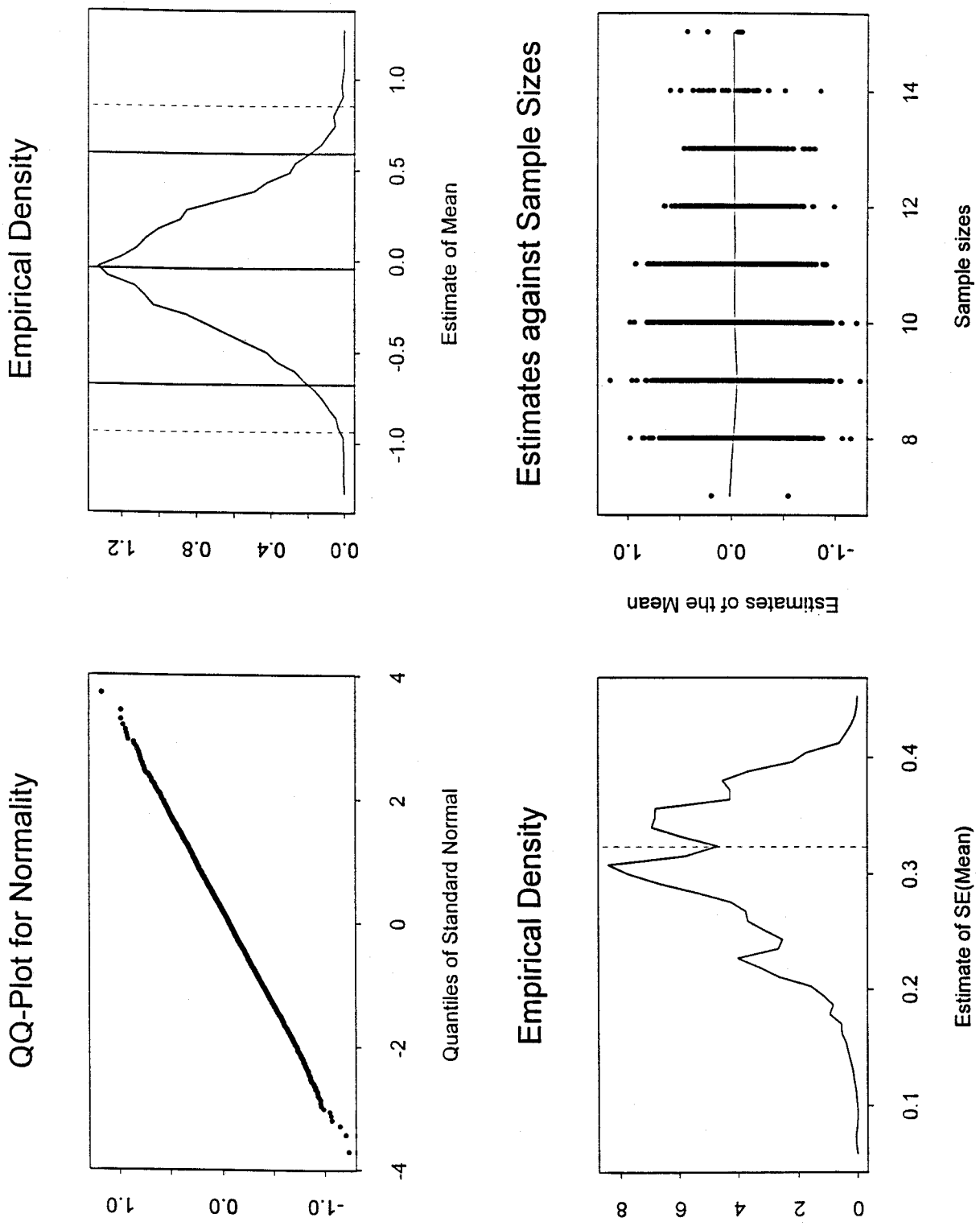


Figure 2. Standard Gaussian distribution, Target = 1, correction = 5.

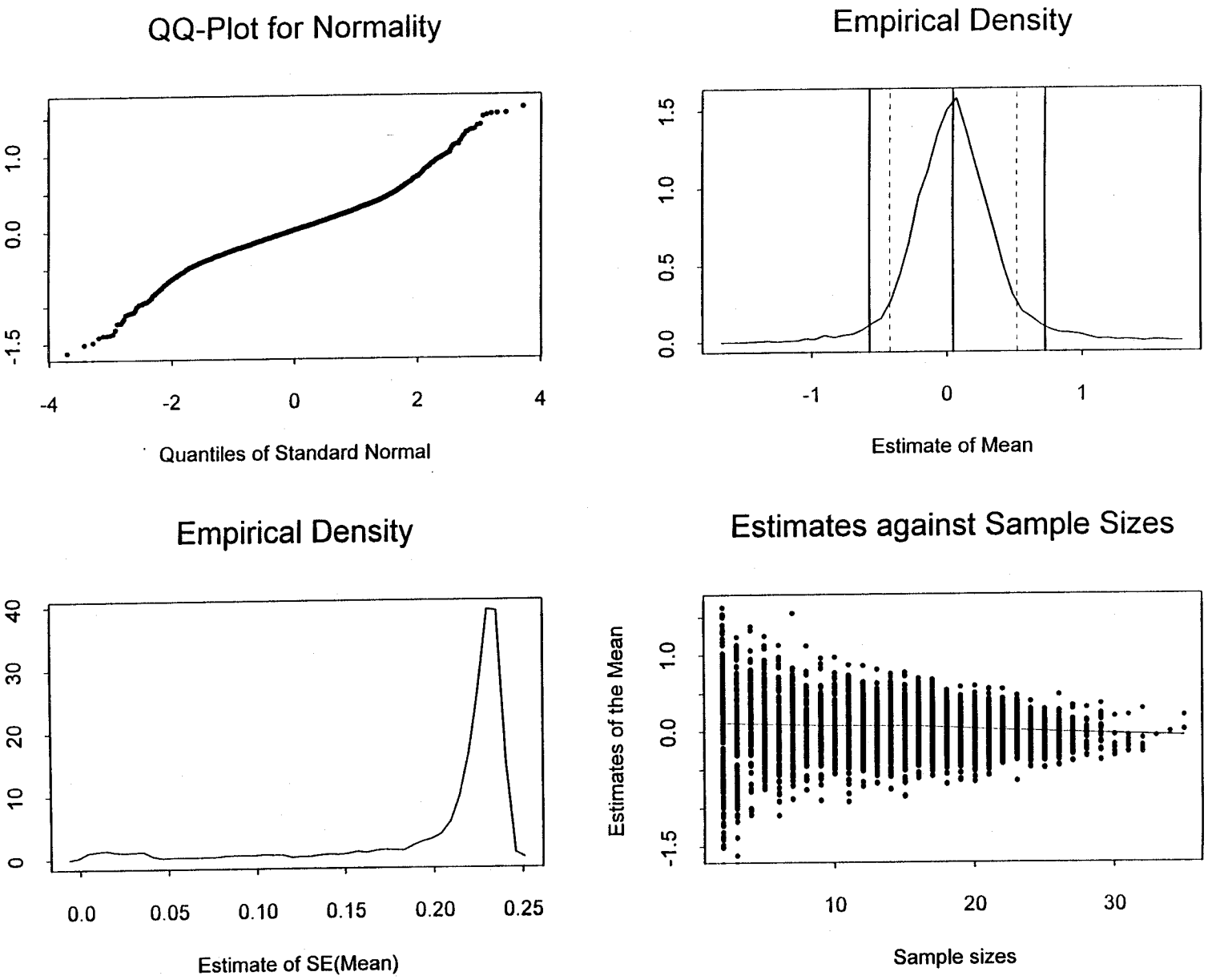


Figure 3. Standard Gaussian distribution. Target = 0.5, no correction.

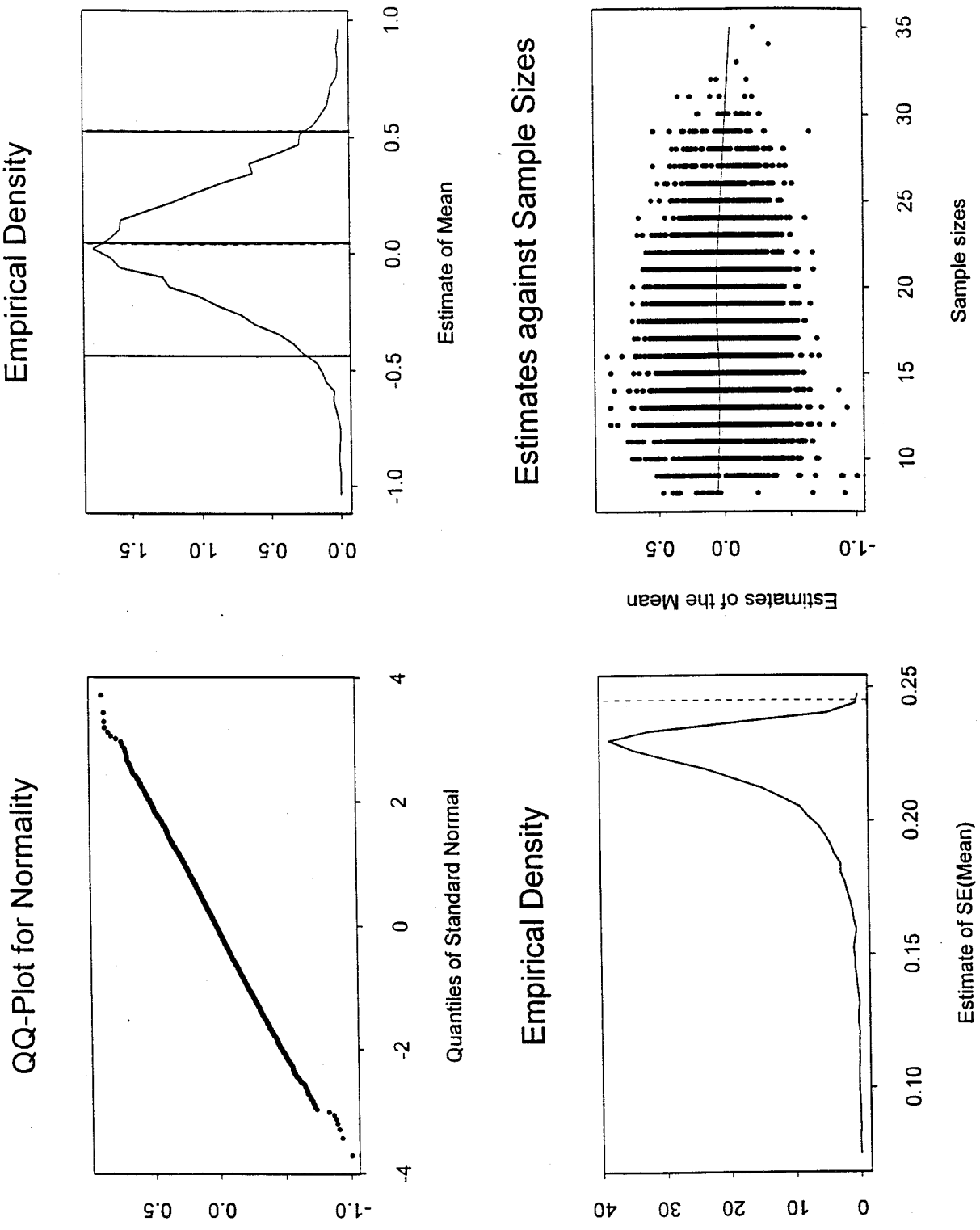


Figure 4. Standard Gaussian distribution, Target = 0.5, correction = 5.

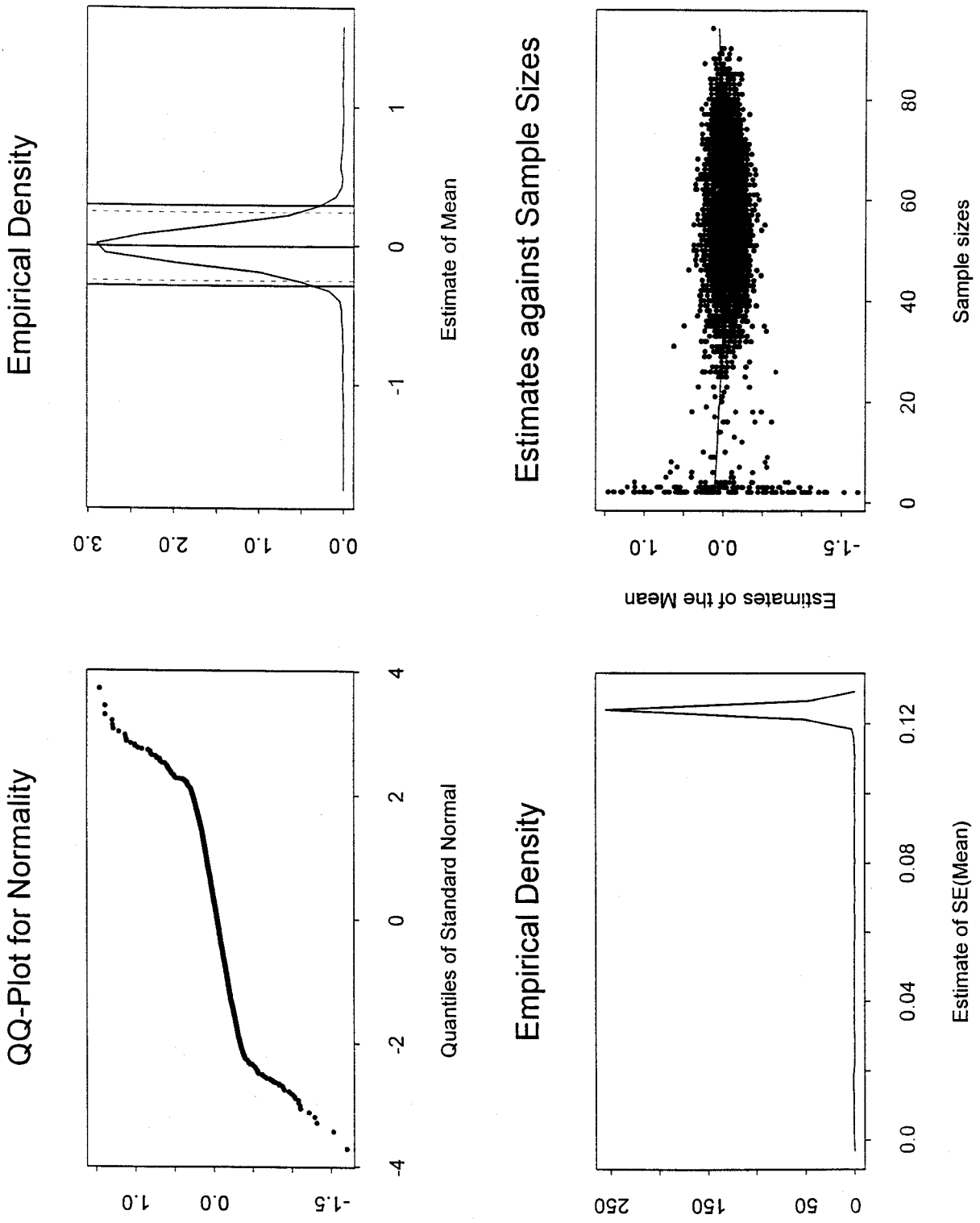


Figure 5. Standard Gaussian distribution, Target = 0.25, no correction.

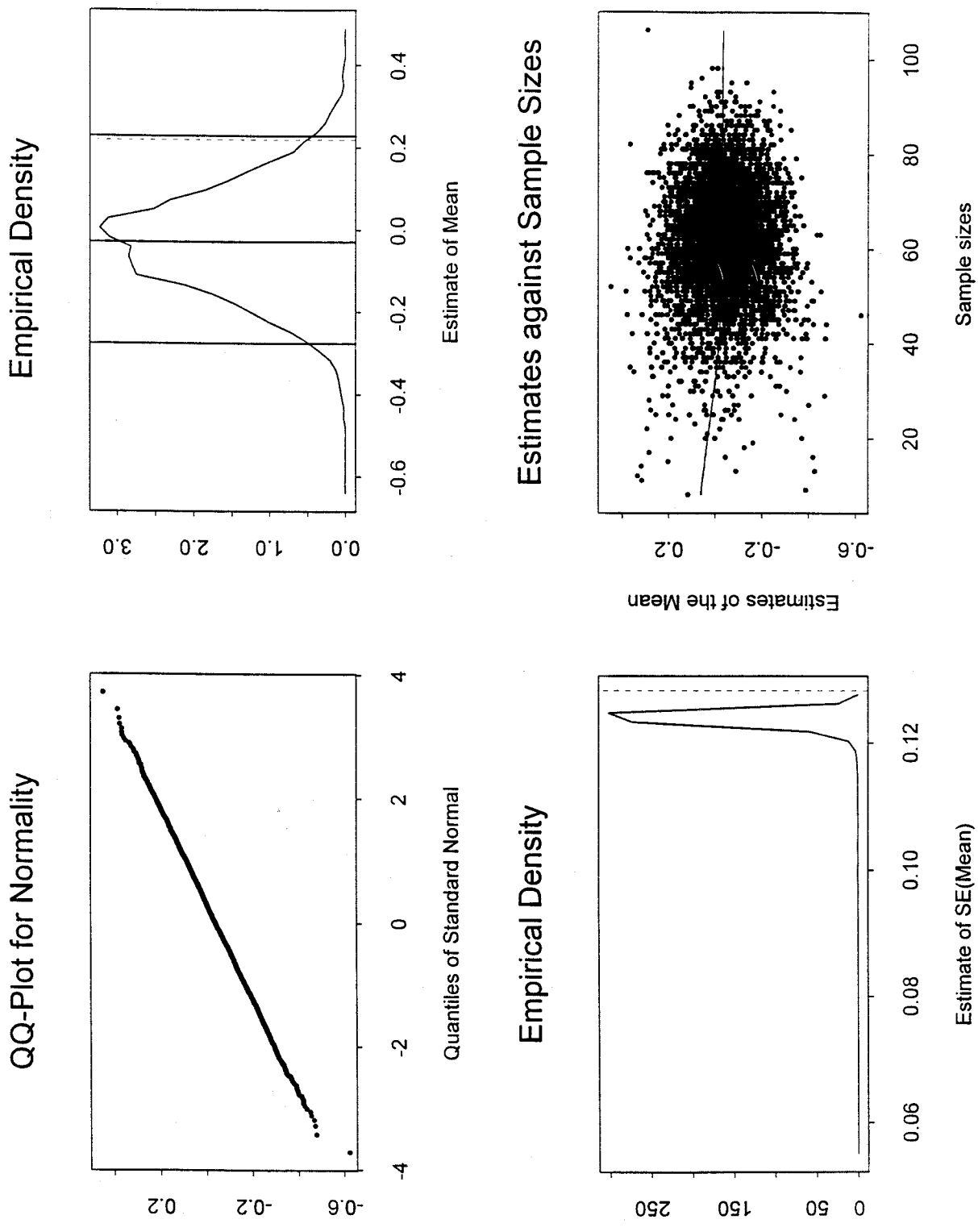


Figure 6. Standard Gaussian distribution, Target = 0.25, correction = 5.

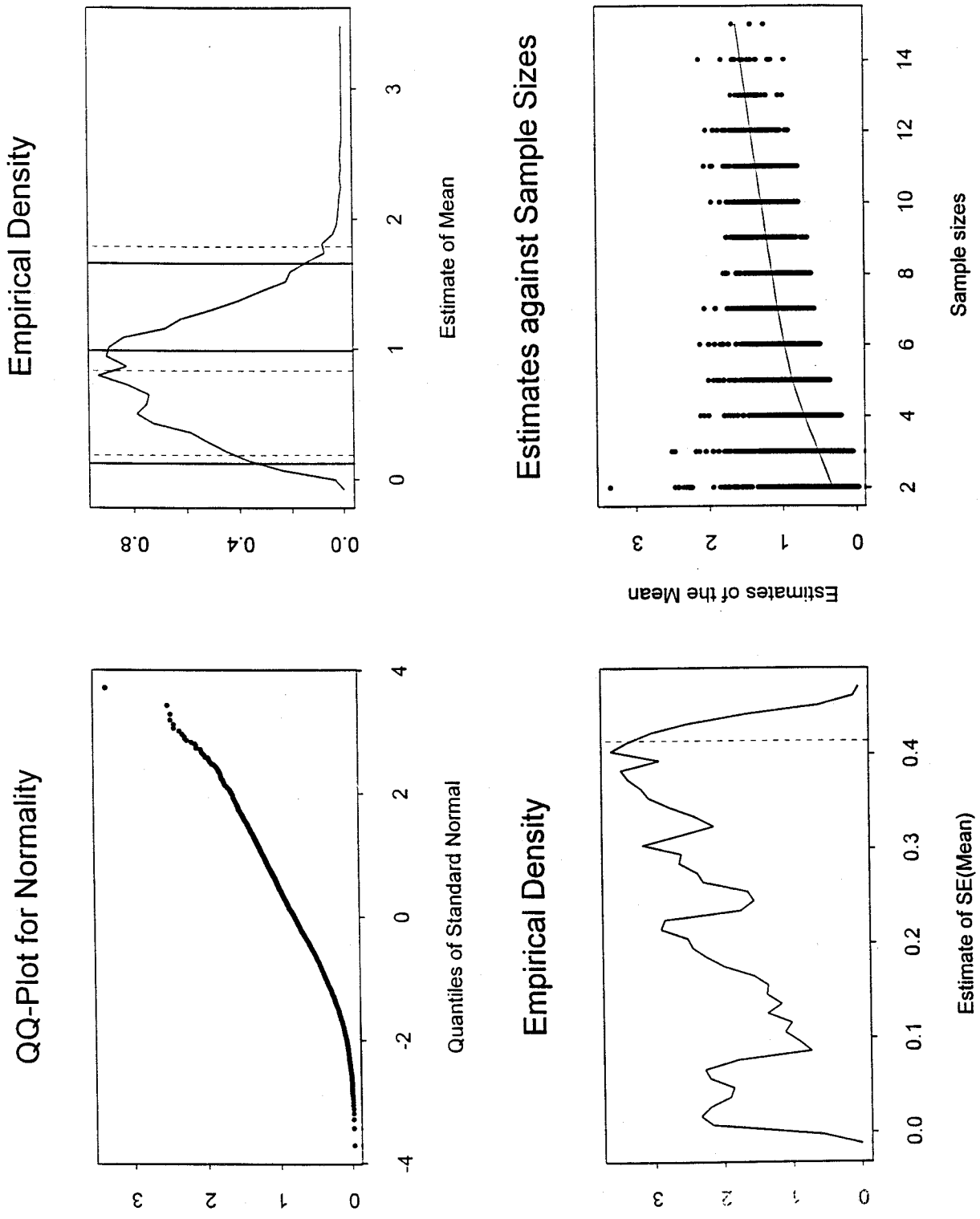


Figure 7. Exponential distribution (1), Target = 1, no correction.

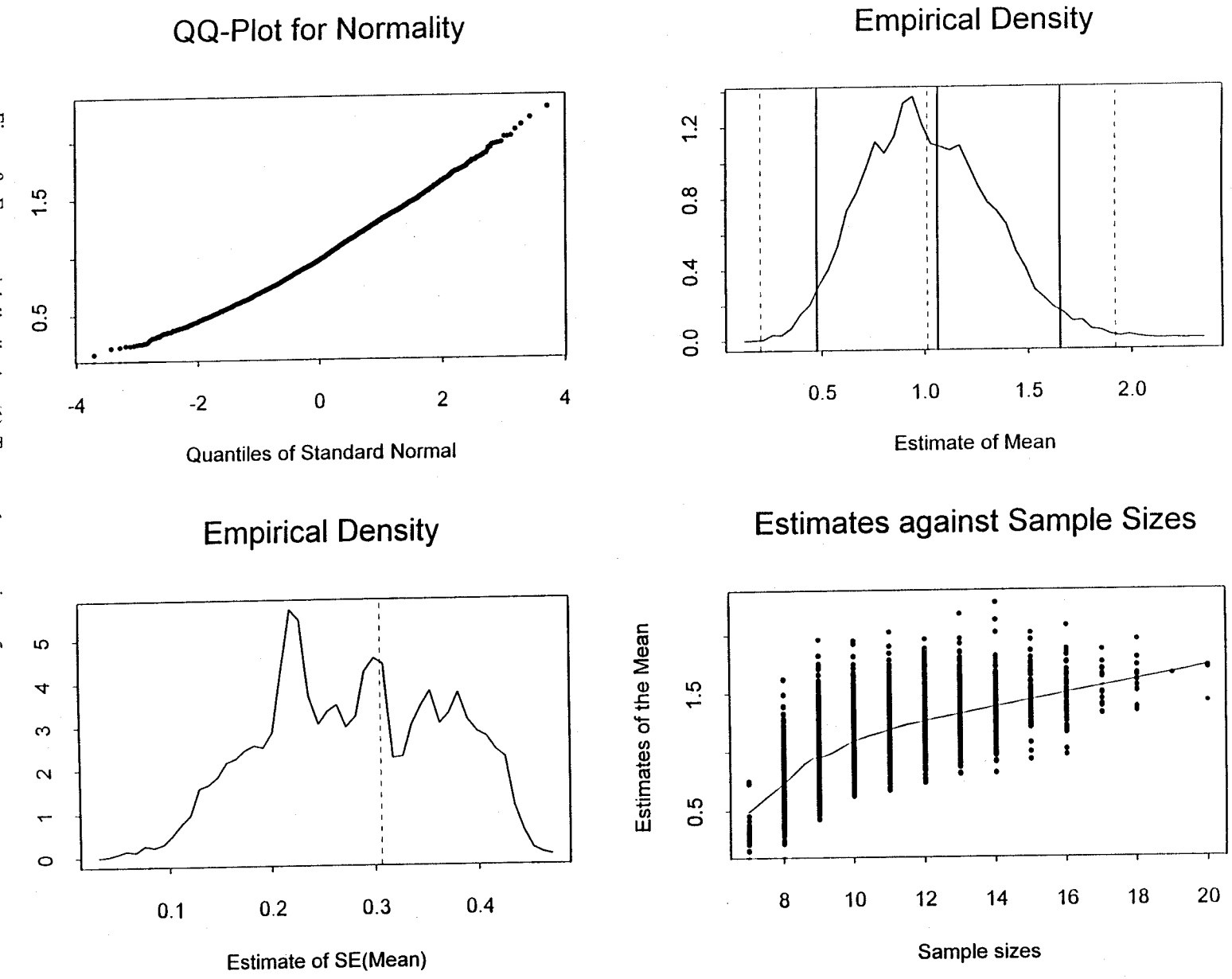


Figure 8. Exponential distribution (1), Target = 1, correction = 5.

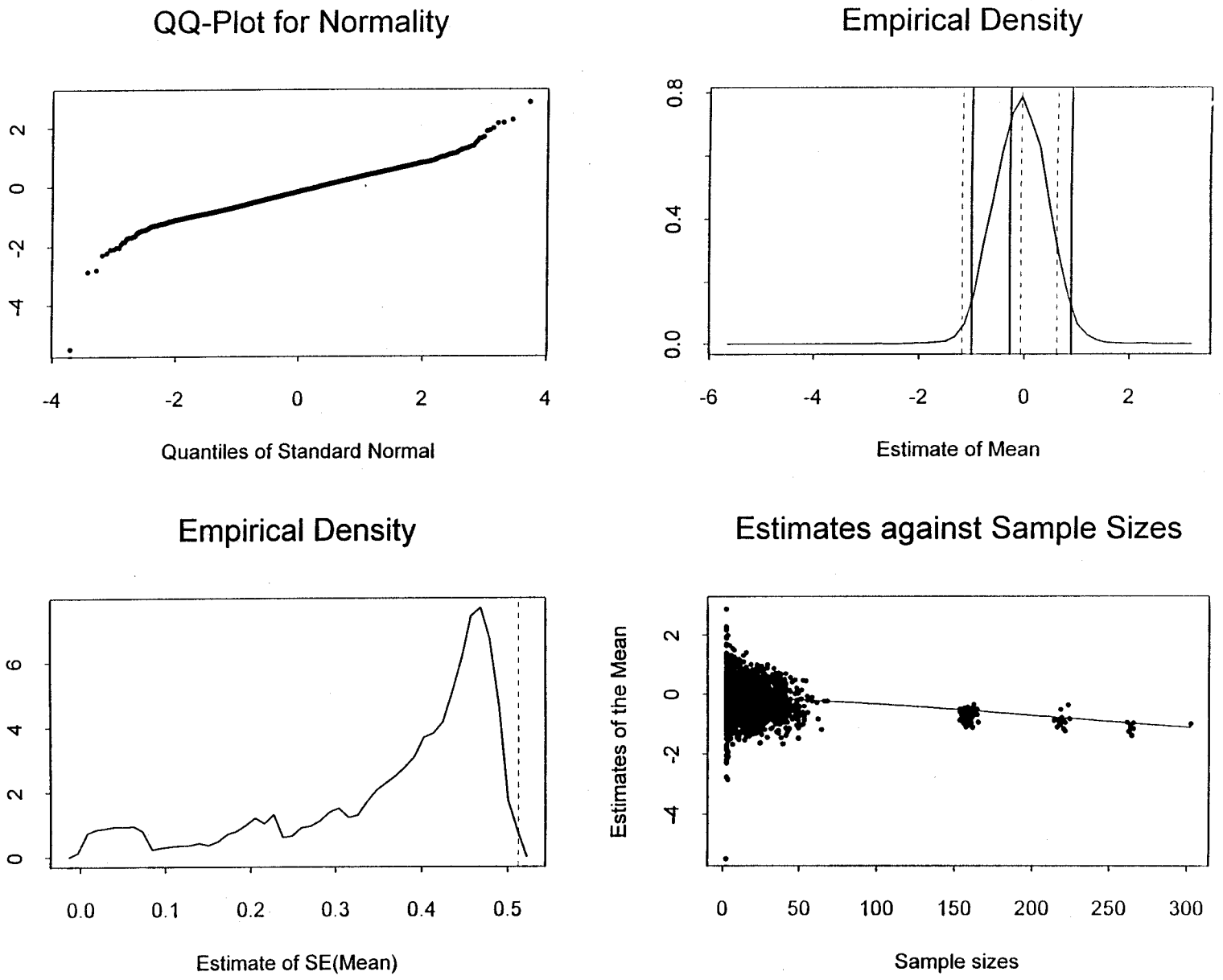


Figure 9. T-distribution (2), Target = 1, no correction.

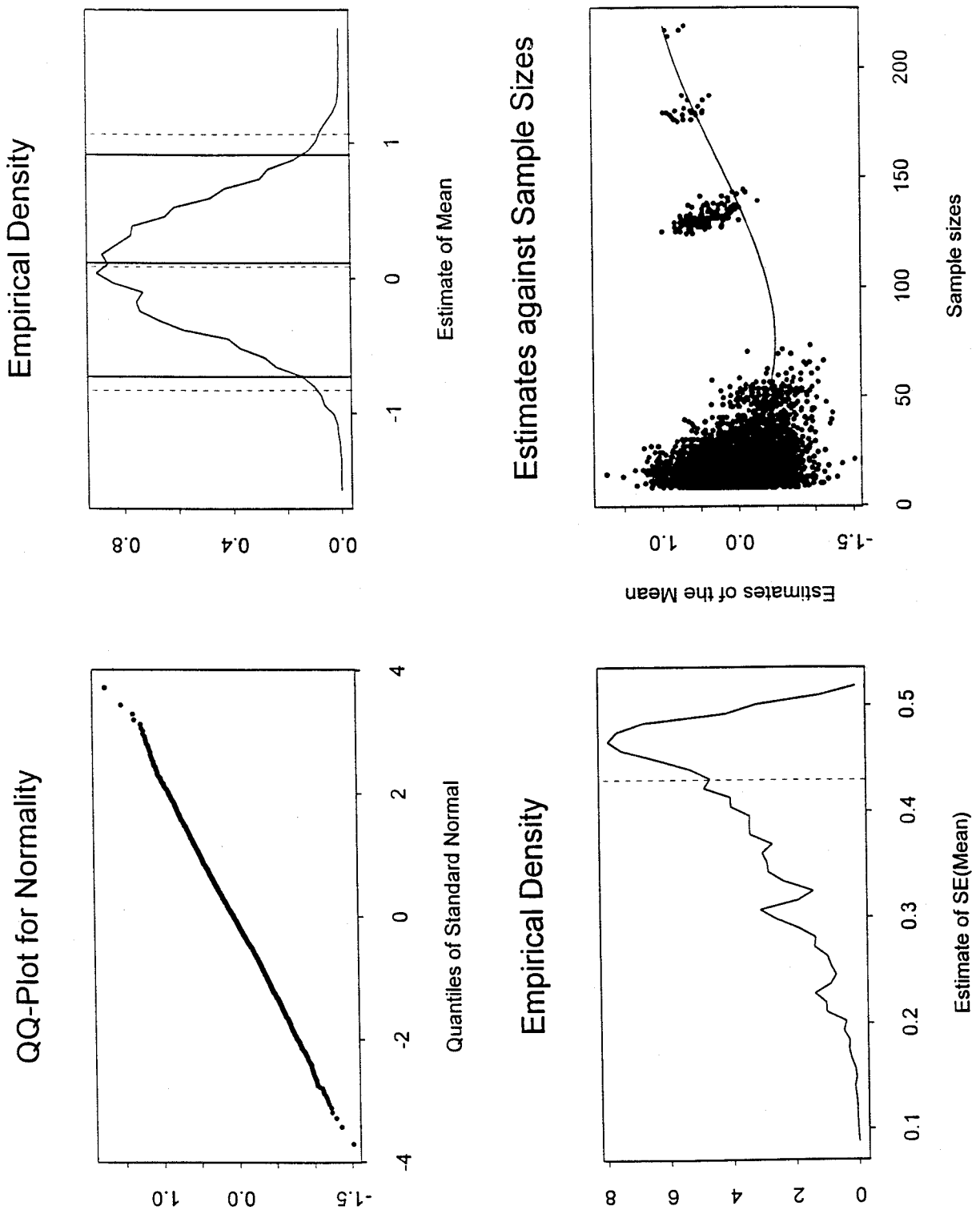


Figure 10. T-distribution (2), Target = 1, correction = 5.

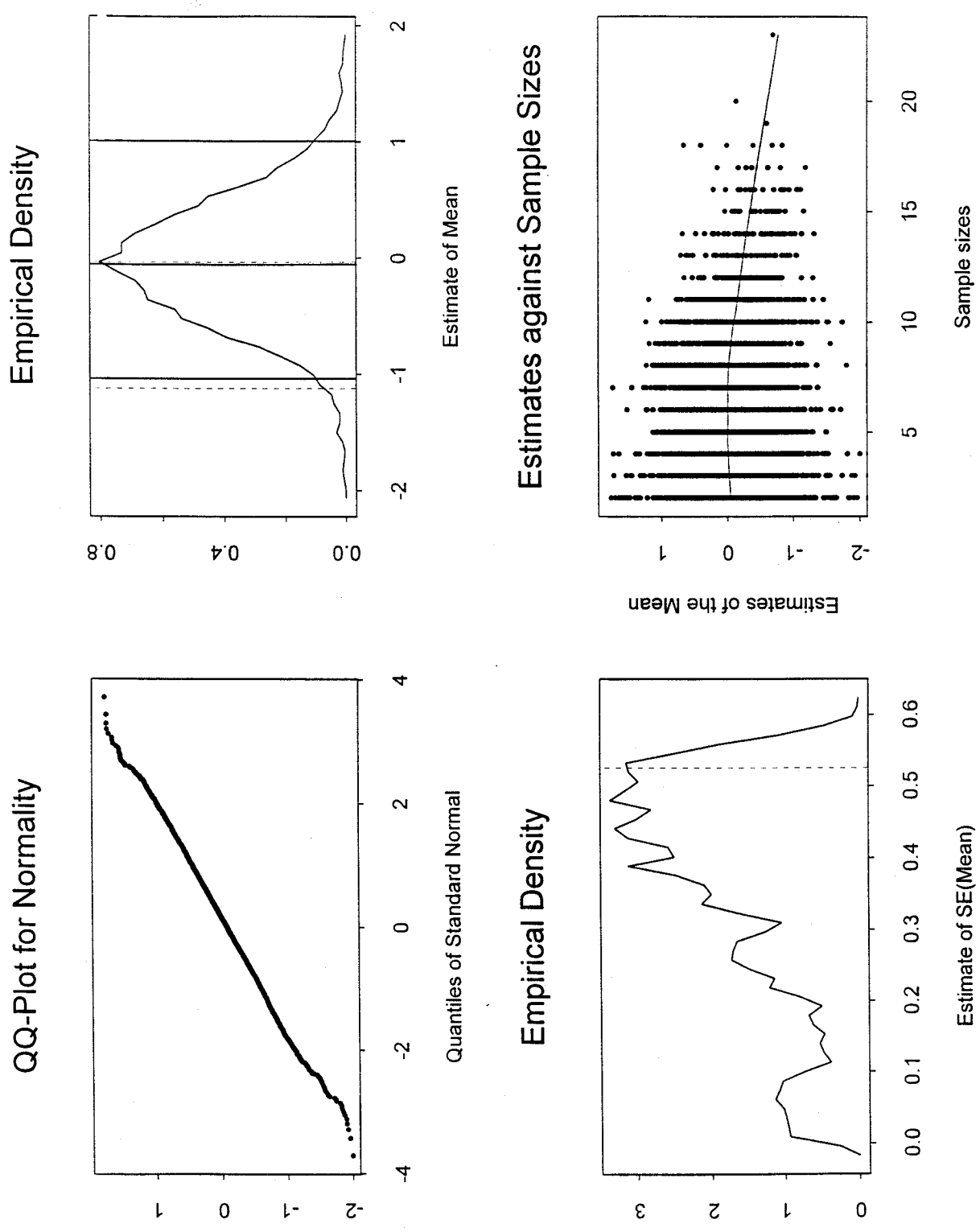


Figure 11. T-distribution (5), Target = 1.265, no correction.

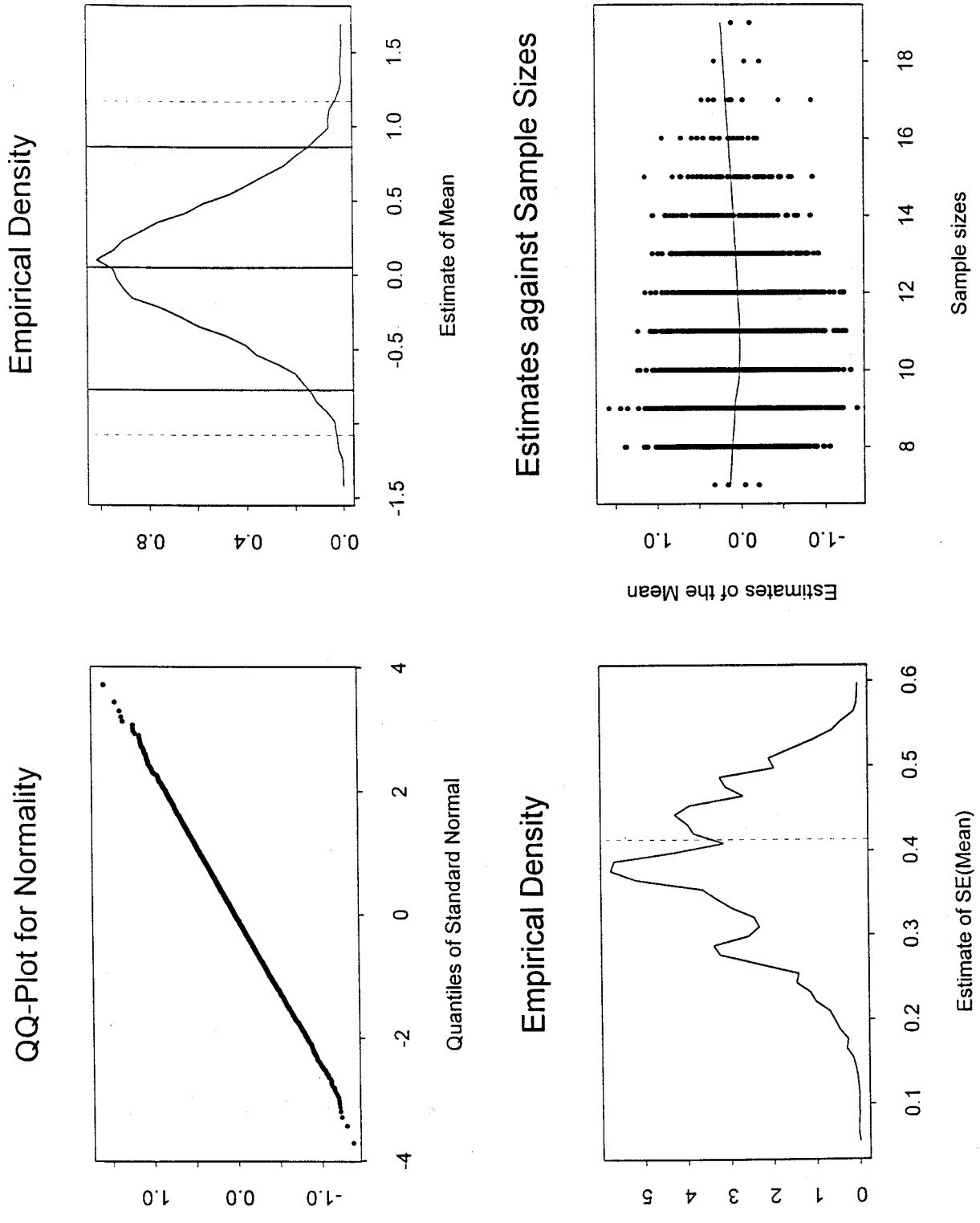


Figure 12. T-distribution (5), Target = 1.265, correction = 5.

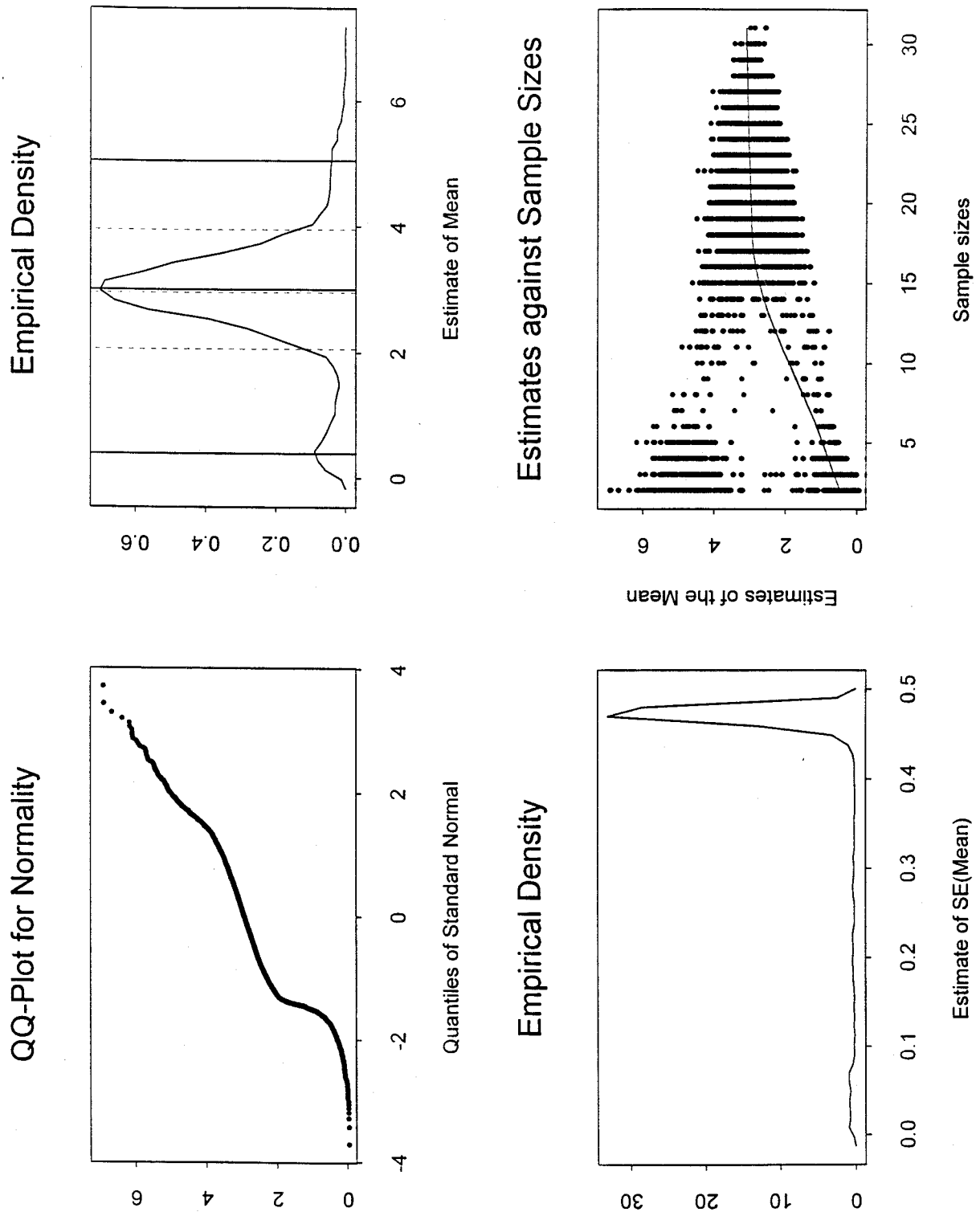


Figure 14. Bimodal distribution, Target = 1, no correction.

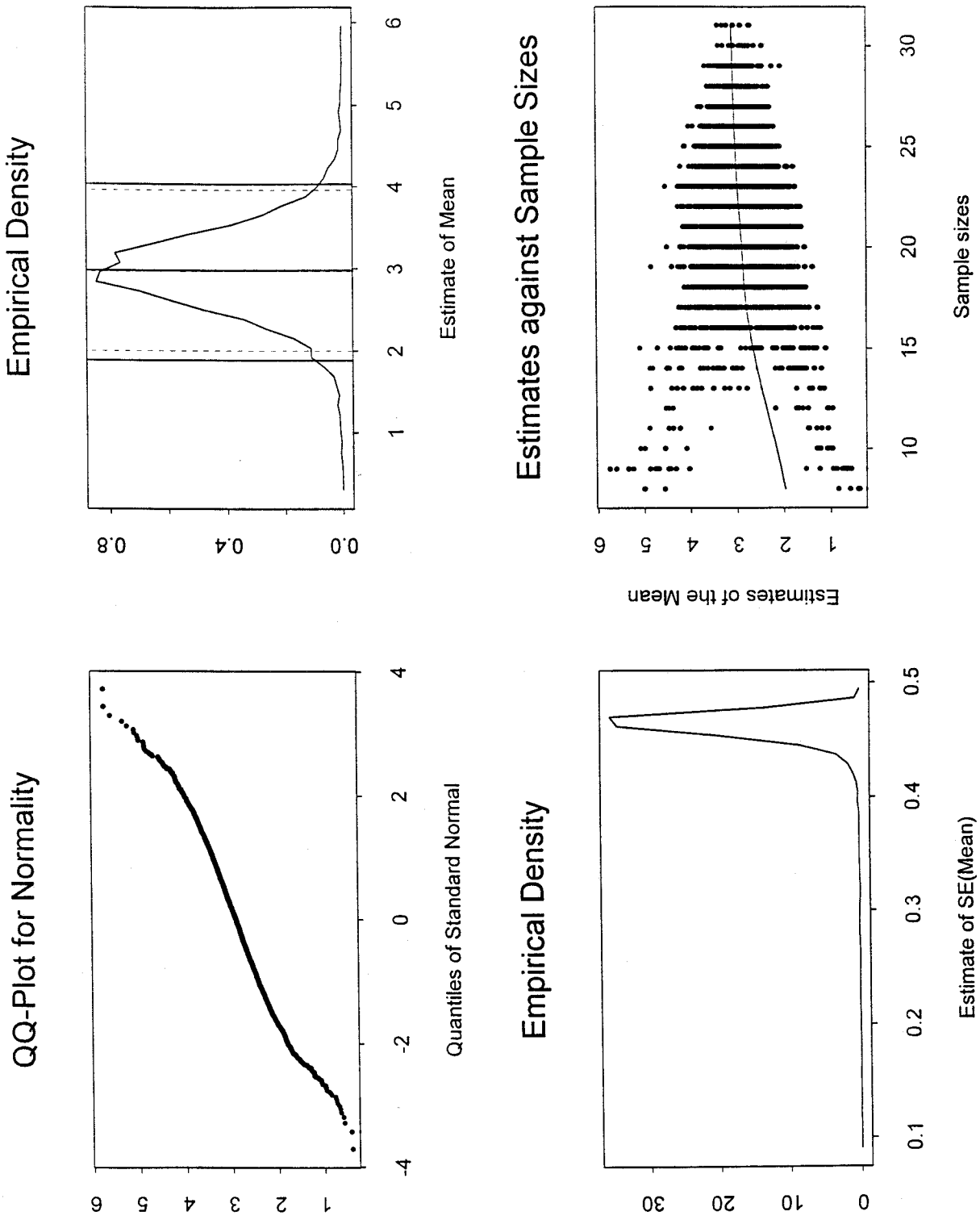


Figure 15. Bimodal distribution, Target = 1, correction = 5.