2.0 Statistical Methods Review

(Chpt's 2 & 3 in Text, Part II, §viii in Bell & Dilworth, and Freese §1-2)

2.1 Variables, Populations, Parameters, Statistics

"Variable" - any characteristic that might vary from one sample unit to another. There would be few sampling problems if there were no variation in characteristics between sample units within populations.

"Variate" - a specific value of a named variable.

"Continuous" - capable of exhibiting every possible value within a given range

"Discrete" - (or discontinuous) values jump from one number or position to the next.

All nominal and ordinal scaled variables are discrete.

Some ratio scaled variables can be discrete, such as counts.

"Population" - the aggregate of all arbitrarily defined non-overlapping sample units. If trees are the sample unit, then all trees on a given area of land could be a population.

"Sample" - the aggregate of sample units chosen from the population from which measurements or observations are taken (a subset of the population).

"Parameters" - a certain constant that describes some aspect of the population as a whole. Typically use Greek letters, such as β, ρ, etc., to denote parameters. Typically, population mean is denoted μ, population variance is denoted σ^2, population standard deviation is denoted σ, which is the square root of the variance.

"Statistic" - a quantitative characteristic of a sample taken from a population. It is an estimate of a parameter. Typically use lower case letters, such as b, r, etc., to denote statistics. Typically, the sample mean, $\bar{x}$, is the best estimator of μ, sample variance, s^2, is the best estimator of σ^2, and sample standard deviation, s, is the best estimator of σ.

Two classes of populations:

1. Finite
2. Infinite

"Finite" - total number of sample units can be expressed as a finite number.
Examples.
The number of trees on a certain tract of land.
The number of sawmills in a geographic region.

"Infinite" - Not finite or sample units cannot feasibly be counted, also includes any finite population that is sampled with replacement.

Example: All gray squirrels in N. America.

From a statistical standpoint, the distinction between finite and infinite populations is important when a large number of sample units is taken from a finite population. Typically, N, the total number of sample units within the population, denotes finite population size. Sample size is typically denoted by n, the number of sample units upon which variables are measured.

Precision and Accuracy are frequently used interchangeably in non-technical parlance.

"Precision" - refers to the degree of agreement in a series of measurements, or how sample values are clustered around their own mean (average). This is the most common definition used in mensuration.

"Accuracy" - refers to closeness of a measurement to the true value, or success in estimating the true value of a quantity.

"Bias"- systematic distortion (or errors) resulting from flawed measurement, instrumental error, or incorrect sampling technique, computation errors, etc. For example, measuring off 100-foot units with a 99-foot chain. Another example is a caliper that is out of adjustment.

Biased estimates can be quite precise, but they cannot be accurate

In the context of statistical sampling, Bias, B, Accuracy, A, and Precision, P are all related:

$$A^2 = B^2 + P^2$$

Thus, if bias is nonexistent or negligible, Accuracy is equal to Precision.

"Probability" - relative frequency with which an event takes place "in the long run"

Example. Let n designate sample size (or number of "trials"), A is our "event", and x is the number of times event "A" occurs.
Then the probability that event "A" occurs would be calculated as

$P(A) = x / n$,

or the ratio of the number of times "A" occurred to the total number of trials.

All probabilities are positive numbers and they range from zero to one, i.e., $0 < P < 1$. Immediately, we can say that the probability of NOT A (the chance that "A" will not occur) is given by 1 - P(A). The event “NOT A” is also known as the *complement* of event A.

"Independence" - the outcome of one trial does not affect the outcome of the next trial.

For independent events, the probability of them all occurring is found by multiplying each of the independent event probabilities.

2.2 Frequency Distributions and Statistical Computations

The frequency distribution defines the relative frequency of (or, the probability of) occurrence of different values of a particular variable.

Each population has its own distinct frequency distribution. One very common way of displaying a frequency distribution is through the histogram.

If the theoretical form of the distribution is known, it is possible to predict the proportion of individuals that are [likely to occur] within any specified limits.

The most commonly employed distributions for modeling data are the Normal, the binomial, and the Poisson.

Poisson - originally used to characterize the number of "changes" in a given continuos interval. It is useful for modeling queues, number of stocked quadrats, etc.

Binomial - originally used to characterize the number of successes in a fixed number of independent trials with the same probability of success for each trial.

Normal - Many natural phenomenon are well described by the Normal frequency distribution. The Normal distribution is associated with continuous variables.

Other statistical distributions of note are the Standard Normal (or Z), Student's *t,* and the *F* distribution.

Statistical data can be obtained through a sample survey or an experiment such as coin tossing. However, before the data can be used to make inferences about a population or phenomenon of interest or as a basis for making decisions, the data must be summarized. Measures of central tendency (or location), measures of dispersion, i.e., scatter, or variation, and measures of association between variables are a few of the important tools for analyzing and interpreting sample data.

Rounding Numbers serves to minimize personal bias and assures a certain amount of consistency in computations and it is desirable to adopt a systematic technique. The chief rounding rule pertains, of course, when the value falls apparently on the halfway mark.

Example. We want the result reported to the nearest 1/10 (0.1)
$98.66 \rightarrow 98.7$
$98.64 \rightarrow 98.6$
$27.65 \rightarrow 27.6$
$114.15 \rightarrow 114.2$

Generally it is best to round off only the number you report (the final result). Carry all intermediate calculations to at least two decimals beyond that which is desired for the final rounded figure.

Measures of Central Tendency (i.e., Location)

Mode - the most frequently occurring value or class of values in a set of observations (or sample).

Median - middle value of a series of sorted observations (or sample). The middle position is found from (n + 1) / 2. If the number of observations is even, it is the arithmetic average of the two middle observations.

Mean - arithmetic average of a set of observations (or sample).

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where Σ denotes "sum of" or summation over observations indicated by subscripts
y = observed value of variable of interest
n = number of sample units measured, or sample size.

Example. Diameters of 26 randomly chosen trees in a forest stand. What are mean, median mode?

Listing of DBH's (cm)		
50	60	70
50	60	60
20	40	40
70	20	50
60	50	60
70	40	50
50	40	40
20	50	50
40	50	

Frequency Table		Relative Frequency
DBH	No. Trees	P
20	3	0.115
30	0	0.000
40	6	0.231
50	9	0.346
60	5	0.193
70	3	0.115
	26	1.000

Measures of dispersion

variance, s^2 -

A measure of how observations are spread out from the average. It considers the position of each observation relative to the mean of the set, by utilizing the difference, or deviation, of each observation from the mean. Specifically, it is the sum of all the squared deviations from the mean. The unbiased estimate of a sample's variance, s^2, is computed from the following formula

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \ .$$

Note that the units for variance are the square of the original units used to measure the observations.

standard deviation, s -

Also a measure of dispersion, but in the same units as the original observations. Standard deviation, *s*, is the positive square root of the variance, *i.e.*,

$$s = \sqrt{s^2} \ .$$

Coefficient of Variation, CV -

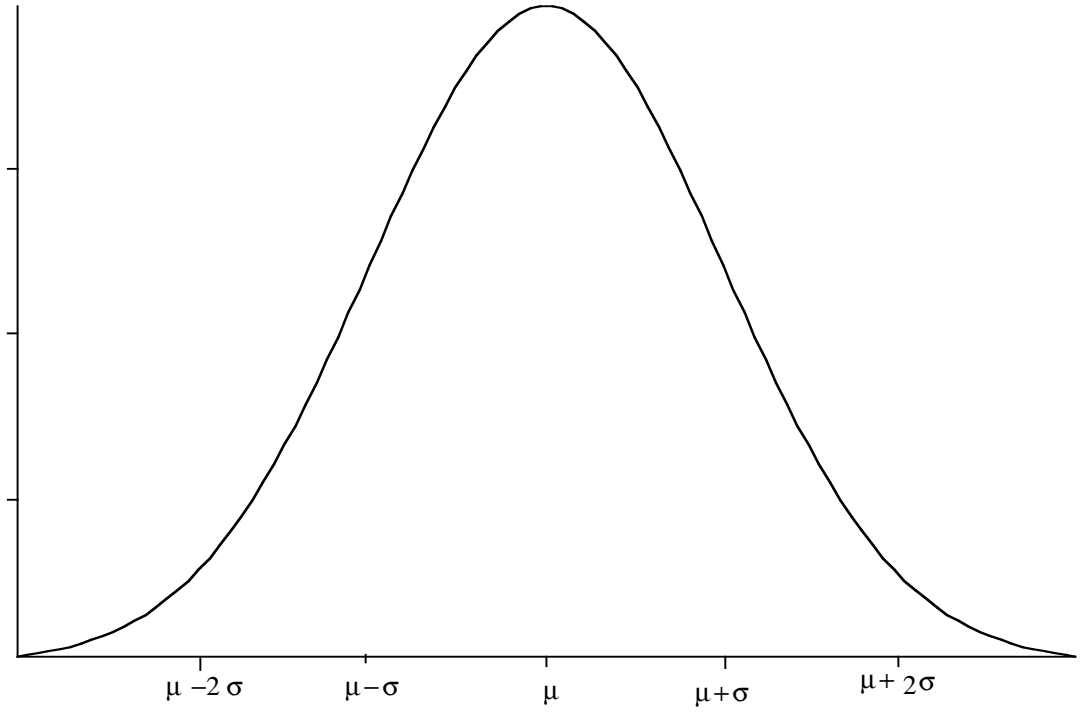
Defined as the ratio of the standard deviation to the mean. It is usually expressed as a percentage. Because populations with large means tend to have larger standard deviations than those with small means, the coefficient of variation (CV) permits comparison of (relative) variability between populations with different sized means. It is computed as

$$CV = \frac{s}{\bar{x}} \cdot (100) \ .$$

Further, the magnitude of the standard deviation (and variance) will be different for the same population when different units of measure are used, but the CV will be the same for a given set of observations regardless of units used.

"Error" & the Normal distribution

The Normal distribution is a symmetrical, Bell-shaped curve described by two parameters: μ, the population mean, and σ^2, the population variance.



Useful probability ranges to remember about data that are distributed as Normal:

68% of all observed values lie within $\pm 1\sigma$ of μ

95% of all observed values lie within $\pm 1.96\sigma$ of μ (about 2 standard deviations)

99% of all observed values lie within $\pm 2.58\sigma$ of μ

If a random variable "Y" is normally distributed, we designate this as

$$Y \sim N(\mu, \sigma^2)$$

This reads "Variable 'Y' is distributed as a Normal Random Variable with mean, μ, and variance, σ^2.

The Normal distribution is handy for three chief reasons.

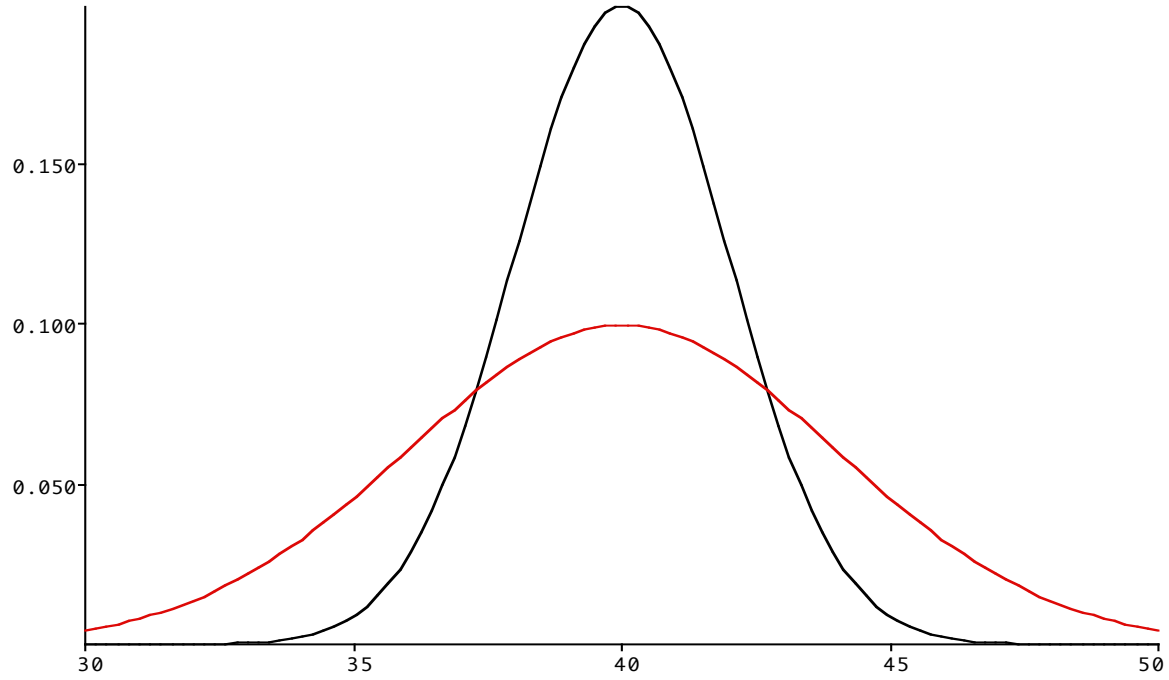
1. Many natural phenomena, or attributes of individuals in a population can be described by it.

2. Arithmetic techniques for handling Normally distributed data are relatively simple (compared to other distributions)

3. Regardless of the underlying theoretically true distribution of individuals in a population, the means of large samples from that population tend to be normally distributed.

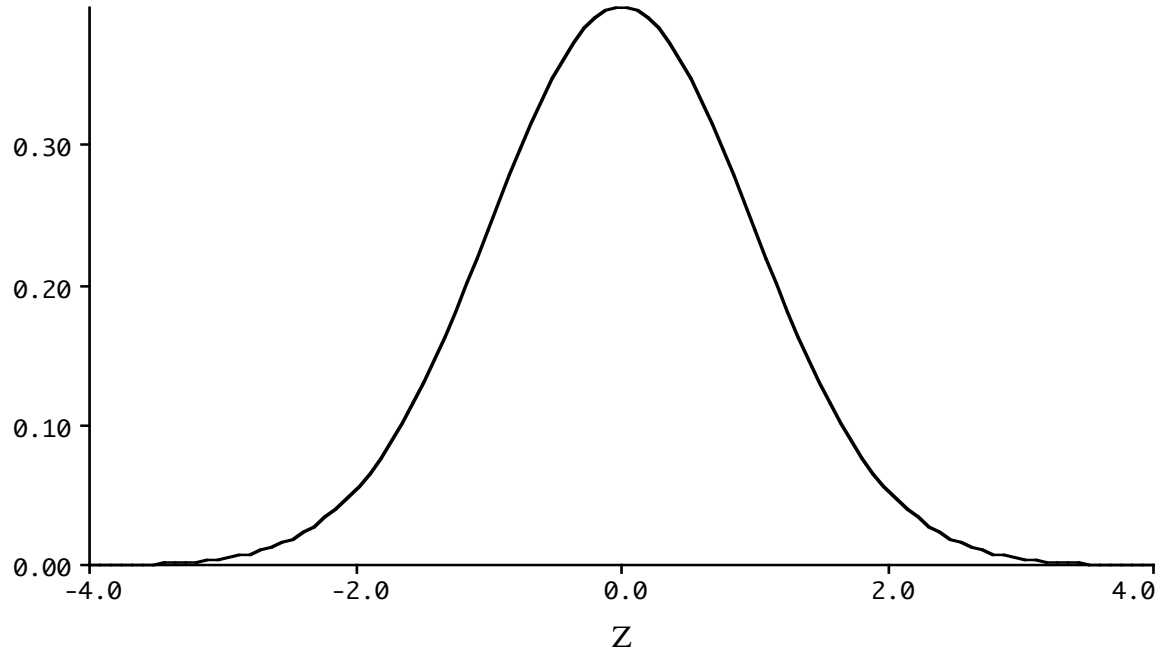
A closer look at each of these three things:

1. Lots of things are normally distributed
   - root collar diameters of seedlings in a nursery
   - weights of large-mouth bass in a lake
   - Example Normal distributions : same means different variances



Examples (above) of two normal distributions with the same mean ($\mu = 40$) and different variances (σ^2)

2. Arithmetic techniques are simple
   - Through a simple transformation, any normal distribution can be made into (mapped onto) a Standard Normal, or Z distribution.
   - If $Y \sim N\left(\mu,\sigma^2\right)$, then $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$ –> Published tables are readily available



The Standard Normal Curve

3. The mean of sample is itself a random variable and its dispersion is characterized with the standard error ($\sigma_{\bar{X}} = \sigma/\sqrt{n}$). The distribution of sample means from sample to sample tends toward normality as the sample becomes increasingly larger (the so-called Central Limit Theorem).

2.3 Additional Sampling Concepts

Central Limit Theorem. If $\bar{Y}$ is the mean of a random sample $Y_1, Y_2, ..., Y_n$ of size *n* from any distribution with a finite mean μ and a finite positive variance σ, then the statistic *W*, defined as

$$W = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$$

is Normally distributed with mean 0 (zero) and variance 1 (one) "in the limit," i.e., as $n \to \infty$.

The chief practical use of the central limit theorem is in approximating the distribution of W with the normal distribution when n is "sufficiently" large. Thus, estimates and inferences may be based on the assumption that given a large enough sample, the mean is distributed approximately as a Normal random variable.

Student's *t* distribution

When we don't know population variance (σ^2), which typically happens when we are sampling from normally distributed populations using small sample sizes, or when we are sampling from a population with an unknown underlying distribution, we cannot use the Normal distribution itself for making probability statements. Student's *t* distribution is the theoretically correct distribution for sample means of a given size when sampling from a population that is normally distributed, regardless of sample size when we don't know population variance. Also, we can use Student's *t* distribution when we have small sample sizes and in the case of an unknown underlying distribution by relying on the Central Limit Theorem, if our sample is "large" enough. The means of large samples tend to be normally distributed. The mean of any sample can be "mapped" to the *t* distribution using the following formula:

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

where $s_{\overline{X}}$ is the standard error of the sample mean. (Note the use of the *sample* standard error in the denominator.)

standard error -

Given that there is variability between individuals in a population, there will also be variation between sample means calculated from several different similarly sized samples drawn from the same population. The *standard error* quantifies this variation among sample means. It may be regarded as a standard deviation among sample means of fixed size, *n*. Standard error, $s_{\bar{x}}$, for infinite populations is given by

$s_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$, and for finite populations it is given $s_{\bar{x}} = \frac{s}{\sqrt{n}} \cdot \sqrt{\left(\frac{N-n}{N}\right)}$.

The standard error of the mean can be used to compute confidence limits for a population mean.

degrees of freedom –

govern the height and width of the t distribution. Generally, the degrees of freedom are equal to the sample size minus the number of parameters estimated from the population. In simple random sampling, we are usually trying to estimate the mean, μ, of the population, so degrees of freedom, df = n - 1.

confidence limits –

Confidence intervals for the t distribution can be derived as follows.

Before sampling $P\left(-t_{\alpha/2,\nu} \le \frac{\overline{X} - \mu}{s_{\overline{X}}} \le t_{\alpha/2,\nu}\right) = 1 - \alpha$

$$P\left(-\overline{X} - t_{\alpha/2,\nu} \cdot s_{\overline{X}} \le -\mu \le -\overline{X} + t_{\alpha/2,\nu} \cdot s_{\overline{X}}\right) = 1 - \alpha$$

$$P\left(\overline{X} - t_{\alpha/2,\nu} \cdot s_{\overline{X}} \le \mu \le \overline{X} + t_{\alpha/2,\nu} \cdot s_{\overline{X}}\right) = 1 - \alpha .$$

Then, after sampling, $(1-\alpha)100\%CI\left(\bar{x} - t_{\alpha/2,\nu} \cdot s_{\bar{x}} \le \mu \le \bar{x} + t_{\alpha/2,\nu} \cdot s_{\bar{x}}\right)$

Before sampling, we speak in terms of probability, but after, in terms of confidence.

2.4 Expanding Means and Standard Errors (Estimating Totals)

Given a sample of measurements from a population, we may need to expand our sample estimate to make it refer to the whole population.

Let's say we want to know the total volume in a tract of land known to contain 200 trees. We've randomly selected and measured 20 trees from the 200 tree population to determine mean volume per tree.

If mean volume per tree was found to be 33 ft^3 and the standard error of the mean was 3 ft^3 then the total volume in the entire treed tract would be 200 times that. The standard error would need similar expansion!

Thus, the estimate for the entire tract would be 6,600 ft^3 with a standard error of 600 ft^3.

Note: the variance of the sample mean is the square of the standard error of the mean, or 9 $(ft^3)^2$, thus, the variance of the total volume is the square of the standard error of the total, or 360,000 $(ft^3)^2$.

2.5 Measures of association between two variables

covariance -

The covariance is a measure of how two variables vary in relation to each other, that is, how they co-vary. If there is little or no association between the two variables, the covariance will be close to zero. In cases where large values of one variable are associated with small values of another, covariance will be negative. It will be positive if large values of one are associated with large values of another. Sample covariance, s_{xy}, is given by

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} \quad .$$

Covariance suffers from the fact that its magnitude depends on the units of measure used for x and y. We often prefer, therefore, to talk about *correlation*, which is just a rescaled version of covariance.

correlation -

The simple correlation coefficient, r, scales the covariance to be in the interval [-1, 1].

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

The closer r is to 1 in magnitude (i.e., | r |, or absolute value of r), the stronger the association (relationship). If variables are independent, then $r = 0$. However, just because $r = 0$, does not indicate that variables are independent, just that there is apparently little or no linear association between the variables. Indeed, there could still be a strong *curvilinear* relationship.

The distribution of Student's t.

Degrees of Freedom	Two-Tailed Probability of Obtaining a Larger Value								
	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	1.0000	1.3764	1.9626	3.0777	6.3137	12.7062	31.8210	63.6559	636.5776
2	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	6.9645	9.9250	31.5998
3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8408	12.9244
4	0.7407	0.9410	1.1896	1.5332	2.1318	2.7765	3.7469	4.6041	8.6101
5	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321	6.8685
6	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	5.9587
7	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.9979	3.4995	5.4081
8	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554	5.0414
9	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693	4.5868
11	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058	4.4369
12	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123	4.2209
14	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768	4.1403
15	0.6912	0.8662	1.0735	1.3406	1.7531	2.1315	2.6025	2.9467	4.0728
16	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208	4.0149
17	0.6892	0.8633	1.0690	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651
18	0.6884	0.8620	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784	3.9217
19	0.6876	0.8610	1.0655	1.3277	1.7291	2.0930	2.5395	2.8609	3.8833
20	0.6870	0.8600	1.0640	1.3253	1.7247	2.0860	2.5280	2.8453	3.8496
25	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251
30	0.6828	0.8538	1.0547	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460
35	0.6816	0.8520	1.0520	1.3062	1.6896	2.0301	2.4377	2.7238	3.5911
40	0.6807	0.8507	1.0500	1.3031	1.6839	2.0211	2.4233	2.7045	3.5510
45	0.6800	0.8497	1.0485	1.3007	1.6794	2.0141	2.4121	2.6896	3.5203
50	0.6794	0.8489	1.0473	1.2987	1.6759	2.0086	2.4033	2.6778	3.4960
55	0.6790	0.8482	1.0463	1.2971	1.6730	2.0040	2.3961	2.6682	3.4765
60	0.6786	0.8477	1.0455	1.2958	1.6706	2.0003	2.3901	2.6603	3.4602
70	0.6780	0.8468	1.0442	1.2938	1.6669	1.9944	2.3808	2.6479	3.4350
80	0.6776	0.8461	1.0432	1.2922	1.6641	1.9901	2.3739	2.6387	3.4164
90	0.6772	0.8456	1.0424	1.2910	1.6620	1.9867	2.3685	2.6316	3.4019
100	0.6770	0.8452	1.0418	1.2901	1.6602	1.9840	2.3642	2.6259	3.3905
150	0.6761	0.8440	1.0400	1.2872	1.6551	1.9759	2.3515	2.6090	3.3565
200	0.6757	0.8434	1.0391	1.2858	1.6525	1.9719	2.3451	2.6006	3.3398
∞	0.6745	0.8416	1.0364	1.2816	1.6449	1.9600	2.3263	2.5758	3.2905

Source: Table was generated using the Splus Statistical Software Package (Insightful Corp., Seattle, WA).