V. ESTIMATION

The Sample Mean and Standard Deviation:

Let $X_1, X_2, \cdots, X_n$ be a sample from the distribution of a population variable *x*. Recall that such a sample arises by choosing an ordered random sample from the population, with replacement, and observing the value of *x* for each member of the sequence of individuals chosen. Since these observations result from the outcome of a random experiment, they are random variables and that is why we have denoted them with upper-case letters. Also remember that if the population size is extremely large in comparison to the sample size, it matters very little whether sampling is done with or without replacement.

The sample mean, sample variance, and sample standard deviation were defined previously as

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]$$

$$S = \sqrt{S^2}$$

We denote these by upper-case letters because they too are random variables. They are important because they are estimates of the population mean μ, the population variance σ^2, and the population standard deviation σ of the variable *x*. Each one has a distribution of values, which is derived from the distribution of values of *x* in the population. This derived distribution may be very complicated and almost impossible to write down. Nevertheless, it can be proved that the theoretical expected values of the sample mean and variance are equal to the population mean and variance, respectively. In symbols,

$$E(\overline{X}) = \mu,$$

and

$$E(S^2) = \sigma^2.$$

In technical jargon, we say that the sample mean and sample variance are *unbiased estimators* of the population mean and variance. Unfortunately, the sample standard deviation is not an unbiased estimator of the population standard deviation. In general, $E(S) \neq \sigma$. However, for reasonably large values of *n* it is almost unbiased, so this is not

a practical concern. Incidentally, one of the reasons we divide by $n-1$ rather than n in the definition of the sample variance is so that it will be an unbiased estimator of σ^2.

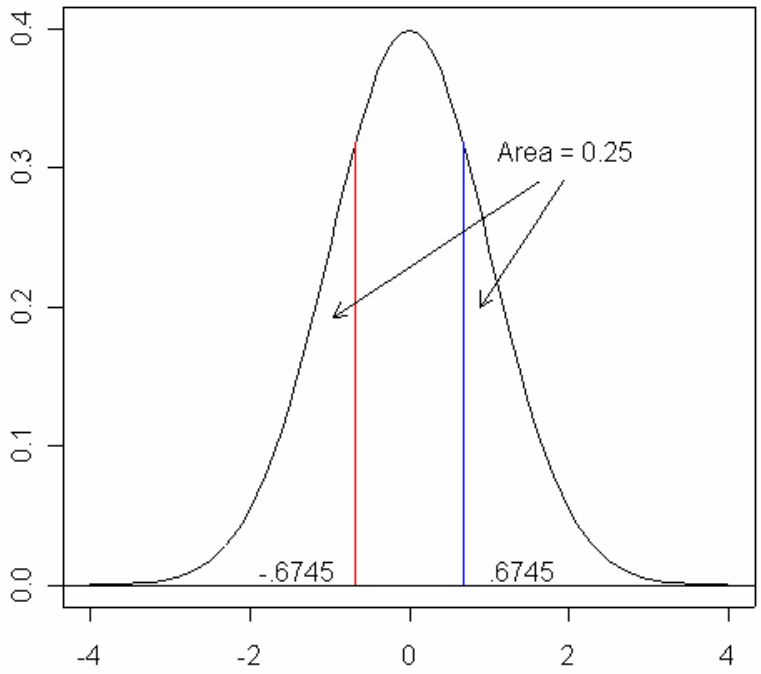
The standard deviation of $\overline{X}$ is related to the population standard deviation σ. In fact, the standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$.

The Standard Normal Distribution:

The normal distributions are theoretical ideal distributions. They are characterized by their *density functions* or density curves. Real population variables may have distributions that are well approximated by a normal distribution, but the normal distribution is only an approximation. It may help to think of the density curve as a probability histogram for an infinite population. Like the probability histograms discussed previously, the total area between the curve and the horizontal axis is 1. Normal distributions are extremely important, not only for their mathematical convenience, but also because of a famous theorem - the central limit theorem, which is discussed below.

There are many normal distributions, but they all bear a simple relationship to one special distribution called the *standard normal distribution.*

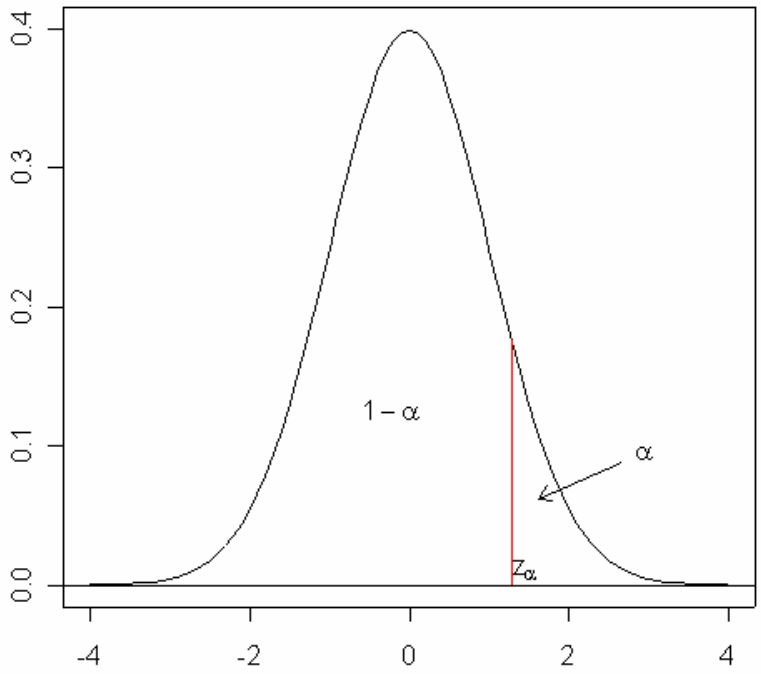
The figure below shows the standard normal density curve. The curve actually extends from $z = -\infty$ to $z = +\infty$ but outside the interval from –4 to 4 its height is so small that it would be indistinguishable from 0 on the graph. The quartiles –0.6745 and 0.6745 are indicated by the red and blue vertical lines. The curve is symmetric about the vertical line $x = 0$, so its median is 0.



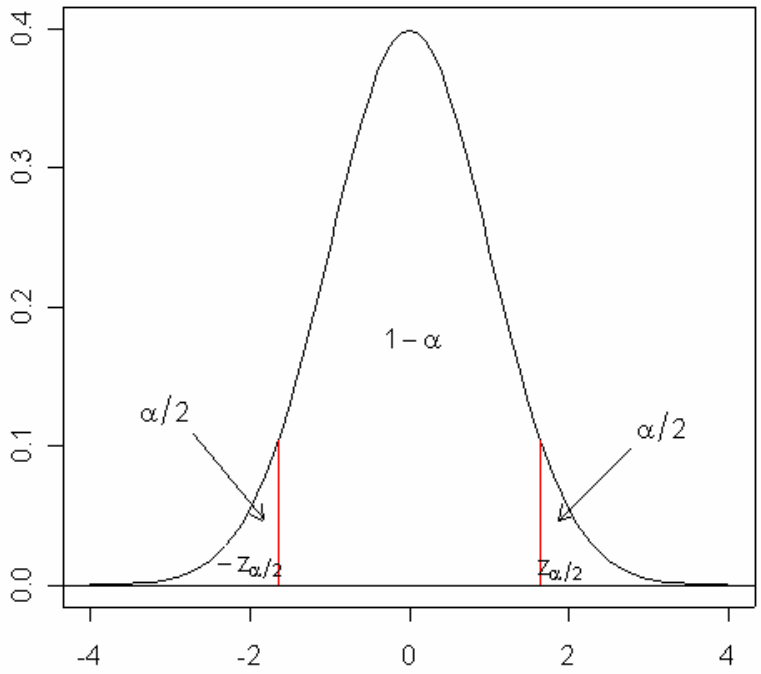
The interquartile range is 2×0.6745, or about 1.35.

The mean and standard deviation of the standard normal distribution are defined using techniques of calculus. Their values are $\mu = 0$ and $\sigma = 1$.

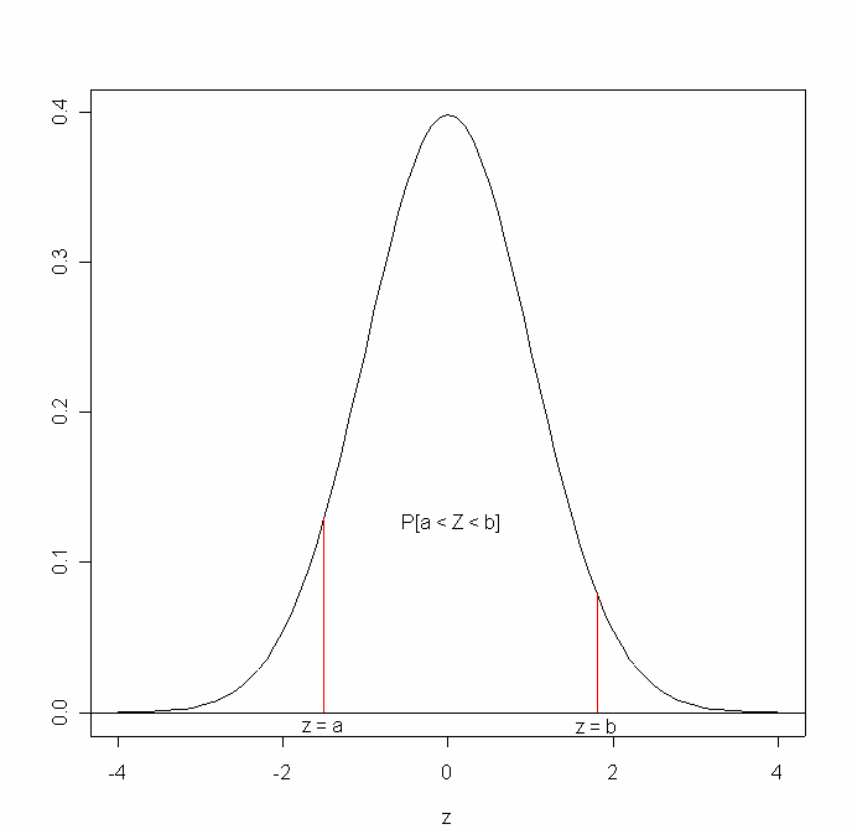
For a number α between 0 and 1, let z_α denote that point on the horizontal axis such that the area under the standard normal curve to the right of z_α is α. The area to the left of z_α is $1-\alpha$. In other words, z_α is the $100(1-\alpha)^{th}$ percentile. The picture below illustrates this.



The area under the density curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is 1-α. This can be seen from the diagram below.



Let Z denote a random variable that has the standard normal distribution. For any two numbers $a < b$, the probability that Z is between a and b is the area under the curve between the points $z = a$ and $z = b$ on the horizontal z axis. This is shown below.



In particular, $P[-z_{\alpha/2} < Z < z_{\alpha/2}] = 1-\alpha$.

Exercise: Using the table of the standard normal distribution or a spreadsheet program, find z_α for several different α and find P [$a < Z < b$] for several choices of a and b.

Other Normal Distributions:

To discuss other normal distributions we first need to know a basic property of means and standard deviations. Suppose a variable x has mean μ_x and standard deviation σ_x. Suppose also that a variable y is related to x by a simple linear equation

$$y = a + bx,$$

where a and b are constants and $b > 0$. Then the mean and standard deviation of y are given by

$$\mu_y = a + b\mu_x$$

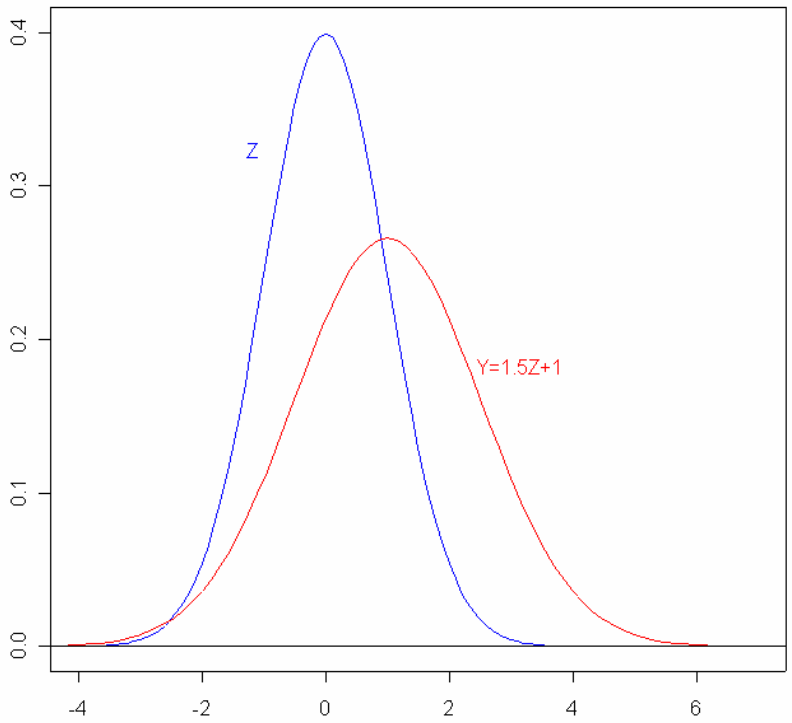
and

$$\sigma_y = b\sigma_x.$$

In fact, we call the mean a measure of location and the standard deviation a measure of scale precisely because they transform in this way. Other measures of location and scale, such as the median and *IQR*, transform in the same way.

Exercise: The variable *x* is the daily low temperature measured in degrees Celsius. Its mean is 20 and its standard deviation is 5. The variable *y* is the daily low temperature in degrees Fahrenheit. What are the mean and standard deviation of *y*? The median temperature in degrees C is 18. What is the median temperature in degrees F? If the *IQR* of Fahrenheit temperatures is 12, what is the *IQR* of Celsius temperatures?

Now let *Z* denote a random variable that has a standard normal distribution, as described above. The mean of *Z* is $\mu_z = 0$ and the standard deviation of *Z* is $\sigma_z = 1$. If μ is any number, if σ is any positive number and if $Y = \sigma Z + \mu$, then *Y* has mean μ and standard deviation σ. The density curve of Y will have the same general "bell" shape as the standard normal density curve, but it will be shifted to a new central location μ and will be spread out or compressed about that location by a factor of σ. The picture below shows what happens for $\mu = 1$ and $\sigma = 1.5$.



Whenever Y is related to Z as described, we say that Y *has the normal distribution* (or is normally distributed) with mean μ and standard deviation σ. Thus the red curve in the preceding figure depicts the normal density curve with mean 1 and standard deviation 1.5.

Let Y be a variable with mean μ and standard deviation σ. To *standardize* Y means to subtract its mean and divide by its standard deviation. This gives a new variable Z, called the *z-score* of Y:

$$Z = \frac{Y - \mu}{\sigma}.$$

According to the transformation equations for means and standard deviations, the mean of Z is 0 and the standard deviation of Z is 1. Furthermore, Y is normally distributed if and only if Z has the standard normal distribution. This allows us to use the table of the standard normal distribution to calculate probabilities for any normal distribution. The rule is

$$P[a < Y < b] = P[\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}]$$

Exercises:

1. Let Y be normally distributed with mean 1 and standard deviation 1.5. Find the following probabilities:

(a) $P[0 < Y < 2.5]$ (b) $P[Y < -1]$ (c) $P[Y > 0]$.

2. Let Y be normally distributed with mean 1 and standard deviation 1.5. Find the 90th percentile of the distribution of Y.

The Central Limit Theorem:

Let $\overline{X}$ be the sample mean from a random sample of size n from a population variable with population mean μ and population standard deviation σ. Previously, we noted that the random variable $\overline{X}$ has mean μ also and standard deviation $\sigma/\sqrt{n}$. The z-score for $\overline{X}$ is

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}.$$

The central limit theorem asserts that for any population distribution whatever, for large sample sizes n the distribution of Z is approximately standard normal. Indeed, the standard normal distribution is the limit of the distribution of Z as n grows without bound. This theorem accounts for the tremendous importance of the normal distributions in probability and statistics.

The central limit theorem does not specify how large n must be for the standard normal distribution to be a good approximation of the distribution of Z. That depends on the underlying distribution of the population variable. However, it is almost always safe to assume that the approximation is good if the sample size is 50 or more. In many cases, fewer than 50 samples is enough.

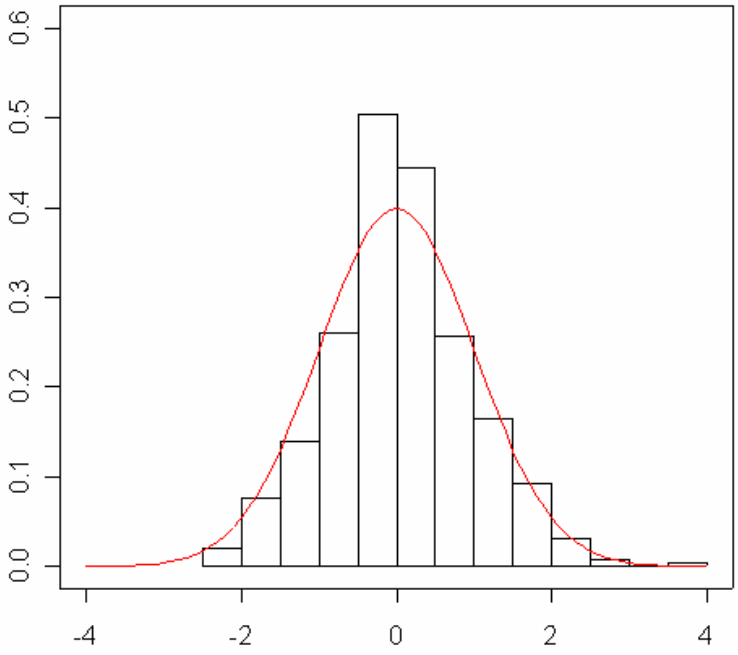
To illustrate the central limit theorem, we will take random samples of size $n = 30$ from the population variable x whose frequencies are tabulated below. For each sample, we will calculate the sample mean $\overline{X}$. We will do this a large number of times and construct a histogram of the z-scores for the different values of $\overline{X}$.

x_j'	0	1	2	3	4	5	6
f_j	.36	.33	.19	.08	.02	.01	.01

To two-place accuracy, the mean of x is $\mu = 1.14$ and the standard deviation is $\sigma = 1.20$. Therefore, the mean of the random variable $\overline{X}$ is 1.14 and its standard deviation is $1.20/\sqrt{30} = 0.22$. The z-scores for the sample averages we generate will be

$$Z = \frac{\overline{X} - 1.14}{0.22}.$$

The histogram of z-scores for 500 repetitions of this experiment is shown below. Superimposed on the histogram is the standard normal curve. The distribution of sample means for samples of size *n* = 30 is fairly close to normal. It is slightly skewed to the right, but not nearly as much as the distribution of *x*. For larger sample sizes the approximation would be better.



Estimating the Population Mean:

If the population mean μ is unknown, the value of $\overline{X}$ provides an estimate of it. However, researchers usually want more than just a single estimated value. They prefer to have an interval of values. The interval should be constructed in such a way that in repetitions of the experiment, a specified probability can be assigned to the event that the interval contains the true population mean. Such an interval is called a *confidence*

interval for the population mean. Assuming that the sample size is large enough for the central limit theorem to apply, we can construct confidence intervals as follows.

First assume that the population standard deviation σ is known. Let $1-\alpha$ be the desired probability that the interval includes the mean μ. Expressed as a percentage, $100(1-\alpha)\%$ is called the *confidence level*. Even though μ is unknown, we can work symbolically with the z-score of $\overline{X}$ and say that

$$1-\alpha = P[-z_{\alpha/2} < Z < z_{\alpha/2}]$$

$$= P[-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}]$$

By rearranging these inequalities,

$$1-\alpha = P[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Thus, a 100(1-α) % confidence interval for μ is the interval with endpoints $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

It is important to interpret confidence intervals correctly. The probability $1-\alpha$ refers to the probability that the method used will produce an interval containing the true population mean. Once the sample has been observed and $\overline{X}$ calculated, the resulting interval either does or does not contain μ, but it is incorrect at that point to assign a probability to either of these statements.

Example 5.2:

Find a 95% confidence level for the mean of the variable *x* in the preceding example.

Solution: We will use 50 samples from the distribution of *x*. The population standard deviation is $\sigma = 1.20$. After sampling, the sample mean turns out to be $\overline{X} = 1.04$. For a 95% confidence interval, $\alpha = .05$ and $z_{\alpha/2} = z_{.025} = 1.96$. The 95% confidence interval has endpoints $1.04 \pm 1.96\frac{1.20}{\sqrt{50}} = 1.04 \pm 0.33$. Therefore, it is the interval from 0.71 to 1.37. Note that it is <u>not</u> correct to say now that $P[0.71 < \mu < 1.37] = .95$. What we can

say is that the method we used to construct the confidence interval leads to success 95% of the time in the long run. Since the true population mean is 1.14 the confidence interval does indeed contain the true mean in this instance.

The procedure just described is slightly unrealistic in that it assumes that the population standard deviation σ is known. That is not likely to be the case. However, σ can be estimated by the sample standard deviation S. It can be shown that σ in the formula for the endpoints of the confidence interval can be replaced by S without affecting the validity of the procedure, provided the sample size is large enough. Therefore, a large sample 100(1-α)% confidence interval for μ is $\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$. The phrase "large sample" does not necessarily mean that the sample has to be very large. Usually, $n \geq 50$ is large enough.

Example 5.3:

We will repeat the example above, this time without using the knowledge that $\sigma = 1.20$. For $n = 50$ samples of x, it happened that $\overline{X} = 1.32$ and $S = 1.39$. Therefore, the confidence interval $\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$ is $1.32 \pm 1.96\frac{1.39}{\sqrt{50}} = 1.32 \pm 0.39$. This gives us the interval from 0.93 to 1.71. Again, the true mean $\mu = 1.14$ lies in the interval.

Sometimes sampling is difficult or expensive and one must make do with a small sample. In this case, there is a procedure, called the student-t procedure, for constructing confidence intervals with small samples. However, it depends on a strong assumption, namely that the underlying population variable x is normally distributed, or nearly so. The population we sampled from in the last two examples was far from normally distributed, so it is questionable whether the student-t procedure would be valid.

Exercise: Open the data file on reading comprehension test scores. Two of the variables are *Pre1* and *Post1*. They are the pre-test and post-test scores on the first reading comprehension measure for individual students before and after instruction. Subtract each pre-test score from the corresponding post-test score and put the result into a new column labeled *Improvement.* Assume that these 66 results are a random sample from a much larger population and find a 95% confidence interval for the population mean improvement in reading comprehension due to the method of instruction.

Estimating a Proportion:

Let x be a population variable that has only two values, 1 and 0. In typical applications these would be numerical codes for two mutually exclusive categories such as "Female" and "Male" or "Approves" and "Disapproves". Let the relative frequency with which

$x = 1$ be denoted by *p*. Thus, *p* is the relative proportion of the population which shares the attribute denoted by *x* = 1, e.g., the proportion of the population which approves the administration's stance on a current political issue. In fact, *p* is the mean of *x* and the standard deviation of *x* is $\sqrt{p(1-p)}$. Sampling from the population and counting the number of samples for which $X_j = 1$ amounts to a binomial experiment with *n* trials and "success" probability *p*. The objective of this section is to develop a confidence interval procedure for *p*.

The sample mean $\overline{X}$ is the proportion of "successes" among the sampled values of *x*. Therefore, it is a natural estimator of *p*. We will denote it by $\hat{p}$ instead of $\overline{X}$ to emphasize its role as an estimator of *p*. The central limit theorem applies to the standardized value of $\hat{p}$

$$Z = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}}.$$

Arguing as before, we obtain as a 100(1-α) % confidence interval for *p* the interval with end-points $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$. Unfortunately, this expression depends on the unknown *p*. It can be shown that *p* may be replaced by $\hat{p}$ in this expression without seriously affecting its validity, provided the sample size is large enough. Therefore, we obtain the interval $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ as a 100(1-α) % confidence interval.

Another approach is to replace $\sqrt{\frac{p(1-p)}{n}}$ by the larger number $\frac{1}{2\sqrt{n}}$. This gives a wider interval $\hat{p} \pm \frac{z_{\alpha/2}}{2\sqrt{n}}$ whose confidence level is actually larger than the nominal level of 100(1-α) %.

Example 5.4:

A public opinion research organization polled 1000 randomly selected state residents. Of these, 413 said they would vote for a 1¢ sales tax increase dedicated to funding higher education. Find a 90% confidence interval for the proportion of all voters who would vote for such a proposal.

Solution: For 90% confidence, $\alpha = .10$ and $z_{\alpha/2} = 1.645$. Thus a confidence interval has endpoints $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.413 \pm 1.645\sqrt{\frac{0.413 \times 0.587}{1000}} = 0.413 \pm 0.026$. The interval extends from 0.387 to 0.439.

The wider interval $\hat{p} \pm \frac{z_{\alpha/2}}{2\sqrt{n}}$ gives the same answer to 3-place accuracy. $\hat{p} \pm \frac{z_{\alpha/2}}{2\sqrt{n}} = 0.413 \pm \frac{1.645}{2\sqrt{1000}} = 0.413 \pm 0.026$. This is because $\hat{p}$ is near 1/2, where the two procedures coincide.

Exercise: Open the data file CPS85Wages. One of the variables is *MARR*, which is a designation of the marital status of the individual. These 534 individuals are a random sample from a much larger population. Find a 95% confidence interval for the proportion of the population that is married.