STT 200 Reading assignment for 8-2-10

On 7-30-10 we examined the method of z-based confidence interval for a population mean μ based on independent samples $x_1, \ldots, x_n$.

Each student independently sampled n = 30 from the population whose probability distribution is

x	0	2	1
p(x)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

For this distribution

$$\mu = E\,X = 0\,(\tfrac{1}{4}) + 1\,(\tfrac{1}{2}) + 2\,(\tfrac{1}{4}) = 1$$

$$E\,X^2 = 0^2\,(\tfrac{1}{4}) + 1^2\,(\tfrac{1}{2}) + 2^2\,(\tfrac{1}{4}) = 1.5$$

$$\sigma^2 = E\,X^2 - (E\,X)^2 = 1.5 - 1^2 = 0.5$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.5} = 0.707.$$

The method consists of forming the sample mean $\bar{x}$ (page 3) and sample standard deviation s (page 4) and then

$\bar{x}$ (sample mean) is an estimate of μ (usually not known)

s is an estimate of σ (usually not known)

$\frac{s}{\sqrt{n}}$ is an estimate of the standard deviation of $\bar{x}$

$P(\mu \text{ in } \bar{x} \pm z\,\frac{s}{\sqrt{n}}) \sim P(Z \text{ in } [-z, z])$ for n large enough

From the line above, for given data either

μ is within $[\,\bar{x} - z\,\frac{s}{\sqrt{n}},\ \bar{x} + z\,\frac{s}{\sqrt{n}}]$

or an event of rarity 2 $P(Z > z)$ has occurred.

MOE. News accounts of statistical studies commonly report the "margin of error" or MOE

$$\text{MOE (margin of error for } \bar{x}) = 1.96 \frac{s}{\sqrt{n}}$$

In that setup either

$$\mu \text{ is within } [\bar{x} - 1.96 \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \frac{s}{\sqrt{n}}]$$

or an event of rarity 0.05 has occurred. The interval

$$[\bar{x} - 1.96 \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \frac{s}{\sqrt{n}}]$$

is entirely known from the data and will vary from one sample of n to the next.

For the classroom exercise 7-30-10 I chose z = 1.0 in which case either

$$\mu \text{ is within } [\bar{x} - 1.00 \frac{s}{\sqrt{n}},\ \bar{x} + 1.00 \frac{s}{\sqrt{n}}]$$

or an event of probability around 0.32 has occurred (failure of the 68% interval to cover μ).

Empirical results of 7-30-10 (posted). In the 8am section 9 of 13 students (69%) had their 68% interval cover μ. In the 12:40 section 8 of 12 students (66.6%) had their 68% interval cover μ.

Terminology. An interval of the form

$$[\bar{x} - z \frac{s}{\sqrt{n}},\ \bar{x} + z \frac{s}{\sqrt{n}}]$$

is called a confidence interval (CI) for unknown population mean μ. You may think of it as a widening of $\bar{x}$ from a point estimate (which is extremely unlikely to hit μ on the nose) into an interval having approximate probability P(Z in [-z, z]) of covering μ, provided n is large.

Other z. Using the z-table we have a 99.8% CI for μ of

$$[\bar{x} - 3.08 \frac{s}{\sqrt{n}},\ \bar{x} + 3.08 \frac{s}{\sqrt{n}}]$$

In exchange for widening the interval (using 3.08 instead of 1.96) we increase the chance that unknown μ is covered from around 95% to around 99.8%.

Precision. The wider interval above is less precise about where μ may be located. However, if $\frac{s}{\sqrt{n}}$ is small the overall interval width 2 3.08 $\frac{s}{\sqrt{n}}$ may be acceptably small.

Modification for samples drawn with equal probability but without replacement. The modification is a simple one. It employs what is known as the finite population correction (FPC)

$$FPC = \sqrt{\frac{N-n}{N-1}}$$

$$[\bar{x} - z \sqrt{\frac{N-n}{N-1}} \frac{s}{\sqrt{n}},\ \bar{x} + z \sqrt{\frac{N-n}{N-1}} \frac{s}{\sqrt{n}}]$$

As with the original CI (sampling with replacement) the version for sampling without replacement satisfies

$$P(\mu \text{ in } [\bar{x} - z \sqrt{\frac{N-n}{N-1}} \frac{s}{\sqrt{n}},\ \bar{x} + z \sqrt{\frac{N-n}{N-1}} \frac{s}{\sqrt{n}}]) \sim P(Z \text{ in } [-z, z])$$

but provided both n and N-n are large, where **N denotes the size of the population**.

Student's t distribution and CI for small n provided the population distribution is itself normal or nearly so. This version has the important advantage that we may be able to apply CI even when the sample size is as small as n = 2. However, it only applies **when the population distribution is normal** or nearly so (this includes many natural processes, as well as much industrial and scientific data and business and manufacturing data for processes having been brought under statistical control.

Form of the t-interval and degrees of freedom (df) (see Key 50). Again, only a simple modification is made to the original CI. One simply replaces the z-score by a counterpart t-score from the t-table (page 211). For example, if n = 3 (from a normal population) the 95% t-based CI for μ replaces z = 1.96 (seen at the bottom of the table) with t = 12.706 (seen at the first row of the table in the column having 1.96 at the bottom). Why row one and not row 2? Because the table is arranged by "degrees of freedom" (df) which for the case being considered is always figured as df = n-1. So taking df = 2-1 = 1 and 95% (right tail area = 0.025) we arrive at t = 12.706. So the 95% CI

for μ based on a sample of 2 from a normal population will be

$$[\bar{x} - 12.706 \frac{s}{\sqrt{n}},\ \bar{x} + 12.706 \frac{s}{\sqrt{n}}]$$

A severe price is paid for using only n = 2 because s is forced to estimate σ based on the smallest of samples for which it is possible to do so. So the interval needs to be very wide by comparison with z = 1.96 applicable to large n. You can see from the t-table that t begins to converge to z as you move down the rows towards large n.