PREP for Project 7

I really want you to dig into this material. We have two weeks to get it down so I’m setting it all out to begin with in Project 7 on the web.

Below refer to i.i.d. sampling from a population. We will use the example of a population of retail outlets. Score x is the selling price of a particular model of shoe.

**PROBLEM:** xBAR is not going to be exactly equal to µ. Can we tell something about how close it is to µ by just looking at the sample data?

**ANSWER:** Yes we can. Not only will a large n sample be nearly certain to place xBAR near µ, it will also tell you something about how close xBAR is to µ!

**SOME NOTATION AND TERMINOLOGY:**

µ is the population mean  
xBAR is an estimator of µ  
E xBAR = µ so xBAR is an unbiased estimator for µ  
xBAR → µ with increasing n so xBAR is consistent for σ  
If we take data and find xBAR = 5.6 then 5.6 is an estimate of µ

If the population s.d. σ is 5.1 then we say the true value of σ is 5.1  
s is an estimator of σ defined on page 24  
E s is not equal to σ so s is not an unbiased estimator of σ  
However, E s → σ with increasing n, so its asymptotically unbiased for σ  
Also s → σ with increasing n so s is consistent for σ  
If our data has s = 4.2 then 4.2 is an estimate of σ

The s.d. of r.v. xBAR is σ / √n  
s / √n is an estimator of σ / √n

The usual 95% conf. interval is " xBAR +/- (1.96 s / √n) "  
The c.i. is a random interval constructed entirely from the data
PERFORMANCE: \( P( \mu \text{ is covered by 95\% c.i.} ) = 0.95 + (\text{order } 1/n) \)
That is, the actual coverage probability will be close to 0.95 since for large \( n \) the difference is of order \( 1/n \) which is nearly zero.

PERFORMANCE IN ORDINARY TERMS: In around 95\% of many independent attempts the 95\% c.i. covers \( \mu \).

PERFORMANCE for other \( z \): \( P(\mu \text{ covered by } \text{xBAR} +/- (z s / \sqrt{n})) = \text{(area under Z curve between } [-z, z]) + (\text{order } 1/n). \) So a 68\% c.i. is \text{xBAR} +/- (1.0 s / \sqrt{n}).

REMEMBER: If for your data the 95\% c.i. works out to say \{23.4, 24.1\} we cannot say whether or not the c.i. has covered \( \mu \) in this instance.

Example 1: (95\% c.i.) A sample of \( n=37 \) retail outlets finds the sample average \text{xBAR} of selling price for a model of shoe is 87.23. The sample s.d. \( s \) of these 37 selling prices is found to be \( s = 16.8 \). The 95\% c.i. for \( \mu \) is then
\[
87.23 +/- (1.96 16.8 / \sqrt{37}) = \{81.8167, 92.6433\}
\]
We do not know whether the average \( \mu \) for all retail outlets is in this interval or not. But the METHOD of 95\% c.i. will produce an interval that covers \( \mu \) around 95\% of the times it is tried.

NORMAL POPULATION: In this case
\[
P(\mu \text{ covered by } \text{xBAR} +/- (t s / \sqrt{n}) )
= \text{(area under t curve with n-1 degrees of freedom) (exactly!)}
\]

Example 2: A population of retail selling prices is (nearly) NORMAL. A sample of only \( n=3 \) is selected from which we find xBAR = 87.23 and \( s = 16.8 \). The sample is too small to justify the usual c.i.
\[
87.23 +/- (1.96 16.8 / \sqrt{3}) = \{68.219,106.241\}
\]
However, since the population is NORMAL we are entitled to use the 95\% c.i. based upon Student’s t with n-1 = 2 degrees of freedom
\[
\text{xBAR} +/- ( t s / \sqrt{n} ).
\]
From the t-table on page 242 we find for 95\% confidence we must leave
0.025 in each tail so looking under \( t_{0.025} \) we replace 1.96 of the usual method by \( t = 4.303 \). So our 95% c.i. is instead
\[
87.23 \pm (4.303 \times 16.8 / \sqrt{3}) = \{45.4931, 128.967\}.
\]
Based upon only \( n = 3 \) we are forced to use a t-method and get a very wide c.i. which is maybe too wide to be of much use. But that is the price we pay for this small sample. In some cases, if \( s \) is small, we will do alright with a small sample. It is nice to have the method available to us for NORMAL populations.

**Example 3:** (Stein’s method) We wish to obtain a 95% c.i. of the form \( \overline{x} \pm 2 \) for the mean selling price in dollars. A PRELIMINARY sample of \( n_0 = 50 \) is taken from which we find that its sample s.d. \( S_0 = 16.8 \). In order to achieve our objective Stein’s method requires us to continue the sampling to a total sample size of \( n = (1.96 S_0 / \Delta)^2 = (1.96 \times 16.8 / 2)^2 = 271.063 \) or around 271. We do so. Suppose that we find our \( \overline{x} \) from all of the 271 in the expanded sample is 84.6. Then our Stein’s 95% c.i. for the mean selling price in all retail outlets is around 84.6 +/- 2. We say "around" because we rounded 271.063 to 271. In exchange for taking the larger sample we are entitled to the greater precision of +/- 2 (dollars) in our 95% c.i..

**Example 4:** (0-1 scores) The proportion \( p \) of retail outlets discounting our shoes is to be estimated. But REMEMBER \( \mu = p \) and \( \sigma = \sqrt{pq} \). We sample \( n = 55 \) retail outlets and find that for our sample there are 12 outlets discounting our shoes. That is \( \overline{x} = \hat{p} = 12 / 55 \). Our 95% c.i. for \( p \) (i.e. \( \mu \)) could be just as before
\[
\overline{x} \pm 1.96 s / \sqrt{n} \quad \text{(where \( \overline{x} \) is 12/55)}
\]
But it is more usual to use the n-divisor form
\[
\overline{x} \pm 1.96 \sqrt{\frac{(\hat{p}Q\hat{p})}{n}} / \sqrt{n} \\
= 12 / 55 \pm \sqrt{((12/55)(43/55)) / 55} \\
= \{0.109029, 0.327335\}