Recitation assignment 10-20-09. You will hand this in as well as complete a multiple choice bubble on these questions. A few of these questions will be modified in recitation by Nick (different data). Be prepared to work with figures in recitation.

1. z - based CI for mu. An equal probability with-replacement sample of 200 accounts is selected from hospital records for a given month. The purpose is to estimate the mean outstanding balance of all accounts for the period. The data have sample average balance xBAR = $72.55 with sample standard deviation s = $83.49. Give the **99%** confidence interval for the population mean account balance due.

2. Claim made for #1. Around what percentage of random samples of 200 recoreds should resullt in a 99% confidence interval (CI) that covers (encloses) the population mean account balance?

3. T - based CI for mu (requires approximately normal population). Suppose that the same data as problem #1 had been obtained from a sample of only 8 records. Assuming that the distribution of all accounts in the population is approximately normal distributed.

4. Finite population correction in #1. How would you modify CI in #1 if you had sampled *without* replacment and there are 7200 hospital records in the population?

5. No finite population correction applies to #3 since approximately normal population as used here effectively means finite population correction~1. Just note our convention on this.

6. z - based CI for p in independent samples (1, 0 as in 1 = dem, 0 = rep). A with-replacement equal probability sample of 100 records is selected from the population of hospital records for a given month. The purpose is to estimate the fraction p of accounts having a non-zero outstanding balance for the period. The data have 62 accounts of the 100 with an outstanding balance. Give the **95%** confidence interval for p, the fraction of all hospital accounts with a non-zero balance.

pHAT = 62/100 = .62

95% CI for p = fraction of hospital records having non-zero balance:

$.62 +/- 1.96 \ \dfrac{\sqrt{.62 \ .38}}{\sqrt{100}}$

7. This example shows how we may exploit correlation between 2009 tax and 2010 taxes paid by individuals in order to get a narrower CI for μ_y = population mean 2010 tax. What we do is sample individuals, auditing them to anticipate their 2010 tax. This is done at no cost to them and not linked to their name or other

$$\overline{x} \qquad\qquad \mu_x \qquad\qquad\qquad \overline{y}$$
$$s_x = 4.3 \qquad\quad \sigma_x \qquad\qquad\qquad s_y$$
$$\overline{xy} \quad \overline{x}\,\overline{y}$$

$$\dfrac{s_y}{\sqrt{n}}$$

$$\mu_x \qquad \sigma_x \qquad\qquad\qquad\qquad\qquad\qquad \sigma_x \qquad\qquad\qquad s_x$$
$$\mu_y$$
$$\mu_y$$
$$s_y$$

$$\frac{\sqrt{.02 \cdot .98}}{\sqrt{100}}$$

$$\mu_y$$

personal information. Suppose the data on 300 sample individuals gives (data in dollars)

$\bar{x} = 22.43$	$\mu_x = 22$	$\bar{y} = 26.5$
$s_x = 4.3$	$[[\sigma_x = 4.5]]$	$s_y = 5$
$\overline{xy} - \bar{x}\,\bar{y} = 17.3$		$R = 0.77$

I'VE DECIDED TO CHANGE THE FORMULAS BELOW SLIGHTLY. WE SHOULD INSERT KNOWNS FOR ESTIMATES WHERE POSSIBLE BUT WILL NOT DO THAT IN ORDER TO KEEP THE FORMU-LAS SIMPLE.

The usual z - based 95 % CI for mu (y) is yBAR + /-1.96 $\frac{s_y}{\sqrt{n}}$.   WORK THIS OUT.

But wait THERE's MORE. We can at NO COST look up each sample individual's 2009 tax. Moreover, we know μ_x and σ_x from the 2009 tax rolls already paid (if we don't know σ_x just use its estimate s_x). Since x, y are statistically correlated we should be able to exploit that fact instead of just relying on yBAR to estimate μ_y as in the usual CI above. An "improved" estimator of μ_y is

$$\text{yBAR* defined} = \bar{y} - (\bar{x} - \mu_x)\, R\, \frac{s_y}{s_x} \quad \text{(NOTE CHANGE FROM EMAIL)}$$

For this "improved" estimator we are entitled to use a 95% CI given by

$$\text{yBAR*} \pm 1.96 \, \frac{s_y}{\sqrt{n}} \sqrt{1 - R^2} \, .$$

In this context estimate R in the usual way by $\dfrac{\overline{xy} - \bar{x}\,\bar{y}}{s_x \; s_y}$.   It looks complicated but works! Do Do it to see for yourself that the resulting CI is indeed narrower than the one ignoring x scores altogether.

Usual CI for μ_y on y data alone:

In[276]:= **26.5 + 1.96 {-1, 1} 5** $\Big/$ $\sqrt{300}$

Out[276]= **{25.9342, 27.0658}**

"Improved" estimator of μ_y:

In[277]:= **26.5 - (22.43 - 22) 0.77** $\times$ $\dfrac{5}{4.3}$

Out[277]= **26.115**

95 % CI using improved estimator :

In[278]:= **26.115 + 1.96 {-1, 1}** $\Big(5\Big/\sqrt{300}\Big)$ $\sqrt{1 - .77^2}$

Out[278]= **{25.754, 26.476}**

#6.

In[279]:= $.62 + \{-1, 1\} \, 1.96 \, \dfrac{\sqrt{.62 \; .38}}{\sqrt{100}}$

Out[279]= $\{0.524864, \, 0.715136\}$