1. The following describe contexts in which we may employ one of seven different CI procedures. They have not been matched up correctly. Give the correct matchups.

Our purpose is to compare the rate of support for a particular provision of the health care reform bill among registered Republicans as opposed to the rate among registered Democrats. We sample independently from each group by with (or without) replacement method and develop a CI for the difference "rate of Republican support — rate of Democrat support." The sample sizes and population sizes are large in the usual sense.

(a) $\mathbf{P(\mu_x \text{ in } \overline{x} \pm 1.96 \frac{s_x}{\sqrt{n}}*) \to .95}$ ******

Our purpose is to estimate the mean income of MSU electrical engineering undergraduates 5 years out (in inflation-adjusted dollars). This is fairly costly to do because we have to find a randomly sampled graduate, work with them to learn what was or is their income 5 years out, adjust the amount and take special precautions to ensure that random sampling has been followed, missing data is properly dealt with, etc. In addition to enquiring about income we also look up their GPA (likewise adjusted for inflation). A CI is prepared for the inflation-adjusted mean income 5 years out.

(b) $\mathbf{P(\mu_x \text{ in } \overline{x} \pm 2.306 \frac{s_x}{\sqrt{n}}*) \equiv .95}$

The process of machining latches has been brought under statistical control to the extent that sampling variations in the amount x of energy used in machining each part follows an approximately normal distribution from part to part. As part of quality assurance operations we periodically select a random sample of 9 latches at random from daily production and from that determine a 95% CI for the mean energy use per part for the days' production.

(c) $\mathbf{P(p_x \text{ in } \hat{p}_x \pm 1.96 \frac{\sqrt{\hat{p}_x(1-\hat{p}_x)}}{\sqrt{n}}*) \to .95}$

We are keeping track of the fraction of weekly emergency admisssions to our hospital that result in billing difficulties not resolved within 30 days. Each week a random sample of 100 admisssions (there are a great many admissions) is selected without-replacement and examined to determine the sample rate at which such payment difficulties occur, a CI then being prepared for the weekly rate.

(d) $\mathbf{P(\mu_x\text{-}\mu_y \text{ in } \overline{x}\text{-}\overline{y} \pm 1.96\sqrt{\frac{s_x{}^{*2}}{n_x} + \frac{s_y{}^{*2}}{n_y}}) \to .95}$

A new protocol is being considered to guide the timing of hospital stays. Such protocols are very dependent on how the patients respond to care but the flow of patients and the nature of their conditions is regarded as statistically stable over time. Patients are randomly allocated to one or the other of two protocols and the duration of their hospital stays are measured. A CI is prepared for the difference "avg stay if everyone were to be assigned to protocol one - avg stay if everyone were to be assigned to protocol two."

(e) $\mathbf{P(p_x\text{-}p_y \text{ in } \hat{p}_x\text{-}\hat{p}_y \pm 1.96\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)\,*^2}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)\,*^2}{n_y}}) \to .95}$

A sampling of 100 family practice doctors in a particular health services organization has been completed utilizing (equal-probability) without-replacement method, each doctor being scored by x = the number of pills of a partiular type prescribed per patient presenting a particular symptom to them in the last year. A CI for the population mean of x will be prepared, however someone points out that we should take into account that prescription of such pills is thought to be more in line with the DO's than the MD's and we do know the rate of DO's among the population of doctors.

(f) $\mathbf{P(\mu_y \text{ in } (\overline{y} + (\mu_x\text{-}\overline{x})\, \mathrm{R}\frac{s_y}{s_x}) \pm 1.96 \frac{s_y *}{\sqrt{n}}\sqrt{1 - R^2}) \to .95}$

Contaminants have been leaked into a lake due to a ground fault fracture that relaeased material from a long-abandoned underground deposit. We need a quick and reliable estimate of the amount being released while the event is fresh. A random sample of water is taken at 100 locations at various depths and distances proximate to the site. A CI is prepared for the population mean ppm (parts per million) of the contaminant.

(g) $\mathbf{P(\mu_x \text{ in } (\sum_{i=1}^{i=k} W_i\, \overline{x_i}) \pm 1.96\sqrt{\sum_{i=1}^{i=k} W_i{}^2 \frac{s_i{}^{2}\,*^2}{n_i}}) \to .95}$

* Times a factor $\frac{\sqrt{N-n}}{\sqrt{N-1}}$(**or its square**, *for the appropriate population or stratum*) if sampling is without-replacement. Note "**or its square**" above.

** As $n_x \to \infty$ (and/or n_i or y counterpart) in with-replacement equal-probability sampling. When sampling without-replacement one also requires population sizes $N \to \infty$ but the conditions are subtle and will not be given here.

2. Which of the following illustrate sampling with equal-probability and **without**-replacement from the list {Jim, Sal, Henry, Albert, Bill, Willa}?

a. These six are the names of patients by order of arrival and the sample is {Jim, Sal, Henry}.

b. The names are placed on paper slips and placed in a bowl. The slips are randomly mixed in the bowl and three are selected without looking. These are found to be {Jim, Sal, Henry}.

c. Each face of a six-faced die is marked with a different one of the six names. The die is tossed three times. We will take the three names that result. The sample is {Jim, Willa, Jim}.

d. A table of random digits has entries 03679 80822 01407. We set up the correspondence

Jim ↔ 0 Sal ↔ 1 Henry ↔ 2 Albert ↔ 3 Bill ↔ 4 Willa ↔5

The sample is {Jim, Albert, Jim}.

3. It is known that ~62% of a particular population are women. A random with-replacement equal-probability sample of 200 is selected from the population for the purpose of estimating the population mean annual income μ. The data has sample mean $\overline{x}$ = 3.635 (thousand) and sample standard deviation s = 2.48. Looking further at the data, the sample is found to have 92 women whose sample mean income is 2.61 and whose sample standard deviation of income is 3.11. For men sample mean and standard deviation are 4.49 and 4.88.

a. Around how many women are expected in a sample of 200?

b. Verify that the overall sample mean is the *sample-weighted average of the sample means* for the men and women: 3.635 = (92 / 200) 2.61 + (108 / 200) 4.49.

c. We know that women comprise 62% of the population yet they are only 46% of the sample. The post-stratified estimate $\overline{x}^*$ of the overall mean is not the overall *sample mean* but instead combines the sample means of women and men in the *known population proportions* 0.62 and 0.38. That is $\overline{x}^*$ = 0.62 2.61 + 0.38 4.49 is the *population-weighted average of sample means*. This represents women more strongly than did the sample. Does this new estimate $\overline{x}^*$ differ much from $\overline{x}$? Has it increased or decreased from $\overline{x}$? Does the direction of the change make sense to you?

d. Use the information given together with the appropriate formula from the list to determine a 95% CI for overall population mean μ based on estimator $\overline{x}^*$.

e. Ignoring the stratification by sex, just use the overall sample mean and sample standard deviation s and provide the 95% CI for μ.

f. Which of intervals (d) or (e) is the narrower?

4. A random (equal-probability) without-replacement sample of 400 hospitals is selected from the very large population of hospitals *of a particular type* operating in the U.S. Each sample hospital is scored for y = amount of federal dollars (for patient care) billed by the hospital in 2008. It is desired to estimate the population mean of y for all hospitals of the given type. Someone points out that we can probably improve upon the estimation by simply making note of each sample hospital's demographic in the form of μ_x = 22.2 = mean per captia income for the region served by the hospital. The following data are collected.

x sample mean = 21.5 (thousand)	sx = 16.2
y sample mean = 1641.8	sy = 421.7
sample x, y correlation R = 0.46	

a. Give the point estimate of y population mean ignoring x.

b. Give the point estimate of y population mean using regression-based estimator.

c. Give the 95% CI for y population mean ignoring x.

d. Give the 95% CI for y population mean using regression-based method.

e. Give the estimated margin of error (MOE) for (c).

f. Give the estimated margin of error (MOE) for (d).

To achieve MOE (d) by method (c) (ignoring x) would (other things being equal) require sampling 400 / (1 - R^2) ~ 507 hospitals instead of 400, potentially a large additional cost.