CI Compared.

1. Approximate z-Based CI for μ for large n, equal-pr, with-repl.

$$P(\mu \text{ in } (\overline{X} \pm z \frac{s}{\sqrt{n}})) \to \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{r^2}{2}} dr \text{ as } n \to \infty.$$

2. Exact t-Based CI for μ for every n > 1 *provided the population is normal distributed ("in statistical control").*

$$P(\mu \text{ in } (\overline{X} \pm t_{\alpha, df} \frac{s}{\sqrt{n}})) = 1 - 2\alpha$$

For DF = ∞ we recover the z-based normal approximation CI.

$$P(\mu \text{ in } (\overline{X} \pm z \frac{s}{\sqrt{n}})) \to \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{r^2}{2}} dr \text{ as } n \to \infty,$$

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{z}^{\infty} e^{-\frac{r^2}{2}} dr.$$

3. Approximate z-Based CI for μ for n→∞ based on a proportionally stratified equal-pr, with-repl, random sample. We suppose the *relative* sizes of sub-populations (strata) i are $W_i = \frac{N_i}{N}$ where N is the size of the overall population, and N_i is the size of stratum i. Weights W_i are obviously probabilities summing to one. A proportionally stratified, equal-pr, with-repl, sample of n is selected, taking sub-samples of sizes $n_i = W_i$ n (we'll assume these are whole numbers for the sake of the exposition). Simple algebra confirms that the overall sample mean is also equal to the weighted average of strata sample means, $\overline{X} = \sum_i W_i \overline{X_i}$, so has variance

$$\text{Var } \overline{X} = \text{Var} \sum_i W_i \overline{X_i} \overset{\text{indep}}{=} \sum_i \text{Var}(W_i \overline{X_i}) = \sum_i W_i^2 \text{ Var } \overline{X_i}$$

$$= \sum_i W_i^2 \frac{\sigma_i^2}{n_i} = \sum_i W_i^2 \frac{\sigma_i^2}{W_i n} = \sum_i W_i \frac{\sigma_i^2}{n} = \frac{\text{within component of } \sigma^2}{n},$$

where σ^2 is the overall population variance and σ_i^2 is the variance for stratum i. This leads to the large n, z-based, CI for μ, *based on proportionally stratified sampling*:

$$P(\mu \text{ in } (\overline{X} \pm z \frac{\sqrt{\sum_i W_i s_i^2}}{\sqrt{n}})) = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{r^2}{2}} dr \text{ as } n \to \infty.$$

Note that $\overline{X}$ is an ordinary sample mean, but the sample is not an unrestricted random sample of the full population as it is in (1). *CI (3) tends to be narrower for any given sample confidence level, and overall sample size n, to the degree that the "between" component is large. This means that stratification does better to the degree that the sub-population means differ.*

4. Approximate z-Based CI for μ for n→∞ based on a *regression-based* estimator. We must avoid a potential conflict of notation since we've used x for sample but regression treats x as the independent variable. Suppose that the population mean of interest is μ_y, we KNOW μ_x, and we have equal-pr and with-repl samples of n *pairs* (X_i, Y_i). An example would be trying to estimate the population average 2008 income tax μ_y due our municipality when we KNOW the average tax μ_x collected in 2007, and we have a random sample of individuals i whose tax X_i in 2007 we can look up, but whose tax Y_i for 2008 has to be determined by an audit (that we will pay for). There may also be an incentive we pay each participant i. This will be a costly study so every way we have of reducing the sample size required for the same information is important to us. Regression-based estimator of μ_y is not $\overline{Y}$ but instead:

$$\hat{\mu}_{y\,\text{regr}} = \overline{Y} + (\mu_x - \overline{X})\ \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}$$ (remember, we assume μ_x is known)

$$\text{P}(\mu_y \text{ in } (\hat{\mu}_{y\,\text{regr}} \pm \text{z}\ \sqrt{1 - r^2}\ \frac{s_y}{\sqrt{n}})) \to \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{r^2}{2}}\, dr \text{ as n}\to\infty.$$

If, for example, r ~ 0.866 then the CI produced using regression will be around half as wide as that produced by the CI using y-scores alone. To accomplish that greater precision using y-scores alone we would have to increase n to 4n (since $\frac{1}{\sqrt{4n}} = \frac{1}{2}\frac{1}{\sqrt{n}}$). If it costs the municipality $600 to audit each sample 2008 tax (y-score) regression affords great savings.

Note: *None of the usual assumptions of the regression model (normal errors, etc.) is used. This CI is simply a consequence of the fact that 2007 tax is correlated with 2008 tax and we draw a large number n of random samples. We need not know the population correlation before selecting the sample, as only the sample correlation* $r = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{\overline{X^2} - \overline{X}^2}\ \sqrt{\overline{Y^2} - \overline{Y}^2}}$ *is used.*

5. Approximate z-Based CI for μ for n→∞ based on a *multiple regression-based* estimator. The data is $(1, X_{i\,2}, \ ..., \ X_{i\,d}, \ Y_i)$

modify the usual estimator $\overline{Y}$ of μ_y as below (assume μ_i are known)

$$\hat{\mu}_{y\,\text{regr}} = \overline{Y} + \left((1, \ \mu_1, \ \ ..., \ \mu_{d-1}) - \left(1, \ \overline{X}_2, \ \ ..., \ \overline{X}_d\right)\right) \bullet \left(\hat{\beta}_0, \ \ ..., \ \hat{\beta}_{d-1}\right)$$

$$\text{P}(\mu_y \text{ in } (\hat{\mu}_{y\,\text{regr}} \pm \text{z}\ \sqrt{1 - r^2_{\text{mult}}}\ \sqrt{\tfrac{n-1}{n-2}}\ \frac{s_{y\,\text{residuals}}}{\sqrt{n}})) \to \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{r^2}{2}}\, dr$$

r_{mult} = sample correlation between Y_i and fitted values $\hat{Y}_i$.

Multiple correlation r_{mult} is the same as |r| in the straight line regression setup, is never negative, ranges in [0, 1], and its square plays the usual role.