

**University of Toronto
Summer 2014**

**STA304/1003 H1F:
Surveys, Sampling, and Observational Data**

Data Analysis Assignment # 2

Data analysis assignments are for practice only and do NOT need to be handed in.

The file "baseball.csv" has specifications on 797 baseball players from the rosters of all major league teams in November, 2004. (See Assignment 1 for details about variables).

Treat the data in the file as the population of all baseball players in 2004.

1. You want to take a stratified sample of size 25 using stratifying variable *leagueID*. If proportional allocation is to be used, how many observations should be taken in each stratum? Show your work and display the final answer.
2. Use 'R' to take a STRS of size 25 of the player *salary*. Use proportional allocation and stratify by *leagueID*. Copy and paste your sample data.

Under your sample data, clearly indicate your answers, including output where necessary:

- a) Estimate the population mean using your sample data.
 - b) Estimate the variance and the standard error of the sample mean, for $n=25$.
 - c) Create an approximate 95% CI for the mean player *salary*.
 - d) Find the population mean. Does your CI include the parameter?
-
3. You want to take a stratified sample of size 50 using stratifying variable *leagueID*. Assume the standard deviation of SRS for AL players is twice that for NL players. If Neyman allocation is to be used, how many observations should be taken in each stratum? Show your work and display the final answer.

4. Use 'R' to take a STRS of size 50 of the player *salary*. Use Neyman allocation and stratify by *leagueID*. Copy and paste your sample data.

Under your sample data, clearly indicate your answers, including output where necessary:

- a) Estimate the mean *salary* for:
 - i. all players in 2004
 - ii. National League (NL) players in 2004
 - iii. American League (AL) players in 2004
- b) Estimate the variance and standard error of the 3 sample means described in part a).
- c) Create an approximate 95% CI for each of the 3 population means described in part a).
- d) Find the 3 population means described in part a). Do the CIs contain the appropriate parameters?