University of Toronto
Summer 2014

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

Data Analysis Assignment # 1

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. The variables are:

Column	Name	Value
1	team	team played for at beginning of the season
2	leagueID	AL or NL
3	player	a unique identifier for each baseball player
4	salary	player salary in 2004
5	POS	primary position coded as P, C, 1B, 2B, 3B, SS, RF, LF, or CF
6	G	games played
7	GS	games started
8	InnOuts	number of innings
9	PO	Put Outs
10	A	number of assists
11	E	Errors
12	DP	number of double plays
13	PB	number of passed balls (only applies to catchers)
14	GB	number of games that player appeared at bat
15	AB	number of at bats
16	R	number of runs scored
17	H	number of hits
18	SecB	number of doubles
19	ThiB	number of triples
20	HR	number of home runs
21	RBI	number of runs batted in
22	SB	number of stolen bases
23	CS	number of times caught stealing
24	BB	number of times walked
25	SO	number of strikeouts
26	IBB	number of times intentionally walked
27	HBP	number of times hit by pitch
28	SH	number of sacrifice hits
29	SF	number of sacrifice flies
30	GIDP	grounded into double play

Data Source: Forman, S. L. (2004). *Baseball-reference.com. Major league statistics and information.* Retrieved November 2004 from www.baseball-reference.com.

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

1. Use 'R' to take a SRS of size of 50 of the player *salary* in 2004 . Copy and paste your sample data.

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

   a) Estimate the population mean of salary using your sample data.

   b) Estimate the variance and the standard error of the sample mean, for n=50.

   c) Create an approximate 95% confidence interval (CI) for the mean salary.

   d) Find the mean salary using the population. Does your CI include the parameter?

2. Calculate by hand/using 'R', the sample size required to estimate the population percent of baseball players who are pitchers within 3% of its true value using a 95% CI .

   a) Using 'R', take a SRS of the required size of the appropriate variable and copy and paste your sample data. Hint: look at how the data is coded and create an appropriate indicator variable.

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

   b) Estimate the population proportion and the standard error of the sample proportion, for n.

   c) Create an approximate 95% CI for population percent of pitchers.

   d) Find the population percent. Does your CI include the parameter?