# STAT 480 Project 1

**Instructions:**

1. Please read this project quickly now and then again more carefully later, so that you will understand what your group needs to manage this project. If you have any general questions, such as the description of the problems below, you can send an email to me or TA at least one day ahead of the due date.

2. Type your group report using Microsoft Word or latex, and use single spacing and a 12 point font in the main body of your report.

3. Every report should have the title (STAT 480 Project 1), authors and date on the first page.

4. All pages should be numbered, and all Tables and Figures should be clearly labeled and numbered, too.

5. Make sure you proofread your report before submitting it.

6. For each group, you need to submit both your group report (in .pdf) and the corresponding R codes through Blackboard. If the R code is all you hand in, with no output result or explanation/discussion of your output result, you will not receive full credit for the question!

7. Please do not ask me or TA to debug your codes.

1. (15 pts) **The Central Limit Theorem (CLT).** CLT is considered to be one of the most important results in statistical theory. It states that means of an arbitrary finite distribution are always distributed according to a normal distribution, provided that the sample size, $n$, for calculating the mean is large enough. To see how big $n$ needs to be we can use the following simulation idea, generate a sample of size $n$ drawn repeatedly (say 1000 times) from a $Uniform(0,1)$ distribution. We want to verify that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normally distributed, and we can draw a QQ-plot for the 1000 $\bar{X}$s to judge the normality. (Hint: `runif()` - generates values from a random uniform distribution between 0 and 1; use the `for()` loop to repeat the sampling)

   (a) (12 pts) Write an R function called `CLTdemo(n)` to illustrate the above CLT, put the number of observations, $n$, as calling argument to the function. When you draw your QQ-plot, make sure that title shows the sample size, $n$. For example, when $n = 100$, the QQ-plot title should be "QQplot for sample size n=100"

   **Hint:** to make the title, you can use

   ```
   title(paste("QQ-plot for
   sample size n=", as.character(n)))
   ```

   (b) (3 pts) Use the `CLTdemo(n)` function in part (a) for $n$ = 2, 10, 25 and 100 and show your results.

2. (30 pts) **A Random Walk.** A symmetric simple random walk starting at the origin is defined as follows. Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables with the distribution

$$\begin{cases} +1, & \text{with probability } 1/2; \\ -1, & \text{with probability } 1/2. \end{cases}$$

   Define the sequence $\{S_n\} \geq 0$ by

$$\begin{aligned} S_0 &= 0 \\ S_n &= S_{n-1} + X_n \text{ for } \quad n = 1, 2, \ldots \end{aligned}$$

   Then $\{S_n\}$ is a symmetric simple random walk starting at the origin. Note that the position of the walk at time n is just the sum of the previous $n$ steps: $S_n = X_1 + \cdots + X_n$.

   (a) (10 pts) Write a function `rwalk(n)` which takes a single argument n and returns a vector which is a realisation of $(S_0, S_1, \ldots, S_n)$, the first $n$ positions of a symmetric random walk starting at the origin.

   **Hint:** the code `sample(c(-1,1), n, replace=TRUE, prob=c(0.5,0.5))` simulates $n$ steps.

(b) (10 pts) Now write a function `rwalkPos(n)` which simulates one occurrence of the walk which lasts for a length of time $n$ and then returns the length of time the walk spends above the x-axis. (Note that a walk with length 6 and vertices at 0, 1, 0, -1, 0, 1, 0 spends 4 units of time above the axis and 2 units of time below the axis.)

(c) (10 pts) Now suppose we wish to investigate the distribution of the time the walk spends above the x-axis. This means we need a large number of replications of `rwalkPos(n)`. Write two functions: `rwalkPos1(nReps,n)` which uses a loop and `rwalkPos2(nReps,n)` which uses `replicate` or `sapply`. Compare the execution times of these two functions by using `system.time`.

3. (15 pts) **Simple Linear Regression.** For the simple linear regression, we know that we can use the least squared method to estimate the intercept and slope. According to the statistical theory, our estimation will be more accurate when the sample size, $n$, is larger or the measurement error is smaller (i.e. the variance of the error $\varepsilon$ is smaller).

   To see this, generate data $(X_i, Y_i)$, $i = 1, 2, ..., n$ from the following linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, ..., n$$

   where $\beta_0 = 1.5$, $\beta_1 = 3$, $X_i$'s are generated from $N(1, 0.5^2)$ and the errors $\varepsilon$'s are generated from $N(0, \sigma^2)$. For the generated dataset, do a linear regression of $y$ on $x$ and obtain the estimates of the coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$. Calculate the squared errors: $(\hat{\beta}_0 - \beta_0)^2 + (\hat{\beta}_1 - \beta_1)^2$. Repeat the above for 200 times, then you will have 200 squared errors $(\hat{\beta}_0 - \beta_0)^2 + (\hat{\beta}_1 - \beta_1)^2$. We would like to look at the average of these 200 squared errors (called MSE for mean squared error) to determine how accurate the estimators are, which is expected to decrease when we increase $n$ or decrease $\sigma_0$.

   (a) (12 pts) Create an R function `MSE(n,sigma)` with the arguments `n` and `sigma` to illustrate the effect of the number of observations $n$ and the standard deviation $\sigma$ in the above and report MSE.

   (b) (3 pts) Use the R function in (a) part for different combinations of $n$ and $\sigma$: $n = 50, 100, 200$, $\sigma = 0.1, 0.5$. Report your MSEs for each combination.

4. (40 pts) **Softball Team Data**. The STAT/MATH/CS co-ed softball team has participated in intramural co-ed softball for 32 years, playing over 450 games during that time. Considering the composition of the team, it is perhaps not surprising that the team has compiled some statistics from those games. Their team captain team was wondering what relationship, if any, there was between how many hits and errors the team made in a game and how well they did. He consulted the data-set, and determined that the team had played 466 games over its history. For each game are listed **RUNS** (his team's score), **ORUN** (the opposing

team's score), **HIT** (the number of hits obtained by his team), **ERR** (the number of errors committed by his team), and **RES** (the result; 'W', 'L', or 'T', depending on whether his team won, lost, or tied the game). The data set is attached in a file called 'softball.txt'. It contains a header line and 466 lines of data.

(a) (5 pts) Create a new variable called **DIFF**, where DIFF=RUNS-ORUN. Then, execute appropriate commands to create and print out the 5-by-5 matrix of Pearson correlations between the five numerical variables (**DIFF, RUNS, ORUN, HIT, ERR**) and discuss briefly what this output reveals.

(b) (10 pts) Since **HIT** is a measure of offensive prowess, one would suspect that the team would score more runs as **HIT** increases. Run a simple linear regression to predict **RUNS** from **HIT**. Use it to predict how many runs would be scored in a game when the team obtained 15 hits, and obtain a 95% Prediction Interval for this estimate. Of the 45 games in which the team achieved exactly 15 hits, in what proportion of these games was the team's **RUNS** actually in the interval calculated in the previous sentence?

(c) (5 pts) Since **ERR** is a measure of defensive ineptitude, one would suspect that the opposing team would score more runs as **ERR** increases. Run a simple linear regression to predict **ORUN** from **ERR** to confirm this relationship.

(d) (10 pts) Run a multiple regression to predict **DIFF** from **HIT** and **ERR**, fitting the model:

$$DIFF = \beta_0 + \beta_1 \times HIT + \beta_2 \times ERR + \epsilon.$$

Softball fans would expect $\beta_1$ to be significantly positive and $\beta_2$ to be significantly negative. Test each of these alternatives at the $\alpha = .01$ level.

(e) (10 pts) Suppose that one uses the regression equation of (d) to predict the outcome of a game by declaring the game a victory if the predicted DIFF $> 0$ and a loss if predicted DIFF $< 0$. Compare these results with the actual results and fill in the table below. (The 9 games which ended in ties should be counted as 1/2 wins and 1/2 losses). What % of all games are predicted correctly? (A game is predicted correctly if it falls in the upper left or lower right boxes in the table below.)