Economics 142 Problem Set #1
Due Thursday January 31

Note: It is recommended that you use the R programming language for assignments in this class. You are welcome to use other programming languages (e.g., Python) or software packages (e.g., Stata). However, the problem sets will contain instructions assuming you are using R.

For problem sets that require coding (probably all of them), you are required to submit your **reproducible code separately** along with a write up with answers to the questions. We will discuss the online submission procedure in section this week.

1. Set up an R program to conduct a simulation of drawing a sample size of ***n*** from a Bernoulli distribution with mean ***p***. In each "replication" ***r*** you will draw a sample, construct the estimated mean from that replication (which we will denote as $\overline{Y}_r$), and calculate the 95% confidence interval $(\overline{Y}_r - 1.96 s_r / n^{1/2}, \overline{Y}_r + 1.96 s_{nr} / n^{1/2})$ where s_r is the estimated standard deviation in that replication, $s_r = (\overline{Y}_r (1 - \overline{Y}_r))^{1/2}$. In each replication, record the length of the confidence interval, and whether or not the true mean is inside the interval.

For given (***n***, ***p***), conduct 1,000 replications and report the following statistics:

a) the mean estimate of ***p***
b) the mean estimate of the true standard deviation
c) the fraction of time that the confidence interval contains the true ***p***. This is called the "coverage rate"

Conduct the analysis for the cases ***n***=30 using ***p***=0.05 and ***p***=0.25, and again for ***n***=60 using ***p***=0.05 and ***p***=0.25 (a total of 4 cases). It is often claimed that n of 30 or larger is enough to ensure that asymptotic confidence intervals work well. Do you agree or not?

2. Go to the course bcourses page and download the data set "ps1.csv" (Files -> ps1). This is a simulated data set based very closely on real data, giving a new born baby's weight (in grams) and his/her mother's weight before pregnancy (in pounds) and her height (in inches). It has 48,871 rows (one for each baby) and 3 columns. We are going to develop some descriptive graphs using these data.

a) Imagine you want to convey to a non-statistical reader the relationship between mother's weight and baby's weight. Develop the following graphs
(i) find the deciles of mother's weight. Then put each mother in a decile, and get the mean birthweight of babies for mothers in each decile. Plot the mean baby weight against the mean mother birthweight in the decile. This is called a "binscatter" plot in economics.
(ii) in addition to the mean baby's weight for each decile, construct the 10, 25, 50, 75 and 90 percentiles. Plot all these against mean mother's weight for mothers in each decile. Do you see any interesting pattern?
(iii) instead of using 10 deciles of mother's weight, redo the analysis in (i) using the individual (1-pound) values of mother's weight. How does this compare to the decile (bin-scatter) plot?

Problem Set 1, continued

2 b) Divide mothers into 5-pound buckets, starting at 95 pounds. Calculate baby weights separately for 4 groups of mother's within each bucket:

- momheight ≤ 60 inches
- momheight between 61 and 63 inches
- momheight between 64 and 66 inches
- momheight ≥ 67 inches

Now plot the mean baby weights for the four groups against the mid-point weight in the bucket. (So if your first bucket is 95-99 points, the midpoint is 96.5). What can you conclude about how mother's weight and height affect baby's weight?

2c) Think about how to construct a 3-dimensional bin-scatter. For example, you can put mother's weight and height into 25 bins (5 x 5) and plot mean baby's weight on a 3-d graph. Try to be creative!