36-303: Sampling, Surveys & Society
HW03: Due Tue Feb 14, 2012 on Blackboard

Reminders:

- Things coming up:

  Due Thu Feb 16: Team Assig. I.3
  - Choose single project to do, *based on my feedback for I.2*.
  - Turn in revised versions of A, B, C, D, E, F, and G for the single project you will do this semester. Details in "project-schedule.pdf" handout (available at the class website, `http://www.stat.cmu.edu/~brian/303`).

  Due Thu Feb 23: Team Working Agreement.
  - Read and do the exercises in the twa.pdf file available at `http://www.stat.cmu.edu/~brian/303`.
  - Turn in a completed team working agreement
    - On paper and signed by everyone on your team. Every team member should also have a signed copy.
    - Submit a copy on Blackboard to me.

- Things to read:
  - You should be reading Lohr, Appendix B (handout, also available at class website).
  - For next week: Groves, Ch's 7 & 8.

- Clear, careful writing and interpretation of results is an important part of both weekly homeworks and the projects. *I always expect neatly typed or neatly handwritten work.*

- Always be judicious about including computer output and graphs: show enough that we can clearly see what you are doing, but not so much that we will get lost or bored leafing through your work!

- Mostly homeworks will be submitted on Blackboard.

Exercises to Turn In (there are 5 exercises):

1. Consider a set of n trials involving independent and identically distributed random variables $X_1, X_2, \ldots, X_n$, where the outcome of 1 corresponds to success and the outcome of 0 corresponds to failure, and the probability of success on each trial is p.

   (a) Show that $E(X_i) = p$, and $\text{Var}(X_i) = p(1-p)$ for $i = 1, 2, \ldots, n$.
   (b) Let $Y = X_1 + X_2 + \ldots + X_n$. Show that $E(Y) = np$, and $\text{Var}(Y) = np(1-p)$.
   (c) Let $\hat{p} = Y/n$. Show that $E(\hat{p}) = p$, and $\text{Var}(\hat{p}) = p(1-p)/n$.

   Use facts from lecture notes and/or the Lohr Appendix B handout to justify your reasoning for these problems.

2. We say that two discrete random variables are *independent* if $P[X = x, Y = y] = P[X = x] \times P[Y = y]$ for all specific values x and y that X and Y could be.

   (a) The table below shows the distribution for two random variables X abd Y. For example, $P[X = 1, Y = 3] = 3/8$.

x	y	$P[X = x, Y = y]$
1	4	$\frac{1}{8}$
1	3	$\frac{3}{8}$
2	4	$\frac{3}{8}$
2	3	$\frac{1}{8}$

   Use the information in the table to show whether X and Y are independent, or not.

   (b) For any two discrete random variables X and Y (not necessarily the ones in part (a)), show that **If** X and Y are independent, **then** $P[X = x|Y = y] = P[X = x]$, for any specific x and y that X and Y could be.

3. Let X and Y be discrete random variables with finite sample spaces $\{x_1, \ldots, x_K\}$ and $\{y_1, \ldots, y_K\}$, and let $p_{ij} = P[X = x_i \text{ and } Y = Y_j]$. Use the definitions of $E[\ ]$, $Var(\ )$, and summation notation ($\sum_{i=1}^{K}$), to show

   (a) $E[aX + bY + c] = aE[X] + bE[Y] + c$

   (b) $Var(aX + bY + c) = a^2 Var(X) + 2abCov(X, Y) + b^2 Var(Y)$

   (c) If X and Y are independent, then $E[X|Y = y] = E[X]$, for any y.

4. Which of the following modes of data collection is most likely to produce question-order effects:

   - Face-to-face
   - telephone
   - Mail-out/mail-back
   - Web-based survey (like surveymonkey.com)

   Why?

5. The Current Population Survey (CPS, http://www.census.gov/cps/), sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS), is the primary source of labor force statistics for the population of the United States. The CPS is the source of numerous high-profile economic statistics, including the national unemployment rate, and provides data on a wide range of issues relating to employment and earnings.

   (a) Answer items E, and F from the class project outline, for the CPS. Be particularly clear about the target population, and use terms from Groves Ch 5 to specify the mode of data collection as carefully as possible.

   (b) The unemployment rate is one of the key variables that the CPS produces. By exploring the CPS website (or other websites), find the exact definition of the unemployment rate that the CPS uses, including who is eligible to be included in this calculation, and who is not. Comment on whether this is a suitable measure of unemployment. Give a list or references or websites you used for this exercise.