Worksheet on Sampling (Chapter TAN.3) Fall 2013
AGEC 4213

(1) You have 5 observations of Y for males whose values are: Y_{males} = 1, 2, 3, 4, 5; you also have 3 observations for females: $Y_{females}$ = 6, 7, 8. Show that the sample average $\frac{i=males,females\, Y_i}{8}$ can be written as the weighted average of the males and females, where the weight equals the percent of the sample comprised of each gender.

(2) Regarding (1), suppose you are using these data to predict the average value of Y for the entire population of males and females, where each gender comprises 50% of the population. Show how the weights in (1) can be modified to give an accurate projection of the population average of Y. What is this projection?

(3) Kinsey wished to estimate X, the percent of the male population who has had a homosexual experience. He did so with the following calculations, except that instead of estimating X_N he simply assumed it was equal to X_V. He did have an estimate for P_N though.

$X = (X_V)(P_V) + (X_N)(P_N)$.

X_V = true average of X for all who will share their sexual past
X_N = true average of X for all who will NOT share their sexual past
P_V = the percent of the population of interest who will share their sexual past
P_N =the percent who will NOT share their sexual past

True value of X for ALL Americans (including those who will and will NOT share their sexual past is: $X = X_V P_V + X_N P_N$

(3.1) Do you believe that X_N is greater than, equal to, or less than X_V?

(3.2) What does this mean for Kinsey's calculation of X?

(3.3) How could Kinsey have modified his sampling procedure to estimate X_N?

(3.4) Kinsey was told he did not have to worry about the statistics of his calculations because his sample size was very large. How does the calculation of $X = X_V P_V + (X_N \text{ set} = X_V)P_N$ change with a larger sample? Does it eliminate the bias? Why?

(3.5) A regular average is like a weighted average when all the weights equal _____.

(3.6) How might the Kinsey Reporter app help sex researchers collect less biased data?

(4) Suppose you are interested in the percent of people who cook with lard, and assume that the percent of <u>all</u> U.S. households using lard equals X. You, of course, cannot calculate X because you cannot collect data from all U.S. households. However, you can collect data from a sample of households using an internet survey and calculate a sample percentage, denoted $\hat{X}$. Suppose that out of the entire population of households there is some portion of people that does not use the internet, and so cannot take your survey. Describe what this implies for your estimate of X by circling the right answers and filling in the blanks.

The people who do not have internet access and do not use the internet even at work are (*circle one*)

LESS / MORE likely to use lard than households who do have access to the internet (using

your best judgment). This suggests that your estimate $\hat{X}$ calculated from your internet survey is

biased, and is expected to be (*circle one*) LARGER / SMALLER than X. To correct for this

biased, you use a different strategy to collect data from households who do not use the internet—

you pay people to visit households in poor and rural households and ask about their usage of lard.

This type of sampling, where you use one strategy for sampling one type of people and a different

strategy for other types of people, is referred to as _____ sampling.

(5) Download the data from our survey on dieting at:
http://seeds.okstate.edu/SeedsPPP/TAN,3/StratifiedSampling,MeatSurvey,2013.xlsx

(5.1) First, identify rows with any missing data. Use the ISBLANK inside of an IF statement to identify missing data (you would do this if your data contained thousands of observations.

(5.2) Use DATA, SORT to sort observations according to whether they have missing data, then by gender, then by age.

(5.3) Delete rows with missing observations.

(5.4) Calculate the sample average response for each question.

One can easily lose weight (and keep the weight off) by consuming less carbs, replacing the carbs with more meat, dairy, and/or eggs.	**One can easily lose weight (and keep the weight off) by consuming less carbs, replacing the carbs with more vegetables.**	**One can easily lose weight (and keep the weight off) by consuming less carbs and consuming less calories.**	**One can easily lose weight (and keep the weight off) by consuming less calories, but eating the same types of foods.**	**One can easily lose weight (and keep the weight off) by consuming less fatty foods.**	**One can easily lose weight (and keep the weight off) by consuming less fatty foods and consuming less calories.**	**One can easily lose weight (and keep the weight off) by eating the same kinds and amounts of food but exercising more.**

(5.5) Now, suppose you know that the population of interest is comprised of the following groups.

Category		Percent of Population
Gender	Age	
Male	18-24	0.10
Male	Older than 24	0.40
Female	18-24	0.15
Female	Older than 24	0.35

Using weighted-averages to make the sample better resemble the population, calculate the estimated population average for the following questions. The SUMPRODUCT formula might be particularly useful.

One can easily lose weight (and keep the weight off) by consuming less carbs, replacing the carbs with more meat, dairy, and/or eggs.	**One can easily lose weight (and keep the weight off) by consuming less carbs, replacing the carbs with more vegetables.**	**One can easily lose weight (and keep the weight off) by consuming less carbs and consuming less calories.**	**One can easily lose weight (and keep the weight off) by consuming less calories, but eating the same types of foods.**	**One can easily lose weight (and keep the weight off) by consuming less fatty foods.**	**One can easily lose weight (and keep the weight off) by consuming less fatty foods and consuming less calories.**	**One can easily lose weight (and keep the weight off) by eating the same kinds and amounts of food but exercising more.**

(6) Watch Video 4 to answer the following questions. To help yourself think through the questions, consider that the percent of votes cast in favor of Mitt Romney can be calculated as

Percent Votes for Romney =

(% of voters who are Democrats)(% voting Democrats who vote for Romney)
+ (% of voters who are Republicans)(% voting Republicans who vote for Romney)
+ (% of voters who are Independents)(% voting Independents who vote for Romney)

(6.1) Why would a poll pay more attention to people who are "extremely interested" in the election than people who are "likely voters"?

(6.2) When a poll is said to "over-sample" Democrats, what exactly does that mean and why is it important?

(6.3) What does mean to "unskew" a poll results?

(6.4) If Brett Bear mentions the Fox News poll contains 20,000 responses, what does that large sample size convey?