IGP 304 (Spring 2006) Homework 1 Keys:

1. List 8 or more major problems associated with the spreadsheet from hell.

i) Drug type is the major focus of the study and should be a variable.
ii) For some quantitative variables, different scales/units are used, making the values not comparable.
iii) For some quantitative variables, even when the values are using the same scale, they are recorded in different formats, making them incompatible.
iv) For some quantitative variables, values are recorded in different precisions.
v) For some quantitative variables, some values were recorded as ranges (e.g. 65+, >350), categories (e.g. obese), or with partial missing information (e.g. 130 as blood pressure, Sep 14th as date), making them less informative and difficult to analyze.
vi) In some categorical variables, the same category is recorded using different descriptions, making them appear to be different categories.
vii) Missing data are recorded inconsistently (?, 0, NA, none, unknown, not staged, blank).
viii) Impossible or unlikely values (e.g. 1120/80 in blood pressure, 2/30/99 as a date).
ix) Confusing recordings (e.g. in Race, "NA" is for Native American or not available?).

2. Describe the difference between a continuous variable and a categorical variable, the advantages and/or disadvantages of categorizing a continuous variable.

A continuous variable can take many possible values and the values often describe quantity. Because of this, values can be compared and the difference between two values often is meaningful. A categorical variable can take only a few possible outcomes and the outcomes often are categories. There may or may not be a natural order among the categories; even if the categories can be ordered, the difference between two categories often is meaningless.

Categorization of a continuous variable is useful for summarizing/reporting data or analysis results. It generally is a bad idea to do analysis on a categorized variable instead of the original variable. This is because categorization makes the information less rich and distorts the information. Categorization also is very rigid, allowing only a limited number of possible effects for a variable.

3. Describe what sensitivity analysis is and what scenarios a sensitivity analysis can lead to.

A sensitivity analysis often is used to explore the effect on the results of various factors that are relatively subjective. Examples of these factors include analysis assumptions on distributions and model structures, data preparation parameters (as in Ryan's question), specification of prior distributions (as in a Bayesian analysis).

In general, two scenarios can result from a sensitivity analysis. In one scenario, the results and conclusions are effectively unchanged for all potential alternative choices of the factors. In the other scenario, the results may be quite different depending on the choices of some factors; then interpret the results with caution.

4. Describe Simpson's paradox. A helpful reading is here.

Simpson's paradox describes the seemingly contradicting effects of variable 1 on variable 2 depending on whether we ignore variable 3 or take variable 3 into account. This often is because the subsets of data stratified by variable 3 are not comparable (i.e. heterogeneous) with respect to variables 1 and/or 2. Thus, pooling the heterogeneous subsets together while ignoring variable 3 often leads to misleading results.

5. Nashville's December 2005 daily mean temperatures had mean 37.7 degrees (Fahrenheit) and standard deviation 7.0 degrees. The formula between Fahrenheit and Celcius is F = C * 9/5 + 32. Now, do you have enough information to get the mean and SD in Celcius? If yes, what are these? If no, what else do you need?

The formula can be re-written to be C = (F – 32) * 5/9. In Celsius, the mean is (37.7 – 32) * 5/9 = 3.2 degrees and the SD is 7.0 * 5/9 = 3.9 degrees.

6. A binary outcome can only take two possible values. Examples include coin flipping, sex of newborn babies, having a type of cancer or not, etc. We always can denote one outcome as "1" and the other as "0". Let the probability of having "1" be p. Then $0 < p < 1$ and $q = 1 - p$ is the probability of having "0". Suppose there are n outcomes. The number x of outcome "1" can vary from 0 to n, with varying probabilities. These possible outcomes together with their associated probabilities are called a binomial distribution. The parameter p can be estimated by x/n. The coefficient of variation of this estimator is sqrt[(1-p)/(np)] x 100%.
    - Calculate the CV for n=10, 100, 1000 and p=.01, .05, .1, .3, .5.
    - Comment on how CV changes as n increases with p fixed and as p changes with n fixed.
    - Suppose you want to estimate a cancer rate with accuracy measured as CV < 10%. What sample size do you need if the real rate is about 10%? What sample size do you need if the real rate is about 1%?

	P = .01	P = .05	P = .1	P = .3	P = .5
N = 10	315%	138%	95%	48%	32%
N = 100	99%	44%	30%	15%	10%
N = 1000	31%	14%	9%	5%	3%

With p fixed, CV decreases as sample size increases. With a fixed sample size, CV decreases as the probability of outcome “1” increases from relatively unlikely to moderately likely.

If p = .1, to achieve CV < 10%, we need sqrt[(1 – .1) / (n*.1)] < 10% = 0.1, which resolves to 9/n < 0.01 and n > 9/0.01 = 900. So, 900 subjects should give us the desired accuracy.

If p = .01, to achieve CV < 10%, we need sqrt[(1 – .01) / (n*.01)] < 10% = 0.1, which resolves to 99/n < 0.01 and n > 99/0.01 = 9900. In this situation, 9,900 subjects are needed to give us the desired accuracy.