PS 699: Problem Set 3

1. An article in the News/Free Press (true story) reported on a study concluding that there was no evidence of a link between a woman having had an abortion and her risk of breast cancer. 1.5 million women were in the study altogether. The concluding paragraph of the article states "Among the 280,965 women in the study who had 370,715 abortions, 1,338 cases of breast cancer were diagnosed by 1992. In the remaining group of women who had not had an abortion, 8,908 cases of breast cancer occurred, according to the study." Test the hypothesis that the mean rate of cancer occurrence is the same regardless of having had an abortion. In particular...

   a) What assumption(s), if any, do you need to make to test the hypothesis?
   b) At what p-level could you just reject the hypothesis? (show your work; make whatever assumptions necessary, indicating which)
   c) Does the conclusion of "no evidence of any difference" seem warranted to you given this evidence?

2. In our data base of characteristics of developed democracies, take the sample of countries with both upper and lower chambers of their legislatures. Assume that these available observations are a random sample from the universe of possible developed democracies with bicameralism (two legislative chambers). Test the hypothesis that the mean number of seats in each house is equal. In particular...

   a) What assumption(s), if any, would you make before proceeding?
   b) At what p-level could you just reject the hypothesis? (show your work)
   c) On this basis, what would you conclude regarding their equality or inequality?
   d) Would you be comfortable concluding that, rather than equal, one chamber has higher mean number of seats? Which (if either) would you conclude has higher mean number of seats and on the basis of what test?

3. In each of the following, assume the available observations in our data set represent a random sample from the population of possible developed democracies. (You may use a spreadsheet or any useful computer program you like.) For each variable mentioned...

   a) give an estimate from the sample of its mean in the population,
   b) give a 90% confidence interval for the sample mean
   c) test the hypothesis that the mean is μ_0 (I'll give μ_0 in parentheses after each variable); report a p-level, don't merely say reject / don't reject
   d) test the hypothesis that the variance is σ_0^2 (I'll give σ_0^2 in parentheses); report a p-level...
   e) state whatever distributional assumptions, if any, you are making in answering a-d
      3i) Lower House Seats (LSEATS, μ_0=300, σ_0^2=31,000)
      3ii) Lower House Proportionality Index (LPROP, μ_0=90, σ_0^2=25)
      3iii) Number of Secondary Government Units (SECGOV, μ_0=35, σ_0^2=250)
      3iv) Average Number of Parties in Government in Postwar Era (NPGOVPW, μ_0=1.5, σ_0^2=1)
      3v) Number of Governments in the Postwar Era (NGOVPW, μ_0=22.5, σ_0^2=100)

4a. What are the sample covariance and correlation between NPGOVPW and NGOVPW?
4b. Test the hypothesis that the variance of the Vanhanen Index of Democracy is greater in 1970 than in 1960 (VDEM60 and VDEM70).

5a. Show that sample variance measured by, $s_a^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ is...

   i) biased
   ii) asymptotically unbiased (unbiased in the limit as $n\rightarrow\infty$) and consistent. It may help you to know that the variance of s^2 across repeated samples is $2\sigma^4/n$.
   iii) and find its M.S.E.

5b. Show that sample variance measured by, $s_a^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ is...

i) unbiased
ii) Since it is unbiased, it is asymptotically unbiased; is it consistent? It may help you to know that the variance of s^2 across repeated samples is $2\sigma^4/(n-1)$.
iii) and find its M.S.E.

5c. Given these results, do you think s^2 or s_a^2 is *better*? Why? On what might your answer depend?

6. The Bernoulli distribution is given by: $f(y_i \mid \pi_i) = \Pr(Y = y_i \mid \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$.
(This is a Bernoulli, so each y_i is 0 or 1.)
Assume you have simple random sample (*s.r.s.*) of N Bernoulli random variables with parameter π_i, and that each $\pi_i = (1 + e^{-x_i\beta})^{-1}$.

a) What is the joint distribution of the y_i's? *I.e.*, what is $f(y_1, y_2, \dots, y_n \mid x)$?

b) This joint distribution is the likelihood function of our outcome data, y, given our input data, x, and our model (theory, assumption) of how x relates to π_i. What, then, is the log-likelihood; i.e., what is $ln\{f(\cdot)\}$?

c) Take the derivative of the log-likelihood with respect to β and set it equal to zero. *I.e.*, set $\partial[ln\{f(\cdot)\}]/\partial\beta = 0$. β^* as a function of x and y cannot be found analytically, but simplify the implicit function given by $\partial[lnf(\cdot)]/\partial\beta = 0$ as much as possible.

You just set up a logit estimation.

7. You estimate two regression models. In one, you estimate $y=b_0+b_1X_1+b_2X_2+b_3X_3+e$; in the other, you estimate $y=a_0+a_1X_1+u$.

a. If the regression output gave the log-likelihood of the 1st model as 17.3 and, of the 2nd, as 15.4; how might you test the hypothesis that the 2nd model was an insignificant restriction on the 1st?

b. Suppose, instead, your statistical package gives you that $b_2=.05$ and $b_3=.025$ and that the estimated variance-covariance matrix of those two estimated coefficients is:

$$\widehat{V(\mathbf{b})} = \begin{bmatrix} .02 & -.01 \\ -.01 & .04 \end{bmatrix}$$

Now how might you test the hypothesis that the 2nd model was an insignificant restriction on the 1st?

c. Suppose the t-stat for b_2 was 1.2 and on b_3 it was 0.5; can you tell if the 2nd model was an insignificant restriction on the 1st based only on this info? How? or Why not?