Homegrown Exercises around M&M Chapter 10

If going on to Course 621 (Data Analysis I) use SAS (INSIGHT or PROC REG) or SPSS to do the following analyses. If not, you may use whatever program you wish (or even fit the equation, and calculate the residual variation, manually).

To set up the SAS dataset, you can either (i) download the already created sas file directly to your sasuser directory or (ii) or download the sas program that contains the data, bring it into the Program Editor, and run the data step from there to create it. No matter which route you take, you can then perform analyses from (a) INSIGHT or (b) running PROCedure steps from within the program Editor To get help on the syntax for a procedure (e.g. the CORR procedure), type HELP CORR in the command box)

-1- The 1970 Draft Lottery (data under Resources for Ch 10)

Run the "new" lottery 30 times and make a stem and leaf plot (by hand, on paper, is sufficient) of (a) the correlations (b) the p-values. Comment on the shapes for (a) and (b).

-2- Correlations in twins and of bone density measures at different anatomical sites (data under Resources for Ch 10)

Determine the point and (by hand) the 95% interval estimate for the correlation of the heights of dizygotic twin pairs (Interpolation using the CI nomogram, rather than calculating it from the formula, is sufficient. Despite M&M's comment at the bottom of p693, the calculation isn't *that* "tedious" and they could easily have provided a nomogram)

Are the correlations of the heights of dizygotic and monozygotic twins pair significantly different at the alpha=0.05 level(two-sided)? Use a direct test on the difference, rather than comparing for overlap of the two CI's [cf notes on Ch 2]

Determine the correlations — for dizygotic twin 1—between (a) tea and coffee consumption (b) the bone density at the 3 sites (spine, and 2 femoral) within the same twin. Interpret these coefficients in words. (If you have time, check whether the same patterns hold up for twin2, and for twin1 and twin2 in the monozygotic pairs)

-3- Correlations: heights of parents (100 from Galton's dataset) (data under Resources for Ch 10)

Determine the (Pearson) correlation between fathers' and mothers' heights. Contrast this with M&M's assumption (without data I suspect) in question 2.46. Answer M&M's question 2.46.

-4- Variability of, and trends in, proportions (SAS program and data, and data in an Excel sheet, are available under Resources for Ch 10)

Refer again to the data on the proportion of Canadian adults responding YES to the question "Have you yourself smoked any cigarettes in the past week?" in Gallup Polls for the years 1974 to 1985.

a Fit a linear regression to these data (regress Rate on Year).

b Identify and interpret the 2 regression coefficients (parameter estimates)

c Calculate a 95% CI to accompany each coefficient.
[Can use respective SE's, together with appropriate t value from the $t_{(n-2)df}$ *table, to construct them]*

d Regress Rate on (Year minus1974) *[a new variable already set up in SAS program... this new variable would also be easy to create "after the fact" within INSIGHT: EditMenu->Variables->Other... Apply the a+bY transformation to "Y"=Year, using a= –1974 and b=1. Notice that the use of the names "X" and "Y" within the "Edit Variables" dialog box bears no relation to "X" and "Y" used in the regression. The transformation will be applied to whatever you designate as "X" and "Y", but this designation is local and is forgotten once the variables are created.]*

e Identify and interpret the 2 coefficients of this new equation.

f Use the coefficients of the new fitted equation to double check the Intercept you obtained under the original equation. From this, state a general rule about the effect of shifting the X Variable on the regression coefficients.

g Why do you think the SE for the intercept is much smaller under the new formulation? Why hasn't the SE for the slope (the coefficient of Year or "Year minus 1974") changed from one formulation to the other?

h Identify and interpret a measure of residual variation from the fitted line (Since the "y" variable is on a percentage scale, make sure you measure the residual variation in this same scale)

Note that this measure of residual variation (which is a mix of sampling variation and any inaccuracies in specifying the form of the curve) does not use the n's (1050 or so) from which these proportions were estimated, or their stated margins of error. Yet is comes close to the value on which the stated margins of error are based.