1. A study aimed to investigate whether caffeine can influence finger speed. The study involved asking sixty volunteers to tap a sequence of letters on a keyboard as fast as they could, twenty minutes after drinking a type of cola. Twenty volunteers were randomly assigned caffeine–free cola, twenty were randomly assigned regular cola and the remainder were assigned a mixture of the two. Each volunteer was timed to complete the sequence of key taps. The volunteers were not told which cola they were given, but the researchers prepared the cups of cola in another room directly before passing them to the volunteers.

   (a) What type of study is described?
   Experiment

   (b) What is the response variable in this study?
   The time taken to tap out the sequence of letters.

   (c) What is the explanatory variable here? How many levels does the explanatory variable have?
   The explanatory variable is the level of caffeine consumed. There are three levels.

   (d) Circle which, if any, of the following were involved in the study:

   - **a control group**
   - simple random sampling
   - **randomization**
   - **blinding**
   - matched pairs

   (e) Indicate a likely flaw in the study, explaining your answer clearly.
   The study appears to take no account of the accuracy of the sequence of letters typed. So speed may be attained at the expense of accuracy.

2. Every Winter Olympics there is a debate about whether the judging in the ice dancing competition is fair. Suppose the correlation between the British judge and the Russian judge was 0.82, taking data across the entire competition. Indicate by circling T or F whether the following statements are true or false:

(a) We can be sure the British judge tends to score higher than the Russian judge.

T **F**

Justify your choice:
The correlation does not tell us that one variable tends to be larger than the other.

(b) When the British judge awards a relatively low score, the Russian judge tends to award a relatively low score also.

T F

Justify your choice:
The correlation of 0.82 indicates that a relatively high score from the British judge tends to arise with a relatively high score from the Russian judge.

(c) We could accurately predict the Russian judge's score given the British judge's score.

T **F**

Justify your choice:
The value of the correlation does not enable us to predict one of the variables given a value of the other (even with perfect correlation).

(d) There is evidence that the British and Russian judges are influencing each other with regards their scoring.

T **F**

Justify your choice:
It is not possible to infer influence based purely on correlation.

3. In a large population, the distribution of annual income is very skewed, with a long right tail. We take a simple random sample of n people from this population and record the n incomes. We expect a histogram of the n incomes in the sample

(a) **will not resemble a Normal distribution whatever the value of n.**

(b) will resemble a Normal distribution provided n is large.

(c) will resemble a Normal distribution for all values of n.

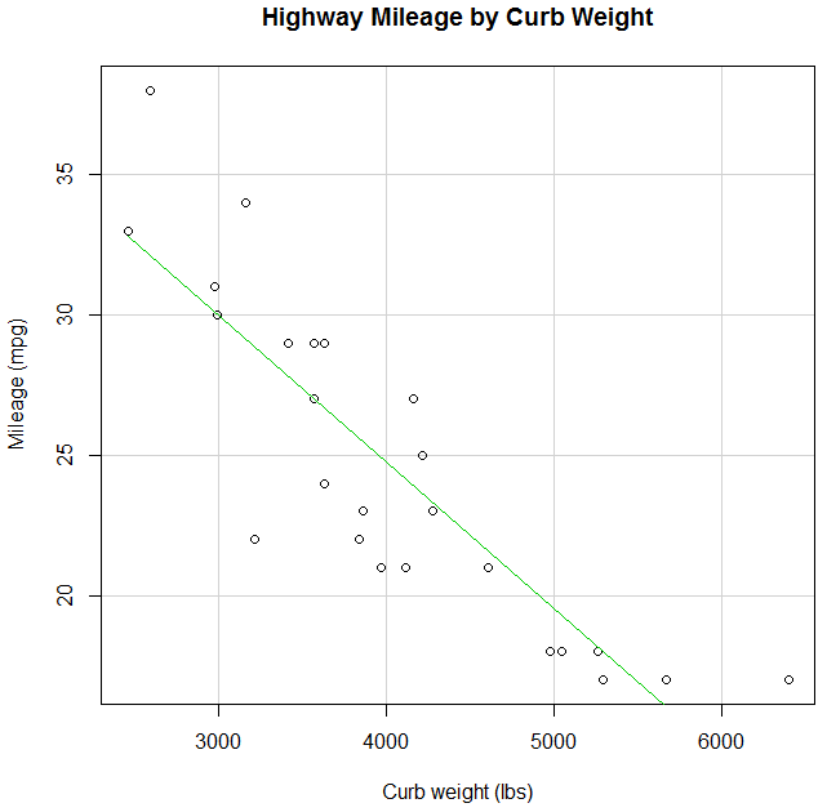
(d) will resemble a Uniform distribution provided n is large.

(e) will resemble a Uniform distribution for all values of n.

Explain your choice clearly:

The sample will resemble the "parent" distribution for all sample sizes, at least approximately. The distribution described is skewed and not Normal.

4. An engineer is interested in whether the fuel economy of a car depends linearly on its weight. She obtains data on 25 vehicle models. For each, she finds the fuel economy in "highway" driving (in miles per gallon), and also the "curb" weight (in pounds). A graph of the data is shown below, including the fitted regression line:



(a) Which of the following is the correlation for the data above?

i. 0
ii. 0.10
iii. –0.40
iv. –**0.80**
v. –0.95

(b) The regression line fitted above is

$$Y = 45.645 - 0.00522X$$

where Y is the fuel economy (in mpg) and X the weight (in lbs). Give a clear interpretation of the estimate of the slope in the above model.
The model suggests that for every pound increase in (curb) weight, miles per gallon decreases by 0.00522.

(c) If investigating the engineer's research question, explain clearly what you would take as the null hypothesis.
We could take the correlation between mpg and curb weight is zero (in the larger population of cars). Alternatively we may take that the slope in the regression line between the two variables is zero.

(d) Describe clearly what you would take as your alternative hypothesis. Justify your choice.
As reasonable alternative may be that there is a negative correlation between the mpg and weight (or equivalently that the linear relationship has negative slope). We may think it reasonable that heavier cars will tend, on average, to need more fuel to run than comparatively lighter cars. (A case could perhaps be made for a two–sided alternative too. Any justification involving eye-balling the data is not acceptable.)

(e) Suppose you were to conduct a simulation–based exploration of the research question with these data. Explain clearly exactly how you would conduct your simulation–based exploration. You can assume you have a lot of time to do your study and, should you require it, access to the internet. You should indicate how you would form your conclusion from your exploration.

What you would do and how you would draw your conclusion:
Physical: Put each mpg value on a small piece of paper. Write down the curb weights in order. Shuffle the papers together. Deal out papers onto the curb weights. Compute the correlation (or regression coefficient), using software or otherwise. Repeat these steps lots of times, each time computing the correlation for the re-shuffled data. Find the proportion of data sets that have correlations that are below (or farther from zero than) –0.8, or those that have regression slopes below (or farther from zero than) –0.00522. If the proportion appears low, we may conclude there is evidence there is a (negative) correlation between the two variables.
Computer–based: Enter the data in pairs into an online applet (such as the Tintal et al. applets used in class). Use the software to shuffle the mpg values and compute the correlation or regression line for the re-shuffled data. Repeat these steps lots of times, each time computing the correlation (or slope estimate) for the re-shuffled data. Find the proportion of data sets that have correlations that are below (or farther from zero than) –0.8, or those that have regression slopes below (or farther from zero than) –0.00522. If the proportion appears low, we may conclude there is evidence there is a (negative) correlation between the two variables.