SSSI Assignment 2009

Due: Tuesday 15th December 2009

Key things to note:

- More marks will be given for answers that show the details of working and which demonstrate understanding.

- Answers should not exceed 3,000 words (excluding tables, equations and appendices). The assignment should be completed using SPSS (as opposed to an alternative statistics package).

- Note that the material being assessed also includes that covered by the recommended reading, not just the content of lectures, labs and tutorials.

- For the marking scale used by examiners, see the current version of the LBSS Graduate School Handbook.

Background to the Data:

Your PhD research seeks to explore the stereotypes embodied in super hero/villain characters and how these compare to superhero characters invented by primary school children. You analyse the ratings recorded for Marvel characters as listed in the *Top Trumps Marvel Comic Heroes* cards. You also encourage a group of primary school children from various year to play with the cards to help them understand the rating system. You then ask them to develop their own superhero characters and to describe the characteristics of their creations using the *Top Trumps* rating system for each of the variables (height, intelligence, strength, speed, agility, fighting skills, and gender).

The data is entered into SPSS and saved as **SuperHeroes09.sav** (available from the Statistics and SPSS page of www.gpryce.com).

For the purposes of your analysis, all characters (whether good or evil) are described as "superheroes".

(A) Summary Statistics (10%)

1. Central Tendency and distributions:

   (a) What does it imply if the median is greater than the mean? Plot a histogram of *fighting skills* to see whether this supports your answer. What would the histogram look like for a variable for which the mean was greater than the median? Demonstrate using a good example of a variable of this kind from the data.

   (b) What is a density curve? Explain its properties and how it is used in statistics.

2. Spread:

   (a) Use an appropriate measure to compare the variability of superhero height and superhero intelligence. Comment on your results.

   (b) How is the sampling distribution of the mean affected by the variance of a variable?

3. Understanding Inference:
By looking at histograms and/or summary statistics of each of the following variables, consider which ***test formula*** you would use for a hypothesis test on the mean (or proportion). Explain you answer in each case.

(a) Height of Marvel superheroes
(b) Height of female superheroes created by Primary School children.
(c) Fighting skills of male superheroes created by Primary School children.
(d) Fighting skills of female superheroes created by Primary School children.
(e) Gender of all superheroes in the data

(B) Inference (40%)

1. For each of the following characteristics, calculate the % difference in sample means between male and female Marvel superheroes:
   - height
   - intelligence
   - strength
   - speed
   - agility
   - fighting skills

2. Repeat question B.1 for the characters created by primary school children.

3. Calculate also the 95% and 99% confidence intervals for:

   (a) the difference in agility scores for male and female Marvel characters;

   (b) the difference in agility scores for male and female characters created by primary school children;

   (c) the difference in mean intelligence scores for male and female Marvel characters;

   (d) the difference in mean intelligence scores for male and female characters created by primary school children;

   (e) the proportion of characters created by primary school children that are male.

   In each case, comment on the implications of your results.

4. Test the following hypotheses about superheroes and comment on your results.

   (a) Male Marvel superheroes tend to have the same agility as female Marvel superheroes.

   (b) Male superheroes created by children tend to have the same agility as female superheroes created by children.

   (c) Male Marvel superheroes tend to be more intelligent than female Marvel superheroes.

   (d) The intelligence gender bias against females holds true for superheroes created by children.

   (e) Primary school children tend to think of superheroes as being male.

5. How do each of the results of your hypothesis tests in question (iii) relate (if at all) to the confidence intervals you calculated in question (ii)? Comment on each test separately.

(C) Relationships Between Categorical Variables (10%)

Looking at the dataset as a whole, do you think whether the creator of a superhero character was Marvel or a Primary School Child has any bearing on whether the superhero is likely to be male or female? Run a cross-tab and chi-square test to support your answer. How is the methodology you use here different to the methodology used in section B?

(D) Regression Analysis (40%) *

NB You may be elligible to one of the alternatives to Section (d) – see info on p.4 below

1. Using the characters created by primary school children, construct a regression model explaining the intelligence score assigned to a superhero character in terms of height, intelligence, strength, speed, agility, fighting skills and gender. Explain your results in detail, paying particular attention to the meaning of the coefficients, t-values and significance levels.

2. Are the results different if the regression is run on Marvel comic characters?

* Alternative questions for Section D are listed below for those who want to use a dataset more closely related to their subject area/research interests. Alternatives are also provided for those who do not need to learn regression analysis (i.e. those for whom neither progression to Social Science Statistics Module 2, *nor* training in basic regression is a requirement for their degree). Most (if not all) students will not need to consider any of these alternatives – you will simply do the question set out above. However, if you do opt for an alternative question, you should only do ***one*** of these three alternatives.

Alternative Questions for Section D

If the section D question listed above is not appropriate, choose ONE out of the following alternative questions:

- Section (D): Alternative Question 1 For Students Wanting to Use Other Data
- Section (D): Alternative Question 2 For Students Not Required to Learn Regression

Details of each of these alternatives are given below:

Section (D): Alternative Question 1 For Students Wanting to Use Other Data:

(i) Choose a data set from the following list (you can select a dataset outwith this list but bear in mind that it may be unsuitable for the kind of analysis required below, or it may need considerable preparatory work):

- Accident Speed Data Set
- BCSagecrime
- BHPSagepolitics
- BHPSincomehheffects
- BPPSincome
- BSASageattitudes
- LOTRnov05
- Universities

(ii) Having chosen a dataset, state the title and source and provide a summary description of the contents of the dataset, highlighting issues to look out for when analysing the data (such as the coding of missing values, inconsistencies, etc.).

(iii) Choose and comment on your dependent variable – i.e. one that you want to explain using other variables in the dataset. Your dependent variable should be approximately continuous. Now choose 3 or 4 explanatory variables. Briefly explain the intuition behind your model – how the explanatory variables determine the dependent variable, and what sign you would expect their regression coefficients to be.

(iv) Prepare your data for regression analysis. This will involve converting categorical variables to dummy variables and recoding the data (e.g. converting all missing values to system missing). Briefly summarise the steps you have taken to prepare your data for regression analysis.

(v) Run a regression of your dependent variable on the various explanatory variables you have selected (you may end up with more than 4 explanatory variables if one or more of them is categorical converted to a dummies – remember to avoid the dummy variable trap!)

(vi) Comment on your regression results, paying particular attention to the meaning of the coefficients, t-values and significance levels.

Section (D): Alternative Question 2 For Students Not Required to Learn Regression:

This option is only available to students who are not required to learn basic regression analysis, and are not intending to sit Social Science Statistics 2. It is recommended that you confirm with your supervisor/course coordinator that this option is compatible with the training requirements of your degree:

(i) Find a refereed journal article in your subject area published since November 2008 that uses empirical data based on a sample (i.e. rather than the entire population, which is rarely the case anyway).

(ii) Write up a critique of the study, focussing on the following issues:

- The extent to which the research shows an understanding of the difference between population and sample;
- The extent to which the research considers the impact of sampling variation by presenting information on means and proportions as confidence intervals (or uses hypothesis tests);
- The extent to which the paper considers the sampling method used and the potential biases that ensue (you might want to discuss the alternatives and consider how the method chosen compares);
- Any other problems of reliability and validity in the data.

(iii) If possible, calculate the appropriate confidence intervals for variables reported in the data and compare with the author's findings.

(iv) Propose an alternative research brief, explaining how your suggested approach improves on the methods and data used in the article.