**STATS 747, 2nd Semester 2010 Assignment 1, Due Wednesday 4th August**

**1)** In your own words, write a brief description (with illustrative examples) of the following concepts to an interested marketing manager who has never heard of these concepts before.

* What is sampling error?
* What is the difference between sampling error and non-sampling error?
* What is a p-value?
* What is a focus group and what are they used for?

Note: write no more than a paragraph (2-3 sentences). Marks will be awarded for being clear and concise.

**2)** You are asked to design a survey to estimate the percentage of Aucklanders who intend to vote in the next election for the new ‘super city’ local body election.

* What sampling frame would you use?
* What variables (if any) would you consider using for stratification and/or clustering?
* What information would aid in the design of this survey, and for weighting the survey results?
* As a consequence, describe a survey design that you think would work well in this situation, along with the weighting procedures you think would be appropriate.

**3)** Suppose you have a continuous explanatory and response variables that you suspect have an underlying linear relationship. Suppose also, that there was substantial missing data, accounting for 5-20% of respondents on each variable (this level of ‘missingness’ is very common in market research data).

(a) Suppose that cases with missing data are simply deleted (list-wise deletion of missing data). Briefly describe the possible effects of this approach on the regression coefficient and its standard error.

(b) Suppose that cases with missing data are imputed with mean values for each variable. Briefly describe the possible effects of this approach on the regression coefficients and its standard error.

(c) Suggest an alternative way of imputing missing values that address any deficiencies you may have found in parts (a) and (b) above.

Hint: Simulate, say, 1000 values with a underlying linear relationship and create missing values etc…

**4)**  The Excel file ‘Assn1.xls’, available on , is a data set for the following questions (a selection from a recent study) asked of 202 respondents about some new snacking concepts (concepts J,K,L,M,N,P,Q & R):

**Questionnaire:**

|  |
| --- |
| **Age of Respondent:** |
| 20-29 1 |
| 30-39 2 |
| 40-49 3 |
| 50-59 4 |

|  |
| --- |
| **Respondent Gender**  |
| Male 1 |
| Female 2 |

SQ2. Are there any children living in your household?

Record number of Children

\_\_\_\_\_\_\_\_

SQ2a. **How many are aged:**

\_\_\_\_\_\_\_\_\_

Under 5years

\_\_\_\_\_\_\_\_\_

5-10 years

\_\_\_\_\_\_\_\_\_

11+ years

…..

 Q4. Now, taking into account your own feelings about snacking and what others in your house like to snack on, I’d like you to consider how relevant each of these ideas is to your household.

 **SHOW CONCEPTS J-Q AGAIN IN RANDOM ORDER**

 **FOR EACH CONCEPT ASK:**

 How *relevant* is this to your household?

 **CODE ONE ONLY IN CORRECT ROW BELOW**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CONCEPT NAME** | **Very Relevant** | **Quite Relevant** | **Not Particularly Relevant** | **Not at all Relevant** | **Don’t Know** |
|  Concept J |  1 |  2 |  3 |  4 |  5 |
| Concept K |  1 |  2 |  3 |  4 |  5 |
| Concept L |  1 |  2 |  3 |  4 |  5 |
| Concept M |  1 |  2 |  3 |  4 |  5 |
| Concept N |  1 |  2 |  3 |  4 |  5 |
| Concept P |  1 |  2 |  3 |  4 |  5 |
| Concept Q |  1 |  2 |  3 |  4 |  5 |
| Concept R |  1 |  2 |  3 |  4 |  5 |

The client wants you to calculate the mean relevance score and standard deviation for each of the concepts after you have:

* Recoded Q4 with the following numbers:

 Very relevant =10

 Quite relevant = 5

 Not particularly relevant = 1

 All others =0

* and re-weighted this recoded data according to these three user groups, whose actual proportion, in a sample of n=200 is:

Under 40, no kids aged 5-10 years   70

Have kids aged 5-10 years           80

Aged 40+, have no kids aged 5-10 years 50