CGS603A: Methods and Tools in Cognitive Science

Assignment #1

Max marks: 290
Due on/before:23.59, 17-Feb.-2019. 7-Feb.-2019

Wherever necessary use R (preferred) or statistical tables to get the necessary values.

1. Popular belief is that when one is angry it helps to express one's anger physically on inanimate objects like a punching bag, pillow, etc.

   Design an experiment to test the above hypothesis. Carefully explain and justify how your experiment will actually achieve this. Make any assumptions you make explicit. Also, clearly indicate your dependent and independent variables and inclusion criteria for subjects. [60]

2. Tuberculosis has become a serious problem in India. There are strains of the TB that are antibiotic resistant. Suppose that 0.2% of individuals in a population suffer from tuberculosis. Of those who have the disease, 98% test positive when administered a diagnostic test. Of those who do not have the disease, 85% test negative in the test. Choose an individual at random and administer the test. Let E be the event that a person has tuberculosis and F the event that the test is positive.

   (a) Construct a tree diagram with two branches: infected and not infected. From each of these branches, show two branches: test positive and test negative. Show the appropriate probabilities on each of the four branches.

   (b) What is P(E and F)?

   (c) What is P(F)?

   (d) What is P(E—F)?

   [10,5,5,5=25]

3. Use the data set FORBES2000, available in the R package HSAUR2 available at:
   `https://cran.r-project.org/web/packages/HSAUR2/index.html`
   The data has company data on 8 variables for 2000 companies in the Forbes list.

   Graphically display information using R that will give answers to the following (you can use more than one way to display the information if it helps):

   (a) How are companies distributed with respect to country?

   (b) How are companies distributed to category?

   (c) What is the distribution for sales and percent profits as a percentage of sales?

   (d) Is there any correlation between assets and sales, market value, profit percentage?

   [5,5,5,10=25]

4. There are different ways to estimate a statistic. Given a sample, consider the following ways to estimate the mean μ:

1. Sample mean.
2. A randomly chosen value from the sample.
3. The average of the largest and smallest value in the sample.
4. $\bar{X} + \frac{2}{n}$.

With respect to properties of estimates (unbiasedness, consistency, efficiency) what can be said about each estimator? Note that in some cases it may not be possible to say something about a property since it may not be measurable. [40]

5. This question looks at the connection between smoking and lung cancer.

   The table below gives the data published by Doll in 1955 for average number of cigarettes smoked per year in 1930 and the average number of deaths per million due to lung cancer in 1950. The data is only for males. Note that this does not establish causation but is suggestive early evidence.

Country	**1930 cigarettes smoked per yr.**	**1950 deaths/million of lung cancer**
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1100	350
Great Britain	1100	460
Holland	490	240
Iceland	230	60
Norway	250	90
Sweden	300	110
Switzerland	510	250
United States	1300	200

   (a) Using R plot a graph for the two variables using suitable coordinates.

   (b) Using R compute the Pearson correlation coefficient r.

   (c) What is your conclusion based on the correlation value.

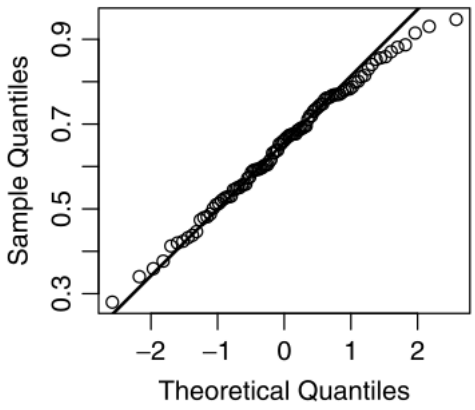
   (d) What could be confounding factors that can be invoked to explain the correlation?

   (e) Give at least 3 independent examples of variables that are likely to be highly correlated (positively or negativly) with almost certainly no causation.

[5×5=25]

6. Suppose the population consists of the 5 measurements 3, 2, 2, 1, 1 with a mean of 1.8. Assume you do not know this mean and want to estimate it by drawing samples of size 3. The statistics you want to experiment with are mean, median and average of largest and smallest of the sample. There are 10 possible samples. Construct the sampling distribution for each statistic. Which statistic will you choose to estimate the mean? Why? [25]

7. One intuitively simple test for normality is to plot quantiles of a theoretically normal distribution on the X-axis versus the same quantiles for the sample normal distribution on the Y-axis. If the sample is normally distributed then one should get a graph very close to a straight line. See figure below.



Write an R script to test out normality for the sample distribution for the mean statistic for differnt discrete and continuous distributions and for different sample sizes. The theoretically normal distribution can be the standard normal distribution. You can start with the R script we discussed in class. [50]

8. Assume that 25% calls to 100 are false alarms. Let X denote the number of false alarms in a random sample of 100 alarms. What are the approximate values of the following probabilities?
   (a) $P(20 \leq X \leq 30)$
   (b) $P(20 < X < 30)$
   (c) $P(35 \leq X)$
   (d) The probability that X is more than 2 standard deviations from its mean.

   [5×4=20]

9. Assume that the mean height of the Indian population is 160cm with a σ of 15cm. Assuming the distribution for the height is normal (in reality it will be a mixture of two Gaussians for male and female) find:
   (a) The probability that the height for a randomly selected person exceeds 197.5cm.
   (b) The probability that the height of a randomly selected person is between 145cm and 175cm and between 135cm and 185cm.
   (c) Suppose that a person is said to be short if his/her height is less than 145cm. What is the probability that at least one person in a sample is short?

   [20]