Assignment 2

WORKSHOP

Dataset

- Topic: statistics on infectious diseases
- Location: US
- Period: 2001 – 2014
- Number of samples: 141777
- Group by gender (male, female, total)

Guidelines for questions

In the period of 2001 – 2014,

(a) Which county had the upward trend of *Amebiasis* disease regardless of gender? (You are required to create a suitable plot to support the statistical analysis)

- *How to extract suitable subsets of data?*
- *How to know the trend is upward?*
- *Which plot should be used to display the trend?*

(b) What can you infer from your findings?

- *What type of upward trend (e.g.,. fluctuation, steady rise, …)*
- *Is the trend of total infected cases is different from that of male/female infected?*
- *Is the trend periodic, random, or unspecified?*

Guidelines for questions

Q2: In the year of 2005, which county had the highest rate of infected females for HIV?

- *How to extract suitable subsets of data?*
- *Which rate should be used (e.g., male, female, total)*
- *Is the highest rate of infected associated with the largest population?*

Q3: In the year of 2010, which county had at least 10 infected cases for *Malaria*?

Q4: In the period of 2010 - 2012, which county had at NO infected case for *Tuberculosis*?

- *How to extract suitable subsets of data?*
- *What is the ratio of counties with at-least-10/NO case to the other counties?*
- *Is the higher (or lower) number of infected case associated with the larger (or smaller) population?*

Guidelines for questions

Q5: Over the whole period,

(a) What is the correlation (R) between the rates of *Chlamydia* and *Salmonellosis* diseases in California? (You are required to create a suitable plot to support the statistical analysis)

- *Which metric should be used to investigate the relationship between variables.*
- *Which plot should be used to display the relationship?*

(b) What can you infer from your findings?

- *How strong is the correlation?*
- *How is the scattering of the data?*
- *Is the analysis affected by outliers?*

Guidelines for questions

Q6: Over the whole period,

(a) Are the rates of *Dengue* disease in San Diego and in San Francisco statistically different?

- *Which test should be used to investigate the statistical difference?*
- *How many samples in each set?*
- *Are these sample paired?*

(b) Write a short paragraph explaining your findings and the reasons for choosing testing methods.

- *Why the testing method chosen?*
- *Is the outcome reliable based on the test's assumptions?*
- *Is there any suggestion to improve the test/outcome?*

Guidelines for questions

Q7: Over the whole periodM

(a) Are the rates of *Cryptosporidiosis* in California, Lake, San Diego, and San Francisco statistically different from each other (ignoring the year)?

- *Which test should be used to investigate the statistical difference?*

(a) Which of these counties are exactly different from each other? Which test did you use to determine this?

- *Which test should be used to find the pair of different groups?*

(a) Create a suitable plot to indicate the changes in the rate of *Cryptosporidiosis* in these four counties.

- *Which plot should used to indicate the changes?*

(a) Write a short paragraph explaining your findings.

- *Which can be drawn from (a), (b), and (c)*

THANK YOU