Hypothesis testing and Data Analysis Assignment 3

Issued: Friday, 12/1/2018 **Due**: Thursday, 25/1/2018

Exercise 1 (50%)

You are part of a software engineering team developing a video surveillance and security system. The system includes several subsystems (modules). Namely

- A video capturing and pre-processing module
- A face detection module
- A face tracking module and
- A decision making and alerting module, in case the system detects a security risk.

Your project manager has assigned to you the task of designing and developing the face detection submodule (i.e. algorithm).

You are given the following specifications:

1. The selected camera captures images at 25 frames per second.
2. The video acquisition and pre-processing module has a maximum execution time of 10ms
3. The face tracking and the decision making and alerting subsystems together have a maximum execution time of 10ms.
4. Your subsystem (i.e. algorithm) reads data (i.e. video frames) and writes (the coordinates of the detected face – if any) to volatile memory.

Once you have implemented your algorithm you will obviously need to test as to whether its performance is within the required specifications. In order to test this you create an experimental setup in which to test your implementation.

Answer the following:

a) Briefly describe what this experimental setup might be.

b) What is the variable(s) you will measure?

c) What do you think are the confounding/external variables?

d) How would you control or alleviate any potential effects of your confounding/external variables on you independent variable? Discuss.

In your experimental setup you execute your submodule and register the observed execution time, t_i for i=1 to 36 and you get the following set of data (in milliseconds)

$[t_1 \dots t_{12}]$ = [20.8, 21.0, 18.0, 19.0, 18.3, 19, 24.0, 22.3, 21.5, 22.8, 21.6, 20.5]

$[t_{13} \dots t_{24}]$ = [22.0, 22.0, 19.5, 19.0, 24.0, 18.0, 20.4, 20.8, 25, 18.9, 20.5, 21.3]

$[t_{25} \dots t_{36}]$ = [19.1, 22.3, 21.4, 21.5, 22.8, 21.6, 20.5, 19.0, 18.9, 22.0, 21.3, 20.4]

Do the following:

i. Estimate the sample mean of the sample set $[t_1 \dots t_{36}]$

ii. Estimate the sample standard deviation

iii. State an appropriate null hypothesis.

iv. State an alternative hypothesis you want to test if the null hypothesis is not true.

v. Test the null hypothesis using the test statistic at a 99% confidence level (i.e. you want to be 99% confident about your answer).

vi. Write a short report that you plan to deliver to your project manager together with the software code of your implementation, in which you should discuss the following "*Given that the original specifications hold (i.e. they are satisfied), does your implementation of the face recognition submodule guarantees the real time performance of the whole system?"*

NOTE: Use the Standard Normal table to find the critical value(s) if you use a probability other than 90%, 95% or 99% - for which the critical values have been given in the classroom (see relevant presentation).

Exercise II (50%)

You are asked to develop and evaluate the performance of algorithms for text retrieval. You have implemented two such different algorithms. Algorithm I belongs to the Supervised Classification family of algorithms (e.g. Neural networks, Decision trees or Genetic algorithms) whereas Algorithm II belongs to the Unsupervised, or Clustering family of algorithms (e.g. K-means).

You also have at your disposal a database of containing a benchmarking literature dataset with 5,000 individual records (articles). Assume the following:

- A database contains 240 records on a particular topic.
- A search was conducted on that topic and 180 records were retrieved by Algorithm I and 200 records by Algorithm II.
- Of the 180 records retrieved by Algorithm I only 145 were relevant.
- Of the 200 records retrieved by Algorithm II only 150 were relevant.

Answer the following questions:

1) Which of the two algorithms performs better? Discuss and justify your answer.
2) If your algorithms are part of a literature mining suite, and you are using them to execute a Systematic Literature Review on the specific topic, which of the two, in your opinion, is best suited for the task? Justify your answer.