Experimental Methods
in Computer Science (and Informatics Engineering)
2017-2018

Training exercises 3

This set of training exercises is meant to help students in the preparation for the written exams. If, for any reason, you conclude that it is not possible to answer a given question explain what is wrong and indicate the reason why the question cannot be answered.

1) Assume you are measuring the response time of a web service method and you are taking several measurements. The web service is installed in a server in another continent and the measurements are taken from a clock in the local machine you used to evoke the web service (your machine). In this scenario explain how can you distinguish random uncertainties in the measurements from the uncertainties related to changes in the conditions of your experiment. Explain also how should you deal with these two types of uncertainties.

2) Assume that in an experiment with a very large number of variables you calculate the Pearson's correlation coefficient using 11 samples of two of these variables (named here as V1 and V2) and you obtain r = 0.663. Knowing that you want to simplify the experiment as much as possible to reduce cost, explain what you can do to achieve such simplification. Justify your answer.

3) You are involved in the development of the new information systems for a large lawyers' office in Lisbon. One of the functionalities defined in the requirements is a search feature using free text. The goal is to have text search functionalities over the data stored in the database (legal processes, customer profiles, contracts, legislation, etc.) that supports the information systems. In another words, the goal is to have a kind of Google search engine working inside the information systems you are developing for the lawyer's office.

   In order to implement this feature, you want to compare two search engines, named here as SA and SB, to select the best one for the project. Both search engines can be easily integrated in the system you are developing. Thus, your choice is totally determined by the features of each search engine. To make up your mind you decided to perform an experiment to compare both search engines.

   Search engines use a classic information retrieval algorithm based on inverted file indexes to find quickly the relevant documents for a given search. The response time of each search engine is obviously an important aspect. However, the response time depends on many factors such as the query (i.e., the set of keywords used to search for the documents) and the number of documents indexed, among other factors.

   Another aspect relevant for the selection of the search engine is the accuracy of the results. Typically, not all the relevant documents (i.e., the ones related to the query) are retrieved and the search engines also retrieve some documents that are not relevant. In practice, there are two figures of merit, precision and recall, that are often used to characterize this aspect: 1) precision represents the fraction of retrieved documents that are relevant and 2) recall is the fraction of relevant documents that are retrieved by the search engine. It is worth noting that the modern search engines include in the inverted file index not only the words but also use synonyms, temporal expressions, etc. For example, if you search for "Lisbon" the search engine is able to retrieve documents that do not include the word "Lisbon" but include the expression "capital of Portugal". These features, although very positive, have great impact on the precision and recall of search engines.

   Considering the scenario presented above, describe the following basic elements of the experiment:

   a) Problem statement.

b) Variables

c) Examples of levels for the different independent variables

d) Hypotheses

e) Assumptions and hypothesis testing technique that should be used in this experiment

4) Software inspections proposed by Michael Fagan is a technique to find errors during software development process. Inspections can be applied to any phase of the software development process but Fagan's inspections are mainly used to verify the requirement specifications (because unlike software code that can be run and tested, requirements cannot be tested automatically). Very briefly, in a requirement inspection a group of inspectors (experienced software developers) examine the document describing the requirements in order to find possible errors, inconsistencies or missing requirements. It is a collective effort and during an inspection session one of the inspectors (the reader) is actually reading aloud the requirement specification while the others (typically 2 or 3 inspectors) verify the correction and consistency of the requirement description that are being read by the reader. Obviously, the inspectors are not the authors of the requirement specification under inspection. Whenever one of the inspectors thinks there is an error he/she raises the issue and the error is recorded by one of the inspectors (the recorder). The errors are simply recorded for posterior analysis of the possible solution (i.e., the inspectors do not discuss how the error found can be solved; that will be done afterwards by the authors of the requirement specification).

   Your company uses requirement inspection very often. But recently the company has decided to introduce several changes in the Fagan's inspection process and create a new inspection methodology called FaganT++. In order to verify that the new methodology is superior to the traditional Fagan's inspections, the company selected a group of 10 developers to perform a controlled experiment to compare the two inspection methodologies. The main point under comparison (dependent variable) is the capacity of error detection of each inspection method. Although the inspections also detect some false positives (i.e., record errors that are not real errors) that is not very important, as the authors of the requirements can easily identify them as false positives later on, when they process the results of the inspection to solve the problems found.

   The group of 10 developers was asked to inspect two sets of requirement specifications where someone has previously inserted 50 errors (i.e., small modifications in the requirement descriptions that simulate real requirement errors) in each set. The first set was inspected using the Fagan's method and the second set was inspected using the FaganT++. The group of 10 inspectors performed first the inspection using the Fagan's method. After that, the inspectors received specific training during several days and when the inspectors are already proficient in the FaganT++'s method they perform the inspection of the second set of requirements. As mentioned, both sets have been injected with 50 errors.
   The following table shows the errors detected by each inspector in the two inspections. In real inspections what is recorded is the performance of the group of inspectors. But in this controlled experiment it was decided to analyze the individual performance of each inspector. The inspector I8 does not participate in the second inspection (the one using FaganT++) because of health reasons.

Inspectors →	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
Fagan inspection	33	45	28	32	26	36	40	31	32	42
FaganT++ inspection	40	42	34	32	34	40	35		38	39

   Considering these results, can you say that the new the FaganT++ method is better than the classic Fagan's inspections with 95% of confidence? Show your calculations. (5.5/20)

5) The tuning of a very complex Relational Database Management Systems (RDBMS) such as Oracle 12c to achieve the best performance possible is a very difficult task. The number of configuration options is immense, making the task of deciding which ones are the best for the specific database application at hand quite difficult. Above all, it is difficult to evaluate if a given configuration has a real impact on the performance and conclude which configuration is the best.

   Assume that you are the database administrator (DBA) of a large banking organization and you want to define the production configuration of the new version of the RDBMS acquired by the Bank, the Oracle 12c. As the Bank runs several database applications in different instances of the same the RDBMS (and each application needs a specific configuration, so the configuration that improves the performance of one database application may reduce the performance of the others), you decided to use a performance benchmark to tune up the Oracle 12c in a general way.

   After some procurement, you selected the TPC-E performance benchmark, which has been designed to broadly represent modern online transaction processing (OLTP) applications such as the ones used by your Bank. In practice, TPC-E is a calibrated database application designed to compare the performance of RDBMS and the underlying server machines. You simple installed the TPC-E application in your Oracle 12c and run the TPC-E a given number of times for each configuration you want to test. The result (performance) is given in the form of transactions per second (tps) executed by the Oracle 12c in each configuration. The following table shows a summary of the results you obtained in the experiments:

RDBMS configuration	TPC-E benchmark execution results		
	Average tps	St. deviation	No. executions
Oracle 12c conf. 1	128.6	5.83	7
Oracle 12c conf. 2	127.4	4.88	9
Oracle 12c conf. 3	132.2	4.59	6
Oracle 12c conf. 4	131.6	5.27	9

   Considering these results and the experimental scenario described above, answer to the following questions:

   a) Are the different Oracle 12c configurations relevant for the RDBMS performance? Consider 90% confidence in your answer. Explain in detail your analysis and justify your answer taking into account the four configurations used in the experiments.

   b) Suppose that instead of asking whether the configurations have impact on the performance or not (as in question 5 a)), your goal was to identify the best configuration from the performance point of view. Explain how could you identify the best configuration with a given confidence level. Note that what is asked is the type of analysis, technique and procedure you would use to identify the best configuration (and not what is actually the best configuration). Additionally, assume that your goal is only to identify the best configuration and that is not necessary to rank the configurations ordered by performance.

6) A given company is developing a new application for smartphones and is particularly concern with three features of the future application: speed, energy consumption and memory used by the application. The company decided to perform a controlled experiment to evaluate several software development environments to select the one that can produce the best results, concerning the three application features mentioned above. The key aspects of each software development environment evaluated by the company are the programming language and the compiler, as there are several compilers available for each possible programming language. The target smartphone operating system is also relevant for the speed, energy consumption and memory usage of the resulting application. Naturally, the experiment developed by the company considers several programs with different features (size, complexity, etc.) to test each software development environment, to be sure that the experimental evaluation is general enough to draw robust conclusions.

   Considering the scenario presented above, describe the following basic elements of the experiment:

a) Problem statement

b) Variables

c) Examples of levels for the different independent variables

d) Hypotheses

7) A company specialized in web design is investigating different styles of web interfaces to define a set of best styles to use consistently in future applications. The most important aspect is the subjective user interface experience quality, which must be evaluated by potential users. In order to perform this study, the company selected a group of evaluators (representative users including user with different demographic features such as age, sex, profession, etc.) and asked these evaluators to perform a set of predefined operations in two examples web interfaces: 1) a supermarket web page and 2) a bank web page. These two interfaces and the related functionalities are very different. The evaluators are simply asked to score the experience of using each web interface in a scale from 1 to 100, after performing the tests with the predefined operations. The table below shows the results provided by the evaluators. The evaluator Ev8 only tested the supermarket interface.

Evaluators →	Ev1	Ev2	Ev3	Ev4	Ev5	Ev6	Ev7	Ev8	Ev9	Ev3
Supermarket web page	88	95	78	91	67	69	94	78	49	84
Bank web page	90	98	82	89	55	73	93		55	78

Considering that in a very first step of the study you just want to test the following hypothesis:

H0 – The user interface experience in using both web interfaces is similar.
H1 – The user interface experience in using both web interfaces is different.

In this context, would you use a two-sample dependent T test or a two-sample independent T test? Explain clearly the reasons behind your choice.

8) The general intuition is that female code inspection teams are better than male inspection teams, when performing classic Michel Fagan's code inspections (see, for example, https://en.wikipedia.org/wiki/Fagan_inspection to know the basics about Fagan's software inspections). In other words, the general intuition is that female inspection teams can find more bugs in code inspections than male teams. However, as simple intuition is not enough, your company has decided to find out whether there is a difference in the performance of male and female code inspection teams.

Suppose you have been in charge of designing such experiment and you recorded the number of bugs found by male and female inspection teams in the inspections performed in your company in the last month. The code units inspected represent a variety of the code developed by the company and all the code units have been developed using the same software development methodology (i.e., the chances of having residual bugs are identical).

The following table shows the number of bugs found in the code inspection performed in the last month:

Male teams	3	5	2	3	2	3	8	5
Female teams	6	4	6	4	7	8	2	

Based on these results, can you state, with 95% of confidence, that the performance of male and female code inspection teams is different? Note that it is not realistic to assume that the number of bugs found in code inspections follows a normal (or close to normal) distribution.