Experimental Methods
in Computer Science (and Informatics Engineering)
2018-2019

Training exercises

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) Explain the differences between laboratory experiments and pilot studies, providing simple and concise examples to clarify the key differences between them.

2) What is the main difference between measuring and benchmarking?

3) One of the first steps in the design of an experiment is to define the problem statement (or research question). Give a concrete example of problem statement (i.e., provide the actual sentence that express the problem statement) and briefly explain the experiment context related to each problem statement.

A good (i.e., relevant) problem statement should be focused enough to allow the clear identification of the variables of the problem but, at the same time, should be sufficiently open to allow different hypothesis to answer the problem/question. For example:

Is the time necessary to sort a given number of items in an array mainly dependent on the number of items to be sorted or the size of the items plays a major role as well?

The context of this research question is performance of sorting algorithms and programs.

4) For the example of problem statement provided in your answer to the previous question, indicate the following:

a) Dependent variable(s)

- Sorting time

b) Independent variables

- Size of the array
- Size of each element of the array (or average size)
- Sorting algorithm
- Programming language
- Discuss other: computer, operating system, etc.

c) Examples of possible levels for the independent variables in the experiments

- Size of the array → 100, 10.000, 100.000, 1,000.000, 100.000.000
- Size of each element of the array (or average size) → 1, 2, 3, 5, 10, 20, 50, 100, 200
- Sorting algorithm → Quick sort, merge sort, shell sort
- Programming language → C, Java, Python

d) Hypothesis that could be tested (indicate if the hypothesis is directional or non-directional).

H0 – The size of the items to be sorted have no impact on the sorting time
H1 – The size of the items to be sorted has impact on the sorting time

5) An engineer conducts a hypothesis test and concludes that his hypothesis is correct. Explain why this conclusion is never an appropriate decision in hypothesis testing.

The goal of hypothesis testing is to prove that the null hypothesis is not true, with a given level of confidence. It is not possible to prove that the null hypothesis is correct; you can only prove that it should be rejected.

6) The definition of the levels taken by the independent variables is one of the most important decisions for the success of an experiment. Explain the key points/goals that should be taken into account in the choice of the levels for the independent variables.

7) Outliers obtained in the measurements should be reported but, in general, are removed from the analysis. Explain what should be taken into account in the decision of ignoring or not the outliers.

Outliers can be caused by random variability in the measurements, by errors in the design and implementation of the experiments, by features of the experimental setup, or may simply indicate something scientifically and technically interesting. The most important thing is to determine if the outliers represent bad data (and in that case, they should be removed from the analysis) oy represent something relevant (although unusual). The analyst should try to systematically identify all the possible outliers causes that represent bad data. If outliers are not caused by any error or variability in the measurements, then it's fair to analyze the outliers as they may represent some relevant aspect.

8) Explain what should be done to deal with the two types of measurement uncertainties: random and systematic uncertainties.

Random uncertainties: treated statistically, using averages, variance and calculating confidence intervals.

Systematic uncertainties: should be understood and removed before analysis.

9) Suppose you are the data center administrator of a big organization and you are about to decide if your organization should accept the new cooling system that was recently installed in the data center. As the cooling system is very expensive, the contract defines precisely the conditions that should be met, concerning the temperature inside the server racks. Quoting the contract: "The cooling system assures that the temperature in the racks is always in the range of 16.00 ± 0.80 Celsius degrees, with a confidence of 95%". In order to be sure that the cooling system is operating under the values defined in the contract, you decided to measure the temperature in the racks using a high precision electronic thermometer. To assure representative measurements, you took 100 measurements, including a variety of server loading scenarios and covering the 24 hours of the day. The results obtained show an average temperature of 16.245 Celsius with a standard deviation of 2.234. Do you think the conditions defined by the contract are met and your organization should accept the cooling system as is?

Considering that the measurements are independent (which is the case, as one measurement is not dependent on previous measurements) and that the small differences observed when the same measurement is repeated are due to random uncertainties (i.e., you have normal distributions), the formula to obtain the confidence interval is:

$$\bar{x} \pm z * s / \sqrt{n}$$

where $\bar{x}$ is the mean (arithmetic average) of the samples, z is confidence coefficient obtained form the Z table, s is the standard deviation of the samples and n is the number of samples. That is:

n = 100
z = 1.96 for 95% confidence interval (see the normal table)
s = 2.234
$\bar{x}$ = 16.245

Doing the calculation, the confidence interval is:

$$z * s / \sqrt{n} = 1.96 * 2.234/\sqrt{100} = 0.44$$

The confidence interval for the measured temperature, with 95% of confidence, is

Minimum value: 16.245 – 0.44 = 15.81
Maximum value: 16.245 + 0.44 = 16.68

Based on these results, the conclusion is that the cooling system is operating under the specifications defined in the contract.

10) Consider the scenario described in the previous question but now suppose that you could only take 15 measurements. Explain what is different in this case. Consider both the arguments of the data center administrator (your perspective) and the vending representative of the cooling system.

Solution:

The calculation is the same as for the previous exercise with the differences that confidence coefficient is now obtained from a T table (T Student distribution) and we consider the df (degree of freedom) of n-1 instead of n samples. The t value for df = 14 is 2.1448 (in a T table for upper tail probability, you will have to look at $\alpha = 0.05/2 = 0.025$, as the confidence interval consider two tails.

Doing the calculation, the confidence interval is:

$$t * s / \sqrt{n} = 2.1448 * 2.234/\sqrt{15} = 1.24$$

The confidence interval for the measured temperature, with 95% of confidence and only 15 measurements, is

Minimum value: 16.245 – 1.24 = 15.01
Maximum value: 16.245 + 1.28 = 17.48

Based on these results, the conclusion is that the cooling system is NOT operating under the specifications defined in the contract, with 95% of confidence. Obviously, the problem here is that due to the small number of measurements, the level of confidence is smaller. The vendor could argue that you would need at least 30 measurements to take a solid decision.

11) Consider the following problem statement: the number of software bugs found in the tests of program units developed by programmers is dependent on the average number of sleeping hours of the programmers. Assume that you have the detailed specifications of a set of program units to be developed, and consider that the program units include units of high, medium and low complexity. Additionally, you have comprehensive unit test suits to test each program unit.

In these circumstances, describe how you would organize an experiment to answer the proposed problem statement. Your answer should be as complete as possible, focusing on the experiment design steps (obviously, it does not make sense to speculate about the experiment results and conclusions), and indicate the dependent and independent variables, the levels you would consider for the independent variables, the hypothesis under evaluation and the hypothesis testing technique you would use. Also describe, very briefly, the experimental setup and take into account in your answer to the whole question that the experiment deals with people (the programmers).

The answer should follow the main steps involved in the design of an experiment. As the problem statement is given, we start by identifying the variables.

Variables:
Dependent variable:
- Number of bugs detected by the test suite (this is what we want to observe).

Independent variables:
- Average number of sleeping hours
- Complexity of the program units

Levels:
After defining the variables, an important step is to select the levels for the independent variables, as these levels represent the set points of the experiments. The levels could be:
- Average number of sleeping hours (2, 4, 6, 8, 10 sleeping hours per day)
- Complexity of the program units (small, medium, and high complexity[1])

The two independent variables mentioned above are the most obvious ones (and would be enough for a correct answer). However, there are other independent variables that could be referred. For example:
- Expertise of the programmer (novice, qualified, expert)
- Programming language (Java, C++, Python)

Hypothesis:
As the hypothesis is a tentative answer to the problem, there are several alternatives. In any case, one of the most obvious hypotheses is:
H0 – The number of bugs done by programmers does not depend on the average sleeping hours of the programmers.
H1 – The number of bugs done by programmers increase when the average sleeping hours of the programmers increases.

Note that the definition of the hypothesis includes both the null hypotheses (H0) and the alternate hypothesis (H1).

Considering that the number of measurements/samples in experiments with people is normally limited, the appropriate hypothesis testing technique would be two-sample T-test, applied successively to the measurements obtained with the different levels of the independent variables, or ANOVA (one-way or even two-ways ANOVA, depending on the number of independent variables and levels used in the experiment). In this case we are treating the two samples as independent samples. However, the experiment would be much better if you could use a two dependent sample setup. In this case, the measurement would be taken considering the same programmer and different sleeping times.

[1] The complexity of program units can be measured using classic software complexity metrics such as McCabe's cyclomatic complexity or Halsteads metrics (if you are not familiar with programming complexity metrics have a look at: http://en.wikipedia.org/wiki/Programming_complexity).

It is worth noting that although the raw data has a binomial nature (bug vs no bug) we do not have measurements based on propositions. That is, the data is viewed as numerical data (bug counts) and not proportions, so it is better to use T-test instead of test for proportions.

Experimental setup/scenario

Very briefly, as requested… Considering the simple scenario with only two independent variables, the experiment setup consists of a group of programmers that develop a set of program units of different complexity levels. Each programmer develops all the program units involved in the experiment (including low, medium and high complexity program units), covering all the possibilities for the levels defined for the sleeping hours (at least one program unit of each complexity level should be developed in each level of sleeping hours). In order to full control the experiment, the sleeping hours of the programmers need to be controlled using a simple alarm clock, in order to achieve the levels defined for the variable "average number of sleeping hours". When a programmer declares that a given program unit is finished and ready for testing, the program is tested using the test suit (mentioned in the question). The number of bugs detected (and corrected until the test suit tells that the program unit is free of bugs) is recorded. After performing all the individual experiments with each programmer and gather the results, the analysis will use the selected hypothesis testing technique (T-test or ANOVA) to draw a conclusion (retain or reject the hypothesis).

12) You want to specify the response time of a Web service you have developed for the Lisbon Stock Exchange, showing the average response time in milliseconds and the confidence interval. The Web service receives as input parameters the identifier of a company (VAT) and a date, and provides as output the total amount in euros for the buying and selling transactions of the papers of the company in the specified date. You have performed a set of experiments with a representative load (set of invocations of the service using companies from the PSI20 and a variety of dates) and measured the response time of the Web service. The Web service was used inside the same network where the servers of the Lisbon Stock Exchange are connected to, in order to avoid the unpredictable delays of remote networks. The results obtained are the following (in milliseconds):

- Average response time: 45.28
- Number of tests of the web service: 180
- Standard deviation: 8.23

a) Calculate the confidence interval with a confidence level of 95% and indicate the response time of the Web service in a technically correct and complete way.

Assuming that the samples are independent and follow the normal distribution, we basically need to apply the formula $\bar{x} \pm z * s/\sqrt{n}$, where $\bar{x}$ is the mean of the samples, z is the point in the X axis where the area under the standard normal curve corresponds to 95% (for a confidence level of 95%, as requested in the question), s is the standard deviation of the samples and n is the number of samples. The value of z can be found in the standard normal tables and, for a confidence level of 95% (i.e., $\alpha = 5\%$, two-tailed) is $z = 1.96$.
Doing the calculations:

$$45.28 \pm 1.96 * 8.28 /\sqrt{180} = 45.28 \pm 1.20$$

The response time of the Web service is 45.28 ± 1.20 milliseconds with 95% of confidence. That is, the response time is in the interval from 44.08 to 46.48 milliseconds, with a probability of 95%.

b) In the experiments performed you noticed that in some executions the response time of the Web service was considerably longer than the average. For example, in the first time the

web service is invoked with the VAT of a given company, the response time could be three or four times longer than the response time obtained in the subsequent invocations of the service for that company. Do you think these cases in which the response time was much longer should be considered as outliers and excluded from the calculation of the confidence interval or not? Justify your answer.

Considering the conditions mentioned in the question, the measurements longer than the average should not be considered outliers and ignored in the calculation. In this case, the reason for a longer response time is clearly identified and represents a known feature of the system. The example mentioned says that "in the first time the web service is invoked with the VAT of a given company, the response time could be three or four times longer than the response time obtained in the subsequent invocations of the service for that company", which means that the response time longer than the average is not caused by spurious or random causes, but it is caused by cache or buffering behavior of the system. As it is a stable and recurrent feature of the system, the values obtained in the first invocation should be considered in the calculation of the average response time.

13) In hypothesis testing there are two types of errors: type II errors (false negatives) and type I errors (false positive). Explain the differences between these two errors and compare the consequences of each one.

The type I error (false positive) is when you reject the null hypothesis but the null hypothesis is true and a type II error (false negative) is when you retain the null hypothesis but the null hypothesis is false.

Type II errors are equivalent to doing nothing. That is, instead of rejecting the null hypothesis (the correct decision) you retain it. For example, if you are optimizing the configuration of a server to improve response time, it means that you keep the current configuration, although the alternative configuration is better. In any case, you can do additional tests and improve the server configuration later on.

The consequences of type I error are much more severe, as rejecting a true null hypothesis means that you change something in the system under evaluation but that change is bad. In the example of tuning up the response time of a server, in a type I error you reject the current configuration (null hypothesis) and accept a new configuration, but this one is worse than the previous configuration.

14) Code inspections (a specific form of software inspections proposed by Michael Fagan more than four decades ago) is a technique to find bugs in the software under development in order to produce high quality software. Very briefly, in a code inspection a group of inspectors (experienced programmers) examine the code of a given code module (in general, short blocs of code with less than 100 lines of code) in order to find possible bugs. Obviously, the inspectors are not the authors of the code under inspection.

The goal is to perform an experiment to evaluate the effectiveness of code inspection in finding bugs. The idea is to use a set of code modules that have been very well tested and are considered totally correct (i.e., with no bugs) and insert a small number of bugs (3 or 4 for bugs per module). The inspectors will perform the code inspection on these modules and the goal is to check if they can detect the bugs that have been inserted in the modules.

In each individual code inspection, the inspector indicates the number of bugs found in the module. As the person that is managing the experiment knows exactly the bugs that have been inserted in each module, it is possible to confirm the quality of the code inspection done by each inspector. In practice, the result of each individual inspection is the number of bugs found and the number of mistakes done by the inspector. The mistakes are of two

types: bugs that are wrongly identified (i.e., the code is correct but the inspector thinks there is a bug; this is a false positive) and real bugs that are not identified by the inspector (i.e., he looked at the code but did not find anything wrong; this is a false negative). These are the results that will be used to evaluate the effectiveness of the code inspections.

Assume that the modules (blocks of code) available for the experiment are written in C, C++ and Java, and that the authors of the modules have inserted the bugs in the code, to assure that the bugs are realistic (i.e., not obvious bugs, very easy to find). Additionally, consider that the group of inspectors available for the experiment fill a questioner with some demographic data (age, sex, etc.) and state the number of years of experience as programmer/code inspector.

In these circumstances, and considering the problem statement of knowing if code inspections are effective in finding bugs, indicate the following elements of the experiment:

a) Hypothesis

There are several possibilities for the definition of hypothesis. The most obvious would be to set a given assumed goal for the performance of the code inspectors and test if the hypothesis can be rejected. For example:

H0 – The code inspectors can detect in average up to 70% of the bugs existing in the inspected code. That is, $\mu \leq 70\%$.

H1 – The code inspectors can detect more than to 70% of the bugs existing in the inspected code. That is, $\mu > 70\%$.

But many other hypotheses could be used in this experiment. For example:

H0 – The probability of bug detection is the same, no matter the programing language in which the inspected module is written (C, C++ or Java).

H1 – The probability of bug detection is depended on the programing language in which the inspected module is written.

b) Variables (dependent and independent) and levels used for the independent variables, justifying your choices

Dependent variables:
- Number of bugs detected
- Number of false positives
- Number of false negatives

The choice of the dependent variable or variables depends on the needs and goals of the experiment. You could just focus in one of these dependent variables or you could study all of them.

Independent variables:
- Programming language → levels: C, C++, Java
- Experience of the inspectors → novice, knowledgeable, expert

Many other possible independent variables could be considered. For examples, variables related to demographic aspects of the inspectors (age or sex), variables related to the duration of the inspection (to assess the impact of lack of concentration after some time), variables related to the metrics of the modules (lines of code, complexity, etc.).

15) The following summary table shows the results of the execution time of a package of benchmark programs compiled with four different compilers, named as A, B, C and D for

experiment purposes. Show whether the compilers used have any effect on the execution time of the compiled programs or not considering 95% of confidence.

Compiler	Number of runs	Exec. time mean	Exec. time standard deviation
A	8	4.35	2.25
B	11	3.79	1.63
C	10	4.91	1.89
D	9	6.33	2.21

1 - State the hypothesis:

H_0 – The compiler **is not relevant** for the execution time.

H_1 – The compiler **is relevant** for the execution time.

2 - Compute the test statistic using ANOVA:

Using the formulas for the sum of squares (see slides used in the lectures) and the elements of the one-way ANOVA table, and making the calculations, the ANOVA summary table is:

Formulas:

$$GM = \frac{\sum n_i \bar{x}_i}{\sum n_i} \qquad SS_{(W)} = \sum df\, s^2 \qquad SS_{(B)} = \sum n_A (\bar{X}_A - \bar{X}_G)^2$$

Table after calculations:

SOURCE	SS	df	MS	F
Between	34.030	3	11.343	2.895
Within	133.228	34	3.918	
TOTAL	167.258	37		

The critical value for α = 5% and $F_{3,34}$ (df_1 = 3 for the numerator and df_2 = 34 for the denominator) can be obtained from a F table. Looking at $F_{3,34}$ in a F table for α = 5%, we see that for df_1 = 3 the table does not show a line with df_2 = 34. But we can observe that the table shows df_2 = 30 and df_2 = 40 with the following values of F:

- $F_{3,30}$ → Fc = 2.9223
- $F_{3,40}$ → Fc = 2.8387

The critical value for α = 5% and $F_{3,34}$ is between these two values. Unfortunately, the value of F = 2.895 we obtained in a) is also inside this interval, so the only conclusion we can draw is that the value of F = 2.895 is very close to the critical value. We would need a more detailed F table to find the accurate critical value for $F_{3,34}$.

3 - Calculate the p value:

As you know, if the value of F obtained from the ANOVA table is greater than the critical value, it means that the probability p is lower than alpha (α = 5% in this exercise) and we should reject the null hypothesis. With the results obtained from the ANOVA table (and without access to a calculator that gives the accurate value of p for $F_{3,34}$) we would need to use the F table for α = 5% and the F table for α = 2.5% to obtain an approximate p value. It is worth noting that most of the F tables (including the ones available at Inforestudante) are organized to find the critical value in a F distribution and not to find the p value for a given value of F. In any case, there are tables that provide for each $F_{df1, df2}$ the values for different probabilities, which simplifies finding the value of p for a given $F_{df1, df2}$.

4 - Make a decision

In the written exam, only very few students provide a detailed characterization of the situation for this exercise, showing the problem of lack of detail in the F table to provide

conclusions in this borderline case. However, I considered as 100% correct a diverse range of answers, provided that it was clear the student fully understands the problem and the solution.

In order to provide a complete answer, using a more detailed table or an online calculator (e.g., https://www.easycalculation.com/statistics/f-test-p-value.php), the value of p for $F_{3,34}$ is 4.93%. This means that we should reject the null hypothesis. As we can see, the value is very close to 5%.

16) A Bank has developed a new database system with the goal of improving the speed of typical web banking transactions done by the Bank customers. There are several types of transactions (e.g., viewing account balances, funds transfers between the customer's accounts, viewing transactions in a given period indicated by the customer, etc.) The response time of each transaction depends on several factors related to both the queries issued by the customer and the load of the system. In order to assess the new database response, the Bank asked its Information Systems Department to perform an experiment to test the speed of the new database and compare the speed with the current database. The experiment includes a set of representative accounts with typical transactions profiles to assure that the measurements of the speed in both databases (the new one and the already existing) are meaningful.

a) State the different elements of the experiment (problem definition, variables, levels, hypothesis,…) and justify your choices.

Problem
Is the new database faster than the old one for a balanced mix of transactions?

Dependent variable
Number of transactions per minute.

Independent variables
- Database system → Levels: old systems; new system
- Transaction profiles → Levels: Profile 1 (e.g., viewing intensive, Profile 2 (e.g., transfer intensive), etc.
- Other independent variable not relevant for the study, given the way the problem is described: Server (machine), Operating System, DBMS tuning, etc.

Hypothesis
H_0 – The speed of the new database system is not higher than the old one.
H_1 – The new database system is faster.

b) In very few cases it was observed the response time of the transactions was several times larger than the average response time. Explain the choices you have to deal with this problem.

1 – Try to understand the source of these unexpected measurements and decide whether they should be considered as part of the nominal response, or should be considered an outlier and ignored.

2 – If they are really rare, assume (best guess) they are outliers; report them but ignore them in the data analysis.

17) A company has developed a new word-searching algorithm in files in smartphones and wants to test the speed of the new algorithm in three operating systems (Android, iOS, and Ubuntu Touch) and in two smartphones (named A and B for the experiment purposes). After a set of measurements, the data collected originated the following ANOVA table.

	SS	df	MS	F	P
Factor OS	23.333	2	11.6667	17.5	1.384e-06
Factor Smartphone	1.667	1	1.667	2.5	0.1197
Interaction	23.333	2	11.6667	17.5	1.384e-06
Within (error)	36.000	54	0.6667		

Based on this table, what conclusions can you take considering a significance level of 5%.

First, we need to make it clear the hypothesis under test. It must be something like this:

- **H_0**
  H_{0a} – The OS has no impact on the speed of the searching algorithm
  H_{0b} – The smartphone has no impact on the speed of the searching algorithm
  H_{0c} – There is no interaction between OSs and smartphone with impact on the speed of the searching algorithm
- **H_1** – Either the OS or the smartphone have impact on the speed of the searching algorithm or there is interactions

Based on the ANOVA table, we can conclude the following:

- The influence of the OS is highly significant (p is very small) → reject **H_0**
- The influence of the smartphone is not significant (p is 11.97%)
- The interaction is highly significant (p is very small) → reject **H_0**

  ➔ We should reject **H_0**

18) In the design of a large-scale experiment involving virtualized infrastructures such as the ones that support the Cloud you have identified a very large number of variables. You are aware of the fact that a large number of variables complicates the experiments and increases the whole experimental cost. Because of that, you are reanalysing the elements of your experiment to simplify (and to focus) the experiment as much as possible. In particular, should the following two variables (A and B) be considered in the experiment (as independent variables) or can you use just one of them? Justify your answer.

Variable A	Variable B
2.34	3.36
3.98	2.23
7.89	-2.98
12.13	-7.56
13.45	-7.98

The solution is to verify if the two variables are correlated. If they have a strong correlation then we can use only one of them as independent variable, as the conclusions we can extract from the experiments with one variable will be similar to the conclusion that can be drown with the other.

In order to verify if the two variables are correlated we can use the Pearson's test and calculate the Pearson's coefficient *r*.

The formula for the Pearson's coefficient is:

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2]\,[n\sum Y^2 - (\sum Y)^2]}}$$

In order to facilitate the calculation, we build the following table with the intermediate results:

	A (X)	B (Y)	XY	X^2	Y^2
	2.34	3.36	7.86	5.48	11.28
	3.98	2.23	8.88	15.84	4.97
	7.89	-2.98	-23.51	62.25	8.88
	12.13	-7.56	-91.70	147.14	57.15
	13.45	-7.98	-107.33	180.90	63.68
Sum	**39.79**	**-12.93**	**-205.81**	**411.61**	**145.96**

Doing the calculations, the result is r = - 0.995. The conclusion is that there is a strong negative correlation between the two variables.

However, in order to exclude one of the variables, we must test the hypothesis that A and B are correlated with a given significance level, for example 95%. The test statistic in this case is:

$$t = \frac{r}{\sqrt{\frac{1-(r)^2}{n-2}}}$$

1 – Hypothesis

H_0 – A and B are not correlated with a high level of significance (e.g., 95%).
H_1 – A and B are correlated

2 – Compute the test statistic

t = - 0.995/SQR((1- (- 0.995)^2)/(5-2)) = - 0.995/0,057662813 = -17.25548839

3 – Obtain p value

Obtain the p value for t = -17.255 for df = 5-2 = 3
→ we need the probability for t_3= -17.255

Searching the T table, we can see that p = 0.05% for t_3 = 12.92. This means that the probability for t_3= -17.255 is even lower.

4 – Make a decision

Since p is very low (lower than 0.05%), we can reject H_0
We can use only one variable, either A or B

19) Your company decided to try a training course meant to improve the productivity of code developers. The course is provided by an international software-consulting firm and is quite expensive, thus your company decided to run a pilot study with a group of 8 programmers randomly selected among the employees, to assess the effectiveness of the course before extending the course to the nearly 300 programmers of the company. In order to measure the productivity of code developers your company use "function points", which provide a more reliable and realistic measure of programmers productivity than the "classic lines of code" metrics. You have been assigned to the task of planning and managing the pilot study. In order to have a reference point you measure the productivity of the 8 programmers selected for the pilot study before and after the training course. To assure that the function point measurements are comparable, you used the specification of a set of benchmark programs especially design to assess code

development productivity. The following table shows the productivity (in function points per unit of time) of each programmer before and after the training.

Developer ID	1	2	3	4	5	6	7	8
Function points (before)	12	11	9	16	12	11	7	13
Function points (after)	11	15	12	14	16	12	11	9

Can you state with 95% of confidence that the training has improved the code development productivity of the programmers, so you can confidently recommend the extension of the course to all the programmers of your company? Justify your answer.

20) A company is deciding between two programming languages (named here as languages A and B) for a given project. Due to the nature of the project, the cost of the software maintenance is very important, as it is expected a long period of changes and adjustments in the code after the deployment of the first version of the product. The company decided to perform a preliminary experiment to evaluate whether there is a difference between the languages A and B concerning the time needed to introduce modifications in the program modules. Describe the following basic elements of the experiment:

c) Problem statement

d) Variables

e) Examples of levels for the different independent variables

f) Hypothesis

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

21) After a series of security incidents that compromised the web pages of the University of Conimbriga, campus of Condeixa, the administration of the university has decided to install a new intrusion detection system (IDS) and redefined in full the security procedures and configurations in the data centre and in the university network. In order to verify if the new IDS and security procedures represent an improvement concerning the previous scenario, the university decided to hire a Red Team from a specialized security company to simulate attacks (using penetration tests and multi-layered attack simulation) and compare the level of protection of the University of Conimbriga network and data infrastructure in the two scenarios (i.e., before the installation of the new IDS and new set of security policies and after the new IDS and new security policies). The report produced by the Red Team is quite comprehensive and detailed but the results can be summarized as follows:

- None of the 100 simulated attacks requiring "low effort" have succeeded in any of the two security configurations (i.e., the original scenario and the new IDS and security configuration).
- The simulated attacks requiring "very high effort" produced the following results:
    - Original scenario: 5 of the 75 simulated attacks compromised the University of Conimbriga network and data infrastructure.
    - New IDS and security configuration: 3 of the 79 simulated attacks compromised the University of Conimbriga network and data infrastructure.

Is it possible to state that the new IDS and security configuration has improved the security of University of Conimbriga network and data infrastructure with 95% of confidence? Explain your answer.

This is a case of hypothesis testing to compare two proportions.

1 – Hypothesis

H0 – The new scenario is not better than the previous one (p2 ≤ p1)

H1 – The new scenario is better than the previous one (p2 > p1)

2 – Compute the test statistic

The formula for the test statistic for the comparison of two proportions is:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p\,(1 - p)}{n_1 + n_2}}}$$

Where p is the pooled sample proportion:

$$p = \frac{p_1' \, n_1 + p_2' \, n_2}{n_1 + n_2}$$

After the calculations, we obtained z = 1.60

3 – Obtain p value

Using the Z table the value for p for one tailed test is 0.0548

4 – Make a decision

As p > 5% we cannot reject H0. That is, we cannot claim that the new scenario is better than the previous one with 95% of confidence.

Discuss this result. Consider the case where the purchase decision has already been done and the case where such decision has not been taken yet.

22) In an experiment, you suspected that two variables are correlated. You applied a Pearson's test to samples of the two variables and the result for the Pearson's coefficient was r = 0.989. Based on this result you concluded that there is a strong positive correlation between the two variables. But you are interested in knowing if this correlation does in fact imply causation.

a) Why is it important to know that correlation also implies causation? In another words, what can you do if you are sure that the high positive correlation also implies causation for the above-mentioned variables?

b) Explain the type of analysis you would perform to confirm whether the correlation also implies causation or not.

23) The University of Coimbra (UC) has decided to reshape the interface of Inforestudante and Infordocente to improve the general quality user interaction and improve the effectiveness of the some tasks (especially in the Infordocente side) that take a long time

to the user. After launching the versions of Inforestudante and Infordocente with the new interface the UC wants to assess whether the new interface is better than the previous one or not. Suppose you are in charge of performing such assessment; what type of experimental technique would you use: 1) Observation and data analysis, 2) Controlled experiment, or 3) Interviews and surveys? Explain your answer and if you consider that the situation would recommend more than one of theses techniques explain your point of view.

24) In experiments with people you can obtain the subjective participant's opinion using interviews and surveys but you can also assess the subjective participant's opinion when you perform controlled experiments with people. Explain the differences between these two situations.

25) What is the problem of treating two dependent samples as two independent samples in a hypothesis testing? Explain the situation clearly.

26) You are trying to improve the average execution time of a long aggregation query in a decision support database (normally know as data warehouses). These types of queries typically scan big portions of the main table of the data warehouse (called the fact table) and filter the registers using several satellite dimension tables. As the execution time of such queries can take minutes or even hours, all the improvements are very welcome. After generating a new set of indexes and materialized views (the typical "tricks" used to speed up such queries) you tried several variants of the query and you observed some reduction in the average execution time. However, when comparing the samples (i.e., execution time measurements) obtained before and after the introduction of the new indexes and materialized views, you concluded that you could not reject the null hypotheses that the average execution time is the same in both scenarios with 90% of confidence. In any case, based on your calculations, you would be able to reject the null hypothesis with 75% of confidence. In these circumstances, do you think you should leave the data warehouse as it was or should you consider the new set of indexes and materialized views as the next production version of the data warehouse?

27) 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.

28) 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.

29) 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

30) 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.

31) 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.

32) 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

33) 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.

34) 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.

35) Your company (a big multinational company) is installing a big datacenter to sell cloud services all over the world. The datacenter includes more than one hundred servers and has thousands of processors and tens of thousands of cores. The memory size needed is huge, including the main memory of the servers and memory to install state solid discs (SSD) in each node. In order to reduce the cost of memory, the company has decided to use the same memory board (a 4 GByte DRAM board)

for both the main memory and for the SSDs, having ordered many thousands of memory modules.

Currently, the increased scale and the constant reduction of the size of the bit cells of dynamic memory (DRAM) chips are causing manufacturing failures that are very difficult to detect by the quality control. The most noticeable example is the data-dependent retention failures that consist of memory cells that may show errors due to interference from the data patterns stored in physically adjacent cells, due to parasitic capacities that result from the manufacturing process. The quality tests at the end of the manufacturing lines do not guaranty that the chips are 100% free of this problem, as the time needed to test each chip with all the data pattern combinations would be months (or more). The memory manufacturers estimate by sampling the maximum number of bit cells susceptible to data-dependent retention failures in each GBit. For the case of the chips your company has ordered, the manufacturer states a maximum of 23 bit cells susceptible to data-dependent retention failures in each GBit.

Considering the huge amount of memory ordered, your company decided to test exhaustively a sample of 16 modules. After several months of test, the results were the following:

- Average number of *bit cells* susceptible to data-dependent retention failures in each GBit: 26
- Standard deviation: 7.53

Based on these results, do you think your company should accept the memory modules ordered or not, considering a confidence level of 95%? Your answer should indicate the hypothesis you are testing and should explain all steps to reach your conclusion.

This is a simple one-sample T test. Note that we do not know the standard deviation associated to the number of data-dependent retention failures per GBit, and the manufacturer does not provide confidence intervals for that number (i.e., we cannot use an ad hoc approach of just checking if the observations with the 16 modules are inside the confidence interval stated by the manufacturer). We just know that "the manufacturer states a maximum of 23 bit cells susceptible to data-dependent retention failures in each GBit".

As we cannot test all the memory chips (it would take many years), we want to know if the results obtained with the sample of 16 modules allow us to reject the hypothesis that the maximum number of bit cells susceptible to data-dependent retention failures per GBit is 23, as stated by the manufacturer.

We use the following steps (the pragmatic hypothesis testing approach proposed in slides 173-180):

1. State the hypothesis or claim to be tested
2. Compute the test statistic
3. Obtain p value
4. Make a decision

State the hypothesis:

H_0 – The maximum number of bit cells susceptible to data-dependent retention failures in each GBit is 23.

H_1 – The number of bit cells susceptible to data-dependent retention failures in each GBit is greater than 23.

Compute the test statistic:

Let us assume that only 1 GBit was tested in each memory module and not the complete 32 GBits (4 GBytes) available in each module. In other words, we consider 16 samples, which means a T distribution, considering that the data-dependent retention failures in each module are independent for the other modules. Computing the test statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{26 - 23}{\frac{7.53}{\sqrt{16}}} = 1.594$$

Obtain p value

Looking at a T table (one-tailed) for t = 1.594 and a degree of freedom df = 15 (df = n -1) we easily conclude that there is no such t value in the table. But looking at the line for df = 15, we can easily conclude that the probability p for t = 1.594 is in the interval 5% < p < 10%. In fact, looking at the table for the line representing df = 15, we find the following values:

- t = 1.753 for p = 5%
- t = 1.341 for p = 10%

As our value of t is 1.594, it is obvious that the probability p is between 5% and 10%. In fact, using an online calculator for t = 1.594 and df = 15, we obtain P = 6.59% for a one-tailed p value.

Make a decision

Considering a confidence level of 95%, as the value of p > 5%, we cannot reject the null hypothesis. This means that we cannot complain about the memory modules.