**MEI**

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

**Experiências de Laboratório:**

***Propósito***: Os experimentos de laboratório são conduzidos para testar hipóteses e estabelecer relações de causa e efeito num ambiente controlado.

***Controlo***: Os investigadores têm um elevado nível de controlo sobre as variáveis, e o ambiente é cuidadosamente manipulado para isolar fatores específicos de interesse.

***Participantes***: O tamanho da amostra é tipicamente maior, e os participantes são frequentemente atribuídos aleatoriamente a diferentes condições.

***Recolha de Dados***: Os dados são recolhidos em condições controladas, frequentemente utilizando medidas precisas e procedimentos padronizados.

***Exemplo***: Testar o impacto de um novo medicamento na pressão arterial, atribuindo aleatoriamente participantes para receber o medicamento ou um placebo num ambiente de laboratório controlado.

**Estudos Piloto:**

***Propósito***: Os estudos piloto são realizados como investigações preliminares em pequena escala para testar a viabilidade, métodos e design de um estudo maior.

***Controlo***: Há menos controlo em comparação com os experimentos de laboratório, e o foco está em refinar os métodos de pesquisa em vez de estabelecer relações causais.

***Participantes***: O tamanho da amostra é geralmente menor, e os participantes podem ser selecionados de uma população específica ou comunidade.

***Recolha de Dados***: Os métodos de recolha de dados são testados e ajustados com base nos resultados do estudo piloto para melhorar a eficiência e eficácia do estudo principal.

***Exemplo***: Antes de realizar um inquérito em grande escala sobre a opinião pública, os investigadores podem realizar um estudo piloto com um grupo menor para refinar as questões do inquérito e identificar potenciais problemas.

Em resumo, os experimentos de laboratório têm como objetivo estabelecer a causalidade através do rigoroso controlo e manipulação de variáveis, enquanto os estudos piloto são investigações preliminares que ajudam a refinar os métodos de pesquisa e garantir a viabilidade de estudos maiores.

Medição:

1. **What is the main difference between measuring and benchmarking?**

**Medição:**

***Definição:*** A medição envolve a avaliação da quantidade, dimensão, extensão ou qualidade de algo usando critérios específicos, normas ou instrumentos.

***Propósito***: O principal objetivo da medição é quantificar ou avaliar um atributo ou característica específica. Proporciona uma representação numérica ou qualitativa de uma variável ou propriedade específica.

***Foco***: A medição é frequentemente um processo independente focado na obtenção de dados precisos e exatos relacionados com um parâmetro específico.

**Benchmarking:**

***Definição***: O benchmarking é um processo sistemático de comparação dos processos, desempenho, produtos ou serviços de uma organização com os de líderes reconhecidos ou as melhores práticas da indústria.

***Propósito***: O principal objetivo do benchmarking é a melhoria. Ao comparar com benchmarks estabelecidos, as organizações procuram identificar áreas onde podem melhorar o seu desempenho, eficiência ou qualidade.

***Foco***: O benchmarking envolve não apenas medir o desempenho interno, mas também compará-lo com normas externas ou concorrentes para impulsionar a melhoria contínua.

**Principal Diferença:**

A diferença chave reside no foco e propósito. A medição está relacionada com a avaliação de atributos ou variáveis específicas para recolher dados, enquanto o benchmarking é um processo mais amplo de comparação do desempenho em relação a padrões externos ou melhores práticas para impulsionar a melhoria. A medição é um componente do benchmarking, pois o benchmarking frequentemente envolve a recolha de dados através de processos de medição para tomada de decisões informadas.

1. **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.**

Uma declaração de problema adequada (ou seja, relevante) deve ser suficientemente focada para permitir a identificação clara das variáveis do problema, mas, ao mesmo tempo, deve ser suficientemente aberta para permitir diferentes hipóteses na resposta ao problema/pergunta. Por exemplo:

*O tempo necessário para ordenar um determinado número de itens num array depende principalmente do número de itens a serem ordenados ou o tamanho dos itens também desempenha um papel importante?*

O contexto desta questão de pesquisa é o desempenho de algoritmos e programas de ordenação.

1. **For the example of problem statement provided in your answer to the previous question, indicate the following:**
   1. **Dependent variable(s)**
      1. Tempo de sorting
   2. **Independent variables**
      1. Tamanho do array
      2. Algoritmo de sort
      3. Tipo de cada elemento do array
      4. Línguagem de programação
   3. **Examples of possible levels for the independent variables in the experiments**
      1. Niveis de tamanho do Array : Baixo=[100], Médio=[10000],Grande=[1000000]
      2. Tamanho de cada elemento do array : 2, 4, 8, 16, 32, 64, 128, 252… etc
      3. Algoritmo de sort : bubble-sort, merge-sort, quick-sort
      4. Línguagem : C, C++, Java, Python
   4. **Hypothesis that could be tested (indicate if the hypothesis is directional or non-directional).**
      1. H0: O tamanho do array NÃO tem impacto no tempo de ordenação
      2. H1: O tamanho o array tem impacto no tempo de ordenação.
2. **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.**
   * 1. O objetivo dos testes de hipóteses é demonstrar que a hipótese nula não é verdadeira, com um determinado nível de confiança. Não é possível comprovar que a hipótese nula está correta; só pode comprovar que ela deve ser rejeitada.
3. **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.**
   1. **Relevância para a Questão de Pesquisa:** Os níveis devem ser escolhidos com base na sua relevância para a pergunta ou hipótese de pesquisa. Certifique-se de que os níveis selecionados abordem os aspectos específicos da variável independente que pretende investigar.
   2. **Amplitude e Variação:** Inclua uma gama de níveis que cubra adequadamente a variação presente na população. Isso permite uma análise mais abrangente da relação entre a variável independente e a variável dependente.
   3. **Variáveis Controladas:** Os níveis devem ser selecionados de forma a minimizar o impacto de variáveis confundidoras. Garanta que os níveis escolhidos ajudem a controlar ou considerar fatores extrínsecos potenciais que poderiam influenciar os resultados.
4. **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.**
   1. Os outliers podem ser causados pela variabilidade aleatória nas medições, por erros no design e implementação das experiências, por características da configuração experimental ou podem simplesmente indicar algo científica e tecnicamente interessante. O mais importante é determinar se os **outliers representam dados incorretos** (e, nesse caso, devem ser removidos da análise) ou **representam algo relevante** (embora incomum). O analista deve tentar identificar sistematicamente todas as possíveis causas de outliers que representam dados incorretos. Se os outliers não são causados por nenhum erro ou variabilidade nas medições, **então é justo** analisá-los, pois podem representar algum aspecto relevante.
5. **Explain what should be done to deal with the two types of measurement uncertainties: random and systematic uncertainties.**
   1. **Incertezas aleatórias:** são tratadas estatisticamente, utilizando médias, variância e calculando intervalos de confiança.
   2. **Incertezas sistemáticas:** devem ser compreendidas e eliminadas antes da análise.
6. **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?**

**H0**: O sistema de refrigeração **NÃO** garante que a temperatura nos racks esteja sempre dentro da faixa de 16,00 ± 0,80 graus Celsius, com uma confiança de 95%.

**H1**: O sistema de refrigeração garante que a temperatura nos racks esteja sempre dentro da faixa de 16,00 ± 0,80 graus Celsius, com uma confiança de 95%.

Logo podemos rejeitar a H0. Porque 0.44 < 0.80

1. **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.**

Com base nestes resultados, a conclusão é que o sistema de refrigeração NÃO está a operar de acordo com as especificações definidas no contrato, com 95% de confiança. Obviamente, o problema aqui é que, devido ao reduzido número de medições, o nível de confiança é menor. O fornecedor poderia argumentar que seria necessário pelo menos 30 medições para tomar uma decisão sólida.

1. **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).**

**Variaveis Dependentes** -> Numero de bugs detetados

**Variaveis Independentes** -> Número de horas de sono, Complexidade

**Níveis**:

Após a definição das variáveis, um passo importante é selecionar os níveis para as variáveis independentes, uma vez que esses níveis representam os pontos de ajuste dos experimentos. Os níveis podem ser:

Número médio de horas de sono (2, 4, 6, 8, 10 horas de sono por dia)

Complexidade das unidades do programa (baixa, média e alta complexidade)

As duas variáveis independentes mencionadas acima são as mais óbvias (e seriam suficientes para uma resposta correta). No entanto, existem outras variáveis independentes que poderiam ser mencionadas. Por exemplo:

Experiência do programador (iniciante, qualificado, especialista)

Linguagem de programação (Java, C++, Python)

Como sabemos, quanto maior o número de variáveis independentes, maior o custo (por exemplo, esforço necessário) do experimento.

H0 – O numero de bugs detetado nos programas NÃO depende na média de horas de sono dos programadores.

H1 – O numero de bugs detetado nos programas depende na média de horas de sono dos programadores.

Dado que o número de medições/amostras em experiências com pessoas é normalmente limitado, a técnica apropriada de teste de hipóteses seria o teste T de duas amostras, aplicado sucessivamente às medições obtidas com os diferentes níveis das variáveis independentes, ou ANOVA (ANOVA de um fator ou mesmo ANOVA de dois fatores, dependendo do número de variáveis independentes e níveis usados no experimento). Neste caso, estamos a tratar as duas amostras como amostras independentes. No entanto, o experimento seria muito melhor se pudesse ser utilizado um conjunto de duas amostras dependentes. Neste caso, a medição seria feita considerando o mesmo programador e diferentes tempos de sono.

É de salientar que, embora os dados brutos tenham uma natureza binomial (erro vs sem erro), não temos medições baseadas em proporções. Ou seja, os dados são considerados como dados numéricos (contagens de erros) e não proporções, sendo assim melhor utilizar o teste T em vez do teste de proporções.

Considerando o cenário simples com apenas duas variáveis independentes, a configuração do experimento consiste em um grupo de programadores que desenvolvem um conjunto de unidades de programa de diferentes níveis de complexidade. Cada programador desenvolve todas as unidades de programa envolvidas no experimento (incluindo unidades de baixa, média e alta complexidade), abrangendo todas as possibilidades para os níveis definidos para as horas de sono (pelo menos uma unidade de programa de cada nível de complexidade deve ser desenvolvida em cada nível de horas de sono). Para ter controle total sobre o experimento, as horas de sono dos programadores precisam ser controladas usando um simples despertador, a fim de atingir os níveis definidos para a variável "média de horas de sono". Quando um programador declara que uma determinada unidade de programa está concluída e pronta para teste, o programa é testado usando o conjunto de testes (mencionado na pergunta). O número de bugs detectados (e corrigidos até que o conjunto de testes indique que a unidade de programa está livre de bugs) é registrado. Após realizar todos os experimentos individuais com cada programador e reunir os resultados, a análise utilizará a técnica de teste de hipóteses selecionada (T-test ou ANOVA) para chegar a uma conclusão (manter ou rejeitar a hipótese).

1. 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

1. 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.

**Resposta:**

Portanto a resposta seria que o tempo de resposta do serviço é de 45.28+- 1.20 milliseconds com 95% de confiança. Dentro do intervalo de 44.08 a 46.48 millisegundos

1. 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.

Resposta : Dadas as condições mencionadas na questão, as medições mais longas do que a média não devem ser consideradas como valores atípicos e ignoradas no cálculo. Neste caso, a razão para um tempo de resposta mais longo é claramente identificada e representa uma característica conhecida do sistema. O exemplo mencionado afirma que 'na primeira vez que o serviço da web é invocado com o NIF de uma determinada empresa, o tempo de resposta pode ser três ou quatro vezes mais longo do que o tempo de resposta obtido nas invocações subsequentes do serviço para essa empresa', o que significa que o tempo de resposta mais longo do que a média não é causado por causas espúrias ou aleatórias, mas sim pelo comportamento de cache ou buffering do sistema. Como é uma característica estável e recorrente do sistema, os valores obtidos na primeira invocação devem ser considerados no cálculo do tempo médio de resposta

1. 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.

Resposta: O erro do Tipo I (falso positivo) ocorre quando se rejeita a hipótese nula, mas a hipótese nula é verdadeira, e o erro do Tipo II (falso negativo) ocorre quando se mantém a hipótese nula, mas a hipótese nula é falsa.

Erros do Tipo II são equivalentes a não tomar nenhuma ação. Ou seja, em vez de rejeitar a hipótese nula (a decisão correta), você a mantém. Por exemplo, ao otimizar a configuração de um servidor para melhorar o tempo de resposta, significa que você mantém a configuração atual, embora a configuração alternativa seja melhor. De qualquer forma, é possível realizar testes adicionais e aprimorar a configuração do servidor posteriormente.

As consequências do erro do Tipo I são muito mais graves, pois rejeitar uma verdadeira hipótese nula significa que você faz uma alteração no sistema em avaliação, mas essa mudança é prejudicial. No exemplo de ajustar o tempo de resposta de um servidor, em um erro do Tipo I, você rejeita a configuração atual (hipótese nula) e aceita uma nova configuração, mas esta é pior do que a configuração anterior.

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.

Resposta:

H0: O compilador não é relevante para o tempo de execução

H1: O compilador é relevante para o tempo de execução

1º Passo: Calcular o GM(Grande média)

2º Passo: calcular o SS(B)

3ºPasso: Calcular o SS(w)

4ºPasso

Agora total é só somar 34.030 + 133.228 = 167.258

5º Passo

Sabemos que para total df = n – 1 é nº de runs totais -1 logo:

6º Passo

K = ao número de compiladores ( Linhas da tabela ) logo K = 4 e

Então k-1 = 4-1 = 3

7º Passo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SOURCE** | **SS** | **Df** | **MS** | **F** |
| Between | 34.030 | **3** | 11.343 | 2.895 |
| Within | 133,228 | **34** | 3,918 |  |
| Total | 167.258 | 37 |  |  |

8º Passo

Sabendo o valor de F(3,34), vamos ver à tabela e não conseguimos saber exatamente o valor de F(3,34) mas sabemos que este valor encontra-se entre df1=3, df2=30 <-> df2=40