**COORDENADORIA DO CURSO DE CIÊNCIA DA COMPUTAÇÃO**

**ESTATÍSTICA E PROBABILIDADE**

**Prof. Peter de Matos Campos**

**ROTEIRO DE ESTUDOS 1**

Prazo máximo para entrega: **15/09**

**Ao final deste roteiro de estudos você deverá ser capaz de:**

* Diferenciar modelo determinístico de modelo probabilístico.
* Caracterizar as diferenças entre população e amostra.
* Estabelecer diferenças entre as variáveis categóricas e numéricas.
* Construir e interpretar tabelas de distribuição de frequências e gráficos.
* Calcular e interpretar medidas de posição e tendência central de um conjunto de dados.
* Medir a variabilidade de um conjunto de dados por meio das medidas de dispersão.
* Utilizar medidas de assimetria e de curtose para caracterizar a forma da distribuição dos elementos da população amostrados em torno da média.
* Calcular medidas-resumo por meio do RStudio.

**Atividades preparatórias**

* **Fazer a leitura do texto de referência**: Apostila, págs.1-19.

**Problema 1**: **Quem fala mais: o homem ou a mulher?** O assunto é polêmico, a fama de tagarela persegue o sexo feminino. Mas, a mulher realmente fala mais do que o homem? As pesquisas disponíveis divergem nos resultados e, muitas, tem metodologias questionáveis. Mas uma das pesquisas mais citadas foi realizada pela Universidade de Maryland e os resultados apontam que as mulheres falam, em média, 20 mil palavras por dia, enquanto os homens, apenas 7 mil (Fonte: Triola, 2014).

**Atividades preparatórias**

* **Fazer a leitura do texto de referência**: Apostila, págs.20-22.

**Problema 2**: Sem dúvida a desigualdade social é um assunto que chama a atenção de todos nós. Mas como medir a desigualdade social? A Rede Nossa São Paulo (RNSP), organização da sociedade civil que luta contra a desigualdade na cidade de São Paulo publica anualmente o Mapa da Desigualdade. Interessados em quantificar as desigualdades regionais a RNSP mede 53 características nas 90 regiões existentes na cidade de São Paulo. Vamos trabalhar com uma dela: a proporção de matrículas no ensino básico em escolas públicas, no total de matrículas (%), de 20 distritos (2 regiões). **Qual será o nível de desigualdade que encontraremos?**

**Atividades preparatórias**

* **Fazer a leitura do texto de referência**: Fávero, L.P.; Belfiori, P., págs.63-69.

**Problema 3**: **Qual a forma da distribuição de frequências para o consumo de álcool?**

Um dos interesses dos pesquisadores é definir um modelo para fazer futuras previsões sobre um evento. Pensando nisto, queremos saber qual a forma da distribuição para a quantidade de álcool ingerida por dia (em ml) por 41 pacientes entrevistados no pré-tratamento, do programa de atendimento do Núcleo de Ensino e Pesquisa sobre Alcoolistas (NEPA) do Hospital Psiquiátrico da Fundação Hospitalar do Estado de Minas Gerais (FHEMIG), em Barbacena. Os dados estão no arquivo “alcool.txt”, disponível no Portal Didático.

**EXERCÍCIOS**

* **Introdução**

1. Escreva com suas palavras, por que precisamos estudar Estatística no curso de Ciência da Computação?

**Resposta:** Estudar Estatística no curso de Ciência da Computação é importante porque ajuda a lidar com dados, algo essencial na área. A Estatística permite tomar decisões informadas ao analisar dados, o que é crucial para projetos de software e algoritmos. Resumindo, a Estatística é fundamental para compreender e utilizar dados eficazmente na Ciência da Computação.

1. Para cada situação descrita a seguir marque D se o modelo que explicaria o fenômeno em estudo for determinístico ou P caso o modelo seja classificado como probabilístico ou não-determinístico.

( P ) Avaliação de um serviço de assistência técnica.

( D ) (reação química entre ácido muriático e soda cáustica).

( P ) Resultado de uma partida de futebol.

( P ) Resultado de pesquisas eleitorais para presidente do Brasil.

( D ) Tempo gasto para uma esfera de aço tocar o chão, quando solta de uma altura h, no vácuo.

( D ) Avaliação entre algoritmos num contexto de aprendizado de máquina.

( D ) (2ª lei de Newton).

( P ) Avaliação de um experimento computacional.

( D ) O Volume do cilindro é

( P ) Previsão do número de acessos a uma página específica.

1. Qual a dinâmica da Estatística e como representar as características e os dados de um fenômeno de interesse?

**Informações contidas na amostra.**

**X, Y, Z, T = Características de interesse (Variáveis).**

Valores observados na amostra

**X, Y, Z, T = Características de interesse (Variáveis).**

Valores observados de uma população

**População (N)**

**Amostra (n)**

**Estimadores**

**Amostragem/Experimentação**

**Inferência/Probabilidade**

**Parâmetros**

**Conclusões sobre as características da população**

**Figura 1.1** – Dinâmica da Estatística (Adaptado de Oliveira e Oliveira, 2018).

**Obs**: Note que, o estudo de uma população é feito sobre as características que o pesquisador, ou quem estiver fazendo a análise de dados, define como sendo importantes, isto é, sobre as **variáveis**. Essas variáveis, por facilidade, **são representadas por qualquer letra latina maiúsculas**. Exemplo: o gênero de um grupo de pessoas pode ser representado, por exemplo pelas letras **X, Y, Z, W, H** (letra latina maiúscula). Já os valores observados dessas variáveis são representados por uma das letras latinas minúsculas, isto é, x, y, z, w, h, respectivamente.

* **Usando a notação sigma**
* **O que simboliza (sigma maiúscula)?** A letra grega significa soma.
* **Por que utilizar a letra ?** Sua utilização tem por objetivo simplificar a notação de uma soma com muitos termos, sujeitos ou observações.

**Exemplo**: considere uma amostra da variável com observações, isto é, . Para resumir a escrita da soma de seus elementos, define-se:

* **Índice do somatório** : indica onde começa a soma (número abaixo do Σ) e onde termina a soma (número acima de Σ). **Obs**: qualquer letra pode ser usada para denotar o índice, mas são as mais comuns.

1. Para as situações descritas a seguir, identifique a população, a amostra e a característica de interesse. E tente julgar se a amostra coletada representa a população desejada para se cumprir o objetivo da pesquisa.
2. Uma amostra de sangue foi retirada de um paciente com suspeita de anemia. Para diagnosticar a anemia é necessário fazer um exame de sangue para avaliar a quantidade de glóbulos vermelhos e hemoglobina, sendo normalmente indicativo de anemia quando os valores de hemoglobina estão abaixo de 12 g/dL no caso das mulheres e 14 g/dL no caso dos homens.

**Resposta:** População: A população neste caso seria todos os pacientes com suspeita de anemia.

Amostra: A amostra é a quantidade de sangue retirada do paciente específico com suspeita de anemia.

Característica de Interesse: A característica de interesse é a quantidade de glóbulos vermelhos e hemoglobina no sangue do paciente, que é usada para diagnosticar a anemia.

Julgamento: A amostra coletada (sangue do paciente com suspeita de anemia) representa a população desejada. Isso ocorre porque a amostra foi retirada de um paciente que já é parte da população de interesse, ou seja, aqueles com suspeita de anemia. O exame de sangue é o método padrão para diagnosticar a anemia, e a amostra (sangue) é apropriada para cumprir o objetivo da pesquisa, que é determinar se o paciente tem anemia com base nos níveis de glóbulos vermelhos e hemoglobina no sangue.

1. O Diretório Central dos Estudantes da UFSJ conduziu um estudo sobre o abuso do uso do álcool entre os estudantes. Cem dos cerca de 11.000 membros do corpo discente foram amostrados e foi solicitado que respondessem a um questionário. Uma pergunta feita foi: “Em quantos dias da semana passada você consumiu pelo menos um drinque com álcool?”.

**Resposta:** População: A população neste caso é composta por todos os 11.000 estudantes da UFSJ (Universidade Federal de São João del-Rei).

Amostra: A amostra consiste em 100 membros aleatórios do corpo discente da UFSJ.

Característica de Interesse: A característica de interesse é o consumo de álcool entre os estudantes da instituição UFSJ.

Julgamento: A amostra coletada, que consiste em 100 alunos selecionados aleatoriamente de um total de 11.000, representa adequadamente a população desejada, que são os alunos da instituição. Isso sugere que a pesquisa pode fornecer informações relevantes sobre o consumo de álcool entre os estudantes da UFSJ com base nas respostas dos 100 alunos amostrados.

1. Para avaliar a opinião dos alunos da UFSJ sobre como tem sido a condução feita pelo presidente do Brasil para controlar a Pandemia de Covid-19, 10 alunos do 2º período de Ciência da Computação da UFSJ foram entrevistados.

**Resposta:** População: A população neste caso é composta por todos os alunos da UFSJ.

Amostra: A amostra consiste em 10 alunos do 2º período de Ciência da Computação da UFSJ.

Característica de Interesse: A característica de interesse é a opinião dos alunos da UFSJ sobre como tem sido a condução feita pelo presidente do Brasil para controlar a Pandemia de Covid-19.

Julgamento: Neste caso, a amostra coletada, que consiste em 10 alunos do 2º período de Ciência da Computação da UFSJ, pode não representar adequadamente a população desejada para cumprir o objetivo da pesquisa. Isso ocorre porque a opinião dos alunos do 2º período de Ciência da Computação pode não ser representativa da opinião de todos os alunos da UFSJ. Além disso, a amostra pode ser tendenciosa em direção a uma área específica de estudo (Ciência da Computação). Para obter uma visão mais abrangente da opinião de todos os alunos da UFSJ sobre o assunto, seria necessário utilizar uma amostra mais diversificada que incluísse estudantes de diferentes cursos e períodos.

1. Principais tipos de Variáveis: complete os espaços.

**VARIÁVEL**

**Característica.**

**Atributo.**

**No R:**  .

**Quantitativa/Numérica.**

Expressa em .

Variável **métrica**.

**Discreta.**

**Finitos** em um intervalo.

**No R:**\_\_\_\_\_\_\_\_\_\_\_ .

Números de casos/dia.

Compartilhamentos/min.

**Contínua.**

**Infinitos** valores em

.

**No R:**\_\_\_\_\_\_\_\_\_\_\_ .

Tempo de deslocamento.

Rendimento/mês.

**Qualitativa/Categórica.**

Expressa em .

Variável **Categórica**.

**Nominal.**

**Sem** \_\_\_\_\_\_\_\_\_\_ .

**Rótulo**.

**No R:**\_\_\_\_\_\_\_\_\_\_\_ .

Números de casos/dia.

Compartilhamentos/min.

**Ordinal.**

**Com** \_\_\_\_\_\_\_\_\_\_ .

**Grau** ou **Nível.**

**No R:**\_\_\_\_\_\_\_\_\_\_ .

Grau de satisfação serviço.

Grau de Jiu-Jitsu.

**Figura 1.2** – Principais tipos de variáveis ou tipos primários (Zeviani, 2021).

1. Classifique cada uma das seguintes variáveis em categórica nominal (CN), categórica ordinal (CO), numérica discreta (ND) e numérica contínua (NC).

( ) Tipo de transporte utilizado para ir à universidade.

( ) Idade, em anos, dos alunos de uma determinada universidade.

( ) Gênero (Masculino e Feminino).

( ) Sua altura.

( ) Nível de concordância com uma determinada atitude ou política institucional.

(Concordo Totalmente, Concordo, Não Concordo Nem Discordo, Discordo, Discordo Totalmente).

( ) Número de cliques em uma página específica.

( ) Nível de instrução (Fundamental, Médio e Superior).

( ) Linguagens/softwares para análise de dados.

( ) Tempo de processamento de um algoritmo.

* **Amostragem probabilística**

1. Admita que o número de alunos de graduação da UFSJ seja de 8.500 estudantes e que se deseja estudar a percepção destes estudantes sobre o período remoto. Considerando que esta população seja homogênea e que esteja ordenada, utilizando o Software Estatístico R, crie uma função que selecione uma amostra sistemática de tamanho 50 para estudar as características definidas no estudo.
2. Considere, somente para fins deste exercício, que a UFSJ tenha os seguintes cursos:

|  |  |
| --- | --- |
| **Curso** | **Nº de alunos** |
| Administração | 1.500 |
| Arquitetura | 650 |
| Ciência da Computação | 800 |
| Jornalismo | 550 |
| Economia | 750 |
| **Total** | **4250** |

Utilizando o Software Estatístico R, crie uma função que selecione uma amostra estratificada proporcional de tamanho 50 para estudar a percepção destes estudantes sobre o período remoto da UFSJ.

* **Estudo descritivo**

1. O Exame Nacional de Desempenho dos Estudantes (ENADE) é um dos pilares da avaliação do Sistema Nacional de Avaliação da Educação Superior (SINAES), criado pela Lei nº 10.861, de 14 de abril de 2004. Além do ENADE, os processos de Avaliação de Cursos de Graduação e de Avaliação Institucional constituem o ‘tripé’ avaliativo do SINAES; os resultados desses instrumentos avaliativos, reunidos, permitem conhecer, em profundidade, o modo de funcionamento e a qualidade dos cursos e Instituições de Educação Superior (IES) de todo o Brasil.

Recentemente o INEP apresentou o Relatório com os resultados do ENADE/2018 para um determinado Curso da UFSJ, cuja nota foi 5. As informações constantes deste relatório traduzem os resultados obtidos a partir da análise dos dados dos estudantes desse curso. A prova foi resolvida por 88 estudantes concluintes, teve duração total de 4 (quatro) horas, apresentou um componente de avaliação da Formação Geral, comum aos cursos de todas as áreas, e um componente específico de cada Área. Os dados de algumas variáveis de uma amostra de 30 alunos concluintes de um curso da UFSJ são apresentados pela Tabela 1.1.

**Tabela 1.1** – Nota obtida, percepção da dificuldade na prova, tipo de escola onde cursou o ensino

médio e faixa de renda da família de uma amostra de 30 alunos concluintes de

um determinado curso da UFSJ que fizeram o ENADE/2018.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **n** | **Nota** | **Percepção** | **Ensino médio** | **Renda** |
| **1** | 40,5 | Muito difícil | Pública | < 3 |
| **2** | 42,7 | Muito difícil | Pública | < 3 |
| **3** | 46,3 | Difícil | Pública | entre 3-6 |
| **4** | 51,2 | Difícil | Pública | < 3 |
| **5** | 59,8 | Difícil | Particular | entre 3-6 |
| **6** | 54,4 | Difícil | Pública | < 3 |
| **7** | 61,5 | Médio | Pública | entre 3-6 |
| **8** | 62,6 | Médio | Pública | < 3 |
| **9** | 64,7 | Médio | Particular | entre 3-6 |
| **10** | 66,1 | Médio | Particular | entre 3-6 |
| **11** | 41,3 | Muito difícil | Pública | < 3 |
| **12** | 43,2 | Muito difícil | Pública | < 3 |
| **13** | 45,1 | Difícil | Pública | entre 3-6 |
| **14** | 52,1 | Difícil | Pública | < 3 |
| **15** | 53,2 | Difícil | Particular | entre 3-6 |
| **16** | 60,1 | Difícil | Pública | < 3 |
| **17** | 60,4 | Médio | Pública | entre 3-6 |
| **18** | 61,2 | Médio | Pública | < 3 |
| **19** | 78,5 | Fácil | Particular | entre 3-6 |
| **20** | 62,9 | Médio | Particular | entre 3-6 |
| **21** | 63,8 | Médio | Pública | < 3 |
| **22** | 62,1 | Médio | Pública | < 3 |
| **23** | 65,6 | Médio | Pública | entre 6-10 |
| **24** | 64,8 | Médio | Pública | entre 3-6 |
| **25** | 67,0 | Fácil | Particular | entre 6-10 |
| **26** | 71,5 | Fácil | Particular | entre 3-6 |
| **27** | 73,1 | Fácil | Particular | >10 |
| **28** | 75,2 | Fácil | Particular | < 3 |
| **29** | 80,8 | Muito fácil | Particular | >10 |
| **30** | 90,5 | Muito fácil | Particular | entre 6-10 |

**Fonte:** adaptado INEP. Brasil, 2019.

Para as variáveis **faixa de renda** e **percepção** faça um estudo descritivo e interprete os resultados obtidos.

1. Planejou-se um experimento educacional para a comparação do estudo cooperativo *versus* o tradicional estilo de aulas. Usaram-se duas seções de uma turma de estatística introdutória, e sete alunos foram selecionados aleatoriamente de cada uma delas. Os dados do segundo teste (um teste com 29 itens) são apresentados a seguir.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tradicional** | 21 | 28 | 25 | 25 | 21 | 19 | 23 |
| **Cooperativo** | 25 | 27 | 28 | 25 | 24 | 24 | 29 |

Qual grupo de estudantes se saiu melhor? Justifique sua resposta.

1. Uma empresa quer ampliar suas vendas on-line e, para isso, vai investir na página de sua loja virtual. Como este é um projeto de custo elevado, duas páginas foram selecionadas e o tempo que 15 usuários passam navegando nelas foi anotado.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pagina A** | 206 | 141 | 50 | 132 | 91 | 71 | 118 | 115 | 85 | 146 | 79 | 227 | 101 | 107 | 143 |
| **Pagina B** | 101 | 154 | 128 | 178 | 214 | 186 | 222 | 257 | 282 | 93 | 81 | 101 | 218 | 227 | 260 |

1. Calcule “na unha” a média, a mediana, a moda e os quartis para a página A. Interprete os resultados. Crie uma função no R para fazer os mesmos cálculos para página B, compare os resultados.
2. Trace o Boxplot para as vendas obtidas pela página A. Utilize a função boxplot( ) do R para comparar as duas páginas. Qual das páginas você recomendaria para a direção da empresa? Justifique sua resposta.
3. No Censo 2010, realizado pelo IBGE, mais de 190 mil recenseadores visitaram 67,6 milhões de domicílios nos 5.565 municípios brasileiros levantando-se inúmeras características sobre a vida cotidiana da população. Estas informações têm sido úteis para gestores e pesquisadores. Interessados em estudar o comportamento feminino no que diz respeito ao momento de ter filhos, pesquisadores selecionaram uma amostra de 30 mulheres e anotaram o nível de escolaridade, o número de filhos e a idade que cada uma delas teve o primeiro filho. Uma das questões que se desejava responder era: **A escolaridade é um dos condicionantes do comportamento da fecundidade feminina?**

Por facilidade a variável Nível de Escolaridade foi codificada da seguinte maneira: (1) Ensino fundamental incompleto; (2) Ensino fundamental completo e ensino médio incompleto; (3) Ensino médio completo e ensino superior incompleto e (4) Ensino superior completo.

Os dados são apresentados na Tabela 1.2.

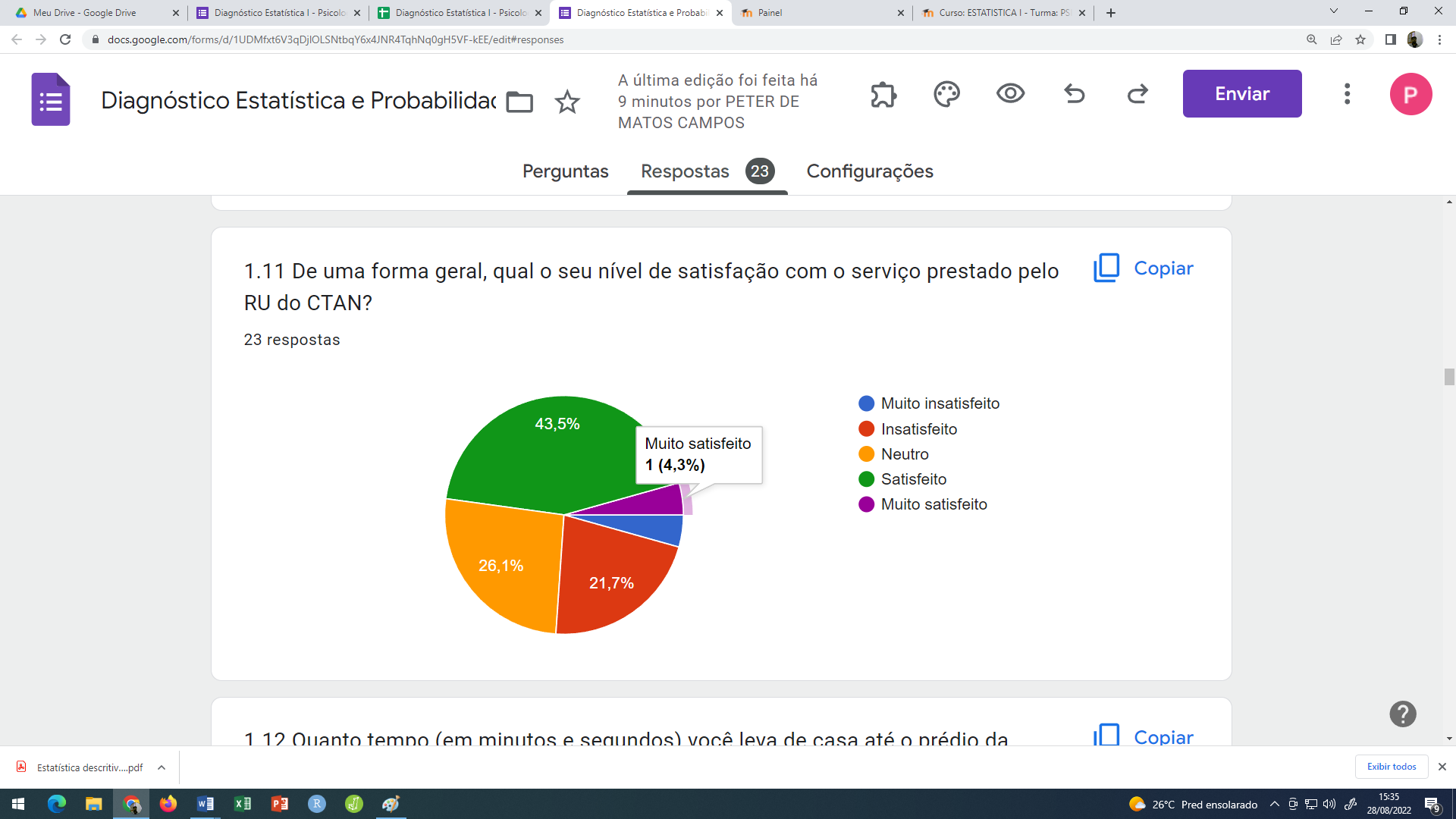
**Tabela 1.2** - Nível de escolaridade, número de filhos e idade da primeira gravidez

de uma amostra de 30 mulheres em idade fértil – Dados retirados do Censo 2010.

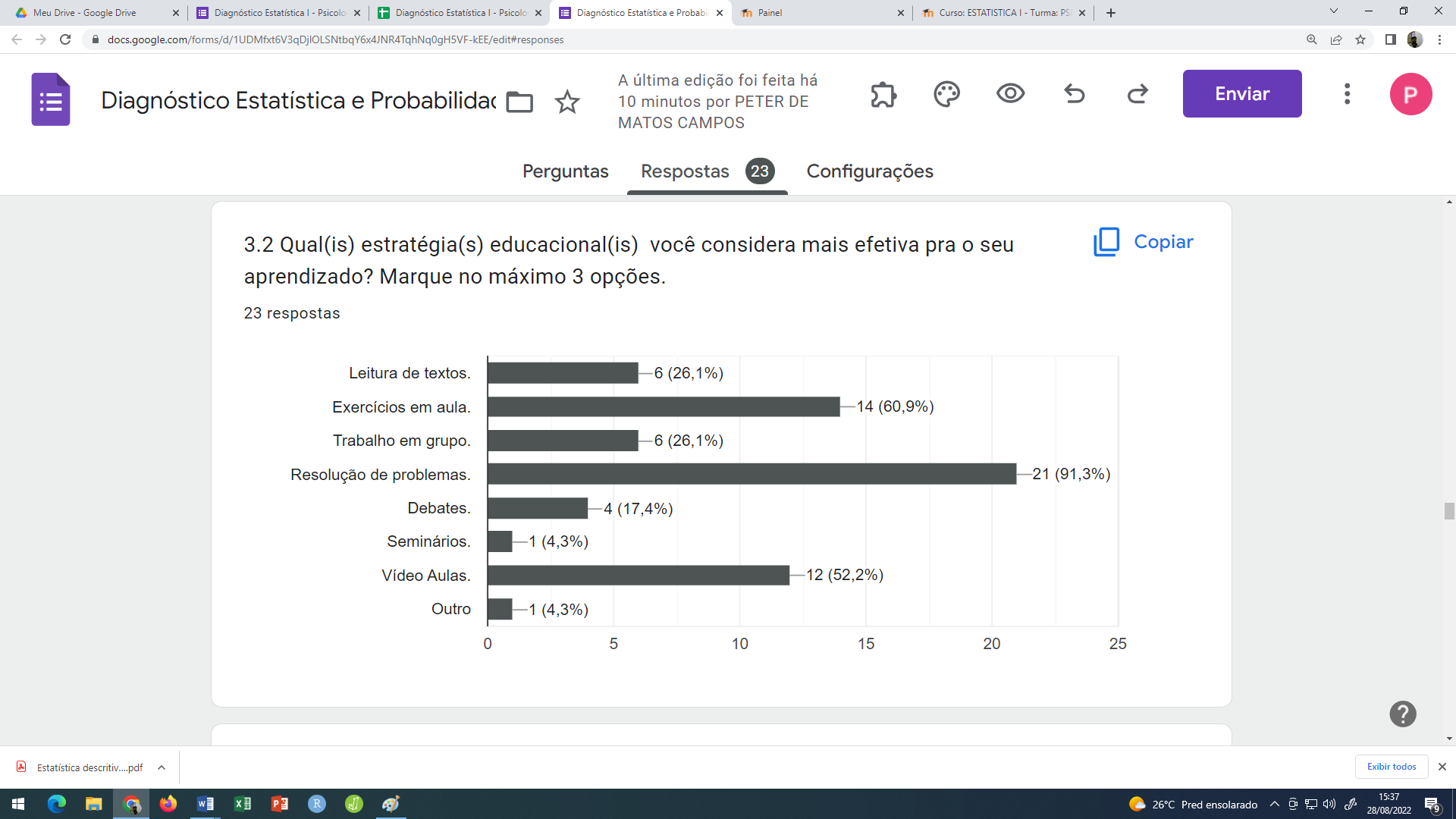
|  |  |  |  |
| --- | --- | --- | --- |
| **N** | **Escolaridade** | **Nº de Filhos** | **Idade 1º Filho** |
| 1 | 1 | 4 | 15 |
| 2 | 1 | 5 | 13 |
| 3 | 1 | 3 | 16 |
| 4 | 3 | 2 | 23 |
| 5 | 1 | 4 | 18 |
| 6 | 4 | 1 | 27 |
| 7 | 1 | 3 | 20 |
| 8 | 3 | 2 | 25 |
| 9 | 2 | 2 | 19 |
| 10 | 4 | 1 | 28 |
| 11 | 1 | 5 | 19 |
| 12 | 3 | 1 | 22 |
| 13 | 1 | 6 | 17 |
| 14 | 4 | 1 | 25 |
| 15 | 1 | 3 | 22 |
| 16 | 1 | 4 | 23 |
| 17 | 3 | 2 | 26 |
| 18 | 2 | 3 | 19 |
| 19 | 1 | 5 | 17 |
| 20 | 1 | 4 | 16 |
| 21 | 4 | 2 | 32 |
| 22 | 3 | 2 | 24 |
| 23 | 1 | 3 | 20 |
| 24 | 1 | 4 | 18 |
| 25 | 3 | 2 | 23 |
| 26 | 2 | 3 | 21 |
| 27 | 1 | 3 | 17 |
| 28 | 2 | 2 | 18 |
| 29 | 3 | 1 | 21 |
| 30 | 1 | 5 | 19 |

**Fonte**: IBGE, Censo Demográfico 2010.

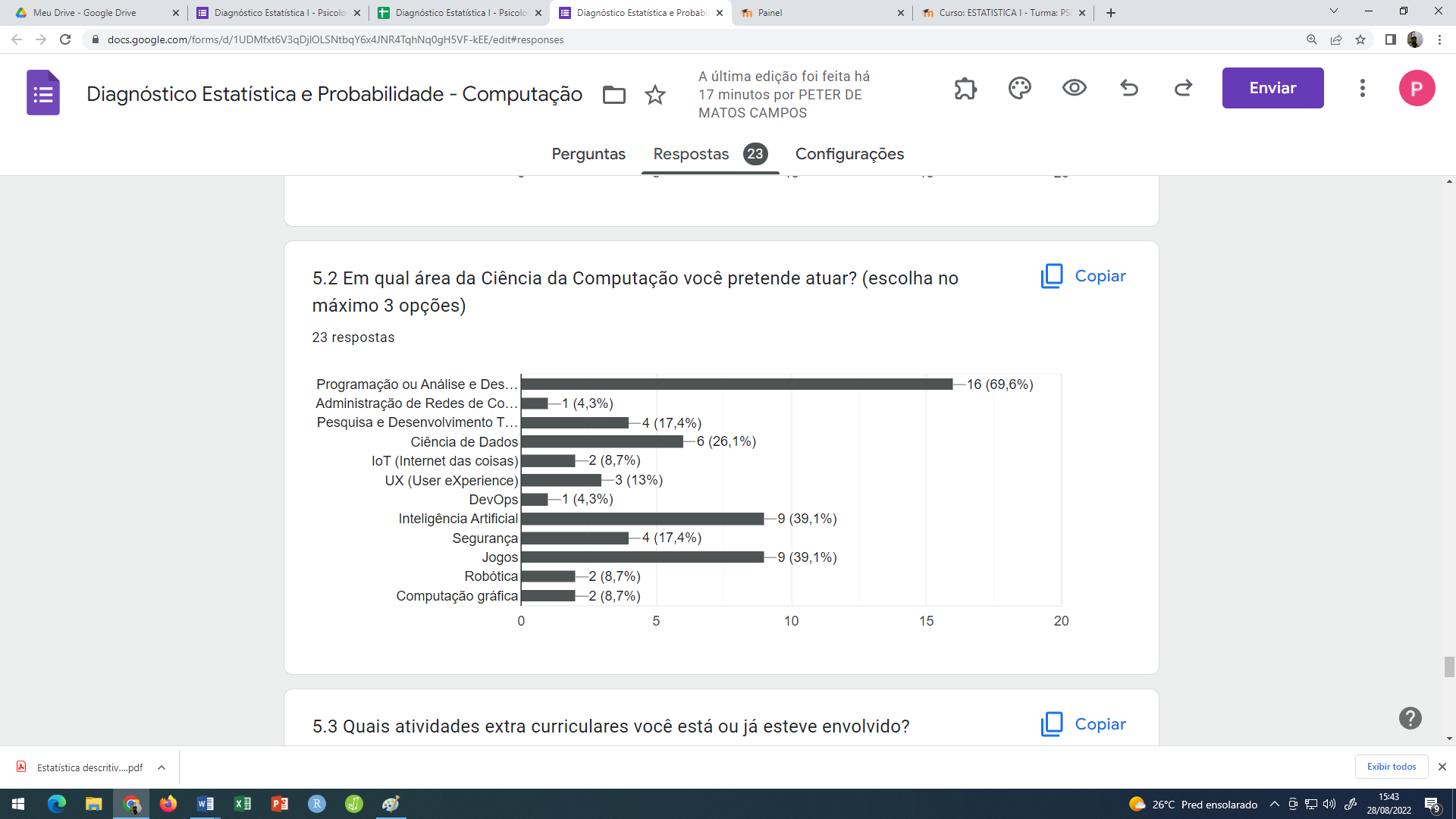
1. Para a variável **número de filhos** faça “na unha” um estudo descritivo e, baseado nos resultados obtidos podemos dizer que a escolaridade é um dos condicionantes do comportamento da fecundidade feminina? A qual conclusão podemos chegar?
2. Utilize o R para fazer um estudo descritivo sobre a variável **idade do primeiro filho**, interprete os resultados.
3. Utilize o R, faça uma análise conjunta entre as variáveis **Escolaridade** e **Número de filhos**, calcule as frequências absolutas conjuntas e as frequências percentuais conjuntas. Trace o gráfico de barras conjuntas para o caso, interpretando os resultados.
4. Analise os gráficos a seguir. Os são gráficos adequados? Caso seja, interprete-o.

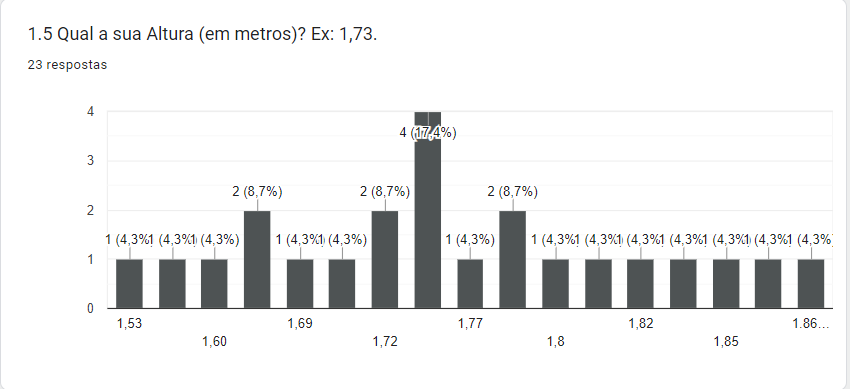
a)

b)



c)



d)

1. Os dados abaixo referem-se à taxa de defeitos (Ponto de Função) descobertos em 50 projetos de softwares.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0,711 | 0,595 | 0,794 | 0,886 | 0,876 | 0,880 | 0,909 | 0,594 | 0,895 | 0,896 | 0,896 | 0,855 | 0,954 |
| 0,993 | 0,889 | 0,894 | 1,010 | 0,892 | 0,900 | 0,870 | 0,901 | 0,921 | 0,842 | 0,896 | 0,999 | 0,894 |
| 0,905 | 0,892 | 0,871 | 0,779 | 0,877 | 0,750 | 0,873 | 0,774 | 0,885 | 0,886 | 0,610 | 0,892 |  |
| 1,059 | 0,893 | 0,892 | 1,180 | 0,812 | 0,896 | 1,180 | 0,892 | 0,839 | 0,995 | 0,909 | 0,958 |  |

Faça o devido estudo descritivo, que conclusão pode-se chegar?

**Nota**: Ponto de função é uma medida de tamanho de softwares. Não se preocupe com seu significado, apenas entenda que esta taxa é uma variável que pode ser estudada estatisticamente.

* **Medidas de forma**

1. Retornando aos dados do exercício 14, taxas de defeitos (Ponto de Função) descobertos em 50 projetos de softwares. Podemos admitir que os dados seguem nossa distribuição de referências, isto é, uma distribuição simétrica ou em forma de sino? Crie uma função no R para traçar o respectivo histograma e calcular o coeficiente de Assimetria e Curtose de Fisher, interprete os resultados.