UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

DISCIPLINA: MATD40 – ANÁLISE DE DADOS

ATIVIDADE EM SALA

1. Consideremos um conjunto de dados como aquele apresentado na Tabela 1, abaixo, obtido de um questionário respondido por 50 alunos de uma disciplina ministrado na Fundação Getúlio Vargas em São Paulo. Os dados estão disponíveis no arquivo ceagfgv.xls. Segue abaixo uma figura dos primeiros 10 registros dessa base de dados.

ident	Salário (R$)	Fluência inglês	Anos de formado	Estado civil	Número de filhos	Bebida preferida
1	3500	fluente	12.0	casado	1	outra alcoólica
2	1800	nenhum	2.0	casado	3	não alcoólica
3	4000	fluente	5.0	casado	1	outra alcoólica
4	4000	fluente	7.0	casado	3	outra alcoólica
5	2500	nenhum	11.0	casado	2	não alcoólica
6	2000	fluente	1.0	solteiro	0	não alcoólica
7	4100	fluente	4.0	solteiro	0	não alcoólica
8	4250	algum	10.0	casado	2	cerveja
9	2000	algum	1.0	solteiro	2	cerveja
10	2400	algum	1.0	solteiro	0	não alcoólica

a) Apresente representações tabular e gráfica univariadamente para cada tipo de variável da base de dados. Escreva uma breve descrição
b) Calcule as medidas de resumo para cada variável quantitativa sem esquecer de interpretar os resultados
c) Realize uma análise bivariada utilizando gráfico e alguma medida sintetização que pode ser coeficiente de correlação de Pearson, alguma medidas de associação. Essas análises devem compreender: duas variáveis quantitativas, duas qualitativas e uma mistura (quantitativa e qualitativa). Interprete-as

2. Os dados apresentados são provenientes de um estudo sobre endometriose realizado na Faculdade de Medicina da Bahia. A endometriose se caracteriza pela localização ectópica do tecido endometrial e é responsável por diversos problemas em mulheres na fase reprodutiva, incluindo dor pélvica e infertilidade. Um dos objetivos do estudo é saber se alguma das características observadas (ou combinação delas) pode ser utilizada como fator prognóstico de endometriose.

   Definição de variáveis:
   Gestação = número de gestações
   Partos = número de partos
   Abortos = número de abortos
   Dismenorréia (dor na menstruação): N = não tem, L = leve, M = moderada, I = intensa
   Dispareunia (dor na relação sexual): N = não tem, P = penetração, PRO = profunda, 2 = penetração e profunda
   AFSr (medida de gravidade da endometriose): 0 = menor gravidade, 4 = maior gravidade
   CA125/A = concentração de CA125 durante a menstruação
   CA125/B = concentração de CA125 10 dias após a menstruação
   PCRa = concentração de PCR (proteína C-reativa) durante a menstruação
   PCRb = concentração de PCR 10 dias após a menstruação

O pesquisador responsável pelo estudo tem a seguinte pergunta:

a) As pacientes doentes apresentam mais dor na menstruação do que as pacientes não doentes? Que tipo de análise você faria para responder essa pergunta utilizando as técnicas vista em disciplinas anteriores.
b) Compare as distribuições das variáveis idade e concentração de PCR durante a menstruação (PCRa) para pacientes dos grupos controle e doente utilizando medidas resumo (mínimo, máximo, quartis, mediana, média, desvio padraõ, etc.), boxplots, histogramas. Como você considerou os valores < 0;5 da variável PCRa nesses cálculos? Você sugeriria uma outra maneira para considerar tais valores?