Pesquisa sobre Incidência de Câncer

Os itens seguintes referem-se aos dados contidos no arquivo de nome *câncer.txt (www.ime.usp.br/~noproest)*. Esse arquivo contém os dados de uma pesquisa sobre incidência de câncer e é apresentado em 9 colunas representando as seguintes variáveis de interesse:

coluna 1:	identificação do paciente
coluna 2:	diagnóstico: 1 = falso-negativo: diagnosticados como não tendo a doença quando na verdade a tinham. 2 = negativo: diagnosticados como não tendo a doença quando de fato não a tinham. 3 = positivo: diagnosticados corretamente como tendo a doença. 4 = falso-positivo: diagnosticados como tendo a doença quando na verdade não tinham.
coluna 3:	idade.
coluna 4:	espectro químico da análise do sangue-alkaliine phosphatose (AKP).
coluna 5:	concentração de fosfato no sangue (P)
coluna 6:	enzima, lactate de dehydrogenase (LDH).
coluna 7:	albumina (ALB).
coluna 8:	nitrogênio na uréia (N).
coluna 9:	glicose (GL).

Fonte: Noções de Probabilidade e Estatística, Marcos N. Magalhães e Antonio C. P. de Lima, Edusp.

1. Escolha quatro variáveis dentre as colunas 2 a 9. Classifique-as, faça o gráfico da distribuição e a tabela de freqüência para cada uma delas. Analise as variáveis quanto a seu formato, posição, dispersão, pontos discrepantes (atípicos) e aglomerados. Comente sobre os resultados encontrados.
2. Uma afirmação feita por alguns médicos é a de que o grupo dos falso-positivos é mais jovem do que o dos falso-negativos. Para os dados dessa amostra, o que você diria a respeito? Justifique sua resposta baseando-se em gráficos e tabelas de freqüência.
3. Obtenha as medidas de posição e de variabilidade para as variáveis Idade e Glicose (GL). Comente os resultados obtidos
4. Repita o item (3) para cada tipo de diagnóstico. Compare as respostas obtidas e comente os resultados obtidos.
5. Utilizando a mediana da variável GL, classifique os pacientes em dois grupos, de alta e de baixa taxa de glicose. Denote essa nova variável por Clagl e construa uma tabela de dupla entrada entre Clagl e ALB. Você diria que as duas variáveis estão relacionadas de alguma forma? Justifique resumidamente a razão de sua resposta.

6. Considere os valores da variável Idade em três grupos: *jovem* com até 25 anos (inclusive), *meia idade* para indivíduos com idades entre 25 e 55 anos (inclusive) e *sênior* para maiores de 55 anos. Construa uma tabela de dupla entrada para estudar o comportamento desses grupos em relação à concentração de fosfato, tirando as conclusões pertinentes. Comente os resultados obtidos

7. Escolhendo-se um paciente ao acaso, qual a probabilidade de que ele seja do grupo *falso-negativo*, dado que tem mais de 50 anos? E ter acima de 50 anos, dado que não é do grupo *falso-negativo*? Utilize tabelas de dupla entrada para apoiar sua resposta.

8. Considere a variável LDH para os pacientes com pelo menos 40 anos de idade.
    a. Obtenha o histograma e algumas medidas descritivas. Justifique suas escolhas e comente;
    b. Você diria que os dados são simétricos? Qual a percentagem de observações compreendidas no intervalo entre a média mais ou menos 1 desvio-padrão? E, no intervalo entre a média mais ou menos 2 desvios-padrão? E, no intervalo entre a média mais ou menos 3 desvios-padrão?

9. Deseja-se verificar se conforme aumenta a idade, muda a concentração de nitrogênio na uréia.
    a. Suponha que selecionamos apenas os pacientes que têm a doença (isto é, consideramos o grupo formado por pacientes cujo diagnóstico é falso-negativo ou positivo). Construa um gráfico de dispersão para idade e concentração de nitrogênio. O que pode ser dito?
    b. Considere agora, os pacientes que não têm a doença (diagnóstico negativo ou falso-positivo). Construa um gráfico de dispersão para idade e concentração de nitrogênio. Compare com o gráfico obtido no item (a). Comente os resultados obtidos.
    c. Nos dois casos, ajuste as retas de regressão. Interprete os coeficientes angulares e os interceptos obtidos. Você diria que o efeito da idade, na concentração de nitrogênio, é um dado importante para discriminar entre pacientes com e sem a doença?