Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo

SME0620 - Estatística I

Professor: Francisco A. Rodrigues

Quinta Lista de exercícios: Análise exploratória de dados

1 - (Use o computador) Considere os dados da tabela 1 (download no site da disciplina).
a) Construa a distribuição de frequências, histograma ou gráficos de setores das variáveis:
i) Estado civil ii) Região de procedência iii) Número de filhos iv) Idade.
b) Construa a função de distribuição acumulada para as variáveis Idade e Salário.
c) Construa o histograma da variável Salário e determine a media, Q1, Q3 e a distância interquantil (Q3-Q1).

2 - Quer se estudar o número de erros de impressão de um livro. Para isso escolheu-se uma amostra de 50 páginas, encontrando-se o número de erros por página da tabela abaixo.

Erros	Frequência
0	25
1	20
2	3
3	1
4	1

(a) Qual o número médio de erros por página? (R: 0.66)
(b) E o número mediano? (R:0,5)
(c) Qual é o desvio padrão? (R: 0.8393)
(d) Se o livro tem 500 páginas, qual o número total de erros esperados no livro? (R: 330)

3 - Mostre que:
a) $\sum_{i=1}^{n}(x_i-\bar{x})=0$ b) $\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$
c) $\sum_{i=1}^{k}n_i(x_i-\bar{x})^2=\sum_{i=1}^{k}n_ix_i^2-n\bar{x}^2$ d) $\sum_{i=1}^{k}f_i(x_i-\bar{x})^2=\sum_{i=1}^{k}f_ix_i^2-\bar{x}^2$

4 - O Departamento Pessoal de uma certa firma fez um levantamento dos salários dos 120 funcionários do setor administrativo, obtendo os resultados (em salários mínimos) da tabela abaixo

Faixa salaria	Frequência relativa
$0 \vdash 2$	0,25
$2 \vdash 4$	0,40
$4 \vdash 6$	0,20
$6 \vdash 10$	0,15

a) Esboce o histograma correspondente.
b) Calcule a média, a variância e o desvio padrão.
c) Calcule o primeiro quartil e a mediana.
d) Se for concedido um aumento de 100% para todos os 120 funcionários, haverá alteração na média? E na variância? Justifique a sua resposta.
e) Se for concedido um abono de dois salários mínimos para todos os 120 funcionários, haverá alteração na média? E na variância? Justifique a sua resposta.

5 - O que acontece com a mediana, média e o desvio padrão de uma série de dados quando:
a) cada observação é multiplicada por 2?
b) soma-se 10 a cada observação?
c) subtrai-se a média geral $\bar{x}$ de cada observação?
d) de cada observação subtrai-se $\bar{x}$ e divide-se pelo desvio padrão σ?

6 - A idade média dos candidatos a um determinado curso de aperfeiçoamento sempre foi baixa, da ordem de 22 anos. Como esse curso foi planejado para atender a todos as idades, decidiu-se fazer uma campanha de divulgação. Para se verificar se a campanha foi ou não eficiente, fez-se um levantamento da idade dos candidatos à última promoção e os resultados estão na tabela a seguir:

Idade	Frequência	Porcentagem
$18 \vdash 20$	18	36
$20 \vdash 22$	12	24
$22 \vdash 26$	10	20
$26 \vdash 30$	8	16
$30 \vdash 36$	2	4

a) Baseando-se nesses resultados, você diria que a companha produziu algum efeito (isto é, aumentou a idade média)? (R:não, $\bar{x} = 22,5$)
b) Um outro pesquisador decidiu usar a seguinte regra: se a diferença $\bar{x} - 22$ fosse maior que o valor $2\sigma/\sqrt{(}n)$, então a campanha teria surtido efeito. Qual a conclusão dele, baseada nos dados?(R:Não)
c) Faça um histogram da distribuição.

7 - Considere os dados da tabela 1 (download no site da disciplina). Verifique se há associação entre as variáveis abaixo através do coeficiente de contingência (C) e coeficiente de contingência modificado ($\tilde{C}$).
a) Grau de instrução e região de procedência.
a) Grau de instrução e estado civil.

8 - Considere os dados da tabela 1 (download no site da disciplina). Construa o *box plots* da variável Salário segundo a Região de Procedência. Faça o mesmo com a variável Salário e Grau de Instrução. Os salários são mais influenciados pela região ou grau de instrução? Justifique sua resposta.

9 - Considere os dados da tabela 1 (download no site da disciplina).
a) Calcule o coeficiente de associação (R^2) entre as variáveis Grau de Instrução e Salário (R:0,415) e Região de Procedência e Salário (R:0,013). Os salários são mais influenciados pela região ou grau de instrução? Compare os resultados com o exercício anterior.
b) Calcule o coeficiente e associação entre as variáveis Estado Civil e Idade. Discuta os resultados.

10 - A seguir estão os dados referentes à porcentagem da população economicamente ativa empregada no setor primário e o respectivo índice de analfabetismo para algumas regiões metropolitanas brasileiras.

Regiões Metropolitanas	Setor Primário (Y)	Índice de Analfabetismo (X)
São Paulo	2,0	17,5
Rio de Janeiro	2,5	18,5
Belém	2,9	19,5
Belo Horizonte	3,3	22,2
Salvador	4,1	26,5
Porto Alegre	4,3	16,6
Recife	7,0	36,6
Fortaleza	13,0	38,4

a) Faça um diagrama de dispersão.
b) Você acha que existe uma dependência linear entre as duas variáveis?
c) Calcule o coeficiente de correlação de Pearson. (R:0,86)
d) Existe alguma região com comportamento diferente das demais? Se existe, elimine o valor correspondente e recalcule o coeficiente de correlação. (R: Porto Alegre e Fortaleza apresentam comportamento diferente das demais.)

11 - O departamento de vendas de certa companhia foi formado há um ano com a admissão de 15 vendedores. Nessa época, foram observados para cada um dos vendedores os valores de três variáveis: T: resultado em um teste apropriado para vendedores, E: anos de experiência em vendas, G: conceito do gerente de vendas, quanto ao currículo do candidato. O direto da companhia resolveu agora ampliar o quadro de vendedores e pede a sua colaboração para responder a algumas perguntas. Para isso, ele lhe dá informações adicionais sobre as duas variáveis: V: volume mensal de vendas em s.m.; Z: zona da capital para o qual o vendedor foi designado. O quadro foi o seguinte:

Mais especificamente, o diretor pede que as seguintes tarefas sejam realizadas:
a) Histograma das vendas em classes de amplitude 10, tendo por limite inferior da primeira classe o valor 15.
b) Média e variância das vendas. (R: $\bar{v} = 30,2$ e $\sigma^2 = 130,6$)
c) Suponha que um vendedor seja considerado excepcional se sua venda está acima de dois desvios-padrões da média. Há vendedores excepcionais entre os 15 iniciais? (R: há um vendedor)
d) O diretor de vendas anunciou que transferirá para outra praça todos os vendedores cujas vendas foram inferior ao

T: Teste	E: Experiência	G : Conceito	V: Vendas	Z: Zona
8	5	Bom	54	Norte
9	2	Bom	50	Sul
7	2	Mau	48	Sul
8	1	Mau	32	Oeste
6	4	Bom	30	Sul
8	4	Bom	30	Oeste
5	3	Bom	29	Norte
5	3	Bom	27	Norte
6	1	Mau	24	Oeste
7	3	Mau	24	Oeste
4	4	Bom	24	Sul
7	2	Mau	23	Norte
3	3	Mau	21	Sul
5	1	Mau	21	Oeste
3	2	Bom	16	Norte

primeiro quartil da distribuição. Qual o mínimo de vendas para não ser transferido? (R: $q_1 = 23,5$)
e) Os vendedores argumentam com o diretor que esse critério não é justo, pois há zonas de venda privilegiadas. A quem você daria razão? Justifique a sua resposta. (R: Há diferenças entre as zonas. Construa um *box plots* das variáveis zonas e vendas.)
f) Qual das três variáveis iniciais é mais importante para julgar o futuro candidato ao emprego? Justifique. (R: $\rho_{T,V} = 0,71$, $\rho_{E,V} = 0,26$)
g) Qual é o grau de associação entre o conceito do gerente e a zona a que o vendedor foi designado? Você tem alguma explicação para esse resultado?
h) Qual é o grau de associação entre o conceito do gerente e o resultado no teste? E entre zona e vendas?