LISTA DE EXERCÍCIOS 1 – INE 7001 - PROF. MARCELO MENEZES REIS
ANÁLISE EXPLORATÓRIA DE DADOS – GABARITO

1) Identificar nas pesquisas a seguir quais são as variáveis independentes e as dependentes. JUSTIFIQUE suas respostas.
a) *Variável independente: sexo do respondente. Variável dependente: preferência declarada. O sexo do respondente PODE influenciar a sua preferência eleitoral.*
b) *Variável independente: curso do CSE. Variáveis dependentes: sexo, idade, situação laboral. O levantamento pretende caracterizar os alunos por curso, que então PODE influenciar os valores das variáveis sexo, idade e situação laboral.*
c) *Variáveis independentes: sexo e faixa etária do consumidor. Variável dependente: preferência pelo refrigerante. As variáveis sexo e faixa etária do consumidor PODEM influenciar a preferência pelo refrigerante.*
d) *Variáveis independentes: temperatura, percentual de ferro e tempo de fundição. Variável dependente: dureza do aço. A variação dos valores de percentual de ferro, tempo de fundição e temperatura PODERÃO influenciar diretamente os valores da dureza do aço.*
e) *Variável independente: tipo de universidade (pública ou privada). Variável dependente: salário dos professores. O tipo de universidade PODE influenciar nos valores dos salários dos professores.*
f) *Variável independente: curso do CTC. Variável dependente: notas finais em Estatística. O fato do aluno estar no curso A, B, ou C (com maior ou menor carga horária de matemática) PODE influenciar a sua nota na disciplina de Estatística.*
g) *Variável independente: regiões econômico-geográficas. Variável dependente: PIB per capita. O fato de um país estar na região A, B, ou C (com maior ou menor desenvolvimento) PODE influenciar o valor do PIB.*

2) Identificar qual é o grau de mensuração das variáveis descritas a seguir. JUSTIFIQUE suas respostas.
a) *Variável qualitativa nominal. As realizações da variável curso são atributos (categorias), e não podem ser ordenadas.*
b) *Variável qualitativa nominal. As realizações da variável preferência declarada são atributos (categorias), e não podem ser ordenadas.*
c) *Variável quantitativa contínua. As realizações da variável consumo em km/l são números, e podem assumir (teoricamente) uma infinidade de valores.*
d) *Variável quantitativa discreta. As realizações da variável número de filhos são números, e podem assumir apenas alguns valores (números inteiros).*
e) *Variável quantitativa discreta. As realizações da variável número de residentes são números, e podem assumir apenas alguns valores (números inteiros).*
f) *Variável qualitativa ordinal. As realizações da variável são atributos (categorias), e podem ser ordenadas.*
g) *Variável qualitativa ordinal. As realizações da variável são atributos (categorias), e podem ser ordenadas.*
h) *Variável quantitativa contínua. As realizações da variável temperatura em graus Celsius são números, e podem assumir (teoricamente) uma infinidade de valores.*
i) *Variável quantitativa discreta. As realizações da variável nível de instrução (em número de anos completos) são números, e podem assumir apenas alguns valores (números inteiros).*
j) *Variável qualitativa ordinal. As realizações da variável velocidade (neste caso) são atributos (categorias), e podem ser ordenadas.*

3) Para os casos a seguir indique qual é o método estatístico (tabela, gráfico, medida de síntese) mais apropriado para resumir e interpretar os dados. JUSTIFIQUE suas respostas.
a) *Há apenas uma variável envolvida. Esta variável (peso) pode considerada quantitativa contínua (por causa da precisão das medidas). O objetivo do estudo é ter uma idéia do peso dos pacotes, o*

que poderia ser interpretado como descrição da tendência central. O conjunto apresenta mais de 100 dados (3000), há pouco tempo para a apresentação (apenas 5 minutos), e o público não conhece Estatística. Por essas razões pode ser construído um histograma agrupado em classes para os pesos: podemos observar onde está o valor do tipo maior de caixa na escala horizontal e verificar se há uma grande quantidade de encomendas com peso acima dele.

b) *Há apenas uma variável envolvida. Esta variável (valores investidos) pode ser considerada quantitativa contínua (suas realizações são números que podem assumir uma infinidade de valores, por causa das grandes diferenças entre os clientes, alguns aplicando dezenas de reais e outros dezenas de milhões). O objetivo do estudo é ter uma idéia dos valores investidos inclusive dos discrepantes, o que pode ser interpretado como uma descrição completa. O conjunto apresenta mais de 100 dados (450), há pouco tempo para a apresentação (apenas 3 minutos), e o público conhece Estatística. Por essas razões pode ser construído um diagrama em caixas para os valores investidos pelos clientes: pode-se observar a tendência central, a dispersão e os valores discrepantes em um diagrama único.*

c) *Há apenas uma variável envolvida. Esta variável (sexo) é qualitativa nominal (pois assume apenas 2 valores, que são atributos e não podem ser ordenados). O objetivo do estudo é apresentar a distribuição da variável sexo, o que poderia ser interpretado como "descrição de tendência central". O conjunto apresenta mais de 100 dados (851), não há restrições de tempo para a apresentação, e como se trata do público em geral, presume-se que não conheça Estatística. Por essas razões há várias opções possíveis: tabela de freqüências da variável sexo, gráfico em barras ou gráfico em setores para a variável sexo. Lembrando que a apresentação gráfica de resultados costuma facilitar a apreensão da informação por parte do público alvo, o que talvez recomendasse um dos dois gráficos citados.*

d) *Há duas variáveis envolvidas: processo e granulometria. Processo pode assumir dois valores, e será considerada independente, pois se julga que ela* ***pode*** *influenciar os valores de granulometria, que seria a dependente. Processo é qualitativa nominal, e granulometria é quantitativa contínua, pois suas realizações são números que podem assumir uma infinidade de valores. O objetivo do estudo é identificar qual dos dois processos é o mais homogêneo, o que significa a descrição de tendência central e dispersão da granulometria por processo, para que seja possível a comparação. O conjunto apresenta menos de 100 dados (40), não há restrições de tempo para a apresentação, e nós conhecemos Análise Exploratória de Dados. Por essas razões recomenda-se o cálculo de média, desvio padrão e coeficiente de variação da granulometria nos dois processos: aquele que apresentar o menor coeficiente de variação será o mais homogêneo.*

e) *Há apenas uma variável envolvida. Esta variável (IDH) pode considerada quantitativa contínua (pois suas realizações são números, que, por resultarem de operações de divisões, podem assumir uma infinidade de valores). O objetivo do estudo é fazer uma análise a mais completa possível, identificando valores típicos e discrepantes, o que pode ser interpretado como uma descrição completa. O conjunto apresenta mais de 100 dados (200), exige-se uma apresentação gráfica, e o público conhece Estatística. Por essas razões pode ser construído um diagrama em caixas para os IDHs: pode-se observar a tendência central, a dispersão e os valores discrepantes em um diagrama único.*

f) *Há apenas uma variável envolvida. Esta variável (peso das embalagens) pode considerada quantitativa contínua (pois suas realizações são números, que podem assumir uma infinidade de valores, visto que é utilizada uma balança de precisão). O objetivo do estudo é ter uma rápida idéia dos pesos, procurando identificar problemas: os problemas podem ser pesos muito pequenos ou muito grandes, ou com assimetria, o que caracteriza uma descrição completa. O conjunto apresenta mais de 100 dados (800), exige-se uma apresentação rápida (o que significaria o uso de algum gráfico), e o público não conhece Estatística. Por essas razões pode ser construído um histograma agrupado em classes, o que evidenciaria a quantidade de embalagens acima e abaixo dos limites e uma eventual assimetria nos pesos.*

g) *Há duas variáveis envolvidas: preferência por modelo e renda mensal. Preferência pode assumir vários valores, e será considerada independente, pois se pretende usá-la como variável de*

agrupamento, em função dos seus valores serão avaliados os valores da renda mensal, que será a dependente. Preferência por modelo é qualitativa nominal (seria apenas o modelo de automóvel), Renda mensal pode ser considerada quantitativa contínua, especialmente se for medida em salários mínimos (mas não há informação a respeito), pois suas realizações são números e podem assumir uma infinidade de valores. O objetivo do estudo é ter uma idéia a mais completa possível da distribuição de renda por preferência, o que configura uma descrição completa. Não há informações sobre o tamanho do conjunto de dados, mas podemos supor que, sendo uma pesquisa patrocinada por uma montadora, haja mais de 100 observações. A apresentação precisa ser feita de forma rápida, e a diretoria é versada em Estatística. Por essas razões recomenda-se a construção de um diagrama em caixas múltiplo, um diagrama para a renda de cada modelo preferido, sendo todos colocados em um gráfico com a mesma escala. Se o número de modelos for muito grande, podemos agrupá-los em categorias, criando uma nova variável qualitativa: por exemplo, montadora FIAT, Palio, Siena, e Weekend seriam "pequenos", Punto e Estilo, "médios", Dobló e Linea "grandes", reduzindo então o número de diagramas e facilitando a análise.

h) *Há apenas uma variável envolvida: dimensão da peça. É razoável imaginar que se disponha de equipamentos de medição razoavelmente precisos, o que permitiria considerar a dimensão como quantitativa contínua, pois suas realizações são números que podem assumir uma infinidade de valores. O objetivo é apresentar um relatório sobre a dimensão da peça, sendo que a técnica escolhida não pode dar margem a interpretações conflitantes se suas características forem modificadas. Exige-se que a apresentação seja rápida, e o público alvo conhece Estatística. Por essas razões recomenda-se a construção de um diagrama de pontos para a dimensão da peça, que permite a visualização do comportamento completo dos dados, e é único para um determinado conjunto de dados.*

i) *Há duas variáveis envolvidas: opinião sobre o banco atual e renda. Opinião será considerada independente, pois se pretende usá-la como variável de agrupamento, em função dos seus valores serão avaliados os valores da renda dos clientes, que será a dependente. Opinião é qualitativa ordinal (poderia ser classificada como "ótima", "ruim", etc.), renda pode ser considerada quantitativa contínua, especialmente se for medida em salários mínimos (mas não há informação a respeito), pois suas realizações são números e podem assumir uma infinidade de valores. O objetivo do estudo é apresentar o relacionamento entre as duas variáveis: não há obrigatoriedade de uma descrição completa, mas enriqueceria mais os resultados. O conjunto tem MENOS de 100 observações, há apenas 5 minutos para realizar a apresentação e a diretoria do banco conhece Estatística. Por essas razões recomenda-se a construção de um diagrama de pontos múltiplo, um diagrama para a renda de cada valor de opinião, sendo todos colocados em um gráfico com a mesma escala.*

j) *Há duas variáveis envolvidas: opinião sobre o banco atual e renda. Opinião será considerada independente, pois se pretende usá-la como variável de agrupamento, em função dos seus valores serão avaliados os valores da renda dos clientes, que será a dependente. Opinião é qualitativa ordinal (poderia ser classificada como "ótima", "ruim", etc.), renda pode ser considerada quantitativa contínua, especialmente se for medida em salários mínimos (mas não há informação a respeito), pois suas realizações são números e podem assumir uma infinidade de valores. O objetivo do estudo é apresentar o relacionamento entre as duas variáveis: não há obrigatoriedade de uma descrição completa, mas enriqueceria mais os resultados. O conjunto tem MENOS de 100 observações, há apenas 5 minutos para realizar a apresentação e a diretoria do banco NÃO conhece Estatística. Por essas razões recomenda-se calcular média e intervalo para a renda de cada valor de opinião sendo todos colocados em um gráfico com a mesma escala: atenção, se o número de observações for muito pequeno, menos de 30, por exemplo, talvez simplesmente apresentar os dados em ordem crescente (rol) fosse mais apropriado.*

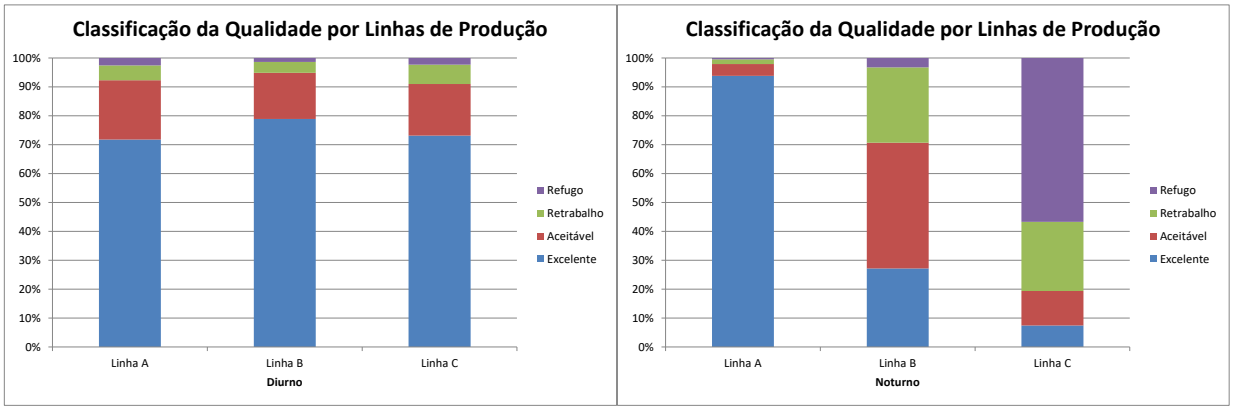
k) *Há duas variáveis envolvidas: opinião sobre o banco atual e renda. Opinião será considerada independente, pois se pretende usá-la como variável de agrupamento, em função dos seus valores serão avaliados os valores da renda dos clientes, que será a dependente. Opinião é qualitativa ordinal (poderia ser classificada como "ótima", "ruim", etc.), renda pode ser considerada*

quantitativa contínua, especialmente se for medida em salários mínimos (mas não há informação a respeito), pois suas realizações são números e podem assumir uma infinidade de valores. O objetivo do estudo é apresentar o relacionamento entre as duas variáveis: não há obrigatoriedade de uma descrição completa, mas enriqueceria mais os resultados. O conjunto tem MAIS de 100 observações, há apenas 5 minutos para realizar a apresentação e a diretoria do banco conhece Estatística. Por essas razões recomenda-se a construção de um diagrama em caixas múltiplo, um diagrama para a renda de cada valor de opinião, sendo todos colocados em um gráfico com a mesma escala.

l) *Há duas variáveis envolvidas: opinião sobre o banco atual e renda. Opinião será considerada independente, pois se pretende usá-la como variável de agrupamento, em função dos seus valores serão avaliados os valores da renda dos clientes, que será a dependente. Opinião é qualitativa ordinal (poderia ser classificada como "ótima", "ruim", etc.), renda pode ser considerada quantitativa contínua, especialmente se for medida em salários mínimos (mas não há informação a respeito), pois suas realizações são números e podem assumir uma infinidade de valores. O objetivo do estudo é apresentar o relacionamento entre as duas variáveis: não há obrigatoriedade de uma descrição completa, mas enriqueceria mais os resultados. O conjunto tem MAIS de 100 observações, há apenas 5 minutos para realizar a apresentação e a diretoria do banco NÃO conhece Estatística. Por essas razões recomenda-se a construção de um histograma agrupado em classes múltiplo, um histograma para a renda de cada valor de opinião, sendo todos colocados em um gráfico com a mesma escala.*

4) Certo fabricante de ferramentas dispõe de três linhas de produção (A, B e C), que são operadas em dois turnos de 8 horas (diurno e noturno). Houve reclamações sobre a qualidade dos produtos, então a direção resolveu intensificar a vigilância avaliando a qualidade (classificada como excelente, aceitável, retrabalho ou refugo) das peças produzidas nas três linhas nos dois turnos. Os resultados estão nas tabelas e gráficos a seguir.

Diurno		Qualidade				Total
Linha		Excelente	Aceitável	Retrabalho	Refugo	
A	Frequência	420	120	30	15	585
	% por linha	71,79%	20,51%	5,13%	2,56%	100,00%
	% por coluna	31,44%	37,50%	33,71%	41,67%	32,85%
B	Frequência	568	115	27	10	720
	% por linha	78,89%	15,97%	3,75%	1,39%	100,00%
	% por coluna	42,51%	35,94%	30,34%	27,78%	40,43%
C	Frequência	348	85	32	11	476
	% por linha	73,11%	17,86%	6,72%	2,31%	100,00%
	% por coluna	26,05%	26,56%	35,96%	30,56%	26,73%
Total	Frequência	1336	320	89	36	1781
	% por linha	75,01%	17,97%	5,00%	2,02%	100,00%
	% por coluna	100%	100%	100%	100%	100%
Noturno		Qualidade				Total
Linha		Excelente	Aceitável	Retrabalho	Refugo	
A	Frequência	575	25	10	3	613
	% por linha	93,80%	4,08%	1,63%	0,49%	100,00%
	% por coluna	79,31%	9,43%	4,76%	1,44%	43,54%
B	Frequência	125	200	120	15	460
	% por linha	27,17%	43,48%	26,09%	3,26%	100,00%
	% por coluna	17,24%	75,47%	57,14%	7,21%	32,67%
C	Frequência	25	40	80	190	335
	% por linha	7,46%	11,94%	23,88%	56,72%	100,00%
	% por coluna	3,45%	15,09%	38,10%	91,35%	23,79%
Total	Frequência	725	265	210	208	1408
	% por linha	51,49%	18,82%	14,91%	14,77%	100,00%
	% por coluna	100%	100%	100%	100%	100%



a) Qual é a qualidade predominante no turno diurno? E no noturno? JUSTIFIQUE.

Observar a linha Total nas duas tabelas. No diurno a qualidade predominante é a Excelente com 75,01% (percentual por linha). No noturno é a Excelente com 51,49% do total.

b) O ideal era a produção total distribuir-se igualmente entre as três linhas de produção. Isso ocorre no diurno? E no noturno? JUSTIFIQUE.

Observar a coluna Total nas duas tabelas. No diurno a distribuição é desigual: 32,85% (percentual por coluna) na linha A, 40,43% na B e 26,73% na C. No noturno a distribuição também é desigual: 43,54% na A, 32,67% na B e 23,79% na C. Uma distribuição igual exigiria em torno de 33% de cada linha de produção nos dois turnos.

c) Existe associação entre a qualidade das peças e a linha de produção no diurno? E no noturno? JUSTIFIQUE.

Observar a linha total nas duas tabelas. Se não houvesse relação os percentuais por linha nas linhas de produção deveriam ser semelhantes aos do total.

No turno diurno deveriam ser 75,01% de excelente, 17,97% de aceitável, 5% de retrabalho e 2,02% de refugo: isso realmente ocorre, os % por linha são semelhantes aos citados (por exemplo, 71,79% de excelente na linha A, 78,89% na linha B e 73,11% na C, não se afastando 5% de 75,01%[1]). Então NÃO HÁ relação entre as variáveis linha de produção e qualidade no turno diurno.

No turno noturno deveriam ser 51,49% de excelente, 18,82% de aceitável, 14,91% de retrabalho e 14,77% de refugo: há grandes diferenças de uma linha para outra, na A há 93,80% de excelente, contra 17,24% na B e apenas 7,46% de excelente na C. Então HÁ relação entre as variáveis linha de produção e qualidade no turno noturno.

É possível visualizar as semelhanças das linhas de produção do diurno pelo gráfico de colunas 100% empilhadas à esquerda: observem a grande semelhança entre as barras, com mais de 70% da produção considerada excelente. Já no gráfico do noturno as barras são muito diferentes entre si: quase a totalidade da produção da linha A é considerada excelente, e mais de 50% da produção da linha C é considerada refugo.

5) O caso do percentual de vendas.

a) *Uma tabela de freqüências agrupada em classes pode ser usada para resumir o conjunto, porque a variável é quantitativa e apresenta grande variação nos valores (pode ser considerada contínua).*

Classes	Freqüência	Ponto Médio
-3,15 |-- 0,51	6	-1.32
0,51 |-- 4,17	7	2.34
4,17 |-- 7,83	10	6.00
7,83 |-- 11,49	7	9.66
11,49 |-- 15,15	1	13.32
15,15 |-- 18,81	1	16.98
Total	32	-

[1] Raciocínio semelhante pode ser obtido pelos percentuais por coluna.

A esmagadora maioria dos percentuais de crescimento está abaixo de 10% (23 filiais). Conclui-se que a promoção não obteve o resultado esperado.

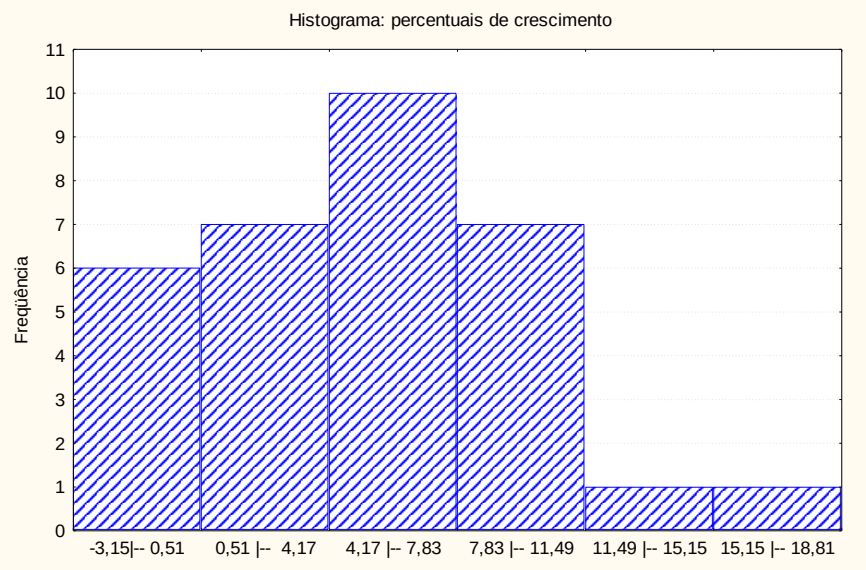
Não obstante poder utilizar a tabela agrupada em classes neste caso ela é mais indicada quando há mais de 100 observações.

b) *Poderia utilizar duas ferramentas:*

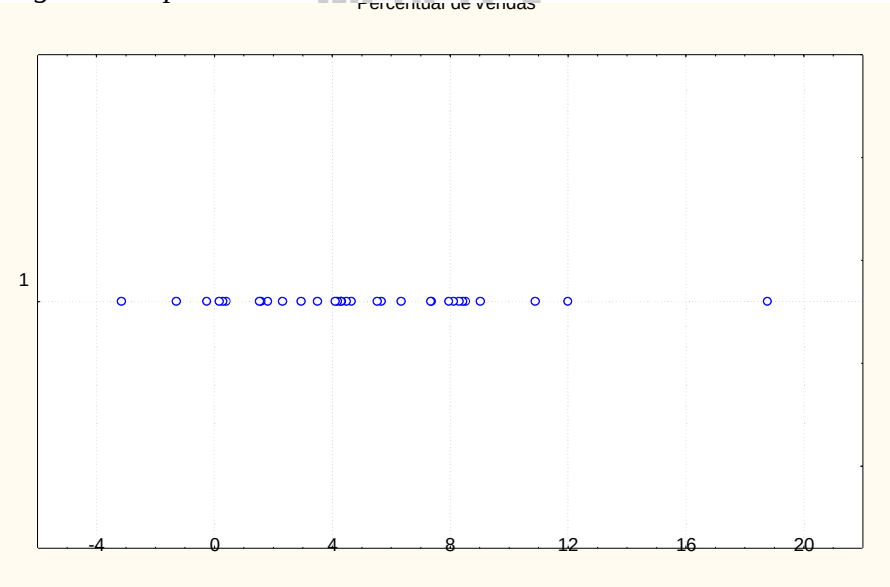
b.1 – *Histograma para dados agrupados em classes: porque a variável é quantitativa e apresenta grande variação nos valores (pode ser considerada contínua), e há uma tabela agrupada em classes disponível.*

b.2 – *Diagrama de pontos: porque a variável é quantitativa e apresenta grande variação nos valores (pode ser considerada contínua) e trata-se de um conjunto pequeno de dados.*

c) *Construindo o histograma:*



Construindo o diagrama de pontos:

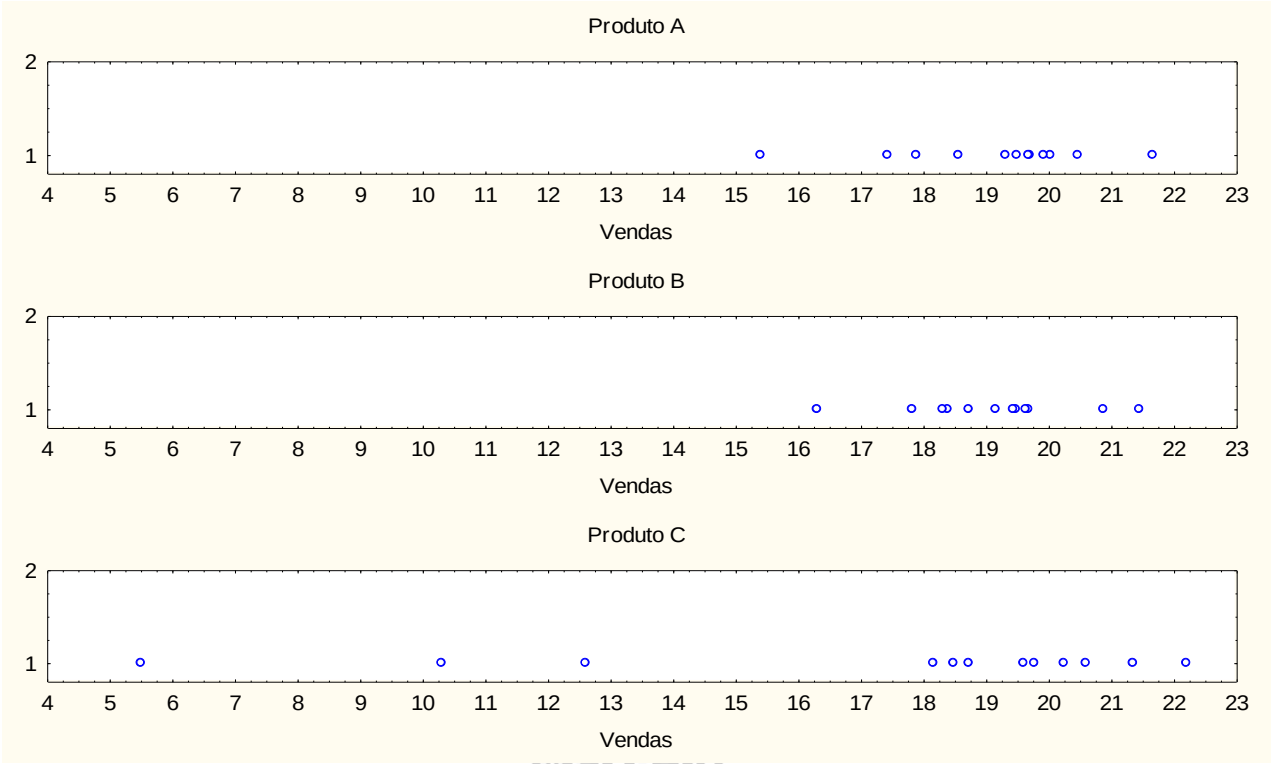


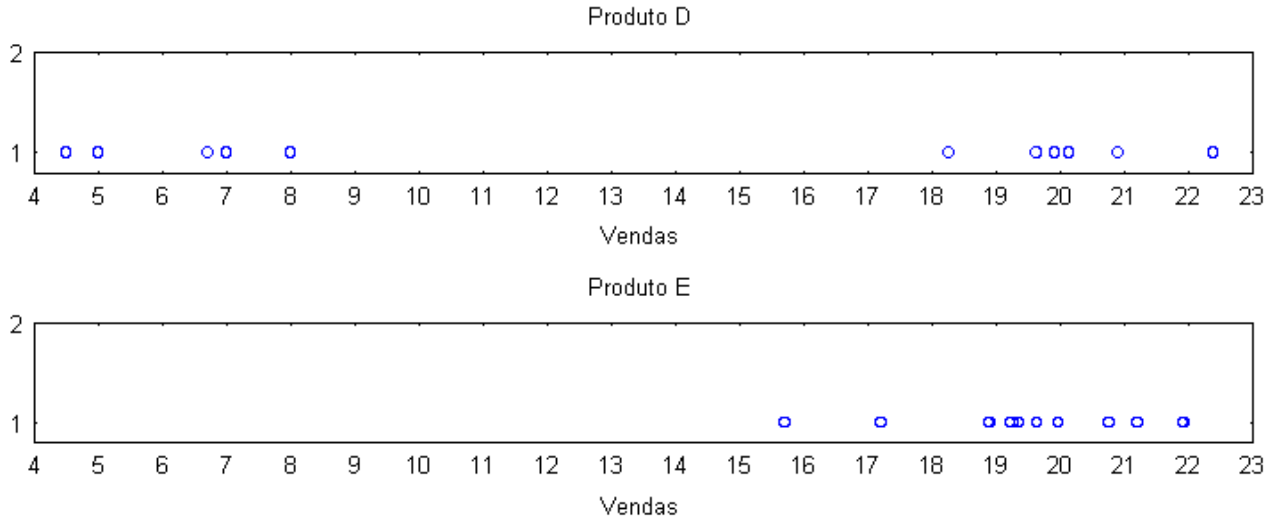
Pelo histograma, ou pelo diagrama de pontos as conclusões são semelhantes às da tabela de frequências agrupadas em classes. A maioria dos percentuais de crescimento está abaixo de 10%, portanto a promoção não teve o efeito esperado.

d) Medidas de síntese. Como há valores discrepantes, tanto superiores quanto inferiores, que podem vir a distorcer a média dos percentuais opta-se pela mediana.

e) A mediana vale 4,405. O valor típico dos percentuais é 4,4%, substancialmente menor do que o 10% considerado mínimo aceitável, portanto conclui-se que a promoção não teve o efeito esperado.

6) *O problema da escolha dos 2 produtos que precisam ser retirados de produção.*
a) *Pode ser usado o diagrama de pontos, porque a variável é quantitativa (com grande variação, podendo ser considerada contínua), e o conjunto de dados é pequeno, apenas 12 observações. É importante ressaltar que é preciso construir um diagrama para cada produto, com a mesma escala, para que seja possível fazer a comparação.*





b) Pelos diagramas de pontos podemos observar que os produtos C e D tem demandas substancialmente diferentes ao longo do ano: em alguns meses são bastante próximas das dos outros produtos A, B e E, mas em outros elas caem muito. Um destes precisa ser escolhido: o produto D tem as vendas menos homogêneas, há 5 valores substancialmente mais baixos, enquanto o produto C tem apenas 3. Como se buscam os produtos com demanda mais homogênea, devemos recomendar a retirada do produto D.
c) Podemos calcular os CV% para detectar os produtos com vendas MENOS homogêneas: os que apresentarem maiores CV% devem ser retirados.
d) Calculando média, desvio padrão e CV% para os produtos, obtemos:

Medida	*A*	*B*	*C*	*D*	*E*
Média	*19,125*	*19,102*	*17,291*	*14,337*	*19,334*
D.padrão	*1,627*	*1,358*	*5,096*	*7,262*	*1,674*
CV%	*8,51%*	*7,11%*	*29,47%*	*50,65%*	*8,66%*

Como C e D têm os maiores CV% (29,47% e 50,65%) apresentam vendas MENOS homogêneas, e o D apresenta as menos homogêneas, por ter o maior V% (50,65%), portanto deve ser o retirado.

7) *O problema da discriminação por sexo e grupo étnico.*

a) *Baseando-se apenas no primeiro gráfico a diferença entre homens e mulheres na empresa (54%, 46%) não é muito diferente da diferença na população (aproximadamente 50% - 50%). Então, observando apenas este gráfico não há discriminação com base no sexo.*

b) *Baseando-se apenas no segundo gráfico a diferença entre pessoas do grupo branco e outro na empresa (78%, 22%) não é muito diferente da diferença na população (70%, 30%). Então, observando apenas este gráfico não há discriminação com base no grupo étnico do funcionário.*

c) *Avaliando o terceiro gráfico (que relaciona função por sexo e grupo étnico) observa-se que no mesmo grupo étnico as oportunidades de homens e mulheres são diferentes: apenas 5,7% das mulheres do grupo branco ocupam cargos de gerência (frente a 36% dos homens), e nenhuma mulher do grupo outro ocupa cargo de gerência.*

d) *Comparando homens dos grupos branco e outro observa-se que 36% dos brancos ocupam cargos de gerência contra apenas 6,3% dos outros (quase 6 vezes menos...). Quando são comparadas as mulheres a diferença é ainda maior: 5,7% contra 0%! Portanto não há oportunidades iguais para pessoas de diferentes grupos étnicos.*

8) As tabelas a seguir apresentam: distribuição de frequências agrupada em classes dos gastos diários de turistas (em R$) por 3 balneários brasileiros (A, B e C); medidas de síntese dos gastos diários de turistas (em R$) para os mesmos dados da distribuição de frequências. Com base nestas tabelas responda os itens a seguir, JUSTIFICANDO suas respostas.

Gastos diários (R$)	Balneário						Total	
	A		B		C			
Classes	Freq.	%	Freq.	%	Freq.	%	Freq.	%
92|-114	0	0,00%	0	0,00%	2	0,80%	2	0,27%
114|-136	0	0,00%	0	0,00%	19	7,60%	19	2,53%
136|-158	0	0,00%	0	0,00%	35	14,00%	35	4,67%
158|-180	1	0,40%	6	2,40%	64	25,60%	71	9,47%
180|-202	6	2,40%	131	52,40%	66	26,40%	203	27,07%
202|-224	44	17,60%	112	44,80%	48	19,20%	204	27,20%
224|-246	97	38,80%	1	0,40%	11	4,40%	109	14,53%
246|-268	82	32,80%	0	0,00%	5	2,00%	87	11,60%
268|-290	19	7,60%	0	0,00%	0	0,00%	19	2,53%
290|-312	1	0,40%	0	0,00%	0	0,00%	1	0,13%
Total	250	100,0%	250	100,0%	250	100,0%	750	100,0%

Medidas de Gastos (R$)	Balneário			Total
	A	B	C	
Média	240,37	200,12	181,32	207,27
Mediana	238,67	200,48	181,86	204,95
Desvio padrão	20,07	9,71	30,63	32,93
CV%	8,41%	4,84%	16,84%	16,07%
Quartil inferior	226,65	193,95	162,25	190,43
Quartil superior	254,61	206,44	203,19	228,36

a) APENAS PELA DISTRIBUIÇÃO AGRUPADA EM CLASSES, há evidência de relação entre as variáveis Gasto diário e Balneário? Use percentuais na justificativa.
a) SIM. Somando os percentuais das classes de R$ 158 a R$ 246 observam-se grandes diferenças: no total, os gastos nessas classes representam 78,27%; no balneário A 59,20% (quase 20% de diferença); no balneário B 100% dos gastos estão nessas classes (mais de 20% de diferença); e no balneário C 75,20% (pequena diferença) dos gastos estão nas classes mencionadas. Somando os percentuais das três últimas classes também se observam grandes diferenças: no total, essas classes compreendem 14,26% dos gastos; no balneário C apenas 2% estão nessa faixa; no balneário A 40,8% dos gastos estão nas últimas três classes; e no balneário B não há nenhum gasto registrado nas últimas três classes. Percebe-se que no balneário C estão os gastos mais baixos, os mais altos no bairro A, e no bairro B os valores são intermediários.
b) APENAS PELAS MEDIDAS DE SÍNTESE, analise as medidas média, mediana e quartis para o TOTAL de balneários. Caracterize a tendência central dos gastos diários.
b) A tendência central do total dos gastos está entre R$ 204,95 (mediana, 50% dos gastos estão abaixo deste valor e 50% acima) e R$ 207,27 (média). Além disso, 25% dos gastos são menores do que R$ 190,43 (quartil inferior), e 25% dos gastos são maiores do que R$ 228,36 (quartil superior).
c) APENAS PELAS MEDIDAS DE SÍNTESE, em qual dos balneários os gastos diários são mais homogêneos em torno da média?
c) Avaliando os coeficientes de variação percentual, pois as médias de gastos nos balneários são diferentes: o balneário B tem os gastos mais homogêneos, pois tem o menor cv%, que vale 4,84% (no balneário A vale 8,41% e no C 16,84%).

9) *Salários anuais dos funcionários.*
a) *O valor típico (mediana) dos salários do sexo masculino é superior ao do sexo feminino. Ambos os conjuntos são assimétricos, mas a assimetria do sexo masculino é mais acentuada. Os salários dos funcionários do sexo masculino apresentam maior dispersão do que os do sexo feminino, e há valores discrepantes superiores em ambos os grupos.*
b) *Há diferença entre os salários de homens e mulheres: o valor típico (mediana) dos homens, além dos quartis inferior e superior, são substancialmente maiores do que os das mulheres.*
c) *Os valores típicos dos salários dos dois grupos estão próximos. Mas o grupo branco apresenta maior assimetria (o grupo outro é praticamente simétrico) e maior dispersão. Há valores discrepantes superiores em ambos, embora em maior quantidade no grupo branco.*
d) *Há diferença: embora o valor típico do grupo branco seja pouco superior ao do grupo outro, o quartil superior do grupo branco é significativamente maior. Além disso a dispersão nos salários acima da mediana no grupo branco é substancialmente maior do que a do grupo outro, ou seja há salários mais altos no grupo branco.*
e) *Diagrama em caixas do salário por sexo e função. O valor típico (mediana) dos homens em cargos de gerência é o maior, sendo que o das mulheres é sempre menor do que o dos homens, independente da função. Os salários dos homens em cargos de gerência e serviços gerais são praticamente simétricos, os outros grupos apresentam assimetria. A dispersão também é maior nos salários dos homens em posições de gerência. Apenas o grupo das mulheres em posições de gerência não apresenta valores discrepantes.*
f) *Obviamente há diferença, especialmente nos cargos de gerência onde o valor típico (mediana) do salário dos homens (63750) é consideravelmente maior do que o das mulheres (45187). Nos cargos de escritório a diferença também existe, embora seja menor (29850 para 24000). Nos serviços gerais a comparação não pode ser feita pois não há mulheres nesta função.*
g) *Os valores típicos (medianas) do salário das funções de escritório e serviços gerais são semelhantes, mas nos cargos de gerência o do grupo outro é superior ao do grupo branco. Há assimetria no grupo outro, tanto para escritório quanto para gerência, nos demais há praticamente assimetria. A dispersão é maior nos salários de gerência, sendo maior no grupo branco. Somente não há valores discrepantes no grupo outro, cargos de gerência.*

h) *Há diferença entre os grupos. Nas funções escritório e serviços gerais as medianas são bastante semelhantes, mas em gerência a mediana do salário do grupo outro é superior a do grupo branco (72375 contra 60187,50), embora os maiores salários do grupo branco sejam maiores do que os maiores salários do grupo outro. Observe que este resultado aparentemente contradiz o obtido na letra d: porque lá não havia a segmentação por função, e os salários das funções de escritório do grupo branco são ligeiramente maiores do que os do grupo outro, assim como os discrepantes no grupo branco ocorrem em maior número e são maiores.*

10) *Estão disponíveis apenas as informações referentes à média e desvio padrão dos salários nas duas faixas etárias.*
a) *Com base apenas na tabela a faixa etária "50 ou mais" tem os maiores salários, pois sua média é 398.610 dólares, acima daquela apresentada pela outra faixa etária.*
b) *Para avaliar a homogeneidade precisamos calcular os coeficientes de variação dos salários em cada faixa etária, pois tanto médias quanto desvios padrões são diferentes, impossibilitando a comparação direta:*

$$\mathbf{cv}\%_{\mathbf{Menos}\ \text{de}\ 50} = \frac{\mathbf{s}}{\bar{\mathbf{x}}} \times 100 = \frac{219{,}98}{382{,}2} \times 100 = 55{,}56\% \quad \mathbf{cv}\%_{50\ \text{ou mais}} = \frac{\mathbf{s}}{\bar{\mathbf{x}}} \times 100 = \frac{197{,}92}{398{,}61} \times 100 = 49{,}65\%$$

Como o coeficiente de variação da faixa etária "50 ou mais" é MENOR do que o da outra, os seus salários são mais homogêneos (há menor dispersão).

11) *A análise dos diagramas em caixa é semelhante aos casos anteriores:*
a) *Valor típico: "50 anos ou mais" apresenta maior valor típico, mediana pouco acima de 350 mil, enquanto o outro grupo apresenta mediana pouco acima de 300 mil.*
Assimetria: ambas as distribuições apresentam assimetria, pois as alturas das caixas são diferentes (Qs - Md ≠ Md - Qi), mas o grupo "menos de 50 anos" é mais assimétrico pois a diferença é maior.
Dispersão: o grupo "menos de 50 anos" apresenta maior dispersão, pois a diferença entre os quartis superior e inferior é maior do que no outro grupo.
Valores discrepantes: ambos os grupos possuem valores discrepantes, um no "menos de 50 anos" e dois no outro.
b) *Para avaliar qual das faixas de idade devemos levar em conta as três separatrizes principais: quartil inferior, mediana e quartil superior. O grupo "50 ou mais" tem mediana e quartil inferior maiores do que o outro, mas seu quartil superior é menor. Não obstante é possível tomar a decisão apenas pela mediana, indicando que 50% do grupo "50 ou mais" ganham mais do que 350 mil, enquanto que no outro grupo a mediana vale em torno de 300 mil: logo, o grupo "50 ou mais" está sendo melhor remunerado.*