1ª Lista de PE – Solução

1. a) Qualitativa nominal.
   b) Quantitativa discreta.
   c) Quantitativa discreta.
   d) Quantitativa contínua.
   e) Quantitativa contínua.
   f) Qualitativa ordinal.
   g) Quantitativa contínua.
   h) Qualitativa nominal.

2. O item (c) seria a melhor opção.

3. O ambiente a ser escolhido deveria ser um bar, item (b). Isto porque nos outros ambientes não deve haver fumantes em potencial para observação.

4. a) Segue abaixo a distribuição de frequências para a variável *Estado civil*.

Estado civil	**Frequência (n_i)**	**Frequência relativa (f_i)**
Solteiro	16	0,4444
Casado	20	0,5556
Total	36	1,0000

   b) Segue abaixo a distribuição de frequências para a variável *Região de procedência*.

Região de procedência	**Frequência (n_i)**	**Frequência relativa (f_i)**
Interior	12	0,3333
Capital	11	0,3056
Outra	13	0,3611
Total	36	1,0000

   c) Segue abaixo a distribuição de frequências para a variável *Número de filhos*.

Número de filhos	**Frequência (n_i)**	**Frequência relativa (f_i)**
0	4	0,20
1	5	0,25
2	7	0,35
3	3	0,15
5	1	0,05
Total	20	1,00

d) Segue abaixo a distribuição de frequências para a variável *Idade*.

Idade	**Frequência (n_i)**	**Frequência relativa (f_i)**
$20 \vdash 25$	2	0,0556
$25 \vdash 30$	6	0,1667
$30 \vdash 35$	10	0,2778
$35 \vdash 40$	8	0,2222
$40 \vdash 45$	8	0,2222
$45 \vdash 50$	2	0,0556
Total	36	1,0000

5. a) **Seção:** Variável qualitativa nominal.
Administração: Variável quantitativa contínua.
Direito: Variável quantitativa contínua.
Redação: Variável quantitativa contínua.
Estatística: Variável quantitativa contínua.
Política: Variável quantitativa contínua.
Economia: Variável quantitativa contínua.
Inglês: Variável qualitativa ordinal.
Metodologia: Variável qualitativa ordinal.

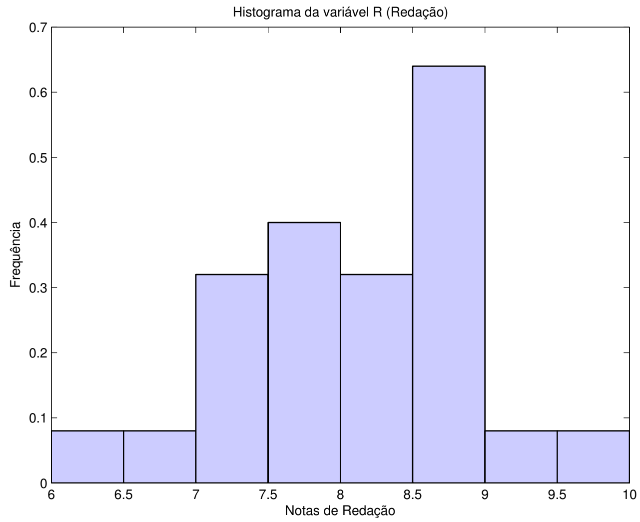
b) A distribuição dos dados para a variável *Direito* é uniforme (todos os funcionários obtiveram mesmo desempenho) enquanto a distribuição dos dados para as variáveis *Estatística* e *Política* é mais dispersa.

c) Para construirmos o histograma com 8 classes para a variável *Redação* precisamos antes determinar a amplitude das classes para em seguida obtermos a frequência relativa e a altura de cada um dos blocos. Para facilitar a visualização desses resultados é sempre interessante recorrermos à distribuição de frequências dos dados em questão.

Lembre-se que calculamos a altura dos blocos pela fórmula $h_i = f_i/\Delta_i$, onde Δ_i é a amplitude das classes. Em nosso caso, deveremos ter 8 classes de amplitude 0,5, ou seja, Δ_i é constante e igual a 0,5 ou $\Delta_i = \Delta = 0,5$.

Redação	**Frequência (n_i)**	**Frequência relativa (f_i)**	**Altura (h_i)**
$6,0 \vdash 6,5$	1	0,04	0,08
$6,5 \vdash 7,0$	1	0,04	0,08
$7,0 \vdash 7,5$	4	0,16	0,32
$7,5 \vdash 8,0$	5	0,20	0,40
$8,0 \vdash 8,5$	4	0,16	0,32
$8,5 \vdash 9,0$	8	0,32	0,64
$9,0 \vdash 9,5$	1	0,04	0,08
$9,5 \vdash 10,0$	1	0,04	0,08
Total	25	1,00	

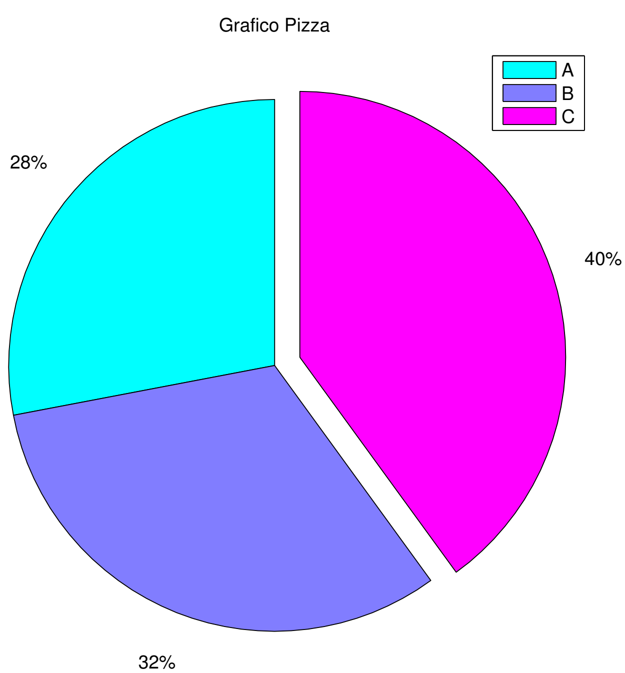
Segue abaixo o histograma para a variável *Redação*.



d) O primeiro passo é obtermos a distribuição de frequências da variável em questão, ou seja, *Metodologia*.

Metodologia	**Frequência** (n_i)	**Frequência relativa** (f_i)
A	7	0,28
B	8	0,32
C	10	0,40
Total	25	1,00

Tendo em vista que a nossa variável é qualitativa ordinal, tanto o gráfico em barras quanto o gráfico em pizza são recomendáveis para representar a distribuição dos dados. Como exemplo, iremos construir o gráfico em pizza da variável.



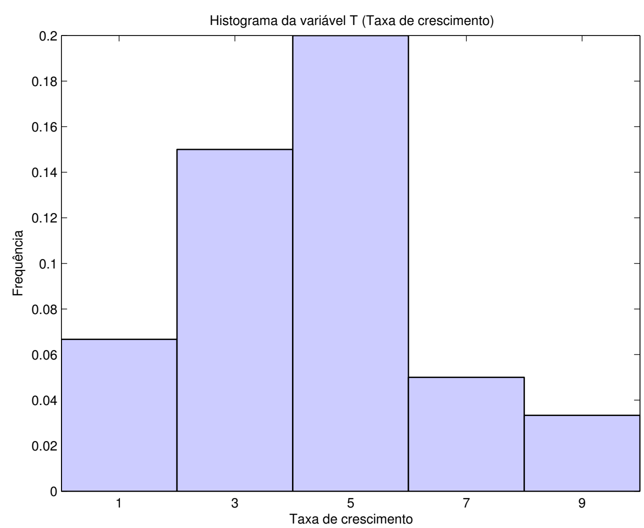
e) Sorteando ao acaso um dos 25 funcionários, a probabilidade de que ele tenha obtido grau A em *Metodologia* está explicitado na tabela de distribuição de frequências, ou seja, 28%.

f) A probabilidade é menor, pois seriam 42 possibilidades num total de 600, ou seja, a probabilidade seria de 7%.

6. a) É fácil observar que a taxa de crescimento geométrico anual é uma variável quantitativa contínua. Como a amplitude dos intervalos já foi especificada, vamos construir um histograma com 5 classes de mesma amplitude e, novamente, precisaremos da frequência relativa e da altura.

Taxa de Cresc.	**Frequência (n_i)**	**Frequência relativa (f_i)**	**Altura (h_i)**
$0 \vdash 2$	4	0,1333	0,0667
$2 \vdash 4$	9	0,3000	0,1500
$4 \vdash 6$	12	0,4000	0,2000
$6 \vdash 8$	3	0,1000	0,0500
$8 \vdash 10$	2	0,0667	0,0333
Total	30	1,0000	

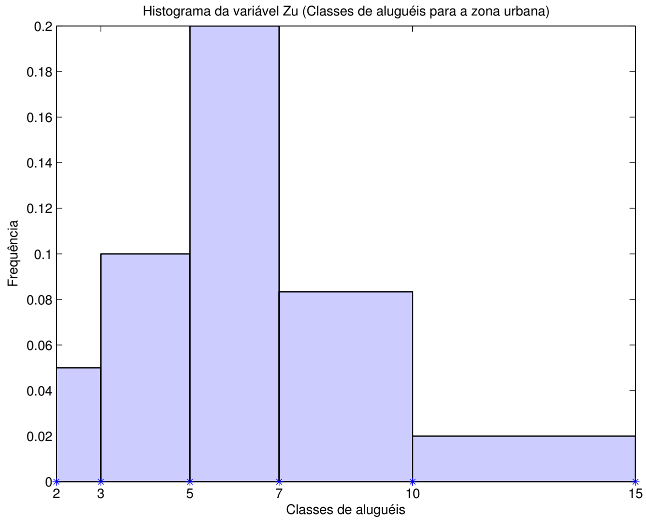
b) À partir dos resultados obtidos no **item a)** segue o histograma para a variável *Taxa de crescimento*.



7. a) Como em todos os exemplos anteriores, utilizamos a distribuição de frequências para facilitar a construção do histograma:

Classes de Aluguéis	**Frequência (n_i)**	**Frequência relativa (f_i)**	**Altura (h_i)**
$2 \vdash 3$	10	0,05	0,05
$3 \vdash 5$	40	0,20	0,10
$5 \vdash 7$	80	0,40	0,20
$7 \vdash 10$	50	0,25	0,08
$10 \vdash 15$	20	0,10	0,02
Total	200	1,00	

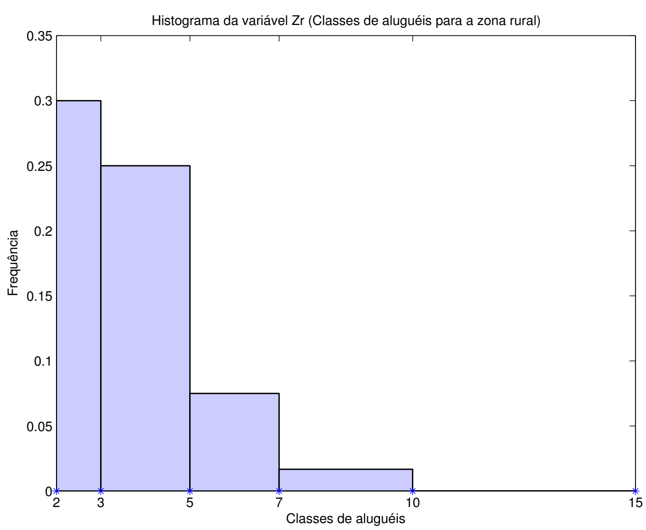
A diferença neste caso é que as classes não possuem mesma amplitude. Desse modo teremos $h_i = f_i/\Delta_i$ onde $\Delta_1 = 1$, $\Delta_2 = 2$, $\Delta_3 = 2$, $\Delta_4 = 3$ e $\Delta_5 = 5$. O histograma resultante segue abaixo:



De maneira análoga obtemos a distribuição de frequências para a Zona rural:

Classes de Aluguéis	**Frequência (n_i)**	**Frequência relativa (f_i)**	**Altura (h_i)**
$2 \vdash 3$	30	0,30	0,30
$3 \vdash 5$	50	0,50	0,25
$5 \vdash 7$	15	0,15	0,08
$7 \vdash 10$	5	0,05	0,02
$10 \vdash 15$	0	0,00	0,00
Total	100	1,00	

e o histograma resultante segue abaixo:



b) Pode-se observar claramente à partir dos histogramas que as classes de aluguéis para a zona urbana possuem distribuição bem mais simétrica em relação à zona rural, onde as classes de aluguéis decaem muito rapidamente à medida que aumentamos o seu valor. Outro ponto a se destacar é a frequência zero para classes maiores do que 10.

8. Para respondermos grande parte dos itens, basta construirmos a tabela de distribuição de frequências. Sendo x_i's os pontos médios dos intervalos de classe, segue abaixo a tabela:

x_i	f_i	$f_i x_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_i(x_i - \overline{x})^2$
3	0,56	1,68	-3,9	15,21	8,52
9	0,28	2,52	2,1	4,41	1,23
15	0,12	1,80	8,1	65,61	7,87
21	0,03	0,63	14,1	198,81	5,96
27	0,01	0,27	20,1	404,01	4,04
$\sum$	1,00	$\overline{x} = 6,90$			$\text{Var}(X) = 27,62$

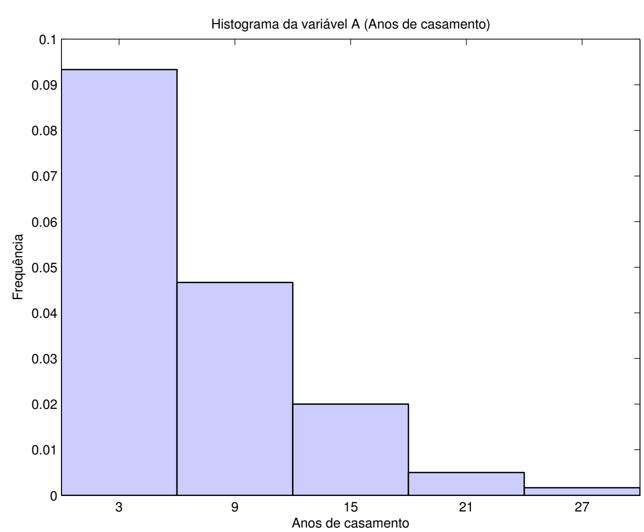
a) Pela tabela acima podemos ver que a média é $\overline{x} = 6,90$. Para o cálculo da mediana, note que a mesma está situada no primeiro intervalo de classe e assim aplicamos o princípio estudado sobre as alturas dos histogramas de tal forma que

$$\frac{56\%}{6-0} = \frac{50\%}{\tilde{x}-0}.$$

Resolvendo a equação obtemos $\tilde{x} = 5,35$.

b) Da tabela temos que $\text{Var}(X) = 27,62$, então $\mathcal{D}(X) = 5,26$.

c) Lembre-se que calculamos a altura dos blocos pela fórmula $h_i = f_i/\Delta_i$, onde Δ_i é a amplitude das classes. Em nosso caso, temos 5 classes de amplitude 6, ou seja, $\Delta_i = \Delta = 6$. As respectivas alturas são $h_1 = 0,093$, $h_2 = 0,047$, $h_3 = 0,020$, $h_4 = 0,005$ e $h_5 = 0,002$. O histograma segue abaixo:



d) Note que $q_1 = q(0,25)$ está situado no primeiro intervalo de classe e assim aplicamos o princípio estudado sobre as alturas dos histogramas de tal forma que

$$\frac{56\%}{6-0} = \frac{25\%}{q_1 - 0}.$$

Resolvendo a equação obtemos $q_1 = 2,68$. Agora, veja que $q_3 = q(0,75)$ está situado no segundo intervalo de classe e aplicando o mesmo princípio temos que

$$\frac{28\%}{12-6} = \frac{19\%}{q_3 - 6}.$$

Resolvendo a equação obtemos $q_3 = 10,07$.

e) Ao observamos o histograma já podemos verificar que os dados são assimétricos. Se desejarmos uma medida algébrica, basta notar que $\tilde{x} - q_1 = 2,67$ é bem menor que $q_3 - \tilde{x} = 4,72$.

9. a) Se cada observação é multiplicada por 2, as estatísticas de ordem permanecem nas mesmas posições, porém multiplicadas por 2 e, portanto, a mediana é $2\tilde{x}$.
A média será dada por

$$\frac{2x_1 + \cdots + 2x_n}{n} = 2\overline{x}.$$

O desvio padrão será

$$\sqrt{\frac{(2x_1 - 2\overline{x})^2 + \cdots + (2x_n - 2\overline{x})^2}{n}} = 2\sqrt{\frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}} = 2\mathcal{D}(X).$$

b) Somando-se 10 unidades à mediana, a mesma apenas será transladada assim como as outras estatísticas de ordem, mantendo a posição central e, portanto, seu valor é $10 + \tilde{x}$.
A média, por sua vez, será

$$\frac{(10 + x_1) + \cdots + (10 + x_n)}{n} = \frac{10n + x_1 + \cdots + x_n}{n} = 10 + \overline{x}.$$

Por fim, o desvio padrão será

$$\sqrt{\frac{(10 + x_1 - (10 + \overline{x}))^2 + \cdots (10 + x_n - (10 + \overline{x}))^2}{n}} = \mathcal{D}(X).$$

c) Por argumentos análogos aos itens anteriores a mediana será $\tilde{x} - \overline{x}$.
A média dos dados será

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x} = 0$$

O desvio padrão obviamente não será alterado, visto que a média acima é zero.

d) Após padronizarmos os dados teremos o conjunto $\left(\frac{x_1 - \overline{x}}{\mathcal{D}(X)}, \cdots, \frac{x_n - \overline{x}}{\mathcal{D}(X)}\right)$. A mediana será simplesmente dada por $\frac{\tilde{x} - \overline{x}}{\mathcal{D}(X)}$.
A média, como no caso anterior, será zero.

O desvio padrão, por sua vez, será

$$\sqrt{\frac{\left(\frac{x_1-\overline{x}}{\mathcal{D}(X)}\right)^2+\cdots+\left(\frac{x_n-\overline{x}}{\mathcal{D}(X)}\right)^2}{n}}=\frac{1}{\mathcal{D}(X)}\sqrt{\frac{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}}=1$$

10. a) Z é a nota dos funcionários com posição e escala corrigidas, ou seja, normalizada. Pelo **item d)** da **Questão 9** é de se esperar que a média seja zero e o desvio padrão seja 1.

    b) A nota de cada funcionário normalizada segue na tabela abaixo em ordem crescente segundo as colunas:

0,58	1,35	1,35	-0,95	-0,95
0,58	-0,18	-0,95	-0,18	-0,18
-0,18	-0,18	-0,95	0,58	1,35
-0,18	0,58	0,58	-3,26	0,58
0,58	-0,18	0,58	-0,95	0,58

    c) $\overline{z}=0$ e $\mathcal{D}(Z)=1$.

    d) Considerando o desvio padrão obtido, o intervalo em questão será $(-2,2)$ e, portanto, o funcionário 19 é um caso atípico, pois $z_{19}=-3,26$. Para os dados originais, sua nota foi 4,0 e está bem abaixo das demais.

    e) Para a variável *Direito* não temos variabilidade, então devemos observar as variáveis *Estatística* e *Política*. Para *Estatística* já calculamos sua nota normalizada no **item b)** e esta é dada por 0,58. A média da variável *Política* é $\overline{p}=7,76$ e o seu desvio padrão é $\mathcal{D}(P)=1,64$. Segue que a nota normalizada do funcionário em *Política* é 0,76 e esta representa seu melhor desempenho relativo.

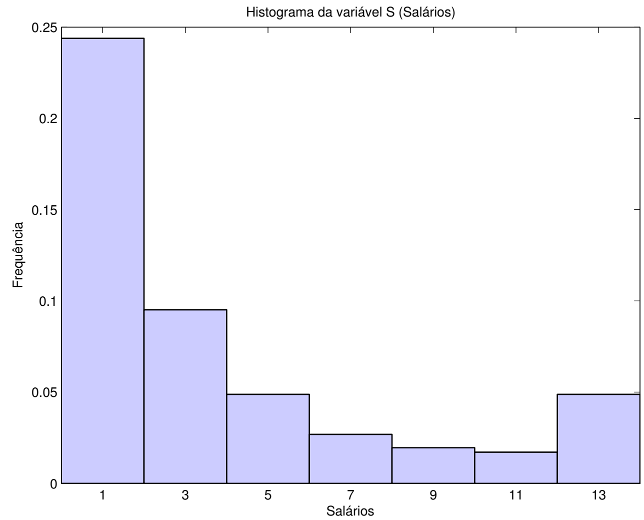
11. Primeiramente, observe que

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{\sum_{i=1}^n(x_i-\overline{x})^2}{n}=\frac{\sum_{i=1}^n\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right)}{n}\\
&= \frac{\sum_{i=1}^n x_i^2}{n}-2\overline{x}\frac{\sum_{i=1}^n x_i}{n}+\frac{\sum_{i=1}^n \overline{x}^2}{n}\\
&= \frac{\sum_{i=1}^n x_i^2}{n}-2\overline{x}^2+\overline{x}^2\\
&= \frac{\sum_{i=1}^n x_i^2}{n}-\overline{x}^2
\end{aligned}$$

Se separássemos os valores segundo suas frequências, teríamos

$$\mathrm{Var}(X)=\frac{\sum_{i=1}^k n_i x_i^2}{n}-\overline{x}^2=\sum_{i=1}^k f_i x_i^2-\overline{x}^2.$$

12. a) Temos sete intervalos de classe com amplitude $\Delta=2$ e as respectivas alturas dos blocos são $h_1=0,2439$, $h_2=0,0951$, $h_3=0,0488$, $h_4=0,0269$, $h_5=0,0195$, $h_6=0,0171$ e $h_7=0,0488$. O histograma segue abaixo:



b) Sendo x_i's os pontos médios dos intervalos de classe, a tabela com os valores necessários segue abaixo:

x_i	f_i	$f_i x_i$	x_i^2	$f_i x_i^2$
1	0,4878	0,4878	1	0,4878
3	0,1902	0,5706	9	1,7118
5	0,0976	0,4880	25	2,4400
7	0,0537	0,3759	49	2,6313
9	0,0390	0,3510	81	3,1590
11	0,0341	0,3751	121	4,1261
13	0,0976	1,2688	169	16,4944
$\sum$	1,0000	$\overline{x} = 3,9172$		31,0504

A média já está na tabela e seu valor é 3,9172. Pela fórmula encontrada na **Questão 11**, a variância é dada por $\text{Var}(X) = 31,0504 - (3,9172)^2 = 15,7059$ e, então, o valor do desvio padrão é 3,9631.

c) O Bairro A é mais homogêneo, pois tem desvio padrão menor.

d) Para encontrarmos a faixa dos 10% mais ricos basta encontrarmos o quantil $q(0,90)$. Note que este quantil está situado no sexto intervalo de classe e então

$$\frac{3,41\%}{12-10} = \frac{3,17\%}{q(0,90)-10}.$$

Resolvendo a equação encontramos $q(0,90) = 11,8592$ e a faixa salarial dos 10% mais ricos é $[11,8592\ ;\ 14,0000]$.

e) Para encontrarmos a riqueza total basta multiplicarmos a média pelo total de observações. No nosso caso, esse valor é aproximadamente 80.302,6.

13. a) $\overline{x}(0,10) = 10,84$ e $\overline{x}(0,25) = 10,52$.

b) A intenção do autor é eliminar a contribuição exagerada que valores atípicos proporcionam à média aritmética.

14. a) Ordenando ambas as variáveis, as medianas encontradas são

$$\tilde{x} = \frac{x_{(20)} + x_{(21)}}{2} = \frac{2+2}{2} = 2$$

e

$$\tilde{y} = \frac{y_{(20)} + y_{(21)}}{2} = \frac{2+3}{2} = 2,5.$$

Desse modo iremos considerar *baixa rotatividade* valores menores que 2 e *baixo salário* valores menores que 2,5. A distribuição conjunta das duas variáveis segue abaixo:

$Y \setminus X$	**Baixo**	**Alto**	**Total**
Baixo	1	19	20
Alto	7	13	20
Total	8	32	40

b) $1/40 = 2,5\%$.

c) $20/40 = 50\%$.

d) $1/8 = 12,5\%$.

e) Bastante modificada; observe que maioria das pessoas que ganham pouco tem alta rotatividade.

f) Parece haver uma associação entre *alta rotatividade* e *baixos salários.*

15. Para avaliar essa afirmação fazemos as tabelas com a distribuição relativa. Primeiramente, segue abaixo a tabela em relação ao total de colunas:

	Duração do efeito de dedetização			
Companhia	**Menos de 4 meses**	**De 4 a 8 meses**	**Mais de 8 meses**	**Total**
A	33%	35%	38%	34%
B	53%	51%	50%	52%
C	14%	14%	12%	13%
Total	100%	100%	100%	100%

Agora, temos a tabela em relação ao total de linhas:

	Duração do efeito de dedetização			
Companhia	**Menos de 4 meses**	**De 4 a 8 meses**	**Mais de 8 meses**	**Total**
A	32%	60%	8%	100%
B	35%	58%	7%	100%
C	34%	60%	6%	100%
Total	34%	59%	7%	100%

É fácil notar que não há diferença entre as 3 empresas.

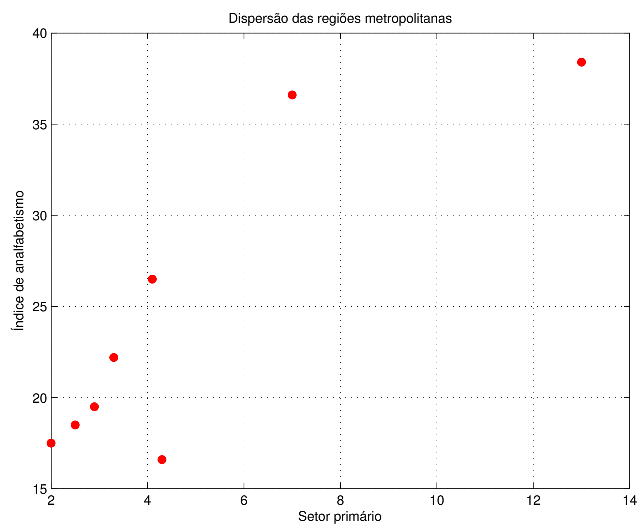
16. Lembrando, novamente, o resultado obtido na **Questão 11**, ou seja

$$\mathrm{Var}(x) = \frac{1}{n}\sum x_i^2 - \overline{x}^2,$$

temos que

$$\begin{aligned}
\frac{1}{n}\sum_i \left(\frac{x_i-\overline{x}}{\mathcal{D}(x)}\right)\left(\frac{y_i-\overline{y}}{\mathcal{D}(y)}\right) &= \frac{1}{n}\frac{\sum(x_iy_i - \overline{x}y_i - \overline{y}x_i + \overline{x}\,\overline{y})}{\sqrt{\left(\frac{1}{n}\sum x_i^2 - \overline{x}^2\right)\left(\frac{1}{n}\sum y_i^2 - \overline{y}^2\right)}} \\
&= \frac{1}{n}\frac{\sum x_iy_i - \overline{x}\sum y_i - \overline{y}\sum x_i + \sum \overline{x}\,\overline{y}}{\sqrt{\left(\frac{\sum x_i^2 - n\overline{x}^2}{n}\right)\left(\frac{\sum y_i^2 - n\overline{y}^2}{n}\right)}} \\
&= \frac{\sum x_iy_i - n\overline{x}\,\overline{y} - n\overline{y}\,\overline{x} + n\sum \overline{x}\,\overline{y}}{\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}} \\
&= \frac{\sum x_iy_i - n\overline{x}\,\overline{y}}{\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}}.
\end{aligned}$$

17. a) Abaixo segue o diagrama de dispersão do *Setor primário* versus o *Índice de analfabetismo* para as *Regiões metropolitanas* em questão:



b) O gráfico indica uma dependência linear positiva entre as variáveis.

c) Utilizando a equação obtida na **Questão 16**, o valor do coeficiente de correlação é aproximadamente 0,87.

d) Se observarmos a linha de tendência linear, as regiões de Porto Alegre e Fortaleza parecem fugir à curva. Calculando o coeficiente de correlação ao suprimir esses valores obtemos aproximadamente 0,99.