Instituto de Relações Internacionais
Estatística Aplicada I
Prof.ª Natalia Dus Poiatti
Monitores: Laura Gama Duarte
& Thiago Hideo Nakamura de Laurentis

Lista 01 - Gabarito

8. Intervalos de Classes Desiguais - É muito comum o uso de classes com tamanhos desiguais no agrupamento dos dados em tabelas de frequências. Nestes casos, deve-se tomar alguns cuidados especiais quanto à análise e construção do histograma.

A tabela abaixo fornece a distribuição de 250 empresas classificadas segundo o número de empregados. Uma análise superficial pode levar à conclusão de que a concentração vem aumentando até atingir um máximo na classe de 40-60, voltando a diminuir depois, mas não tão acentuadamente. Porém, um estudo mais detalhado revela que a amplitude da classe 40-60 é o dobro da amplitude das classes anteriores. Assim, espera-se que mais elementos caiam nessa classe, mesmo que a concentração seja levemente inferior. Então, um primeiro cuidado é construir a coluna que indica as amplitudes Δ_i de cada classe. Estes valores estão representados na terceira coluna da tabela.

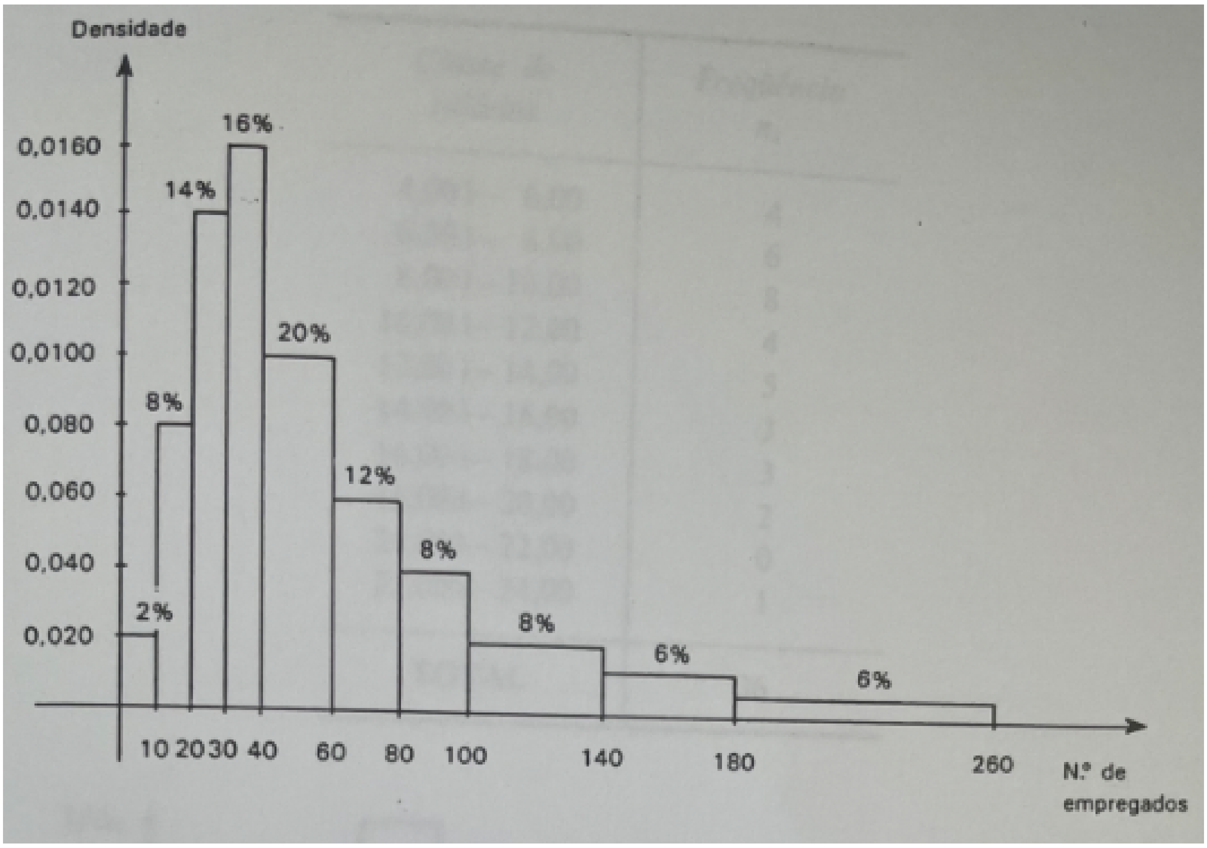
Distribuição de 250 empresas segundo o número de empregados

Número de empregados	*Freqüência* n_i	*Amplitude* Δ_i	*Densidade* n_i/Δ_i	*Proporção* f_i	*Densidade* f_i/Δ_i
0 ⊢ 10	5	10	0,50	0,02	0,0020
10 ⊢ 20	20	10	2,00	0,08	0,0080
20 ⊢ 30	35	10	3,50	0,14	0,0140
30 ⊢ 40	40	10	4,00	0,16	0,0160
40 ⊢ 60	50	20	2,50	0,20	0,0100
60 ⊢ 80	30	20	1,50	0,12	0,0060
80 ⊢ 100	20	20	1,00	0,08	0,0040
100 ⊢ 140	20	40	0,50	0,08	0,0020
140 ⊢ 180	15	40	0,38	0,06	0,0015
180 ⊢ 260	15	80	0,19	0,06	0,0008
TOTAL	250	—	—	1,00	—

Um segundo passo é a construção da coluna das densidades de frequências em cada classe que é obtida dividindo as frequências n_i pelas amplitudes Δ_i. Ou seja, à medida que indica qual a concentração por unidade da variável. Assim, observando-se os números da quarta coluna, vê-se que a classe de maior concentração passa a ser a 30-40, enquanto que a última é de menor concentração. Para compreender a distribuição, estes dados são muito mais informativos do que as frequências absolutas simplesmente.

De modo análogo, pode-se construir a densidade da proporção (ou porcentagem) por unidade da variável (verifique a construção através da 5° e 6° colunas). A interpretação para $f_i \,/\, \Delta_i$ é muito semelhante àquela dada para as sobre $n_i \,/\, \Delta_i$.

Para a construção do histograma basta lembrar que a área total deve ser igual a um ou 100%, o que sugere usar no eixo das ordenadas os valores de $f_i \,/\, \Delta_i$. O histograma para estes dados está na figura abaixo.



Não há nada para se resolver aqui, é só leitura.

9. Dispomos de uma relação de 200 aluguéis de imóveis urbanos e uma relação de 100 aluguéis rurais.

Classes de aluguéis (codificados)	*Zona urbana*	*Zona rural*
2 ⊢ 3	10	30
3 ⊢ 5	40	50
5 ⊢ 7	80	15
7 ⊢ 10	50	5
10 ⊢ 15	20	0
TOTAL	200	100

(a) Construa os histogramas das duas distribuições.

Primeiro, vamos construir as tabelas com as densidades. Em seguida, podemos construir o histograma.

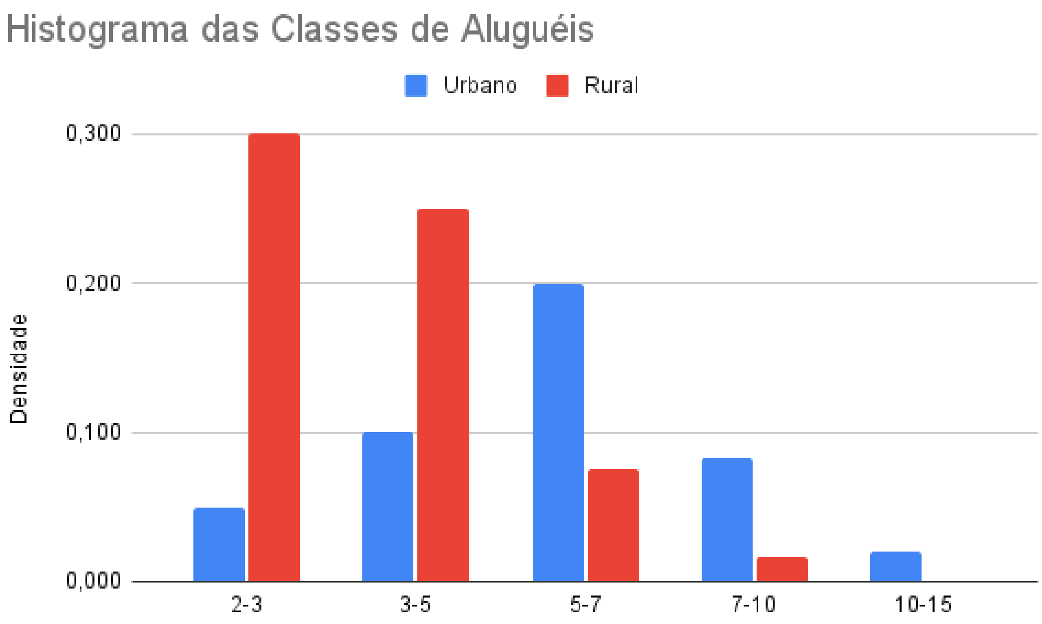
Zona Urbana

Classe de Aluguéis	Frequência n_i	Amplitude Δi	Densidade $n_i \,/\, \Delta i$	Proporção f_i	Densidade $f_i \,/\, \Delta i$
2-3	10	1	10	0,05	0,050
3-5	40	2	20	0,20	0,100
5-7	80	2	40	0,40	0,200
7-10	50	3	16,67	0,25	0,083
10-15	20	5	4	0,10	0,020
Total	200			1,00	

Zona Rural

Classe de Aluguéis	Frequência n_i	Amplitude Δi	Densidade $n_i \ / \ \Delta i$	Proporção f_i	Densidade $f_i \ / \ \Delta i$
2-3	30	1	30	0,30	0,300
3-5	50	2	25	0,50	0,250
5-7	15	2	7,5	0,15	0,075
7-10	5	3	1,67	0,05	0,017
10-15	0	5	0	0,00	0,000
Total	100			1,00	

O histograma da densidade $f_i \ / \ \Delta i$ por classes de aluguéis é



(b) Com base nos histogramas, discuta e compare as duas distribuições.

A densidade de frequência no gráfico da zona urbana demonstra que a classe de aluguéis predominante é 5-7. No caso das famílias provenientes da zona rural, a classe predominante de aluguéis é 2-3, demonstrando um valor mais baixo entre a população observada. É possível notar que a distribuição da densidade das classes de aluguéis é mais simétrica, em torno da classe 5-7 (onde se encontra a mediana), na zona urbana, e mais assimétrica para esquerda na zona rural.

1. Quer se estudar o número de erros de impressão de um livro. Para isso, escolheu-se uma amostra de 50 páginas, encontrando-se o seguinte número de erros por página:

Erros	*Freqüência*
0	25
1	20
2	3
3	1
4	1

(a) Qual o número médio de erros por página?

Seja x_i o número de erros na página i da nossa amostra. A média amostral de erros nessa amostra de 50 páginas é

$$\overline{X} = \frac{1}{50}\sum_{i=1}^{50} x_i = \frac{1}{50}\Big[(25)(0) + (20)(1) + (3)(2) + (1)(3) + (1)(4)\Big] = 0,66.$$

(b) E o número mediano?

Vamos ordenar x_i do menor para o maior, tal que $x_i < x_{i+1}$. Assim, $x_{50} > x_{49} > \ldots > x_2 > x_1$. Assim, como o tamanho da amostra é 50, ou seja, é par, a mediana amostral é dada por

$$md_x = \frac{x_{50/2} + x_{(50/2)+1}}{2} = \frac{x_{25} + x_{26}}{2} = \frac{0+1}{2} = 0,5.$$

(c) Qual é o desvio padrão?

O desvio padrão amostral é dado por

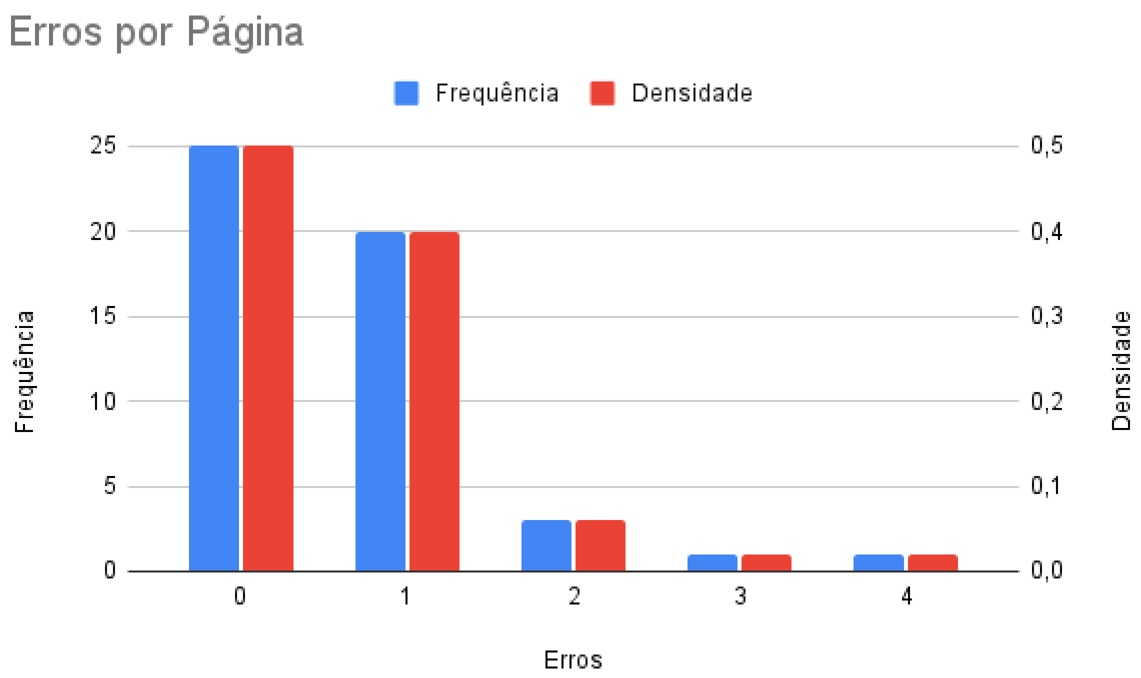
$$S_x = \sqrt{\frac{1}{50-1}\sum_{i=1}^{50}(x_i - \overline{X})^2} =$$

$$= \frac{1}{7}\sqrt{25(0-0,66)^2 + 20(1-0,66)^2 + 3(2-0,66)^2 + (3-0,66)^2 + (4-0,66)^2} =$$

$$= \frac{1}{7}\sqrt{35,22} \approx 0,8478.$$

(d) Faça uma representação gráfica para a distribuição.

Note que a frequência e a densidade nos retornam informações equivalentes.



(e) Se o livro tem 500, qual o número total de erros esperado no livro?

Partindo do pressuposto de que a nossa amostra é não viesada (por exemplo, se escolhermos só páginas com imagens, isto é, haveria menos palavras por página, subestimando o número de erros), espera-se que encontremos 330 erros nas 500 páginas (lembre-se que a média amostral de erros é 0,66 por página).

2. As taxas de juros recebidas por 10 ações durante um certo período foram (medidas em porcentagem) 2,59; 2,64; 2,60; 2,62; 2,57; 2,55; 2,61; 2,50; 2,63; 2,64. Calcule a média, a mediana e o desvio padrão.

Aqui não se faz a distinção de amostra e população. Assim, não diferenciaremos o cálculo do desvio padrão, podemos fazer de ambas as formas, amostral e populacional. Denotamos a taxa de juros como r.

Se considerarmos os dados como uma amostra, então,

- a média amostral é dada por

$$\overline{r} = \frac{1}{10}(2,59 + 2,64 + 2,60 + 2,62 + 2,57$$

$$+2,55 + 2,61 + 2,50 + 2,63 + 2,64) = 2,595;$$

- ordenando a amostra (2,5; 2,55; 2,57; 2,59; **2,6; 2,61**; 2,62; 2,63; 2,64; 2,64), vemos que a mediana amostral é

$$md_r = \frac{2,6 + 2,61}{2} = 2,605;$$

- a variância amostral é

$$S_r^2 = \frac{1}{10-1}\Big[(2,5 - 2,595)^2 + (2,55 - 2,595)^2 + (2,57 - 2,595)^2$$

$$+(2,59 - 2,595)^2 + (2,6 - 2,595)^2 + (2,61 - 2,595)^2 + (2,62 - 2,595)^2$$

$$+(2,63 - 2,595)^2 + (2,64 - 2,595)^2 + (2,64 - 2,595)^2\Big] \approx 0,001983;$$

- e o desvio-padrão amostral é

$$S_r \approx \sqrt{0,001983} \approx 0,0445.$$

Se tratarmos os dados como nossa população de interesse (a média e a mediana ficam iguais), a variância populacional é

$$\sigma_r^2 = \frac{1}{10}\Big[(2,5-2,595)^2+(2,55-2,595)^2+(2,57-2,595)^2$$

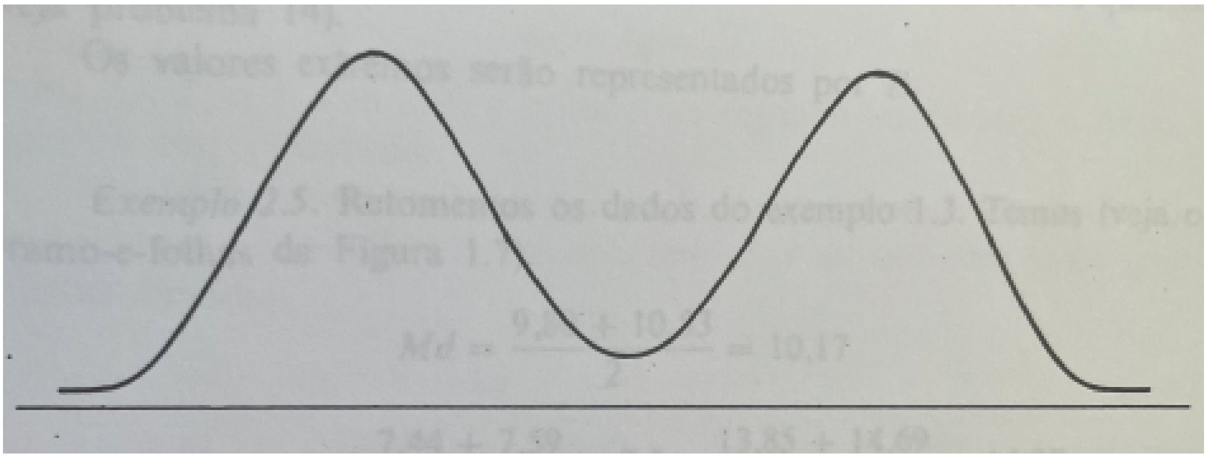
$$+(2,59-2,595)^2+(2,6-2,595)^2+(2,61-2,595)^2+(2,62-2,595)^2$$

$$+(2,63-2,595)^2+(2,64-2,595)^2+(2,64-2,595)^2\Big] \approx 0,001785.$$

e o desvio padrão populacional é

$$\sigma_r \approx \sqrt{0,001785} \approx 0,0422.$$

5. Suponha que a variável abaixo tenha a distribuição como na figura abaixo (suponha que é simétrica - a página escaneada ta torta).



Você acha que a média é uma boa medida de posição? E a mediana? Justifique.

Observe que só com a média e a mediana não dá para saber muita coisa da nossa variável de interesse, dado que essas medidas de posição pouco informam sobre a distribuição, não sendo possível identificar se a distribuição é simétrica, nem o quão dispersa é.

Note que uma medida de posição que traria mais informação acerca da distribuição da variável de interesse seria a moda (no caso, se for simétrica), uma vez que, sabendo que há duas modas (a isto damos o nome de distribuição bimodal), saberíamos que haveria dois picos na distribuição (ainda não seria possível aferir nada sobre o quão simétrica é, nem sobre sua dispersão).

6. Numa pesquisa realizada com **100 famílias**, levantaram-se as seguintes informações:

Número de filhos	0	1	2	3	4	5	mais que 5
Freqüência de famílias	17	20	28	19	7	4	5

(a) Qual é a mediana do número de filhos?

Denotamos x_i o número de filhos da família i. Se ordenarmos a amostra, então, $x_{50} = 2$ e $x_{51} = 2$, ou seja, a mediana é 2. Note que da família 38 até a 55, todas têm dois filhos.

(b) E a moda?

A moda é 2, dado que a maior frequência é de famílias que têm 2 filhos.

(c) Que problemas você enfrentaria para calcular a média? Faça alguma suposição e encontre-a.

O problema é não saber quanto é "mais que 5" especificamente, dado que precisamos dessa informação para computar a média. Supondo que "mais que 5" seja 6, então,

$$\overline{X} = \frac{1}{100}\Big[0(17) + 1(20) + 2(28) + 3(19) + 4(7) + 5(4) + 6(5)\Big] = 2,11,$$

ou seja, cada família tem em média 2,11 filhos.