Capítulo 6
População e amostra

Este capítulo constroi a ponte entre a parte do livro que trata da Teoria de Probabilidade e a que trata da Estatística. Quem trafega pela ponte, fazendo a ligação, é o conceito de amostra. Mas a construção dos pilares da ponte exige a introdução de muitos outros conceitos. Esses conceitos é o cerne do presente capítulo.

Cabe aquí um comentário sobre a posição do presente capítulo entre os outros capítulos do livro. Por um lado, o capítulo deve preceder o Teorema Central de Limite e, consequentemente, todos os métodos estatísticos a serem ensinados no livro que usam tal teorema. A obrigação da referida precedência motiva-se pelo fato que os conceitos de população e de distribuição populacional apresentados no presente capítulo são imprescindível na discussão dos aspectos práticos do Teorema de Limite Central. Já falando do material do livro que deve anteceder ao presente capítulo, ele precisa contar ao leitor sobre variáveis aleatórias e conceitos subjacentes pois tudo isso é imprescindível para a compreenção da relação entre população e amostra, que é um dos principais assuntos do presente capítulo.

Vale também notar que uma das vantagem do posicionamento desse capítulo na sequencia de apresentação do livro é o ênfas que consigo dar à distinção entre população e amostra, e, consequentemente, à toda a lógica sobre a estimação da segunda pela primeira, o que por sua vez, esclarece a necessidade e a construção de estimadores.

6.1 População, distribuição populacional, sua caraterização e apresentação

6.1.1 Definição de conceitos primordiais

Não achei as definições cirtas, objetivas e rigorosas dos conceitos primordiais que aquí preciso. Mas achei um exemplo que dá a noção indúbia de tais conceitos. Ele está abaixo e os conceitos estão destacados em negrito.

Exemplo 58 emprestado por mim do livro "Elementary Statistics" escrito por Ron Larson e Betsy Farber (Prentice Hall, Inc. 2003).

Num remoto vilarejo da Alaska, chamado Akhiok, cuja imagem você vé na Figura 1.1 há exatamente 77 habitantes; as fotos de alguns deles estão na Figura 1.2.

Figura 6.1: A foto de algumas casa do vilarejo Akhiok. É aqui que residem os 77 moradores sobre os quais falamos no Exemplo 1 e em exemplos subsequentes.

Essas 77 pessoas fornecem uma excelente exemplo para podermos falar de população e de suas caraterísticas. Então, em primeiro lugar, notamos que essas pessoas podem ser chamadas de **população**, pois fica claro que são os moradores daquele vilarejo e nada mais e nada menos.

Imagine que perguntamos e anotamos a idade de cada morador. Especificamente, imagine que encontramos com cada morador e perguntamos sua idade. Nesse caso, assim como na maioria de casos que aparecerão no texto, não exporemos e não discutimos a maniera que foi seguida para garantir que perguntarmos de todos mesmo. Podemos imaginar que usamos a ordem alfabética.

A idade aquí é um excelente exemplo daquilo que chama-se de **atributo**; outros possíveis atributos são: altura, peso, sexo, a informação se nasceu em Akhiok (respondida por sim/não).Esses são exemplos de atributos que os indivíduos da população de Akhiok possuem. Para outras populações existem outros atributos. Isso é óbvio se comprendemos "atributo de populaçã" como uma caraterística qualquer que cada individuo da população possui. Vamos ficar com essa compreenção.

Na continuação, vamos trabalhar com o arrendondamento por inteiros das idades dos maradores de Akhiok. O motivo para o arredondamento é só obter certa comodidade na exposição.

Figura 6.2: As fotos de alguns de moradores do vilarejo Akhiok.

Esses valores estão apresentados na Figura 1.3. Já na Figura 1.4 temos os mesmos valores após sua ordenação. Prosseguiremos com os valorer ordenados, fato que descarta a necessidade de descrever a ordem da apresentação do conjunto não ordenado. Mas, se deseja saber, posso lhe informar que a ordem seguida é alfabética.

28, 6, 17, 48, 63, 47, 27, 21, 3, 7, 12,
39, 50, 54, 33, 45, 15, 24, 1, 7, 36, 53,
46, 27, 5, 10, 32, 50, 52, 11, 42, 22, 3,
17, 34, 56, 25, 2, 30, 10, 33, 1, 49, 13,
16, 8, 31, 21, 6, 9, 2, 11, 32, 25, 0,
55, 23, 41, 29, 4, 51, 1, 6, 31, 5, 5,
11, 4, 10, 26, 12, 6, 16, 8, 2, 4, 28

Figura 6.3: As idades de todos os 77 moradores de Akhiok (em ordem alfabética) arredondados por inteiros e apresentados em ordem alfabética.

0, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4,
4, 5, 5, 5, 6, 6, 6, 6, 7,7, 8,
8, 9, 10, 10, 10, 11, 11, 11, 12, 12, 13,
15, 16, 16, 17, 17, 21, 21, 22, 23, 24, 25,
25, 26, 27, 28, 28, 28, 29, 30, 31, 31, 32,
32, 33, 33, 34, 36, 39, 41, 42, 45, 46, 47,
48, 49, 50, 50, 51, 52, 53, 54, 55, 56, 63

Figura 6.4: As idades de todos os 77 moradores de Akhiok (em ordem alfabética) arredondados por inteiros e apresentados em ordem alfabética.

O conjunto de dados ordenados (chamado também conjunto ordenado de dados) pode ser representado via tabela, cuja única vantagem sobre a apresentação como conjunto é que cada

valor aparece uma vez só, mas junto com seu contador que reflete o número de vezes que este valor está repetido no conjunto (este contador chama-se, naturalmente, por **frequencia absoluta**):

0	1	2	3	4	5	6	7	8	9	10	11
1	3	3	2	3	3	4	2	2	1	3	3
12	13	15	16	17	21	22	23	24	25	26	27
2	1	1	2	2	2	1	1	1	2	1	2
28	29	30	31	32	33	34	36	39	41	42	45
2	1	1	2	2	2	1	1	1	1	1	1
46	47	48	49	50	51	52	53	54	55	56	63
1	1	1	1	2	1	1	1	1	1	1	1

A tabela poderia ter sido feita para conjunto não ordenado, porém a ordenação ajuda à visulaização do conjunto de dados e ao cálculo de suas caraterísticas.

O contador (a contagem) pode ser feita por **frequencias absolutas**, como na tabela acima, ou por **frequencias relativas**, como na de baixo:

0	1	2	3	4	5	6	7	8	9	10	11
$\frac{1}{77}$	$\frac{3}{77}$	$\frac{3}{77}$	$\frac{2}{77}$	$\frac{3}{77}$	$\frac{3}{77}$	$\frac{4}{77}$	$\frac{2}{77}$	$\frac{2}{77}$	$\frac{1}{77}$	$\frac{3}{77}$	$\frac{3}{77}$
12	13	15	16	17	21	22	23	24	25	26	27
$\frac{2}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{2}{77}$	$\frac{2}{77}$	$\frac{2}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{2}{77}$	$\frac{1}{77}$	$\frac{2}{77}$
28	29	30	31	32	33	34	36	39	41	42	45
$\frac{2}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{2}{77}$	$\frac{2}{77}$	$\frac{2}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$
46	47	48	49	50	51	52	53	54	55	56	63
$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{2}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$

A tabela da distribuição de frequencia relativa perde uma informação em relação com a que está na tabela da distribuição de frequencia absoluta. De fato, ao receber o valor $\frac{3}{77}$ referente a idade 1, não saberemos se há 3 crianças com a idade 1 na população de total de 77 pessoas, ou se são 30 crianças com esta idade na população de 770 pessoas. Entretanto, as vezes, o que interessa é só a distribuição de frequencia relativa. E se um estatístico foi chamado para trabalhar com dados, ele pode pedir a tabela de frequencia relativa junto com o **tamanho da população**. Assim nada será perdido.

As tabelas chamam-se, em Inglês **frequency distribution**, o que traduz-se como **a distribuição de frequencia**. Ao ouvir esse termo, qualquer um perguntaria: "A frequência está distribuida sobre o que?" A resposta é: "Sobre os valores observados do atributo "idade" na população do vilarejo Akhiok." Isso explica que os nomes completos para as tabelas seriam assim:

distribuição da frequencia absoluta sobre os valores observados do atributo "idade" na população do vilarejo Akhiok

e

distribuição da frequencia relativa sobre os valores observados do atributo "idade" na população do vilarejo Akhiok

só que no dia-a-dia esses termos são encurtadas. Sobre isso falo no parágrafo abaixo.

Confesso que se alguém apresentasse para mim a primeira tabela, contasse tudo acerca da maneira como a mesma foi obtida e pedisse de mim sugerir o nome, eu diria: **a tabela que apresenta a distribuição de idade pela população dos moradores de Akhiok**. O nome profano que dei – e creio não estou sozinho nessa abordagem cotidiana – insinua que é o atributo

que está sendo distribuido. Em contraste com isso, o nome oficial insinua que é a frequencia que está sendo distribuida. Vive com as duas, mas use aquela que é oficial quando for escrever um documento oficial.

Meu último aviso sobre a nomenclatura é assim: vou usar a seguinte versão encurtada **distribuição populacional**. Combino com você que está subentendido que ela é de uma frequencia sobre um atributo. Ainda mais, sempre terei em mente a frequencia relativa, pois a frequencia absoluta nos interessa muito menos, e quando vier a interessar, serei explicito sobre isso.

6.2 O que desejamos fazer com distribuições populacionais

Ao receber/construir uma distribuição populacional, o objetivo é identificar suas caraterísticas e/ou propriedades. Quais? – perguntaria você. A resposta é: Aquelas que podem ser úteis. Felizmente, não somos nós quem vai definir o que é util e que não é. Usaremos aqueles que a Estatística Teórica desenvolveu durante mas que 100 anos de sua vida. Logo, o objetivo da seção é apresentar e mostrar aplicação de certos simples métodos que permitem analizar distribuições populacionais.

6.2.1 Função de distribuição acumulada

Todos os métodos dos quais conversaremos nas seções proximas ofuscam certas propriedades de distribuições populacionais em prol de destacar uma outras. Isso fez necessário falar agora sobre uma ferramenta específica que não omite nada sobre as distribuições. Ela chama-se **função de distribuição acumulada**.

Começamos a exposição mostrando como a distribuição de frequencia pode ser representada por gráfico do tipo de "função de probabilidade". Abaixo está o gráfico da distribuição de frequencias relativa do Exemplo 1.

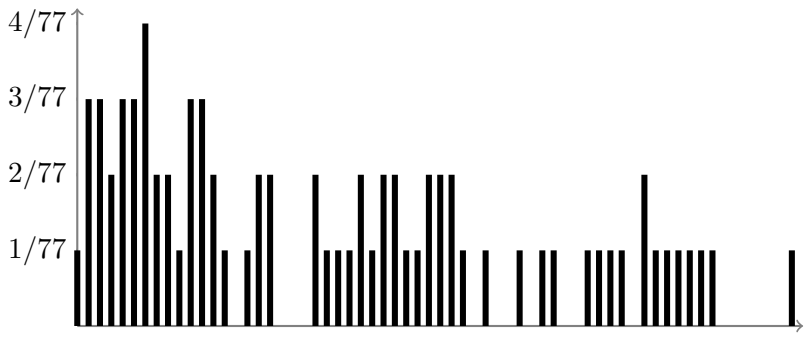


Figura 6.5: Apresentação da distribuição da freq. relativa por atributo "idade" em forma de Função de Probabilidade.

Essa seção será complatada no futuro. Seu assunto principal (a função de distribuição acumulada) não será cobrada em exercícios e provas.

6.2.2 Apresentação por histogramas e sua utilização

A ideia a ser realizada nessa seção é agrupar os valores de atributo, calcular as frequencias relativas de classes e compará-las. Uma das maneiras comodas de comparação é a visual.

Para fornecer a visualização útil e eficiente inventou-se histograma. Informalmante falando, histograma apresenta as frequencias em forma de retângulos, sendo que a base de cada retângulo está posicionada em cima da correspondente classe.

1.2.2.1. Observações relevantes sobre a construção. Vou falar dos histogramas por frequencia relativa. Aqueles, que são por frequencia absoluta, são muito menos usados, e você conseguira entedé-los se enteder direitinho so que serão apresentados por mim.

No histograma as frequencias relativas estão apresentadas por área, diferentemente daquilo que acontece no caso de gráfico da função de probabilidade, onde as frequencias estão apresentadas pela altura do "palito" como você viu na Figura 1.5.

Histogramas serão então desenhados no espaço euclidiano $\mathbb{R}^2$. Seu eixo horizontal representa os valores do atributo em interesse. Já o eixo vertical não possui interpretação imediata. Nos desenhos a seguir esse eixo é chamado "densidade". Esse nome será explicado mais tarde.

1.2.2.2. As regras de construção de histogramas via exemplos. Na Figura 1.6 está um dos possiveis histogramas feito para o conjunto de dados que apresenta as idades dos moradores de Akhiok. Os cálculos auxiliares que determinam seu formato estão no texto e também na legenda da figura.

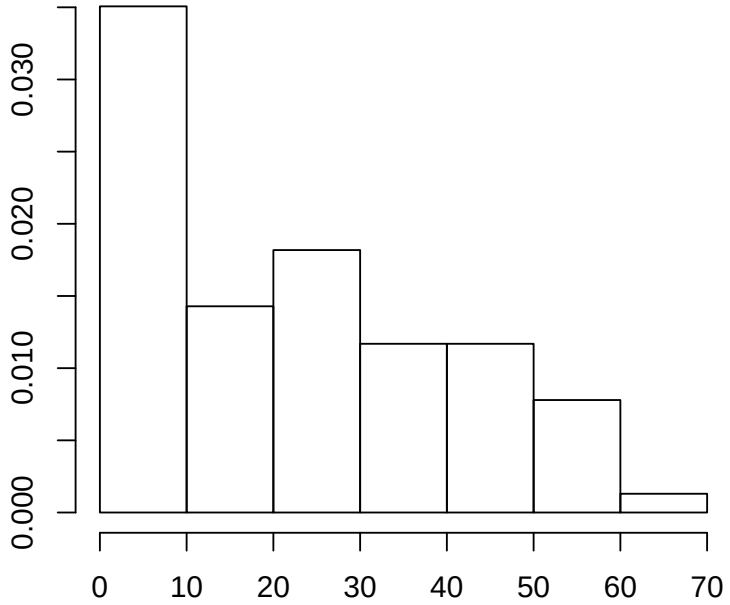


Figura 6.6: O histograma para a distribuição de frequencia relativa por atributo idade na população dos moradores de Akhiok. O correspondende conjunto de dados (idades dos moradores, no caso) foi separado por 0, 10, 20, 30, 40, 50, 60, 70. As correspondentes classes são $[0, 10]$, $(10, 20]$, $(20, 30]$, $(30, 40]$, $(40, 50]$, $(50, 60]$, $(60, 70]$. As alturas dos correspondentes "prédios" do histogram são, de esquerda para a direita, 0.035064935, 0.014285714, 0.018181818, 0.011688312, 0.011688312, 0.007792208, 0.001298701. O cálculo das alturas está apresentado pelas fórmulas (1.4).

A construção do histograma segiu aos seguintes passos. Em primeiro lugar, o conjunto de dados foi separado em seguintes classes

$$[0, 10],\ (10, 20],\ (20, 30],\ (30, 40],\ (40, 50],\ (50, 60],\ (60, 70] \tag{6.1}$$

Os separadores são chamados **separadores de classes**, e, no caso, eles têm os seguintes valores: $0, 10, 20, 30, 40, 50, 60, 70$. Observo que quando um valor do conjunto de dados coincidir com um dos separadores, surge a questão: o valor deve ser atribuído à classe à esquerda ou à da

direita? A regra de atribuição depende do objetivo de toda a construção. No presente texto, sou eu quem vai ditar essa regra. Minhas indicações usarão a notação "(" e "]" da maneira que você vé em (1.1). Em todos os casos, não darei justificativa pois ela não acrescenta nada de importante.

A quantidade de valores que pertencem a cada classe (respeitando a indicação "(" e "]" usada para definir as classes conforme se vê na Eq. (1.1)) chamam-se **frequencias absolutas**, ou **counts** em Inglês. Elas são:

$$27, 11, 14, 9, 9, 6, 1 \tag{6.2}$$

Elas seguem por direta contagem. A contagem é simples se for usado a conjunto ordenado de dados.

As frequencias absolutas são úteis para o cálculo de **frequencias relativas**, as quais ficam um passo só até a obtenção das alturas dos prédios do histograma; a relação usa o conceito óbvio **amplitude de classe** e é assim:

$$\text{frequencia relativa} = \frac{\text{frequencia absoluta}}{\text{quantidade de dados}}$$

No caso considerado, a quantidade de dados é 77. Por aplicação essa fórmula temos, portanto, as seguintes frequencias relativas; a apresentação segue a ordem das correspondentes classes:

$$27/77{=}0.35064935,\ 11/77{=}\ 0.14285714,\ 14/77{=}0.18181818,\ 9/77{=}0.11688312,$$
$$9/77{=}\ 0.11688312,\ 6/77{=}\ 0.07792208,\ 1/77{=}0.01298701$$

Tudo que fizeram até o momento nos serve para calcular as alturas dos prédios, informação que permite completar o desenho que chamamos por histograma. A fórmula é assim

$$\text{altura do prédio apoiado por classe} = \frac{\text{frequencia relativa da classe}}{\text{amplitude de classe}} \tag{6.3}$$

No caso considerado, todas as classes têm a aplitude 10. Portanto, a aplicação da fórmula de cima dá:

$$\begin{aligned} &\frac{0.35064935}{10} = 0.035064935,\ \frac{0.14285714}{10} = 0.014285714,\ \frac{0.18181818}{10} = 0.018181818,\\ &\frac{0.11688312}{10} = 0.011688312,\ \frac{0.11688312}{10} = 0.011688312,\ \frac{0.07792208}{10} = 0.007792208,\\ &\frac{0.01298701}{10} = 0.001298701 \end{aligned} \tag{6.4}$$

Com essas alturas, construímos os prédios do histograma.

1.2.2.3. A propriedade princial de histogramas. É de extrema importânica para as futuras exposições que sejam destacadas a seguinte proposição a seus corolários e derivados.

Proposição 19 *(sobre uma propriedade central de histogramas).*
(a) - *a propriedade assegurada pela propria construção.* O prédio elevado em cima de qualquer das classes dum histograma tem sua área igual à frequencia relativa dos valores de atributo que encontram-se nessa classe.
(b) - *uma consequencia direta do* (**a**). A frequencia relativa dos valores de atributo que pertencem a um conjunto qualquer de classes é igual à soma das áreas dos prédios de histograma correspondentes às classes do conjunto.

Uma tradicional suposição usada na interpretação de histogramas. É muito raro que um estatístico receba um histograma sem ser acompanhado pelo conjuto de dados para qual foi cosntruido, mas quando isso acontece, o estatístico não tem como saber algo acerca da distribuição de valores em cada classe do histograma; se tal informação for necessária, as vezes assume-se que os valores estão uniformamente espalhados pelo intervalo de classe.

Corolário ao Prop. 1. Ao juntar a proposição com a tradicional suposição acima formuladas, concluí-se o seguinte: Se $[a, b]$ for qualquer intervalo então a frequencia relativa dos valores de atributo que encontram-se entre a e b é a área do histograma cuja base é $[a, b]$ e cuja altura está composta dos teros dos prédios do histograma que encontram-se entre as linhas verticais $x = a$ e $x = b$ (observe: não é necessariamente que a e/ou b sejam separador de classes).

Interpretação de histograma como densidade. O corolário mostra que os tetos dos prédios de histograma, sendo considerados como uma curva contínua, determina as frequencias relativas dos valores de atributo da mesma maneira que uma função-densidade determina probabilidades de uma variável aleatória contínua. Esse é o motivo de chamar o perfil de tetos por **densidade da frequencia relativa**. Naturalmente o nome veio pela assicoação e essa baseia-se no fato que probabilidades e frequencias relativas são parentes conforme impusemos na nossa definição do conceito Probabilidade.

1.2.2.4. Outros exemplos de histograma com ilustração de aplicação e propriedade princial de histogramas. Agora veremos outros histogramas, todos para aquele mesmo conjunto de dados. Suas construções não trazem novidades em relação daquilo que foi lhe ensinado na construção do histograma da Figura 1.6. Os resultados estão nas Figuras 1.7–1.9.

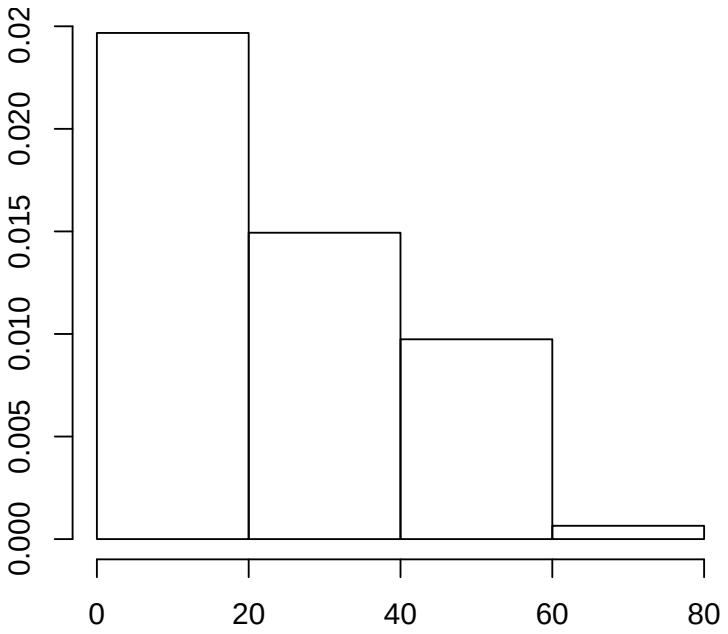


Figura 6.7: O histograma para a distribuição de frequencia relativa por atributo idade na população dos moradores de Akhiok. O correspondende conjunto de dados (idades dos moradores, no caso) foi separado por 0, 20, 40, 60, 80. As correspondentes classes são $[0, 20]$, $(20, 40]$, $(40, 60]$, $(60, 80]$, e suas frequencias absolutas são $38, 23, 15, 1$. As alturas dos correspondentes "prédios" do histogram são, de esquerda para a direita: $\frac{38}{77\times 20} = \frac{0.4935065}{20} = 0.0246753247, 0.0149350649, 0.0097402597, 0.0006493506$.

Comentário 1 *exemplos de aplicação de histograma.* O histograma por classes da amplitude 20 sugere que as quantidades de pessoas por faxas etárias de 20 em 20 anos diminui com

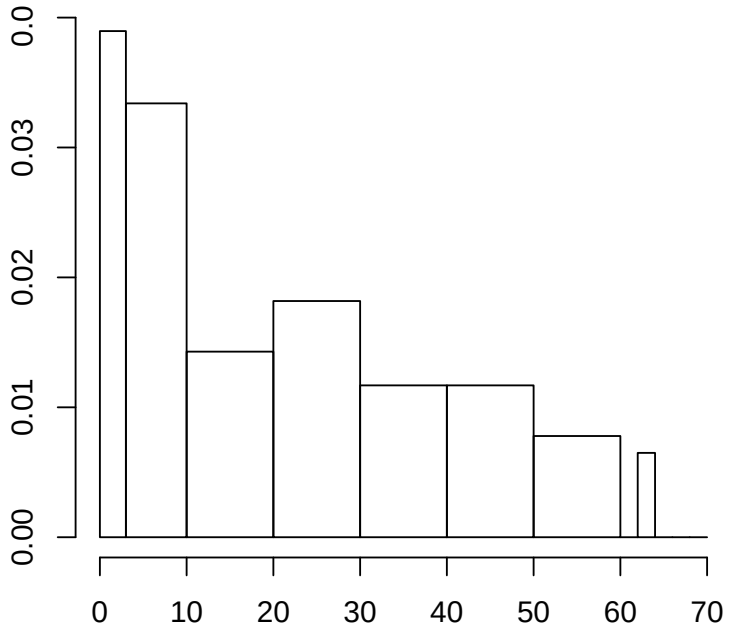


Figura 6.8: O histograma para a distribuição de frequencia relativa por atributo idade na população dos moradores de Akhiok. O correspondende conjunto de dados (idades dos moradores, no caso) foi separado por 0, 3, 10, 20, 30, 40, 50, 60, 62, 64, 66, 68, 70. As frequencias absolutas das classes correspondentes são: 9, 18, 11, 14, 9, 9, 6, 0, 1, 0, 0, 0. As alturas dos correspondentes prédios são $0.038961039, 0.033395176, 0.014285714, 0.018181818, 0.011688312, 0.011688312, 0.007792208,$ $0.000000000, 0.006493506, 0.000000000, 0.000000000, 0.000000000$. Observe que as amplitudes de classes são diferentes. Por exemplo, a da primeira classe é 3 e isso repercurte no cálculo da primeira altura assim: $9/(3 \times 77) = 0.038961039$.

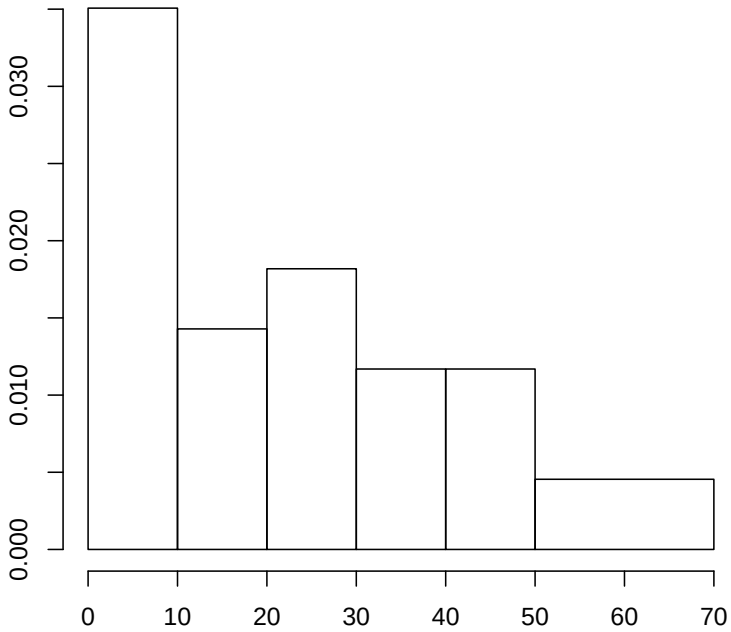


Figura 6.9: O histograma para a distribuição de frequencia relativa por atributo idade na população dos moradores de Akhiok. O correspondende conjunto de dados (idades dos moradores, no caso) foi separado por $0, 10, 20, 30, 40, 50, 70$; sua escolha foi motivada no Comentário 2.

quase que o mesmo coeficiente de "evasão"(morte). Esta é uma sugestão. A verificação dessa

é assunto de métodos quantitativos de Estatística.

O historgrama por classes da amplitude 10 não indica claramente o fenômento de diminuição sugerido pelo histograma com amplitudes 20. Aquele histograma só mostra a diminuição, mas não indica que taxa de evasão/mortalidade possa ser a mesma, se for contada a cada 20 anos.

Em compensação, o histograma por classes da amplitude 10 mostra que *(a)* a proporção dos moradores com idade na faxa 0–10 despara muito acima de todas as outras frequencias (calculadas por faxa de 10 anos), e que *(b)* a proporção das pessaos na faxa entre 30 e 40 anos é a mesma que a na faxa entre 40 e 50. Isso não dá para ver no histograma por classes de aplitude 20.

O historgrama por classes da amplitude 10 também mostra que há menos pessoas na faxa etária 10-20 que as na faxa 20-30. O histograma não revela se isto é algo intrinsico, ou se tivemos azar de analizar a população no momento quando as pessoas que estariam distribuidas por igual na faxa entre 19 e 21 ano ficaram "deslocadas" para 21. O deslocamento do separador 20 poderia revelar a razão.

Comentário 2 *sobre histogramas por classes desiguais.* Existe infinitude de razões para construir histogramas com classes de amplitudes desiguais. Por exemplo, no caso do conjunto de dados sobre as idades dos moradores de Akhiok, queremos ver as pessoas mais jovens (até 10 anos) separadas por duas classes (até 3 anos e entre 3 e 10), e achamos que uma pessoa com 63 anos não pode representar classe na faxa $(60, 70]$. Seja por estas razões, ou seja por outras qualquer, escolhemos os seguintes separadores:

$$0,\ 3,\ 10,\ 20,\ 30,\ 40,\ 50,\ 60,\ 62,\ 64,\ 66,\ 68,\ 70$$

Desconheço regras rígidas para a escolha de amplitudes para calsses de separação. Por exemplo, há quem acha que a única pessoa com a idade na faxa 60-70 deve ser juntada com as da classe anterior, isto é, que os separadores devem ser assim:

$$0, 10, 20, 30, 40, 50, 70$$

Isto dá o histograma apresentada na Figura 1.9.

Você precisa dar atenção especial e redobrada à construção de histogramas com amplitudes desiguais. É muito comum que alunos erram na tal construção. O erro típico é é apresentar a frequencia relativa de classe pela altura do seu "prédio". O correto é calcular a altura da maneira tal que a área do prédio seja igual à frequencia relativa.

Comentário 3 *sobre os dispensados.* Além de histograma, existem outras maneiras gráficas para visualização e apresentação de conjuntos de dados (por exemplo, pizza, diagrama de barras, etc.) Algumas são superadas devido ao avanço do desenho gráfico de programas de computador, outras ainda são a vir. Você vai facilmente aprender qualquer de tais maneiras se e quando for necessário. Eu prefiro não gastar o tempo de minhas aulas para discutí-las.

6.2.3 Caraterísticas/medidas de posição e de dispersão de conjuntos de dados

Começo lembrando notações e introduzindo as novas:
N denota a quantidade de individuos na população que foi observada, ela chama-se **tamanho da população**;
$x_1, x_2, \ldots, x_N$ denotam as observações; é natural que cada x_i chame-se **observação**.

Por exemplo, no exemplo da população de Akhiok, $N = 77$, e $x_1 = 28, x_2 = 6, \ldots, x_{76} = 4, x_{77} = 28$. A atribuição de índices de x's corresponde à ordem com a qual as observações vieram para mim; recorde: o quem fez as observações, apresentou-as de acordo com a ordem

alfabética das pessoas. Se a apresentação fosse diferente, a indexação seria diferente também. Mas a mudança de índices naõ afeta os valores da média e da variânica.

Observe a escolha e segue-a: N é maiúsculo, pois n minúsculo está reservado para o tamanho de amostra. Cada observação é uma letra minúscula do alfabeto latino, pois letras maiúsculas foram usadas para denotar variáveis aleatórias.

O valor de

$$\frac{x_1 + \ldots + x_N}{N}$$

chama-se **média populacional** e denota-se por $\bar{x}$, enquanto que o valor de

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N}$$

chama-se **variância populacional** e denota-se por σ_x^2.

No momento estamos discutindo as situações nas quais há só populações; nestas, é permitido usar os termos **média** e **variância**. No futuro, estaremos discutindo situações nas quais há também amostras. Uma amostra sempre está associada a uma populações, mesmo quando nosso conhecimento sobre essa for limitado ou nulo. Em tais situações, é imprescindível carregar a palavra "populacional" ou "amostral", pois a mesma permite identificar se trata-se da média e variância advindas de população ou de amostra.

As duas fórmulas em notações mais curtas aparecem assim:

$$\frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{e} \quad \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

e a da variância ainda tem duas suas irmás-gémeas

$$\frac{\sum_{i=1}^{N}(x_i)^2}{N} - (\bar{x})^2$$

$$\frac{\left(\sum_{i=1}^{N}(x_i)^2\right) - N\,(\bar{x})^2}{N}$$

Quanto às notações $\bar{x}$ e σ_x^2, elas têm sentido só se x foi usado como a notação genérica para as observações. Isso torna-se quase que obrigatório caso consideramos duas populações distintas. Nesse caso, é cómodo definir que uma é "x" e a outra é "y", e com isso, fica claro que $\bar{x}$ e σ_x^2 referem-se à primeira, enquanto que $\bar{y}$ e σ_y^2 à segunda. Na análise de uma população só, as notações $\bar{x}$ e σ_x^2 não se justificam por completo, mas já que não são totalmente erradas, então estão comumente usadas.

No caso do Exemplo sobre os moradores do vilarejo Akhiok,

$$media = \frac{0 + 1 + \cdots + 54 + 55 + 56 + 63}{77} = 22,67532 \approx 22,7$$

Note que o mesmo valor dá-se por

$$\frac{1\times\ 0 + 3\times\ 1 + 3\times\ 2 + 2\times\ 3 + \ldots + 1\times\ 63}{77}$$

que é a soma dos produtos de (idade)×(sua frequencia absoluta), dividida pelo tamanho da população, e ainda por

$$\frac{1}{77}\times\ 0 + \frac{3}{77}\times\ 1 + \frac{3}{77}\times\ 2 + \ldots + \frac{1}{77}\times\ 63$$

que é a soma dos produtos de (idade)×(frequencia relativa).

Isso dá-lhe mais duas maneiras de cálculo da média populacional (e também da média amostral, cuja definição é semelhante à da média populacional, conforme veremos adiante).

Para que serve a média e variância populacionais? Existem diversas aplicações. É frequente que médias e variâncias se usam para comparar duas ou mais que duas populações.

Na disciplina "Noções de Estatística" a principal aplicação de média e de variância está na aproximação por distribuições normais. Acontece que algumas (embora não todas) distribuições populacionais podem ser muito bem aproximadas por distribuição Normal. Tal aproximação é o assunto de nossas aulas no futuro próximo, e no momento, só digo que essa é possível e é muito útil. Então, quando a aproximação existe, a escolha da distribuição da familia das distribuições Normais, que aproxime-se melhor de todas à distribuição populacional dá-se com o auxílio de exclusivamente dos valores da média e da variância correspondentes à população aproximada.

Medidas (caraterísticas) de posição

Média, mediana, quantis (e em particular, decis, quartis, etc.), o máximo e o mínimo chama-se **medidas de posição** do conjunto de dados para o qual foram calculados.

Medidas (caraterísticas) de dispersão

Já foi dito que a variância de uma variáveil aleatória pode ser interpretada como a medida de dispersão da distribuição dessa variável. Fiz um desenho na lousa para explicar tal interpretação.

O mesmo procedimento pode ser aplicado à distribuição de frequencias por atributo de uma população, e assim conclui-se que a variância (σ_x^2) é uma medida de dispersão.

Essa não é a única possível (e nunca ninghém falou que seja a melhor de todas as outras). Eis algumas outras (as que você deve conhecer):

- sua **amplitude** defina-se como $\max - \min$ (o que é igual a $x_{(N)} - x_{(1)}$); a aplitude e denota-se tipicamente por A;
- o **intervalo interquartil** defina-se por $Q3 - Q1$;
- a **variânica** denota-se por σ_x^2 e defina-se pelo

$$\sigma_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N},$$

- o **desvio padrão** denota-se por σ_x e defina-se por

$$\sigma_x = \sqrt{\sigma_x^2}$$

- o **coeficiente de variação** denota-se por CV_x e defina-se por

$$CV_x = \frac{\sigma_x}{\bar{x}} \times 100\%$$

Explicação sobre o coeficiente de variação

O CV_x merece explicação:

Imagine a população de duas pessoas. Suponha que eu meço suas alturas em centímetros:

170 e 190

Então, a média é $\bar{x} = 180$ e a variância é

$$\sigma_x^2 = \frac{(170-180)^2 + (190-180)^2}{2} = 100$$

Suponha que uma outra pessoa mede as alturas em metros; eis as medições:

$$1,70 \text{ e } 1,90$$

Então a média $\bar{y} = 1,80$ e a variância é

$$\sigma_y^2 = \frac{(1,70-1,80)^2 + (1,90-1,80)^2}{2} = 0,01$$

Outras caraterísticas de distribuição populacional. Existem diversas caraterísticas de distribuições populacionais.

Vamos considerar aqui só aquelas que são numéricas; elas chama-se alternativamente **medidas** (de distribuição).

Vamos introduzir algumas das medidas de duas classes específicas: a chamada de classe de **medidas de posição**, e a chamada de classe de **medidas de dispersão**.

Para seu conhecimento (sem a cobrança nas provas do curso), existem outras medidas, como, por exemplo, a que mede a assimetria da distribuição.

A notação para observações ordenadas. Para falar de quantis e de outros conceitos derivados desses, é precisa introduzir uma notação.

Recorde que $x_1, x_2, \ldots, x_N$ era a notação para as observações de um atributo qualquer numa população qualquer. Suponha que tais observações foram ordenadas da menor para a maior. Então, a primeira dela, a menor, quer dizer, adquira a notação $x_{(1)}$. A segunda menor está denotada por $x_{(2)}$. E assim por diante, até $x_{(N)}$, a qual é, obviamente, a maior observação de todas.

Por exemplo, as idades dos 77 moradores de Akhoik

$$28,\ 6,\ 17,\ 48,\ 63,\ 47,\ 27,\ 21,\ 3,\ 7,\ 12,$$
$$\ldots$$
$$4,\ 10,\ 26,\ 12,\ 6,\ 16,\ 8,\ 2,\ 4,\ 28$$

quando ordenadas, deram

$$0,\ 1,\ 1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 3,\ 4,\ 4,$$
$$\ldots$$
$$48,\ 49,\ 50,\ 50,\ 51,\ 52,\ 53,\ 54,\ 55,\ 56,\ 63$$

Nas notações introduzidas, $x_{(1)} = 0, x_{(2)} = 1, x_{(3)} = 1, x_{(4)} = 1, x_{(5)} = 2, \ldots, x_{(77)} = 63$,

6.2.4 Quantis. Discussão preliminar

Considere, como uma motivação para a introdução do conceito "quantil", o seguinte

EXEMPLO: os salários dos 120 empregados contratados numa certa empresa, já ordenados, em mil R$, estão na tabela da transparência seguinte.

A pergunta é achar o teto salarial dos 10% dos empregados menos pagos.

É óbvio que a resposta é o valor do observação tal que a quantidade das observações cujos valores são menores que ele ou igual a ele seja 10%. Como no caso, há 120 observações, então 10% de 120 é 12, e a resposta, portanto, é o valor da 12-a observação do conjunto ordenado de dados, quer dizer, $x_{(12)}$; contando até a 12-a observação no conjunto abaixo, acha-s a resposta numérica: 5.2 (mil R$).

3.1 3.1 3.5 4.3 4.4 4.5 4.7 4.9 4.9 5.0 5.1 5.2
5.3 5.3 5.4 5.6 5.7 5.8 5.8 5.8 5.9 5.9 6.0 6.1
6.1 6.2 6.2 6.2 6.2 6.4 6.5 6.5 6.6 6.6 6.6 6.6
6.6 6.7 6.7 6.8 6.8 6.8 6.8 6.8 6.9 6.9 6.9 7.0
7.0 7.1 7.1 7.1 7.1 7.2 7.2 7.3 7.3 7.3 7.3 7.3
7.3 7.3 7.3 7.4 7.5 7.6 7.7 7.8 7.8 7.8 7.9 7.9
7.9 7.9 8.0 8.0 8.0 8.0 8.0 8.0 8.1 8.1 8.2 8.2
8.2 8.2 8.3 8.3 8.3 8.3 8.4 8.5 8.5 8.5 8.5 8.5
8.5 8.6 8.6 8.6 8.6 8.7 8.7 8.8 8.9 8.9 8.9 8.9
9.0 9.0 9.1 9.1 9.1 9.2 9.3 9.3 9.4 9.6 9.8 9.8

O limiar procurado (e achado) no exemplo tem nome: quantil de ordem $0,1$.

O "valor da ordem" (quer dizer "$0,1$") corresponde à proporção das observações à esquerda da "corte" feita no conjuto ordenado pelo quantil, ou, falando com maior precisão, o "valor da ordem" corresponde à proporção daquelas observações que são menores que ou iguais ao quantil.

Infelizmente, não é que para qualquer conjunto de observações e para qualquer p, podemos cortar o conjunto em proporções p e $1-p$, sendo que na primeira dessas incluam-se as observações cujos valores coincidem com o da corte.

Por exemplo, pela lógica da "corte" não existe o quantil da ordem 0,27 para o conjunto das observações de salário (pois $0,27 \times 120 = 32,4$ - valor não inteiro).

O problema com a não existência de quantis de certas ordens para certos conjuntos não é algo grave pois a construção de quantis é comumente guiada pelo bom senso (com vista em aplicações específicas) que permite ignorar as situações problemáticas justificando isso pela futilidade na perspectiva de aplicabilidade.

Entretanto, é bom que exista uma regra da construção de quantis que seja aplicável a qualquer p. Tal regra está apresentada abaixo para um caso especial que é muito usado (e por isso, que exige uma regra). O uso é na construção de *Q-Q plot* que é uma ferramenta estatística útil mas excluida do escopo do presente texto.

Seja q um número inteiro. Para cada $k = 1, 2, \ldots, q$, definiremos o **k-ésimo q-quantil** de um conjunto de observações duma população da seguinte maneira:

- calcula-se o valor de $N \times \frac{k}{q}$ e se esse for inteiro, então o valor do k-ésimo q-quantil declara ser o valor de $x_{\left(N \times \frac{k}{q}\right)}$, quer dizer, o valor da observação que está na posição $N \times \frac{k}{q}$ das observações ordenadas (da menor para a maior);

- já se $N \times \frac{k}{q}$ não for inteiro, toma-se o inteiro M imediatamente superior a $N \times \frac{k}{q}$, e o valor do k-ésimo q-quantil declara ser o valor de $x_{(M)}$.

Na definição acima, exclui-se a possibilidade de $k = 0$ pois esse valor, sendo colocado na fórmula, daria

$$N \times \frac{0}{q} = 0$$

e como não há $x_{(0)}$, então a definição não generaliza-se para $k = 0$.

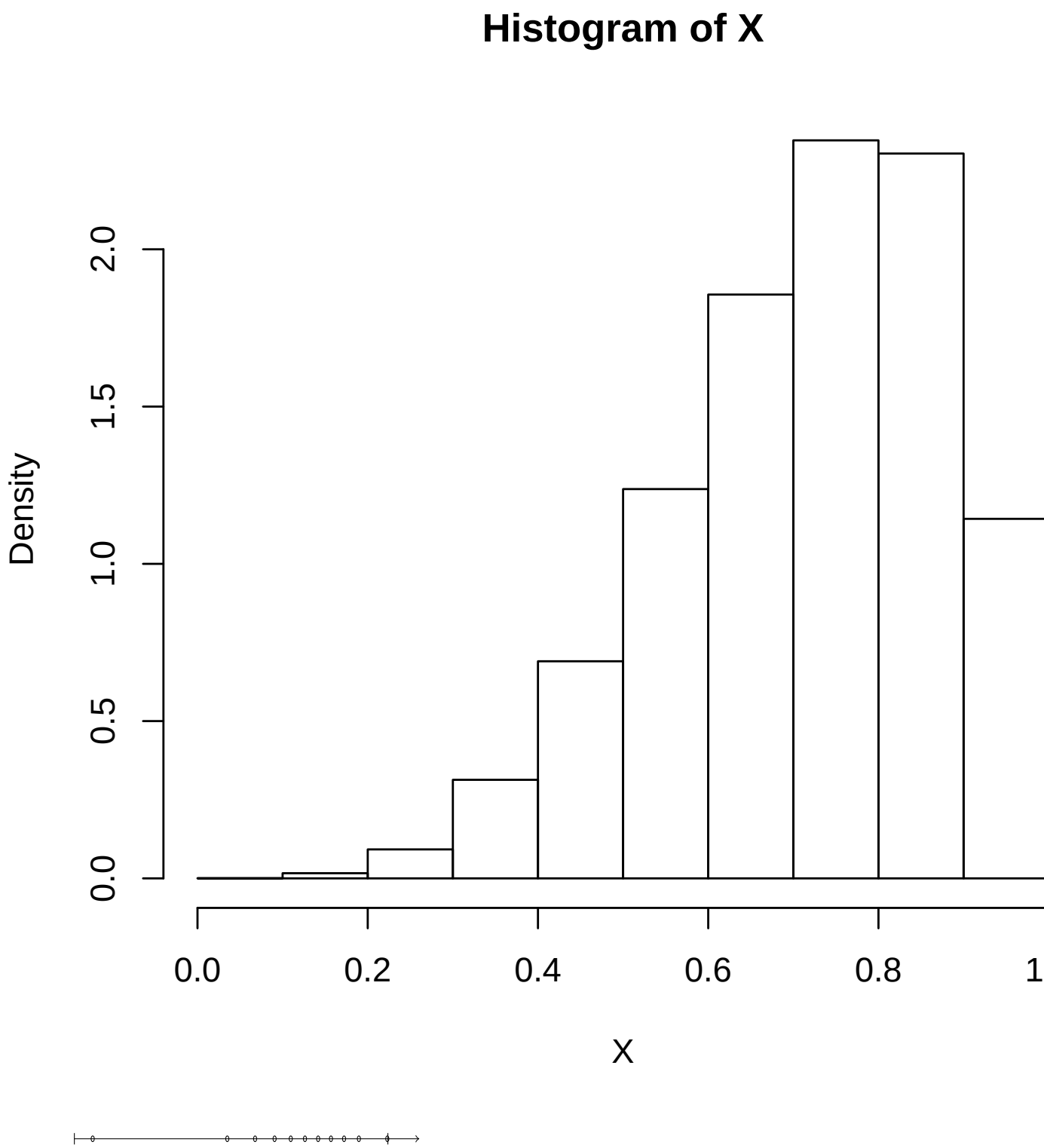
É cómodo definir que 0-ésimo q-quantil seja $x_{(1)}$, a observação mínima do conjunto.

Oba! Sem querer, acabei de introduzir o conceito **mínimo** de conjunto. Introduzo então também o **máximo**, e os correspondentes símbolos min e max; obsserve o óbvio: $\min = x_{(1)}$ e $\max = x_{(N)}$.

UM EXEMPLO de utilização de quantis.

Criei um conjunto de dados de tamanho $N = 100.000$. Vamos considerá-lo como observações de um certo atributo numa certa população.

Na transparência seguinte, apresentei o histograma da distribuição de frequencia relativa pelo atributo (usando classes de amplitudes iguais), e também os decis.



Observe que o histograma permite (entre outras coisas) comparar as frequencias por classes, e portanto, permite obter uma caraterização qualitativa da forma da distribuição; algo do tipo: a frequencia cresce mais rápido que linearmente (você vé tal fato se tentar passar uma regua pelos telhados dos prédios do histograma) até os valores do atributo na faxa de 0,7 – 0,8, depois estabelce, e depois decresce.

Uma conlcusão semelhante pode ser derivada a partir da observação das distâncias entre os

decis, pois entre um decil e o próximo, há 10% de todas as observações do conjunto.

A revelação a partir de histograma é mais fácil, mas a vantagem do desenho de decis é que ele é unidimensional.

É muito comum o uso do desenho do tipo que foi mostrado na transparência anterior, só que não para decil, mas sim para **quartis**. As notações e nomes usados em tais desenhos são e suas interpretações são:

Entree todos os quantis de ordem p, os que a gente mais usa são:

min	para o mínimo;
$Q1$	para o 1-o 4-quantil, chamado de **primeiro quartil**;
$Q2$	para o 2-o 4-quantil, chamado de **segundo quartil** e também de **mediana**;
$Q3$	para o 3-o 4-quantil, chamado de **terceiro quartil**;
max	para o máximo.

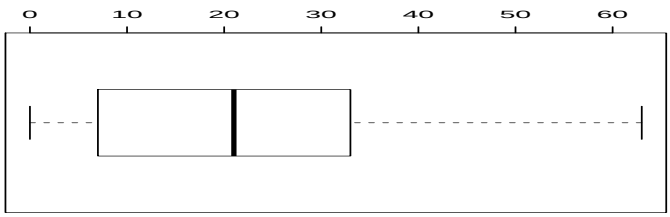
Os nomes são autoexplicativos: os $Q1$, $Q2$ e $Q3$ dividem um conjunto ordenado de dados em quatro partes (quase) iguais, sendo que a igualdade entende-se aqui no sentido da quantidade de dados; em cada parte há (quase) 25% de todos os dados.

Os "quases" acontecem por causa das observações cujos valores coincidem com os valores de quartis. Para conjuntos de dados grandes com poucas repetições, isso não atrapalha, fato que faz a palavra "quase" estar esquecida.

6.2.5 Box-plot

Um conjunto de dados pode ser respresentado em diversas maneiras. Uma delas, chama-se *Box Plot*. BoxPlot não transmite toda a informação sobre o conjunto para qual foi feito. Só apresenta *max*, *min*, e $Q1$, $Q2$, $Q3$.

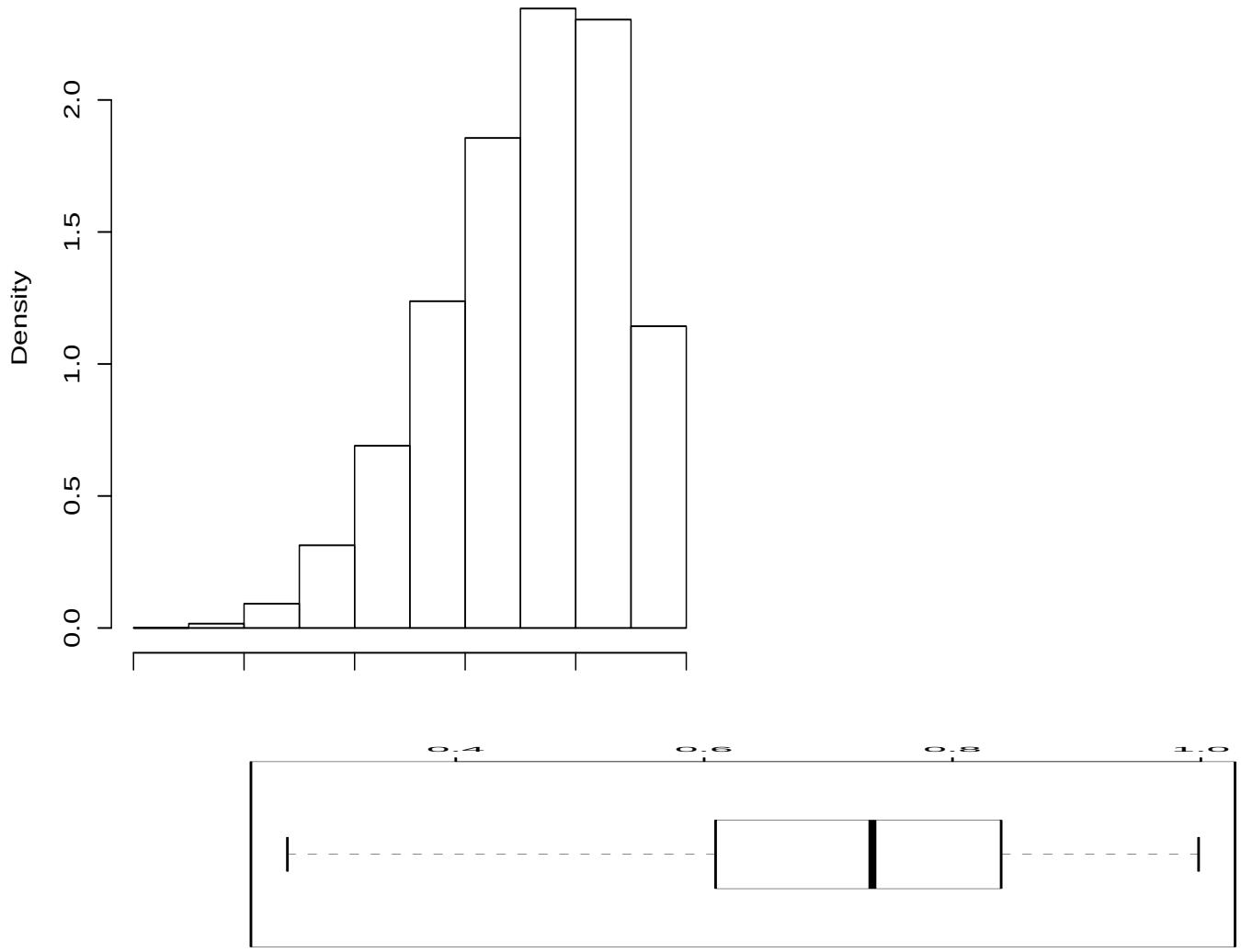
Abaixo, você vé o Box-Plot para as observações da idade dos moradores do vilarejo frequencia Akhiok:



O acordo sobre o formato do desenho é: $Q1$ e $Q2$ formam um retângulo (cuja altura não importa), dentro do qual, fica o separador correspondente ao valor de $Q3$. Para fora do retângulo, "crescem bigodes" até, respectivamente min e max. Por isto que a tradução de BoxPlot é "caixinha com bigodes".

Com a divisão do conjunto de observações em quatro partes, podemos chamar por **caudas de distribuição** a primeira e a última, e podemos chamar por **valores centrais de distribuição** às observações que estão na segunda e na terceira parte da divisão.

Com essa linguagem, podemos usar o Box-Plot para falar da distribuição como um todo. Veja o examplo nas duas transparências a seguir. Na primeira delas, há o histograma de uma distribuição populacional e, abaixo dessa, o Box-Plot feito para o mesmo conjunto de dados. Na segunda transparência, eu "falo" sobre a forma do hisograma a partir da consideração da forma do Box-Plot.



Olhando ao Box-Plot, podemos sugerir que:

(a) a cauda esquerda da distribuição é mais comprida que a cauda direita;

(b) os 50% dos valores centrais estão ligeiramente deslocados à direita;

(c) os 50% dos valores centrais,sendo divididos pela mediana no meio, apresentam a seguinte propriedade: os que estão à direita da mediana são mais "agrupados" ou "densos", em outras palavras, do que os que estão à esquerda da mediana.

Do ponto de vista da variância, as medições "y" têm dispersão menor. Mas é claro que isso é a consequencia da mudança de escala. O coeficiente de variçaõ "corrige" a distorção:

$$CV_x = \frac{\sqrt{100}}{180} \times 100\% = \frac{\sqrt{0,01}}{1,80} \times 100\% = CV_y$$

6.1 6.2 6.2 6.2 6.2 6.4 6.5 6.5 6.6 6.6 6.6 6.6
6.6 6.7 6.7 6.8 6.8 6.8 6.8 6.8 6.9 6.9 6.9 7.0
7.0 7.1 7.1 7.1 7.1 7.2 7.2 7.3 7.3 7.3 7.3 7.3
7.3 7.3 7.3 7.4 7.5 7.6 7.7 7.8 7.8 7.8 7.9 7.9
7.9 7.9 8.0 8.0 8.0 8.0 8.0 8.0 8.1 8.1 8.2 8.2
8.2 8.2 8.3 8.3 8.3 8.3 8.4 8.5 8.5 8.5 8.5 8.5
8.5 8.6 8.6 8.6 8.6 8.7 8.7 8.8 8.9 8.9 8.9 8.9
9.0 9.0 9.1 9.1 9.1 9.2 9.3 9.3 9.4 9.6 9.8 9.8

6.3 Amostragem, amostra e sua relação com população

6.3.1 Um par de avisos de caráter genérico

Amostras serão vistas nessa seção por um ângulo que não é muito tradicional em cursos de Estatística Básica, mas que, em compensação, é muito útil na discussão das propriedades de amostras que estão na base da maioria de métodos estatísticos. Uma das vantagens proporcionada pelo ponto de vista aqui adotado é a facilidade na demonstração e explicação dos seguintes fatos

- (i) que a média amostral pode ser tomada como aproximação para média populacional;
- (ii) que a variância amostral pode ser tomada como aproximação para variância populacional;
- (iii) que um histograma construído para amostra pode ser visto como aproximação daquele histograma que seria construída para a população via a separação de suas observações por as mesmas classes que são as do histograma amostral;
- (iv)que qualquer quantil da distribuição amostral pode quer tomado como aproximação para o correspondente quantil da distribuição populacional (e, em particular, que o Box-Plot construído para amostra pode ser tomado como aproximação para Box-Plot construído para população).

Aplicar os fatos (i)-(iv) para inferir sobre propriedades de populacional a partir da análise de amostra é o que será cobrado de vocês nos exercícios referentes ao conteúdo da presente seção. Verdade é, entretanto, que a solução dos referidos exercícios exigem seu conhecimento dos fatos (i)-(iv) mas não exige das razões que os amparam. Por isso, sua primeira leitura da seção pode ser superficial.

6.3.2 Amostragem, amostra, frequencia amostral

Vamos considerar a população dos moradores do vilarejo Akhiok. Eu lhe digo que pretendo retirar ao acaso uma pessoa dessa população, e lhe pergunto: "Qual é a probabilidade da pessoa retirada ter idade 3 anos?"

Antes de discutir a resposta, temos que alinhar nossos entendimentos acerca o significado do termo **retirada de um indivíduo ao acaso de uma população**. Ao usar as palavras "ao acaso" nesse termo eu quero lhe dizer que cada pessoa tem a mesma probabilidade de ser escolhida. Já que tais palavras apareceram nos capítulos anteriores e nestes carregavam exatamente o sentido de "com a mesma probabilidade" então você capta facilmente também o sentido do termo todo. O problema é como construir o procedimento de retirada que garanta essa igualdade de probabilidades. No caso da população de moradores de Akhiok imaginamos o vilarejo deles e percebemos que "a mesma probabilidade" não será alcançada se escolhermos a primeira casa na entrada do vilarejo, ou se passarmos pela praça cetral ao meio dia e apanhamos a primeira pessoa que se aproximar a nos. Bom, sem perder tempo contando outras opções inadequadas, vou apresentar uma que funciona e que será canonizada a partir de agora. Nesse procedimento, eu tomo 77 bolas idênticas (recorde, 77 é o tamanho da população) e escrevo nomes dos moradores nas bolas. Depois, coloco as bolas numa urna e retiro uma delas ao acaso. O nome de morador escrito nela é "a pessoa retirada ao acaso da população". Como preciso observa a idade dela, vou a sua casa e pergunto. Na realidade, eu poderia escrever em cada bola o nome e a idade de cada pessoa. Desse modo não precisaria "ir a casa para perguntar a idade". Pensando mais um pouco, descobre-se que o nome de morar escrito em "sua" bola é irrelevante e pode ser apagado.

O procedimento descrito agora atende à desejada qualidade de "com mesma probabilidade" pois essa qualidade está no experimento aleatório "retirada ao acaso de uma bola de urna"

(recorde que o fato citado por último foi deduzido da nossa compreenção intuitiva do conceito de probabilidade; isso foi feito no Capítulo 1, naquela sua parte que introduziu o experimento aleatório "retirada ao acaso de uma bola de urna"). Então, como o procedimento descrito nos convém, não vamos procurar por outros. Com isso eu quero dizer que toda vez que eu dizer "retiro de um indivíduo ao acaso de uma população" você entende que faço o procedimento descrito acima, quer dizer, associo cada individuo com uma das bolas numa urna cuja quantidade é igual ao tamanho da população e retiro uma bola ao acaso.

Agora que associamos a retirada de um indivíduo ao acaso de uma população à retirada ao acaso de uma bola de urna, podemos, com ajuda de tudo que aprendemos acerca de bolas em urnas responder à pergunta "Qual é a probabilidade da pessoa retirada ter idade 3 anos?" Naturalmente, é a frequencia relativa das pessoas com 3 anos de idade; essa frequencia é 2/77. (Se você deseja detalhes da derivação da resposta, conto-lhe que isso é a consequencia direta do fato de que na urna há três bolas correspondentes às pessoas com idade 3 na toda a população.)

É óbvio que se no lugar de "3" fosse qualquer outra idade, a resposta virá pelo mesmo raciocínio que aplicamos agora para o caso 3. É óbvio tambem, que o mesmo raciocínio aplica-se a qualquer população e qualquer atributo. Esse fato é importante para os argumentos futuros. Vamos destacá-lo como uma proposição.

Proposição 20 *(sobre probabilidade de retirada de um indivísuo ao acaso de uma população).*
Suponha que $(x_1, f_1), (x_2, f_2), \ldots, (x_M, f_M)$ é a distribuição de frequencia relativa sobre um atributo de uma população.

▷ Por exemplo, no caso da população dos moradores de Akhiok, a distribuição da frequencia relativa sobre o atributo "idade" foi apresentada pela Tabela DARNOME. Na primeira coluna dela, há 0 para "idade" e 1/77 para a respectiva frequencia relativa; então, 0 é o que foi chamado por x_1 acima, e 1/77 é o que foi chamado por f_1. A Tabela DARNOME possui 48 colunas. Cada coluna é um par (x, f) na notacão acima, e M usado na notação é 48 no caso (evitei usar N pois esse associa-se usualmente ao tamanho de população).

Suponha que um individuo será retirado da população ao acaso e seu atribuito será observado. Então, a observação pode ser qualquer um dos valores $x_1, x_2, \ldots, x_M$, e a probabilidade que será valor x_i é f_i, $i = 1, 2, \ldots, M$.

A próxima proposição é uma reformulação da Proposição 2 com uso do termo "variável aleatória". A motivação para a reformulação é que, conforme avisado no Capítulo 3 que introduziu variáveis aleatórias, essas nos provem da linguagem que é muito cómoda.

Proposição 21 *(sobre a distribuição da variável aleatória que representa o valor do atributo medido para o indivíduo ao acaso de uma população).*
Suponha que $(x_1, f_1), (x_2, f_2), \ldots, (x_M, f_M)$ é a distribuição de frequencia relativa sobre um atributo de uma população. Assume que x's são números (isso é necessário pois usaremos x's como valores da variável aleatória a ser construída abaixo, e, conforme acordado, variáveis aleatórias só assumem valores numéricos). Represente a distribuição em forma de tabela para facilitar a compreenção da afirmação da presente proposição:

o valor do atributo	x_1	x_2	$\cdots$	x_i	$\cdots$	x_M
a respectiva frequencia relativa	f_1	f_2	$\cdots$	f_i	$\cdots$	f_M

Considere o seguinte experimento aleatório: um individuo será retirado da população

ao acaso e seu atributo será observado. Denote por X a variável aleatória que em cada resultado desse experimento aleatório assume o valor observado.

Então, a variável aleatória X pode assumir qualquer um dos valores $x_1, x_2, \ldots, x_M$, e a probabilidade, p_i, de assumir o valor x_i é igual a f_i, $i = 1, 2, \ldots, M$. Em outras palavras, a tabela da distribuição de X é como se segue:

valor	x_1	x_2	$\cdots$	x_i	$\cdots$	x_M
a probabilidade de X assumir o respectivo valor	f_1	f_2	$\cdots$	f_i	$\cdots$	f_M

Agora fixe um número natural qualquer e chame-o por n (com o intuito de dar concretude á exposição, pode pensar que $n = 10$). Repita n vezes o experimento aleatório "retirada de um indivíduo ao acaso da população o observação do valor de seu atributo". (Para a concretude, pode pensar que a população é os moradores de Akhiok e que o atributo observado é a idade.) Observe que para que o experimento aleatório seja repetido nas mesmas consições precisa que após cada rodada, o indivíduo escolhido seja retornado à população, ou, em termos de bolas-em-urna, seja devolvida à urna bola selecionada. Tal procedimento chama-se **amostragem com reposição**. Observe que somos nós quem determinou que os experimentos devem acontecer nas mesmas condições. Essa exigência tem suas razões de ser, que revelar-se-ão no futuro. Outras razões, quando se aplicam, obrigam nos a querer que não hajam reposições. Tal esquema chama-se **amostragem sem reposição**.

▷ Eis uma das situações nas quais, pela própria natureza, a amostragem é sem reposição. Imagine a população de todos os pacientes com colesterol alta tratados atualmente na clínica Aiaidoeó. A gente escolha um deles ao acaso e aplica um tratamento que deve baixar o colesterol. Depois, pega mais um e aplica o tratamento nele. Isso repete-se n vezes. Naturalmente, o paciente já tratado não pode ser devolvido à população.

Então, suponha que a amostragem com reposição foi executada. Vamos usar a seguinte notação genérica para o resultado:

$$a_1, a_2, \ldots, a_n \tag{6.5}$$

onde – obviamente – a_i significa o valor de atributo do indivíduo escolhido na i-ésima retirada. O nome completo para a sequência em (1.5) a ser usado no meu texto é **amostra resultande de amostragem com reposição**. Mas como na maior parte do texto surgem só essas amostras, então vou me referir a elas por simples **amostra**. Enquanto que se falar de caso diferente, vou ser cuidadoso na explicação euhastiva e didática da amostragem que gera o caso considerado.

▷ Aquilo chamado acima por "amostra" tem também o nome oficial na Teoria Estatística: **amostra aleatória simples com reposição**. Eu não vou usar esse nome. A razão para ele cair em disuso no meu livro é a disnecessidade da necessidade de explicação de sua composição e formação. Mas se você deseja ouvir a explicação, eis essa. Na Estatística, a **amostra aleatória simples** está definida como resultado de qualquer método de amostragem probabilística que dá a cada elemento da população alvo e a cada possível amostra de um tamanho determinado, a mesma probabilidade de ser selecionado. Então, se você procurar por todos os métodos que provêm tal igualdade de probabilidade e exigir que haja no método também a reposição, você vai descobrir que existe um e único método e que esse é exatamente a amostragem com reposição conforme definida acima por mim. Espero que você concorda comigo que tudo isso é pouco exagerado para o nível de apresentação do meu livro.

Para a população de Akhiok, eu fiz $n = 10$ retiradas e obtive a seguinte sequencia:

$$31, 60, 3, 3, 12, 6, 47, 55, 57, 34 \tag{6.6}$$

Declaro publicamente que esses números foram por mim obtidos obedecendo honestamente e veemente o esquema de ações que geram amostras com reposição; isto é: Coloquei na urna 77 bolas idênticas, marquei nelas as idades dos moradores, misturei bem e retirei uma; ela deu o valor 31 que você vé no início da sequencia apresentada. Devolvi essa bola, misturei todas as bolas de novo e retirei a segunda; essa mostrou 60. Repeti esse procedimento 10 vezes. Se você não acredita que não minto, pode vir visitar a universidade onde trabalho: a urna foi colocada no museu, onde ainda pode ser vista, se ninguém a tirou de lá.

▷ É importante que os valores de amostra estejam apresentados na ordem das suas respectivas retiradas? Um muitos casos a resposta é "não". Posso até garantir que todos procedimentos estatísticos estudados nesse texto são os casos de "não". Isso implica no que os valores obtidos podem ser chamados por **conjunto** de valores, pois o sentido matemático do termo "conjunto" é justamente "a ordem não importa", diferentemente do termo **sequencia** que enfatisa a importânica da ordem. Toda essa informação é importante para você pois o porquê os termos **conjunto de dados** e **sequencia de dados** aparecem como sinônimos. E já que estamos falando de sinônimos, vale observar que a palavra "dados" pode ser substituida por **valores** assim como por **observações**.

▷ Gostaria de apresentar uma observação que não é importante para minha exxposição mas pode despertar curiosidade. A observação é que tudo que eu expliquei em detalhes pode ser expresso na seguinte definição que eu achei numa apostila de ensino de Estatística: "Amostragem aleatória simples é um método de amostragem probabilística que dá a cada elemento da população alvo e a cada possível amostra de um tamanho determinado, a mesma probabilidade de ser selecionado". Se gostou do laconismo dessa definição, fique com ela, mas anote que o objeto por ela definido coincide com o definido por mim acima.

6.3.3 Aproximação de frequencias populacionais por frequencias amostrais e aplicações

Voltaremos agora à situação envolvendo moradores de Akhiok, e imaginamos que foi feita uma amostra $a_1, \ldots, a_n$. Pergunto: Com que frequencia relativa o valor 3 está presente na amostra? Antes de prosseguir para a derivação de resposta, gostaria de apresentar a expressão matemática para o termo "a frequencia relativa do valor 3 na amostra". A expressão bate com aquilo que sua intuição imaginou quando ouviu esse termo. Por outro lado, se sua intuição estava ou está na dúvida, você deve aceitar a expressão abaixo como a definição formal do termo.

$$\text{a expressão}\rightarrow\quad \frac{\text{quantidades de } a\text{'s da amostra iguais a 3}}{n} = f^{\text{amostra}}_{\text{de tamanho } n}(\text{valor 3}) \quad \leftarrow\text{a notação} \tag{6.7}$$

▷ A frequencia relativa, sobre a qual versa a pergunta acima, é a **frequencia relativa amostral**. Você facilmente adivinha o modo de seu cálculo fazendo o paralelo com o cálculo da frequencia relativa introduzida para populações (alías, essa será chamada a partir de agora **frequencia relativa populacional**). Por certo, os adjetivos "amostral" e "populacional" devem ser obrigatoriamente presentes pois só assim poderei garantir que você entenda plenamente sobre qual das duas estou falando. Entretanto, é frequente que se a frequencia relativa referida no texto é amostral ou populacional será óbvio do contexto. Quando isso for o caso, eu realmente vou omitir os adjetivos.

Para responder na pergunta acima colocada, argumentamos assim: Primeiramente, temos que $f^p(\text{valor 3}) = 2/77$ (temporareamente, f^p aponta que estamos falando da frequencia populacional). Isso junto com Proposição 2 dizem que se retirarmos ao acaso um indivíduo

da população e observarmos sua idade, então a probabilidade de observar valor 3 (a ser denotada por $\mathbb{P}[\text{valor 3}]$) será igual a $f^p(\text{valor 3}) = 2/77$; concluindo: $\mathbb{P}[\text{valor 3}] = 2/77$. Por último recordamos que Definição 1 formalizou nossa intuição acerca do conceito "probabilidade" da seguinte maneira: $\mathbb{P}[\text{valor 3}]$ é o valor assintótico da frequencia relativa do aparecimento deste resultado numa sequencia de repetições do experimento. Mas os valores $a_1, a_2, \ldots, a_n$ da amostra são exatamente os resultados de n repetições do experimento "retirada ao acaso dum indivíduo da população e observação de sua idade". Portanto,

$$\mathbb{P}[\text{valor 3}] = \lim_{n\to\infty} \frac{\text{quantidades de } a\text{'s da amostra iguais a 3}}{n} \tag{6.8}$$

Lembrando agora a identidade entre $\mathbb{P}[\text{valor 3}]$ e a frequencia relativa populacional, reescrevemos (1.8) assim (nesta reescrita, substituimos a fração por $f_n^a(\text{valor 3})$):

$$f^p(\text{valor 3}) = \lim_{n\to\infty} f_n^a(\text{valor 3}) \tag{6.9}$$

É óbvio que o raciocínio que culminou na afirmação (1.9) pode ser proferido para qualquer população e para qualquer valor dos possíveis valores do atributo observado na população. O resultado genérico está formulado abaixo em forma de uma proposição já que sua afirmação é importante para tudo que segue-se e pode ser provado rigorosamente (eu diria que o argumento que deduziu (1.9) para população específica e valor específico pode ser facilmente transformada na demonstração rigorosa).

Proposição 22 *(sobre frequencias populacionais como limite de frequencias amostrais).*
Seja $(x_1, f_1), (x_2, f_2), \ldots, (x_M, f_M)$ a distribuição de frequencia relativa sobre um atributo de uma população. Considere uma sequencia infinita de amostragens da população com reposição. Denote por

$$a_1, a_2, \ldots$$

os resultados, quer dizer, a sequencia na qual a_i é o valor do atributo do indivíduo retirado na i-ésima amostragem.

Para cada $x_i \in \{x_1, x_2, \ldots, x_M\}$ e cada $n = 1, 2, \ldots$, intoruza $f_n^a(x_i)$, frequencia relativa de aparência de x_i na sequencia $a_1, a_2, \ldots, a_n$, da seguinte maneira:

$$f_n^a(x_i) := \frac{\text{quantidades de } a\text{'s da amostra iguais a } x_i}{n}$$

Então,

$$f_i = \lim_{n\to\infty} f_n^a(x_i), \text{ para cada } x_i \in \{x_1, x_2, \ldots, x_M\} \tag{6.10}$$

Já que na prática nossas amostras são finitas, então, para fim prático, precisamos interpretar a relação assintôtica (1.10) da forma que aplique-se as amostras finitas. Uma das interpretações está sugerida abaixo. O rigor matemático dela não é perfeito, mas está de acordo com o nível do presente texto, e, na forma como está, ela pode ser usada para amparar as propriedades que estão na base dos métodos estatísticos a serem apresentados no texto a seguir. Eis esta:

$$\begin{array}{ll} (a) & \text{para } n \text{ suficientemente grande, } f_n^a(x_i) \text{ está próxima a } f_i \text{ para cada valor } x_i \\ & \text{entre os valores do atribuito observado na população} \\ (b) & \text{conforme } n \text{ cresce, para cada } x_i,\ f_n^a(x_i) \text{ está cada vez mais próxima a } f_i \end{array} \tag{6.11}$$

A proximidade das frequencias amostrais $f_n^a(x_1), f_n^a(x_2), \ldots, f_n^a(x_M)$ as correspondentes frequencias populacionais $f_1, f_2, \ldots, f_M$ permite concluir sobre a proximidade de caraterísticas das

duas distribuições. As conclusões podem ser formuladas de uma maneira precisa, parecida com a da Proposição 4 mas eu prefiro as formulações menos formais e mais práticas. A lista delas está abaixo, e já aviso que as fórmulas da lista são exemplificadas por exemplos numéricos na Seção 1.3.4.

▷ Nas fórmulas da lista usarei quase todos os símbolos que foram definidos no texto acima. Mas como tais definições são bem expalhadas, torna-se então útil uma lebrete que junta todas. Eis esta. O tamamnho da população foi denotado por N. O que interessa acerca dos indivíduos da população são os valores de um certo atributo (a idade, por exemplo, como no caso da população de Ahkiok. A quantidade de valores diferentes foi denotado por M já que essa quantia não tem a obrigação de coincidir com N. Os próprios valores adquiram a notação genérica $x_1, x_2, \ldots, x_M$. A frequencia relativa populacional do valor x_i denota-se por f_i; naturalmente, há M frequencias f_i. Nesse enredo todo, há uma amostra. O tamanho dela foi denotado por n, e os valores observados são $a_1, a_2, \ldots, a_n$. Obviamente, cada a_j é um valor do conjunto $\{x_1, x_2, \ldots, x_M\}$. Isso permite nos falar da frequencia relativa amostral de cada x_i. Tal frequencia está denotada por $f_n^a(x_i)$.

(I) A média populacional cuja definição formal é

$$\bar{x} := \tfrac{1}{M}\left(x_1 f_1 + x_2 f_2 + \cdots + x_M f_M\right) \quad \begin{array}{l}(\bar{x} \text{ aqui é o símbolo constumeiramente}\\ \text{usado para denotar a média populacional)}\end{array} \tag{6.12}$$

pode ser aproximada pela média amostral cuja definição replica a da média populacional conforme mostrado abaixo

$$\frac{1}{M}\left(x_1 f_n^a(x_1) + x_2 f_n^a(x_2) + \cdots + x_M f_n^a(x_M)\right) \tag{6.13}$$

mas cujo cálculo faz-se tradicionalmente pela seguinte fórmula (óbvio, que as fórmulas (1.13) e (1.14) são equivalentes entre si)

$$\bar{a}_n := \tfrac{1}{n}\left(a_1 + a_2 + \cdots + a_n\right) \quad \begin{array}{l}(\bar{a}_n \text{ aqui é o símbolo constumeiramente}\\ \text{usado para denotar a média de amostra de tamanho } n)\end{array} \tag{6.14}$$

(II) A variância populacional cuja definição formal é

$$\sigma_x^2 := \tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left(x_i - \bar{x}\right)^2 f_i \quad \begin{array}{l}(\bar{x} \text{ aqui é o símbolo constumeiramente}\\ \text{usado para denotar a média populacional)}\end{array} \tag{6.15}$$

ou, equivalentemente,

$$\sigma_x^2 := \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \tag{6.16}$$

pode ser aproximada pelo valor calculado a partir de amostra de acordo com fórmulas semelhantes; as fórmulas são duas já que há acima duas definições para σ_x^2, mas a segunda delas é de valor infinitamente maior que a primeira tanto pela amplitude de sua uso na prática quanto pela importância didático-teórica valorizada pelo presente texto; eis esta:

$$\frac{1}{n}\sum_{i=1}^{n}\left(a_i - \bar{a}_n\right)^2 \tag{6.17}$$

Observe que a fórmula (1.17) não foi chamada de "variância amostral". Esse termo pertence a outra expressão na qual, diferentemente da expressão (1.17) o denominador é $n-1$. Isso, naturalmente, causa-lhe uma supresa, e eu aprovieto de sua curiosidade para amarrar a resposta a um argumento importantíssimo da Estatística. Isso é o conteúdo da Seção 1.12.

(III) Qualquer q-quantil da distribuição de frequencia relativa amostral pode ser tomada como aproximador para q-quantil da distribuição da frequencia relativa populacional.
(IV) O histograma

6.3.4 Um exemplo numérico da aproximação de população pela amostra

6.4 Conceitos básicos de simulação e a aplicação no caso de uma variável aleatória discreta

Ver o material do probabilidade/Aulas/MinhaAulaSimulacaoAnaliseExploratoria/AulaSimulacao.tex

Para que a explicação do conceito "simulação" corra solta e fácil, precisamos voltar ao Capítulo ??. Espero que nele e refrescar o conceito "experimento aleatório" e alguns outros relacionados a este. Tomo como garantida a sua concepção do conceito "experimento aleatório" e do "proferir ou realizar um exerimento aleatório".

Tudo torna-se muito simples, se ao falar de um experimento aleatório, o locutor concebe claramente se ele esteja no eixo de tempo antes da proferência do experimento aleatório ou depois. Se estiver antes, então qualquer resultado do experimento aleatório é possível a acontecer ou, ao ser observado, em outros termos, e vale ainda lembrar que as possibilidades expressam-se formalmente em termos de probabilidade de acontecer.

Se estiver depois, então um dos resultados possíveis aconteceu e foi observado. O nome para este será **realização** ou **observação**.

Mas não deixe de ser enganado pela última frase e reclame: "As palavras realização e observação também são usadas nas falas de locutor – inclusive, do próprio autor deste livro –, quando este está na posição antes do acontecer de experimento aleatório." Concordo com a reclamação. Mas em defesa, chamo sua atenção que há uma ligeira diferença na terminologia que permite identificar o ponto de vista: na posição de antes, falamos "realização possível" e "observação possível", enquanto que na posição de depois, falamos "realização" e "observação". Infelizmente, a confusão tem uma chance, mas esta à confusão, mas esta chance Desta posição do locutor, não há mais chances para outros resultados, e, consequentemente, não se fala mais da probabilidade

Gostaria ainda mencionar, que o importante na verdade não é onde está o locutor em relação ao momento da proferência de experimento aleatório, mas sim, do qual ponto ele deseja analisá-lo. Por exemplo, nas aplicações de Estatística na vida real,

Suponha que foi formulado um experimento aleatório, e nele foi definida uma variável aleatória. Para o título de esclarecimento, recordo que definir uma variável aleatória num experimento aleatório significa, em palavras simples embora precisas, fornecer uma tabela que associa um número a cada resultado possível do experimento aleatório. Vale ainda notar e lembrar que os números podem se repetir, e que o conjunto de todos eles, sem as reperições, chama-se o conjunto de valores da variável aleatória. Estes lembrete serão valiosos no que se segue.

Suponha agora que você conduziu o experimento aleatório formulado, obteve sua realização e, ao consultar a tabela supramencionada, quer dizer, aquela que define a variável aleatória X, você identificou o seu valor correspondente. Este, então, chama-se *o valor simulado da variável aleatória* X, e todo o processo supradescrito que gerou-lo chama-se **simulação da variável aleatória** X, ou, genericamente, **simulação**. Seria útil notar também que o termo **realização de variável aleatória** usa-se frequentemente como sinônimo do valor simulado.

No parâgrafo acima, introduzi a terminologia principal relacionada ao assunto do presente capítulo. Como se vé, é nada complicado, e até, nada de novo. Se tudo isto te deixou tranquilo, pode passar para a seção seguinte. Ou, alternativamente, pode ler o exemplo abaixo que fecha a presente seção apresentando ilustrando o significado da terminologia introduzida.

Imagine que você segura em sua mão um dado equlibrado e pretende lançá-lo e anunciar (para mim, digamos) o valor da sua façe superior, quando o dado parar. Enquanto você estar com o dado na mão, nós nos referimos ao resultado a ser observado por "variável aleatória que pode assumir valores 1, ou 2, ou 3, ou 4, ou 5, ou 6 com as respectivas probabilidades 1/6, 1/6, 1/6, 1/6, 1/6 e 1/6´´; tal referência e termos nela usados são parte de nossos definições e

acordos formuladas e assumidos nos capítulos anteriores. Agora, quando você lançou o dado e observou o resultado, nós nos referimos a este por "valor simulado daquela variável aleatória´´.

O exemplo acima indica que "o valor simulado´´ de uma variável aleatória é um dos valores possíveis da mesma. Isto está certo, porém isto não implica que ao ser solicitado a dar um valor simulado da variável aleatória supracitada você pode simplesmente escolher, ao seu gosto, um dos números entre 1, 2, 3, 4, 5 e 6. O certo é proferir (ou, realizar, em outras palavras) o experimento aleatório e dar o valor que a variável aleatória assume na realização desse.

Do exemplo acima, você entendeu que se eu lhe pedir um valor simulado de uma variável aleatória, então você precisa realizar o experimento aleatório, no qual a desejada variávela aleatória foi por mim definida, observar sua realização e, seguindo minha definição, achar o valor que minha variável aleatória assume para a realização observada; este valor é sua resposta ao meu pedido. Se é assim que você entendeu, então devo te parabenizar, pois você está certo. Permita-me só descrever a mesma coisa mas com palavras diferentes: valor simulado de uma variável aleatória é um dos seus valores possíveis que essa assume numa realização do experimento aleatório no quel foi construida.

Mas sendo comprovada sua exatidão na compreenção, lhe surge, naturalmente, a dúvida existencial: "Tu precisa de um capítulo inteiro para coisa tal simples?´´ Dúvida esta vem do seguinte pensamento: "Tu quer de mim um valor de uma variável aleatória. Certo. Para tal, você descreve para mim o experimento aleatório e a definição daquela variável aleatória. E ai, eu vou ter que realizar seu experimento, e ver e te dizer o valor dessa variável na realização do experimento que der. É só isto e nada mais. Inclusive, se me permitir uma má educação, vou te mandar realizar seu experimento com forças próprias, pois já que tu inventou esse então ninghem outro consegue realizar sua invenção melhor que tu!´´

De novo, você está correta, menos num pequeno detalhe: temos, e ainda teremos, diversas variáveis aleatórias definidas por tabela de sua distribuição, quer dizer, sem a presença de um experimento aleatório por tráz delas. Parece um caso complicado, mas não é, na verdade; mostrarei adiante uma saída simples. O segundo ponto não tal-tal óbvio para você neste momento é como simular variáveis aleatórias contínuas. Este problema também terá uma solução simples a ser apresentada. E haverão mais detalhes miudinhos, os quais junto com os dois maiores acima mencionados já dão assunto para um capítulo, que é o presente.

Sua segunda dúvida é sobre a utilidade da simulação como descrita acima.

Por fim, gostaria de apresentar um dicionário de sinónimos ligados ao assunto Simulação de Variáveis Aleatórias. Você pode dar pouca atenção à esta informação, pois eu tentarei usar linguagem padronizado. Mas como o padrão é meu, os textos de outros autores podem te supreender pela nomenclatura diferente, e é por isto que lhe apresneto esta. valor simulado (de uma variável aleatória) = realização (de uma variável aleatória) =

simular simular variável aleatória gerar um valor de uma variável aleatória sacar um valor de uma variável aleatória simular obter realização gerar realização obter valor gerar valor número aleatório resultado de simulação sacar um valor da distribuição sacar um valorda va com uma distribuição sacar um valor distribuido tal-e-tal

6.5 Gerador de Números Aleatórios

Gerador de números aleatórios é o nome genérico para programa de computador que, como o próprio nome indica, gera números aleatórios. Quase todo pacote possui um de tal gerador. No EXEL, o comando que invoca o programa é ????.

Para entender o que o gerador faz e, o mais importante, como usá-lo para simulação, você precisa primeiramente passar pela exposição dos dois próximos parágrafos.

Começamos a exposição voltando nossa atenção ao objeto já conhecido: a variável alaeatória "o número observado no lançamento de uma dado equilibrado". Você reconhece a tabela abaixo

como a da distribuição desta variável aleatória:

Table to be put here

Esta distribuição chama-se *uniforme em números 1, 2, 3, 4, 5 e 6.* O termo "unifmorme" significa aqui e em toda a exposição a seguir, que a distribuição atribui a mesma probabilidade a todo valor.

Agora ficou fácil para você entender ao que me refiro quando digo "distribuição uniforme nos valores de uma grade regular do intervalo $[0, 1]$".

6.6 Simulação de uma variável aleatória discreta

O nome correto no título deveria ser "simulação de uma variável aleatória discreta dada via distribuição", pois se fosse dada via experimento aleatório, então as idéias acima apresentadas serveriam para a simulação.

Então, suponha que lhe peço simular um valor da variável aleatória Y e lhe digo sobre esta que sua distribuição é assim:

y	2	3	7
$\mathbb{P}[Y = y]$	0,2	0,5	0,3

Para executar esta tarefa é preciso construir um experimento aleatório, construir neste uma variável aleatória que tenha a mesma distribuição que Y acima, e simular um valor da variável construida. Este valor será a resposta à tarefa original.

Você não precisa da justificativa para o programa bolado acima.

Vamos ao passo (a) do programa. Tem-se uma liberdade absoluta na escolha de experimento aleatório. Porém, deve-se ter em mente que no passo seguinte, este experimento... Por exemplo, o lançamento de um dado equilibrado não seria compatível com a distr ibuição de Y pois é impossível nele criar eventos com probabilidades 0,2 e 0,3.

Pensando nesta compatibilidade, torna-se atrativa a sugestão para o experimento aleatórioque daremos a seguir, pois, como veremos, a sugestão funcionará para qualquer distribuição.

Então, nossa sugestão é que o experimento aleatórioseja: obtenção de um valor por um Gerador de Números Aleatórias. Neste experimento aleatório, vamos definir uma variável aleatória da sorte tal que sua distribuição seja idêntica à de Y. Vamos chamar a variável aleatória a ser contruida por $\mathcal{Y}$, assinalando pela diferença com símbolo Y o fato que as variáveis $\mathcal{Y}$ e Y podem ser diferentes por terem origem diferentes.

Chamaremos por U o valor a receber do Gerador de Números Aleatórias. Por favor, certifique-se que você entendeu a mensagem por completo; a mensagem é que, sendo visto da posição de antes da chamada do Gerador, o valor a receber está incerto e isto repercurte no nome e notação que demos a ele: "variável aleatória U".

Se você não entendeu meu comentário acima, pode então pular a exposição a seguir, e ir diretamente ao texto que mostra a parte técnica do assunto, quer dizer, mostra o que deve ser feito para chegar ao resultado final.

O ponto levantado dois parágrafos acima não é que podemos chamar por U o resultado a receber do Gerador, pois o nome é de menos que importa. O ponto é que sabemos a distribuição do valor a receber. Este nosso conhecimento veio da propriedade (???) de geradores de números aleatórios. Desta propriedade, segue-se que

Propriedade 23 Para qualquer intervalo $[a, b]$ escolhido antemão no intervalo $[0, 1]$, a probabilidade do valor gerado cair nesse é o seu comprimento; isto, sendo colocado em termos precisos soa da seguinte maneira:

$$\mathbb{P}\left[U \in [a, b]\right] = b - a, \text{ para qualquer } [a, b] \text{ contido por inteiro em } [0, 1]. \tag{6.18}$$

Esta propriedade ajudar-nos-á quando formos justificar nosso algoritmo de simulação. No presente, vamos executar este algoritmo para nosso caso especial. O cerne do algoritmo é a função que receberá o valor do Gerador e entregará o valor simulado da variável em interesse, quer dizer, de Y. A função depende da distribuição de Y. Isto é quase óbvio, mas, para garantia, está destacado agra. Sua cara está na Figura caraFy e ela surge assim: como Y pode assumir 3 valores, então dividimos o intervalo $[0, 1]$ em 3 intervalos, sendo que o comprimento de cada intervalo é igual à respectiva probabilidade, isto é, o comprimento

6.7 Simulação de uma variável aleatória contínua

6.8 Simulação simultânea de duas ou mais variáveis aleatórias

6.9 Exercícios

Exercício 65. Você pode lançar duas vêzes uma moeda honesta. Explique como usando os resultados destes lançamentos você vai obter uma realização da variável aleatória X que assume valor 1 com probabilidade 1/4 e valor 4 com probabilidade 3/4.

Exercício 66. Você pode lançar uma vez um dado equilibrado e uma vez uma moeda honesta. Explique como usando os resultados destes lançamentos você vai obter uma realização da variável aleatória X que assume valor 1 com probabilidade 1/12, valor 3 com probabilidade 3/12, e valor 5 com probabilidade 8/12.

Exercício 67. Use os números aleatórios das primeiras duas linhas de seu bloco[1] para simular 20 valores da variável aleatória $X_{0,5}$ com a seguinte distribuição

$$I\!P\left[X_{0,5} = 1\right] = 0,5, \qquad I\!P\left[X_{0,5} = 0\right] = 1 - 0,5 = 0,5 \tag{6.19}$$

(O índice $0,5$ da variável aleatória X assinala a probabilidade dela assumir valor 1.)

Calcule a média e a variância amostrais da amostra dos valores simulados. Compare os valores calculados com a esperança matemática e com a variância da variável aleatória $X_{0,5}$.

Exercício 68. Repita Exercício 3 usando outras 4 pares de linhas de números aleatórios (de seu bloco e do bloco que está debaixo dele na tabela). Para cada 20 valores simulados calcule a média e a variância amostrais. Juntando os resultados deste exercícios com o do exercício anterior, você tem 5 médias amostrais. Quantas delas se distanciam da esperança matemática de variável aleatória simulada por menos que $0,05$?
Juntando os resultados deste exercícios com o do exercício anterior, você tem 5 variâncias amostrais. Quantas delas se distanciam da variância de variável aleatória simulada por menos que $0,25$?

Exercício 69. **(a)** Repita os Exercícios 3 e 4 para variável aleatória $X_{0,9}$, quer dizer, para a variável aleatória com a seguinte distribuição:

$$I\!P\left[X_{0,9} = 1\right] = 0,9, \qquad I\!P\left[X_{0,9} = 0\right] = 1 - 0,9 = 0,1 \tag{6.20}$$

(b) Repita os Exercícios 3 e 4 para variável aleatória $X_{0,2}$, quer dizer, para a variável aleatória com a seguinte distribuição:

$$I\!P\left[X_{0,2} = 1\right] = 0,2, \qquad I\!P\left[X_{0,2} = 0\right] = 1 - 0,2 = 0,8 \tag{6.21}$$

[1]Para as simulações use a Tabela de Números Aleatórios. Esta tabela agrupa números aleatórios em blocos; são 24 blocos ao todo. O número de "seu" bloco é igual ao soma dos últimos dois algoritmos de seu número USP. Você vai usar "seu" bloco e o bloco que está debaixo dele na tabela.

Exercício 70. Usando a tabela de números aleatórios, construa três realizações da variável aleatória X que tem a seguinte distribuição:

x	1	3	5	7	9
$\mathbb{P}[X = x]$	1/20	2/20	3/20	5/20	9/20

Exercício 71. Usando a tabela de números aleatórios, construa três realizações da variável aleatória X que tem a seguinte distribuição:

x	1	3	5
$\mathbb{P}[X = x]$	1/5	3/10	1/2

Exercício 72. Usando a tabela de números aleatórios, construa três realizações da variável aleatória normal padrão Z. Use a tabela para gerar variável aleatória que tem distribuição uniforme nos valores

$$0;\ 0,0001;\ 0,0002;\ \ldots;\ 0,9999$$

Exercício 73. Use os dois blocos de números aleatórios para construir 25 numeros aleatórios com 4 casas depois de vírgula. Use estes números para simular 25 realizações da variável aleatória normal com média 3 e variância 4 (quer dizer, variável aleatória $Y \sim \mathcal{N}(3; 4)$). Calcule a média e a variância amostrais dos valores simulados. Ache a distância entre a média amostral e a esperança matemática da variável aleatória simulada. Ache a distância entre a variância amostral e a variância da variável aleatória simulada.

Acrescido enquanto trabalhava com outros capítulos: Aqui podemos e devemos falar da função de distribuição acumulada. Como um dos exercícios sobre este tema, pode ser feito o seguinte: considerar as variáveis aleatórias geométricas com parâmetros diferentes, desenhar as suas funções-probabilidade e mostrar que as funções não carregam muito informação acerca da comparação: a com q mener, começa com valor menos, mas, em compensação decai masi devagar que outra; então parece que não dá comparação entre as duas. Mas, quando apresentarmos por funções de distribuição acumulada, veemos que uma fica abaixo da outra, e para sempre.

Função de distribuição acumulada

Os próximos exercícios vieram de uma das listas do curso MAE112. O enfoque daquela lista eram as funções de distribuição acumulada. Precisa verificar se a inclusao dos exercícios no presente capítulo faz sentido.

3. Seja X tal que $\mathbb{P}[X = 0] = 1/3$ e $\mathbb{P}[X = 1] = 2/3$. Definimos nova variável aleatória Y como $Y = 2X - 1$. Desenhe no mesmo desenho as funções de probabilidade de X e de Y. Desenhe no mesmo desenho as funções de distribuição acumulada de X e de Y.

4. Seja X a variável aleatória com a seguinte distribuição:

x	1	3	5
$\mathbb{P}[X = x]$	1/6	2/6	3/6

Faça no mesmo desenho as funções de probabilidade de X, de $2X$ e de $X + \frac{1}{2}$. Faça no mesmo desenho as funções de distribuição acumulada de X, de $2X$ e de $X + \frac{1}{2}$. Observe que apesar de que $2X$ é o dobro de X, a função da distribuição acumulada de $2X$ não é o dobro da de X. Já sobre as função de probabilidade destas variáveis aleatórias, podemos dizer que a segunda é o dobro da primeira, entendendo com isso que a segunda é a primeira esticada duas vezes ao longo do eixo de x.

5. Sejam X e Y duas variáveis aleatórias independentes e distribuidas da seguinte maneira:

x	0	1
$\mathbb{P}[X = x]$	1/2	1/2

y	0	1
$\mathbb{P}[Y = y]$	1/2	1/2

Definimos nova variável aleatória Z como a soma das duas: $Z = X+Y$. Faça no mesmo desenho as funções de probabilidade de X, de Z. Faça no mesmo desenho as funções de distribuição acumulada de X, de Z. Observe que apesar de que Z é a soma das cópias independentes de X, a função da distribuição acumulada de Z não é o dobro da de X.

Note que a independências das variáveis aleatórias X e Y é essencialmente usada no cálculo da distribuição de Z. Por exemplo,

$$\begin{aligned}
\mathbb{P}[Z = 1] &= \mathbb{P}[X = 1 \text{ e } Y = 0 \text{ ou } X = 0 \text{ e } Y = 1] \\
&= \mathbb{P}[X = 1 \text{ e } Y = 0] + \mathbb{P}[X = 0 \text{ e } Y = 1] \\
&= \mathbb{P}[X = 1] \cdot \mathbb{P}[Y = 0] + \mathbb{P}[X = 0] \cdot \mathbb{P}[Y = 1] = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}
\end{aligned}$$

6. Repita Ex. 5 substituindo a distribuição de Y (somente de Y!) por

y	0	2
$\mathbb{P}[Y = y]$	1/2	1/2

6.10 Exercícios sobre populações e amostras

Aviso. Na prática, quando se analisa um conjunto de dados, é importante saber se ele contém as observações duma população ou duma amostra. A importânica advém da diferença entre os objetivos da análise. A análise de uma amostra tem por seu objetivo conseguir diversas inferências sobre aquela população de onde a amostra foi retirada. A construção dos métodos das inferências desejadas devem tomar em conta os fatores aleatórios inevitavelmente presentes no processo de amostragem. No momento, não falaremos sobre tais fatores, mas a menção deles agora é importante pois são eles que causam as diferenças nos tratamentos de conjuntos de dados de amostra e de população, já que na obtenção de dados de população qualquer fator aleatório está ausente por completo.

No que se tange ao conteúdo desse capítulo, as diferenças supramencionadas afetam a construção de quartis (assim como todos os q-quantis sobre os quais não foi dito nada nas aulas), a construção de box-plot e o cálculo da variância de conjunto (e todas as medidas de posição e de dispersão que dependem dessa variância).

As três diferenças destacadas no parágrafo anterior não são importantes para a aprendizagem das idéias e dos métodos que tenciono ensinar no presente capítulo. Por isso, com o intuíto de assingelar nossas vidas, sugiro que em todas os exercícios do capítulo, os alunos tratem todos os conjuntos de dados seguindo os seguintes padrões:

(**i**) Que os quartis (melhor dizer, os valores numéricos dos quartis) $Q1, Q2, Q3$ sejam calculados de acordo com a regra abaixo. Na formulação da regra, eu uso a notação $x_{(k)}$ para me referir à k-ésima observação do conjunto ordenado (da menor para maior) de observações; note a diferença entre x_k e $x_{(k)}$: o primeiro é a k-ésima observação na ordem de chegada (que tipicamente não está determinada pois–realmente–não desempeno papel algum) enquanto que a segunda é a observação que está no k-ésimo lugar após que todas as observações foram ordenadas.

$$Q1 = \begin{cases} x_{((N+1)/4)}, \text{ se } (N+1)/4 \text{ for inteiro} \\ \frac{1}{2}\left\{x_{\left(\underline{(N+1)/4}\right)} + x_{\left(\overline{(N+1)/4}\right)}\right\}, \text{ se } (N+1)/4 \text{ não for inteiro} \end{cases} \tag{6.22}$$

$$Q2 = \begin{cases} x_{((N+1)/2)}, \text{ se } (N+1)/2 \text{ for inteiro} \\ \frac{1}{2}\left\{x_{\left(\underline{(N+1)/2}\right)} + x_{\left(\overline{(N+1)/2}\right)}\right\}, \text{ se } (N+1)/2 \text{ não for inteiro} \end{cases} \tag{6.23}$$

$$Q3 = \begin{cases} x_{(3(N+1)/4)}, \text{ se } 3(N+1)/4 \text{ for inteiro} \\ \frac{1}{2}\left\{x_{\left(\underline{3(N+1)/4}\right)} + x_{\left(\overline{3(N+1)/4}\right)}\right\}, \text{ se } 3(N+1)/4 \text{ não for inteiro} \end{cases} \tag{6.24}$$

(Nas fórmlas acima $\underline{r}$ significa o maior dos inteiros que são menores que r, enquanto que $\overline{r}$ significa o menor dos inteiros que são maiores que r. Por exemplo, $\underline{9,25} = 9$ e $\overline{9,25} = 10$.)

(**ii**) Que os box-plots estejam desenhados com base nos valores dos quartis calculados de acordo com o item acima, e nos valores do máximo e do mínimo (e sem a identificação de valores aberrantes). Observo que a diferença na construção de box-plot supramencionada não é somente a consequencia da diferença dos cálculos de quartis para populações e para amostras; a diferença expressa-se também na identificação e marcação de valores aberrantes do conjunto representado pelo box-plot. Especificamente, nos exercício do capítulo, eu não espero/solicito que os valores aberrantes sejam identificados e marcados.

(**iii**) Que a variância seja calculado de acordo com a fórmula

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{6.25}$$

que foi introduzida para tratar conjuntos de observações de populações. Recordo para o título de completude de minha apresentação que a variância de conjunto de observações que forma uma amostra calcula-se pela fórmula

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{6.26}$$

Noto ainda que a diferença principal entre as duas está no denominador; ela tem sua razão a ser, e conforme já mencionado acima, por tráz dela há a aleatoriadade presente na obtenção das observações de amostra que combina-se com o objetivo de usar a amostra a fim de estimar a variância da população. Quanto ao uso de letra minúscula ou maiúscula para o tamanho do conjunto de observações, essa diferença é o tributo à tradição: para uma população, a tradição manda denotar seu tamanho por N, enquanto que para uma amostra–por n.

(**iv**) Por último, decidi que vale declarar explicitamente que o método de histogramas é o mesmo tanto no caso de população quanto no de amostra.

Ainda com o intuíto de evitar suas surpresas desagradáveis, devo lhe avisar de que alguns pacotes de programas fazem o cálculo dos valores de quantis pelo caminho que é pouco diferente daquele que expliquei acima. Para entender uma das diferenças, oberva que o valor de a parte após a vírgula de $(N+1)/4$ pode ser "$,25$", ou "$,5$", ou "$,75$". Quando essa parte for "$,25$", alguns programas calculam $Q1$ não pela fórmula acima, mas pela seguinte regra:

$$Q1=\frac{3}{4}x_{(\underline{(N+1)/4})}+\frac{1}{4}x_{(\overline{(N+1)/4})}$$

Essa fórmula atende à concepção de que a posição de $Q1$ no conjunto de dados ordenados está mais próxima aos primeiros 25% de dados do que aos 75% de dados maiores, e, de acordo com essa concepção, a fórmula desloca o valor de $Q1$ na direção do conjunto de 25% de dados menores por intermédio de aumento do peso junto ao $x_{(\underline{(N+1)/4})}$ para 3/4 em relação de 1/2, que era o valor do peso na fórmula (1.22). As alterações semelhantes ocorrem quando a posição acaba em "$,75$", só que nesse caso, o peso diminui-se para 1/4. As mesmas regras aplicam-se ao cálculo do valor de $Q3$. Concluíndo: você deve fazer os cálculos dos valores de quantis usando o método descrito no item (**i**) acima, mas se seus resultados forem pouco diferentes dos resultados fornecidos por aplicativos do tipo EXEL, R, SAS, você não deve ficar surpreso pela diferença. Inclusive, alguns gabaritos dos exercícios dessa seção foram feitos com o uso desses aplicativos; é bom que você esteja avisado sobre isso.

Exc. 74. *(Esse exercício trata conjuntos pequenos de dados para proporcioná-lhe a possibilidade de acompanhar visualmente o funcionamento da regra que calcula quartis de conjunto de dados.)*

Calcule

(a)	Q1, Q2 e Q3 para		1	3	6	13			
(b)	Q1, Q2 e Q3 para		1	3	5	6	13		
(c)	Q1, Q2 e Q3 para		1	3	5	5	6	13	
(d)	Q1, Q2 e Q3 para	1	1	3	5	5	6	13	
(e)	Q1, Q2 e Q3 para	1	2	3	4	5	6	7	
(f)	Q1, Q2 e Q3 para	1	2	3	4	5	6	7	8

Exc. 75. *(Fazendo esse exercício, você desenvolve a sensibilidade acerca da mudança de valores dos quartis de conjunto em resposta à pequenas mudanças nos valores de algumas poucas observações ou em resposta ao acrescimo de algumas poucas observações.)*

(a) Para o conjunto de 20 observações dado abaixo, calcule a média, a variância, os quartis e construa o box-plot:

10 13 21 28 29 31 39 50 52 54 54 67 75 79 80 83 86 87 88 97

(b) No conjunto de observações do item (a) foram feitas as seguintes modificações: adicionou-se 3 unidades nas observações 1, 19 e 20:

13 13 21 28 29 31 39 50 52 54 54 67 75 79 80 83 86 87 **91** **100**

Calcule a média, a variância, os quartis e construa o box-plot do conjunto acima. Compare com os correspondentes resultados obtidos no item (a).
(c) Agora, no conjunto de observações do item (a) foram feitas modificações diferentes, a saber: adicionou-se 2 unidades nas observações 5 e 15 e subtraiu-se 2 unidades das observações 10 e 11:

10 13 21 28 **31** 31 39 50 52 **52** **52** 67 75 79 **82** 83 86 87 88 97

Calcule a média, a variância, os quartis e construa o box-plot do conjunto acima. Compare com os correspondentes resultados obtidos no item (a).
(d) Considere agora as observações abaixo que resultaram das modificações do conjunto do item (a) da seguinte natureza: adicionou-se 3 unidades nas observações 19 e 20 e subtraiu-se 3 unidades da observações 1 e 2:

7 **10** 21 28 29 31 39 50 52 54 54 67 75 79 80 83 86 87 **91** **100**

Calcule a média, a variância, os quartis e construa o box-plot do conjunto acima. Compare com os correspondentes resultados obtidos no item (a).

Exc. 76. *(O objetivo desse exercício é similar ao do Exc. 11, só que no presente exercício foi colocada uma observação atípica com o intuito de chamar sua atenção à forma como essa deve estar apresentada no box-plot do conjunto de todas as observações.)*
Os dados abaixo representam velocidades do vento (km/h) num determinado aeroporto para os primeiros 15 dias de dezembro de 2008:

22,2 61,1 13,7 27,8 22,7 7,4 8,7 6,3 20,4 25,6 23,2 11,1 13,0 7,2 14,8

(a) Calcule a média, a mediana, o desvio padrão e os quartis da velocidade. Faça o box-plot.
(b) Note que o dia 2 de dezembro apresenta um valor atípico devido a uma tempestade forte com chuva e vento. Remova esse valor e refaça o item anterior. Comente as diferenças encontradas.
(c) Construa os box-plots dos dados originais e dos dados sem a observação do dia 2 de dezembro.

Exc. 77. *(Fazendo esse exercício, você desenvolve a precepção acerca do "mecanismo" de mudança de valores dos quartis de conjunto em resposta mudança que envolve todas as observações do conjunto.)*
(a) Considere o conjunto 1, 3, 7, 8, 11. Para este, calcule a média, o mínimo, o máximo, os quartis, a variância, o desvio padrão, o intervalo interquartil (recordo que este é $Q3 - Q1$ e que seu valor pode ser usado como a medida de dispersão de dados), e o coeficiente de variação (recordo: $CV =$ (desvio padrão/média)).
(b) Acrescente 2 a cada valor do conjunto e repita os cálculos. Analise como o acrescimo afetou os valores de média, mínimo, máximo, quartis, variância, desvio padrão, intervalo interquartil e o coeficiente de variação.
(c) Multiplique por 3 cada valor do conjunto e repita os cálculos. Analise como o multiplicador afetou os valores de média, mínimo, máximo, quartis, variância, desvio padrão, intervalo interquartil e o coeficiente de variação.

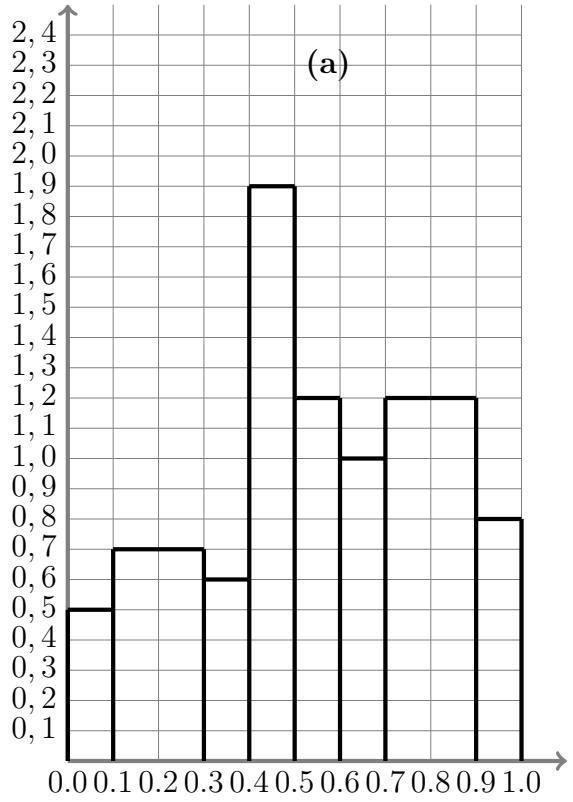
Exc. 78. *(Esse exercício é o último do ciclo de exercícios que trabalham com a mudança dos valores de quartis de um conjunto em resposta à mudança dos valores de observações deste conjunto. Diferentemente dos exercícios anteriores, no presente, é você quem precisa alterar os valores de observações com o objetivo de alcançar a mudança de quartis solicitada pelo enunciado.)*

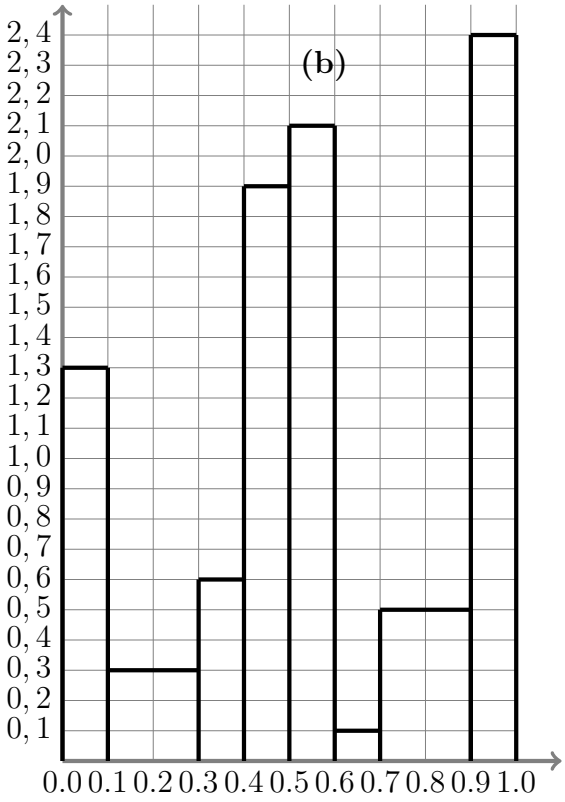
Considere o seguinte conjunto de observações:

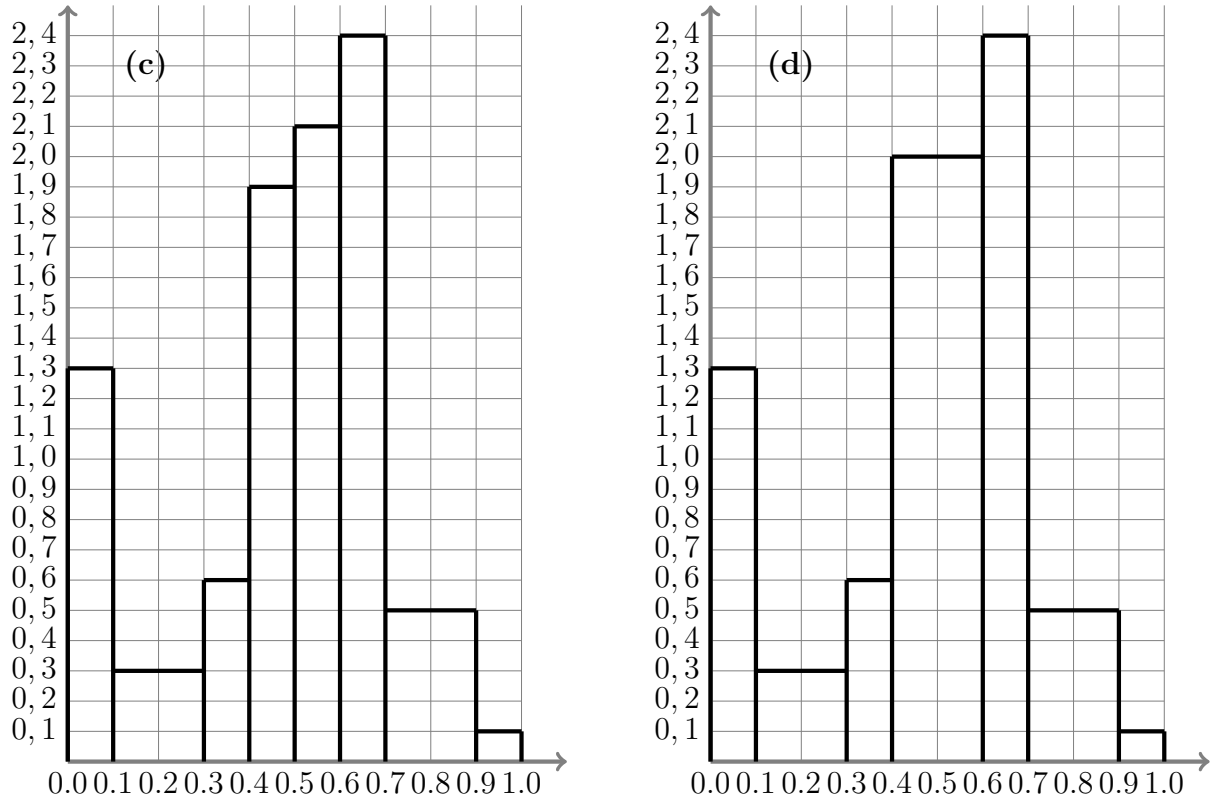
$$1 \quad 4 \quad 6 \quad 7 \quad 9 \quad 10 \quad 12 \quad 15$$

(a) Acrescente 4 observações da maneira tal que os valores dos $Q1, Q2$ e $Q3$ não se alterem, e o valor da média seja o mesmo que o do conjunto original.
(b) Acrescente 4 observações da maneira tal que os valores dos $Q1, Q2$ e $Q3$ não se alterem, mas o valor da média seja maior que o do conjunto original.
(c) Modifique o valor de uma das observações da maneira tal que os valores dos $Q1, Q2$ e $Q3$ não se alterem, mas o valor da média seja maior que o do conjunto original.
(d) Modifique os valores das observações ou acrescente novas observações da maneira tal que os valores dos $Q1, Q2, Q3$ e a média não se alterem, mas a variância aumente.

Exc. 79. Em cada um dos quatro desenhos abaixo, você vé o histograma feito para um conjunto de dados. Com base na informação apresentada pelo histograma, o que você pode afirmar sobre os valores dos quartis do conjunto em cada desenho?







Exc. 80. Valores de porosidade (em %) de 57 arenitos encontrados a uma profundidade de 1100 a 1105 metros são mostrados a seguir:

19,8	20,0	21,0	21,1	21,2	21,5	21,6	21,7	21,8	21,8	21,9
22,0	22,1	22,1	22,1	22,2	22,3	22,3	22,3	22,3	22,4	22,4
22,4	22,5	22,6	22,6	22,8	23,0	23,0	23,1	23,2	23,3	23,5
23,7	23,7	23,7	23,8	23,9	23,9	23,9	24,0	24,1	24,2	24,2
24,4	24,9	25,0	25,2	25,3	25,3	25,6	26,0	26,6	27,3	27,6
27,9	28,6									

(a) Construa o histograma correspondente ao seguinte conjunto de amplitudes:

$$[19, 20), [20, 21), [21, 22), [22, 23), [23, 24), [24, 25), [25, 26), [26, 27), [27, 28), [28, 29]$$

Quais caraterísticas da distribuição populacional são sugeridas por esse histograma?
(b) Construa o histograma correspondente ao seguinte conjunto de amplitudes:

$$[19, 21), [21, 22), [22, 23), [23, 24), [24, 25), [25, 26), [26, 29]$$

Quais caraterísticas da distribuição populacional são sugeridas por esse histograma?
(c) Construa o box-plot da amostra. Quais caraterísticas da distribuição populacional são sugeridas por esse histograma?
Observação: Certifique-se que você entende plenamente ao que refere-se o termo "a distribuição populacional" que apareceu nas questões do exercício.

Exc. 81. (da apostila do MAE0116) Os dados a seguir referem-se a medidas de prostaglandina (pg/ml) e cálcio (ml/dl) em pacientes com câncer apresentando ou não hipercalcemia (os pacientes do 1-o a 11-o são os que apresentam a hipercalcemia, o os do 12-o ao 22-o são os que não). Desenhe o boxplot para as variáveis prostaglandina e cálcio, separando por grupos com e sem hipercalcemia, e use-los para concluir sobre o efeito da hipercalcemia nas medidas de prostaglandina e cálcio.

Com hipercalcemia			Sem hipercalcemia		
Paciente	prostag-landina	cálcio	paciente	prostag-landina	cálcio
1	500	13.3	12	254	10.1
2	500	11.2	13	172	9.4
3	301	13.4	14	168	9.3
4	272	11.5	15	150	8.6
5	226	11.4	16	148	10.5
6	183	11.6	17	144	10.3
7	183	11.7	18	130	10.5
8	177	12.1	19	121	10.2
9	136	12.5	20	100	9.7
10	118	12.2	21	88	9.2
11	60	18	22	60	11.2

Exc. 82. Considere o seguinte conjunto de dados:

3.1,	3.1,	3.5,	4.3,	4.4,	4.5,	4.7,	4.9,	4.9,	5.0,	5.1,	5.2,
5.3,	5.3,	5.4,	5.6,	5.7,	5.8,	5.8,	5.8,	5.9,	5.9,	6.0,	6.1,
6.1,	6.2,	6.2,	6.2,	6.2,	6.4,	6.5,	6.5,	6.6,	6.6,	6.6,	6.6,
6.6,	6.7,	6.7,	6.8,	6.8,	6.8,	6.8,	6.8,	6.9,	6.9,	6.9,	7.0,
7.0,	7.1,	7.1,	7.1,	7.1,	7.2,	7.2,	7.3,	7.3,	7.3,	7.3,	7.3,
7.3,	7.3,	7.3,	7.4,	7.5,	7.6,	7.7,	7.8,	7.8,	7.8,	7.9,	7.9,
7.9,	7.9,	8.0,	8.0,	8.0,	8.0,	8.0,	8.0,	8.1,	8.1,	8.2,	8.2,
8.2,	8.2,	8.3,	8.3,	8.3,	8.3,	8.4,	8.5,	8.5,	8.5,	8.5,	8.5,
8.5,	8.6,	8.6,	8.6,	8.6,	8.7,	8.7,	8.8,	8.9,	8.9,	8.9,	8.9,
9.0,	9.0,	9.1,	9.1,	9.1,	9.2,	9.3,	9.3,	9.4,	9.6,	9.8,	9.8

(a) Construa o histograma (para a frequencia relativa) correspondente as seguintes classes:

$$[3, 5), [5, 8), [8, 10]$$

(b) Construa o histograma (para a frequencia relativa) correspondente as seguintes classes:

$$[3, 4), [4, 5), [5, 6), [6, 7), [7, 8), [8, 9), [9, 10]$$

Quais caraterísticas da distribuição populacional são sugeridas por esse histograma? (Esta pergunta é opcional; nada parecido estará na prova.)

(d) Construa o box-plot desse conjunto de dados.

6.11 Soluções dos exercícios sobre população e amostra

Solução do Exc. 11

(a)

Min.	Q. 1	Mediana	Média	Q. 3	Máx.	Var.
10.00	30.50	54.00	56.15	80.75	97.00	766.23

(b)

Min.	Q. 1	Mediana	Média	Q. 3	Máx.	Var.
13.00	30.50	54.00	56.60	80.75	100.00	775.83

(c)

Min.	Q. 1	Mediana	Média	Q. 3	Máx.	Var.
10.00	31.00	52.00	56.15	82.25	97.00	767.29

(d)

Min.	Q. 1	Mediana	Média	Q. 3	Máx.	Var.
7.00	30.50	54.00	56.15	80.75	100.00	819.29

Desenhos do exercícios que devem ser colocados como figura nessa solução:
`V-EstatisticaDescritivaExercicios-boxplot1.pdf`,
`V-EstatisticaDescritivaExercicios-boxplot2.pdf`,
`V-EstatisticaDescritivaExercicios-boxplot3.pdf`,
`V-EstatisticaDescritivaExercicios-boxplot4.pdf`

Solução do Exc. 14. Precisaremos saber que $Q1 = 5, Q2 = 8, Q3 = 11, media = 8$. Tais valores calculam-se diretamente do conjunto de dados.
(a) Basta adicional 1 observação a cada 25% das observações e coloca-las simetricamente em torno da média do conjunto original (que era 8): 2 e 14, e 6,5 e 9,5. Aquí por "cada 25% das observações" me referi aos quatro subconjuntos de observações que encopntram-se entre *min* e $Q1$, entre $Q1$ e $Q2$, entre $Q2$ e $Q3$, e entre $Q3$ e *max*.

(b) Basta adicional 1 observação a cada 25% das observações e coloca-las assimetricamente em torno da média do conjunto original, dando assim um deslocamento da mesma à direita. Uma das possibilidades é: 2, 14, 6.5 e 9.6. Esta solução "aproveita" da solução do item(a), mas desloca 9.5 para 9.6.

Outra possibilidade é "aproveitar" da mesma solução do item (a), mas deslocar a observação 14 acrescida ao bigode direito do box-plot. Já que não é precisa preservar o valor do extremo direito do bigode, deslocamos 14 para 16. Isto dá a seguinte solução: os valored as quatro observações acrescidas são 2, 16, 6.5 e 9.5.

É claro que as duas soluções apresentadas não são únicas. Por exemplo, 2, 16, 6.5 e 9.6 também resolvem a tarefa.

(c) Aqui o mais fácil é aumentar o valor da observação que dá o extremo direito do bigode direito. Esta observação não participa diretamente na conta dos valores de $Q1, Q2$ e $Q3$, mas entra na conta da média com o mesmo peso que todas as outras. Portanto, uma das soluções é: 1 4 6 7 9 10 12 17

(d) A maneira mais fácil aquí é afastar as observações extremais para mais longe da média: 0 4 6 7 9 10 12 16

Já sabemos que estas observações não conseguem afetar os valores dos quantis. Também, nossa alteração não afetou a média pois as duas observações se posicionavam da menira simêtrica em torno da média, e nos mativemos esta simetria ao afastar as duas da média. Mas este

afastamento aumentou a variânica pois o valor desta é a soma ponderada dos quadrados das distânicas entre a média e as observações.

Solução Para a construção dos gráficos box-plot, calcula-se os valores da mediana, primeiro quartil, terceiro quartil, LI e LS da mesma forma como feito no exercicio 1 para os conjuntos de dados A e B. Os valores encontram-se dispostos na tabela 1.1.

Tabela 6.1: Medidas para construção do box-plot.

	Conjunto A	Conjunto B
Mediana	3,66	4,42
Primeiro quartil	2,345	2,845
Terceiro quartil	4,69	5,845
LI	-1,1725	-1,655
LS	8,2075	10,345

Dessa forma, observando-se que os conjuntos de dados não apresentam valores atípicos, construiu-se o box-plot exposto na Figura ??

Verifica-se pela Figura ??que os dados do conjunto B possuem maior variação do que os dados do conjunto A, assim como um maior valor de mediana. O box-plot indica também que o conjunto B possui distribuição semelhante à distribuição do conjunto A deslocada para valores maiores.

Solução ao Exc. 12. (a) Cálculo da média:

$$\overline{x} = \frac{\sum_{j=1}^{n} x_i}{n} = \frac{\sum_{j=1}^{15} x_i}{15} = \frac{285,2}{15} = 19,0133$$

Cálculo da mediana: Para calcular a mediana é preciso primeiro calculae sua posição, considerando que ela é o quartil de ordem 0,5, a posição segue-se da aplicação da fórmula $pos = 0,5(n+1)$, o que dá $(15+1)0,5 = 16 * 0,5 = 8$. Isto quer dizer que o valor da mediana é igual ao valor da oitava observação dos dados ordenados. O resultado da ordenação é assim:

$$6,3\ 7,2\ 7,4\ 8,7\ 11,1\ 13\ 13,7\ 14,8\ 20,4\ 22,2\ 22,7\ 23,2\ 25,6\ 27,8\ 61,1$$

Portanto, Mediana= $14,8$.
Cálculo do primeiro e terceiro quartis segue o mesmo caminho que o do cálculo da mediana, com a única diferença que na fórmula para a posição usa-se o fator $0,25$ (para Q1) ou $0,75$ (para Q3) no lugar de $0,5$. O resultado é: $Q1 = 9,9$ e $Q3 = 22,95$.

Calculo do desvio padrao: A variância (que precisa ser calculada para descobrir o desvio padrão, já que este é a raíz quadrada da variância) é calculada pela fórmula:

$$\frac{\sum_{j=1}^{n} (x_i - \overline{x})^2}{n-1}$$

já que trata-se da variância de uma distribuição amostral (se fosse a de distribuição populacional, usariamos n no lugar de $n-1$ no denuminador da fórmula). Temos, então:

$$s = \sqrt{\frac{\sum_{j=1}^{n} (x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{j=1}^{n} x_i^2 - n\overline{x}^2}{n-1}} = \sqrt{\frac{8044,86 - 15 * 361,5068}{15-1}} = 13,6859$$

É de custume colocar numa tabela os valores calculados até o momento e chamar a tabela por Tabela de Estatísticas descritivas. No presente caso, esta está abaixo:

	Valor
Média	19,0133
Mediana	14,8
Desvio Padrão	13,6859
Primeiro quartil	9,9
Terceiro quartil	22,95

Para a construção do box-plot são necessários os valores de mediana, primeiro e terceiro quartis, e além destes, é necessário verificar a presença de pontos atípicos nos dados. Executaremos primeiramente a última tarefa. Calculamos LS, o limiar tal que qualquer observação à sua direita declarar-se-á "atípica". Temos: $LS = Q3 + 1,5(Q3 - Q1) = 42,525$. E verificamos em seguida que a observação 2 de valor 61,1 é a única que é maior que LS; ela será marcada na desenho por $\circ$, assinalando com isto a "atipicidade" dela (há pacotes computacionais de Estatística que produzem desenhos nos quais os valores atípicos são marcados por $*$). É importante observar que o bigode superior do box-plot não chega obrigatoriamente à altura LS; o bigode acaba na maior observação do conjunto entre as que não ultrapassam LS (reforço: as que ultrapassam são "atípicas" e são marcadas como tal). De acordo com esta regra, bigode superior estica-se do $Q3 = 22,95$ a $27,8$.

Vamos agora ao bigode inferior e as valores atípicos pequenos da amostra tartada. Assim como acima, calcula-se preimeiramente o limiar tal que os valores abaixo dele serão considerados como atíticos: $LI = Q1 - 1,5(Q3 - Q1) = -9,675$. No caso da amostra tratada, não há valores inforeir a LI, logo não há valores atípicos pequenos. É importante observar que o bigode inferior do box-plot não desce obrigatoriamente ao nível LI; o bigode acaba na menor observação do conjunto entre as que não caem abaixo de LI. Esta observação é $6,3$.

Dessa forma, a Figura ??apresenta o box-plot dos dados originais.

(b) Da mesma forma como calculado no item A, as medidas para o conjunto de dados sem a observação 2 se encontram na tabela abaixo:

	Valor
Média	16,0071
Mediana	14,25
Desvio Padrão	7,465
Primeiro quartil	9,3
Terceiro quartil	22,575

Pode-se verificar que, ao retirar a segunda observação, os 3 quartis não obtiveram seu valor muito alterado; porém, a média e o desvio padrão reduziram significantemente, sendo que o desvio padrão reduziu quase que pela metade.

Solução ao Exc. 16. A Tabela de Esatatísticas descritivas encontra-se abaixo. Esta foi calculada com uso do pacote computacional R; é possível que R calcula $Q1, Q2, Q3$ sa maneira pouco diferente daquela que foi explicada na aula e realizada na solução do Exc. 12.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.80	22.10	23.00	23.38	24.20	28.60

Solução (a) Obs.: o texto foi elborado pelo monitor. O box-plot mostra que os dados estão concentrados na "extremidade direita": os 50% das maiores observações são mais concentrados que os 50% dos menores. Ainda mais: já que as amplitudes dos 3-o e 4-o subconjuntos são relativamente iguais, podemos acreditar que as observações à direita da mediana ($Q2$) são distribuidas mias ou menos uniforme. Podemos também acreditar que a causa esquerda da

conjunto de dados é muito fina, quer dizer, há poucas observações espalhadas num intervalo relativamente extenso; a indicação para isto é a extenção do bigode esquerda do box-plot.

Solução (b) Obs.: o texto foi elborado pelo monitor. Neste caso, o box-plot indica que as menores e as maiores observações são as mais concentradas. Como as aplitudes do primeiro e do querto subconjuntos são quase iguais, podemos sugerir que a distribuição possui duas modas: uma delas fica entre os 25% das menores observações, e a outra entre os 25% das maiores. Já os 50% das observações centrais são mais dispersas. O box-plot não permite deduzir se estas são distribuidas uniformamente entre $Q1$ e $Q3$ ou não. O histograma mostra que não: há uma significativa diminuição da densidade logo à direita da médiana. O box-plot foi incapáz de sentir isto.

Solução (c) Obs.: o texto foi elborado pelo monitor. O formato do box-plot indica que há ima única moda da distribuição e que essa econtra-se entre $Q1$ e a mediana. A densidade das observações diminui conforme das observações afastam-se da moda, sendo que a cauda direita é mais longa que a cauda esquerda.

Obs. do professor: Observaria que o conjunto de dados foi criado por nós da maneira tal que sua histograma apresentasse duas cumes: uma – a mais alta – com a coordenada 0.15 e a outra – mais baixinha – com a coordenada 0.55. Podemos agora ver que o box-plot não sentiu esta particularidada de distribuição.

Solução (d) Obs.: o texto foi elborado pelo monitor. Neste conjunto de dados, as observações encontram-se concentrados na extremidade superior, e a concentração diminui com a diminuição de seus valores. É provável – conforme o box-plot segere – que a moda esteja à direita de $Q3$.

Solução (e) Obs.: o texto foi elborado pelo monitor. O box-plot deste conjunto de dados sugere que a concentração de observações diminui com o aumento de seus valores, que a moda fica perto de 0 e que a distribuição não possui cauda esquerda. A igualdade (aproximada) das extenções do terceiro e quarto subconjuntos pode indicar a distribuição uniforme das observações à direita da mediana. Porem, os subconjuntos são muito extensos, e isto permite que a densidade das observações flutue dentro de cada um sem que o box-plot consegue sentir estas flutuações. Na verdade, é bem isto que aconteceu: há uma concentração relativa de observações em torno de 0.6 (expressada no histograma por uma elevação acima de 0.6), cuja existência não foi indicada pelo box-plot.

6.12 Porque a fórmula para a variância amostral tem $n-1$ no seu denominador

6.12.1 Formulando a pergunta

Suponhamos que alguém colocou bolas idênticas numa urna, sendo que cada bola carrega um dos números $k_1, k_2, \ldots, k_M$. Denotaremos por p_i a proporção das bolas da urna que carregam o número k_i. Com isso, a seguinte tabela

k_1	k_2	$\cdots\cdots$	k_{M-1}	k_M
p_1	p_2	$\cdots\cdots$	p_{M-1}	p_M

é exatamente aquilo ao que chamamos por "distribuição de frequencia relativa por atributo "numero carregado" da população de bolas na urna". Esse distribuição será chamada no que se segue por "distribuição populacional".

Recordo, para as necessidades dos argumentos a vir, que a quantia

$$k_1p_1 + k_2p_2 + \cdots + k_{M-1}p_{M-1} + k_Mp_M \tag{6.27}$$

calculada com valores da distribuição populacional, chama-se *média populacional*; denotaremos essa por μ.

Outrossim recordo, também para as necessidades futuras, que a quantia

$$(k_1-\mu)^2p_1 + (k_2-\mu)^2p_2 + \cdots + (k_{M-1}-\mu)^2p_{M-1} + (k_M-\mu)^2p_M \tag{6.28}$$

chama-se *variância populacional*; denotaremos essa por σ^2.

Imagine agora que não sabemos nada sobre a distribuição populacional: não sabemos nem M, nenhum dos k_i e nenhum dos p_i.

Imagine que desejamos estimar o valor de σ^2, quer dizer, desejamos estimar a variânica populacional, e para o fim dessa estimação, somos autorizados fazer amostra simples com reposição. Isso quer dizer, que podemos retirar ao acaso uma bola da urna e observar o número carregada pela bola retirada, e podemos repetir esse procedimento tantas vezes quantas desejamos, desde que após cada ratirada, a bola seja devolvida à urna, e todas as bolas sejam bem misturadas.

Então, ao fixar o número de retiradas a serem feitas e ao denotar esse por n (esse n é o que chama-se *o tamanho da amostra* a ser feita), sugerimos o seguinte procedimento: após fazer n retiradas e ver os números

$$x_1, x_2, \ldots, x_n \tag{6.29}$$

fazer a média delas, isso é,

$$\frac{1}{n}\{x_1 + x_2 + \cdots + x_n\} \tag{6.30}$$

e usar esse para calcular o valor da expressão

$$\frac{1}{n-1}\left\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right\} \tag{6.31}$$

onde $\bar{x}$ é a notação para a média. O valor da expressão (1.31) é a nossa estimativa para a variância populacional. Na verdade, ela é nossa pois aceitamos a imposição de Estatistica Teórica que ela é boa por um motivo sensato e cómodo. De acordo com meu plano didático, não preciso discutir esse motivo, embora na Seção 1.12.2, em sua Equação (1.35), aparecerá a propriedade que dá lúz ao motivo, e esse será comentado em (1.12). Mas para você, mau leitor, isso tudo deve estar no segundo plano. O que você deve perguntar é: *Por que o denominador da expressão (1.31) é $n-1$ e não n, como seria de se esperar com base em considerações intuitivas e genéricas?* O presente texto é sobre isso e aresposta virá na prossima seção.

6.12.2 Respondendo a pergunta

Para que possamos dar a a explicação ao $n-1$, precisamos olhar na amostra antes dela acontecer. Desse ponto de vista,

$$\begin{array}{ccccccc}
\text{o número a ser visto na 1-a retirada é uma variável aleatória, a qual denotamos por } X_1, & & & & & & \\
\text{o número a ser visto na 2-a retirada é uma variável aleatória, a qual denotamos por } X_2, & & & & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{o número a ser visto na } n\text{-ésima retirada é uma variável aleatória, a qual denotamos por } X_n & & & & & &
\end{array}$$

E então, a média (1.30) adquira a expressão

$$\frac{1}{n}\{X_1 + X_2 + \cdots + X_n\} \tag{6.32}$$

Fica claro que isso é uma variávela aleatória; denotamos ela por $\bar{X}$. Ainda mais, do nosso ponto de vista de antes, a expressão (1.31) adquira a expressão

$$\frac{1}{n-1}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\} \tag{6.33}$$

Para justificar que a fórmula sugerida para estimar σ^2 é boa, vou mostrar que

$$\mathbb{E}\left[\frac{1}{n-1}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\}\right] = \sigma^2 \tag{6.34}$$

o que interpreta-se dizendo que

$$\text{o estimador acerta em média no valor estimado} \tag{6.35}$$

que, de acordo com considerações cujas elucidações não está na nossa programação, significa que (1.33) é um bom estimador.

Para completarmos o objetivo de mostrar a igualdade (1.34), precisamos dos seguintes fatos:

Fato 1: A distribuição de cada variável aleatória X_i é exatamente a distribuição populacional, quer dizer,

x	k_1	k_2	$\cdots\cdots$	k_{M-1}	k_M
$\mathbb{P}[X_i = x]$	p_1	p_2	$\cdots\cdots$	p_{M-1}	p_M

e, como consequencia disso,

Fato 2:

$$\mathbb{E}[X_i] = \mu, \qquad \text{Var}[X_i] = \sigma^2, \text{ para cada } X_i \tag{6.36}$$

Fato 3: As variávei aleatórias $X_1, X_2, \ldots, X_n$ são independentes em conjunto.

Confere que você entende bem os Fatos 1-3: O primeiro deles segue-se da definição de experimento aleatório, da construção de modelo probabilístico e da definição de variável aleatória como a expressão do resultado a vir num experimento aleatório. É importante entender que a suposição, segundo a qual não conhecemos nada sobre a distribuição populacional, não impede da mesma ser a distribuição de cada variável aleatória que definimos; é só não conhecemos a distribuição. O comentário do Fato 2 vem no mesmo sentido: apesar de não conhecer os valores da média populacional e da variância populacional, sabemos que esses valores coincidem com,

respectivamente, a esperança e a variância de cada uma das variáveis aleatórias que definimos. O Fato 3 segue-se de um comentário que fiz quando expliquei o conceito de independencia entre variáveis aleatórias. É suficiente que você acredite na independência. Creio que isso não é difícil a ser concebido pois as bolas retiradas são devolvidas na urna, e, portanto, o resultado de i-ésima retirada não pode dependenr dos resultados das retiradas anteriores.

Vamos agora introduzir notações que ajudarão a simplificar a escrita da demostração que pretendemos fazer. Em primeiro lugar, observe que

$$X_i - \bar{X} = (X_i - \mu) - (\bar{X} - \mu) = (X_i - \mu) - \frac{(X_1 - \mu) + (X_2 - \mu) + \ldots + (X_n - \mu)}{n}$$

Por isso, ao introduzir

$$Y_i = X_i - \mu, \text{ para cada } i$$

podemos reescrever nosso objetivo da seguinte maneira:

$$\text{provar que } \mathbb{E}\left[\frac{1}{n-1}\left\{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2\right\}\right] = \sigma^2 \tag{6.37}$$

onde

$$Y_1, Y_2, \ldots, Y_n \text{ são variáveis aleatórias independentes em conjunto,} \tag{6.38}$$

$$\mathbb{E}[Y_i] = 0 \text{ e } \mathrm{Var}[Y_i] = \sigma^2 \text{ para cada } i, \tag{6.39}$$

$$\bar{Y} \text{ é a notação para } \frac{1}{n}\left\{Y_1 + Y_2 + \cdots + Y_n\right\} \tag{6.40}$$

sendo que (1.38) é a consequencia do Fato 3, e (1.39) é a consequencia do Fato 2; óbvio, que poderiamos desenhar a distribuição de cada Y_i a partir da distribuição das X's, mas não fizeram isso pois estas não serão importantes para as contas a seguir.

O segundo passo na nossa presente derivação de fatos e notações auxiliares é o fato de que

$$\mathbb{E}[Y_i^2] = \sigma^2 \tag{6.41}$$

Esse segue-se da expanção $\mathrm{Var}[Y_i] = \mathbb{E}[Y_i^2] - (\mathbb{E}[Y_i])^2$ junto com as duas relações da Eq. (1.39). Já no terceiro passo deduzimos que

$$\mathbb{E}[Y_i Y_j] = \mathbb{E}[Y_i]\,\mathbb{E}[Y_j] = 0 \times 0 = 0, \text{ para todos } i \text{ e } j \text{ diferentes entre si} \tag{6.42}$$

Observe que a primeira igualdade da sequencia (1.42) sustenta-se pela independência entre Y_i e Y_j, a qual, por sua vez, segue-se da independência das Y's em conjunto. Essa conta é o único lugar onde usamos a independência, mas o resultado da conta é a chave principal para tudo, de modo que a ausência da independência quebraria a conta e, em sequencia, toda a demonstração.

Agora, juntamos (1.41) e (1.42) para derivar que

$$\begin{aligned}
&\mathbb{E}\left[(Y_1 + Y_2 + \ldots + Y_n)^2\right] = \mathbb{E}\left[Y_1^2 + Y_2^2 + \ldots + Y_n^2 + 2Y_1Y_2 + \ldots + 2Y_{n-1}Y_n\right] = \\
&= \left(\mathbb{E}[Y_1^2] + \ldots + \mathbb{E}[Y_n^2]\right) + \left(2\mathbb{E}[Y_1Y_2] + \ldots + \mathbb{E}[Y_{n-1}Y_n]\right) = \\
&= (\sigma^2 + \ldots + \sigma^2) + (0 + \ldots + 0) = n\sigma^2
\end{aligned}$$

Esse resultado será usado para fechar a demonstração em conjunto com o resultado da seguinte conta, puramente algébrica:

$$\begin{aligned}
&\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \\
&= \left(Y_1^2 - 2Y_1 \times \bar{Y} + \left(\bar{Y}\right)^2\right) + \ldots + \left(Y_n^2 - 2Y_n \times \bar{Y} + \left(\bar{Y}\right)^2\right) = \\
&= \left(Y_1^2 + \ldots + Y_n^2\right) - 2n\bar{Y} \times \bar{Y} + n\left(\bar{Y}\right)^2 = \\
&= \left(Y_1^2 + \ldots + Y_n^2\right) - n\left(\bar{Y}\right)^2 = \left(Y_1^2 + \ldots + Y_n^2\right) - \frac{1}{n}\left(Y_1 + \ldots + Y_n\right)^2
\end{aligned}$$

Portanto,

$$\mathbb{E}\left[\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2\right]=\mathbb{E}\left[Y_1^2+\ldots+Y_n^2\right]-\frac{1}{n}\mathbb{E}\left[(Y_1+\ldots+Y_n)^2\right]=$$
$$=\mathbb{E}\left[Y_1^2\right]+\ldots+\mathbb{E}\left[Y_n^2\right]-\frac{1}{n}\left(n\sigma^2\right)=$$
$$=n\sigma^2-\sigma^2=(n-1)\sigma^2$$

Agora ficou claro que (já passando de volta para as variáveis aleatórias X's)

$$\mathbb{E}\left[\frac{1}{n-1}\left\{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots+(X_n-\bar{X})^2\right\}\right]=\sigma^2 \tag{6.43}$$

enquanto que, se tomassemos n e vez de $n-1$ no denominador, teriamos

$$\mathbb{E}\left[\frac{1}{n}\left\{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots+(X_n-\bar{X})^2\right\}\right]=\frac{n-1}{n}\sigma^2 \tag{6.44}$$