DISTRIBUIÇÃO POPULACIONAL SUAS CARATERÍSTICAS E MANEIRAS DE VIZUALISAÇÃO/APRESENTAÇÃO

Ministrante Prof. Dr. Vladimir Belitsky, IME-USP

Sobre o conceito "população".

A definição do termo **população**, tentei formular....
.....mas não consegui.

Para que podermos continuar, só nos resta a torcer que sua intuição entenda este termo da mesma maneira que a minha. Para testar a unissonância das intuições, vejamos exemplos:

(1) a população de todos os moradores no prédio onde eu moro;

(2) a população de todos que passaram pela cirugia bariátrica nos últimos 5 anos e não morreram até o presente momento.

Os exemplos sugerem que a **população** poderia ser definida como **todos os** ou **todos que**, mas creio que o formalismo da definição é dispensável pois os exemplos acima já bastam.

Sobre o conceito “atributo”.

A definição do termo **atributo**, tentei formular....
.....mas não consegui.

Para que podermos continuar, só nos resta a torcer que sua intuição entenda este termo da mesma maneira que a minha. Para testar a unissonância das intuições, vejamos exemplos:

(a) salário mensal;

(b) alteração de peso do momento de antes da cirugia até o presente momento.

Continuo sem saber como definir **atributo** mas sinto que isto é totalmente disnecessário.

Um exemplo, onde meu aluno encontra exemplo de “população”, exemplo de “atributo”, e também aprende três novos conceitos. (1/4)

O exemplo a seguir é do livro
Ron Larson, Betsy Farber, “Elementary Statistics”, Prentice Hall, Inc. (2003)

Um exemplo, onde meu aluno encontra exemplo de "população", exemplo de "atributo", e também aprende três novos conceitos. (2/4)

Num remoto vilarejo da Alaska, chamado Akhiok,....

Google Maps Old BIA homes

Captura da imagem: out 2015 As imagens podem ter direitos autorais. Panoramio

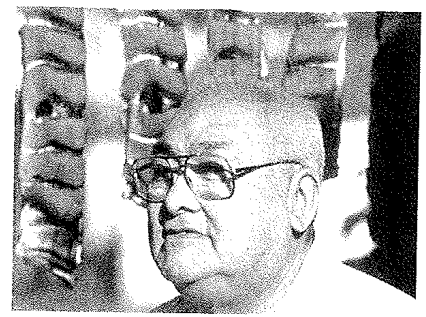
Jessie Lynn Huff

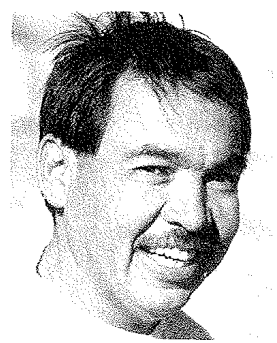
Foto - out 2015

Um exemplo, onde meu aluno encontra exemplo de "população", exemplo de "atributo", e também aprende três novos conceitos. (3/4)

... há exatamente 77 habitantes; conheça alguns deles:








Um exemplo, onde meu aluno encontra exemplo de "população", exemplo de "atributo", e também aprende três novos conceitos. (4/4)

As idades de todos os 77 são (em ordem 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

Cada um dos números acima pode (e de fato, vai) ser chamado **observação** (pois observamos idade de individuo), ou **dado** (pois é dado, quer dizer, caraterística de individuo).

E todos os números podem (e de fato, serão) chamados por **conjunto de dados** ou **conjunto de observações**.

Como o grosso da apresentação é sobre um atributo quantitativo, e bom adiantar a apresentação com a menção sobre outros tipos de atributo. (1/5)

Vale notar que no Exemplo apresentado, o atributo "idade" está sendo medido por números. Neste caso, o atributo chama-se **quantitativo**, embora o termo mais tradicional e muito mais usado é **variável quantitativa**. É claro que não tem nada a ver com "variável aleatória".

Ainda mais, as variáveis quantitativas separam-se em **discretas** e **continuas**. Tipicamente, as discretas são aqueles que dão-se por contagem (por exemplo, o número de aparelhos de televisão em domicílio), e as continuas são aqueles em cuja medição pode surgir qualquer número de um intervalo de números reais.

Como o grosso da apresentação é sobre um atributo quantitativo, e bom adiantar a apresentação com a menção sobre outros tipos de atributo. (2/5)

A definição de discreta/contínua, dada na transparência anterior, tem que ser aceita e interpretada com certa cautela. Por exemplo, dizemos que a altura de individuo é variável quantitativa contínua, pois imaginamos que esta pode ser qualquer número entre 0,1m e 3m. Entretanto, nossa medição dá tipicamente valores com duas casas após a vírgula (tipo, 1,76m). Portanto, se anotassemos a altura em centímetros, os resultados de medição para qualquer população pareceria como se fosse que medimos variável quantitativa discreta.

Como o grosso da apresentação é sobre um atributo quantitativo, e bom adiantar a apresentação com a menção sobre outros tipos de atributo. (3/5)

Além de variáveis quantitativas existem **variáveis qualitativas** ou **categóricas**.
Dois exemplos mais simples e esclarecedores de variáveis **qualitativas** são: (1) sexo e (2) escolaridade. Observe que se desejarmos, a escolaridade pode ser ordenada; ela é um exemplo de variável qualitativa **ordenal**. O sexo não pode ser ordenado, por isto esta variável qualitativa chama-se **nominal**. Na hora de guardar dados num computador, você pode representar mulhers por 1 e homens por 2, mas isto não significa que você ordenou o sexo.

Como o grosso da apresentação é sobre um atributo quantitativo, e bom adiantar a apresentação com a menção sobre outros tipos de atributo. (4/5)

Abaixo, estão alguns exemplos que esclarecem por completo a classificação de atributos/variáveis por qualitativa nominal, qualitativa ordinal, quantitativa discreta e quantitativa contínua:

Vitamina (A, B1, B2, B6, B12)$\rightarrow$ Qualitativa nominal
Quantidade de caloria na batata frita$\rightarrow$ Quantitativa contínua
Desfecho de uma doença (curado ou não)$\rightarrow$ Qualitativa nominal
Classificação de lesão (fatal, severa, moderada, pequena) $\rightarrow$ Qualitativa ordinal
Grupo sangüíneo (A,B,AB,O) $\rightarrow$ Qualitativa nominal
Paridade (primeira gestação, segunda gestação, terceira ...) $\rightarrow$ Quantitativa discreta

Como o grosso da apresentação é sobre um atributo quantitativo, e bom adiantar a apresentação com a menção sobre outros tipos de atributo. (5/5)

Continuamos com exemplos:

Estado geral de um paciente (bom, regular, ruim)$\rightarrow$ Qualitativa ordinal

Número de nascidos em certo hospital no certo mês/ano $\rightarrow$ Quantitativa discreta

Idade $\rightarrow$ Quantitativa discreta

Concentração de flúor na água $\rightarrow$ Quantitativa contínua

Atividade esportiva preferida $\rightarrow$ Qualitativa nominal

O conceito "ordenação de dados" (via Exemplo).

Abaixo, há o conjunto de dados de Exemplo está em sua forma **ordenada**. A ordenação facilita a apresentação a seguir (na realidade, a ordenação é um tributo à tradição de apresentar dados numéricos em ordem crescente ou descrescente).

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

Apresentação de dados por tabela de frequencias absolutas (exposição com uso do Exemplo). (1/2)

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

Apresentação de dados por tabela de frequencias absolutas (exposição com uso do Exemplo). (2/2)

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.

Apresentação de dados por tabela de frequencias absolutas (exposição com uso do Exemplo).

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
$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{2}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$	$\frac{1}{77}$

O conceito "tamanho da população" e um comentário sobre a relação entre a distribuição de frequencia absoluta e a de frequencia relativa.

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.

Um comentário sobre a termo “distribuição de frequencia”. (1/2)

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 raramente vi algo do tipo na literatura.

Um comentário sobre a termo “distribuição de frequencia”. (2/2)

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.

Viva com as duas, mas use aquela que é oficial quando for escrever um documento oficial.

O tema da aula.

A presente aula é sobre a **distribuição de frequência pelo atributo observado em todos os membros de uma população** (chamada também **distribuição populacional de frequência**), sobre algumas caraterísticas de tal distribuição e sobre algumas formas de sua apresentação.

Exemplos do termo em negrito acima:

(i) a distribuição de frequência pelos valores do atributo "salário mensal" observado na população de moradores do prédio onde moro;

(ii) a distribuição de frequência pelos valores do atributo "alteração de peso" observado na população das pessoas que passaram pela cirugia bariátrica nos últimos 5 anos e ainda não morreram.

(iii)) a distribuição de frequência pelos valores do atributo "idade" observado na população de moradores de um vilarejo (visto no Exemplo).

Apesentação por gráficos.

Ao receber um conjunto de dados (observações) e sendo perguntado

(i) o que o conjunto tem de interessante,

(ii) de importante,

(iii) se tem um propriedade específica,

(iv) ou se posso sugerir uma estratégia com base nos dados, que seja útil, agradável, lucrativa, etc.,

faço apresentação gráfica para a distribuição de frequencia do conjunto e torço que análise visual do gráfico segira uma resposta ou o caminho que possa levar à resposta.

Apesentação por gráficos.

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.

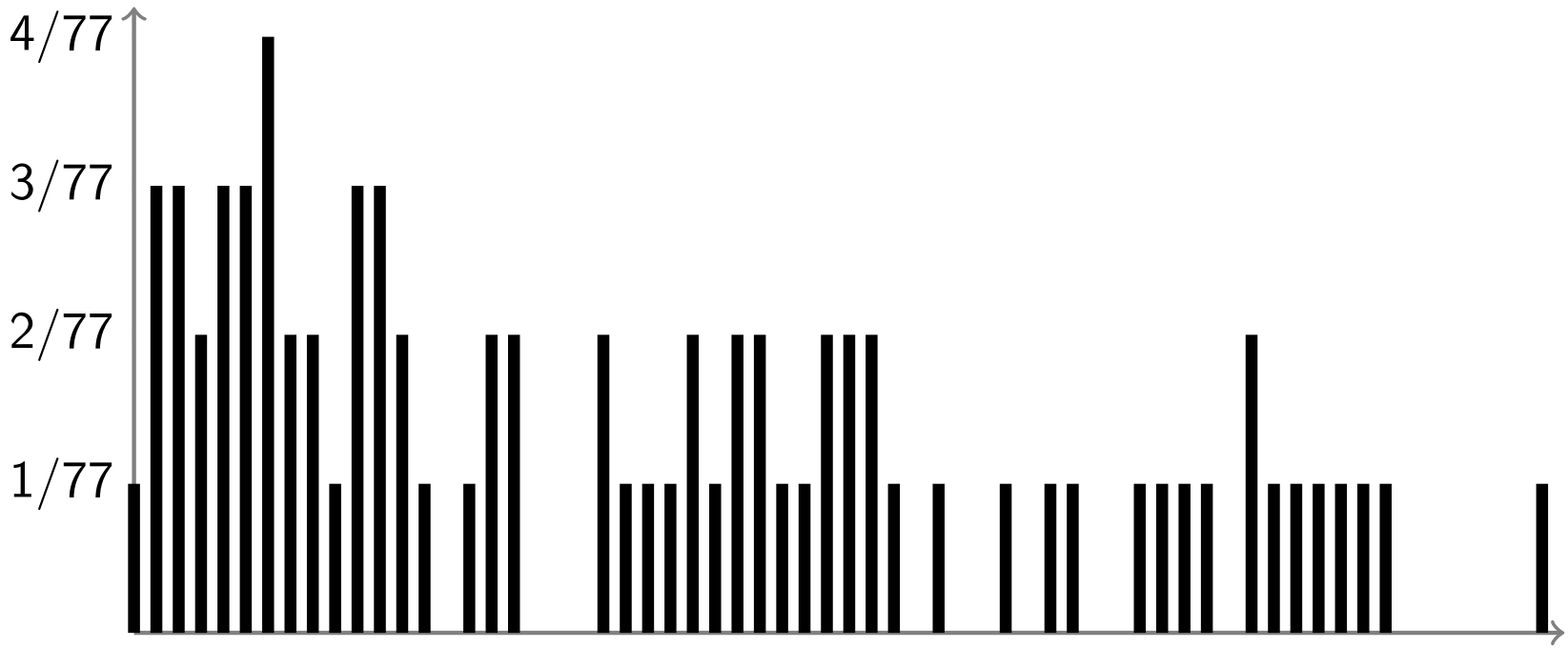


Figura: Apresentação da distribuição da freq. relativa por atributo "idade" em forma de Função de Probabilidade.

Apesentação por gráficos.

As vezes, o agrupamento de valores observados do atributo facilita a revelação de propriedades do conjunto de dados. O agrupamento pode ser feito por acumulo de frequencia (isso vai levar aos gráficos do tipo de **Box-Plot** a ser discutido posteriomente), ou por agrupamento dos valores do atributo. No segundo caso, o resultado é o gráfico chamado **histograma**.

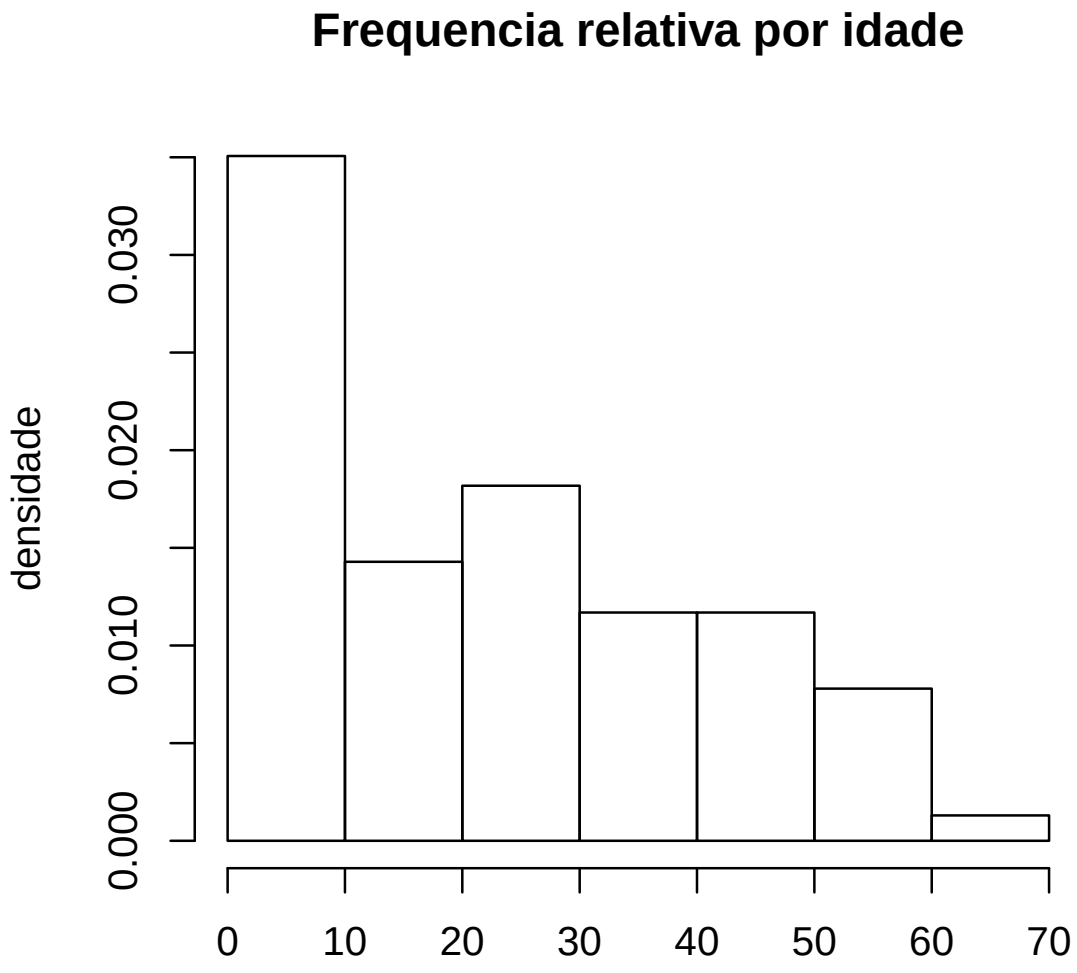
Vou falar dos histogramas por frequencia relativa. Aqueles, que são por frequencia absoluta, usam-se muito menos, e você conseguira entedé-los se enteder direitinho so que serão apresentados por mim.

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".

Uma das consequencias dessa maneira de apresentação é que o eixo vertical não possui interpretação imediata. Nos desenhos a seguir esse eixo é chamado "densidade". Esse nome pode ser explicado, mas no momento, tal explicação é disnecessária e pouco esclarecidora. Desconsidere o nome "densidade" se esse aparecer.

Histograma.

Abaixo, 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 na transparência seguinte.



Histograma.

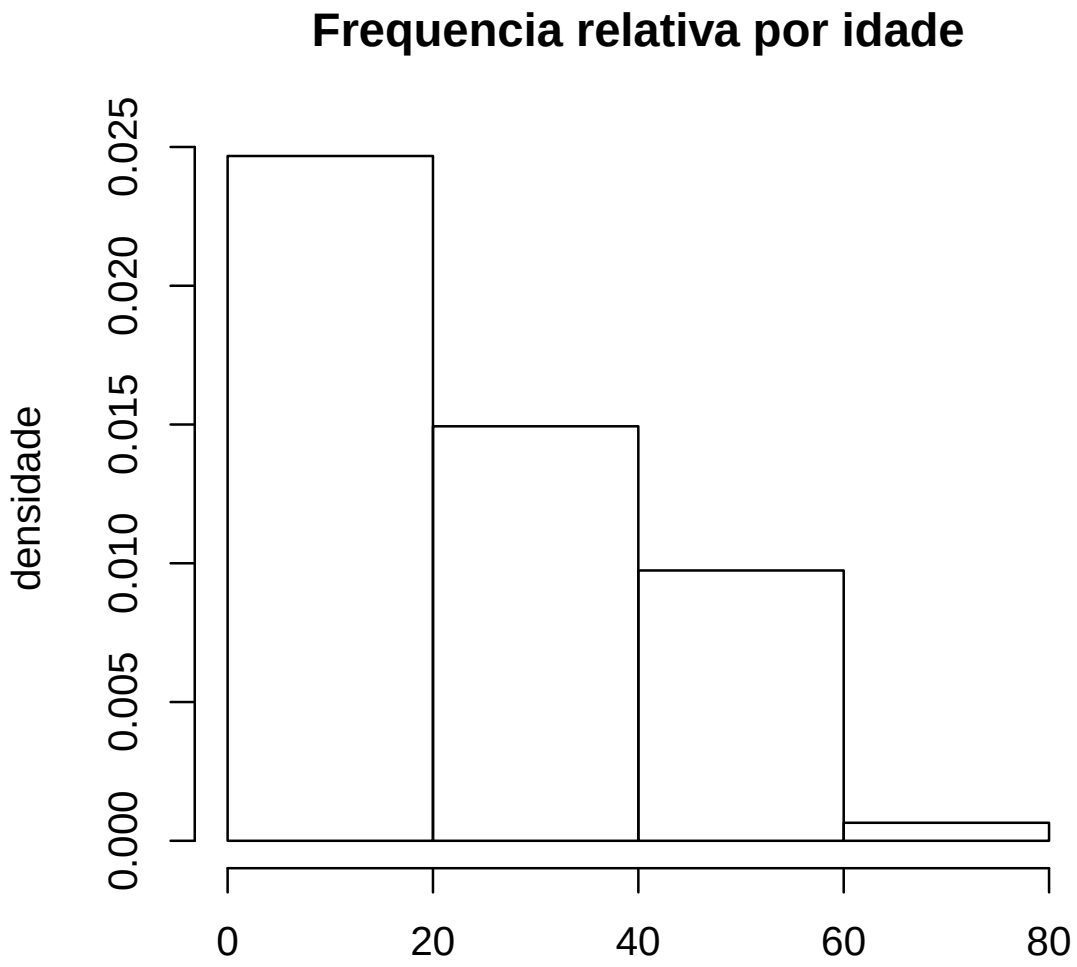
Separadores dos classes: 0, 10, 20, 30, 40, 50, 60, 70. Classes (observe "(" e "]"):
$[0, 10]$, $(10, 20]$, $(20, 30]$, $(30, 40]$, $(40, 50]$, $(50, 60]$, $(60, 70]$
Frequencias absolutas (quantidade de dados em cada classe) chamadas de **counts** em Inglês: $27, 11, 14, 9, 9, 6, 1$
Frequencias relativas: $27/77 = 0.35064935$, $11/77 = 0.14285714$, $14/77 = 0.18181818$, $9/77 = 0.11688312$, $9/77 = 0.11688312$, $6/77 = 0.07792208$, $1/77 = 0.01298701$.
As frequencias relativas são representadas por áreas dos "prédios" correspondentes. As alturas dos prédios do histograma calcula-se pela regra:

$$\text{freq.relativa} = \text{área do prédio} = \text{altura} \times \text{base}$$

Obs: "base" chama-se também por **amplitude da classe**.
$\frac{0.35064935}{10} = 0.035064935$, $\frac{0.14285714}{10} = 0.014285714$,
$\frac{0.18181818}{10} = 0.018181818$, $\frac{0.11688312}{10} = 0.011688312$,
$\frac{0.11688312}{10} = 0.011688312$, $\frac{0.07792208}{10} = 0.007792208$,
$\frac{0.01298701}{10} = 0.001298701$.

Histograma.

Agora veja outro histograma feito para o mesmo conjunto de dados. As classes desse são diferentes das do histograma anterior. Os cálculos auxiliares que determinam seu formato estão na transparência seguinte.



Histograma.

Os cálculos usados para a construção do histograma da transparência anterior:

Separadores dos classes: 0 20 40 60 80. Classes: $[0, 20]$, $(20, 40]$, $(40, 60]$, $(60, 80]$.

Frequencias absolutas (quantidade de dados em cada classe): 38, 23, 15, 1.

Freq. relativas: $\frac{38}{77} = 0.4935065, \cdots$

Alturas dos prédios: $\frac{0.4935065}{20} = 0.0246753247$, 0.0149350649, 0.0097402597, 0.0006493506.

Histograma.

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

Histograma.

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.

Essa foi uma das razões para construção de histograma por amplitudes desiguais. Existe infinitude de outras razões. 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:

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

Histograma.

Então, os separadores são:

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

As frequencias absolutas das classes correspondentes são:

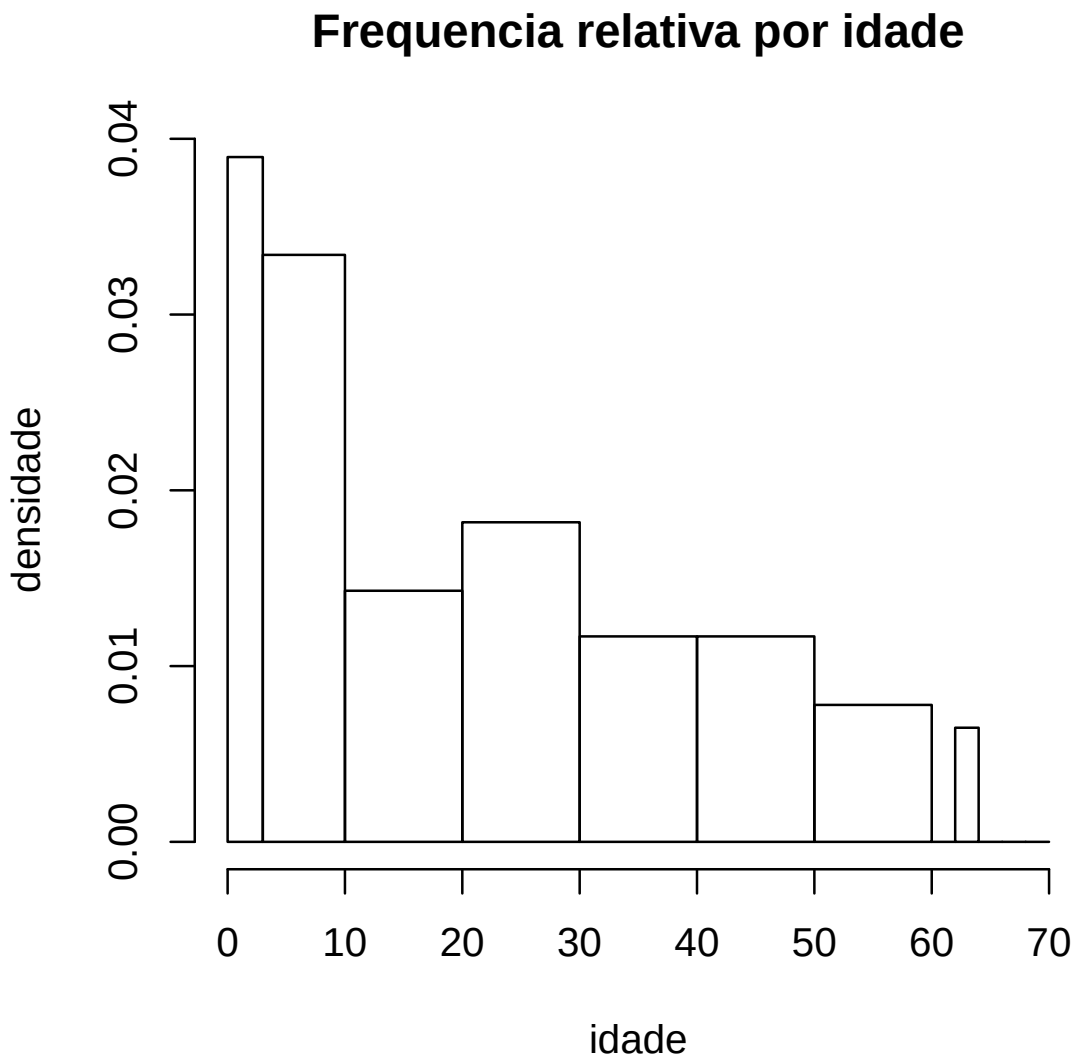
$$9,\ 18,\ 11,\ 14,\ 9,\ 9,\ 6,\ 0,\ 1,\ 0,\ 0,\ 0$$

Observe: a área do primeiro prédio é a frequencia relativa dos pessoas na primeira classe: $9/77 = 0.1168831$, mas a altura deste prédio é tal que, sendo multiplicada pela base, dá esta área, quer dizer, a altura é $9/(3 \times 77) = 0.038961039$. Eis a lista das alturas:

$$0.038961039, 0.033395176, 0.014285714, 0.018181818,$$
$$0.011688312, 0.011688312, 0.007792208, 0.000000000,$$
$$0.006493506, 0.000000000, 0.000000000, 0.000000000$$

Histograma por amplitudes desiguais.

Eis o histograma resultante.



Histogramas; duas observações.

(1) Desconhe´ço 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:

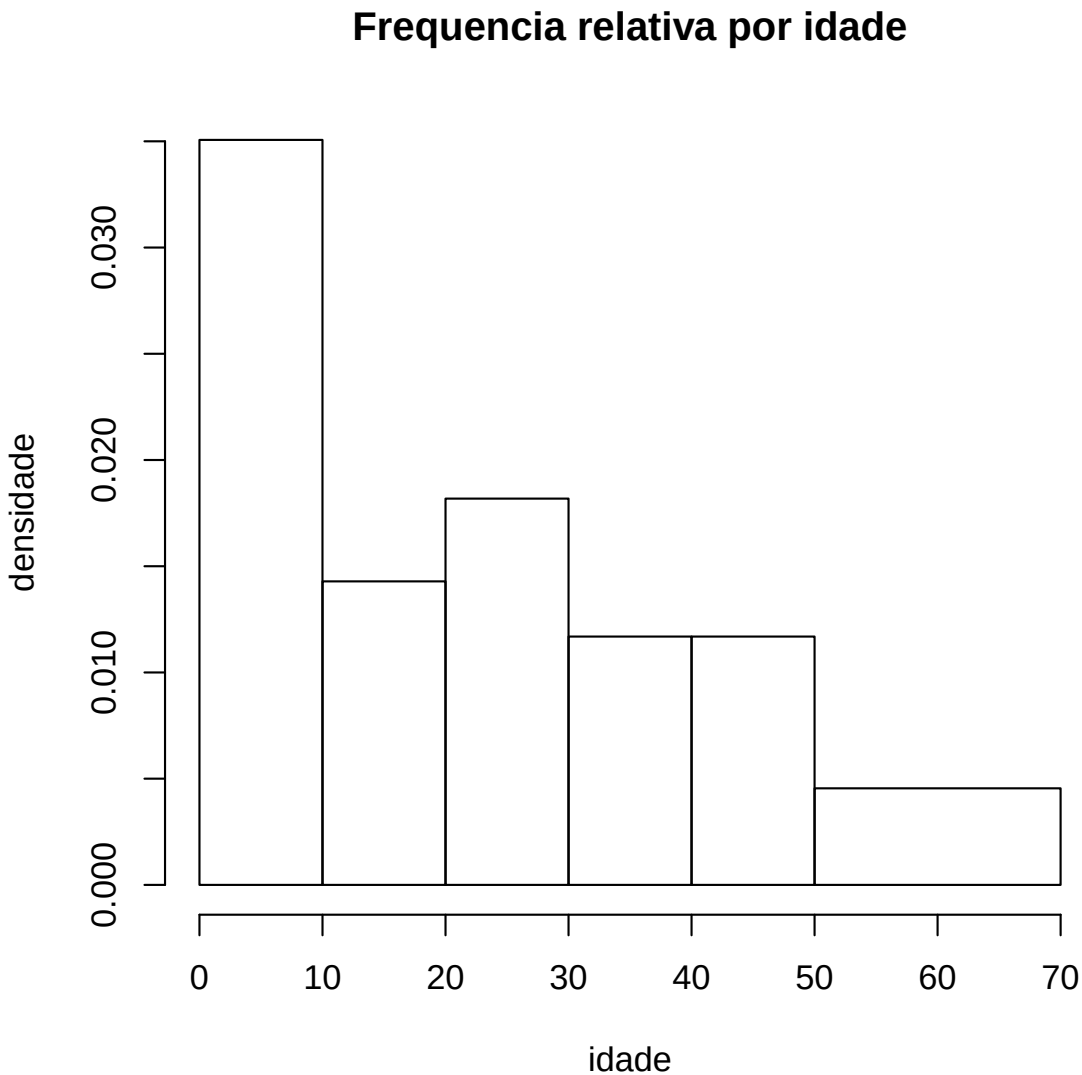
$$0, 10, 20, 30, 40, 50, 70$$

Isto dá o histograma apresentada na transparência a seguir.

(2) 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.

Histograma por amplitudes desiguais.

O histograma correspondente aos separadores $0, 10, 20, 30, 40, 50, 70$ cuja escolha foi motivada na transparência anterior.



Histogramas; mais duas observações.

(3) 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.

(4) 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.

Média e variânica populacionais.

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.

Média e variânica populacionais.

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.

Média e variânica populacionais.

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

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

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

Média e variânica populacionais.

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.

Média e variânica populacionais.

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)$\times$(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)$\times$(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).

Média e variânica populacionais.

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 variãncia 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.

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.

A notação para observações ordenadas.

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

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.

Quantis. Discussão preliminar.

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

Quantis. Discussão preliminar.

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).

Quantis. Discussão preliminar.

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.

Quantis. Definição para conjunto populacional.

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 $x_{(M)}$.

Quantis.

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

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

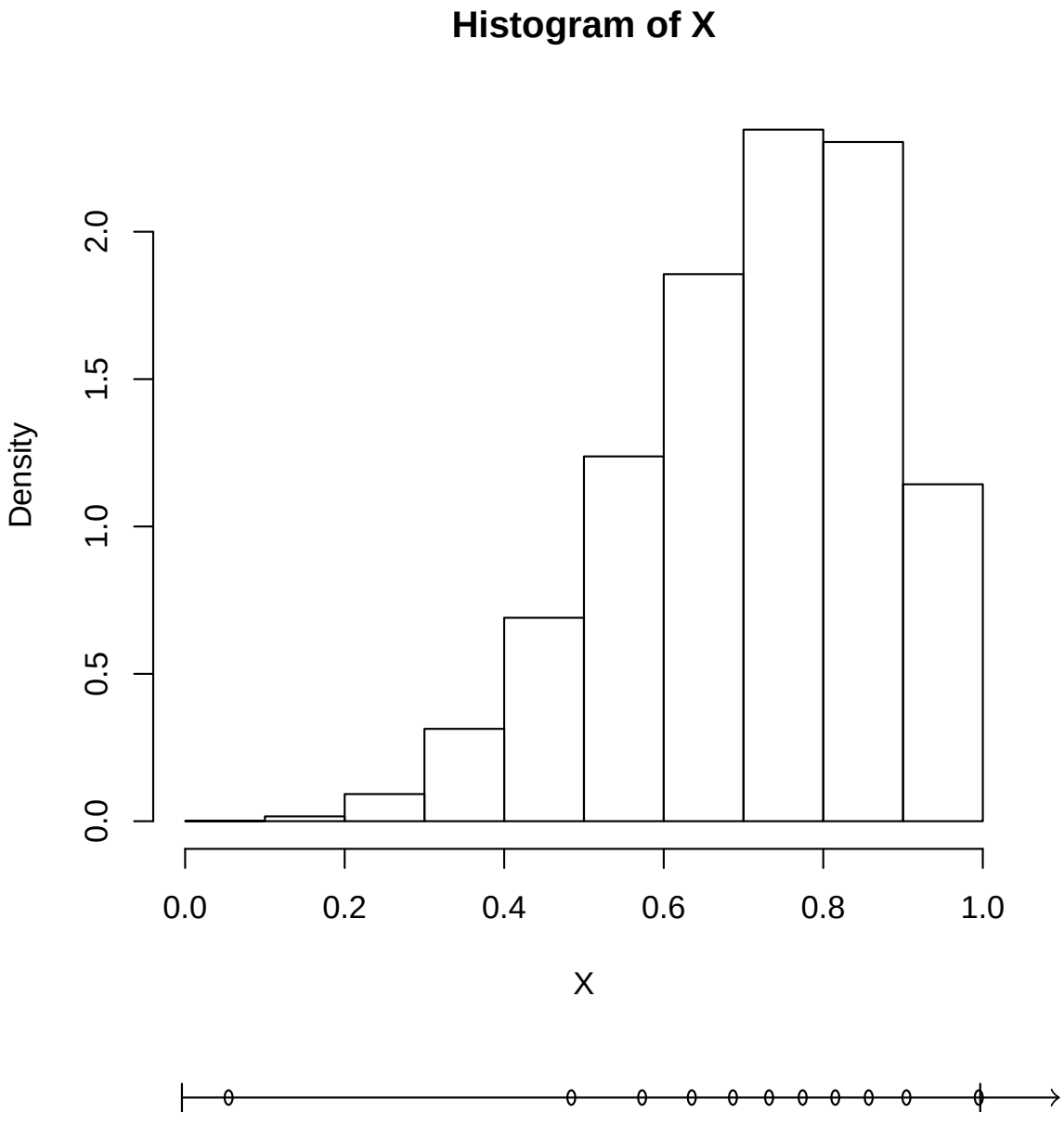
e como não há $x_{(0)}$, então a definição não generaliza-se para $k = 0$. É cómodo definir que 0-ésimo q-quantil seja $x_{(1)}$, a observação mínima do conjunto.

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 $\max = x_{(N)}$.

Quantis.

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.

Quantis.



Quantis.

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.

Quartis.

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.

Quartis.

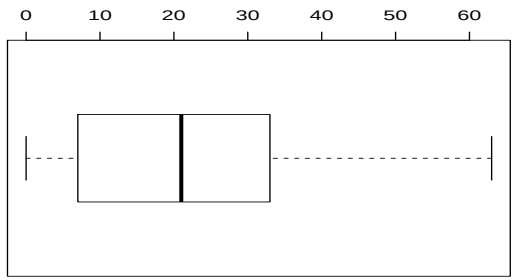
Os nomes são autoexplicativos: os $Q1$, $Q2$ e $Q3$ dividem um conjunto ordenado de dados em quatro partes (quase) iguais, sendo que a igualdade entende-se aqui no sentido da quantidade de dados; em cada parte há (quase) 25% de todos os dados.

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.

Quartis. BoxPlot.

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:



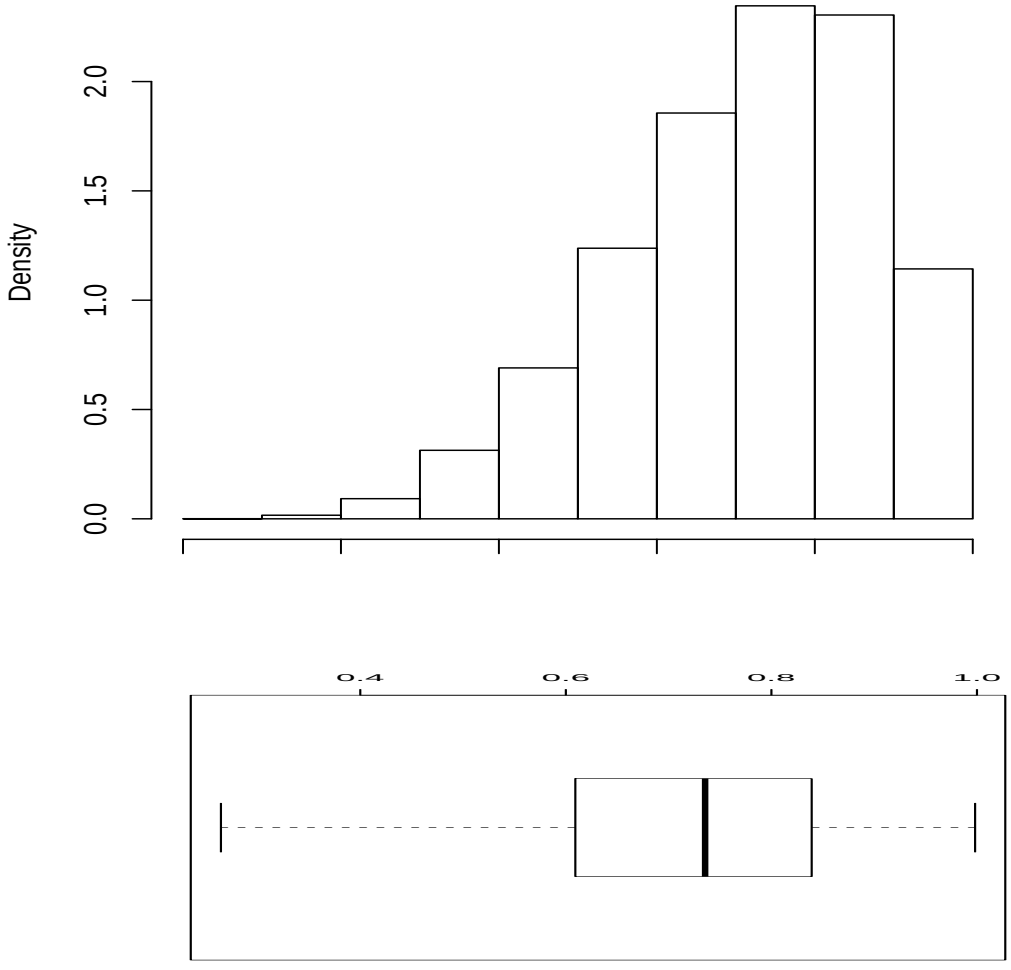
O acordo sobre o formato do desenho é: $Q1$ e $Q2$ formam um retângulo (cuja altura não importa), dentro do qual, fica o separador correspondente ao valor de $Q3$. Para fora do retângulo, "crescem bigodes" até, respectivamente min e max. Por isto que a tradução de BoxPlot é "caixinha com bigodes".

Quartis. BoxPlot.

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.

Quartis. BoxPlot.



Quartis. BoxPlot.

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.

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):

Medidas (caraterísticas) de dispersão.

- 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:

$$170 \text{ e } 190$$

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

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

Explicação sobre o coeficiente de variação.

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:

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