Material Teórico - Módulo de ESTATÍSTICA BÁSICA I

Estatística Básica: O Início

Primeiro Ano do Ensino Médio

Autor: Prof. Francisco Bruno Holanda
Revisor: Prof. Antonio Caminha Muniz Neto

1 Introdução

Podemos dizer que a *Estatística* é a ciência que coleta, organiza e analisa **dados** visando responder certas questões cotidianas utizando um método científico. Sua principal função é evitar determinados erros analíticos que são comuns quando utilizamos métodos *heurísticos*.

Por exemplo, muitas pessoas resolvem incentivar seus filhos a treinarem basquete durante a infância esperando que esse esporte os tornem mais altos quando chegarem à vida adulta. O raciocínio simplista dessas pessoas está baseado na constatação de que a maioria dos atletas profissionais de basquete é formada por atletas muito altos. Na verdade, o que ocorre é exatamente o contrário, sendo chamado **viés de sobrevivência**: apenas as crianças que começam a ficar mais altas do que os colegas ganham destaque nos times juvenis de basquete e, com isso, têm maiores chances de chegar às ligas profissionais, enquanto as crianças de estatura mediana tentam escolher outras profissões. De outra forma, vários estudos médicos comprovaram que a maioria dos jogadores de basquete que são altos também possuem os pais altos, o que aponta fatores genéticos como principais influenciadores da altura de uma pessoa na vida adulta.

Outro exemplo comum que podemos destacar é o uso da Estatística para analisar se determinadas políticas públicas atingiram ou não seus objetivos.

Hoje em dia, os métodos estatísticos são usados em diversos campos de investigação científica, como Medicina, Demografia, Meteorologia, Economia etc.

2 Estatística

A Estatística está dividida em dois ramos principais: a Estatística Descritiva e a Estatística Inferencial, também conhecida como Inferência Estatística. Conforme veremos a seguir, tais ramos correspondem a duas fases distintas da análise estatística de um conjunto de dados.

A Estatística Descritiva objetiva coletar, descrever e sumarizar um determinado conjunto de dados através de um conjunto de tarefas como:

1) Encolher um método apropriado de coletar dados numéricos, evitando um *viés de seleção*. Um **viés de seleção** é um erro comum, que ocorre quando escolhemos uma amostra que não representa uma população como um todo. Por exemplo, suponha que desejamos saber qual é o esporte favorito dos alunos de uma escola, mas, ao entrevistar uma amostra desses alunos, selecionamos somente garotos. Nesse caso, teremos claramente incorrido em um viés de seleção.

2) Determinar como os dados serão apresentados: utilizando uma tabela, um gráfico de pizza, histograma ou outra forma na qual o comportamento global dos dados seja observado com relativa facilidade.

3) Utilizar certas *medidas estatísticas* para descrever um conjunto de dados. Exemplos de medidas estatísticas são as *medidas de tendência central* e as *medidas de dispersão*. Medidas de tendência central incluem a *média*, a *mediana* e a *moda*. Medidas de dispersão incluem o *desvio padrão*, a *variância*, o *valor máximo*, o *valor mínimo*, a *obliquidade* e a *curtose*.

A Inferência Estatística baseia-se na Teoria das Probabilidades para estabelecer conclusões sobre todo um grupo, ao qual nos referiremos doravante como a **população** sob estudo. Para tanto, ela baseia-se nas observações coletadas apenas de uma parte representativa dessa população, parte esta à qual nos referimos como a **amostra** da população em questão.

Neste primeiro módulo, abordaremos os fundamentos da Estatística Descritiva e o uso de alguns *softwares* que servem de apoio para essa disciplina.

3 Coletando e organizando dados

Quando coletamos dados para realizar um estudo, as observações coletadas são chamadas de **dados brutos**. Um exemplo de dados brutos corresponde ao número de multas de trânsito dadas a motoristas de uma determinada região, em um certo mês.

Os dados foram obtidos em uma pesquisa do DETRAN e apresentados na forma em que foram coletados (veja a Tabela 1); por este motivo são denominados *dados brutos*. Geralmente, este tipo de dado traz pouca ou nenhuma informação ao analista, sendo necessário primeiramente organizar os dados com o intuito de aumentar sua capacidade de informação.

Tabela 1: Dados Brutos

Código da Cidade	Número de Multas
AB01	230
CD03	187
GR23	167
FR26	256
GT78	69
FT56	392
AF02	59
GF02	56
AH55	139
VT67	408

Em relação aos dados brutos acima, o primeiro passo que podemos tomar pode ser colocar os dados em ordem crescente, de acordo com o número de multas.

Na Tabela 2, fica claro qual é o menor número de multas (56) e o maior (408). A diferença entre esses dois valores é chamada de **amplitude** dos dados coletados, que nesse caso específico é igual a $408 - 56 = 352$.

Tabela 2: Dados Organizados

Código da Cidade	Número de Multas
GF02	56
AF02	59
GT78	69
AH55	139
GR23	167
CD03	187
AB01	230
FR26	256
FT56	392
VT67	408

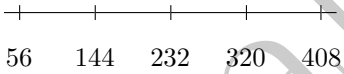
A amplitude representa a medida do intervalo no qual se encontram os dados coletados. Esse intervalo pode ser dividido em subintervalos de mesma medida, chamados de **classes de frequência**.

Em relação ao exemplo acima, se quisermos dividir o intervalo de amplitude total em quatro subintervalos de mesmo tamanho, cada um desses intervalos será uma classe de frequência de **amplitude parcial**

$$\frac{408-56}{4} = \frac{352}{4} = 88.$$

Para evitar ambiguidades, vamos considerar os intervalos que compõem as classes de frequência como fechados à esquerda e abertos à direita. A exceção será o último intervalo, que consideramos como fechado à esquerda e à direita, a fim de não excluir dado algum.

Assim, no caso do exemplo que vimos discutindo, as quatro classes de frequência em que os dados foram divididos são os intervalos $[56, 144)$, $[144, 232)$, $[232, 320)$ e $[320, 408]$. Podemos visualizar esta subdivisão da seguinte forma



Observação
Quando dividirmos um intervalo total de dados em quatro classes de frequência, cada uma dessas classes será chamada de **quartil**.

Uma vez que tenhamos dividido um conjunto de dados em classes de frequência, podemos construir a chamada **tabela de distribuição de frequências de classes**, a qual consiste em três colunas: na coluna da esquerda escrevemos os intervalos que compõem as classes de frequência. Na coluna do centro escrevemos a quantidade de observações (denominada **frequência absoluta**) que pertence ao intervalo correspondente a cada classe. Por fim, na coluna da direita escrevemos a **frequência relativa** de cada classe, que corresponde à razão entre o número de observações situadas em uma determinada classe e o total de observações de que dispomos.

A Tabela 3 a seguir é a tabela de distribuição de frequências de classes relativa aos quartis do conjunto de dados sobre multas. Observe a notação utilizada para cada um dos intervalos de classes de frequência, bem como a última linha, onde listamos o total dos dados coletados, o qual sempre corresponde a uma frequência relativa igual a 1.

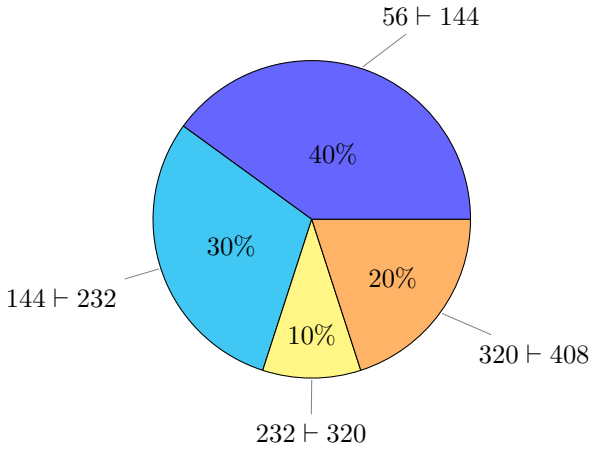
Tabela 3: Distribuição de Frequência de Classes

Classes	Fr. Absoluta	Fr. Relativa
$56 \vdash 144$	4	0,4
$144 \vdash 232$	3	0,3
$232 \vdash 320$	1	0,1
$320 \vdash 408$	2	0,2
Total	10	1

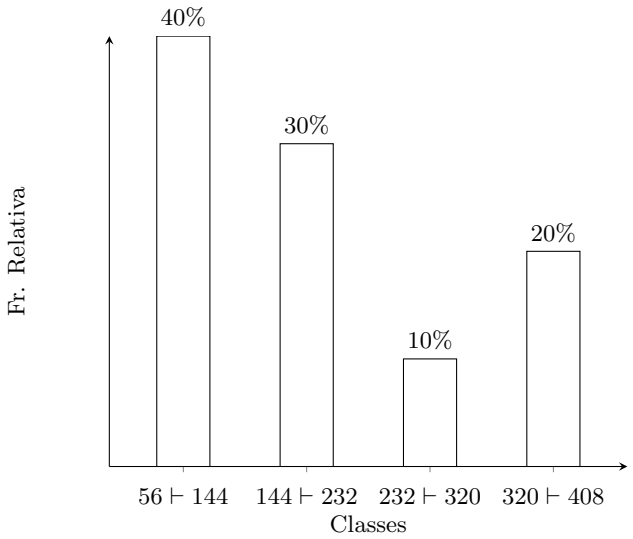
De posse dos valores encontrados na tabela de frequências podemos construir dois tipos de gráficos, os quais nos darão uma perspectiva *visual* dos dados coletados.

Começamos pelo **gráfico de pizza**, cuja construção é relativamente simples. Lembrando que um círculo corresponde a um ângulo de 360°, representaremos cada classe por um setor circular com ângulo central proporcional à sua frequência relativa.

Por exemplo, ainda em relação aos dados sobre multas, a classe $56 \vdash 144$ (cuja frequência relativa é $0,4$) é representada por um setor circular com ângulo central igual a $0,4 \times 360° = 144°$. Calculando os ângulos centrais correspondentes aos demais quartis da amostra de forma análoga, obtemos o gráfico de pizza abaixo:



Outra forma usual de representar os dados graficamente é através de um **histograma**, também conhecido como um **gráfico de barras**. Neste tipo de gráfico, cada barra é proporcional à barra mais alta, que por sua vez corresponde à classe de maior frequência.



Revisemos os conceitos aprendidos até aqui resolvendo o exercício a seguir.

Exercício 1. *A tabela a seguir coleciona o número de pessoas que habitam cada um dos* 25 *apartamentos de um certo condomínio.*

2	4	3	3	4
1	3	4	2	1
1	1	1	2	2
3	2	5	4	2
1	1	6	3	1

a) Construa a tabela de frequências absolutas e frequências relativas, com número de classes igual ao número de diferentes observações.

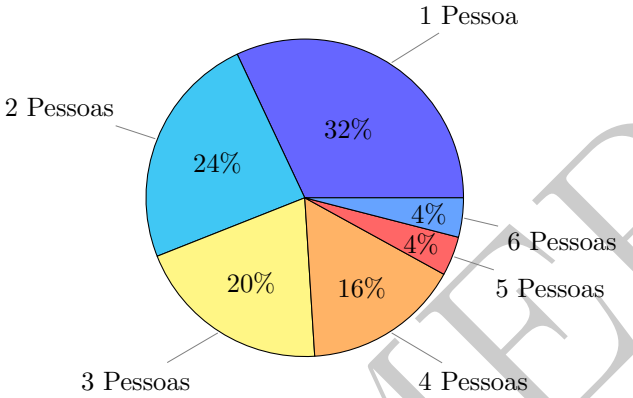
b) Construa um gráfico de pizza e um histograma correspondentes às observações do item (a).

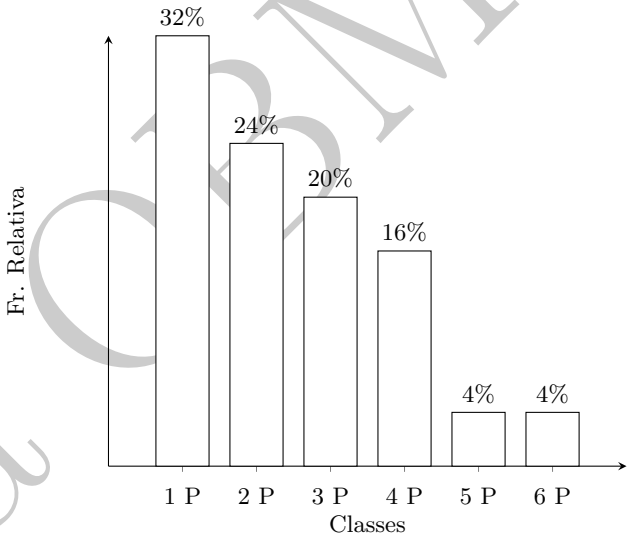
c) Informe qual classe possui a maior frequência.

Solução.
(a) Existem seis tipos de observações distintas: 1, 2, 3, 4, 5, 6. Daí, teremos seis classes, organizadas na tabela de frequências abaixo:

Classes	Fr. Absoluta	Fr. Relativa
1	8	0,32
2	6	0,24
3	5	0,2
4	4	0,16
5	1	0,04
6	1	0,04
Total	25	1

(b) O gráfico de pizza e o histograma correspondentes são esboçados a seguir:





(c) A classe de maior frequência é 1. Ou seja, nesse condomínio predominam pessoas que vivem sozinhas. □

Observação
Quando construímos a tabela de frequências de um determinado conjunto de dados, a classe de maior frequência é chamada de **moda**.

Nos exemplos anteriores, aprendemos como construir uma tabela de frequências simples. Além desse tipo de tabela de frequências, também podemos construir a chamada **tabela de frequências acumuladas**. Montamos essa tabela da seguinte forma:

Considere uma tabela de frequências simples com k classes, em que fa_i e fr_i denotam as frequências absolutas e relativas correspondentes à i-ésima classe, respectivamente.

A **frequência absoluta acumulada** para a i-ésima classe, denotada Fa_i, é definida por

$$Fa_i = fa_1 + fa_2 + \cdots + fa_i.$$

Por outro lado, a **frequência relativa acumulada** para a i-ésima classe, denotada Fr_i, é dada por

$$Fr_i = fr_1 + fr_2 + \cdots + fr_i.$$

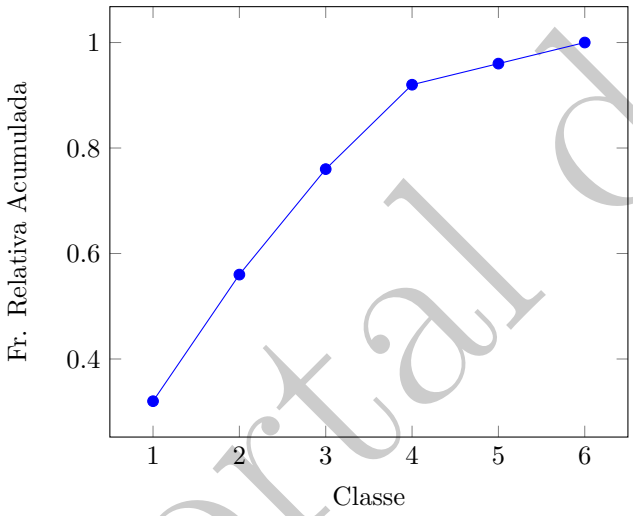
Em palavras, a frequência absoluta (resp. relativa) acumulada para a $i-$ésima classe é igual à soma das frequências absolutas (resp. relativas) correspondentes a todas as classes, desde a classe 1 até a classe i.

Abaixo, apresentamos a tabela de frequências acumuladas relativa ao exercício anterior.

	Fr. Simples		Fr. Acumulada	
Classes	fa	fr	Fa	Fr
1	8	0,32	8	0,32
2	6	0,24	14	0,56
3	5	0,2	19	0,76
4	4	0,16	23	0,92
5	1	0,04	24	0,96
6	1	0,04	25	1

Observe que a frequência absoluta acumulada da última classe é igual ao total de observações, ao passo que a frequência relativa acumulada da última classe é sempre igual a 1.

A partir dessa nova organização dos dados, podemos construir um **gráfico poligonal** no plano cartesiano $\mathbb{R}^2$ para as frequências relativas acumuladas, marcando e conectando os pontos (i, Fr_i) através de segmentos de reta. Abaixo, mostramos tal gráfico correspondente à tabela de frequências acumuladas acima.



4 Sugestões ao professor

Separe dois encontros de 50 minutos cada para desenvolver o conteúdo desta aula. No primeiro encontro, apresente as definições e exemplos. Na segunda, verifique se os alunos aprederam os conceitos através de exercícios propostos.

Do ponto de vista didático, também pode ser interessante elaborar algum projeto no qual os alunos tenham que ir a campo coletar dados. *A posteriori*, esses dados podem ser organizados em tabelas de frequências e utilizados para elaborar gráficos pizza, histogramas e gráficos poligonais. A turma poderá ser dividida em grupos de no máximo cinco alunos, cada grupo ficando responsável por coletar dados sobre algum tópico específico, tais como: coleta de lixo, nível de escolaridade, acesso à Internet em casa etc. O objetivo principal é fazer com que os alunos utilizem os conhecimentos adquiridos para entender melhor a comunidade em que vivem, percebendo, assim, a importância da Estatística no cotidiano.