Material Teórico - Módulo de INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

Conceitos Básicos

Segundo Ano do Ensino Médio

Prof. Francisco Bruno Holanda
Prof. Antonio Caminha Muniz Neto

12 de março de 2020

PORTAL DA MATEMÁTICA
OBMEP

1 Introdução

Nas aulas anteriores, estudamos as principais medias (centrais e de dispersão) de um conjunto de dados. Além das diversas formas de visulização dos dessas informações. Também vimos que a análise de dados é fundamental em diversas áreas da ciência pois permite que as pessoas possam compreender melhor uma determinada situação sem ter que recorrer às suas próprias experiências pessoais.

Neste material vamos aprender um pouco sobre a chamada **Estatística Inferencial**. Esta disciplina permite obter informações aproximadas de uma *população* analisando apenas dados obtidos de uma *amostra*. Por exemplo, suponha que o governo tenha implantado um programa que visa aumentar a frequência e o desempenho escolar dos alunos do ensino fundamental oferecendo um auxílio alimentação àqueles alunos com frequência acima de 90%. Para saber se o programa está realmente atingindo os resultados pretendidos, ao invés de realizar uma pesquisa que envolva todos os alunos - que teria um custo elevado, os pesquisadores podem se restringir a coleta de dados a um subgrupo menor (digamos 5%) da população total de alunos que foi contemplada com o programa.

Agora, antes de apresentarmos o método básico de Estatística Inferencial para o caso geral, iremos analisar um caso particular bem simples a fim de entendermos a abordagem formal de maneira ingênua. Considere o exemplo a seguir:

Exemplo 1. *Em uma família formada por seis pessoas, foi feita uma pesquisa sobre o grau de escolariedade de seus participantes. Os dados foram colecionados na tabela a seguir: É fácil calcular que a média de anos estudados*

Nome	Anos de estudo
Arnaldo	8
Bernaldo	10
Ceraldo	5
Dernaldo	15
Etevaldo	12
Floraldo	10

pelos membros dessa família é de

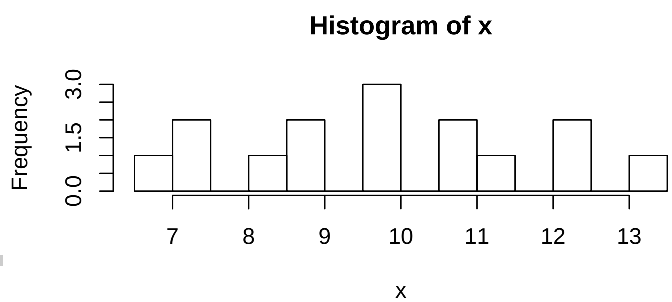
$$Média = \frac{8+10+5+15+12+10}{6} = \frac{60}{6} = 10$$

anos. Agora, suponha que seja possível apenas coletar os dados de duas dessas pessoas. O que acontecerá se calcularmos a média apenas dos dois indivíduos dessa amostra?

Solução. A resposta imediata para esta questão é: *depende dos dois indivíduos escolhidos.* Veja que existem um total de 15 pares de pessoas que podem ser escolhidas. Para cada um destes pares, vamos calcular a *média da amostra.*

Par	Média	Par	Média	Par	Media
A e B	9	B e C	7,5	C e E	8,5
A e C	6,5	B e D	12,5	C e F	7,5
A e D	11,5	B e E	11	D e E	13,5
A e E	10	B e F	10	D e F	12,5
A e F	9	C e D	10	E e F	11

Note que apesar de alguns valores da média amostral estarem "distantes" da média verdadeira, a maioria dos valores estão "próximos" a esta. Por exemplo, a probabilidade de obtermos uma média amostral igual a 9, 10 ou 11 é igual a $\frac{7}{15} \approx 47\%$. Para finalizar, vamos construir o histograma das possíveis médias amostrais.



$\square$

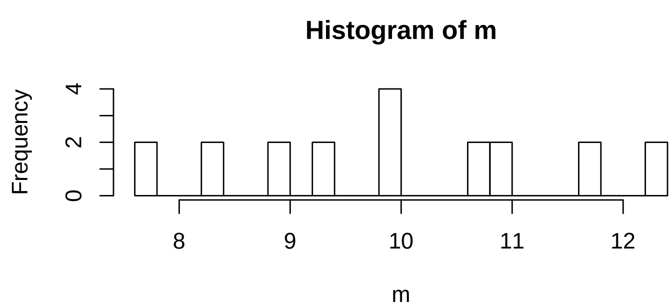
Exemplo 2. *Agora vamos fazer um experimento ligeiramente diferente: ao invés de analisarmos as amostras com dois indivíduos, iremos analisar a média amostral de todos os subgrupos com três pessoas.*

Demonstração. Teremos um total de 20 possíveis amostras. As médias em cada uma dessas amostras serão:

$$10, 10, 8.33, 11.67, 9.33, 7.67, 11, 7.67, 11,$$

$$9.33, 10.67, 9, 12.33, 9, 12.33, 10.67, 8.33, 11.67, 10, 10$$

Veja que a probabilidade de obtermos um valor entre 9 e 11 nesse caso é de $\frac{12}{20} \approx 60\%$. E para finalizar, a figura a seguir mostra o histograma de todos estes valores.



$\square$

A partir destes dois exemplos podemos "intuir" dois conceitos fundamentais de estatística:

1. Não precisamos conhecer os dados de toda uma população para **inferir** sobre sua média. Basta calcular a média de uma amostra. Existe uma probabilidade alta de estarmos "próximo" do valor real.

2. A medida que aumentarmos o tamanho da amostra, mais "seguros" estaremos de que o valor da média amostral está próximo da média populacional.

Infelizmente, a formalização das ideias acima estão fora do escopo deste material. Para entendê-los de forma rigorosa, precisaríamos de dois cursos de nível superior: Cálculo e Probabilidade Avançada. Porém, isso não nos impede de entender os conceitos fundamentais da Estatística.

No próximo exemplo, faremos algo semelhante ao que foi feito nos dois exemplos anteriores. Contudo, em escala maior: iremos considerar uma população com 30 observações. Nesse caso, torna-se muito trabalhoso computar as médias amostrais manualmente. Portanto, faremos uso de um *software* estatístico para realizar os cálculos, expondo apenas os resultos no texto.

Exemplo 3. *Considere uma população com trinta observações. Digamos os valores sejam os inteiros de* 1 *a* 30. *Faça o histograma das possíveis médias amostrais quando os tamanhos das amostras forem:*

a) $n = 2$*;*

b) $n = 3$*;*

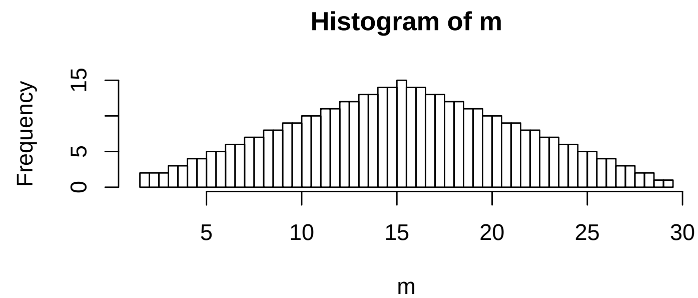
c) $n = 5$.

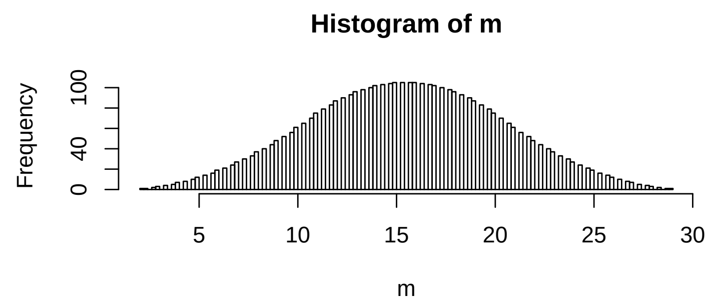
Solução. Como mencionamos anteriormente, utilizaremos um *software* estatístico chamado R para resolver este problema. Os comandos para cada um dos itens são:

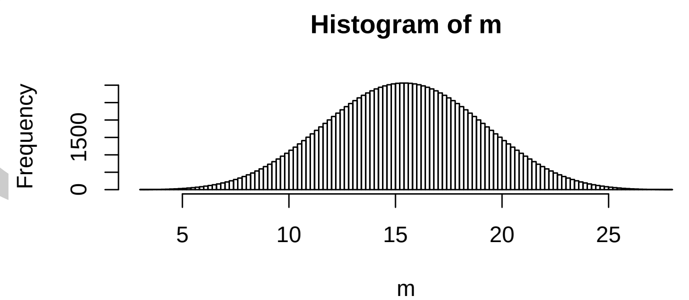
```
a) x <- seq(1,30,1)
   M <- t(combn(x, 2))
   m <- round(rowSums(M)/2,2)
   hist(m,breaks = 100)


b) x <- seq(1,30,1)
   M <- t(combn(x, 3))
   m <- round(rowSums(M)/3,2)
   hist(m,breaks = 100)


c) x <- seq(1,30,1)
   M <- t(combn(x, 5))
   m <- round(rowSums(M)/5,2)
   hist(m,breaks = 100)
```







E os histogramas gerados são:
Sabemos que a média populacional é

$$\mu = \frac{1+2+3+\cdots+30}{30} = \frac{31}{2} = 15,5.$$

Agora, observe que os histogramas mostram que a maioria dos possíveis valores para a média amostral encontram-se "ao redor" da média populacional. Além disso, a medida que aumentamos o tamanho da amostra, aumentamos a probabilidade dessa média amostral estar próxima de μ. $\square$

Além das observações já mencionas, no exemplo anterior é possível perceber que os histogramas das médias amostrais possuem um formato que parecem de um sino quando o tamanho da população e da amostra são relativamente altos. Este fato não é mera coincidência.

Em um curso de probabilidade avançada, os alunos aprendem o conhecido **Teorema Central do Limite**.

Apesar de ser um resultado técnico e de difícil demonstração, podemos entender sua mensagem principal.

A curva em formato de sino é chamada de **curva normal de Gauss** e possui uma série de propriedades que possibilitam o cálculo da probabilidade do valor da média amostral encontrar-se próximo de uma vizinhança da média populacional.

Falaremos de forma mais detalhada sobre estes resultados em uma próxima aula. No momento, o que o aluno deve ter em mente é o seguinte resultado ingênuo:

A média amostral tem alta probabilidade de ser estar próxima da média real (populacional). E quanto maior for o tamanho da amostra, maior será essa probabilidade.

Um exemplo prático trata-se da pesquisa eleitoral.

Exemplo 4. *Em uma cidade com* 100.000 *foi feita uma pesquisa eleitoral para o segundo turno de uma votação para prefeito. Foram coletados dados de um total de* 800 *participantes. Os resultados desta pesquisa foram: Calcule*

Nome	Número de votos
Candidato A	450
Candidato B	300
Nulo	40
Não Sabem	10

uma estimativa para:

a) o total de votos no candidato A;

b) o total de votos no candidato B;

Solução. a) A partir dos dados da amostra, podemos inferir que de cada 800 pessoas, 450 irão votar no candidato *A*. Logo, a estimativa para o número total de votos no candidato A é de

$$\frac{450}{800} \times 100.000 = 56.250$$

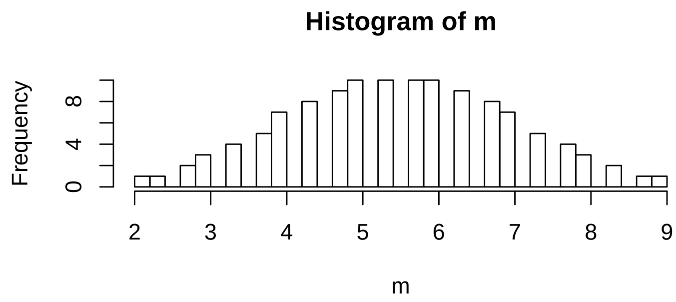
b) De modo análogo, a estimativa para o número total de votos no candidato B é de

$$\frac{300}{800} \times 100.000 = 37.500$$

□

Para finalizar, considere o exemplo a seguir:

Exemplo 5. *Em um condomío há* 10 *adultos e* 10 *crianças. Nessa população, as crianças têm* $1, 2, ..., 10$ *anos de estudo, enquanto que os adultos têm* $11, 12, ..., 20$. *É fácil ver que a média populacional é de* 10.5 *anos de estudo por pessoa. Porém, veja o que ocorre com o histograma da média amostral quando analisamos os subgrupos formados por exatamente três crianças.*



Veja que, neste caso a probabilidade da média amostral estar próxima da média populacional é bem pequena. Isso ocorre pois houve um **viés de seleção**. Ou seja, a amostra escolhida não é uma boa representação da população como um todo. Para evitar este problema, a pesquisa deve coletar os dados de maneira que qualquer indivíduo da população tenha a mesma probabilidade de ser selecionado para a amostra. Por exemplo, no caso do exemplo 4, se a pesquisa eleitoral tiver sido realizada em apenas um único bairro da cidade, a estimativa computada estará comprometida. Nesse caso, a escolha de um bairro em que o candidato A é mais popular irá *viesar* a estatística.

2 Sugestões aos Professores

Separe pelo menos dois encontros de 50 minutos para ensinar o conteúdo deste material. Se possível, ministre a aula em um único encontro de 140 minutos. Um ideia interessante é realizar um experimento de coleta de dados com a turma. Escreva em pedaços de papel os valores correspondentes à tabela do primeiro exemplo. Em seguida, peça para alguns alunos escolherem aleatoriamente dois pedaços, calculando a média em seguida. Anote os resultados no quadro. Faça o histograma das médias amostrais obtidas e compare o resultado com o histograma apresentado no material.

Durante esta aula realizamos uma série de exemplos utilizando o R. Este é um programa livre que está disponível para *download* no site `www.r-project.org` e é compatível com os principais sistemas operacionais. Além disso, vários site na internet simulam este *sotfware* de maneira online e grátis. Um deles é `www.r-fiddle.org`. Copie os códigos apresentados no exemplo 3 e veja o que acontece! Tente também modificar os códigos para obter histogramas para outros exercícios.

Referências

[1] João Ismael Pinheiro et al. *Estatística Básica: a arte de trabalhar com dados.* Campus, 2009.

[2] Pedro A. Morettin and Wilton de O. Bussab. *Estatística Básica*. Saraiva, 2010.