**Material Teórico - Módulo de INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA**

# Aplicações de Inferência Estatística

**Segundo Ano do Ensino Médio**

**Prof. Francisco Bruno Holanda**
**Prof. Antonio Caminha Muniz Neto**

**12 de abril de 2020**

# 1 Introdução

Neste material, iremos apresentar algumas situações que podem ser estudadas utilizando-se os conceitos básicos sobre estatística inferencial aprendidos na aula anterior.

# 2 Estimando a quantidade de pessoas em uma multidão

Você já deve ter percebido que, em geral, os jornais costumam noticiar eventos que reúnem uma grande quantidade de pessoas: protestos, festivais de rua, shows, dentre outros. Também deve ter notado que é comum que essas notícias sejam acompanhadas de *estimativas* dos números de pessoas que participaram desses eventos. Assim, uma pergunta natural é questionar como essas estimativas são feitas e qual o grau de confiabilidade que podemos ter em tais estatísticas.

Figura 1: Aglomerado de pessoas. Foto: `@chuttersnap`

O método mais conhecido para estimar o tamanho de uma multidão é chamado de Método de Jacobs, em homenagem a seu criador, Herbert Jacobs, professor universitário de jornalismo. Ele desenvolveu o método de estimativa do tamanho de uma multidão depois de observar inúmeros protestos contra a Guerra do Vietnã da janela do escritório onde trabalhava.

A ideia geral do método é relativamente simples: calcula-se a área do local, estima-se o número de pessoas por $m^2$, e multiplica-se esses dois números. Se as pessoas concentrarem-se de forma desigual (o que quase sempre ocorre), leva-se isso em consideração, dividindo-se a área de concentração em setores menores.

Jacobs descobriu que, nas multidões mais densas, cada pessoa ocupava cerca de $0,2m^2$. Por outro lado, em uma multidão densa, mas administrável, ele observou que os participantes tinham $0,4m^2$ de espaço para se locomover, ao passo que, em multidões leves, os participantes chegavam a ter até $1m^2$ para locomoção.

Assim, ao particionar a região de concentração de pessoas em setores, é possível estimar quantos indivíduos estão presentes em cada setor. Um exemplo de como essa estimativa é feita pode ser visualizado na foto a seguir:

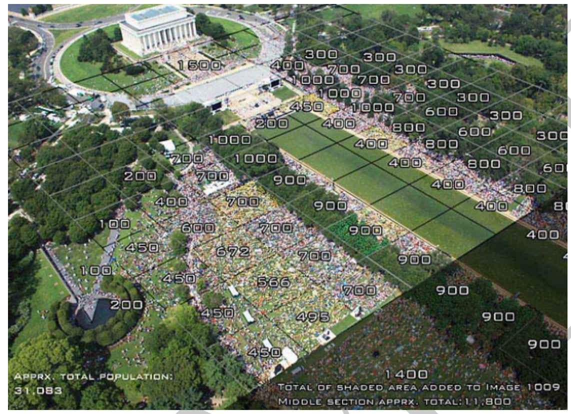



Figura 2: Multidão particionada em setores.

O Método de Jacobs pode parecer uma solução simples e sem rigor científico, mas ele é surpreendentemente eficiente quando é feito por observadores não tendenciosos, ainda mais levando-se em consideração que, com a ajuda de ferramentas modernas, como o Google Earth, tornou-se bastante fácil calcular o tamanho exato da área de determinado local.

# 3 Quantos peixes tem este tanque?

Em muitas regiões do Brasil, é possível encontrar produtores rurais que especializaram-se na criação de peixes em cativeiro. A piscicultura consiste na criação de peixes em tanques, nos quais eles nascem, são alimentados, reproduzem-se e são retirados para o consumo.

Suponha que um piscicultor deseje estimar quantos peixes há em seu tanque. *Como ele pode fazer isso?* Uma solução viável consiste no seguinte processo:

1. No primeiro dia, ele retira uma quantidade $n$ de peixes do tanque, os marcando com algum sinalizador que não comprometa a capacidade de nado dos peixes; uma vez feito isso, ele retorna os peixes marcados ao tanque.

2. No segundo dia, ele retira uma quantidade $2n$ de peixes e calcula a proporção de peixes marcados dentre o total de peixes retirados. Em seguida, ele supõe que essa proporção coincide com a real e usa esta suposição para estimar o total de peixes.

Por exemplo, suponha que o piscicultor tenha retirado (e marcado) 50 peixes no primeiro dia. No segundo dia, ele retira 100 peixes e percebe que 4 peixes estavam marcados. Assim, ele estima que a quantidade de peixes marcados no tanque seja aproximadamente 4% do total. Com esses dados em mãos, estimamos o número total $x$ de peixes no tanque da seguinte forma: uma vez que a fração exata de peixes marcados é $\frac{50}{x}$ e a fração presumida de peixes marcados é 4%, temos

$$\frac{50}{x} \cong \frac{4}{100},$$

equação que nos dá $x \cong 1250$.

É possível demonstrar que, à medida que o valor $n$ de peixes retirados no primeiro dia aumenta, a proporção de peixes marcados estimada com a retirada no segundo dia torna-se cada vez mais próxima da proporção real de peixes marcados.

Você e sua turma podem fazer uma simulação desse experimento utilizando uma caixa de sapatos e feijões, para estimar quantos grãos há em um saco de um quilo. Para tanto, pinte um total de $n$ feijões do saco (tente pintar pelo menos $n = 50$ deles) e, após a tinta secar, coloque-os em uma caixa, junto com os demais feijões do saco. Em seguida, balance bem a caixa e depois retire, sem olhar, 100 feijões. Conte quantos desses 100 feijões estavam marcados, obtendo uma proporção estimada de feijões marcados. Por fim, estime o número total de feijões, igualando essa proporção estimada à proporção real. É instrutivo repetir o processo para $n = 100$ e depois para $n = 200$, comparando as três estimativas assim obtidas.

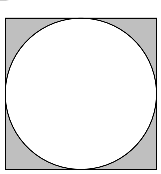
## 4 Estimando o valor de $\pi$ através de simulações

O irracional $\pi$ é um dos números mais importantes da Matemática. Na Geometria, o $\pi$ é definido como a razão entre a circunferência (isto é, o comprimento) de um círculo e seu diâmetro.

Os matemáticos da Antiguidade Clássica grega já eram capazes de calcular as primeiras casas decimais de $\pi$ forma correta, utilizando aproximações dos círculos por polígonos regulares. Com o passar dos anos (de fato, séculos), novos métodos analíticos surgiram para calcular aproximações cada vez melhores de $\pi$, cada vez com um número maior de casas decimais corretas.

Um desses métodos utiliza simulação estatística e será descrito através do seguinte experimento mental:

1. Admita que você tenha um círculo de raio 1 inscrito em um quadrado de lado 2, conforme mostrado na figura a seguir:



2. Suponha que sorteemos, aleatoriamente, um ponto do interior do quadrado. Se todos os pontos tiverem a mesma probabilidade de serem sorteados, a probabilidade de escolhermos um ponto no interior do círculo é de
$$\frac{\text{Área do círculo}}{\text{Área do quadrado}} = \frac{\pi \cdot 1^2}{2^2} = \frac{\pi}{4}.$$

3. Agora, imagine que sorteemos, também aleatoriamente, $n$ pontos no interior do quadrado. Se contarmos quantos pontos foram escolhidos no interior do círculo, podemos esperar que:
$$\frac{\text{Número de pontos no interior do círculo}}{\text{Número de total de pontos escolhidos}} \cong \frac{\pi}{4}. \quad (1)$$

4. Além disso, utilizando métodos avançados de probabilidade, é possível demonstrar que, à medida que a quantidade total de pontos escolhidos aumentar, o valor computado da razão
$$\frac{\text{Número de pontos no interior do círculo}}{\text{Número total de pontos escolhidos}}$$
irá se aproximar cada vez mais do valor real de $\frac{\pi}{4}$. (Este fato é conhecido como o **Teorema do Limite Central**).

Vejamos como esse experimento mental pode ser implementado através de uma simulação computacional. Faremos isso usando os comandos a seguir, escritos na linguagem `R`.

```
n<-100
x.pos <- runif(n, min=-1, max=1)
y.pos <- runif(n, min=-1, max=1)
local.pos <- ifelse(x.pos^2 + y.pos
   ^2 <= 1, TRUE, FALSE)
dentro <- length(which(local.pos ==
   TRUE))
```

Note que o algoritmo consiste de cinco linhas de comando:

1. Na primeira linha, criamos uma variável $n$ e atribuímos um valor inteiro positivo a essa variável $n$. Nesse caso específico, escolhemos $n = 100$.

2. Na segunda linha, criamos uma sequência de $n$ valores aleatórios no intervalo $[-1, 1]$, os quais irão representar a coordenada $x$ dos pontos selecionados. Denote esses valores por $(x_1, x_2, ..., x_n)$.

3. Na terceira linha, criamos uma sequência de $n$ valores aleatórios no intervalo $[-1, 1]$, os quais irão representar a coordenada $y$ dos pontos selecionados. Denote esses valores por $(y_1, y_2, ..., y_n)$.

4. Na quarta linha, estabelecemos um critério de validação para saber se cada um dos $n$ pontos gerados está no interior ou no exterior do círculo. Esse critério é intuitivo: a distância do ponto $(x_n, y_n)$ até o centro da figura (o ponto $(0, 0)$) for menor do que ou igual a 1 se, e só se, o ponto está no interior do círculo.

5. Por fim, na quinta linha, contamos a quantidade de pontos que foram classificados como estando no interior do círculo. Essa quantidade é armazenada na variável `dentro`.

Agora, graças a (1), para estimar o valor de $\pi$, bastar calcularmos o valor de

$$4 \times \frac{\texttt{dentro}}{n}.$$

Usando esses comandos, realizamos três blocos com três simulações cada; em cada bloco, variamos a quantidade de pontos $n$.

I. Com $n = 100$, ao realizarmos três simulações diferentes para estimarmos o valor de $\pi$ encontramos os seguintes resultados:

`3,48     3,04     3,4`

II. Com $n = 1000$, ao realizarmos três simulações diferentes para estimarmos o valor de $\pi$ encontramos os seguintes resultados:

`3,08     3,004     3,172`

III. Com $n = 10000$, ao realizarmos três simulações diferentes para estimar o valor de $\pi$ encontramos os seguintes resultados:

`3,1224     3,0996     3,1448`

Assim, percebemos que, à medida que aumentamos o valor de $n$, mais os valores estimados tornam-se próximos do verdadeiro valor de $\pi$, que, com seis casas decimais corretas, é $3{,}141592\ldots$.

**Observação 1.** *Ao repetir essas simulações em um outro computador, os valores estimados serão diferentes pois são frutos de um experimento aleatório. Porém, a aproximação dessas estimativas do valor real de $\pi$ continuará sendo observada ao aumentarmos o valor de $n$.*

## 5 Sugestões aos Professores

Ao longo desta aula, criamos algoritmos em R. Este é um *software* gratuito e de código aberto, que pode ser encontrado no site `www.r-project.org`, sendo compatível com os principais sistemas operacionais. Além disso, vários sites na Internet simulam este *sotfware* de maneira online. Um deles é `www.r-fiddle.org`.

Recomendamos que o professor utilize esse site para apresentar o funcionamento do algoritmo apresentado nesse material, demonstrando como a simulação ocorre na prática. Se não for possível fazer a apresentação no computador, uma alternativa é construir uma caixa quadrada de isopor, desenhar um círculo inscrito nesse quadrado e utilizar grãos de milho para simular os pontos sorteados aleatoriamente.

A simulação que fizemos em `R` é conhecida no meio acadêmico como *simulação de Monte Carlo*. Esse método é usado em diversas áreas científicas para estimar a probabilidade de eventos futuros que não podem ser previstos com exatidão, tais como: a probabilidade de chuva nos próximos dias, a probabilidade de uma doença infecciona se espalhar por diversos países ou a probabilidade da inflação aumentar nos próximos meses.

## 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.