

Introdução / motivação

①

Não parece errado afirmar que hoje, muitas pessoas, das mais variadas profissões, têm necessidade de trabalhar com dados e assim, envolvem-se com a Estatística.

Definição de Estatística

A estatística é um conjunto de técnicas que permite, de forma sistemática,

1^a etapa: recolher e organizar;

2^a etapa: explorar e descobrir e

3^a etapa: analisar e interpretar dados oriundos de estudos ou experimentos,

realizados em qualquer área do conhecimento.

Áreas da Estatística

(2)

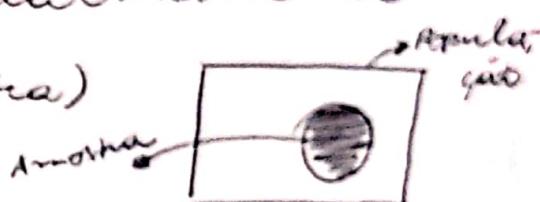
De modo bem geral, a estatística é composta por três áreas:

- * Estatística descritiva (ou Análise Exploratória de dados) ; etapa inicial da análise, quando tomamos contactos com cálculos de medidas descritivas (média e variância) ou dados seja prima vez.

VISÃO MODERNA:

TUKEY (1977) : técnicas gráficas além do cálculo de medidas descritivas.

- * Probabilidade : a teoria de probabilidade nos permite descrever os fenômenos aleatórios, ou seja, aqueles em que está presente a incerteza.

* Inférença Estatística: É o estudo de técnicas que possibilitam a extrapolação,
a um grande ^(população) conjunto de dados, das
informações e conclusões obtidas a partir
de subconjuntos de valores, usualmente de
dimensão menor. → (amostra) 
Permite fazer afirmações sobre as características de uma população, com base em informações dadas por amostras. ~ D. A. R.

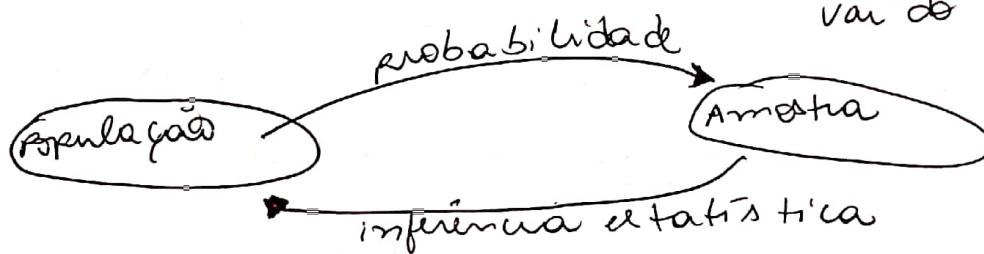
O uso de informações de uma amostra para concluir sobre o todo faz parte da atividade diária da maioria das pessoas.

Basta observar como uma engenheira verifica se o prato que ela está separando tem ou não a quantidade adequada de sal.

④ Relações entre probabilidade e inferência estatística

A probabilidade faz consideração da população para a amostra (raciocínio deductivo) e a inferência estatística faz consideração da amostra para a população (raciocínio indutivo)

→ por meio de qual se vai do específico ao geral.



Raciocínio deductivo : no qual se argumenta das premissas para chegar à conclusão.

↓
fatores considerados na dedução

Exemplo : O ferro condiz eletricidade.
O ferro é metal.
O ouro condiz eletricidade.
O ouro é metal.

premissas

O ouro condiz eletricidade.
O ouro é metal.
Logo, os metais condizem eletricidade.

conclusão

Raciocínio deductivo:

(3) a

premissa maior: todos os alunos de inferência estatística têm mais de 18 anos.

premissa menor: José é aluno de inferência estatística.

Conclusão: José tem mais de 18 anos.

Ol, andai, quando um comprador, apoi ④
experimentar um sedaço de laranja numa
banca de feira, decide se vai comprar ou
não as laranjas.

Decisão baseadas em procedimentos amostrais.

Neste objetivo nesta disciplina é procurar
dar a concetração formal a esses princí-
pios intuitivos do dia-a-dia para que
possam ser utilizados cientificamente
em situações mais complexas.

População e Amostra

Definição População é o conjunto de todos os
elementos ou resultados sob investigação.
(BUSSAB)

Amostra é qualquer subconjunto da
população.

Vejamos exemplos para melhor entender
essas definições.

População

40

* Definição: O conjunto de valores de uma característica (observável) associada a uma coleção de indivíduos ou objetos de interesse é dito ser uma população.

(Belfiore e Santos)

Exemplo 1 : consideremos uma pesquisa⁽²⁾ para estudar os salários dos 500 funcionários de uma companhia. Selecionam-se 36 indivíduos e anotam-se os seus salários.

A população é formada pelos 500 funcionários da companhia. A amostra é constituída pelos 36 indivíduos selecionados.

A variável aleatória a ser observada é o salário. Na realidade, estaremos interessados nos salários, portanto, para sermos mais precisos, devemos considerar como a população os 500 salários correspondentes aos 500 funcionários. Conseqüentemente, a amostra será formada pelos 36 salários dos indivíduos selecionados.

Exemplo 2 : mostrar que a distribuição dos salários na amostra reflete a distribuição de todos os salários, desde que a amostra tenha sido escolhida com cuidado.

queremos estudar a proporção de indivíduos⁽⁶⁾ na cidade A que são favoráveis a certo projeto governamental. 200 pessoas foram sorteadas e a opinião de cada uma é registrada como sendo a favor ou contra o projeto.

A população consiste de todos os moradores da cidade, e a amostra é formada pelas 200 pessoas selecionadas.

Neste contexto, podemos definir a variável X que assume o valor 1, se a resposta de um morador for favorável e o valor 0 se a resposta for contrária ao projeto. Assim, a população teria a distribuição de X e a amostra seria constituída de uma sequência de 200 zeros e uns.

Exemplo 3 : Para investigar a "honestidade" de uma moeda, nós a lançamos 50 vezes e contarmos o número de caídas observadas.

Como no exemplo anterior, podemos definir a variável X que assume o valor 1 se cairer cara e 0 se cairer coroa.

Assim, a população seria a distribuição de X e a amostra seria uma sequência de 50 números zeros ou uns.

Ainda, podemos definir uma variável X assumindo o valor 1, com probabilidade p , se cairer cara, e assumindo o valor 0, com probabilidade $1-p$, se cairer coroa. Ou seja, a população pode ser considerada como tendo distribuição de Bernoulli (p).

A amostra seria uma sequência de 50 números zeros ou uns.

Observação: Estes exemplos mostram
uma ampliação do conceito definido de
populações, ou seja, designamos agora a
população como sendo a função probabi-
lidade ou função densidade de probabili-
dade de uma variável aleatória X , modela-
ndo a característica de interesse.

P

Nestes casos, simplificaremos a linguagem,
dizendo: "seja a população $f(x)$ " ou
"a população X " significando que a
variável de interesse X , definida sobre a
população, segue uma distribuição $f(x)$.

Como selecionar uma amostra?

A maneira de se obter uma amostra é
muito importante e existem diversos ~~metodos~~
~~procedimentos~~ especialidades dentro
da estatística para tal fim, por
exemplo amostragem e planejamento
de experimentos.

- * Levantamentos amostrais;
- * Levantamentos observacionais;
- * Planejamento de experimentos;

Levantamentos Observacionais

Os dados são coletados sem que o pesquisador tenha controle sobre as informações obtidas, exceto eventualmente sobre possíveis erros grosseiros. As séries de dados temporais são exemplos típicos destes levantamentos. Por exemplo, queremos prever as vendas de uma empresa em função de vendas passadas. O pesquisador não pode selecionar dados, estes são as vendas efetivamente ocorridas.

No caso de uma série temporal, podemos pensar que a série efetivamente observada é uma das infinitas possíveis realizações. A população hipotética aqui seria o conjunto de todas essas realizações e a série observada seria a amostra.

Planejamento de experimentos:

102

Principal objetivo é o de analisar o efeito de uma variável sobre outra. Requer, portanto, interferência do pesquisador sobre o ambiente em estudo (população), bem como o controle de fatores externos, com o intuito de medir o efeito desejado.

Nesta disciplina, iremos nos concentrar principalmente em levantamentos amostrais.

Levantamentos Amostrais

A amostra é obtida de uma população bem definida e por meio de métodos controlados pelo pesquisador.

Poderemos dividir-las em dois grupos:

Levantamentos probabilísticos e não-probabilísticos.

Levantamentos probabilísticos reúnem todas aquelas técnicas que usam mecanismos aleatórios de seleção dos elementos de uma amostra, atribuindo a cada um delas uma probabilidade, conhecida a priori, de pertencer a amostra.

Levantamentos não-probabilísticos reúnem os demais procedimentos, fazendo com que amostras intencionais, nas quais os elementos não se Selecionados com o auxílio de especialistas e

amostra de voluntários.

(10)

vantagem dos levantamentos probabilísticos:
permite medir a precisão da amostra obtida
baseando-se no resultado contido na própria
amostra.

Amostragem probabilística e não probabilística

Amostragem probabilística é o processo de seleção de uma amostra na qual cada unidade amostral da população tem probabilidade diferente de ser e ir a considerar de pertencer à amostra.

Na amostragem não-probabilística, a probabilidade de seleção é desconhecida para alguns ou todos os elementos da população, podendo alguns destes elementos ter probabilidade nula de pertencer à amostra, como em amostras intencionais ou de voluntários.

VIÉS $|E(\hat{\theta}) - \theta|$

viés de seleção

O melhor modo de evitar o viés de seleção é o uso do sorteio, seja ele manual ou por meio de uma tabela de números aleatórios, ou então pelo gerador de números aleatórios por computador.

A amostragem probabilística é isenta de viés de seleção.

Nesta disciplina, iremos nos concentrar principalmente em levantamentos amostrais e mais ainda num caso específico de amostragem probabilística, a amostragem aleatória simples (a.s.)

Amostragem Aleatória Simples

A amostragem aleatória simples é a maneira mais simples (fácil) para selecionarmos uma amostra probabilística de uma população.

Utilizando - se um procedimento aleatório, sorteia - se um elemento da população, sendo que todos os elementos têm a mesma probabilidade de serem selecionados. Repete - se o procedimento até que sejam sorteadas as n unidades da amostra.

*** (pág 11)

Poderemos ter uma a.s. com reposição, se for permitido que uma unidade possa ser sorteada

mai de uma vez e sem reposição, se a ⑪ unidade sorteada for removida da população.

Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado.

Contudo, a amostragem com reposição conduz a um tratamento teórico mais simples, pois ela implica que tentarmos independência entre as unidades selecionadas. Esta independência facilita o desenvolvimento da inferência estatística.

Portanto, o plano amostral considerado será o de (amostragem aleatória simples com reposição), que denotaremos simplesmente por $a.a^1$. *(Nesse caso as n unidades que compõem a amostra são selecionadas de tal forma que todas as possíveis amostras têm a mesma probabilidade de serem escolhidas.)* Reparemos com algum detalhe o significado mais preciso de uma amostra.

Exemplo : Numa urna têm-se cinco tiras de papel numeradas 1, 3, 5, 5, 7. Uma tira é sorteada e recolocada na urna; então, uma segunda tira é sorteada.

Considere o problema acima, em que colhemos todas as amostras possíveis de tamanho 2, com reposição, da população $\{1, 3, 5, 5, 7\}$.

Defina a variável X : valor assumido pelo elemento na população. Então, a distribuição de X é dada por

| x | 1 | 3 | 5 | 7 |
|----------|---------------|---------------|---------------|---------------|
| $p(x=x)$ | $\frac{1}{5}$ | $\frac{1}{5}$ | $\frac{2}{5}$ | $\frac{1}{5}$ |

Sejam X_1 e X_2 o primeiro e o segundo número sorteado, é possível escrever a distribuição conjunta de X_1 e X_2 .

$$S = \{(1, 1), (1, 3), (1, 5), (1, 5), (1, 7), \\ (7, 1), (7, 3), (7, 5), (7, 5), (7, 7)\}$$

Neste exemplo,
 são amostras
 com reposição e
 probabilidades
 iguais.
 Existe

| $x_2 \setminus x_1$ | 1 | 3 | 5 | 7 | Total |
|---------------------|----------------|----------------|-----------------|----------------|----------------|
| 1 | $\frac{1}{2}5$ | $\frac{1}{2}5$ | $\frac{2}{2}5$ | $\frac{1}{2}5$ | $\frac{1}{2}5$ |
| 3 | $\frac{1}{2}5$ | $\frac{1}{2}5$ | $\frac{2}{2}5$ | $\frac{1}{2}5$ | $\frac{1}{2}5$ |
| 5 | $\frac{2}{2}5$ | $\frac{2}{2}5$ | $4\frac{1}{2}5$ | $\frac{2}{2}5$ | $\frac{2}{2}5$ |
| 7 | $\frac{1}{2}5$ | $\frac{1}{2}5$ | $\frac{2}{2}5$ | $\frac{1}{2}5$ | $\frac{1}{2}5$ |
| Total | $\frac{1}{2}5$ | $\frac{1}{2}5$ | $\frac{2}{2}5$ | $\frac{1}{2}5$ | 1 |

As variávies x_1 e x_2 são independentes?

$$P(X_1 = x_1, X_2 = y_2) = P(X_1 = x_1) \cdot P(X_2 = y_2)$$

Identify common distributional patterns.

Desse modo, cada uma das 25 amostras de tamanho 2 que podemos extrair dessa população corresponde a observar uma particular realização das variáveis x_1 e x_2 , com x_1 e x_2 independentes e

$$P(x_1=x) = P(x_2=x) = P(x=x), \forall x$$

Essa é a caracterização de a a 1.

Exercício: Resposta a exemplo, considerando sem reprodução.

As variáveis x_1 e x_2 são independentes?

As variáveis x_1 e x_2 são identicamente distribuídas?

Comentário: O modelo de amostragem aleatória na definição, algumas vezes, é demandado de amostragem de uma população infinita.

Pense na obtenção dos valores de x_1, \dots, x_n sequencialmente. Primeiro, o experimento é realizado e se pode observar $x_1 = x_1$. Então, o experimento é repetido e se observa $x_2 = x_2$. A suposição de independência na amostragem aleatória implica que a distribuição de probabilidade para x_2 não é afetada pelo fato de que se observou anteriormente $x_1 = x_1$. "Remover" x_1 da população infinita não modifica a população, desse modo, $x_2 = x_2$ ainda é uma observação aleatória da mesma população.

(Casella págs 186 e 187).

Uma amostra se diz casual simples quando $P(x_j = x_i) = \frac{1}{N}$, para qualquer que sejam $i = 1, \dots, n$ e $j = 1, 2, \dots, N$.

↑ tamanho da população
tamanho da amostra

Isto significa que em uma amostra casual simples todos os elementos da população têm a mesma probabilidade de serem selecionados.

a) Quando a amostragem é feita com reposição, para $n=\infty$, temos

$$P(X_1=x_1, X_2=x_2) = \frac{1}{N^2} \quad \text{e} \quad P(X_2=x_2 | X_1=x_1) =$$

$$\frac{\frac{1}{N^2}}{\frac{1}{N}} = \frac{1}{N}$$

b) Quando a amostragem é sem reposição, para $n=2$, temos

X_1 e X_2 não são mutuamente independentes.

$$P(X_1=x_1, X_2=x_2 | \dots) = \frac{1}{N(N-1)}$$

$$P(X_2=x_2 | X_1=x_1) = \frac{1}{N-1}, \text{ portanto}$$

$$P(X_1=x_1, X_2=x_2) = \frac{1}{N(N-1)}.$$

Podemos formar o quadro a seguir.

| ^{1-a seleção} | x_1 | x_2 | \dots | x_N | $p(x)$ |
|------------------------|--------------------|--------------------|----------|--------------------|---------------|
| x_1 | 0 | $\frac{1}{N(N-1)}$ | \dots | $\frac{1}{N(N-1)}$ | $\frac{1}{N}$ |
| x_2 | $\frac{1}{N(N-1)}$ | 0 | \dots | $\frac{1}{N(N-1)}$ | $\frac{1}{N}$ |
| \vdots | \vdots | \vdots | \vdots | \vdots | \vdots |
| x_N | $\frac{1}{N(N-1)}$ | $\frac{1}{N(N-1)}$ | \dots | 0 | $\frac{1}{N}$ |
| $r(x)$ | $\frac{1}{N}$ | $\frac{1}{N}$ | \dots | $\frac{1}{N}$ | 1 |

E é interessante notar que x_1, \dots, x_n são identicamente distribuídas, ou seja, a distribuição marginal de x_i é a mesma para cada $i = 1, \dots, n$.

Logo, tanto para amostragem com reposição como para sem reposição, teremos

$$P(x_j = x_i) = \frac{1}{N} \text{ e } P(x_{j+1} = x_i) = \frac{1}{N}.$$

DEFINIÇÃO: uma amostra aleatória simples 14 de tamanho n de uma "população X ", é o conjunto de n variáveis aleatórias independentes x_1, x_2, \dots, x_n , cada uma com a mesma distribuição de X . Se reja, a amostra será a n -upla (x_1, x_2, \dots, x_n) . $\rightarrow *$ atras

DEFINIÇÃO: Diz-se que as variáveis x_1, x_2, \dots, x_n formam uma amostra aleatória (simples) de tamanho n se

- 1- os x_i 's são variáveis aleatórias independentes.
- 2- todo x_i possui a mesma distribuição de probabilidades.

As condições 1 e 2 podem ser parafraseadas, dizendo-se que os x_i 's são independentes e identicamente distribuídos (iid).

Comontâns: Quando a população é caracterizada por uma ~~distribuição~~ de probabilidades, o modo mais simples para sortear uma amostra é usar os procedimentos de simulação.

Para retirar uma amostra (com reposição) de n indivíduos da população X , basta gerar n números aleatórios independentes dessa distribuição.

Exemplo: Vamos retirar uma amostra de 5 alturas (em cm) de uma população de mulheres cujas alturas X seguem distribuição $N(167, 25)$.

\hookrightarrow média \hookrightarrow variância

Usando, por exemplo, o gerador de números aleatórios do R, rnorm ($n=5$, mean = 167, sd = 5),

obtemos os valores

$$x_1 = 167,9224 \quad x_2 = 174,9222 \quad x_3 = 165,8441$$

$$x_4 = 167,6694 \quad \text{e} \quad x_5 = 167,8116.$$

Desse modo, obtemos uma amostra

x_1, x_2, x_3, x_4 e x_5 de uma população $X \sim N(167, 25)$.

$$N(167, 25).$$

Assim, $x_1 \sim N(167, 25)$

$$x_2 \sim N(167, 25)$$

$$x_3 \sim N(167, 25)$$

$$x_4 \sim N(167, 25)$$

$$x_5 \sim N(167, 25)$$

e x_1, x_2, x_3, x_4 e x_5 são independentes.

$\Rightarrow x_1, x_2, x_3, x_4$ e x_5 são variáveis aleatórias iid.

1-a amostra

$$x_1 = 30,7 \quad x_2 = 29,4 \quad x_3 = 31,1$$

2-a amostra

$$x_1 = 28,8 \quad x_2 = 30,0 \quad x_3 = 31,1$$

Antes de obter os dados, há incerteza acerca do valor de cada x_i . Devido a essa incerteza, considera-se que os dados estejam disponíveis, considerando cada observação como uma variável aleatória e denotando a amostra por X_1, X_2, \dots, X_n .

Comparação da inferência estatística com método de simulação da contagem

Quando "obtemos" a amostra, estamos apenas observando o resultado da simulação, não conhecemos nada do processo gerador dos dados.

↳ (distribuição de X).

O objetivo da inferência estatística é fornecer critérios para nos ajudar a descobrir a forma da distribuição e/ou parâmetros.