CE 301 - ESTATÍSTICA BÁSICA

• Inferência estatística

- População: É o conjunto de todos os elementos que possuem alguma característica comum que temos interesse em estudar.
- Amostra: É um subconjunto da população.
- Inferência: Ramo da Estatística que tem como objetivo estudar a população por meio de evidências fornecidas por uma amostra.

Geralmente estamos interessados em quantidades populacionais, contudo trabalhar com a população pode ser custoso, para solucionar isso trabalhamos com uma amostra. O objetivo das técnicas de amostragem é gerar um subconjunto que seja representativo em relação a população para estimar as quantidades de interesse. No entanto, caso se repita o processo de amostragem, uma amostra diferente da inicial será obtida e consequentemente, as medidas de interesse calculadas em diferentes amostras não serão iguais. Desse modo, como há aleatoriedade envolvida, os valores calculados com base na amostra são candidatos à quantidade da população.

Os objetivos da Inferência enstatística são:

1- Estimar quantidades com base apenas na amostra (estimativa pontual).

2- Avaliar o quão preciso ou creditável é o valor estimado (intervalo de confiança).

3- Decidir sobre possíveis valores da quantidade baseado apenas na amostra (teste de hipóteses).

⟶ Conceitos importantes

• Parâmetro: Uma medida numérica que descreve alguma característica da população. Geralmente são desconhecidos. Por exemplo, uma média, uma proporção, variância, etc.

Representados por letras gregas ($\theta, \mu, \sigma, \ldots$).

• Espaço paramétrico: Conjunto de valores que um parâmetro pode assumir.

• Estimador: Função da variável aleatória.

Cálculo efetuado com os elementos da amostra com a finalidade de representar (estimar) um parâmetro da população.
Usualmente representados por letras gregas com acento circunflexo $(\hat{\theta}, \hat{\mu}, \hat{\sigma}, \ldots)$.

- Estimativa : Valores numéricos assumidos pelos estimadores. Uma função dos valores observados da variável. Um número.
- Estimativa pontual: Um único valor numérico como candidato para o parâmetro de interesse.
- Estimativa intervalar: Intervalo de conjunto de valores "possíveis" para o parâmetro de interesse.

Exemplo: Suponha que temos interesse em estimar a média da idade dos alunos de um curso de graduação. Calcular a média populacional é muito custoso, por isso, tomou-se uma amostra da população. Com essa amostra foi usado um estimador para chegar a uma estimativa da média populacional. Complementar à estimativa pontual foi construído um intervalo de confiança para estimativa.
A coordenação do curso tem interesse em avaliar se existe evidência suficiente nos dados que permite afirmar que a idade média é menor que 22 anos, para isso, pode ser feito um teste de hipóteses.

Exemplo (População de domicílios): Considere a população formada por 3 domicílios. Observou-se as seguintes variáveis:

Variável	Valores		
Unidade	1	2	3
Nome do chefe	Ada	Beto	Ema
Idade	20	30	40
Renda bruta (salários mínimos)	12	30	18
Nº de trabalhadores	1	3	2

→ Parâmetros: Idade média $= \mu_I = \dfrac{20+30+40}{3} = \dfrac{90}{3} = 30$.

Renda média (por trabalhador) $= \mu_R = \dfrac{12+30+18}{1+3+2} = \dfrac{60}{6} = 10$.

→ Amostras de tamanho 2: $s_1 = \{1,2\}$, $s_2 = \{1,3\}$, $s_3 = \{2,3\}$.

Considerando a amostra $s_1 = \{1,2\}$.

Média da idade (x): $\bar{x} = \frac{20+30}{2} = 25$.
↳ estimativa pontual da média da idade

média da renda bruta (y): $\bar{y} = \frac{12+30}{2} = 21$.

média do nº de trabalhadores (T): $\bar{t} = \frac{1+3}{2} = 2$

• Estatística: Qualquer característica numérica dos dados correspondentes à amostra s, é chamada de estatística, ou seja, é qualquer função de uma variável calculada na amostra s.

→ Distribuição Amostral

Suponha que estamos interessados em uma variável aleatória na população, denotada por Y (por exemplo, o peso dos indivíduos).

Desta variável aleatória tomamos uma amostra de tamanho n, que denotaremos $Y_1, Y_2, \ldots, Y_n$ (por exemplo, uma amostra de pesos dos indivíduos da população). Em geral, consideramos que esta amostra é aleatória simples com reposição, de modo a garantir que os elementos da amostra sejam independentes e identicamente distribuídos (iid).

Suponha que temos interesse em uma quantidade populacional θ (por exemplo, o peso médio dos indivíduos). Não conseguimos obter o valor real de θ, então vamos estimar θ por meio de um estimador $\hat{\theta}$ que é uma função das variáveis aleatórias constituintes da amostra, isto é, $\hat{\theta} = f(Y_1, Y_2, \ldots, Y_n)$.

Veja que o estimador é uma variável aleatória (sabemos o que pode acontecer, mas não o que vai acontecer). E variáveis aleatórias têm distribuição de probabilidade.

A distribuição de probabilidade de estatísticas é chamada de distribuição amostral.

Ou seja, imagine que coletamos diversas amostras. Em cada amostra calculamos o estimador de interesse. Se obtivermos a distribuição empírica desse estimador, podemos fazer inferência.

A distribuição amostral pode ser usada para avaliar o que aconteceria se o estudo fosse replicado um grande número de vezes.

A estimativa pontual é um resumo da distribuição amostral. Contudo, na prática, temos apenas uma amostra. Mas diversas estatísticas de interesse têm distribuições amostrais conhecidas, como a média, variância e proporção.

• Distribuição amostral: É a distribuição de uma estatística obtida a partir de várias amostras tiradas de uma população. Ou seja, descreve como a estatística se comporta quando você coleta várias amostras diferentes da população.

Exemplo (População de domicílios): Queremos determinar a distribuição amostral da estatística definida como a razão entre a renda familiar e o nº de trabalhadores.

Considere a variável D: renda média por trabalhador

Possíveis amostras: $S = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$.

Como é uma amostra aleatória simples, todas têm a mesma probabilidade de se observar, ou seja, $P(s) = 1/9$.

Calculando a estatística D para todas as amostras:

s	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
d	12	10,5	10	10,5	10	9,6	10	9,6	9
P(s)	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

para $s = (3,1)$: $d = \frac{18+12}{2+1} = \frac{30}{3} = 10$.

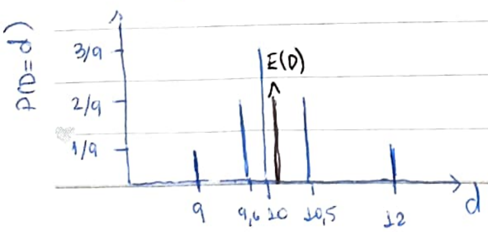
Então, a distribuição amostral para estatística D é:

d	9	9,6	10	10,5	12
P(D=d)	1/9	2/9	3/9	2/9	1/9

Podemos resumir a distribuição amostral de D usando a esperança e variância dela:

$$E(D) = \sum_d d \cdot P(D=d) = 9 \cdot \frac{1}{9} + 9{,}6 \cdot \frac{2}{9} + 10 \cdot \frac{3}{9} + 10{,}5 \cdot \frac{2}{9} + 12 \cdot \frac{1}{9} = 10{,}13.$$

$$Var(D) = \sum_d [d - E(D)]^2 \cdot P(D=d) = (9-10{,}13)^2 \cdot \frac{1}{9} + (9{,}6-10{,}13)^2 \cdot \frac{2}{9} + \dots + (12-10{,}13)^2 \cdot \frac{1}{9} = 0{,}63$$



- Para populações pequenas é fácil obter a distribuição amostral.
- Mas e para populações grandes?
- Veja que não assumimos nenhuma distribuição para variável aleatória de interesse.

→ Distribuição amostral da média amostral: variáveis aleatórias normais

Considere Y_i com distribuição normal de média μ e variância σ^2, $i = 1, \dots, N$, ou seja, $Y_i \sim N(\mu, \sigma^2)$.

Suponha que uma amostra aleatória de tamanho n foi obtida, os valores observados são $y_1, \dots, y_n$.

A distribuição amostral da média $\bar{y} = \sum_{i=1}^{n} y_i / n$ é dada por:

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

, ou seja, a média amostral tem distribuição normal com mesma média da população e variância igual a variância da população dividida pelo tamanho da amostra.