8.1-Noções gerais.

(Introdução, amostragem, parâmetro, estimador, estimativa, distribuição de amostragem e Teorema do limite central.(TLC)).

Introdução.

Na **inferência estatística**, analisamos e interpretamos amostras com o objetivo de tirar conclusões acerca da população de onde se extraiu a amostra.

População- é o conjunto de todos os elementos em estudo.

Amostra- é um subconjunto finito da população.

Nota: Usamos amostras porque, pode demorar demasiado tempo a recolher os dados da população, ter demasiados custos ou até ser impossível.

Exemplos: Obter a média dos pesos de todas as pedras de uma determinada praia de calhau. Pode não ser impossível pesar todas as pedras, mas seria muito difícil e pouco adequado realizar essa tarefa.

Exemplo: medir a temperatura do ar em todos os pontos da atmosfera terrestre. Seria impossível.

Exemplo: Provar todas as bolachas de uma determinada remessa de uma fábrica, para saber se têm bom sabor. Não seria impossível, mas não ficariam bolachas para vender.

Nota: A amostra deverá ser representativa da população, caso contrário, não poderemos tirar conclusões fiáveis. Quando uma amostra não é representativa, dizemos que é **enviesada**.

Exemplo : Para sabermos a média das alturas dos alunos de uma escola, se apenas analisássemos alunos que jogam basquetebol.

Métodos de amostragem probabilística- Qualquer elemento da população tem alguma probabilidade de fazer parte da amostra.

Exemplo: Para saber a opinião dos cidadãos acerca do partido que tenciona votar, escolher uma amostra a partir da lista de todas as pessoas que estão registadas nos cadernos eleitorais. Ir à procura das pessoas escolhidas.

Métodos de amostragem não probabilística- alguns elementos da população podem não ter possibilidade de ser selecionados para a amostra.

Exemplo: Para saber a opinião dos cidadãos acerca do partido que tenciona votar, escolher uma amostra a partir dos leitores de um determinado jornal online. Muitos eleitores não consultam esse site, e alguns nem têm internet.

Métodos de amostragem (probabilística).

1- **Amostragem <u>aleatória simples</u> de n elementos**.

Exemplo
Se tivermos uma população de 1000 pessoas e quisermos uma amostra de 50 elementos, podemos escrever os nomes das 1000 pessoas em papelinhos. Misturamos bem e retiramos 50 aleatoriamente. Fazemos esta escolha <u>sem reposição</u>!...

Alternativamente, podemos numerar as pessoas de 1 a 1000 e pedir numa calculadora ou num computador uma amostra de 50 número aleatoriamente e escolher as pessoas com esses números. (Função a explorar: *Random* ou #*Rand*)

CG: Gerar aleatoriamente 50 números inteiros entre 1 e 1000
O comando fundamental é **"random".**

<u>Casio:</u>
Tecla **OPTN/** PROB/RAND/Int.
RanInt#(1,1000,50)

<u>Texas:</u>
Tecla **MATH**/ PROB/randInt (inteiro)
randint(1, 1000, 50) ou introduzir por esta ordem…

2- **Amostragem <u>aleatória de n elementos com reposição</u>**.

Se tivermos uma população de 1000 pessoas e quisermos uma amostra de 50 elementos, podemos escrever os nomes das 1000 pessoas em papelinhos. Misturamos bem e retiramos 50 aleatoriamente. Fazemos esta escolha <u>com reposição</u>!...

3- **Amostragem <u>aleatória sistemática</u>-** criamos uma regra para extrair os números.

Exemplo
Se tivermos uma população de 1000 pessoas e quisermos uma amostra de 50 elementos, podemos começar por numerar as pessoas de 1 a 1000. Depois escolhemos ao acaso um dos números entre 1 e 1000. A partir daí, saltamos de 20 em 20.

Por exemplo, se sair o número 900, escolhemos: 900; 920; 940; 960; 980; 1000; 20; 40; ... até obter 50 números. repare que o primeiro número foi aleatório, mas os seguintes seguiram uma regra pré definida, mas garantindo que iriam percorrer toda a amplitude de valores. Repare que 1000/50=20.

Podíamos escolher outra regra para saltar entre os números.

4- **Amostragem aleatória estratificada**- Escolhemos a amostra respeitando alguns estratos da população que acreditamos influenciar as respostas ao inquérito.

Nota: Esta é a amostragem mais utilizada em estudos oficiais tais como eleições, onde temos em conta a o género (m/f), a região, a idade, o nível de escolaridade, o nível socioeconómico, etc… estas variantes costumam influenciar as escolhas dos partidos políticos… (sugestão: quando aceder a uma sondagem, consulte a ficha técnica.)

Exemplo
Numa sondagem para umas eleições em Portugal, podemos escolher uma amostra com 5000 eleitores de entre os 10 milhões que estão inscritos em Portuga.
Como acreditamos que o género(masculino/feminino) tem influência na escolha do partido em que vota, podemos consultar o censos e procurar o percentagem de homens e de mulheres na população portuguesa. Supondo que existiam 52 % de mulheres e 48 % de homens, então na nossa amostra de 5000 pessoas devem constar $5000 \times 0.52 = 2600$ mulheres e 2400 homens.

Do mesmo modo, devemos analisar a percentagem de eleitores que vivem em cada região do nosso país, e escolher uma amostra que respeite essas proporções.

Também é frequente termos o cuidado de considerar a faixa etárias, isto é, escolher uma amostra representativa que apresente a mesma proporção de indivíduos de cada faixa etária, com a verificada na população.

5- Amostragem aleatória por conglomerados. Quando analisamos grupos de indivíduos que correspondam ao modo como se agrupam naturalmente na população em que estão inseridos.

Exemplo: Quando estudamos a vida escolar dos estudantes, é costume estudar turmas inteiras. Isso permite compreender as interações. Ao estudar peixes, estudamos cardumes. Ao estudar lobos, estudamos alcateias, etc…

Nota importante: mesmo respeitando os métodos de amostragem, a generalização dos resultados à população tem sempre algum erro associado.

Parâmetro e estatística. Estimativa pontual

O **parâmetro** é referente à população. A estatítica ou **estimador** ou **estimativa** é referente à amostra.

Símbolos a utilizar:

Dimensão da população **N** / Tamanho da amostra: ***n***

Valor médio populacional: $\boldsymbol{\mu}$ / média amostral: $\overline{X}$ ou $\overline{x}$

Proporção populacional: ***p*** / Proporção amostral: $\widehat{p}$

Desvio padrão populacional: $\boldsymbol{\sigma}$ / Desvio padrão amostral: ***S***

Nota: por vezes, usamos um acento circunflexo (**^**) para indicar que se refere à amostra.

*Um **Parâmetro,** $\boldsymbol{\theta}$* carateriza a população. Uma estatística ou estimador $\widehat{\boldsymbol{\theta}}$ carateriza a amostra.

Nota: **Parâmtro, estimador, estimativa.**

No caso da média, temos, como vimos μ **Parâmetro** (populacional)

$\overline{X}$ **Estimador**(fórmula ou processo para estimar). O estimador é uma variável aleatória, pois os seus valores variam de amostra para amostra.

$\bar{x}$ **Estimativa**- resultado concreto de uma amostra particular.

Estimativa pontual é o valor numérica assumido pelo estimador, para a amostra selecionada.

Exemplo

Considerando a nossa população como os 1200 alunos da Escola A, e admitindo que pretendíamos estudar as alturas de todos os alunos, suponha que foi possível medir todos e analisar os dados obtidos.

*Para a dimensão da nossa população, indicamos **N**=1200.*

Admitindo que a média das alturas foi 168 cm, com um desvio padrão de 11 cm, indicamos **μ**=168 e **σ**= 11. (Valores populacionais ou parâmetros da população)

Exemplo
Considerando novamente a nossa população como os 1200 alunos da Escola A, e admitindo que pretendíamos estudar as alturas dos alunos, suponha que não foi possível medir todos, e que recorremos a uma amostra com 50 elementos.
*Para o tamanho da nossa amostra, indicamos **n=50.***

Admita que a média das alturas da nossa amostra foi 167 cm, com um desvio padrão amostral de 11 cm, indicamos $\bar{x} = 167$ e **s**= 10. (valores amostrais).

Neste caso o $\bar{x} = 167$ é uma estimativa, pois é um valor concreto.

O estimador ($\bar{X}$)é a fórmula usada para o calcular, isto é, a soma de 50 elementos, a dividir por 50:

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_{50}}{50}$$

Exemplo
Considerando a nossa população como os 1200 alunos da Escola A, e admitindo que queríamos saber qual é a proporção de raparigas da escola. Suponha que foi possível consultar os dados de todos os alunos e que, ao todo eram 750 raparigas.
A proporção de raparigas é $\boldsymbol{P} = \frac{750}{1200} = 0.625$ ou 62.5%. (proporção amostral).

Exemplo
Considerando a nossa população como os 1200 alunos da Escola A, e admitindo que queríamos saber qual é a proporção de raparigas da escola. Suponha que não foi possível consultar os dados de todos os alunos. Escolhemos uma amostra com 60 alunos e analisámos, tendo obtido 36 raparigas.
A proporção amostral de raparigas é $\widehat{\boldsymbol{p}} = \frac{36}{60} = 0.6$ ou 60%.

Distribuição de amostragem de um estimador.

Nota: Em exemplos anteriores, se recolhêssemos outras amostras com a mesma dimensão, muito provavelmente iríamos obter diferentes valores para a proporção amostral. No entanto, e por desconhecermos o valor exato do parâmetro, é impossível dizer qual dessas estimativas é a melhor.

Um **estimador** é uma variável aleatória.

A **distribuição de amostragem de um estimador** para estimar determinado parâmetro é a distribuição de todos os valores que ele pode tomar na estimação desse parâmetro para todas as amostras possíveis de igual dimensão.

No tópico seguinte, vamos concretizar esta variabilidade de estimativas para o caso de um valor médio.

Estimação do valor médio.

Exemplo(Extra)

Para melhor compreender a distribuição de amostragem referida no tópico anterior, consideremos uma **população** só com 3 elementos: **{1; 2; 3}.**

Para esta população, vamos agora obter todas as **amostras** com 2 elementos**(n=2)**, com reposição:

{1, 1}; {1, 2; {1, 3}; {2, 1}; {2, 2}; {2, 3}; {3, 1}; {3, 2}; {3, 3}.

Ao todo são 9 amostras possíveis. Repare que são 3×3=9, ou 3^2=9.

Questão: como relacionar o valor médio populacional com as médias amostrais?

Reparemos que a média populacional é dada por $\mu = \frac{1+2+3}{3} = \frac{6}{3} = 2$

Valor médio populacional: $\boldsymbol{\mu = 2}$

Calculemos as médias de cada uma das amostras, isto é, as médias amostrais:

	{1, 1}	{1, 2}	{1, 3}	{2, 1}	{2, 2}	{2, 3}	{3, 1}	{3, 2}	{3, 3}
$\bar{x}$	1	1.5	2	1.5	2	2.5	2	2.5	3

A média de todas as médias amostrais é dada por

$$E(\bar{X}) = \frac{1 + 1.5 + 2 + 1.5 + 2 + 2.5 + 2 + 2.5 + 3}{9} = \frac{18}{9} = 2$$

Repare que $\boldsymbol{E(\bar{X}) = \mu}$ a média das médias amostrais é igual à média populacional.

Questão: como relacionar o desvio padrão populacional, com o desvio padrão das médias amostrais?

O desvio padrão populacional é dado por

$$\sigma = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}} = \sqrt{\frac{1+0+1}{3}} = \sqrt{\frac{2}{3}} \approx 0.816$$

Nota: pode obter este mesmo valor pelo desvio padrão populacional da calculadora.

Se tentar fazer o desvio padrão populacional de todas as médias amostrais, obtemos:

(pode usar a calculadora- indique uma coluna com os valores das médias amostrais obtidas acima:

1	1.5	2	1.5	2	2.5	2	2.5	3

Obtém um desvio padrão populacional de $\sigma_{\bar{X}}$ ≈0.577, que não é igual a 0.816.

No entanto, se dividir 0.816 pela raiz quadrada de 2, obtém $\frac{0.816}{\sqrt{2}} \approx 0.577$.

Neste caso, escolhemos dividir por $\sqrt{2}$ porque 2 é a dimensão das amostras.

De um modo geral,

$$\boldsymbol{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

onde ***n*** é a dimensão da amostra.

Nota: Depois de analisado o exemplo 10, podemos tirar as seguintes conclusões:

$$\boldsymbol{E(\bar{X}) = \mu}$$

Dizemos que $\bar{X}$ é um **estimador centrado** ou cêntrico ou não enviesado, pois o seu valor médio é igual ao parâmetro que pretendemos estimar.

O desvio padrão da distribuição de amostragem da média, isto é, o desvio padrão das médias amostrais pode ser dado por

$\boldsymbol{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$ também se designa **erro padrão.**

Nota: Tal como se pode deduzir a partir da expressão anterior, o aumento da dimensão da amostra, a partir da qual é calculado o estimador, faz diminuir a variabilidade das estimativas. Isto é, se n aumentar, o valor de $\boldsymbol{\sigma_{\bar{X}}}$ diminui e, como tal as médias amostrais são mais próximas umas das outras.

Teorema do Limite Central.

Seja X uma população com valor médio μ e desvio padrão σ, da qual se recolhem amostras de dimensão n.

Então, se $n \geq 30$, a distribuição de amostragem da média $\bar{X}$ pode ser aproximada a uma distribuição normal com valor médio μ e desvio padrão $\frac{\sigma}{\sqrt{n}}$,

isto é, para $\mathbf{n \geq 30}$

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Nota: este teorema diz-nos que as médias amostrais, apesar de variarem de amostra para amostra, tendem a concentrar-se em torno do valor médio da população à medida que a dimensão das amostras aumenta, uma vez que o desvio-padrão $\frac{\sigma}{\sqrt{n}}$ diminui.

Exemplo

Da produção de embalagens de um dado produto alimentar, sabemos que o peso médio é 150 gramas, com um desvio padrão de 10 gramas. Recolhe-se uma amostra com 50 elementos. Como será de esperar que seja a distribuição de amostragem da média?

Resposta:

Atendendo tratar-se de uma amostra com mais de 30 elementos, o teorema do limite central garante que:

-as médias amostrais têm distribuição normal, isto é, $\bar{X} \sim N$,

- o valor médio, ou "média das médias amostrais" será 150, pois $E(\bar{X}) = \mu = 150$

-O desvio padrão das "médias amostrais" $\boldsymbol{\sigma_{\bar{X}}} = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \frac{\mathbf{10}}{\sqrt{\mathbf{50}}} \approx 1.414$

Exemplo

A Mariana trabalhou numa agência de viagens, na qual analisava os gastos mensais de clientes. Em maio de 2017, a Mariana analisou os gastos em viagens dos clientes da agência e verificou que, nesse mês, os clientes gastaram, em média, 1200 euros, com um desvio padrão de ***a*** euros.

Nessas condições, para uma amostra de dimensão 324, o desvio padrão da distribuição de amostragem da média é bem aproximado pelo valor 6.

Qual é o valor de ***a***?

Apresente todos os cálculos e as justificações necessárias.

Resolução:

$\bar{x} = 1200 \quad \sigma = a \quad n = 324$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{a}{\sqrt{324}} = \frac{a}{18} \qquad \frac{a}{18} = 6 \Leftrightarrow a = 6 \times 18 \Leftrightarrow a = 108$