

UNIDADE

8

Inferência estatística e distribuição amostral

Objetivo

Nesta Unidade, você vai conhecer os conceitos de inferência estatística e de distribuição amostral, que são a base para o processo de generalização usado pelos administradores em várias tomadas de decisão.

Conceito de inferência estatística

Caro estudante, vamos relembrar um pouco nossa trajetória ao longo da disciplina de Estatística.

Na Unidade 1, vimos que, através da **inferência estatística**, usando os conceitos de probabilidade (e variáveis aleatórias, Unidades 5, 6 e 7), podemos generalizar os resultados de uma pesquisa por amostragem (Unidade 2) para a população da qual a amostra foi retirada.

Lembre-se, estamos supondo que a amostra foi retirada por meio de **amostragem probabilística ou aleatória**; temos, então, um **experimento aleatório**: não sabemos quem fará parte da amostra antes do sorteio (Unidade 5).

Uma vez retirada a amostra, fazemos Análise Exploratória dos Dados (Unidades 3 e 4): por exemplo, calculamos a média de uma variável quantitativa. Esta média e todas as demais estatísticas serão variáveis aleatórias (pois estão associadas ao **espaço amostral** de um experimento aleatório), e poderemos tentar identificar o modelo probabilístico mais apropriado para elas (Unidades 6 e 7). Mas, neste caso, o modelo probabilístico de uma estatística da amostra é chamado de **distribuição amostral**.

Conhecer a distribuição amostral das principais estatísticas vai nos ser muito útil quando estudarmos os tipos particulares de inferência estatística: estimação de parâmetros (Unidade 9) e testes de hipóteses (Unidade 10).

Vamos continuar aprendendo? É muito bom ter você conosco!

Estatística é a ciência que se ocupa de organizar, descrever, analisar e interpretar dados para que seja possível a tomada de decisões e/ou a validação científica de uma conclusão. Os dados são coletados para estudar uma ou mais características de uma população: conjunto

Na Unidade 2 enumeramos as principais razões para usar amostragem.

GLOSSÁRIO

***Amostra aleatória, casual ou probabilística** – amostra retirada por meio de um sorteio não viciado, que garante que cada elemento da população terá uma probabilidade maior do que zero de pertencer à amostra. Fonte: Barbetta (2006).

***Parâmetros** – característica numérica do modelo probabilístico da variável de interesse na população, tais como média, variância, proporção. Fonte: Barbetta, Reis e Bornia (2004).

***Estimação de parâmetros** – forma de inferência estatística que busca estimar os parâmetros do modelo probabilístico da variável de interesse na população, a partir de dados de uma amostra probabilística desta mesma população. Fonte: adaptado pelo autor de Barbetta, Reis e Bornia (2004).

das medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m).

Uma população pode ser representada através de um modelo probabilístico: este apresenta condições para uso, forma para a distribuição de probabilidades e parâmetros.

Os dados necessários para a obtenção do modelo podem ser obtidos através de um censo (pesquisa de toda a população) ou através de uma **amostra** (subconjunto finito) da população.

A amostra deve ser: representativa da população, suficiente (para que o resultado tenha confiabilidade) e aleatória (retirada por sorteio não viciado).

*A inferência estatística consiste em fazer afirmações probabilísticas sobre as características do modelo probabilístico, que se supõe representar uma população, a partir dos dados de uma **amostra aleatória (probabilística)** desta mesma população.*

Fazer uma afirmação probabilística sobre uma característica qualquer é associar à declaração feita uma probabilidade de que tal declaração esteja correta (e, portanto, a probabilidade complementar de que esteja errada). Quando se usa uma amostra da população, sempre haverá uma probabilidade de se estar cometendo um erro (justamente por ser usada uma amostra): a diferença entre os métodos estatísticos e os outros reside no fato de que os métodos estatísticos permitem calcular essa probabilidade de erro. E para que isso seja possível, a amostra da população precisa ser aleatória.

As afirmações probabilísticas sobre o modelo da população podem ser basicamente:

- estimar quais são os possíveis valores dos **parâmetros*** – **Estimação de parâmetros***;
- qual é o valor da média de uma variável que segue uma distribuição normal?;

- qual é o valor da proporção de um dos dois resultados possíveis de uma variável que segue uma distribuição binomial?;
- testar hipóteses sobre as características do modelo: parâmetros, forma da distribuição de probabilidades, entre outros – **Testes de hipóteses***;
- o valor da média de uma variável que segue uma distribuição é maior do que um certo valor?;
- o modelo probabilístico da população é uma distribuição normal?; e
- o valor da média de uma variável que segue uma distribuição normal em uma população é diferente da mesma média em outra população?

Estudaremos estimação de parâmetros na Unidade 9 e testes de hipóteses na Unidade 10.

Parâmetros e estatísticas

Vamos imaginar uma pesquisa como a da Unidade 1, opinião dos registrados no CRA-SC sobre os cursos em que se graduaram, desde que tenham se graduado em Santa Catarina. Naquela Unidade, e depois na Unidade 2, declaramos que era possível realizar uma amostragem probabilística, e vimos um exemplo de como fazer isso.

Independente da pesquisa, uma vez que tenha sido realizada por amostragem probabilística, os dados podem ser estatisticamente generalizados para a população.

Uma vez tendo coletado os dados, é preciso resumi-los e organizá-los de maneira a permitir uma primeira análise e posterior uso das informações. As técnicas estatísticas que se ocupam desses aspectos constituem a Análise Exploratória de Dados, que estudamos detalhadamente nas Unidades 3 e 4.

GLOSSÁRIO

*Testes de hipóteses

– forma de inferência estatística que busca testar hipóteses sobre características (parâmetros, forma do modelo) do modelo probabilístico da variável de interesse na população, a partir de dados de uma amostra probabilística desta mesma população. Fonte: adaptado pelo autor de Barbetta, Reis e Bornia (2004).

Esta última está relacionada aos percentuais de ocorrência dos valores em uma distribuição de frequências de uma variável qualitativa.

O conjunto de dados pode ser resumido (e apresentado) através das distribuições de frequências, que relacionam os valores que a variável pode assumir com a frequência (contagem) com que foram encontrados naquele conjunto. Esta distribuição pode ser apresentada na forma de uma tabela ou através de um gráfico (estes dois métodos podem ser usados tanto para variáveis qualitativas quanto para variáveis quantitativas).

Há uma terceira forma de resumir o conjunto de dados, quando a variável sob análise é quantitativa: as medidas de síntese ou **estatísticas***. As principais estatísticas são a média, o desvio-padrão, a variância e a proporção.

GLOSSÁRIO

***Estatísticas** – medidas de síntese da variável calculadas com base nos resultados de uma amostra da população. Se a amostra for probabilística (aleatória), as estatísticas podem ser consideradas variáveis aleatórias. Fonte: Barbetta, Reis e Bornia (2004).

Atenção, vamos lembrar o que cada uma dessas significa.

Média: *média aritmética simples (ver Unidade 4, seção 4.1.1). Trata-se de uma estatística que caracteriza o “centro de massa” do conjunto de dados (valor esperado – ver Unidade 6, seção 6.4). Quando é a média populacional, recebe o símbolo μ ; quando é a média amostral, recebe o símbolo \bar{x} .*

Variância: *trata-se de uma estatística (ver Unidade 4, seção 4.2.2) que mede a dispersão em torno da média do conjunto (em torno do valor esperado – Ver Unidade 6.4), possuindo uma unidade que é o quadrado da unidade da média (e dos valores do conjunto). Quando é a variância populacional, recebe o símbolo σ^2 , quando é a variância amostral, recebe o símbolo s^2 .*

Desvio-padrão *é a raiz quadrada positiva da variância, tendo, portanto, uma unidade que é igual à unidade da média, sendo muitas vezes preferida para efeito de mensuração da dispersão. Quando é o valor populacional recebe o símbolo σ ; e quando é o amostral, recebe o símbolo s .*

Proporção: *consiste em calcular a razão entre o número de ocorrências do valor de interesse de uma variável qualitativa e o número total de ocorrências registradas no conjunto (de todos os valores que a variável pode assumir); quando é uma proporção populacional, recebe o símbolo π ; quando é uma proporção amostral, recebe o símbolo p .*

Os valores das medidas de síntese, além de resumirem o conjunto de dados, constituem uma indicação dos prováveis valores dos parâmetros. Assim, em estudos baseados em amostras, é comum utilizar tais medidas de síntese como estatísticas que serão utilizadas para estimar os parâmetros do modelo probabilístico que descreve a população.

A Tabela 5 resume os parâmetros e as estatísticas.

Tabela 5: Parâmetros e estatísticas mais comuns

Medidas de síntese	Parâmetros (População)	Estatísticas (Amostra)
Média	$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$
Variância	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$
Proporção	$\pi = \frac{f_a}{N}$	$p = \frac{f_a}{n}$

Fonte: elaborada pelo autor

Onde N é o número de elementos da população, n é o número de elementos da amostra, e f_a é a frequência de ocorrência de um dos valores de uma variável qualitativa na população ou na amostra.

As estatísticas são variáveis aleatórias, pois seus valores podem variar dependendo do resultado da amostra. Se forem variáveis aleatórias, podem ser caracterizadas através de algum modelo probabilístico. Este modelo recebe o nome de distribuição amostral.

Distribuição amostral

Seja uma população qualquer com um parâmetro θ de interesse, correspondendo a uma estatística T em uma amostra. Amostras aleatórias são retiradas da população; e, para cada amostra, calcula-se o valor t da estatística T .

NÃO confundir com o t da distribuição t de Student, seção 7.2.4, Unidade 7.

Os valores de t formam uma nova população que segue uma distribuição de probabilidades, que é chamada de **distribuição amostral** de T .

Vamos ver um exemplo.

Exemplo 1: seja a população abaixo, constituída pelos pesos em kg de oito pessoas adultas:

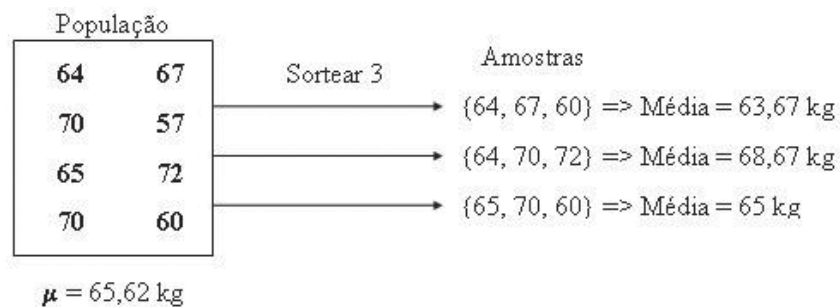


Figura 75: Distribuição amostral – Exemplo 1

Fonte: elaborada pelo autor

Observe que foram retiradas três amostras. Para cada amostra, foi calculada a média, visando a estimar a média populacional, que vale 65,62 kg. Observe que há uma variação na estatística média, pois o processo de amostragem é aleatório: é um experimento aleatório. Esta variação precisa ser considerada quando são realizadas as inferências sobre os parâmetros.

Assim sendo, o conhecimento das distribuições amostrais das principais estatísticas é necessário para fazer inferências sobre os parâmetros do modelo probabilístico da população. Por ora, basta co-

nhecer as distribuições amostrais das estatísticas médias de uma variável quantitativa qualquer e a proporção de um dos dois únicos resultados de uma variável qualitativa.

Distribuição amostral da média

Vamos observar as particularidades da distribuição amostral da média.

Neste segundo exemplo, suponha uma variável quantitativa cujos valores constituem uma população com os seguintes valores: **(2, 3, 4, 5)**.

Para esta população, que tem uma distribuição uniforme, podemos observar que os parâmetros são: $\mu = 3,5$ $\sigma^2 = 1,25$ (usou-se **n** no denominador por ser uma população).

Se retirarmos todas as amostras aleatórias de dois elementos (com reposição) possíveis desta população (**n = 2**), teremos os seguintes resultados:

Há 16 amostras possíveis.

(2, 2)	(2, 3)	(2, 4)	(2, 5)
(3, 2)	(3, 3)	(3, 4)	(3, 5)
(4, 2)	(4, 3)	(4, 4)	(4, 5)
(5, 2)	(5, 3)	(5, 4)	(5, 5)

O cálculo das médias de todas as amostras acima resultará na matriz abaixo:

$$\bar{X} \begin{Bmatrix} (2,0) & (2,5) & (3,0) & (3,5) \\ (2,5) & (3,0) & (3,5) & (4,0) \\ (3,0) & (3,5) & (4,0) & (4,5) \\ (3,5) & (4,0) & (4,5) & (5,0) \end{Bmatrix}$$

Se estas médias forem plotadas em um histograma (Figura 76):

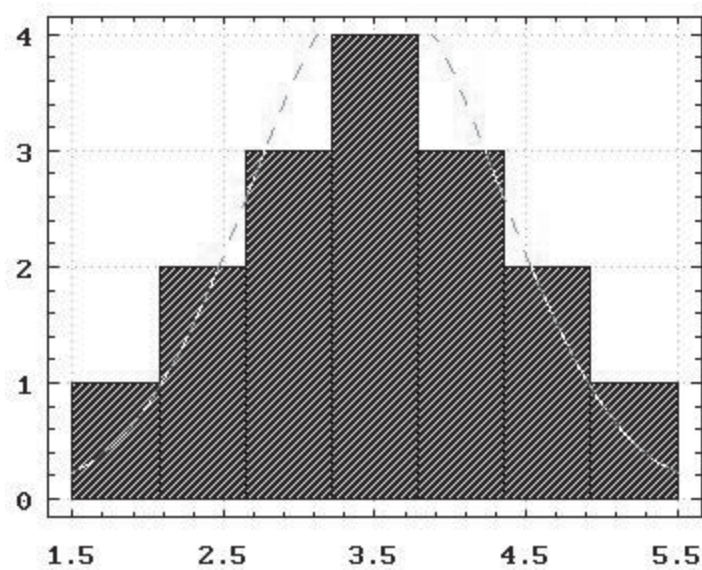


Figura 76: Histograma de médias amostrais

Fonte: adaptada pelo autor de Statsoft®

Se forem calculadas a média e a variância das médias de todas as amostras, o resultado será:

$$\bar{X} = 56/16 = 3,5 = \mu \quad V(\bar{x}) = 0,625 = \frac{1,25}{2} = \frac{\sigma^2}{n}$$

Observe como a distribuição das médias amostrais da variável pode ser aproximada por um modelo normal (não obstante a distribuição da variável na população não ser normal) e que o valor esperado das médias amostrais (média das médias) é igual ao valor da média populacional da variável, e a variância das médias amostrais é igual ao valor da variância populacional da variável dividida pelo tamanho da amostra. Quanto maior o tamanho da amostra (quanto maior **n**), mais o histograma acima vai se aproximar de um modelo normal, independentemente do formato da distribuição da variável na população.

Vamos ver outro exemplo.

Na Figura 77, temos a distribuição populacional de uma variável quantitativa qualquer de interesse. Ela apresenta média populacional (μ) igual a **416,99**, e variância populacional (σ^2) igual a **89554,51264**.

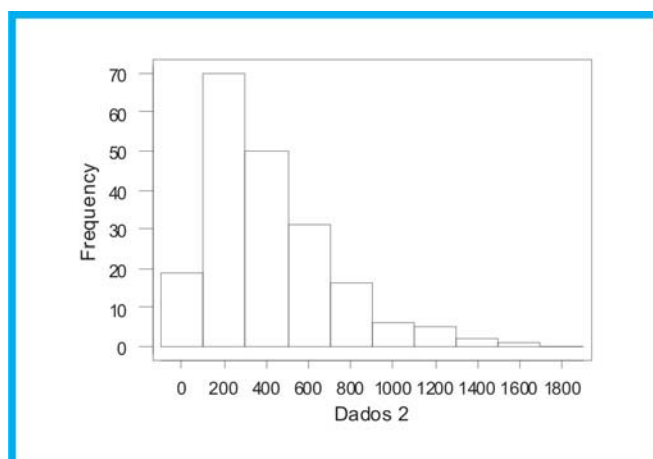


Figura 77: Distribuição populacional de uma variável quantitativa
Fonte: adaptada pelo autor de Minitab®

Observe que a distribuição é assimétrica, ou seja, não é normal! Vamos imaginar que seja possível retirar várias amostras aleatórias (com reposição) desta população, medir os valores da variável e calcular a média da variável em cada amostra. Posteriormente, construiremos um histograma das médias das amostras, e calcularemos a média das médias e a variância das médias.

Vamos começar com **40 amostras** aleatórias de dois elementos cada. Veja a Figura 78.

A retirada das amostras foi efetuada através do pacote estatístico Minitab®.

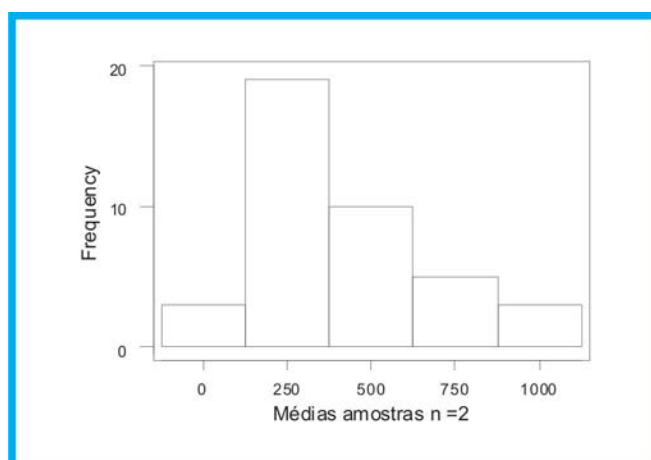


Figura 78: Distribuição amostral da média ($n = 2$)
Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 423,8875, e a variância das médias amostrais vale 67528,98666. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: $\sigma^2/n = 89554,51264/2 = 44777,25632$.

Observando o histograma, vemos que a distribuição das médias, para amostras de dois elementos, continua assimétrica, e o valor da média das médias amostrais (423,8875) não está muito próximo da média populacional (416,99), bem como a variância das médias amostrais (67528,98666), distante de $2/n = 44777,25632$.

Obviamente, o tamanho da amostra utilizada (dois elementos) ainda não foi grande o bastante para levar aos resultados obtidos no Exemplo 2 (provavelmente porque a distribuição da população é assimétrica). Vamos agora ver os resultados obtidos para 40 amostras aleatórias de quatro elementos cada. O histograma das médias está na Figura 79.

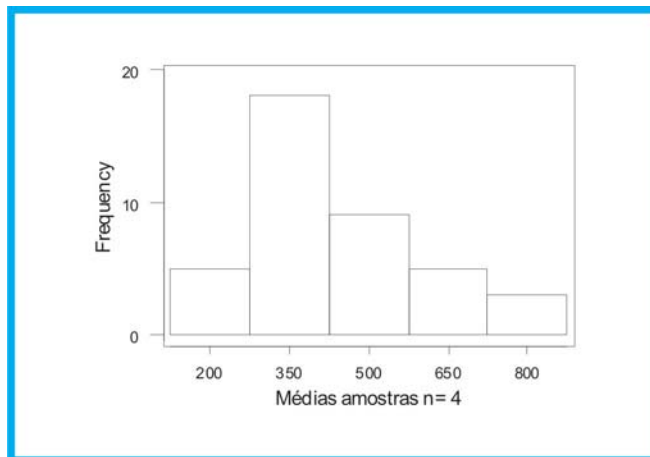


Figura 79: Distribuição amostral da média (n = 4)
Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 444,5375, e a variância das médias amostrais vale 26464,3269. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: $2/n = 89554,51264/4 = 22388,62816$.

Observando o histograma, vemos que a distribuição das médias, para amostras de quatro elementos, continua assimétrica, e o valor da média das médias amostrais (444,5375) não está muito próximo da

média populacional (416,99), mas a variância das médias amostrais (26464,3269) aproxima-se mais de $\sigma^2/n = 22388,62816$.

Novamente, o tamanho da amostra utilizada (quatro elementos) ainda não foi o bastante para levar aos resultados obtidos no Exemplo 2. Vamos agora ver os resultados obtidos para 40 amostras aleatórias de 16 elementos cada. O histograma das médias está na Figura 80.

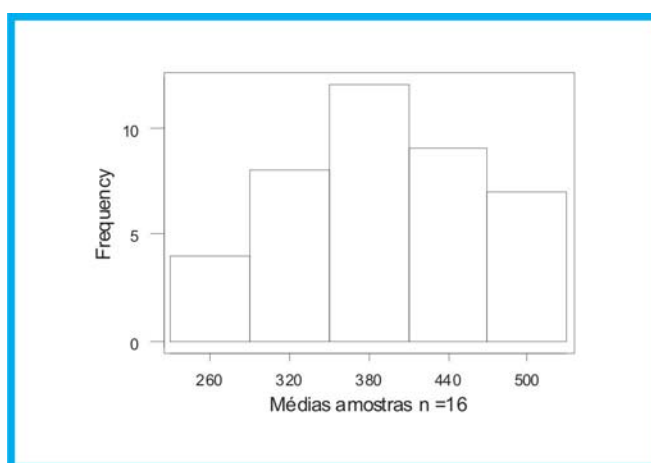


Figura 80: Distribuição amostral da média (n = 16)

Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 394,4922, e a variância das médias amostrais vale 5568,3945. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: $\sigma^2/n = 89554,51264/16 = 5597,1577$.

Observando o histograma, vemos que a distribuição das médias, para amostras de 16 elementos, está mais próxima da simetria, o valor da média das médias amostrais (394,4922) está mais próximo da média populacional (416,99), e a variância das médias amostrais (5568,3945) aproxima-se bastante de $\sigma^2/n = 5597,1577$.

Estamos muito próximos de obter um comportamento simétrico e aproximadamente normal para o histograma das médias amostrais. Se retirarmos mais 40 amostras, mas agora com 30 elementos em cada, o resultado poderá ser visto na Figura 81.

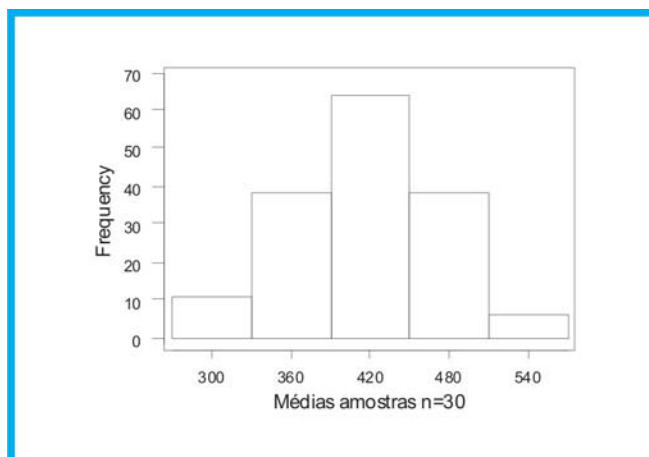


Figura 81: Distribuição amostral da média (n = 16)

Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 421,9217, e a variância das médias amostrais vale 2945,1326. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: $\sigma^2/n = 89554,51264/30 = 2985,1508$.

Observando o histograma, vemos que a distribuição das médias, para amostras de 30 elementos, é virtualmente normal, o valor da média das médias amostrais (421,9217) está bem próximo da média populacional (416,99), e a variância das médias amostrais (2945,1326) também é muito próxima de $\sigma^2/n = 2985,1508$.

Podemos, então, enunciar os teoremas:

Teorema das Combinações Lineares

Se a variável de interesse segue uma distribuição normal na população, a distribuição amostral das médias de amostras aleatórias retiradas desta população também será normal, independentemente do tamanho destas amostras.

Teorema Central do Limite

Se a variável de interesse não segue uma distribuição normal na população (ou não se sabe qual é a sua distribuição), a distribuição

amostral das médias de amostras aleatórias retiradas desta população será normal se o tamanho destas amostras for **suficientemente grande**, com uma média igual à média populacional e uma variância igual à variância populacional dividida pelo tamanho da amostra.

Para o caso da Proporção, podemos chegar a uma conclusão semelhante.

Distribuição amostral da proporção

Vamos estudar as particularidades da distribuição amostral da proporção através de um exemplo.

Neste Exemplo 4, pense agora em uma variável qualitativa que pode assumir apenas dois valores e que constitui a seguinte população: (□, □, □, □, ■)

Vamos supor que há interesse no valor † (este valor seria o nosso “sucesso”). A proporção deste valor na população (o valor do parâmetro) será $\pi = 1/5$.

Se retirarmos todas as amostras aleatórias de dois elementos (com reposição) possíveis desta população ($n = 2$), teremos os seguintes resultados:

(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, ■)	(□, ■)	(□, ■)	(□, ■)	(■, ■)

Figura 82: Amostras de tamanho 2 para proporção

Fonte: elaborada pelo autor

Observe que, se definirmos a variável como o número de “sucessos” (número de †), esta seguirá um modelo binomial: há apenas

Este “suficientemente grande” varia de distribuição para distribuição; como foi visto, uma distribuição uniforme precisa de uma amostra pequena ($n = 2$ no caso) para que a aproximação seja possível, outras distribuições precisam de amostras maiores. Alguns autores costumam chamar de “grandes amostras” aquelas que possuem mais de 30 elementos; a partir deste tamanho, a aproximação poderia ser feita sem maiores preocupações. Há 25 amostras possíveis.

dois resultados possíveis para cada realização, há um número limitado de realizações ($n = 2$ no caso), e cada realização independe da outra (porque a amostra é aleatória com reposição).

Calculando a proporção de π em cada uma das amostras, e chamando esta proporção amostral de p , teremos os seguintes resultados:

$$p = \begin{matrix} (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (1/2) & (1/2) & (1/2) & (1/2) & (1) \end{matrix}$$

Calculando a média (valor esperado) e a variância das proporções acima, teremos:

$$\bar{X} = E(p) = \frac{1}{5} = \pi \quad s^2 = 0,08 = \frac{\left(\frac{1}{5}\right) \times \left(1 - \frac{1}{5}\right)}{2} = \frac{\pi \times (1 - \pi)}{n}$$

Observe que o valor esperado (média) das proporções amostrais é igual ao valor da proporção populacional de π , e que a variância das proporções amostrais é igual ao produto da proporção populacional de π por seu complementar, dividido pelo tamanho da amostra.

Lembre-se de que um modelo binomial pode ser aproximado por um modelo normal se algumas condições forem satisfeitas: se o produto do número de realizações pela probabilidade de “sucesso” ($n \times p$) e o produto do número de realizações pela probabilidade de “fracasso” ($n \times [1 - p]$) forem ambos maiores ou iguais a 5. E esta distribuição normal teria média igual a $n \times p$ e variância igual a $n \times p \times (1 - p)$. Se estivermos interessados apenas na proporção (probabilidade de “sucesso”), e não no número de “sucessos”, as expressões anteriores podem ser divididas por n (o tamanho da amostra): média = p e variância = $[p \times (1 - p) / n]$.

Voltaremos a analisar o significado deste resultado quando estudarmos estimação por ponto.

Isto também é decorrência do Teorema Central do Limite.

Por causa do Teorema Central do Limite é que o modelo normal é tão importante. É claro que ele representa muito bem uma grande variedade de fenômenos, mas é devido à sua utilização em inferência estatística que o seu estudo é imprescindível. Ressalte-se, porém, que a sua aplicação costuma se resumir ao que se chama de inferência paramétrica, inferências sobre os parâmetros dos modelos probabilísticos que descrevem as variáveis na população. Para fazer inferências sobre outros aspectos que não os parâmetros, ou quando as amostras utilizadas não forem suficientemente grandes para assumir a validade do Teorema Central do Limite, é preciso usar técnicas de inferência não paramétrica (que nós não veremos nesta disciplina).

Saiba mais...

■ Sobre distribuição amostral:

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 7.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 7.

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. *Estatística Aplicada à Administração e Economia*. 2. ed. São Paulo: Thomson Learning, 2007, capítulo 7.

■ Sobre a utilização do Microsoft Excel® para estudar distribuições amostrais, veja:

LEVINE, D. M.; et al. *Estatística: teoria e aplicações – usando Microsoft Excel em português*. 5. ed. Rio de Janeiro: LTC, 2006, capítulo 5.

RESUMO

O resumo desta Unidade está mostrado na Figura 83:

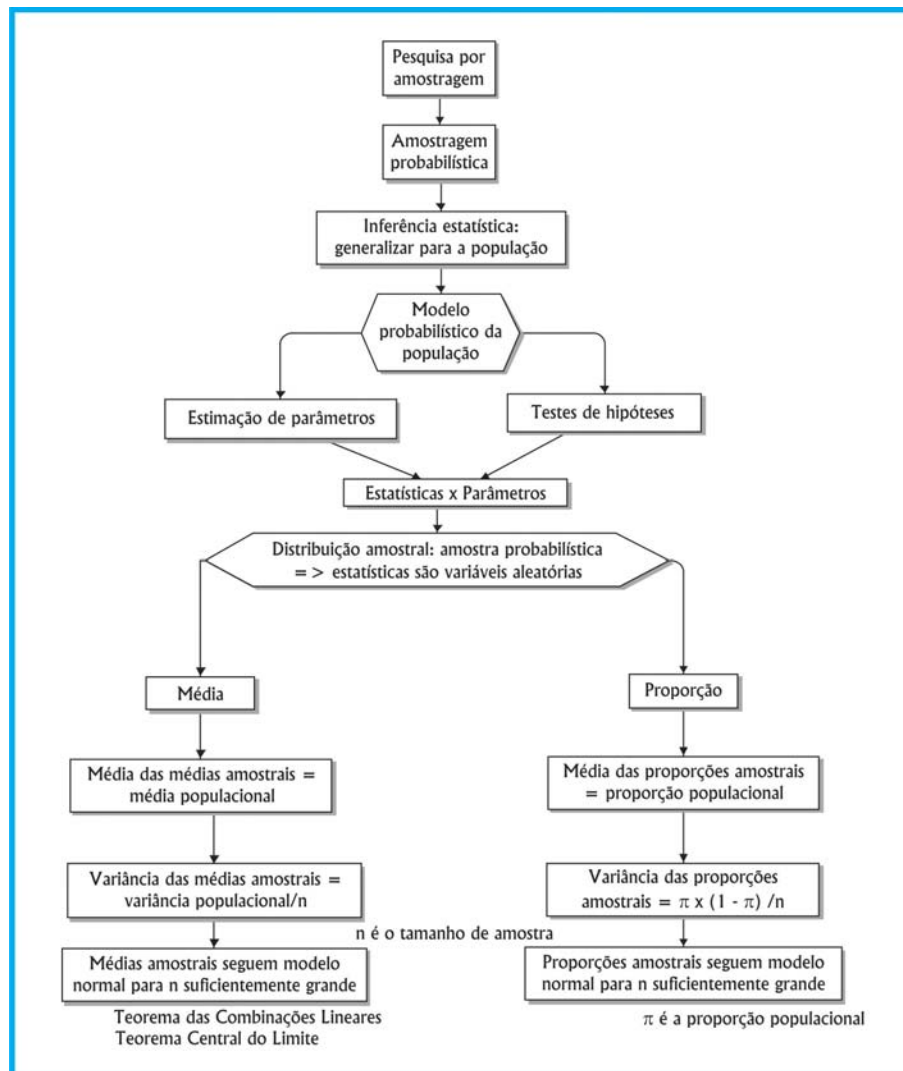


Figura 83: Resumo da Unidade 8

Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

Esta Unidade foi muito importante para o seu aprendizado, pois lhe dará base para chegar à inferência estatística propriamente dita, assunto que será tema de discussão nas Unidades 9 e 10. Vimos até agora sobre a inferência estatística e distribuição amostral, seu modelo probabilístico e testes de hipóteses. Chegamos ao final desta Unidade e à continuidade da aprendizagem proposta desde o início deste material. Interaja com seus colegas, responda as atividades de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem, espaço este que contemplará suas possíveis dúvidas. Procure seu tutor e solicite todas as informações necessárias para o seu aprendizado. Bons estudos!