By @kakashi_copiador

DISTRIBUIÇÃO AMOSTRAL

Nesta seção, estudaremos a distribuição de probabilidade dos principais estimadores, chamada de **Distribuição Amostral**. Vale ressaltar que os **estimadores** são **variáveis aleatórias**.

Inicialmente, cabe pontuar que a **distribuição** de uma **amostra aleatória** qualquer segue a **mesma distribuição populacional**.

Para entender melhor esse conceito, vamos considerar uma moeda com 2 faces, que vamos denominar face 0 e face 1. Tratando-se de uma moeda equilibrada, a probabilidade de cada face é de $\frac{1}{2} = 0{,}5$ e a esperança e variância são, respectivamente:

$$\mu = E(X) = \sum x.P(X = x)$$

$$E(X) = 0 \times 0{,}5 + 1 \times 0{,}5 = 0{,}5$$

$$V(X) = \sum (x - \mu)^2.P(X = x)$$

$$V(X) = (0 - 0{,}5)^2 \times \frac{1}{2} + (1 - 0{,}5)^2 \times \frac{1}{2} = \frac{0{,}25 + 0{,}25}{2} = 0{,}25$$

Suponha que vamos extrair uma amostra aleatórias de tamanho 3, ou seja, vamos lançar a moeda 3 vezes. Podemos representar essa amostra por X_1, X_2, X_3.

Considerando que os possíveis resultados das amostras são os mesmos da população (0 ou 1) e com as mesmas probabilidades de $\frac{1}{2} = 0{,}5$ para cada face, então temos:

$$E(X_1) = E(X_2) = E(X_3) = 0 \times 0{,}5 + 1 \times 0{,}5 = 0{,}5$$

$$V(X_1) = V(X_2) = V(X_3) = (0 - 0{,}5)^2 \times \frac{1}{2} + (1 - 0{,}5)^2 \times \frac{1}{2} = \frac{0{,}25 + 0{,}25}{2} = 0{,}25$$

Ou seja, a esperança e a variância de cada amostra são **iguais** às da população:

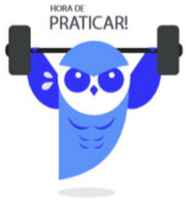
$$E(X_i) = E(X) = \mu$$

$$V(X_i) = V(X) = \sigma^2$$

Mais precisamente, **toda** a **distribuição de probabilidade da amostra** é **igual** à **distribuição de probabilidade da população** (as esperanças e as variâncias são iguais como consequência desse fato).

No exemplo do lançamento da moeda, a **população** é **infinita**, pois podemos lançar a moeda infinitas vezes. Quando a população é **infinita** (ou muito grande em comparação com o tamanho da amostra) **ou** quando a **amostra** é extraída **com reposição** dos elementos selecionados, então as amostras são **independentes**.

Nesses casos, dizemos que as variáveis que representam as amostras, $X_1, X_2, X_3, \ldots$, são **independentes e identicamente distribuídas** (**i.i.d.**), isto é, são independentes e apresentam a mesma distribuição.



(FGV/2021 – FunSaúde/CE) Se X_1, X_2, ... X_n é uma amostra aleatória simples de uma determinada distribuição de probabilidades f(x), avalie se as afirmativas a seguir estão corretas.

I. X_1, X_2, ... X_n são independentes.

II. X_1, X_2, ... X_n são identicamente distribuídos.

III. Nem sempre cada X_i, i = 1,..., n, tem distribuição f(x).

Está correto o que se afirma em

a) I, apenas.

b) I e II, apenas.

c) I e III, apenas.

d) II e III, apenas.

e) I, II e III.

Comentários:

Essa questão trabalha com conceitos de distribuição amostral. A base da distribuição amostral é que os elementos de uma amostra (que o enunciado denotou por X_1, X_2, ... X_n) são variáveis aleatórias independentes que apresentam a mesma distribuição da população (consequentemente, apresentam a mesma média e a mesma variância).

Assim, as afirmativas I e II estão corretas, enquanto a afirmativa III está incorreta, pois cada variável X_i apresenta a mesma distribuição f(x) da população (sempre).

Gabarito: B

Agora, estudaremos a distribuição dos estimadores mais utilizados, quais sejam, a média, a proporção e a variância amostrais.

Distribuição Amostral da Média

Para estimarmos a **média da população μ**, utilizamos como **estimador** a **média amostral $\bar{X}$**. Sendo $X_1, X_2, \dots, X_n$ os valores observados da amostra, a média amostral é a razão entre a soma dos valores observados, $X_1 + X_2 + \cdots + X_n$, e o número de elementos observados, n:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Assim como os demais estimadores, a média amostral é uma **variável aleatória**, uma vez que $\bar{X}$ **varia** de acordo com os valores observados da amostra $X_1, X_2, \dots, X_n$.

Vamos ao exemplo da moeda lançada 3 vezes. Se o resultado for {0, 0, 1}, a média amostral será:

$$\bar{X} = \frac{0 + 0 + 1}{3} = \frac{1}{3} \cong 0{,}33$$

Se o resultado for {0, 1, 1}, por exemplo, a média amostral será:

$$\bar{X} = \frac{0 + 1 + 1}{3} = \frac{2}{3} \cong 0{,}67$$

E qual seria a esperança desse estimador? Bem, sabendo que as faces possíveis são 0 e 1, cada uma com 50% de chance, esperamos que as médias desse experimento estejam em torno de 0,5.

Ou seja, a **esperança da média amostral** é igual à **média populacional**?

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

E quanto à variância? A **variância da média amostral** é dada por:

$$V(\bar{X}) = \frac{V(X)}{n} = \frac{\sigma^2}{n}$$

Para o exemplo dos 3 lançamentos da moeda, em que V(X) = 0,25 e n = 3, a variância da média amostral é:

$$V(\bar{X}) = \frac{0{,}25}{3} \cong 0{,}08$$

Note que a **variância da média amostral, $V(\bar{X})$**, é **menor** do que a **variância populacional, $V(X)$**. Além disso, quanto **maior o tamanho da amostra, n**, **menor será a variância da média amostral**.

Isso ocorre porque a média das observações, $\bar{X}$, costuma ser um valor bem mais próximo da média μ (ou seja, apresenta bem menos valores extremos) do que as observações X por si só. Isso significa que a **variância** da média amostral, $\bar{X}$, é **menor** do que a variância das observações, X. Lembrando que as observações da amostra seguem a mesma distribuição da população, então a **variância da média amostral**, $V(\bar{X})$, é **menor** do que a variância da população.

O **desvio padrão** (raiz quadrada da variância) de um estimador pode ser chamado de **erro padrão**.

> **O erro padrão (ou desvio padrão) da média amostral** é dado por:
>
> $$EP(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Ou seja, o erro padrão (ou devio padrão) da média amostral pode ser calculado como a **raiz quadrada da variância da média amostral,** $\sqrt{\boldsymbol{V(\bar{X})}}$, ou como a **razão entre o desvio padrão populacional e a raiz do número de elementos da amostra,** $\boldsymbol{\frac{\sigma}{\sqrt{n}}}$**.**

Também podemos denotar o erro padrão (ou desvio padrão) da média amostral por $\sigma_{\bar{X}}$. Para o nosso exemplo, o erro padrão da média amostral pode ser calculado como:

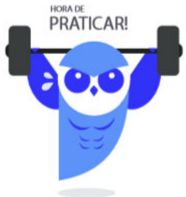
$$EP(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{\frac{0{,}25}{3}} \cong 0{,}29$$

INDO MAIS FUNDO!

> As fórmulas da esperança e variância podem ser obtidas a partir das respectivas propriedades. Sendo $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$, a esperança $E(\bar{X})$ é dada por:
>
> $$E(\bar{X}) = E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = E\left(\frac{X_1}{n}\right) + E\left(\frac{X_2}{n}\right) + \cdots + E\left(\frac{X_n}{n}\right)$$
>
> $$E(\bar{X}) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n)$$
>
> Vimos que a esperança amostral é igual à esperança populacional, $E(X_i) = E(X) = \mu$:
>
> $$\boldsymbol{E(\bar{X})} = \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = n \times \frac{1}{n} \times \mu = \boldsymbol{\mu}$$
>
> Considerando que as variáveis $X_1, X_2, \dots, X_n$ são **independentes**, a **variância** $V(\bar{X})$ é:
>
> $$V(\bar{X}) = V\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = V\left(\frac{X_1}{n}\right) + V\left(\frac{X_2}{n}\right) + \cdots + V\left(\frac{X_n}{n}\right)$$
>
> $$V(\bar{X}) = \frac{1}{n^2}V(X_1) + \frac{1}{n^2}V(X_2) + \cdots + \frac{1}{n^2}V(X_n)$$
>
> Sabendo que a variância amostral é igual à variância populacional, $V(X_i) = V(X) = \sigma^2$:
>
> $$\boldsymbol{V(\bar{X})} = \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = n \times \frac{1}{n^2} \times \sigma^2 = \boldsymbol{\frac{\sigma^2}{n}}$$

A **variância amostral** é **igual** à **variância populacional**, $V(X_i) = V(X) = \sigma^2$. O que é **diferente** é a **variância da média amostral**, $V(\bar{X}) = \frac{\sigma^2}{n}$.

Note que a **variância da média amostral** é **menor** do que a **variância populacional**. Isso faz sentido, certo? As médias sempre variam menos do que os valores individuais, pois os valores extremos acabam sendo compensados quando calculamos a média. Logo, **a distribuição da média amostral é menos dispersa do que a distribuição populacional**.



(CESPE/2016 – TCE/PA) Uma amostra aleatória, com n = 16 observações independentes e identicamente distribuídas (IID), foi obtida a partir de uma população infinita, com média e desvio padrão desconhecidos e distribuição normal. Tendo essa informação como referência inicial, julgue o seguinte item.

Para essa amostra aleatória simples, o valor esperado da média amostral é igual à média populacional.

Comentários:

Vimos que, de fato, o valor esperado da média amostral (para uma amostra aleatória) é igual à média populacional.

Gabarito: Certo.

(2019 – UEPA) Considere uma amostra aleatória $X_1, X_2, \dots, X_n$ de uma população normal de média μ e variância σ^2 = 9. Então, a média e a variância de $\overline{X_n} = \frac{X_1, X_2, \dots, X_n}{n}$, são, respectivamente,

a) μ e $\frac{3}{n}$.

b) $\frac{\mu}{n}$ e $\frac{9}{n}$.

c) μ e $\frac{9}{n}$.

d) μ e $\frac{n}{9}$.

Comentários:

Vimos que a média e a variância da média amostral são, respectivamente:

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

$$V(\bar{X}) = \frac{V(X)}{n} = \frac{9}{n}$$

Gabarito: C.

(VUNESP/2015 – TJ-SP) Resultados de uma pesquisa declaram que o desvio padrão da média amostral é 32. Sabendo que o desvio padrão populacional é 192, então o tamanho da amostra que foi utilizada no estudo foi

a) 6.

b) 25.

c) 36.

d) 49.

e) 70.

Comentários:

O desvio padrão (ou erro padrão) da média amostral pode ser calculado como a razão entre o desvio padrão da população e a raiz do número de elementos da amostra:

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

O enunciado informa que o desvio padrão da média amostral é $\sigma_{\bar{X}} = 32$ e o desvio padrão populacional é $\sigma = 192$. Logo:

$$32 = \frac{192}{\sqrt{n}}$$

$$\sqrt{n} = \frac{192}{32} = 6$$

$$n = (6)^2 = 36$$

Gabarito: C.

(FCC/2012 – TRE-SP) Uma variável aleatória U tem distribuição uniforme contínua no intervalo $[\boldsymbol{\alpha}, \boldsymbol{3\alpha}]$. Sabe-se que U tem média 12. Uma amostra aleatória simples de tamanho n, com reposição, é selecionada da distribuição de U e sabe-se que a variância da média dessa amostra é 0,1. Nessas condições, o valor de n é

a) 80.

b) 100.

c) 120.

d) 140.

e) 150.

Comentários:

O que essa questão exige a respeito da matéria que acabamos de estudar é a fórmula do desvio padrão da média amostral:

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

Assim, podemos escrever o tamanho amostral como:

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

$$n = \left(\frac{\sigma}{\sigma_{\bar{X}}}\right)^2 = \frac{\sigma^2}{\sigma_{\bar{X}}^2}$$

Ou seja, o tamanho amostral é a razão entre a variância populacional σ^2 (quadrado do desvio padrão populacional σ) e a variância da média amostral $\sigma_{\bar{X}}^2$ (quadrado do desvio padrão da média amostral $\sigma_{\bar{X}}$).

O enunciado informa que a variância da média amostral é $\sigma_{\bar{X}}^2 = 0{,}1$.

Pronto! A matéria desta aula acabou. Agora, para calcular a variância populacional, precisamos saber calcular a média e a variância da distribuição contínua uniforme.

A média (esperança) dessa distribuição é igual à média aritmética dos limites do intervalo, a e b. Sabendo que a = $\boldsymbol{\alpha}$, b = 3.$\boldsymbol{\alpha}$ e E(X) = 12, conforme dados do enunciado, temos:

$$E(X) = \frac{a+b}{2}$$

$$12 = \frac{\alpha + 3\alpha}{2} = \frac{4\alpha}{2} = 2\alpha$$

$$\alpha = 6$$

Ou seja, a = 6 e b = 3x6 = 18. Então, a variância é dada por:

$$V(X) = \frac{(b-a)^2}{12} = \frac{(18-6)^2}{12} = \frac{(12)^2}{12} = 12$$

Voltando à nossa fórmula para encontrar o tamanho amostral, temos:

$$n = \frac{\sigma^2}{\sigma_{\bar{X}}^2} = \frac{12}{0{,}1} = 120$$

Gabarito: C

Fator de Correção para População Finita

Os resultados da variância e do desvio padrão da média amostral são válidos para variáveis $X_1, X_2, \dots, X_n$ **independentes**, ou seja, quando a população é **infinita ou** quando as amostras são extraídas **com reposição**.

Quando isso não ocorre, ou seja, quando a **população é finita e** as amostras são extraídas **sem reposição**, precisamos fazer um **ajuste**. Sendo n o tamanho da amostra e N o tamanho da população, precisamos multiplicar a **variância** da média amostral, pelo **fator de correção de população finita** $\frac{\boldsymbol{N-n}}{\boldsymbol{N-1}}$.

$$V(\bar{X}_*) = \frac{\sigma^2}{n} \times \frac{\boldsymbol{N-n}}{\boldsymbol{N-1}}$$

Observe que o fator de correção é **menor do que 1**, pois $N - n < N - 1$, logo, o ajuste para populações finitas, cuja amostras são extraídas com reposição, **diminui** a variância da média amostral.

Distribuição da Média Amostral e a Curva Normal

Quando a **população** segue **distribuição normal (ou gaussiana)**, a **média amostral** também seguirá **distribuição normal**. Como vimos anteriormente, a esperança, a variância e o desvio padrão (chamado erro padrão) dessa distribuição serão:

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

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

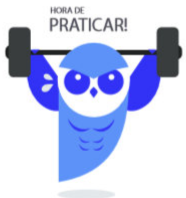
$$EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Ainda que a população **não** siga **distribuição normal**, pelo **Teorema Central do Limite**, é possível **aproximar** a distribuição da média amostral, a uma **normal**, também com média $E(\bar{X}) = \mu$, variância $V(\bar{X}) = \frac{\sigma^2}{n}$ e desvio padrão (erro padrão) $EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

De modo equivalente, também podemos dizer que a variável $Z_{\bar{X}}$, definida abaixo, segue distribuição **normal padrão (ou reduzida)**, isto é, com média igual a 0 e desvio padrão igual a 1, o que representamos como $N(0,1)$:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

A possibilidade de tal aproximação **depende** do **tamanho da amostra** e da **distribuição da população**. Quanto maior o tamanho da amostra e quanto mais próximo de uma normal for a distribuição da população, **melhor** será a aproximação. Usualmente, considera-se que para uma amostra grande, com $\boldsymbol{n \geq 30}$, a aproximação será satisfatória, para **qualquer distribuição** populacional.



(CESPE/2014 – ANATEL) Com base no teorema limite central, julgue o item abaixo.

Sendo uma amostra aleatória simples retirada de uma distribuição X com média μ e variância 1, a distribuição da média amostral dessa amostra, $\bar{X}$, converge para uma distribuição normal de média nμ e variância 1, à medida que n aumenta.

Comentários:

Vimos que a distribuição da média amostral, $\bar{X}$, converge para uma distribuição normal à medida que n aumenta. Porém, a média dessa distribuição é $E(\bar{X}) = \mu$ e variância $V(\bar{X}) = \frac{V(X)}{n}$. Sabendo que $V(X) = 1$, a variância da média amostral é $V(\bar{X}) = \frac{1}{n}$.

Gabarito: Errado.

(CESPE 2018/PF) O tempo gasto (em dias) na preparação para determinada operação policial é uma variável aleatória X que segue distribuição normal com média M, desconhecida, e desvio padrão igual a 3 dias. A observação de uma amostra aleatória de 100 outras operações policiais semelhantes a essa produziu uma média amostral igual a 10 dias. Com referência a essas informações, julgue o item que se segue, sabendo que $P(Z > 2) = 0{,}025$, em que Z denota uma variável aleatória normal padrão.

O erro padrão da média amostral foi inferior a 0,5 dia.

Comentários:

O erro padrão da média amostral é dado por:

$$EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Na fórmula acima, temos: n é o tamanho da amostra (=10); σ é o desvio padrão amostral (=3).

$$EP(\bar{X}) = \frac{3}{\sqrt{100}} = \frac{3}{10} = 0{,}3$$

Gabarito: Certo.

(CESPE/2019 – Analista Judiciário TJ) Um pesquisador deseja comparar a diferença entre as médias de duas amostras independentes oriundas de uma ou duas populações gaussianas. Considerando essa situação hipotética, julgue o próximo item.

Para que a referida comparação seja efetuada, é necessário que ambas as amostras tenham N ≥ 30.

Comentários:

Quando a população segue uma distribuição normal (ou gaussiana), a média amostral também seguirá uma distribuição normal, independente do tamanho da amostra. Logo, o item está errado.

Para fins de complementação, se a amostra for grande o suficiente (normalmente, consideramos isso para $n \geq 30$), a média amostral seguirá aproximadamente uma distribuição normal, independente da distribuição populacional.

Gabarito: Errado.

(FCC 2015/SEFAZ-PI) Instrução: Para responder à questão utilize, dentre as informações dadas a seguir, as que julgar apropriadas. Se Z tem distribuição normal padrão, então:

P(Z < 0,4) = 0,655; P(Z < 1,2) = 0,885; P(Z < 1,6) = 0,945; P(Z < 1,8) = 0,964; P(Z < 2) = 0,977.

Uma auditoria feita em uma grande empresa considerou uma amostra aleatória de 64 contas a receber. Se a população de onde essa amostra provém é infinita e tem distribuição normal com desvio padrão igual a R$

200,00 e média igual a R$ 950,00, a probabilidade da variável aleatória média amostral, usualmente denotada por $\boldsymbol{X}$, estar situada entre R$ 980,00 e R$ 1.000,00 é dada por

a) 18,4%

b) 9,2%

c) 28,5%

d) 47,7%

e) 86,2%

Comentários:

O enunciado informa que a população segue distribuição normal, logo, a média amostral também terá distribuição normal. Assim, utilizamos a seguinte transformação para a normal padrão:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

O enunciado informa que o tamanho da amostra é n = 64, logo, $\sqrt{n} = 8$; que a média populacional é $\mu = 950$ e que o desvio padrão populacional é $\sigma = 200$. Substituindo esses valores, a transformação para $\overline{X_{inf}} = 980$ é:

$$Z_{\bar{X}} = \frac{980 - 950}{\frac{200}{8}} = 30 \times \frac{8}{200} = \frac{6}{5} = 1{,}2$$

Para $\overline{X_{sup}} = 1000$, temos:

$$Z_{\bar{X}} = \frac{1000 - 950}{\frac{200}{8}} = 50 \times \frac{8}{200} = 2$$

Ou seja, a probabilidade desejada corresponde a $P(1{,}2 < Z < 2)$, que pode ser calculada pelos dados fornecidos no enunciado:

$$P(1{,}2 < Z < 2) = P(Z < 2) - P(Z < 1{,}2) = 0{,}977 - 0{,}885 = 0{,}092 = 9{,}2\%$$

Gabarito: B

Distribuição Amostral da Proporção

Agora, vamos trabalhar com uma população em que **determinada característica** está presente em uma **proporção $\boldsymbol{p}$** dessa população, por exemplo, 15% da população apresenta olhos azuis; 20% da população está doente, 1% da produção apresenta defeito, etc.

Um elemento qualquer da população X pode **apresentar** a característica estudada, o que chamamos de **sucesso** ($X = 1$), ou **não**, o que chamamos de **fracasso** ($X = 0$). A probabilidade de sucesso é $\boldsymbol{p}$ e a probabilidade de fracasso é $q = 1 - p$. Assim, essa população apresenta uma distribuição de **Bernoulli**, com parâmetro $\boldsymbol{p}$.

Sendo essa proporção populacional desconhecida, precisamos **estimá-la** a partir da proporção de sucessos na **amostra**, denotada por $\boldsymbol{\hat{p}}$. Considerando que cada observação X_i da amostra será $X_i = 0$ ou $X_i = 1$ (como para a população), então a proporção de sucessos na amostra pode ser calculada por:

$$\boxed{\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}}$$

Vamos supor, por exemplo, que estejamos interessados na proporção de defeitos em uma produção de medicamentos. Para estimá-la, extraímos uma amostra de 10 medicamentos, que apresentou o seguinte resultado, em que 0 representa um item não defeituoso e 1 representa um item defeituoso:

{0, 0, 1, 0, 0, 0, 1, 0, 0, 0}

Logo, a proporção encontrada nessa amostra é:

$$\hat{p} = \frac{0+0+1+0+0+0+1+0+0+0}{10} = \frac{2}{10} = 0{,}2$$

Note que o estimador $\hat{p}$ é calculado da mesma forma que $\boldsymbol{\bar{X}}$**.** Logo, a **esperança** de $\hat{p}$ é calculada da mesma forma que para $\bar{X}$, utilizando $\boldsymbol{p}$ no lugar de μ:

$$\boxed{E(\hat{p}) = \boldsymbol{p}}$$

Ou seja, a **esperança** do estimador é igual à **proporção populacional**. Em outras palavras, a proporção amostral **tende** à **proporção populacional**.

A **variância** de $\hat{p}$ também é calculada da mesma forma que para $\bar{X}$, ou seja:

$$V(\hat{p}) = \frac{V(p)}{\boldsymbol{n}}$$

Sabendo que X segue distribuição de Bernoulli, temos $V(p) = p.q$, então:

$$\boxed{V(\hat{p}) = \frac{p.q}{n}}$$

E o **erro padrão** (ou desvio padrão) para $\hat{p}$, raiz quadrada da sua variância é dado por:

$$\boxed{EP(\hat{p}) = \sqrt{\frac{p.q}{n}}}$$

Sem conhecer a proporção populacional, p, não podemos calcular a variância da população, $V(p) = p.q$. Assim, utilizamos o estimador $\hat{\boldsymbol{p}}$ calculado a partir da amostra para **estimar** a variância e o desvio padrão.

A **estimativa da variância populacional** é dada por:

$$V(p) = \hat{p}.\hat{q}$$

Em que $\hat{q} = 1 - \hat{p}$. E a **estimativa da variância do estimador** $\hat{\boldsymbol{p}}$ é:

$$V(\hat{p}) = \frac{\hat{p}.\hat{q}}{n}$$

Para o nosso exemplo, em que encontramos $\hat{p} = 0{,}2$ (logo, $\hat{q} = 1 - \hat{p} = 0{,}8$). A estimativa da **variância** da proporção **populacional** é:

$$V(p) = \hat{p}.\hat{q} = 0{,}2 \times 0{,}8 = 0{,}16$$

E a estimativa da **variância** da proporção **amostral** é:

$$V(\hat{p}) = \frac{\hat{p}.\hat{q}}{n} = \frac{0{,}2 \times 0{,}8}{10} = 0{,}016$$

Logo, a estimativa para o **erro padrão** (ou desvio padrão) da **proporção amostral** é:

$$EP(\hat{p}) = \sqrt{V(\hat{p})} = \sqrt{0{,}016} \cong 0{,}126$$

Considerando que cada elemento da população segue distribuição de Bernoulli, então o número de elementos com o atributo sucesso encontrados em uma **amostra** de tamanho n segue uma distribuição **binomial**, com parâmetros n e p. Para o nosso exemplo, temos uma distribuição binomial com $n = 10$ e proporção estimada $\hat{p} = 0{,}2$.

Porém, assim, como vimos para a média amostral, também podemos aproximar, pelo **Teorema Central do Limite**, essa distribuição a uma **normal**, com média $E(\hat{p}) = p$ e variância $V(\hat{p}) = \frac{p.q}{n}$, quando o tamanho da amostra, $\boldsymbol{n}$, é **grande**.

Por outro lado, se a população for **finita** e a amostra for extraída **sem reposição**, será necessário aplicar o fator de correção para população finita, multiplicando a **variância** por $\frac{N-n}{N-1}$:

$$V(\hat{p}) = \frac{p.q}{n} \times \frac{N-n}{N-1}$$

(CESPE 2016/TCE-PA) Em estudo acerca da situação do CNPJ das empresas de determinado município, as empresas que estavam com o CNPJ regular foram representadas por 1, ao passo que as com CNPJ irregular foram representadas por 0.

Considerando que a amostra

{0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1}

Foi extraída para realizar um teste de hipóteses, julgue o item subsequente.

A estimativa pontual da proporção de empresas da amostra com CNPJ regular é superior a 50%.

Comentários:

O estimador da proporção $\hat{p}$ é dado por:

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$\hat{p} = \frac{0+1+1+0+0+1+0+1+0+1+1+0+0+1+1+0+1+1+1+1}{20} = \frac{12}{20} = 0{,}6$$

Logo a proporção é de 60%.

Gabarito: Certo

(CESPE/2019 – TJ-AM) Para estimar a proporção de menores infratores reincidentes em determinado município, foi realizado um levantamento estatístico. Da população-alvo desse estudo, constituída por 10.050 menores infratores, foi retirada uma amostra aleatória simples sem reposição, composta por 201 indivíduos. Nessa amostra foram encontrados 67 reincidentes. Com relação a essa situação hipotética, julgue o seguinte item.

A estimativa do erro padrão da proporção amostral foi inferior a 0,04.

Comentários:

O erro padrão da proporção é dada pela relação:

$$E = \sqrt{\frac{p \times q}{n}}$$

A questão nos diz que a amostra é composta por 201 indivíduos, sendo 67 deles reincidentes. Assim, temos que $n = 201$ e $p = 67/201 = 1/3$. Logo, temos $q = 1 - p = 2/3$. Como consequência, o erro padrão é de:

$$E = \sqrt{\frac{\frac{1}{3} \times \frac{2}{3}}{201}} = \sqrt{\frac{2}{1809}} \cong 0{,}033$$

Gabarito: Certo.

RESUMINDO

Estimador para a média: $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$

Esperança: $E(\bar{X}) = \mu$; Variância: $V(\bar{X}) = \frac{\sigma^2}{n}$; Erro Padrão: $EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Estimador para a proporção: $\hat{p} = \frac{X_1+X_2+\cdots+X_n}{n}$

Esperança: $E(\hat{p}) = p$; Variância: $V(\hat{p}) = \frac{p.q}{n}$; Erro Padrão: $EP(\bar{X}) = \sqrt{\frac{p.q}{n}}$

Distribuição Amostral da Variância

Quando a variância da população é desconhecida, precisamos estimá-la a partir da amostra, assim como fizemos com a média e a proporção. O **estimador da variância** que utilizamos para uma amostra de tamanho n é:

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{\boldsymbol{n-1}}$$

Utilizamos esse estimador, com a divisão por $\boldsymbol{n-1}$, pelo fato de ele ser melhor do que o estimador que apresenta a divisão por n.

EXEMPLIFICANDO

Vamos supor que a variância da altura de determinado grupo de adultos seja desconhecida, assim como a sua média. Para estimar esses parâmetros, obtemos uma amostra de 5 pessoas com os seguintes resultados:

{1,65; 1,75; 1,8; 1,85; 1,95}

Primeiro, precisamos calcular a média da amostra (que continua sendo calculada pelo somatório das observações, dividido por n):

$$\bar{X} = \frac{\sum X_i}{n} = \frac{1{,}65+1{,}75+1{,}8+1{,}85+1{,}95}{5} = \frac{9}{5} = 1{,}8$$

Agora, calculamos o estimador da variância, somando os desvios em relação à média, elevados ao quadrado, e dividindo o somatório por $\boldsymbol{n-1}$:

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{\boldsymbol{n-1}}$$

$$s^2 = \frac{(1{,}65-1{,}8)^2+(1{,}75-1{,}8)^2+(1{,}8-1{,}8)^2+(1{,}85-1{,}8)^2+(1{,}95-1{,}8)^2}{4}$$

$$s^2 = \frac{(-0{,}15)^2+(-0{,}05)^2+(0)^2+(0{,}05)^2+(0{,}15)^2}{4}$$

$$s^2 = \frac{0{,}0225+0{,}0025+0{,}0025+0{,}0225}{4} = \frac{0{,}05}{4} = 0{,}0125$$

INDO MAIS FUNDO!

Sabe-se que uma maneira alternativa de calcular a variância (populacional) é:

$$\sigma^2 = E(X^2) - [E(X)]^2$$

Para a variância amostral, também podemos utilizar uma fórmula similar a essa, mas com as devidas adaptações. No lugar de $E(X)$, utilizamos a média amostral $\bar{X}$; e, no lugar de $E(X^2)$, utilizamos $\overline{X^2}$, que é a média dos valores elevados ao quadrado:

$$\overline{X^2} = \frac{\sum X_i^2}{n}$$

Para o exemplo anterior, teríamos:

$$\overline{X^2} = \frac{(1{,}65)^2+(1{,}75)^2+(1{,}8)^2+(1{,}85)^2+(1{,}95)^2}{5} = \frac{16{,}25}{5} = 3{,}25$$

Por fim, ajustamos o denominador. Para calcular a variância populacional, dividimos por n e para a variância amostral, dividimos por $\boldsymbol{n-1}$. Logo, precisamos multiplicar o resultado por $\frac{n}{\boldsymbol{n-1}}$:

$$s^2 = \left[\overline{X^2} - (\bar{X})^2\right] \times \frac{n}{\boldsymbol{n-1}}$$

Para o nosso exemplo, em que $\overline{X^2} = 3{,}25$, $\bar{X} = 1{,}8$ e $n = 5$, a variância amostral pode ser calculada como:

$$s^2 = [3{,}25 - (1{,}8)^2] \times \frac{5}{4} = [3{,}25 - 3{,}24] \times \frac{5}{4} = \frac{0{,}01 \times 5}{4} = 0{,}0125$$

Para esse estimador, a sua **esperança** é igual à **variância populacional**, análogo ao que ocorreu com os demais estimadores. A sua variância e erro padrão são dados por:

$$E(s^2) = \sigma^2$$

$$V(s^2) = \frac{2.\sigma^4}{n-1}$$

$$EP(s^2) = \sqrt{\frac{2}{n-1}}\sigma^2$$

Sem conhecer a variância populacional, σ^2, não podemos calcular esses parâmetros. Para isso, utilizamos, no lugar de σ^2, a própria estimativa s^2.

Para o nosso exemplo, em que calculamos $s^2 = 0{,}0125$, para a amostra com $n = 5$ observações, o desvio padrão (ou erro padrão) de s^2 pode ser estimado como:

$$EP(s^2) = \sqrt{\frac{2}{n-1}s^2} = \sqrt{\frac{2}{4}} \times 0{,}0125 \cong 0{,}009$$

Se a **população** seguir uma **distribuição normal**, então o estimador s^2, multiplicado pelo fator $\boldsymbol{\frac{n-1}{\sigma^2}}$, segue uma **distribuição qui-quadrado** com $\boldsymbol{n-1}$ **graus de liberdade**:

$$\mathcal{X}^2_{n-1} = \left(\frac{n-1}{\sigma^2}\right).s^2$$

Em outras palavras, o estimador s^2 é uma variável com distribuição qui-quadrado, com $n-1$ graus de liberdade, multiplicada por $\frac{\sigma^2}{n-1}$:

$$s^2 = \left(\frac{\sigma^2}{n-1}\right).\mathcal{X}^2_{n-1}$$

Vamos supor que, para uma população normal, a variância populacional seja $\sigma^2 = 1$, e que vamos extrair amostras de tamanho $n = 5$. Nesse caso, a variância amostral s^2 terá a seguinte distribuição:

$$s^2 = \left(\frac{1}{5-1}\right).\mathcal{X}^2_{5-1} = \frac{\mathcal{X}^2_4}{4}$$

Ou seja, a variância amostral seguirá uma distribuição qui-quadrado com $n - 1 = 4$ graus de liberdade, dividida por 4.

Assim, inserimos a tabela da distribuição qui-quadrado com 4 graus de liberdade, que apresenta os valores de probabilidade $P(\mathcal{X}^2_4 < x)$ e os respectivos valores de x. Como a variância amostral segue essa distribuição, dividida por 4, criamos uma terceira coluna, dividindo os valores de x por 4:

$P(\mathcal{X}_4^2 < x)$	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
x	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
$x/4$	0,05	0,07	0,12	0,18	0,27	0,48	0,84	1,35	1,94	2,37	2,79	3,32	3,72

Ou seja, a probabilidade de a variância amostral observada ser inferior a 0,48 é:

$$P(s^2 < 0{,}48) = \mathbf{0{,}25}$$

A partir desse resultado, podemos calcular a esperança e a variância do estimador. A esperança é dada por:

$$s^2 = \left(\frac{\sigma^2}{n-1}\right).\mathcal{X}_{n-1}^2$$

$$E[s^2] = E\left[\left(\frac{\sigma^2}{n-1}\right).\mathcal{X}_{n-1}^2\right] = \left(\frac{\sigma^2}{n-1}\right).E[\mathcal{X}_{n-1}^2]$$

Considerando que a média de uma distribuição qui-quadrado com k graus de liberdade é igual a k, então fazendo $k = n - 1$, temos:

$$E[\mathcal{X}_{n-1}^2] = k = n - 1$$

Substituindo no resultado anterior, temos:

$$E[s^2] = \left(\frac{\sigma^2}{n-1}\right).(n-1) = \sigma^2$$

Esse é o resultado que vimos no início da seção. A variância do estimador é:

$$V[s^2] = V\left[\left(\frac{\sigma^2}{n-1}\right)\mathcal{X}_{n-1}^2\right] = \left(\frac{\sigma^2}{n-1}\right)^2.V[\mathcal{X}_{n-1}^2] = \left(\frac{\sigma^4}{(n-1)^2}\right).V[\mathcal{X}_{n-1}^2]$$

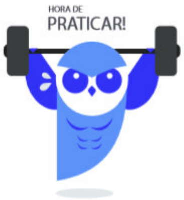
Considerando que a variância de uma distribuição qui-quadrado com k graus de liberdade é igual a $2k$, então fazendo $k = n - 1$, temos:

$$V[\mathcal{X}_{n-1}^2] = 2.k = 2.(n-1)$$

Substituindo no resultado anterior, temos:

$$V[s^2] = \left(\frac{\sigma^4}{(n-1)^2}\right).2.(n-1) = \frac{2.\sigma^4}{n-1}$$

Vale acrescentar que uma população normal depende dos dois parâmetros, variância e média, os quais são independentes. Consequentemente, os estimadores correspondentes também serão **independentes**.



(FGV/2016 – IBGE) Suponha que uma amostra de tamanho $\boldsymbol{n} = \mathbf{5}$ é extraída de uma população normal, com média desconhecida, obtendo as seguintes observações:

$$\boldsymbol{X_1} = \mathbf{3},\ \boldsymbol{X_2} = \mathbf{5},\ \boldsymbol{X_3} = \mathbf{6},\ \boldsymbol{X_4} = \mathbf{9} \text{ e } \boldsymbol{X_5} = \mathbf{12}$$

São dados ainda os seguintes valores, retirados da tabela da distribuição qui-quadrado:

- $P(\chi_4^2 < \mathbf{5}) \cong \mathbf{0{,}713}$
- $P(\chi_4^2 < \mathbf{12{,}5}) \cong \mathbf{0{,}986}$
- $P(\chi_5^2 > \mathbf{5}) \cong \mathbf{0{,}854}$
- $P(\chi_5^2 > \mathbf{12{,}5}) \cong \mathbf{0{,}971}$

Se a população tem variância verdadeira $\boldsymbol{\sigma^2} = \mathbf{4}$, em nova amostra ($\boldsymbol{n} = \mathbf{5}$), a probabilidade de se observar uma variância amostral maior do que a anterior é de:

a) 0,014

b) 0,029

c) 0,146

d) 0,287

e) 0,713

Comentários:

Para resolver essa questão, vamos primeiro calcular a variância amostral obtida nessa primeira amostra:

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{\boldsymbol{n - 1}}$$

Para isso, precisamos da média amostral:

$$\bar{X} = \frac{3 + 5 + 6 + 9 + 12}{5} = \frac{35}{5} = 7$$

Agora, podemos calcular a variância amostral dessa primeira amostra:

$$s_1^2 = \frac{(3-7)^2 + (5-7)^2 + (6-7)^2 + (9-7)^2 + (12-7)^2}{4}$$

$$s_1^2 = \frac{(-4)^2 + (-2)^2 + (-1)^2 + (2)^2 + (5)^2}{4} = \frac{16 + 4 + 1 + 4 + 25}{4} = \frac{50}{4} = 12{,}5$$

Para calcular a probabilidade de a variância amostral ser $s^2 > 12{,}5$, consideramos que esse estimador é uma distribuição qui-quadrado com n – 1 graus de liberdade, multiplicada por $\left(\frac{\sigma^2}{n-1}\right)$:

$$s^2 = \left(\frac{\sigma^2}{n-1}\right).\mathcal{X}_{n-1}^2$$

Sabendo que a variância populacional é $\sigma^2 = 4$ e que o tamanho da amostra é $n = 5$, então:

$$s^2 = \left(\frac{4}{5-1}\right).\mathcal{X}_{5-1}^2 = \mathcal{X}_4^2$$

Portanto, a variância amostral segue a mesma distribuição de $\mathcal{X}_4^2$. A probabilidade de $s^2 > 12{,}5$ é, portanto:

$$P(s^2 > 12{,}5) = P(\mathcal{X}_4^2 > 12{,}5)$$

O enunciado informa que $P(\mathcal{X}_4^2 < 12{,}5) = 0{,}986$. A probabilidade $P(\mathcal{X}_4^2 > 12{,}5)$ é complementar:

$$P(\mathcal{X}_4^2 > 12{,}5) = 1 - P(\mathcal{X}_4^2 < 12{,}5) = 1 - 0{,}986 = 0{,}014$$

Gabarito: A

Distribuições para Amostragem Estratificada

Para uma amostragem estratificada, com k estratos, a **média amostral** será dada por:

$$\bar{X} = \sum_{i=1}^{k} \frac{N_i}{N} \bar{x_i}$$

Nessa expressão, N_i é o tamanho de cada estrato; N é o tamanho total da população; e $\bar{x_i}$, a média amostral observada para cada estrato.

Ou seja, calculamos a média $\bar{x_i}$ para cada estrato i, multiplicamos cada valor pelo tamanho do estrato N_i e dividimos pelo tamanho total N.

Para ilustrar, vamos supor uma população dividida em 3 estratos, com os seguintes tamanhos N_i e os seguintes valores de média amostral $\bar{x_i}$ para cada estrato i:

Estrato	N_i	n_i	$\bar{x_i}$
1	50	5	2
2	30	3	3
3	20	2	4

A média amostral para toda a população corresponde, então, às médias amostrais dos estratos, ponderadas pelos respectivos tamanhos dos estratos:

$$\bar{X} = \frac{50 \times 2 + 30 \times 3 + 20 \times 4}{100} = \frac{100 + 90 + 80}{100} = 2{,}7$$

Considerando a fórmula da média amostral, podemos calcular **a variância da média amostral**:

$$V(\bar{X}) = V\left(\sum_{i=1}^{k} \frac{N_i}{N} \bar{x}_i\right)$$

Pelas propriedades da variância, temos:

$$V(\bar{X}) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 V(\bar{x}_i)$$

Sendo a variância populacional de cada estrato **desconhecida**, precisamos estimá-la a partir da estimativa da variância para cada estrato encontrada na amostra estratificada. Substituindo, na fórmula acima, $V(\bar{X})$ por $s_{\bar{x}}^2$ e $V(\bar{x}_i)$ por $s_{\bar{x}_i}^2$, temos:

$$s_{\bar{x}}^2 = \sum_{h=1}^{k} \left(\frac{N_i}{N}\right)^2 \times s_{\bar{x}_i}^2$$

Além disso, a estimativa da variância da média amostral para cada estrato i, $s_{\bar{x}_i}^2$, considerando uma amostra de tamanho n_i para tal estrato, já com a correção para população finita é dada por:

$$s_{\bar{x}_i}^2 = \frac{s_{x_i}^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)$$

Vamos supor que as estimativas da variância para cada estrato sejam:

Estrato	N_i	n_i	$\bar{x}_i$	$s_{x_i}^2$
1	50	5	2	2
2	30	3	3	1,5
3	20	2	4	1

Assim, as variâncias das médias amostrais para cada estrato são:

$$s_{\bar{x}_1}^2 = \frac{2}{5}\left(\frac{50-5}{50-1}\right) \cong 0{,}367$$

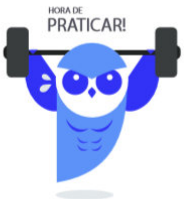
$$s_{\bar{x}_2}^2 = \frac{1{,}5}{3}\left(\frac{30-3}{30-1}\right) \cong 0{,}466$$

$$s_{\bar{x}_3}^2 = \frac{1}{2}\left(\frac{20-2}{20-1}\right) \cong 0{,}474$$

E a variância da média amostral global é dada por:

$$s_{\bar{x}}^2 = \left(\frac{50}{100}\right)^2 \times 0{,}367 + \left(\frac{30}{100}\right)^2 \times 0{,}466 + \left(\frac{20}{100}\right)^2 \times 0{,}474$$

$$s_{\bar{x}}^2 = 0{,}25 \times 0{,}367 + 0{,}09 \times 0{,}466 + 0{,}04 \times 0{,}474 = 0{,}092 + 0{,}042 + 0{,}019 = 0{,}153$$



(CESPE/2018 – STM) Um estudo acerca do tempo (x, em anos) de guarda de autos findos em determinada seção judiciária considerou uma amostragem aleatória estratificada. A população consiste de uma listagem de autos findos, que foi segmentada em quatro estratos, segundo a classe de cada processo (as classes foram estabelecidas por resolução de autoridade judiciária.) A tabela a seguir mostra os tamanhos populacionais (N) e amostrais (n), a média amostral ($\bar{x}$) e a variância amostral dos tempos (s^2) correspondentes a cada estrato.

Estratos	Tamanhos Populacionais (N)	Tamanhos Amostra (n)	$\bar{x}$	s^2
A	30.000	300	20	3
B	40.000	400	15	16
C	50.000	500	10	5
D	80.000	800	5	8
Total	200.000	2.000	-	-

Considerando que o objetivo do estudo seja estimar o tempo médio populacional (em anos) de guarda dos autos findos, julgue os itens a seguir.

(CESPE/2018 – STM) A estimativa do tempo médio populacional da guarda dos autos findos é maior ou igual a 12 anos.

Comentários:

Em uma amostra aleatória por estratificação, a média da população é calculada pela relação

$$\bar{X} = \sum_{i=1}^{k} \frac{N_i}{N} \overline{x_i}$$

Em que k é a quantidade de estratos; N_i o tamanho de cada estrato; e $\overline{x_i}$ a média de cada estrato.

Vinculando os estratos A, B, C e D aos números 1, 2, 3 e 4, respectivamente, temos:

$$\bar{X} = \frac{N_1 \times \overline{x_1} + N_2 \times \overline{x_2} + N_3 \times \overline{x_3} + N_4 \times \overline{x_4}}{N}$$

$$\bar{X} = \frac{30000 \times 20 + 40000 \times 15 + 50000 \times 10 + 80000 \times 5}{200000}$$

$$\bar{X} = \frac{60 + 60 + 50 + 40}{20}$$

$$\bar{X} = 10{,}5$$

Gabarito: Errado.

(CESPE/2018 – STM) Combinando-se todos os estratos envolvidos, a estimativa da variância do tempo médio amostral da guarda dos autos findos é inferior a $0{,}005\ ano^2$.

Comentários:

Em uma amostragem estratificada, a estimativa da variância da média da amostragem é dada pela relação:

$$s_{\bar{x}}^2 = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \times s_{\bar{x}_i}^2$$

Em que

$$s_{\bar{x}_i}^2 = \frac{s_{x_i}^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)$$

Pelos valores apresentados na tabela, teremos:

$$s_{\bar{x}_A}^2 = \frac{3}{300}\left(\frac{30000 - 300}{30000 - 1}\right) = 0{,}0099$$

$$s^2_{\bar{x}_B} = \frac{16}{400}\left(\frac{40000-400}{40000-1}\right) = 0{,}0396$$

$$s^2_{\bar{x}_C} = \frac{5}{500}\left(\frac{50000-500}{50000-1}\right) = 0{,}0099$$

$$s^2_{\bar{x}_D} = \frac{8}{800}\left(\frac{80000-800}{80000-1}\right) = 0{,}0099$$

Além disso, temos que:

$$\left(\frac{N_A}{N}\right)^2 = \left(\frac{30000}{200000}\right)^2 = 0{,}0225$$

$$\left(\frac{N_B}{N}\right)^2 = \left(\frac{40000}{200000}\right)^2 = 0{,}04$$

$$\left(\frac{N_C}{N}\right)^2 = \left(\frac{50000}{200000}\right)^2 = 0{,}0625$$

$$\left(\frac{N_D}{N}\right)^2 = \left(\frac{80000}{200000}\right)^2 = 0{,}16$$

Portanto:

$$s^2_{\bar{x}} = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \times s^2_{\bar{x}_h}$$

$$s^2_{\bar{x}} = (0{,}0225 \cdot 0{,}0099 + 0{,}04 \cdot 0{,}0396 + 0{,}0625 \cdot 0{,}0099 + 0{,}16 \cdot 0{,}0099)$$

$$s^2_{\bar{x}} = (0{,}0225 + 0{,}0625 + 0{,}16) \cdot 0{,}0099 + 0{,}04 \cdot 0{,}0396$$

$$s^2_{\bar{x}} \cong 0{,}0041.$$

Gabarito: Certo.

ESTIMAÇÃO PONTUAL

A **estimação pontual** é o **valor** (número) calculado para o estimador. Os principais estimadores são a média amostral $\bar{X}$, a proporção amostral $\hat{p}$ e a variância amostral s^2.

Agora, vamos entender por que esses estimadores são utilizados. Vamos estudar as propriedades que estimadores devem apresentar e os métodos utilizados para obtê-los.

Propriedades dos Estimadores

Não é qualquer estimador que vamos querer utilizar para estimar os parâmetros populacionais. Eles devem apresentar determinadas características desejáveis.

Suficiência

Uma **estatística** (isto é, uma **função dos dados observados**) é considerada **suficiente** se ela captura, a partir da amostra obtida, **toda a informação possível sobre o parâmetro populacional desconhecido**, de modo que qualquer outra informação não contribuirá com a estimação do parâmetro populacional.

Por exemplo, sabemos que a média amostral é utilizada para estimar a média populacional. Essa estatística é considerada suficiente porque ela captura toda a informação, disponível na amostra, necessária para estimar a média populacional. Qualquer outra informação, como a média geométrica ou a variância da amostra, não influencia na estimativa da média populacional.

Mais especificamente, sendo X e Y duas amostras que fornecem o mesmo valor para a estatística, $T(X) = T(Y)$, (por exemplo, a mesma média amostral), então essa estatística será considerada suficiente se a inferência sobre o parâmetro populacional θ for a **mesma**, independente de X ou Y. Dizemos que uma estatística $T(X)$ é **suficiente** para o parâmetro populacional θ, se a **distribuição da amostra, condicionada ao parâmetro $\boldsymbol{T(X)}$,** for **independente de $\boldsymbol{\theta}$**.

Em outras palavras, uma estatística suficiente é uma forma de **resumir** as informações presentes na amostra, **sem perder as informações** necessárias para estimar o parâmetro populacional. Por exemplo, a **soma** dos valores presentes em uma amostra de tamanho n é uma estatística **suficiente** para a média da população, pois, para estimá-la, basta dividirmos essa estatística por n. Não importa quais sejam os valores exatos de cada variável presente na amostra; se a soma for a mesma, a estimativa para a média será sempre a mesma.

Todos os estimadores que mencionamos anteriormente (média amostral $\bar{X}$, proporção amostral $\hat{p}$ e variância amostral s^2) são considerados suficientes.

INDO MAIS FUNDO!

Há, ainda, as seguintes definições relacionadas à suficiência:

- Uma estatística é dita **suficiente minimal** se ela for uma **função** de qualquer outra estatística suficiente.
- Por outro lado, uma estatística $S(X)$ é dita **anciliar** se ela **não trouxer informação** alguma a respeito do parâmetro θ, isto é, se $S(X)$ for **independente** de θ.
- Por fim, uma estatística T é dita **completa** para o parâmetro θ, se o fato de a **esperança** de uma função $g(T)$ ser igual a **zero** para todo θ, $E[g(T)] = 0$, implicar necessariamente em $\boldsymbol{g(T) = 0}$ para todo θ.

Por exemplo, para uma amostra $X_1, X_2, \dots, X_n$ com distribuição de Bernoulli e parâmetro p, por exemplo, $p = 0{,}2$, a estatística $T = X_1 - X_2$ **não é completa**. Isso porque a esperança é $E[X_1] = E[X_2] = p = 0{,}2$, logo a esperança da estatística é:

$$E[X_1 - X_2] = E[X_1] - E[X_2] = 0$$

E isso vale para **qualquer valor de $\boldsymbol{p}$**. Porém, $T = X_1 - X_2$ **não** é necessariamente igual a **zero**, pois é possível ter $X_1 = 1$ e $X_2 = 0$ ou também $X_1 = 0$ e $X_2 = 1$.

HORA DE PRATICAR!

(CESPE/2018 – EBSERH) $X_1, X_2, \dots, X_{10}$ representa uma amostra aleatória simples retirada de uma distribuição normal com média μ e variância σ^2, ambas desconhecidas. Considerando que $\hat{\mu}$ e $\widehat{\sigma^2}$ representam os respectivos estimadores de máxima verossimilhança desses parâmetros populacionais, julgue o item subsecutivo.

A soma $X_1 + X_2 + \dots + X_{10}$ é uma estatística suficiente para a estimação do parâmetro μ.

Comentários:

Observe que a soma dos elementos captura todas as informações disponíveis na amostra para a estimação da média, de modo que **qualquer outra informação** (média, variância,...) a respeito dessa amostra **não irá influenciar na estimação do parâmetro populacional**.

Afinal, a estimativa para o parâmetro μ será a **mesma**, sempre que a **soma** das 10 variáveis for a **mesma**, independentemente dos seus resultados exatos. Logo, a soma dos elementos é uma estatística suficiente.

Gabarito: Certo.

(CESPE/2013 – Telebras) A respeito de inferência estatística, julgue o item que se segue.

Considerando uma amostra aleatória simples X_1, X_2, X_3 retirada de determinada distribuição de Bernoulli, com parâmetro p desconhecido, é correto afirmar que $X_1+X_2.X_3$ é estatística suficiente.

Comentários:

Observe que ao multiplicarmos $X_2.X_3$ há uma possibilidade de perda de informação, pois sendo um dos elementos iguais a zero, ficamos sem a informação do outro elemento, ou seja, há **perda de informação**.

Assim, ter uma outra informação a respeito dessa amostra irá contribuir com a estimação do parâmetro p.

Por exemplo, se tivermos $X_1 = 1$, $X_2 = 1$ e $X_3 = 0$ e, portanto, $X_1 + X_2.X_3 = 1$, a estimativa para a proporção p será p = 2/3; e se tivermos $X_1 = 1$, $X_2 = 0$ e $X_3 = 0$, e, portanto, o mesmo valor para a estatística, $X_1 + X_2.X_3 = 1$, a estimativa para p será p = 1/3.

Ou seja, a **estimação** do parâmetro p desconhecido pode ser **diferente**, mesmo se o resultado da **estatística** for **igual**.

Logo, essa estatística **não é suficiente**.

Gabarito: Errado.

Não viés

Dizemos que um estimador $\hat{\theta}$ é **não viesado** (também chamado de **não viciado** ou **não tendencioso**) quando a sua **esperança** é igual ao **parâmetro populacional** θ sendo estimado:

$$E\left(\widehat{\theta_{NT}}\right) = \theta$$

Em relação aos principais estimadores que mencionamos anteriormente, temos:

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

$$E(\hat{p}) = p$$

$$E(s^2) = \sigma^2$$

Logo, podemos afirmar que a média amostral $\bar{X}$ é um estimador não viesado (ENV) para a média populacional; o parâmetro amostral $\hat{p}$ é um ENV para o parâmetro populacional; e s^2 é um ENV para a variância populacional.

O estimador para a variância populacional é definido como:

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{\boldsymbol{n-1}}$$

Dissemos, ainda, que esse estimador é **melhor** do que o estimador $s_*^2 = \frac{\sum (X_i - \bar{X})^2}{\boldsymbol{n}}$. Isso decorre do fato de o estimador s_*^2 ser **tendencioso**, ou seja:

$$E(s_*^2) \neq \sigma^2$$

Outra forma de descrever essa propriedade é a partir do **erro**, dado pela diferença entre o estimador e o parâmetro populacional:

$$e = \hat{\theta} - \theta$$

A **esperança do erro** é chamada de **viés do estimador** (ou **tendenciosidade**), denotado por $b(\hat{\theta})$:

$$b(\hat{\theta}) = E(e)$$

O viés pode ser calculado pelas propriedades da esperança:

$$b(\hat{\theta}) = E(e) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - E(\theta)$$

$$\boxed{b(\hat{\theta}) = E(\hat{\theta}) - \theta}$$

Ou seja, o viés representa a diferença entre a média (ou esperança) das estimativas e o parâmetro a ser estimado. Para um estimador $\widehat{\theta^*}$ **não tendencioso**, a sua esperança é $E\left(\widehat{\theta^*}\right) = \theta$. Logo, o **viés do estimador (esperança do erro)** é **nulo**:

$$b\left(\widehat{\theta_{NT}}\right) = E\left(\widehat{\theta_{NT}}\right) - \theta = \theta - \theta = 0$$

EXEMPLIFICANDO

Vamos supor uma amostra de n = 5 elementos X_1, X_2, X_3, X_4 e X_5 de uma variável X com média μ desconhecida. Agora, vamos calcular o viés do seguinte estimador para a média μ:

$$\hat{\theta} = X_1 + X_2 + X_3 - X_4 - X_5$$

$$b(\hat{\theta}) = E(\hat{\theta}) - \mu$$

$$b(\hat{\theta}) = E(X_1 + X_2 + X_3 - X_4 - X_5) - \mu$$

Pelas propriedades da esperança, temos:

$$b(\hat{\theta}) = E(X_1) + E(X_2) + E(X_3) - E(X_4) - E(X_5) - \mu$$

Considerando que a amostra segue a mesma distribuição da população, temos $E(X_i) = E(X) = \mu$, logo:

$$b(\hat{\theta}) = \mu + \mu + \mu - \mu - \mu - \mu = 0$$

Portanto, esse estimador é não viesado. Vejamos um outro exemplo de estimador para a média μ:

$$\widehat{\theta'} = X_1 - X_2$$

$$b(\widehat{\theta'}) = E(\widehat{\theta'}) - \mu$$

$$b(\widehat{\theta'}) = E(X_1 - X_2) - \mu = E(X_1) - E(X_2) - \mu = \mu - \mu - \mu = -\mu$$

Portanto, o estimador $\widehat{\theta'}$ é viesado, com viés no valor de $-\mu$.

Já, a **variância do erro** é chamada de **Erro Quadrático Médio** (EQM):

$$EQM(\hat{\theta}) = V(e)$$

Desenvolvendo a expressão da variância do erro, podemos obter o seguinte resultado:

$$EQM(\hat{\theta}) = V(\hat{\theta}) + [b(\hat{\theta})]^2$$

Ou, simplesmente:

$$EQM(\hat{\theta}) = \sigma^2 + b^2$$

Ou seja, o Erro Quadrático Médio (EQM) é a soma da **variância** do estimador com o **quadrado do viés** do estimador.

Para estimadores **não viesados**, o viés do estimador é $b(\widehat{\theta_{NT}}) = 0$, logo, o EQM é igual à **variância** do estimador:

$$EQM(\widehat{\theta_{NT}}) = V(\widehat{\theta_{NT}})$$

Para ilustrar, inserimos abaixo a expressão do EQM de um estimador em que a primeira parcela corresponde à variância do estimador $V(\hat{\theta})$ e a segunda parcela corresponde ao quadrado do viés do estimador $[b(\hat{\theta})]^2$ (como na prova da FGV/2017 – IBGE):

$$EQM(\hat{\theta}_1) = \frac{3.\sigma^2}{n} + \left(\frac{\theta - 1}{n}\right)^2$$

Podemos observar que o **viés**, $\left(\frac{\theta-1}{n}\right)$, **não é nulo**, logo, concluímos que o estimador é **tendencioso**.

EXEMPLIFICANDO

Agora, vamos calcular o EQM dos estimadores para a média μ, que vimos anteriormente.

Para isso, precisamos da variância e do viés do estimador. Como já calculamos o viés, falta calcular a sua variância.

$$\hat{\theta} = X_1 + X_2 + X_3 - X_4 - X_5$$

$$V(\hat{\theta}) = V(X_1 + X_2 + X_3 - X_4 - X_5)$$

Se as amostras forem **independentes** (população infinita ou amostra extraída com reposição), então, pelas propriedades da variância, temos:

$$V(\hat{\theta}) = V(X_1) + V(X_2) + V(X_3) + V(X_4) + V(X_5)$$

Considerando que a amostra segue a mesma distribuição da população, temos $V(X_i) = V(X) = \sigma^2$, logo:

$$V(\hat{\theta}) = \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2 = 5\sigma^2$$

Como o viés é nulo, o EQM desse estimador é igual variância:

$$EQM(\hat{\theta}) = V(\hat{\theta}) = 5\sigma^2$$

Para o outro exemplo de estimador, a variância é (considerando as amostras **independentes**):

$$\widehat{\theta'} = X_1 - X_2$$

$$V(\widehat{\theta'}) = V(X_1 - X_2) = V(X_1) + V(X_2) = \sigma^2 + \sigma^2 = 2\sigma^2$$

Já calculamos o viés do estimador $b(\widehat{\theta'}) = -\mu$.

Portanto, o EQM é:

$$EQM(\widehat{\theta'}) = V(\hat{\theta}) + \left[b(\hat{\theta})\right]^2 = 2\sigma^2 + [-\mu]^2 = 2\sigma^2 + \mu^2$$

Se as amostras **não** forem **independentes**, a variância da soma e da subtração de duas variáveis é dada por:

$$V(X_1 + X_2) = V(X_1) + V(X_2) + 2.Cov(X_1, X_2)$$

$$V(X_1 - X_2) = V(X_1) + V(X_2) - 2.Cov(X_1, X_2)$$

Podemos também **combinar estimadores**, formando um **novo** estimador, por exemplo:

$$\breve{\theta} = \hat{\theta} + \widehat{\theta}'$$

Como saberemos se o novo estimador é tendencioso ou não?

Como fizemos anteriormente, aplicando a fórmula do viés:

$$b(\breve{\theta}) = E(e) = E(\breve{\theta}) - \theta$$

EXEMPLIFICANDO

Em relação ao estimador $\breve{\theta} = \hat{\theta} + \widehat{\theta'}$ para a média populacional μ, o seu viés é:

$$b(\breve{\theta}) = E(e) = E(\breve{\theta}) - \mu = E(\hat{\theta} + \widehat{\theta'}) - \mu$$

Sabendo que a esperança da soma é a soma das esperanças (propriedade), temos:

$$b(\breve{\theta}) = E(\hat{\theta}) + E(\widehat{\theta'}) - \mu$$

Para esse exemplo, vimos que o primeiro estimador é não tendencioso, logo, a sua esperança é $E(\hat{\theta}) = \mu$.

Já o segundo estimador é tendencioso com viés $b(\widehat{\theta'}) = -\mu$. Assim, a sua esperança pode ser calculada pela fórmula do viés:

$$b(\widehat{\theta'}) = E(\widehat{\theta'}) - \mu = -\mu$$

$$E(\widehat{\theta'}) = 0$$

Substituindo os valores de $E(\hat{\theta}) = \mu$ e $E(\widehat{\theta'}) = 0$ na equação $b(\breve{\theta}) = E(\hat{\theta}) + E(\widehat{\theta'}) - \mu$, temos:

$$b(\breve{\theta}) = \mu + 0 - \mu = 0$$

Como o viés é nulo, esse estimador é **não tendencioso**!

HORA DE PRATICAR!

(2019/Analista Censitário – *Adaptada*) É conhecido que o erro quadrático médio (EQM($\hat{\theta}$)) mede, em média, quão perto um estimador $\hat{\theta}$ chega ao valor real do parâmetro θ. Diante do exposto, julgue os itens seguintes:

I – EQM($\hat{\theta}$) é uma medida que combina viés e variância de um parâmetro;

II – para estimadores viesados, EQM($\hat{\theta}$) é igual à variância;

III – EQM($\hat{\theta}$) = $\hat{\theta}$ – Viés($\hat{\theta}$).

Comentários:

Em relação ao primeiro item, o EQM($\hat{\theta}$) combina viés e variância de um **estimador** (não do parâmetro populacional), logo o item I está incorreto.

Em relação ao segundo item, o EQM($\hat{\theta}$) é igual à variância para um estimador **não** viesado, logo o item II está incorreto.

Em relação ao terceiro item, a fórmula do erro quadrático médio é:

$$EQM(\hat{\theta}) = V(\hat{\theta}) + [Viés(\hat{\theta})]^2$$

Logo, o item III está incorreto.

Resposta: todos os itens errados.

(FGV/2022 - EPE – *Adaptada*) Considere uma amostra aleatória de n variáveis X_1, X_2, ..., X_n, normalmente distribuídas com média μ e variância σ^2. Considere o seguinte estimador da média populacional:

$$\bar{T} = \frac{1}{n^2}\sum_{i=1}^{n} X_i$$

Sobre as propriedades desse estimador, julgue o seguinte item.

O estimador é não-tendencioso.

Comentários:

Para um estimador não tendencioso, a sua esperança é igual ao parâmetro populacional estimado, no caso, a média μ. Vamos então calcular a esperança do estimador $\bar{T}$:

$$E(\bar{T}) = E\left(\frac{1}{n^2}\sum_{i=1}^{n} X_i\right)$$

Quando multiplicamos uma variável por uma constante, a sua esperança será multiplicada por essa constante (propriedade da esperança), logo:

$$E(\bar{T}) = \frac{1}{n^2}.E\left(\sum_{i=1}^{n} X_i\right)$$

Ademais, a esperança da soma de variáveis é igual à soma das esperanças (outra propriedade da esperança):

$$E(\bar{T}) = \frac{1}{n^2}.\sum_{i=1}^{n} E(X_i)$$

Considerando que cada variável da amostra segue a mesma distribuição da população, temos $E(X_i) = \mu$. Portanto, a soma $\sum_{i=1}^{n} E(X_i)$ para as n variáveis corresponde a $n.\mu$:

$$E(\bar{T}) = \frac{1}{n^2}.n.\mu = \frac{\mu}{n}$$

Que é diferente de μ. Logo, o estimador $\bar{T}$ é **tendencioso**.

Resposta: Errado.

(CESPE 2020/TJ-PA) Um estimador que fornece a resposta correta em média é chamado não enviesado. Formalmente, um estimador é não enviesado caso seu valor esperado seja igual ao parâmetro que está sendo estimado. Os possíveis estimadores para a média populacional (μ) incluem β, média de uma amostra, α, a menor observação da amostra, e π, a primeira observação coletada de uma amostra. Considerando essas informações, julgue os itens subsequentes.

I A média de uma amostra (β) é exemplo de um estimador enviesado para a média populacional (μ), pois seu valor esperado é igual à média populacional, ou seja, E(β) = μ.

II A menor observação da amostra (α) é um exemplo de estimador não enviesado, pois o valor da menor observação da amostra deve ser inferior à média da amostra; portanto, E(α) < μ.

III A primeira observação coletada de uma amostra equivale a tomar ao acaso uma amostra aleatória da população de tamanho igual a um e, portanto, é considerado um estimador não enviesado.

Assinale a opção correta.

a) Nenhum item está certo.

b) Apenas o item I está certo.

c) Apenas o item II está certo.

d) Apenas o item III está certo.

e) Todos os itens estão certos.

Comentários:

A questão trabalha com alguns estimadores: a média amostral β, que normalmente chamamos de $\bar{X}$; a menor observação da amostra, α; e a primeira observação, π. Em relação ao item I, sabemos que o valor esperado da média amostral é igual à média populacional, E(β) = μ, o que defini um estimador **não** enviesado. Logo, o item I está errado.

Em relação ao item II, o valor esperado da menor observação da amostra é inferior à média da amostral, E(α) < μ, o que define um estimador **enviesado**. Logo, o item II está errado. Em relação ao item III, de fato, a primeira observação pode ser considerada uma amostra com tamanho n = 1, ou seja, E(π) = μ. Logo, esse estimador é não enviesado. Logo, o item III está certo.

Gabarito: D.

Eficiência

Vimos que, para um estimador não tendencioso, o erro quadrático médio é igual à sua **variância**:

$$EQM(\widehat{\theta_{NT}}) = V(\widehat{\theta_{NT}})$$

Assim, a **variância** de um estimador não tendencioso (que é igual ao **erro** quadrático médio) está inversamente relacionada à **precisão da estimativa**: quanto **menor** a variância, **maior** a precisão.

Por sua vez, a precisão da estimativa está diretamente relacionada à eficiência do estimador. Ou seja, quanto **menor** a **variância** do estimador, **maior** será sua **eficiência**. Essa comparação também pode ser feita para estimadores com o **mesmo viés**.

Dizemos que um estimador é **eficiente** se for **não viesado** e apresentar a **menor variância possível**. Vale pontuar que tanto a **média amostral** quanto a **proporção amostral** são estimadores **eficientes**.

INDO MAIS FUNDO!

Um estimador é dito **assintoticamente eficiente** quando a sua matriz de variância-covariância assintótica não é maior que a de qualquer outro estimador, ou seja, quando a sua variância **converge mais rapidamente**.

Vamos com TUDO!

Existe um **limite inferior** para a variância de um estimador não tendencioso, chamado limite de **Cramér-Rao**, calculado a partir da função log-verossimilhança da amostra. Nenhum estimador não tendencioso terá uma variância inferior a esse limite.

Quando um estimador não tendencioso apresenta variância igual ao limite de Cramér-Rao, então podemos concluir que tal estimador é **eficiente**, pois apresenta a menor variância possível.

Por exemplo, para uma população com distribuição normal, o limite de Cramér-Rao para estimadores não tendenciosos da média populacional é $\frac{\sigma^2}{n}$, que é justamente a variância da média amostral, $\bar{X}$. Por se tratar de um estimador não tendencioso com a menor variância possível, concluímos que tal estimador é eficiente.

No entanto, nem sempre o limite de Cramér-Rao pode ser atingido por um estimador não tendencioso.

Por exemplo, para uma população com distribuição normal, o limite de Cramér-Rao para estimadores da variância populacional é $\frac{2.\sigma^4}{n}$, porém **não** há estimador não tendencioso com essa variância - todos apresentam variância maior que esse limite.

Em outras palavras, a variância de um estimador **eficiente**, isto é, de um estimador não tendencioso com a menor variância possível, **não** será **necessariamente** igual ao limite de **Cramér-Rao**.

(CESPE/2019 – TJ-AM) Com relação aos parâmetros estatísticos e suas estimativas, julgue o item que se segue.

Entre dois estimadores, A e B, com consistência, viés e demais características iguais, o estimador mais útil é aquele que possui menor variância.

Comentários:

Para estimadores com todas as demais características iguais, o melhor estimador será aquele com a menor variância, que está associada ao erro quadrático médio (EQM).

Gabarito: Certo

(2018 – Petrobras) Seja (Y_1 , Y_2 , Y_3) uma amostra aleatória simples extraída de modo independente de uma população com média μ e variância σ^2, ambas desconhecidas. Considere os dois estimadores da média da população definidos abaixo

$\widehat{\mu_1} = \frac{Y_1+2.Y_2+3.Y_3}{6}$ e $\widehat{\mu_2} = \frac{Y_1+Y_2+Y_3}{3}$

Relativamente a esses dois estimadores, conclui-se que:

a) apenas o primeiro estimador é não tendencioso.

b) apenas o segundo estimador é não tendencioso.

c) os dois estimadores são não tendenciosos, mas a eficiência não pode ser determinada sem a estimação da variância σ^2 .

d) os dois estimadores são não tendenciosos, mas o primeiro é mais eficiente por apresentar a variância inferior à do segundo estimador.

e) os dois estimadores são não tendenciosos, mas o segundo é mais eficiente por apresentar a variância inferior à do primeiro estimador.

Comentários:

O viés do estimador é a esperança do seu erro:

$$b(\hat{\theta}) = E(e) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

O viés do primeiro estimador é:

$$b(\widehat{\mu_1}) = E\left(\frac{Y_1 + 2.Y_2 + 3.Y_3}{6}\right) - \mu$$

Pelas propriedades da esperança, temos:

$$b(\widehat{\mu_1}) = \frac{1}{6}[E(Y_1) + 2.E(Y_2) + 3.E(Y_3)] - \mu$$

Sabendo que a distribuição da amostra é igual à distribuição da população, temos $E(Y_1) = E(Y) = \mu$:

$$b(\widehat{\mu_1}) = \frac{1}{6}[\mu + 2.\mu + 3.\mu] - \mu = \frac{1}{6}[6.\mu] - \mu = \mu - \mu = 0$$

Logo, o primeiro estimador é não viesado (alternativa B errada).

Similarmente, o viés do segundo estimador é dado por:

$$b(\widehat{\mu_2}) = E\left(\frac{Y_1 + Y_2 + Y_3}{3}\right) - \mu$$

$$b(\widehat{\mu_2}) = \frac{1}{3}[E(Y_1) + E(Y_2) + E(Y_3)] - \mu$$

$$b(\widehat{\mu_2}) = \frac{1}{3}[\mu + \mu + \mu] - \mu = \frac{1}{3}[3.\mu] - \mu = \mu - \mu = 0$$

Ou seja, o segundo estimador também não é viesado (alternativa A errada).

A eficiência do estimador não viesado é medida por sua variância. A variância do primeiro estimador é dada por:

$$V(\widehat{\mu_1}) = V\left(\frac{Y_1 + 2.Y_2 + 3.Y_3}{6}\right)$$

Pelas propriedades da variância, aplicável a variáveis independentes, temos:

$$V(\widehat{\mu_1}) = \frac{1}{6^2}[V(Y_1) + 2^2.V(Y_2) + 3^2.V(Y_3)] = \frac{1}{36}[V(Y_1) + 4.V(Y_2) + 9.V(Y_3)]$$

Sabendo que a distribuição da amostra é igual à distribuição da população, temos $V(Y_1) = V(Y) = \sigma^2$:

$$V(\widehat{\mu_1}) = \frac{1}{36}[\sigma^2 + 4.\sigma^2 + 9.\sigma^2] = \frac{1}{36}[14.\sigma^2] = \frac{7}{18}\sigma^2$$

Similarmente, a variância do segundo estimador é dada por:

$$V(\widehat{\mu_2}) = V\left(\frac{Y_1 + Y_2 + Y_3}{3}\right) = \frac{1}{3^2}[V(Y_1) + V(Y_2) + V(Y_3)] = \frac{1}{9}[V(Y_1) + V(Y_2) + V(Y_3)]$$

$$V(\widehat{\mu_2}) = \frac{1}{9}[\sigma^2 + \sigma^2 + \sigma^2] = \frac{1}{9}[3.\sigma^2] = \frac{3}{9}\sigma^2 = \frac{6}{18}\sigma^2$$

Como a variância do segundo estimador é menor que a variância do primeiro estimador, concluímos que o segundo estimador é mais eficiente.

Gabarito: E

Consistência

Para um estimador consistente $\widehat{\theta_C}$, as suas estimativas **convergem** para o parâmetro populacional θ com o **aumento do tamanho amostral**.

INDO MAIS FUNDO!

Matematicamente, a definição de consistência é:

$$\lim_{n \to \infty} P\left(\left|\widehat{\theta_C} - \theta\right| > \varepsilon\right) = 0$$

para algum valor $\varepsilon > 0$ pequeno.

Ou seja, quando o tamanho da amostra n tende a infinito, a probabilidade de um estimador consistente $\hat{\theta}_C$ **diferir** do parâmetro verdadeiro θ (em mais do que ε) é **nula**.

O estimador $\hat{\theta}$ será consistente se a sua esperança $E\left(\hat{\theta}\right)$ tende ao parâmetro populacional θ e sua variância $V\left(\hat{\theta}\right)$ tende a zero, quando o **tamanho da amostra** n tende ao **infinito**:

$$\lim_{n \to \infty} \boldsymbol{E\left(\hat{\theta}\right) = \theta}$$

$$\lim_{n \to \infty} \boldsymbol{V\left(\hat{\theta}\right) = 0}$$

Todos os estimadores que mencionamos anteriormente (média amostral $\bar{X}$, proporção amostral $\hat{p}$ e variância amostral s^2) são consistentes.

Ademais, a estimativa tendenciosa para a variância amostral $s_*^2 = \frac{\sum(X_i - \bar{X})^2}{\boldsymbol{n}}$ também é consistente.

Podemos deduzir se essas duas características estão presentes ou não, a partir da fórmula do **Erro Quadrático Médio** do estimador:

$$EQM\left(\hat{\theta}\right) = V\left(\hat{\theta}\right) + \left[b\left(\hat{\theta}\right)\right]^2$$

Vamos considerar o mesmo exemplo dado na prova da FGV/2017 (IBGE), em que a primeira expressão corresponde à variância do estimador, $V(\widehat{\theta_1})$, e a segunda expressão ao quadrado do viés do estimador, $[b(\widehat{\theta_1})]^2$:

$$EQM(\hat{\theta}_1) = \frac{3.\sigma^2}{n} + \left(\frac{\theta - 1}{n}\right)^2$$

Podemos deduzir que o viés é:

$$b(\hat{\theta}_1) = \frac{\theta - 1}{n}$$

Apesar de não ser nulo, na sua expressão, consta uma divisão por n. Por isso, conforme n aumenta, o viés diminui. No limite, quando n tende a infinito, o viés é igual a zero, portanto, pode-se dizer que o estimador é **assintoticamente não tendencioso**.

E a variância é:

$$V(\hat{\theta}_1) = \frac{3.\sigma^2}{n}$$

Nessa expressão, também consta uma divisão por n. Logo, quando n tende a infinito, a sua variância é igual a zero e o estimador é **assintoticamente eficiente**.

Como o estimador apresenta as duas características, podemos concluir que é **consistente**.

Se o estimador apresentar as duas propriedades assintóticas, então podemos concluir que é **consistente**.

Porém, se o estimador **não** apresentar essas propriedades assintóticas, **não** podemos concluir que o estimador é **inconsistente**.

Em outras palavras, um estimador pode ser assintoticamente tendencioso e/ou assintoticamente ineficiente e, ainda assim, ser consistente.

Ou seja, essas propriedades são condições **suficientes**, que nos permitem concluir que o estimador é consistente; porém, **não** são **necessárias** para que um estimador seja consistente.

(CESPE/2019 – TJ-AM) Com relação aos parâmetros estatísticos e suas estimativas, julgue o item que se segue.

Situação hipotética: A e B são dois estimadores não viciados e diferentes utilizados para estimar um mesmo parâmetro. A variância de A é menor que a variância de B. Assertiva: Em relação à consistência desses estimadores, é correto afirmar que o estimador A é mais consistente que o B.

Comentários:

Um estimador é dito consistente quando a esperança tende ao estimador e a variância tende a zero, quando a amostra cresce.

Logo, o fato de a variância de A ser menor do que a variância de B não nos permite afirmar que A é mais consistente que B.

Gabarito: Errado

(FGV/2019 – DPE-RJ – *Adaptada*) Sobre as principais propriedades dos estimadores pontuais, para pequenas e grandes amostras, julgue os itens a seguir.

I – Se $\lim_{n\to\infty} EQM(\hat{\theta}) = +\infty$, então $\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) \neq 0$.

II – Um estimador que seja assintoticamente tendencioso não poderá ser consistente.

Comentários:

Esses 2 itens trabalham com o fato de que as propriedades assintóticas são condições **suficientes**, porém **não necessárias** para a consistência do estimador.

O item I descreve um estimador cujo erro quadrático médio (EQM) cresce com o aumento do tamanho amostral n.

Como o EQM é a soma da variância com o quadrado do viés, $EQM(\hat{\theta}) = V(\hat{\theta}) + [b(\hat{\theta})]^2$, então **ou** o viés cresce com o aumento de n (tendencioso, mesmo em termos assintóticos) **ou** a variância cresce com o aumento de n (não é assintoticamente eficiente), ou ambos.

Portanto, não podemos concluir que o estimador é consistente, pois ele não segue uma ou ambas as propriedades assintóticas.

Por outro lado, também não podemos concluir que o estimador é inconsistente, pois é possível que um estimador não siga as propriedades assintóticas, mas ainda assim seja consistente.

Ou seja, não podemos afirmar que $\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) \neq 0$. Por isso, o item I está errado.

Similarmente, em relação ao item II, um estimador pode ser consistente, ainda que seja assintoticamente tendencioso.

Resposta: Ambos os itens errados.

ESQUEMATIZANDO

Estatística Suficiente: Captura **todas as informações** disponíveis na amostra para estimar o parâmetro populacional

Estimador Não Viesado (ENV), ou **Não Viciado**, ou **Não Tendencioso**: Esperança do estimador igual ao parâmetro; Esperança do erro igual a zero:

$$E(\hat{\theta}) = \theta \leftrightarrow E(e) = 0$$

$s_*^2 = \frac{\sum(X_i - \bar{X})^2}{\boldsymbol{n}}$ é **tendencioso**; $\bar{X}$, $\hat{p}$ e s^2 são **não tendenciosos**

Estimador Eficiente: apresenta a **menor variância** $V(\hat{\theta})$ possível.

$\bar{X}$ e $\hat{p}$ são **eficientes**

Estimador Consistente: estimativas convergem para parâmetro populacional se:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta, \lim_{n\to\infty} V(\hat{\theta}) = 0$$

Todos os estimadores que estudamos são consistentes.

Métodos de Estimação

Para obtermos bons estimadores, de acordo com os critérios que vimos, utilizamos **métodos de estimação**. Estudaremos agora alguns desses métodos.

Método dos Momentos

O Método dos Momentos se baseia nos **momentos teóricos e amostrais** das variáveis aleatórias. O k-ésimo **momento teórico** da variável X, denotado por μ_k, é:

$$\mu_k = E(X^k)$$

Se a variável for discreta, temos:

$$\mu_k = E(X^k) = \sum x^k . P(X = x)$$

Em particular, o **primeiro momento teórico** é igual à esperança da variável:

$$\mu_1 = E(X) = \sum x.P(X = x)$$

Se a variável for contínua, substituímos o somatório pela integral e a função de probabilidade $P(X = x)$ pela função densidade de probabilidade $f(x)$.

Pontue-se que é possível que distribuições **distintas** apresentem os **mesmos** momentos teóricos.

Os **momentos amostrais** são calculados de maneira análoga. O k-ésimo **momento amostral**, denotado por m_k, é a soma dos valores observados elevados a k, dividida pelo tamanho da amostra, n:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Em particular, o **primeiro momento amostral** ($k = 1$) consiste na **média amostral**:

$$m_1 = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

E o **segundo momento amostral** ($k = 2$) corresponde ao que definimos por $\overline{X^2}$:

$$m_2 = \frac{\sum_{i=1}^{n} X_i^2}{n} = \overline{X^2}$$

Para o exemplo das alturas, em que os $n = 5$ valores da amostra observada foram {1,65; 1,75; 1,8; 1,85; 1,95}, o primeiro momento amostral é igual a média amostral:

$$m_1 = \bar{X} = \frac{\sum X_i}{n} = \frac{1{,}65 + 1{,}75 + 1{,}8 + 1{,}85 + 1{,}95}{5} = \frac{9}{5} = 1{,}8$$

E o segundo momento amostral é a soma dos valores observados, elevados ao quadrado, divididos pelo tamanho da amostra:

$$m_2 = \overline{X^2} = \frac{\sum {X_i}^2}{n} = \frac{(1{,}65)^2 + (1{,}75)^2 + (1{,}8)^2 + (1{,}85)^2 + (1{,}95)^2}{5} = \frac{16{,}25}{5} = 3{,}25$$

Para calcular os estimadores pelo **método dos momentos (EMM)**, que podemos denotar por $\hat{\theta}_{MM}$, **igualamos** os **momentos teóricos (ou populacionais)** aos **momentos amostrais**:

$$\mu_k = m_k$$

Essa equação é feita para todos os parâmetros do modelo. Por exemplo, se houver 2 parâmetros, fazemos:

$$\mu_1 = m_1$$

$$\mu_2 = m_2$$

Em seguida, **resolvemos o sistema de equações**.

Por exemplo, o EMM para a **média populacional**, $\hat{\mu}_{MM}$, é dado por:

$$\hat{\mu}_{MM} = m_1 = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

TOME NOTA!

Ou seja, a **média amostral** é o **estimador** para a **média populacional**, obtido pelo **método dos momentos**.

Agora, vamos calcular o estimador de θ pelo método dos momentos para uma variável **uniforme** no intervalo $(0, \theta)$. Nesse caso, temos um único parâmetro, então basta calcularmos o primeiro momento:

$$\mu_1 = m_1 = \bar{X}$$

Para uma variável uniforme em um intervalo (a, b), a média populacional (ou esperança da variável) corresponde à média aritmética entre os extremos do intervalo:

$$\mu_1 = \frac{a+b}{2} = \frac{0+\theta}{2} = \frac{\theta}{2} = \bar{X}$$

$$\hat{\theta}_{MM} = 2.\bar{X}$$

Assim, o estimador para o parâmetro θ obtido pelo método dos momentos é o **dobro da média amostral**.

Em relação à variância, sabemos que $\sigma^2 = E(X^2) - [E(X)]^2$, ou seja, a variância populacional é a **diferença** entre o **segundo momento** e o **quadrado do primeiro momento**. Logo, o **EMM** para a **variância** é a **diferença** entre o **segundo momento amostral** e o **quadrado do primeiro momento amostral** (o qual corresponde à média amostral):

$$\hat{\sigma}_{MM}^2 = m_2 - (m_1)^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2$$

Com manipulações algébricas, obtemos como resultado:

$$\hat{\sigma}_{MM}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\boldsymbol{n}}$$

Entretanto, sabemos que esse estimador é **tendencioso**, ou seja, a esperança desse estimador é **diferente** do parâmetro populacional estimado (variância):

$$E(\hat{\sigma}_{MM}^2) \neq \sigma^2$$

Portanto, **nem sempre** o estimador obtido pelo método dos momentos apresenta **boas propriedades**. Inclusive, tais estimadores podem **não ser suficientes**.

Também é possível ter **mais de um estimador** para um **mesmo parâmetro** populacional. Para a distribuição de Poisson, por exemplo, temos $E(X) = V(X) = \lambda$. Assim, o parâmetro λ pode ser estimado tanto por $\hat{\mu}_{MM}$, quanto por $\hat{\sigma}^2_{MM}$, o que pode resultar em **estimativas bem diferentes**.

INDO MAIS FUNDO!

Ademais, é possível obter valores negativos para os parâmetros de uma distribuição binomial, segundo método dos momentos. Vejamos porquê:

Como há 2 parâmetros para essa distribuição, precisamos de 2 momentos. O primeiro momento populacional é:

$$\mu_1 = E(X) = n.p$$

Sabemos que a variância da binomial é $V(X) = n.p.(1-p)$. Lembrando que a variância pode ser calculada como $V(X) = E(X^2) - [E(X)]^2$, podemos calcular $E(X^2)$ para essa distribuição:

$$\mu_2 = E(X^2) = V(X) + [E(X)]^2 = n.p.(1-p) + [n.p]^2$$

$$\mu_2 = E = n.p.(1-p) + n^2.p^2$$

Para calcular os EEMs, fazemos $m_1 = \mu_1$ e $m_2 = \mu_2$. Sabendo que o primeiro momento amostral é $m_1 = \bar{X}$, então precisamos resolver o seguinte sistema de equações, supondo que foi obtida uma amostra de tamanho k:

$$\left\{\begin{matrix} \bar{X} = n.p \\ \left(\frac{1}{k}\right)\sum_{i=1}^{k} {X_i}^2 = n.p.(1-p) + n^2.p^2 \end{matrix}\right\}$$

Resolvendo esse sistema (que não vamos demonstrar aqui), obtemos as seguintes expressões para os estimadores dos parâmetros n e p:

$$\hat{n} = \frac{\bar{X}^2}{\bar{X} - \left(\frac{1}{k}\right)\sum_{i=1}^{k}(X_i - \bar{X})^2} \qquad \hat{p} = \frac{\bar{X}}{\hat{n}}$$

Como o valor de $\left(\frac{1}{k}\right)\sum_{i=1}^{k}(X_i - \bar{X})^2$ pode superar o valor de $\bar{X}$, é possível que o denominador de $\hat{n}$ seja negativo e assim obtermos tanto $\hat{n}$ quanto $\hat{p}$ negativos, o que é mais um indicativo de que o método dos momentos pode não resultar em bons estimadores.

(2017 – TRF-2ª Região) Considerando o método de estimação conhecido como Método dos Momentos, assinale a afirmativa INCORRETA.

a) Estimadores obtidos utilizando o k-ésimo momento não têm propriedades assintóticas para k ≥ 2.

b) O k-ésimo momento amostral é dado por $m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ e o k-ésimo momento populacional é dado por E(X^k).

c) Duas variáveis aleatórias com funções densidade de probabilidade distintas podem ter os mesmos momentos.

d) Este método iguala os momentos da população aos momentos amostrais. Os estimadores são obtidos resolvendo a equação ou o sistema de equações resultante.

Comentários:

Em relação à alternativa A, sabemos que para k = 2, temos $m_2 = \frac{\sum_{i=1}^{n} X_i^2}{n}$.

Conforme n aumenta, esse valor tende a zero. Logo, **não** podemos dizer que tais estimadores não possuem propriedades assintóticas. Assim, a alternativa A está INCORRETA.

Em relação à alternativa B, vimos que o k-ésimo momento amostral é, de fato, $m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ e que o k-ésimo momento populacional é $\mu_k = E(X^k)$. Logo, a alternativa B está CORRETA.

Em relação à alternativa C, de fato, distribuições distintas podem apresentar momentos iguais, logo a alternativa C está CORRETA.

Em relação à alternativa D, vimos que o método dos momentos iguala os momentos amostrais aos momentos teóricos (ou populacionais) aos momentos amostrais.

Em seguida, os estimadores são encontrados resolvendo a equação (quando houver apenas 1 parâmetro) ou o sistema de equações (quando houver mais parâmetros).

Gabarito: A

(2007 – TCE/RO – *Adaptada*) Seja (y_1, y_2, y_3, y_4) uma amostra aleatória independente e identicamente distribuída, de tamanho n = 4, extraída de uma população cuja característica estudada possui distribuição de probabilidade

$$f_Y(y,\theta) = \begin{Bmatrix} {}^{1}/_{\theta}, & 0 \le y \le \theta \\ 0, & caso\ contrário \end{Bmatrix}$$

Se a amostra selecionada foi (6, 2, 14, 8), a estimativa do parâmetro θ pelo método dos momentos é:

a) 7

b) 7,5

c) 14

d) 15

e) 14,5

Comentários:

Para calcular o estimador pelo método dos momentos, igualamos os momentos amostrais aos momentos teóricos:

$$\mu_k = m_k$$

Ou seja, o primeiro momento teórico, que é a esperança (ou média populacional), é igual à média amostral:

$$\bar{X} = \frac{\sum X}{n} = \frac{6 + 2 + 14 + 8}{4} = \frac{30}{4} = 7{,}5$$

Pela função densidade fornecida, podemos identificar que se trata de uma distribuição uniforme, no intervalo (0, θ).

A média dessa distribuição é a média aritmética dos extremos:

$$\mu = \frac{0 + \theta}{2} = \frac{\theta}{2}$$

Igualando os dois momentos, temos:

$$\frac{\widehat{\theta_{MM}}}{2} = 7{,}5$$

$$\widehat{\theta_{MM}} = 15$$

Resposta: D

Método da Máxima Verossimilhança

Para estimar o parâmetro populacional desejado, o **método da máxima verossimilhança** busca a estimativa para a qual a **probabilidade** de se obter os valores observados é a **maior possível**.

Em outras palavras, suponha que as observações tenham sido $X_1, X_2, \dots, X_n$.

Nesse caso, considerando o estimador de máxima verossimilhança, a probabilidade associada às observações $X_1, X_2, \dots, X_n$ é **maior** do que se considerássemos **qualquer outro estimador**.

Esse método terá uma função diferente, de acordo com a distribuição da população.

O estimador de máxima verossimilhança para um parâmetro θ é calculado a partir de uma função de **probabilidade conhecida** dependente desse parâmetro $f(\theta, x)$ e das observações de uma amostra aleatória $x_1, x_2, \ldots, x_n$.

Para calcular o estimador, precisamos da **função de máxima verossimilhança** $L(\theta, x_i)$, dada pelo produto da função de probabilidade, aplicada para cada resultado da amostra (função densidade conjunta):

$$L(\theta, x_i) = \prod_{i=1}^{n} f(\theta, x_i) = f(\theta, x_1) \times f(\theta, x_2) \times \ldots \times f(\theta, x_n)$$

O estimador de máxima verossimilhança será aquele que **maximizar** essa função. Para isso, **derivamos** a função e a igualamos a **zero**.

Normalmente, é mais simples trabalhar com o **logaritmo natural** da função $\ln L(\theta, x_i)$.

Para uma variável com **distribuição normal**, o estimador de máxima verossimilhança (EMV) para a média é:

$$\widehat{\mu_{MV}} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

Ou seja, para uma população com **distribuição normal**, a **média amostral** é o **estimador de máxima verossimilhança (EMV)** para a média populacional.

E o EMV para a **variância,** para uma **distribuição normal**, é:

$$\widehat{\sigma^2_{MV}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Esse é o mesmo estimador obtido pelo método dos momentos, que é **tendencioso**:

$$E(\hat{\sigma}^2_{MV}) \neq \sigma^2$$

Para uma variável aleatória com distribuição de **Poisson**, o estimador de máxima verossimilhança (EMV) também é a **média amostral**:

$$\widehat{\lambda_{MV}} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{\boldsymbol{X}}$$

Por exemplo, vamos supor que queiramos estimar o número médio (λ) de pessoas que chegam em um ponto de ônibus por hora, considerando que essa variável segue distribuição de Poisson.

Para isso, vamos considerar uma amostra de $n = 6$ horas (ou seja, vamos ficar 6 horas observando o ponto de ônibus e anotando o número de pessoas que chegam a cada hora).

Vamos supor que os valores observados sejam os seguintes:

{10, 12, 15, 10, 8, 5}

Para estimar o parâmetro λ pelo método de máxima verossimilhança, calculamos a média amostral:

$$\widehat{\lambda_{MV}} = \overline{\boldsymbol{X}} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{10 + 12 + 15 + 10 + 8 + 5}{6} = \frac{60}{6} = 10$$

Ou seja, a estimativa é que chegam, em média, 10 pessoas por hora nesse ponto de ônibus.

Para uma variável com distribuição **exponencial**, a média é o inverso do parâmetro: $E(X) = \frac{1}{\lambda}$, então a estimativa de máxima verossimilhança para o parâmetro λ é o inverso da média amostral:

$$\widehat{\lambda_{MV}} = \frac{1}{\overline{\boldsymbol{X}}}$$

Por exemplo, vamos supor que queiramos estimar o tempo de duração de uma lâmpada incandescente. Foram selecionadas $n = 5$ lâmpadas com as seguintes durações (em horas):

{50, 52, 46, 48, 54}

A média dessa amostra é:

$$\overline{\boldsymbol{X}} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{50 + 52 + 46 + 48 + 54}{5} = \frac{250}{5} = 50$$

Logo, o parâmetro estimado dessa distribuição, pelo método da máxima verossimilhança, é:

$$\widehat{\lambda_{MV}} = \frac{1}{\overline{\boldsymbol{X}}} = \frac{1}{50} = 0{,}02 \; por \; hora$$

Vamos efetivamente calcular o estimador de máxima verossimilhança para a distribuição **exponencial**. A f.d.p. dessa variável é:

$$f(\theta, x) = \theta . e^{-\theta . x}$$

A **função de máxima verossimilhança** $L(\theta, x_i)$ corresponde ao produto da função de probabilidade, aplicada para cada resultado da amostra:

$$L(\theta, x_i) = \prod_{i=1}^{n} f(\theta, x_i) = f(\theta, x_1) \times f(\theta, x_2) \times \ldots \times f(\theta, x_n)$$

$$L(\theta, x_i) = \prod_{i=1}^{n} \theta . e^{-\theta . x_i}$$

Para simplificar, vamos trabalhar com o **logaritmo natural** dessa função:

$$\ln L(\theta, x_i) = \ln\left(\prod_{i=1}^{n} \theta . e^{-\theta . x_i}\right)$$

O logaritmo do **produto** corresponde à **soma** dos logaritmos ($\ln a . b = \ln a + \ln b$):

$$\ln L(\theta, x_i) = \sum_{i=1}^{n} \ln\left(\theta . e^{-\theta . x_i}\right) = \sum_{i=1}^{n} \ln \theta + \sum_{i=1}^{n} \ln e^{-\theta . x_i}$$

Sabendo que o logaritmo da potência é igual ao produto do logaritmo ($\ln a^x = x . \ln a$) e que $\ln e = 1$, temos:

$$\ln L(\theta, x_i) = \sum_{i=1}^{n} \ln \theta + \sum_{i=1}^{n} -\theta x_i = n . \ln \theta - \theta \sum_{i=1}^{n} x_i$$

Agora, igualamos a **derivada** dessa função a **zero** (**equação de log-verossimilhança**):

$$\frac{\partial \ln L(\theta, x_i)}{\partial \theta} = 0$$

Pontue-se que a derivada de $\ln x$ é $\frac{1}{x}$:

$$n \times \frac{1}{\theta} - \sum_{i=1}^{n} x_i = 0$$

$$\frac{n}{\theta} = \sum_{i=1}^{n} x_i$$

$$\theta = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}}$$

Que é justamente o estimador da distribuição exponencial!

Para uma variável aleatória com distribuição de **Bernoulli**, o estimador de máxima verossimilhança para p é a **proporção amostral** (que segue a mesma definição de média amostral):

$$\widehat{p_{MV}} = \frac{\sum_{i=1}^{n} X_i}{n} = \hat{p}$$

Por exemplo, vamos supor que queiramos estimar a proporção p de peças defeituosas produzidas por uma fábrica. Para isso, foi selecionada uma amostra de $n = 10$ elementos que apresentou o seguinte resultado (em que 1 representa o defeito e 0 representa a peça boa):

{0, 0, 1, 0, 0, 0, 0, 1, 0, 0}

O estimador de máxima verossimilhança para essa distribuição é a proporção amostral:

$$\widehat{p_{MV}} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{0+0+1+0+0+0+0+1+0+0}{10} = \frac{2}{10} = 0{,}2$$

Para uma distribuição **geométrica**, que representa o número de tentativas até o primeiro sucesso, a média é o inverso da probabilidade de sucesso: $E(X) = \frac{1}{p}$, então o estimador de máxima verossimilhança para a proporção p é o inverso da média amostral:

$$\widehat{p_{MV}} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}$$

Vamos supor que 5 pessoas estejam jogando para ver quem consegue primeiro a face CARA lançando uma mesma moeda viciada (não equilibrada). Os resultados foram os seguintes:

{3, 4, 2, 5, 1}

A média amostral desses resultados é:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{3+4+2+5+1}{5} = \frac{15}{5} = 3$$

Logo, a probabilidade de obter a face CARA nessa moeda, estimada pelo método da máxima verossimilhança, é:

$$\widehat{p_{MV}} = \frac{1}{\bar{X}} = \frac{1}{3}$$

Para uma variável **uniforme** no intervalo $(0, \theta)$, o estimador de máxima verossimilhança para θ é o **maior valor observado na amostra**. (Observe como os estimadores são diferentes pelos diferentes métodos: lembra que o estimador de θ nesse caso pelo método dos momentos é o dobro da média amostral?)

(CESPE/2015 – Telebras) Considerando que os principais métodos para a estimação pontual são o método dos momentos e o da máxima verossimilhança, julgue o item a seguir.

Para a distribuição normal, o método dos momentos e o da máxima verossimilhança fornecem os mesmos estimadores aos parâmetros μ e σ.

Comentários:

Para uma distribuição normal, o estimador da média, segundo o método da máxima verossimilhança, é a média amostral, $\bar{X}$, o mesmo estimador obtido pelo método dos momentos. E o estimador da variância, segundo o método da máxima verossimilhança, para uma população normal, é $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, que também é o mesmo estimador obtido pelo método dos momentos.

Gabarito: Certo.

(CESPE/2015 – Telebras) Considerando que os principais métodos para a estimação pontual são o método dos momentos e o da máxima verossimilhança, julgue o item a seguir.

O estimador da máxima verossimilhança para a variância da distribuição normal é expresso por $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ e este estimador é não viciado.

Comentários:

De fato, o estimador $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ é o estimador de máxima verossimilhança para a variância de uma população com distribuição normal, porém esse estimador é viciado.

Gabarito: Errado.

(CESPE/2016 – TCE/PA) Uma amostra aleatória com n = 16 observações independentes e identicamente distribuídas (IID) foi obtida a partir de uma população infinita, com média e desvio padrão desconhecidos e distribuição normal. Tendo essa informação como referência inicial, julgue o seguinte item.

Caso, em uma amostra aleatória de tamanho n = 4, os valores amostrados sejam A = {2, 3, 0, 1}, a estimativa de máxima verossimilhança para a variância populacional será igual a $\frac{5}{3}$.

Comentários:

A estimativa de máxima verossimilhança para a variância de uma população normal, é:

$$\widehat{\sigma^2_{MV}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

Para calculá-la, precisamos primeiramente da média amostral, $\bar{X}$:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{2+3+0+1}{4} = \frac{6}{4} = 1{,}5$$

A estimativa é, portanto:

$$\widehat{\sigma^2_{MV}} = \frac{(2-1{,}5)^2 + (3-1{,}5)^2 + (0-1{,}5)^2 + (1-1{,}5)^2}{4} = \frac{(0{,}5)^2 + (1{,}5)^2 + (-1{,}5)^2 + (-0{,}5)^2}{4}$$

$$\widehat{\sigma^2_{MV}} = \frac{0{,}25 + 2{,}25 + 2{,}25 + 0{,}25}{4} = \frac{5}{4}$$

Observe que a questão exigiu conhecimento justamente da diferença entre o estimador de máxima verossimilhança e o estimador que utilizamos (não tendencioso).

Gabarito: Errado.

(FGV/2022 – SEFAZ/ES) Uma amostra aleatória simples de tamanho 4 de uma população normalmente distribuída forneceu os seguintes dados:

2,1 3,8 3,1 3,0

As estimativas de máxima verossimilhança da média e da variância populacionais são respectivamente

a) 3,0 e 0,486

b) 2,8 e 0,386

c) 2,8 e 0,535

d) 3,0 e 0,544

e) 3,0 e 0,365

Comentários:

Essa questão trabalha com o método de máxima verossimilhança para estimar a média e a variância de uma população com distribuição normal.

Em relação à média, temos a média amostral:

$$\bar{X} = \frac{\sum x}{n} = \frac{2{,}1 + 3{,}8 + 3{,}1 + 3{,}0}{4} = \frac{12}{4} = 3$$

Já a estimativa de máxima verossimilhança para a variância é o estimador tendencioso, em que dividimos a soma dos quadrados dos desvios por n, e não por $n-1$:

$$\widehat{\sigma^2_{MV}} = \frac{\sum(x_i - \bar{X})^2}{\boldsymbol{n}}$$

Para calcular o numerador, vamos construir uma tabela com os desvios:

x_i	2,1	3,8	3,1	3,0
$x_i - \bar{X}$	-0,9	0,8	0,1	0
$(x_i - \bar{X})^2$	0,81	0,64	0,01	0

E o numerador é a soma dos valores da última linha:

$$\sum(x_i - \bar{X})^2 = 0{,}81 + 0{,}64 + 0{,}01 + 0 = 1{,}46$$

Para calcular o estimador da variância, dividimos esse resultado por $n = 4$:

$$\widehat{\sigma_{MV}^2} = \frac{1{,}46}{4} = 0{,}365$$

Gabarito: E

(FGV/2017 – IBGE) Seja X uma variável aleatória com função de probabilidade dada por $P(X = x) = p.(1 - p)^{x-1}$ para x = 1, 2, 3,..., onde p é um parâmetro desconhecido. Dispondo de uma amostra de tamanho n, x_1, x_2, x_3,...., x_n, o estimador de Máxima Verossimilhança de p é:

a) $\hat{p} = \sum x_i$

b) $\hat{p} = \frac{1}{\sum x_i}$

c) $\hat{p} = \frac{\sum x_i}{n}$

d) $\hat{p} = \frac{n}{\sum x_i}$

e) $\hat{p} = \sqrt[n]{\sum x_i}$

Comentários:

Podemos observar que a função de probabilidade fornecida na questão é de uma variável com distribuição geométrica. Para essa variável, o estimador de máxima verossimilhança é o inverso da média amostral:

$$\widehat{p_{MV}} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}$$

Gabarito: D

(CESPE/2013 – MPU) Suponha que x_1, ..., x_n seja uma sequência de cópias independentes retiradas de uma distribuição com função densidade de probabilidade $f(x) = \alpha.x.e^{\frac{-\alpha x^2}{2}}$, em que x ≥ 0 e $\alpha > 0$ é seu parâmetro. Com base nessas informações, julgue o item a seguir.

Supondo que (x_1, ..., x_5) = (3, 4, 4, 6, 6), a estimativa de máxima verossimilhança do parâmetro α é inferior a 1/10.

Comentários:

O primeiro passo é calcular a função de máxima verossimilhança $L(\alpha, x_i)$, dada pelo produto da função de probabilidade, aplicada para cada resultado da amostra:

$$L(\theta, x_i) = \prod_{i=1}^{n} f(\alpha, x_i) = f(\alpha, x_1) \times f(\alpha, x_2) \times \ldots \times f(\alpha, x_n)$$

Para o nosso caso, temos:

$$L(\alpha, x_i) = \prod_{i=1}^{n} \alpha.x_i.e^{\frac{-\alpha x^2}{2}}$$

Agora, calculamos o logaritmo natural dessa função:

$$\ln L(\alpha, x_i) = \ln\left(\prod_{i=1}^{n} \alpha . x_i . e^{\frac{-\alpha x^2}{2}}\right)$$

O logaritmo do produto corresponde à soma dos logaritmos ($\ln a.b = \ln a + \ln b$):

$$\ln L(\alpha, x_i) = \sum_{i=1}^{n} \ln\left(\alpha . x_i . e^{\frac{-\alpha x^2}{2}}\right) = \sum_{i=1}^{n} \ln \alpha + \sum_{i=1}^{n} \ln x_i + \sum_{i=1}^{n} \ln e^{\frac{-\alpha x^2}{2}}$$

Sabendo que o logaritmo da potência é igual ao produto do logaritmo ($\ln a^x = x.\ln a$) e que $\ln e = 1$:

$$\ln L(\alpha, x_i) = \sum_{i=1}^{n} \ln \alpha + \sum_{i=1}^{n} \ln x_i + \sum_{i=1}^{n} \frac{-\alpha x^2}{2} \ln e = n.\ln \alpha + \sum_{i=1}^{n} \ln x_i - \frac{\alpha}{2} \sum_{i=1}^{n} x^2$$

E a derivada dessa função em relação a α é:

$$\frac{\partial \ln L(\alpha, x_i)}{\partial \alpha} = n \times \frac{1}{\alpha} - \frac{1}{2} \times \sum_{i=1}^{n} x^2 = \frac{n}{\alpha} - \frac{\sum_{i=1}^{n} x^2}{2}$$

Agora, essa função a **zero** (**equação de log-verossimilhança**):

$$\frac{n}{\alpha} - \frac{\sum_{i=1}^{n} x^2}{2} = 0$$

$$\frac{n}{\alpha} = \frac{\sum_{i=1}^{n} x^2}{2}$$

$$\alpha = \frac{2n}{\sum_{i=1}^{n} x^2}$$

Encontramos o estimador de máxima verossimilhança para α. Substituindo os resultados da amostra descrita no enunciado, temos:

$$\alpha = \frac{2 \times 5}{3^2 + 4^2 + 4^2 + 6^2 + 6^2} = \frac{10}{9 + 16 + 16 + 36 + 36} = \frac{10}{113}$$

Que é inferior a 1/10.

Gabarito: Certo.

Método de Mínimos Quadrados

O Método de Mínimos Quadrados busca a estimativa $\widehat{\theta_{MQ}}$ que resulta no **menor valor** para o **quadrado das diferenças** entre os valores observados X_i e o estimador $\widehat{\theta_{MQ}}$:

$$\text{Min} \sum_{i=1}^{n} \left(X_i - \widehat{\theta_{MQ}}\right)^2$$

Ou seja, o método busca **minimizar o erro quadrático total da amostra**. Assim, $\widehat{\theta_{MQ}}$ é chamado de Estimador de Mínimos Quadrados (EMQ) para o parâmetro populacional θ. Esse método é utilizado quando **não se conhece o tipo de distribuição da variável**.

TOME NOTA!

O **estimador de mínimos quadrados** para a **média populacional**, $\widehat{\mu_{MQ}}$, é igual à **média amostral**:

$$\widehat{\mu_{MQ}} = \bar{X}$$

Similarmente, o **estimador de mínimos quadrados** para a **proporção populacional**, $\widehat{p_{MQ}}$, é a **proporção amostral**:

$$\widehat{p_{MQ}} = \hat{p}$$

RESUMINDO

A **média amostral** é o estimador para a média populacional obtido pelos **métodos dos momentos** (EMM) e dos **mínimos quadrados** (EMQ). Se a população tiver **distribuição normal**, a média amostral também é o estimador obtido pelo **método da máxima verossimilhança** (EMV).

A **proporção amostral** é o estimador para a proporção populacional obtido pelos métodos dos **mínimos quadrados** (EMQ) e da **máxima verossimilhança** (EMV).

O estimador para a **variância** obtido pelos **métodos dos momentos** (EMM) e da **máxima verossimilhança** (EMV) é **tendencioso**.

Estimação Intervalar

Após obtermos uma estimativa para o parâmetro populacional desejado (estimação pontual), calculamos o intervalo dentro do qual esse parâmetro deve variar. Ou seja, na **estimação intervalar** (ou **estimação por intervalos**), a estimativa deixa de ser um ponto (isto é, um valor único) e passa a ser um **intervalo**. Esse intervalo, chamado **intervalo de confiança**, fornece uma noção de **precisão** da estimativa.

O **intervalo de confiança** é construído em torno da estimativa pontual $\hat{\theta}$, indicado da forma $\left(\hat{\theta} - E; \hat{\theta} + E\right)$ ou como $\hat{\theta} \pm E$. Esse intervalo indica que o parâmetro populacional θ deve estar entre o limite inferior, $\hat{\theta} - E$, e o limite superior $\hat{\theta} + E$. O valor E corresponde à **metade da amplitude** do intervalo, podendo ser chamado de **margem de erro**, **erro de precisão**, **erro máximo**.

Como assim o parâmetro "deve" estar no intervalo? É possível que o parâmetro θ esteja fora desse intervalo? Por se tratar de um intervalo construído em torno de uma variável aleatória, sim! Em Inferência Estatística, sempre convivemos com um nível de dúvida. Mas, felizmente, é possível dimensionar esse nível de dúvida.

Mais precisamente, atribuímos ao intervalo um **nível** (ou **grau**) **de confiança $1 - \alpha$**, indicado em forma percentual, por exemplo, 95%. A interpretação desse nível é a seguinte: repetindo o procedimento para a construção do intervalo muitas vezes, em $(1 - \alpha)\%$ (por exemplo, 95%) dessas vezes, o intervalo construído incluirá o parâmetro populacional.

Ou seja, o **nível de confiança** é uma **probabilidade**. Porém, **não** se trata de uma probabilidade de o **parâmetro populacional pertencer ao intervalo**, uma vez que o parâmetro populacional é **fixo** (embora seja desconhecido). O nível de confiança representa a **probabilidade de o intervalo**, o qual é construído a partir de variáveis aleatórias, **incluir o parâmetro populacional**.

Quanto **maior o nível de confiança** $(1 - \alpha)$, ou seja, quanto maior a certeza necessária de que o intervalo calculado inclui o parâmetro populacional desejado, maior terá que ser o **tamanho do intervalo**, se mantivermos todas as demais características iguais. Logo, maior será a **margem de erro** (isto é, a semi-amplitude do intervalo).

Se a margem de erro não puder aumentar, então teremos que aumentar o **tamanho da amostra**. Ou seja, para termos uma estimativa mais precisa (com menor margem de erro e/ou com maior nível de confiança), teremos que investigar um número maior itens. Veremos as fórmulas dessas relações adiante, mas é desejável entender essa lógica por trás delas.

E o valor de α? α é chamado de **nível de significância** e corresponde à probabilidade dos valores que **não estão no intervalo de confiança**. Falaremos bastante dele na aula de Teste de Hipóteses.

Nas próximas seções, veremos como construir o intervalo de confiança para os parâmetros populacionais (média, proporção e variância), a partir dos respectivos estimadores.

(CESPE/2019 – TJ/AM) Acerca de métodos usuais de estimação intervalar, julgue o item subsecutivo. Um intervalo de confiança de 95% descreve a probabilidade de um parâmetro estar entre dois valores numéricos na próxima amostra não aleatória a ser coletada.

Comentários:

Existem alguns erros neste item. A estimação, de modo geral, trabalha com amostras **aleatórias já coletadas**. Além disso, não se trata de uma probabilidade de o parâmetro estar entre dois valores, mas sim de os dois valores englobarem o parâmetro.

Gabarito: Errado.

(CESPE/2019 – TJ/AM) Acerca de métodos usuais de estimação intervalar, julgue o item subsecutivo. É possível calcular intervalos de confiança para a estimativa da média de uma distribuição normal, representativa de uma amostra aleatória.

Comentários:

De fato, é possível calcular intervalos de confiança para a média, a partir de uma amostra aleatória.

Gabarito: Certo.

(FGV/2019 – DPE-RJ – *Adaptada*) Para a aplicação de técnica de estimação por intervalos, há uma série de requisitos e recomendações. Sobre essas condições, julgue os seguintes.

I – A amplitude do intervalo varia positivamente com o grau de confiança e o tamanho da amostra.

II – A ideia da técnica é a da construção de um intervalo ao qual seja possível associar uma probabilidade, justamente aquela de que o parâmetro de interesse esteja nele contido.

III – É inquestionável que, antes da seleção da amostra, o grau de confiança é a probabilidade de o intervalo teórico conter de fato o verdadeiro valor do parâmetro de interesse.

Comentários:

Em relação ao item I, quanto maior o nível de confiança, maior será a amplitude do intervalo; porém, quanto maior o tamanho da amostra, mais precisa será a estimativa e, portanto, menor será a amplitude do intervalo (e a margem de erro). Logo, o item I está errado.

Em relação ao item II, não se trata de uma probabilidade de o parâmetro de interesse (que é fixo) estar contido no intervalo construído, mas sim de o intervalo construído conter o parâmetro de interesse.

Em relação ao item III, o grau de confiança $(1-\alpha)$, de fato, representa a probabilidade de o intervalo conter o parâmetro de interesse.

Resposta: Itens I e II Errados; Item III Certo.

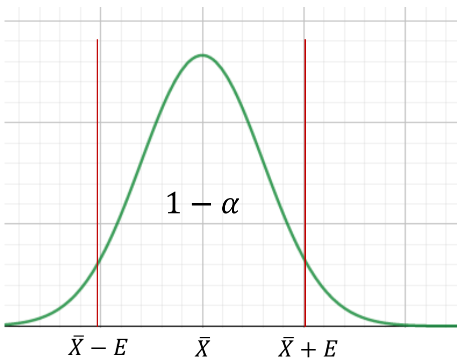
Intervalo de Confiança para a Média

Para estimarmos a média populacional, utilizamos a média amostral como estimador. Porém, para construirmos um intervalo de confiança em torno desse estimador, é necessário saber se a variância populacional é **conhecida ou não**.

População com Variância Conhecida

Se a população tiver **variância conhecida** σ^2 e **distribuição normal** (ou se apresentar outra distribuição, mas o tamanho da amostra for suficientemente **grande**), a média amostral $\bar{X}$ terá **distribuição normal**.

Sendo assim, vamos considerar a curva normal para construir um intervalo que delimite uma probabilidade de $\mathbf{1-\alpha}$ em **torno de** $\bar{\boldsymbol{X}}$. Ou seja, devemos encontrar os valores de X que delimitam uma área sob a curva normal, correspondente a $1-\alpha$:



Para encontrar os limites, devemos utilizar a tabela normal padrão. Logo, precisamos utilizar a transformação para a distribuição normal padrão Z (com média 0 e desvio padrão igual a 1):

$$z = \frac{valor\ procurado - \boldsymbol{média}\ da\ distribuição}{\boldsymbol{desvio\ padrão}\ da\ distribuição}$$

Aqui, os valores procurados são $\bar{X}+E$ e $\bar{X}-E$; a média da distribuição é $\bar{X}$ e o desvio padrão é o erro padrão de $\bar{X}$:

$$EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Ou seja, o limite superior do intervalo de confiança, $\bar{X}+E$, é calculado como:

$$z = \frac{\bar{X}+E\ -\bar{X}}{\frac{\sigma}{\sqrt{n}}} = \frac{E}{\frac{\sigma}{\sqrt{n}}}$$

$$\boxed{E = z.\frac{\sigma}{\sqrt{n}}}$$

O intervalo de confiança, $\bar{X} \pm E$, é, portanto:

$$\left(\bar{X} - z.\frac{\sigma}{\sqrt{n}}; \bar{X} + z.\frac{\sigma}{\sqrt{n}}\right)$$

Esses limites podem ser chamados de **valores críticos**.

E qual é o valor de z? Depende do **nível de confiança**.

Por exemplo, suponha um nível de confiança $1 - \alpha = 95\%$ (normalmente, a prova fornece o valor do nível de confiança mesmo, ou seja, o valor de $1 - \alpha$).

Para que toda a região indicada no gráfico acima delimite uma área de $1 - \alpha = 95\%$, então a área entre 0 e z deve ser a metade: $P(0 < Z < z) = 0{,}475$.

Z	...	0,05	0,06	0,07	...
...	...	...	...	...	...
1,8	...	0,4678	0,4686	0,4693	...
1,9	...	0,4744	0,475	0,4756	...
2	...	0,4798	0,4803	0,4808	...
...	...	...	...	...	...

Pela tabela da normal padrão, observamos que $z = 1{,}96$. Logo, se o nível de confiança for de 95%, o intervalo construído será:

$$\bar{X} \pm 1{,}96.\frac{\sigma}{\sqrt{n}}$$

Agora podemos visualizar o raciocínio que construímos no início desta seção: quanto **maior o nível de confiança**, maior será o valor de ***z*** e, portanto, **maior o tamanho do intervalo de confiança** (e a margem de erro), mantendo as demais características constantes.

E quanto **maior o tamanho da amostra analisada**, **menor será o tamanho do intervalo**, mantendo as demais características constantes. Além disso, também podemos observar que quanto **maior a variabilidade** da população, medida pelo desvio padrão (σ), **maior será o intervalo**.

Para exemplificar os cálculos, vamos supor que uma população com média desconhecida e variância igual $\sigma^2 = 9$ (portanto, o desvio padrão é $\sigma = \sqrt{\sigma^2} = \sqrt{9} = 3$). Para estimar a média, foi considerada uma população de 36 indivíduos. Nesse caso, a margem de erro, a um nível de $1 - \alpha = 95\%$ de confiança (em que $z = 1{,}96$), é dada por:

$$E = 1{,}96.\frac{\sigma}{\sqrt{n}} = 1{,}96.\frac{3}{\sqrt{36}} = 1{,}96.\frac{3}{6} = \frac{1{,}96}{2} = 0{,}98$$

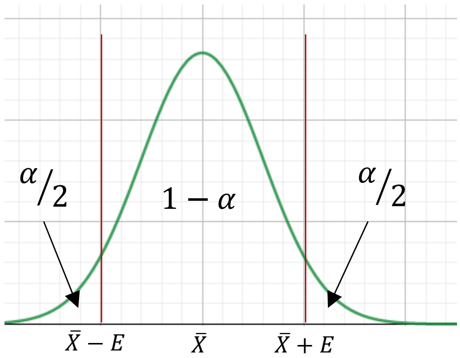
Supondo que a média amostral calculada foi $\bar{X} = 10$, o intervalo de confiança é dado por:

$$\bar{X} + E = 10 + 0{,}98 = 10{,}98$$

$$\bar{X} - E = 10 - 0{,}98 = 9{,}02$$

$$(9{,}02; 10{,}98)$$

Note que a probabilidade associada aos valores não incluídos no intervalo de confiança é complementar, ou seja, igual a α. Por se tratar de uma distribuição simétrica, a probabilidade associada aos valores superiores ao intervalo é ${}^{\alpha}/_{2}$ e aos valores inferiores é também ${}^{\alpha}/_{2}$:



Assim, é comum utilizar a notação $z_{\alpha/2}$ para indicar o limite do intervalo na distribuição normal padrão (no nosso exemplo, temos $z_{\alpha/2} = 1{,}96$).

Tamanho Amostral

Pode ser que a questão forneça, além do nível de confiança $(1 - \alpha)$, o valor do erro máximo, ou seja, o valor de E, e indague a respeito do **tamanho** necessário da **amostra** n. Para isso, podemos utilizar a mesma fórmula do erro que vimos há pouco:

$$E = z.\frac{\sigma}{\sqrt{n}}$$

Lembre-se que o valor de z é obtido a partir do nível de confiança desejado e que o **erro** corresponde à **metade da amplitude do intervalo de confiança**.

Reorganizando essa fórmula, temos:

$$\boxed{n = \left(z.\frac{\sigma}{E}\right)^2}$$

Podemos observar que, quanto **maior** o **nível de confiança** desejado (z), **maior** será o **tamanho amostral**; e quanto **maior** o **erro máximo** (E), **menor** será o **tamanho amostral**.

Para o exemplo que vimos anteriormente, em que obtivemos uma margem de erro de $E = 0{,}98$, com uma amostra de $n = 36$. Vamos supor que o erro máximo tolerável seja a metade, ou seja, $E = 0{,}49$. Considerando que os demais valores permaneceram constantes, ou seja, $z = 1{,}96$ e $\sigma = 3$, o tamanho da nova amostra n_2, necessário para obter a margem de erro desejada é:

$$n_2 = \left(z.\frac{\sigma}{E}\right)^2 = \left(1{,}96 \times \frac{3}{0{,}49}\right)^2 = (4 \times 3)^2 = (12)^2 = 144$$

Note que **o tamanho da amostra quadruplicou** para que a **margem de erro** fosse reduzida à **metade**, mantendo o mesmo nível de confiança, para a mesma população (portanto, o mesmo desvio padrão).

Na verdade, poderíamos saber que isso aconteceria, sem conhecer os dados do problema. Vejamos: a amostra inicial é dada por:

$$n_1 = \left(z.\frac{\sigma}{E}\right)^2$$

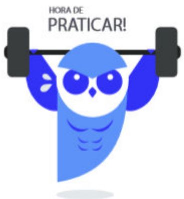
E a nova amostra necessária para que o erro se reduza à metade é:

$$n_2 = \left(z.\frac{\sigma}{E/2}\right)^2 = \left(z.\frac{\sigma}{E}.2\right)^2 = 4.\underbrace{\left(z.\frac{\sigma}{E}\right)^2}_{n_1} = 4.n_1$$

Pontue-se que tais fórmulas pressupõem uma população infinita ou amostras extraídas com reposição. Caso a **população seja finita e as amostras extraídas sem reposição**, então será necessário aplicar o **fator de correção para população finita**.

Para isso, devemos multiplicar a fórmula do erro pelo fator $\sqrt{\frac{N-n}{N-1}}$, o que influencia no **tamanho da amostra**:

$$E = z.\frac{\sigma}{\sqrt{n}}.\sqrt{\frac{N-n}{N-1}}$$



(CESPE 2020/TJ-PA) Uma equipe de engenheiros da qualidade, com vistas a estimar vida útil de determinado equipamento, utilizou uma amostra contendo 225 unidades e obteve uma média de 1.200 horas de duração, com desvio padrão de 150 horas.

Considerando-se, para um nível de confiança de 95%, z = 1,96, é correto afirmar que a verdadeira duração média do equipamento, em horas, estará em um intervalo entre

a) 1.190,00 e 1.210,00.

b) 1.185,20 e 1.214,80.

c) 1.177,50 e 1.222,50.

d) 1.180,40 e 1.219,60.

e) 1.174,20 e 1.225,80.

Comentários:

Vamos listar as informações do enunciado:

$Média\ amostral\ \ \rightarrow \bar{X} = 1.200$

$Escore\ associado\ ao\ nível\ de\ confiança \rightarrow z = 1{,}96$

$Desvio\ padrão \rightarrow \sigma = 150$

$Tamanho\ da\ amostra \rightarrow n = 225$

Agora, podemos apenas aplicar na fórmula do intervalo de confiança para a média (como a questão já forneceu o valor de z, não precisamos consultar a tabela para obtê-lo, a partir do nível de confiança):

$$\bar{X} \pm z \times \frac{\sigma}{\sqrt{n}}$$

$$1200 \pm 1{,}96 \times \frac{150}{\sqrt{225}}$$

$$1200 \pm 1{,}96 \times \frac{150}{15}$$

$$1200 \pm 1{,}96 \times 10$$

$$1200 \pm 19{,}6$$

Logo, o intervalo é $(1200 - 19{,}6 = 1180{,}4;\ 1200 + 19{,}6 = 1219{,}6)$

Gabarito: D.

(FGV/2022 – TJDFT) Uma grande amostra foi selecionada para estimar o tempo médio de tramitação de um tipo particular de ação em uma comarca. Essa amostra demonstrou que o intervalo bilateral de 95% de confiança para o tempo médio de tramitação estava entre 8 e 10 anos.

Com o objetivo de aumentar a precisão dessa estimativa, um estatístico resolveu diminuir a confiança para 85%. O novo intervalo de confiança passou a ser, aproximadamente, igual a:

a) $9 \pm 0{,}26$

b) $9 \pm 0{,}34$

c) $9 \pm 0{,}72$

d) $9 \pm 0{,}88$

e) $9 \pm 1{,}44$

Nessa prova, foi fornecida a tabela normal padrão da forma $P(Z > Z_0)$, parcialmente replicada a seguir.

Z_0	Segunda decimal de Z_0									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,00	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,10	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,20	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,30	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,40	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,50	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,60	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,70	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,80	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,90	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,00	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183

Comentários:

A questão informa que para um nível de 95% de confiança, o intervalo para estimar a média foi (8; 10), ou seja, 9 ± 1. O erro portanto é dado por:

$$E = z_{95\%}.\frac{\sigma}{\sqrt{n}} = 1$$

E a questão pede o novo intervalo, quando reduzimos o nível de confiança para 85%, sabendo que os demais parâmetros continuarão os mesmos.

Assim, precisamos comparar o valor de z para 95% de confiança e o valor de z para 85% de confiança.

A prova apresenta a tabela normal da forma $P(Z > Z_0)$. Para um nível de 95% de confiança, resta 2,5% abaixo do limite inferior do intervalo e 2,5% acima do limite superior do intervalo, logo, precisamos do valor de Z_0 associado a uma probabilidade $P(Z > Z_0) = 2,5\% = 0,025$. Pela tabela fornecida, temos que $Z_0 = 1,96 \cong 2$.

Para um nível de 85% de confiança, temos 7,5% (metade) abaixo do limite inferior e 7,5% acima do limite superior, logo, precisamos do valor de Z_0 associado a uma probabilidade $P(Z > Z_0) = 7,5\% = 0,075$. Pela tabela fornecida, temos que $Z_0 = 1,44$.

Agora, vamos calcular a razão entre os valores de z:

$$\frac{z_{85\%}}{z_{95\%}} \cong \frac{1{,}44}{2} = 0{,}72$$

Considerando que o erro é diretamente proporcional ao valor de z, a razão entre os erros segue essa mesma proporção:

$$\frac{E_{85\%}}{E_{95\%}} \cong 0{,}72$$

Sabendo que o erro a 95% de confiança é igual a 1, então o erro a 85% de confiança é:

$$E_{85\%} \cong 0{,}72 \times 1 = 0{,}72$$

Logo, o intervalo de confiança é da forma $9 \pm 0,72$.

Gabarito: C

(FCC/2014 - TRT/MA) Para responder às questões use, dentre as informações dadas abaixo, as que julgar apropriadas.

Se Z tem distribuição normal padrão, então: P(Z < 0,25) = 0,599, P(Z < 0,80) = 0,84, P(Z < 1) = 0,841, P(Z < 1,96) = 0,975, P(Z < 3,09) = 0,999

Considere X_1, X_2, ...X_n uma amostra aleatória simples, com reposição, da distribuição da variável X, que tem distribuição normal com média μ e variância 36. Seja X a média amostral dessa amostra. O valor de n para que a distância entre X e μ seja, no máximo, igual a 0,49, com probabilidade de 95% é igual a:

a) 256

b) 225

c) 400

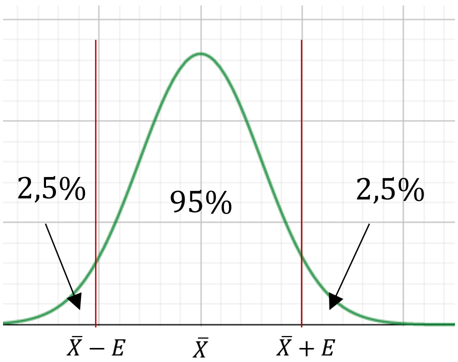
d) 144

e) 576

Comentários:

Ao dizer que a distância entre a média amostral (ou seja, estimativa da média populacional) e a média populacional seja no máximo igual a 0,49 com probabilidade de 95%, a questão forneceu o erro máximo $(\bar{X} - E; \bar{X} + E), E = 0{,}49$, e o nível de confiança de 95%.

Para que 95% da distribuição esteja no intervalo $(\bar{X} - E; \bar{X} + E)$, 2,5% da distribuição estará acima desse intervalo e 2,5% abaixo:



Assim, precisamos do valor de z cuja probabilidade P(Z < z) seja igual a 2,5% + 95% = 97,5%.

Pelos valores fornecidas observamos que z = 1,96. Sabendo que a variância é $\sigma^2 = 36$, ou seja, o desvio padrão é $\sigma = \sqrt{\sigma^2} = \sqrt{36} = 6$, tendo em vista que o erro é E = 0,49, temos:

$$n = \left(z.\frac{\sigma}{E}\right)^2$$

$$n = \left(1{,}96.\frac{6}{0{,}49}\right)^2 = (24)^2 = 576$$

Gabarito: E.

(FGV/2022 – MPE/SC) O tempo, em horas diárias, que homens com idades entre os 40 e 50 anos acessam redes sociais segue uma distribuição Normal com média 2,5 e desvio padrão 1,5. Para o mesmo grupo etário de mulheres, esse tempo segue também uma distribuição Normal com média 3 e desvio padrão 1. Serão retiradas duas amostras casuais e independentes, uma de homens e outra de mulheres.

O tamanho mínimo da amostra da população das mulheres que se pretende com probabilidade pelo menos 0,95 e cuja diferença em valor absoluto entre a média amostral e a média populacional não exceda 0,1 é, aproximadamente:

a) 20;

b) 100;

c) 250;

d) 385;

e) 500.

Comentários:

Essa questão trabalha com o tamanho amostral, para um intervalo de confiança para a média, com variância conhecida:

$$n = \left(\frac{z}{E}.\sigma\right)^2$$

Em que z é o valor da tabela normal padrão associado ao nível de confiança desejado; E é a margem de erro e σ é o desvio padrão.

A questão pede o tamanho mínimo para o tempo das mulheres, em que o desvio padrão é $\sigma = 1$. O enunciado informa, ainda, que a probabilidade do intervalo (nível de confiança) é de 95%, logo, z = 1,96. Ademais, informa que a diferença máxima entre a média amostral e a média populacional, que corresponde à definição de margem de erro, é E = 0,1. Substituindo esses dados na fórmula, temos:

$$n = \left(\frac{1{,}96}{0{,}1}.1\right)^2 = (19{,}6)^2 = 384{,}16$$

Logo, o menor tamanho amostral, que corresponde ao menor número inteiro maior que o valor calculado, é 385.

Gabarito: D

(CESPE 2016/TCE-PA) Considerando uma população finita em que a média da variável de interesse seja desconhecida, julgue o item a seguir.

Se uma amostra aleatória simples, sem reposição, for obtida de uma população finita constituída por N = 45 indivíduos, o fator de correção para população finita não será considerado na definição do tamanho da amostra para a estimação da média.

Comentários:

Quando uma população é finita, é feito um ajuste na equação para as médias amostrais. Esse ajuste é chamado fator de correção. Portanto, usamos sim o fator de correção.

Gabarito: Errado.

População com Variância Desconhecida

Sendo a variância populacional desconhecida, não podemos calcular a variância da média amostral como $V(\bar{X}) = \frac{V(X)}{n}$.

Então, precisamos **estimar a variância populacional**, a partir da **variância amostral**.

O **estimador não tendencioso para a variância** é:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Esse estimador vale para **populações infinitas OU amostras extraídas com reposição**.

Caso a população seja **finita**, de tamanho N, **E** a **amostra seja extraída sem reposição**, é necessário aplicar o fator de correção, multiplicando o resultado da variância amostral por $\frac{N-n}{N-1}$.

Com a estimativa para a variância populacional, calculamos a variância da média amostral (basta substituir σ^2 por s^2, na fórmula da variância da média amostral, que conhecemos):

$$V(\bar{X}) = \frac{s^2}{n}$$

Logo, o desvio padrão (ou erro padrão) da média amostral será:

$$D(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{\frac{s^2}{n}}$$

$$D(\bar{X}) = \frac{s}{\sqrt{n}}$$

O método para a construção do intervalo de confiança será similar, porém, em vez de considerarmos a distribuição normal, utilizaremos a distribuição **t-Student**, considerando $\boldsymbol{n-1}$ **graus de liberdade**.

Pontue-se que essa distribuição é similar à normal, porém mais achatada no centro e com caudas mais largas, ou seja, apresenta **maior variabilidade**.

Precisamos utilizar uma outra distribuição, que não a distribuição normal, porque agora a variância considerada deixou de ser um valor fixo e passou a ser também uma estimativa, o que justifica o uso de uma distribuição com maior variabilidade do que a normal.

Observe a fórmula da transformação, considerando os estimadores $\bar{X}$ e s^2:

$$\frac{x-\bar{X}}{\sqrt{s^2}}$$

Se a variância fosse fixa, teríamos a transformação para a normal padrão. Porém, estamos dividindo pela raiz da estimativa da variância $\sqrt{s^2}$.

Vale pontuar que o estimador da variância $\boldsymbol{s^2}$ tem **distribuição qui-quadrado**. Com isso em mente, compare a fórmula da transformação indicada acima com a definição da variável t-Student:

$$T = \frac{Z}{\sqrt{\frac{\boldsymbol{x_k^2}}{\boldsymbol{k}}}}$$

Percebeu a similaridade? Na fórmula da transformação acima, temos uma divisão pela raiz de uma distribuição qui-quadrado, assim como na definição da variável de t-Student.

Considerando que t_{n-1} representa o valor da tabela da distribuição t-Student padrão (com média igual a zero e desvio padrão igual a 1) para $n-1$ graus de liberdade, em que n é o tamanho da amostra, temos:

$$t_{n-1} = \frac{\bar{X} + E \; - \bar{X}}{\frac{s}{\sqrt{n}}} = \frac{E}{\frac{s}{\sqrt{n}}}$$

$$E = t_{n-1}.\frac{s}{\sqrt{n}}$$

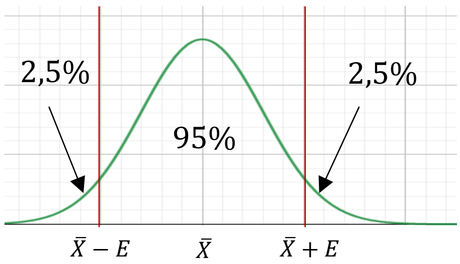
Ou seja, o erro é calculado assim como fizemos para a população com variância conhecida, apenas substituindo a variável da normal padrão pela variável t-Student. O intervalo de confiança será da forma $\left(\bar{X} - t_{n-1}.\frac{s}{\sqrt{n}} ; \bar{X} + t_{n-1}.\frac{s}{\sqrt{n}}\right)$.

O valor de $\boldsymbol{t_{n-1}}$ também dependerá do **nível de confiança** e do **tamanho da amostra $\boldsymbol{n}$.**

Por exemplo, suponha o mesmo nível de confiança do nosso exemplo anterior de $1 - \alpha = 95\%$ e uma amostra de tamanho $n = 5$. Logo, precisamos buscar o valor de t considerando $n - 1 = 4$ graus de liberdade. Abaixo, inserimos parte da tabela de t-Student, que apresenta os valores da função acumulada, ou seja, da forma $P(T < t)$.

ν	$t_{0,55}$	$t_{0,60}$	$t_{0,70}$	$t_{0,75}$	$t_{0,80}$	$t_{0,90}$	$t_{0,95}$	$t_{0,975}$	$t_{0,99}$	$t_{0,995}$
1	,158	,325	,727	1,000	1,376	3,08	6,31	12,71	31,82	63,66
2	,142	,289	,617	,816	1,061	1,89	2,92	4,30	6,96	9,92
3	,137	,277	,584	,765	,978	1,64	2,35	3,18	4,54	5,84
4	,134	,271	,569	,741	,941	1,53	2,13	2,78	3,75	4,60

Se a probabilidade associada ao intervalo de confiança é de 95%, então a probabilidade associada aos valores acima e abaixo desse intervalo é de 2,5%. Ou seja, a probabilidade associada aos valores abaixo do valor crítico superior é $P(X < \bar{X} + E) = 95\% + 2{,}5\% = 97{,}5\%$.



Assim, devemos buscar o valor de t_8 para o qual $P(T < t_4) = 0{,}975$. Pela tabela acima, temos $t_4 = 2{,}78$. Logo, o intervalo será da seguinte forma:

$$\bar{X} \pm 2{,}78.\frac{s}{\sqrt{5}}$$

O valor de s é a raiz quadrada da variância amostral observada. No nosso exemplo anterior, em que estimamos a variância da altura populacional, com base em uma amostra de 5 pessoas, calculamos a variância amostral em $s^2 = 0{,}0125$. Nesse caso, o desvio padrão é dado por:

$$s = \sqrt{s^2} = \sqrt{0{,}0125}$$

Portanto, a margem de erro para esse exemplo é:

$$E = 2{,}78.\frac{s}{\sqrt{5}} = 2{,}78.\frac{\sqrt{0{,}0125}}{\sqrt{5}} = 2{,}78 \times \sqrt{0{,}0025} = 2{,}78 \times 0{,}05 = 0{,}139$$

Para esse mesmo exemplo, a média encontrada foi $\bar{X} = 1{,}8$. Logo, o intervalo de confiança é:

$$\bar{X} + E = 1{,}8 + 0{,}139 = 1{,}939$$

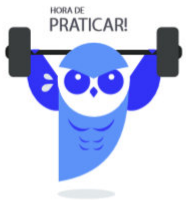
$$\bar{X} - E = 1{,}8 - 0{,}139 = 1{,}661$$

$$(1{,}661; 1{,}939)$$

Pontue-se que **o tamanho amostral** necessário para se obter determinada margem de erro, é calculado de maneira análoga ao caso da variância conhecida, ou seja:

$$E = t_{n-1}.\frac{s}{\sqrt{n}}$$

$$\boldsymbol{n = \left(t_{n-1}.\frac{s}{E}\right)^2}$$



(2019/SEFAZ-BA) Para obter um intervalo de confiança de 90% para a média p de uma população normalmente distribuída, de tamanho infinito e variância desconhecida, extraiu-se uma amostra aleatória de tamanho 9 dessa população, obtendo-se uma média amostral igual a 15 e variância igual a 16. Considerou-se a distribuição t de *Student* para o teste unicaudal tal que a probabilidade P(t - t_0) = 0,05, com n graus de liberdade. Com base nos dados da amostra, esse intervalo é igual a

Dados:

n	7	8	9	10	11
$t_{0,05}$	1,90	1,86	1,83	1,81	1,80

a) (12,56; 17,44)

b) (13,76; 16,24)

c) (12,47; 17,53)

d) (12,59; 17,41)

e) (12,52; 17,48)

Comentários:

Pelo enunciado, sabemos que o tamanho amostral é $N = 9$. Logo, temos $N - 1 = 8$ graus de liberdade. Pela tabela fornecida, para 8 de graus de liberdade, observamos que $t = 1{,}86$. Considerando, ainda, que a variância é $s^2 = 16$, logo $s = \sqrt{s^2} = \sqrt{16} = 4$, temos:

$$E = t.\frac{s}{\sqrt{N}} = 1{,}86 \times \frac{4}{3} = 2{,}48$$

Assim, o intervalo de confiança para a média $\bar{X} = 15$ é:

$$\bar{X} - E = 15 - 2{,}48 = 12{,}52$$

$$\bar{X} + E = 15 + 2{,}48 = 17{,}48$$

Gabarito: E.

(FGV/2022 - TJDFT) Em um modelo de simulação de uma fila com apenas um servidor para atendimento, foram realizadas 9 replicações para determinar o número médio de pessoas em fila. Os resultados obtidos para cada replicação estão no quadro a seguir.

Replicação	1	2	3	4	5	6	7	8	9	Média	Desvio padrão	Variância
Média de pessoas na fila	3,8	3,5	4,5	1,4	2	1,5	1,1	3,2	2,5	2,61	1,20	1,44

O intervalo bilateral de confiança de 95% para a média é, aproximadamente:

a) (1,83; 3,39)

b) (1,73; 3,50)

c) (1,69; 3,53)

d) (0,33; 4,83)

e) (0,04; 5,12)

Nessa prova, foi fornecida a tabela de t-Student da forma P(T > t_0), parcialmente replicada a seguir.

Grau de liberdade	Área da cauda superior				
	0,1	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169

Comentários:

Essa questão trabalha com um intervalo de confiança para a média, com variância desconhecida:

$$\bar{X} \pm t.\frac{s}{\sqrt{n}}$$

Em que s é a estimativa para o desvio padrão. Pela tabela fornecida na questão, observamos que $\bar{X} = 2{,}61$, que $s = 1{,}2$ e $n = 9$ (logo, $\sqrt{n} = \sqrt{9} = 3$).

Agora, precisamos encontrar o valor de t. Um intervalo com 95% de confiança deixa 2,5% abaixo do limite inferior e 2,5% acima do limite superior. Logo, precisamos do valor de t_0 associada a uma probabilidade P(T > t_0) = 2,5% = 0,025.

Pela tabela de t-Student fornecida, observamos que, para n - 1 = **8 graus de liberdade** e **0,025 de área da cauda superior**, temos t_0 = 2,306 ≅ 2,3. Substituindo esses dados na fórmula do intervalo de confiança:

$$2{,}61 \pm 2{,}3.\frac{1{,}2}{3} = 2{,}61 \pm 2{,}3 \times 0{,}4 = 2{,}61 \pm 0{,}92 = (1{,}69; 3{,}53)$$

Gabarito: C

Intervalo de Confiança para a Proporção

Para estimar a proporção populacional, p, utilizamos a proporção encontrada na amostra, $\hat{p}$. A variância desse estimador é dada por:

$$V(\hat{p}) = \frac{\hat{p}.\hat{q}}{n}$$

Lembre-se que $\hat{q} = 1 - \hat{p}$.

O desvio padrão é, portanto:

$$D(\hat{p}) = \sqrt{V(\hat{p})} = \sqrt{\frac{\hat{p}.\hat{q}}{n}}$$

Assim, a transformação para a normal padrão é:

$$z = \frac{\hat{p} + E \; - \hat{p}}{\sqrt{\frac{\hat{p}.\hat{q}}{n}}} = \frac{E}{\sqrt{\frac{\hat{p}.\hat{q}}{n}}}$$

$$\boxed{E = z.\sqrt{\frac{\hat{p}.\hat{q}}{n}}}$$

E o intervalo de confiança será da forma $\left(\hat{p} - z.\sqrt{\frac{\hat{p}.\hat{q}}{n}} ; \hat{p} + z.\sqrt{\frac{\hat{p}.\hat{q}}{n}}\right)$.

Vamos supor que tenhamos estimado a proporção amostral de defeito em $\hat{p} = 0{,}2$ (logo, $\hat{q} = 1 - \hat{p} = 0{,}8$), considerando uma amostra de $n = 100$ peças.

Nesse caso, o desvio padrão (ou erro padrão) é dado por:

$$\sqrt{\frac{\hat{p}.\hat{q}}{n}} = \sqrt{\frac{0{,}2 \times 0{,}8}{100}} = \frac{\sqrt{0{,}16}}{\sqrt{100}} = \frac{0{,}4}{10} = 0{,}04$$

Considerando um nível de confiança de 95% (z = 1,96), a margem de erro será:

$$E = z.\sqrt{\frac{\hat{p}.\hat{q}}{n}} = 1{,}96 \times 0{,}04 = 0{,}0784 \cong 0{,}8$$

Então, o intervalo de confiança para a proporção será aproximadamente:

$$\hat{p} + z.\sqrt{\frac{\hat{p}.\hat{q}}{n}} \cong 0{,}2 + 0{,}08 = 0{,}28$$

$$\hat{p} - z.\sqrt{\frac{\hat{p}.\hat{q}}{n}} \cong 0{,}2 - 0{,}08 = 0{,}12$$

$$(0{,}12; 0{,}28)$$

Tamanho Amostral

Também podemos determinar o tamanho amostral, dado um nível de confiança $(1 - \alpha)$ e um valor de erro máximo E. Reorganizando a fórmula do erro que acabamos de ver, temos (lembre-se que $\hat{q} = 1 - \hat{p}$):

$$\boxed{n = \left(\frac{z}{E}\right)^2 \hat{p}.\hat{q}}$$

Vamos supor que o erro máximo tolerável seja E = 0,04, mantendo os demais parâmetros iguais. Nesse caso, o tamanho da amostra n_2 necessário é:

$$n_2 = \left(\frac{1{,}96}{0{,}04}\right)^2 \times 0{,}2 \times 0{,}8 = (49)^2 \times 0{,}16 = 2401 \times 0{,}16 \cong 384$$

Vale pontuar que o **maior valor de n** é obtido com $\hat{p} = \hat{q} = 0{,}5$, considerando o mesmo nível de confiança (mesmo z) e o mesmo erro máximo E, pois essa proporção está associada ao **maior desvio padrão/variância**.

Assim, **mesmo sem ter uma estimativa para a proporção amostral**, é possível determinar um **valor máximo para o tamanho amostral** (também chamado de **valor seguro**), considerando $\hat{p} = \hat{q} = 0{,}5$ na fórmula acima.

Supondo, então, $z = 1{,}96$ e o erro máximo tolerável $E = 0{,}04$, o valor seguro para o tamanho amostral, n_*, sem termos uma estimativa para a proporção amostral é:

$$n_* = \left(\frac{1{,}96}{0{,}04}\right)^2 \times 0{,}5 \times 0{,}5 = (49)^2 \times 0{,}25 = 2401 \times 0{,}25 \cong 600$$

De modo geral, o valor de n será maior quanto mais próximo de 0,5 forem essas proporções.

Novamente, tal fórmula pressupõe uma população infinita ou amostras extraídas com reposição. Caso a população seja finita e as amostras extraídas sem reposição, então será necessário aplicar o fator de correção para população finita. Para isso, devemos multiplicar a fórmula do erro pelo fator $\sqrt{\frac{N-n}{N-1}}$:

$$E = z.\sqrt{\frac{\hat{p}.\hat{q}}{n}}.\sqrt{\frac{N-n}{N-1}}$$

(CESPE 2018/PF) Determinado órgão governamental estimou que a probabilidade p de um ex-condenado voltar a ser condenado por algum crime no prazo de 5 anos, contados a partir da data da libertação, seja igual a 0,25.

Essa estimativa foi obtida com base em um levantamento por amostragem aleatória simples de 1.875 processos judiciais, aplicando-se o método da máxima verossimilhança a partir da distribuição de Bernoulli.

Sabendo que P(Z < 2) = 0,975, em que Z representa a distribuição normal padrão, julgue o item que se segue, em relação a essa situação hipotética.

A estimativa intervalar 0,25 ± 0,05 representa o intervalo de 95% de confiança do parâmetro populacional p.

Comentários:

O intervalo de confiança de uma distribuição de proporção é dado por:

$$\hat{p} \pm Z_0 \times \sqrt{\frac{\hat{p} \times \hat{q}}{n}}$$

Vamos aos dados do problema:

$\hat{p} = 0{,}25 \rightarrow proporção\ amostral$

$\hat{q} = 1 - \hat{p} = 0{,}75$

$n = 1.875 \rightarrow tamanho\ da\ amostra$

$Z_0 = 2 \rightarrow nível\ de\ confiança\ de\ 95\%\ na\ tabela\ normal$

$$0{,}25 \pm 2 \times \sqrt{\frac{0{,}25 \times 0{,}75}{1.875}}$$

$$0{,}25 \pm 2 \times \sqrt{\frac{0{,}1875}{1.875}}$$

$$0{,}25 \pm 2 \times \sqrt{0{,}0001}$$

$$0{,}25 \pm 2 \times 0{,}01$$

$$0{,}25 \pm 0{,}02$$

Gabarito: Errado.

(2019/FMS) Uma pesquisa tem como finalidade conhecer a proporção de pessoas em Teresina que teriam interesse em frequentar uma nova franquia de lanchonete vinda do exterior. O empreendedor diz que só vale a pena a instalação da franquia, se pelo menos 10% da população tivesse interesse em frequentar o estabelecimento. Supondo que a proporção máxima da população não será maior que 30%, qual tamanho de amostra (aproximado) tal que a diferença entre a proporção populacional e proporção amostral não tenha um erro maior que três pontos percentuais, com uma confiança de 95%. Obs.: z_γ=1,96.

a) 897

b) 683

c) 700

d) 300

e) 654

Comentários:

O tamanho amostral é dado por:

$$n = \left(\frac{z}{E}\right)^2 \hat{p}.\hat{q}$$

O enunciado informa que $z = 1{,}96$ e $E = 0{,}03$. Devemos considerar, ainda, $\hat{p} = 0{,}3$. (Esse é o valor máximo para a proporção. Quanto mais próximos de 50%, maior será o valor de n. Logo, ao utilizar o valor de 30%, estamos calculando um valor seguro para o tamanho amostral). Sabendo que $\hat{q} = 1 - \hat{p} = 0{,}7$, temos:

$$n = \left(\frac{1{,}96}{0{,}03}\right)^2 \times 0{,}3 \times 0{,}7 \cong 897$$

Gabarito: A.

(FGV/2022 – MPE/SC) Uma empresa recebeu um lote muito grande, milhões de peças de refugo, e deseja saber quantas peças deverá examinar para estimar a proporção de itens defeituosos, de modo que o erro de estimação seja no máximo 2%. Será empregada uma seleção aleatória de itens onde cada um será classificado como defeituoso ou não defeituoso. Deseja-se extrair uma amostra aleatória de tamanho n.

Tendo como padrão um grau de confiança de 95%, o tamanho da amostra necessário para garantir o processo é:

a) 189;

b) 384;

c) 600;

d) 1681;

e) 2401.

Comentários:

Essa questão também trabalha com o tamanho amostral, para a estimação de proporções:

$$n = \left(\frac{z}{E}\right)^2 .\hat{p}.\hat{q}$$

Em que z é o valor da tabela normal padrão associado ao nível (ou grau) de confiança desejado; E é a margem de erro (ou erro máximo de estimação); $\hat{p}$ é a estimativa para a proporção de sucesso e $\hat{q}$ é a estimativa para a proporção de fracasso, sendo $\hat{q} = 1 - \hat{p}$.

O enunciado informa que o erro máximo é $E = 2\% = 0{,}02$; e que o grau de confiança é de 95%, logo, z = 1,96. A questão não informa a estimativa para a proporção de sucesso, logo, devemos considerar aquela que maximiza o tamanho da amostra, qual seja, $\hat{p} = \hat{q} = 0{,}5$. Substituindo esses dados na fórmula, temos:

$$n = \left(\frac{1{,}96}{0{,}02}\right)^2 \times 0{,}5 \times 0{,}5 = (98)^2 \times 0{,}25 = 2401$$

Gabarito: E

Intervalo de Confiança para a Variância

O estimador da variância s^2, multiplicado pelo fator $\left(\frac{n-1}{\sigma^2}\right)$, segue uma distribuição **qui-quadrado** com $\boldsymbol{n-1}$ **graus de liberdade**:

$$\mathcal{X}_{n-1}^2 = \left(\frac{n-1}{\sigma^2}\right) s^2$$

Como essa distribuição é **assimétrica**, utilizaremos a tabela da distribuição qui-quadrado para encontrar tanto o limite superior, $\mathcal{X}_{SUP}^2$, quanto o limite inferior, $\mathcal{X}_{INF}^2$, do intervalo de confiança:

$$\mathcal{X}_{INF}^2 < \left(\frac{n-1}{\sigma^2}\right) s^2 < \mathcal{X}_{SUP}^2$$

Isolando σ^2 nessa expressão, temos:

$$\frac{(n-1).\,s^2}{\mathcal{X}_{SUP}^2} < \sigma^2 < \frac{(n-1).\,s^2}{\mathcal{X}_{INF}^2}$$

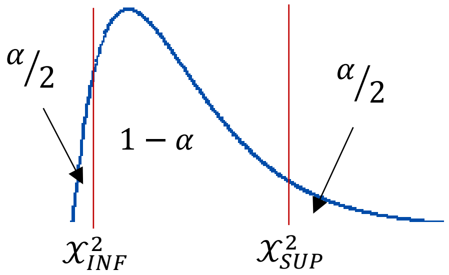
Ou seja, para o limite **inferior**, dividimos por $\boldsymbol{\mathcal{X}_{SUP}^2}$ e para o limite **superior** dividimos por $\boldsymbol{\mathcal{X}_{INF}^2}$.

Atente-se que $\mathcal{X}_{SUP}^2$ é um valor **maior** que $\mathcal{X}_{INF}^2$. Assim, quando dividimos por $\mathcal{X}_{SUP}^2$ (limite inferior) obtemos um valor menor do que quando dividimos por $\mathcal{X}_{INF}^2$ (limite superior).

Logo, o intervalo de confiança para a variância é da forma:

$$\boxed{\left(\frac{(n-1).s^2}{\mathcal{X}_{SUP}^2};\frac{(n-1).s^2}{\mathcal{X}_{INF}^2}\right)}$$

Novamente, os valores $\mathcal{X}_{SUP}^2$ e $\mathcal{X}_{INF}^2$ são obtidos a partir da **tabela**, dependendo do **nível de confiança** $1-\alpha$ desejado e do **tamanho da amostra**, conforme ilustrado abaixo.



Assim, como para as distribuições anteriores, o valor $\mathcal{X}_{INF}^2$ deixa $\frac{\alpha}{2}$ da distribuição abaixo; e o valor $\mathcal{X}_{SUP}^2$ deixa $\frac{\alpha}{2}$ da distribuição acima (a diferença é que agora não estamos mais lidando com uma distribuição simétrica):

$$P(\mathcal{X}^2 < \mathcal{X}_{INF}^2) = \frac{\alpha}{2}$$

$$P(\mathcal{X}^2 < \mathcal{X}_{SUP}^2) = 1 - \frac{\alpha}{2}$$

Para ilustrar, vamos supor uma amostra de tamanho n = 5. Abaixo, replicamos os valores da tabela para n – 1 = 4 graus de liberdade:

$P(\mathcal{X}_4^2 < x)$	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
x	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86

Supondo um nível de confiança de $1-\alpha = 0{,}95$ (portanto, $\alpha = 0{,}05$), podemos observar na tabela que o valor de $\mathcal{X}_{INF}^2$ associado à probabilidade $P(\mathcal{X}^2 < \mathcal{X}_{INF}^2) = \frac{\alpha}{2} = 0{,}025$ é $\mathcal{X}_{INF}^2 = 0{,}48$; e o valor $\mathcal{X}_{SUP}^2$ associado à probabilidade $P(\mathcal{X}^2 < \mathcal{X}_{SUP}^2) = 1 - \frac{\alpha}{2} = 0{,}975$ é $\mathcal{X}_{SUP}^2 = 11{,}14$.

Considerando que a variância amostral seja $s^2 = 0{,}0125$, os limites do intervalo de confiança são:

$$\frac{(n-1).s^2}{\mathcal{X}_{SUP}^2} = \frac{4 \times 0{,}0125}{11{,}14} = \frac{0{,}05}{11{,}14} \cong 0{,}004$$

$$\frac{(n-1).s^2}{\mathcal{X}_{INF}^2} = \frac{4 \times 0{,}0125}{0{,}48} = \frac{0{,}05}{0{,}48} \cong 0{,}104$$

E o intervalo é, então:

$$(0{,}004; 0{,}104)$$

(FGV/2019 – DPE-RJ) Com o objetivo de produzir uma estimativa por intervalo para a variância populacional, realiza-se uma amostra de tamanho n = 4, obtendo-se, após a extração, os seguintes resultados:

$X_1 = 6$, $X_2 = 3$, $X_3 = 11$ e $X_4 = 12$

Informações adicionais:

$P\,(X^2{}_4 < 0{,}75\,) = 0{,}05$ $P\,(X^2{}_3 < 0{,}40\,) = 0{,}05$

$P\,(X^2{}_4 < 10{,}8\,) = 0{,}95$ $P\,(X^2{}_3 < 9\,) = 0{,}95$

Então, sobre o resultado da estimação, e considerando-se um grau de confiança de 90%, tem-se que:

a) $5 < \sigma^2 < 72$;

b) $8 < \sigma^2 < 180$;

c) $6 < \sigma^2 < 135$;

d) $4 < \sigma^2 < 22$;

b) $6 < \sigma^2 < 24$.

Comentários:

O intervalo de confiança para a variância é dado por:

$$\left(\frac{(n-1).s^2}{\mathcal{X}^2_{SUP}}; \frac{(n-1).s^2}{\mathcal{X}^2_{INF}}\right)$$

Primeiro, precisamos calcular a estimativa para variância s^2:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

O valor da média amostral é:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{6+3+11+12}{4} = \frac{32}{4} = 8$$

Logo, o estimador da variância é:

$$s^2 = \frac{(6-8)^2 + (3-8)^2 + (11-8)^2 + (12-8)^2}{4-1} = \frac{(-2)^2 + (-5)^2 + (3)^2 + (4)^2}{3}$$

$$s^2 = \frac{4+25+9+16}{3} = \frac{54}{3} = 18$$

Sabendo que são n − 1 = 3 graus de liberdade e um intervalo de confiança de 95%, temos $\mathcal{X}^2_{INF} = 0{,}4$ e $\mathcal{X}^2_{SUP} = 9$, logo o intervalo de confiança é:

$$Inf = \frac{3 \times 18}{9} = 6$$

$$Sup = \frac{3 \times 18}{0{,}4} = 135$$

Gabarito: C.

ESQUEMATIZANDO

Estimação Intervalar

Intervalo para a Média – Variância conhecida (distribuição normal)**:**

Margem de Erro: $E = z.\frac{\sigma}{\sqrt{n}}$; **Tamanho amostral:** $n = \left(z.\frac{\sigma}{E}\right)^2$

Intervalo para a Média – Variância desconhecida (distribuição t-Student)**:**

Margem de Erro: $E = t_{n-1}.\frac{s}{\sqrt{n}}$; **Tamanho Amostral:** $n = \left(t.\frac{s}{E}\right)^2$

Intervalo para a Proporção:

Margem de Erro $E = z.\sqrt{\frac{\hat{p}.\hat{q}}{n}}$; **Tamanho Amostral:** $n = \left(\frac{z}{E}\right)^2 \hat{p}.\hat{q}$

Intervalo para a Variância: $\left(\frac{(n-1).s^2}{\chi^2_{SUP}}; \frac{(n-1).s^2}{\chi^2_{INF}}\right)$

INFERÊNCIA BAYESIANA

A **inferência Bayesiana** (ou **estatística Bayesiana**) é um assunto que vem recebendo mais atenção, tanto no mundo acadêmico quanto no mundo prático das análises Estatísticas. Como reflexo disto, também estão surgindo algumas questões em concursos sobre o tema.

Normalmente, estudamos a **inferência clássica** (ou **frequentista**) que considera que **toda a informação** disponível está **representada na amostra** e que o **parâmetro populacional** é **fixo**.

Já na **inferência Bayesiana**, além das **informações disponíveis na amostra**, que chamamos de **distribuição a posteriori**, incorporamos informações que sabemos ser verdadeiras, mas que não estarão representadas na amostra (**não observáveis**). Essas informações podem ser denominadas **subjetivas** e compõem a **distribuição a priori**.

Por exemplo, no tocante a eleições democráticas, sabemos que um candidato não terá 100% dos votos. Essa informação pode ser incorporada ao modelo para aprimorar a estimativa. Por outro lado, a inserção de informações que não são verdadeiras tornam o modelo **inapropriado**. Por exemplo, se for considerado que nenhum candidato terá 70% dos votos, mas se isso for de fato possível, a estimativa estará enviesada.

Na **inferência Bayesiana**, considera-se também que o **parâmetro populacional** a ser estimado é uma **variável aleatória** (e **não fixo** como na inferência clássica). Ademais, o que chamávamos de **intervalo de confiança** na inferência clássica passa a ser chamado de **intervalo de credibilidade**.

RESUMINDO

Inferência Clássica ou Frequentista	Inferência Bayesiana
• Parâmetro populacional é **fixo** • Toda a informação está **disponível** na amostra • Apenas distribuição a **posteriori** • Intervalo de **confiança**	• Parâmetro populacional é **variável aleatória** • Depende de informações **não observáveis** (além das informações da amostra) • Depende de uma **distribuição a priori** (além da distribuição a posteriori) • Intervalo de **credibilidade**

O nome dessa abordagem decorre do fato de ela ser baseada no Teorema de Bayes (da Teoria de Probabilidade):

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

A probabilidade $P(A|B)$ representa a **probabilidade a posteriori**; $P(A)$ representa a **probabilidade a priori**; e $P(B|A)$ representa a **função de verossimilhança**.

INDO MAIS FUNDO!

Mais especificamente, a Inferência Bayesiana considera que os dados observados na amostra $X_1, X_2, \dots, X_n$ possuem uma **distribuição condicionada** ao **parâmetro** de interesse (ou vetor de parâmetros) θ **aleatório** e não observável, $f(X_i|\theta)$.

Assim, a partir dos valores observados na amostra, define-se a **função de verossimilhança**, $L(\theta; x)$. E, aplicando a fórmula de Bayes, tem-se:

$$P(\theta|X = x) = \frac{L(\theta; x) \times P(\theta)}{P(X = x)}$$

Em que $P(\theta|X = x)$ corresponde à distribuição a posteriori do parâmetro θ (de interesse), dada a amostra observada; $P(\theta)$ é a distribuição a priori do parâmetro θ e $P(X = x)$ é chamada de distribuição preditiva de X, que representa a distribuição associada às observações.

O denominador tem pouca importância nos cálculos e um fundamento relevante do método é que a distribuição a posteriori do parâmetro é proporcional ao produto entre da **versomilhança** (que carrega a distribuição da amostra) com a **distribuição a priori** (que carrega a distribuição prévia, antes dos dados).

$$P(\theta|X = x) \propto L(\theta; x) \times P(\theta)$$

(2017 – TRF – 2ª Região – *Adaptada*) Sobre a abordagem bayesiana para estimar um parâmetro θ, julgue as afirmativas a seguir.

I – Uma distribuição de probabilidade é atribuída para esse parâmetro.

II – A definição da distribuição priori pode ser totalmente subjetiva.

Comentários:

Na inferência bayesiana, considera-se que o parâmetro de interesse não é fixo, mas sim, uma variável aleatória. Ou seja, de fato, atribui-se uma distribuição ao parâmetro de interesse e a afirmativa I está correta.

Ademais, considera-se uma distribuição a priori, com informações subjetivas, que não podem ser observada na amostra. Logo, a afirmativa II também está correta.

Resposta: I e II certas.

(CESPE/2019 – TJ-AM) A respeito dos diferentes métodos de estimação de parâmetros, julgue o item a seguir.

A estimação de parâmetros pelo método bayesiano independe da distribuição a priori utilizada.

Comentários:

Na inferência bayesiana, são consideradas informações não observáveis na amostra (subjetivas) que compõem a chamada distribuição a priori. Logo, o item está incorreto.

Gabarito: Errado.

RESUMO DA AULA

Estimadores

Erro Padrão é o **desvio padrão** (raiz quadrada da variância) do estimador

⇒ **Média amostral** $\bar{X}$**:** $E(\bar{X}) = \mu$, $V(\bar{X}) = \frac{\sigma^2}{n}$, $EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

⇒ **Proporção amostral** $\hat{p}$**:** $E(\hat{p}) = p$, $V(\hat{p}) = \frac{p.q}{n}$, $EP(\hat{p}) = \sqrt{\frac{p.q}{n}}$

⇒ Estimador da **variância amostral:** $s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$

Propriedades dos Estimadores

⇒ Suficiente (contempla todas as informações para estimar o parâmetro populacional)

⇒ Não Tendencioso (esperança do estimador é igual ao parâmetro populacional)

⇒ Eficiente (menor variância possível)

⇒ Consistente (estimativas convergem com o aumento do tamanho amostral)

Métodos de Estimação

⇒ Método dos Momentos: Resulta em $\bar{X}$ como estimador para a média; e em $\widehat{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{n}$ como estimador para a variância ($\widehat{\sigma^2}$ é tendencioso)

⇒ Método da Máxima Verossimilhança: Resulta em $\hat{p}$ como estimador para a proporção; para uma população normal, em $\bar{X}$ como estimador da média e $\widehat{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{n}$ como estimador para a variância ($\widehat{\sigma^2}$ é tendencioso)

⇒ Método dos Mínimos Quadrados: Resulta em $\bar{X}$ como estimador para a média; e em $\hat{p}$ como estimador para a proporção

Estimação Intervalar

Erro Máximo (**metade** da **amplitude do intervalo**)

⇒ Erro depende do nível de confiança desejado e do tamanho da amostra

⇒ Intervalo de confiança para população com variância conhecida: $\bar{X} \pm z.\frac{\sigma}{\sqrt{n}}$

⇒ Intervalo de confiança para população com variância desconhecida: $\bar{X} \pm t_{n-1}.\frac{\sigma}{\sqrt{n}}$