Insper

PROJETO 3

TEOREMA DO LIMITE CENTRAL (TLC)

DISTRIBUIÇÃO DE $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ E DE $\frac{\bar{X}-\mu}{S/\sqrt{n}}$

ESTUDO DA DISTRIBUIÇÃO DA MÉDIA AMOSTRAL ($\overline{X}$) VIA SIMULAÇÃO

OBJETIVO:

O objetivo deste projeto é que o aluno seja capaz de **compreender e explicar o resultado do Teorema do Limite Central (TLC) por meio de simulação**, uma vez que a demonstração teórica é complexa e requer conhecimento de outras teorias avançadas.

A seguir, apresentamos a definição da média amostral e o resultado descrito no Teorema do Limite Central.

Definição:

Seja X uma variável aleatória de interesse com distribuição qualquer, cuja média populacional e variância populacional são, usualmente, denotadas por $E(X) = \mu$ e $Var(X) = \sigma^2$.

Para estimar a **média populacional,** μ, utiliza-se o estimador **Média Amostral** definido por:

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

Teorema do Limite Central:

Para uma amostra aleatória $(X_1, \ldots, X_n)$ retirada de uma população cuja variável de interesse X tem média μ e variância σ^2, então a distribuição de da média amostral ($\overline{X}$) se aproxima de uma distribuição normal com média μ e variância σ^2/n, quando n tende ao infinito (suficientemente grande).

Parafraseando, seja X uma variável aleatória com média μ e variância σ^2 e seja $(X_1, \ldots, X_n)$ uma amostra aleatória retirada de X, então $\overline{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$, quando n tende ao infinito (suficientemente grande). Mesmo que X não tenha distribuição normal

O QUE DEVE SER FEITO:

Nesse projeto, você e sua equipe (até DUPLA) deverão fazer um estudo de **SIMULAÇÃO** para visualizar graficamente a distribuição da média amostral, calcular o valor médio e a variância dessas médias e interpretar esses resultados.

Para tanto, siga os seguintes passos:

1. Construir os valores da população:
    a. Definir uma distribuição teórica para variável X e definir valor(es) para representar o(s) parâmetro(s) da sua distribuição teórica escolhida.
        i. Por exemplo, se definir que $X \sim \text{Poisson}(\lambda)$, então, para essa distribuição teórica, tem-se que $E(X) = \text{Var}(X) = \lambda$. Logo, será necessário definir apenas um valor para λ, por exemplo, $\lambda = 4$.
    b. Sortear aleatoriamente um número muito grande de valores da sua distribuição teórica definida no item 1.a. Esses valores gerados aleatoriamente irão caracterizar todos os valores da sua população. Pesquise o método `.rvs` da sua distribuição no Python.
        i. Continuando o exemplo acima da Poisson, aqui seria `stats.poisson.rvs(mu=4, loc=0, size=1000000)`.
    c. Calcule a média e a variância desses valores, os quais irão representar, respectivamente, a média populacional μ e a variância populacional σ^2.
2. Construir a distribuição da média amostral:
    a. Inicialmente, considere $n = 2$.
    b. Sorteie n valores entre os muitos gerados no item 1b. Para tanto, pesquise o método `.random.choice` do `numpy`.
    c. Calcule a média amostral dos valores obtidos no item anterior.
    d. Repita os itens 2.b e 2.c por 10.000 vezes. Nesse caso, deverá ter 10.000 médias amostrais.
    e. Considerando essas 10.000 médias amostrais, construa um histograma com esses valores, calcule a média e a variância dessas médias amostrais.
    f. Interprete os resultados confrontando com o resultado aguardado via TLC.
3. Repita o item 2 para $n = 4$; $n = 20$; $n = 50$.
4. Faça uma conclusão geral explicando tudo o que aconteceu, o que é para ser visto e como se relaciona com o TLC. Faça essa explicação como se alguém pudesse aprender o TLC a partir de sua simulação.

Ainda, para compreender o resultado do TLC, é necessário observar que a sequência $(X_1, \dots, X_n)$ é chamada de **amostra aleatória**, ou seja, as variáveis dessa sequência são independentes e identicamente distribuídas a X.

Para tanto, siga os seguintes passos:

i. Considerando as simulações de $n = 50$, por exemplo, guarde sempre o terceiro e o décimo quinto valores sorteados em cada uma das 10.000 vezes. Ao final, deverá ter duas listas de 10.000 valores cada, uma que irá representar X_3 e outra que irá representar X_{15}.

ii. Faça o histograma apenas com os valores X_3.

iii. Faça o histograma apenas com os valores X_{15}.

iv. Compare com a distribuição de X com que escolheram no item 1 acima. Elas são iguais? Ou seja, X_3.e X_{15} são identicamente distribuídas a X?

v. Calcule a correlação entre X_3.e X_{15} e verifique se essa correlação é próxima de zero, indicando a independência entre elas.

ESTUDO DA DISTRIBUIÇÃO DE $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ E DE $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ VIA SIMULAÇÃO

OBJETIVO:

Outro objetivo deste projeto é que aluno seja capaz de **compreender a distribuição da padronização da média amostral utilizando o verdadeiro desvio padrão σ e a distribuição da padronização da média amostral utilizando o desvio padrão amostral S, ambos por meio de simulação**.

As distribuições dessas padronizações são importantes na Estatística e muito utilizadas em técnicas inferenciais. Assim, as distribuições que seguem tais padronizações são:

Resultado 1:

Para uma amostra aleatória $(X_1, \dots, X_n)$ retirada de uma população cuja variável de interesse X tem média μ e variância σ^2, então a distribuição de:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

quando n tende ao infinito (suficientemente grande).

Se X tiver distribuição normal, então o resultado acima segue uma distribuição normal padrão exata, não importando o tamanho amostral.

Resultado 2:

Para uma amostra aleatória $(X_1, \dots, X_n)$ retirada de uma população cuja variável de interesse X tem distribuição Normal com média μ e variância σ^2, então a distribuição de:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)},$$

sendo S o desvio padrão amostral da sequência $(X_1, \dots, X_n)$. Assim, a padronização da média amostral assumindo o desvio padrão amostral segue distribuição t-Student com $n-1$ graus de liberdade.

O QUE DEVE SER FEITO:

Aqui, devem visualizar graficamente a distribuição da padronização da média amostral considerando desvio padrão populacional e a distribuição da padronização considerando desvio padrão amostral.

Para tanto, siga os seguintes passos:

1. Construir os valores da população:
   a. Definir valores para a distribuição normal atribuída à variável X e definir valor(es) para representar o(s) parâmetro(s) da sua distribuição teórica escolhida.
      i. Por exemplo, se definir que $\mathrm{X} \sim \mathrm{Normal}(\mu, \sigma^2)$, então, para essa distribuição teórica, será necessário definir um valor para μ e outro para σ^2. Por exemplo, $\mu = 10$ e $\sigma^2 = 4$.
   b. Sortear aleatoriamente um número muito grande de valores da sua distribuição teórica definida no item 1.a. Esses valores gerados aleatoriamente irão caracterizar todos os valores da sua população. Pesquise o método `.rvs` da sua distribuição no Python.
      i. Continuando o exemplo acima da Normal, aqui seria `stats.norm.rvs(loc=10, scale=2, size=1000000)`.
   c. Calcule a média e a variância desses valores, os quais irão representar, respectivamente, a média populacional μ e a variância populacional σ^2.
2. Construir a distribuição da média amostral:
   a. Inicialmente, considere $\mathrm{n} = 4$.
   b. Sorteie n valores entre os muitos gerados no item 1b. Para tanto, pesquise o método `.random.choice` do `numpy`.
   c. Calcule a média amostral $\bar{x}$ e o desvio padrão amostral s dos valores obtidos no item anterior.
   d. Calcule $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ e de $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$.
   e. Repita os itens 2.b a 2.d por 10.000 vezes. Nesse caso, deverá ter 10.000 valores z e 10.000 valores t.
   f. Considerando esses 10.000 valores de z, construa um histograma com os valores de z e coloque a f.d.p. da distribuição normal padrão (N(0,1)).
   g. Considerando esses 10.000 valores de t, construa um histograma com os valores de t e coloque a f.d.p. da distribuição t-Student com $\mathrm{n} - 1$ graus de liberdade.
   h. Interprete os resultados confrontando com os Resultados 1 e 2.

3. Repita o item 2 para $\mathrm{n} = 20$; $\mathrm{n} = 50$; $\mathrm{n} = 300$.
4. Faça uma conclusão geral explicando todos os resultados obtidos com a simulação confrontados com os resultados teóricos.

ENTREGÁVEIS ESPERADOS E DATAS:

Item	Data	Descrição
Entrega 1	13/04/2017 Até às 15h30	1. Criar pasta Projeto 3 2. Publicar arquivo Python sobre o Projeto 3 Entrega 1 trabalhado em sala de aula. 3. Os dois integrantes do grupo devem fazer isso.
Entrega Final	18/04/2017 Até às 23h59	1. Publicar, na pasta Projeto 3, o arquivo Python sobre o Projeto 3 FINAL. 2. Os dois integrantes do grupo devem fazer isso.

RUBRICS DE AVALIAÇÃO DO OBJETIVO DE APRENDIZADO

Objetivo de aprendizado	Insatisfatório (I)	Em desenvolvimento (D)	Essencial (C)	Proficiente (B)	Avançado (A)
Compreender o resultado do TLC via simulação	Apresentou entregas insuficientes ou atrasadas fora do prazo combinado com o professor.	Escolheu um modelo para a variável X considerando valores para a média e a variância dessa variável. A simulação não só é incompleta como também não dá o menor indicativo que seria possível simular o TLC.	Apresenta simulações mas não estão conforme as instruções deste enunciado deixando apenas ideia do comportamento da distribuição da média amostral.	Simulou exatamente como no enunciado e explica basicamente o TLC	Simulou exatamente como no enunciado Procurou complementar a simulação com detalhes extra que reforcem a compreensão do TLC (fórmulas, gráficos, etc) Explica o TLC a ponto de se poder estudar pelo seu notebook.
Compreender a distribuição da padronização da média amostral utilizando desvio padrão populacional e utilizando o desvio padrão amostral	Apresentou entregas insuficientes ou atrasadas fora do prazo combinado com o professor.	A simulação não só é incompleta como também não dá o menor indicativo que seria possível compreender a distribuição das padronizações da média amostral.	Apresenta simulações mas não estão conforme as instruções deste enunciado deixando apenas ideia do comportamento da distribuição das padronizações.	Simulou exatamente como no enunciado e explica basicamente a padronização da média amostral em ambos os contextos.	Simulou exatamente como no enunciado Procurou complementar a simulação com detalhes extra que reforcem a compreensão das padronizações da média amostral e explica com clareza tais resultados..