CURSO DE ESTATÍSTICA - PARTE 2

Trabalho sobre Probabilidades, Amostragem e Estimações

Utilizando os conhecimentos adquiridos em nosso treinamento execute as tarefas abaixo. Siga o roteiro proposto e vá completando as células vazias.

DATASET DO PROJETO

Pesquisa Nacional por Amostra de Domicílios - 2015

A **Pesquisa Nacional por Amostra de Domicílios - PNAD** investiga anualmente, de forma permanente, características gerais da população, de educação, trabalho, rendimento e habitação e outras, com periodicidade variável, de acordo com as necessidades de informação para o país, como as características sobre migração, fecundidade, nupcialidade, saúde, segurança alimentar, entre outros temas. O levantamento dessas estatísticas constitui, ao longo dos 49 anos de realização da pesquisa, um importante instrumento para formulação, validação e avaliação de políticas orientadas para o desenvolvimento socioeconômico e a melhoria das condições de vida no Brasil.

Fonte dos Dados

https://ww2.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2015/microdados.shtm

Variáveis utilizadas

Renda

Rendimento mensal do trabalho principal para pessoas de 10 anos ou mais de idade.

Idade

Idade do morador na data de referência em anos.

Altura (elaboração própria)

Altura do morador em metros.

UF

Código	Descrição
11	Rondônia
12	Acre

Código	Descrição
13	Amazonas
14	Roraima
15	Pará
16	Amapá
17	Tocantins
21	Maranhão
22	Piauí
23	Ceará
24	Rio Grande do Norte
25	Paraíba
26	Pernambuco
27	Alagoas
28	Sergipe
29	Bahia
31	Minas Gerais
32	Espírito Santo
33	Rio de Janeiro
35	São Paulo
41	Paraná
42	Santa Catarina
43	Rio Grande do Sul
50	Mato Grosso do Sul
51	Mato Grosso
52	Goiás
53	Distrito Federal

Sexo

Código	Descrição
0	Masculino
1	Feminino

Anos de Estudo

Código	Descrição
1	Sem instrução e menos de 1 ano
2	1 ano
3	2 anos
4	3 anos
5	4 anos

Código	Descrição
6	5 anos
7	6 anos
8	7 anos
9	8 anos
10	9 anos
11	10 anos
12	11 anos
13	12 anos
14	13 anos
15	14 anos
16	15 anos ou mais
17	Não determinados
	Não aplicável

Cor

Código	Descrição
0	Indígena
2	Branca
4	Preta
6	Amarela
8	Parda
9	Sem declaração

Observação

Os seguintes tratamentos foram realizados nos dados originais:

1. Foram eliminados os registros onde a **Renda** era inválida (999 999 999 999);
2. Foram eliminados os registros onde a **Renda** era missing;
3. Foram considerados somente os registros das **Pessoas de Referência** de cada domicílio (responsável pelo domicílio).

Utilize a célula abaixo para importar as bibliotecas que precisar para executar as tarefas

Sugestões: pandas, numpy, scipy etc.

In [1]:
```
import pandas as pd
import numpy as np
from scipy.stats import norm
from scipy.special import comb
from scipy.stats import binom
```

Importe o dataset e armazene o conteúdo em uma DataFrame

In [2]:
```
dados = pd.read_csv("dados.csv")
```

Visualize o conteúdo do DataFrame

In [3]:
```
dados.head()
```

Out[3]:

	UF	Sexo	Idade	Cor	Anos de Estudo	Renda	Altura
0	11	0	23	8	12	800	1.603808
1	11	1	23	2	12	1150	1.739790
2	11	1	35	8	15	880	1.760444
3	11	0	46	2	6	3500	1.783158
4	11	1	47	8	9	150	1.690631

Problema A

Avaliando nosso dataset é possível verificar que a **proporção de homens** como chefes de domicílios é de quase **70%**. Precisamos **selecionar aleatoriamente grupos de 10 indivíduos** para verificar as diferenças entre os rendimentos em cada grupo. Qual a **probabilidade de selecionamos um grupo que apresente a mesma proporção da população**, ou seja, selecionarmos um grupo que seja **composto por 7 homens e 3 mulheres**?

Como tarefa extra, verifique a real proporção de homens e mulheres em nosso dataset (vimos como fazer isso em nosso primeiro curso de estatística).

Verifique que tipo de distribuição de probabilidade se encaixa neste experimento.

Solução

In [4]:
```
n = 10
p = 0.7
k = 7
probabilidade = binom.pmf(k, n, p)
print(f"{probabilidade:.8f}")
```

```
0.26682793
```

In [5]:
```
print(f"Homens: {(dados.Sexo.value_counts(normalize = True)[0] * 100):.2f}%")
print(f"Mulheres: {(dados.Sexo.value_counts(normalize = True)[1] * 100):.2f}%")
```

```
Homens: 69.30%
Mulheres: 30.70%
```

Distribuição Binomial

Problema B

Ainda sobre a questão anterior, **quantos grupos de 10 indivíduos** nós precisaríamos selecionar, de forma aleatória, para conseguir **100 grupos compostos por 7 homens e 3 mulheres**?

Lembre-se da forma de cálculo da média de uma distribuição binomial

Solução

In [6]:
```
n = 100 / probabilidade
print(f"{int(n.round())} grupos")
```

```
375 grupos
```

Problema C

Um cliente nos encomendou um estudo para avaliar o **rendimento dos chefes de domicílio no Brasil**. Para isso precisamos realizar uma nova coleta de dados, isto é, uma nova pesquisa de campo. Após reunião com o cliente foi possível elencar o seguinte conjunto de informações:

> A. O resultado da pesquisa precisa estar pronto em **2 meses**;
>
> B. Teremos somente **R$ 150.000,00** de recursos para realização da pesquisa de campo; e
>
> C. Seria interessante uma **margem de erro não superior a 10% em relação a média estimada**.

Em nossa experiência com estudos deste tipo, sabemos que o **custo médio por indivíduo entrevistado fica em torno de R$ 100,00**. Com este conjunto de fatos avalie e obtenha o seguinte conjunto de informações para passar ao cliente:

> 1. Para obter uma estimativa para os parâmetros da população (renda dos chefes de domicílio no Brasil), realize uma amostragem aleatória simples em nosso conjunto de dados. Essa amostra deve conter 200 elementos (utilize random_state = 101 para garantir que o mesmo experimento posso ser realizado novamente). Obtenha a média e o desvio-padrão dessa amostra.
> 2. Para a **margem de erro** especificada pelo cliente obtenha os **tamanhos de amostra** necessários para garantir os **níveis de confiança de 90%, 95% e 99%**.
> 3. Obtenha o **custo da pesquisa** para os três níveis de confiança.
> 4. Para o maior nível de confiança viável (dentro do orçamento disponível), obtenha um **intervalo de confiança para a média da população**.
> 5. Assumindo o **nível de confiança escolhido no item anterior**, qual **margem de erro** pode ser considerada utilizando todo o recurso disponibilizado pelo cliente?
> 6. Assumindo um **nível de confiança de 95%**, **quanto a pesquisa custaria ao cliente** caso fosse considerada uma **margem de erro de apenas 5%** em relação a média estimada?

Solução do item 1

Seleção de uma amostra aleatório simples

Lembre-se de utilizar *random_state = 101*

In [7]:
```
amostra_simples = dados.Renda.sample(200, random_state = 101)
amostra_simples.head()
```

Out[7]:
```
29042     480
62672     250
29973     788
```

```
22428    1680
55145    2500
Name: Renda, dtype: int64
```

In [8]:
```
media_amostra = amostra_simples.mean()
media_amostra
```

Out[8]: 1964.205

In [9]:
```
desvio_padrao_amostra = amostra_simples.std()
desvio_padrao_amostra
```

Out[9]: 3139.8855167452157

Dados do problema

In [10]:
```
recursos = 150000
custo_entrevista = 100
```

Solução do item 2

Obtenha a margem de erro

Lembre-se que a margem de erro deve estar na mesma unidade da variável que está sendo estudada (R$)

In [11]:
```
e = media_amostra * 0.1
print(f"A margem de erro é de R$ {e:.2f} para mais ou para menos.")
```

```
A margem de erro é de R$ 196.42 para mais ou para menos.
```

Tamanho da amostra ($1 - \alpha = 90\%$)

In [12]:
```
z = norm.ppf(0.5 + (0.9 / 2))
z
```

Out[12]: 1.6448536269514722

In [13]:
```
n_90 = (z * (desvio_padrao_amostra / e)) ** 2
n_90 = int(n_90.round())
print(f"Tamanho da amostra para nível de confiança de 90%: {n_90}")
```

```
Tamanho da amostra para nível de confiança de 90%: 691
```

Tamanho da amostra ($1 - \alpha = 95\%$)

In [14]:
```
z = norm.ppf(0.5 + (0.95 / 2))
z
```

Out[14]: 1.959963984540054

In [15]:
```
n_95 = (z * (desvio_padrao_amostra / e)) ** 2
n_95 = int(n_95.round())
print(f"Tamanho da amostra para nível de confiança de 95%: {n_95}")
```

```
Tamanho da amostra para nível de confiança de 95%: 982
```

Tamanho da amostra ($1 - \alpha = 99\%$)

In [16]:
```
z = norm.ppf(0.5 + (0.99 / 2))
z
```

Out[16]: 2.5758293035489004

In [17]:
```
n_99 = (z * (desvio_padrao_amostra / e)) ** 2
n_99 = int(n_99.round())
print(f"Tamanho da amostra para nível de confiança de 99%: {n_99}")
```

```
Tamanho da amostra para nível de confiança de 99%: 1695
```

Solução do item 3

Custo da pesquisa para o nível de confiança de 90%

In [18]:
```
custo_90 = n_90 * custo_entrevista
print(f"Custo da pesquisa para um nível de confiança de 90%: R$ {custo_90:,.2f}")
```

```
Custo da pesquisa para um nível de confiança de 90%: R$ 69,100.00
```

Custo da pesquisa para o nível de confiança de 95%

In [19]:
```
custo_95 = n_95 * custo_entrevista
print(f"Custo da pesquisa para um nível de confiança de 95%: R$ {custo_95:,.2f}")
```

```
Custo da pesquisa para um nível de confiança de 95%: R$ 98,200.00
```

Custo da pesquisa para o nível de confiança de 99%

In [20]:
```
custo_99 = n_99 * custo_entrevista
print(f"Custo da pesquisa para um nível de confiança de 99%: R$ {custo_99:,.2f}")
```

```
Custo da pesquisa para um nível de confiança de 99%: R$ 169,500.00
```

Solução do item 4

In [21]:
```
intervalo = norm.interval(alpha = 0.95, loc = media_amostra, scale = desvio_padrao_amostra / np
intervalo
```

Out[21]: (1767.820973280509, 2160.589026719491)

Solução do item 5

In [22]:
```
n_95 = recursos / custo_entrevista
n_95
```

Out[22]: 1500.0

In [23]:
```
z = norm.ppf(0.5 + (0.95 / 2))
e = z * (desvio_padrao_amostra / np.sqrt(n_95))
e
```

```
158.89721122673737
```

Out[23]:

In [24]:
```
e_percentual = e / media_amostra
print(f"A nova margem de erro é de {(e_percentual * 100):.2f}%")
```

```
A nova margem de erro é de 8.09%
```

Solução do item 6

In [25]:
```
e = media_amostra * 0.05
print(f"A margem de erro é de R$ {e:.2f} para mais ou para menos.")
```

```
A margem de erro é de R$ 98.21 para mais ou para menos.
```

In [26]:
```
z = norm.ppf(0.5 + (0.95 / 2))
z
```

Out[26]:
```
1.959963984540054
```

In [27]:
```
n_95 = (z * (desvio_padrao_amostra / e)) ** 2
n_95 = int(n_95.round())
print(f"Tamanho da amostra para um nível de confiança de 95%: {n_95}")
```

```
Tamanho da amostra para um nível de confiança de 95%: 3927
```

In [28]:
```
custo_novo = n_95 * custo_entrevista
print(f"A pesquisa com nível de confiança de 95%, com margem de erro de 5% em relação á media c
```

```
A pesquisa com nível de confiança de 95%, com margem de erro de 5% em relação á media custaria
R$ 392,700.00
```