**Estatística Com Python**

**Parte I – Frequências e Medidas:**

Todos os arquivos serão baixados, mas deixo o link do drive de qualquer forma: <https://drive.google.com/drive/folders/1MoAf3dLinpNBXVGIIyetaY5qfSwZ31Gt>.

1. **Aula 1 – Qual Seu Tipo de Dado:**
   1. Dependendo do tipo de dados análises diferentes serão feitas. Importante saber quais são os tipos.
      1. Quantitativos: Expressa quantidade de dados. Contagem, mensurações. Ex.:
         1. Idade;
         2. Renda;
         3. Altura.
      2. Qualitativos: Expressa qualidade dos dados. Ex.:
         1. UF;
         2. Sexo;
         3. Cor;
         4. Anos de Estudo.
      3. Podemos dividir os qualitativos em ordinais e nominais, e os quantitativos em discretos e contínuos.
   2. O que aprendemos:
      1. A identificar a versão de uma biblioteca do Python;
      2. A ler um dataset no formato CSV e criar um DataFrame pandas, com o conteúdo deste dataset;
      3. A identificar e classificar as variáveis de um dataset como quantitativas ou qualitativas;
      4. Que as variáveis qualitativas se dividem em ordinais e nominais;
      5. Que as variáveis quantitativas se dividem em discretas e contínuas.
2. **Aula 2 – Distribuição de Frequência:**
   1. S.value\_counts(): conta a quantidade de vezes que uma variável aparece na series. Podemos passar o parâmetro normalize = True, onde ele devolve a porcentagem de cada variável, onde ele soma elas e divide cada um pela soma:

dados['Sexo'].value\_counts(normalize=True) \* 100



* + 1. Podemos usar a porcentagem e a frequência para criar um df de distribuição e frequência qualitativas:

dist\_freq\_qualitativas = pd.DataFrame({'Frequência': frequencia, 'Porcentagem (%)': percentual})



* + 1. Df.rename\_axis(‘nome\_index’, axis=’columns’/’rows’, inplace=True): Renomeia a coluna do index:

dist\_freq\_qualitativas.rename(index= {0: 'Masculino', 1: 'Feminino'}, inplace=True)

dist\_freq\_qualitativas.rename\_axis('Sexo', axis='columns', inplace=True)

dist\_freq\_qualitativas



* 1. Pd.crosstab(variável\_linha, variável\_coluna): Cria um df onde faz o cruzamento dos dados da linha e da coluna passadas. No caso utilizamos cor e sexo, então ele irá mostrar na tabela quantas pessoas do sexo masculino tiveram de cada cor, bem como os femininos:

sexo = {0: 'Masculino',

        1: 'Feminino'}

cor = {0: 'Indígena',

        2: 'Branca',

        4: 'Preta',

        6: 'Amarela',

        8: 'Parda',

        9: 'Sem declaração'}

frequencia = pd.crosstab(dados.Sexo, dados.Cor)

frequencia.rename(index=sexo, columns=cor, inplace=True)

frequencia



* + 1. Podemos passar o mesmo parâmetro (normalize=True) que para value\_counts() e já receber o valor das porcentagens:

frequencia = pd.crosstab(dados.Sexo, dados.Cor, normalize=True) \* 100

frequencia



* + 1. Podemos ainda fazer algo bem mais legal. Podemos passar o valor de outra variável do nosso dataset para ser analisado nesses cruzamentos, ou seja, se escolhermos média da renda, por exemplo, ao invés de mostrar quantos homens e mulheres existem de cada cor, ele passa a exibir, em média, quanto os homens e mulheres ganham de renda. Para isso utilizaremos os parâmetros aggfunc=análise\_estat’ e values = dados[‘index’]:

frequencia = pd.crosstab(dados.Sexo, dados.Cor, aggfunc = 'mean', values = dados.Renda)

frequencia



* 1. Pd.cut(variável, bins=[’distribuição’], labels=labels, include\_lowest=True/False): criamos uma lista com o intervalo colocado nos bins a partir da variável escolhida fazendo relação com as labels que determinamos, como no caso da renda, onde temos a classificação de A a E da renda populacional dos nossos dados analisados. Utilizamos essa função para mostrar quais pessoas estão em qual classificação a partir da renda:

classes = [0, 1576, 3152, 7880, 15760, 200000]

labels = ['E', 'D', 'C', 'B', 'A']

pd.cut(dados.Renda, bins = classes, labels = labels, include\_lowest=True)



* + 1. Porém Podemos notar que ele classifica indivíduo por indivíduo. Para saber a frequência podemos colocar o cut dentro do pd.value\_counts(), sabendo quantas pessoas estão em cada classificação:

frequencia = pd.value\_counts(pd.cut(dados.Renda,

                                    bins = classes,

                                    labels = labels,

                                    include\_lowest=True))

frequencia



* + 1. Para saber o percentual, basta colocar o normalize=True e multiplicar por 100, como estávamos fazendo com os outros:

percentual = pd.value\_counts(pd.cut(dados.Renda,

                                    bins = classes,

                                    labels = labels,

                                    include\_lowest=True), normalize=True) \* 100

percentual



* + 1. Criamos a tabela de frequência e porcentagem com pd.DataFrame():

dist\_freq\_quantitativas\_personalizadas = pd.DataFrame({'Frequência': frequencia,

                                                       'Porcentagem (%)': percentual})

dist\_freq\_quantitativas\_personalizadas



* + 1. E alteramos a ordem do index para crescente, ou seja, de A para baixo com o sorted\_index(ascending=True):

dist\_freq\_quantitativas\_personalizadas.sort\_index(ascending=False)



* 1. Biblioteca numpy possuí diversas fórmulas matemáticas que podemos usar.
     1. Com isso podemos fazer vários cálculos como log10 dentre vários outros, permitindo com que façamos cálculos como tamanho de classes:



* + - 1. N = valor de linhas do nosso df. Basta atribuir o df.shape[0] em n.
    1. Tendo a fórmula nós podemos executar a conta:

n = dados.shape[0]

k = 1 + (10/3) \* np.log10(n)

k = int(k.round(0))

* + - 1. Arredondamos porque não tem como fazer 17.26.... classes, apenas números inteiros.
    1. Com isso podemos fazer um cut passando k como as bins para ter essa classe de amplitude fixa. De resto, a criação da frequência é a mesma:

frequencia = pd.value\_counts(pd.cut(x = dados.Renda, bins = k, include\_lowest = True), sort = False)

frequencia



* + 1. O mesmo vale para o percentual:

percentual = pd.value\_counts(pd.cut(x = dados.Renda, bins = k, include\_lowest = True), sort = False, normalize=True) \* 100

percentual



dist\_freq\_quantitativas\_amplitude\_fixa = pd.DataFrame({'Frequência': frequencia, 'Porcentagem (%)': percentual})

dist\_freq\_quantitativas\_amplitude\_fixa



* 1. Sns.distplot(dados, kde=bool): cria um histograma com a variável do df passada.
     1. O kde é a exibição de densidade do gráfico. Quando true mostra a linha traçada em cima das colunas no fundo, quando false, só as colunas.
     2. Podemos atribuir esse gráfico a uma variável (ax, por exemplo) para ficar com fácil acesso.
     3. A partir disso podemos configurar o tamanho de exibição do gráfico com ax.set\_size\_inches(x, y).
     4. O Título com ax.set\_titles(‘titulo’, fontsize=n).
     5. E as labels com ax.set\_xlabels(‘label’, fontsize=n) e ax.set\_ylabels(‘label’, fontsize=n).
     6. Exemplo:

ax = sns.displot(dados.Altura, kde=True)

ax.figure.set\_size\_inches(12,6)

ax.set\_titles('Distribuição de Frequências - Altura - KDE', fontsize = 18)

ax.set\_xlabels("Metros", fontsize = 14)

ax



Kde = false:



* + 1. Podemos obter um resultado parecido com o pandas usando a função dados.Altura.hist.(bins=n\_barras, figsize=(x,y)):

dados.Altura.hist(bins = 50, figsize=(12,6))



* + 1. Com pandas ainda podemos fazer um df[‘variavel’].plot.bar(width=n, color=’color’, alpha=n). Alpah é a transparência da cor das barras:

dist\_freq\_quantitativas\_personalizadas.Frequência.plot.bar(width = 1, color = 'blue', alpha = 0.2)



* 1. O que aprendemos:
     1. A criar distribuições de frequências (tabelas de frequências) com a função value\_counts() do pandas;
     2. A criar distribuições de frequências, com o cruzamento de duas variáveis, utilizando a função crosstab() do pandas;
     3. A criar distribuições de frequências, com classes personalizadas, utilizando as funções value\_counts() e cut() conjuntamente;
     4. A utilizar a regra de Sturges para obter um número de classes ótimo para determinado tamanho de amostra;
     5. A plotar o histograma, que é a representação gráfica de uma distribuição de frequências.

1. **Aula 3 – Medidas de Tendência Central:**
   1. A média é o ponto de equilíbrio da variável.
      1. As vezes não é o melhor dado para representar um grupo, uma vez que, no caso da renda, se tiver muita gente ganhando pouco e pouca gente ganhando muito, a média não representa essa variável, pois apesar de ser uma média grande, o desvio padrão e a distribuição dos dados são altos, tendo uma baixa representatividade. Cálculo da média:



* + 1. A média nada mais é do que a soma dos valores dividido pelo seu n total, ou seja:

(8 + 10 + 4 + 8 + 6 + 10 + 8) / 7



* + 1. Felizmente não precisamos fazer isso na mão desse jeito sempre, uma vez que seria impossível fazer isso com 70k de dados, por exemplo. Principalmente se forem floats.
    2. Nesse caso o pandas nos ajuda e muito com suas função sum() e seu shape. Com eles podemos pegar apenas uma coluna de informação e passar para o sum que irá somar todos os valores, em seguida podemos fazer a divisão dessa soma pela quantidade de observações do nosso dado, ou seja, quantidade de linhas dele. Para isso usamos o atributo shape[0], que nos retorna exatamente essa informação. No final obtemos o mesmo resultado que feito na mão, assim:

df.Fulano.sum() / df.shape[0]



* + 1. Ou então podemos fazer de uma forma ainda MUITO mais fácil: simplesmente pedir a média dessa coluna para o pandas com o mean():

df.Fulano.mean()



* + 1. Utilizando esse conhecimento podemos utilizar nossos dados e pegar a média de todas as rendas, ou então usar a função de agrupamento do pandas e fazer a média diretamente a partir de uma variável para esse grupo df.groupby(‘variável\_para\_agrupar\_por’)[‘variável\_analisar’].mean():

dados.groupby(['Sexo'])['Renda'].mean()



* 1. Mediana é o valor que divide a series exatamente no meio, por exemplo, a idade de 10 pessoas, organizamos em ordem crescente e vemos qual é a idade que divide exatamente no meio, com os mais novos pra um lado e os mais velhos para o outro.



* + 1. Cálculos – Identificando Elemento:
       1. Quando n for ímpar:



* + - 1. Quando n for par:



* + 1. Cálculos – Obtendo a Mediana:
       1. Quando n for ímpar:



* + - 1. Quando n for par:



* + 1. Exemplo no python:

df = pd.DataFrame(data = {'Fulano': [8, 10, 4, 8, 6, 10, 8],

                          'Beltrano': [10, 2, 0.5, 1, 3, 9.5, 10],

                          'Sicrano': [7.5, 8, 7, 8, 8, 8.5, 7]},

                  index = ['Matemática', 'Português', 'Inglês',

                           'Geografia', 'História', 'Física', 'Química'])

df.rename\_axis('Matérias', axis = 'columns', inplace = True)

df



notas\_fulano = df.Fulano

notas\_fulano = notas\_fulano.sort\_values()

notas\_fulano = notas\_fulano.reset\_index()

n = notas\_fulano.shape[0]

elemento\_md = int((n + 1) / 2)

notas\_fulano.loc[elemento\_md - 1]



Ou então:

notas\_fulano.median()



* + 1. Com n par:

notas\_beltrano = df.Beltrano.sample(6, random\_state=101)

notas\_beltrano



notas\_beltrano.median()



* 1. A moda é sempre o que mais aparece, ou seja, no caso das notas do df acima, temos que a nota modal de fulano é 8, uma vez que ela aparece 3 vezes.
     1. Se tivermos 2 modas chamamos de bimodal.
     2. Mais de 3, multimodal.
     3. Para pegar as modas use df.mode(). Será retornado um df com as modas:

df.mode()



* + 1. Multimodal:

exemplo = pd.Series([1, 2, 2, 3, 4, 4, 5, 6, 6])

exemplo.mode()



* 1. Relação entre média, mediana e moda:
     1. Moda maior que mediana maior que média: Assimetria à direita:

ax = sns.distplot(dados.query('Renda < 20000').Renda)

ax.figure.set\_size\_inches(12,6)

ax



* + - 1. Tirando a prova nos dados:

moda = dados.Renda.mode()[0]

moda



mediana = dados.Renda.median()

mediana



media = dados.Renda.mean()

media



moda < mediana < media



* + 1. Todas iguais (com uma margem de erro): Simétrica:

ax = sns.distplot(dados.Altura)

ax.figure.set\_size\_inches(12,6)

ax



* + - 1. Prova nos dados:

moda = dados.Altura.mode()

moda



mediana = dados.Altura.median()

mediana



media = dados.Altura.mean()

media



* + 1. Média menor que mediana menor que moda: Assimetria à esquerda:

ax = sns.distplot(dados['Anos de Estudo'], bins = 17)

ax.figure.set\_size\_inches(12,6)

ax



moda = dados['Anos de Estudo'].mode()[0]

moda



mediana = dados['Anos de Estudo'].median()

mediana



media = dados['Anos de Estudo'].mean()

media



moda > mediana > media



* 1. O que aprendemos:
     1. A calcular as principais medidas de tendência central: média aritmética, mediana e moda;
     2. A identificar características importantes de uma distribuição, como a presença de assimetria e sua direção a partir da relação entre as medidas de tendência central.

1. **Aula 4 – Medidas Separatrizes:**
   1. Há uma série de medidas de posição semelhantes na sua concepção à mediana, embora não sejam medidas de tendência central. Como se sabe, a mediana divide a distribuição em duas partes iguais quanto ao número de elementos de cada parte. Já os quartis permitem dividir a distribuição em quatro partes iguais quanto ao número de elementos de cada uma; os decis em dez partes e os centis em cem partes iguais.
      1. S.quantile(.25, .75): Devolve os quartis todos, dependendo do valor passado no parâmetro. Se não passar nada, ele devolve o 0.5, ou seja, a mediana.
      2. Podemos ainda passar uma lista para receber todos os quartis de uma vez:

dados.Renda.quantile([0.25, 0.5, 0.75])



* + 1. Utilizamos a mesma função para fazer os decis, mas ao invés dos parâmetros que passamos antes, fazemos uma list comprehensions com um for para repetir esse processo 9x:

dados.Renda.quantile([i / 10 for i in range(1, 10)])



* + - 1. Desse modo temos que 10% das pessoas estão abaixo de 350 enquanto 90 estão acima. O mesmo raciocínio serve para o resto dos decis.
    1. A mesma lógica se aplica aos percentis, mas ao invés de 10 colocamos 100 em tudo:

dados.Renda.quantile([i / 100 for i in range(1, 100)])



* + - 1. A principal diferença é na quantidade de informações que temos, já que ele varia de 1% em 1%. Podemos inclusive notar que 4% das pessoas estão abaixo de 50, mas ao 5% estão abaixo de 100 e assim por diante.
    1. Podemos criar um histograma e visualizar em gráfico o que vimos em números, passando para os parâmetros sns.distplot(dados, hist.\_kws= {‘cumulative’:True}, kde\_kws = {‘cumulative’:True}, bins=10). Ao passar o bins ele vai mostrar 10 barras representando os decis:

ax = sns.distplot(dados.Idade, hist\_kws= {'cumulative':True},

                  kde\_kws = {'cumulative':True}, bins = 10)

ax.figure.set\_size\_inches(12,6)

ax.set\_title('Distribuição de Frequências Acumuladas', fontsize=18)

ax.set\_ylabel('Acumulado', fontsize=14)

ax.set\_xlabel('Anos', fontsize=14)

ax



* 1. Box-plot são representações gráficas bem importantes pois mostra muito da distribuição que estamos estudando, como informações de posição, dispersão, simetria, dados discrepantes, dentre outros, e são construídos a partir de medidas separatrizes.



* + 1. Podemos criar sns.boxplot(x, y~~opt~~, data=data, orient=’h’) passando mais de 1 parâmetro, como x e y, por exemplo, sendo 2 dados diferentes. Parâmetro orient=’h’ para deixar os gráficos horizontais:

ax = sns.boxplot( x = 'Altura', y = 'Sexo', data = dados, orient = 'h')

ax.figure.set\_size\_inches(12, 4)

ax.set\_title('Altura', fontsize=18)

ax.set\_xlabel('Metros', fontsize=14)

ax



* + 1. Conseguimos passar uma query para fazer uma seleção de dados, como no caso da renda onde não conseguimos ver nada por causa do outlier:

ax = sns.boxplot( x = 'Renda', data = dados, orient = 'h')

ax.figure.set\_size\_inches(12, 4)

ax.set\_title('Altura', fontsize=18)

ax.set\_xlabel('Metros', fontsize=14)

ax



ax = sns.boxplot( x = 'Renda', data = dados.query('Renda < 10000'), orient = 'h')

ax.figure.set\_size\_inches(12, 4)

ax.set\_title('Altura', fontsize=18)

ax.set\_xlabel('Metros', fontsize=14)

ax



* + 1. Podemos ainda colocar 2 dados como acima:

ax = sns.boxplot( x = 'Renda', y = 'Sexo', data = dados.query('Renda < 10000'), orient = 'h')

ax.figure.set\_size\_inches(12, 4)

ax.set\_title('Altura', fontsize=18)

ax.set\_xlabel('Metros', fontsize=14)

ax



* + 1. Comparação de simetria:



* 1. O que aprendemos:
     1. A obter os quartis, decis e percentis de uma distribuição;
     2. Também conhecidas como medidas separatrizes, pois dividem uma distribuição em partes iguais;
     3. A construir e interpretar um boxplot, com a utilização dos quartis.

1. **Aula 5 – Medidas de Dispersão:**
   1. Embora as medidas de posição forneçam uma sumarização bastante importante dos dados, elas podem não ser suficientes para caracterizar conjuntos distintos, especialmente quando as observações de determinada distribuição apresentarem dados muito dispersos.
      1. Desvio Médio Absoluto: Para fazer esse cálculo podemos fazer na mão como demonstrado abaixo, ou utilizar o s.mad(), o retorno será o mesmo. Cálculo:



* + 1. Como podemos ver pela fórmula, dm é igual a média do valor absoluto, ou seja, positivo, da somatória dos desvios, sendo este o xi (x índice) menos a média geral:

notas\_fulano = df[['Fulano']]

nota\_media\_fulano = notas\_fulano.mean()[0]

notas\_fulano['Desvio'] = notas\_fulano['Fulano'] - nota\_media\_fulano

notas\_fulano['|Desvio|'] = notas\_fulano.Desvio.abs()

notas\_fulano['|Desvio|'].mean()



* + 1. Usando o s.mad():

desvio\_medio\_absoluto = notas\_fulano.Fulano.mad()

desvio\_medio\_absoluto



* + 1. A soma dos desvios em relação a média é 0.
    2. Quando vamos fazer esse cálculo pode ser que tenham números negativos, nesse caso podemos usar o abs() para ignorar o sinal negativo tornando todos positivos:

notas\_fulano['|Desvio|'] = notas\_fulano.Desvio.abs()

notas\_fulano



* 1. Variância: A variância é construída a partir das diferenças entre cada observação e a média dos dados, ou seja, o desvio em torno da média. No cálculo da variância, os desvios em torno da média são elevados ao quadrado.
     1. Variância Populacional: Calcula a variância, ou seja, o desvio em torno da média de uma população toda. Cálculo:



* + 1. Variância Amostral: Utiliza-se de um cálculo para ver a variância de uma amostra que representa a população. Cálculo:



* + 1. Podemos usar o s.pow(n) do pandas para fazer cálculos de elevados. Substituir n pelo valor a elevar:

notas\_fulano['(Desvio)^2'] = notas\_fulano.Desvio.pow(2)

notas\_fulano



* + 1. Utilizamos isso para fazer a variância, agora basta fazer a somatória geral do desvio2 e dividir por n – 1:

notas\_fulano['(Desvio)^2'].sum() / (len(notas\_fulano) - 1)



* + 1. Ou então, simplificando e MUITO com o pandas, usamos o s.var():

variancia = notas\_fulano.Fulano.var()

variancia



* 1. Desvio Padrão: Uma das restrições da variância é o fato de fornecer medidas em quadrados das unidades originais - a variância de medidas de comprimento, por exemplo, é em unidades de área. Logo, o fato de as unidades serem diferentes dificulta a comparação da dispersão com as variáveis que a definem. Um modo de eliminar essa dificuldade é considerar sua raiz quadrada.
     1. É a raiz quadrada da variância, ou seja, retira o 2 da nossa variância que tem isso como unidade de medida, obtendo uma medida de dispersão média dos dados na mesma unidade de medida do dado que o gerou.
     2. Desvio Padrão Populacional:



* + 1. Desvio Padrão Amostral:



* + 1. Podemos extrair a raiz quadrada de 2 formas: utilizamos a np.sqrt(variância), ou então a s.std():

np.sqrt(variancia)



desv\_pad = notas\_fulano.Fulano.std()

desv\_pad



* 1. O que aprendemos:
     1. A obter três medidas de dispersão importantes. O desvio médio absoluto, a variância e o desvio padrão;
     2. Que, embora as medidas de posição forneçam uma sumarização bastante importante dos dados, elas podem não ser suficientes para caracterizar conjuntos distintos, especialmente quando as observações de determinada distribuição apresentarem dados muito dispersos:
        1. Por isso, para complementar nossas análises e poder caracterizar melhor o conjunto de dados, utilizamos as medidas de dispersão.

1. **Aula 6 – Projeto Final:**
   1. O que aprendemos:
      1. A realizar uma análise descritiva em um conjunto de dados;
      2. Funcionalidades extras da função crosstab() com o parâmetro aggfunc;
      3. A construir boxplots com o cruzamento de três informações distintas.

**Parte II – Probabilidade e Amostragem:**

1. **Aula 1 – Distribuição Binominal:**
   1. Distribuição binomial: Um evento binomial é caracterizado pela possibilidade de ocorrência de apenas duas categorias. Estas categorias somadas representam todo o espaço amostral, sendo também mutuamente excludentes, ou seja, a ocorrência de uma implica na não ocorrência da outra. Em análises estatísticas o uso mais comum da distribuição binomial é na solução de problemas que envolvem situações de sucesso e fracasso. Cálculo:





* + 1. Espaço amostral é a possibilidade de ocorrência de uma probabilidade, como no caso de jogar um dado de 6 lados, a probabilidade de cair 1 deles é 1 em 6.
    2. Requisitos para podermos usar o experimento binomial:
       1. Realização de n ensaios idênticos;
       2. Os ensaios são independentes;
       3. Somente dois resultados são possíveis, exemplo: Verdadeiro ou falso; Cara ou coroa; Sucesso ou fracasso;
       4. A probabilidade de sucesso é representada por p e a de fracasso por 1−p=q. Estas probabilidades não se modificam de ensaio para ensaio.
    3. Média da distribuição binomial: O valor esperado ou a média da distribuição binomial é igual ao número de experimentos realizados multiplicado pela chance de ocorrência do evento:



* + 1. Desvio padrão da distribuição binomial: O desvio padrão é o produto entre o número de experimentos, a probabilidade de sucesso e a probabilidade de fracasso:



* + 1. Combinações: Número de combinações de $n$ objetos, tomados $k$ a cada vez, é:





Por definição:



* + - 1. Para calcular as combinações, precisamos importar o comb da biblioteca scipy.special:

from scipy.special import comb

* + 1. Exemplo de combinações: Em um volante de loteria da Mega Sena temos um total de 60 números para escolher onde a aposta mínima é de seis números. Você que é curiosa(o) resolve calcular a probabilidade de se acertar na Mega Sena com apenas um jogo. Para isso precisamos saber quantas combinações de seis números podem ser formadas com os 60 números disponíveis:



* + 1. Resolvendo com o comb do scipy:

combinacoes = comb(60, 6)

combinacoes



* + - 1. Para calcular a probabilidade de acerto basta pegar o n de números de apostas, por exemplo 1 bilhete de aposta, e dividir pelo resuldado de c606, que é 50063860 nesse caso:

probabilidade = 1 / combinacoes

probabilidade

 = 0,0000000199744...

* + - 1. Podemos fazer uma formatação com o print:

print(f'{probabilidade:.15f}')



* 1. Quando temos um problema precisamos coletar os itens disponibilizados e ver em que distribuição de probabilidade ele se adequa.
     1. Exemplo: Em um concurso para preencher uma vaga de cientista de dados temos um total de 10 questões de múltipla escolha com 3 alternativas possíveis em cada questão. Cada questão tem o mesmo valor. Suponha que um candidato resolva se aventurar sem ter estudado absolutamente nada. Ele resolve fazer a prova de olhos vendados e chutar todas as respostas. Assumindo que a prova vale 10 pontos e a nota de corte seja 5, obtenha a probabilidade deste candidato acertar 5 questões e também a probabilidade deste candidato passar para a próxima etapa do processo seletivo:
        1. Distribuição do tipo binominal: Ao analisar as características colocadas acima para caracterizar essa distribuição, vemos que nosso exemplo se enquadra por ter 10 questões com 3 alternativas e 1 concurso. Cada questão é um experimento diferente com 3 alternativas cada, ou seja, temos n=10. Só existem 2 possibilidades: apesar de 3 alternativas, temos apenas a possibilidade de acertar ou errar.
        2. Probabilidade de acertar: p = 1 em 3 ou seja:

numero\_de\_alternativas\_por\_questao = 3

p = 1 / numero\_de\_alternativas\_por\_questao

p

 \* 100 = 33.333...%

* + - 1. Probabilidade de errar: q = 1 - p

q = 1 - p

q

 \* 100 = 66.666...%

* + - 1. Total de eventos que se deseja sucesso: k = 5. 5 pois queremos que ele acerte 5 das 10 questões para passar.
      2. Solução: Aplicamos a fórmula do começo – a primeira.:

probabilidade = (comb(n, k) \* (p\*\*k) \* (q\*\*(n-k)))

print(f'{probabilidade:.8f}')

 \* 100 = 13.65%

* + - 1. Resolvendo com o python: Usamos a biblioteca binom do scipy.stats, passando binom.pmf(k, n, p) :

probabilidade = binom.pmf(k, n, p)

print(f'{probabilidade:.8f}')



* + 1. Entretanto, apesar de termos feito tudo isso essa foi somente a probabilidade de ele acertar 5, não a dele passar, uma vez que 5 acertos é o mínimo necessário para que ele passe, ele ainda pode acertar 6, 7, 8, 9 e 10 questões:



* + - 1. Podemos fazer esse cálculo de várias formas, sendo a primeira delas exatamente igual a figura acima:

binom.pmf(5, n, p) + binom.pmf(6, n, p) + binom.pmf(7, n, p) + binom.pmf(8, n, p) + binom.pmf(9, n, p) + binom.pmf(10, n, p)

 = 21.31%

* + - 1. O segundo seria passando ao invés de k, uma lista com todas as outras chances e somando tudo no final:

binom.pmf([5, 6, 7, 8, 9, 10], n, p).sum()



* + - 1. A terceira é utilizar o binom.cdf(k, n, p), sendo essa função acumulativa, ou seja, ele vai fazer o cálculo e a soma de todas as probabilidades desde 0 até o valor passado em k:

binom.cdf(4, n, p)



* + - 1. Note que não é o valor que queríamos, mas ao colocar o 1 – binom ele nos retorna exatamente o valor de antes. Isso ocorre por ele ter o caráter acumulativo, então precisamos subtrair 100%, ou seja 1, do valor que ele nos retorna:

1 - binom.cdf(4, n, p)



* + - 1. Temos ainda o binom.sf(k, n, p) como quarta opção, fazendo exatamente a conta acima: 1 – binom.cdf(k, n, p):

binom.sf(4, n, p)



* + 1. **1-cdf e sf são as probabilidades de acertar 5 questões ou mais. Pmf é a probabilidade de acertar até 5 questões.**
  1. Aplicação da média:
     1. Exemplo: Uma cidade do interior realiza todos os anos uma gincana para arrecadar fundos para o hospital da cidade. Na última gincana se sabe que a proporção de participantes do sexo feminino foi de 60%. O total de equipes, com 12 integrantes, inscritas na gincana deste ano é de 30. Com as informações acima responda: Quantas equipes deverão ser formadas por 8 mulheres?
     2. Nosso p = 0.6, n = 12 e k = 8:

probabilidade = binom.pmf(k, n, p)

print(f'{probabilidade:.8f}')



* + 1. Nossa pergunta foi quantas equipes poderiam ser formadas por 8 mulheres, para isso fazemos n \* p como vimos na fórmula no começo:

equipes = 30 \* probabilidade

equipes



* + - 1. Não existe número de equipes quebrado, portanto podemos arredondar, mas se quiser a resposta exata é a acima.
  1. O que aprendemos:
     1. Os conceitos básicos da distribuição de probabilidades binomial;
     2. A identificar e solucionar problemas que envolvam um experimento binomial;
     3. A obter a probabilidade de eventos que envolvam apenas duas possibilidades de resultado.

1. **Aula 2 – Distribuição de Poisson:**
   1. É empregada para descrever o número de ocorrências em um intervalo de tempo ou espaço específico. Os eventos são caracterizados pela possibilidade de contagem dos sucessos, mas a não possibilidade de contagem dos fracassos. Como exemplos de processos onde podemos aplicar a distribuição de Poisson temos a determinação do número de clientes que entram em uma loja em determinada hora, o número de carros que chegam em um drive-thru de uma lanchonete na hora do almoço, a determinação do número de acidentes registrados em um trecho de estrada etc. Cálculo:



* + 1. O que é cada elemento:



* + - 1. O numero de euri, ou seja, o *e* conseguimos obter a partir da numpy.e:

import numpy as np

np.e



* + 1. O experimento:



* + 1. A média da distribuição é justamente o u da formula acima:



* + 1. Desvio padrão de Poisson:



* 1. Vamos resolver o seguinte exemplo com utilizando essa distribuição:
     1. Um restaurante recebe em média 20 pedidos por hora. Qual a chance de que, em determinada hora escolhida ao acaso, o restaurante receba 15 pedidos?
     2. O próprio exemplo já nos dá no enunciado os elementos da fórmula que precisaremos utilizar:
        1. Mi/u/média = 20;
        2. K = 15;
     3. Depois disso é só fazer a fórmula:

probabilidade = ((np.e \*\* (-media)) \* (media \*\* k)) / (np.math.factorial(k))

probabilidade



* + - 1. Podemos fazer o fatorial de forma muito simples utilizando a np.math.factoria(k).
    1. Para resolver o mesmo problema de modo muito mais simples, podemos utilizar o poisson.pmf(k, mi/média) da biblioteca scipy.stats:

from scipy.stats import poisson

probabilidade = poisson.pmf(k, media)

probabilidade



* + 1. Ou seja, a probabilidade de um restaurante receber 15 pedidos em 60 minutos, onde já recebe 20 em média, é de 5,16%.
  1. O que aprendemos:
     1. Os conceitos básicos da distribuição de probabilidades Poisson;
     2. A obter as probabilidades em problemas, como os seguintes:
        1. Determinação do número de clientes que entram em uma loja em determinada hora;
        2. Número de carros que chegam em um drive-thru de uma lanchonete na hora do almoço;
        3. Determinação do número de acidentes registrados em um trecho de estrada.

1. **Aula 3 – Distribuição Normal:**
   1. A distribuição normal é uma das mais utilizadas em estatística. É uma distribuição contínua, onde a distribuição de frequências de uma variável quantitativa apresenta a forma de sino e é simétrica em relação a sua média.



* 1. Características importantes:
     1. É simétrica em torno da média;
     2. A área sob a curva corresponde à proporção 1 ou 100%;
     3. As medidas de tendência central (média, mediana e moda) apresentam o mesmo valor;
     4. Os extremos da curva tendem ao infinito em ambas as direções e, teoricamente, jamais tocam o eixo x;
     5. O desvio padrão define o achatamento e largura da distribuição. Curvas mais largas e mais achatadas apresentam valores maiores de desvio padrão;
     6. A distribuição é definida por sua média e desvio padrão;
     7. A probabilidade sempre será igual à área sob a curva, delimitada pelos limites inferior e superior.
  2. Cálculo:



* + 1. Cada elemento:



* + 1. A probabilidade é obtida a partir da área sob a curva, delimitada pelos limites inferior e superior especificados. Um exemplo pode ser visto na figura abaixo.



* + 1. Para obter a área acima basta calcular a integral da função para os intervalos determinados. Conforme equação abaixo:



* + 1. Elementos:



* 1. Tabelas padronizadas:
     1. As tabelas padronizadas foram criadas para facilitar a obtenção dos valores das áreas sob a curva normal e eliminar a necessidade de solucionar integrais definidas.
     2. Para consultarmos os valores em uma tabela padronizada basta transformarmos nossa variável em uma variável padronizada Z.
     3. Esta variável Z representa o afastamento em desvios padrões de um valor da variável original em relação à média. Cálculo:



* + 1. Elementos:



* + 1. Podemos criar a tabela com o seguinte código, mas é mais fácil consultar o livro:

import pandas as pd

import numpy as np

from scipy.stats import norm

tabela\_normal\_padronizada = pd.DataFrame(

    [],

    index=["{0:0.2f}".format(i / 100) for i in range(0, 400, 10)],

    columns = ["{0:0.2f}".format(i / 100) for i in range(0, 10)])

for index in tabela\_normal\_padronizada.index:

    for column in tabela\_normal\_padronizada.columns:

        Z = np.round(float(index) + float(column), 2)

        tabela\_normal\_padronizada.loc[index, column] = "{0:0.4f}".format(norm.cdf(Z))

tabela\_normal\_padronizada.rename\_axis('Z', axis = 'columns', inplace = True)

tabela\_normal\_padronizada



* + 1. A tabela acima fornece a área sob a curva entre −∞ e Z desvios padrão acima da média. Lembrando que por se tratar de valores padronizados temos μ=0.



* 1. Exemplo para resolver: Em um estudo sobre as alturas dos moradores de uma cidade verificou-se que o conjunto de dados segue uma distribuição aproximadamente normal, com média 1,70 e desvio padrão de 0,1. Com estas informações obtenha o seguinte conjunto de probabilidades:
     1. Probabilidade de uma pessoa, selecionada ao acaso, ter menos de 1,80 metros.
     2. Probabilidade de uma pessoa, selecionada ao acaso, ter entre 1,60 metros e 1,80 metros.
     3. Probabilidade de uma pessoa, selecionada ao acaso, ter mais de 1,90 metros.
  2. Primeiro caso:



* + 1. Calculando Z:

media = 1.7

desvio\_padrao = 0.1

Z = (1.8 - media) / desvio\_padrao

Z



* + 1. Solução 1: Usando a tabela:
       1. Pegamos o valor de Z e consultamos na tabela procurando esse valor, ou seja, se foi 1.00, olhamos na linha 1.00 e na coluna 0.00. Mas se nosso Z tivesse dado 0.15, por exemplo, precisaríamos olhar na linha 0.10 e na coluna 0.05, a intercessão da linha com a coluna seria nossa probabilidade:



* + - 1. No nosso caso, com o resultado 1.00 temos que a probabilidade é igual a 0.8413, ou seja, a probabilidade de uma pessoa selecionada ao acaso ter menos de 1.8m é de 84.13%.
    1. Solução 2: Usando scipy:

from scipy.stats import norm

norm.cdf(Z)



* + - 1. Utilizando o scipy basicamente nos poupa tempo de procurar o resultado na tabela que é deveras bem grande, mas ainda assim temos que calcular Z.
  1. Segundo caso:



* + 1. Solução 1: Tabela
       1. Calculamos Z para 1.8 como no caso 1 e o valor da tabela que recebemos pegamos o 0.5, ou seja, 50%, e subtraímos do valor da tabela encontrada pelo Z. Fizemos isso pois queremos saber qual é a área da média, sendo ela 1.7, até 1.8, o resto da curva à direita, que corresponde 50%, ou seja, 0.5, não nos interessa, portanto subtraímos um do outro:

Z = (1.8 - media) / desvio\_padrao

Z



probabilidade = 0.8413 - 0.5

probabilidade



* + - * 1. Com isso temos que nossa área de 1.7 à 1.8 é 0.3413.
      1. Como nesse caso também queremos a área correspondente de 1.7 à 1.60 também, e sabemos que equivale ao mesmo valor que 1.7 à 1.8 por ambos variarem 10 da média, podemos só multiplicar a área que descobrimos acima por 2:

probabilidade = (0.8413 - 0.5) \* 2

probabilidade



* + 1. Solução 2: Scipy
       1. Precisamos calcular o Z inferior e superior para esse cálculo:

Z\_inferior = (1.6 - media) / desvio\_padrao

Z\_inferior



Z\_superior = (1.8 - media) / desvio\_padrao

Z\_superior



* + - 1. Para calcular a área precisamos subtrair o valor da tabela do limite superior de 1 menos o limite superior novamente:

probabilidade = norm.cdf(Z\_superior) - (1 - norm.cdf(Z\_superior))

probabilidade



* + - 1. Para deixar ainda mais fácil, podemos simplesmente passar o norm.cdf() do z superior menos o do inferior:

probabilidade = norm.cdf(Z\_superior) - norm.cdf(Z\_inferior)

probabilidade



* + 1. Solução 3:



* + - 1. Quando obtemos o Z e vemos o valor na tabela temos exatamente o valor da área da esquerda para a direita até o ponto que nos é pedido, ou seja, se obtermos o Z de 1.9 e olharmos na tabela, o valor que vamos receber vai ser o total da esquerda até o tracejado do 1.9, e não a área pintada que é o que queremos. Sabendo disso, e que a área total abaixo da curva vale 1, para obter a área pintada basta subtrair 1 do valor de Z na tabela:

probabilidade = 0.9767 # Valor da área branca

probabilidade = 1 - 0.9767

probabilidade



* + - * 1. Ou seja, temos 2,32% de chance de ter alguém com mais de 1.9 de altura.
    1. Solução 2: Scipy
       1. Segue o mesmo raciocínio, 1 – norm.cdf(Z):

probabilidade = 1 - norm.cdf(Z)

probabilidade



* + - 1. O scipy é inteligente e sabe que se a gente passar um número negativo pra ele estamos querendo calcular a área igual a desse caso:

probabilidade = norm.cdf(-Z)

probabilidade



* 1. O que aprendemos:
     1. Os conceitos básicos da distribuição normal de probabilidades;
     2. A trabalhar com a tabela padronizada Z;
     3. A obter as probabilidades utilizando a distribuição normal em um conjunto de situações.

1. **Aula 4 – Técnicas de Amostragem:**
   1. **População**: Conjunto de todos os elementos de interesse em um estudo. Diversos elementos podem compor uma população, por exemplo: pessoas, idades, alturas, carros etc. Com relação ao tamanho, as populações podem ser limitadas (populações finitas) ou ilimitadas (populações infinitas).
      1. Populações Finitas: Permitem a contagem de seus elementos. Como exemplos temos o número de funcionário de uma empresa, a quantidade de alunos em uma escola etc.
      2. Populações Infinitas: Não é possível contar seus elementos. Como exemplos temos a quantidade de porções que se pode extrair da água do mar para uma análise, temperatura medida em cada ponto de um território etc.
         1. **Quando os elementos de uma população puderem ser contados, porém apresentando uma quantidade muito grande, assume-se a população como infinita.**
   2. **Amostra**: Subconjunto representativo da população.
      1. Os atributos numéricos de uma população como sua média, variância e desvio padrão, são conhecidos como parâmetros. O principal foco da inferência estatística é justamente gerar estimativas e testar hipóteses sobre os parâmetros populacionais utilizando as informações de amostras.
   3. **Quando utilizar uma amostra**:
      1. Populações Infinitas: O estudo não chegaria nunca ao fim. Não é possível investigar todos os elementos da população.
      2. Testes Destrutivos: Estudos onde os elementos avaliados são totalmente consumidos ou destruídos. Exemplo: testes de vida útil, testes de segurança contra colisões em automóveis.
      3. Resultados Rápidos: Pesquisas que precisam de mais agilidade na divulgação. Exemplo: pesquisas de opinião, pesquisas que envolvam problemas de saúde pública.
      4. Custos Elevados: Quando a população é finita, mas muito numerosa, o custo de um censo pode tornar o processo inviável.
   4. **Amostragem Aleatória Simples:** É uma das principais maneiras de se extrair uma amostra de uma população. A exigência fundamental deste tipo de abordagem é que cada elemento da população tenha as mesmas chances de ser selecionado para fazer parte da amostra.
      1. Para pegar uma amostra utilizamos o df.sample(n = n\_amostra, Random\_state = n) do pandas. O segundo parâmetro é uma seed para sempre obtermos a mesma aleatorização, ou seja, sempre com as mesmas amostras:

amostra = dados.sample(n = 100, random\_state = 101)

amostra



* + 1. Ao fazermos a comparação da frequência do sexo dos indivíduos da nossa população e da nossa amostra, podemos observar que é praticamente a mesma:

dados.Sexo.value\_counts(normalize = True)



amostra.Sexo.value\_counts(normalize = True)



* + - 1. Podemos ainda alterar o tamanho da nossa amostra e, apesar de variar um pouco os resultados, continuam estando ao redor da população:

amostra = dados.sample(n = 1000, random\_state = 101)

dados.Sexo.value\_counts(normalize = True)

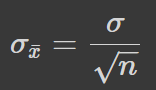


amostra.Sexo.value\_counts(normalize = True)



* 1. **Amostragem Estratificada:** É uma melhoria do processo de amostragem aleatória simples. Neste método é proposta a divisão da população em subgrupos de elementos com características similares, ou seja, grupos mais homogêneos. Com estes subgrupos separados, aplica-se a técnica de amostragem aleatória simples dentro de cada subgrupo individualmente.
  2. **Amostragem por Conglomerados:** Também visa melhorar o critério de amostragem aleatória simples. Na amostragem por conglomerados são também criados subgrupos, porém não serão homogêneas como na amostragem estratificada. Na amostragem por conglomerados os subgrupos serão heterogêneos, onde, em seguida, serão aplicadas a amostragem aleatória simples ou estratificada. Um exemplo bastante comum de aplicação deste tipo de técnica é na divisão da população em grupos territoriais, onde os elementos investigados terão características bastante variadas.
  3. O que aprendemos:
     1. Os conceitos de população e amostra;
     2. A identificação de populações finita e infinita;
     3. Quando utilizar a técnica de amostragem em um estudo;
     4. Técnicas de seleção de amostra, como:
        1. Amostragem aleatória simples;
        2. Amostragem estratificada;
        3. Amostragem por conglomerados.

1. **Aula 5 – Nível e Intervalo de Confiança:**
   1. Estimação: Estimativa dos parâmetros de uma população a partir dos dados de uma amostra.
   2. Problema para resolver: Suponha que os pesos dos sacos de arroz de uma indústria alimentícia se distribuem aproximadamente como uma normal de desvio padrão populacional igual a 150 g. Selecionada uma amostra aleatória de 20 sacos de um lote específico, obteve-se um peso médio de 5.050 g. Construa um intervalo de confiança para a média populacional assumindo um nível de significância de 5%.
   3. É a forma de se fazer suposições generalizadas sobre os parâmetros de uma população tendo como base as informações de uma amostra.
      1. Parâmetros são os atributos numéricos de uma população, tal como a média, desvio padrão etc.
      2. Estimativa é o valor obtido para determinado parâmetro a partir dos dados de uma amostra da população.
   4. **Teorema do Limite Central**: O Teorema do Limite Central afirma que, com o aumento do tamanho da amostra, a distribuição das médias amostrais se aproxima de uma distribuição normal com média igual à média da população e desvio padrão igual ao desvio padrão da variável original dividido pela raiz quadrada do tamanho da amostra. Este fato é assegurado para n maior ou igual a 30. Cálculo:



* + 1. O desvio padrão das médias amostrais é conhecido como erro padrão da média.
    2. O Teorema do Limite Central afirma que, com o aumento do tamanho da amostra, a distribuição das médias amostrais se aproxima de uma distribuição normal com média igual à média da população e desvio padrão igual ao desvio padrão da variável original dividido pela raiz quadrada do tamanho da amostra. Este fato é assegurado para n maior ou igual a 30.
    3. Para comprovar fizemos um df com 2k observações e 1.5k amostras da idade dos nossos dados e chamamos isso de amostras:

n = 2000

total\_de\_amostras = 1500

amostras = pd.DataFrame()

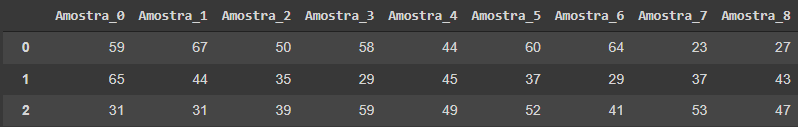
for i in range(total\_de\_amostras):

  \_ = dados.Idade.sample(n)

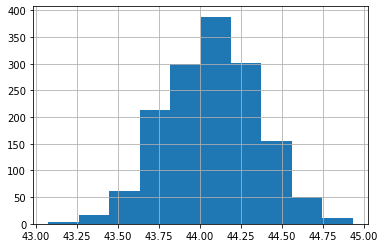
  \_.index = range(0, len(\_))

  amostras['Amostra\_' + str(i)] = \_

amostras

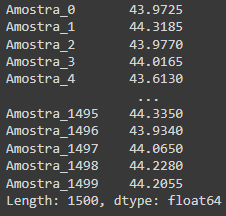


amostras.mean().hist()



* + 1. Se assemelhando bastante a uma distribuição normal.
    2. O Teorema do Limite Central afirma que, com o aumento do tamanho da amostra, a distribuição das médias amostrais se aproxima de uma distribuição normal com média igual à média da população e desvio padrão igual ao desvio padrão da variável original dividido pela raiz quadrada do tamanho da amostra. Este fato é assegurado para n maior ou igual a 30.
    3. Para comprovar tiramos a média das médias da nossa amostra e a média da nossa população:

amostras.mean()



amostras.mean().mean()



dados.Idade.mean()



* + 1. O Teorema do Limite Central afirma que, com o aumento do tamanho da amostra, a distribuição das médias amostrais se aproxima de uma distribuição normal com média igual à média da população e desvio padrão igual ao desvio padrão da variável original dividido pela raiz quadrada do tamanho da amostra. Este fato é assegurado para n maior ou igual a 30.

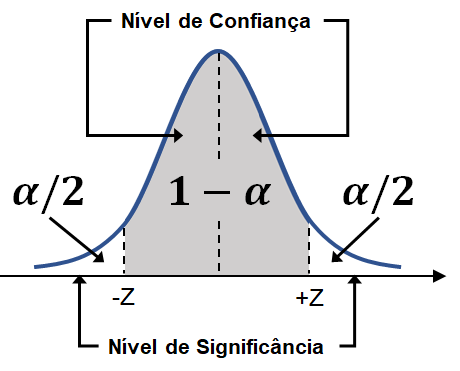
amostras.mean().std()



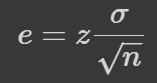
dados.Idade.std() / np.sqrt(n)



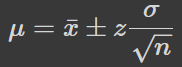
* 1. Níveis de Confiança e significância:
     1. O nível de confiança (1−α) representa a probabilidade de acerto da estimativa. De forma complementar o nível de significância (α) expressa a probabilidade de erro da estimativa.
     2. O nível de confiança representa o grau de confiabilidade do resultado da estimativa estar dentro de determinado intervalo. Quando fixamos em uma pesquisa um nível de confiança de 95%, por exemplo, estamos assumindo que existe uma probabilidade de 95% dos resultados da pesquisa representarem bem a realidade, ou seja, estarem corretos.
     3. O nível de confiança de uma estimativa pode ser obtido a partir da área sob a curva normal como ilustrado na figura abaixo.



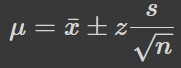
* 1. **Erro Inferencial**: O erro inferencial é definido pelo desvio padrão das médias amostrais σx¯ e pelo nível de confiança determinado para o processo. Cálculo:



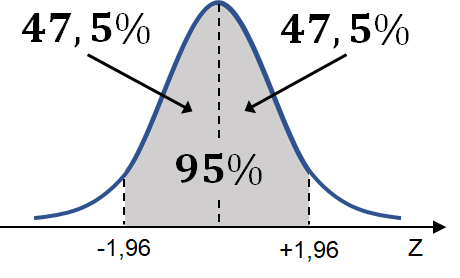
* 1. **Intervalo de Confiança**:
     1. Intervalo de confiança para a média da população:
        1. Com desvio padrão populacional conhecido:



* + - 1. Com desvio padrão populacional desconhecido:



* + 1. Exemplo: Suponha que os pesos dos sacos de arroz de uma indústria alimentícia se distribuem aproximadamente como uma normal de desvio padrão populacional igual a 150 g. Selecionada uma amostra aleatória de 20 sacos de um lote específico, obteve-se um peso médio de 5.050 g. Construa um intervalo de confiança para a média populacional assumindo um nível de significância de 5%.
       1. media\_amostral = 5050;
       2. significancia = 0.05;
       3. confianca = 1 – significância = 0.95;
       4. Obtendo Z:



0.95/2



0.5 + (0.95/2)



1.9 + 0.06

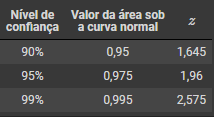


z = norm.ppf(0.975)

z



* + 1. Valores de z para os níveis de confiança mais utilizados:



* + 1. Obtendo σx¯:

desvio\_padrao = 150

n = 20

raiz\_de\_n = np.sqrt(n)

raiz\_de\_n



sigma = desvio\_padrao / raiz\_de\_n

sigma



* + 1. Obtendo *e*:

e = z \* sigma

e



* + 1. Solução 1: Manual

intervalo = (

    media\_amostral - e,

    media\_amostral + e

)

intervalo



* + 1. Solução 2: Python

norm.interval(alpha = 0.95, loc = media\_amostral, scale = sigma)

