Estatística Inferencial

Quem sou eu?

Alexsandro Pompeu

Data Product Manager na Farfetch Portugal

Minha formação

Bacharel em Sistemas da Informação pela Universidade Presbiteriana Mackenzie.

Pós Graduado em Análise em Big Data pela FIA

Minha carreira

SAP – Estagiário de Redes

Santander – Estagiário de Banco de Dados

Itaú Unibanco – Analista de Dados

PicPay – Tech Lead de Analytics

Via – Coordenador de Business Analytics

Estatística Inferencial

Conteúdo do Curso

Amostragem

Probabilidade

Intervalo de confiança

Teste de hipótese

Aula 1

Introdução à Estatística Inferencial

Introdução à Estatística Inferencial

Estatística descritiva

Descreve quantitativamente as características importantes do conjunto de dados (Amostra e população)

Estatística inferencial

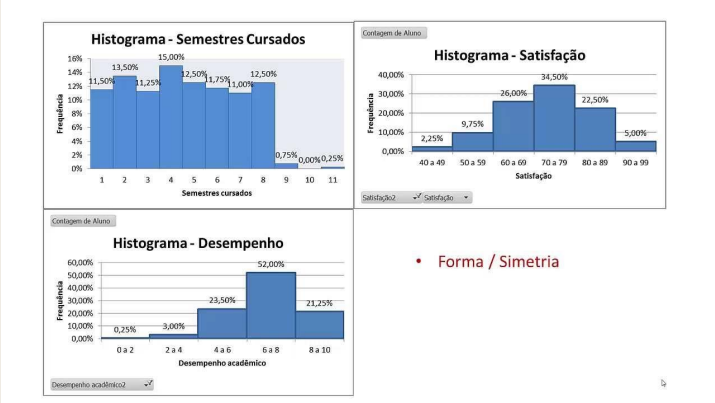
Estimar e fazer generalizações sobre características da população a partir de amostra

Exemplos

Estatística descritiva

- Média
- Mediana
- Desvio Padrão

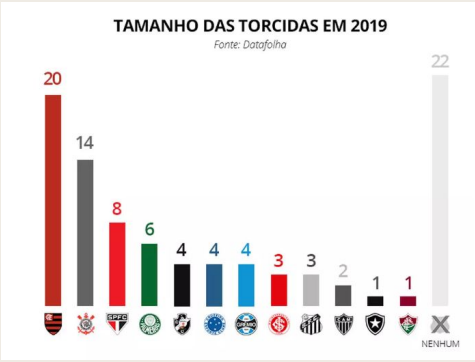
Visualização



- Forma / Simetria

Estatística inferencial

- Características da amostra para estimar a população



Introdução à Estatística Inferencial

Estatística causal

Busca determinar se mudanças em uma variável causam mudança em outra

Correlação não implica em causa

Aula 2

População e Amostra

Estatística Inferencial

População

Definição

Conjunto de elementos com uma característica comum

Exemplo

População do Brasil – Todas as pessoas que residem dentro do território brasileiro

Infectados pela Covid-19 no Brasil – Todas as pessoas que residem dentro do território brasileiro e que foram infectadas com o vírus da Covid-19

Amostra

Definição

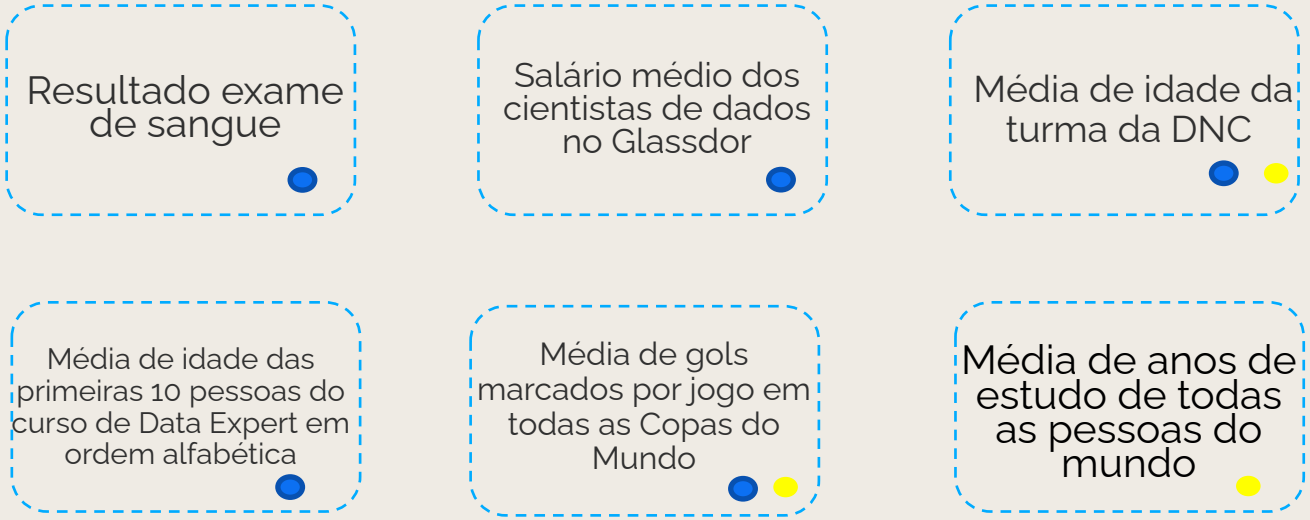
Subconjunto da população que é estudado para investigar as características ou o comportamento da população

Exemplo

Pesquisas eleitorais – São escolhidas pessoas aleatoriamente para responder qual candidato eles irão votar e assim estimar o comportamento da população como um todo

Estatística Inferencial

Exemplos de Amostra e População



Aula 3

Introdução à Estatística Inferencial

Estatística Inferencial

Estimar parâmetros

Definição

Medida que descreve certa característica dos elementos da população

Exemplo

Estatística Inferencial

Teste de hipótese

Definição

Usado para verificar se um determinado valor hipotético representa positivamente ou não em uma determinada ocasião.

Ele é baseado na utilização de uma amostra aleatória extraída de uma população de interesse com a finalidade de testar uma afirmação sobre um parâmetro ou característica desta população.

Exemplo

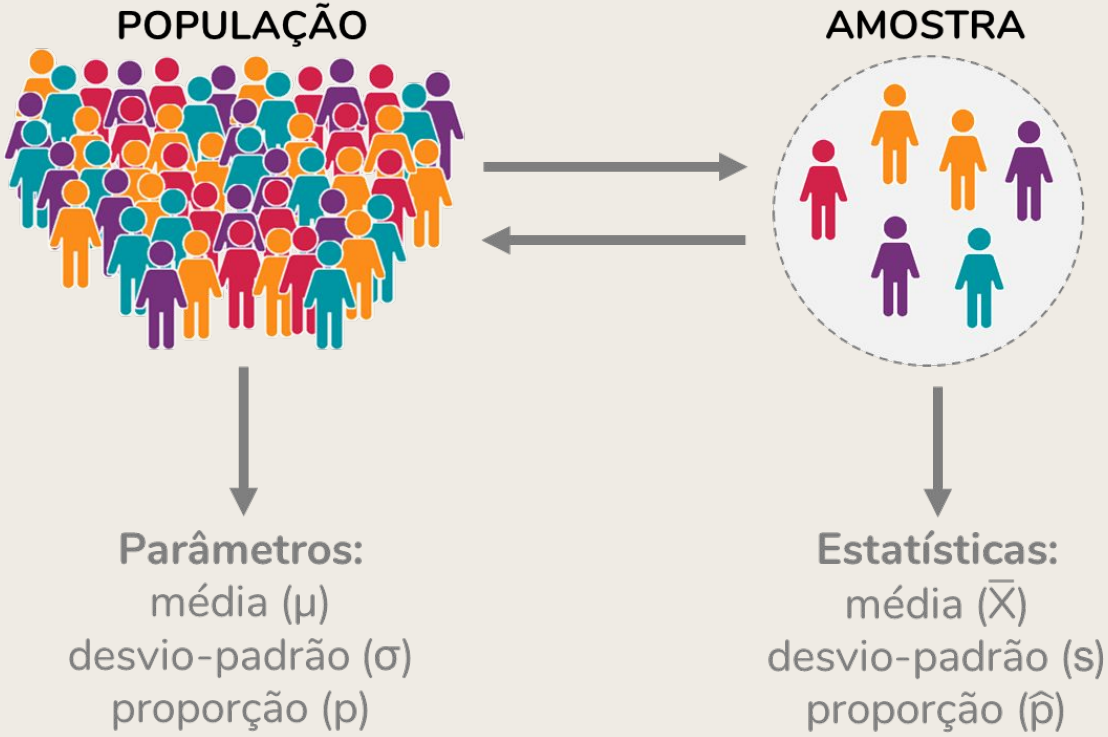
- Possibilidade do candidato A vencer as eleições
- A proporção dos motoristas habilitados de SP que tiveram suas carteiras apreendidas após a nova lei é maior que 2% ou não?

Estatística Inferencial

Exemplos

Estimação	Teste de hipótese
Qual é a probabilidade de “cara” no lançamento de uma moeda?	A moeda é honesta ou é desequilibrada?
Qual é a proporção de votos que o candidato A terá na próxima eleição?	O candidato A vencerá a eleição?
Qual é a proporção de motoristas habilitados de SP que tiveram suas carteiras apreendidas após a vigência da nova lei de trânsito?	A proporção dos motoristas habilitados de SP que tiveram suas carteiras apreendidas após a nova lei é maior que 2% ou não?

Estatística Inferencial



Aula 4

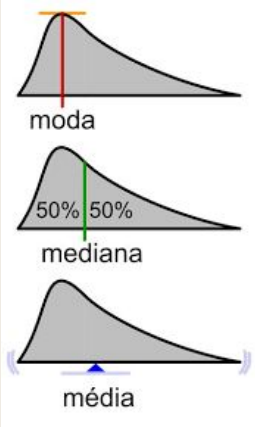
Retomando conceitos de estatística descritiva

Retomando conceitos....

Estatística descritiva

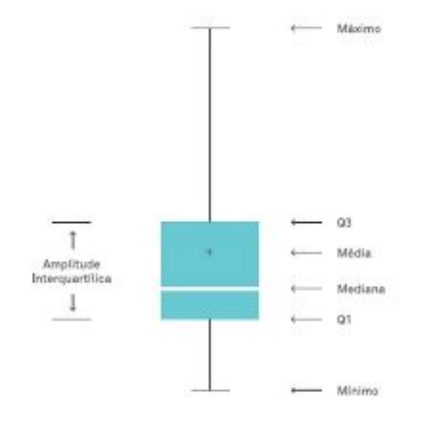
Medidas de tendência central

- Média
- Mediana
- Moda



Medidas de variabilidade

- Desvio padrão
- Variância
- Quartis
- Amplitude



Aula 5

Amostragem

Amostragem

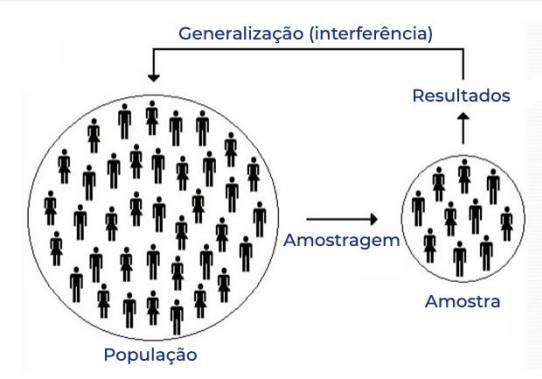
Definição

Área da estatística que estuda métodos de como determinar o tamanho de uma amostra e técnicas de seleção dessa amostra para se atingir um determinado objetivo.

Por que utilizar amostragem?

- Requer menos tempo que selecionar toda população
- Eficiente em termos de custo
- Prática, ágil e eficiente

Etapas do processo



1 - Identificação e definição da População Alvo (Target Population)

2 - Escolha do Método de Amostragem

3 - Determinar o tamanho da Amostra

4 – Coleta do dado necessário

Amostragem

Viés

O viés de seleção amostral ocorre quando um tipo de indivíduo da população tem maior ou menor chance de aparecer na amostra do que esperado pela teoria.

Causado por:

Dificuldade em selecionar a população

Seleção inadequada da amostra

Erro amostral

Erro estatístico que ocorre quando **a pessoa pesquisadora seleciona uma amostra que não representa a população alvo**. Sempre vai existir um erro mesmo que mínimo, até que a amostra seja a própria população, mas esse **erro pode ser minimizado**.

Aula 6

Amostragem Aleatória - Teórica

Amostragem

Amostragem Simples Aleatória

Cada indivíduo é escolhido de forma aleatória na população

Vantagem

Forma mais simples e direta de seleção de amostra

Desvantagem

Não tirar proveito do conhecimento sobre a população

Subgrupos da população tem interesses particulares e não podem ser incluídos com um número suficiente na amostra.

Aula 7

Amostragem Aleatória - Prática

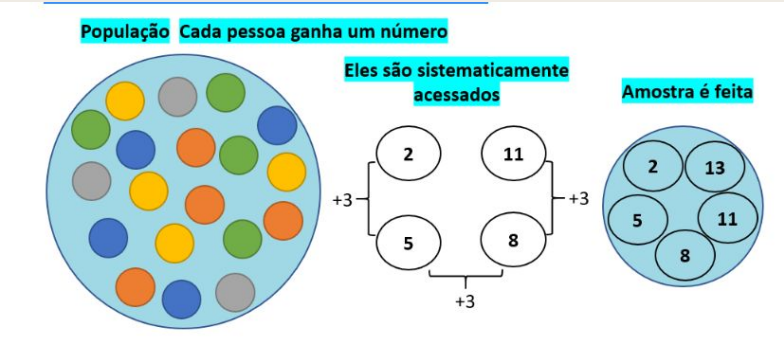
Aula 8

Amostragem Sistemática - Teórica

Amostragem

Amostragem Sistemática

É uma variação da amostragem simples. Após a identificação dos participantes, um determinado critério é eleito (por exemplo, a cada 5) e a seleção segue este formato.



Vantagem

Mais rápida de se implementar do que a amostragem aleatória simples

Desvantagem

Eventualmente, pode não representar bem subgrupos populacionais.

Aula 9

Amostragem Sistemática - Prática

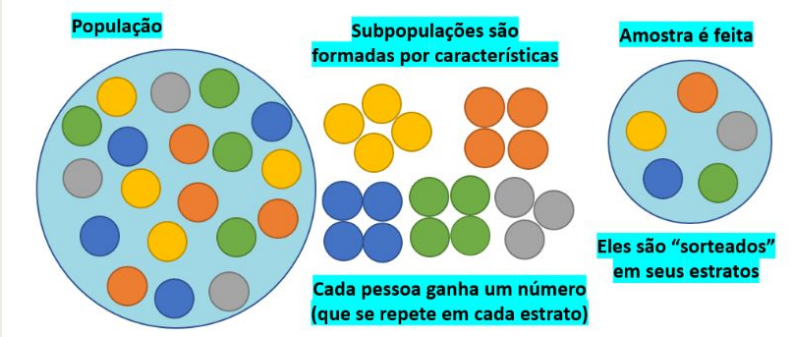
Aula 10

Amostragem Estratificada - Teórica

Amostragem

Amostragem Estratificada

Neste tipo de amostragem, a população é dividida em subpopulações em função de características em comum, o que é chamado de estrato. Em seguida, cada participante recebe uma identificação dentro de seu estrato e o processo de amostragem aleatória simples é feito dentro em cada estrato.



Vantagem

Reduz o viés da amostra

Cada subgrupo da população recebe uma representação adequada dentro da amostra

Desvantagem

Não pode ser usada em todos os estudos por obrigar ao pesquisador classificar os subgrupos

Dificuldade em classificar com precisão cada membro da população em um único estrato

Aula 11

Amostragem Estratificada - Prática

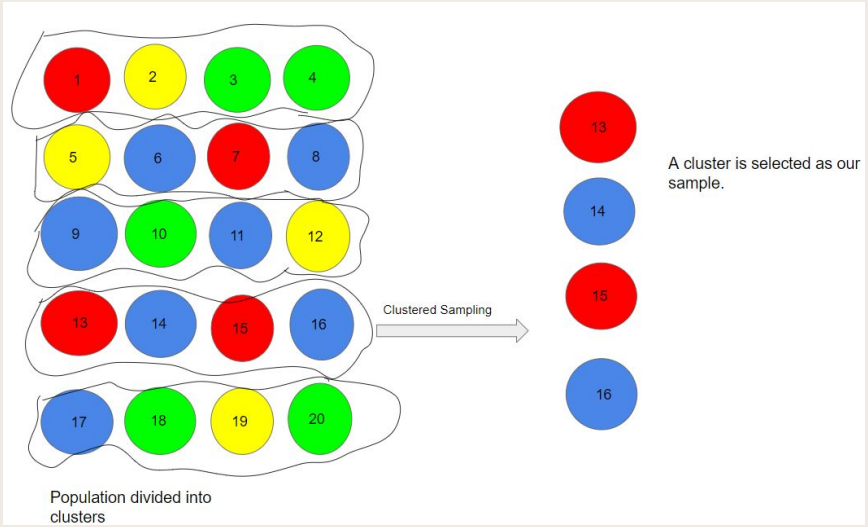
Aula 12

Amostragem Clusterizada - Teórica

Amostragem

Amostragem Clusterizada

Neste tipo de amostragem, a unidade de amostragem são os próprios clusters (isto é, todos os hospitais, ou todas as escolas, ou todas as empresas de camionagem, etc.), enquanto na amostragem estratificada, a unidade de amostragem são os sujeitos da população



Vantagem

Útil quando queremos estudar certas características de uma população muito grande

Reduzir custos em relação à estratificada

Desvantagem

Maior erro amostral

Requer tamanho parecido dos subgrupos

Aula 13

Amostragem Clusterizada - Prática

Aula 14

Probabilidade

Probabilidade

A probabilidade de um evento se refere à possibilidade ou a quão provável é que um evento aleatório aconteça.

Ao lançarmos um dado não viciado a probabilidade de cair a face com valor 5 é de 1/6 (0.167). E é a mesma para todas as faces.

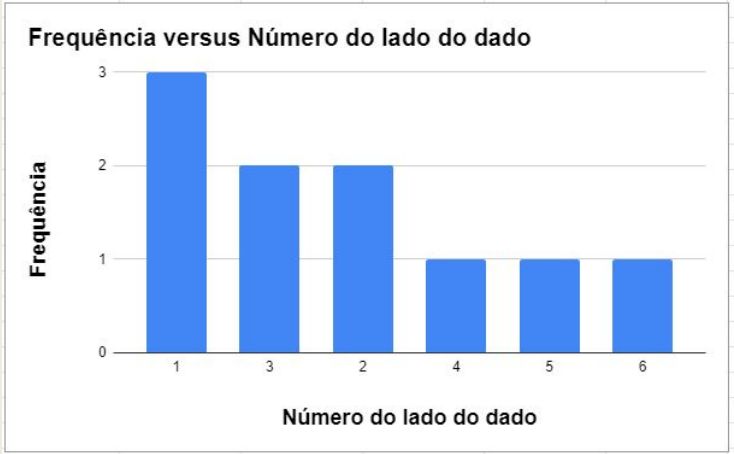
Probabilidade x Frequência

Uma distribuição descrever um agrupamento de dados e como esses dados se distribuem em um intervalo.

Probabilidade

Probabilidade de ocorrência de resultados em um experimento aleatório

Frequência



Contagem de ocorrência dentro dos intervalos

Aula 16

Tipos de variáveis

Variáveis algébricas x aleatórias

Algébricas

"Definimos como expressão algébrica uma expressão que contém letras e números, separados por operações básicas da Matemática, como a adição e a multiplicação.

$x + 1 = 5$

"x" é desconhecido mas pode ser encontrado

$x = y + 2$

Um valor foi atribuído a "X"

Aleatória

Uma variável aleatória é aquela cujo valor é sujeito a variações devido a aleatoriedade.
Há dois tipos: **Discreta e Contínua**.

Quando "x" é uma variável aleatória e possui um conjunto de valores podendo assumir qualquer desses valores aleatoriamente.

Exemplo:

- A frequência de vezes que o dado caiu com a face nº 2 para cima. Ele pode assumir qualquer valor dentro da amostra

Variáveis aleatórias

Tipo de variável aleatória	Característica de valores que pode assumir	Exemplos
Discreta	Valores distintos (“separados”) ou finitos (contáveis)	•Jogar uma moeda (cara ou coroa) •Quantidade de pessoas que visitam uma loja •Teste de Covid (positivo ou negativo)
Contínua	Valores em intervalo contínuo (infinitos)	• Distância que uma moeda viaja ao ser arremessada (1cm, 1.1 cm, 1.11 cm) • pH médio de rios e oceanos • Temperatura em um dia

Variáveis aleatórias

Discreta

- Número de filhos
- Número de acessos ao APP
- Número de acidentes

Contínua

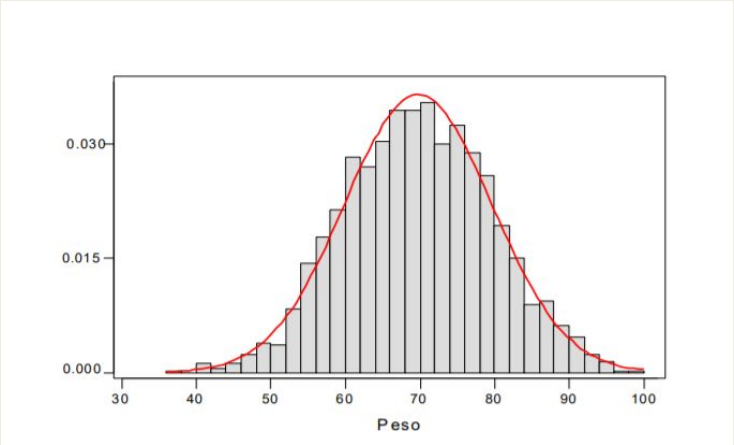
´- Altura
- Peso
- Salário

Aula 17

Teorema do Limite Central

Teorema do Limite Central

Distribuição Normal ou Gaussiana



Eventos aleatórios que seguem um padrão

Média, mediana e moda possuem o mesmo valor

Quanto maior a curva, mais dispersos os dados estão da média

Teorema do Limite Central

População

150 150 150

Amostra **Amostra** **Amostra**

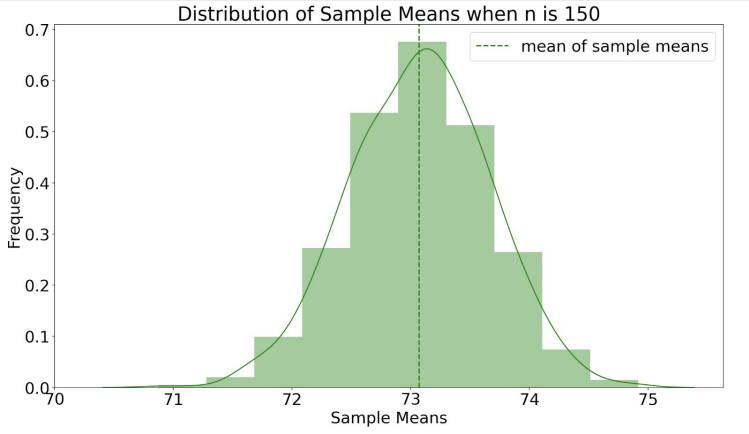
1.000 amostras

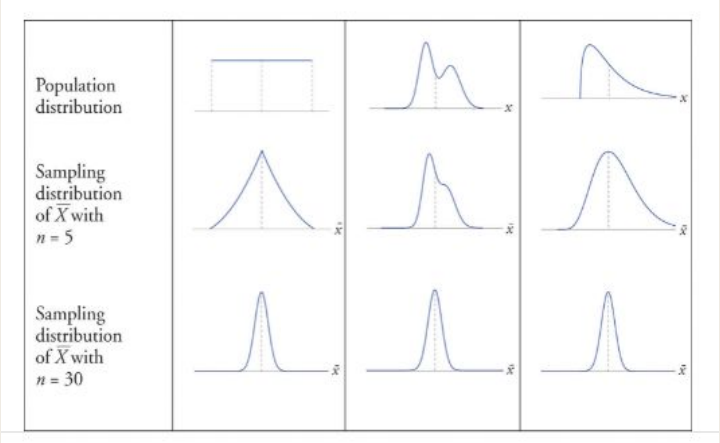
Média das amostras

Amostra (n=150)	Média (em anos)
#1	85
#2	70
#3	66
#4	62
...	...
#1000	80

Teorema do Limite Central

Distribuição Amostral das médias Amostrais





Qualquer distribuição amostral de médias de uma população com qualquer distribuição é aproximadamente uma distribuição normal.*
*se o tamanho da amostra for pelo menos 30.

Independente da forma inicial da distribuição populacional a **distribuição amostral da média vai aproximar uma distribuição normal**. Quando o tamanho da **amostra aumenta a distribuição amostral vai ficar mais estreita e mais normal** (centrada na média)..

Aula 18

Teorema do Limite Central - Teórico

Aula 19

Intervalo de confiança – Estimar parâmetros

Estimar Parâmetros

Usar a estatística para tirar conclusões sobre parâmetros populacionais.

Estimativa Pontual

Fazemos uma única estimativa (um valor) para um determinado parâmetro populacional.

Exemplo prático: Expectativa média de vida de um brasileiro: 75 anos.

Estimativa Intervalar

Fazemos uma estimativa de um intervalo de valores possíveis, no qual se admite esteja o parâmetro populacional.

Exemplo prático: Expectativa média de vida de um brasileiro: entre 70 anos e 80 anos, isto é, uma estimativa pontual (75 anos), com margens de erro de 5 anos para mais ou para menos.

Intervalo de Confiança

Intervalo de confiança (Margem de erro)

A questão aqui é **quanto de erro será tolerado na pesquisa**. A margem de erro, também conhecida como intervalo de confiança, é **expressa em média** – e é possível definir quanta diferença você permitirá **entre a média da sua amostragem e a média da sua população**

Exemplos

Média de longevidade no Brasil está entre 72 – 80 anos com 95% de **nível de confiança.**

Desvio padrão da renda per capita está entre 2.500 – 12.400 com 99% de **nível de confiança.**

Premissas

- Amostragem simples aleatória da população
- População normal
- Conhecimento do desvio populacional*

*há formas de calcular se o desvio for desconhecido

Cálculo

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

Em que:

CI = Intervalo de confiança
$\bar{x}$ = média da amostra
z = valor do nível de confiança
s = desvio padrão da amostra*
n = tamanho da amostra

Nível de Confiança

Nível de confiança

A questão aqui é o quanto você quer **estar confiante de que a média real** está dentro da sua **margem de erro.**

Exemplos

Se você tiver coletado uma centena de amostras, e tiver calculado 95% de intervalos de confiança, você esperaria que aproximadamente 95 dos intervalos contivesse o parâmetro populacional, como a média da população.

Nível de Confiança	Valor de Z*-
80%	1.28
90%	1.645 (convencional)
95%	1.96
98%	2.33
99%	2.58

Desvio Padrão

Definição

Deverá ser estimado o quanto as respostas que você receber variarão umas das outras e da média. Um desvio padrão baixo significa que todos os valores estarão próximos da média, e um desvio padrão alto significa que eles estarão mais espalhados em uma faixa mais longa com números bem baixos e bem altos nas extremidades.

Exemplos

Quando ainda não sabemos essa medida, assumimos o valor de 0,5, que vai ajudar a garantir que a sua amostragem seja suficiente.

Intervalo de Confiança - Estimar Parâmetros

Exemplo

Empresa está fazendo um estudo sobre o comprimento de lâmpadas produzidas para otimizar materiais

População

200 mil

N= 135

Amostra

Qual o intervalo de confiança para a média de comprimento dessas lâmpadas?

$\bar{x}$ = 0.988 cm

Nível de confiança 95% e desvio padrão (sigma = 0.028)*
***nem sempre conhecido**

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

IC = 0.988 +- 1.96 x (0.028/sqrt(135))
IC = 0.988 +- 0.0047
(0.983, 0.993)

O que significa o 95%? Se repetirmos o processo, a média populacional estará no intervalo estimado em 95% das vezes (19 de 20)

Tamanho da amostra - Estimar Parâmetros

Exemplo – População conhecida

Tamanho da amostra =

$$\frac{\frac{z^2 \times p\,(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p\,(1-p)}{e^2 N}\right)}$$

z = escore z
p = desvio padrão
e = margem de erro
N = tamanho da população

Nível de confiança de 95%, um desvio padrão de 0,5 e uma margem de erro (intervalo de confiança) de +/- 5% e uma população de 2000.

Nível de Confiança	Valor de Z*-
80%	1.28
90%	1.645 (convencional)
95%	1.96
98%	2.33
99%	2.58

((1,96)^2 x 0,5(0,5)) / (0,05)^2 / 1 + (1,96)^2 x 0,5(0,5) / 0,05^2 x 2000)

323 participantes

Tamanho da amostra - Estimar Parâmetros

Exemplo – População não conhecida

$$\text{Tamanho da amostra} = \frac{(\text{Z-score})^2 * \text{DesvioPadrão} * (1 - \text{DesvioPadrão})}{(\text{margem de erro})^2}$$

Nível de confiança de 95%, um desvio padrão de 0,5 e uma margem de erro (intervalo de confiança) de +/- 5%.

Nível de Confiança	Valor de Z*-
80%	1.28
90%	1.645 (convencional)
95%	1.96
98%	2.33
99%	2.58

((1,96)2 x 0,5(0,5)) / (0,05)2

(3,8416 x 0,25) / 0,0025

0,9604 / 0,0025

384,16

385 participantes

Aula 20

Teste de hipótese

Teste de hipótese

Definição

Decidir, com base na estatística amostral, se uma hipótese sobre um parâmetro populacional deve ou não ser rejeitada (se está certa ou errada e com qual probabilidade

H0 - hipótese nula normalmente retrata que não existe uma relação no fenômeno medido.

H1 - hipótese alternativa fala sobre haver efeito

Homens e mulheres têm salários diferentes quando saem da graduação:

H0: média salarial de homens = média salarial de mulheres
H1: média salarial de homens =/= média salarial de mulheres
* Hipóteses são sobre parâmetros, nunca estatísticas

Peso médio de produção de queijo difere do peso alvo de 750g?

H0: peso médio = 750
H1: peso médio != 750