

dnc>class



Introdução à Estatística Inferencial

Consultora: Amanda

Olá, meu nome é Amanda

- Formação na área de humanas, RI > muito espaço pra quem não é de exatas
- IBM, 4 anos (manutenção preditiva) – Advanced Analytics Consultant
- Aluna especial de mestrado em Computação > interesse em pesquisa
- Senior Data Scientist Loft (valuation)
- Entusiasta de D&I em tech: Espero que tenha mulheres no curso! Sempre disponível pra falar do assunto.

Retomando conceitos...

1. O que já foi visto sobre **estatística descritiva**

Medidas de tendência central

Média, mediana, moda

Medidas de variabilidade

Desvio padrão, variância, quartis, amplitude

Medida	Sensibilidade a outliers	Alteração quando dado muda	Identificação visual
Média	S	S	N
Mediana	N	N	N
Moda	N	N	S

Entender distribuições e fenômenos: distribuições com mesmas medidas centrais, mas uma mais espalhadas/extensa

S= Sim, característica presente
N= Não, característica ausente

O que veremos neste módulo

01.
Introdução a
Estatística
Inferencial –
Parte 1

02.
Amostra e
População

03.
Introdução a
Estatística
Inferencial –
Parte 2

04.
Abordagem
Frequentista e
Bayesiana

05.
Probabilidade e
Distribuição

06.
Teorema do
Limite Central

07.
Intervalo de
Confiança

08.
Testes de
Hipótese

09.
Pacotes em
Python e R

10.
Projeto Final

Introdução a Estatística Inferencial

1. Estatística descritiva vs Estatística inferencial

Estatística Descritiva (Dedutiva)

- organização, descrição, sumarização de dados **tanto de população como de amostras**

- **Dedução:** método de raciocínio lógico que parte de uma certeza para a interpretação de dados ou fatos (da causa para os efeitos) -> geral para particular

Estatística Inferencial (Indutiva)

- estimar / criar **inferências** e fazer generalizações sobre características de uma **população** baseadas nos dados de **amostra**

- **Indução:** parte-se de dados ou fatos semelhantes para a definição de uma certeza comum (dos efeitos para as causas) -> particular para geral

Introdução a Estatística Inferencial

1. Estatística descritiva vs estatística inferencial

INFERÊNCIA

uma definição feita com **base em informações** ou um **raciocínio que usa dados** disponíveis para se **chegar a uma conclusão**. Inferir é chegar a um **resultado, por lógica**, com base na interpretação de outras informações.

raciocínio concluído ou desenvolvido **a partir de indícios**

A seguir...

POPULAÇÃO E AMOSTRA

Inferência causal

Inferência Estatística

- estimar / criar **inferências** e fazer generalizações sobre características de uma **população** baseadas nos dados de **amostra**.
- Coletar mais dados deixa mais próximo da realidade

Inferência Causal

- estimar **contrafactuais** – resultados que teriam ocorrido se o **tratamento** tivesse sido diferente (dados não disponíveis). Causalidade é comparar resultados obtidos com os contrafactuais. Ex. testes de medicamentos – jamais teremos resultados de não tratamento.
- Usa inferência estatística para estimar tratamento médio, por exemplo.
- Coletar mais dados não ajuda necessariamente

Recap - Introdução a Estatística Inferencial

Estatística Descritiva

(Dedutiva): organização, descrição, sumarização de dados **tanto de população como de amostras**

Estatística Inferencial

(Indutiva): estimar / criar **inferências** e fazer generalizações sobre características de uma **população** baseadas nos dados de **amostra**

Dedução:

método de raciocínio lógico que parte de uma certeza para a interpretação de dados ou fatos (da causa para os efeitos) -> geral para particular

Indução:

parte-se de dados ou fatos semelhantes para a definição de uma certeza comum (dos efeitos para as causas) -> particular para geral

Inferência: Inferir é chegar a um **resultado, por lógica**, com base na **interpretação de outras informações**.

Raciocínio desenvolvido a partir de **indícios**

Inferência

Causal: estimar **contrafactuais** – resultados que teriam ocorrido se o **tratamento** tivesse sido diferente (dados não disponíveis). Causalidade é comparar resultados obtidos com contrafactuais.

dnc>class



Amostra e População

Amostra e População

1. Estatística descritiva vs estatística inferencial

Estatística Descritiva

- organização, descrição, sumarização de dados **tanto de população como de amostras**

Estatística Inferencial

- estimar / criar **inferências** e fazer generalizações sobre características de uma **população** baseadas nos dados de **amostra**

Amostra e População

POPULAÇÃO:

conjunto de elementos
com uma característica
comum

AMOSTRA:

subconjunto da
população

Definição de população alvo quem
faz é quem conduz análise!

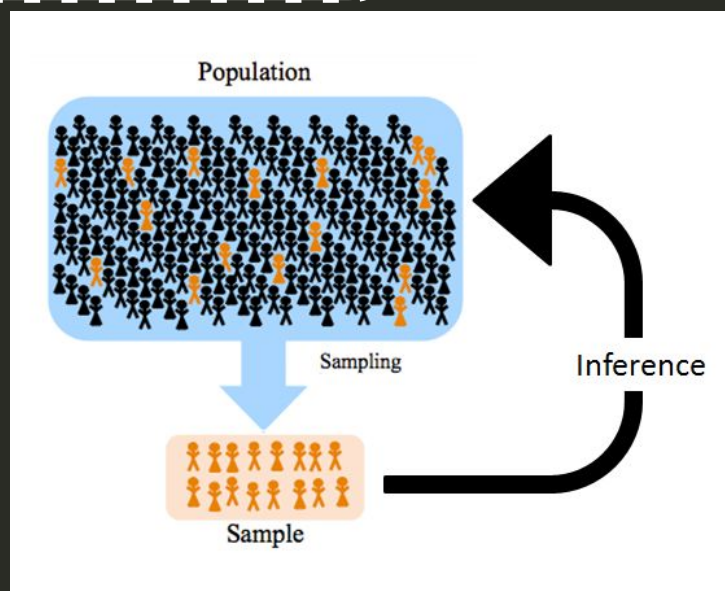
Amostra e População

POPULAÇÃO:

conjunto de elementos
com uma característica
comum

AMOSTRA:

subconjunto da
população



Amostra e População

POPULAÇÃO:

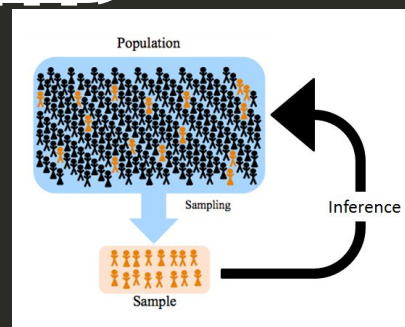
conjunto de elementos
com uma característica
comum

AMOSTRA:

subconjunto da
população

PARÂMETRO

*medida que descreve
certa característica dos
elementos da população*



ESTATÍSTICA

*medida associada
aos dados de uma
amostra extraída da
população*

Amostra e População

POPULAÇÃO:

conjunto de elementos com uma característica comum

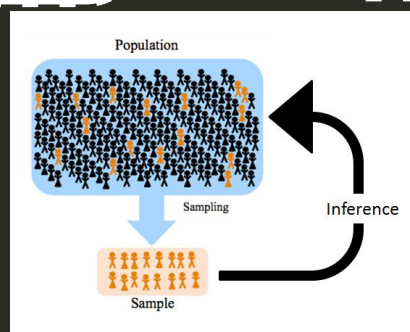
AMOSTRA: subconjunto da população

PARÂMETRO

medida que descreve certa característica dos elementos da população

ERRO AMOSTRAL

diferença entre uma estatística e o parâmetro que se quer estimar



ESTATÍSTICA

medida associada aos dados de uma amostra extraída da população

ESTIMADOR

uma função que calcula uma estimativa de um determinado parâmetro populacional. Ex. média aritmética amostral

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro(população)**?

Resultado de colesterol em exame de sangue

Renda per capita com base no Censo 2010

Salário médio de cientistas de dados publicados no Glassdoor

Média de idade da turma de Data Expert

Média de idade das primeiras 10 pessoas do curso de Data Expert em ordem alfabética

Média de gols marcados por jogo em todas as Copas do Mundo

Média de anos de estudos de todas as pessoas do mundo

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro(população)**?

Resultado de colesterol em exame de sangue ●

Renda per capita com base no Censo 2010

Salário médio de cientistas de dados publicados no Glassdoor

Média de idade da turma de Data Expert

Média de idade das primeiras 10 pessoas do curso de Data Expert em ordem alfabética

Média de gols marcados por jogo em todas as Copas do Mundo

Média de anos de estudos de todas as pessoas do mundo

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro(população)**?

Resultado de
colesterol em
exame de
● sangue

Renda per
capita com
base no
● censo 2010

Salário médio
de cientistas
de dados
publicados no
Glassdoor

Média de
idade da
turma de
Data Expert

Média de idade das
primeiras 10
pessoas do curso
de Data Expert em
ordem alfabética

Média de gols
marcados por
jogo em todas as
Copas do Mundo

Média de anos
de estudos de
todas as pessoas
do mundo

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro (população)**?

Resultado de
colesterol em
exame de
sangue ●

Renda per
capita com
base no Censo
2010 ●

Salário médio de
cientistas de
dados
publicados no
Glassdoor ●

Média de idade
da turma de
Data Expert

Média de idade das
primeiras 10 pessoas
do curso de Data
Expert em ordem
alfabética

Média de gols
marcados por jogo
em todas as Copas
do Mundo

Média de anos de
estudos de todas
as pessoas do
mundo

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro (população)**?

Resultado de
colesterol em
exame de
sangue ●

Renda per
capita com
base no Censo
2010 ●

Salário médio de
cientistas de
dados
publicados no
Glassdoor ●

Média de idade
da turma de
Data Expert ●●

Média de idade das
primeiras 10 pessoas
do curso de Data
Expert em ordem
alfabética

Média de gols
marcados por jogo
em todas as Copas
do Mundo

Média de anos de
estudos de todas
as pessoas do
mundo

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro (população)**?

Resultado de
colesterol em
exame de
sangue ●

Renda per
capita com
base no Censo
2010 ●

Salário médio de
cientistas de
dados
publicados no
Glassdoor ●

Média de idade
da turma de
Data Expert ●●

Média de idade das
primeiras 10 pessoas
do curso de Data
Expert em ordem
alfabética ●

Média de gols
marcados por jogo
em todas as Copas
do Mundo

Média de anos de
estudos de todas
as pessoas do
mundo

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro (população)**?

Resultado de
colesterol em
exame de
sangue ●

Renda per
capita com
base no Censo
2010 ●

Salário médio de
cientistas de
dados
publicados no
Glassdoor ●

Média de idade
da turma de
Data Expert ●●

Média de idade das
primeiras 10 pessoas
do curso de Data
Expert em ordem
alfabética ●

Média de gols
marcados por jogo
em todas as Copas
do Mundo ●●

Média de anos de
estudos de todas
as pessoas do
mundo

Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro (população)**?

Resultado de
colesterol em
exame de
sangue



Renda per
capita com
base no Censo
2010



Salário médio de
cientistas de
dados
publicados no
Glassdoor



Média de idade
da turma de
Data Expert



Média de idade das
primeiras 10 pessoas
do curso de Data
Expert em ordem
alfabética



Média de gols
marcados por jogo
em todas as Copas
do Mundo



Média de anos de
estudos de todas
as pessoas do
mundo



Amostra e População

2. Exemplos: **estatística (amostra)** ou **parâmetro (população)**?

Resultado de
colesterol em
exame de
sangue ●

Renda per
capita com
base no Censo
2010 ●

Salário médio de
cientistas de
dados
publicados no
Glassdoor ●

Média de idade
da turma de
Data Expert ●●

Média de idade das
primeiras 10 pessoas
do curso de Data
Expert em ordem
alfabética ●

Média de gols
marcados por jogo
em todas as Copas
do Mundo ●●

Média de anos de
estudos de todas
as pessoas do
mundo ●

SUA VEZ! Consegue pensar em uma
estatística/parâmetro e definir qual seria a
amostra e população? Manda o seu desafio
pra turma adivinhar!

Amostra e População

3. Amostra e população no mundo real

Raramente teremos dados completos de uma população!

Precificação de apartamentos: não temos todos os dados de preço de todos os apartamentos de SP

Conquista de clientes pessoa jurídica: não temos todas as informações de todas as empresas que existem no Brasil

Envio de e-mails de marketing: não temos dados de todas as pessoas que têm interesse no produto

Validação e feedback sobre um novo jogo de vídeo game: nem todas as pessoas que instalaram o jogo responderam ao questionário

Insights de People Analytics: empresa tem os dados de toda a população (pessoas que trabalham na empresa)

Se você trabalha ou desenvolve projetos na faculdade, já teve que trabalhar com dados? Que tipo de dados a sua empresa ou um empresa que você gosta precisa? São dados amostrais ou populacionais?

Recap - Amostra e População

POPULAÇÃO:

conjunto de elementos com uma característica comum

PARÂMETRO

medida que descreve certa característica dos elementos da população

ERRO AMOSTRAL

diferença entre uma estatística e o parâmetro que se quer estimar

AMOSTRA: subconjunto da população que se usa para gerar inferência

ESTATÍSTICA

medida associada aos dados de uma amostra extraída da população

ESTIMADOR

uma função que calcula uma estimativa de um determinado parâmetro populacional. Ex. média aritmética amostral

dnc>class

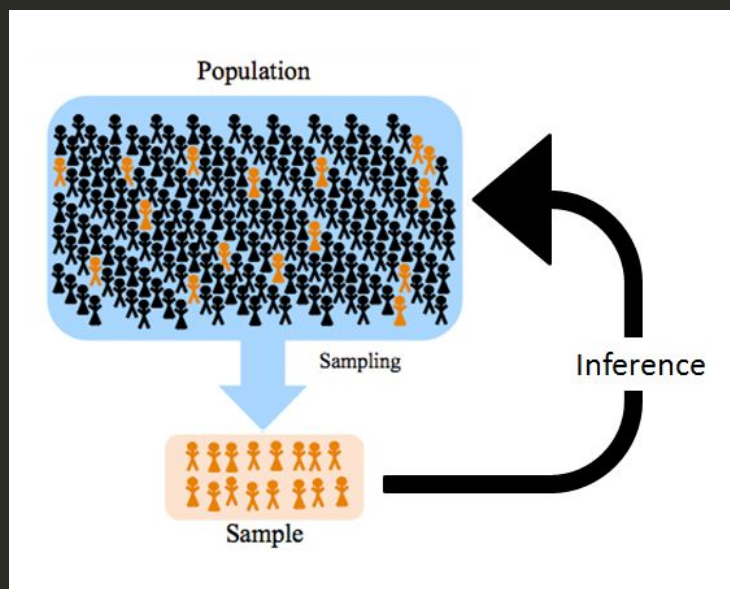


2.2 Amostragem

Amostra e População

4. Amostragem

COMO CHEGAMOS NA AMOSTRA?



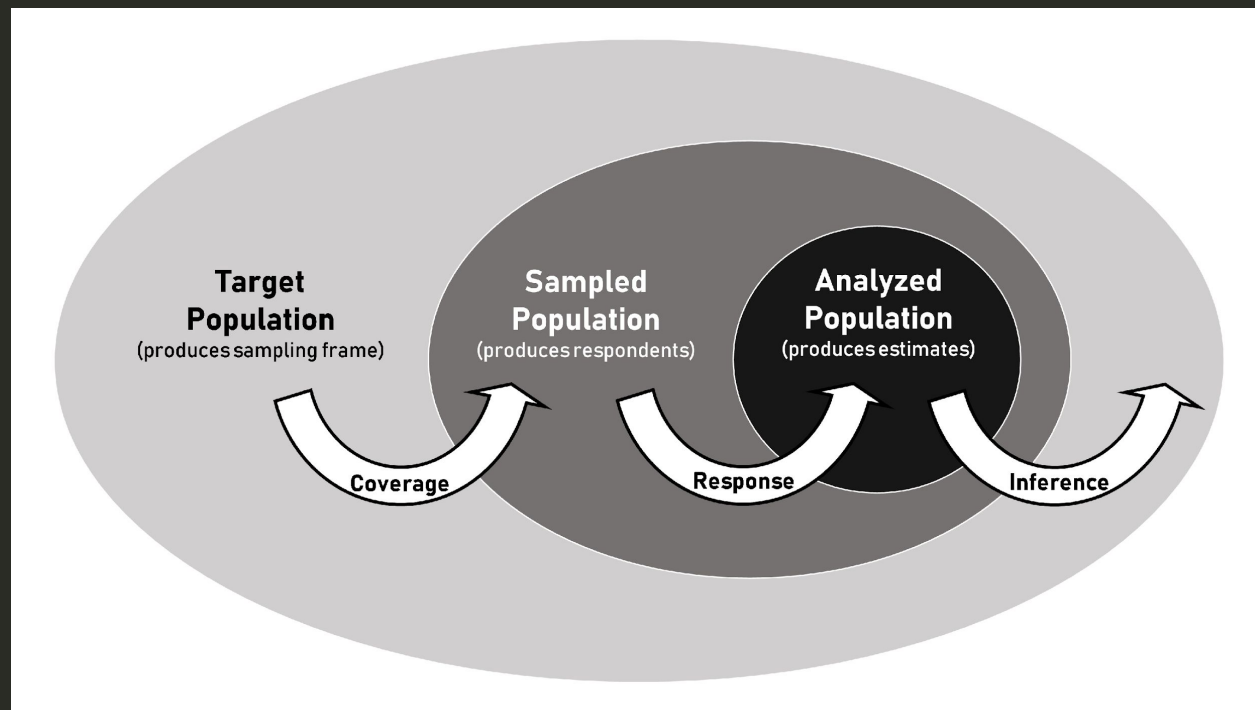
Amostragem é um processo de **seleção de um subconjunto** da população de interesse que **gera a amostra**. A amostragem é uma área da estatística que estuda **métodos de como determinar o tamanho** de uma amostra e **técnicas de seleção dessa amostra** para se atingir determinado objetivo.

Por que Amostragem?

- **Seleção de amostra requer menos tempo que selecionar toda a população**
- **É uma forma eficiente em termos de custos**
- **Análise de amostra é mais prática, ágil e eficiente**

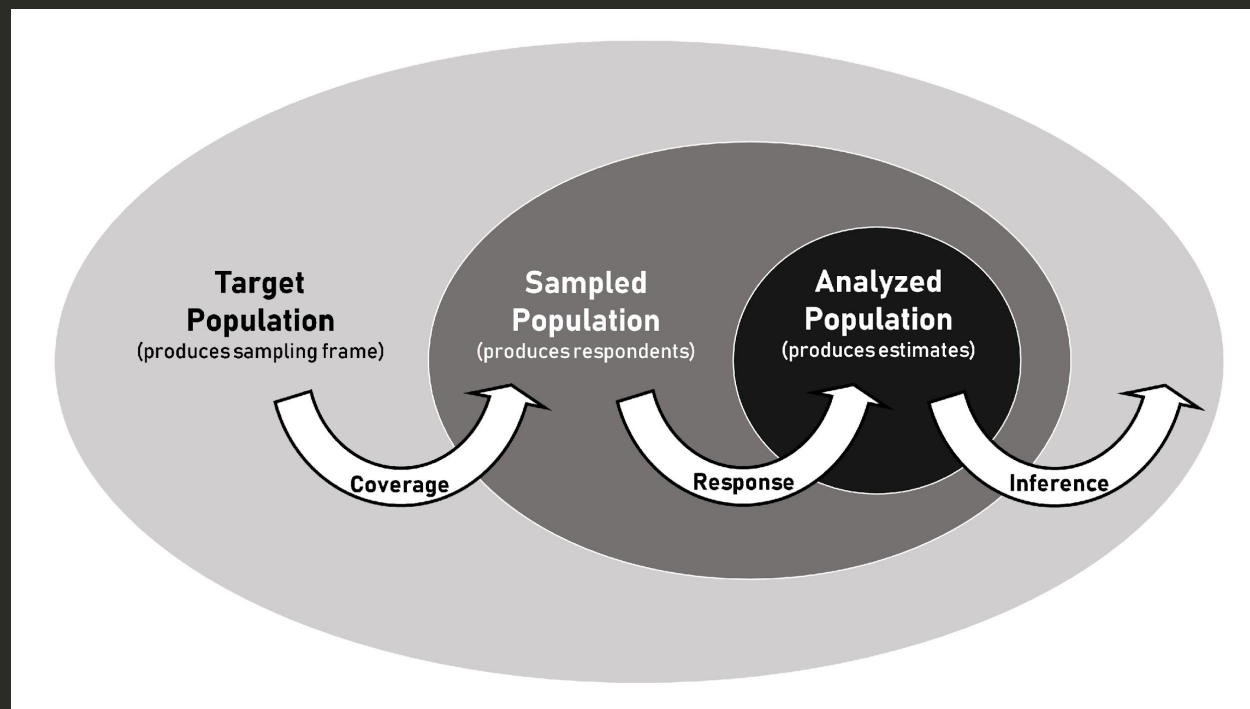
Amostra e População

4. Amostragem: passos do processo



Amostra e População

4. Amostragem: passos do processo



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

2.
Seleção da
“Sampling
Frame”

3.
Escolha do
Método de
Amostragem

4.
Determinar o
tamanho da
Amostra

5.
Coleta do
dato
necessário

Amostra e População

4. Amostragem: desafios

Viés de seleção (Selection bias):

Viés potencial introduzido por **seleção de itens** (método de amostragem) **que não é aleatória**. A amostra se torna pouco representativa da população alvo que se pretende analisar.

Origem: baixa taxa de resposta, substituição de indivíduos no processo de amostragem, uso de vocabulário enviesado (“wording”), etc.

Erro amostral (Sampling error):

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**.

Amostra e População

4. Desafios de Amostragem: Eleição de 1948



Jornal Tribuna > amostra superdimensionada de Republicanos em seus dados por uma simples razão: a pesquisa foi conduzida totalmente por telefone. Como pessoas ricas eram mais propensas a ter telefone e a se identificarem como Republicanas a pesquisa foi fortemente distorcida pendendo para Dewey.

Leitura: conta o caso, como isso é um problema para cientista de dados ainda hoje e como fazer diferente.

Recap - Amostra e População

AMOSTRAGEM:

é um processo de **seleção de um subconjunto** da população de interesse que **gera a amostra**

ETAPAS DE AMOSTRAGEM

1.

Identificação e definição da População Alvo

2.

Seleção da "Sampling Frame"

3.

Escolha do Método de Amostragem

4.

Determinar o tamanho da Amostra

5.

Coleta do dado necessário

VIÉS DE SELEÇÃO: Viés potencial introduzido por **seleção de itens** (método de amostragem) **que não é aleatória**.

ERRO AMOSTRAL: Erro estatístico que ocorre quando a **pessoa pesquisadora seleciona uma amostra que não representa a população alvo**.

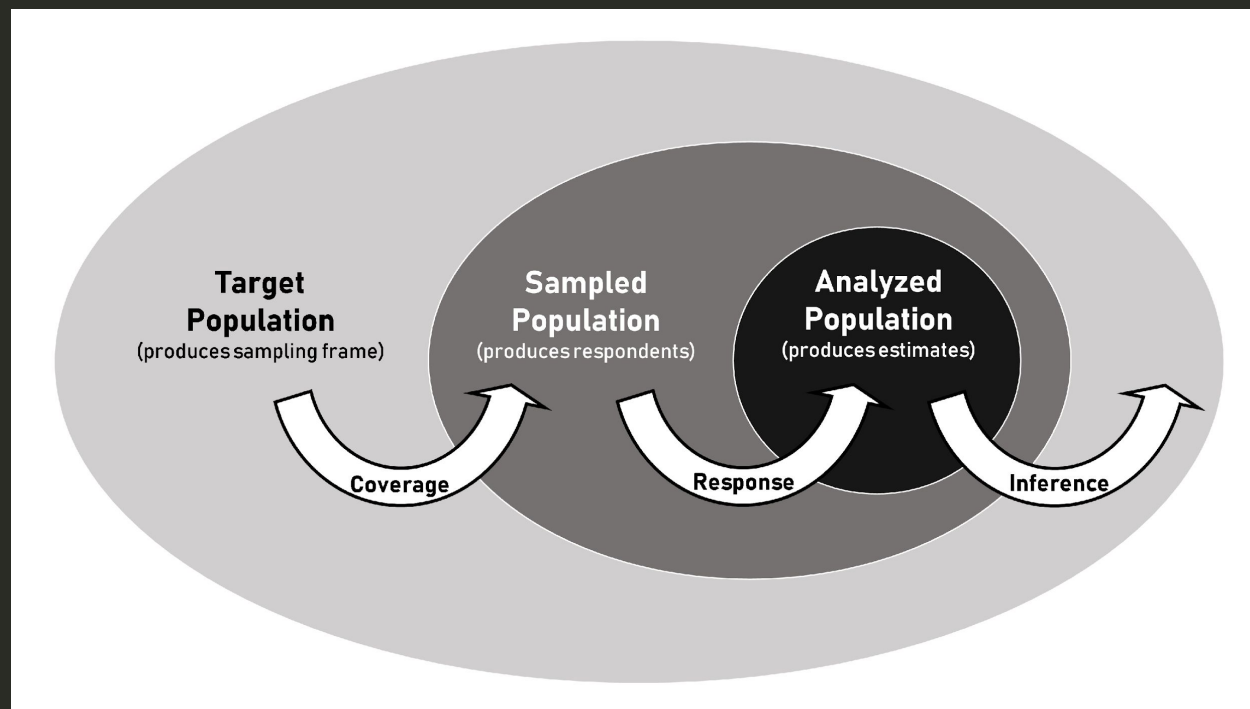
dnc>class



2.3 Amostragem: Etapas 1 e 2

Amostra e População

4. Amostragem: passos do processo



1.

Identificação e
definição da
População Alvo
(Target
Population)

2.

Seleção da
“Sampling
Frame”

3.

Escolha do
Método de
Amostragem

4.

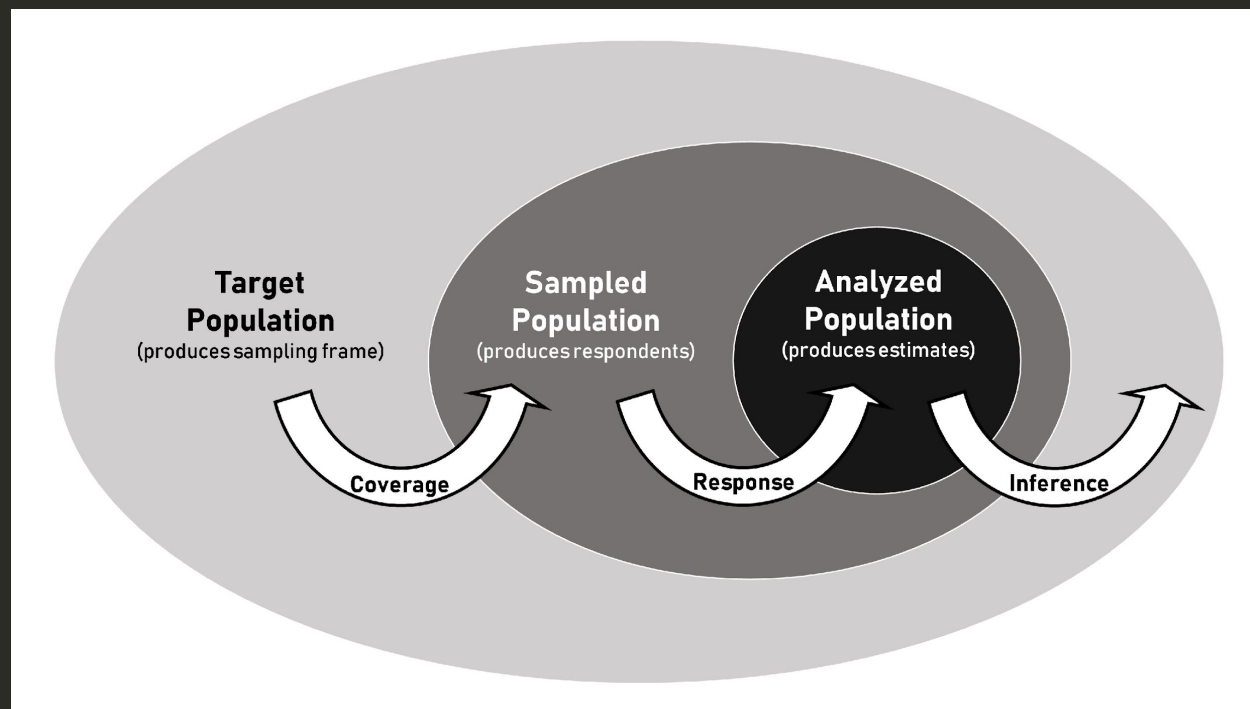
Determinar o
tamanho da
Amostra

5.

Coleta do
dado
necessário

Amostra e População

4. Amostragem: passos do processo



1.

Identificação e
definição da
População Alvo
(Target
Population)

2.

Seleção da
“Sampling
Frame”

3.

Escolha do
Método de
Amostragem

4.

Determinar o
tamanho da
Amostra

5.

Coleta do
dado
necessário

Amostra e População

4. Amostragem: passos do processo

1.

Identificação e
definição da
População Alvo
(Target
Population)

2.

Seleção da
“Sampling
Frame”

Uma **sampling frame** é uma lista de todos os itens da sua população alvo. A diferença entre a população alvo e a sampling frame é que a **população** é um conceito genérico enquanto a **sampling frame** é concreta, específica e contabiliza itens acessíveis para a amostragem.

População	Sampling Frame
Pessoas que estão fazendo o curso da DNC de Data Expert	Beatriz Salgado, Bruno Santos, João Oliveira, Luana Dias, etc.
Estados do Brasil	AC, AL, AM, AP, BA, CE, DF, ES, GO, MA, MT, MS, MG, PA, PB, PR, PE, PI, RJ, RN, RS, RO, RR, SC, SP, SE, TO

Amostra e População

4. Amostragem: passos do processo

2.

**Seleção da
“Sampling
Frame”**

QUALIDADE DA SAMPLING FRAME

- Organização e ordem lógica
- Itens únicos e sem duplicação
- Apenas elementos da população de interesse
- Todos os elementos podem ser acessados

Frame digitalizado

**Lista em
papel**

**Base de
dados
digital**

**Lista
telefônica**

**Mapa ou
rua**

**Registro
de
funcionári
os**

Amostra e População

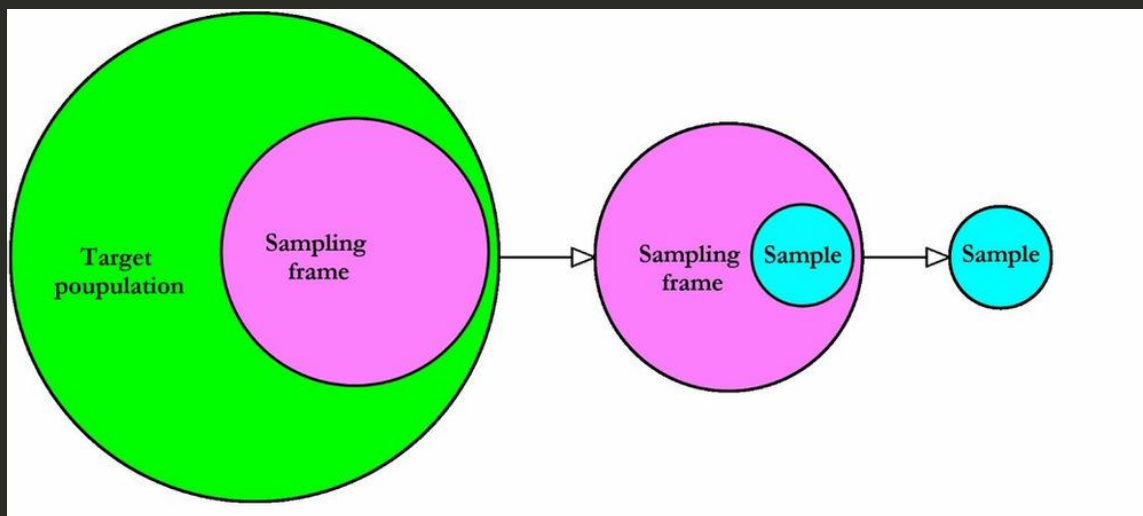
4. Amostragem: passos do processo

2.

Seleção da
"Sampling
Frame"

QUALIDADE DA SAMPLING FRAME

- Organização e ordem lógica
- Itens únicos e sem duplicação
- Apenas elementos da população de interesse
- Todos os elementos podem ser acessados



Recap - Amostra e População

AMOSTRAGEM:

é um processo de **seleção de um subconjunto** da população de interesse que **gera a amostra**

ETAPAS DE AMOSTRAGEM

1.

Identificação
e definição da
População
Alvo

2.

Seleção da
“Sampling
Frame”

3.

Escolha
do Método
de
Amostragem

4.

Determinar
o tamanho
da
Amostra

5.

Coleta
do dado
necessário

ETAPA 1 VS ETAPA 2

QUALIDADE DE SAMPLING
FRAME

dnc>class

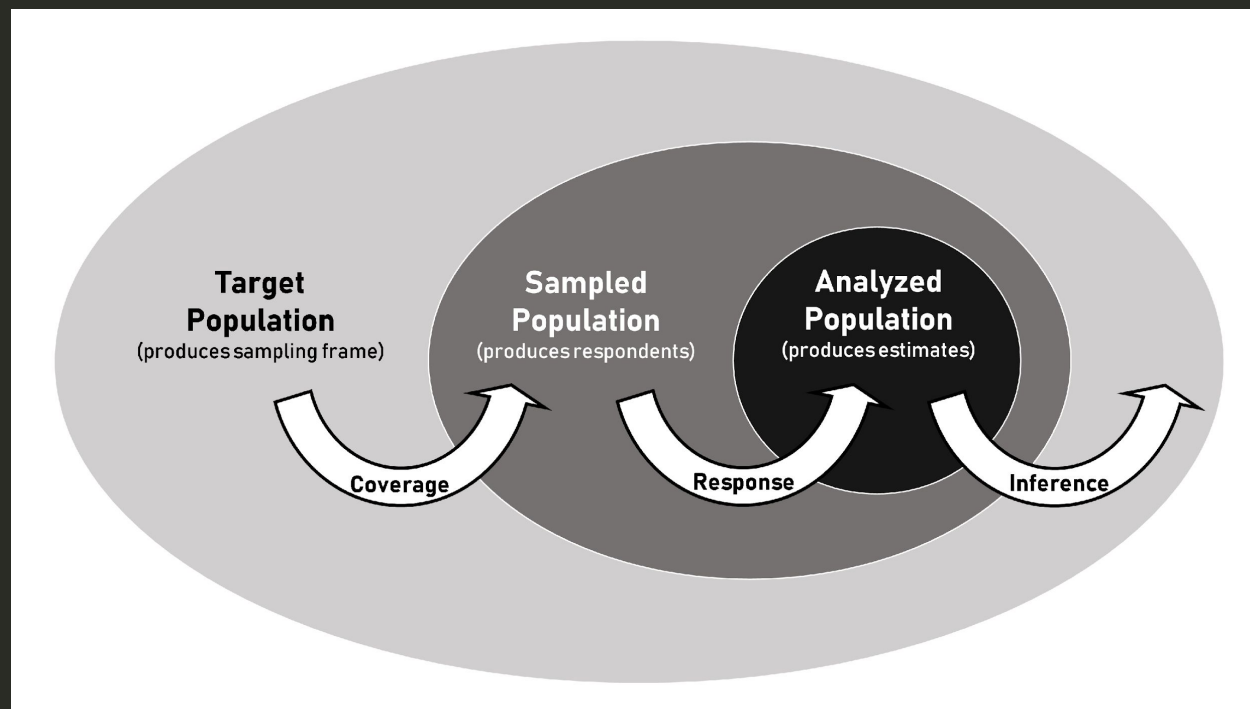


2.3 Amostragem: Etapas 3

Tipos de Amostragem

Amostra e População

4. Amostragem: passos do processo



1.

Identificação e
definição da
População Alvo
(Target
Population)

2.

Seleção da
“Sampling
Frame”

3.

Escolha do
Método de
Amostragem

4.

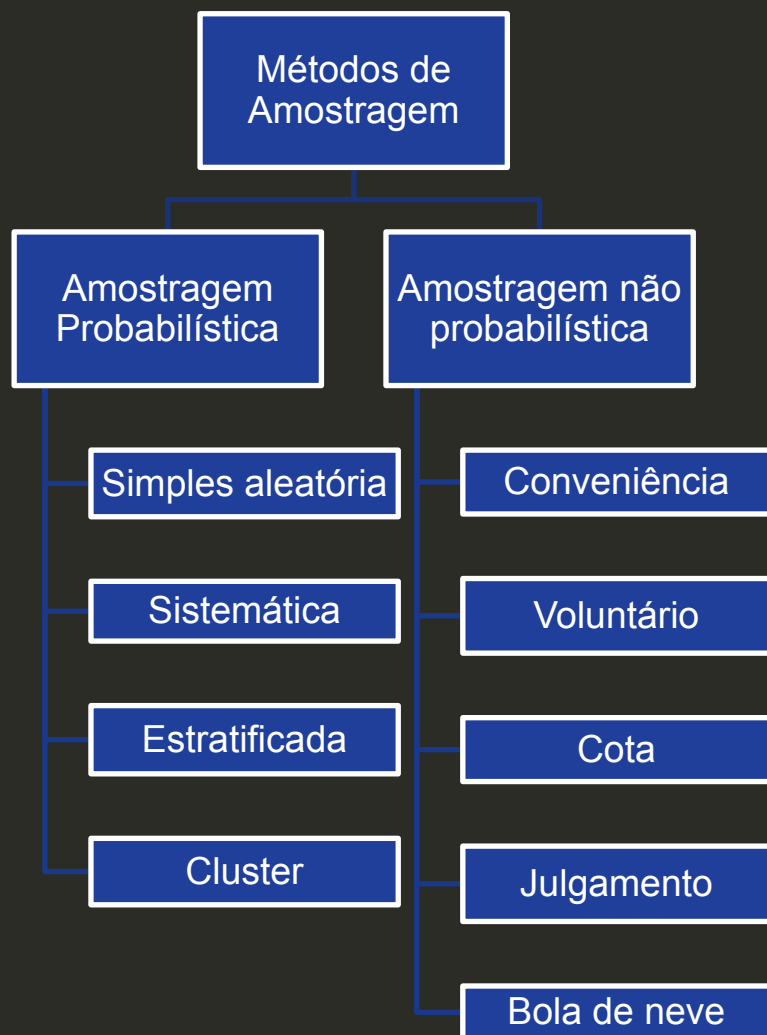
Determinar o
tamanho da
Amostra

5.

Coleta do
dado
necessário

Amostra e População

5. Amostragem: Tipos de amostragem



Amostragem Probabilística

Cada elemento ou item da “*sampling frame*” tem probabilidade definida e não nula de ser incluído na amostra. É o método de amostragem mais indicado para criar amostras mais representativas da população.

Amostragem Não Probabilística

Os elementos ou itens da “*sampling frame*” não tem a **mesma chance de ser selecionado para a amostra**. Consequentemente, existe um risco de gerar uma amostra não representativa que não produz resultados generalizáveis.

Amostra e População

5. Tipos de amostragem Probabilística

Simple aleatória (Random sampling)

Cada indivíduo é escolhido de forma aleatória na população.



Vantagem: forma mais simples e direta de seleção de amostra, reduz viés de seleção

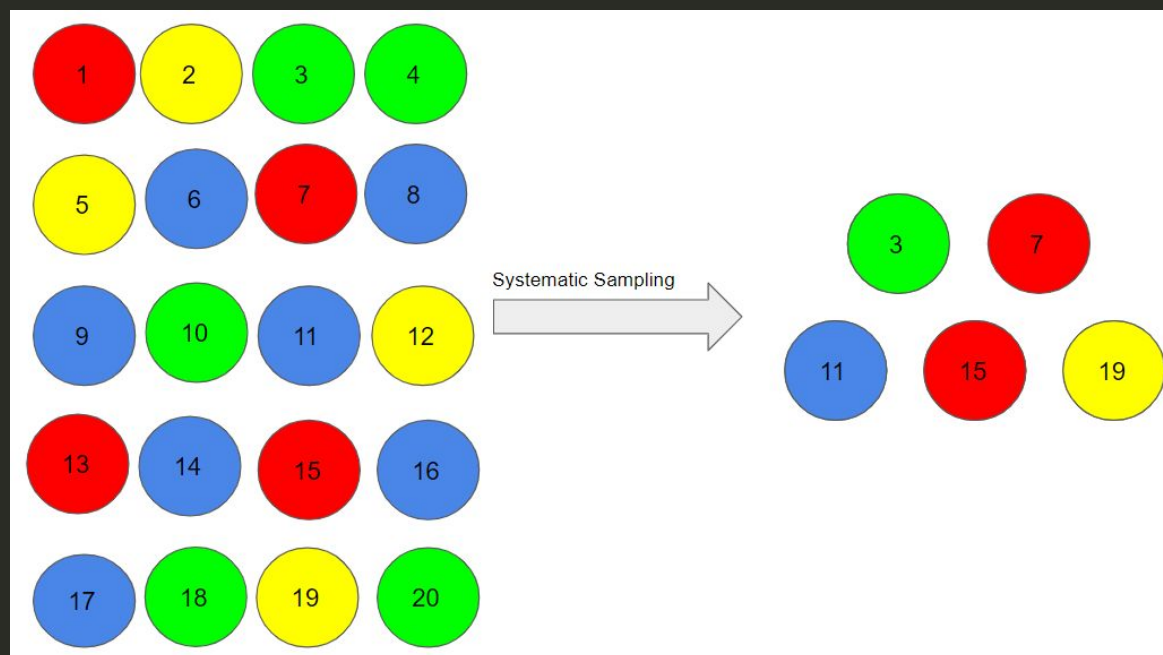
Desvantagem: pode não coletar indivíduos suficientes com características desejáveis, especialmente características incomuns. Trabalhoso se cobertura geográfica da população é alta e formas de coleta de dados variam. Ex.: email, telefone, carta para atingir cobertura.

Amostra e População

5. Tipos de amostragem Probabilística

Sistemática (Systematic)

O primeiro indivíduo é escolhido de forma aleatória na população e demais indivíduos são escolhidos usando um intervalo fixo.



Vantagem: forma mais fácil de organizar do que a amostra aleatória

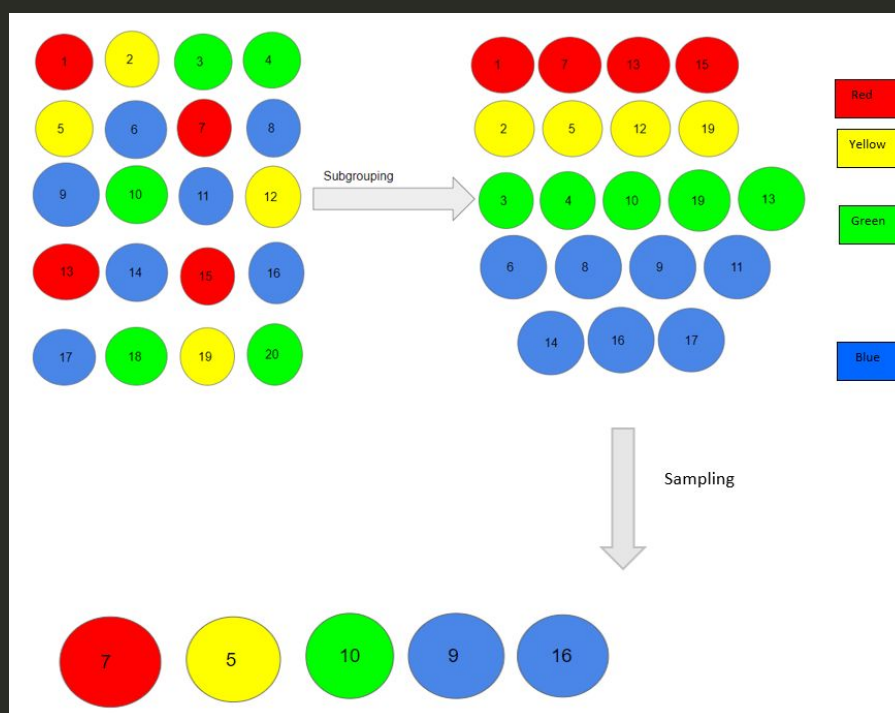
Desvantagem: carregar viés se existe um padrão nos dados. Ex.: média de idade de uma população, seleção de 2 em 2, mas homens e mulheres intercalados na sampling frame.

Amostra e População

5. Tipos de amostragem Probabilística

Estratificada (Stratified)

Divisão da população em subgrupos,
seguida de seleção proporcional



Vantagem: representação de subgrupos. Ex.: análise de renda no Brasil por estado: se aleatória algum estado poderia ficar de fora. Mas, ao mesmo tempo, estados maiores e com maior população são mais representativos.

Desvantagem: requer conhecimento prévio sobre as características da população. Pode ser difícil decidir qual característica deve ser estratificada.

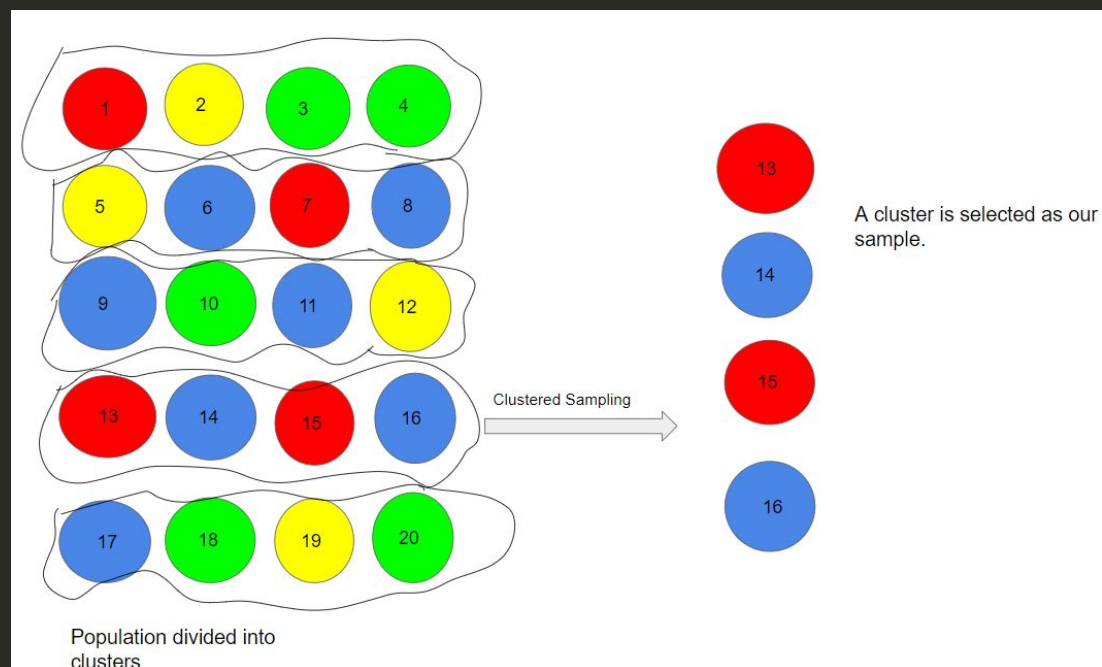
Amostra e População

5. Tipos de amostragem Probabilística

Cluster

Etapa única - Divisão da população em subgrupos (não baseado em característica específica), seguida de seleção aleatória do cluster

Duas etapas – etapa única mais seleção aleatória dentro do cluster



Vantagem foco em área ou região específica.

Clusters geralmente já são unidades que existem: exemplos, cidades, estados, escola.

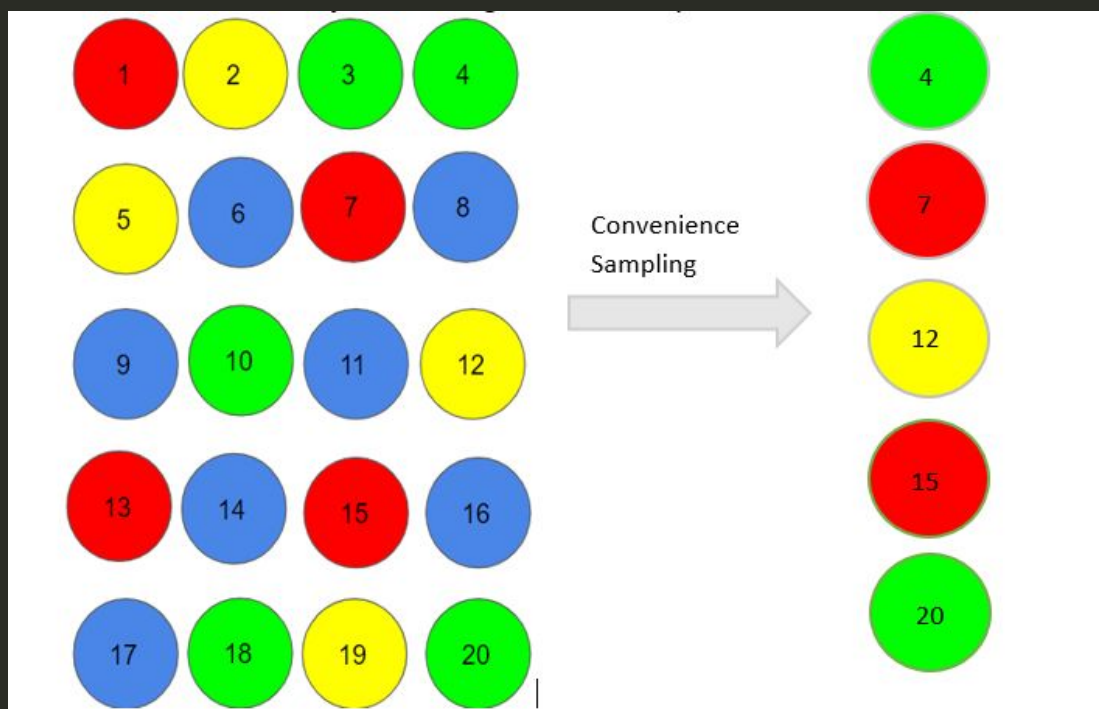
Desvantagem: viés se o cluster não for representativo da população.

Amostra e População

5. Tipos de amostragem Não Probabilística

Conveniência (Convenience)

indivíduos que estão disponíveis são incluídos na amostra



Vantagem: indivíduos com disponibilidade, fácil coleta. Ex. pessoas que são paradas nas ruas, as 50 primeiras pessoas que chegarem, pacientes que atendem requisitos de pesquisas clínicas

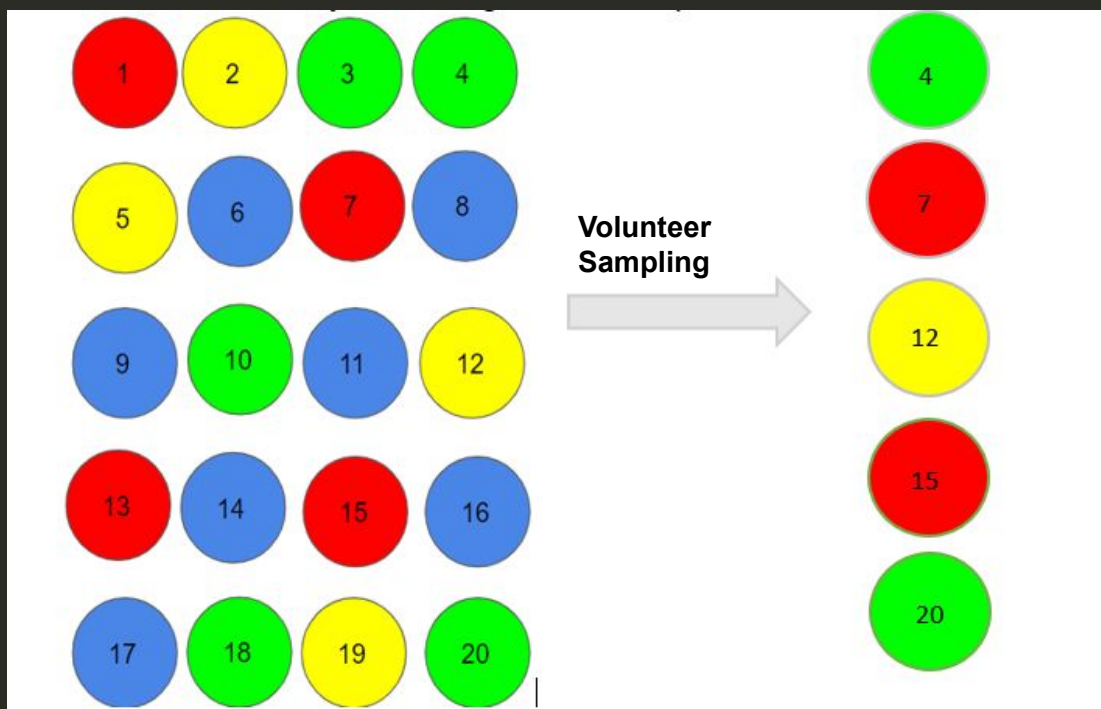
Desvantagem: risco de viés significativo na seleção – respondentes, horário, local, e pouca representatividade da população

Amostra e População

5. Tipos de amostragem Não Probabilística

Voluntário (Volunteer)

indivíduos que estão disponíveis e querem participar são incluídos na amostra



Vantagem: indivíduos com disponibilidade e dispostos, fácil coleta. Ex.: Pedido de resposta em formulários para TCC – responde quem quer.

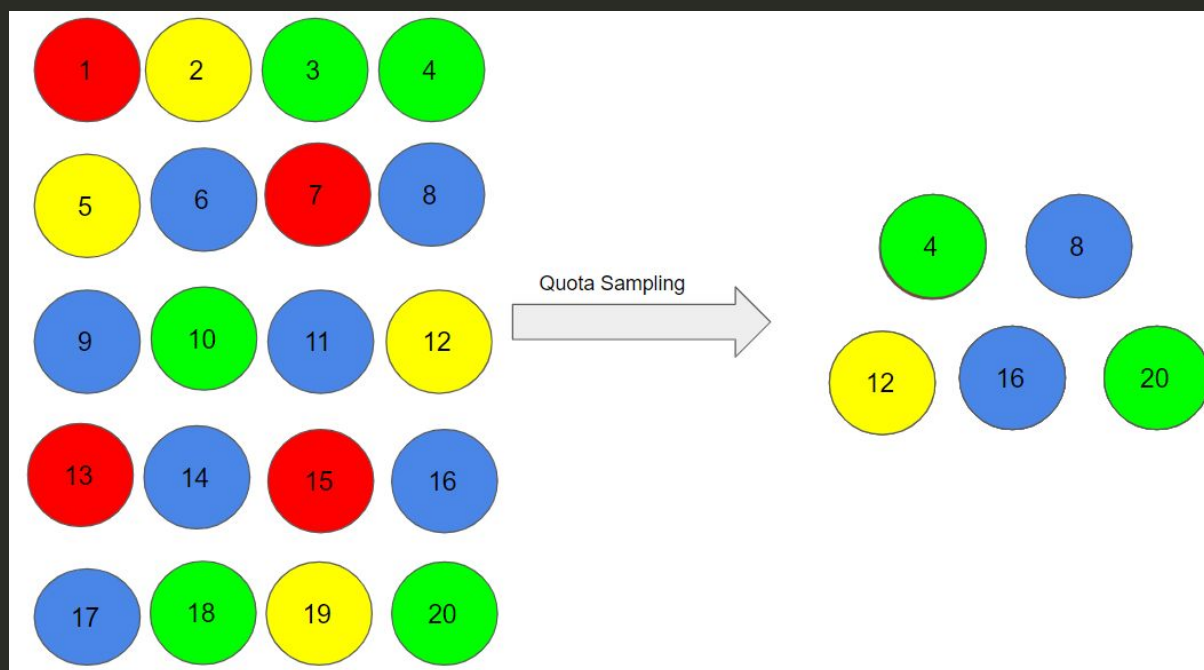
Desvantagem: risco de viés significativo, pois, quem decide não responder pode ser bem diferente de quem decide responder. Talvez só as pessoas que se identificam com o tema? Temas mais polêmicos como religião, ou aspectos de gêneros não representados.

Amostra e População

5. Tipos de amostragem Não Probabilística

Cota (Quota)

indivíduos são selecionados de acordo com uma cota e característica predeterminada. Exemplo: 100% de múltiplos de 4.



Vantagem: fácil coleta e potencialmente mais representativo do que outras abordagens não probabilísticas. Ex.: selecionar 20 homens, 20 mulheres, 10 crianças, 12 idosos. Idealmente representando as proporções populacionais.

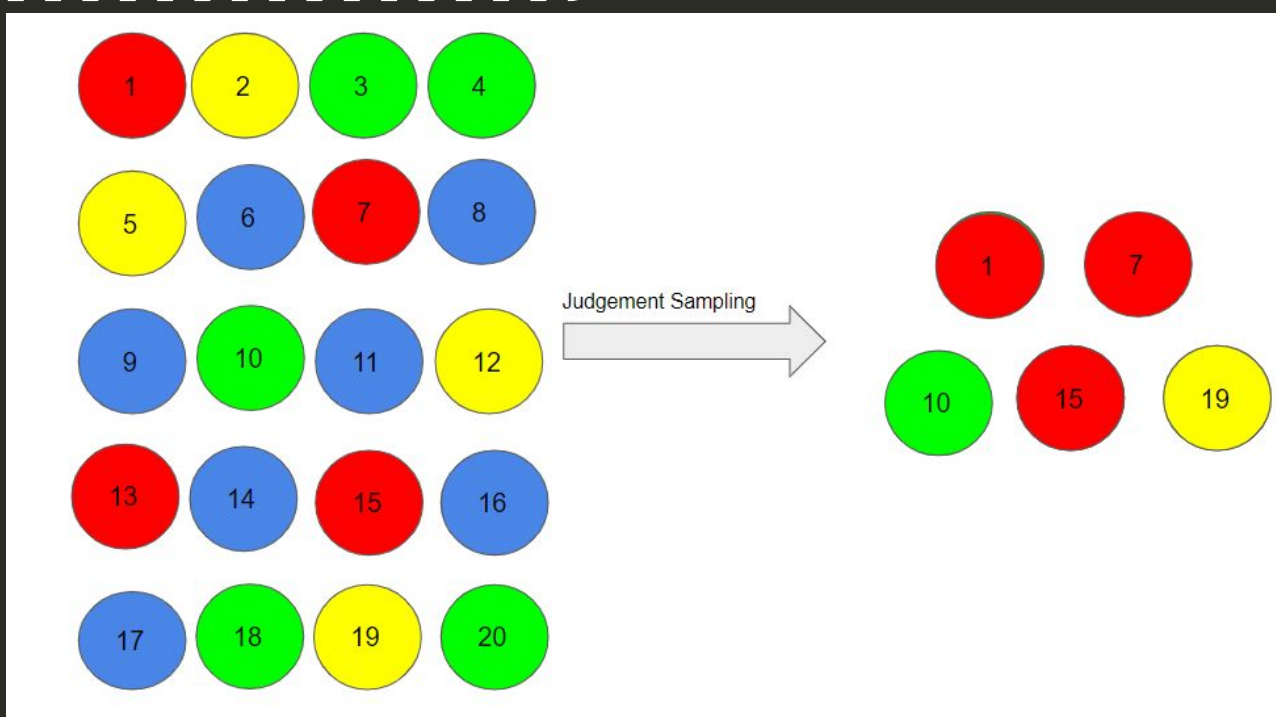
Desvantagem: pouca representatividade comparado a métodos probabilísticos.

Amostra e População

5. Tipos de amostragem Não Probabilística

Julgamento ou seletiva (Judgement)

pesquisadores conduzindo as pesquisas ou experts
selecionam quem deveria participar



Vantagem: economiza tempo e custos, pode ser representativo se o conhecimento do juiz sobre a população estiver correto.

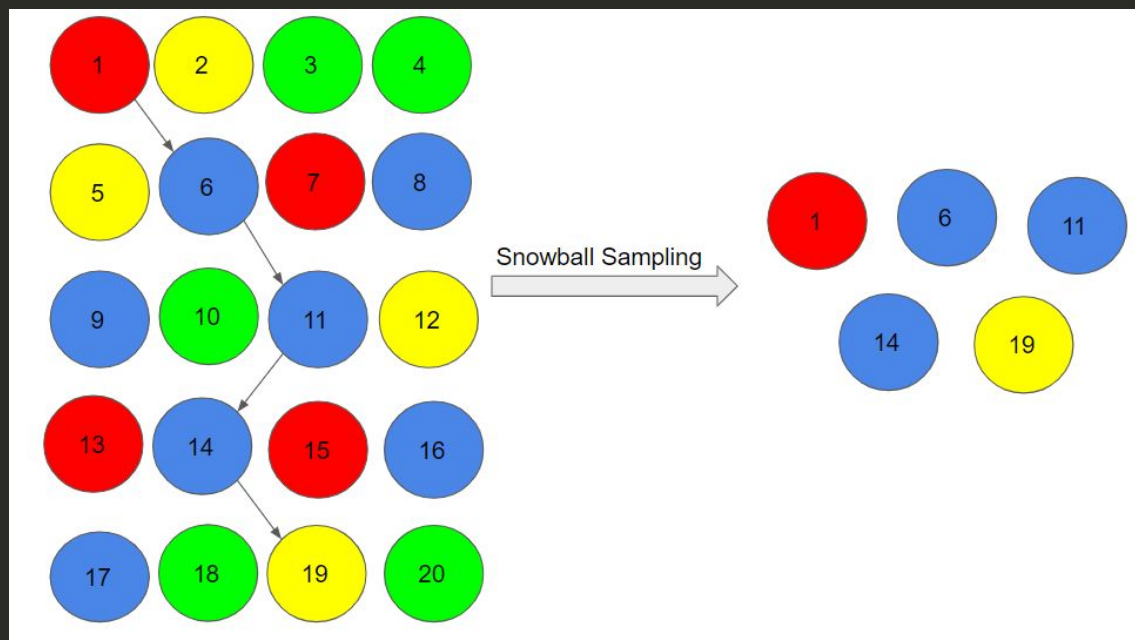
Desvantagem: sujeito a pouca representatividade e erros do juiz.

Amostra e População

5. Tipos de amostragem Não Probabilística

Bola de Neve (Snowball)

indivíduos selecionados indicam outros indivíduos para a amostra.



Vantagem: efetivo quando a população alvo é difícil de ser identificada/contatada. Ex.: população usuária de drogas e entorpecentes, indicações em processos seletivos.

Desvantagem: viés de seleção já que tendem a indicar conhecidos e amigos, baixa representatividade. Sampling frame não disponível.

Recap - Amostra e População

AMOSTRAGEM:

é um processo de **seleção de um subconjunto** da população de interesse que **gera a amostra**

AMOSTRAGEM PROBABILÍSTICA:

existe probabilidade associada à seleção

ALEATÓRIA
SIMPLES

SISTEMÁTICA

ESTRATIFICADA

CLUSTER

AMOSTRAGEM NÃO

PROBABILÍSTICA: não existe probabilidade associada à seleção

CONVENIÊNCIA

VOLUNTÁRIO

QUOTA

JULGAMENTO

BOLA DE NEVE

dnc>class

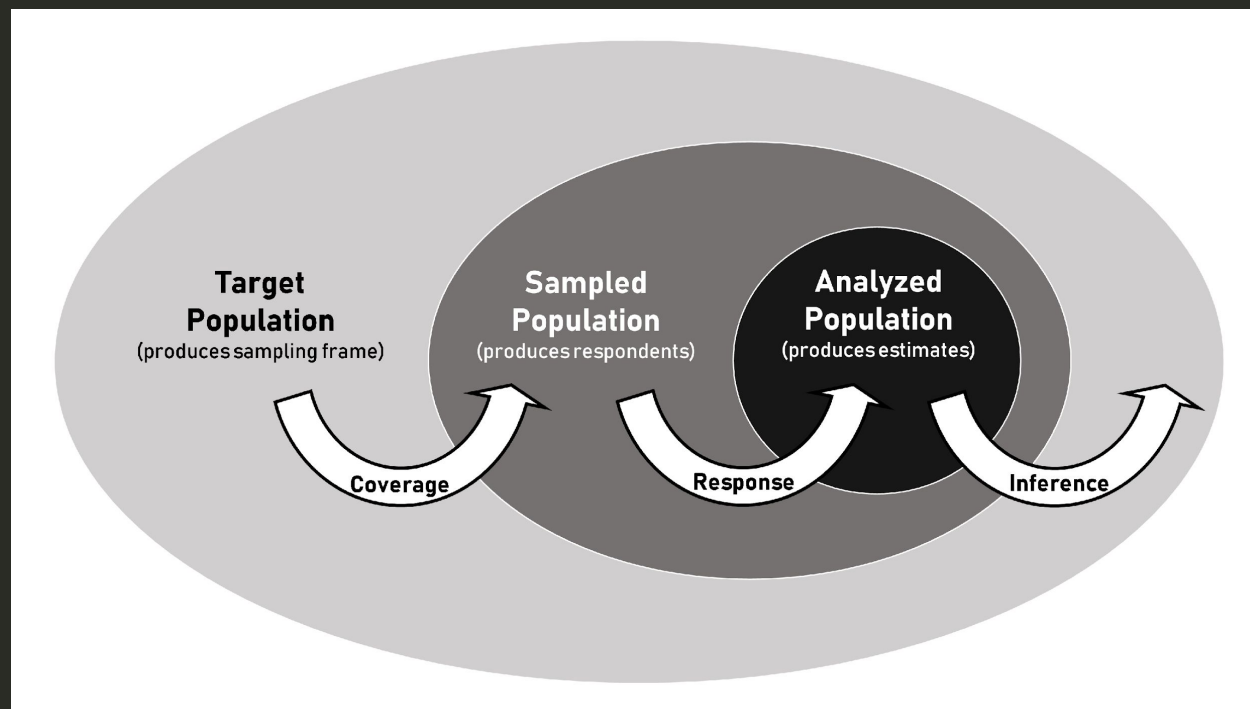


2.3 Amostragem: Etapas 4 e 5

Tamanho de Amostra

Amostra e População

4. Amostragem: passos do processo



1.

Identificação e
definição da
População Alvo
(Target
Population)

2.

Seleção da
“Sampling
Frame”

3.

Escolha do
Método de
Amostragem

4.

Determinar o
tamanho da
Amostra

5.

Coleta do
dado
necessário

Amostra e População

6. Tamanho de amostra

4.

Determinar o tamanho da Amostra

- Decisões sobre acurácia dos seus resultados

1) **Margem de Erro:** quanto de erro é aceitável?

Ex. erro de 4 e resultado de que a % das pessoas que têm computador em casa é de 30%. A porcentagem real deverá estar entre 26% (30 – 4) e 34% (30+4).

2) **Nível de Confiança:** quanto de certeza é desejável ter? Níveis mais comuns são 90%, 95% e 99%.

População grande com tamanho desconhecido

$$n = \frac{z^2}{4m^2}$$

População com tamanho conhecido

$$n = \frac{(pz^2)}{4m^2(p - 1) + z^2}$$

- Sample Size Calculator e - pwr() em R

“n” é o tamanho da amostra recomendado,
 “p” é o tamanho da população,
 “m” é a margem de erro,
 “z” é o z-score.

Amostra e População

6. Tamanho de amostra

4.

Determinar o tamanho da Amostra

	Confidence level	Z-score
1	90%	1.645
2	95%	1.96
3	99%	2.576

$$n = \frac{z^2}{4m^2}$$

Margem de 5% (m)
Nível de Confiança de 95%(z)

$$n = (1.96)^2 / 4(0.05)^2$$

$$n = 384.16$$

$$n = \frac{(pz^2)}{4m^2(p-1) + z^2}$$

Margem de 5% (m)
Nível de Confiança de 95%(z)
População(p) = 200

$$n = 200 \times (1.96)^2 / 4(0.05)^2$$

$$x(200-1) + (0.05)^2$$

$$n = 131.75$$

384 é o tamanho de amostra padrão e relativamente seguro para populações grandes de tamanho desconhecido.

**** Para populações homogêneas.** Se há diversidade, aumentar para 400-1000 é indicado.

Amostra e População

6. Tamanho de amostra

4.

Determinar o tamanho da Amostra

	Confidence level	Z-score
1	90%	1.645
2	95%	1.96
3	99%	2.576

$$n = \frac{z^2}{4m^2}$$

Margem de 5% (m)
Nível de Confiança de 95%(z)

$$n = (1.96)^2 / 4(0.05)^2$$

$$n = 384.16$$

$$n = \frac{(pz^2)}{4m^2(p-1) + z^2}$$

Margem de 5% (m)
Nível de Confiança de 95%(z)
População(p) = 200

$$n = 200 \times (1.96)^2 / 4(0.05)^2$$

$$x(200-1) + (0.05)^2$$

$$n = 131.75$$

Determine Sample Size

Confidence Level: ☒ 95% ☐ 99%

Confidence Interval:

Population:

Calculate

Clear

Sample size needed:

384 é o tamanho de amostra padrão e relativamente seguro para populações grandes de tamanho desconhecido.

**** Para populações homogêneas.** Se há diversidade, aumentar para 400-1000 é indicado.

Amostra e População

6. Tamanho de amostra

4.

Determinar o
tamanho da
Amostra

Usar inferência bayesiana quando não tem “n”
suficiente para significância estatística

Amostra e População

6. Tamanho de amostra

5.

Coleta do
dato
necessário

- **Survey design:** perguntas, resultados e indicadores (SMART).
- **Métodos de coleta de dados:** observação, questionário, entrevista, discussão em grupo.
- **Tipos de perguntas:** dissertativas, binárias, múltipla escolha, numérica.
- **Escrita:** objetividade, uma pergunta por vez, perguntas negativas confundem.

Recap - Amostra e População

AMOSTRAGEM:

é um processo de **seleção de um subconjunto** da população de interesse que **gera a amostra**

ETAPAS DE AMOSTRAGEM

1.
Identificação
e definição da
População
Alvo

2.
Seleção da
“Sampling
Frame”

3.
Escolha
do Método
de
Amostragem

4.
Determinar
o tamanho
da
Amostra

5.
Coleta
do dado
necessário

ETAPA 4: MARGEM DE ERRO E
NÍVEL DE CONFIANÇA

QUALIDADE DE COLETA DE
DADOS

dnc>class



2.4 Amostragem e Ciência de Dados na Vida Real

Amostra e População

7. Exemplos de Amostragem na vida real de ciência de dados

Volume de dados muito grande

Alto volume de dados para ser organizado

Processamento computacional custoso (tempo e investimento)

Ex. Dados de Censo, redes sociais, navegação de e-commerce ou transações bancárias

Aquisição de dados custosa (adquirir apenas o necessário)

Dados populacionais têm um custo de aquisição

Coleta dos itens da população fragmentada e trabalhosa

Ex. Dados de matrículas de apartamentos

Amostra e População

7. Exemplos de Amostragem na vida real de ciência de dados

Volume de dados muito grande

Alto volume de dados para ser organizado

Processamento computacional custoso (tempo e investimento)

Ex. Dados de Censo, redes sociais, navegação de e-commerce ou transações bancárias

Aquisição de dados custosa (adquirir apenas o necessário)

Dados populacionais têm um custo de aquisição

Coleta dos itens da população fragmentada e trabalhosa

Ex. Dados de matrículas de apartamentos

Alguns métodos que aplicam amostragem

Separação de treino, teste e validação (train test split) – aleatório ou julgamento

Bootstrap (aleatório com reposição) – gera medidas de acurácia (bias, variância, intervalo de confiança)- das estimativas das amostras.

Intervalo de confiança de modelos de aprendizado de máquina!

Oversampling (SMOTE - Synthetic Minority Oversampling Technique) – usa amostragem aleatória e uma forma de sistemática.

dnc>class



Intro Estatística Inferencial

Parte II

O que é Estatística Inferencial

1. Estatística inferencial

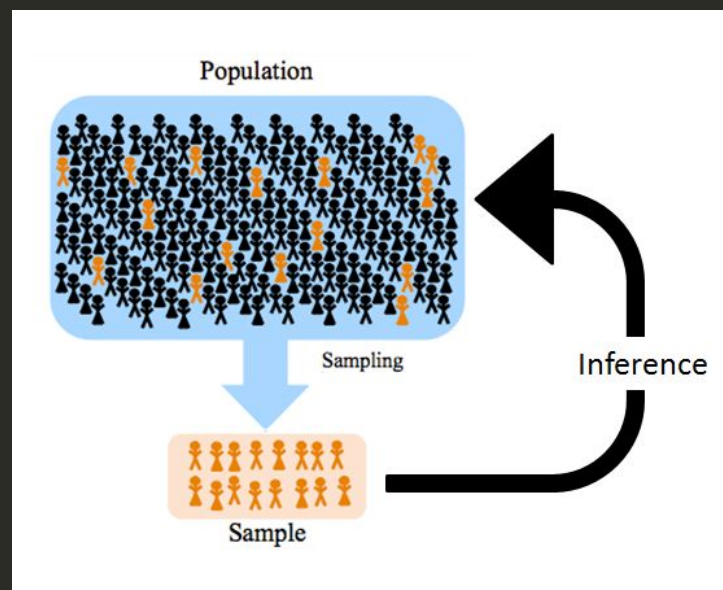
Estatística Inferencial

estimar / criar **inferências** e fazer generalizações sobre características de uma **população** baseadas nos dados de **amostra**

Como?

- Estimar parâmetros
- Testar hipóteses

Base teórica em...



O que é Estatística Inferencial

1. Estatística inferencial

Estatística Inferencial

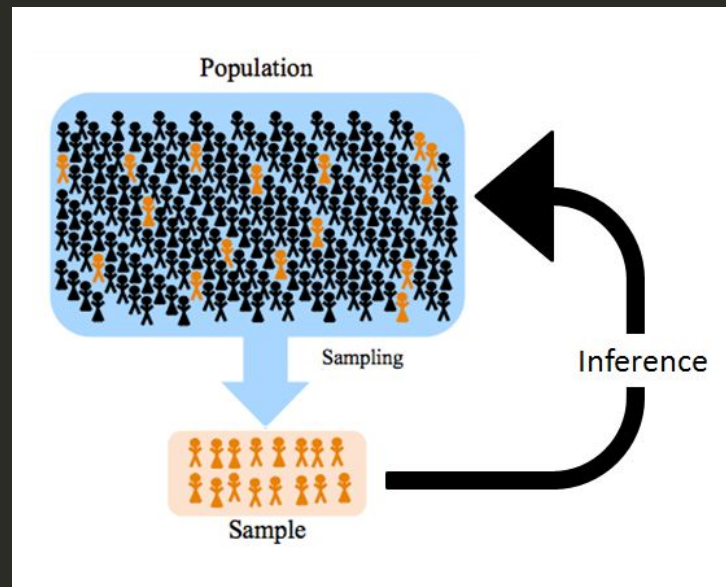
estimar / criar **inferências** e fazer generalizações sobre características de uma **população** baseadas nos dados de **amostra**

Como?

- Estimar parâmetros
- Testar hipóteses

Base teórica em...

Probabilidade!

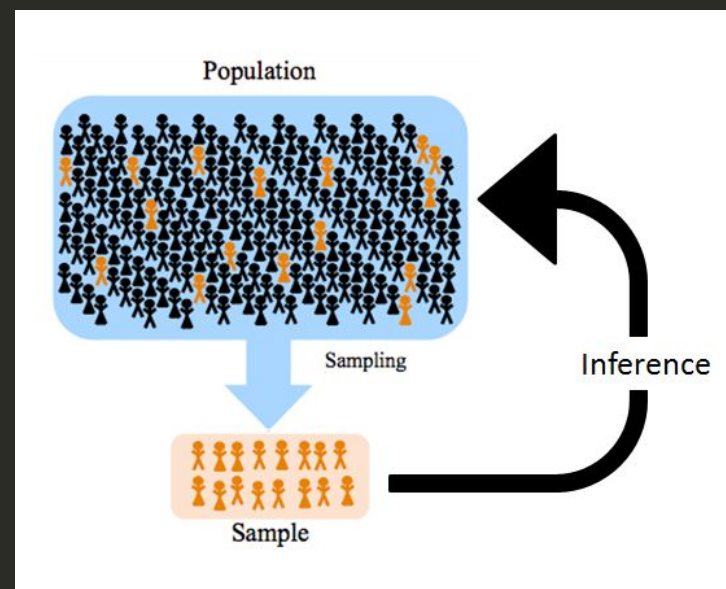


O que é Estatística Inferencial

1. Estatística inferencial

Estatística Inferencial

estimar / criar **inferências** e fazer generalizações sobre características de uma **população** baseadas nos dados de **amostra**



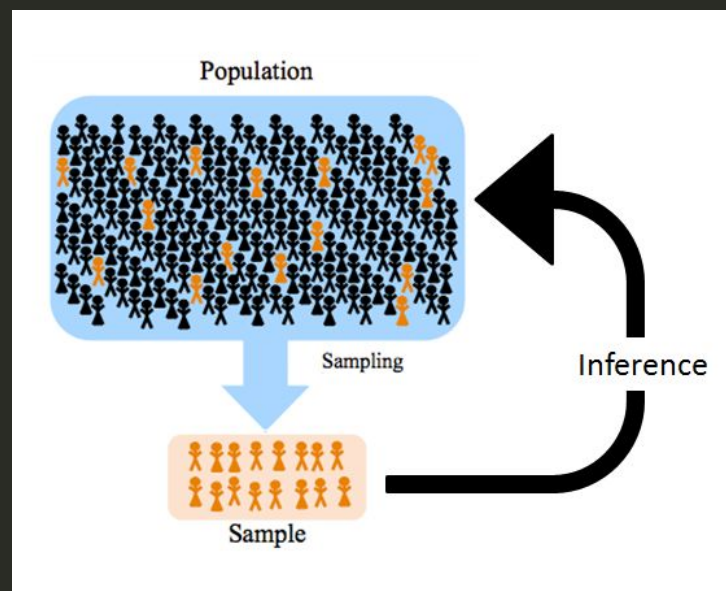
Estimar parâmetros: usar estatística amostral para tirar conclusões sobre parâmetros populacionais.

Testar hipóteses: Testar hipóteses: 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)

O que é Estatística Inferencial

Estimar Parâmetros Populacionais

Estimar parâmetros populacionais, isso é, média, mediana, variância. Após cálculo da estatística amostral, determinar o parâmetro populacional como um valor (“point estimate”) ou como um intervalo, entre x e y (“interval estimate”).

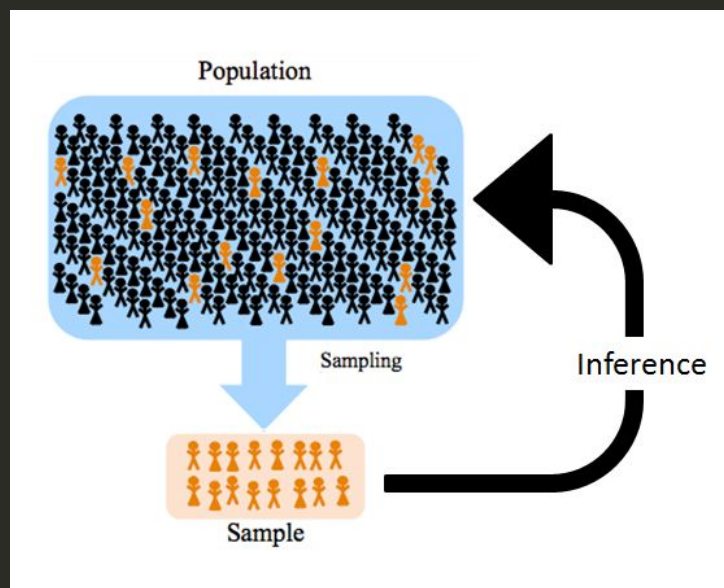


O que é Estatística Inferencial

Teste de Hipótese

Investigar uma conjectura sobre parâmetros populacionais, isso é, média, mediana, variância. Após cálculo da estatística amostral, determinar se a conjectura sobre o parâmetro populacional é refutada ou confirmada.

****** Hipótese deve ser definida antes da análise.



O que é Estatística Inferencial

Estimação vs. Teste de Hipótese – exemplo prático

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?

O que é Estatística Inferencial

2. Exemplo de uso de estatística inferencial:

Horas de estudos por pessoa da Turma de Data Expert da DNC: **80% das pessoas estuda 10 hrs ou mais na semana**

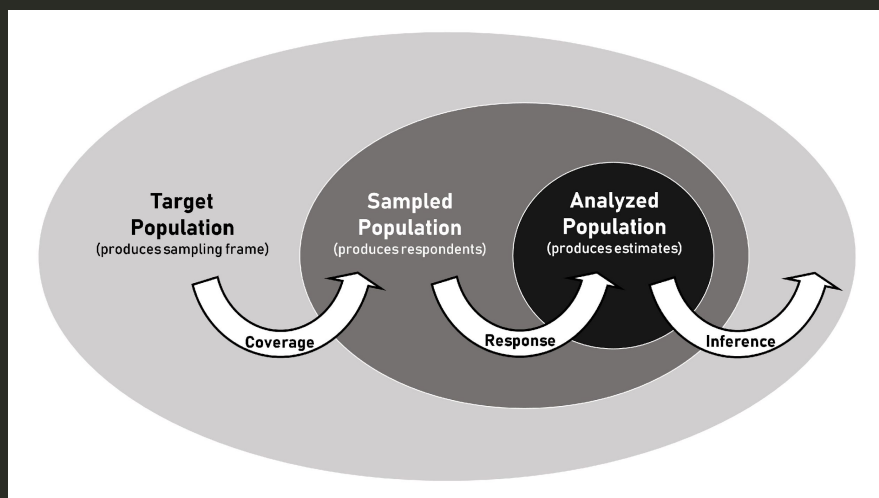
Posso afirmar que 80% das pessoas alunas da DNC estudam 10hrs ou mais? Não!

Amostragem: pessoas de diferentes cursos e turmas – **média de horas de estudos** é uma **estatística**

Essas pessoas representam a população? Qual a probabilidade dessa hipótese (80% das pessoas estudam 10 hrs ou mais) estar certa?

O que é Estatística Inferencial

2. Exemplo de uso de estatística inferencial:



População Alvo: pessoas que estão fazendo cursos da DNC

População amostrada: pessoas que estão fazendo o curso de Data Expert

População Analisada: pessoas que estão fazendo o curso Data Expert e responderam ao questionário sobre horas de estudos

Recap – Intro parte 2

BASE TEÓRICA EM PROBABILIDADE

ESTIMAR PARÂMETROS

TESTES DE HIPÓTESE

dnc>class



Abordagem Frequentista e Bayesiana

Abordagens de (Inferência) Estatística

Frequentista

Bayesiana

**Diferentes
premissas**

**Diferentes
fundações
teóricas**

**Diferentes
crenças
filosóficas**

Abordagem Frequentista

Século XX

Repetibilidade

Incerteza é derivada de erro amostral
<> Mais dados

Interpretação de probabilidade é a frequência de longo prazo de experimentos repetíveis.

EXEMPLO: probabilidade de uma moeda ser lançada e cair em Cara ser 0.5 significa que repetir o lançamento da moeda várias vezes viríamos Cara em 50% do tempo.

Abordagem Bayesiana

Século XVIII

Teorema de Bayes

Determinismo

Usada muito em
análises de saúde

Prior e Posterior

Atualização de
crenças com base
em dados mais
recentes

Interpretação de probabilidade é uma incerteza relacionada ao desconhecimento humano e menos à incerteza do mundo em si. Probabilidade representa o grau de crença em algo.

Abordagem Bayesiana

Teorema de Bayes

$$P(H|D) = P(D|H) * P(H) / P(D)$$

H -> hipótese que se deseja investigar

D -> dados amostrados para o experimento

$P(H|D)$ -> posterior : probabilidade que se quer calcular, de a hipótese ser verdadeira

$P(D|H)$ -> “likelihood”: probabilidade de selecionar esses dados, dado que a hipótese é verdadeira. (Único termo que importaria na abordagem frequentista)

$P(H)$ -> prior: probabilidade de a hipótese ser verdadeira, antes de se ver os dados. Aqui entra o contexto, experiências anteriores e atualização de crenças.

$P(D)$ -> termo normalizador

Abordagens de (Inferência) Estatística

Frequentista

Falta de contexto

Falsos positivos

**Dependência
exclusiva na
amostra**

Bayesiana

**Prior seria um
viés?**

**Como estimar o
prior?**

DID THE SUN JUST EXPLODE? (IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.

DETECTOR! HAS THE
SUN GONE NOVA?

ROLL

YES.



FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.



Recap – Freq e Bayesiana

Frequentista

Bayesiana

Repetibilidade

Determinismo

Incerteza é derivada de erro amostral
<> Mais dados

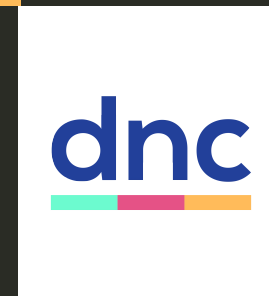
Atualização de crenças
com base em dados
mais recentes

Falta de contexto

Falsos positivos

Prior seria um
viés?

dnc>class



Probabilidade e Distribuições

Probabilidade

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



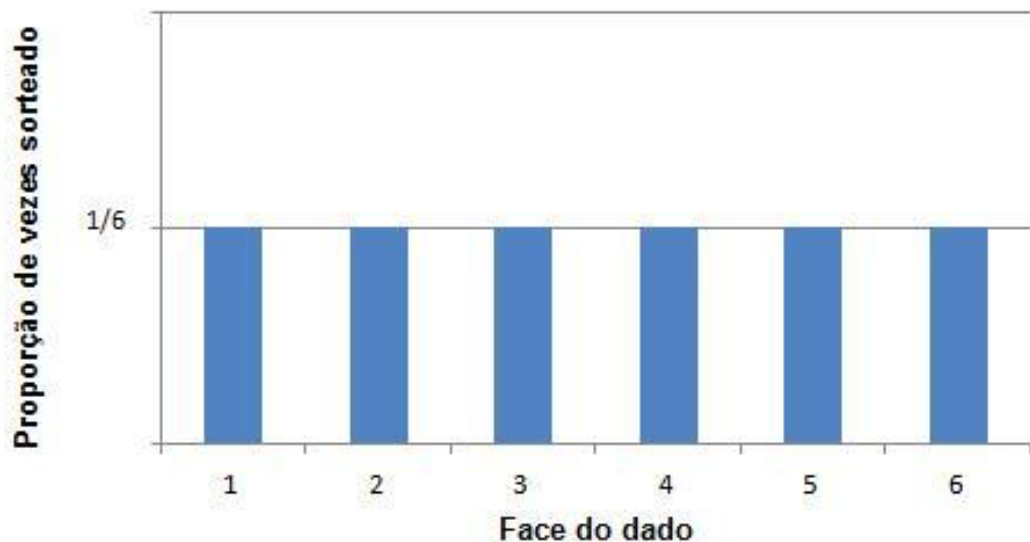
Probabilidade

CONCEITOS



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

Infinitos lançamentos de um dado



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.

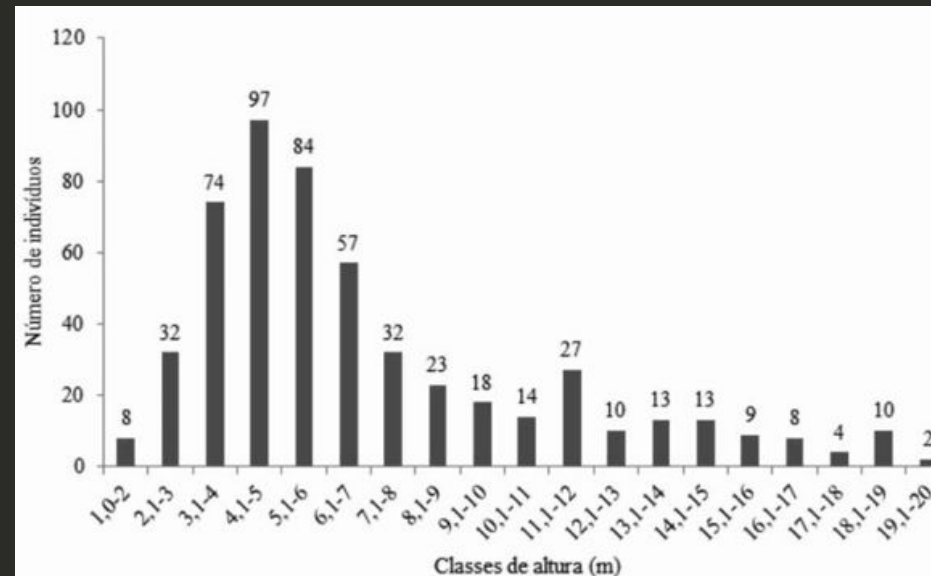
Probabilidade

CONCEITOS

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

Probabilidade

Frequência



Probabilidade de ocorrência de resultados em um experimento aleatório

Contagem de ocorrências dentro de intervalos

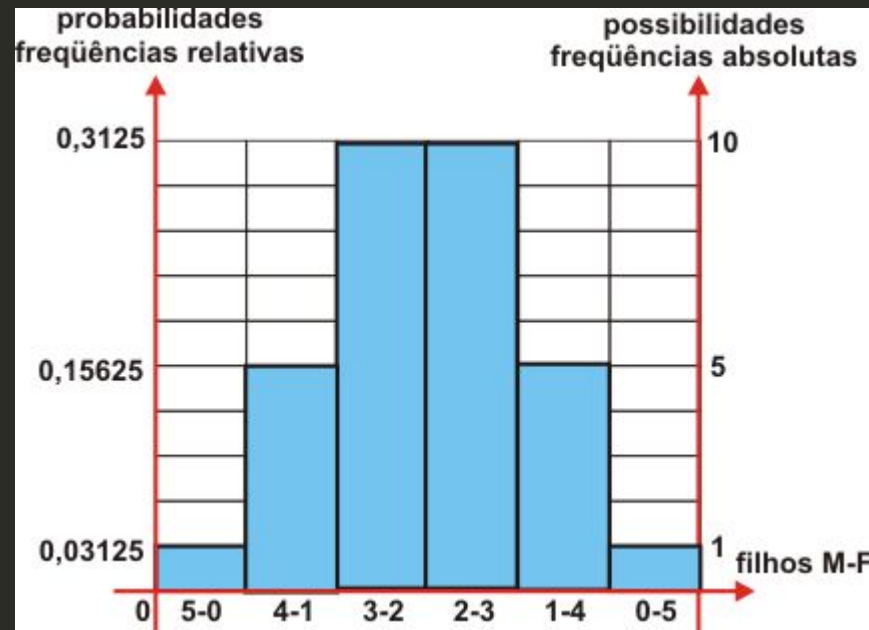
Probabilidade

CONCEITOS

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

Probabilidade

Frequência



Probabilidade de ocorrência de resultados em um experimento aleatório

Contagem de ocorrências dentro de intervalos

Probabilidade

CONCEITOS

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

COMO ASSIM?

**Variável
algébrica**

$$x + 1 = 5$$

“x” é
desconhecido
mas pode ser
encontrado

$$x = y + 2$$

Um valor foi
atribuído a x

**Variável
aleatória**

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

Probabilidade

CONCEITOS

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

Tipo de variável aleatória	Característica de valores que pode assumir	Exemplos
Discreta	Valores distintos ("separados") ou finitos (contáveis)	<ul style="list-style-type: none">- 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)	<ul style="list-style-type: none">- 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

Probabilidade

Quiz: Variável aleatória Discreta ou Contínua?

Ex. 1

O número de carros que uma empresa consegue fabricar em um dia

DISCRETA

Recap – Intro Probabilidade e Dist

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

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

Probabilidade

Frequência

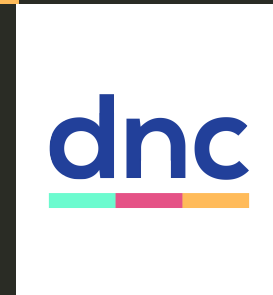
**Variável
algébrica**

**Variável
aleatória**

Discreta

Contínua

dnc>class



Distribuições de Probabilidades

Probabilidade

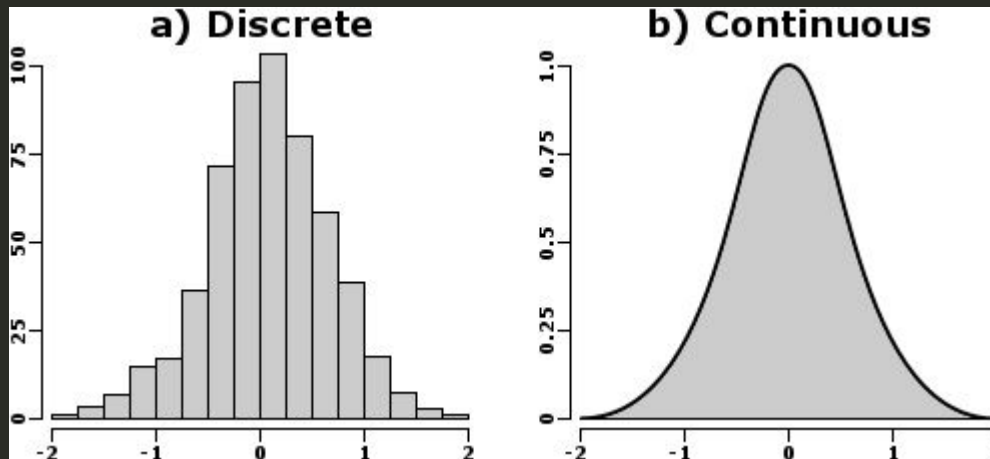
Distribuição de Probabilidades



Distribuição de probabilidades é uma função matemática que **descreve a aleatoriedade de variáveis aleatórias**. É uma representação de todos os **possíveis resultados** de uma variável aleatória **e suas probabilidades associadas**.

Probabilidade

Distribuição de Probabilidades



Distribuição de Probabilidade Discreta:

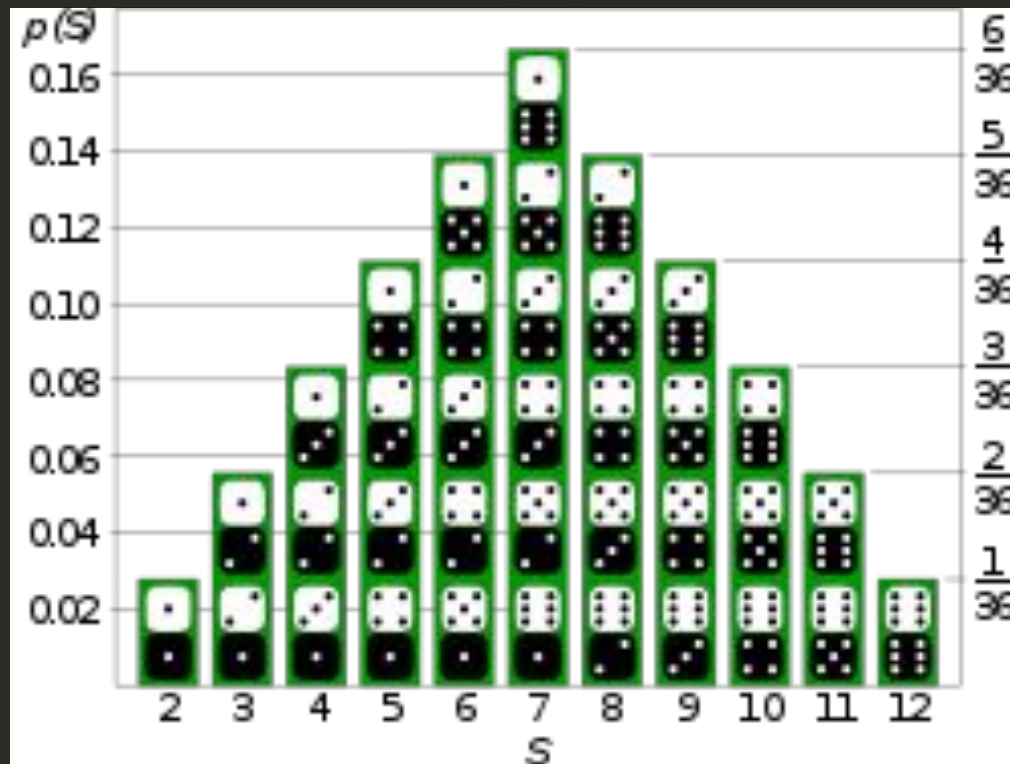
Probability mass function - Pmf
(função massa de probabilidade)

Distribuição de Probabilidade Contínua:

Probability density function - Pdf
(função densidade de probabilidade)

Probabilidade - Pmf

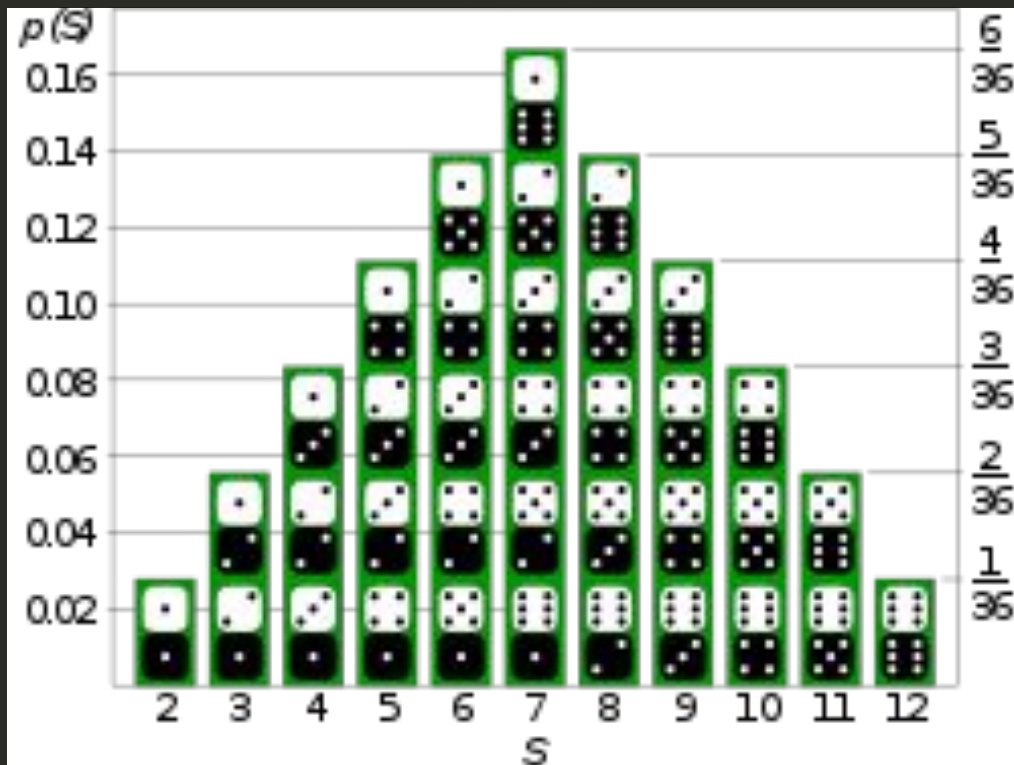
Exemplo: Função massa de probabilidade que especifica a distribuição de probabilidade da soma de dois lançamentos de dados (variável aleatória discreta).



Probabilidade - Pmf

Exemplo:

Função massa de probabilidade que especifica a distribuição de probabilidade da soma de dois lançamentos de dados (variável aleatória discreta).



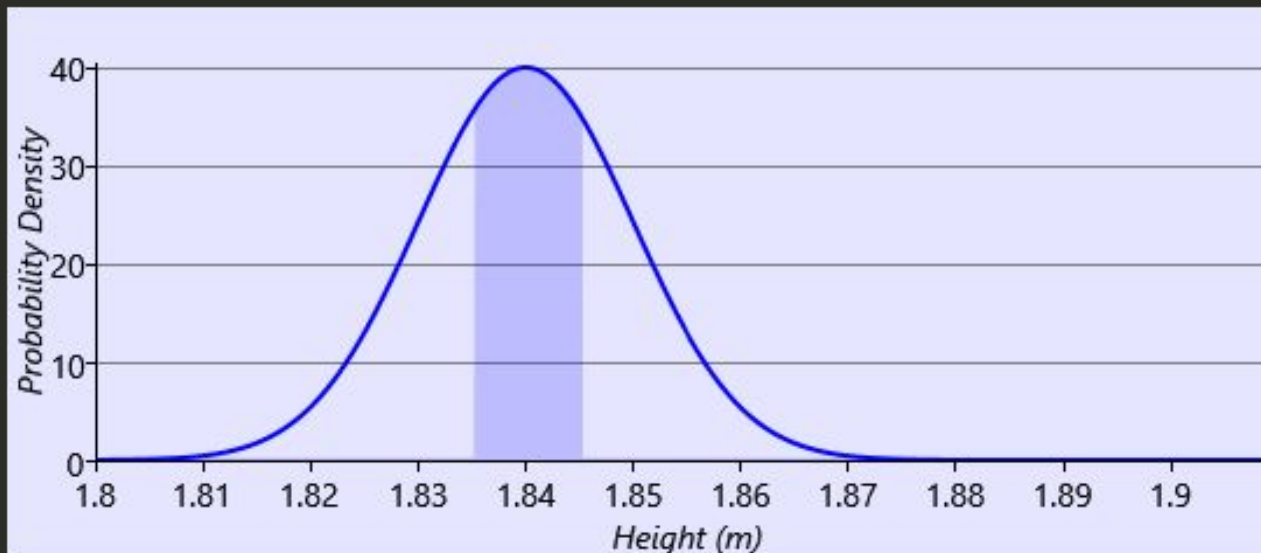
$$P(S=11) = 2/36 = 1/18$$

$$P(S>9) = 3/36 + 2/36 + 1/36 = 1/6$$

Probabilidade - Pdf

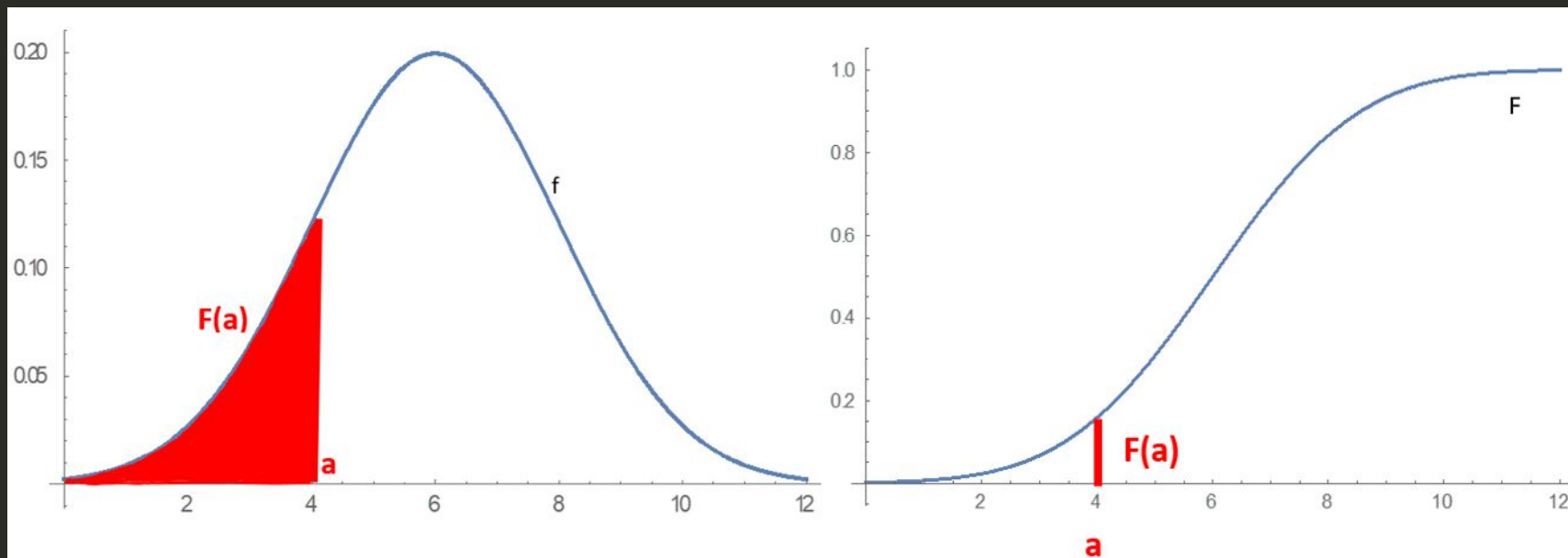
Exemplo:

Função densidade de probabilidade que especifica a **probabilidade infinitesimal (muito pequena) de um valor específico** de altura, e a probabilidade de o valor estar em um intervalo é computada calculando a integral da área no intervalo.

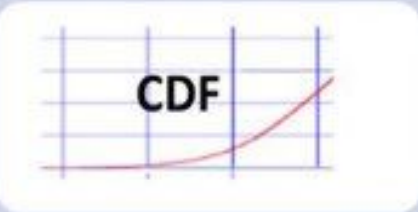

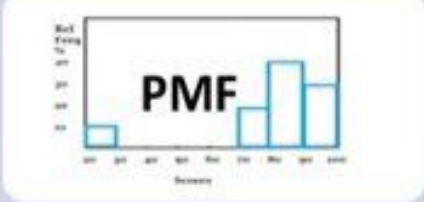


Probabilidade - Cdf

Função distribuição acumulada que especifica a **probabilidade** de a variável aleatória ser menor ou igual a um determinado valor. Em funções densidade de probabilidade (distribuições contínuas) a probabilidade acumulada (função à direita na figura abaixo) é a área abaixo da curva até o determinado valor (na figura abaixo, o valor determinado é 4).



Funções de probabilidade – resumo de características

			
	Cumulative Density Function	Probability Density Function	Probability Mass Function
Purpose	Cumulative probability associated with a function.	Probabilities for continuous random variables .	Probabilities for discrete random variables .
Example	Cumulative value from negative infinity up to a random variable X (i.e. $x < 10$)	Probability of a range of outcomes (e.g. $X = 5$ to 6)	Probability of a certain outcome (e.g. $X = 6$)
Properties	Integral of the PDF. A CDF has [2]: a/ Left limit = 0, right limit = 1 b/ Nondecreasing c/ Right continuous (defined up to a point) [3].	Derivative of the CDF. A PDF satisfies the following [4]: a/ It is positive everywhere b/ $AUC = 1$ c/ Total probability = integral of $f(x)$	Satisfies the following[4]: a/ It is positive everywhere b/ $AUC = 1$ c/ Total probability = summations of individual probabilities.

Recap – Distribuições de Probabilidade

Probability mass function – distribuição discreta

Probability density function – distribuição contínua

Cumulative density function – ambas distribuições

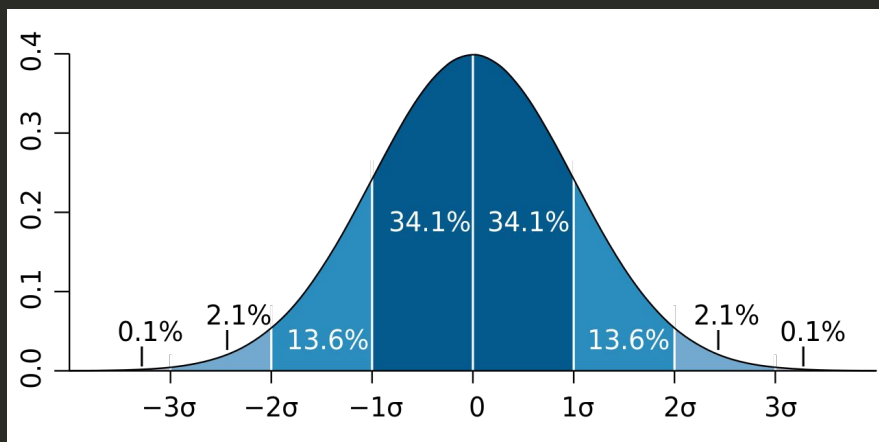
dnc>class



Distribuições Normal

Distribuição Normal

A **distribuição normal** é uma distribuição de probabilidade contínua, uma das distribuições, senão a mais, conhecida e importante. A pdf da distribuição normal também é chamada de “bell curve” ou gaussiana.



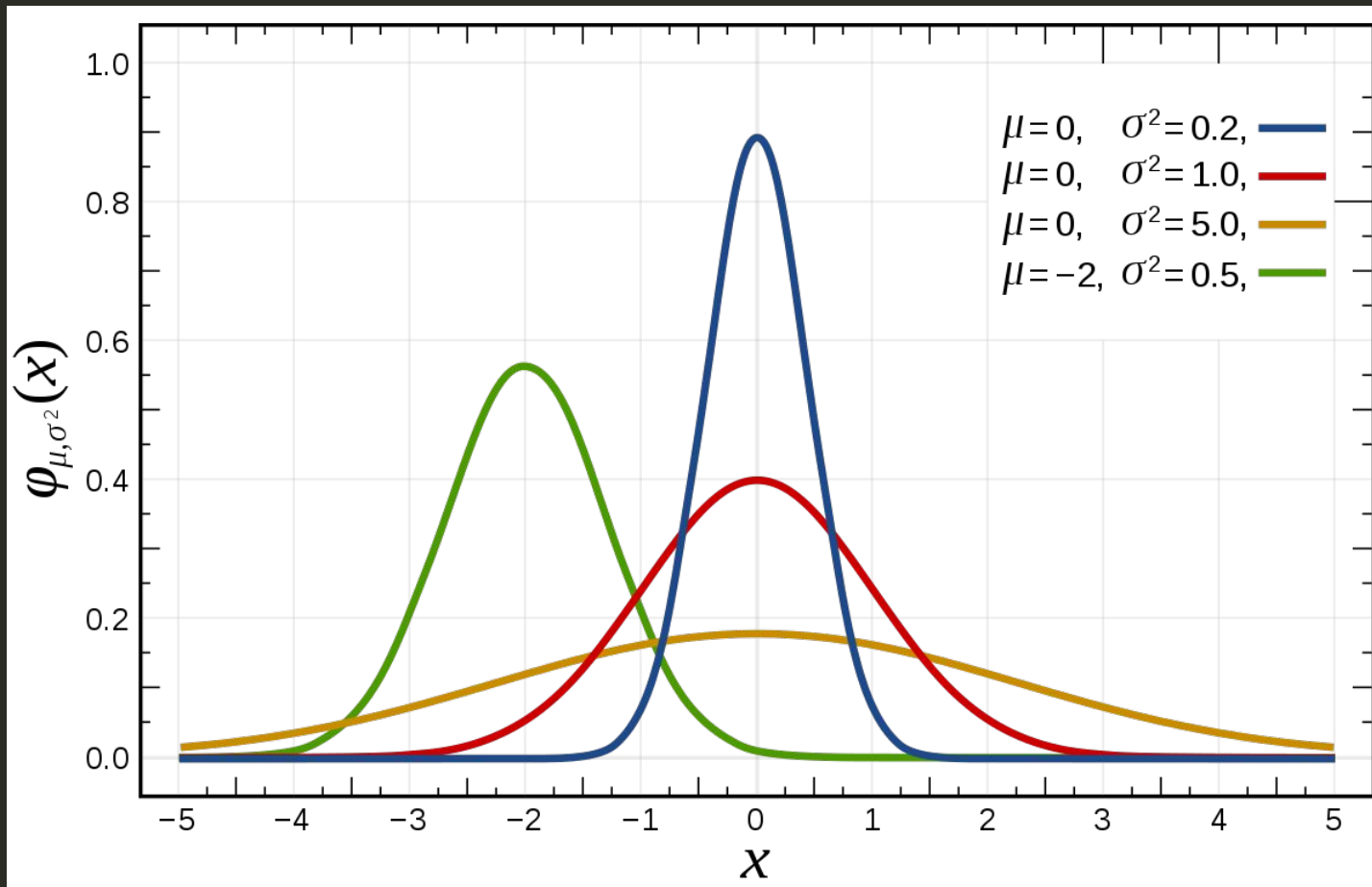
A área abaixo da distribuição representa probabilidades e as posições número de desvios padrões da média.

Dados com distribuição normal, a média, mediana e moda são aproximadamente iguais

No modelo teórico, medidas de tendência central são exatamente iguais e distribuição é simétrica

Distribuição Normal

A **distribuição normal** pode assumir formatos diferentes dependendo de medidas de tendência central e variabilidade. Mas todas possuem área abaixo da curva de 1.



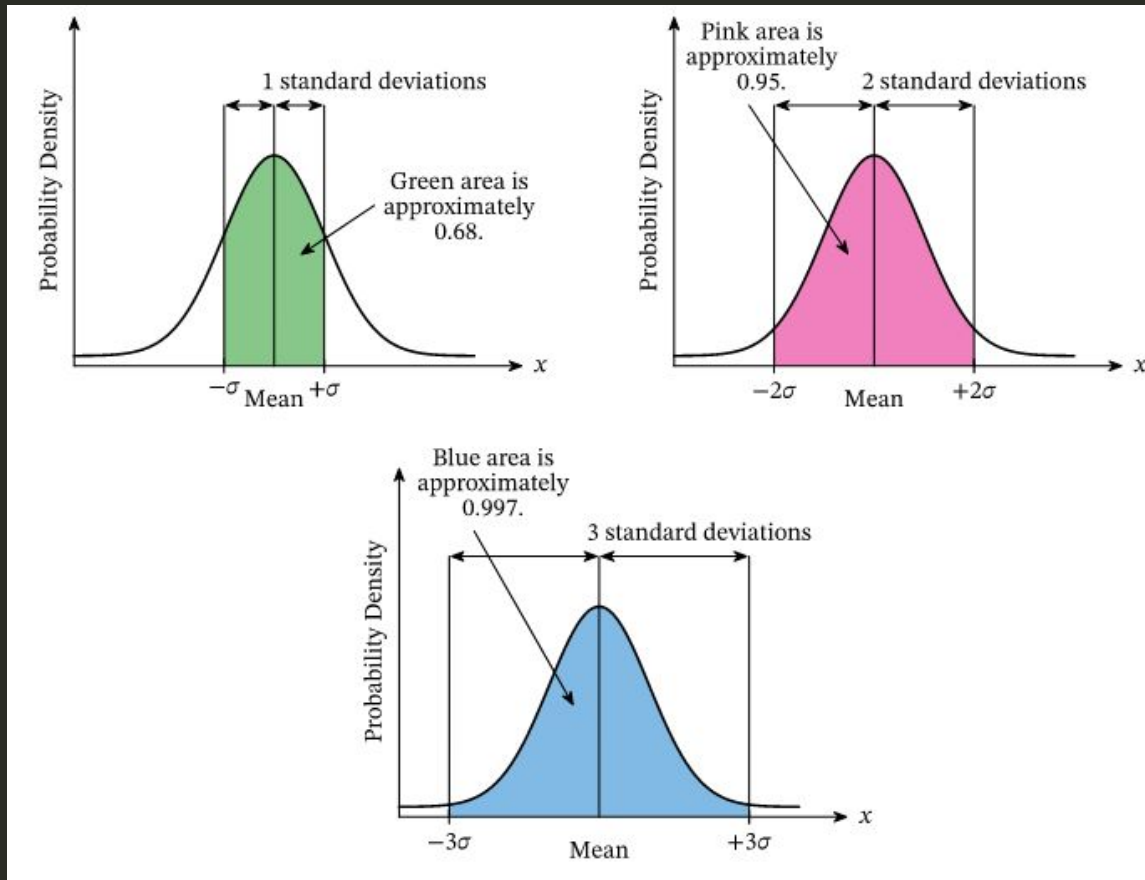
Distribuição Normal

Toda **distribuição normal** segue a “Empirical Rule” (Lei empírica) de que:

68% dos dados caem dentro de 1 desvio padrão da média

95% dos dados caem dentro de 2 desvios padrões da média

99.7% dos dados caem dentro de 3 desvios padrões da média



Distribuição Normal

A **distribuição normal** é uma das distribuições de probabilidades mais significativas da estatística.

Muitos dados contínuos na natureza e psicologia seguem o formato de curva de sino (bell curve)

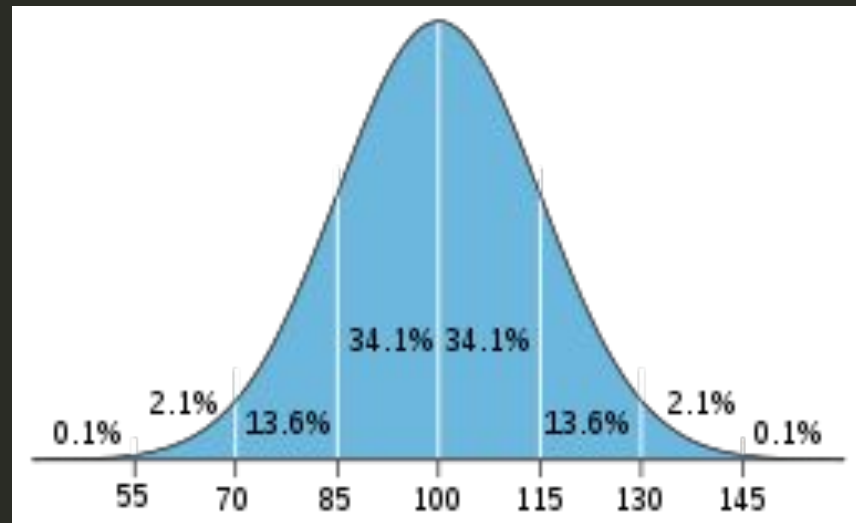
VARIÁVEIS NORMAIS:

Altura

Peso

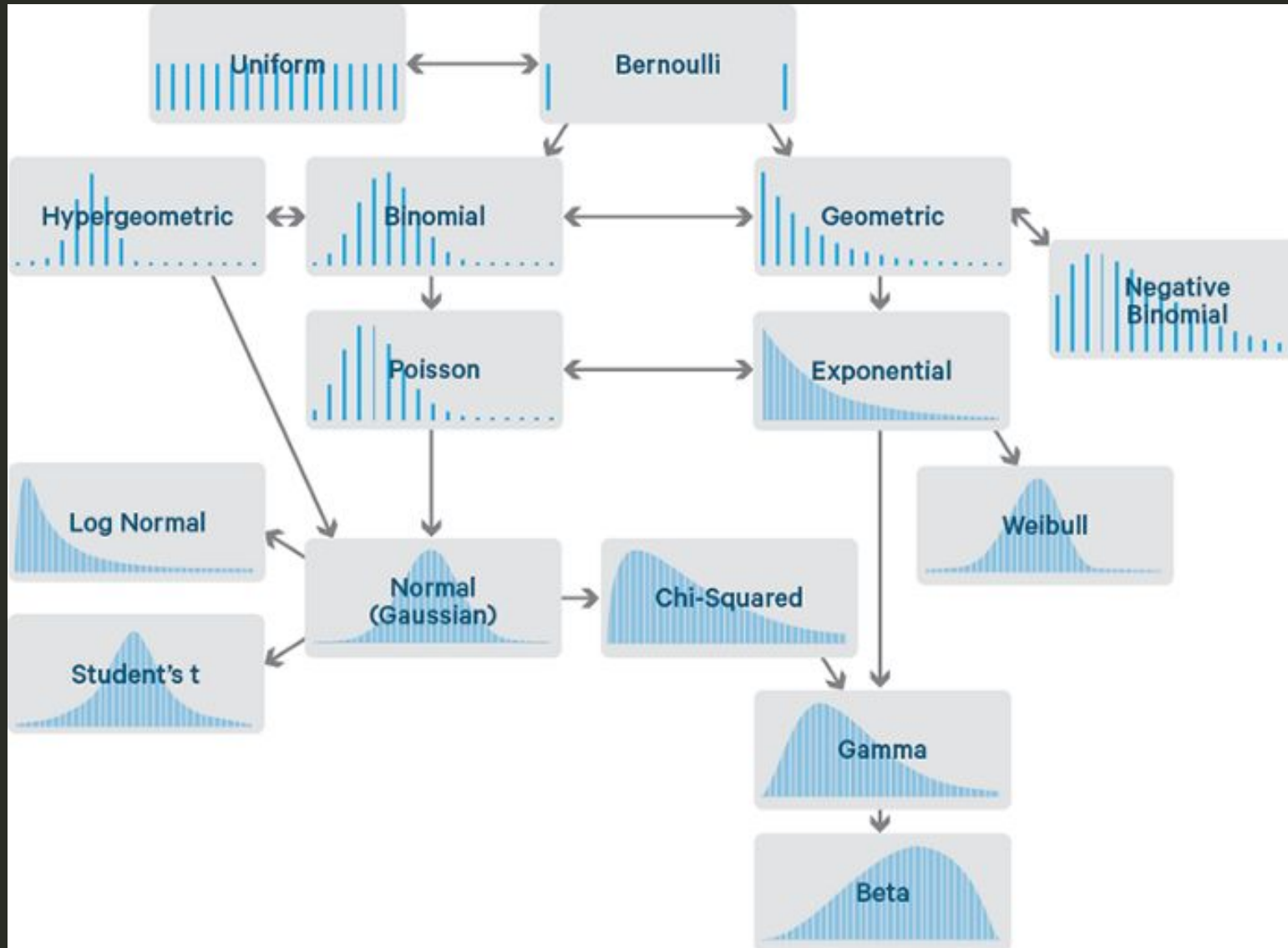
Pressão
sanguínea

QI



Distribuição de QI.

Outras Distribuições



Recap – Distribuições de Probabilidade

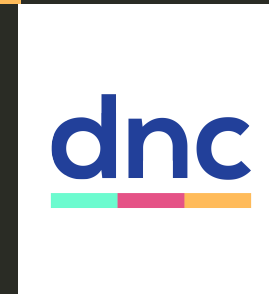
Distribuição normal: simetria e tendência central

Empirical rule

Importância de distribuição normal

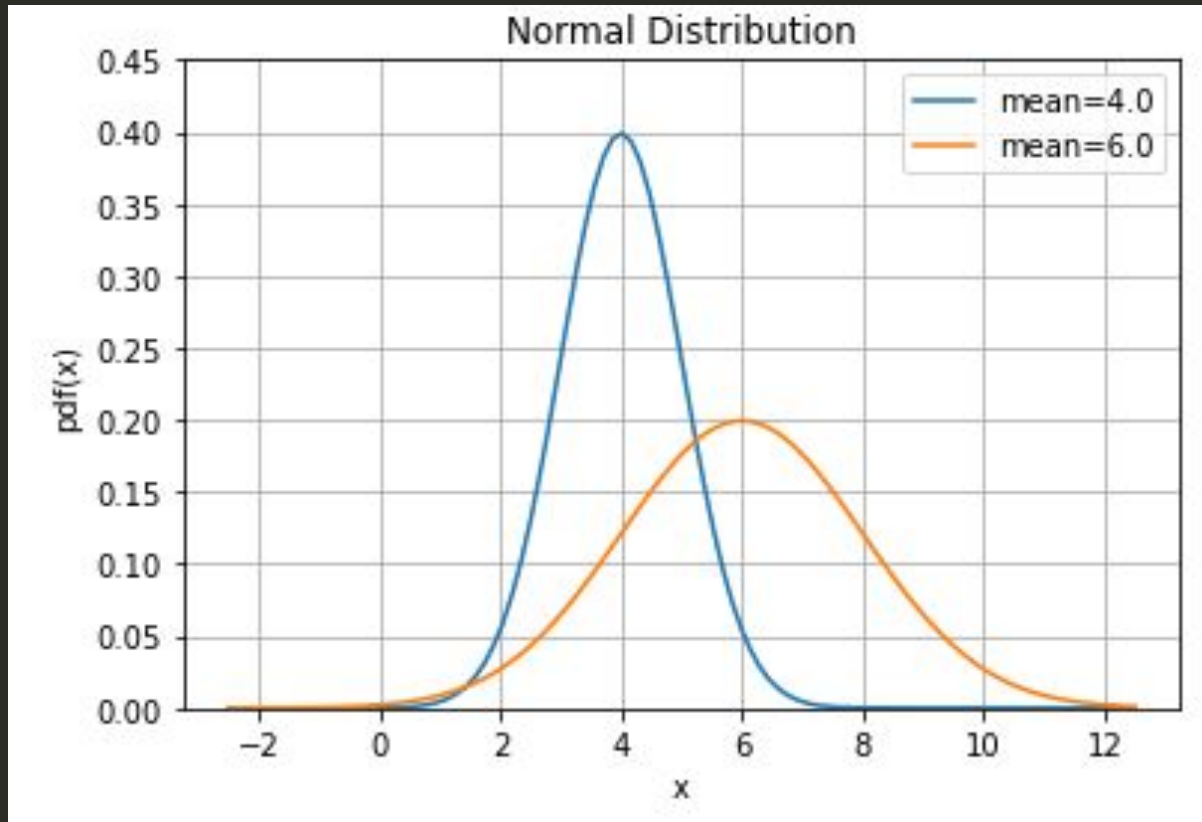
Outras distribuições

dnc>class



Distribuição Normal Standard

Distribuição Normal



Média de ph: 4.0
Desvio: 1.0
Medida: 1.8

Média de ph: 6
Desvio: 2
Medida: 1.6

Qual é a amostra relativamente mais ácida?

Distribuição Normal

Z-score: o número de desvios padrões que qualquer valor está da média. É possível **converter qualquer valor de uma distribuição** em um z-score, desde que o desvio não seja nulo, e, com isso, estamos **“standardizando”** a distribuição.

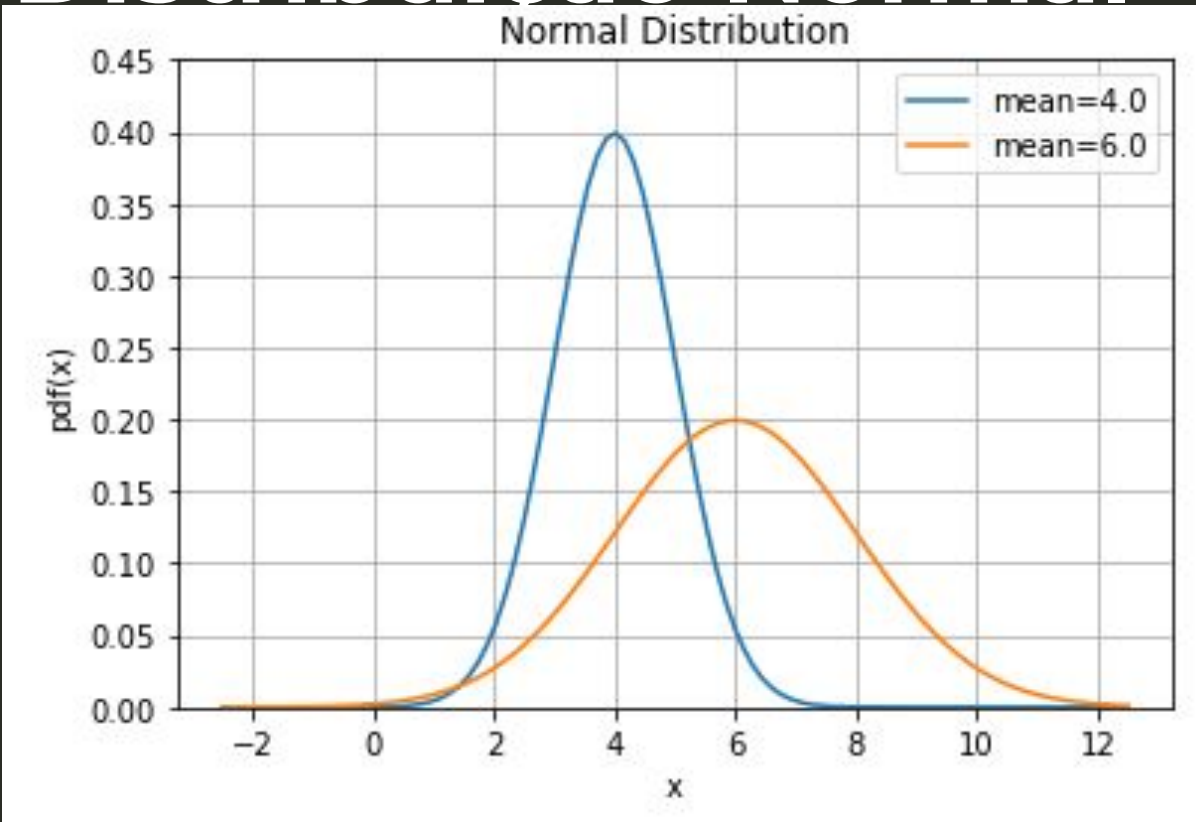
$$Z = \frac{x - \mu}{\sigma}$$

Em que:

μ é a **média** da população.

σ é o **desvio padrão** da população.

Distribuição Normal



Média de ph: 6
Desvio: 2
Medida: 1.6

Média de ph: 4.0
Desvio: 1.0
Medida: 1.8

$$Z = (1.8 - 4.0) / 1.0$$
$$Z = -2.2$$

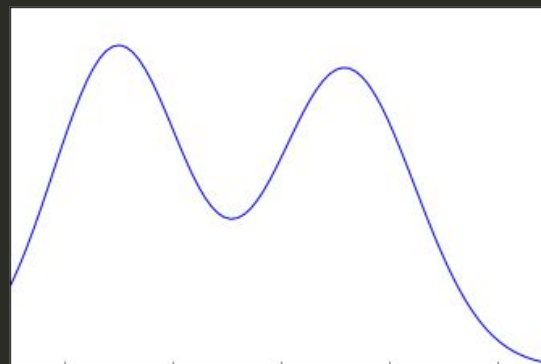
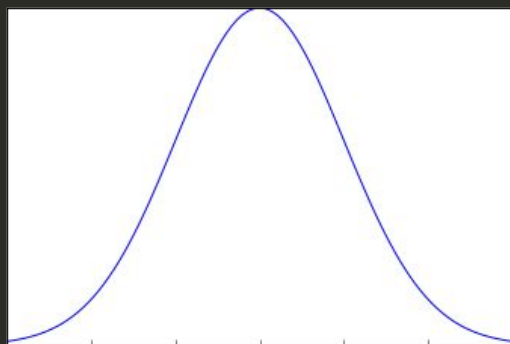


$$-2.2 = (x - 6.0) / 2.0$$
$$x = 1.6$$

Distribuição amostral

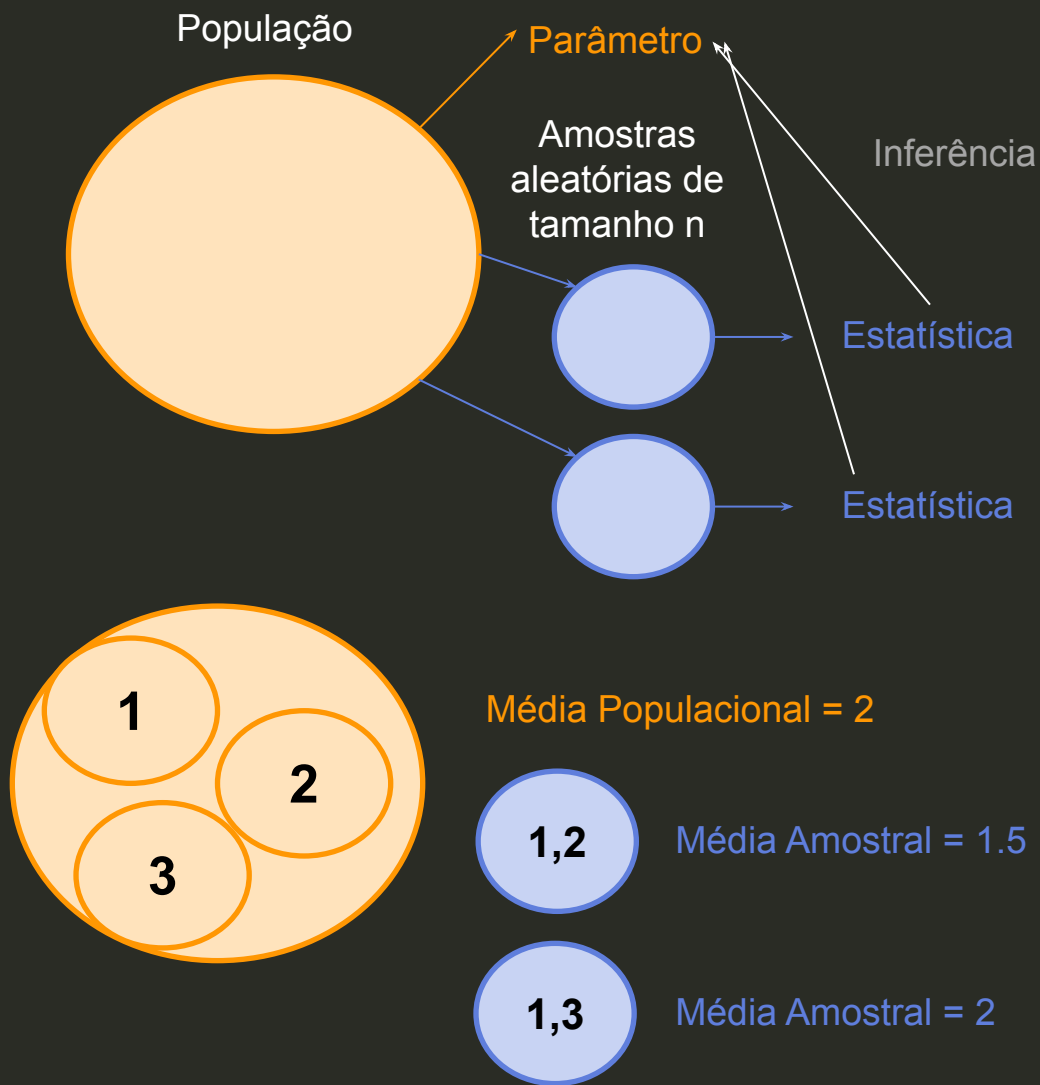
A **distribuição amostral** é a distribuição de probabilidades de uma amostra grande selecionada a partir da população.
Depende do tamanho da amostra e tipo de amostragem.

A forma da Distribuição amostral não revela nada sobre a forma da distribuição da população. Exemplo abaixo: distribuição amostral (esq.) e distribuição populacional (dir.).



A **distribuição amostral ajuda a estimar parâmetros populacionais**.
Como? A seguir... Teorema do Limite Central.

Distribuição amostral - Média



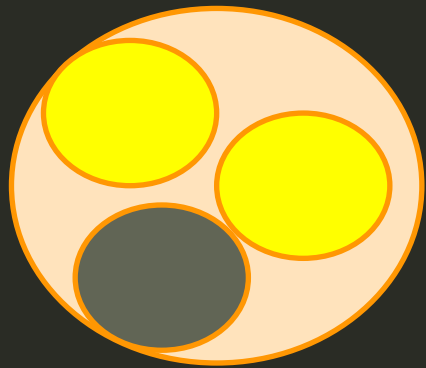
Itens amostrados	Média amostral
1,2	1,5
1,3	2
2,3	2.5
3,2	2.5

Distribuição Amostral da média de 4 amostras de tamanho 2

Distribuição amostral - Proporção

Distribuição amostral da média de amostra

Distribuição amostral da proporção de amostra (variáveis categóricas)



Proporção Populacional amarelo = $2/3$

A,C

Proporção Amostral = $1/2$

A,A

Proporção Amostral = 1

Distribuição amostral – Erro padrão (Standard Error)

Cálculo do erro padrão – “standard error” (que é o desvio padrão da distribuição amostral) e representa uma medida de incerteza da amostra

Média

$$SE = \frac{\sigma}{\sqrt{n}}$$

“ σ ” (Sigma) é o desvio padrão populacional e “ n ” é o tamanho da amostra

Proporção

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

“ p ” é a proporção populacional e “ n ” é o tamanho da amostra

Erro padrão (Standard Error)

- Intuição

Cálculo do erro padrão – “standard error” (que é o desvio padrão da distribuição amostral) e representa uma medida de incerteza da amostra

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Quanto maior o tamanho da amostra maior será o denominador e menor o erro.

Se queremos diminuir o erro podemos aumentar o tamanho da amostra.

Recap – Distribuições de Probabilidade

Distribuição normal standard

Z-score

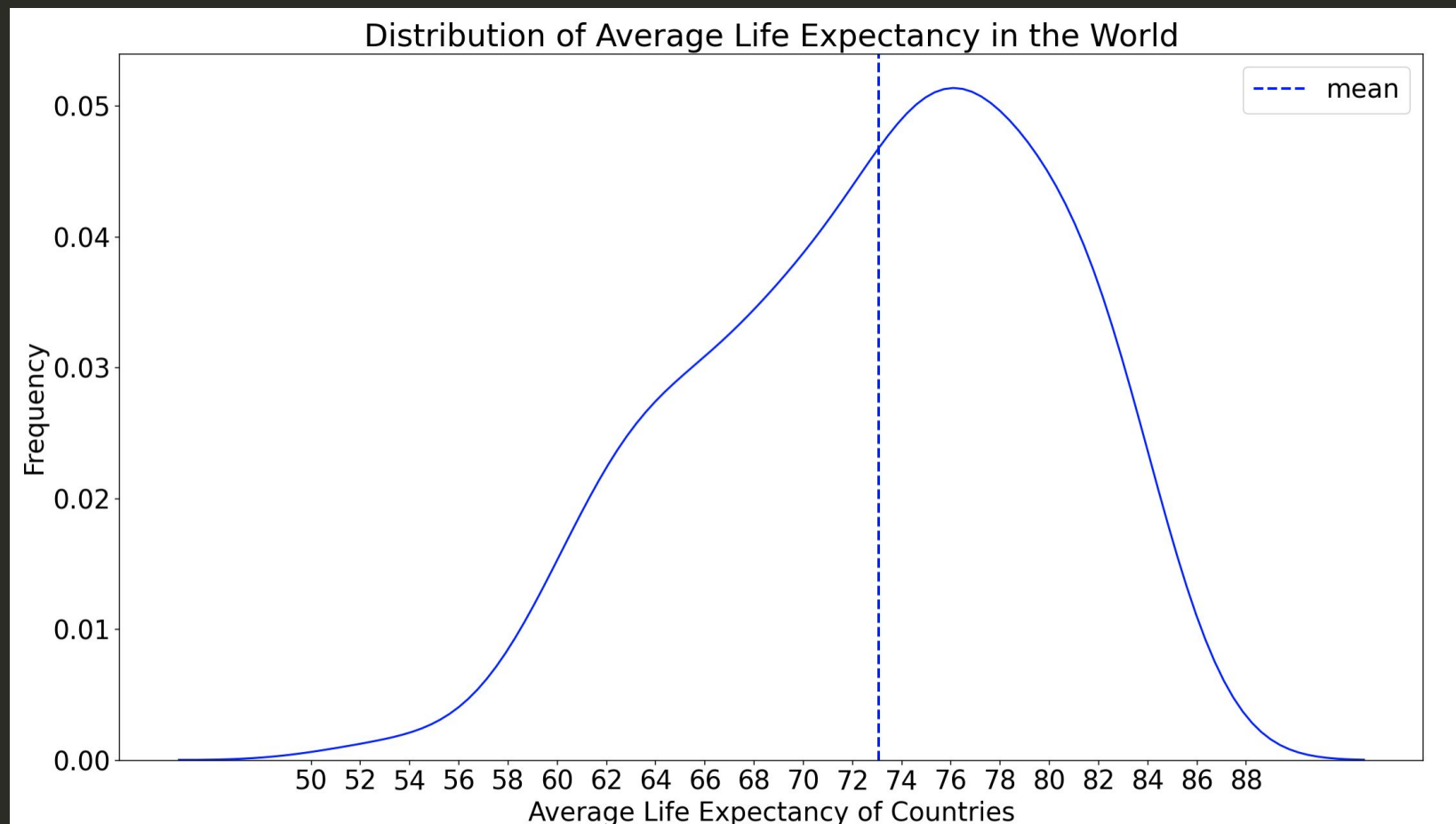
Distribuição amostral

Erro padrão

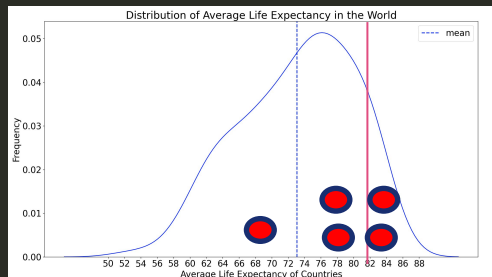
dnc>class

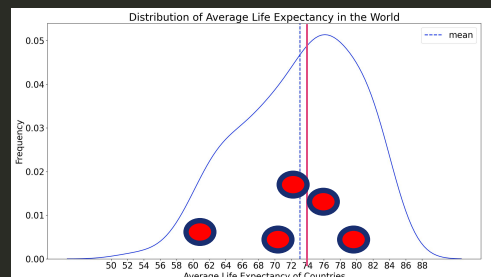
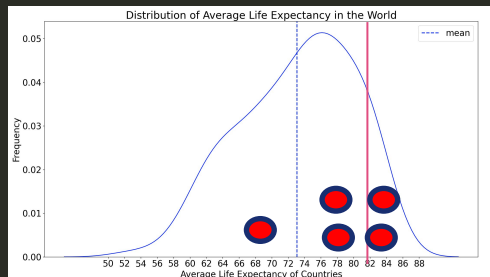


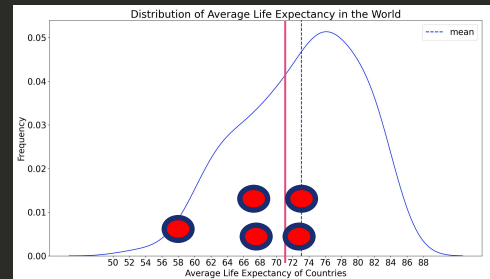
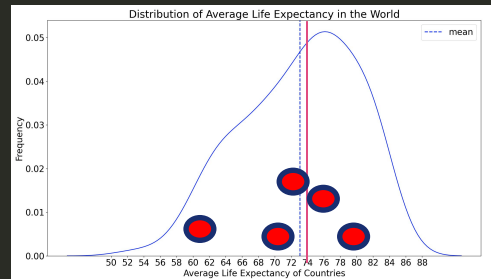
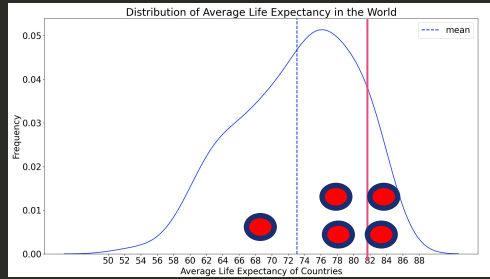
Teorema do Limite Central

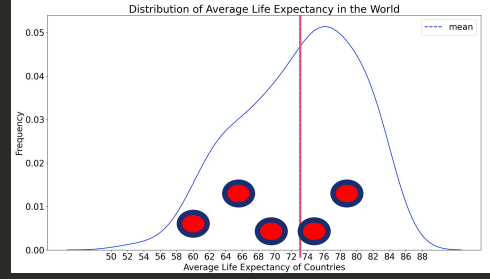
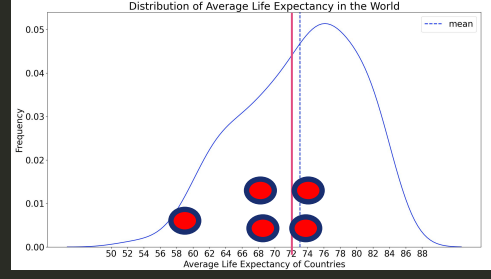
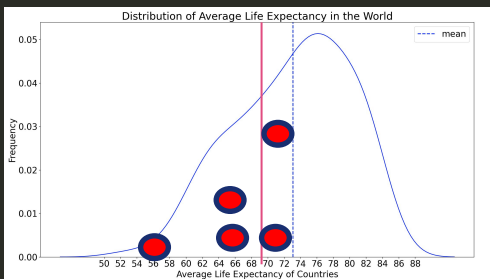
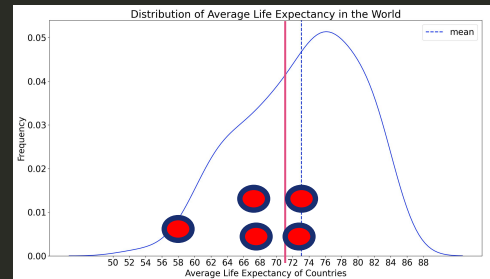
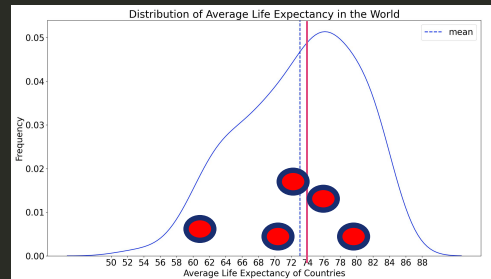
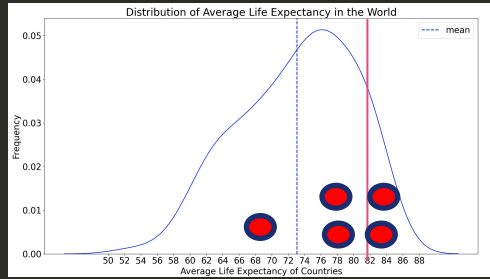


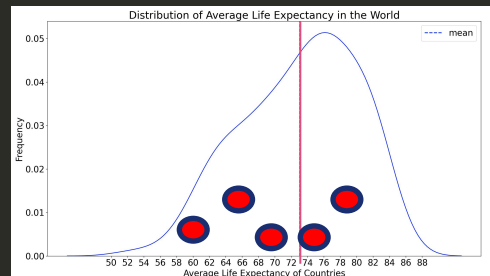
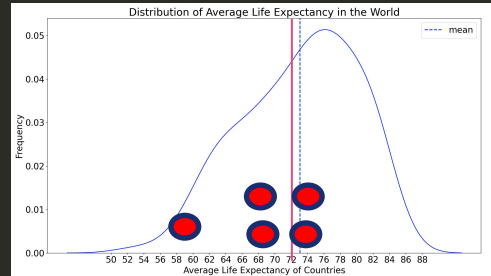
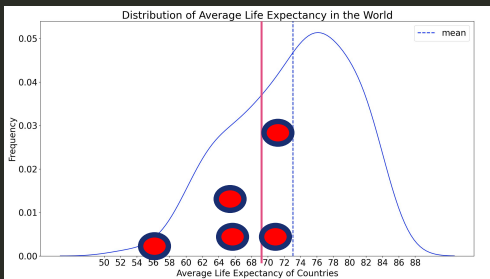
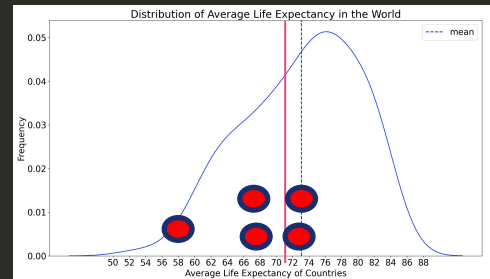
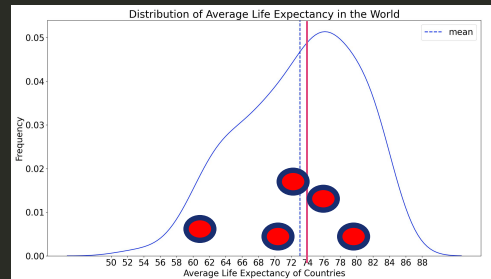
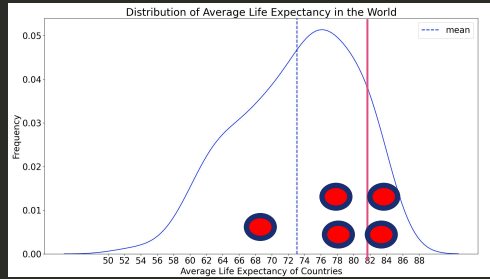
Processo de obtenção de uma **distribuição amostral**: 1000 subconjuntos de tamanho 150 cada serão extraídos da população. Plotar a média de cada um desses subconjuntos (amostra).











Amostra (n=150)

Média (em anos)

#1

85

#2

70

#3

66

#4

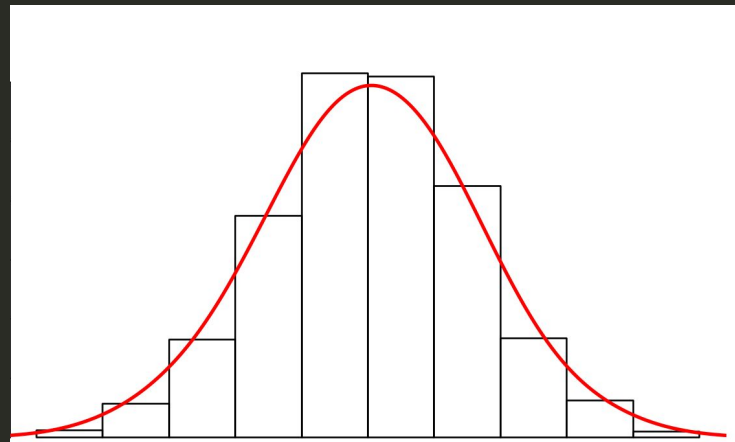
62

...

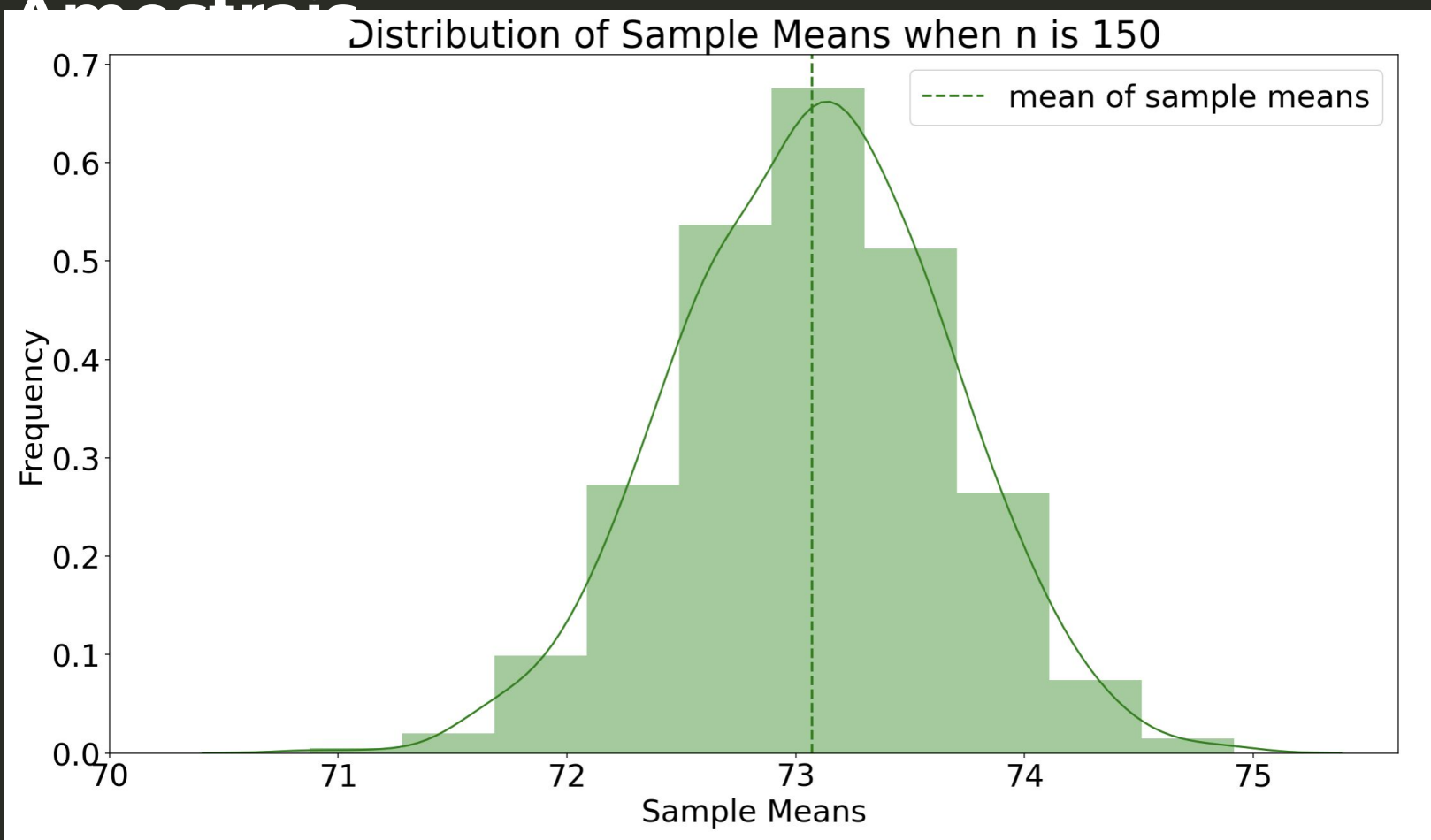
...

#1000

80



Distribuição Amostral das médias

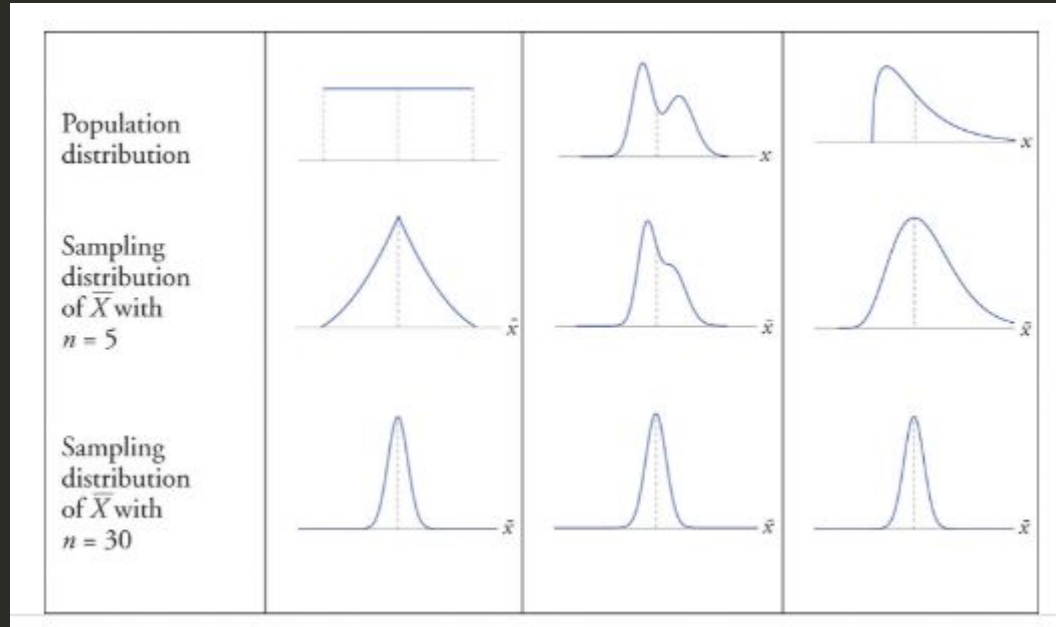


Qualquer distribuição amostral de médias, de uma população com qualquer distribuição, é aproximadamente uma distribuição normal.*

*se o tamanho da amostra for pelo menos 30.

Teorema do Limite Central

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).



Teorema do Limite Central

IMPLICAÇÕES

Uma forma de estimar a média populacional é através de observações repetidas de amostras de um tamanho fixo.

Mesmo que a distribuição original seja desconhecida, ou não normal, é possível utilizar técnicas de inferência bem desenvolvidas que são baseadas na distribuição normal.

Quanto maior o tamanho amostral, menor o erro padrão e maior a acurácia em determinar a média populacional a partir da estatística amostral.

Teorema do Limite Central

Exemplo:

Em uma empresa muito grande o salário médio é de 6.200 reais com desvio padrão de 3.200 reais.

Se uma das pessoas funcionárias é selecionada aleatoriamente qual a probabilidade do salário dessa pessoa exceder 6.600?

Teorema do Limite Central

Exemplo:

Em uma empresa muito grande o salário médio é de 6.200 reais com desvio padrão de 3.200 reais.

Se uma das pessoas funcionárias é selecionada aleatoriamente qual a probabilidade do salário dessa pessoa exceder 6.600?

X = variável aleatória que representa o salário de uma pessoa selecionada aleatoriamente

$P(X > 6.600)$

$P(X > 6.600) = P(z > (6600 - 6200)/3200) = P(z > 0.125)$

$$z = \frac{x - \mu}{\sigma}$$

Encontrar a probabilidade do z usando a distribuição normal standard...

Teorema do Limite Central

Exemplo:

Em uma empresa muito grande o salário médio é de 6.200 reais com desvio padrão de 3.200 reais.

Se uma das pessoas funcionárias é selecionada aleatoriamente qual a probabilidade do salário dessa pessoa exceder 6.600?

X = variável aleatória que representa o salário de uma pessoa selecionada aleatoriamente

$$P(X > 6.600)$$

$$P(X > 6.600) = P(z > (6600 - 6200)/3200) = P(z > 0.125)$$

Encontrar a probabilidade do z usando a distribuição normal standard...

... Seria um **ERRO**.

Não sabemos se a distribuição de salários é normal. Não é possível estimar sem mais informações sobre a distribuição de salários.

Teorema do Limite Central

Exemplo:

Em uma empresa muito grande o salário médio é de 6.200 reais com desvio padrão de 3.200 reais.

Se 100 pessoas funcionárias são selecionadas aleatoriamente qual a probabilidade da média do salário dessas pessoas exceder 6.600?

Teorema do Limite Central

Exemplo:

Em uma empresa muito grande o salário médio é de 6.200 reais com desvio padrão de 3.200 reais.

Se 100 pessoas funcionárias são selecionadas aleatoriamente qual a probabilidade da média do salário dessas pessoas exceder 6.600?

\bar{x} = média de salários da amostra

\bar{x} > distribuição amostral aproximadamente normal (CLT)

$$z = \frac{x - \mu}{\sigma} \quad SE = \frac{\sigma}{\sqrt{n}}$$

$P(\bar{x} > 6.600)$

$P(z > 6.600) = P(z > (6600 - 6200)/(3200/\sqrt{100})) = P(z > 1.25)$

Encontrar a probabilidade do z usando a distribuição normal standard...

Teorema do Limite Central

Standard Normal Probabilities

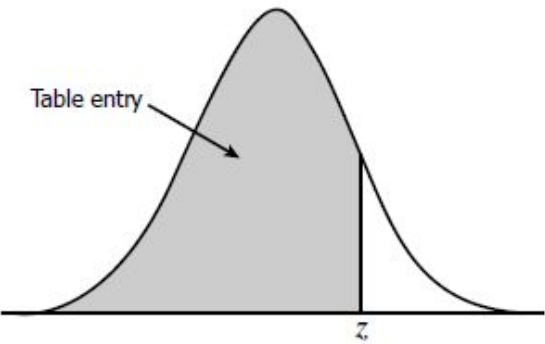


Table entry for z is the area under the standard normal curve to the left of z .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

$$\frac{\sigma}{\sqrt{n}}$$

Teorema do Limite Central

Exemplo:

Em uma empresa muito grande o salário médio é de 6.200 reais com desvio padrão de 3.200 reais.

Se 100 pessoas funcionárias são selecionadas aleatoriamente qual a probabilidade da média do salário dessas pessoas exceder 6.600?

\bar{x} = média de salários da amostra

\bar{x} > distribuição amostral aproximadamente normal (CLT)

$$z = \frac{x - \mu}{\sigma} \quad SE = \frac{\sigma}{\sqrt{n}}$$

$P(\bar{x} > 6.600)$

$P(z > 6.600) = P(z > (6600 - 6200)/(3200/\sqrt{100})) = P(z > 1.25)$

Encontrar a probabilidade do z usando a distribuição normal standard...

$P(z > 1.25) = 1 - P(z < 1.25) = 1 - 0.8944 = 0.1056$

$P(\bar{x} > 6.600)$ é aproximadamente 0.106.

Teorema do Limite Central

Em problemas de **aprendizado de máquina** o conjunto de dados representa uma amostra da população. E, ao estudar essa amostra, buscamos **captar os principais padrões nos dados e generalizar para a população**.

O **CLT** auxilia nesse processo de **inferência de amostra para população e na construção de modelos de aprendizado melhores**.

Ajuda a informar se a **amostra pertence a uma população**, olhando para a distribuição amostral.

Recap – Teorema do Limite Central

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

Uma forma de estimar a média populacional é através de observações repetidas de amostras de um tamanho fixo.

Coletamos amostras e plotamos as médias, sabemos aproximadamente onde a média populacional se encontra – mas é importante mensurar o quanto temos confiança nessa medida – que é estimado usando **Intervalos de Confiança**

dnc>class



Intervalo de Confiança

Intervalo de Confiança

Estimar parâmetros: usar estatística amostral para tirar conclusões sobre parâmetros populacionais.

Estimativa pontual (“Point Estimate”)

Estimativa de Intervalo (“Interval Estimate”)

As estimativas são feitas a partir de amostras da população, por isso, são suscetíveis a erros – tanto para mais como para menos. Não há muito valor em uma estimativa pontual se não for possível estimar a incerteza desse ponto.

Intervalo de Confiança

Estimar parâmetros: usar estatística amostral para tirar conclusões sobre parâmetros populacionais.

Estimativa pontual (“Point Estimate”)

Estimativa de Intervalo (“Interval Estimate”)

Nível de confiança (“level of confidence” – valor de probabilidade)

Estimativa de intervalo + nível de confiança = intervalo de confiança

Intervalo de Confiança

Estimar parâmetros: usar estatística amostral para tirar conclusões sobre parâmetros populacionais.

Intervalo de confiança é uma estimação de intervalos ou limites para **valores plausíveis de um parâmetro**, ou seja, uma variável populacional. Representa a quantificação **da incerteza de uma estimativa**.

- 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**.

Intervalo de Confiança

Estimar parâmetros: usar estatística amostral para tirar conclusões sobre parâmetros populacionais.

Intervalo de confiança é uma estimação de intervalos ou limites para **valores plausíveis de um parâmetro**, ou seja, uma variável populacional. Representa a quantificação **da incerteza de uma estimativa**.

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

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

*(ou sigma, desvio populacional, quando conhecido)

Intervalo de Confiança

Estimar parâmetros: usar estatística amostral para tirar conclusões sobre parâmetros populacionais.

Intervalo de confiança é uma estimação de intervalos ou limites para **valores plausíveis de um parâmetro**, ou seja, uma variável populacional. Representa a quantificação **da incerteza de uma estimativa**.

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

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

*(ou sigma, desvio populacional, quando conhecido)

INTERVALO DE CONFIANÇA:

POINT ESTIMATE +/- MARGEM DE ERRO (DADO NÍVEL DE CONFIANÇA E ERRO PADRÃO)

Intervalo de Confiança

Confidence Interval for the Population
Mean if **SD σ known** or **$n \geq 30$**

$$\overline{X} \pm z \frac{\sigma}{\sqrt{n}} \longrightarrow \sigma \text{ known}$$

$$\overline{X} \pm z \frac{s}{\sqrt{n}} \longrightarrow n \geq 30$$

INTERVALO DE CONFIANÇA:

**POINT ESTIMATE +/- MARGEM DE ERRO (DADO NÍVEL DE
CONFIANÇA E ERRO PADRÃO)**

Intervalo de Confiança - Premissas

Intervalo de confiança é uma estimacão de intervalos ou limites para **valores plausíveis de um parâmetro**, ou seja, uma variável populacional. Representa a quantificação **da incerteza de uma estimativa**.

INTERVALO DE CONFIANÇA:

POINT ESTIMATE +/- MARGEM DE ERRO (DADO NÍVEL DE CONFIANÇA E ERRO PADRÃO)

Premissas:

- Amostragem simples aleatória da população
- População normal (não importante se o tamanho da amostra for grande, pelo CLT)
- Conhecimento do desvio populacional*

*há formas de calcular se o desvio for desconhecido

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

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

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

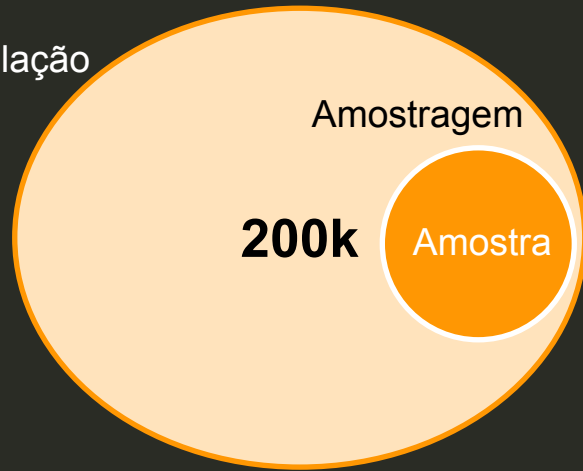
200k

População

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

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

População



Amostragem

200k

Amostra

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

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

População

200k

Amostra

N =
135

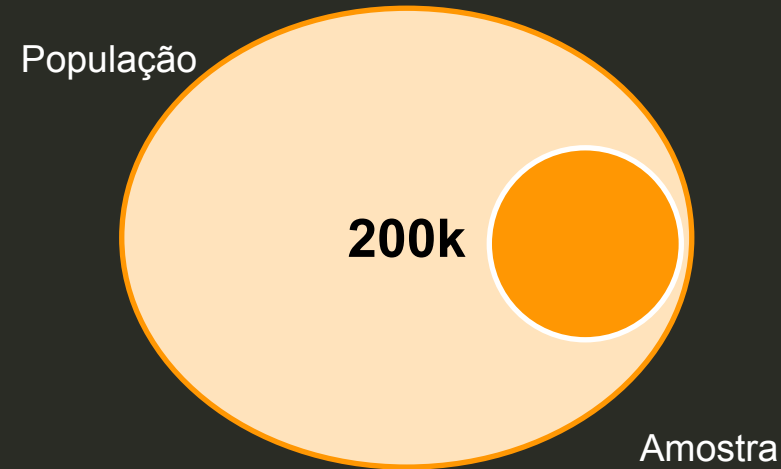


$\bar{x} = 0.988 \text{ cm}$

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

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

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



N =
135



$\bar{x} = 0.988 \text{ cm}$

Nível de confiança 95% e desvio populacional ($\sigma = 0.028$)*

*nem sempre conhecido

**premissas validadas

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

$$IC = 0.988 \pm 1.96 \times (0.028/\sqrt{135})$$

$$IC = 0.988 \pm 0.0047$$

$$(0.983, 0.993)$$

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

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

Desvio populacional (sigma) desconhecido

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

200k

N =
135

$\bar{x} = 0.988$ cm

Nível de confiança 95%

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

Desvio populacional (sigma) desconhecido

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

200k

N =
135

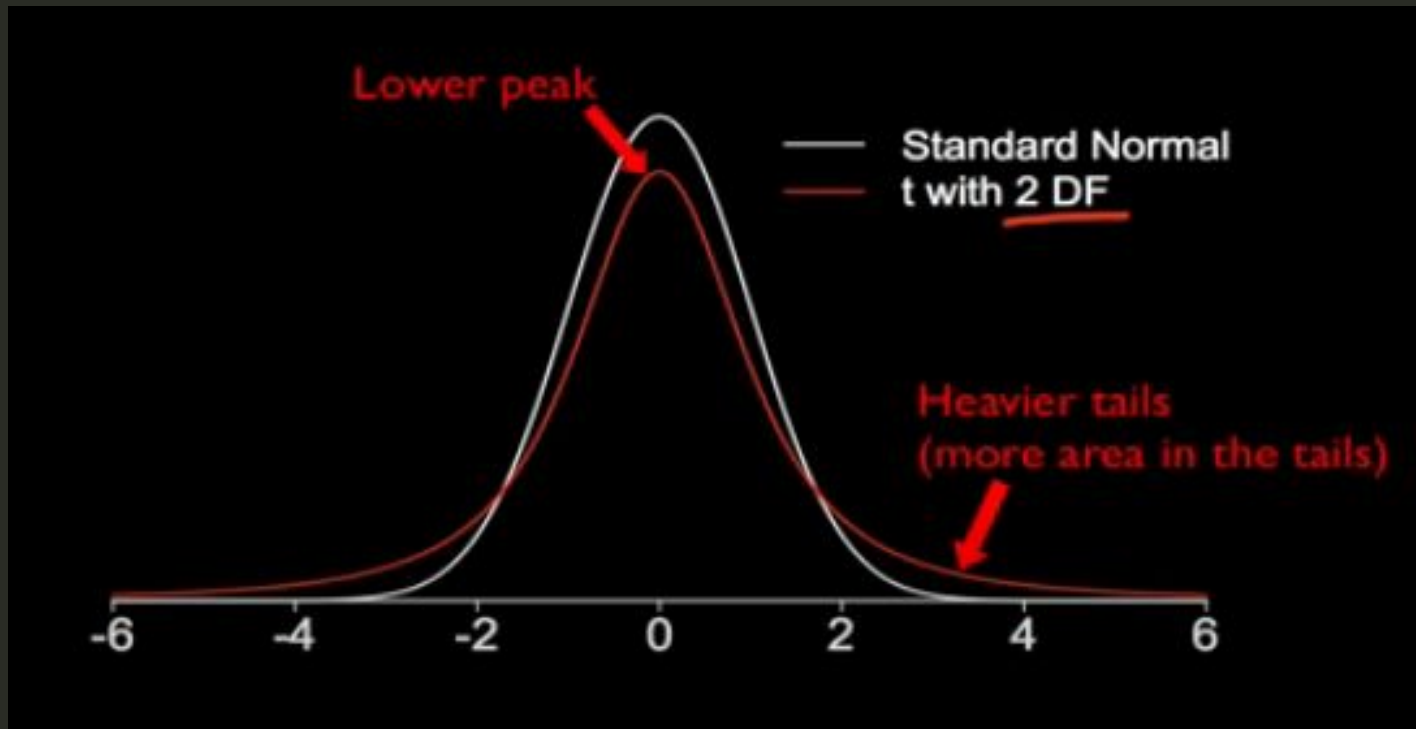
$\bar{x} = 0.988$ cm

Amostra

Nível de confiança 95%

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

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$



Degrees of freedom (DF) – grau de liberdade: relacionado ao tamanho da amostra

Quanto **maior DF**, maior a amostra e mais próxima de uma normal a distribuição t caminha.

Produtor de queijos está produzindo itens com 750g

Desvio populacional (sigma) desconhecido

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

200k

N = 7

$\bar{x} = 795.3g$

Nível de confiança 95%
s = 17.8g

T = n-1 DFs = 6 (t-table: 2.447)

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

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$\begin{aligned} IC &= 795.3 \pm 2.447 \times (17.8/\sqrt{7}) \\ IC &= 795.3 \pm 16.46 \\ &(778.8, 811.8) \end{aligned}$$

**Eleição:
pessoas candidatas A e B**

**Intervalo de confiança para
proporção**

**Eleição:
pessoas candidatas A e B**

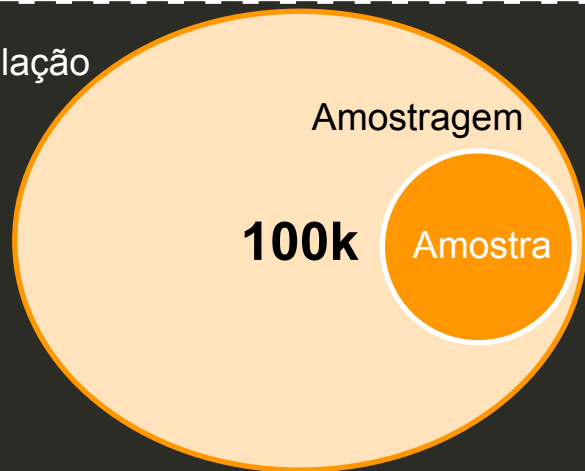
População

100k

**qual a proporção de pessoas que votará
para A?**

Eleição:
pessoas candidatas A e B

População



qual a proporção de pessoas que votará
para A?

Eleição:
pessoas candidatas A e B

qual a proporção de pessoas que votará
para A?

População

100k

Amostra

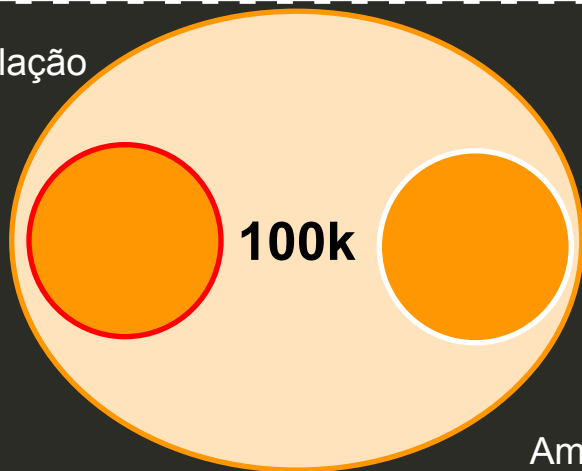
N =
100



$p = 0.54$

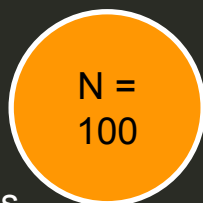
Eleição:
pessoas candidatas A e B

População

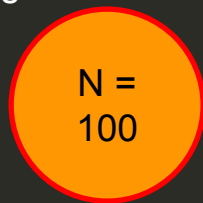


qual a proporção de pessoas que votará
para A?

Amostras



$p = 0.54$

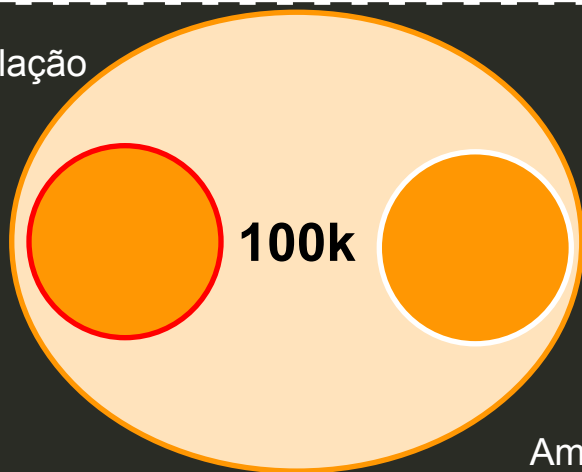


$p = 0.58$

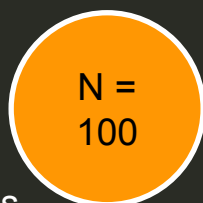
Eleição:
pessoas candidatas A e B

qual a proporção de pessoas que votará
para A?

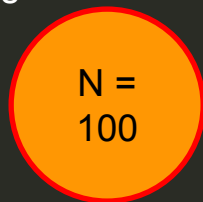
População



Amostras

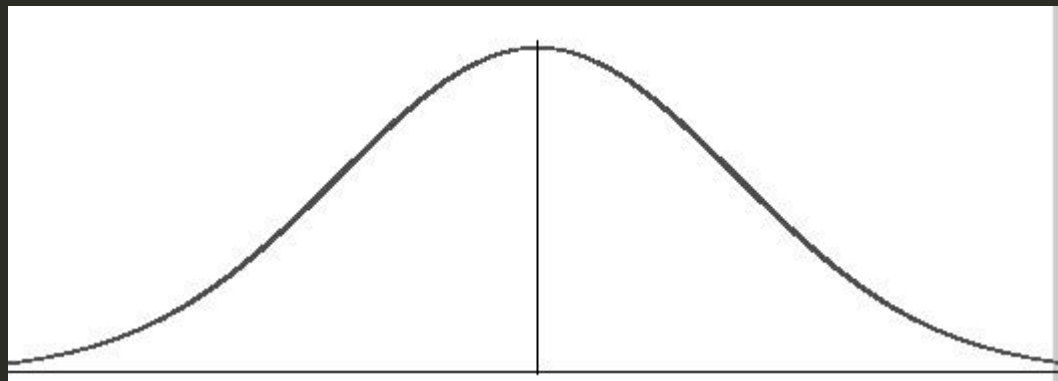


$P_{\text{amostra}} = 0.54$



$P_{\text{amostra}} = 0.58$

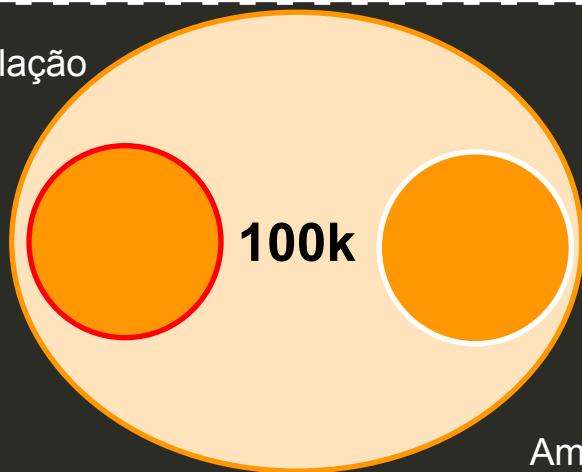
Distribuição amostral das proporções de amostras ($n=100$)



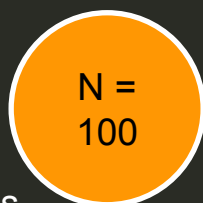
Eleição:
pessoas candidatas A e B

qual a proporção de pessoas que votará
para A?

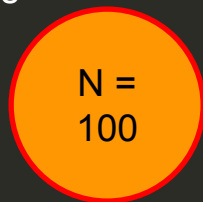
População



Amostras

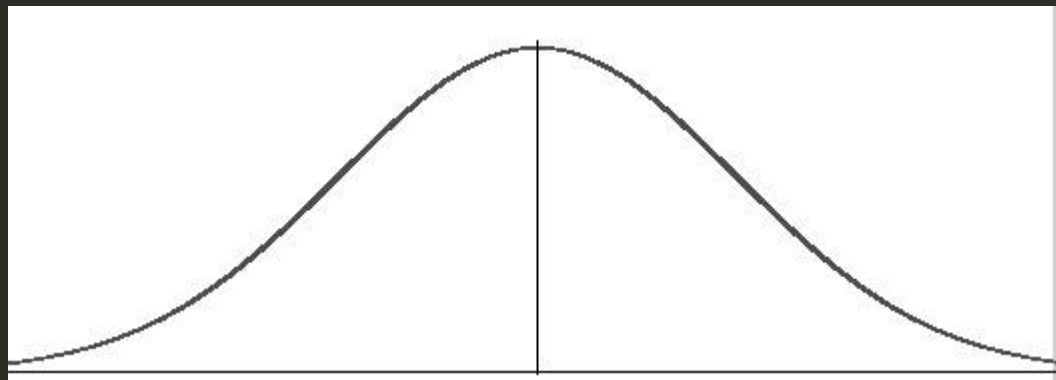


$P_{\text{amostra}} = 0.54$



$P_{\text{amostra}} = 0.58$

Distribuição amostral das proporções de amostras ($n=100$)

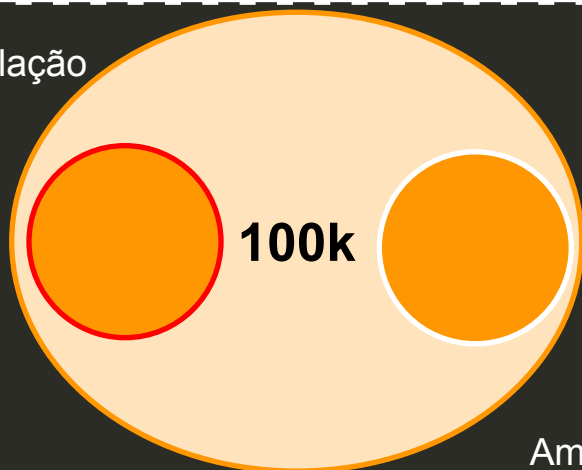


P populacional

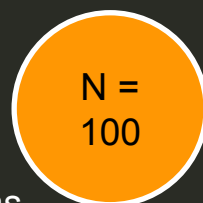
Eleição:
pessoas candidatas A e B

qual a proporção de pessoas que votará
para A?

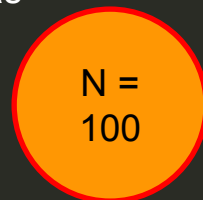
População



Amostras

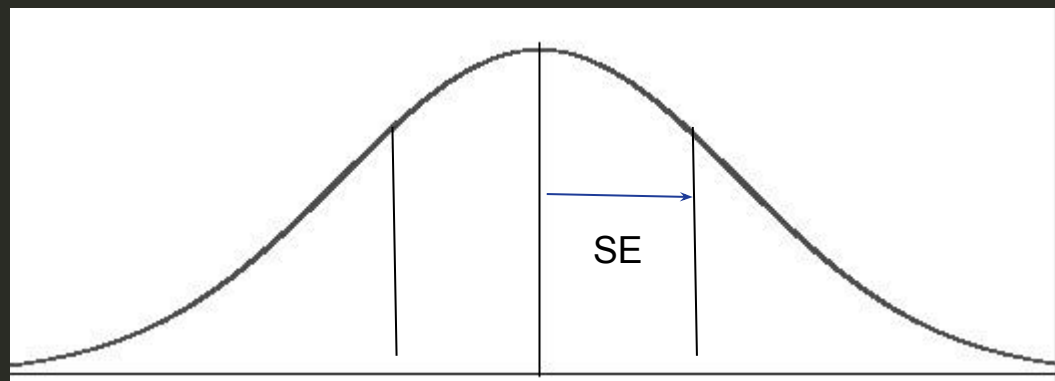


P amostra = 0.54



P amostra = 0.58

Distribuição amostral das proporções de amostras (n=100)



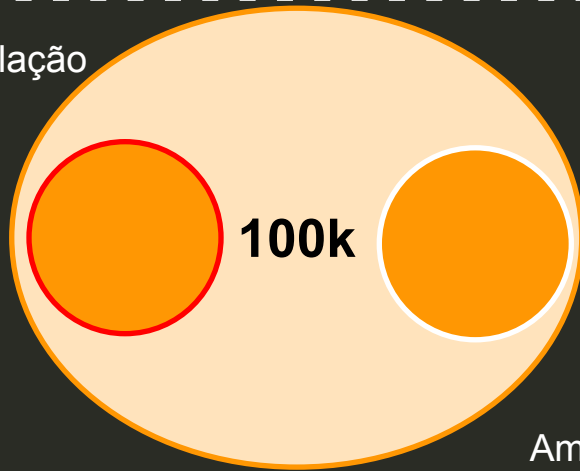
SE (standard error) = desvio
da distribuição amostral

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Eleição:
pessoas candidatas A e B

qual a proporção de pessoas que votará
para A?

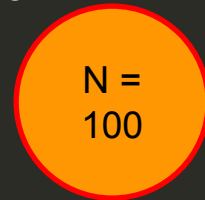
População



Amostras



P amostra = 0.54



P amostra = 0.58

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

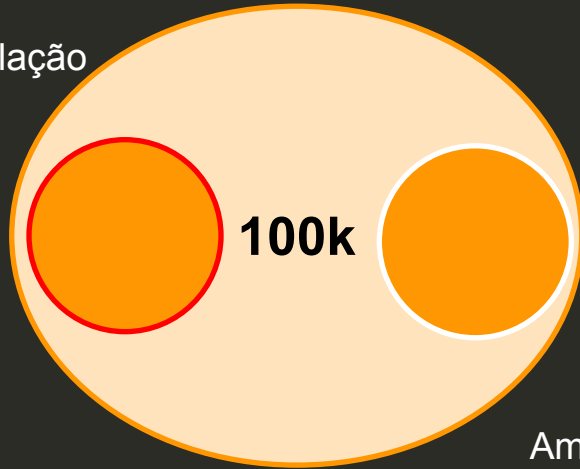
SE = sqrt(0.54x0.46)/100)

SE~ 0.05

Eleição:
pessoas candidatas A e B

qual a proporção de pessoas que votará
para A?

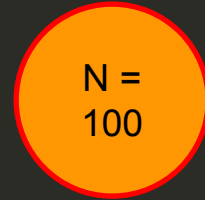
População



Amostras



P amostra = 0.54



P amostra = 0.58

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$SE = \text{sqrt}(0.54 \times 0.46) / 100$$

$$SE \sim 0.05$$

$$IC = 0.54 \pm 1.96 \times 0.05$$

$$IC = 0.54 \pm 0.098$$

$$(0.442, 0.638)$$

Pessoas que pretendem votar em A

dnc>class



Testes de Hipóteses

Teste de Hipóteses

Testar hipóteses: Testar hipóteses: 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)

Teste de Hipóteses

Estimação
(ex. Intervalos de
confiança)

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?

Teste de Hipóteses

Teste de Medicamentos



A



B

Média: 6 hrs	Média: 20 hrs
7	10
9	9
2	41

14 horas de
diferença

Teste de Hipóteses

Teste de Medicamentos



A

Média: 6 hrs	Média: 20 hrs
7	10
9	9
2	41



B

Hipótese:

peessoas que tomam o medicamento A se recuperam, em média, com 14 horas a menos que as pessoas que tomam o medicamento B.

Média: 20 hrs	Média: 10 hrs
15	8
20	12
25	10

Média: 15 hrs	Média: 6 hrs
20	8
13	4
12	6

Média: 14 hrs	Média: 8 hrs
11	7
13	9
18	8

14 horas de
diferença

Teste de Hipóteses

Teste de Medicamentos



A

Média: 6 hrs	Média: 20 hrs
7	10
9	9
2	41

14 horas de
diferença



B

Hipótese:

pessoas que tomam o medicamento A se recuperam, em média, com 14 horas a menos que as pessoas que tomam o medicamento B.

Média: 20 hrs	Média: 10 hrs
15	8
20	12
25	10

Média: 15 hrs	Média: 6 hrs
20	8
13	4
12	6

Média: 14 hrs	Média: 8 hrs
11	7
13	9
18	8

Oposto resultado da hipótese original

Teste de Hipóteses

Teste de Medicamentos



A

Média: 6 hrs	Média: 20 hrs
7	10
9	9
2	41

14 horas de
diferença



B

Hipótese:

~~pessoas que tomam o medicamento A se recuperam, em média, com 14 horas a menos que as pessoas que tomam o medicamento B.~~

Rejeitamos a hipótese!

Média: 20 hrs	Média: 10 hrs
15	8
20	12
25	10

Média: 15 hrs	Média: 6 hrs
20	8
13	4
12	6

Média: 14 hrs	Média: 8 hrs
11	7
13	9
18	8

Oposto resultado da hipótese original

Teste de Hipóteses

Teste de Medicamentos



Z



Y

Média: 6 hrs	Média: 20 hrs
7	10
9	9
2	41

14 horas de
diferença

Hipótese:

peessoas que tomam o medicamento Z se recuperam, em média, com 14 horas a menos que as pessoas que tomam o medicamento B.

Resultados não são diferentes o suficiente para rejeitarmos a hipótese, mas também não nos convence de a hipótese ser verdadeira. Melhor opção é falhar em rejeitar a hipótese!

Média: 7 hrs	Média: 20 hrs
15	8
20	12
25	10

13 horas de
diferença

Média: 5 hrs	Média: 19.5 hrs
6	16
5	18
4	24.5

14.5 horas de
diferença

Média: 8 hrs	Média: 21.5 hrs
8	22
9	17.5
7	25

13.5 horas de
diferença

Teste de Hipóteses

Teste de Medicamentos



Z

Y

Média: 6 hrs	Média: 20 hrs
7	10
9	9
2	41

14 horas de
diferença

Hipótese:

peessoas que tomam o medicamento Z se recuperam, em média, com 14 horas a menos que as pessoas que tomam o medicamento B.

Várias hipóteses poderiam ser razoáveis:
14hrs, 13hr, 13.5hrs, 13.4hrs? Como escolher a melhor hipótese?

Média: 7 hrs	Média: 20 hrs
15	8
20	12
25	10

13 horas de
diferença

Média: 5 hrs	Média: 19.5 hrs
6	16
5	18
4	24.5

14.5 horas de
diferença

Média: 8 hrs	Média: 21.5 hrs
8	22
9	17.5
7	25

13.5 horas de
diferença

Teste de Hipóteses

**Teste de
Medicamentos**



Z



Y

Hipótese:

Não há diferença no tempo de recuperação entre os medicamentos Z e Y.

Teste de Hipóteses

Teste de Medicamentos



Z



Y

Hipótese Nula (H_0) :

Não há diferença no tempo de recuperação entre os medicamentos Z e Y.

Não precisamos nos preocupar se a diferença é exatamente 6 ou 6.5 hrs, por exemplo. Sem a hipótese nula, seria necessário dados preliminares para construir uma hipótese.

Teste de Hipóteses

Teste de Medicamentos



Z



Y

Hipótese Nula (H_0):

Não há diferença no tempo de recuperação entre os medicamentos Z e Y.

Não precisamos nos preocupar se a diferença é exatamente 6 ou 6.5 hrs, por exemplo. Sem a hipótese nula, seria necessário dados preliminares para construir uma hipótese. **Hipótese “Status quo”.**

Hipótese Alternativa (H_a , H_1):

Há diferença no tempo de recuperação entre os medicamentos Z e Y. **Hipótese de pesquisa.**

Contrário da hipótese nula (quando temos apenas dois grupos em comparação).

Teste de Hipóteses - Exemplos

Teste de Medicamentos



Hipótese nula: não há diferença entre os tempos de recuperação dos medicamentos

Hipótese alternativa: não há diferença entre os tempos de recuperação dos medicamentos X e Y, mas há diferença entre eles para o medicamento Z.

Independente da hipótese alternativa, sempre rejeitamos ou falhamos em rejeitar a hipótese nula apenas.

Teste de Hipóteses - Exemplos

Homens e mulheres têm salários diferentes quando saem da graduação:

H0: média salarial de homens = média salarial de mulheres

Ha: média salarial de homens \neq média salarial de mulheres

* Hipóteses são sobre parâmetros, nunca estatísticas

Três processos produtivos de queijos produzem produtos com a mesma variação de peso?

H0: variação 1 = variação 2 = variação 3

Ha: variação 1 \neq variação 2 \neq variação 3

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

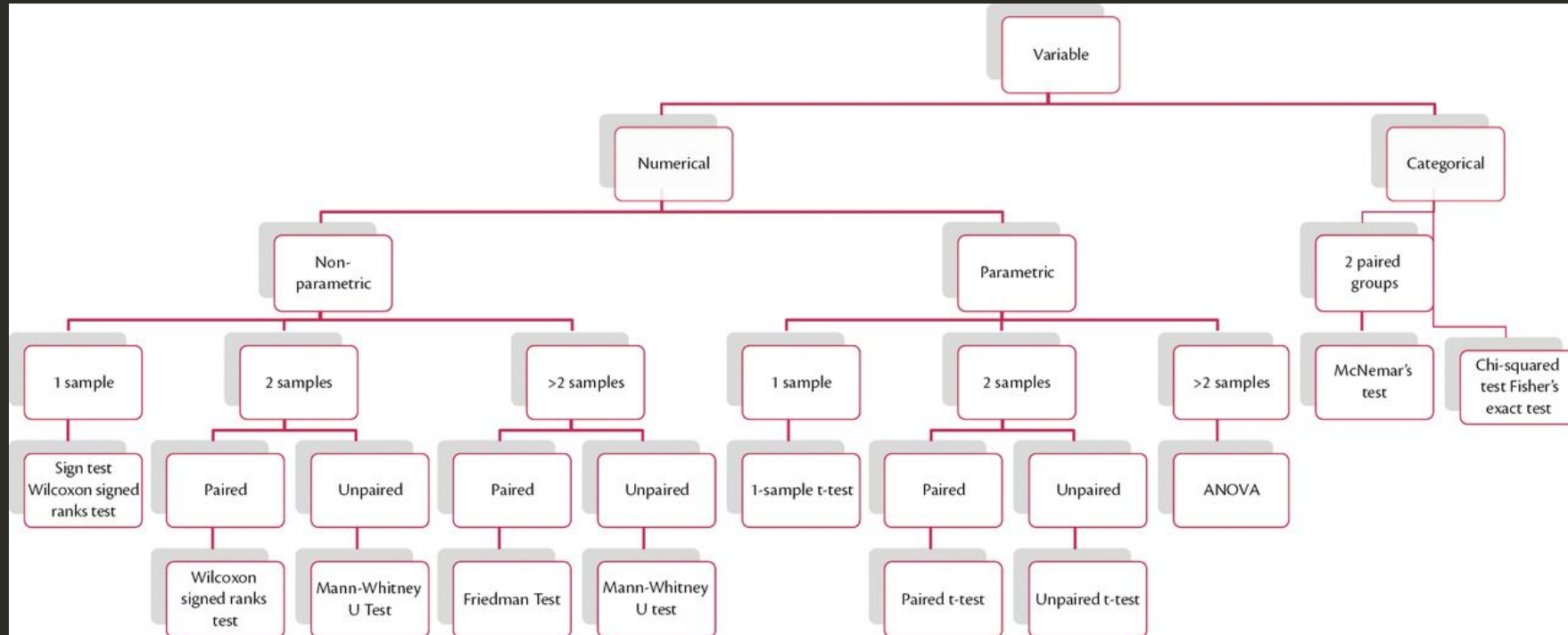
H0: peso médio = 750

Ha: peso médio \neq 750

ou

Ha: peso médio > 750

Testes estatísticos



Teste para média populacional

Desvio populacional sigma conhecido: z-test
Desvio populacional sigma desconhecido: t-test

Existe evidências de que a média populacional é diferente de um valor que nos interessa (valor hipotético)?

Teste para média populacional

Desvio populacional sigma conhecido: z-test
Desvio populacional sigma desconhecido: t-test

Existe evidências de que a média populacional é diferente de um valor que nos interessa (valor hipotético)?

H_0 : média populacional = valor hipotético

H_a : média populacional < valor hipotético

ou

H_a : média populacional > valor hipotético

ou

H_a : média populacional \neq valor hipotético

One-sided alternatives
One-tailed test

Two-sided alternatives
Two-tailed test

Teste para média populacional

Concentração de vitamina C em um composto manipulado. Testar se a manipulação está com os níveis acima de 0.40mg que é a necessidade de um grupo de clientes.

H_0 : média concentração = 0.40mg

H_a : média > 0.40mg

Sigma (desvio padrão populacional é 0.08mg)

16 amostras com média de 0.74mg

Teste para média populacional

Concentração de vitamina C em um composto manipulado. Testar se a manipulação está com os níveis acima de 0.40mg que é a necessidade de um grupo de clientes.

H0: média concentração = 0.40mg
Há: média > 0.40mg

Sigma (desvio padrão populacional é 0.08mg)

16 amostras com média de 0.74mg

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

$\bar{x} = 0.74$, sigma = 0.08, n=16, sigma da amostra (SE)=0.08/sqrt(16)

Teste para média populacional

Concentração de vitamina C em um composto manipulado.
Testar se a manipulação está com os níveis acima de 0.40mg
que é a necessidade de um grupo de clientes.

H0: média concentração = 0.40mg
Há: média > 0.40mg

Sigma (desvio padrão populacional é 0.08mg)

16 amostras com média de 0.74mg

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

$\bar{x} = 0.74$, sigma = 0.08, n=16, sigma da amostra (SE)=0.08/sqrt(16)

$$Z = 0.74 - 0.40 / 0.02 = 17$$

Teste para média populacional

Concentração de vitamina C em um composto manipulado.
Testar se a manipulação está com os níveis acima de 0.40mg
que é a necessidade de um grupo de clientes.

H0: média concentração = 0.40mg
Há: média > 0.40mg

Sigma (desvio padrão populacional é 0.08mg)

16 amostras com média de 0.74mg

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

$\bar{x} = 0.74$, sigma = 0.08, n=16, sigma da amostra (SE)=0.08/sqrt(16)

$$Z = 0.74 - 0.40 / 0.02 = 17$$

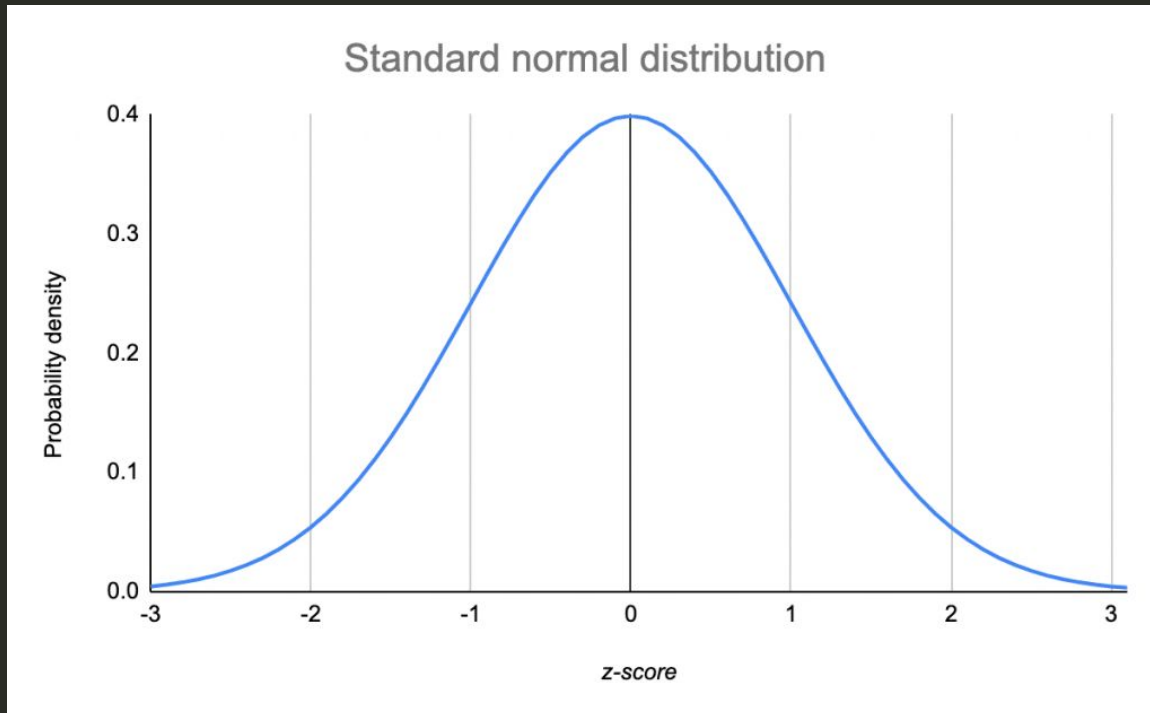
Se H0 é verdadeira, essa estatística do teste
terá uma distribuição normal standard, e 17
deverá ser uma amostra aleatória dessa
distribuição.

Teste para média populacional

$\bar{x} = 0.74$, $\sigma = 0.08$, $n=16$, σ da amostra (SE) = $0.08/\sqrt{16}$

$$Z = 0.74 - 0.40 / 0.02 = 17$$

Se H_0 é verdadeira, essa estatística do teste terá uma distribuição normal standard, e 17 deverá ser uma amostra aleatória dessa distribuição.



$$P(Z \geq 17) = 4.1 \times 10^{-65}$$

**Evidência forte
contra H_0 . Mas
quanto é
significante?**

Teste para média populacional

P-valor:

Probabilidade de obter o valor do teste estatístico ou um valor como evidência contra a H_0 , se a hipótese nula é verdadeira.

Probabilidade condicional se H_0 é verdadeira.

$$Z = 1.53$$

$$P\text{-valor} = P(Z \geq 1.53) = 0.063 \text{ (tabela standard normal)}$$

Outras possibilidades dependendo da hipótese alternativa

$$P\text{-valor} = P(Z \leq 1.53)$$

$$P\text{-valor} = P(Z \geq 1.53) + P(Z \leq -1.53)$$

Quanto menor o p-valor maior a evidência contra a hipótese nula.

Teste para média populacional

P-valor:

Probabilidade de obter o valor do teste estatístico ou um valor como evidência contra a H_0 , se a hipótese nula é verdadeira.

Probabilidade condicional se H_0 é verdadeira.

$$Z = 1.53$$

$$P\text{-valor} = P(Z \geq 1.53) = 0.063 \text{ (tabela standard normal)}$$

Quanto menor o p-valor maior a evidência contra a hipótese nula.

Mas quanto é o suficiente? Nível de significância (alpha)!

Se o p-valor for menor que o nível de significância, rejeitamos a H_0 em favor da hipótese alternativa com um nível de significância de 0.05.

dnc➤class



Pacotes em Python e R



SciPy (pronunciado “Sigh Pie”) é um software de código aberto para matemática, ciência e engenharia.

Statsmodel fornece classes e funções para a estimativa de diferentes modelos estatísticos, assim como para a condução de testes estatísticos e exploração estatística de dados..

Funções Estatísticas (scipy.stats)

SciPy User Guide

- Introduction
- Special functions (**scipy.special**)
- Integration (**scipy.integrate**)
- Optimization (**scipy.optimize**)
- Interpolation (**scipy.interpolate**)
- Fourier Transforms (**scipy.fft**)
- Signal Processing (**scipy.signal**)
- Linear Algebra (**scipy.linalg**)
- Sparse eigenvalue problems with ARPACK
- Compressed Sparse Graph Routines (**scipy.sparse.csgraph**)
- Spatial data structures and algorithms (**scipy.spatial**)
- Statistics (**scipy.stats**)
- Multidimensional image processing (**scipy.ndimage**)
- File IO (**scipy.io**)

Statistics stats

statsmodels v0.13.2	Power and Sample Size Calculations
Installing statsmodels	The power module currently implements power and sample size calculations and Chi-square goodness of fit test. The implementation includes shortcut functions, <code>tt_solve_power</code> , <code>tt_ind_solve_power</code> , and <code>gof_chisquare</code> parameters of the power equations.
Getting started	
User Guide	
Background	TTestIndPower(**kws)
Regression and Linear Models	TTestPower(**kws)
Time Series Analysis	GofChisquarePower(**kws)
Other Models	NormalIndPower([ddof])
Statistics and Tools	FTestAnovaPower(**kws)
Statistics stats	FTestPower(**kws)
Contingency tables	normal_power_het(diff, nobs, alpha[, ...])
Multiple Imputation with Chained Equations	normal_sample_size_one_tail(diff, power, alpha)
Empirical Likelihood	
emlike	
Distributions	
Graphics	
Input-Output io lib	
Tools	
Working with Large Data	



Em R a maioria das funções estatísticas já é nativa, ou seja, não é necessário instalar e importar pacotes/bibliotecas.

Por que não usar apenas R?

May 2021	May 2020	Change	Programming Language	Ratings	Change
1	1		C	13.38%	-3.68%
2	3	⬆	Python	11.87%	+2.75%
3	2	⬇	Java	11.74%	-4.54%
4	4		C++	7.81%	+1.69%
5	5		C#	4.41%	+0.12%
6	6		Visual Basic	4.02%	-0.16%
7	7		JavaScript	2.45%	-0.23%
8	14	⬆	Assembly language	2.43%	+1.31%
9	8	⬇	PHP	1.86%	-0.63%
10	9	⬇	SQL	1.71%	-0.38%
11	15	⬆	Ruby	1.50%	+0.48%
12	17	⬆	Classic Visual Basic	1.41%	+0.53%
13	10	⬇	R	1.38%	-0.46%