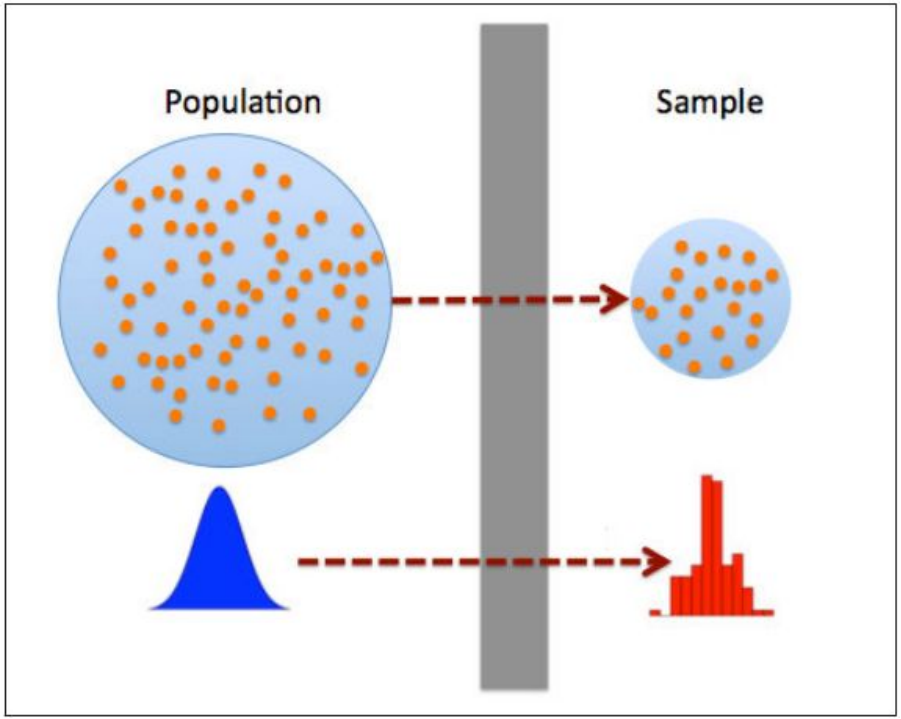
Big Data

Wellington Franco
Universidade Federal do Ceará – UFC
Campus de Crateús
wellington@crateus.ufc.br

Distribuição dos Dados

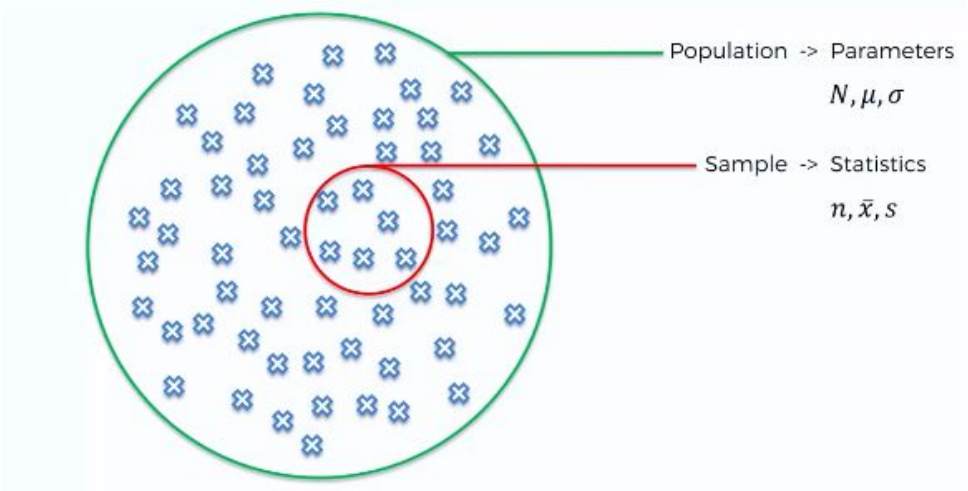
Amostra dos Dados



- O lado esquerdo representa uma **população** com sua distribuição desconhecida;
- A única coisa disponibilizada é uma pequena **amostra**, exibida no lado direito;
- Técnicas de amostragem ajudam a identificar uma distribuição a partir de uma amostra do total.

Por que estudar distribuição é importante?

Porque quando trabalhamos com subconjuntos de uma população (amostras), precisamos entender se suas amostras irão se comportar da mesma forma que o conjunto completo dos dados (população), para que não sejam gerados modelos de predição de forma equivocada.



Por que não usar o conjunto completo dos dados?

- A população (conjunto completo dos dados) pode ser muito grande:
  - Ex: Pesquisas eleitorais;
- A população pode ser desconhecida:
  - Ex: Identificar o comportamento dos sonegadores. Não conhecemos todos os sonegadores.

Definições-chave para uso em *Data Science*

- **População:** o total do *dataset*;
- **Amostra:** um subconjunto do *dataset*;
    - **Sinônimo:** *sample*
  - Isso será um ponto importante no futuro, pois os modelos preditivos são treinados utilizando um subconjunto dos dados disponíveis (conjunto de treinamento).
- **N :** o tamanho de uma população;
- ***n*:** o tamanho de uma amostra (*sample*);
- **Amostra aleatória:** amostra (suas instâncias) selecionada aleatoriamente;
- **Amostra estratificada:** dividir a população em *stratas* (grupos) e selecionar aleatoriamente amostras de cada *strata*.
    - *Stratified sample*.
- **Amostra *enviesada*:** Uma amostra que deturpa a população.
    - Sinônimo: *sample bias*

Observação: Qualidade Amostral

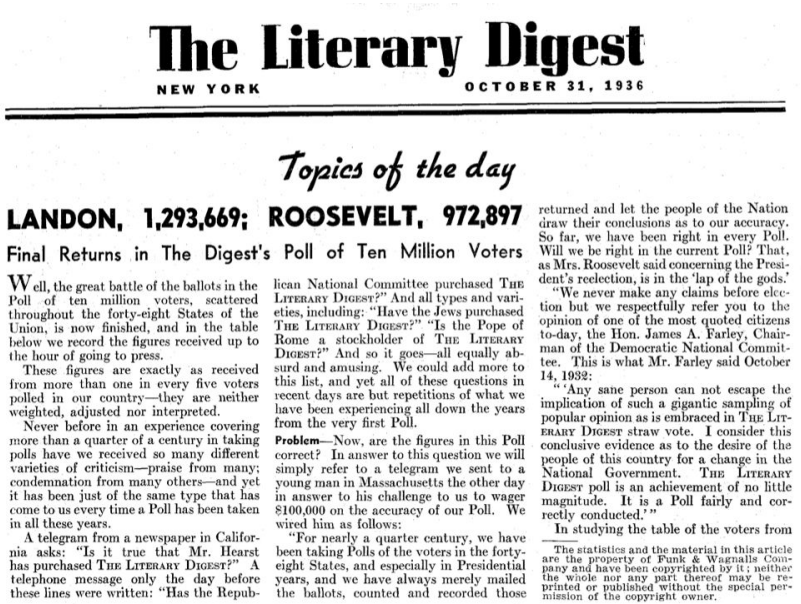
- **Amostragem aleatória:** processo pelo qual cada elemento da amostra, retirado da população, teve a mesma chance de ser selecionado.
- **Amostragem com reposição:** após o processo de seleção de uma amostra (subconjunto), as instâncias (itens) selecionadas para compor a amostra são devolvidas à população, para uma possível "*resseleção*".

Observação: Qualidade Amostral

- A qualidade do dado oferece muito mais informação ao modelo gerado em *Machine Learning* do que a quantidade de dados;
- Qualidade de dados em *Data Science* envolve:
  - Completude;
  - Consistência (formato);
  - Limpeza;

Exemplo de amostragem ruim (*enviesada)*

- Em 1936 a revista The *Literary Digest* previu a vitória de Al Langdon contra Franklin Roosevelt;
  - Eles levaram em conta em sua predição a opinião de seus 10 milhões de assinantes;
- George Gallup, dono de outra revista, entrevistou 2000 pessoas de outros nichos e previu a vitória de Roosevelt;
- Evidentemente, Roosevelt foi o eleito.
- A amostra da The *Literaty Digest* estava completamente enviesada (*bias*) por uma classe social.



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased THE LITERARY DIGEST?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased THE LITERARY DIGEST?" "Is the Pope of Rome a stockholder of THE LITERARY DIGEST?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in THE LITERARY DIGEST straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. THE LITERARY DIGEST poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

Cuidado: ERRO x BIAS

Em *Machine Learning*, quando geramos modelos (classificadores) eles contém erros devido à sua amostragem, mas também podem conter erros devido ao *bias*.

- O **erro** está associado à sua limitada representação dos dados referente à população.
  - O modelo não aprende ou não generaliza.
- O ***bias*** está associado à uma possível coleta ruim dos dados (como foi o caso da revista *The Literary Digest*).
  - O modelo aprendeu a partir de dados enviesados.

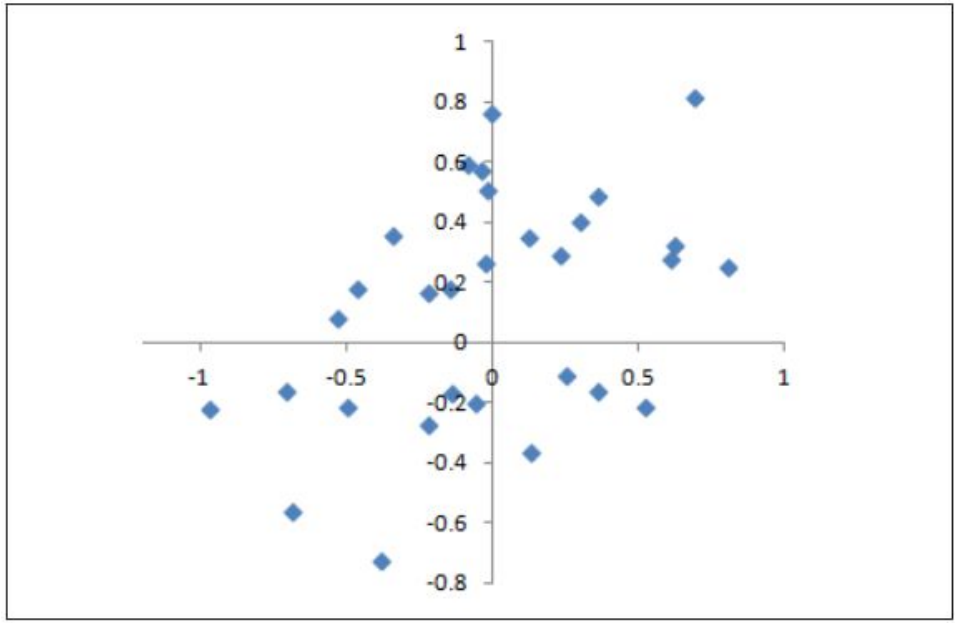
Exemplo (*Erro X Bias*)

Considere o processo de tiro ao alvo sob a tentativa de acertar o centro da mira (coordenadas 0,0);

Retirar uma amostra da quantidade de tiros disparados produz a seguinte imagem.

Quanto mais distante do centro, maior o **erro**.

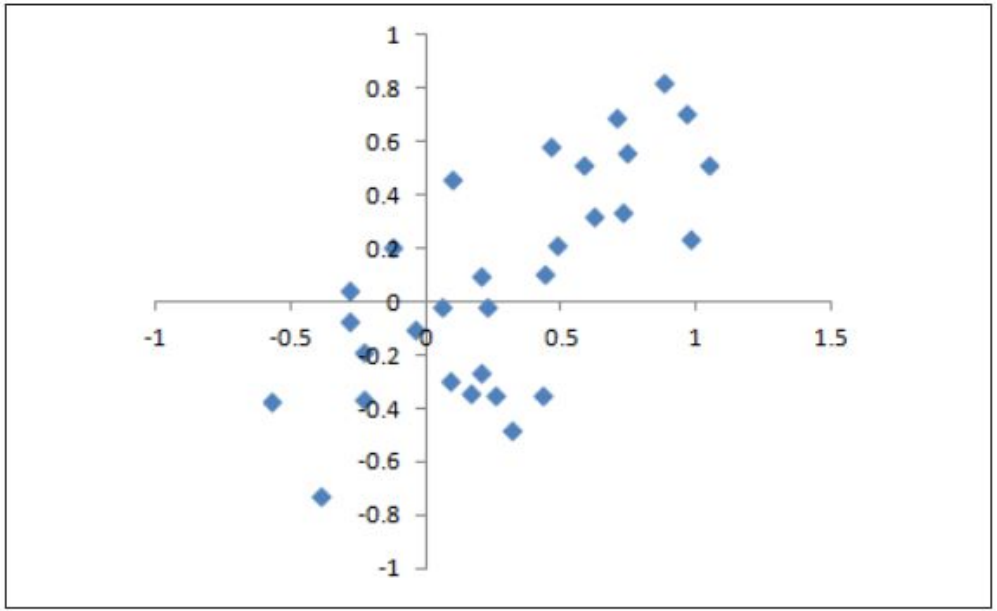
Várias retiradas aleatórias (com ou sem reposição) tenderá a mostrar esse comportamento geral.



Exemplo

Entretanto, como mostra essa imagem, o modelo ainda contém erros;

Contudo, poderíamos concluir somente baseado nessa amostra que os tiros tenderiam sempre a ser disparados para o quadrante superior direito.



Seleção Aleatória

- Embora *George Gallup* tenha entrevistado ¼ de pessoas comparada a *The Literary Digest*, o sucesso deu-se à **seleção aleatória** da população.


Today

THE NATIONAL WEEKLY POLL of PUBLIC OPINION

Next Sunday

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York is F.D.R. 'Sure' Column

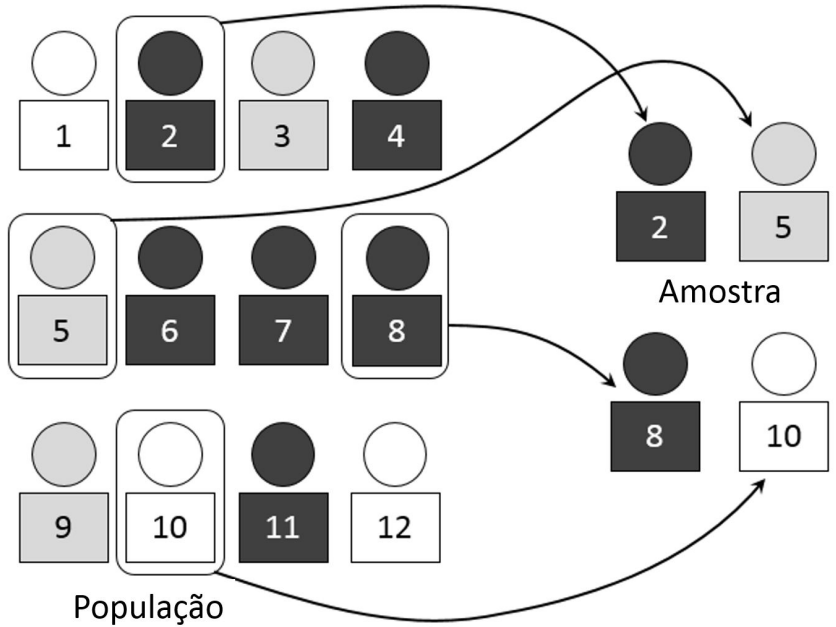
Election Forecast

Election Will Test Gauging Poll Methods

Seleção Aleatória: exemplo

A amostragem aleatória nem sempre é fácil.

Suponha que desejamos gerar uma amostra com um perfil representativo de clientes de um estabelecimento a fim de realizar uma pesquisa-piloto com esses clientes.



Seleção Aleatória: exemplo

- Primeiro, precisamos definir quem é o cliente:
  - Devemos selecionar apenas clientes com os registros de gastos com valor > 0?
    - Há casos em que clientes se cadastraram em um estabelecimento mas nunca consumiram nada.
  - Incluímos todos os clientes anteriores a um determinado período?
    - Um período pode influenciar diretamente na amostragem.
  - Incluímos reembolsos?
  - Incluímos revendedores?

Seleção Aleatória: exemplo

- Em seguida, precisamos especificar um procedimento de amostragem: pode ser, por exemplo, "*selecionar 100 clientes aleatoriamente*".
- Quando uma amostra de um fluxo está envolvida (por exemplo, transações com clientes em tempo real ou visitantes da Web), considerações de tempo podem ser importantes (por exemplo, um visitante da Web às 10h da manhã de um dia da semana pode ser diferente de um visitante da Web às 22h da noite de um final de semana).

Seleção Aleatória: exemplo

- Verificar se há grupos distintos (amostragem *estratificada*)
- Na amostragem *estratificada*, a população é dividida em grupos e são coletadas amostras aleatórias de cada grupo.
  - Os diferentes perfis de clientes podem mostrar comportamentos diferentes de acordo com região, classe social, distância etc.

Quantidade X Qualidade

O tempo e o esforço gastos em amostragem aleatória não apenas reduzem o viés, mas também permitem maior atenção à qualidade dos dados.

Por exemplo, dados ausentes e valores discrepantes podem conter informações úteis. Pode ser proibitivamente caro rastrear valores ausentes ou avaliar valores discrepantes em milhões de registros, mas fazê-lo em uma amostra de vários milhares de registros pode ser viável.

Quantidade X Qualidade: dados *esparsos*

Um cuidado adicional é necessário quando você tem dados esparsos (grande quantidade de valores zero na matriz).

Exemplo: Considere os dados dos usuários da Netflix, onde as colunas representam os filmes e as linhas indicam as notas dos usuários que assistiram esses filmes.

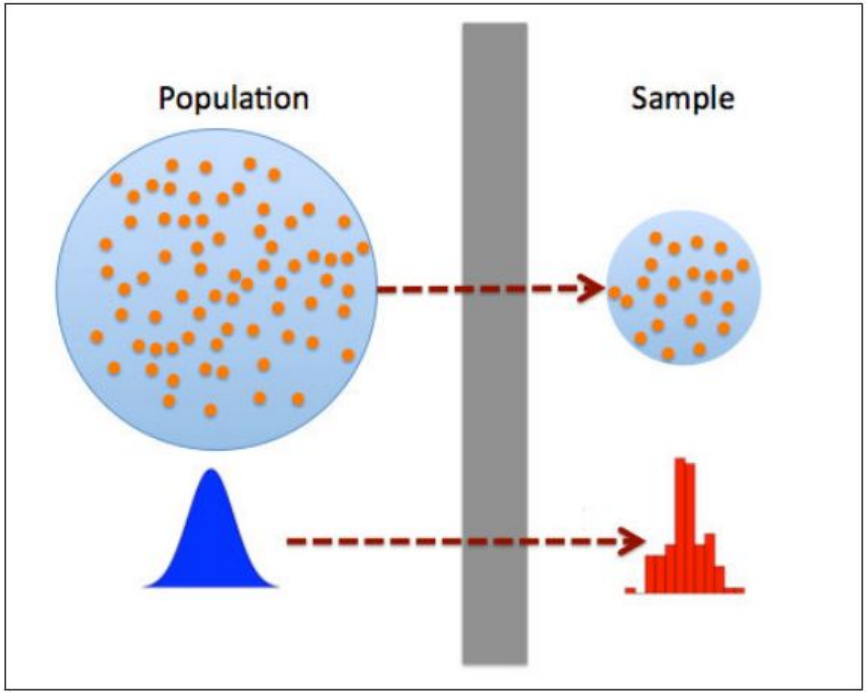
↓ Users / Movies →	Toy Story	Os Incríveis	O Poderoso Chefão	E.T.
João	5	4		3
Maria	2		4	5
Francisco		5	1	3
Joaquim	1		5	
Ana		2	4	4

Quantidade X Qualidade: dados *esparsos*

- A Netflix possui uma enorme quantidade de filmes.
- Isso gera uma enorme matriz, cuja grande maioria das entradas é "0".

Users	Movie A	Movie B	Movie C	Movie D
User 1	5		2	
User 2		4	3	
User 3		1	2	
User 4	3	5	1	
User 5	2			3

Média Populacional X Média Amostral



A média amostral é dada pelo símbolo: $\overline{x}$

(Também pode ser usado como **x-barra**).

A média populacional é dada pelo símbolo: **μ**

Data Snooping

Data Snooping é uma extensiva busca nos dados a fim de encontrar algum resultado interessante.

Cuidado: existe uma diferença entre o fenômeno de você descobrir algo vasculhando os dados e o fenômeno de torturar os dados até eles mostrarem o que você quer ver.

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Exemplo de amostragem aleatória

Imagine que você pede uma pessoa para jogar uma moeda e fazer com que ela dê ‘cara’ nos próximos 10 lançamentos.

Você a desafia (o equivalente a um experimento), e ela passa a lançar a moeda 10 vezes seguidas, e em todos os lançamentos o resultado é ‘cara’.

Claramente, você atribui algum talento especial a ela - a probabilidade de que 10 lançamentos de moedas dê cara por acaso é de 1 em 1024.

Exemplo de amostragem aleatória

Agora imagine que o locutor de um estádio esportivo solicite às 20.000 pessoas presentes que joguem uma moeda 10 vezes seguidas e se reporte a um fiscal se conseguirem 10 ‘caras’ consecutivas.

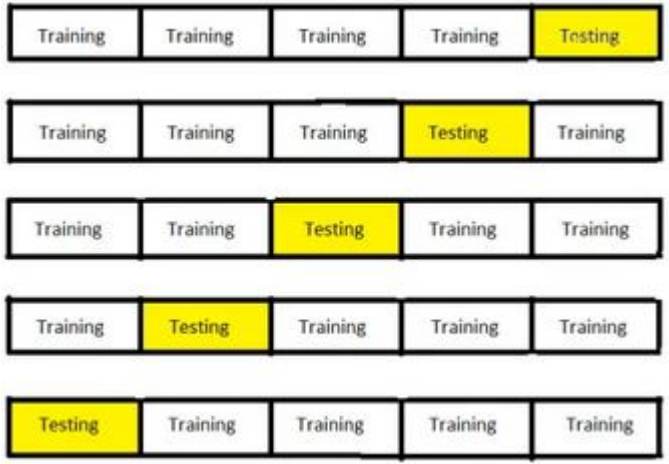
A chance de haver uma pessoa no estádio que consiga as 10 ‘caras’ consecutivas é mais alta.

Claramente, selecionar uma pessoa (ou mais) que obtiveram 10 ‘caras’ no estádio não indica que eles têm algum talento especial - é provavelmente sorte (*aleatoriedade*).

Observações

Isso mostra que é interessante testarmos a partir de um grande conjunto de dados várias amostras para tentar entender como funciona uma pergunta em diferentes amostragens do todo.

Em *Machine Learning*, é como se nós quiséssemos **generalizar** o modelo.



SPOILER: Cenas para os próximos módulos..

Distribuição Amostral

O termo distribuição amostral de uma estatística refere-se à distribuição amostral de uma - das muitas - amostras de uma população.

- **Amostra Estatística:** uma métrica calculada para uma amostra dos dados retirada de uma grande população;
- **Distribuição dos Dados:** A distribuição de frequência de valores individuais em um conjunto de dados.
- **Distribuição Amostral:** refere-se à distribuição de alguma amostra estatística sobre muitas amostras extraídas da mesma população.
    - Sinônimo: *sample statistic*
- **Teorema do Limite Central:** é a tendência de uma distribuição amostral seguir o formato normal a medida que a amostra aumenta.

Variabilidade Amostral

A **variabilidade amostral** (ou ***standard error***) de uma amostra estatística sob muitas amostras.

- Nosso objetivo como cientista de dados é coletar dados suficientes para representar a população do nosso problema de forma mais fiel possível;
- Como nosso conjunto de dado é uma amostra do todo, pode acontecer de haver "bias", ou seja, dada uma outra amostra da mesma população, a distribuição pode ser outra.

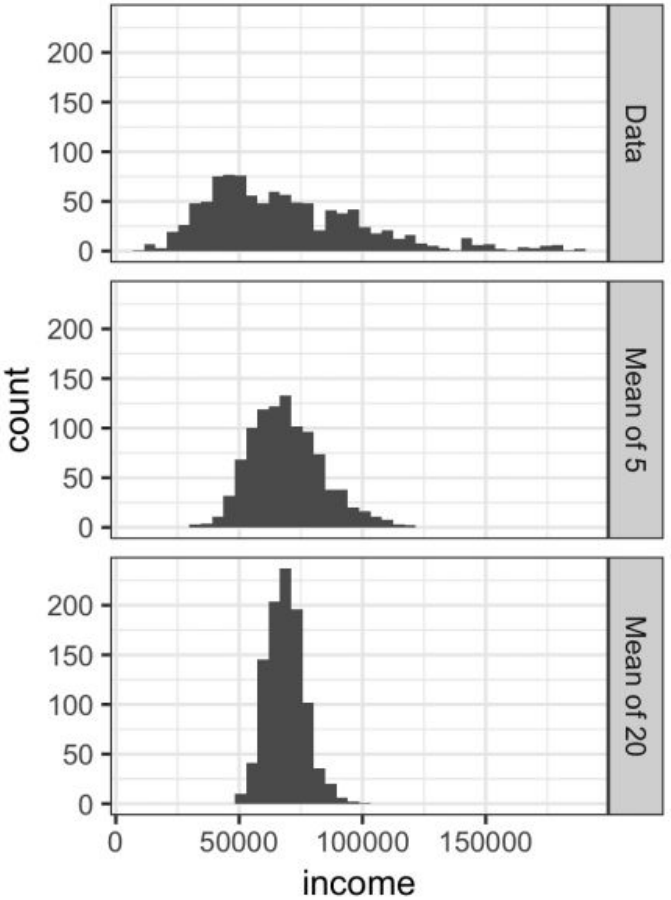
Variabilidade Amostral

A distribuição de uma amostra estatística, como a média, provavelmente será mais regular e em forma de sino (normal) do que a distribuição dos próprios dados em si.

Quanto maior a amostra em que a estatística se baseia, mais isso é verdade. Além disso, quanto maior a amostra, menor a variabilidade da amostra estatística.

Exemplo

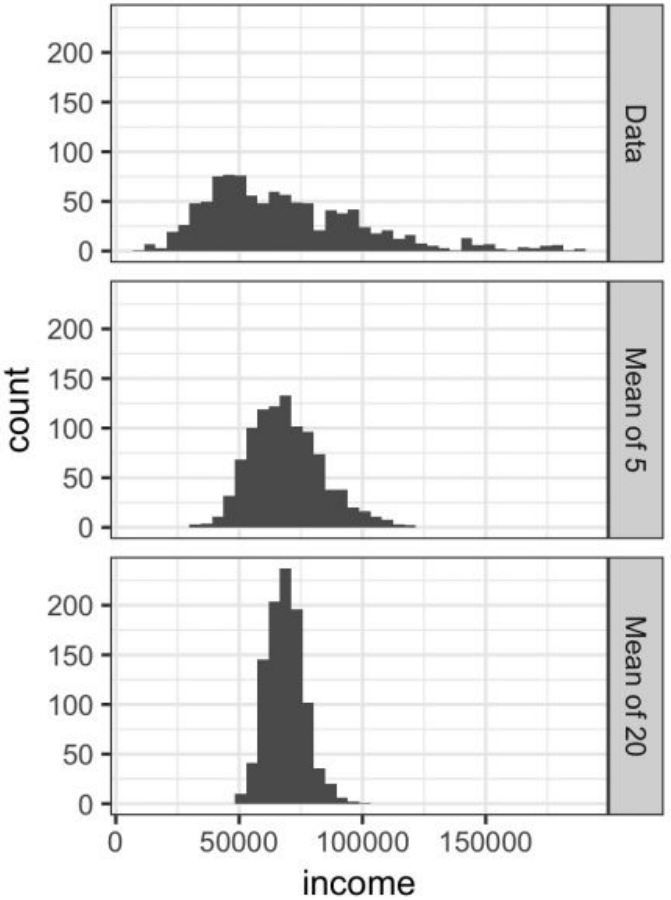
- Seja o seguinte cenário de uma renda anual para solicitantes de empréstimos a para um determinado local.
- Três amostras desses dados:
  - uma amostra de 1000 valores;
  - uma amostra com médias a cada 5 valores;
  - uma amostra com médias a cada 20 valores.



renda

Exemplo

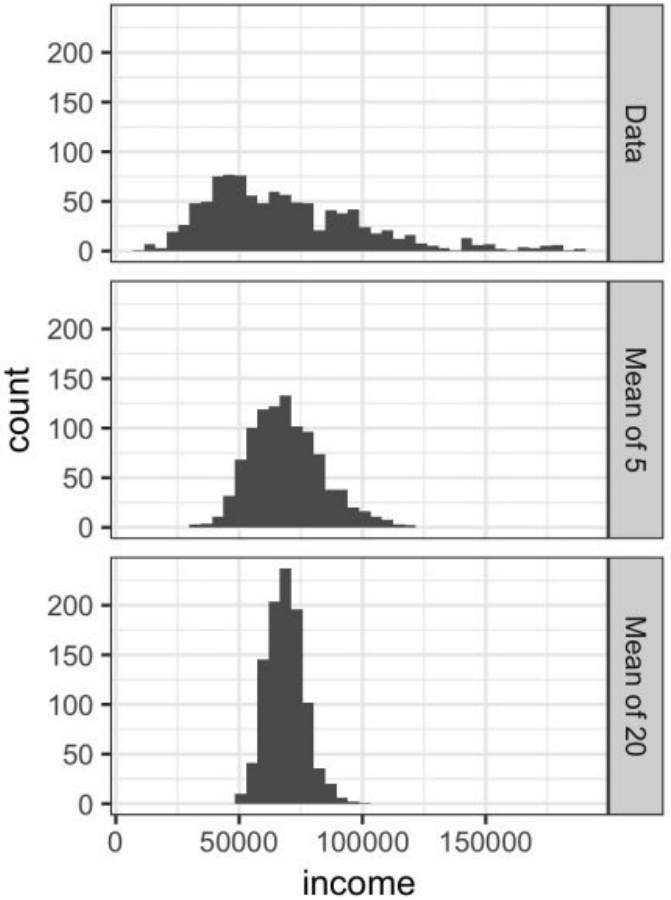
- O histograma dos valores de dados individuais é amplamente distribuído e inclinado para valores mais altos (cauda), como é de se esperar com os dados de renda.
- Os histogramas das médias de 5 e 20 são cada vez mais compactos e mais em forma de sino (normal).



renda

Exemplo

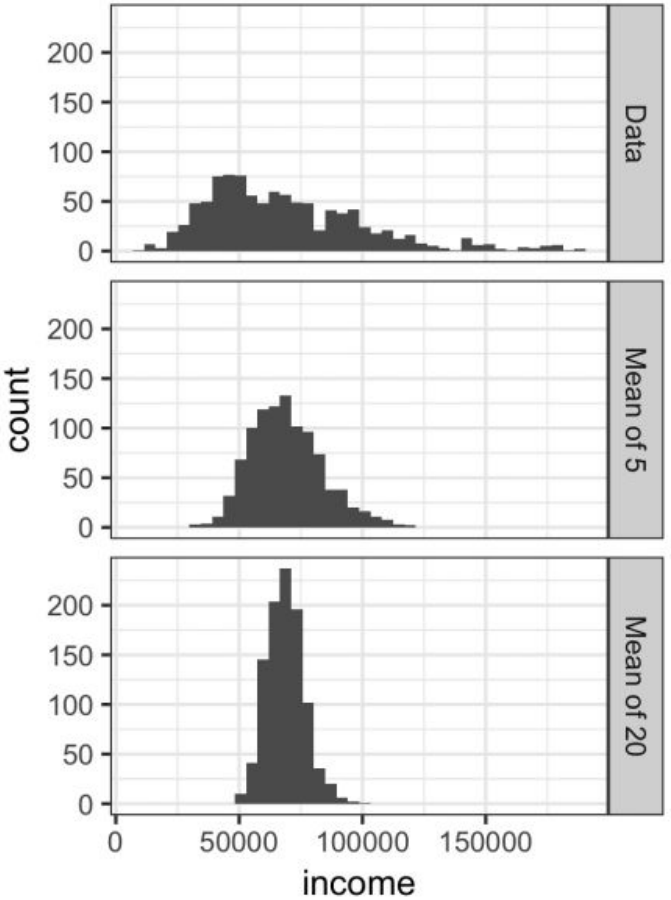
- Isso é causado graças a um fenômenos chamado teorema central do limite;
- Esse teorema diz que a média de múltiplas amostras tenderá a ter um formato de sino (normal), mesmo que a distribuição da população não seja normalmente distribuída.



renda

Exemplo

- O teorema central do limite é muito comum na estatística, mas na ciência de dados é pouco desempenhado devido a testes de hipóteses e intervalos de confiança;
- Uma das saídas para o cientista de dados é o ***bootstrap*** (mais adiante).



renda

Variabilidade (*Standard Error*)

- O ***standard error*** é uma métrica única que resume a variabilidade na distribuição amostral de uma estatística
- Algoritmo:
  a. Colete uma quantidade de amostras de uma população;
  b. Para cada amostra, calcule sua amostra estatística (por exemplo, a média);
  c. Calcule o desvio padrão de cada amostra da estatística calculada no passo anterior; use essa estimativa como *standard error*.
- Quanto maior o tamanho da amostra, menor será o valor do erro (visto que ela tenderá a variar menos da população).

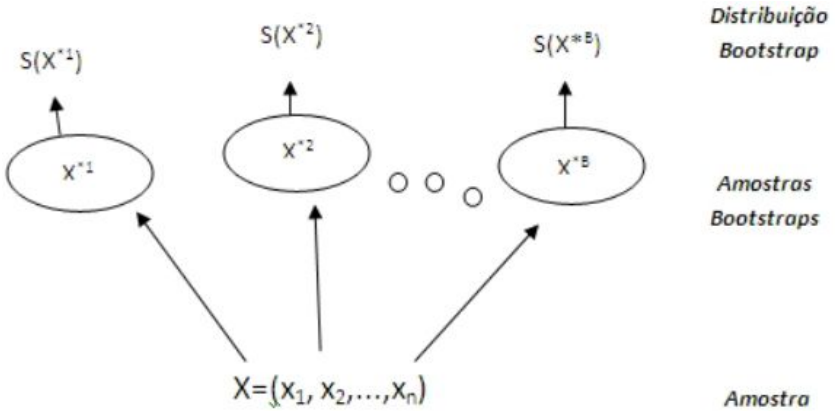
$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

Observações

- A distribuição de frequência de uma estatística de amostra nos diz como essa métrica seria diferente de amostra para amostra;
- Essa distribuição amostral pode ser estimada através do *bootstrap* ou de fórmulas que se baseiam no teorema do limite central;
- Uma métrica chave que resume a variabilidade de uma estatística de amostra é seu erro padrão;
- Em *Machine Learning* pode ser válido, pois dado um conjunto de dados extenso, basta gerar modelo preditivo usando um subconjunto desses dados.

Bootstrap

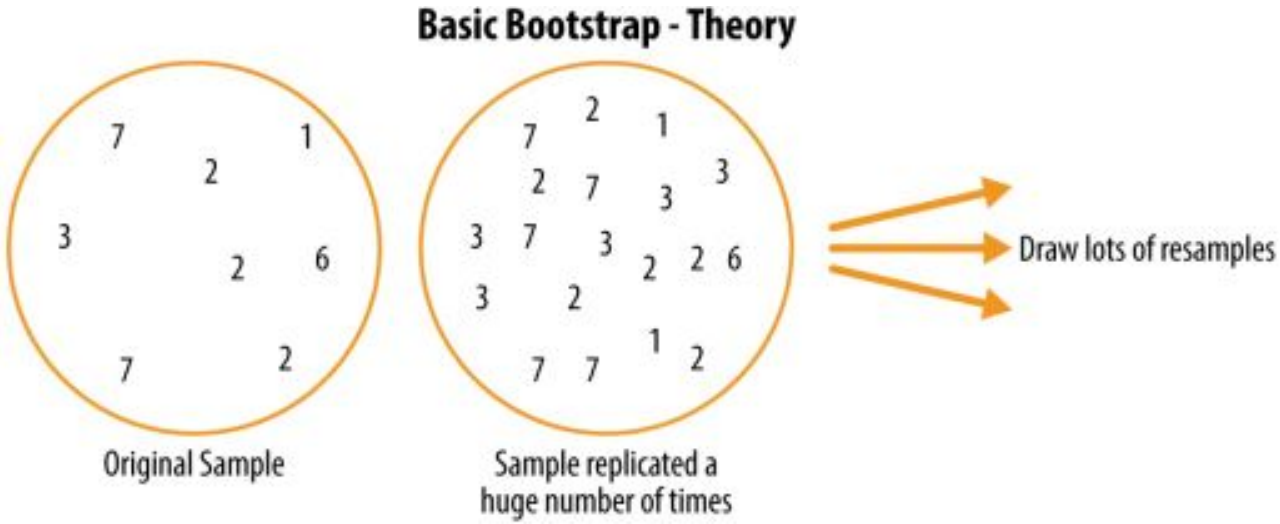
- Uma maneira fácil e eficaz de estimar a distribuição amostral de uma estatística é coletar amostras adicionais, com reposição da própria amostra, e recalcular a estatística amostral ou o modelo para cada nova amostra.
- Esse procedimento é chamado de ***bootstrap*** e não envolve necessariamente nenhuma suposição sobre a distribuição normal dos dados ou da estatística de amostra.



Bootstrap

- **Amostra de *bootstrap*:** uma amostra tirada com elementos que foram repostos no dataset observado;
    - Sinônimo: *Bootstrap sample*
- **Reamostragem:** é o processo de pegar repetidas amostras dos dados observados, incluindo ambos os procedimentos de *bootstrap* e permutação (*shuffling*);
- **Estimativa de Parâmetro:** é um método de estimativa de parâmetros para a população usando amostras. Um parâmetro é uma característica mensurável associada a uma população. Por exemplo, a altura média dos residentes em uma cidade, a contagem de glóbulos vermelhos etc.

Bootstrap

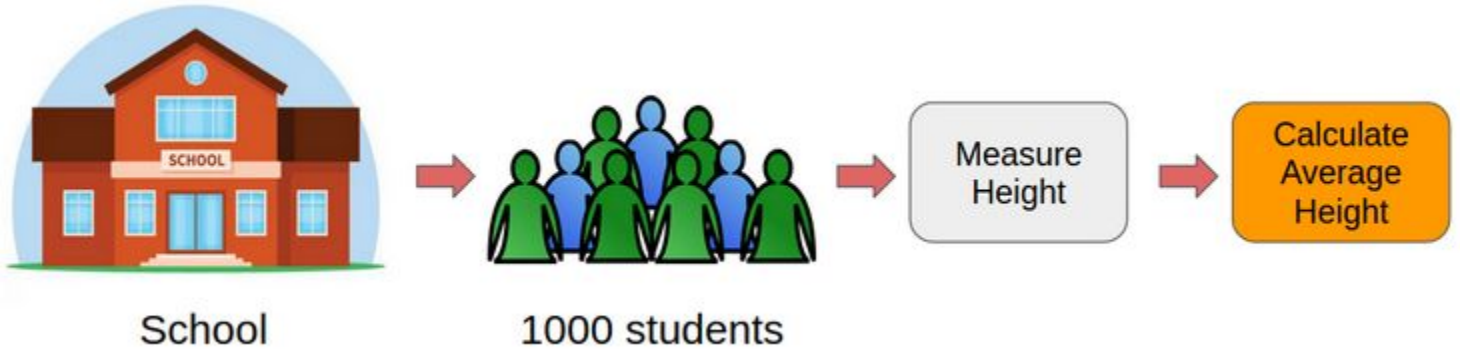


Na prática, não é necessário replicar a amostra muitas vezes. Simplesmente devolvemos cada observação após cada sorteio - fazemos uma amostra com a reposição. Desse modo, criamos efetivamente uma população infinita na qual a probabilidade de um elemento ser extraído permanece inalterada de extração a extração.

Por que *Bootstrap* é necessário?

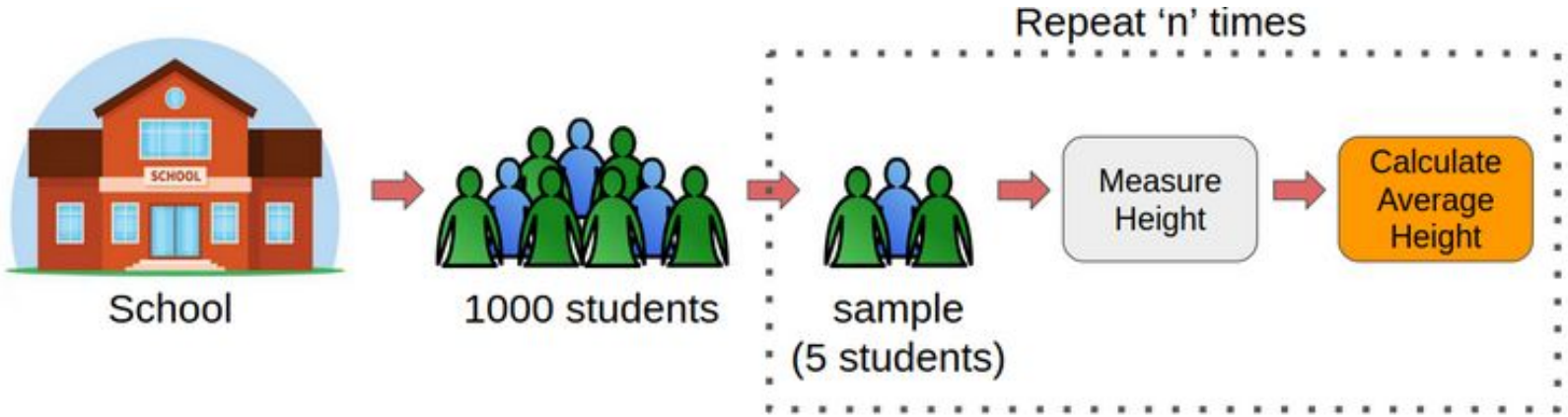
Suponha que queremos encontrar a média dos pesos de todos os (1000) alunos de uma escola (**estimativa de parâmetro**).

Obter o peso de cada aluno seria um trabalho que demandaria um certo tempo.



Por que *Bootstrap* é necessário?

Ao invés disso, pegam-se 20 grupos de 5 alunos (amostras) e coletamos o peso dos 100 alunos escolhidos. Ao final, tira a média de cada grupo e, por fim, a média geral. Isso será a média de todos os estudantes da escola.



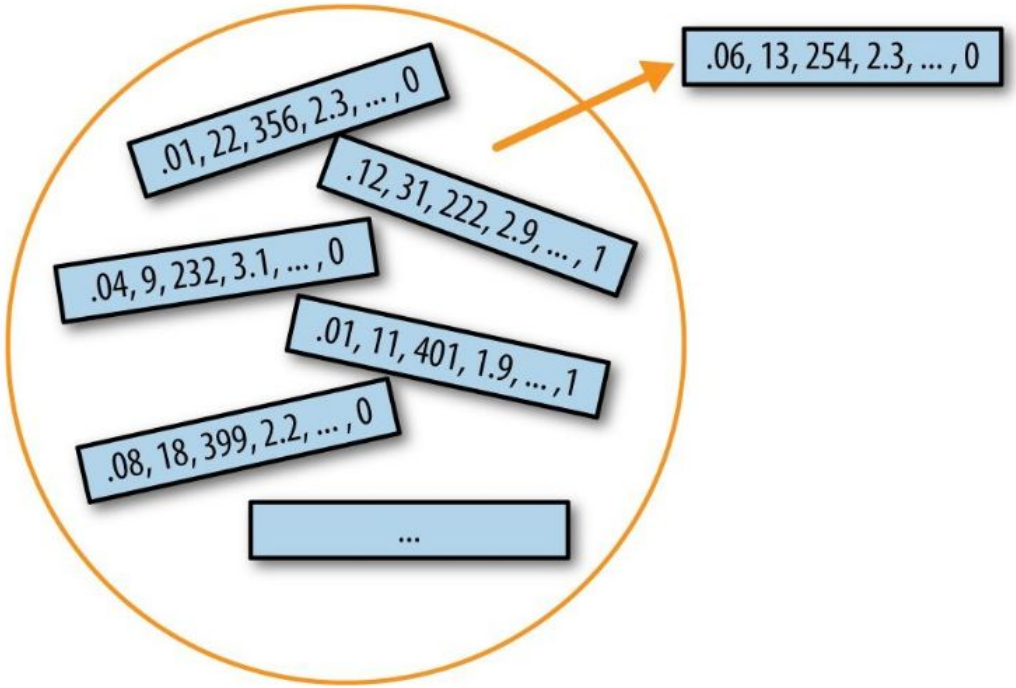
Algoritmo

1. Pegue um elemento da amostra, registre esse valor e reponha-o na amostra
2. Repita o passo anterior N vezes
3. Mantenha registrado a média dos N valores reamostrados
4. Repita os passos 1-3 B vezes
5. Use os B resultados para:
    a. Tirar a média das médias
    b. Calcular seus respectivos desvios padrão (isso estima o erro padrão amostral médio)
    c. produza um histograma ou boxplot

- B é o número de iterações de bootstrap - referente à quantidade de amostras que se deseja ter;
- Quanto mais iterações você fizer, mais preciso (*accurate*) será a estimativa do erro padrão ou do intervalo de confiança.

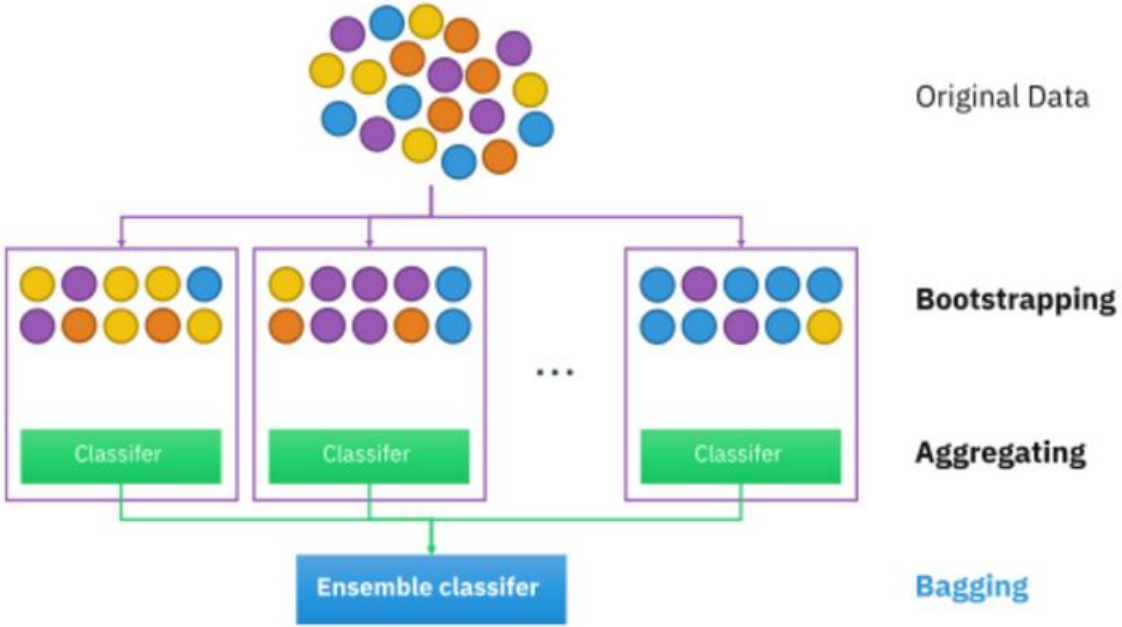
Bootstrap para Dados Multivariados

- Para dados multivariados, cada linha das amostras são vistas como uma unidade;



Bagging (*Bootstrap Aggregating*)

Em *Machine Learning*, o processo de *bootstrap* é conhecido como *bagging* (*bootstrap aggregating*) pode ser feito na hora da geração de modelos preditivos;



Bagging (*bootstrap aggregating*): exemplo

```
# normal distribution
x = np.random.normal(loc= 500.0, scale=1.0, size=10000)

np.mean(x)
```

← Gerando uma amostra Gaussiana (normal) de tamanho 10k com a média populacional = 500

Output: 500.00889503613934 ← Média populacional

```
sample_mean = []

# Bootstrap Sampling
for i in range(40):
  y = random.sample(x.tolist(), 5)
  avg = np.mean(y)

  sample_mean.append(avg)
```

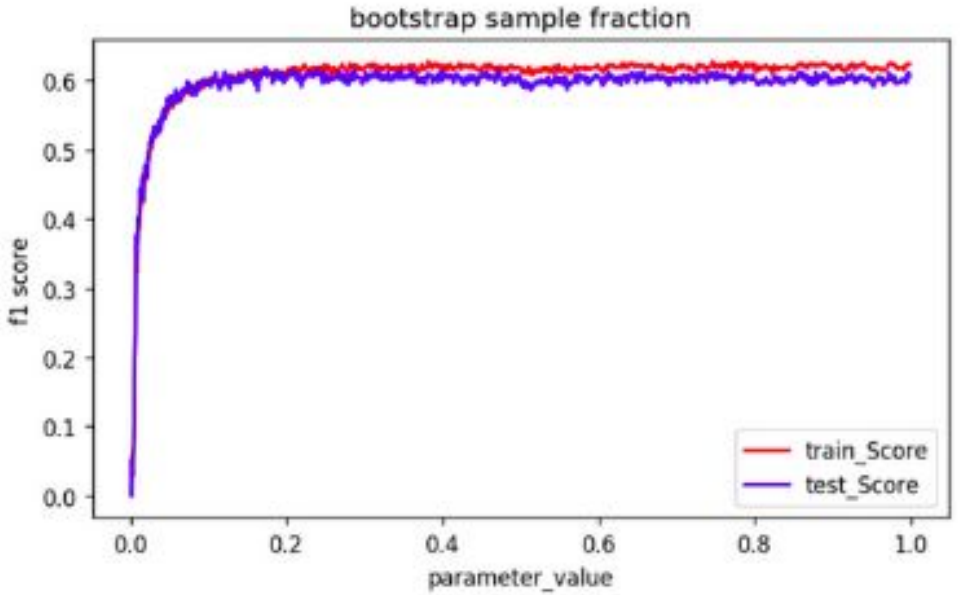
← Gerando 40 amostras de tamanho 5 da população e calculando a média de cada amostra

```
np.mean(sample_mean)
```

Output: 500.024133172629 ← Média dos valores das médias amostrais

Na prática

O gráfico ao lado mostra a métrica de performance *f1-score* (eixo y) sobre a proporção do tamanho da amostra sob a população (eixo x). É possível observar que em 20% da dos dados populacionais já geram ótimas amostras para classificadores igualmente como se estivessem usando a população inteira.

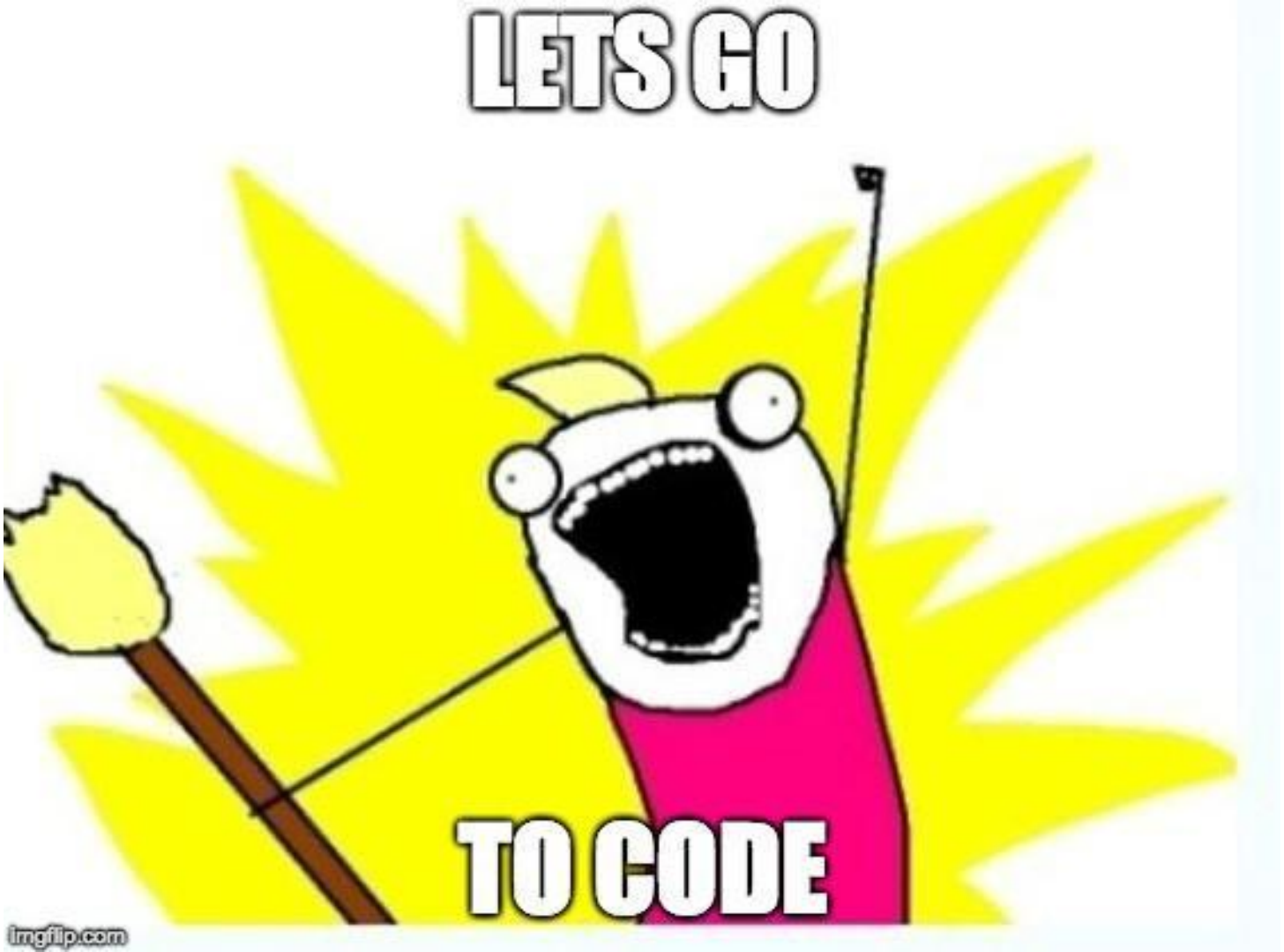


Observação

- O *Bootstrap* pode ser usado para determinar o tamanho da amostra - experimente diferentes valores para N para ver como a distribuição da amostra é afetada;
- O *Bootstrap* não compensa um tamanho pequeno de amostra - ele não cria novos dados nem preenche furos em um conjunto de dados existente. Apenas nos informa que muitas amostras adicionais teriam, quando extraídas de uma população como a nossa amostra original.

Conclusão de *Bootstraping*

- O *bootstrap* (ou seja, amostragem de dados com reposição de um dataset) é uma poderosa ferramenta para avaliar a variabilidade de uma amostra estatística;
- Ele nos permite estimar a distribuição amostral sem a necessidade de aproximações matemáticas desenvolvidas;
- Quando usados em modelos preditivos, agregar múltiplas amostras de *bootstrap* (*bagging*) pode superar o uso de um único modelo.


LETS GO
TO CODE
imgflip.com