Data Scientist - Curso

Manoel Teles

16 de maio de 2018

1 Índice

2 Estatística

2.1 Amostragem

Conceitos

População: Alvo do estudo.

Censo: Pesquisa com toda a população, pode ser caro ou impossível inferir sobre toda a população.

Enviesamento: Você subestima ou superestima o parâmetro da população.
causas: Pesquisa de pessoas próximas ou de fácil acesso, pesquisas pela internet, sem uso de mecanismo de seleção aleatório.

Amostra - Parte de uma população (Subconjunto da população), selecionada usando alguma técnica que de chances iguais a todos os elementos da população de serem selecionados.

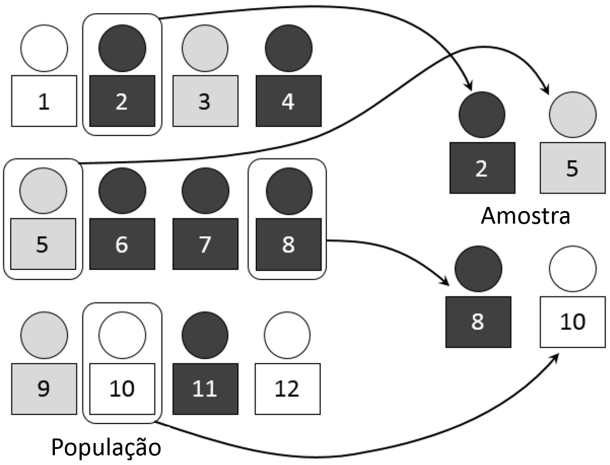
Uma amostra feita corretamente deve representar as mesmas características da população de onde foi retirada.

Se ela não representa a população, dizemos que ela é enviesada.

"Custo"da Amostra
Margem de erro e nível de confiança.
Variação: amostrar diferentes podem apresentar resultados diferentes.
podemos "medir"a variação esperada.

Principais tipos de amostras

Aleatória simples - Um determinado número de elementos é retirado da população de forma aleatória. Todos os elementos da população alvo do processo de amostragem, devem ter as mesmas chances de serem selecionado para fazer parte da amostra.



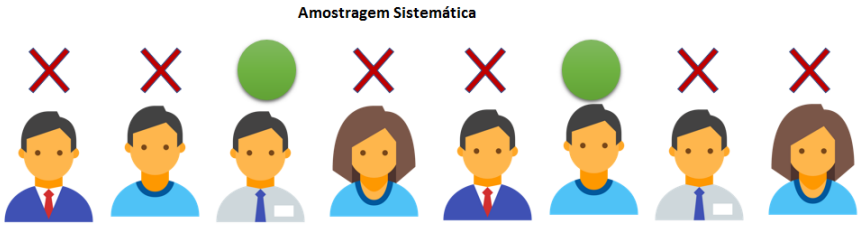
Há duas formas de trabalhar com amostra aleatória simples, com reposição e sem reposição.

Com reposição: Quer dizer que uma vez que o elemento da população é selecionado, ele volta a fazer parte da população e ele passa a ter as mesmas chances de ser selecionado novamente, exemplo(Exame de dopping em atletas das olimpíadas).

Sem reposição: Uma vez que o elemento da população é selecionado, ele não faz mas parte da população e não tem mais chance ser selecionado novamente, exemplo(Pesquisa de intensão de votos).

Estratificada - As vezes as populações estão divididas nos chamados estratos, características comuns que os elementos tem, podem ser relacionados a raça, escolaridade ou religião.

Sistemática - Neste tipo de amostragem, é escolhido um elemento aleatório, e a partir daí, a cada N elementos um novo membro é escolhido.



Por Unidade Monetária - Neste tipo de amostra é informada uma coluna de ordenação, e uma coluna numérica que sera utilizada para gerar os cálculos para produzir a amostra.

Amostragem por Unidade Monetária

Total de Registros: 50

Total de Débitos: R$ 108.465,00

Calcula o intervalo da amostra:
Valor Total / registros = 2169

Ordena por Clientes

Seleciona um número aleatório entre 1 e 2169: 1800 (Gerado no R)

Clientes	Débito	Cumulativo		
Austin College Austin College	R$ 701,00	R$ 701,00	!	
Algonquin College	R$ 1.200,00	R$ 1.901,00	R$ 3.969,30	(Número Aleatório + Int. Amostra)
Beijing Foreign Studies BFSU	R$ 453,00	R$ 2.354,00	!	
Blackboard Inc.	R$ 290,00	R$ 2.644,00	!	
Boise State	R$ 1.340,00	R$ 3.984,00	!	
Bowling Green State	R$ 245,00	R$ 4.229,00	!	
Blue Dots Consultancy Services B	R$ 780,00	R$ 5.009,00	R$ 6.138,60	(3.969,30 + Int.Amostra)
Brigham Young	R$ 1.456,00	R$ 6.465,00	!	
British Columbia Institute of Tech	R$ 820,00	R$ 7.285,00	R$ 8.307,90	(6.138,30 + Int.Amostra)
Campus Management	R$ 1.432,00	R$ 8.717,00	!	
CAESY Education Systems	R$ 2.344,00	R$ 11.061,00	R$ 10.477,20	(8.307,90 + Int.Amostra)
Carnegie Mellon Carnegie Mello	R$ 1.029,00	R$ 12.090,00	R$ 12.646,50	(10.477 + Int.Amostra)

2.2 Amostragem - R

2.2.1 Aleatória simples

Utilizando o R vamos demonstrar a amostragem utilizando o dataset Iris. Mais informações sobre o dataset: dataset iris

bibliotecas que devem ser importadas:

library(datasets) - importa os Datasets
data(iris) - importa o Dataset Iris
summary(iris) - produz o resultado sumarizado do dataset
dim(iris) - Retorna o conjunto da dimensão de um objeto

Para o exemplo vamos criar um modelo separando os dados do dataset IRIS em dois conjuntos.
Passos
1 - Gerar 150 números aleatórios, que poderão ser 0 ou 1 gerar 150 números aleatórios.
2 - Usar esses números aleatórios e dividi-los por números aleatórios.

Função sample é composta por 4 parametros, onde o primeiro parametro indica de onde ele vai buscar a amostra, neste caso ele ira escolher apenas 0 ou 1, o segundo parametro indica o tamanho da amostra, o terceiro(replace) indica que sera uma amostra com reposição e por ultimo o vetor de probabilidade indicando q cada numero tera 50de chance de ser escolhido.

```
> amostra = sample(c(0,1),150,replace = TRUE, prob = c(0.5,0.5))
> amostra
  [1] 1 0 1 1 1 0 0 1 0 1 1 1 1 1 1 0 1 0 1 0 1 0 0 1 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 1 1 1 0 0 1 0 1 1 0 0 1 1 0
 [62] 0 1 1 0 1 1 1 0 0 1 0 1 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0 0 0 1 0 0 1 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 0 0 1 0
[123] 0 1 1 1 0 0 1 1 0 1 0 1 1 0 1 1 0 1 0 0 0 1 1 0 0 0 0 1
```

Total da amostra

Total da Amostra que é igual a 1
length(amostra[amostra==1])

Total da Amostra que é igual a 0
length(amostra[amostra==0]

```
> length(amostra[amostra==1])
[1] 72
> length(amostra[amostra==0])
[1] 78
>
```

Repetir um experimento

Cria uma semente de aleatoriedade que permite repetir o experimento.
set.seed(2345)
sample(c(100),1)

```
> set.seed(2345)
> sample(c(100),1)
[1] 12
```

2.2.2 Estratificada

Para gerar uma amostra estratificada, vamos utilizar o Dataset iris.

O comando summary mostra as informações do Dataset de forma sumarizada

.

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Será necessário instalar o pacote sampling(Função para desenhos e calibração de amostras.) Package sampling

Carregar o pacote na memória.
library(sampling)

Gerar Estrato
Função strata : primeiro passa-se o conjunto de dados depois o vetor com as colunas(neste caso vamos utilizar apenas uma coluna(species)), e por fim o vetor com o tamanho de cada estrato.

amostrairis2 = strata(iris, c("Species"), size=c(25,25,25), method="srswor")

```
> summary(amostrairis2)
      Species      ID_unit          Prob        Stratum
 setosa    :25   Min.   :  2.0   Min.   :0.5   Min.   :1
 versicolor:25   1st Qu.: 40.0   1st Qu.:0.5   1st Qu.:1
 virginica :25   Median : 74.0   Median :0.5   Median :2
                 Mean   : 75.0   Mean   :0.5   Mean   :2
                 3rd Qu.:113.5   3rd Qu.:0.5   3rd Qu.:3
                 Max.   :149.0   Max.   :0.5   Max.   :3
```

.

Caso desejarmos gerar uma amostra estratificada em que haja uma proporção de elementos. para isso utilizaremos o Dataset Infert.

```
> summary(infert)
   education       age            parity         induced            case          spontaneous
 0-5yrs : 12   Min.   :21.00   Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 6-11yrs:120   1st Qu.:28.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 12+ yrs:116   Median :31.00   Median :2.000   Median :0.0000   Median :0.0000   Median :0.0000
               Mean   :31.50   Mean   :2.093   Mean   :0.5726   Mean   :0.3347   Mean   :0.5766
               3rd Qu.:35.25   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
               Max.   :44.00   Max.   :6.000   Max.   :2.0000   Max.   :1.0000   Max.   :2.0000
    stratum      pooled.stratum
 Min.   : 1.00   Min.   : 1.00
 1st Qu.:21.00   1st Qu.:19.00
 Median :42.00   Median :36.00
 Mean   :41.87   Mean   :33.58
 3rd Qu.:62.25   3rd Qu.:48.25
 Max.   :83.00   Max.   :63.00
```

Precisamos gerar uma amostra com 100 elementos, portanto precisamos calcular o numero de elementos do estrato dividido pela população versus o tamanho da amostra.

Entre 6 e 11 anos temos 120.
length(infert[infert==’6-11yrs’])

```
> length(infert[infert=='6-11yrs'])
[1] 120
```

Numero de elementos do Dataset.
length(infert$education)

```
> length(infert$education)
[1] 248
```

Tamanho da amostra que desejo gerar.
100

Cálculo.
120 / 248 * 100

```
> 120 / 248 * 100
[1] 48.3871
```

Resultados
unique(infert$education)
round(length(infert[infert==’0-5yrs’])/length(infert$education)*100)
round(length(infert[infert==’6-11yrs’])/length(infert$education)*100)
round(length(infert[infert==’12+ yrs’])/length(infert$education)*100)

```
> unique(infert$education)
[1] 0-5yrs  6-11yrs 12+ yrs
Levels: 0-5yrs 6-11yrs 12+ yrs
> round(length(infert[infert=='0-5yrs'])/length(infert$education)*100)
[1] 5
> round(length(infert[infert=='6-11yrs'])/length(infert$education)*100)
[1] 48
> round(length(infert[infert=='12+ yrs'])/length(infert$education)*100)
[1] 47
```

Utilizando a função strata.
’0-5yrs’ = round(length(infert[infert==’0-5yrs’])/length(infert$education)*100)
’6-11yrs’ = round(length(infert[infert==’6-11yrs’])/length(infert$education)*100)
’12+ yrs’ = round(length(infert[infert==’12+ yrs’])/length(infert$education)*100)
amostra = strata(infert, c(”education”), size=c(‘0-5yrs‘,‘6-11yrs‘,‘12+ yrs‘),
method=”srswor”)
summary(amostra)

```
> unique(infert$education)
[1] 0-5yrs  6-11yrs 12+ yrs
Levels: 0-5yrs 6-11yrs 12+ yrs
> round(length(infert[infert=='0-5yrs'])/length(infert$education)*100)
[1] 5
> round(length(infert[infert=='6-11yrs'])/length(infert$education)*100)
[1] 48
> round(length(infert[infert=='12+ yrs'])/length(infert$education)*100)
[1] 47
```

O método srswor - método padrão da função strata que gera uma amostra aleatória sem reposição

2.2.3 Sistemática

Para estudarmos a amostragem sistemática, precisamos instalar o pacote TeachingSampling.
install.packages("TeachingSampling")
Depois carregar o pacote na memória.
library(TeachingSampling)

TeachingSampling vai gerar números aleatórios que poderão ser utilizados pra fazer a amostra, então vamos gerar uma amostragem aleatória sistemática no conjunto de dados iris e essa amostra sistemática pega uma instancia do conjunto de dados iris a cada dez.

Vai gerar uma um numero aleatório a cada dez registros.
amostra = S.SY(150, 10)

```
> amostra = S.SY(150, 10)
> amostra
      [,1]
 [1,]    9
 [2,]   19
 [3,]   29
 [4,]   39
 [5,]   49
 [6,]   59
 [7,]   69
 [8,]   79
 [9,]   89
[10,]   99
[11,]  109
[12,]  119
[13,]  129
[14,]  139
[15,]  149
```

Utilizando a amostra para selecionar os dados do Dataset Iris.

```
> amostrairis = iris[amostra,]
> amostrairis
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
9            4.4         2.9          1.4         0.2     setosa
19           5.7         3.8          1.7         0.3     setosa
29           5.2         3.4          1.4         0.2     setosa
39           4.4         3.0          1.3         0.2     setosa
49           5.3         3.7          1.5         0.2     setosa
59           6.6         2.9          4.6         1.3 versicolor
69           6.2         2.2          4.5         1.5 versicolor
79           6.0         2.9          4.5         1.5 versicolor
89           5.6         3.0          4.1         1.3 versicolor
99           5.1         2.5          3.0         1.1 versicolor
109          6.7         2.5          5.8         1.8  virginica
119          7.7         2.6          6.9         2.3  virginica
129          6.4         2.8          5.6         2.1  virginica
139          6.0         3.0          4.8         1.8  virginica
149          6.2         3.4          5.4         2.3  virginica
```

2.3 Medidas de Centralidade e Variabilidade

2.3.1 Centralidade

Média - soma dos valores dos dados de um conjunto dividido pelo número de dados (elementos) constante nesse conjunto.

$$\mu = \frac{\sum x}{N}$$

$$\overline{X} = \frac{\sum x}{n}$$

40.000
18.000
12.000
250.000
30.000
140.000
300.000
40.000
800.000
181.111

Moda - É o valor mais frequente num conjunto de dados.

Mediana - É o valor que medeia os valores presentes num conjunto ordenado numericamente.

	Impar		Par	
40.000	1	12.000	1	12.000
18.000	2	18.000	2	18.000
12.000	3	30.000	3	30.000
250.000	4	40.000	4	40.000
30.000	5	40.000	5	40.000
140.000	6	140.000	6	40.000
300.000	7	250.000	7	140.000
40.000	8	300.000	8	250.000
800.000	9	800.000	9	300.000
			10	800.000

n = Total da amostra	(n+1)/2	n/2	n/2 + 1
	(9+1)/2 = 5	10/2 = 5	10/2 + 1 = 5 + 1 = 6

Desvio Padrão - Indica o grau de variação de um conjunto de elementos.
Como Calcular:

Passo 1 - Calcular Média

$$\overline{X} = \frac{\sum x}{n}$$

12 18 30 40 40 140 250 300 800

9

$$\overline{X} = 181$$

Passo 2 - Calcular Variância

- s^2 → variância amostral
- σ^2 → variância populacional

$$s^2 = \frac{\sum (x-\overline{x})^2}{n-1}$$

$$\sigma^2 = \frac{\sum (x-\overline{x})^2}{N}$$

$$s^2 = \frac{\sum (x-\overline{x})^2}{n-1}$$

x	$\overline{x}$	$x-\overline{x}$	$(x-\overline{x})^2$
12	-169	28.561	
18	-163	26.569	
30	-151	22.801	
40	-141	19.881	
40	-141	19.881	
140	-41	1.681	
250	69	4.761	
300	119	14.161	
800	619	383.161	/ 521.457

$$s^2 = \frac{28.561 + 26.569 + 22.801 + 19.881 + 19.881 + 1.681 + 4761 + 14.161 + 383.161}{8} = \frac{521.457}{8}$$

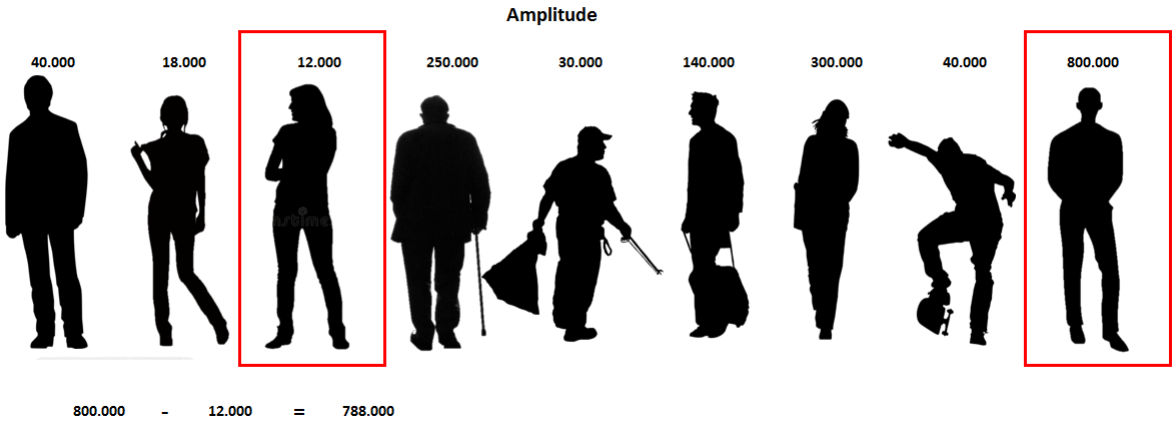
$$s^2 = 65.182$$

Passo 3 - Calcular Desvio Padrão

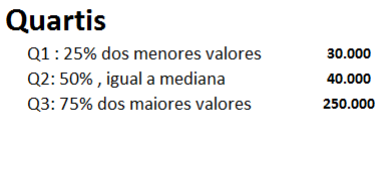
$$s = \sqrt{s^2}$$

$$s = 255$$

Amplitude - Em estatística, a amplitude representa a diferença entre o maior e o menor valor de um conjunto de dados.



Quartis (Q1, Q2 e Q3): São valores dados a partir do conjunto de observações ordenado em ordem crescente, que dividem a distribuição em quatro partes iguais. O primeiro quartil, Q1, é o número que deixa 25% das observações abaixo e 75% acima, enquanto que o terceiro quartil, Q3, deixa 75% das observações abaixo e 25% acima. Já Q2 é a mediana, deixa 50% das observações abaixo e 50% das observações acima. Mais informações sobre quartil Quartis



2.3.2 Centralidade - R

Para realizarmos os testes vamos criar a variável jogadores, com os mesmos salários da imagens.

Variável
jogadores = c(40000, 18000, 12000, 250000, 30000, 140000, 300000, 40000, 800000)

Calcula a média
mean(jogadores)

Calcula a mediana
median(jogadores)

Calcula os Quartis
quartis = quantile(jogadores)

ver os quartis
quartis

ver o terceiro quartil
quartis[4]

Calcula o Desvio Padrão
sd(jogadores)

Mostra o resultado da variável
summary(jogadores)
summary(jogadores)

Resultados:

```
> jogadores = c(40000, 18000, 12000, 250000, 30000, 140000, 300000, 40000, 800000)
> mean(jogadores)
[1] 181111.1
> median(jogadores)
[1] 40000
> quartis = quantile(jogadores)
> quartis
    0%    25%    50%    75%   100%
 12000  30000  40000 250000 800000
> quartis[4]
   75%
250000
> sd(jogadores)
[1] 255307.9
> summary(jogadores)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12000   30000   40000  181111  250000  800000
```

Estatística	Amostra	População
Média	$\bar{X}$	μ
Desvio Padrão	S	σ