Universidade Federal Rural de Pernambuco - Curso de Ciência da Computação

Estatística Exploratória - Prof. Cláudio Cristino - 25/04/2022

Amostragem Estatística

Uma Estatística é uma característica da amostra, pois se `X1, X2, X3, ... Xn` se trata de uma amostra, logo uma função (`f(X1, X2, X3, ..., Xn)`) é uma estatística determinada por tal amostra e que assume um valor determinado. Logo se queremos determinar valores desconhecidos de uma dada população, a estatística se torna uma opção viável, visto que, por muitas vezes analisar todo o conjunto populacional pode ser muito custoso, por exemplo, se queremos analisar todas as pessoas em situação de rua no Brasil. Veja que, esse exemplo não é nada prático, pois não há dados públicos acerca desse público e seria necessário procurar todas as pessoas nessa situação em todo o Brasil, algo inevitavelmente inviável e bastante custoso.

Nesse momento, a Estatística visa "estimar" os valores de uma distribuição, ou características de uma população por meio de uma amostra, dessa forma, os valores estimados são altamente dependentes dos dados de nossa amostra. A amostragem consiste do processo para selecionar os elementos da população, que vão constituir a amostra [1]. Em si, nossa amostra, se trata de um grupo de interesse, que é um subconjunto da população e que também possui uma característica de interesse. Exemplo:

- Característica: Peso dos estudantes da UFRPE;
- Amostra: 100 estudantes selecionados **ao acaso**.

Vale lembrar que não existe nenhuma técnica estatística capaz de salvar uma amostra mal coletada [2], pois ao aplicar as técnicas estatísticas estamos observando um subgrupo específico e tais elementos da amostra possuem características intrínsecas a população, logo erros grosseiros seriam cometidos e os resultados finais serão provavelmente bastante incorretos [2].

Há dois tipos de amostras comuns de amostragem: A Amostragem probabilística e a Amostragem não-probabilística. A amostragem probabilística constitui da amostragem onde todos os elementos da população possuírem uma probabilidade conhecida, e diferente de zero, de pertencer à amostra; do contrário a amostra será não-probabilística [2].

Amostragem probabilística

A amostragem probabilística pode ser dividida em diversas técnicas de amostragem, que são as técnicas de amostragem casual simples, sistemática, por meio de conglomerados, estratificada e múltipla. Vejamos:

Amostragem casual simples

Se trata da amostragem em que todas as possíveis amostras de tamanho `n` tem a mesma chance de serem escolhidas (de uma população com `N` elementos). Por exemplo:

- Selecionar 10 estudantes de uma sala por sorteio e perguntar seu peso;
- Gerar uma amostra aleatória de 1000 estudantes da UFRPE e perguntar seu peso.

Trata-se do método mais simples para selecionar uma amostra probabilística de uma população, que também serve de base para outros procedimentos amostrais, planejamento de experimentos e estudos observacionais.

Amostragem sistemática

A amostragem sistemática consiste de obter elementos dispostos de maneira organizada e **aleatória** (ex. fila, lista), do qual, escolhe-se um ponto de partida e seleciona-se um elemento a cada `n` elementos da população em fila. O exemplo mais comum é numa linha de produção, selecionar-se-à um produto à cada 500 produtos produzidos, do qual os produtos selecionados seriam utilizados para teste.

Sua principal vantagem é a facilidade em na determinação dos elementos da amostra, porém para adotá-la é necessário manter o fator de **aleatoriedade** na ordenação dos elementos, pois poderá ocorrer ciclos de variação em variáveis de interesse, do qual pode prejudicar a fidelidade da amostra com a população.

Amostragem por conglomerados

Se uma população pode ser constituída por conglomerados, por exemplo, bairros de uma cidade ou lotes de fabricação; poderemos selecionar conglomerados aleatoriamente para a nossa amostra, constituindo, dessa forma, a amostragem por conglomerados, onde todos os elementos do conglomerado selecionado fazem parte de nossa amostra.

Amostragem Estratificada

Uma amostragem estratificada, é uma amostra que é indicada por uma população estratificada, ou seja que está dividida em grupos distintos, denominados **estratos**.

Observe que, por exemplo, você deseja observar a comunidade institucional da UFRPE, tal comunidade é constituída por estudantes, técnicos, professores e terceirizados, muitas vezes uma amostragem casual simples pode não representar devidos *extratos* importantes para a comunidade institucional, pois a probabilidade de representar estudantes é muito maior do que os terceirizados, devido ao valor maior do número de estudantes; logo é notório observar que desejamos representar todos os extratos de forma homogênea, constituindo toda a comunidade universitária da UFRPE. Pretende-se assim otimizar a obtenção de informações sobre a população, com base no principio de que, onde a variação é menor, menos elementos são necessários para bem caracterizar o comportamento da variável [2].

Amostragem múltipla

A amostragem múltipla é um tipo especial de amostragem, pois consiste a amostra é realizada em diversas etapas sucessivas. O objetivo é obter uma conclusão de se deve ou não aceitar uma dada hipótese, pois dependendo dos resultados observados, etapas suplementares podem ser dispensadas.

Amostragem não-probabilística

Na impossibilidade de ser obter amostra probabilísticas ou por vias de simplicidade, a amostragem não-probabilística pode ser uma alternativa quando queremos estudar uma população. Em muitos casos, o efeito da utilização de uma amostragem não-probabilística podem ser considerados equivalentes aos de uma amostragem probabilística [2], resultando na importância dos processos de amostragem não-probabilística. Vejamos:

Amostragem sem norma

Se trata de uma amostragem, no qual o amostrador busca ser aleatório, porém não se utiliza de um instrumento ou método que garanta a aleatoriedade do processo de amostragem. Em geral, quando o amostrador escolhe uma população homogênea, os resultados são semelhantes aos da amostragem probabilística, porém qualquer tipo a influência do amostrador interfere no processo e poderá comprometer os resultados.

Amostragem por Julgamento

A amostra por julgamento, ou também conhecida como amostra intencional, é quando o amostrador escolhe deliberadamente os elementos da amostra; por julgar necessário para participar de uma amostra que represente uma população qualquer. Óbvio que esse tipo de amostragem é mais perigosa se comparada a outras amostras, pois pode equivocar um resultado já esperado ou previsível no processo de escolha de elementos da amostra, mas que por sua vez não representam a população como um todo.

Observando amostras da flor `Iris`

Uma das databases mais conhecidas na literatura de reconhecimento de padrões é a database `iris`. Apresentada em um trabalho clássico de [3] referenciado até o dia de hoje. O Data Set contém uma amostra de 150 instâncias do tipo de planta Íris. O Data Set é dividido em 3 grandes classe, porém para essa análise queremos observar a amostra de Íris.

In [ ]:
```
import pandas as pd
import numpy as np
from sklearn import datasets

iris = datasets.load_iris()

df = pd.DataFrame(data=np.c_[iris['data']], columns=iris['feature_names'])
```

In [ ]:
```
df.head()
```

Out[ ]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

`iris` é um Data Set que contém características de flores Íris em relação a dimensão das suas flores, logo duas características são cruciais ao identificar uma flor Iris, a sua sépala (*sepal*) e a pétala (*petal*).

Vamos estudar montar diversas **amostras casual simples** e tirar conclusões acerca de nossas amostras.

In [ ]:
```
samples = []

[samples.append(df.sample(n=50)) for _ in range(5)]
```

Out[ ]:
```
[None, None, None, None, None]
```

Agora, podemos analisar algumas informações interessante sobre nossas amostras!

In [ ]:
```
samples[0].describe()
```

Out[ ]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	50.000000	50.000000	50.000000	50.000000
mean	5.722000	3.124000	3.404000	1.058000
std	0.846672	0.369534	1.878445	0.820177
min	4.400000	2.400000	1.200000	0.100000
25%	5.025000	2.900000	1.500000	0.200000
50%	5.650000	3.100000	3.850000	1.150000
75%	6.300000	3.400000	5.000000	1.775000
max	7.700000	4.000000	6.900000	2.500000

Nessa primeira amostra, podemos observar valores importantes como a média amostral (`mean`) e também o desvio padrão (`std`) da amostra.

In [ ]:
```
samples[1].describe()
```

Out[ ]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	50.000000	50.000000	50.000000	50.000000
mean	5.840000	3.100000	3.688000	1.188000
std	0.886405	0.479796	1.872861	0.812565
min	4.400000	2.000000	1.200000	0.100000
25%	5.100000	2.800000	1.500000	0.300000
50%	5.700000	3.100000	4.300000	1.300000
75%	6.575000	3.400000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Nessa segunda amostra, podemos comparar com a amostra anterior e verificamos semelhanças com o grupo de dados anterior, principalmente na média amostral e o desvio padrão amostral.

In [ ]:
```
samples[2].describe()
```

Out[ ]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	50.000000	50.000000	50.00000	50.000000
mean	5.928000	3.024000	3.97400	1.280000
std	0.719989	0.408861	1.55259	0.711996
min	4.600000	2.200000	1.40000	0.100000
25%	5.425000	2.800000	3.07500	1.000000
50%	5.850000	3.000000	4.35000	1.300000
75%	6.300000	3.300000	5.00000	1.775000
max	7.700000	4.100000	6.70000	2.500000

E assim por diante podemos observar os valores de nossa amostra.

In [ ]:
```
samples[3].describe()
```

Out[ ]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	50.000000	50.000000	50.000000	50.000000
mean	5.770000	3.098000	3.636000	1.162000
std	0.857678	0.482125	1.871021	0.798695
min	4.400000	2.200000	1.000000	0.100000
25%	5.000000	2.800000	1.500000	0.225000
50%	5.700000	3.100000	4.250000	1.300000
75%	6.375000	3.400000	5.075000	1.875000
max	7.900000	4.400000	6.700000	2.500000

In [ ]:
```
samples[4].describe()
```

Out[ ]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	50.000000	50.000000	50.000000	50.000000
mean	5.774000	3.052000	3.566000	1.094000
std	0.864919	0.430112	1.807875	0.747898
min	4.300000	2.200000	1.100000	0.100000
25%	5.000000	2.800000	1.600000	0.200000
50%	5.700000	3.000000	4.150000	1.300000
75%	6.375000	3.400000	4.975000	1.800000
max	7.900000	4.000000	6.700000	2.300000

Legal, analisamos diversas amostras de 50 elementos da nossa população principal, mas qual amostra se aproxima melhor do valor real?

In [ ]:
```
df.describe()
```

Out[ ]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Observe que, por mais que uma amostra se aproxime do parâmetro, ainda sim não foi possível igualar as observações amostrais ao parâmetro estudado. Porém, dessa forma é possível estimar um valor que represente determinado parâmetro dada uma amostra.

No caso do exemplo anterior, de amostra casual simples, talvez se observado uma outra metodologia de seleção amostral iríamos obter resultados mais precisos ou próximos ao nosso parâmetro. Logo, um dos maiores desafios para a estimação dos dados de uma população é a seleção da amostra.

Referências

[1] Martins, E.G.M., (2015) Amostragem (Estatística), Rev. Ciência Elem., V3(1):076, DOI: http://doi.org/10.24927/rce2015.076

[2] NETO, Pedro L. C. Estatística. Ed. Blucher Ltda, 1977.

[3] Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." Annals of eugenics 7.2 (1936): 179-188.