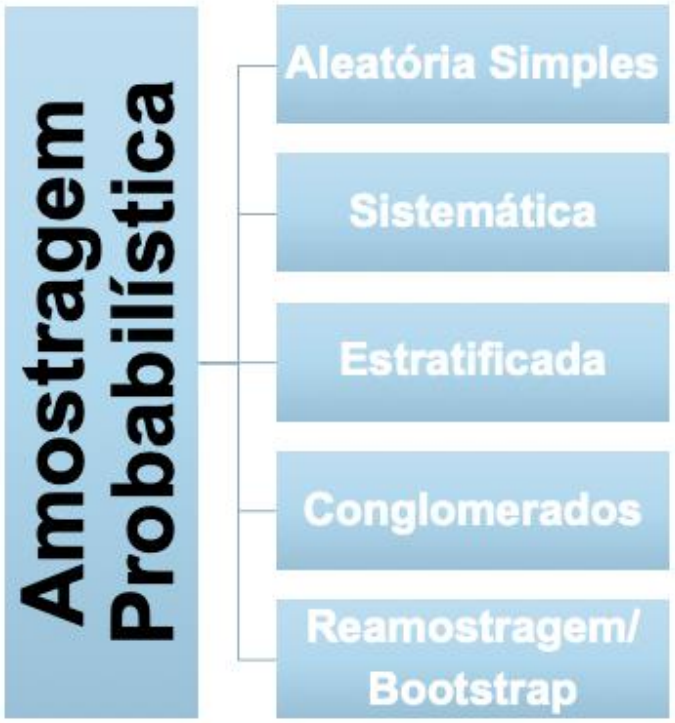
Big Data Analytics com R e Microsoft Azure Machine Learning
Versão 3.0

Descrição dos Tipos de Amostragem Probabilística

Amostragem probabilística é uma amostra em que todos os membros da população têm chance de pertencer a amostra. O que não significa que cada membro tenha exatamente a mesma chance de ser selecionado.

Com Amostragem Probabilística, nós temos a possibilidade de realizar uma variedade de testes de estatística inferencial, que nos permitirá extrair conclusões confiáveis sobre a população. Vejamos uma descrição dos tipos de amostragem probabilística.



Amostragem Aleatória Simples

Uma amostra aleatória simples é a amostra em que cada membro de uma população tem igual chance de ser selecionado. Na amostragem aleatória simples, todos os elementos da população têm a mesma probabilidade de pertencerem à amostra. Essa amostragem pode ser sem reposição, que é quando o elemento que já foi sorteado não continua no sorteio, ou com reposição, quando o elemento sorteado continua no sorteio, podendo ser escolhido novamente.

Amostragem Sistemática

Outra forma de garantir que uma amostra é selecionada randomicamente, é usar Amostragem Sistemática. A amostragem sistemática consiste em selecionar as unidades elementares da população em intervalos pré-fixados. Para funcionar, a técnica requer a listagem prévia da população, que deverá ser homogênea e uma atenção especial na periodicidade dos dados. A amostragem sistemática é semelhante à aleatória simples, mas a listagem é ORDENADA.

O processo basicamente é:

- Divide-se o tamanho da população (N) pelo tamanho da amostra (n), obtendo um intervalo de retirada (k).

- Sorteia-se o ponto de partida. A cada k elementos retira-se um para a amostra (k pode ser o segundo, quarto, vigésimo elemento, etc...)

Uma das vantagens de se utilizar esta técnica é evitar amostra tendenciosa. Como o membro da amostra será selecionado sempre na mesma posição a partir da população, evitamos uma seleção direcionada a membros específicos. Entretanto, deve-se ter cuidado com o conceito de periodicidade, quando um padrão da população coincide com a posição que estamos selecionando.

Por exemplo: Vamos supor que exista a necessidade de fazer uma pesquisa com estudantes universitários, para descobrir quantas horas por semana eles estudam. Para tal experimento, selecionamos sempre a quarta semana de cada mês para a pesquisa. Entretanto, a cada 2 meses a universidade aplica testes e provas, previstos no calendário, na última semana do bimestre. Obviamente, a pesquisa terá valores diferentes nesta semana (caso seja coincidente com a semana do nosso experimento), pois os alunos estarão estudando mais horas para os testes. Isso pode comprometer o resultado da pesquisa.

Amostragem Estratificada

Nós dividimos nossa população em grupos mutuamente exclusivos (chamados estratos) e randomicamente selecionamos membros de cada grupo para nossa amostra.

Por exemplo: Vamos supor que continuamos fazendo um experimento com estudantes universitários, mas desta vez desejamos criar uma amostra de estudantes, com a qual faremos nossa pesquisa de horas de estudo por semana. Nossa amostra deve conter 100 estudantes e sabemos que 30% dos alunos são calouros, 22% estão no segundo ano, 28% no terceiro ano e 20% no quarto ano. Nossa amostra deve ter uma proporção de alunos, condizente com a população.

Para isso, dividimos os alunos em grupos mutuamente exclusivos (um aluno não pode ser calouro e estar no quarto ano ao mesmo tempo) e aplicamos a regra a seguir para criar nossa amostra de 100 alunos, que serão alvo da nossa pesquisa. Com isso, teremos uma amostra (100 alunos) que é representativa da população (todos os alunos da universidade).

Teremos uma amostra (**100 alunos**) que é representativa da população.

Grupo	Número de alunos (% da População)	Amostra
Calouros	540 (30%)	0.30 x 100 = 30
Estudantes no segundo ano	396 (22%)	0.22 x 100 = 22
Estudantes no terceiro ano	504 (28%)	0.28 x 100 = 28
Estudantes no quarto ano	360 (20%)	0.20 x 100 = 20
Total	**1.800 (100%)**	**100**

Cada grupo é homogêneo, ou seja, possui as mesmas características em relação a população.

Amostragem Por Conglomerado

Na amostragem por conglomerados, a população é dividida fisicamente em conglomerados (grupos ou clusters). Selecionam-se aleatoriamente os conglomerados que farão parte da amostra, ao passo que todos os elementos dos conglomerados selecionados serão amostrados. É muito utilizada quando há necessidade de se realizar entrevistas ou observações em grandes áreas.

Na amostragem por conglomerados, também criamos grupos mutuamente exclusivos, cada um representativo da população. Ao invés de selecionarmos membros de cada grupo, selecionamos randomicamente grupos inteiros para nossa amostra. Vamos usar o exemplo anterior, mas desta vez criar uma amostra por conglomerados.

Grupo	Número de alunos (% da População)	Amostra
Calouros	540 (30%)	0.30 x 100 = 30
Estudantes no segundo ano	396 (22%)	0.22 x 100 = 22
Estudantes no terceiro ano	504 (28%)	0.28 x 100 = 28
Estudantes no quarto ano	360 (20%)	0.20 x 100 = 20
Total	**1.800 (100%)**	**100**

Grupo	Tamanho (% da População)	Amostra
Turma de Estatística	540 (30%)	0.30 x 100 = 30
Turma de Filosofia	396 (22%)	0.22 x 100 = 22
Turma de Cálculo I	504 (28%)	0.28 x 100 = 28
Turma de Contabilidade II	360 (20%)	0.20 x 100 = 20
Total	**1.800 (100%)**	**100**

Neste caso, os grupos são heterogêneos, ou seja, é possível que tenhamos alunos do segundo e terceiro ano na turma de Filosofia. Cada grupo representa a população. Temos agora nossa amostra de 100 alunos criada por outra técnica de amostragem.

Reamostragem (Bootstrapping)

Será estudada no próximo item de aprendizagem.