Noções de Estatística Básica

Introdução a Inferência Estatística

* População x Amostra
* Estatísticas x Parâmetros
* Distribuição Amostral da Média

1. População x Amostra

Vamos retomar brevemente tudo o que aprendemos até agora. Aprendemos que a primeira parte de uma análise estatística tem início com uma coleta de dados. Em seguida, os dados devem ser organizados e deve-se analisar as variáveis do estudo. Depois, estudamos uma variedade de métodos que nos auxiliam a resumir as informações obtidas na nossa coleta de dados. Essa parte é a parte de análise descritiva dos dados. Depois nos distanciamos da nossa análise estatística e estudamos conceitos em probabilidade e vimos algumas distribuições teóricas de probabilidade. Por que fizemos tudo isso? Não podemos nos esquecer de um dos objetivos principais de uma análise estatística: inferir informações a respeito de uma população. Quando coletamos dados, estamos na verdade estudando apenas um pedacinho da população. Esse pedacinho se chama Amostra. Como podemos deduzir informações para uma população olhando apenas para uma amostra? Através da análise de inferência estatística.



Definição: População é o conjunto de todos os elementos sob investigação. Amostra é qualquer subconjunto da população.

1. Estatísticas x Parâmetros

Obtida uma amostra, muitas vezes desejamos usá-la para estudar alguma característica específica da população. Por exemplo, se quiséssemos saber a média das alturas da população brasileira, poderíamos coletar uma amostra e começar por estudar a média da nossa amostra. É claro que sabemos que a média da nossa amostra nem sempre vai corresponder ao valor da média populacional, mas é um começo!

Definição: Uma estatística é uma característica da amostra.

Uma estatística é uma função das observações

As estatísticas mais comuns são:

E quanto às características da população?

Definição: Um parâmetro é uma medida usada para descrever uma característica da população.

Abaixo, veremos uma tabela que relaciona as principais e mais comuns estatísticas e com seus respectivos parâmetros.

|  |  |  |
| --- | --- | --- |
| Denominação | População | Amostra |
| Média |  |  |
| Mediana |  |  |
| Variância |  |  |
| Nº de Elementos | N | n |
| Proporção | p |  |
| Quantil | Q(p) | q(p) |
| Quartis |  |  |
| Desvio Padrão | |  |

1. Distribuição Amostral da Média

Antes de avançarmos, é preciso entender uma coisa. Normalmente o objetivo de uma análise estatística é fazer afirmações sobre alguma característica de uma população. Porém, dificilmente teremos acesso a toda a população.

Imagine que uma empresa precise saber, por exemplo, a média da duração de baterias de um determinado modelo de notebooks. Não basta calcular a média de todos os notebooks fabricados. Novos notebooks são fabricados todos os dias. Seria prático ficar testando todas as baterias fabricadas todos os dias?

Imagine que seja necessário calcular a resistência de motores de automóveis de luxo, por exemplo Ferraris. É muito custoso testar muitos desses motores, a Ferrari gostaria de saber a resistência média dos motores utilizando para isso o menor número de motores possíveis.

Para isso, técnicas de análise de inferência podem nos auxiliar.

Primeiramente, vamos falar sobre inferência sobre a média de uma população.

Inicio a discussão com um exemplo prático.

Suponha que tenhamos 5 bolinhas. Cada bolinha possui uma etiqueta com um número.

7

5

5

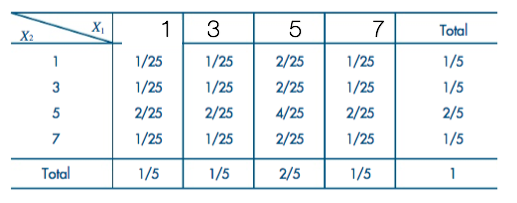
3

1

Vamos sortear duas bolinhas. Quais são as possibilidades de sorteio? E com que probabilidade?

Cada bolinha possui probabilidade de ser sorteada de 1/5. Assumindo independência entre os sorteios, duas bolinhas têm probabilidade 1/25 de serem sorteadas. Exceto se a bolinha tiver número 5, porque ela aparece duas vezes a mais que as demais.

Após algum trabalho, é possível construir a tabela abaixo:

Distribuição das probabilidades das possíveis amostras de tamanho 2 que podem ser selecionadas com reposição da população {1,3,5,5,7}

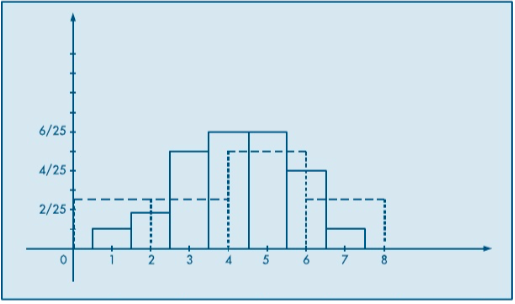
Já que é possível calcularmos as probabilidades do sorteio de 2 bolinhas, podemos calcular também as probabilidades das médias de cada sorteio.

Por exemplo, se sortearmos duas bolinhas de número 5, a média do nosso sorteio será 5. Se sortearmos uma bolinha 1 e uma bolinha 3, a média do nosso sorteio será 2. Procedendo de maneira análoga para todos as combinações possíveis, conseguimos construir uma tabela de probabilidades para a média dos sorteios.

Faça esse exercício com calma. Você deverá obter a seguinte tabela:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|  | 1/25 | 2/25 | 5/25 | 6/25 | 6/25 | 4/25 | 1/25 | 1 |

Observe agora as distribuições de frequências de X (linha tracejada) e (linha contínua) sobrepostas.



Você concorda que o histograma com linhas contínuas parece apresentar um comportamento conhecido? Sim, parece um comportamento que pode ser modelado pela curva normal.

Esse exemplo serve apenas para ilustrar o que acontece com a distribuição amostral da média. Mesmo que a distribuição da população siga uma distribuição diferente, se pudéssemos coletar múltiplas amostras e calcularmos a média de cada uma dessas amostras e em seguida colocássemos essas médias em um histograma como fizemos com o nosso exemplo, obteríamos sempre um comportamento *normal*.

Teorema:

Seja X uma variável aleatória com média e variância e seja uma amostra aleatória simples de X, então:

Esse teorema apenas indica que, independente da distribuição de X, sabemos como se comportam as médias amostrais de X. E as médias amostrais de X seguem uma distribuição normal (\*para n grande\*) com média e variância .

Esse teorema que acabamos de estudar é conhecido como o Teorema do Limite Central. (TLC)

1. Distribuição Amostral de uma Proporção