Noções de Estatística Básica

Introdução a Inferência Estatística

- População x Amostra
- Estatísticas x Parâmetros
- Esperança da Binomial e da Normal
- Distribuição Amostral da Média
- Distribuição Amostral de uma proporção
- Determinação do tamanho de uma amostra

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.

2. 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 $x_1, x_2, \ldots, x_n$.

As estatísticas mais comuns são:

$$\bar{x}: média\ da\ amostra$$
$$S^2: variância\ da\ amostra$$
$$x_{(1)}: o\ menor\ valor\ da\ amostra$$
$$x_{(n)}: o\ maior\ valor\ da\ amostra$$
$$x_{(i)}: a\ i-ésima\ maior\ observação\ da\ amostra$$
$$Quantis\ empíricos\ (q(p))$$

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 com seus respectivos parâmetros.

Denominação	População	Amostra
Média	μ	$\bar{x}$
Mediana	$Md = Q_2$	$md = q_2$
Variância	σ^2	S^2
Nº de Elementos	N	n
Proporção	p	$\hat{p}$
Quantil	Q(p)	q(p)
Quartis	Q_1, Q_2 ...	$q_1, q_2, \ldots$
Desvio Padrão	σ	S

3. Esperança: Valor médio de uma variável aleatória

O que é ***Esperança***?

Para entender o que é esperança, vamos relembrar alguns conceitos que aprendemos sobre o método estatístico até agora.

Primeiramente, coletamos dados. Os dados passam por uma avaliação. Estatísticas são calculadas, tabelas de distribuição de frequência são construídas e também gráficos de distribuição de frequência (histogramas). Em seguida, a partir da observação dos dados e dos histogramas, podemos propor modelos probabilísticos teóricos que se adequem aos nossos dados. Nas nossas aulas vimos três modelos: ***Bernoulli***, ***Binomial*** e ***Normal***.

A esperança é o valor médio de uma distribuição probabilística teórica.

No caso da distribuição Bernoulli, temos que a esperança é ***p*** (probabilidade de um indivíduo ter uma característica).

No caso da distribuição Binomial, temos que a esperança é ***np*** (número de indivíduos multiplicados pela probabilidade de um indivíduo ter uma característica).

No caso da distribuição Normal, por se tratar de uma variável aleatória contínua, o valor médio da variável é calculado com o auxílio de métodos matemáticos mais avançados, o cálculo integral. Por convenção, diz-se que a esperança de uma variável com distribuição Normal é μ.

4. 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?

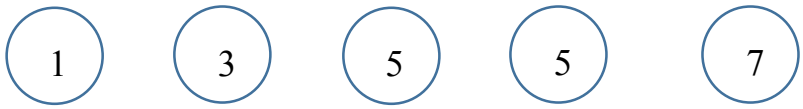
Imagine que seja necessário calcular a resistência de motores de automóveis de luxo, por exemplo Ferrari. É muito custoso testar muitos desses motores, a Ferrari gostaria de saber a resistência média dos motores utilizando para isso o menor número de motores possíveis.

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

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.



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}

X_2 \ X_1	1	3	5	7	Total
1	1/25	1/25	2/25	1/25	1/5
3	1/25	1/25	2/25	1/25	1/5
5	2/25	2/25	4/25	2/25	2/5
7	1/25	1/25	2/25	1/25	1/5
Total	1/5	1/5	2/5	1/5	1

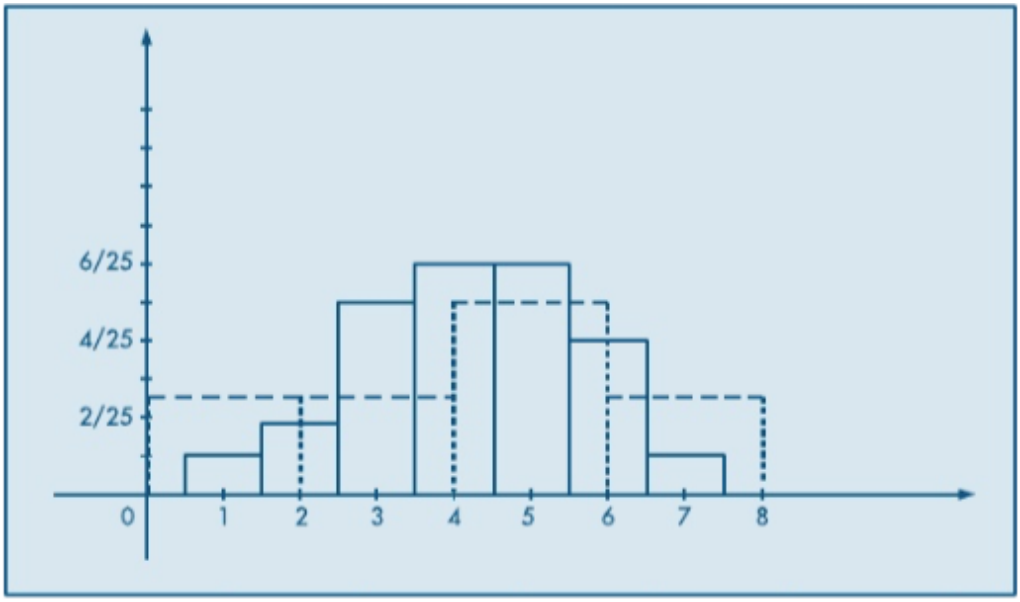
Já que é possível calcularmos as probabilidades do sorteio de 2 bolinhas, podemos calcular também as probabilidades das médias de cada sorteio.

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:

$\bar{x}$	1	2	3	4	5	6	7	Total
$P(\bar{X} = \bar{x})$	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 $\bar{X}$ (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:

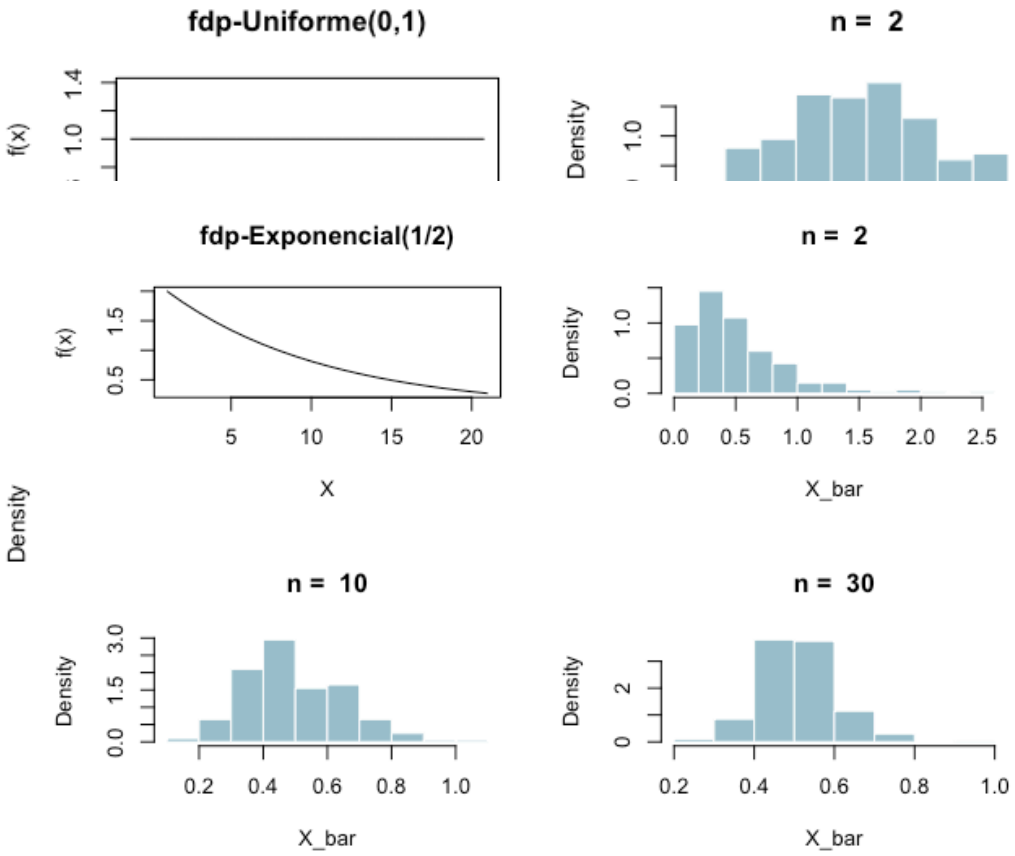
Seja X uma variável aleatória contínua com média μ e variância σ^2 e seja $(x_1 \dots x_n)$ uma amostra aleatória simples de X, então:

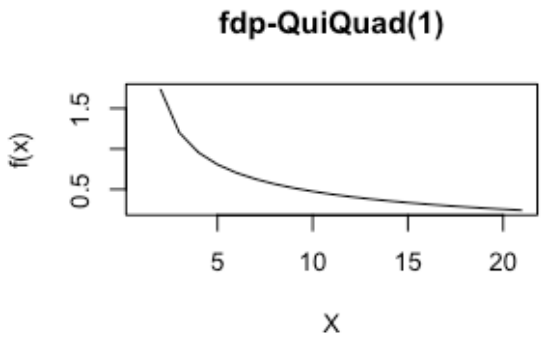
$$E(\bar{X}) = \mu \quad e \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

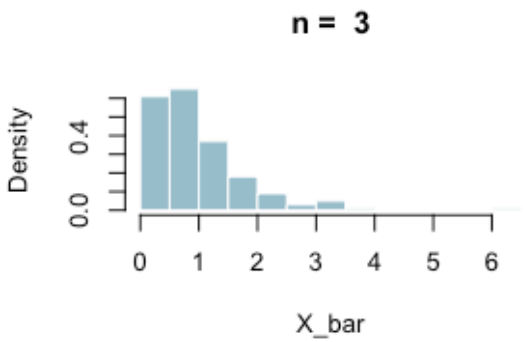
Esse teorema apenas indica que, independente da distribuição de X, sabemos como se comportam as médias amostrais de X. E as médias amostrais de X seguem uma distribuição normal (*para n grande*) com média μ e variância $\frac{\sigma^2}{n}$.

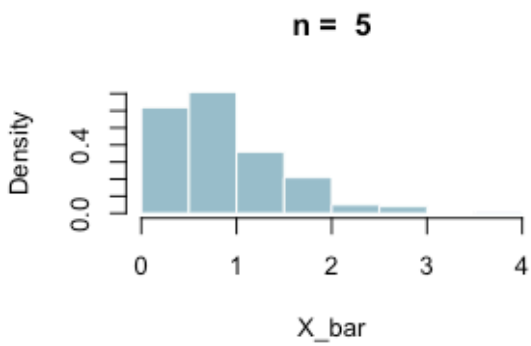
Esse teorema que acabamos de estudar é conhecido como o Teorema do Limite Central. (TLC)

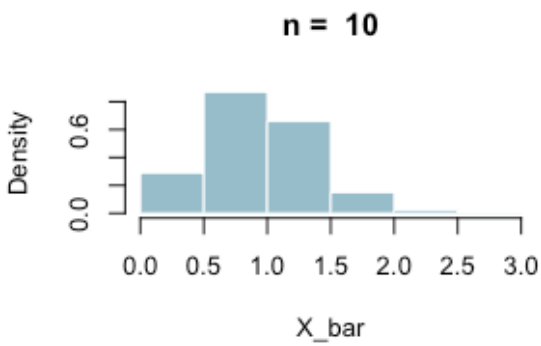
Veja abaixo alguns exemplos de distribuições amostrais da média para amostras extraídas de algumas populações com distribuições diferentes:

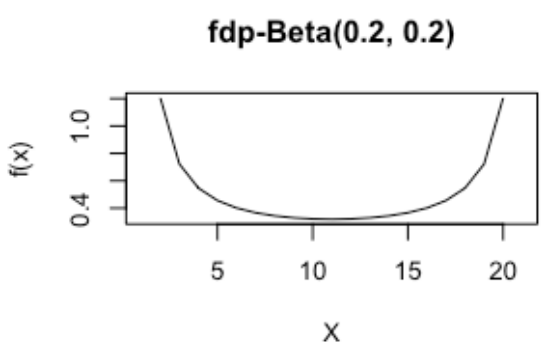


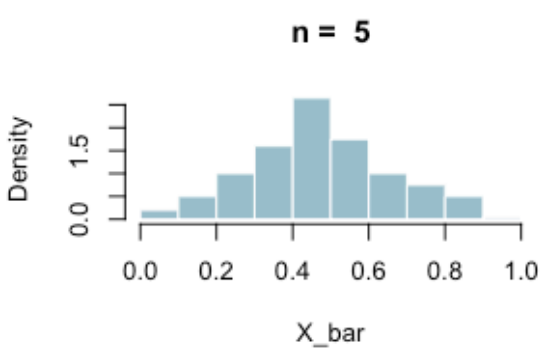

fdp-QuiQuad(1)
f(x)
0.5
1.5
5
10
15
20
X

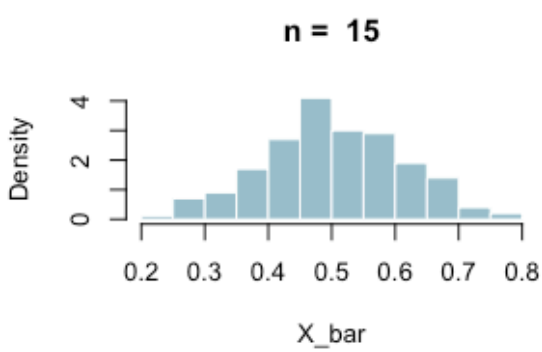

n = 3
Density
0.0
0.4
0
1
2
3
4
5
6
X_bar

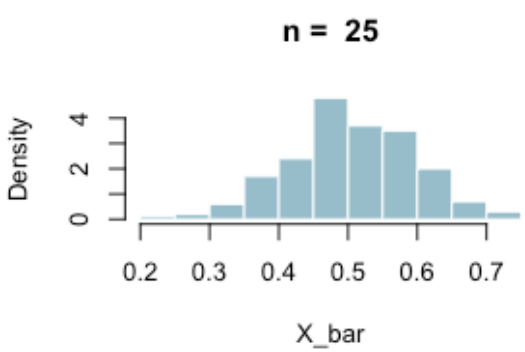

n = 5
Density
0.0
0.4
0
1
2
3
4
X_bar


n = 10
Density
0.0
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
X_bar


fdp-Beta(0.2, 0.2)
f(x)
0.4
1.0
5
10
15
20
X


n = 5
Density
0.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
X_bar


n = 15
Density
0
2
4
0.2
0.3
0.4
0.5
0.6
0.7
0.8
X_bar


n = 25
Density
0
2
4
0.2
0.3
0.4
0.5
0.6
0.7
X_bar

5. Distribuição Amostral de uma Proporção

Vimos acima que para uma variável com distribuição contínua qualquer, as médias amostrais de X seguem uma distribuição aproximadamente Normal com média μ e variância $\frac{\sigma^2}{n}$.

Mas o que acontece quando temos uma variável do tipo discreta, por exemplo, ter uma doença ou não (Bernoulli)?

$$X = \begin{cases} 1, se\ o\ indivíduo\ possui\ uma\ característica \\ 0, se\ o\ indivíduo\ não\ possui\ essa\ característica \end{cases}$$

Nessa situação estamos diante de uma distribuição amostral para uma Proporção.

Porém, pode-se mostrar também que o T.L.C também vale para esse caso, a única diferença é que, ao invés de utilizarmos a esperança μ e variância $\frac{\sigma^2}{n}$, teremos que:

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

6. Determinação do Tamanho de uma Amostra

Uma pergunta muito frequente para os estatísticos é:
Como calcular o tamanho de uma amostra?

Essa pergunta é importante pois, quanto maior a nossa amostra, mais precisão teremos ao fazer estimativas para uma população. Isto é bastante intuitivo! Suponha uma população com 3 mil indivíduos. Agora suponha também que coletamos uma amostra com 2.997 indivíduos e outra com 10 indivíduos. Qual das amostras você escolheria para fazer previsões a respeito da população?

Como podemos escolher o tamanho ideal de uma amostra?

Primeiramente, siga o seguinte raciocínio:

A variável X na população possui uma distribuição desconhecida, porém sabemos que o valor médio da nossa variável é μ. Podemos tentar estimar μ a partir da nossa amostra, para tal, utilizaremos a estatística $\bar{X}$. Porém, sabemos que haverá uma diferença entre o valor obtido na amostra e o valor do parâmetro μ.

À esta diferença, atribuímos o nome ***erro*** representado pelo símbolo ε.

$$|\bar{X} - \mu| = \varepsilon$$

O tamanho da amostra pode ser calculado a partir do momento em que eu escolho um erro amostral máximo, ou seja, existe uma variação na minha estimativa que eu tolero.

Exemplo: Estimar média populacional de alturas no Brasil.

Fará diferença para o meu estudo se eu errar em milímetros? Por exemplo, suponha que a média verdadeira seja 1.767m e nós estimamos 1.761m. E se errarmos em metros? A média populacional é 1.7m e estimamos 2.3m. Cada variável, cada estudo e cada pesquisador irão, em conjunto, determinar o erro máximo tolerável.

Porém sabemos que é impossível ter certeza de que estamos respeitando o erro, afinal de contas, não sabemos o valor da média populacional.

O que se pode fazer é esperar que estejamos respeitando o erro máximo com uma alta probabilidade (95%, por exemplo).

Traduzindo:

$$P(|\bar{X} - \mu| \leq \ \varepsilon) \geq \gamma$$

Resumindo, o tamanho da amostra irá depender de estabelecermos um erro limite bem como uma probabilidade de tolerância.

Sabemos também que $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Logo, $\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$.

Ou seja,

Padronizando

$$P(-\varepsilon < \bar{X} - \mu < \varepsilon) = P\left(\frac{-\sqrt{n}\varepsilon}{\sigma} \leq z \leq \frac{\sqrt{n}\varepsilon}{\sigma}\right) = \gamma$$

Dado γ, podemos obter z_γ da tabela da N(0,1) tal que

$$\frac{\sqrt{n}\varepsilon}{\sigma} = z_\gamma$$

Portanto,

$$n = \frac{\sigma^2 z_\gamma}{\varepsilon^2}$$

Note que para calcularmos o tamanho da amostra precisamos saber a variância populacional. Para isso, precisaremos saber alguma informação prévia sobre a variância ou utilizar uma amostra pequena piloto para estimar σ^2.