POPULAÇÃO, AMOSTRA MEDIDAS DE DISPERSÃO

Estatística Aplicada I
IRI-USP

Maio 2021

Profª. Maria Antonieta Del Tedesco Lins

1

Estrutura da aula

2

- Temas
  - População e amostra
  - Voltando às medidas de posição e dispersão
    - Características da variância
    - Desvio-padrão e suas propriedades
  - Ferramenta estatística descritiva Excel
- Bibliografia básica

  - Agresti, A. e Finlay, B. Métodos Estatísticos para as Ciências Sociais. 4ºed. Porto Alegre: Penso, 2012, Cap. 2.
  - Barrow, M. Estatística para economia, contabilidade e administração. São Paulo: Ática, 2007, Cap. 1
  - Lapponi, J. Estatística usando Excel 5 e 7. São Paulo: Lapponi Treinamento e Editora, 1997. Capítulos 3 e 4
  - Morettin, P. e W. Bussab. Estatística básica. 5. ed. São Paulo: Saraiva, 2005. Cap. 3

2

3 População e Amostra

3

População e amostra

4

- População é o conjunto total de indivíduos (unidades elementares, como definimos antes) sobre os quais queremos obter informações
- Amostra é um subconjunto selecionado de elementos da população
- Uma **amostra representativa** é um conjunto que tem as mesmas características da população da qual foi extraída
- Em geral, os números que vemos divulgados no dia a dia são relativos a amostras
  - Resultados de pesquisas: intenção de voto, aprovação de governo, avaliação de políticas, preferências, etc
  - Indicadores sobre grupos de países, etc.

4

População e amostra

5

- Para que as conclusões sejam corretas, a amostra deve ser representativa
  - ou seja, a composição da amostra deve ser similar à do eleitorado
- Um caminho intuitivo de montar a amostra poderia ser pegar 2000 pessoas ao acaso (aleatoriamente)
  - poderiam ser sorteadas pessoas ou tomadas ao acaso não população
- Mas o procedimento não é tão simples assim...

5

População e amostra

6

- Por agora, devemos saber que existem vantagens de trabalhar com amostras
- Custo
  - Perguntar a cada brasileiro como avalia o governo poderia ser custoso...
  - Por isso o censo é feito a cada dez anos
- Tempo
  - Processar informações referentes a uma população demanda tempo
- Impossibilidade
  - Usar cada produto fabricado por uma indústria para verificar sua qualidade seria impossível e irracional, no mínimo...

6

População e amostra

7

- Amostras aleatórias
  - Permitem aplicar a teoria das probalilidades às amostras
    - Pode-se afirmar com que probabilidade (dependendo do tamanho da amostra) a amostra é representativa
  - Uma amostra aleatória de tamanho n retirada de uma população é uma possível combinação (entre inúmeras) de n indivíduos que podem ser retiradas da população. Qualquer amostra de tamanho n tem a mesma probabilidade de ser retirada

7

População e amostra

8

- No momento, não precisamos nos preocupar com formas de montagem de amostras, mas sim saber que tipo de amostra temos, quando lidamos com alguns dados
- As pesquisas de opinião procuram reproduzir o sentimento de toda a população ao entrevistarem uma amostra
- As condições em que é feita a entrevista importam: lugar, urna eletrônica ou papel, a pergunta, etc

8

Pesquisas de opinião

9

- Realizadas por institutos de pesquisa privados, incluem **pesquisas eleitorais** e p**esquisas de avaliação do governo.**
- Os dados são comumente levantados através de amostragem por cotas, onde a população de interesse é caracterizada a partir de variáveis descritivas (gênero, raça, idade, renda, anos de estudo e local de moradia, por exemplo) que são utilizadas para fixação de cotas de resposta. Assim, entrevistadores dirigem-se a local de realização de entrevistas e aplicam questionários até que a cota de um público específico (Ex. 300 Homens, brancos, entre 30-40 anos, com renda entre R$3000-R$5000, graduado, morador do Rio de Janeiro) seja atingida.
- Nestas pesquisas, há problemas de cobertura e não-resposta da população que não são facilmente mensuráveis e impactam nas estimativas do Intervalo de Confiança, impactando na confiabilidade das inferências.
- Há uma tendência mais atual dos institutos utilizarem a chamada Amostragem Probabilística por Cotas (APC), uma abordagem híbrida de planos amostrais probabilísticos e não-probabilísticos que minimizam problemas com cobertura da população.

9

Exemplo: Pesquisa de opinião

10

- Tomemos o exemplo de uma pesquisa de opinião realizada pelo instituto Datafolha em março de 2021.
- O objetivo da pesquisa é inferir a opinião do conjunto da população brasileira.
- Para isso, se investiga uma amostra representativa da população (com distribuição semelhante à população: gênero, renda, raça, distribuição geográfica, etc.)
- Por causa da pandemia, as pessoas foram entrevistadas por telefone/

10

Ex: Metodologia da pesquisa de opinião

11

- A pesquisa telefônica, utilizada neste estudo, representa o total da população adulta do país. As entrevistas são realizadas por profissionais treinados para abordagens telefônicas e as ligações feitas para aparelhos celulares, utilizados por cerca de 90% da população.
- O método telefônico exige questionários rápidos, sem utilização de estímulos visuais, como cartão com nomes de candidatos, por exemplo. Assim, mesmo com a distribuição da amostra seguindo cotas de sexo e idade dentro de cada macrorregião, e da posterior ponderação dos resultados segundo escolaridade, os dados devem ser analisados com alguma cautela por limitar o uso desses instrumentos.
- Na pesquisa divulgada hoje, feita dessa forma para evitar o contato pessoal entre pesquisadores e respondentes, o Datafolha adotou as recomendações técnicas necessárias para que os resultados se aproximem ao máximo do universo que se pretende representar.
- Todos os profissionais do Datafolha trabalharam em casa, incluídos os entrevistadores, que aplicaram os questionários através de central telefônica remota.
- Foram entrevistados 2023 brasileiros adultos que possuem telefone celular em todas as regiões e estados do país. A margem de erro é de dois pontos percentuais.
- A coleta de dados aconteceu entre os dias 15 e 16 de março de 2021.

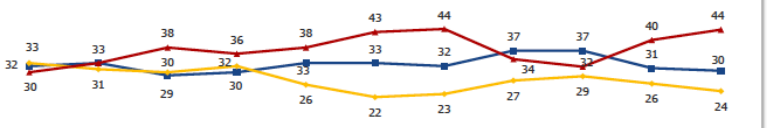
Fonte: Datafolha. Disponível em https://datafolha.folha.uol.com.br/opiniaopublica/2021/03/1989226-maioria-54-agora-reprova-trabalho-de-bolsonaro-na-pandemia.shtm

11

Avaliação do presidente Jair Bolsonaro

Resposta estimulada e única, em %

O presidente Jair Bolsonaro completou um ano e onze meses de governo. Na sua opinião o presidente Jair Bolsonaro está fazendo um governo ótimo, bom, regular, ruim ou péssimo?



	Após 3 meses - 02 e 03/04/2019	Após 6 meses - 04 e 05/07/2019	Após 8 meses - 29 e 30/08/2019	Após 1 ano - 05 e 06/12/2019	Após 1 ano e 4 meses 27/04/2020	Após 1 ano e 4 meses 25 e 26/05/2020	Após 1 ano e 5 meses 23 e 24/06/2020	Após 1 ano e 7 meses 11 e 12/08/20	Após 1 ano e 11 meses 08 a 10/12/20	Após 2 anos 20 e 21/01/21	Após 2 anos e 3 meses 15 e 16/03/21
Ótimo/bom	32	33	29	30	33	33	32	37	37	31	30
Regular	33	31	30	32	26	22	23	27	29	26	24
Ruim/Péssimo	30	33	38	36	38	43	44	34	32	40	44
Não sabe	4	2	2	1	3	2	1	1	3	2	2

Fonte: O presidente Jair Bolsonaro completou um ano e onze meses de governo. Na sua opinião o presidente Jair Bolsonaro está fazendo um governo ótimo, bom, regular, ruim ou péssimo?
Base : Total da amostra

FOLHA DE S.PAULO Datafolha

A partir de uma amostra de 2000 pessoas, obtém-se uma aproximação da opinião de 212 milhões...

12

12

13

Voltando à medidas de posição e dispersão

13

Ainda pensando no ordenamento dos dados: Quantis

14

- Vimos que, tanto a média como o desvio padrão podem apresentar falhas para representar um conjunto de dados, já que eles
  - São afetados pelos valores extremos
  - Não nos dão uma ideia da simetria ou assimetria da distribuição de dados
- Em uma relação entre a escala que vai de 0% a 100% e a série de números naturais que representam uma série dados ordenados de uma amostra, pode-se subdividir a escala em valores fixos, de forma a verificar a concentração dos dados nas referidas faixas

14

Ainda pensando no ordenamento dos dados

15

Quantis

Percentis - medidas que dividem um conjunto de dados em diversas partes são úteis na apresentação da distribuição de seus valores, principalmente se o conjunto de dados é não simétrico.

Os percentis dividem um conjunto de dados em cem partes de igual tamanho

A mediana representa o percentil 50.

Quartis – 1° e 3° Quartis (25% e 75%)

Quintis - 20% , 40%, 60% e 80%.

15

Quantis

16

- Pode-se definir várias formas de referências fixas: 10% ou decil; 25% ou quartil, etc.

- Algumas formas de quantis
  - q(0,25): 1° quartil= 25° percentil
  - q(0,50): mediana= 5° decil = 50° percentil
  - q(0,75): 3° quartil= 75° percentil
  - q(0,40): 4° decil

16

Quantis

17

- Por exemplo, se tivermos uma série de dados referentes a uma variável, como as notas obtidas por uma classe de 9 alunos

  2, 4, 8, 9, 5, 10, 7

  Ordenando os valores, teríamos

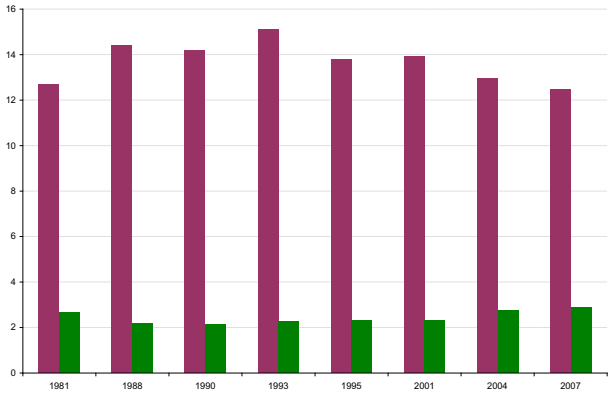
  2<4<5<7<8<9<10

  A mediana seria q(0,50)=7
- Mas, o que seria q(0,20) [o valor que deixa 20% das observações à sua esquerda]? Que valor considerar? 2º ou 3º?
- Aqui teríamos que usar os conceitos que aprendemos vendo as distribuições de frequência e criar intervalos aos quais relacionaríamos o número de observações encontradas em cada faixa de valor

17

Com dados de distribuição de renda, por exemplo, esses conceitos são muito úteis

18



Fonte: Ipeadata

■ Proporção da renda apropriada pelos indivíduos pertencentes ao 1% mais rico da distribuição de indivíduos segundo a renda domiciliar per capita

■ Proporção da renda apropriada pelos indivíduos pertencentes ao 20% mais pobre da distribuição de indivíduos segundo a renda domiciliar per capita

18

As medidas de dispersão vistas

- Desvio médio

$$dm(X) = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n}$$

- Variância

$$\text{var}(X) = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n}$$

- Desvio padrão $dp(X) = \sqrt{\text{var}(X)}$

19

Retomando as medidas já vistas

- Até agora, consideramos sempre as medidas para a população
- Para a amostra, as fórmulas ficam:
  - Variância da amostra ⇨ $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$
  - O desvio padrão da população era ⇨ $\sigma_X = +\sqrt{\sigma_X^2}$
  - Para a amostra é ⇨ $S_X = +\sqrt{S_X^2}$

20

Resultado da variância tem características importantes

21

- Variância é sempre um número positivo
- O numerador, quando se calcula a variância para a população ou uma amostra é o mesmo (soma dos desvios ao quadrado)
- A variância de uma variável considerada como população é uma média aritmética dos quadrados dos desvios
- A variância de uma variável considerada como amostra é uma média, embora seja dividida por n-1
- A variância é afetada por valores extremos, como a média
- A variância traz problemas com a unidade de medida, que é o quadrado da unidade de medida original da amostra

21

Desvio padrão

- A variância é um medida ao quadrado. No nosso exemplo da aula passada, emigrantes da Bósnia em determinado ano ao quadrado, o que é estranho
- Como vimos, o desvio padrão é $dp(X) = \sqrt{\text{var}(X)}$
- Portanto, o resultado está expresso em uma unidade de medida que faz algum sentido
- Alto desvio padrão indica uma maior variação e o valor de S (ou σ) dá uma ideia da distância típica em relação à média

22

Desvio padrão

23

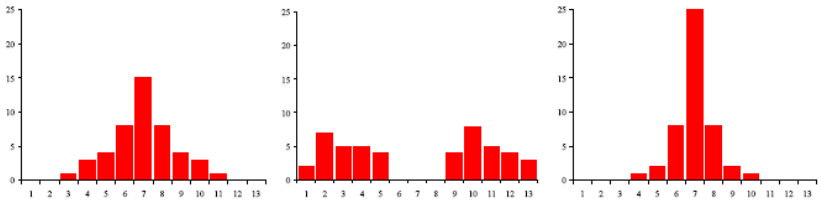
- Pode-se mudar a escala do desvio padrão, assim como se faz para a média, mas sempre tendo em mente a unidade de medida original (imigrantes, número de filhos, pontos na nota, etc.)
- Só são comparáveis desvios padrão que estejam na mesma unidade de medida
- O desvio padrão será 0 quando todas as observações forem iguais
- Quanto maior a variação, maior o desvio padrão
- Uma boa forma de pensar em S é como uma distância média entre uma observação e a média

23

Exemplos de desvio padrão

24

A distribuição dos valores em relação à média pode ser muito distinta.
Dizemos que os valores que assumem a variável dão origem a diferentes distribuições
Duas variáveis com médias iguais e desvio padrão diferentes têm diferentes formas de distribuição de frequência



Distribuição agrupada
S = 1,69

Distribuição dispersa
S = 4,03

Distribuição estreita
S = 1,04

24

Coeficiente de variação: medida relativa de dispersão

- O desvio padrão considera que os desvios são distribuídos de maneira homogênea em torno da média
- Trata-se de uma medida absoluta
- Não podemos comparar medidas de dispersão de duas ou mais distribuições: as unidades podem se diferentes ou os valores de média muito afastados
- O **coeficiente de variação** é uma medida relativa de dispersão, que permite comparar distribuições

$$CV_{pop} = \frac{\sigma_X}{\mu_X} \qquad CV_{amo} = \frac{S_X}{\bar{X}}$$

⇨A variável com menor CV tem menor dispersão ou variabilidade

25

Medida de assimetria

26

- Além da dispersão das observações em um conjunto de dados, podemos saber como se distribuem esses dados, se estão mais dispersos "para um lado ou para outro"
- O coeficiente de assimetria mostra isso
  - Se CA>0 distribuição tem assimetria à direita
  - Se CA<0 distribuição tem assimetria à esquerda
  - Se CA=0 distribuição é simétrica

26

Medida de assimetria

27

- Coeficiente de assimetria

$$CA = \frac{\sum f(x-\mu)^3}{N\sigma^3}$$

$$CA = \frac{\sum f(x-\overline{X})^3}{NS^3}$$

- A medida de assimetria não é sempre útil na observação prática dos dados
- O CA nem sempre dá ideia da forma da distribuição. Duas distribuições diferentes podem ter o mesmo CA
- É sempre bom fazer um gráfico

27

ERRO PADRÃO DA MÉDIA (S_x)

28

Quando se obtém uma amostra aleatória de tamanho n, estima-se a média populacional. É bastante intuitivo supor que se uma nova amostra aleatória for realizada a estimativa obtida será diferente daquela primeira. Desta forma, reconhece-se que as médias amostrais estão sujeitas à variação e formam populações de médias amostrais, quando todas as possíveis amostras são retiradas de uma população.

O erro padrão analisa a variabilidade de uma média

28

Erro padrão

29 Fornece um mecanismo de medir a precisão com que a média populacional foi estimada

$$S\bar{x} = \frac{S}{\sqrt{n}}$$

29

30 Análise estatística descritiva

Seguindo passos propostos em Lapponi, com Excel

30

Estatística descritiva

31

- Seguindo o modelo do pacote Excel apresentado por Lapponi, podemos fazer uma análise estatística
- Devemos registrar a amostra (intervalo de entrada)
- As opções de saída vão nos dar os resultados que queremos
- Informações sobre resultados
  - Erro padrão: é o erro amostral S, calculado
  - Nível de confiança (95%): 95% é percentual de acerto da estimativa da média da população

31