Aula 4 – Estatística – Conceitos básicos

Fundação Itaú Social

Plano de Aula

- Amostra e universo
- Média
- Variância / desvio-padrão / erro-padrão
- Intervalo de confiança
- Teste de hipótese

Amostra e Universo

A estatística nos ajuda a estudar fenômenos de uma população a partir de uma amostra dela.

- **Universo**: é o conjunto completo da população que pretendemos estudar
- **Amostra**: é um subconjunto do universo que selecionamos para o estudo

A ideia é que a **amostra** seja uma fatia da população que **represente bem** o universo.

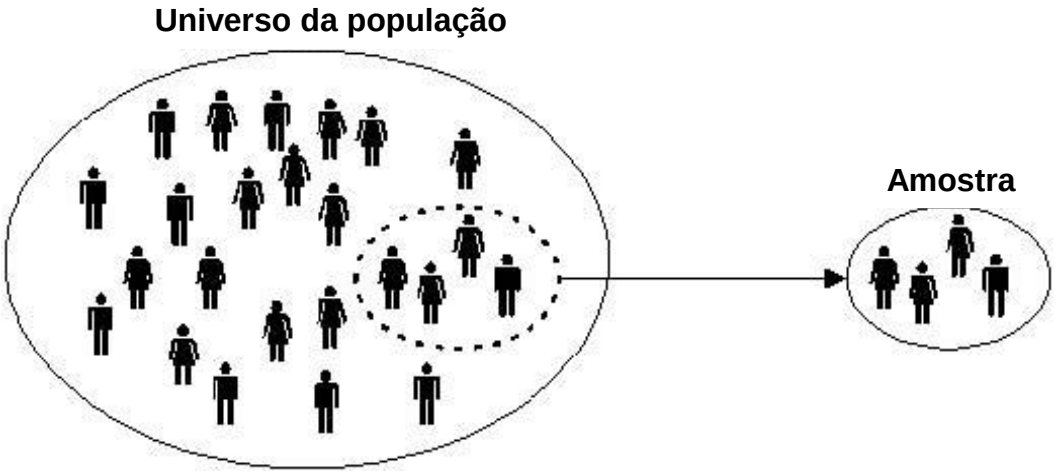
Nenhuma amostra representa exatamente o universo, portanto todo cálculo estatístico está sujeito a **erros**.

Aprenderemos aqui como calcular algumas **estatísticas básicas**
e como lidar com os **erros** dessas estatísticas.

Amostra e Universo

Exemplo, se queremos estudar a intenção de voto para presidente dos eleitores brasileiros, não precisamos entrevistar TODOS os eleitores, basta coletar uma boa amostra e usar a estatística!

Os institutos de pesquisa trabalham com amostras de 2 a 3 mil eleitores e conseguem calcular a intenção de voto para toda população!



Média

- Média = o 'valor esperado' de uma variável.
- É um dos parâmetros que descreve nossa amostra.
- Exemplo:
  - Imagine que temos as notas obtidas pelos alunos de uma sala de aula.
  - A média das notas da sala pode nos dar uma idéia do desempenho esperado daquela sala.

Média - cálculo

Média = soma de todos os valores observados da variável (p. ex., a nota) dividido pelo número de observações (p. ex., os alunos):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Onde:

- x_i = é a nota de cada aluno.
- n é o número de alunos na sala.
- $\sum$ = Somatória

Exemplo:

- Imagine que queremos estudar o desempenho dos alunos de um colégio.
- Para isso pegamos uma **amostra** de 20 alunos e aplicamos uma prova, cujas notas estão tabuladas ao lado.

id_aluno	Nota
201401	4
201402	8
201403	8
201404	3
201405	7
201406	7
201407	5
201408	5
201409	5
201410	9
201411	2
201412	9
201413	2
201414	7
201415	4
201416	5
201417	6
201418	5
201419	4
201420	4

- A média pode ser calculada facilmente com a fórmula:

$$\frac{(4+8+8+\cdots+4+4)}{20} = 5{,}5$$

- Então, em média, os alunos da nossa amostra tiveram nota 5,5.

Fundação Itaú Social

Média

A média dá uma ideia, ainda que geral, do perfil esperado de uma característica da população. Ela nos diz que:

- Se calculamos uma nota média de 8,0, podemos esperar que a população é formada por ótimos alunos.
- Porém, se a nota média for 2, podemos esperar que a população seja formada por alunos de baixo desempenho.

O que a média (sozinha) não nos permite dizer:

- A média não diz nada sobre a característica de cada individuo nem quantos indivíduos estão acima ou abaixo dela.
- Ou seja, usando só a nota media não podemos se João tem bom ou mau desempenho nem podemos afirmar se a maioria dos alunos tem nota boa ou ruim.

Exemplo:

- Imagine agora duas amostras de alunos com suas notas.
- Calculando a nota média das duas amostras chegamos ao mesmo número: 5,5
- Considerando apenas a média, ela nos diz que o desempenho esperado dos dois grupos é idêntico.

id_aluno	Nota 1	Nota 2
201401	8	6
201402	7	5
201403	8	6
201404	10	6
201405	8	5
201406	7	5
201407	1	5
201408	1	5
201409	1	5
201410	7	6
201411	10	5
201412	1	6
201413	3	5
201414	0	5
201415	1	6
201416	4	5
201417	4	6
201418	9	5
201419	10	6
201420	9	6

- Mas veja que o desempenho esperado neste caso vem de dois grupos completamente diferentes.
- Repare que no grupo 1 todos os alunos tem notas distantes de 5,5, enquanto no grupo 2 todas as notas estão muito próximas de 5,5
- Assim, a média que calculamos reflete melhor o desempenho do grupo 2.

Fundação Itaú Social

Variância / Desvio-padrão

- Para saber o quão bem a média representa os integrantes da amostra, usamos as chamadas **medidas de dispersão**, como a variância.

- A variância mede a dispersão dos valores observados em torno de seu valor esperado.
- Para a **turma 1** do exemplo, podemos esperar uma **variância maior** (há muitos alunos longe da média) e para a **turma 2** devemos ter uma **variância menor** (há muitos alunos perto da média)

- Desvio-padrão é a raiz quadrada da variância. Esta será uma medida bastante importante nos nossos cálculos mais à frente.

Variância / desvio-padrão - cálculo

$$s^2 = \frac{\sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right) \right]^2}{(n-1)}$$

variância

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right) \right]^2}{(n-1)}} = \sqrt{s^2}$$

desvio-padrão

Onde:

- x_i = nota de cada aluno
- $\bar{x}$ = nota média
- n = número de alunos

Fundação Itaú Social

Exemplo:

- Voltando ao exemplo anterior e aplicando a fórmula da variância, confirmamos nossa hipótese de que o grupo 1 tem variância maior.

- $Var\ 1 = \frac{(8-5{,}5)^2+\cdots+(9-5{,}5)^2}{(20-1)} = 12{,}8$
- $Var\ 2 = \frac{(6-5{,}5)^2+\cdots+(6-5{,}5)^2}{(20-1)} = 0{,}7$

- *Desvio-padrão 1* = 3,6
- *Desvio-padrão 2* = 0,8

id_aluno	Nota 1	Nota 2
201401	8	6
201402	7	5
201403	8	6
201404	10	6
201405	8	5
201406	7	5
201407	1	5
201408	1	5
201409	1	5
201410	7	6
201411	10	5
201412	1	6
201413	3	5
201414	0	5
201415	1	6
201416	4	5
201417	4	6
201418	9	5
201419	10	6
201420	9	6

Fundação Itaú Social

Variância / Desvio-padrão

- Como essas duas medidas dão ideia da distância que as pessoas estão da média, elas podem ser usadas como medidas de precisão da média.

- Na turma 1, onde a **variância é alta**, a média é **pouco precisa**, isto é, diz pouco sobre as notas dos alunos pois há muitos longe da média
- Na turma 2, onde a **variância é baixa**, a média é **mais precisa**, isto é, diz muito sobre as notas dos alunos pois há muitos perto da média
- Assim, quanto **menor a variância/desvio-padrão**, **maior é a confiança** de que o valor a ser observado será próximo da média.

Erro padrão e intervalo de confiança

- Para avaliarmos melhor a precisão da média, usamos o desvio-padrão para calcular o chamado **erro-padrão da média**, cuja fórmula é:

$$SE = {}^{s}/_{\sqrt{n}}$$

SE = erro padrão dá média

s = desvio- padrão

- Veja que o erro padrão depende do desvio-padrão e do tamanho da amostra, de modo que, quanto **maior a amostra**, **menor o erro-padrão** e maior a precisão da média.

Exemplo:

- Vamos então calcular os erros-padrão para as notas médias das nossas amostras de alunos.

- $SE_1 = {}^{3,6}/_{\sqrt{20}} = 0{,}80$
- $SE_2 = {}^{0,8}/_{\sqrt{20}} = 0{,}18$

- Como já sabíamos, a média da amostra 1 é menos precisa, isto é, tem erro padrão maior, do que na amostra 2.

Como calcular média, variância etc. no Excel

INSTRUÇÕES PARA O EXCEL 2010

O primeiro passo é instalar o pacote de ferramentas de análise de dados do Excel (que servirá para todo o curso):

1. Clique em "Arquivo" > "Opções" > "Suplementos".
2. Na janela, na parte de baixo, em "Gerenciar" selecione "Suplementos do Excel" e clique em "Ir".
3. Na próxima janela, selecione "Ferramentas de Análise" e clique em OK.
4. Vá até a guia "Dados" e verifique se em cima no canto direito aparece o botão "Análise de Dados".

Pronto. As ferramentas estão instaladas. Agora vamos usá-las

Como calcular média, variância etc. no Excel

- ✓ As ferramentas de análise de dados são um ótimo atalho para calcular média, variância etc. de um conjunto de dados.

1. Com a base de dados aberta, clique na guia "Dados" > "Análise de Dados".
2. Na janela que aparece, clique em "Estatística descritiva" > "OK".
3. Na nova janela, em "Intervalo de dados" selecione as células de todas as variáveis que queremos descrever (incluindo os seus títulos).
4. Clique em "Rótulos na primeira linha".
5. Clique em "Resumo estatístico".
6. Clique "OK".

Pronto. O Excel reporta uma nova planilha com todas as estatísticas (inclusive algumas que não vimos por aqui...).

Exemplo:

- Imagine que, junto com as notas de cada turma, coletamos o gênero (menino = 1) e a idade dos alunos:

Nota 1	Menino 1	Idade 1	Nota 2	Menino 2	Idade 2
8	1	10	6	0	11
7	1	10	5	0	10
8	1	10	8	0	13
10	0	11	6	1	12
8	0	10	5	1	11
7	1	10	5	0	11
1	0	11	5	1	12
1	1	10	5	0	11
1	0	11	5	1	12
7	0	11	6	1	12
9	0	10	5	1	11
1	0	10	6	1	10
3	1	10	5	0	10
0	1	10	5	0	10
1	0	12	6	1	13
4	1	12	5	0	13
5	0	10	6	1	13
9	1	10	5	0	11
10	0	10	6	1	10
9	1	11	4	0	12

Exemplo:

- Colocando estes dados no Excel e seguindo os procedimento temos a seguinte saída (editamos um pouco para visualizar melhor):

	Nota 1	Menino 1	Idade 1	Nota 2	Menino 2	Idade 2
Média	5,5	0,5	10,5	5,5	0,5	11,4
Erro padrão	0,8	0,1	0,2	0,2	0,1	0,2
Mediana	7	0,5	10	5	0,5	11
Modo	1	1	10	5	0	11
Desvio padrão	3,6	0,5	0,7	0,8	0,5	1,1
Variância da amostra	12,8	0,3	0,5	0,7	0,3	1,2
Curtose	-1,6	-2,2	0,5	3,9	-2,2	-1,2
Assimetria	-0,3	0,0	1,3	1,4	0,0	0,1
Intervalo	10	1	2	4	1	3
Mínimo	0	0	10	4	0	10
Máximo	10	1	12	8	1	13
Soma	109	10	209	109	10	228
Contagem	20	20	20	20	20	20

Hora de praticar

(item a do exercício da aula 4)

O que sabemos até agora

- Vimos que a estatística nos permite:
  - Estudar uma população a partir de uma amostra
  - Podemos descrever uma amostra a partir de sua média
  - Mas a média não nos diz tudo... Precisamos analisar a variância também
  - A partir da variância, e do desvio-padrão, avaliamos a precisão da média
  - E para isso, calculamos o erro-padrão da média.

Intervalo de confiança

- Com o erro padrão podemos construir o que chamamos de **intervalo de confiança**, que é que um conjunto de valores que a média de uma variável pode assumir com uma certa probabilidade.

- Já sabemos que toda média tem um erro associado devido à variação dos dados e ao tamanho da amostra. O intervalo de confiança nos dará uma ideia melhor de quanto a média que calculamos pode variar.

Intervalo de confiança

- Calculamos os intervalos de confiança somando e subtraindo da média o valor do erro-padrão, multiplicado por um fator que dependerá da probabilidade do intervalo.
  - IC 90%: $[\bar{x} - 1{,}65 \cdot SE \; ; \bar{x} + 1{,}65 \cdot SE]$
  - IC 95%: $[\bar{x} - 1{,}96 \cdot SE \; ; \bar{x} + 1{,}96 \cdot SE]$
  - IC 99%: $[\bar{x} - 1{,}58 \cdot SE \; ; \bar{x} + 1{,}58 \cdot SE]$
- Podemos fazer as contas para qualquer probabilidade, mas estas são as mais usadas.

Nota: estes fatores que multiplicam não vem do nada... São aproximados para uma distribuição Normal. Porém, este detalhe é avançado demais, por enquanto. Para mais detalhes, consulte seu professor.

Exemplo:

- Vamos construir intervalos de confiança para a 95% de probabilidade para nossas amostras de alunos.

- $IC_1: [5{,}5 - 1{,}96 \cdot 0{,}80 \,;\, 5{,}5 + 1{,}96 \cdot 0{,}80] = [4{,}0 \,;\, 7{,}0]$

- $IC_2: [5{,}5 - 1{,}96 \cdot 0{,}18 \,;\, 5{,}5 + 1{,}96 \cdot 0{,}18] = [5{,}2 \,;\, 5{,}8]$

- Então, os ICs dizem que há grandes chances (95%) que a nota média da turma 1 varie de 4,0 a 7,0 e que a nota média da turma 2 varie de 5,2 a 5,8.
- Veja então que a média da amostra 1 pode variar tanto que fica difícil dizer se estamos tratando com uma turma de bom ou de mau desempenho. Já o segundo intervalo nos diz, com mais precisão, que a turma 2 é de alunos com desempenho médio.

Teste de hipótese

- Com todas essas ferramentas podemos testar hipóteses sobre as médias que calculamos.
- Ou seja, a partir da média de uma amostra podemos fazer algumas inferências sobre a média da população.
- Geralmente, estamos interessados em dois tipos de testes:
  - Se a média de uma variável é igual a algum valor
  - Se as médias de duas amostras são iguais

Teste de hipótese

- No caso da amostra de alunos que fizeram uma prova, podemos estar interessados em saber se a nota da população de alunos é igual a 6,0, supondo que esta seja uma nota considerada satisfatória.

- Quando calculamos as médias dos dois grupo vimos que ambas são iguais a 5,5.

- Essa nota média não é igual a 6,0, mas é próxima. Nosso desafio é saber se podemos afirmar que a nota da população de alunos é 6,0.

Teste de hipótese

- Para demonstrar como funcionam os testes de hipótese, mostraremos um procedimento simples, porém aproximado, mas que funciona bem na maioria dos casos.

- O procedimento preciso de como realizar um teste de hipótese envolve conceitos extra, que não vêm ao caso. De qualquer forma, a explicação encontra-se em anexo no final desta aula. Caso tenha interesse, consulte seu professor.

Exemplo:

- Voltando aos intervalos de confiança que calculamos antes, podemos testar nossa hipótese de que os alunos têm nota média satisfatória (6,0) apenas verificando se o valor 6,0 está dentro ou fora do IC:

- IC_1: [4,0 ; 7,0]
- IC_2: [5,2 ; 5,8]

- Veja que na turma 2, a nota 6,0 está acima do intervalo de confiança, então neste caso dizemos que a nota média da turma 2 é estatisticamente diferente de 6,0, com 95% de probabilidade.
- Já para a turma 1, repare que 6,0 está dentro do intervalo de confiança e, neste caso, não podemos afirmar, pela estatística, que a nota da turma seja diferente de 6,0. O teste é inconclusivo.

Teste de hipótese

- Outro teste de hipótese muito útil é o de diferença de média, onde nos perguntamos se as médias de duas amostra são diferentes entre si.

- A mecânica do teste é parecida e, novamente, colocamos em anexo o procedimento mais preciso.

Exemplo:

- Suponha que queremos comparar a nota média de nossas duas turmas com a de uma terceira, que tem as seguintes características:

- Nota média: 4,5
- IC: [4,0 ; 5,0]

- Para fazer o teste, basta verificar se a média desta turma está dentro do intervalo de confiança das outras duas.
- A média 4,5 não está contida no IC da turma dois, então dizemos que as notas dessas duas turma são estatisticamente diferentes entre si.
- Esta média no entanto está contida no IC da turma um, então não podemos afirmar que as notas dessas duas turmas sejam diferentes entre si.

Do teste de hipótese para a avaliação econômica

- Os testes de hipóteses serão úteis para avaliação de impacto, entre outras coisas, na descrição dos grupos de tratamento e controle.

- Como já discutimos, o importante para a avaliação de impacto é comparar o grupo de tratamento com um grupo mais parecido possível.

- Usaremos o teste de hipótese para fazer esta comparação e decidir se o grupo de controle que escolhemos é um bom contrafactual para o grupo de tratamento.

Do teste de hipótese para a avaliação econômica

- Descobrir se tratados e controles são parecidos ou não é o primeiro passo para qualquer avaliação de impacto, pois será determinante na escolha do método a ser aplicado.

- Como você verá na próxima aula, se os grupos de tratamento e controle forem parecidos (em caso de aleatorização, por exemplo), podemos calcular o impacto do programa de forma mais simples.

- Mas, se detectarmos diferenças entre os dois grupos, precisaremos levar em conta estas diferenças para estimar o impacto do programa.

- Por enquanto, vamos ver como descrever dois grupos.

Como testar diferenças de médias no Excel

Usando novamente o suplemento de análise de dados, podemos construir intervalos da seguinte forma:

1. Com a base de dados aberta, clique na guia "Dados" > "Análise de Dados".
2. Na janela que aparece, clique em "Estatística descritiva" > "OK".
3. Na nova janela, em "Intervalo de dados" selecione as células de todas as variáveis que queremos descrever (incluindo os seus títulos).
4. Clique em "Rótulos na primeira linha".
5. Clique em "Nível de confiabilidade p/ a média".
6. Clique "OK".

Exemplo:

- Usando a mesma base de dados para os alunos das duas turmas, temos a seguinte saída:

	Nota 1	Menino 1	Idade 1	Nota 2	Menino 2	Idade 2
Nível de confiança(95,0%)	1,67	0,24	0,32	0,39	0,24	0,51

- O Excel reporta o valor que deve ser somado e subtraído da média para criar o intervalo de confiança. Fazendo as contas:

	Nota 1	Menino 1	Idade 1	Nota 2	Menino 2	Idade 2
Nível de confiança(95,0%)	[3,78 ; 7,12]	[0,26 ; 0,74]	[10,13; 10,77]	[5,06 ; 5,84]	[0,26 ; 0,74]	[10,89; 11,91]

Exemplo:

- Com os intervalos de confiança podemos testar se as duas turmas têm características diferentes em média.

	Nota 1	Menino 1	Idade 1	Nota 2	Menino 2	Idade 2
Média	5,5	0,5	10,5	5,5	0,5	11,4
Nível de confiança(95,0%)	[3,78 ; 7,12]	[0,26 ; 0,74]	[10,13; 10,77]	[5,06 ; 5,84]	[0,26 ; 0,74]	[10,89; 11,91]

- Comparando as médias de uma turma com os ICs da oura turma, podemos ver que as duas turmas não são estatisticamente diferentes em termos de *nota* nem de *gênero* do alunos.
- Porém, é possível ver que as idades médias dos alunos das duas turmas são estatisticamente diferentes entre si, sendo que os alunos da turma 2 são quase um ano mais velhos, em média.

Hora de praticar

(item b do exercício da aula 4)

ANEXO - Teste de hipótese – procedimento mais preciso

Voltando ao exemplo em que queremos testar se a nota média é diferente de 6,0.

- Primeiramente, temos que formular nossa hipótese, que chamaremos de hipótese nula:
  - $H_0: nota\ média = 6{,}0$
- O procedimento do teste dará a probabilidade de esta hipótese ser verdadeira, esta probabilidade se chama p-valor
  - P-valor: probabilidade de uma hipótese nula ser verdadeira
- Assim, quanto menor o p-valor maior a chance de que a nota média seja diferente de 6,0.

Nota: encontrar o p-valor envolve o cálculo de estatísticas de teste, com base em distribuições teóricas etc. Porém, são detalhes avançados demais, por enquanto. Para mais detalhes, consulte seu professor.

ANEXO - Exemplo:

- No nosso exemplo, encontramos (usando Excel) os seguintes p-valores para os dois grupos de alunos:

- P-valor 1: 0,50
- P-valor 2: 0,01

- Ou seja, concluímos que na primeira amostra há uma chance de 50% da nota média ser igual a 6,0. Isso é o mesmo que dizer que há 50% de chance da nota média ser diferente de 6,0.
- Já no segundo caso, há uma chance de 1% da nota média ser igual a 6,0, ou seja, 99% de chance da nota média ser diferente de 6,0.
- O que podemos concluir, então?
- Para chegar à conclusão, precisamos de uma regra de decisão.

ANEXO - Teste de hipótese – procedimento mais preciso

- Regra de decisão consiste em adotar uma probabilidade que consideramos razoável para aceitar ou rejeitar a hipótese nula.
- Um padrão entre o economistas é que rejeitamos as hipóteses nulas caso o p-valor seja menor que 5% ou 0,05.
  - Regra de decisão: se p-valor < 0,05, então rejeitamos a hipótese.

- Caso o p-valor seja maior que 0,05 então podemos aceitar a hipótese?
- Não. A rigor, quando o p-valor é maior que 0,05 o teste de hipótese é inconclusivo, ou seja, não podemos afirmar, estatisticamente, que a média seja de fato igual ou diferente da hipótese.

ANEXO - Exemplo:

- Voltando ao nosso exemplo, podemos tomar as seguintes decisões:
- Amostra 1: como p-valor > 0,05, o teste é inconclusivo, não podemos afirmar se a nota média é igual ou diferente de 6,0.
- Amostra 2: como p-valor < 0,05, concluímos que a nota média é estatisticamente diferente (menor) que 6,0.
- Ou seja, nossa coleta de dados diz que a turma 2 deve ter desempenho médio abaixo do satisfatório. Já para a turma 1 não podemos concluir o mesmo, com os dados coletados.

ANEXO - Teste de hipótese – procedimento mais preciso

- No caso de um teste de diferença de médias, o procedimento é muito parecido com o teste de hipótese anterior.
- A principal diferença é como definimos a hipótese nula.
- Se queremos testar é se duas amostras são iguais em uma variável, definimos a hipótese nula como:
  - $H_0: \bar{x}_1 - \bar{x}_2 = 0$
- De resto, o teste segue o mesmo caminho, encontramos o p-valor para rejeição da hipótese e usamos a regra de decisão.

ANEXO - Exemplo:

- Se quisermos testar se as duas turmas tem desempenho igual ou diferente, encontramos (usando o Excel) encontramos o seguinte p-valor:

- P-valor da diferença: 1,00

- Pela regra de decisão (p-valor > 0,05), não podemos afirmar que as notas das duas turmas sejam estatisticamente diferentes.
- Repare inclusive que o p-valor nos diz que a probabilidade de as médias serem iguais é 100%, ou seja, é realmente difícil dizer que as duas turmas tenham desempenho diferente!