Ensino de Estatística na Escola Básica – 2º. Sem. Letivo de 2018

Produção de Dados

I) Conceitos Preliminares:

- **Mensuração**:

Atribuição de numerais (por exemplo: 3, III, 11) a objetos de acordo com certas regras.

- **Classificação:**

Atribuição de categorias ou classes a objetos de acordo com certas regras.

- **Escalas de Mensuração / Classificação**[1]**:**

Escala de Mensuração	Regras Matemáticas permitidas
1. Nominal	- correspondência um a um
2. Ordinal	- correspondência um a um - relações de ordem com transformação monotônica
3. Intervalar	- correspondência um a um - atribuição de postos (ordenação ou "ranking") - igualdade de diferenças
4. Razão	- correspondência um a um - atribuição de postos (ordenação ou "ranking") - igualdade de diferenças - divisão e multiplicação

1. e 2. são usualmente chamados de **dados categóricos** ou dados qualitativos.

[1] Stevens, S.S. 1951. Mathematics, measurement and psychophysics. In Stevens, S.S. (ed.), *Handbook of Experimental Psychology*. New York: Wiley.

- **Escalonamento**

Tem como objetivo melhorar o nível da escala de mensuração dos dados e capturar toda a informação contida nos mesmos[2].

Ex: dados em escala nominal transformados em ordinal, intervalar ou razão. Dados em escala intervalar ou razão em geral não necessitam de escalonamento. Dados ordinais podem ter um escalonamento ótimo.

- **Classificação de dados categóricos (Nishisato, 2007)**

1. Dados de incidência
2. Dados de dominância

1. Dados de Incidência

Seus *elementos* são a ausência (0) ou presença (1) de um atributo, que nos dão as frequências de tais atributos.

1.1 Tabelas de Contingência (bidimensional)

Os dados representam as frequências conjuntas de **dois** conjuntos de categorias.

Exemplo: Tipos de laxantes e efeitos

	Efeito				
Laxante	Nenhum	Leve	Adequado	Demasiado	TOTAL
A	0	3	6	21	30
B	5	15	9	1	30
C	2	18	10	0	30
TOTAL	7	36	25	22	90

[2] Nishisato, S. 2007. Multidimensional nonlinear descriptive analysis. Boca Raton: CRC Press.

1.2 Dados de Múltipla Escolha

Extensão da tabela de contingência bidimensional, usada para mais de duas variáveis categóricas. Difícil representação em tabelas com mais de três variáveis categóricas e maior possibilidade de células em branco. É preferível utilizar a forma de matriz indicadora (*respondentes ou sujeitos* por categorias de perguntas de múltipla escolha – *objetos,* mutuamente exclusivas e exaustivas).

Ex: **Pergunta1**: Faixa Etária [20-30; 31-40; 41 +]

Pergunta2: Você concorda com nova lei de porte de armas? [sim; não]

Pergunta3: Em que região você mora? [A; B; C; D]

Exemplo de matriz indicadora com os dados:

Obj.	20-30	31-40	41+	Sim	Não	A	B	C	D
1	0	1	0	1	0	1	0	0	0
2	0	0	1	1	0	0	0	1	0
....									
n	1	0	0	0	1	0	1	0	0

1.3 Dados de Separação ("Sorting Data")

Não são largamente utilizados, mas há situações em que são necessários.

Exemplo: Sete disciplinas: A= Inglês; B= História; C= Matemática; D= Física; E=Psicologia; F= Biologia; G= Educação. A um número de alunos ou *sujeitos* (n = 8, por exemplo) é solicitado que comecem com o numeral 1 em qualquer disciplina e então continuar com este numeral para todas as disciplinas similares. Passar então para o numeral 2 e repetir a operação até que

todas as disciplinas estejam agrupadas em “pilhas” de similaridade. A decisão sobre o número de “pilhas” e sobre o tamanho das mesmas é arbitrário (não há restrições quanto ao julgamento).

Disciplinas	Alunos							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
A	1	1	2	3	4	3	1	2
B	1	2	2	3	3	3	1	1
C	2	3	1	2	2	2	2	3
D	2	3	1	2	2	2	2	3
E	3	4	2	1	1	3	1	4
F	4	4	2	2	5	1	2	5
G	1	1	2	1	1	3	1	1
No Cat.	**4**	**4**	**2**	**3**	**5**	**3**	**2**	**5**

Pergunta:

Como representar as respostas para a linha A: [11234312] em termos de matriz indicadora com as linhas sendo as disciplinas e as colunas sendo os respondentes e suas respectivas categorias?

Resposta:

[(1000), (1000), (01), (001), (00010), (001), (10), (01000)]

2. Dados de dominância

Aqui os *elementos dos dados* são as mensurações ordinais e o objetivo da quantificação é distinto daquele para os dados de incidência.

2.1 Dados de comparação par-a-par

Os respondentes (*sujeitos*) decidem para cada par apresentado qual a preferência ou qual elemento do par é mais importante. Para dois *objetos* (X_j, X_k) a resposta do *sujeito i* é codificada assim:

${}_if_{jk}$= 1 se $X_j > X_k$

0 se $X_j = X_k$

-1 se $X_j < X_k$

É também comum usar a codificação 1, 2 e 0, para preferência pelo primeiro elemento, segundo elemento e ausência de preferência, respectivamente.

Exemplo: Comparação par-a-par entre ***quatro*** frutas, feita por ***cinco*** indivíduos, para os pares: A: (maçã, pera), B: (maçã, manga), C: (maçã, uva), D: (pera, manga), E: (pera, uva), F: (manga, uva), utilizando a codificação 1, 2 e 0 vista acima.

Sujeitos	Pares de frutas					
	A	B	C	D	E	F
1	1	2	2	1	0	1
2	2	2	2	1	1	1
3	2	2	1	1	0	1
4	2	2	2	2	2	2
5	1	1	1	1	2	2

2.2 Dados ordenados por postos

Respostas codificadas como 1, 2, 3, etc., onde "1" indica a primeira escolha ou a mais preferida, e o numeral maior corresponde à última escolha.

Exemplo: Cinco gerentes (A, B, C, D, E) classificaram sete pretendentes a uma vaga de emprego após entrevista. No caso de empate, utiliza-se o posto médio. Por exemplo, se dois candidatos são as primeiras preferências, a eles é dados o posto 1,5. Se os três primeiros estão empatados, cada um recebe o posto 2. Desta forma, a soma de todos os postos é fixa, sendo igual a n(n+1)/2

	1	2	3	4	5	6	7
A	3	6	5	4	1	7	2
B	2	7	5	4	3	6	1
C	2	5	6	3	4	7	1
D	3	7	4	5	1	6	2
E	4	6	7	5	2	3	1

2.3 Dados com categorias sucessivas

Em essência, o mesmo que dados de incidência do tipo múltipla-escolha, só que o mesmo conjunto de categorias sucessivas ordenadas é utilizado para o julgamento de todas as perguntas (como é muito utilizado em "*surveys*" sobre atitudes, satisfação, etc.).

Exemplo: Considere o seguinte conjunto de categorias: 1 = Baixa; 2 = Média; 3 = Alta. Cinco indivíduos são questionados quanto à motivação para: fazer exercícios físicos, fazer dieta, fazer tratamento clínico contra obesidade, fazer tratamento cirúrgico contra obesidade.

Sujeitos	Atividade			
	Ex. Fís.	Dieta	Trat. Clín.	Trat. Cir.
1	1	1	3	3
2	1	1	2	3
3	2	1	3	3
4	2	2	2	2
5	1	1	2	2

Por que não nos referimos a esta tabela como contendo dados de múltipla escolha apenas?

Podemos fazer o escalonamento dos "tratamentos" (atividades), através de algum critério, mas também os limites (fronteiras) das categorias, um entre "Baixa" e "Média" e outro entre "Média" e "Alta".

Sendo assim, teremos os dados convertidos a dados *ordenados* por postos, tanto para os limites das categorias quanto para os "tratamentos". Esta é a razão para considerarmos este caso especial de dados de múltipla-escolha como dados de dominância.

Leitura complementar: *A Matemática da Escolha Social*, de Steffenon e Jabuinski (arquivo pdf anexo na página do Prof. Lupércio).

- **Origem dos dados:**

Os dados podem ser coletados pelo pesquisador, de acordo com seus objetivos, constituindo o que se costuma chamar de **dados primários**.

Se os dados utilizados pelo pesquisador não foram coletados tendo como objetivo o estudo realizado pelo mesmo, temos o que se costuma chamar de **dados secundários**.

II) Produção de dados

Uma pesquisa sempre apresenta duas etapas:

- coleta de dados
- análise de dados

Através da identificação do que é gerado nas duas etapas acima podemos entender melhor a natureza da pesquisa realizada / relatada.

- **A coleta de dados**

Os dados coletados podem ser armazenados de forma **estruturada** ou **não-estruturada**.

Os conjuntos de dados **estruturados** são em geral armazenados na forma mostrada na Tabela a seguir, constituindo o que se costuma chamar de “dados em estrutura de casos” (*dataframe*):

		Variáveis			
		Sexo	Idade	ENEM	Classe Social
Casos	Candidato 1	M	17	973,6	A
	Candidato 2	M	19	822,5	C
	Candidato 3	F	18	843,5	B
	Candidato 4	M	21	830,3	C
	Candidato 5	F	20	789,2	D
	---	---	---	---	---

Comparar

Observe que os dados utilizados para preencher as células da estrutura de casos acima podem ter sido originados de um questionário (instrumento de pesquisa), de uma entrevista

estruturada ou semi-estruturada, de observação, de registros obtidos de outras fontes sobre cada candidato, e de muitas outras formas.

Observe também que os casos não necessariamente são pessoas, podendo ser classes de uma escola, escolas de um município, estados, países, anos, etc.. Daí serem chamados também de **unidades de análise**, ou seja, 'os objetos sobre os quais se coleta informação.'

Uma 'pesquisa do tipo *survey*' deve produzir dados que possam ser registrados da forma estruturada acima descrita.

- **Os métodos usuais de pesquisa**

Em geral, a 'Pergunta de Pesquisa' pode ser respondida com base nos métodos abaixo:

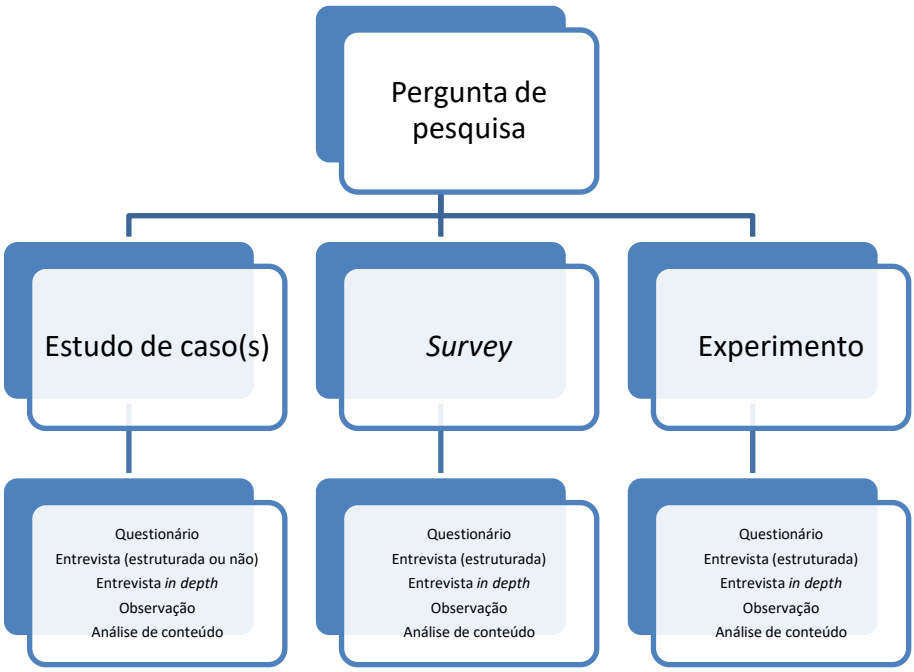
- Estudo de caso(s)

- Pesquisa do tipo *survey*

- Experimento

Veja a seguir um esquema dos métodos mais usuais de pesquisa e suas respectivas técnicas de coleta de dados[3]:

[3] De Vaus, D.A. 2002. *Surveys in Social Research.* Crows Nest: Allen & Unwin.



1) Estudo de caso(s):

Foca no(s) caso(s) e tenta obter uma compreensão detalhada do(s) mesmo(s). Não necessariamente compara os casos.

2) *Survey*:

Coleta dados de forma sistemática sobre vários objetos (casos), tornando-os comparáveis, já que a mesma informação é coletada para todos os casos.

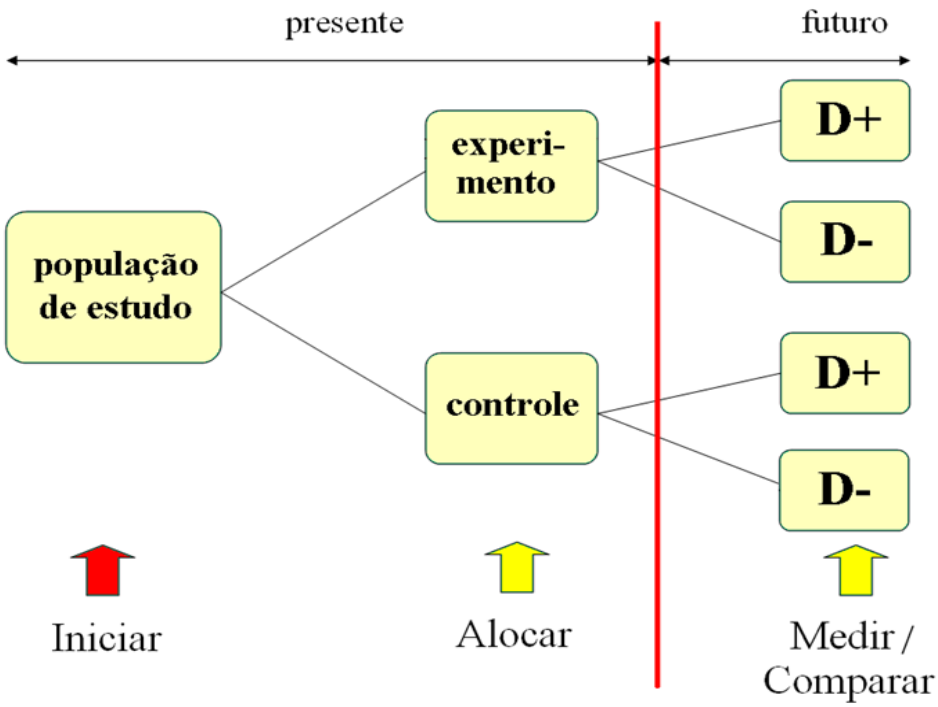
3) Experimento:

Similar ao método *survey*, mas a variação entre os atributos dos objetos é criada por intervenção, de forma aleatória. Aqui o pesquisador está interessado em verificar se uma intervenção leva a uma diferença.

Exemplo: Um experimento poderia se iniciar com duas turmas da mesma série e do mesmo turno, consideradas similares, com a única diferença que uma turma recebe a intervenção ou

‘tratamento’ (por exemplo, uma nova metodologia de ensino para a média aritmética) e a outra tem o ensino de média aritmética pela metodologia tradicional. Desta forma, qualquer diferença entre os resultados de uma avaliação sobre este conteúdo específico deve ser devida à intervenção. Por outro lado, um estudo do tipo *survey* não iria criar a variação, mas apenas observar a variação natural, ou seja, encontrar alunos que não sofreram a intervenção e tiveram boas notas e compará-los com aqueles que tiveram boas notas, mas que sofreram a intervenção.

Veja a Figura a seguir para uma descrição visual da pesquisa experimental, onde D pode ser entendido como o ‘desempenho’, ou seja, o desfecho que se deseja avaliar:



- **A análise de dados:**

O que distingue os métodos de análise de dados é justamente a lógica por trás da análise. Em pesquisa do tipo *survey*, por

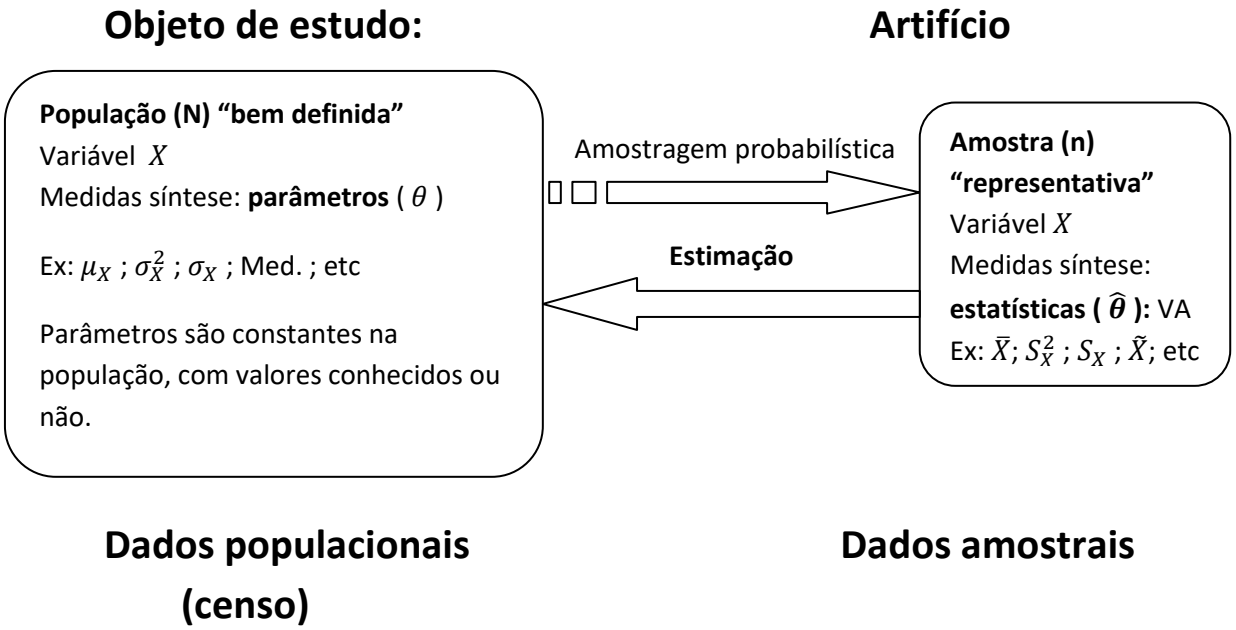
exemplo, a lógica é de que a variação em uma variável é acompanhada por variações sistemáticas em outras variáveis. Além disso, pode-se estar interessado em análise causal (se uma variável afeta a outra).

A lógica das outras formas de pesquisa já foram brevemente citadas anteriormente.

Atividade número 7

III) Inferência Estatística: a ideia geral

- **Estimação:**



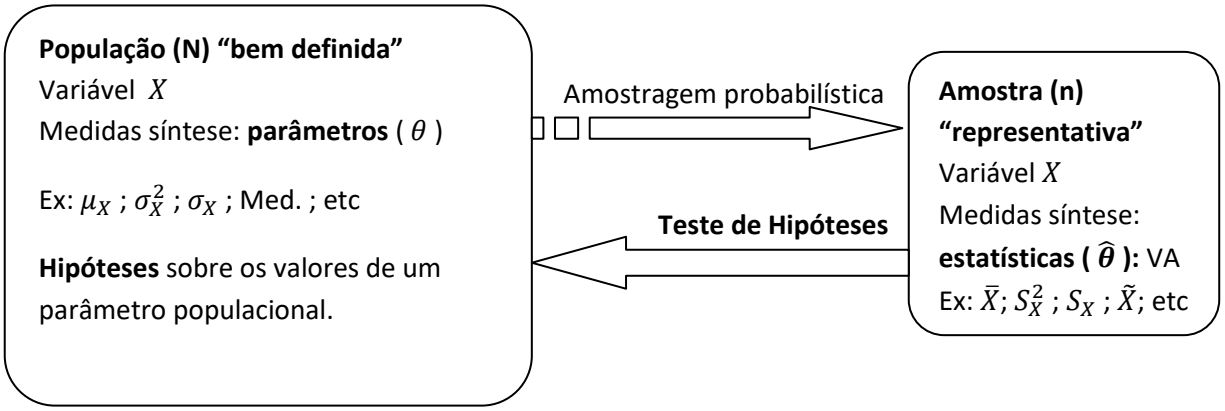
Para estimarmos o valor de um parâmetro θ desconhecido, precisamos de teorias sobre as estatísticas $\hat{\theta}$, denominadas

estimadores do respectivo parâmetro. Devemos conhecer o "comportamento" da distribuição da VA considerada:

$$\begin{cases} -formato \\ -tendência\ central \\ -dispersão \end{cases}$$

- **Teste de hipóteses:**

Objeto de estudo: Artifício



Dados populacionais (censo) **Dados amostrais**

Para testarmos hipóteses sobre o valor de um parâmetro θ desconhecido, precisamos de teorias sobre as estatísticas $\widehat{\theta}$, denominadas estimadores do respectivo parâmetro. Devemos, da mesma forma, conhecer o "comportamento" da distribuição da VA considerada.