

AULA 01

Slides Extras

Ernesto F. L. Amaral

16 de julho de 2012

Análise de Regressão Linear (MQ 2012)

www.ernestoamaral.com/mq12reg.html

Fonte:

Babbie, Earl. 1999. “Métodos de Pesquisas de *Survey*”. Belo Horizonte: Editora UFMG. Capítulos 3-5 (pp.77-158).

Triola, Mario F. 2008. “Introdução à estatística”. 10^a ed. Rio de Janeiro: LTC. Capítulo 1 (pp.2-31).

Wooldridge, Jeffrey M. 2008. “Introdução à econometria: uma abordagem moderna”. São Paulo: Cengage Learning. Capítulo 1 (pp.1-17).

ESTRUTURA DA AULA

- 1) **Pesquisa de survey como método das ciências sociais:**
Babbie (cap.3)
- 2) **Tipos de desenhos de pesquisa:** Babbie (cap.4)
- 3) **Tipos de dados:** Triola (cap.1)
- 4) **Tipos de bancos de dados:** Babbie (cap.4); Wooldridge (cap.1)
- 5) **Econometria:** Wooldridge (cap.1)
- 6) **A lógica da amostragem do survey:** Babbie (cap.5)
- 7) **Amostragem probabilística:** Babbie (cap.5)
- 8) **Amostragem não-probabilística:** Babbie (cap.5)
- 9) **Sejamos críticos:** Triola (cap.1)

1) Pesquisa de *survey* como método das ciências sociais

Babbie, 1999: 77-91 (capítulo 3)

PESQUISA DE *SURVEY*

- Babbie (1999) aborda o método de pesquisa com aplicação de questionários (pesquisa de *survey*).
- Esse método pode ser aplicado a uma grande quantidade de desenhos de pesquisa.
- O método consiste em aplicar um questionário a uma amostra da população de interesse.
- As entrevistas podem ser pessoais, por telefone ou por correio.
- As respostas são codificadas de forma padronizada e registradas de forma quantitativa em programas computacionais.
- Os dados são analisados, correlacionados e generalizados para a população.

CARACTERÍSTICAS CIENTÍFICAS DO *SURVEY*

- É um tipo de ferramenta disponível aos pesquisadores sociais.
- Os métodos de *survey* podem não ser apropriados para determinados tópicos de pesquisa.
- De todo modo, o *survey* pode ser usado em combinação com outros métodos.
- Esse tipo de pesquisa tem função pedagógica, porque suas deficiências ficam mais claras do que em outros métodos, permitindo avaliações mais conscientes de suas implicações.

PESQUISA DE *SURVEY* É LÓGICA

- Mesmo que haja comportamentos racionais e irracionais, os cientistas sociais devem procurar entender o comportamento social racionalmente.
- Os dados de *survey* facilitam a aplicação cuidadosa do pensamento lógico.
- O formato da pesquisa de *survey* permite desenvolvimento e teste rigorosos de explicações lógicas.
- O exame de uma grande quantidade de entrevistas permite testar proposições complexas que envolvam diversas variáveis em interação simultânea.

PESQUISA DE *SURVEY* É DETERMINÍSTICA

- Todo evento ou situação (resultado) tem razões determinadas que o antecederam (causa).
- Sempre que o pesquisador de *survey* procura explicar causas e razões de eventos, características e correlações observadas, a investigação deve assumir postura determinística.
- O sistema determinístico de causa e efeito é clarificado no *survey*, porque é possível elaborar um modelo lógico claro.
- A disponibilidade de numerosos casos e variáveis permite ao analista explicar uma variável (dependente) com base em muitas informações (variáveis independentes).
- É possível ainda analisar efeito de variáveis intervenientes.

PESQUISA DE *SURVEY* É GERAL

- Ao buscar observar e entender padrões gerais de eventos e correlações entre acontecimentos, a teoria social busca a generalização de suas conclusões.
- Pesquisas amostrais são realizadas para entender a população maior da qual a amostra foi selecionada.
- Análises descritivas e explicativas em pesquisas de *survey* visam desenvolver proposições gerais sobre o comportamento humano.
- Grande número de casos permite que achados possam ser replicados entre subconjuntos da amostra.
- Relato cuidadoso da metodologia de um *survey* facilita réplicas posteriores por outros pesquisadores e/ou para outros subgrupos, permitindo testes e retestes.

PESQUISA DE *SURVEY* É PARCIMONIOSA

- O cientista social busca explicar o maior número de fenômenos com o menor número de informações.
- Como um *survey* tem grande número de variáveis, é possível analisar importância relativa de cada uma.
- Ou seja, busca-se obter o máximo de compreensão com menor número de variáveis.
- As variáveis explicativas não precisam ser definidas a priori.
- É possível coletar uma grande quantidade de informações que serão usadas em diferentes modelos explicativos, com o objetivo de selecionar o mais adequado.

PESQUISA DE *SURVEY* É ESPECÍFICA

- É preciso definir exatamente quais são os métodos e conceitos utilizados na pesquisa social.
- A medição de cada variável no *survey* é construída a partir de respostas específicas a itens específicos de questionário, o qual é codificado e quantificado de forma específica.
- Ao explicitar como métodos, medidas e conceitos foram elaborados, o *survey* permite ser avaliado e criticado.
- Se variáveis não são conceituadas e medidas corretamente, correlações observadas podem não fazer sentido.
- A superficialidade e aproximações de todo tipo de pesquisa científica são mais aparentes e explícitas no *survey*.
- É mais difícil saber como conceitos são construídos e medidos na pesquisa qualitativa.

OUTRAS CARACTERÍSTICAS DA PESQUISA DE *SURVEY*

- *Survey* é empiricamente verificável: proposições, teorias e hipóteses podem ser testadas na prática (no mundo real).
- *Survey* é intersubjetivo: detalhamento de métodos e técnicas utilizados permite que outro cientista replique estudo e alcance resultados semelhantes.
- *Survey* é aberto a modificações: coleta e quantificação permite fonte permanente de informações e nova análise dos dados quando há modificações de teorias.

COMPARAÇÃO DO *SURVEY* COM OUTROS MÉTODOS

- Uma investigação abrangente será beneficiada se utilizar métodos diferentes focados em um tópico.
- A pesquisa de *survey* pode se beneficiar de conhecimentos de outros métodos:
 - Experimentação.
 - Análise de conteúdo.
 - Observação participante.

PESQUISA DE *SURVEY* E EXPERIMENTAÇÃO

- Na lógica do experimento controlado (grupos de tratamento e controle) é realizado emparelhamento sócio-econômico de indivíduos, de forma que única diferença entre grupos é o recebimento da política.
- Na pesquisa de *survey*, o pesquisador também realiza o controle de variáveis, permitindo análises de subgrupos.
- Nos dois casos, o objetivo é de excluir influência de variáveis secundárias.

PESQUISA DE *SURVEY* E ANÁLISE DE CONTEÚDO

- A codificação de respostas de questionários é uma forma de análise de conteúdo.
- Se forem coletadas perguntas abertas, o pesquisador de *survey* precisará categorizar as respostas posteriormente.
- Por isso, os pesquisadores de *survey* devem utilizar as experiências e métodos de analistas de conteúdo.

PESQUISA DE *SURVEY* E OBSERVAÇÃO PARTICIPANTE

- Pesquisas de *survey* podem utilizar as técnicas desenvolvidas por pesquisadores que utilizam observação participante:
 - Obtenção de informações sobre interações inconscientes entre pessoas (*rappport*).
 - Manutenção da neutralidade quando a atividade de campo é realizada.
 - Realização de observações precisas ao coletar os dados.

CIÊNCIAS SOCIAIS SÃO CIÊNCIAS DA COMPLEXIDADE

- Problemas de amostragem e de generalização são maiores no estudo do comportamento social.
- A pesquisa de *survey* é excelente para desenvolvimento de métodos úteis e de entendimento amplo.
- A medição de conceitos polêmicos (preconceito, por exemplo) pode ser de difícil aplicabilidade.
- A simples presença de pesquisadores no campo pode afetar entrevistados.
- Alguns participantes podem formar opiniões no momento em que são entrevistados.
- É preciso ter compreensão mais sofisticada sobre medição, conceituação, probabilidade e complexidade de correlações e causalidades nas ciências sociais do que nas naturais.

2) Tipos de desenhos de pesquisa

Babbie, 1999: 93-111 (capítulo 4)

TIPOS DE DESENHOS DE PESQUISA

- Há uma tendência a considerar a análise de dados de *survey* como mais desafiadora e interessante do que o desenho do *survey* e a coleta de dados.
- Durante a análise, começa-se a ganhar entendimento do assunto estudado e pode-se compartilhar as descobertas com colegas.
- No entanto, é preciso que haja a correta elaboração do desenho de pesquisa e sua execução, para que não haja problemas na etapa da análise.
- Portanto, o desenho é igualmente desafiador e importante.
- Há vários tipos de pesquisa de *survey*, com objetivos, custos, tempo e escopo diferenciados.

FINALIDADES DA PESQUISA DE *SURVEY*

- Uma pesquisa de *survey* pode ser realizada para atender diferentes objetivos:
 - Intenção de voto.
 - Venda de produtos.
 - Projetar infra-estrutura.
 - Modificar programa social.
- De uma forma geral, três objetivos gerais permeiam os interesses de pesquisas de *survey*:
 - Descrição.
 - Explicação.
 - Exploração.

DESCRIÇÃO

- *Surveys* são frequentemente realizados para descobrir a distribuição de certos traços e atributos da população, permitindo análise descritiva.
- Há preocupação em **como** é a distribuição de determinada variável (descrição), e não no **porquê** (explicação).
- Exemplos de dados descritivos:
 - Distribuições de candidatos por idade e sexo.
 - Taxas de aprovação do governo.
 - Percentual de políticos na escala esquerda/direita.
- Além de descrever a amostra total (e inferir para a população total), pesquisadores descrevem subgrupos.

EXPLICAÇÃO

- Muitas pesquisas de *survey* têm o objetivo adicional de fazer análises explicativas sobre a população, procurando relações de causa e efeito.
- Explicar quase sempre requer análise multivariada, que é a análise do impacto simultâneo de duas ou mais variáveis (independentes) sobre outra variável (dependente).
- Há uma preocupação em explicar o porquê de determinada distribuição da variável, geralmente utilizando modelos estatísticos.
- Por exemplo, podemos querer explicar escala esquerda/direita de deputados por partido de pertencimento, Estado de residência, partido do governador...

EXPLORAÇÃO

- Pesquisadores podem fazer pesquisa sobre um tema que está começando a ser investigado, e não se sabe exatamente suas causas e consequências.
- Primeiramente, pode ser realizada entrevista em profundidade (qualitativa) para captar o máximo de informações possíveis do público em estudo:
 - Não é necessário ter amostra representativa.
 - Não é preciso coletar dados com questionário padronizado.
- Posteriormente, as informações são analisadas para melhor entender os indivíduos, sem preocupação explicativa, mas visando a elaboração de questionário padronizado.

UNIDADES DE ANÁLISE

- Em uma pesquisa de *survey*, dados são coletados para descrever cada unidade que está sendo analisada.
- As unidades de análise são tipicamente pessoas, mas podem ser domicílios, bairros, municípios, Estados, países, empresas, universidades...
- Os dados são agregados e manipulados, permitindo descrever a amostra estudada, além de realizar análises explicativas.
- Um mesmo *survey* pode envolver mais de uma unidade de análise, como é o caso da Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE), que coleta dados de pessoas e domicílios.

FALÁCIA ECOLÓGICA

- Unidades de análise num *survey* podem ser descritas com base em seus componentes.
- Se o objeto da pesquisa for descrever os municípios e agregar as várias descrições para descrever todos municípios, a unidade básica de análise é o município.
- Pesquisador pode selecionar unidade de análise incorreta, incorrendo em falácia ecológica. Por exemplo:
 - Ao analisar taxas de criminalidade por raça, o correto seria ter dados por indivíduos.
 - Se utilizarmos dados municipais, poderíamos concluir que negros possuem maiores taxas de criminalidade.
 - No entanto, os crimes podem estar sendo realizados por brancos que residem em municípios com maioria negra.

IDENTIFICAÇÃO DAS UNIDADES DE ANÁLISE

- É importante que as unidades de análise sejam identificadas no momento da montagem dos bancos de dados.
- O mais seguro é que em pesquisas que possuam mais de uma unidade de análise, sejam criados arquivos separados de dados para cada unidade.

3) Tipos de dados

Triola, 2008: 2-31 (capítulo 1)

TIPOS DE DADOS

- **Dados** são observações coletadas de um determinado grupo de interesse.
- **Dados quantitativos** são números que representam contagens ou medidas (renda, anos de escolaridade...).
 - **Dados discretos** são aqueles em que o número de valores possíveis são finitos ou “enumeráveis” (número de cômodos em um domicílio...).
 - **Dados contínuos** resultam de infinitos valores possíveis em uma escala contínua (renda per capita...).
- **Dados qualitativos** (ou categóricos ou de atributos) podem ser separados em diferentes categorias que se distinguem por alguma característica não-numérica (sexo, ideologia política).

NÍVEIS DE MEDIÇÃO

- Variáveis sociais possuem diferentes níveis de medição.
- **Nominal:** distingue as categorias que compõem uma variável (sexo, religião, região de residência...). As categorias da variável nominal são mutuamente excludentes.
- **Ordinal:** as categorias de uma variável são ordenadas em uma escala (classe social, religiosidade, alienação...). Os números têm significado somente de indicação de ordem.
- **Intervalo:** usa números para descrever uma variável e distâncias entre pontos têm significado real. Diferença entre 20 e 40 graus Fahrenheit é a mesma que entre 60 e 80. Mas 40 não é necessariamente duas vezes mais quente que 20.
- **Razão:** é o mesmo que a medição de intervalo, mas tem zero real. Uma pessoa de 20 anos tem dobro de idade de uma pessoa de 10 anos.

NÍVEIS DE MENSURAÇÃO DE DADOS

- Nível **nominal** de mensuração possui dados que informam nomes, rótulos ou categorias:
 - Os dados não são ordenados e não devem ser usados para cálculos de médias.
 - Raça e código postal, por exemplo.
- Nível **ordinal** de mensuração engloba dados que podem ser organizados em alguma ordem:
 - Sabemos que há diferenças relativas entre os valores dos dados, mas não sabemos as magnitudes das diferenças.
 - Na escala de frequência (pouco/médio/muito), é possível ordenar os dados, mas não sabemos se a diferença entre “pouco” e “médio” é o mesmo que “médio” e “muito”.

NÍVEIS DE MENSURAÇÃO DE DADOS (cont.)

- Nível **intervalar** de mensuração é similar ao ordinal, mas sabemos as magnitudes das diferenças entre dois valores:
 - Os dados não possuem um ponto inicial zero natural.
 - Sabemos as magnitudes das diferenças entre os anos censitários (1970, 1980, 1991 e 2000), mas o tempo não começou em zero.
- Nível de mensuração de **razão** é similar ao intervalar, mas há um ponto inicial zero natural:
 - Como há um zero que indica nenhuma quantidade, é possível dizer que uma quantidade é maior que outra em X vezes (razões significativas).
 - 30 anos de idade é 6 vezes maior do que 5 anos de idade, por exemplo.

RESUMO DOS NÍVEIS DE MENSURAÇÃO DE DADOS

Nível	Resumo	Exemplo
Nominal	Apenas categorias. Os dados não podem ser arranjados em um esquema de ordem. Há categorias ou nomes apenas.	Município de residência.
Ordinal	As categorias são ordenadas, mas as diferenças não podem ser encontradas ou não têm significado.	Frequência à igreja: pouco, médio, muito.
Intervalar	As diferenças são significativas, mas não existe ponto inicial zero natural e as razões não têm sentido.	Ano censitário (não há tempo zero).
Razão	Há um ponto inicial zero natural e as razões são significativas.	Taxa de desemprego.

4) Tipos de bancos de datos

Babbie, 1999: 93-111 (capítulo 4)

Wooldridge, 2008: 1-17 (capítulo 1)

DESENHOS BÁSICOS DE *SURVEY*

- Após especificar os objetivos e unidades de análise da pesquisa, é preciso escolher entre diversos desenhos diferentes:
 - *Surveys* interseccionais (*cross-sectional*).
 - *Surveys* longitudinais (tendências, coortes ou painel).
 - *Surveys* interseccionais servindo como longitudinais.
- Wooldridge (2008) classifica os dados econômicos em:
 - Dados de corte transversal = *surveys* interseccionais.
 - Cortes transversais agrupados = estudos de tendências.
 - Dados de séries de tempo = estudos de coortes.
 - Dados de painel ou longitudinais = estudos de painel.

DADOS DE CORTE TRANSVERSAL (Wooldridge)

***SURVEYS* INTERSECCIONAIS (Babbie)**

- Um conjunto de dados de corte transversal consiste em uma amostra de uma unidade de análise, tomada em um determinado ponto no tempo.
- Esses dados são muito utilizados em economia e em outras ciências sociais.
- Dados em um determinado ponto do tempo são importantes para testar hipóteses e avaliar políticas.
- Dados podem ter problemas de seleção amostral, no caso de determinados indivíduos não revelarem informações acuradas.
- Amostragem deve ser realizada de forma acurada para evitar que coleta se concentre em unidades com características semelhantes.

EXEMPLO DE DADOS DE CORTE TRANSVERSAL

- Conjunto de dados de corte transversal para o ano de 1976 de 526 trabalhadores (Wooldridge 2008):

Número da observação	Salário por hora	Anos de escolaridade	Anos de experiência no mercado de trabalho	Feminino	Estado civil (casado)
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
...	...	...	...	...	...
525	11,56	16	5	0	1
526	3,50	14	5	1	0

CORTES TRANSVERSAIS AGRUPADOS (Wooldridge)

ESTUDOS DE TENDÊNCIAS (Babbie)

- Uma população pode ser amostrada e estudada em ocasiões diferentes.
- Um mesmo conjunto de variáveis é coletado em diferentes períodos do tempo, em **distintas** amostras aleatórias de uma mesma população (Censo Demográfico, Pesquisa Nacional por Amostra de Domicílios – PNAD).
- Agrupar cortes transversais de diferentes anos é eficaz para analisar os efeitos de uma política pública.
- O ideal é coletar dados de anos anteriores e posteriores a uma importante mudança de política governamental.
- Além de aumentar o tamanho da amostra, a análise de corte transversal agrupada é importante para estimar como uma relação fundamental mudou ao longo do tempo.
- Geralmente são utilizados dados secundários, coletados por outros pesquisadores ou instituições.

EXEMPLO DE CORTES TRANSVERSAIS AGRUPADOS

- Conjunto de dados sobre os preços da moradia em 1993 e 1995 nos Estados Unidos (Wooldridge 2008):

Número da observação	Ano	Preço comercializado	Imppro	Arquad	Quantidade de dormitórios	Quantidade de banheiros
1	1993	85.500	42	1.600	3	2,0
2	1993	67.300	36	1.440	3	2,5
3	1993	134.000	38	2.000	4	2,5
...	...	...	...	...	...	...
250	1993	243.600	41	2.600	4	3,0
251	1995	65.000	16	1.250	2	1,0
252	1995	182.400	20	2.200	4	2,0
253	1995	97.500	15	1.540	3	2,0
...	...	...	...	...	...	...
520	1995	57.200	16	1.100	2	1,5

DADOS DE SÉRIES DE TEMPO (Wooldridge)

ESTUDOS DE COORTES (Babbie)

- Um conjunto de dados de séries de tempo consiste em observações sobre variáveis ao longo do tempo.
- Como eventos passados podem influenciar eventos futuros, o tempo é uma dimensão importante em um conjunto de dados de séries de tempo.
- A análise desses dados pode ser dificultada, porque observações econômicas não são independentes ao longo do tempo (variáveis possuem padrões sazonais).
- Há uma série de frequências possíveis: diárias, semanais, mensais, trimestrais, anuais, decenais...
- Estes dados são também chamados de estudos de coorte, em que mesma população é analisada, mas amostras estudadas podem ser diferentes:
 - Pessoas com 10 anos em 2000, 20 anos em 2010, 30 anos em 2020, 40 anos em 2030...

EXEMPLO DE DADOS DE SÉRIES DE TEMPO

- Conjunto de dados de séries de tempo sobre efeitos do salário mínimo em Porto Rico (apud Wooldridge 2008):

Número da observação	Ano	Salário mínimo médio no ano	Taxa de trabalhadores cobertos pela lei de salário mínimo	Taxa de desemprego	Produto Nacional Bruto (PNB)
1	1950	0,20	20,1	15,4	878,7
2	1951	0,21	20,7	16,0	925,0
3	1952	0,23	22,6	14,8	1.015,9
...	...	...	...	...	...
37	1986	3,35	58,1	18,9	4.281,6
38	1987	3,35	58,2	16,8	4.496,7

DADOS DE PAINEL OU LONGITUDINAIS (Wooldridge)

ESTUDOS DE PAINEL (Babbie)

- Um conjunto de dados de painel consiste em uma série de tempo para **cada** membro do corte transversal.
- Os dados de painel são distintos dos dados de corte transversal agrupados (tendências) e de séries de tempo (coortes), porque as **mesmas** unidades são acompanhadas ao longo de um determinado período.
- Dados de painel podem ser coletados para indivíduos, domicílios, instituições ou unidades geográficas.
- Esses dados são os mais sofisticados para fins explicativos, mas são mais difíceis e caros de se obter.
- Pode haver problema de grande número de não respostas nas últimas ondas de entrevistas.
- A análise dos dados pode se tornar complicada quando se tentar avaliar as mudanças dos indivíduos no tempo.

EXEMPLO DE DADOS DE PAINEL OU LONGITUDINAIS

- Conjunto de dados de painel sobre crime e estatísticas relacionadas em 1986 e 1990 em 150 cidades nos Estados Unidos (Wooldridge 2008):

Número da observação	Cidade	Ano	Homicídios	População	Desemprego	Polícia
1	1	1986	5	350.000	8,7	440
2	1	1990	8	359.200	7,2	471
3	2	1986	2	64.300	5,4	75
4	2	1990	1	65.100	5,5	75
...	...	...	...	...	...	...
297	149	1986	10	260.700	9,6	286
298	149	1990	6	245.000	9,8	334
299	150	1986	25	543.000	4,3	520
300	150	1990	32	546.200	5,2	493

CORTE TRANSVERSAL USADO COMO LONGITUDINAL

- Alguns mecanismos podem ser utilizados num *survey* interseccional (corte transversal) para aproximar o estudo de processo ou mudança (longitudinal).
- Podem ser realizadas perguntas referentes ao passado (renda no ano anterior, local de residência anterior):
 - Há problemas de erro de memória.
 - Os dados devem ser interpretados como amostra da população atual, e não de população passada.
- Por exemplo, é possível utilizar um único banco de dados de corte transversal para comparar pessoas de diferentes idades (jovens e idosos) e coortes (calouros e veteranos).

VARIAÇÕES DOS DESENHOS BÁSICOS

- Os desenhos básicos de pesquisa apresentados anteriormente podem ser modificados para se enquadrarem aos objetivos de um estudo:
 - **Amostras paralelas:** amostras separadas de populações diferentes, utilizando mesmo questionário (exemplo é a pesquisa sobre preconceito na UFMG).
 - **Estudos contextuais:** uso de dados sobre o ambiente ou meio da pessoa para descrever o contexto do indivíduo.
 - **Estudos sociométricos:** intenção é de observar as inter-relações entre membros da população estudada (redes de amizades, por exemplo).

ESCOLHENDO O DESENHO APROPRIADO

- **Dados de corte transversal** são mais apropriados se objetivo é descrição de tempo único.
- **Mudanças ao longo do tempo** são mais difíceis de realizar, porque dados de painel exigem tempo e recursos:
 - É possível utilizar dados de corte transversal e comparar pessoas que passaram por uma experiência no passado, com aqueles que não passaram.
- **Estudos de painel** são mais viáveis economicamente quando o fenômeno estudado tem duração curta (por exemplo, opinião de voto durante uma campanha eleitoral).
- **Estudos de tendências** podem ser realizados quando dados antigos são complementados com dados coletados pelo pesquisador.

5) Econometria

Wooldridge, 2008: 1-17 (capítulo 1)

ECONOMETRIA

- A econometria evoluiu como uma disciplina separada da estatística matemática, porque enfoca problemas inerentes à coleta e à análise de dados econômicos não-experimentais.
- **Dados não-experimentais** não são acumulados por meio de experimentos controlados de indivíduos, firmas ou segmentos da economia.
- Dados não-experimentais são também chamados de **dados observacionais** para enfatizar o fato de que o pesquisador é um coletor passivo de dados.
- **Dados experimentais** são frequentemente coletados em ambientes de laboratório nas ciências naturais, mas são muito mais difíceis de serem obtidos nas ciências sociais.
- O método de análise da **regressão múltipla** é utilizado por econometristas e estatísticos matemáticos, mas o foco e interpretação pode diferir significativamente.

ANÁLISE ECONÔMICA EMPÍRICA

- Os métodos econométricos são usados para testar uma teoria econômica ou para analisar relações que apresentam importância para análises de políticas públicas.
- Uma análise empírica usa dados para testar uma teoria ou estimar uma relação.
- O primeiro passo em qualquer análise empírica é a formulação cuidadosa da questão de interesse, a qual pode ser a de testar efeitos de uma política governamental ou, até mesmo, de testar hipóteses e teorias.
- O modelo econômico formal consiste em equações matemáticas que descrevem relações para testar teorias.

MICROECONOMIA

- Os indivíduos fazem escolhas para maximizar seu bem-estar (**maximização da utilidade**), sujeitas às restrições de recursos.
- Isso oferece um arcabouço para criar modelos econômicos para fazer previsões entre variáveis.
- A maximização da utilidade leva a um conjunto de **equações de demanda**, no contexto das decisões de consumo.
- Em uma equação de demanda, a quantidade demandada de cada produto depende do seu próprio preço, do preço dos bens substitutos e complementares, da renda do consumidor e das características individuais que influem no gosto.

MODELO ECONÔMICO

- O modelo econômico é a formulação teórica de uma relação entre variáveis econômicas.
- A quantidade de tempo gasto na atividade criminosa é uma função de vários fatores (Gary Becker 1968):

$$y=f(x_1, x_2, x_3, x_4, x_5, x_6, x_7),$$

y = horas gastas em atividades criminosas.

x_1 = “salário” por hora ocupada em atividade criminosa.

x_2 = salário-hora em emprego legal.

x_3 = renda de outras atividades que não o crime ou um emprego legal.

x_4 = probabilidade de ser capturado.

x_5 = probabilidade de ser condenado se capturado.

x_6 = sentença esperada se condenado.

x_7 = idade.

MODELO ECONOMETRICO

- Após elaborar o modelo econômico, é especificado um modelo econométrico, que será aplicado a dados existentes.
- A forma da função $f(.)$ deveria ser especificada antes de realizar uma análise econométrica.
- Se uma variável não pode ser obtida, é possível utilizar uma variável que se aproxima desta que se quer medir (**proxy**).
- Outros fatores são considerados no termo de erro u (ou termo de perturbação):
 - **Erro amostral** é a diferença entre o resultado amostral e o verdadeiro resultado da população (devidos ao acaso).
 - **Erro não-amostral** ocorre quando os dados amostrais são coletados, registrados ou analisados incorretamente.
- Modelo econométrico de Becker (1968):

$$\text{crime} = \beta_0 + \beta_1 \text{salário} + \beta_2 \text{outrenda} + \beta_3 \text{freqpris} + \beta_4 \text{freqcond} + \beta_5 \text{sentmed} + \beta_6 \text{idade} + u$$

MODELO ECONOMETRICO NA PRÁTICA

- Na maioria dos casos, a análise econométrica começa pela especificação de um modelo econométrico, sem consideração de detalhes da criação do modelo econômico.
- É comum começar com um modelo econométrico e usar o raciocínio econômico e conhecimentos científicos como guias para escolher as variáveis.
- Após a especificação do modelo econométrico, várias hipóteses podem ser formuladas em termos das direções e influências dos parâmetros desconhecidos (independentes) sobre a variável de interesse (dependente).
- Após os dados terem sido coletados, os métodos econométricos são usados para estimar os parâmetros do modelo econométrico e para testar as hipóteses de interesse.

6) A lógica da amostragem do *survey*

Babbie, 1999: 113-158 (capítulo 5)

CONCEITOS BÁSICOS

- **Estatística** (como ciência) é um conjunto de métodos para:
 - Planejamento de estudos e experimentos.
 - Obtenção de dados.
 - Organização, resumo, apresentação, análise, interpretação e elaboração de conclusões baseadas nos dados.
- **Censo** é um conjunto de dados obtidos de todos os membros da população.
- **Amostra** é um subconjunto de membros selecionados de uma população.

POR QUE FAZER AMOSTRAGEM?

- Há duas razões para selecionar uma parcela (amostra) da população e fazer estimativas sobre toda a população: tempo e custo.
- Muitas vezes *surveys* por amostragem são mais precisos do que entrevistar todos componentes da população, devido à logística do *survey*.
- Um grande número de entrevistas requer muitos pesquisadores, os quais podem não ter boa qualidade.
- A qualidade dos dados coletados se reduz com a menor qualidade dos entrevistadores.
- Uma pesquisa em menor escala permite procedimentos de acompanhamento mais severos.

POR QUE FAZER AMOSTRAGEM? (cont.)

- Se forem realizadas uma grande quantidade de entrevistas, o tempo aumenta, dificultando especificar a data de referência da entrevista.
- O ponto médio da entrevista pode ser utilizado como data de referência.
- Também é possível pedir aos entrevistados para responder sobre uma data específica no passado, mas isso introduz problema de recordação imprecisa.
- Geralmente é recomendado que pesquisas de *survey* tenham período de coleta de dados de no máximo 3 meses.
- A supervisão, relatórios e treinamento são muito mais difíceis em um grande *survey*, e a qualidade dos dados pode ser menor do que aquela obtida com menos entrevistas.

DADOS POR AMOSTRAGEM SÃO REALMENTE EXATOS?

- Se forem tomados procedimentos de amostragem bem estabelecidos, *surveys* por amostragem podem permitir estimativas muito precisas sobre as populações.
- Raramente é possível determinar com exatidão o grau de precisão dos achados de uma amostra.
- Pesquisadores de opinião política e eleitoral representam um grupo de analistas a quem é dada a oportunidade de verificar a precisão de seus dados (resultados reais da eleição).

PROBLEMAS EM AMOSTRAGEM

- Em grande parte, o **objeto da estatística** é o uso de dados amostrais para se fazer generalizações sobre uma população.
- Os **dados amostrais** devem ser coletados de modo apropriado, senão a amostra não será representativa para a população da qual foi obtida.
- **Amostra de resposta voluntária** (ou amostra auto-selecionada) é aquela em que os respondentes decidem, eles mesmos, se serão ou não incluídos.
 - É uma amostra tendenciosa e não deve ser usada para fazer afirmações de uma população maior.
 - Censo americano tenta corrigir esse problema com visita de recenseadores após envio dos questionários.
- **Pequenas amostras** podem gerar resultados não confiáveis.

HOMOGENEIDADE E HETEROGENEIDADE

- **Homogeneidade:** se todos os membros de uma população fossem idênticos uns aos outros em todos os sentidos, qualquer tamanho de amostra seria suficiente. Na realidade, bastaria um caso para estudar as características de toda população.
- **Heterogeneidade:** se há variação na população estudada, o pesquisador deve usar procedimentos de amostragem mais controlados.
 - Uma amostra de indivíduos de uma população deve conter a mesma variação existente na população, para permitir descrições úteis dela.
 - A amostragem probabilística é um método eficiente para extrair uma amostra que reflita corretamente a variação existente na população como um todo.

VIÉS CONSCIENTE E INCONSCIENTE NA AMOSTRAGEM

- É preciso entrevistar pessoas com perfis diferenciados para representarem a população total.
- Porém, sem método de amostragem robusto, não saberemos as proporções adequadas dos tipos diversos de indivíduos que devem ser entrevistados.
- Há então o perigo dos vieses pessoais do pesquisador afetarem a amostra selecionada.
- Esse viés pode acontecer de maneira consciente ou inconsciente na coleta de dados.

REPRESENTATIVIDADE E PROBABILIDADE DE SELEÇÃO

- Como vimos, as amostras de *surveys* devem representar as populações das quais são retiradas.
- Além disso, amostras necessitam ser representativas para os interesses substantivos da pesquisa.
- A representatividade vai depender se queremos fazer análises específicas por:
 - Áreas geográficas.
 - Nível de escolaridade.
 - Nível sócio-econômico.
 - Faixa etária.
 - Sexo...

CONCEITOS E TERMINOLOGIA DE AMOSTRAGEM

- **Elemento** (unidade de análise) é a unidade sobre a qual a informação é coletada e que serve de base para a análise (pessoas, famílias, corporações, países...).
- **Universo** é a agregação teórica e hipotética de todos elementos definidos num *survey*. O universo não é especificado quanto a tempo e lugar.
- **População** é a agregação teoricamente especificada de elementos do *survey*. É a coleção completa de todos os elementos a serem estudados. É preciso definir o elemento e o referencial de tempo da pesquisa. É a especificação teórica do universo.
- **População do *survey*** é a agregação de elementos da qual a amostra do *survey* é de fato selecionada. Nem todos elementos da população terão chance de ser selecionados para a amostra.

CONCEITOS E TERMINOLOGIA DE AMOSTRAGEM (cont.)

- **Unidade de amostra** é o elemento ou conjunto de elementos considerados para seleção em alguma etapa da amostragem.
 - Numa amostra simples, as unidades de amostra são o mesmo que os elementos.
 - Em outros casos, uma amostra pode ter diferentes unidades de amostra (setores censitários, domicílios, pessoas), mas só as pessoas são elementos.
- **Moldura de amostragem** é a lista de unidades de amostra da qual a amostra é selecionada.
 - Na amostra de etapa única, a moldura de amostragem é a lista dos elementos compondo a população de *survey*.
 - Muitas vezes, são as molduras que definem as populações de *survey*, e não o contrário.

CONCEITOS E TERMINOLOGIA DE AMOSTRAGEM (cont.)

- **Unidade de observação** (unidade de coleta de dados) é um elemento (ou agregação de elementos) do qual se coleta dados.
 - Unidade de análise (elemento) e unidade de observação podem ser a mesma coisa (indivíduo, por exemplo).
 - Mas em uma pesquisa podemos entrevistar chefes de domicílio (unidades de observação) para coletar informação sobre todos membros (unidades de análise).
- **Variável** é um conjunto de características mutuamente excludentes, como sexo, idade, emprego...
 - Podemos descrever os elementos de uma população com base nas características individuais das variáveis.

CONCEITOS E TERMINOLOGIA DE AMOSTRAGEM (cont.)

- **Parâmetro** é uma medida numérica que descreve alguma característica de uma população.
- **Estatística** é uma medida numérica que descreve alguma característica de uma amostra.
- **Erro amostral**: pesquisa de *survey* busca estimar parâmetros com base em amostras, o que gera erros.
 - Teoria da probabilidade permite estimar o grau de erro.
- **Níveis de confiança** indicam o percentual de confiança de que sua estatística (amostral) está dentro de um **intervalo de confiança** do parâmetro (populacional).
 - Nível e intervalo são estimados usando o erro amostral.

7) Amostragem probabilística

Babbie, 1999: 113-158 (capítulo 5)

TEORIA DA AMOSTRAGEM PROBABILÍSTICA

- A **finalidade do *survey*** é selecionar um conjunto de elementos de uma população de modo que as estatísticas descrevam com precisão os parâmetros populacionais.
- A **amostragem probabilística** fornece métodos para ampliar a possibilidade de alcançar esta meta, bem como métodos para estimar o grau de sucesso provável.
- Em uma amostra com **seleção aleatória**, todos membros da população têm oportunidade igual de serem selecionados para a amostra.
- Os métodos de seleção aleatória:
 - Diminuem os vieses conscientes ou inconscientes do pesquisador.
 - Permitem utilizar a teoria da probabilidade para estimar parâmetros populacionais e erros amostrais.

DISTRIBUIÇÃO BINOMIAL DA AMOSTRAGEM

- Queremos avaliar uma **variável binomial** em nossa população, em que os valores possíveis são zero e um.
- A **distribuição amostral** é a distribuição das estatísticas das diferentes amostras selecionadas.
- Se **muitas amostras aleatórias** independentes são extraídas de uma população, as estatísticas amostrais fornecidas por estas amostras estarão distribuídas em torno do parâmetro populacional.
- A **teoria da probabilidade** possui uma fórmula para estimar o quão próximo as estatísticas da amostragem estão aglomeradas em torno do valor real.
- O **erro da amostragem** ou erro padrão (S) é medido com base nos parâmetros (P, Q) e tamanho da amostra (n):

$$S = \sqrt{\frac{PQ}{n}}$$

ERRO DA AMOSTRAGEM (ERRO PADRÃO)

- O **erro padrão** (S) indica em que medida as estimativas amostrais estão distribuídas em torno do parâmetro populacional.
- Aproximadamente 34% das estimativas amostrais estão até **um erro padrão** (desvio padrão) acima do parâmetro da população, e outros 34% até um desvio padrão abaixo do parâmetro.
 - Ou seja, 68% das amostras resultarão em estimativas dentro de (mais ou menos) um erro padrão.
- Cerca de 95% das amostras estarão dentro de (mais ou menos) **dois erros padrão** do valor verdadeiro.
- 99,9% das amostras estarão dentro de (mais ou menos) **três desvios padrão**.
- A proporção de amostras dentro de 1, 2 ou 3 erros padrão do parâmetro é constante em amostragens aleatórias.

TEORIA DA PROBABILIDADE NA PRÁTICA

- A teoria da probabilidade depende que se selecione **grandes números de amostras aleatórias**.
- Se **parâmetro for conhecido** e muitas amostras aleatórias forem selecionadas, podemos prever quantas amostras cairão dentro de intervalos específicos a partir do parâmetro.
- No entanto, geralmente **desconhecemos o parâmetro** e selecionamos **somente uma amostra**.
- Sabemos que uma amostra aleatória tem probabilidade de 68% de estar dentro da faixa de um erro padrão.
 - Ou seja, temos 68% de confiança que a estimativa da amostra está dentro de um erro padrão do parâmetro.
- Como não temos parâmetro, **usamos a estimativa da amostragem** na fórmula do erro padrão.
- Estamos x% confiante (**nível de confiança**) de que o parâmetro está entre dois valores (**intervalo de confiança**).

POPULAÇÕES E MOLDURAS DE AMOSTRAGEM

- É preciso discutir as relações entre as **condições de campo** e os pressupostos teóricos.
- A moldura de amostragem é a lista de elementos dos quais é selecionada uma amostra probabilística.
- O *survey* amostral é representativo dos membros desta lista.
- É importante saber até que ponto há semelhança ou diferença entre **populações e molduras de amostragem**.
- Se as molduras de amostragem são realmente representativas da população estudada, amostras corretamente selecionadas fornecem informações apropriadas para descrever a população.

SOBRE QUAL POPULAÇÃO ESTAMOS FALANDO?

- Os resultados dos *surveys* por amostragem só podem ser considerados representativos do conjunto de elementos que compõem a moldura de amostragem.
- As listagens não incluem de fato todos os elementos da população, por isso é preciso avaliar a extensão destas omissões e corrigi-las, se possível.
- No caso de existir omissões, é necessário informar nos relatórios de pesquisa, com o objetivo de expor corretamente qual população está sendo estudada.
- Para poder generalizar a população que compõe a moldura de amostragem, é necessário que todos elementos tenham representação igual (apareçam só uma vez).

MÉTODOS DE AMOSTRAGEM

- **Amostragem probabilística:** indivíduos do universo têm probabilidade conhecida e diferente de zero de serem selecionados para amostra (mas não necessariamente a mesma probabilidade).
 - Amostras probabilísticas são mais representativas do que outros tipos, porque são evitados os vieses de seleção.
 - Além disso, a teoria da probabilidade permite estimar a precisão ou representatividade da amostra.
- **Amostragem não-probabilística:** não são estimadas as probabilidades de seleção dos elementos da amostra.

AMOSTRAGEM ALEATÓRIA SIMPLES (AAS)

- O desenho da pesquisa é realizado com AAS quando toda amostra possível de ser obtida de tamanho n tem a mesma chance de ser escolhida.
- De posse da moldura de amostragem, você enumera cada elemento da lista, atribuindo só um número a cada um, sem saltos.
- Em seguida, é usada uma tabela de números aleatórios ou um programa computacional para selecionar os elementos da amostra.
- AAS é raramente realizada na prática, pois é necessário dispor de uma lista de elementos (moldura de amostragem) bem definida a priori.
- Mesmo de posse desta lista, geralmente os pesquisadores usam o método de amostragem sistemática.

AMOSTRAGEM SISTEMÁTICA

- Na amostragem sistemática, há a escolha de algum ponto inicial e a seguir é selecionado cada k -ésimo elemento da população.
- Por exemplo, se a lista tem 10.000 elementos e se deseja uma amostra de 1.000, é selecionado cada 10º elemento.
- **Amostra sistemática de início aleatório**: primeiro elemento é selecionado aleatoriamente.
- **Intervalo de amostragem** é a distância padrão entre os elementos selecionados na amostra (10 no exemplo acima).
- **Razão amostral** é a proporção de elementos selecionados da população: divisão de 1 pelo intervalo de amostragem.
- Há perigo da **periodicidade**, se lista de elementos estiver organizada com base em alguma informação da população.
- Resultados da AAS e da AS são idênticos, mas a AS é mais fácil e conveniente de ser realizada na prática.

AMOSTRAGEM ESTRATIFICADA

- AAS e AS são importantes para garantir grau de representatividade e permitir estimativa de erro.
- Na **amostragem estratificada**, intuito é de garantir representação das variáveis de estratificação para aumentar representação de outras variáveis.
- AE tem possibilidade de ser **mais representativa** em certas variáveis do que AAS, reduzindo erro amostral.
- Como erro é menor com população homogênea, idéia é de selecionar amostra dentro de **subconjuntos homogêneos** (com heterogeneidade entre subconjuntos).
- População é dividida em subgrupos (estratos) diferentes entre si. Sujeitos de um subgrupo possuem mesmas características. **Variáveis de estratificação** dependem do interesse de estudo e disponibilidade de informação.
- Com base na proporção relativa de cada subgrupo na população, **elementos dos subgrupos são selecionados** na mesma proporção do tamanho da amostra.

ESTRATIFICAÇÃO NA AMOSTRAGEM SISTEMÁTICA

- Se uma lista de elementos estiver agrupada por características da população, a amostragem sistemática realizará uma **estratificação implícita**.
- Por exemplo, se alunos são organizados por turma, sexo e notas em uma lista, uma amostra sistemática selecionaria números adequados de cada subgrupo.
- Uma lista ordenada de elementos pode ser mais útil do que uma lista não ordenada (aleatória).
- Nestas condições, a amostragem sistemática é mais precisa na representatividade dos subgrupos do que a amostragem aleatória simples.
- A lista só deve ser reordenada se o arranjo apresenta algum problema de periodicidade.

AMOSTRAGEM POR CONGLOMERADOS

- Muitas pesquisas não possuem listagens para fins de amostragem, tais como população de um município.
- Realizar amostragem de **grupos de elementos** e depois selecionar elementos de cada conglomerado.
- Idéia é de selecionar amostra dentro de **subconjuntos heterogêneos** (com homogeneidade entre subconjuntos).
- **Passos:** (1) população é dividida em conglomerados (seções); (2) são selecionados alguns dos conglomerados; e (3) seus membros são listados e amostrados.
- **Exemplo** de amostragem por conglomerados:
 - Ao pesquisar um município, pode ser feita lista de bairros (conglomerados) para amostragem AAS ou AS.
 - Após selecionar bairros, pode ser feita lista de quarteirões (conglomerados) para amostragem AAS ou AS.
 - Após selecionar quarteirões, é realizada listagem de domicílios para nova amostragem AAS ou AS.

PRECISÃO NA AMOSTRAGEM POR CONGLOMERADOS

- A amostragem por conglomerados é recomendada pela eficiência, embora não tenha muita precisão.
- Como a amostragem por conglomerados é geralmente realizada em múltiplas etapas, ela terá **um erro amostral para cada etapa**.
- A amostra inicial dos conglomerados representa a população de conglomerados em uma faixa de erro amostral.
- A amostra de elementos extraída de um conglomerado representa todos elementos, com um erro amostral.
- Há o **risco** de selecionar uma amostra de quarteirões desproporcionalmente ricos, assim como amostra de residências ricas dentro dos quarteirões.
- **Solução** é pensar: (1) na quantidade de conglomerados selecionados no início; e (2) no número de elementos selecionados em cada conglomerado.

AMOSTRA E HOMOGENEIDADE NOS CONGLOMERADOS

- Sabemos que erro amostral é menor com o aumento do tamanho da amostra e com maior homogeneidade dos elementos a serem amostrados.
- Uma **amostra por conglomerados** representará melhor todos conglomerados se um **grande número** for selecionado e se todos eles forem bastante **semelhantes entre si**.
- Uma **amostra de elementos** representará melhor todos elementos de um conglomerado se um **grande número** de elementos for selecionado do conglomerado, e se todos os elementos forem bastante **semelhantes entre si**.
- Dado um tamanho máximo de amostra, se o número de conglomerados aumentar, o número de elementos num conglomerado deve diminuir.
- Ou seja, a representatividade dos conglomerados aumenta à custa da menor representação dos elementos de cada um, e vice-versa.

DIRETRIZ GERAL DE AMOSTRA DE CONGLOMERADOS

- Os elementos de um conglomerado natural dentro de uma população são tipicamente mais homogêneos do que todos os elementos da população total.
- Por exemplo, os moradores de um bairro são mais semelhantes entre si do que todos os moradores do município.
- Podem ser necessários **poucos elementos** para representar adequadamente um conglomerado natural, enquanto um **número maior de conglomerados** pode ser necessário para representar adequadamente a diversidade dos conglomerados.
- A **diretriz geral** (científica) no desenho de conglomerados é maximizar o número de conglomerados selecionados, diminuindo o número de elementos de cada um.
- Porém, um número menor de conglomerados pode ser listado com mais agilidade e economia.

REGRA SIMPLES DE AMOSTRA DE CONGLOMERADOS

- Pesquisadores populacionais **convencionalmente** selecionam cinco residências por zona ou quarteirão de recenseamento.
- Por exemplo, querendo entrevistar um total de 2.000 residências, selecionamos 400 quarteirões com cinco entrevistas de residências em cada um.
- De todo modo, amostragem por conglomerados envolve **perda de precisão**.
- Desenho amostral em etapas múltiplas está sujeito a erros amostrais em cada uma das etapas:
 - O erro amostral em cada etapa será maior do que no caso de uma amostra aleatória em etapa única, porque o tamanho da amostra é menor em cada etapa.
 - Isso está relacionado com a fórmula de cálculo do erro amostral: $s = \sqrt{\frac{PQ}{n}}$

AAS, AS E ESTRATIFICAÇÃO NOS CONGLOMERADOS

- Discutimos a amostragem por conglomerados como se fosse selecionada uma amostra aleatória simples a cada etapa do desenho.
- Informações podem ser usadas para **estratificar os conglomerados**:
 - Censo pode ser usado para estratificar quarteirões de recenseamento em termos de composição étnica, classe social, valores das propriedades, qualidade de estruturas, natureza da propriedade das construções e tamanho.
- Após **estratificação** das unidades primárias de amostragem de acordo com as variáveis disponíveis, pode-se usar tanto técnicas de amostragem **aleatória simples** ou amostragem **sistemática** para extrair a amostra.
- **Conglomerados e estratificação**: erro amostral será mais reduzido, quanto mais os conglomerados estiverem combinados em estratos homogêneos.

AMOSTRAGEM SIMPLES POR CONGLOMERADO

- Para garantir a seleção de uma amostra representativa de elementos, é preciso dar a cada elemento da população total uma oportunidade igual de ser selecionada.
- A forma mais simples de fazer isso numa amostra por conglomerados é dar a cada um a mesma chance de seleção e selecionar uma determinada proporção de elementos de cada conglomerado selecionado:
 - Temos 100.000 elementos agrupados em 1.000 conglomerados e queremos amostra de 1.000 elementos.
 - Selecionamos $1/10$ dos conglomerados (100) com igual probabilidade.
 - Depois selecionamos $1/10$ dos elementos de cada conglomerado escolhido.
 - Ou seja, cerca de 1.000 elementos foram selecionados e tiveram a mesma probabilidade de seleção.

AMOSTRAGEM ANTERIOR É SIMPLES E INEFICIENTE

- A técnica anterior é simples, mas não é a mais eficiente.
- A maioria das amostragens por conglomerados envolve conglomerados de tamanhos muito diversos.
- Ao selecionar conglomerados com igual probabilidade, retirando uma proporção fixa de elementos dos conglomerados selecionados:
 - Seleciona-se um número relativamente pequeno de conglomerados grandes na primeira etapa da amostragem.
 - Os elementos selecionados para representar todos os elementos dos grandes conglomerados são retirados de poucos desses conglomerados.
 - No limite, toda população da cidade residindo em dez grandes bairros poderia ser representada pelas pessoas vivendo em apenas um deles.

AMOSTRAGEM PPT

- Como vimos, a amostragem por conglomerados terá maior eficiência se houver a seleção de muitos conglomerados, com poucos elementos retirados de cada um.
- Esse princípio é utilizado pelo método de amostragem de probabilidade proporcional ao tamanho (PPT):
 - Há seleção de maior número de conglomerados.
 - Há garantia de representação dos elementos de conglomerados grandes.
 - Cada elemento na população recebe igual chance de seleção.

PROCEDIMENTOS DA AMOSTRAGEM PPT

- Na primeira etapa, é dada a cada conglomerado uma oportunidade de seleção proporcional ao seu tamanho.
- Na segunda etapa, o mesmo número de elementos é escolhido de cada conglomerado selecionado.
- Esses procedimentos igualam as probabilidades últimas de seleção de todos os elementos, já que:
 - Conglomerados com mais elementos têm maior probabilidade de serem selecionados.
 - Elementos em conglomerados maiores têm menor chance de seleção do que elementos em conglomerados menores.

FÓRMULA E EXEMPLO DE PPT

- A probabilidade de seleção de um elemento num desenho de amostragem PPT é:

$$[probabilidade]_{do\ elemento} = \left[\frac{número\ de\ conglomerados}{selecionados} \right] \times \left[\frac{tamanho\ do\ conglomerado}{tamanho\ da\ população} \right] \times \left[\frac{elementos\ selecionados\ por\ conglomerado}{tamanho\ do\ conglomerado} \right]$$

- Se 100 conglomerados são selecionados e 10 elementos escolhidos de cada um, a partir de uma população total de 100.000, a probabilidade geral de seleção de cada elemento será de 1.000/100.000 ou 1/100.
- **Exemplo 1:** um conglomerado com 100 elementos:

$$Prob_1 = 100 \times (100/100.000) \times (10/100) = 1/100$$

- **Exemplo 2:** um conglomerado com 10 elementos:

$$Prob_2 = 100 \times (10/100.000) \times (10/10) = 1/100$$

- Independente do número de elementos num conglomerado, cada elemento tem a mesma probabilidade de seleção.

SIMPLIFICANDO A FÓRMULA DE PPT

- O tamanho do conglomerado é cancelado:

$$[probabilidade]_{do\ elemento} = \left[\frac{número\ de\ conglomerados}{selecionados} \right] \times \left[\frac{\cancel{tamanho\ do\ conglomerado}}{tamanho\ da\ população} \right] \times \left[\frac{\cancel{tamanho\ do\ conglomerado}}{selecionados\ por\ conglomerado} \right]$$

- Isso é o tamanho da amostra dividido pelo tamanho da população:

$$[probabilidade]_{do\ elemento} = \frac{\left[\frac{número\ de\ conglomerados}{selecionados} \right] \times \left[\frac{selecionados}{por\ conglomerado} \right]}{\left[\frac{tamanho\ da}{população} \right]}$$

MODIFICAÇÕES NO DESENHO PPT

- Costuma-se fazer duas modificações no desenho de amostra PPT.
- **Primeiro**, podemos achar necessário representar conglomerados muito grandes na amostra.
 - Seleccionamos todos esses conglomerados (probabilidade igual a 1) e seus elementos com a probabilidade geral.
- **Segundo**, se um número padrão de elementos for selecionado de cada conglomerado escolhido, conglomerados com menos elementos do que o número padrão representam um problema.
 - Solução é combinar pequenos conglomerados, de forma que cada combinação tenha pelo menos o número padrão a ser selecionado.
- Amostragem por conglomerados é difícil, mas é adequada quando for impossível obter uma lista de todos elementos da população.

AMOSTRAGEM PROBABILÍSTICA E PONDERAÇÃO

- Sabemos que uma amostra probabilística é aquela em que cada elemento da população tem uma probabilidade diferente de zero e conhecida de seleção, mesmo que elementos diferentes tenham probabilidades diferentes.
- A amostra será representativa da população se cada elemento da amostra receber um peso igual ao inverso da sua probabilidade de seleção.
- No caso de todos os elementos da amostra terem a mesma chance de seleção (amostra autoponderada), cada um recebe o mesmo peso (1).

AMOSTRAGEM E PONDERAÇÃO DESPROPORCIONAIS

- É ainda possível fazer amostragem e ponderação desproporcionais:
 - Ou seja, amostramos subpopulações desproporcionalmente para garantir número suficiente de casos de cada subpopulação para análise.
- Podemos sobre-amostrar uma primeira área para analisá-la detalhadamente, ao mesmo tempo em que a segunda área terá uma amostra proporcionalmente menor:
 - Devemos analisar as duas amostras separadamente ou comparativamente.
 - Se quisermos combinar as duas amostras para criar um quadro composto de toda região, devemos realizar um procedimento de ponderação da amostra.

GRAUS DE PRECISÃO NA PONDERAÇÃO

- Em um desenho complexo de amostra, podemos computar pesos para cada elemento até várias casas decimais, ou atribuir pesos aproximados para corrigir uma eventual amostragem desproporcional.
- A precisão que buscamos na ponderação deve ser comensurável com a precisão que queremos nos nossos resultados.
- Se nossos objetivos de pesquisa permitem tolerar erros de alguns pontos percentuais, não nos preocuparemos com cálculo de ponderações exatas.
- O grau de precisão da ponderação terá relação direta com o grau de erro amostral e não amostral de nossa pesquisa.
- Não há uma fórmula precisa para determinar qual a melhor ponderação.
- No entanto, devemos informar nos resultados se nossa ponderação foi aproximada ou precisa.

MÉTODOS DE PONDERAÇÃO

- Para a **ponderação aproximada**:
 - Primeiro criamos tabelas sem pesos de cada uma das amostras das subpopulações.
 - Depois multiplicamos as células da subpopulação que foi subamostrada (segundo grau de subamostragem), antes de adicioná-la à subpopulação que foi sobreamostrada.
- Para uma **ponderação mais extensa e rápida**:
 - Dados para certos casos podem ser copiados. Intenção é de aumentar número de casos, como se tivéssemos aplicado uma maior quantidade de questionários.
- Para **ponderações previamente calculadas**:
 - Podemos utilizar um banco de dados que já possui os pesos de cada elemento. Neste caso, somente informamos ao programa computacional qual variável contém a informação de peso.

PONDERAÇÃO E INFERÊNCIA ESTATÍSTICA

- Os procedimentos de ponderação têm efeitos sérios na maioria dos cálculos de inferência estatística.
- Se finalidade da pesquisa exige inferências estatísticas precisas, feitas com base em dados cuidadosamente ponderados, é preciso consultar um estatístico de amostragem antes da amostra ser desenhada.

PANORAMA DA AMOSTRAGEM PROBABILÍSTICA

- O método chave de amostragem usado na pesquisa de *survey* é a amostragem probabilística.
- Os elementos são escolhidos de uma população, com base na seleção aleatória e com probabilidades diferentes de zero e conhecidas.
- A amostragem probabilística pode ser simples ou complexa, dependendo da situação de campo.
- É o método mais eficiente para seleção de elementos:
 - Evita o viés consciente ou inconsciente na seleção de elementos por parte do pesquisador.
 - Utiliza métodos controlados de seleção que permitem estimar o grau de erro esperado (erro amostral).

8) Amostragem não-probabilística

Babbie, 1999: 113-158 (capítulo 5)

DESENHOS DE AMOSTRAGEM NÃO-PROBABILÍSTICA

- Apesar dos métodos de amostragem probabilísticos serem adequados à generalização, podemos decidir por utilizar amostragem não-probabilística:
 - Porque é mais barata.
 - Porque talvez não seja necessário ter representatividade exata.
- Há alguns métodos de amostragem não-probabilística:
 - Amostragem intencional ou por julgamento.
 - Amostragem por cotas.
 - Amostragem por confiança em sujeitos disponíveis.

AMOSTRAGEM INTENCIONAL OU POR JULGAMENTO

- Na amostragem intencional ou por julgamento, pode-se selecionar a amostra baseado no próprio conhecimento: (1) da população; (2) dos seus elementos; e (3) da natureza das metas de pesquisa.
- Pode-se realizar um pré-teste do questionário com a seleção de maior variedade de entrevistados para testar a aplicabilidade geral das perguntas.
- Este é um pré-teste do instrumento e não um *survey* propriamente dito.
- Ou seja, há a seleção de indivíduos-chave importantes para os interesses da pesquisa.

AMOSTRAGEM POR COTAS

- Na amostragem por cotas, sabemos qual a proporção da população que possui determinadas características (sexo, idade, raça, local de residência, ocupação):
 - São colhidos dados de pessoas com todas as combinações de características.
 - Todas pessoas recebem peso apropriado à sua porção na população total.
 - Quando todos elementos da amostra estiverem ponderados, os dados como um todo devem resultar numa representação razoável da população total.
- Há problemas na amostragem por cotas:
 - A moldura das cotas deve ser precisa.
 - Pode haver vieses na seleção de elementos com determinadas características.

CONFIANÇA EM SUJEITOS DISPONÍVEIS

- Na amostragem por confiança em sujeitos disponíveis, há a seleção de indivíduos que são fáceis de acessar.
- Também é chamada de amostragem de conveniência ou amostragem a esmo.
- Embora sirva à finalidade de pré-testar um questionário, este método de amostragem não deve ser usado para um estudo que pretende descrever uma população.

USOS NÃO-SURVEY DOS MÉTODOS DE AMOSTRAGEM

- Na discussão sobre a lógica e os métodos de amostragem, foi dada maior atenção à pesquisa de **survey**, em que há a seleção de elementos de uma população para entrevista ou questionários auto-administrados.
- Métodos de amostragem podem ter aplicação **não-survey**:
 - **Análise de conteúdo**: volume de documentos para analisar pode ser muito grande, o que exige uso de técnicas de amostragem de seleção de tais documentos.
 - **Experimentos de laboratórios**: sujeitos para experimentos são geralmente voluntários. Pode-se usar métodos de estratificação para satisfazer cotas.
 - **Observação participante**: ao utilizar métodos de amostragem de *survey*, um pesquisador pode obter um conjunto mais representativo de observações.

9) Sejam os críticos

Triola, 2008: 2-31 (capítulo 1)

PENSAMENTO CRÍTICO

- “O sucesso em um curso introdutório de estatística normalmente requer mais *senso comum* do que habilidade matemática.” (Triola 2008, p.9)
- É importante interpretar os dados e resultados com métodos científicos, e saber usar o senso comum.
- “Há três tipos de mentiras: mentiras, mentiras horríveis e estatística.” (Benjamin Disraeli)
- “Números não mentem, mas mentirosos manipulam números.” (Autor desconhecido)
- “Algumas pessoas usam a estatística como os bêbados usam os postes - para apoio mais do que para iluminação.” (Andrew Lang)
- “Estatístico é um especialista que coleta números e depois os desvirtua.” (*Esar's Comic Dictionary*)