Introdução à Tecnologia de Amostragem

Prof. Caio Azevedo

Introdução

- Estatística: área do conhecimento/Ciência que trata de metodologias (análise de dados) apropriadas para se coletar, organizar e analisar dados.
- A Estatística incorpora elementos de Probabilidade, Matemática (Cálculo Diferencial e Integral, Álgebra de Matrizes, Teoria da Medida, Análise Funcional etc), Computação e Ciência de dados (envolve outras formas de análise de dados), desenvolvendo (novas) metodologias.
- A Estatística é uma ferramenta muito importante na resolução de problemas levantados em diversas áreas: Biologia, Psicometria, Educação, Medicina, Física, Computação entre outras.
- É importante que o Estatístico participe de todas as etapas de um estudo (pesquisa/consultoria).

Etapas para a resolução de um problema

1. Determinação do problema/objeto de estudo (incluindo a população de interesse).
2. Determinação dos objetivos (gerais e específicos).
3. **Determinação do tamanho da amostra-delineamento amostral**/experimental.
4. **Levantamento dos dados: entrevistas, experimento, coleta de dados etc**.
5. **Análise Descritiva.**
6. **Análise Inferencial** (Modelos de regressão).
7. **Conclusões e elaboração dos relatórios/artigos/trabalhos pertinentes**.

Pode-se retornar a pontos anteriores ou mesmo avançar, desconsiderando-se alguns pontos, consoante a necessidade.

Pré-requisitos

- Cálculo diferencial e integral: Cálculo I, Cálculo II.
    - Probabilidade I : página do curso de Probabilidade I
    - Probabilidade II : página do curso de Probabilidade II
- Inferência (conceitos e alguns resultados): página do curso de ME 419/ME 420, Inferência para o Mestrado (MI402).

Motivação

- "Surveys" - **levantamento de dados**, obtidos através de: pesquisas de opinião, entrevistas, questionários, e-mail, enquetes etc.
- "Research" - **pesquisas experimentais**, obtidos através de experimentos científicos: desenvolvimento e testes sobre vacinas, mensuração do conhecimento de indivíduos.
- "Dados não planejados/não estruturados (dados orgânicos)" - não possuem estrutura usual de banco de dados ("planilhas"), em geral obtidos sem um delineamento/planejamento rigoroso(s): informações de consumidores em sites de vendas (Amazon), informações sobre professor/alunos em aulas on-line (modelo corrente na Unicamp), textos (veja também o material sobre Big Data a ser enviado via moodle).

Motivação

- Para qualquer um dos três tipos de dados anteriores é possível utilizar técnicas de amostragem. O que seria isso?
  - Coletar/organizar os dados sob planos amostrais apropriados (complexos).
  - Levar em consideração a estrutura desses planos na análise de dados (estimativas pontuais, intervalares, testes de hipóteses, modelos de regressão etc)
- Consideraremos os dois pontos acima, no curso, com o intuito de realizar estimação (pontual e intervalar) bem como testes de hipótese (semelhante ao que é feito no curso de Noções de Estatística).

Motivação

- **Ponto de partida**: qual(ais) é (são) a(s) população(ões) de interesse? Quais são as questões relevantes com relação à população de interesse?
- População: conjunto de elementos com características semelhantes.
- Amostra: subconjunto da população.
- Variável: característica de interesse relacionada a cada elemento da população (amostra).
- Parâmetro: característica de interesse relacionada à população (ou parte dela) como um todo (ou da amostra).

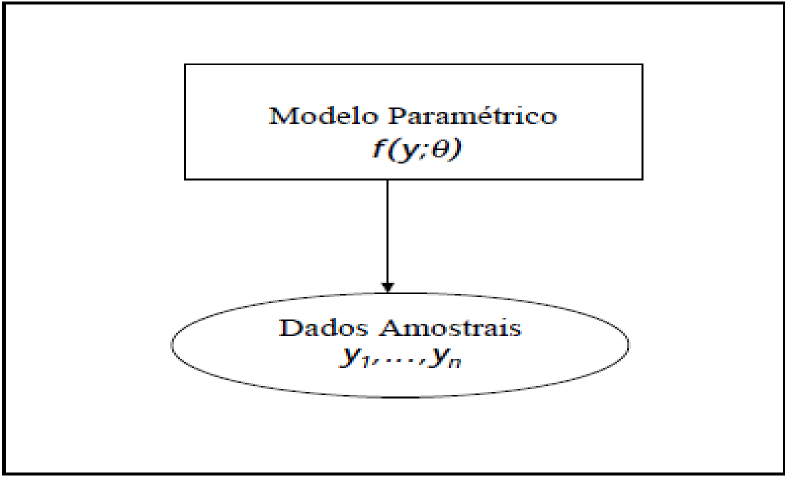
Populações infinitas (modelagem clássica)

- Obter conclusões acerca de parâmetros (de uma ou mais populações) com base em uma amostra (ou mais de uma).
- Usualmente: cada elemento tem a mesma probabilidade de ser selecionado.
  - Interesse: amostra aleatória $Y_1, ..., Y_n$ (conjunto de possíveis valores observados), regida por alguma distribuição de probabilidade (paramétrica).
  - Inferência a respeito dos parâmetros dessa distribuição.
  - Métodos de estimação, construção de IC's e testes de hipótese.
- Usualmente feito na disciplina de Inferência (link 1, link 2).

Populações infinitas (modelagem clássica)

Objeto	Descrição
Dados Amostrais	$Y_1 \quad \ldots \quad Y_n$ $\downarrow \quad \ldots \quad \downarrow$ $y_1 \quad \ldots \quad y_n$
Modelo paramétrico/hipóteses	$Y_1, .., Y_n$ variáveis aleatórias IID com distribuição $f(y; \boldsymbol{\theta})$, onde $\boldsymbol{\theta} \in \Theta$
Objetivo	Inferir sobre $\boldsymbol{\theta}$ usando as observações $y_1, ..., y_n$

Populações infinitas (modelagem clássica)



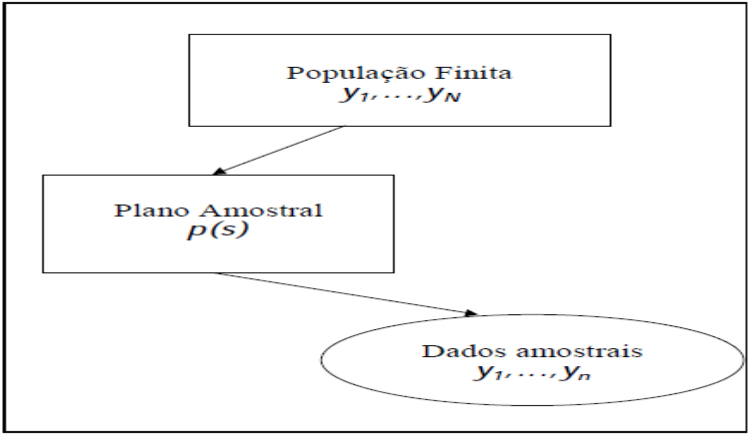
Populações finitas (modelagem probabilística)

- População composta por N elementos.
- Define-se alguma característica de interesse dessa população: média, variância, total, coeficiente de variação etc, associados a alguma variável cujos valores na população são dados por $y_1, ..., y_N$.
- Processo de amostragem: cada elemento pode ou não ter a mesma probabilidade de ser selecionado.
  - Interesse: amostra aleatória (selecionado segundo algum **plano amostral (PA** $\equiv p(s) \equiv P(s)$**))** $Y_1, ..., Y_n$ (conjunto de possíveis valores observados). Pode se considerar, à rigor, que a amostra é aleatória simples com reposição (AAS_c), embora na prática, não o seja.
  - Probabilidades de ocorrência de cada amostra: **função do plano amostral**.
  - Inferência a respeito dos parâmetros da população de interesse.
  - Estimadores definidos com base nos parâmetros populacionais.

Populações finitas (modelagem probabilística)

Objeto	Descrição
Dados Amostrais	$Y_1 \quad \ldots \quad Y_n$ $\downarrow \quad \ldots \quad \downarrow$ $y_1 \quad \ldots \quad y_n$
Modelo/hipóteses	extraídos de $y_1, ..., y_N$
Objetivo	Inferir sobre funções $g(y_1, ..., y_N)$ usando $y_1, ..., y_n$

Populações finitas (modelagem probabilística)



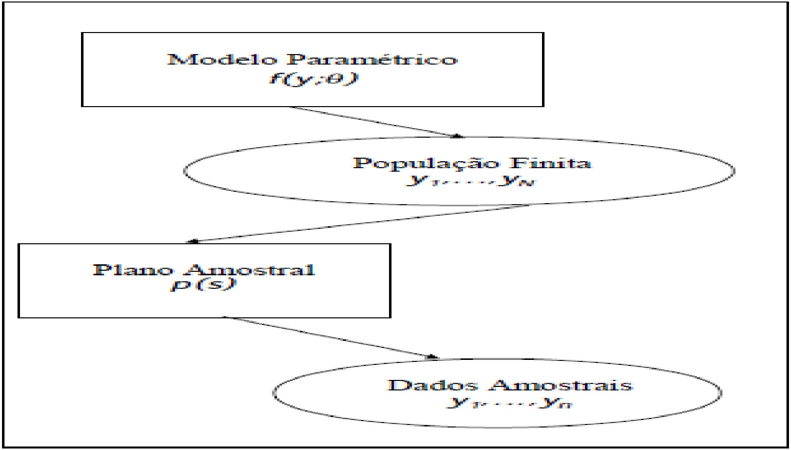
Populações finitas (modelagem de superpopulação)

- Consideramos, essencialmente, a situação anterior (população finita).
- Mas, neste caso, os valores populacionais $y_1, ..., y_N$, são considerados como observações ou realizações das variáveis aleatórias $Y_1, ..., Y_N$ supostas IID com distribuição $f(.;\boldsymbol{\theta})$, em que $\boldsymbol{\theta} \in \Theta$.
- Utilizando um plano amostral definido por $p(s)$, obtemos os valores na amostra $y_1, ..., y_n$.
- A partir de $y_1, ..., y_n$ (não considerados IID, em geral) queremos fazer inferências sobre o parâmetro θ.
- Temos, assim, uma combinação entre as duas modelagens anteriores.

Populações finitas (modelagem de superpopulação)

Objeto	Descrição
Dados Amostrais	$\begin{matrix} Y_1 & \dots & Y_n \\ \downarrow & \dots & \downarrow \\ y_1 & \dots & y_n \end{matrix}$
População e esquema de seleção	Extraídos de $y_1, \ldots, y_N$ segundo $p(s)$
Modelo para a população	$Y_1, \ldots, Y_N$ variáveis aleatórias IID com distribuição $f(y; \boldsymbol{\theta})$, em que $\boldsymbol{\theta} \in \Theta$
Parâmetro-alvo	associar $\boldsymbol{\theta} \leftrightarrow g(Y_1, \ldots, Y_N)$
Objetivo	Inferir sobre $g(y_1, \ldots, y_N)$ a partir de $y_1, \ldots, y_n$ usando $p(s)$

Populações finitas (modelagem de superpopulação)



Estrutura geral (exemplo)

- População alvo: população de interesse. **Exemplo: pessoas residentes no Brasil no mês de Agosto de 2011.**
- População referenciada: subconjunto da população alvo para a qual está disponível um **sistema de referência (cadastro que permita implementar o plano amostral)**. **Exemplo: pessoas residentes no Brasil no mês de Agosto de 2011 das quais se possui o endereço correto.**
- População amostrada: subconjunto da população referenciada da qual, efetivamente, é possível retirar uma amostra. **Exemplo: pessoas residentes no Brasil no mês de Agosto de 2011 das quais se possui o endereço correto e não residem em localidades de difícil acesso.**

Estrutura geral cont.

- Sistema de referência: banco de dados contendo informações sobre os elementos da população referenciada (organizado de modo a permitir implementar o PA).
- Unidade elementar: elemento da população portadora das informações de interesse. **Exemplo: Eleitor brasileiro**.
- Unidade amostral: entidade que será selecionada no processo de amostragem. Pode ser formada por uma ou mais de uma unidades elementares. **Exemplo: Domicílio**.
- Unidade resposta: entidade que fornece as informações de interesse relacionadas à unidade amostral. **Exemplo: Pessoa responsável pelo sustento do domicílio.**

Plano amostral (PA)

- PA: conjunto de procedimentos que definem como a amostra deve ser selecionada. Ela determina a probabilidade de cada elemento ser selecionado.
- O PA pode influenciar o comportamento do estimador: esperança, vício, erro quadrático médio.
- Ignorar o PA pode comprometer o processo de inferência, e.g., subestimar ou superestimar a variância do estimador, subestimar ou superestimar as probabilidades de cometimento de erro em testes de hipótese.

Cont.

- Objetivos:
  - Definir o PA que produza uma amostra que conduza a inferências apropriadas (junto com a escolha de um estimador apropriado).
  - Incorporar a estrutura do PA na construção e obtenção das características (esperança, variância) de estimadores.
- Amostra representativa : definição problemática.
- Amostra probabilística: selecionada segundo algum plano amostral (regido **totalmente** por mecanismos de sorteios aleatórios).
- Amostra sistemática: selecionada segundo algum plano amostral (regido **parcialmente** por mecanismos de sorteio aleatório).

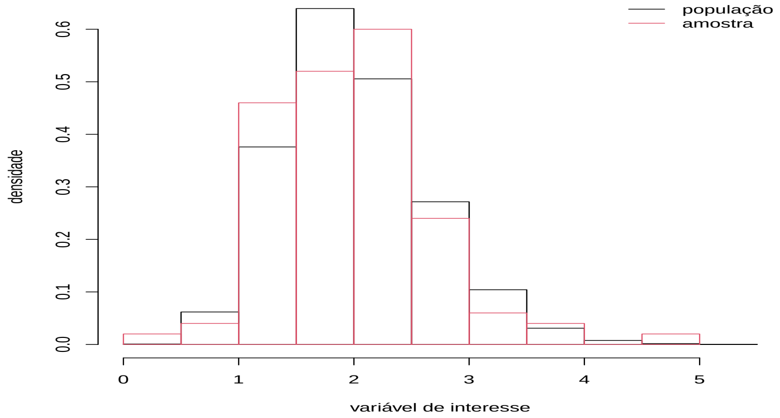
Tipos de amostra

Critério do "amostrista"	**Procedimento de seleção**	
	probabilístico	**não-probabilístico**
objetivo	amostra probabilísticas	amostra criteriosa
subjetivo	amostra quase-aleatória	amostra intencional

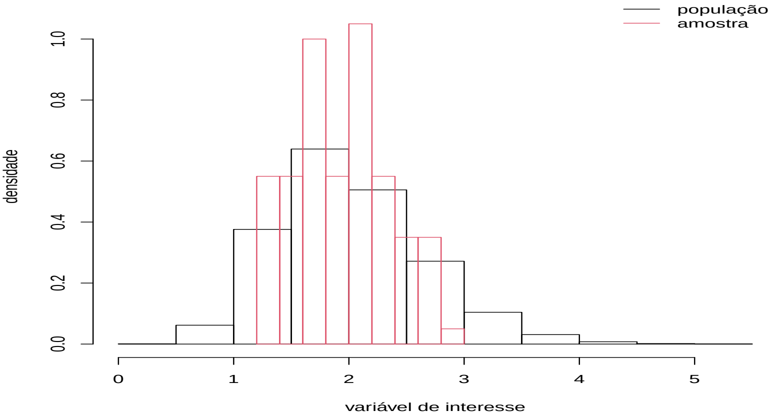
amostrista: responsável por (ao menos) definir o PA.

Nos slides seguintes mostraremos um exemplo de população e de várias amostras.

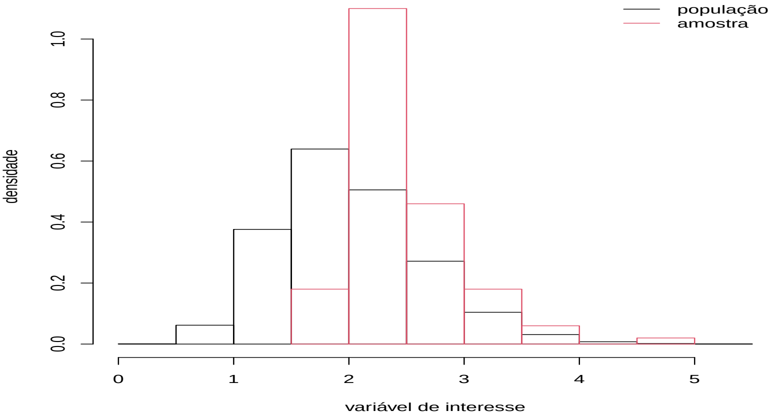
Exemplo de população (amostra 1)



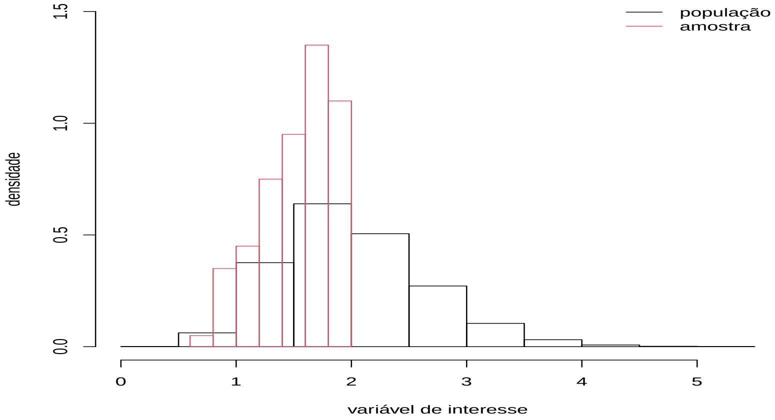
Exemplo de população (amostra 2)



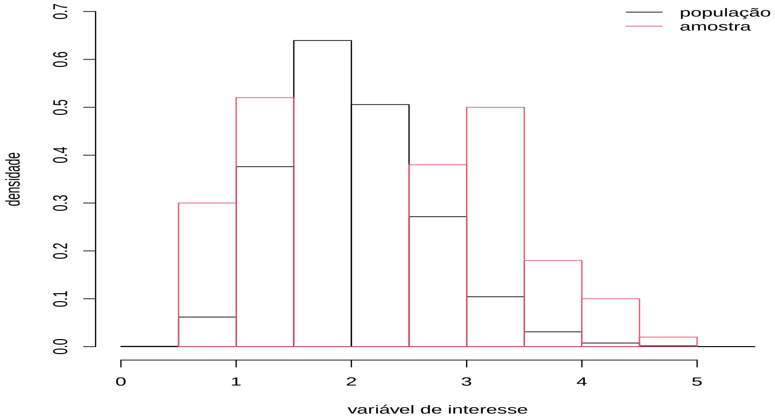
Exemplo de população (amostra 3)



Exemplo de população (amostra 4)



Exemplo de população (amostra 5)



Classificação de amostras probabilísticas

- Probabilidade de seleção da unidade amostral: **igual** ou distinta.
- Unidade amostral: **uma unidade de resposta (elementar)** ou **elementos (conglomerado)**.
- Divisão em estratos: **não estratificada** ou **estratificada**.
- Número de estágios: **um único** ou mais de um (**somente dois**).
- Seleção de unidades: **aleatória** ou sistemática.

Erros (imprecisões) envolvidos

- Censo: analisa-se todos os elementos da população de interesse (não há erros relativos à processos de amostragem).
- Amostragem: analisa-se uma parte da população de interesse (há erros (imprecisões) relativos à processos de amostragem).
- Erros (imprecisões) amostrais: desvios existentes entre as estimativas e os verdadeiros valores dos parâmetros decorrentes do processo amostral (controláveis através do PA).
- Erros não amostrais: desvios existentes entre as estimativas e os verdadeiros valores dos parâmetros decorrentes de fatores não inerentes ao processo amostral (não-controláveis através do PA).

Definições/Notações

- **População ou universo:** $\mathcal{U} = \{1, 2, ..., N\}$, em que N é o tamanho da população, que pode ser conhecido ou desconhecido.
- **Elemento populacional:** elemento pertencente à $\mathcal{U}$, i.e., $i \in \mathcal{U}$.
- **Característica(s) de interesse:** variável ou vetor de variáveis associado a cada elemento da população.
- Representação: y_i (univariado) ou $\boldsymbol{y}_i = (y_{i1}, ..., y_{ip})^t_{p \times 1}$, (multivariado), $i \in \mathcal{U}$ (transposto: $'$ ou ^{t}).
- Usaremos letras maiúsculas para representar variáveis aleatórias (e.g., X) e letras minúsculas para representar variáveis não aleatórias ou valores observados de variáveis aleatórias (e.g., x).

Cont.

- **Parâmetro populacional:** denota o vetor correspondente a todos os valores de uma variável de interesse (ou um vetor de variáveis). Notação: $\boldsymbol{d} = (y_1, ..., y_N)'_{N\times 1}$ (ou $\boldsymbol{d}' = (y_1, ..., y_N)_{1\times N}$) ou $\boldsymbol{d} = (\boldsymbol{y}'_1, ..., \boldsymbol{y}'_N)_{N\times p}$ (ou $\boldsymbol{d} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_N)_{p\times N}$).
- **Função paramétrica populacional (ou parâmetro):** característica numérica qualquer de uma população. Notação: $\theta(\boldsymbol{d})$ ou θ.
- **Amostra ordenada:** qualquer sequência de n unidades de $\mathcal{U}$, ou seja, $\boldsymbol{s} = (k_1, ..., k_n), k_i \in \mathcal{U}$.
- **Espaço amostral:** conjunto de todas as amostras (de todos os tamanhos) de $\mathcal{U}$ ($\mathcal{S}(\mathcal{U})$); subclasse de todas as amostras de tamanho n ($\mathcal{S}_n(\mathcal{U})$).

Exemplo

- Considere uma população formada por três domicílios, $\mathcal{U} = \{1, 2, 3\}$ e que se observam as seguintes variáveis: nome (do chefe), sexo, idade, fumante ou não, renda bruta (mensal em salários mínimos) familiar e número de trabalhadores.

Cont. (População de três domicílios)

Variável	**Valores**			**Notação**
Unidade	1	2	3	i
nome do chefe	Ada	Beto	Ema	a_i
sexo	0	1	0	x_i
idade	20	30	40	y_i
fumante	0	1	1	g_i
renda bruta familiar	12	30	18	f_i
n^o de trabalhadores	1	3	2	t_i

Exemplo

- Parâmetro populacional
  - Variável idade: $\boldsymbol{d} = (20, 30, 40)'$.
  - Variáveis (f, t): $\boldsymbol{d}' = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix}$
- Função paramétrica (chamaremos de parâmetros):
- Média das variáveis renda e número de trabalhadores: $\boldsymbol{\theta} = (20, 2)^t$
- Renda média por trabalhador: $\theta = 10$

Parâmetros de interesse

- Em geral, os parâmetros de interesse associados à uma dada variável são:
  - Total populacional: $\tau = \sum_{i=1}^{N} y_i$
  - Média populacional: $\mu = \frac{\tau}{N} = \frac{1}{N} \sum_{i=1}^{N} y_i$.
  - Variância populacional: $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2$
  - Proporção populacional: $p = \frac{1}{N} \sum_{i=1}^{N} y_i$, quando $y_i \in \{0, 1\}, \forall i$.

Planejamento amostral

- **Planejamento amostral:** mecanismo que associa (estabelece) a probabilidade de ocorrência de cada amostra possível ($P(\boldsymbol{s})$), ou seja

$$P(.) \quad : \quad \mathcal{S}(\mathcal{U}) \rightarrow [0,1]$$

$$P(\boldsymbol{s}) \geq 0, \forall \boldsymbol{s} \in \mathcal{S}(\mathcal{U}) \equiv \mathcal{S} \quad , \quad \sum_{\boldsymbol{s}:\boldsymbol{s} \in \mathcal{S}} P(\boldsymbol{s}) = 1$$

- Conjunto das amostras possíveis sob o plano amostral A: $\mathcal{S}_A$ (nesse caso, o "tamanho da amostra" pode ou não variar entre as amostras possíveis).

Cont. do Exemplo

- Exemplos de amostra: $\boldsymbol{s}_1 = (1, 2)$, $\boldsymbol{s}_2 = (2, 1)$, $\boldsymbol{s}_3 = (1, 1, 3)$, $\boldsymbol{s}_4 = (3)$, $\boldsymbol{s}_5 = (2, 2, 1, 3, 2)$
- Conjunto de todas as amostras possíveis (com reposição): $\mathcal{S}(\mathcal{U}) = \{(1), (2), (3), (1, 1), (1, 2), ..., (2, 2, 1, 3, 2), ...)\}$
- Conjunto de todas as amostras possíveis de tamanho 2 (com reposição): $\mathcal{S}_2(\mathcal{U}) = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}$.
- Exercício: obter todas as amostras possíveis para um sorteio sem reposição.

Cont. do Exemplo: planos amostrais

- Plano A: Sorteia-se com igual probabilidade um elemento de $\mathcal{U}$ e anota-se a unidade sorteada. Este elemento é devolvido à população e sorteia-se um segundo elemento do mesmo modo.

  $\mathcal{S}_A = \{11, 12, 13, 21, 22, 23, 31, 32, 33\}$.

$$P_A(\boldsymbol{s}) = \begin{cases} 1/9, & \text{se } i \in \boldsymbol{s} \\ 0, & \text{se } i \notin \boldsymbol{s} \end{cases}$$

Cont. do Exemplo: planos amostrais

- Plano B: Sorteia-se com igual probabilidade um elemento de $\mathcal{U}$ e anota-se a unidade sorteada. Este elemento é retirado da população e sorteia-se um segundo elemento do mesmo modo.

  $\mathcal{S}_B = \{12, 13, 21, 23, 31, 32\}$.

$$P_B(\boldsymbol{s}) = \begin{cases} 1/6, & \text{se } i \in \boldsymbol{s} \\ 0, & \text{se } i \notin \boldsymbol{s} \end{cases}$$

Cont. do Exemplo: planos amostrais

- Plano C: Sorteia-se um elemento de $\mathcal{U}$ com probabilidade proporcional ao número de trabalhadores. Sem repor o domicílio selecionado, sorteia-se um segundo também com probabilidade proporcional ao número de trabalhadores.

  $\mathcal{S}_C = \{12, 13, 21, 23, 31, 32\}$.

$$P(12) = \frac{1}{6} \times \frac{3}{5} = \frac{1}{10} \quad ; \quad P(21) = \frac{3}{6} \times \frac{1}{3} = \frac{1}{6}$$
$$P(13) = \frac{1}{6} \times \frac{2}{5} = \frac{1}{15} \quad ; \quad P(31) = \frac{2}{6} \times \frac{1}{4} = \frac{1}{12}$$
$$P(23) = \frac{3}{6} \times \frac{2}{3} = \frac{1}{3} \quad ; \quad P(32) = \frac{2}{6} \times \frac{3}{4} = \frac{1}{4}$$

Inferência

- Dada uma amostra $\boldsymbol{s} = (k_1, k_2, ..., k_n)$ tem-se uma característica (y_{k_i}, no caso univariado) ou um vetor de características ($\boldsymbol{y}_{k_i} = (y_{k_i 1}, ..., y_{k_i p})^t$, no caso multivariado), para cada unidade amostral (observação não aleatória).
- **Dados da amostra $\boldsymbol{s}$:** vetor - $\boldsymbol{d_s} = (y_{k_1}, ..., y_{k_n})^t$ ou matriz $\boldsymbol{d_s} = (\boldsymbol{y}_{k_1}^t, ..., \boldsymbol{y}_{k_n}^t)^t$, $k_i \in \boldsymbol{s}$.
- **Conjunto de todas as amostras possíveis**: Nesse caso, teremos quantidades aleatórias

$$
\begin{aligned}
\boldsymbol{D_s} &= \boldsymbol{Y} = (Y_1, ..., Y_i, ..., Y_n)^t \,(\text{vetor}) \\
\boldsymbol{D_s} &= \boldsymbol{Y} = (\boldsymbol{Y}_1^t, ..., \boldsymbol{Y}_i^t, ..., \boldsymbol{Y}_n^t)^t \,(\text{matriz}),
\end{aligned}
$$

em que $Y_i(\boldsymbol{Y}_i)$ é a va (vea) que representa os valores possíveis de ocorrerem na i-ésima posição da amostra.

Exemplo

- Exemplo da população com três domicílios: N=3.
- Variável idade : $\boldsymbol{d} : (y_1, y_2, y_3)' = (20, 30, 40)'$ (variável não aleatória).
- Suponha n= 2 (tamanho da amostra) selecionando esses indivíduos sem reposição, então $\boldsymbol{D_s} = \boldsymbol{Y} = (Y_1, Y_2)'$ (variável aleatória).
- Se por exemplo, $\boldsymbol{s} = (1, 3)$, então $\boldsymbol{d_s} = (y_1, y_3)' = (20, 40)'$.
- Nesse caso, os valores observados de Y_1 e Y_2 são, respectivamente, $y_1 = 20$ e $y_3 = 40$.
- Repetir o raciocínio para, simultaneamente, as variáveis (f, t).

Cont.

- **Estatística (estimador)**: função da amostra $T = h(\boldsymbol{Y})$, $t = h(\boldsymbol{y})$ (valor observado).
- **Distribuição amostral de uma estatística segundo um PA, digamos A:**

$$p_h = P_A\left(\boldsymbol{s} \in \mathcal{S}_A : h(\boldsymbol{y}) = h\right) = P(h) \tag{1}$$

- **Valor esperado de T sob o plano amostral A:** $\mathcal{E}_A(T) = \sum_h h p_h$.
- **Variância de T sob o plano amostral A:** $\mathcal{V}_A(T) = \sum_h \left(h - \mathcal{E}_A(T)\right)^2 p_h = \mathcal{E}_A(T^2) - \mathcal{E}_A^2(T)$.

Cont.

- Sejam T e G duas estatísticas.
- **Covariância entre T e G :** $Cov(T,G) = \mathcal{E}_A(TG) - \mathcal{E}_A(T)\mathcal{E}_A(G)$.
- **Correlação entre T e G:** $Corre(T,G) = \frac{Cov(T,G)}{\sqrt{\mathcal{V}_A(T)\mathcal{V}_A(G)}}$.

Cont. do Exemplo: Estatísticas e distribuições amostrais

- Considere que o interesse reside nas variáveis (f, t):

$$\boldsymbol{d} = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix}$$

- Considere os planos amostrais A e B e a estatística $R = h(\boldsymbol{D_s})$: razão entre o total da renda familiar e o número de trabalhadores.

Cont. do Exemplo: Estatísticas e distribuições amostrais

- Plano amostral A:

s	11	12	13	21	22	23	31	32	33
$P(\boldsymbol{s})$:	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
$h(\boldsymbol{d_s}) = r$:	12	10,5	10	10,5	10	9,6	10	9,6	9

- Distribuição amostral de R sob o plano amostral A:

h	9	9,6	10	10,5	12
p_h	1/9	2/9	3/9	2/9	1/9

Cont. do Exemplo: Estatísticas e distribuições amostrais

- Plano amostral B:

$\mathbf{s}$	12	13	21	23	31	32
$P(\boldsymbol{s})$:	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$
$h(\boldsymbol{d_s}) = r$:	10,5	10	10,5	9,6	10	9,6

- Distribuição amostral de R sob o plano amostral B:

h	9,6	10	10,5
p_h	$1/3$	$1/3$	$1/3$

Tipos de erros

- Erros-amostrais: decorrentes do plano amostral e via de regra quantificáveis através do erro-padrão do estimador (geralmente viável sob amostragem probabilística).
- Erros-não amostrais: quando ocorrem problemas no sistema de referência, na coleta de dados (falta de respostas - "missing data"), transcrição de dados etc. Eventualmente, podem ser identificados numa análise descritiva, caso já não o tenham sido antes. Eventualmente, podem ser contornados (depende de vários fatores).
- Em particular, a "ocorrência" de dados faltantes pode ter impacto nas análises estatísticas (incorporar a modelagem da ocorrência dos dados faltantes).

Incorporação do planejamento amostral

- Levar em consideração a probabilidade de cada indivíduo aparecer na amostra e não a "distribuição de frequências" da população.
- Estimadores que levem em consideração, na sua fórmula/distribuição, a estrutura do planejamento amostral.
- Incorporar o planejamento amostral no cálculo da esperança, variância etc do estimador.
- Exercício: Estudar o Capítulo 1 do livro "Elementos de Amostragem"e os Capítulo 1 e 2 do livro "Amostragem: Teoria e Prática usando R ".