Amostragem

Fernando Ferreira[1]

[1] Departamento de Medicina Veterinária Preventiva e Saúde Animal
Universidade de São Paulo

Epidemiologia

Resumo

Resumo

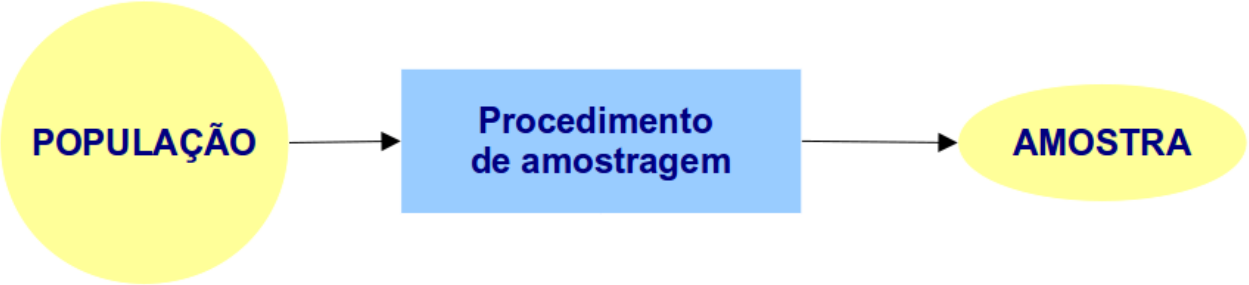
População e Amostra

População

Conjunto de elementos que apresentam, em comum, determinada característica. Pode ser finita ou infinita.

Amostra

Subconjunto da população.



Objetivos da amostragem

Estimar parâmetros populacionais.

- Totais
- Médias
- Proporções
- Variância

Por que trabalhar com amostra?

- Em geral é difícil / impraticável avaliar a população toda
  - Limitações: tempo, recursos, capacidade operacional
- Censos podem custar caro.
- Amostras podem ser obtidas mais rapidamente.
- Censo é a única maneira de se medir exatamente a distribuição de uma variável na população.

Resumo

Definições

- População Alvo
- População Estudada
- Unidades Elementares/Amostrais
- Estratos
- Quadro Amostral
- Fração Amostral

Definições I

População alvo é a população sobre a qual a informação é requerida.

População de estudo é a população da qual a amostra é obtida.
Em geral, a população de estudo deveria ser a população alvo, mas nem sempre é possível. Ex. Deseja-se estimar a incidência de um agravo em cães da raça maltês, mas só é possível realizar a observação em exposições ou em clínicas e hospitais veterinários.

Unidades elementares / amostrais a população de estudo é dividida em unidades elementares. Em geral, a unidade amostral é o animal.

Definições II

Estrato é o agrupamento de unidades amostrais, feito com base em alguma característica comum (sexo, idade, raça, tipo de produção, tamanho da propriedade, ...)

Quadro amostral é constituído, por exemplo, de uma lista que contém os membros da população de estudo. Ex. lista de clínicas veterinárias, lista de propriedades rurais de uma região, etc. Cuidado: O quadro amostral pode estar incompleto!

Fração amostral é a razão entre o tamanho da amostra e o tamanho da população de estudo. Ex. Se 10 animais foram analisados em uma população de 1000, a fração amostral é de 1%.

Vícios

Estimativas tendenciosas podem ser obtidas se:

- Lista de membros do Quadro é incompleta
- Informação é obsoleta
- Segmentos da população não são traçáveis
- Recusa na participação / cooperação
- Procedimentos de amostragem não são aleatórios

O vício (viés / tendenciosidade) introduzido por estes processos não é compensado pelo aumento do tamanho da amostra

Acurácia e Precisão

Acurácia

Conceito associado ao valor verdadeiro da variável. A introdução de viés leva à perda de acurácia.

Precisão

Conceito associado à variabilidade (reprodutibilidade) de uma medida. Frequentemente medida pelo desvio-padrão, nesse caso, quando menor o desvio-padrão maior a precisão.

Acurácia e Precisão

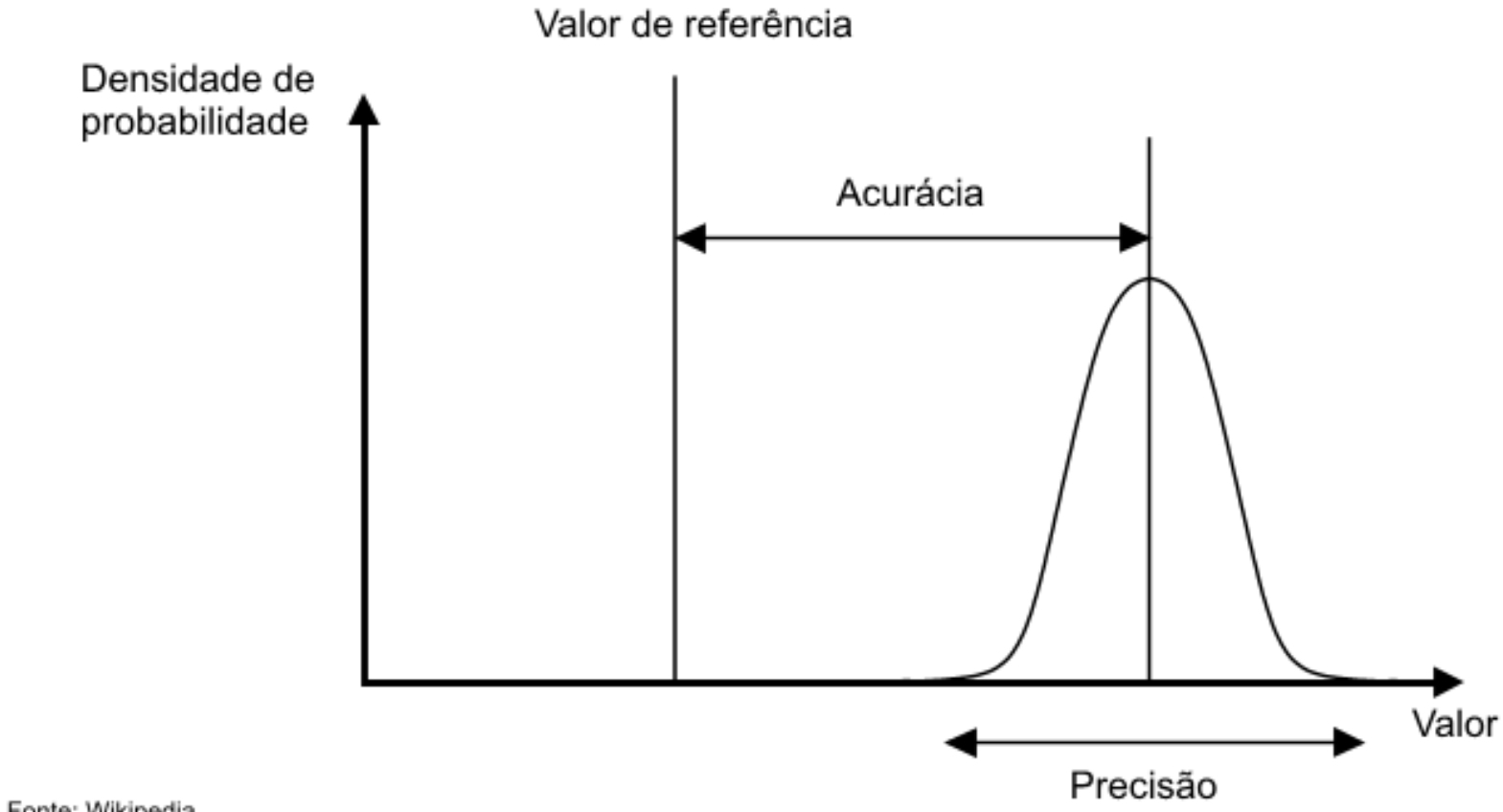
Preciso
Pouco acurado

Pouco preciso
Pouco acurado

Pouco preciso
Acurado

Preciso
Acurado

Acurácia e Precisão



Fonte: Wikipedia

Amostra Representativa

Permite estimar um parâmetro populacional de forma acurada e com precisão conhecida.

Questões I

- Para levantar dados sobre o número de filhos por casal, em uma comunidade, um pesquisador organizou um questionário que enviou, pelo correio, a todas as residências. A resposta ao questionário era facultativa. Perguntava-se qual o número de filhos por casal morador na residência. Os dados obtidos podem apresentar algum tipo de *tendenciosidade*? (Questões propostas por Vieira (1980))
- Um pesquisador pretende levantar dados sobre número de habitantes por domicílio. Para isso, visitará cada domicílio selecionado. Se nenhuma pessoa estiver no domicílio, excluirá o domicílio da amostra. Isto introduz um *vício* na amostra. Por quê?

Questões II

- Para verificar se as famílias se tornaram menores, foi selecionada uma amostra de 2000 casais e perguntou-se quantos filhos eles tinham, quantos filhos tinham seus pais e quantos filhos tinham seus avós. Esse procedimento introduz um *viés* nos dados. Por quê?

Natureza das Unidades Amostrais

- Indivíduos
- Agregados de Indivíduos (rebanhos, fazendas, regiões administrativas, ...)

É importante que as Unidades Amostrais sejam as Unidades Epidemiológicas (grupo de animais de importância epidemiológica para transmissão e manutenção da infecção)

Tipos de Amostragem

Amostragem Probabilística

Cada unidade amostral tem probabilidade conhecida e diferente de zero de ser selecionada.

Amostragem Não Probabilística

Qualquer amostra que não siga o critério acima

Métodos de Amostragem

- Não Probabilísticas
  - Amostra de Conveniência
  - Amostra Intencional
- Probabilísticas
  - Amostra Aleatória Simples
  - Amostra Sistemática
  - Amostra Estratificada
  - Amostra por Conglomerados (Clusters)

Resumo

Amostragem Não Probabilística

Amostra de Conveniência

Seleção de unidades amostrais mais facilmente acessíveis. Ex. pesquisas com animais de um determinado hospital veterinário.

Amostra Intencional

São selecionados indivíduos com características próximas do valor médio da variável da população alvo.
Obs.: a amostra pode não ser “representativa” e pode apresentar viés.

Resumo

Amostra Probabilística

- Amostra Aleatória (Casual) Simples
  - Amostra aleatória selecionada a partir do quadro amostral
- Amostra Sistemática
  - Os elementos são escolhidos não por acaso, mas por um sistema. Ex. entrevistas feitas de 3 em 3 domicílios para aplicação de um questionário sobre população canina.
  - Não requer o conhecimento do tamanho total da população em estudo
- Amostra Estratificada
  - É obtida subdividindo-se a população em grupos (estratos), selecionando-se aleatoriamente os indivíduos em cada estrato
  - Estratificação melhora a acurácia da amostra pois controla a possibilidade de, em uma amostra aleatória simples, super ou sub representar algum estrato do quadro amostral

Amostra Probabilística

- Amostra por Conglomerados (Clusters)
  - Conglomerados definidos por regiões geográficas como bacias, zonas produtoras, vilas, rebanhos, etc. Todos os animais do cluster são amostrados (amostragem em 1 estágio).
- Delineamentos Amostrais Complexos
  - Amostragem por Conglomerados em Múltiplos Estágios
    - Ex. Amostragem aleatória de clusters seguida de amostragem aleatória simples das unidades pertencentes ao cluster.
    - Levantamento de prevalência de brucelose (PNCEBT)

Tamanho da Amostra

- Fatores não estatísticos
  - Infraestrutura (tempo, custos)
  - Existência de quadros amostrais
- Fatores estatísticos
  - Precisão
  - Frequência esperada

Distribuição das Médias Amostrais

Considere a variável aleatória x, medida em indivíduos de uma população que se deseja estudar. Suponha que essa variável apresente média μ e desvio-padrão σ.
O teorema do limite central (TLC) garante que se tomarmos um número suficientemente grande de amostras de tamanho n dessa população, a distribuição das médias amostrais segue uma distribuição normal com média igual a $x_m = \mu$ e desvio-padrão igual a $\sigma_m = \frac{\sigma}{\sqrt{n}}$.
Simulador TLC
`http://onlinestatbook.com/stat_sim/sampling_dist/index.html`

Tamanho de amostra

Considere que desejemos estimar o valor médio da variável aleatória x numa população com base em uma amostra simples aleatória. O TLC nos permite estimar o intervalo de confiança para valor médio da variável como sendo igual a:

$$\overline{x} \pm z_c \sigma_m \quad (1)$$

$$\overline{x} \pm z_c \frac{\sigma}{\sqrt{n}} \quad (2)$$

onde $z_c = 1.96$ para um intervalo com 95% de confiança, $\overline{x}$ é a média da amostra, σ é o desvio-padrão da variável x na população e n é o tamanho da amostra.

Tamanho de amostra

O segundo termo da equação 2 pode ser chamado de erro da estimativa (ϵ)

$$\overline{x} \pm \underbrace{z_c \frac{\sigma}{\sqrt{n}}}_{\epsilon} \tag{3}$$

Dessa forma, podemos estabelecer o tamanho de uma amostra n necessário para estimar o valor médio de uma variável considerando a magnitude do erro associado à estimativa e o nível de confiança desejado. A fórmula é dada por:

$$\begin{aligned} \epsilon &= z_c \frac{\sigma}{\sqrt{n}} \\ n &= z_c^2 \frac{\sigma^2}{\epsilon^2} \end{aligned} \tag{4}$$

Tamanho de amostra

No caso de populações finitas, ou seja, em que o tamanho da amostra (n) não é muito menor que o tamanho da população (N), o tamanho da amostra calculado deve ser corrigido utilizando-se:

$$n_{aj} = \left(\frac{N}{N+n} \right) n \tag{5}$$

Exemplo

Um lote de suínos com 1000 animais apresenta um peso médio de 78.0 kg e desvio padrão de 7.0 kg. Quantos animais devem ser pesados para que tenhamos uma estimativa da média com nível de confiança de 95% e precisão (erro) de 2.0 kg?

Resposta

dados:

$$\begin{aligned} z_c &= 1.96 \\ \sigma &= 7kg \\ \epsilon &= 2kg \end{aligned}$$

solução:

$$\begin{aligned} n &= z_c^2 \frac{\sigma^2}{\epsilon^2} = 1.96^2 \frac{7^2}{2^2} \simeq 48 \\ n_{aj} &= \frac{N}{N+n} n = \frac{1000}{1000+48} 48 \simeq 46 \end{aligned}$$

Variáveis binomiais

Se a variável que estamos trabalhando não é quantitativa mas do tipo binomial, utilizamos a aproximação normal da binomial e de modo semelhante podemos estimar o tamanho da amostra por:

$$\begin{aligned} \overline{p} &\pm \underbrace{z_c \frac{\sigma}{\sqrt{n}}}_{\epsilon} \\ \epsilon &= z_c \frac{\sigma}{\sqrt{n}} \\ n &= z_c^2 \frac{\sigma^2}{\epsilon^2} \end{aligned}$$

como $\sigma = \sqrt{p(1-p)}$, temos:

$$n = z_c^2 \frac{p(1-p)}{\epsilon^2} \quad (6)$$

Tamanho de amostra

No caso de populações finitas, ou seja, em que o tamanho da amostra (n) não é muito menor que o tamanho da população (N), o tamanho da amostra calculado deve ser corrigido utilizando-se:

$$n_{aj} = \left(\frac{N}{N+n} \right) n \tag{7}$$

Exemplo

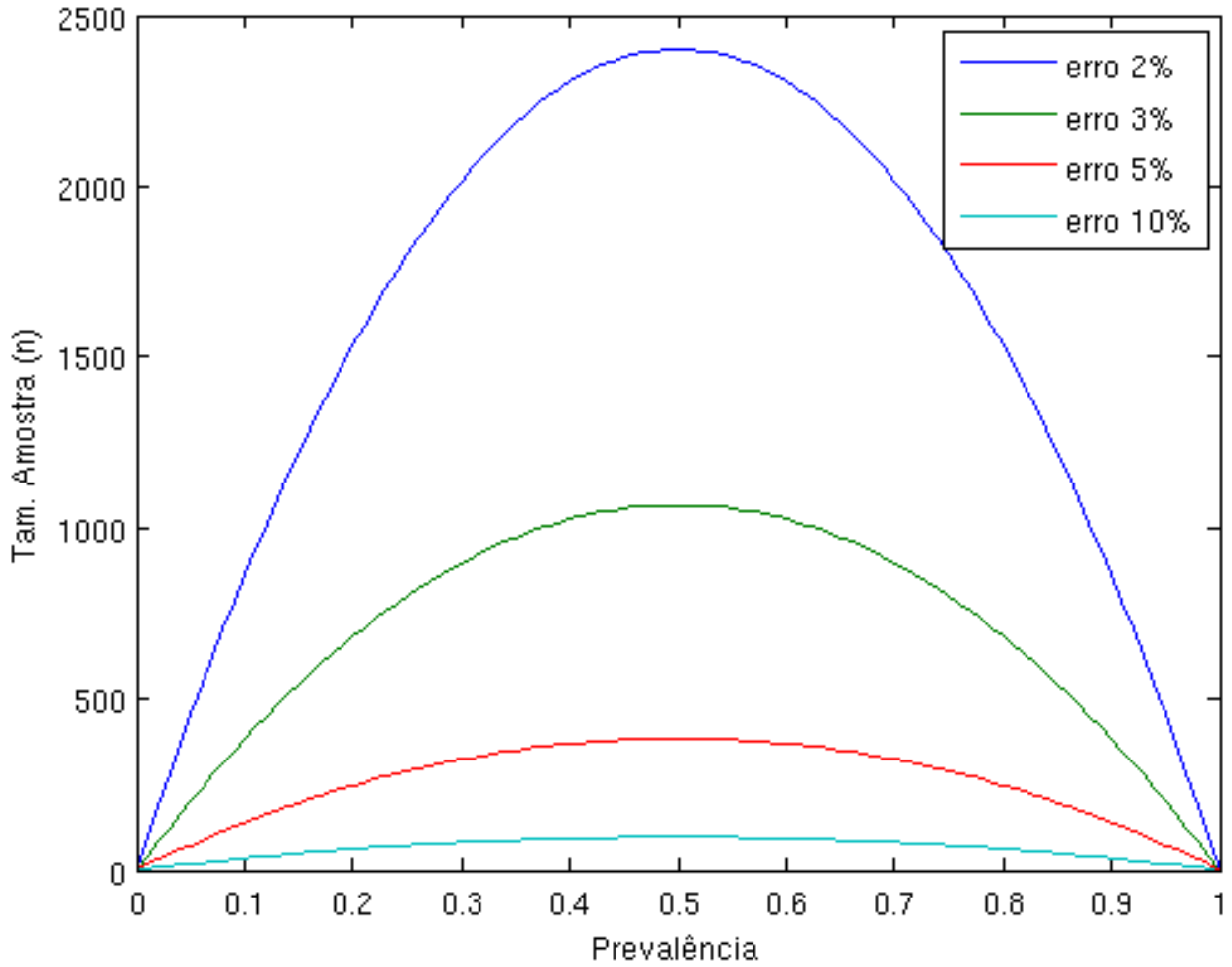
Um proprietário de uma fazenda com 10,000 bovinos deseja selecionar uma amostra para estimar a frequência de animais soropositivos para a brucelose. Qual o tamanho da amostra para estimar a frequência com confiança de 95% e erro de 5%?

Resposta

dados:

$$
\begin{aligned}
z_c &= 1.96 \\
p &= ? \\
\epsilon &= 0.05
\end{aligned}
$$

Que valor de p utilizar?



Resposta

dados (experiência anterior sugere p ao redor de 10%):

$$\begin{aligned} z_c &= 1.96 \\ p &= 0.10 \\ \epsilon &= 0.05 \end{aligned}$$

solução:

$$n = z_c^2 \frac{p(1-p)}{\epsilon^2} = 1.96^2 \frac{0.1(1-0.1)}{0.05^2} \simeq 139$$

$$n_{aj} = \frac{N}{N+n} n = \frac{10000}{10000+139} 139 \simeq 138$$

Amostra em dois estágios

A utilidade deste método, quando comparado ao processo aleatório simples, está associada ao menor custo na coleta das amostras e à possibilidade de realizar a amostragem na ausência de quadros amostrais detalhados, ou seja, a falta de um cadastro de domicílios.
Para o procedimento descrito a seguir será necessário a obtenção da lista de setores censitários do município junto ao IBGE ou, eventualmente, à secretaria de planejamento do município.

Notação I

- M = número de setores censitários no município.
- m = número de setores censitários na amostra.
- N = número de domicílios no município.
- N_i = número de domicílios no setor censitário i.
- n = número de domicílios na amostra.
- $\overline{N} = \frac{\sum_{i=1}^{M} N_i}{M}$ = número médio de domicílios por setor censitário
- n_i = número de domicílios selecionados no setor censitário i.
- $\overline{n} = \frac{\sum_{i=1}^{m} n_i}{m}$ = número médio de domicílios amostrados por setor censitário.
- X = população de cães/gatos total no município.

Notação II

- X_i = número total de cães/gatos no setor censitário i.
- X_{ij} = número de cães/gatos no domicílio j do setor censitário i.
- x_i = número total de cães/gatos nos domicilios amostrados do setor i.
- x'_{ppt} = estimativa da população de cães/gatos total no município.
- $x = \sum_{i=1}^{m} x_i$ = número total de cães/gatos na amostra.
- $\pi'_i = \frac{N_i}{N}$ = probabilidade de seleção do setor censitário i.
- $\overline{X} = \frac{\sum_{i=1}^{M} X_i}{M}$ = número médio de cães/gatos por setor censitário.
- $\overline{\overline{X}} = \frac{X}{N}$ = número médio de cães/gatos por domicílio.

Notação III

- $\sigma^2_{1x} = \frac{\sum_{i=1}^{M}(X_i - \overline{X})^2}{M}$ = variância do número de cães/gatos por setor censitário.
- $\sigma^2_{2x} = \left(\frac{1}{N}\right)\sum_{i=1}^{M}\left(\frac{N_i}{N_i - 1}\right)\sum_{j=1}^{N_i}(X_{ij} - \overline{\overline{X_i}})^2$ = variância do número de cães/gatos por domicílio.
- $1 - \alpha$ = nível de confiança.
- ϵ = erro máximo da estimativa de cães/gatos.

Estimativa da população de cães/gatos I

Considere que no município existam M setores censitários dos quais m serão selecionados aleatoriamente para fazer parte da amostra com probabilidade proporcional ao número de domicílios existentes. A seguir, em cada setor censitário sorteado serão examinados $\overline{n}$ domicílios nos quais o número de cães/gatos será contado.
O número de cães/gatos no domicílio será obtido utilizando-se a equação 8.

$$x'_{ppt} = \frac{N}{n} \sum_{i=1}^{m} x_i \tag{8}$$

O intervalo de confiança de 95% para a X será dado por:

$$x'_{ppt} - 1.96 \times \widehat{EP} \leq X \leq x'_{ppt} + 1.96 \times \widehat{EP} \tag{9}$$

Estimativa da população de cães/gatos II

onde $\widehat{EP}$ é dado por:

$$\widehat{EP} = \sqrt{\frac{\sum_{i=1}^{m} \left(\frac{Nx_i}{\overline{n}} - x'_{ppt} \right)^2}{m(m-1)}} \tag{10}$$

Quantos domicílios devem ser examinados em cada setor censitário? I

Vamos considerar que cada setor será selecionado de forma simples aleatória e que em cada um número fixo $\overline{n}$ dos domicílios serão examinados.
O cálculo do número médio de domicílios a serem examinados em cada setor de modo a se obter o menor desvio padrão considerando-se os custos de visita aos setores (C_1) e de realização de entrevista em cada setor (C_2) é dado por:

$$\overline{n} = \left[\left(\frac{C_1}{C_2} \right) \left(\frac{1 - \delta_x}{\delta_x} \right) \right]^{1/2} \quad (11)$$

Quantos domicílios devem ser examinados em cada setor censitário? II

onde δ_x é uma generalização do coeficiente de correlação intraclasse e é dado por:

$$\delta_x = \frac{[M/(M-1)]\sigma_{1x}^2 - \overline{N}\sigma_{2x}^2}{[M/(M-1)]\sigma_{1x}^2 + \overline{N}(\overline{N}-1)\sigma_{2x}^2} \qquad (12)$$

O problema agora está em determinar os valores de σ_{1x}^2 e σ_{2x}^2. A variância do número de cães/gatos por setor censitário (σ_{1x}^2) e a variância do número de cães/gatos por domicilio podem ser obtidas a partir de uma amostra piloto ou, no impossibilidade de sua realização, utilizando-se, como aproximação, as variâncias encontradas no município de São Paulo.

Quantos setores devemos selecionar?

Uma vez determinada a fração de domicílios a serem amostrados em cada setor, devemos estabelecer o número de setores a serem selecionados de modo que a diferença entre nossa estimativa e o valor real não seja superior a $(100 \times \epsilon)\%$ com uma confiança de $100 \times (1 - \alpha)\%$. Esse valor de m é dado por:

$$m = \frac{z^2_{1-\alpha/2} \sum_{i=1}^{m^*} \left(\frac{N x_i}{\overline{n}} - x'_{ppt} \right)^2}{\epsilon^2 (x'_{ppt})^2 (m^* - 1)} \tag{13}$$

onde, para uma confiança de 95% temos $z_{1-0.025} = 1.96$ e os valores de m^*, $\overline{n}$, x_i e x'_{ppt} são obtidos a partir de um estudo piloto.

Amostra para detecção de doença

$$n \simeq \frac{(1-(1-\alpha)^{1/D})(N-\frac{(SeD-1)}{2})}{Se} \tag{14}$$

Onde:

- α: probabilidade de encontrar pelo menos 1 animal doente
- D: número de doentes
- N: tamanho da população
- Se: sensibilidade da prova

Exemplo

(Martin et al.) Suponha que, em uma população de 1000 suínos, há pelo menos 10 animais com rinite atrófica, se ela estiver presente. O tamanho da amostra necessário para, em 95% dos casos, detectar pelo menos um suíno com rinite considerando a sensibilidade da prova como sendo igual a 90% é:

$$
\begin{aligned}
n &\simeq \frac{(1-(1-\alpha)^{1/D})(N-\frac{(SeD-1)}{2})}{Se} \\
n &\simeq \frac{(1-(1-0.95)^{1/10})(1000-\frac{(0.90\times 10-1)}{2})}{0.90} \\
n &\simeq 287
\end{aligned}
$$

Tabelas de tamanho de amostra

Sample size determination in health studies: a practical manual.
Lwanga, Stephen Kaggwa; Lemeshow, Stanley.
Geneva; World Health Organization; 1991. 80 p.
Disponível em:
`http://whqlibdoc.who.int/publications/9241544058_(p1-p22).pdf`
`http://whqlibdoc.who.int/publications/9241544058_(p23-p80).pdf`

Referências

- Paul S. Levy and Stanley Lemeshow, *Sampling of populations: methods and applications*, third edition, Wiley (1999).
- W.G. Cochran, *Sampling techniques*, third edition, Wiley (1977).
- Nilza Nunes da Silva, *Amostragem probabilística*, primeira edição, EDUSP (1998).
- Sonia Vieira, *Bioestatística*, terceira edição, Campus (1980).