Capítulo 5 Estimação de proporções

5.1 Proporção populacional

Um caso especial de variável de interesse para muitos estudos ou pesquisas é a variável y cujo valor indica se uma determinada unidade populacional tem ou não uma determinada característica ou atributo ou pertencem a um determinado grupo dentro da população. São exemplos desse tipo as investigações sobre:

- migrantes entre os habitantes de determinada região;
- estabelecimentos agropecuários que se dedicam exclusivamente à produção leiteira numa determinada localidade;
- estudantes do sexo feminino em escolas;
- sondagens eleitorais, onde se deseja conhecer qual parcela dos eleitores pretende votar em determinado candidato.

Sendo uma variável indicadora, a variável y irá assumir para cada unidade da população apenas o valor 1, se a unidade possui o atributo pesquisado, ou 0, caso a unidade não possua o atributo.

Então, para cada unidade i da população, a variável y será definida como:

$$y_i = \begin{cases} 1, \text{ se a unidade i possui o atributo} \\ 0, \text{ em outro caso} \end{cases}$$

Dessa forma, o total populacional da variável y será dado pela fórmula:

$$Y = \sum_{i=1}^{N} y_i$$

expressando o número de unidades populacionais que possuem a característica de interesse.

A média populacional de y definida como:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{Y}{N}$$

será o número de unidades da população que tem a característica de interesse dividido pelo número total de unidades da população.

A esse valor dá-se o nome de *Proporção* de unidades na população possuidoras da caracterírtica em estudo.Uma proporção pode assumir valores de 0, quando nenhuma unidade da população tem o atributo investigado, até 1, quando todas as unidades possuem esse atributo. Muitas vezes é interessante expressar a proporção sob forma de porcentagem variando, então, de 0% até 100%.

O parâmetro populacional Proporção será aqui representado pela letra p minúscula, já que a letra P maiúscula já foi escolhida para denotar Probabilidade.

Como y só pode receber valores 1 ou 0 a expressão da sua variância populacional pode ser simplificada:

$$S_y^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - NY^{-2}\right) = \frac{1}{N-1}\left(Np - Np^2\right) = \frac{N}{N-1}p(1-p) = \frac{N}{N-1}pq$$

onde a letra q denota a proporção de unidades da população que *não* possuem a caracteristica em estudo.

A variância populacional de y pode também ser definida como $\sigma_y^2 = pq$. Tanto S_y^2 como σ_y^2 representam a dispersão da distribuição dos valosre de y na população. De acordo com a situação uma ou outra definição pode ser utilizada para simplificar a manitpulação algébrica de alguns resultados.

Para populações com um grande número de unidades, $N \to \infty,$ pode-se considerar $S_y^2 \approx pq = \sigma_y^2$.

Outra medida importante para avaliar a dispersão de uma variável é o seu *Coeficiente de variação* definido como a divisão do *Desvio Padrão* de y por sua média:

$$CV_y = \frac{\sqrt{\sigma_y^2}}{\bar{Y}} = \sqrt{\sigma_y^2 / Y^{-2}} = \sqrt{q/p}$$

Exemplo 5.1: Seja uma escola de ensino fundamental onde deseja-se estudar a composição dos estudantes por sexo. Vamos supor que a escola tenha um total de 1000 estudantes, dos quais 480 são do sexo feminino. Pode-se definir a variável y de interesse como:

$$\begin{cases} 1, \text{ se o estudante for do sexo feminino} \\ \end{cases}$$

O total de meninas da escola será o total da variável y, dado por :

$$Y = \sum_{i=1}^{1000} y_i = 1 + 1 + 0 + 1 + \ldots + 0 + 1 + 1 = 480$$

O *valor esperado* ou *média* da variável y, que neste caso é chamado de *proporção# será:

$$\bar{Y} = p = \frac{1}{N} \sum_{i=1}^{1000} y_i = \frac{Y}{N} = \frac{480}{1000} = 0,48 \text{ ou } 48\%$$

A *variância* da variável y, medida por S_y^2 será:

$$S_y^2 = \frac{1}{N-1} \sum_{i=1}^{1000} (y_i - N\bar{Y})^2 = \frac{1}{999} \left(\sum_{i=1}^{1000} y_i^2 - 1000\bar{Y}^{-2} \right) S_y^2 = \frac{1}{999} \left(1^2 + 1^2 + 0^2 + 1^2 + \ldots + 0^2 +^2 + 1^2 \right.$$

Finalmente, o *coeficiente de variação* de y será:

$$CV_y = \frac{\sqrt{\sigma_y^2}}{\bar{Y}} = \sqrt{q/p} = \sqrt{\frac{0,52}{0,48}} \doteq 1,041$$

Nos ítens que se seguem trata-se do problema da estimação desses parâmetros populacionais a partir da seleção de uma amostra aleatória simples, com ou sem reposição.

5.2 Estimação de uma proporção sob Amostragem Aleatória Simples com reposição - *AASC*

Seja uma *AASC*, s, de tamanho n selecionada de uma população composta de N unidades, onde se observa uma variável y definida como no item anterior. Pode-se definir estimadores para os parâmetros populacionais de y, fazendo analogia com o que foi visto no capítulo anterior, da seguinte forma:

O total de unidades da amostra com a característica de interesse será dado pelo total amostral, ou seja:

$$t_y = \sum_{i \in s} y_i$$

Da mesma forma a proporção amostral de unidades que possuem a característica em estudo será dada pela média amostral:

$$\hat{p} = \bar{y} = \frac{1}{n} \sum_{i \in s} y_i = \frac{t_y}{n}$$

Pode-se facilmente verificar que $\hat{p}$ é um estimador não viciado para a proporção populacional p, pois:

$$E_{AASC}(\hat{p}) = E_{AASC}(\bar{y}) = \bar{Y} = p$$

A variância da proporção amostral pode ser calculada, também, analogamente a variância da média amostral.

$$V_{AASC}(\hat{p}) = \frac{\sigma_y^2}{n} = \frac{pq}{n}$$

A variância amostral de y será dada por:

$$s_y^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}) = \frac{n}{n-1}\hat{p}\hat{q}$$

Utilizando s_y^2 como um estimador não viciado para a variância populacional, σ_y^2, obtem-se um estimador para a variância do estimador $\hat{p}$:

$$v_{AASC}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}$$

5.3 Distribuição amostral de y sob *AASC*

Pode-se chegar aos mesmos resultados do item anterior analisando a distribuição amostral da variável indicadora y no caso de uma seleção aleatória simples sem reposição.

Foi definido que a variável y assume apenas valores iguais 1 ou 0, no caso da existência ou não do atributo de interesse em cada uma das unidades da população, portanto para cada $y_i \in s$ tem-se que:

$$y_i = \begin{cases} 1, & \text{se a unidade i tem o atributo} \\ 0, & \text{em outro caso} \end{cases}$$

Como as unidades amostrais são selecionadas com igual probabilidade e com reposição, as variáveis $y_i \in s$ são independentes e identicamente distribuídas com probabilidades definidas por:

$$P(y_i = 1) = P(y_i \text{ ter o atributo de interesse}) = \frac{Y}{N} = pP(y_i = 0) = P(y_i \text{ não ter o atributo de interesse}) = 1 -$$

Dessa maneira fica configurada uma distribuição de $Bernoulli(p)$:

$$\begin{array}{cc} k & 1 \quad 0 \\ P(y_i = k) & p \quad q \end{array}$$

O total amostral t_y que, neste caso, representa o número de unidades na amostra com a característica de interesse, será dado pelo soma de n variáveis aleatórias com distribuição $Bernoulli(p)$, portanto a variável aleatória t_y segue uma distribuição $Binomial(n, p)$.

Imediatamente tem-se que:

$$E_{AASC}(t_y) = np \quad \text{e} \quad V_{AASC}(t_y) = npq$$

E seguindo o mesmo raciocínio pode-se ter o valor esperado e a variância de $\hat{p}$:

$$E_{AASC}(\hat{p}) = E_{AASC}\left(\frac{t_y}{n}\right) = p \quad \text{e} \quad V_{AASC}(\hat{p}) = V_{AASC}\left(\frac{t_y}{n}\right) = \frac{pq}{n}$$

Outro resultado importante é que se tem a distribuição de probabilidades de $\hat{p}$, pois:

$$P\left(\hat{p} = \frac{k}{n}\right) = P(t_y = k) = \binom{n}{k} p^k q^{n-k}, \qquad \text{para todo } k = 1, 2, \ldots, n$$

5.4 Estimação de uma proporção sob Amostragem Aleatória Simples sem reposição - *AAS*

No caso de uma amostra s com seleção do tipo *AAS*, as expressões do total amostral, t_y, da variância amostral, s_y^2, tem a mesma forma apresentada para *AASC*.

Uma diferença fundamental é que no caso da *AAS* a variância amostral , s_y^2, é um estimador não viciado para S^2y e não para σ_y^2.

Foi visto no capítulo anterior que a variância do estimador da média é calculado pela expressão:

$$V_{AAS}(\bar{y}) = \frac{N-n}{N}\frac{S_y^2}{n}$$

Então, no caso do estimador de uma proporção, por analogia, tem-se que:

$$V_{AAS}(\hat{p}) = \frac{N-n}{N-1}\frac{pq}{n}$$

Utilizando s_y^2 como estimador não viciado para S_y^2 chega-se ao estimador para a variância de $\hat{p}$:

$$v_{AAS}(\hat{p}) = \frac{N-n}{N}\frac{\hat{p}\hat{q}}{n-1}$$

Nota-se que para populações onde o número de unidades, N, é suficientemente grande tem-se que $v_{AAS}(\hat{p}) \approx \frac{\hat{p}\hat{q}}{n-1}$, resultando numa equivalência entre os desempenhos da *AAS* e da *AASC*. Intuitivamente isso ocorre porque a probabilidade de seleção repetida tende a ser muito pequena no caso de populações muito grandes.

5.5 Distribuição amostral de y sob *AAS*

Da mesma forma que na *AASC* quando a seleção da amostra aleatória simples é feita sem reposição, também para cada $y_i \in s$ tem-se que:

$$y_i = \begin{cases} 1, & \text{se a unidade i tem o atributo} \\ 0, & \text{em outro caso} \end{cases}$$

A diferença é que na *AAS*, em uma dada seleção, a probabilidade de ser selecionada uma unidade que já foi sorteada anteriormente é nula. Logo as *n* variáveis aleatórias $y_i \in s$ não são independentes, porém são identicamente distribuídas e, também neste caso $P(y_i = 1) = p$ e $P(y_i = 0) = q$.

O número total de amostras aleatórias simples sem reposição de tamanho n que podem ser selecionadas de uma população com N unidades é dado por $\binom{N}{n}$ e o número de amostras com exatamente k unidades com a característica em estudo e $n-k$ unidades sem essa

característica pode se calculado por $\binom{Y}{k}\binom{N-Y}{n-k}$, onde Y é o número de unidades da população com a característica de interesse.

Utilizando o fato de que a probabilidade de um determinado evento ocorrer é obtido pela divisão do número de casos favoráveis pelo número total de casos possíveis, pode-se verificar que a distribuição da estatística t_y, que é o número de unidadaes da amostra com a característica de interesse, segue uma distribuição Hipergeométrica(N,Y,n), representada pela expressão:

\begin{equation} P(t_y=k)=\frac {\binom{Y}{k} \binom{N-Y} {n-k}} {\binom{N}{n}}, {\quad\quad\quad\quad\quad\quad0\le k \le min (n,Y)}\tag{5.21} \end{equation}
A média ou valor esperado de uma distribuição Hipergeométrica(N,Y,n) é dado por n\frac Y N, portanto tem-se imediatamente que:

\begin{equation} E_{AAS}(t_y)=n\frac Y N =np\implies E_{AAS}(\widehat p)=p\tag{5.22} \end{equation}
A variância da mesma distribuição hipergeométrica é calculada por n\frac Y N\frac{N-Y}{N}\left(1-\frac{n-1}{N-1}\right)=npq\left(\frac{N-n}{N-1}\right), levando à variância do total de unidades na amostra que tem o atributo de interesse:

\begin{equation} V_{AAS}(t_y)=\left(\frac{N-n}{N-1}\right)npq\tag{5.23} \end{equation}
Consequentemente a variância do estimador da proporção p será dada por:

\begin{equation} V_{AAS}(\widehat p)=V_{AAS}(\frac{t_y}{n}) = \left(\frac{N-n}{N-1}\right)\frac {pq}{n}\tag{5.24} \end{equation}
Um estimador da variância de \widehat p será obtido substituindo S^2_y pelo seu estimador não viciado que é a variância amostral, resultando em:

\begin{equation} v_{AAS}(\widehat p)= \left(\frac{N-n}{N}\right)\frac {\widehat p\widehat q}{n-1}\tag{5.25} \end{equation}

5.6 Intervalos de confiança para proporções na Amostragem Aleatória Simples

Foi visto que na Amostragem Aleatória Simples, tanto com ou sem reposição, são conhecidas as distribuições exatas para o estimador \widehat p, sendo, portanto, possívem obter os limites inferior e superior para um intervalo de confiança para a proporção p, com um nível de significância \alpha fixado.

Para isso, no caso de *AASC*, é necessário resolver o siatema de equações seguinte determinando os valores de \widehat p_I e \widehat p_S que o satisfaçam:

\begin{equation} \begin{cases}\displaystyle\sum _{k=0}^{t_y} \binom {n}{k} \widehat p_S^k (1 - \widehat p_S)^{n-k}=\frac{\alpha} {2} \\\\ \displaystyle\sum _{k=t_y}^n \binom {n}{k} \widehat p_I^k (1 - \widehat p_I)^{n-k}=\frac{\alpha} {2}\end{cases}\tag{5.26} \end{equation}
No caso da *AAS* o sistema a ser resolvido é baseado na distribuição Hipergeométrica como se segue:

\begin{equation} \begin{cases}\displaystyle\sum_ {k=0}^{t_y}\frac {\binom{N\widehat p_S}{k} \binom{N-N\widehat p_S} {n-k}}{\binom{N}{n}}=\frac {\alpha}{2} \\\\ \displaystyle \sum_ {k=t_y}^{n}\frac {\binom{N\widehat p_I}{k} \binom{N-N\widehat p_I} {n-k}} {\binom{N}{n}}={\frac {\alpha}{2}}\end{cases}\tag{5.27} \end{equation}
Em ambos os casos \alpha é o nível de significância desejado, correspondendo ao nível de confiança 1-\alpha.

A solução desses sistemas pode ser muito trabalhosa, exigindo aplicação de métodos iterativos que exigem quantidade razoávelmente grande de recursos computacionais. (Cochran 1977) dá alguns exemplos e referências para a solução desses problemas.Na maioria dos casos práticos essa dificuldade leva a opção pelo uso da aproximação pela distribuição Normal de probabilidades. Isso pode ser feito sempre que as condições do problema assim o permitirem.

5.7 Intervalos de confiança utilizando a Aproximação Normal

Como já foi visto no capítulo anterior, a distribuição do estimador da proporção, \widehat p, pode ser aproximada pela distribuição Normal de probabilidade. Esta aproximação pode ser utilizada mesmo no caso da *AAS* onde os y_i observados na amostra não são independentes, desde que se tenha valores de N e n suficientemente grandes e valor da fração amostral, f=\frac n N, pequeno.

Sob estas condições pode-se considerar que:

\begin{equation} \frac {\widehat p - p} {\sqrt {V_{p(s)}(\widehat p)}} \simeq N(0,1)\tag{5.28} \end{equation} Na Figura 5.1 a seguir o gráfico mostra a aproximação Normal para a seleção de 1000 amostras de tamanho n=100, selecionadas com reposição, de uma população de tamanho N=5000, onde exatamente metade das unidades tem a caracterísitica de interesse

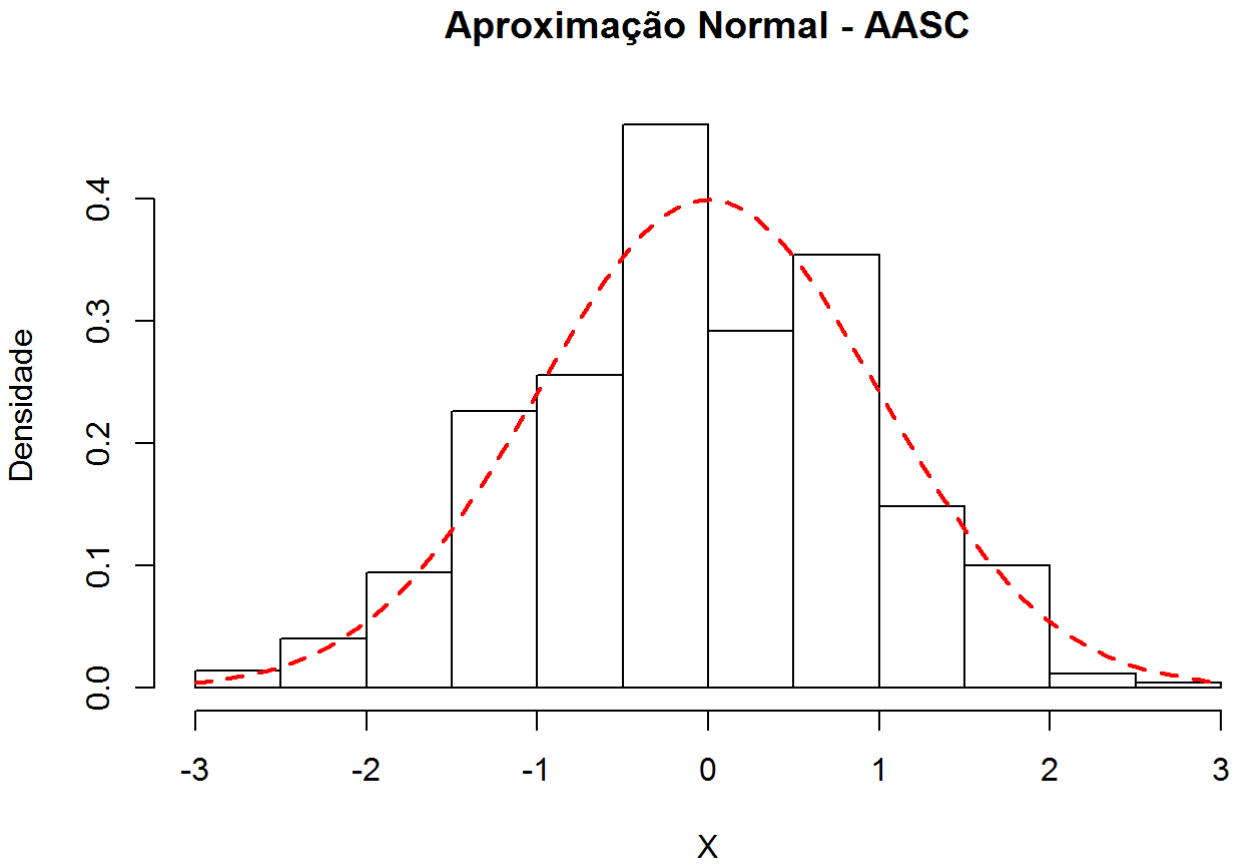

Figura 5.1: Aproximação Normal no caso de uma *AASC*

A Figura 5.2 mostra o gráfico para um exercício similar ao anterior onde as amostras foram selecionadas sem reposição.

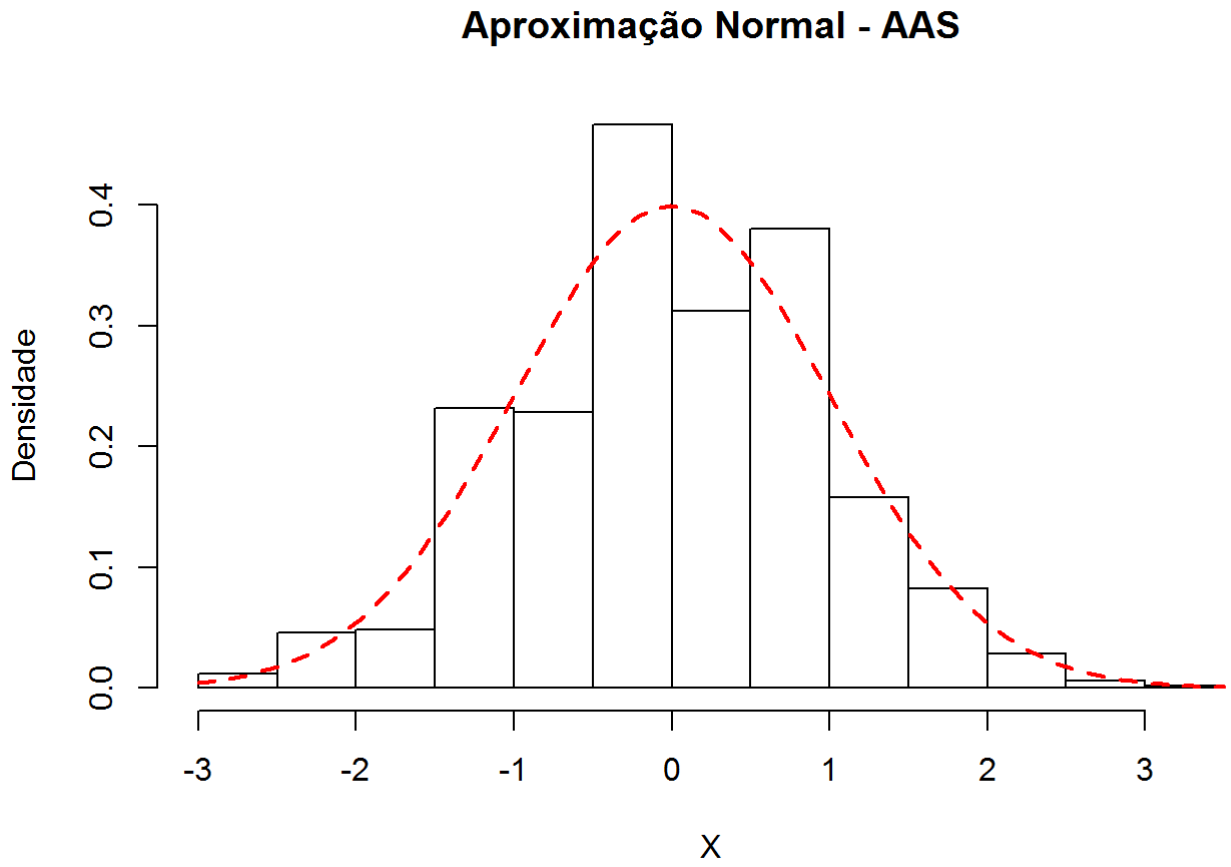

Figura 5.2: Aproximação Normal no caso de uma *AAS*

Tanto nos casos de seleção com ou sem reposição pode-se considerar que as aproximações são satisfatórias, com os histogramas das distribuições amostrais do estimador \widehat p aderindo consideravelmente à curva Normal padrão.

(Cochran 1977) mostra uma tabela, reproduzida na Tabela 5.1, com oalguns valores mínimos do total de unidades observadas na amostra, t_y, onde a aproximação Normal pode se utilizada.

Tabela 5.1 - Valores mínimos de t_y para uso da aproximação Normal

$$\begin{array}{r r r} \hline P & t_y & n \\ \hline 0,5 & 15 & 30 \\ 0,4 & 20 & 50 \\ 0,3 & 24 & 80 \\ 0,2 & 40 & 200 \\ 0,1 & 60 & 600 \\ 0,05 & 70 & 1400 \\ \approx 0 & 80 & \infty \\ \hline \end{array}$$

Essa tabela foi construída considerando um nível de significância de \alpha=0,05, que é um valor comumente utilizado em muitas situações práticas.Tem-se a partir daí critérios práticos para assumir a utilização da aproximação Normal, notando-se que o tamanho mínimo da amostra requerido é de n=30.

Nas condições estabelecidas para a validade da aproximação Normal tem-se que S^2_y \approx \sigma^2_y=pq, portanto, \displaystyle V_{AAS}(\widehat p) \approx V_{AASC} (\widehat p), então para os dois tipos de seleção pode-se considerar o intervalo de confiança

para a proporção como: \begin{equation} \displaystyle IC_{1-\alpha}(p):\left[\widehat p-z_{\alpha/ 2}\sqrt { {pq}/n};\widehat p+z_{\alpha/2}\sqrt {{pq}/n} \right]\tag{5.29} \end{equation} Caso se deseje considerar o fator de correção para populacões finitas, quando a fração amostral não possa ser considerada pequena e a seleção for sem reposição, a expressão do intervalo de confiança passa a ser: \begin{equation} \displaystyle IC_{1-\alpha}(p):\left[\widehat p-z_{\alpha/ 2}\sqrt { \left(\frac{N-n}{N-1} \right)\frac{pq}{n}};\widehat p+z_{\alpha/2}\sqrt { \left(\frac{N-n}{N-1} \right)\frac{pq}{n}} \right]\tag{5.30} \end{equation} Em (Cochran 1977) é apresentada uma *correção de continuidade* acrescentando a fração 1/2n à margem de erro do intervalo de confiança pelo fato de se fazer uma aproximação de uma distribuição discreta (Binomial ou Hipergeométrica) pela distribuição Normal, que é contínua. Desse modo a expressão do intervalo de confiança passa a ser: \begin{equation} \displaystyle IC_{1-\alpha}(p):\left[\widehat p-z_{\alpha/ 2}\sqrt {{pq}/n}-1/2n;\widehat p+z_{\alpha/2}\sqrt{{pq}/n} +1/2n \right]\tag{5.31} \end{equation} Ou considerando a coreção para população finita: \begin{equation} \displaystyle IC_{1-\alpha}(p):\left[\widehat p-z_{\alpha/2}\sqrt { \left(\frac {N-n}{N-1} \right)\frac{pq}{n}}-\frac 1 {2n};\widehat p+z_{\alpha/2}\sqrt { \left(\frac {N-n}{N-1} \right)\frac{pq}{n}}+\frac 1{2n} \right]\tag{5.32} \end{equation} Nas aplicações práticas o valor da variância do estimador da proporção p, geralmente, não é conhecido. Assim o que se pode fazer é estimar um intervalo de confiança, substituindo S^2_y por s^2_y na expressões anteriores: \begin{equation} \displaystyle ic_{1-\alpha}(p):\left[\widehat p-z_{\alpha/ 2}\sqrt {\frac{\widehat p \widehat q} {n-1}}-\frac 1{2n};\widehat p+z_{\alpha/2}\sqrt{\frac{\widehat p \widehat q} {n-1}} +\frac 1{2n} \right]\tag{5.33} \end{equation} \begin{equation} \displaystyle ic_{1-\alpha}(p):\left[\widehat p-z_{\alpha/ 2}\sqrt {\left(\frac{N-n}{N}\right)\frac {\widehat p\widehat q}{n-1}}-\frac 1{2n};\widehat p+z_{\alpha/2}\sqrt {\left(\frac{N-n}{N}\right)\frac {\widehat p\widehat q}{n-1}} +\frac 1{2n} \right]\tag{5.34} \end{equation}

Veja que o efeito da correção de continuidade tende rapidamente a ser nulo quando o tamanho da amostra, n, cresce. Para uma amostra de tamanho n=50 esse fator já é de apenas 1%, o que pode ser desprezível dependendo da proporção que estiver sendo estimada, porém é preciso muito cuidado pois quando se está trabalhando com proporções são valores, às vezes, bastante pequenos.

5.8 Cálculo do tamanho da amostra

O tamanho de uma amostra aleatória simples a ser selecionada, como já foi visto no capítulo anterior, será calculado a partir da definição do erro amostral ou margem de erro admissível para o caso, do nível de confiança desejado e se a seleção for com ou sem reposição.

No caso de seleção com reposição, considerando uma margem de erro máxima admissível de d com um nível de confiança 1-\alpha, basta utilizar expressão da margem de erro: \begin{equation} d \le z_{\alpha/2}\sqrt { \frac{pq}{n}} \implies n \ge \frac{z^2_{\alpha/2}pq} {d^2} \tag{5.35} \end{equation} Para seleção sem reposição o tamanho da amostra será calculado como: \begin{equation} d \le z_{\alpha/2}\sqrt{\left(\frac {N-n}{N-1}\right)\frac{pq}n} \implies n \ge \frac{z^2_{\alpha/2}pq} {d^2 \frac {N-1}N + \frac 1 N {z^2_{\alpha /2}pq}}\approx \frac{z^2_{\alpha/2}pq} {d^2 + \frac 1 N {z^2_{\alpha /2}pq}} \tag{5.36} \end{equation} Uma maneira prática de calcular o tamanho da amostra para uma *AAS* em dois passos é calcular primeiro: \begin{equation} n_0=\frac{z^2_{\alpha/2}pq}{d^2}\tag{5.37} \end{equation} E depois fazer: \begin{equation} n \ge \frac {n_0}{1+n_0 /N}\tag{5.38} \end{equation}
Note que n_0 é equivalente ao tamanho da amostra para uma *AASC* e o valor de n para a *AAS* é obtido pela correção para população finita do valor n_0. Também pode-se concluir que quando o tamanho da população, N, é grande o fator n_0/N tende a se anular fazendo com que n=n_0, ou seja: quando o tamanho da população é grande as amostras aleatórias simples com ou sem reposição são equivalentes.

AS fórmulas apresentadas dependem do nivel de significância \alpha e da margem de erro d que serão definidos pelo pesquisador de acordo com seu conhecimento relativo ao assunto pesquisado, pois esses valores estão diretamente ligados à natureza da pesquisa. Pesquisas que utilizam medidas objetivas para alcançar seus resultados, como instrumentos para se medir fisicamente o fenômeno estudado, podem ser mais exigentes quanto a presisão das estimativas desejadas, enquanto que pesquisas da área social, por exemplo, onde se utilizam questionários que dependem da memória ou até da boa vontade dos entrvistados, frequentemente não podem ter o mesmo nível de exigência. O tamanho da amostra também depende da variância da variável utilizada para seo cálculo, através do produto pq. Como p é a proporção que se deseja estimar, se fosse conhacida não haveria a necessidade de uma amostra. Geralmente, como no caso de se pesquisar variáveis continuas, utiliza-se pesquisas anteriores ou sobre variáveis correlacionadas com a atual variável de interesse, ou mesmo uma pesquisa piloto com um tamanho arbitrário de amostra para se ter uma estimativa inicial do fenomeno a ser medido e poder calcular o tamanho de amostra realmente necessário. Quando se utiliza uma pesquisa piloto existem métodos para utilizar os resultados relativos às unidades já pesquisadas e selecionar outras unidades para complementar o tamanho da amostra.

No caso da estimação de proporções o valor de pq é limitado variando de 0 a 0,25, sendo esse valor máximo atingido quando p=q=0,5. O gráfigo da Figura 5.3 mostra como variam os valores de pq conforme variam os valore de p.

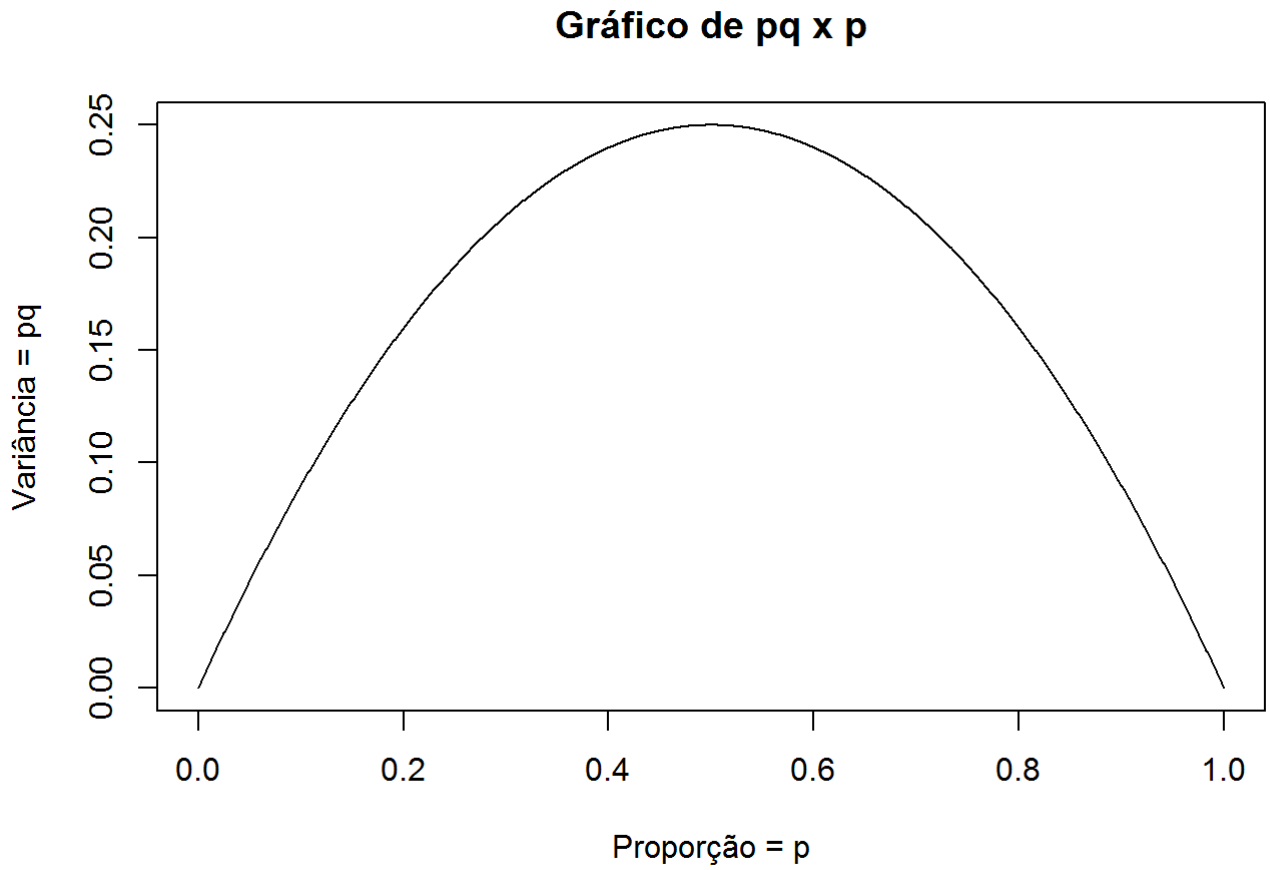

Figura 5.3: Variação de pq em função dos valores de p

Como o valor máximo da variância é atingido quando p=0,5, caso não exista nenhuma informação sobre a proporção a ser estimada, uma maneira de determinar um tamanho de amostra conservador é supor exatamente que p=0,5. Assim pode-se simplificar a fórmula de cálculo de n para uma *AASC*: \begin{equation} n \ge \frac {z^2_{\alpha /2}}{4d^2}\tag{5.39} \end{equation} No caso de uma *AAS* basta fazer: \begin{equation} n_0 = \frac {z^2_{\alpha /2}}{4d^2} \implies n \ge \frac {n_0}{1+n_0/N}\tag{5.40} \end{equation}
Geralmente os resultados das fórmulas para cálculo do tamanho da amostra não são valores inteiros. Em todos esses casos o valor de n calculado deverá ser arredondado para o valor inteiro imediatamente superior, preservando assim a precisão desejada.

5.8.1 Cálculo do n utilizando outras formas de representar o erro amostral

As fómulas apresentadas para o cálculo do tamanho da amostra utilizaram a margem de erro d como parâmetro de entrada, porém o erro amostral pode ser representado de outras maneiras. Pode-se defini-lo como o *coeficiente de cariação*, como a *variância* ou como o *erro relativo* do estimador a ser calculado.

Para calcular um tamanho de amostra de maneira que o coeficiente de variação máximo esperado para o estimador \widehat p seja um valor fixado c, pode-se utilizar a fórmula: \begin{equation} n \ge \frac q {c^2p}\tag{5.41} \end{equation}
no caso da seleção com reposição.

Para chegar a esse resultado basta ver que CV_{AASC}(\widehat p)= \sqrt {V_{AASC}(\widehat p)}/p e substituir na fórmula apresentada para o cálculo de n a partir da margem de erro fixada.

Para a seleção sem reposição pode-se fazer: \begin{equation} n_0 = \frac q {c^2p} \implies n \ge \frac {n_0} {1+n_0/N}\tag{5.42} \end{equation}
Seguindo o mesmo raciocínio pode-se chegar às formulas para calcular *n* fixando a variância máxima esperada de \widehat p em v ou o seu erro relativo máximo em r.

Para a seleção com reposição tem-se: \begin{equation} n \ge \frac {pq} {v} { \quad\textrm { ou } \quad} n \ge \frac {z^2_{\alpha/2}}{r^2} \frac q p \tag{5.43} \end{equation}
As expressões para a seleção sem reposição são derivadas como no caso em que foi fixado o valor máximo esperado para CV(\widehat p).

5.9 Estimação de proporções para variáveis com mais duas categorias

Até o momento foi tratado o caso em que temos uma variável indicadora com duas categorias, definindo se uma determinada unidade na população (ou na amostral) tem ou não determinada característica de interesse. Muitas vezes temos a necessidade de definir mais de duas categorias como, por exemplo:

- estudar a distribuição por faixas etárias de uma localidade ou grupo de pessoas;
- estudar a classificação econômica das empresas de determinado país;
- estimar a intenção de votos dos candidatos em uma eleição com mais de 2 candidatos, além das possibilidades de voto em branco ou nulo ou, ainda, ou eleitores indecisos.

Nestes casos há interesse de estimar a proporção de unidades em cada uma das possíveis categorias e respectiva precisão dessas estimativas.

Exemplo : Seja uma escola com 1000 alunos distribuídos entre as 9 etapas do ensino fundamental como na tabela seguinte: \begin{array}{l|r r r r r r r r r} \hline \text{Etapa de ensino} & \text{1° ano} & \text{2° ano} & \text{3° ano} &\text{4° ano} &\text{5° ano} &\text{6° ano} & \text{7° ano} & \text{8° ano} & \text{9° ano} & \text{Total}\\ \hline \text {Alunos} & 110 &

108 & 110 & 115 & 104 & 119 & 116 & 107 & 111 & 1000\\ \hline \text {Proporção} & 0,110 & 0,108& 0,110 & 0,115 & 0,104 & 0,119 & 0,116 & 0,107 & 0,111 & 1,000\\ \hline \end{array}
Observe que para calcular as proporções em cada uma das categorias na verdade o que se faz é atribuir o valor 1 às unidades da categoria em questão e o valor 0 para as unidades pertencente às demais categorias. Em outras palavras, se a variável tem M categorias é como se fossem M problemas com duas categorias.

A proporção de unidades da população pertencentes à categoria c\in (1, 2,.., M), é dada por:

\begin{equation} p_c=\frac {N_c} {N} \tag{5.44} \end{equation}
0nde N_c é o número de unidades na categoria c e N é o tamanho total da população.

Seja uma amostra aleatória simples (com ou sem reposição) de tamanho n e seja a variável indicadora y_i definida como: \begin{equation} \displaystyle y_i= \begin{cases} 1, \text {se a unidade i pertence à categoria c} \\\\ 0, \text {se a unidade i não pertence à categoria c} \end{cases}\tag{5.45} \end{equation} Com tal definição pode-se ver que o número de unidades da catecoria c na amostra será dado por:: \begin{equation} \displaystyle n_c=\sum_{i=1}^{n} y_i\tag{5.46} \end{equation} Um estimador para a proporção de unidades populacionais pertencentes à categoria c poderá ser obtido por: \begin{equation} \displaystyle \widehat p_c=\frac {1}{n} \sum_{i=1}^{n} y_i =\frac {n_c} {n}\tag{5.47} \end{equation}
O problema foi reduzido ao caso de estimar proporções em variáveis com duas categorias. Pode-se obter, também, estimativas de precisão utilizando as mesmas ferramentas já apresentadas neste capítulo.

Muitas vezes pode-se estar interessado em estimar proporções para agrupamentos das categorias originais.

Voltando ao exemplo da escola do ensino fundamental, pode ser de interesse estudar a proporção de seus alunos que estão matriculados no primeiro segmento do ensino fundamental. Nesse caso seriam contabilizados como pertencentes à categoria c de interesse todos os alunos do 1º até o 5º ano, para os quais y_c=1, sendo y_c=0 para os demais alunos da escola.

Outro caso de interesse ocorre quando na aplicação de um questionário, por exemplo, aparecem respondentes que se recusaram a responder ou,mesmo, disseram que não sabiam a resposta. Num caso como esse pode-se estar interessado em estimar a proporção das pessoas que responderam determinada alternativa, entre as pessoas que efetivamente responderam a pesquisa escolhendo uma das alternativas válidas. Um exemplo prático seria uma pesquisa sobre a intenção de voto numa eleição com apenas dois candidatos. Neste

caso o entrevistado poderia reponder que votará no canditato A, no candidato B, que votará nulo ou em branco, onde apenas as duas primeiras alternativas seriam considerados como votos válidos.

Neste caso poderia ser estimada a proporção para cada uma das quatro categorias iniciais ou apenas a proporção de votos válidos para cada um dos dois candidatos. \begin{equation} \displaystyle \widehat p_A=\frac {n_A} {n_A+n_B} \quad\text { e } \quad\widehat p_B=\frac {n_B} {n_A+n_B}\tag{5.48} \end{equation}
Vale notar que neste caso tanto o numerador como o denominador do estimador da proporçãO são variáveis aleatórias, pois a população (eleitores que efetivamente vão votar num dos candidatos) é desconhecida.

Referências

Cochran, William Gemmell. 1977. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.