Universidade Federal de Lavras

Departamento de Estatística

Prof. Daniel Furtado Ferreira

$9^{\underline{a}}$ Teoria da Estimação

1) Desejando saber a incidência de uma doença em uma região foram amostrados $n = 6$ animais. Destes $y = 2$ estavam doentes. Determinar os intervalos de confiança exato e aproximado (usando à aproximação normal) para a proporção de doença da região, considerando um coeficiente de confiança de 95%. Tirar todas as conclusões de interesse.

   Dados: $F_{0,025;\nu_1=10;\nu_2=4} = 8{,}8437$ e $F_{0,025;\nu_1=6;\nu_2=8} = 4{,}6517$; O intervalo de confiança exato é:

$$IC_{1-\alpha}(p):\left[\frac{1}{1+\dfrac{(n-y+1)F_{\alpha/2;\nu_1=2(n-y+1);\nu_2=2y}}{y}};\frac{1}{1+\dfrac{n-y}{(y+1)F_{\alpha/2;\nu_1=2(y+1);\nu_2=2(n-y)}}}\right]$$

2) Obter uma amostra piloto de $n^* = 10$ alunos e determinar o tamanho da amostra necessário para estimar a média de altura de alunos da Ufla com erro máximo de 0,05 m e 95% de confiança. Se desejarmos diminuir o erro para 0,01 m, qual será o novo tamanho de amostra necessário?

   A amostra piloto obtida é dada por:

   1,83 1,78 1,81 1,60 1,59 1,82 1,75 1,67 1,75, 1,87 cuja variância é igual a $S^2 = 0{,}009401$ m^2.

   Dados:

ν	$t_{0,025;\nu}$	ν	$t_{0,025;\nu}$
9	2,262157	18	2,101000
363	1,966521	480	1,964918
16	2,119905	362	1,966539

3) Determinar o tamanho da amostra necessário para estimar a proporção de animais doentes na região para a qual foi realizada a amostra do exercício 1 com erro máximo de 3% e 95% de confiança. Utilize os resultados do exercício 1 como amostra piloto. Também utilize a fórmula do n_{max} para dimensionar a amostra ($\hat{p} = 0{,}5$). Se o erro for diminuído para 2%, qual deve ser o novo tamanho de amostra? Qual a conclusão que você obtém confrontado os dois valores determinados de n?

   Dado $Z_{0,025} = 1{,}96$

4) O controle químico de uma doença será feito se a proporção de plantas doentes de uma região atingir 3%. Realizando uma amostra de tamanho $n = 500$ plantas, o pesquisador observou $y = 11$ plantas doentes. Se você fosse tomar a decisão com base na estimativa pontual, qual seria sua decisão? Se por outro lado você fizesse o intervalo de 95% de confiança, qual seria sua decisão? Justificar sua decisão realizando os cálculos apropriados utilizando a aproximação normal. Quais seriam suas respostas se em vez de observar $y = 11$ doentes, tivesse sido observado $y = 8$ doentes?

5) Um melhorista irá realizar o melhoramento em uma espécie vegetal se a variância desta população for igual a $\sigma^2 = 16$ (t/ha)^2. Uma amostra de $n = 50$ plantas resultou na seguinte estimativa da variância: $S^2 = 14$ (t/ha)^2. Baseado na estimativa pontual o melhorista não irá investir nesta população. Porém qual deverá ser sua decisão se for considerado o erro de estimação? Obter o intervalo de confiança de 95% para subsidiar sua resposta.

   Dados: $\chi^2_{0,025;\nu=49} = 70{,}222$ e $\chi^2_{0,975;\nu=49} = 31{,}555$.

REFERÊNCIAS BIBLIOGRÁFICAS ADICIONAIS

BLYTH, C.R. Approximate binomial confidence limits. *Journal of the American Statistical Association.* v.81, n.395, p.843-855, 1986.

PRATT, J.W. A normal approximation for binomial , F, beta, and other common, related tail probabilities, II. *Journal of the American Statistical Association.* n.63, p.1457-1483, 1968.

Resolução

1) O intervalo de 95% de confiança exato, sabendo que $y = 2$ e $n = 6$, é dado por:

$$IC_{1-\alpha}(p) : \left[\frac{1}{1+\dfrac{(n-y+1)F_{\alpha/2;\nu_1=2(n-y+1);\nu_2=2y}}{y}};\frac{1}{1+\dfrac{n-y}{(y+1)F_{\alpha/2;\nu_1=2(y+1);\nu_2=2(n-y)}}}\right]$$

$$IC_{0,95}(p) : \left[\frac{1}{1+\dfrac{(6-2+1)F_{0,025;\nu_1=2(6-2+1);\nu_2=2\times 2}}{2}};\frac{1}{1+\dfrac{6-2}{(2+1)F_{0,025;\nu_1=2(2+1);\nu_2=2(6-2)}}}\right]$$

$$IC_{0,95}(p) : \left[\frac{1}{1+\dfrac{5\times F_{0,025;\nu_1=10;\nu_2=4}}{2}};\frac{1}{1+\dfrac{4}{3\times F_{0,025;\nu_1=6;\nu_2=8}}}\right]$$

$$IC_{0,95}(p) : \left[\frac{1}{1+\dfrac{5\times 8{,}8437}{2}};\frac{1}{1+\dfrac{4}{3\times 4{,}6517}}\right]$$

$$IC_{0,95}(p) : [0{,}0433;\ 0{,}7772].$$

Portanto, podemos afirmar com pelo menos 95% de confiança que a verdadeira proporção de animais doentes na região deve estar entre 4,33% e 77,72%. Duas observações devem ser feitas. A primeira diz respeito ao fato de que o uso do intervalo de confiança exato não altera a precisão tornando o erro pequeno, se estivermos usando amostra pequenas. Isso é uma falácia e um erro de interpretação, pois não há mágica com métodos estatísticos. O intervalo exato só garantirá que a confiança seja de pelo menos (e não exatamente) 95%, mas não fará com que o erro de estimação (amplitude do intervalo) seja pequeno. A confiança será de pelo menos 95%, uma vez que trata-se de intervalos para parâmetros de distribuição de variáveis aleatórias discretas, no caso, binomial. Veja que a amplitude total do intervalo é muito grande, pois ninguém espera que uma amostra de apenas 6 animais seja suficiente para termos qualidade na inferência sobre o parâmetro desconhecido referente à proporção de animais doentes da região. Nenhum instituto de pequisa de opinião iria utilizar uma amostra de seis eleitores para fazer inferência a respeito da intenção de votos em um país, estado ou município. A segunda observação é que os erros para menos e para mais podem ser obtidos por $\hat{p} - LI$ e $LS - \hat{p}$, respectivamente. Neste caso, o erro para menos foi de 29 pontos percentuais e o erro para mais foi de 44,39 pontos percentuais. Em geral, o erro para menos e para mais são diferentes quando utilizamos o intervalo de confiança exato, pois se $p \neq 0{,}5$ a binomial é uma distribuição assimétrica.

O intervalo de 95% de confiança utilizando a aproximação normal é dado por:

$$IC_{\approx 1-\alpha}(p) : \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$IC_{\approx 1-\alpha}(p) : 0{,}3333 \pm 1{,}96\sqrt{\frac{0{,}3333\times 0{,}6667}{6}} = 0{,}3333 + 0{,}3772$$

$$IC_{\approx 1-\alpha}(p) : [-0{,}0439;\ 0{,}7105].$$

Logo, podemos afirmar com aproximadamente 95% de confiança que a verdadeira proporção de animais doentes da região deve estar entre 0% e 71,05%. Observe que o limite inferior do intervalo foi negativo, o que é inaceitável, e que o limite superior diferiu do exato em 6,67 pontos percentuais para menos. De uma maneira geral, o intervalo de confiança aproximado é muito deficiente se n é pequeno e p afasta-se de 1/2. Usamos para uma boa aproximação $np > 5$. Independentemente se utilizamos o intervalo de confiança aproximado ou exato, se amostra é pequena, o intervalo resultante é grande e não muito útil. Em grande amostras os intervalos de confiança exato e aproximado praticamente apresentam o mesmo resultado e por isso a aproximação normal é utilizada.

2) Com base na amostra piloto, apresentada no enunciado da questão, estimamos a média e a variância populacionais por $\bar{X} = 1{,}747$ m e $S^2 = 0{,}009401$ m^2, respectivamente. Para dimensionarmo a amostra, inicialmente determinamos o quantil superior 2,5% da distribuição t de *Student*, com $\nu = 10 - 1 = 9$ graus de liberdade. Este valor é $t_{0,025;\nu=9} = 2{,}262157$. Logo, o tamanho da amostra necessário, considerando um erro de 0,05 m, deve ser obtido de forma iterativa da seguinte forma:

$$n = \frac{S^2 t^2_{\alpha/2}}{e^2} = \frac{0{,}009401 \times 2{,}262157^2}{0{,}05^2} = 19{,}24 \approx 19.$$

Como $n \neq 10$, da amostra original, o processo deve continuar. Assim, o valor $t_{0{,}025;\nu=18} = 2{,}101$ e o novo n é:

$$n = \frac{S^2 t^2_{\alpha/2}}{e^2} = \frac{0{,}009401 \times 2{,}101^2}{0{,}05^2} = 16{,}59 \approx 17.$$

Novamente, $n = 17$ difere do valor anterior 19 e o processo é repetido. Assim, o valor $t_{0{,}025;\nu=16} = 2{,}119905$ e o novo n é:

$$n = \frac{S^2 t^2_{\alpha/2}}{e^2} = \frac{0{,}009401 \times 2{,}119905^2}{0{,}05^2} = 16{,}90 \approx 17.$$

Como o valor de n neste passo foi igual ao do passo anterior, o processo é encerrado. Logo precisamos de $n = 17$ alunos para estimar a média com erro para menos e para mais de 0,05 m e com 95% de confiança.

No caso específico de erro dado por $e = 0{,}01$ m, devemos iniciar o processo com $\nu = 9$ graus de liberdade, da amostra piloto. Assim,

$$n = \frac{S^2 t^2_{\alpha/2}}{e^2} = \frac{0{,}009401 \times 2{,}262157^2}{0{,}01^2} = 481{,}09 \approx 481.$$

Como $n \neq 10$, da amostra original, o processo deve continuar. Assim, o valor $t_{0{,}025;\nu=480} = 1{,}964918$ e o novo n é:

$$n = \frac{S^2 t^2_{\alpha/2}}{e^2} = \frac{0{,}009401 \times 1{,}964918^2}{0{,}01^2} = 362{,}97 \approx 363.$$

Novamente, $n = 363$ difere do valor anterior 481 e processo é repetido. Assim, o valor $t_{0{,}025;\nu=362} = 1{,}966539$ e o novo n é:

$$n = \frac{S^2 t^2_{\alpha/2}}{e^2} = \frac{0{,}009401 \times 1{,}966539^2}{0{,}01^2} = 363{,}57 \approx 364.$$

Novamente, $n = 364$ difere do valor anterior 363 e processo é repetido. Assim, o valor $t_{0{,}025;\nu=363} = 1{,}966521$ e o novo n é:

$$n = \frac{S^2 t^2_{\alpha/2}}{e^2} = \frac{0{,}009401 \times 1{,}966521^2}{0{,}01^2} = 363{,}56 \approx 364.$$

Portanto, como o valor de n neste passo iterativo é igual ao do passo anterior, o tamanho da amostra necessário para estimarmos a média é igual a $n = 364$ alunos, considerando um erro de 0,01 m para menos e para mais e confiança de 95%.

3) A estimativa pontual (amostra piloto) foi de $\hat{p} = 0{,}3333$. Considerando o erro de 3% e 95% de confiança, que nos fornece o quantil superior 2,5% da normal $Z_{0{,}025} = 1{,}96$, temos o seguinte tamanho de amostra, para a fórmula que utiliza informação da amostra piloto:

$$n = \frac{Z^2_{\alpha/2}\hat{p}(1-\hat{p})}{e^2} = \frac{1{,}96^2 \times 0{,}3333 \times 0{,}6667}{0{,}03^2} = 948{,}495 \approx 949.$$

Assim devemos amostrar $n = 949$ animais, para estimarmos a proporção de doentes com 95% e erro para mais e para menos de 3 pontos percentuais.

Se não utilizarmos a amostra piloto, devemos dimensionar a amostra por:

$$n_{max} = \frac{Z^2_{\alpha/2}}{4e^2} = \frac{1{,}96^2}{4 \times 0{,}03^2} = 1067{,}11 \approx 1068.$$

Assim, devemos amostrar $n = 1068$ animais, para estimarmos a proporção de doentes com 95% e erro para mais e para menos de 3 pontos percentuais. Pagamos um preço ao não utilizarmos a amostra piloto, de super-dimensionar o tamanho da amostra (amostra maior), o que encarece a pesquisa. Se, no entanto, a informação da amostra piloto for muito precária é um preço que vale a pena pagar.

Para o caso de diminuirmos o erro de 3% para 2%, teremos o seguinte dimensionamento de amostra utilizando a informação da amostra piloto:

$$n = \frac{Z^2_{\alpha/2}\hat{p}(1-\hat{p})}{e^2} = \frac{1{,}96^2 \times 0{,}3333 \times 0{,}6667}{0{,}02^2} = 2134{,}11 \approx 2135.$$

Assim devemos amostrar $n = 2134$ animais, para estimarmos a proporção de doentes com 95% e erro para mais e para menos de 2 pontos percentuais.

Se não utilizarmos a amostra piloto, devemos dimensionar a amostra por:

$$n_{max} = \frac{Z^2_{\alpha/2}}{4e^2} = \frac{1{,}96^2}{4 \times 0{,}02^2} = 2401.$$

Assim, devemos amostrar $n = 2401$ animais, para estimarmos a proporção de doentes com 95% e erro para mais e para menos de 2 pontos percentuais. Pagamos um preço ao não utilizarmos a amostra piloto, de super-dimensionar o tamanho da amostra (amostra maior), o que encarece a pesquisa. Se compararmos cada situação destes últimos dimensionamentos ($e = 2\%$) com a situação de erros iguais a 3 pontos percentuais, veremos que o preço a ser pago é maior ainda. Quanto menor o erro maior será o tamanho da amostra, mas o crescimento do tamanho da amostra é não-linear, ou seja, não é proporcional linearmente à redução do erro. Com a redução de 1/3 no valor do erro, a amostra mais do que dobrou.

4) A estimativa pontual foi de $\hat{p} = y/n = 11/500 = 0{,}022$, o que nos levaria a tomar a decisão de não efetuar o controle, pois 2,2% é menor do que 3%. Devemos, no entanto, tomar a decisão considerando o erro da estimação. Para isso construímos o intervalo de confiança de (95%) para p por:

$$\begin{aligned} IC_{1-\alpha}(p) :& \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ :& 0{,}022 \pm 1{,}96 \times \sqrt{\frac{0{,}022 \times 0{,}978}{500}} \\ :& 0{,}022 \pm 0{,}012910 = [0{,}009090;\ 0{,}034910]. \end{aligned}$$

A verdadeira proporção de plantas doentes, com aproximadamente 95% de confiança (intervalo aproximado), é um valor entre 0,91% e 3,49%. Como o intervalo de confiança abrangeu o nível de dano de 3%, então a recomendação é que se faça o controle, decisão diferente da que seria tomada somente com a estimativa pontual.

Se o número de sucesso fosse $y = 8$, em vez de $y = 11$, então a estimativa pontual seria de $\hat{p} = y/n = 8/500 = 0{,}016 = 1{,}6\%$, o que nos levaria a tomar decisão de não entrar com o controle químico da doença. Essa decisão não considera o erro de estimação. Logo, estimamos o intervalo de confiança de 95% para a verdadeira proporção de doentes, para contemplarmos o erro da estimação e tomarmos uma decisão cientificamente balizada. O intervalo de confiança é

$$\begin{aligned} IC_{1-\alpha}(p) :& \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ :& 0{,}016 \pm 1{,}96 \times \sqrt{\frac{0{,}016 \times 0{,}984}{500}} \\ :& 0{,}016 \pm 0{,}01099838 = [0{,}005002;\ 0{,}026998]. \end{aligned}$$

A verdadeira proporção de animais doentes, com aproximadamente 95% de confiança (intervalo aproximado), é um valor entre 0,50% e 2,70%. Como o intervalo de confiança não abrangeu o nível de dano de 3%, então a recomendação é que não se faça o controle químico, decisão igual a que seria tomada somente com a estimativa pontual. Então, por coincidência poderíamos ter a mesma decisão (sem considerar e considerando o erro da estimação), mas podemos ter decisões totalmente antagônicas. Então, devemos sempre considerar o erro da estimação para tomarmos decisões cientificamente balizadas.

5) O intervalo de 95% de confiança para variância populacional σ^2 é:

$$
\begin{aligned}
IC_{1-\alpha}(\sigma^2) &: \left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}};\ \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right]\\
&: \left[\frac{49\times 14}{70{,}222};\ \frac{49\times 14}{31{,}555}\right]\\
&:[9{,}77;\ 21{,}74].
\end{aligned}
$$

Podemos afirmar com 95% de confiança que a variância populacional (σ^2) deve ser um valor entre 9,77 e 21,74 (t/ha)^2. Assim, como intervalo contempla o valor do melhorista, tecnicamente determinado, não há evidências estatísticas significativas para não realizar o melhoramento da população.