Intervalo de Confiança - Margem de Erro

Tatiene Correia de Souza / UFPB
tatiene@de.ufpb.br

October 26, 2014

Margem de erro - relatórios da mídia

A pesquisa com adolescentes na volta às aulas em 1998 incluiu a afirmação: *margem de erro* $\pm 3.1\%$. A maioria das pesquisas é acompanhada por alguma afirmação semelhante. Além de *margem de erro*, podemos encontrar também: *erro amostral*, *erro máximo da pesquisa*, *erro estatístico*, entre outros.

O que é margem de erro

Tecnicamente a margem de erro é o termo adicionado e subtraído do estimador para formar um intervalo de confiança. Por exemplo, usando um nível de confiança de 95%, a margem de erro para proporção $\hat{p}$, assim, margem é igual $1,96\sqrt{\hat{p}(1-\hat{p})/n}$

De forma geral, podemos escrever do valor do erro amostral máximo como:

$$e_{max} = Z_{tab}\frac{\sigma}{\sqrt{n}}$$

Estimativas por intervalo de confiança

Motivação

Diferentes pesquisadores, selecionando amostras de uma mesma população, poderão obter estimativas obter estimativas pontuais diferentes para o mesmo parâmetro populacional. Isto está relacionado com o que denominamos de variabilidade amostral do estimador pontual. Uma forma mais apropriada seria construir um estimador que levasse em consideração essa variabilidade. Este seria o *estimador por intervalo* que combina o estimador pontual com o erro amostral máximo esperado.

Os limites inferir (LI) e o superior (LS) de um intervalo de confiança para um parâmetro θ é dado por:

$$\mathrm{LI} = \hat{\theta} - e_{\max}$$

e

$$\mathrm{LS} = \hat{\theta} + e_{\max}$$

Intervalo de confiança para média populacional

Aqui precisamos considerar dois casos:

1. Desvio padrão da população é conhecido (usar tabela da normal);
2. Desvio padrão da população não é conhecido (usar tabela da distribuição t).

Consideremos uma amostra aleatória simples $X_1, ..., X_n$ obtida de uma população com distribuição Normal, com média μ e variância σ^2 conhecida. Desta forma, a distribuição amostral da média também é Normal com média μ e variância σ^2, ou seja

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Assim, temos que

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$$

isto é, a variável Z tem distribuição Normal padronizada.
Consideremos que a probabilidade da variável Z tomar valores entre $-Z_{\alpha/2}$ e $Z_{\alpha/2}$ é $1 - \alpha$.

Então, temos que

$$P[-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}] = (1 - \alpha)$$

ou seja,

$$P\left[-Z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\alpha/2}\right] = (1 - \alpha)$$

o que implica que

$$P\left[\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha.$$

Com isso, o intervalo de confiança da média é dado por

$$IC(\mu, 1 - \alpha) = \left(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Exemplo

A distribuição dos pesos de pacotes de sementes de milho, enchidos automaticamente por uma certa máquina, é normal com desvio padrão, σ, conhecido e igual a 0,20kg. Uma amostra de 15 pacotes retirada ao acaso apresentou os seguintes pesos, em kg:
$20,05; 20,10; 20,25; 19,78; 19,69, 19,90; 20,20; 19,89; 19,70;$
$20,30; 19,93; 20,25; 20,18; 20,01; 20,09$
Construir os intervalos de confiança de 95% e 99% para o peso médio dos pacotes de sementes de milho.

Desvio padrão populacional desconhecido

Na maioria das situações práticas, o desvio padrão da população σ não é conhecido. Nessas situações, ele é substituído por pelo seu estimador S, desvio padrão amostral. Essa substituição causa uma alteração na distribuição de probabilidade a ser considerada. O valor a ser utilizado é obtido a partir de distribuição t com $n - 1$ graus de liberdade.

Distribuição t

Assim como a distribuição normal, a distribuição t é simétrica, com média zero, porém apresenta maior quantidade de dados nos extremos da distribuição.

Tendo os conceitos básicos sobre intervalos de confiança, vamos agora tratar uma situação mais realista: quando a variância σ^2 da população é desconhecida.
Consideremos uma amostra aleatória simples $X_1, X_2, ..., X_n$, obtida de uma população com distribuição Normal, com média μ e variância σ^2 desconhecidas. Como neste caso a variância é desconhecida, utilizaremos a variância amostral S^2. Assim, temos que

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

ou seja, a variável T tem distribuição t-Student com $n - 1$ graus de liberdade.
Então, ao fixarmos o nível de significância α, obtemos da Tabela da distribuição t-Student com $(n - 1)$ graus de liberdade, o valor $t_{((n-1),\alpha/2)}$, que satisfaz

$$P[-t_{((n-1),\alpha/2)} \leq T \leq t_{((n-1),\alpha/2)}] = 1 - \alpha$$

Analogamente ao caso anterior, obtemos que

$$P\left[-t_{((n-1),\alpha/2)} \leq \frac{\overline{X}-\mu}{s/\sqrt{n}} \leq t_{((n-1),\alpha/2)}\right] = 1-\alpha$$

ou seja,

$$P\left[\overline{X} - t_{((n-1),\alpha/2)}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{((n-1),\alpha/2)}\frac{s}{\sqrt{n}}\right] = 1-\alpha.$$

Logo, o intervalo com $100(1-\alpha)\%$ de confiança para ?, com variância desconhecida, será dado por

$$IC(\mu, 1-\alpha) = \left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right).$$

Exemplo

Os resíduos industriais jogados nos rios, muitas vezes, absorvem o oxigênio necessário à respiração de peixes e de outras formas de vida aquática. Um lei estadual exige um valor médio não inferior a 5 ppm de oxigênio dissolvido, cujo conteúdo seja suficiente para manter vida aquática. Seis amostras de água retiradas de um rio revelaram os índices: 4,9; 5,1; 4,9; 5,0; 5,0 e 4,7 ppm de oxigênio dissolvido. Construir o intervalo com 95% de confiança para a verdadeira média de oxigênio dissolvido, em ppm e interprete.

Determinação do tamanho amostral

Vimos que o processo de estimação envolve um erro que denominamos de erro amostral. A magnitude desses erros está relacionada com o tamanho da amostra. Temos que

$$n = \left(\frac{z_{\text{tab}} \sigma}{e_{\max}} \right)^2$$

Podemos observar que o tamanho da amostra depende do grau de confiança, do desvio padrão da população e do erro amostral máximo desejado.

Exemplo

Com relação ao exemplo dos pacotes de sementes de milho, qual tamanho de amostra será necessário coletar para garantir um erro amostral de no máximo 0,05 kg, com 95% confiança na estimação do verdadeiro valor médio?

cont. exemplo

Ou seja, desejamos determinar um tamanho amostral de modo que tenhamos 95% de confiança de que a média da amostra difira de (no máximo) 0,05 kg para mais ou para menos da média da população.

$$n = \left(\frac{1,96 0,20}{0,05} \right)^2 \cong 62$$

Portanto, vamos necessitar de uma amostra aleatória de 62 pacotes de milho para estimar a média populacional com precisão e a confiança desejadas.

Executivos-chefe desaprovam romances no ambiente de trabalho?

Ouvimos ao longo dos anos que empresas que desencorajam relacionamentos entre funcionários, até mesmo a ponto de proibir que maridos e esposas trabalhem juntos. Se duas pessoas casassem, um deveria sair. Na Fortune de 1994 foi relatada uma pesquisa feita com 200 executivos-chefe de empresas americanas, que explorou questões ligadas ao relacionamento entre funcionários de uma mesma empresa. Os executivos foram indagados: *Você aprova ou desaprova romances no ambiente de trabalho entre funcionários não-casados? Você diria que a empresa não deve interferir nessa questão?*

70% dos executivos disseram que a empresa não deve interferir neste assunto.

Como poderíamos apresentar esse percentual adotando uma certa confiabilidade?

Executivos-chefe desaprovam romances no ambiente de trabalho?

Ouvimos ao longo dos anos que empresas que desencorajam relacionamentos entre funcionários, até mesmo a ponto de proibir que maridos e esposas trabalhem juntos. Se duas pessoas casassem, um deveria sair. Na Fortune de 1994 foi relatada uma pesquisa feita com 200 executivos-chefe de empresas americanas, que explorou questões ligadas ao relacionamento entre funcionários de uma mesma empresa. Os executivos foram indagados: *Você aprova ou desaprova romances no ambiente de trabalho entre funcionários não-casados? Você diria que a empresa não deve interferir nessa questão?*

70% dos executivos disseram que a empresa não deve interferir neste assunto.

Como poderíamos apresentar esse percentual adotando uma certa confiabilidade?

Executivos-chefe desaprovam romances no ambiente de trabalho?

Ouvimos ao longo dos anos que empresas que desencorajam relacionamentos entre funcionários, até mesmo a ponto de proibir que maridos e esposas trabalhem juntos. Se duas pessoas casassem, um deveria sair. Na Fortune de 1994 foi relatada uma pesquisa feita com 200 executivos-chefe de empresas americanas, que explorou questões ligadas ao relacionamento entre funcionários de uma mesma empresa. Os executivos foram indagados: *Você aprova ou desaprova romances no ambiente de trabalho entre funcionários não-casados? Você diria que a empresa não deve interferir nessa questão?*

70% dos executivos disseram que a empresa não deve interferir neste assunto.

Como poderíamos apresentar esse percentual adotando uma certa confiabilidade?

Com 95% de confiança, a verdadeira proporção de todos os executivos que são de opinião de que a empresa não deve inferir em assuntos de romance no ambiente de trabalho é algo entre 64% e 77%. Como você acha que esses valores foram obtidos?

Suponha que alguém nos peça essa informação como mais confiança, por exemplo, 99%. O que você acha que deve acontecer? E se fosse com 90%?

Intervalo de confiança da proporção

O parâmetro de p - a proporção de todos os indivíduos na população com a característica de interesse. A estimativa de p é a proporção amostral $\hat{p}$, a proporção de indivíduos incluídos na pesquisa com aquela característica. Quando n é grande, temos que a distribuição de $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ é aproximadamente $\mathcal{N}(0,1)$.
Portanto, o intervalo de confiança para proporção é dado por
$\hat{p} \pm z_{\text{tab}}\text{ep}(\hat{p})$

Voltemos ao exemplo dos executivos...

Dos 100 intervalos de confiança construídos é esperado que em 95%, a verdadeira proporção de todos os executivos-chefe americanos que são de opinião que a empresa não deveria interferir em assuntos de romance no ambiente de trabalho é algo entre 64% e 77%.
Pergunta: Como você acha que chegaram a essa conclusão?

Tamanho amostral - proporção

Suponha que queremos estimar a proporção da população portadora de hepatite B, usando uma amostra aleatória dessa população. Queremos que o tamanho da amostra seja grande o suficiente, de modo que a margem de erro de nossa estimativa seja aceitável, digamos, não maior do que 3%.

Lembrem-se: sabemos que o intervalo de confiança da proporção é dado por: $\hat{p} \pm z_{\mathrm{tab}}\mathrm{ep}(\hat{p})$.

Portanto, queremos que $z_{\mathrm{tab}}\mathrm{ep}(\hat{p}) < 0,03$.

Note que não temos informações sobre $\hat{p}$. Neste caso é razoável fazermos $\hat{p} = 0.5$.

Então, considerando os dados da questão, temos:

O tamanho mínimo da amostra é aproximadamente 1067 $((1,96/0,03)^2 x 0,5 x 0,5)$.

cont. Tamanho amostral - proporção

Pergunta: O que você achou o valor do tamanho amostral encontrado? Grande?
Nota: O uso de $p = 0.5$ é uma 'é uma adivinhação segura' que garante que uma margem de erro não maior do que o $e_{\max}$. Se você soubesse que a verdadeira proporção está próxima de 0 ou 1, usar $p = 0.5$ lhe conduzirá a tomar uma amostra muito maior (mais cara) do que o estritamente necessário.

Exercício

Considerando os dados da questão anterior, obtenha o tamanho amostral necessário quando:
a) $p = 0,3$; $n \cong 896$
b) $p = 0,9$; $n \cong 384$

Exercício

Calcule o tamanho amostral necessário para que a margem de erro para uma verdadeira proporção p não seja maior do que 2% sob as seguintes circunstâncias. Adote uma confiança de 95%.

a) Você não quer fazer nenhuma suposição sobre o valor de p.

$n = 2401$

b) Você está seguro de que $p < 0,15$. $n \cong 1225$

c) Você está seguro de que $p > 0,85$. $n \cong 1225$

Comparação de duas médias

Banford et al. [1982] notaram que as concentrações de tiol dentro de células sanguíneas humanas são raramente determinadas em estudos clínicos. Os autores divulgaram um novo método confiável para medir a concentração de tiol e demostraram que, em pelo menos uma doença (artrite reumatoide), a mudança na condição do tiol na lisina das células sanguíneas cheias é substancial. Havia dois grupos voluntários. O primeiro grupo que era ‘normal’ e o segundo, que sofria de artrite reumatoide.

cont. Comparação de duas médias

- Concentração de tiol no grupo ‘normal’ - Grupo 1: 1,84; 1,92; 1,94; 1,92; 1,85; 1,91; 2,07.
- Concentração de tiol no grupo ‘reumatoide’ - Grupo 2: 2,81; 4,06; 3,62; 3,27; 3,27; 3,76.

- Tamanho amostral do Grupo 1 = 7; média amostral = 1,92 e desvio-padrão amostral 0,075.
- Tamanho amostral do Grupo 2 = 6; média amostral = 3,46 e desvio-padrão amostral 0,440.

Intervalo de confiança para diferença de médias - variâncias conhedidas

Intervalos de confiança para uma diferença entre médias populacionais $(\mu_1 - \mu_2)$:

$$(\overline{x}_1 - \overline{x}_2) \pm z_{tab} ep(\overline{x}_1 - \overline{x}_2),$$

em que $ep(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, z_{tab} é o quantil da distribuição normal.

Intervalo de confiança para diferença de médias - variâncias desconhedidas e iguais

Intervalos de confiança para uma diferença entre médias populacionais $(\mu_1 - \mu_2)$:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{tab} ep(\overline{x}_1 - \overline{x}_2),$$

em que $ep(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, t_{tab} é o quantil da distribuição t com $n_1 + n_2 - 2$ graus de liberdade.

Intervalo de confiança para diferença de médias - variâncias desconhedidas e diferentes

Intervalos de confiança para uma diferença entre médias populacionais $(\mu_1 - \mu_2)$:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{tab} ep(\overline{x}_1 - \overline{x}_2),$$

em que $ep(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, t_{tab} é o quantil da distribuição t com ν graus de liberdade, onde

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

Voltando ao exemplo...

Temos que $\overline{x}_R - \overline{x}_N = 3,465 - 1,921 = 1,544$;
$ep(\overline{x}_R - \overline{x}_N) = \sqrt{\frac{s_R^2}{n_R} + \frac{s_N^2}{n_N}} = 0,182$; $\nu = 5.25$; $t_{tab} = 2,571$. Portanto o intervalo de confiança de 95% é dado por: $IC(\mu_R - \mu_N) = [1,08; 2,01]$. Ou seja, com 95% de confiança, a verdadeira média da concentração de tiol para população com artrite reumatoide excede aquela para a população normal por algo entre 1,08 e 2,01.

Podemos dizer ainda que estamos 95% confiantes de que um intervalo entre 1,08 e 2,01 contem a diferença das média populacionais para concentração de tiol em pacientes com artrite e sem artrite. O fato que nos chama atenção é que ambos limites são positivos o que indica que a média da concentração de tiol em pacientes com artrite é maior do que em pacientes sem artrite.

Exemplo das pilhas

Um artigo publicado em 1983 comparou vários tipos de pilhas. A vida útil média das pilhas alcalinas AA Duracell e das pilhas alcalinas AA da marca B foi dada por 4,1 horas e 4,5 horas, respectivamente.

Suponha que $\overline{X}$ seja a média amostral de 100 pilhas Duracell e $\overline{Y}$ seja a média amostral de 100 pilhas da marca B. Qual o valor médio de $\overline{X} - \overline{Y}$? Sua resposta depende do tamanho amostral? Justifique. [a]

Suponha que os desvios padrão da vida útil da população seja, 1,8 hora para pilhas Duracell e 2,0 horas para pilha tipo B. Qual a variância de $\overline{X} - \overline{Y}$? [b]

---

[a] -0.4
[b] 0.0724

cont. exemplo pilha

Como você construiria o intervalo de confiança para diferença das médias da vida útil das pilhas?

Os alunos universitários do sexo masculino se cansam mais do que suas colegas?

Essa pergunta foi avaliada no paper publicado em 1991. Os autores administraram uma escalam chamada de propensão ao cansaço a 97 homens e 148 mulheres de uma universidade dos Estados Unidos. Os dados apoiam a hipótese da pesquisa de que o graus de propensão ao cansaço médio é maior para os homens que para as mulheres? Apresente o intervalo de confiança para a diferença das médias.

- homens: $n_H = 97$; $\overline{x}_H = 10,40$ e $S_H = 4,83$.
- mulheres: $n_M = 148$; $\overline{x}_M = 9,26$ e $S_M = 4,68$.