Distribuição de $\bar{X}$

Exemplo

Uma variável aleatória X tem distribuição normal, com média 100 e desvio padrão 10.

(a) Qual a $P(90 < X < 110)$?

(b) Se $\bar{X}$ for a média de uma amostra de 16 elementos retirados dessa população, calcule $P(90 < \bar{X} < 110)$.

(c) Represente, num único gráfico, as distribuições de X e $\bar{X}$.

(d) Que tamanho deveria ter a amostra para que $P(90 < \bar{X} < 110) = 0.95$?

Fonte: Morettin & Bussab, Estatística Básica 5^a edição, pág 274.

Distribuição de $\bar{X}$

(a) Devemos padronizar o evento, para comparar com a distribuição normal padrão.

$$P(90 < X < 110) = P\left(\frac{90-100}{10} < \frac{X-100}{10} < \frac{110-100}{10}\right)$$

$$= P(-1 < Z < 1) = P(Z < 1) - P(Z < -1) = \Phi(1) - \Phi(-1)$$

Consultando a tabela disponível na página da disciplina[1], vemos que $\Phi(1) = 0.8413$. Para encontrar $\Phi(-1)$, note que a distribuição normal é simétrica e portanto $\Phi(-z) = 1 - \Phi(z)$, daí $\Phi(-1) = 0.1569$ e portanto $\Phi(1) - \Phi(-1) = 0.6844$.

[1] http://www.ime.unicamp.br/~veronica/Coordenadas1s/N.pdf

Distribuição de $\bar{X}$

(b) Se temos uma amostra e tiramos a média, note que

$$\text{Var}\left(\bar{X}\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{(1)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(X_i\right) \overset{(2)}{=} \frac{\text{Var}\left(X\right)}{n}$$

onde a igualdade (1) vale por independência, e a igualdade (2) vale por serem identicamente distribuidas. Consequentemente, o desvio padrão novo será $\sigma/\sqrt{n}$, ou $10/4$. Temos então que

$$P(90 < \bar{X} < 110) = P\left(\frac{90-100}{10/4} < \frac{\bar{X}-100}{10/4} < \frac{110-100}{10/4}\right)$$

(b) Continuando,

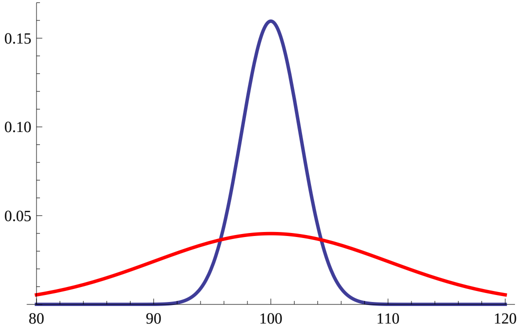
$$= P(-4 < Z < 4) = P(Z < 4) - P(Z < -4) = \Phi(4) - \Phi(-4)$$

Se consultarmos a tabela agora, veremos que a probabilidade $P(Z < 4)$ é tão grande nem está listada. Ela então pode ser considerada como aproximadamente igual a 1.

De fato, com a ajuda de algum método de integração numérica, podemos verificar que $\Phi(4) - \Phi(-4)$ é igual a 0.9999367.

Distribuição de $\bar{X}$

(c) No gráfico, a função de densidade de X está em vermelho, e a de $\bar{X}$ em azul:



Distribuição de $\bar{X}$

(d) Queremos resolver a seguinte equação:

$$P\left(\frac{90-100}{10/\sqrt{n}} < \frac{\bar{X}-100}{10/\sqrt{n}} < \frac{110-100}{10/\sqrt{n}}\right) = 0.95$$

Note que, consultando a tabela, vemos que $P(-q < Z < q) = 0.95$ se $q = 1.96$. Então a equação que queremos resolver pode ser reescrita como:

$$\frac{110-100}{10/\sqrt{n}} = 1.96 \Leftrightarrow \sqrt{n}\frac{110-100}{10} = 1.96 \Leftrightarrow n = 1.96^2$$

Portanto, $n = 4$ é suficiente para obtermos a confiança desejada.

Precisão e Tamanho Amostral

Exemplo

Qual deve ser o tamanho de uma amostra cuja a população da qual ela ser sorteada possui um desvio-padrão igual a 10 para que a diferença da média amostral para a média da população, em valor absoluto, seja menor que 1, com coeficiente de confiança igual a:

(a) 95%

(b) 99%

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 308.

Precisão e Tamanho Amostral

(a) Note que se $X \sim N(\mu, \sigma^2)$, então $X - \mu \sim N(0, \sigma^2)$. Sabemos que $\sigma = 10$, e que o desvio-padrão do estimador da média, $\bar{X}$, será $10/\sqrt{n}$. Queremos que $P(|\bar{X} - \mu| < 1) = 0.95$. Mas o evento é equivalente a

$$P\left(-\sqrt{n}/10 < Z < \sqrt{n}/10\right)$$

Como $P(-1.96 < Z < 1.96) = 0.95$, então $\sqrt{n}/10 = 1.96$ ou $n \approx 385$.

(b) De modo análogo, temos que $P(-2.57 < Z < 2.57) = 0.99$, então $\sqrt{n}/10 = 2.57$ ou $n \approx 665$.

Intervalo de Confiança para proporções

Exemplo

Suponha que $p = 30\%$ dos estudantes de uma escola sejam mulheres. Colhemos uma amostra aleatória simples de $n = 10$ estudantes e calculamos $\hat{p} =$ proporção de mulheres na amostra. Qual a probabilidade de que $\hat{p}$ difira de p em menos de 0.01? E se $n = 50$?
Adaptado de: Morettin & Bussab, Estatística Básica 5ª edição, pág 276.

Intervalo de Confiança para proporções

Temos que a probabilidade que desejamos encontrar é dada por

$$P\left(|\hat{p} - p| < 0.01\right) = P\left(-0.01 < \hat{p} - p < 0.01\right)$$

Onde p é o valor verdadeiro da proporção de mulheres, e $\hat{p}$ a proporção observada na amostra. Sabemos que se n é grande, $\hat{p} - p$ pode ser aproximada por uma normal $N\left(0, p(1-p)/n\right)$. Como $p = 0.3$, temos que

$$\mathrm{Var}\left(\hat{p} - p\right) = \frac{0.3 \cdot 0.7}{10} = 0.021$$

Intervalo de Confiança para proporções

Portanto, a probabilidade pedida é igual a

$$P\left(\frac{-0.01}{\sqrt{0.021}} < Z < \frac{0.01}{\sqrt{0.021}}\right) = P(-0.07 < Z < 0.07) = 0.056$$

Mas $n = 10$ é grande? Podemos comparar essa probabilidade com o resultado exato.

Não sabemos a distribuição de $\hat{p}$, mas o evento $\hat{p} = \alpha$ é igual ao evento $\sum X_i = n\alpha$, onde X_i são v.a. independentes e identicamente distribuidas Bernoulli(0.3). A soma é portanto Binomial$(10, 0.3)$.

Intervalo de Confiança para proporções

O evento $\{|\hat{p} - p| < 0.01\}$ é igual ao evento $\{|\sum X_i - 10 \cdot 0.3| < 0.1\}$. Como $\sum X_i$ assume somente valores inteiros, temos que

$$\left\{\left|\sum_{i=1}^{10} X_i - 10 \cdot 0.3\right| < 0.1\right\} = \left\{\sum_{i=1}^{10} X_i = 3\right\}.$$

Portanto,

$$P\left(\left\{\sum_{i=1}^{10} X_i = 3\right\}\right) = \binom{10}{3} 0.3^3 0.7^7 = 0.267.$$

Temos uma probabilidade que é 5 vezes maior que a aproximação.

Intervalo de Confiança para proporções

Tome $n = 50$, agora. Podemos modificar rapidamente as contas da aproximação normal. A variância agora é 0.0042, e portanto a probabilidade aproximada é

$$P\left(\frac{-0.01}{\sqrt{0.0042}} < Z < \frac{0.01}{\sqrt{0.0042}}\right) = P(-0.154 < Z < 0.154) = 0.12239$$

A probabilidade exata agora é dada pelo evento $\{|\sum X_i - 50 \cdot 0.3| < 0.5\}$, ou simplesmente $\{\sum_{i=1}^{50} X_i = 15\}$.

Intervalo de Confiança para proporções

Observe agora que

$$P\left(\sum_{i=1}^{50} X_i = 15\right) = \binom{50}{15} 0.3^{15} 0.7^{50-15} = 0.12237$$

A diferença agora é muito menor e, é possível demonstrar, medida que $n \to \infty$ ela tende 0. É preciso contudo ter em mente que a aproximação só é válida para grandes tamanhos de amostra, independentes e identicamente distribuidas.

Intervalo de Confiança para proporções

Exemplo

Uma amostra aleatória de 625 donas de casa revela que 70% delas preferem a marca A de detergente. Construir um intervalo de confiança para $p =$ proporção das donas de casa que preferem A com coeficiente de confiança $\gamma = 90\%$.
Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 308.

Intervalo de Confiança para proporções

Temos que em nossa amostra aleatória $\hat{p} = 0.7$. Como $\hat{p} \sim N(p, p(1-p)/n)$, então o intervalo de confiança é dado por

$$\left(\hat{p} - z(\gamma)\sqrt{\hat{p}(1-\hat{p})/n}\ ;\ \hat{p} + z(\gamma)\sqrt{\hat{p}(1-\hat{p})/n}\right)$$

Temos que para $\gamma = 0.90$, $z(\gamma) = 1.68$ e portanto o intervalo de confiança para a proporção de donas de casa que preferem o detergente A é dado por

$$\left(0.7 - 1.68\sqrt{0.7 \cdot 0.3/625}\ ;\ 0.7 + 1.68\sqrt{0.7 \cdot 0.3/625}\right)$$

$$(0.6692\ ;\ 0.7308)$$

Intervalo de Confiança para proporções

Exercício

Suponha que estejamos interessados em estimar a porcentagem de consumidores de um certo produto. Se a amostra de tamanho 300 forneceu 100 indivíduos que consomem o dado produto, determine:

(a) O intervalo de confiança de p, com c.c. de 95%; interprete o resultado.

(b) O tamanho da amostra para que o erro da estimativa não exceda 0.02 unidades com probabilidade de 95%; interprete o resultado.

Fonte: Morettin & Bussab, Estatística Básica 5^a edição, pág 309.

Intervalo de Confiança para proporções

(a) O intervalo de confiança a 95% de confiabilidade é dado por:

$$\mathrm{IC}(p; 0.95) = 0.333 \pm 1.96\sqrt{\frac{0.333 \cdot 0.667}{300}} = 0.333 \pm 0.053$$

Ou simplesmente $(0.280; 0.387)$.

Interpretação: Se pudéssemos construir um grande número de intervalos aleatórios para p, todos baseados em amostras de tamanho n, 95% deles conteriam o parâmetro p.

Intervalo de Confiança para proporções

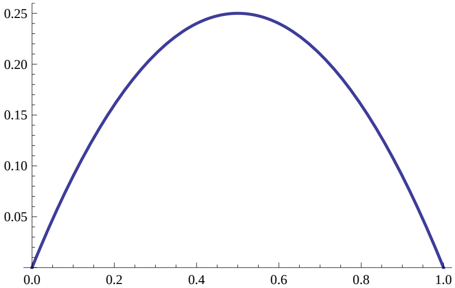
(b) Utilizando a estimativa da amostra observada ($\hat{p} = 0.333$), temos que n é dado por

$$n = \left(\frac{1.96}{0.02}\right)^2 \times 0.333 \times 0.667 \cong 2134.$$

Contudo, frequentemente devemos determinar o tamanho da amostra antes de realizar qualquer experimento, isto é, sem nenhuma informação prévia de p. Se esse for o caso, devemos considerar o caso em que a variância da amostra é a pior possível.

Intervalo de Confiança para proporções

(b) Se olhamos a variância como função de p, obtemos o seguinte gráfico:



Note que a variância é máxima quando $p = 1/2$.

Intervalo de Confiança para proporções

(b) Utilizando o valor máximo de $p(1-p)$, isto é, $1/4$, obtemos

$$n = \left(\frac{1.96}{0.02}\right)^2 \times \frac{1}{4} \cong 2401$$

Interpretação: Utilizando o tamanho amostral encontrado, teremos uma probabilidade de 95% de que a proporção amostral não difira do verdadeiro valor de p em menos que 2%.

Note que a prática de obter amostras pequenas para examinar p, e aí determinar o tamanho amostral sem utilizar o "pior caso", é no que consiste a idéia de *amostras piloto*.

Intervalo de Confiança

Exemplo

Estão sendo estudados dois processos para conservar alimentos, cuja principal variável de interesse é o tempo de duração destes. No processo A, o tempo X de duração segue a distribuição $N(\mu_A, 100)$, e no processo B o tempo Y obedece à distribuição $N(\mu_B, 100)$. Sorteiam-se duas amostras independentes: a de A, com 16 latas, apresentou tempo médio de duração igual a 50, e a de B, com 25 latas, duração média igual a 60.

(a) Construa um IC para μ_A e μ_B, separadamente.

Intervalo de Confiança

Exemplo

(b) Para verificar e os dois processos podem ter o mesmo desempenho, decidiu-se construir um IC para a diferença $\mu_A - \mu_B$. Caso o zero pertença ao intervalo, pode-se concluir que existe evidência de igualdade dos processos. Qual seria sua resposta?

Fonte: Morettin & Bussab, Estatística Básica 5^a edição, pág 318.

Intervalo de Confiança

(a) Para o caso geral, o intervalo de confiança para μ, com coeficiente de confiabilidade γ, é dado por

$$\left(\bar{X} - z(\gamma)\sqrt{\sigma^2/n}\ ;\ \bar{X} + z(\gamma)\sqrt{\sigma^2/n}\right)$$

Repare que $\sigma_A = \sigma_B$. Para o coeficiente de confiança $\gamma = 0.95$, por exemplo, temos $z(\gamma) = 1.96$, e os intervalos de confiança serão, respectivamente:

$$\text{IC}(\mu_A) = \left(50 - 1.96\sqrt{100/16}\ ;\ 50 + 1.96\sqrt{100/16}\right)$$

$$\text{IC}(\mu_B) = \left(60 - 1.96\sqrt{100/25}\ ;\ 60 + 1.96\sqrt{100/25}\right)$$

Intervalo de Confiança

(a) (cont.) Fazendo as contas, obtemos que

$$\text{IC}(\mu_A) = (45.1\ ;\ 54.9)$$

$$\text{IC}(\mu_B) = (56.08\ ;\ 63.92)$$

Observe que os intervalos não se interceptam; temos evidência para dizer que as durações médias serão diferentes, a 95% de confiança.

(b) Temos aqui duas amostras diferentes mas independentes. A diferença $\bar{X}_A - \bar{X}_B$ tem distribuição Normal, com média $\mu_A - \mu_B$ e variância $\sigma_A^2/n_A + \sigma_B^2/n_B$.

(b) (cont.) Então o intervalo de confiança para $\mu_A - \mu_B$ é dado por

$$\left(\bar{X}_A - \bar{X}_B - z(\gamma)\sqrt{\sigma_A^2/n_A + \sigma_B^2/n_B} \;; \right.$$

$$\left. \bar{X}_A - \bar{X}_B + z(\gamma)\sqrt{\sigma_A^2/n_A + \sigma_B^2/n_B} \right)$$

Aplicando os valores conhecidos ou observados, e fixando a confiança em $\gamma = 0.95$ temos:

$$\mathrm{IC}(\mu_A - \mu_B) = \left(50 - 60 - 1.96\sqrt{100/16 + 100/25} \;; \right.$$

$$\left. 50 - 60 + 1.96\sqrt{100/16 + 100/25} \right)$$

Intervalo de Confiança

(b) (cont.) Executando as contas, obtemos finalmente que

$$\text{IC}(\mu_A - \mu_B) = (-16.27 \ ; \ -3.72)$$

Em concordância com o item (a), vemos que 0 não está contido no intervalo e, portanto, rejeitamos a hipótese, a $\gamma = 0.95$ de confiança, das médias μ_A e μ_B serem iguais.