2. Intervalos de Confiança

ANADI

Licenciatura em Engenharia Informática

Instituto Superior de Engenharia do Porto

Ano letivo 2018/2019

Relembrando:

- população: conjunto de todos os objetos cujas características pretendemos estudar;
- amostra: qualquer subconjunto finito da população;
- medidas como a média e o desvio padrão são usadas para descrever amostras e populações.

Estas medidas chamam-se:

- parâmetros: quando se referem às características da população;
- estatísticas: quando se referem às características de uma amostra. As estatísticas estimam o valor dos parâmetros que pretendemos determinar.

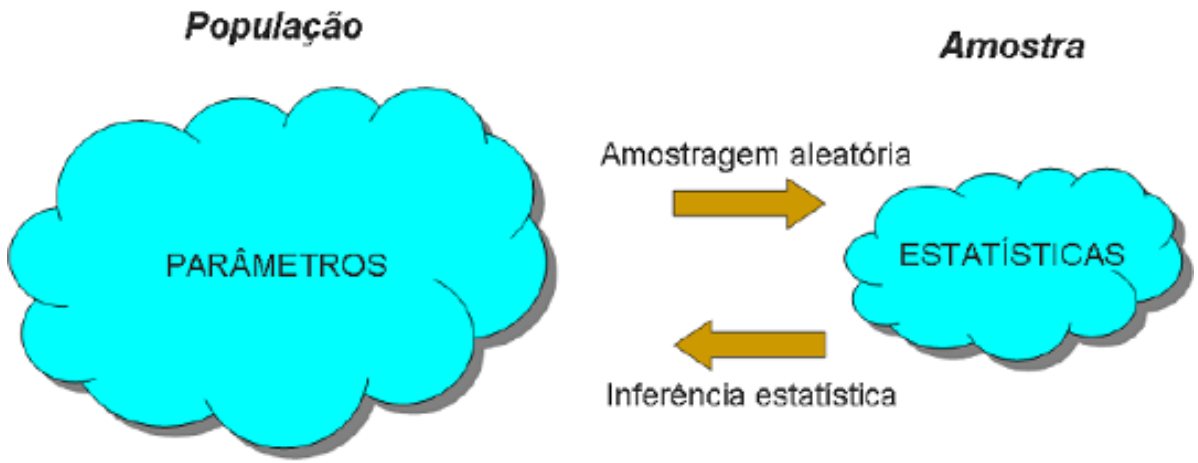
A amostragem é uma técnica de seleção de elementos de uma população para se estimar propriedades e características da população.

Uma amostra aleatória simples de uma população finita $X_1, X_2, \ldots, X_n$ é uma amostra obtida por um processo de amostragem aleatória simples com reposição, isto é, em que as observações $X_1, X_2, \ldots, X_n$

1. são independentes,
2. têm a mesma distribuição de probabilidade,

e, por isso, dizem-se independentes e identicamente distribuídas (i.i.d.).

Uma amostra aleatória simples de uma população infinita é uma amostra em que as observações $X_1, X_2, \ldots, X_n$ são i.i.d..



Estimador e estimativa

Partindo de estatísticas baseadas numa amostra aleatória, é possível fazer inferências acerca do valor de parâmetros de uma população.

- Estimador pontual de um parâmetro θ de uma população: é uma estatística $\hat{\Theta}$ usada para estimar o valor de θ.
- Estimativa pontual de um parâmetro θ de uma população: é um valor $\hat{\theta}$ de uma estatística $\hat{\Theta}$.

parâmetro da população, θ	estimador de θ, $\hat{\Theta}$	estimativa de θ, $\hat{\theta}$
média, μ	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
variância, σ^2	$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$	$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$
proporção, p	$\hat{P} = \frac{\sum_{i=1}^{n} X_i}{n}$, $X_i \sim Bernoulli(p)$	$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$

Intervalos de confiança

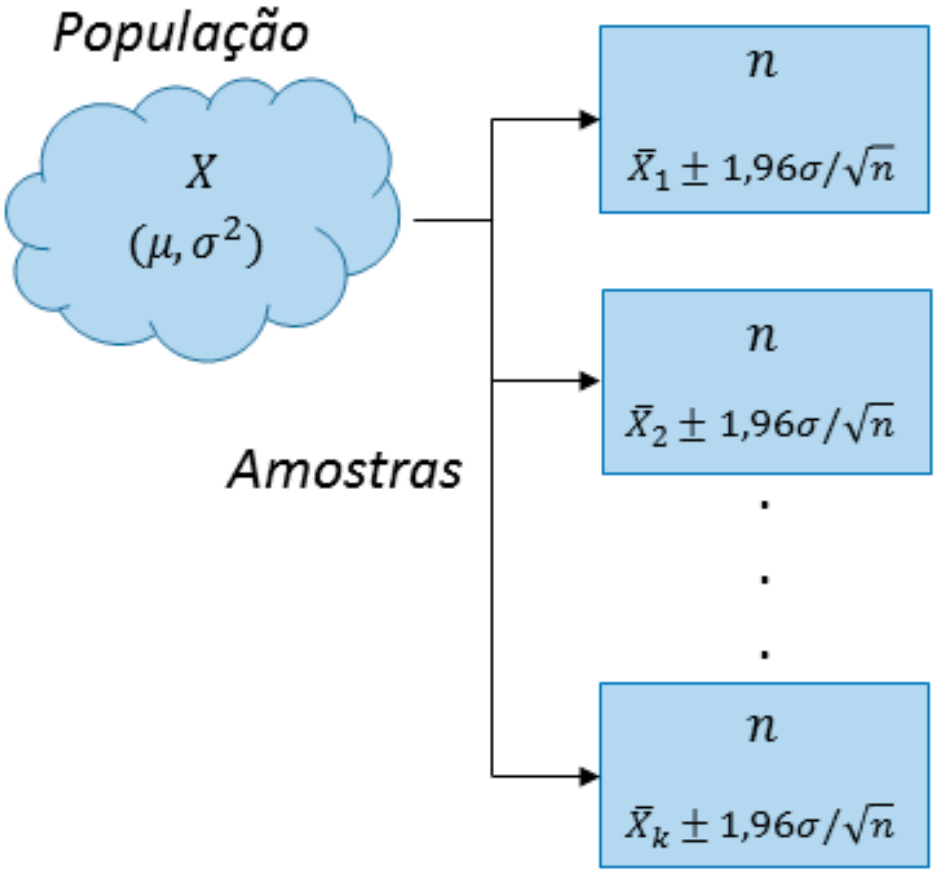
Ao fazer uma estimativa de um parâmetro θ de uma população queremos conhecer a grandeza do erro de amostragem.

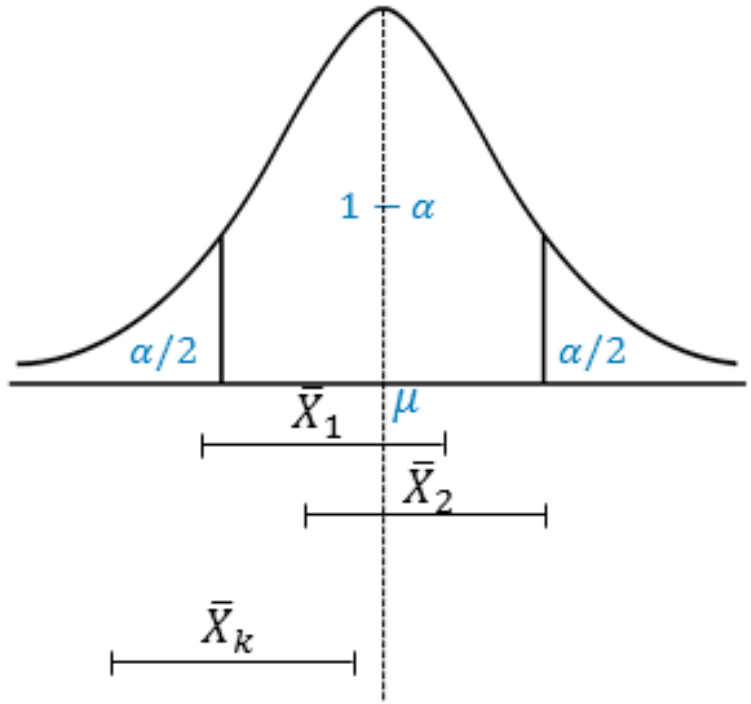
Como os estimadores pontuais não dão essa informação, determina-se um intervalo onde se espera encontrar o valor do parâmetro θ:

Intervalo de confiança (IC) de $(1-\alpha)\times 100\%$ para θ: é um intervalo aleatório $[LI(\hat{\Theta}), LS(\hat{\Theta})]$ em que os limites de confiança $LI(\hat{\Theta})$ e $LS(\hat{\Theta})$ são duas estatísticas amostrais tais que

$$P(LI(\hat{\Theta}) \leq \theta \leq LS(\hat{\Theta})) = 1-\alpha$$

sendo $1-\alpha$ o coeficiente de confiança e $\alpha \in]0,1[$ o nível de significância.





Aproximadamente 95% dos intervalos contêm μ

Idealmente, um intervalo aleatório deverá ter amplitude pequena (grande precisão) e coeficiente de confiança elevado (probabilidade elevada do IC conter o parâmetro desconhecido θ).

Para um tamanho de amostra fixo, o coeficiente de confiança só pode aumentar, se a amplitude do intervalo também aumentar.

Para valores do coeficiente de confiança elevados, a amplitude do IC aumenta rapidamente.

Assim, os valores mais típicos do coeficiente de confiança $1 - \alpha$ são 0.99, 0.95 e 0.90.

O problema de determinar um IC para um parâmetro θ (ou seja, $LI(\hat{\Theta})$ e $LS(\hat{\Theta})$) reduz-se a encontrar um estimador pontual de θ cuja distribuição de probabilidade seja conhecida e não dependa de θ.

Intervalos de confiança para a média populacional (σ^2 conhecido)

Para populações normais ou, pelo Teorema do Limite Central, para amostras de tamanho suficientemente grande, sabemos que

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Na prática, usa-se esta aproximação quando a variância σ^2 é conhecida e as amostras são de tamanho superior ou igual a 30.

Seja z_p, $0 < p < 1$, o percentil $100p$ da distribuição $N(0,1)$. Então

$$P\left(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1 - \alpha,$$

ou seja,

$$P\Big(- z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \Big) = 1 - \alpha,$$

e, portanto,

$$P\Big(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \Big) = 1 - \alpha.$$

Obtemos então o intervalo

$$\Big[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \Big].$$

Este intervalo aleatório tem uma probabilidade $1 - \alpha$ (exata, no caso normal, ou aproximada) de conter o verdadeiro, mas desconhecido, valor da média μ.

Depois de realizarmos a amostragem, substituímos $\bar{X}$ por $\bar{x}$ e obtemos o intervalo determinístico

$$IC_{(1-\alpha)\times 100\%}(\mu) = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

que nos dá $(1-\alpha)\times 100\%$ de confiança do erro cometido (ou seja, o valor absoluto da diferença entre $\bar{x}$ e μ) ser inferior a $z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Este (intervalo centrado na média amostral) não é o único intervalo de $(1-\alpha)\times 100\%$ de confiança, mas é aquele em que a amplitude é mínima.

Intervalos de confiança para a média populacional (σ^2 desconhecida)

Usualmente, a variância da população, σ^2, é desconhecida, pelo que temos de recorrer à estatística S^2 (variância amostral) e usar a variável aleatória $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$.

Teorema

Se $\bar{X}$ é a média e S^2 a variância de uma a.a. i.i.d. de tamanho n, extraída de uma população normal com média μ e variância σ^2, então

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T(n-1),$$

onde T é a distribuição t-Student com $n-1$ graus de liberdade.

Resumindo:

Sejam $X_1, X_2, \ldots, X_n$ uma a.a. i.i.d. de uma população com média μ e variância σ^2. Sejam $\bar{X}$ e S^2 a média e a variância da a.a.. Então:

$X \sim N(\mu, \sigma^2)$	σ^2 conhecido	$n \geq 30$	Estatística de teste	$IC_{(1-\alpha)\times 100\%}(\mu)$
Sim	Sim	Indiferente	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
Não	Sim	Sim	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \underset{aprox}{\sim} N(0,1)$	
Sim	Não	Indiferente	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T(n-1)$	$\left[\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$
Não	Não	Sim	$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \underset{aprox}{\sim} N(0,1)$	$\left[\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$

Nota: Para amostras de tamanho suficientemente grande (tipicamente, $n \geq 30$) os resultados obtidos ao usar-se a distribuição normal standardizada para aproximar a distribuição t-Student são muito próximos (*o IC quando se usa a distribuição normal tem mais precisão*), pelo que, no software R usualmente usa-se a distribuição t-Student.

Intervalos de confiança para a diferença entre médias populacionais

Sejam $X_1, X_2, \ldots, X_{n_X}$, e $Y_1, Y_2, \ldots, Y_{n_Y}$, a.a. i.i.d. de duas populações com médias μ_X e μ_Y e variâncias σ_X^2 e σ_Y^2, respetivamente. Sejam $\bar{X}$ e $\bar{Y}$ e S_X^2 e S_Y^2 as médias e as variâncias respetivas das a.a.. Então:

popul. normais	σ_X^2, σ_Y^2 conhecidas	$n_X \geq 30$ $n_Y \geq 30$	Estatística de teste	Limites $IC_{(1-\alpha)\times 100\%}(\mu_X - \mu_Y)$
Sim	Sim	Indif.	$Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \sim N(0,1)$	$(\bar{X}-\bar{Y}) \mp z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$
Não	Sim	Sim	$Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \underset{aprox}{\sim} N(0,1)$	
Sim	Não, iguais	Indif.	$T = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim T(n_X+n_Y-2)$	$(\bar{X}-\bar{Y}) \mp t_{1-\alpha/2}S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}$
Não	Não, iguais	Sim	$Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \underset{aprox}{\sim} N(0,1)$	$(\bar{X}-\bar{Y}) \mp z_{1-\alpha/2}S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}$
Ind.	Não, difer.	Sim	$Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}} \underset{aprox}{\sim} N(0,1)$	$(\bar{X}-\bar{Y}) \mp z_{1-\alpha/2}\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}$

onde $S^2 = \frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}$.

Intervalos de confiança para uma proporção

Para amostras de tamanho n suficientemente grande ($np \geq 5$ e $n(1-p) \geq 5$), sabemos, pelo Teorema do Limite Central, que a proporção amostral

$$\hat{P} = \frac{\sum_{i=1}^{n} X_i}{n} \underset{aprox}{\sim} N\left(p, \frac{p(1-p)}{n}\right) \Leftrightarrow Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{aprox}{\sim} N(0,1).$$

Usando um procedimento análogo, torna-se agora simples determinar o intervalo de confiança a $(1-\alpha) \times 100\%$ para a proporção p:

$$\left[\hat{P} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right].$$

Este intervalo aleatório tem uma probabilidade aproximada $1-\alpha$ de conter o verdadeiro, mas desconhecido, valor da proporção p.

Intervalos de confiança para a diferença entre proporções

Dadas duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, suficientemente grandes, foi visto que, pelo Teorema do Limite Central, a diferença entre as proporções amostrais

$$\hat{P}_1 - \hat{P}_2 \underset{aprox}{\sim} N(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})$$

e (...), portanto, o intervalo

$$\left[(\hat{P}_1 - \hat{P}_2) - z_{1-\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}, (\hat{P}_1 - \hat{P}_2) + z_{1-\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}\right].$$

Este intervalo aleatório tem uma probabilidade aproximada $1-\alpha$ de conter o verdadeiro, mas desconhecido, valor da proporção $p_1 - p_2$.