Matemática Computacional

Capítulo 4

Distribuições amostrais e intervalos de confiança (IC)

Licenciatura em Engenharia Informática
ISEP
(2024/2025)

Conteúdo

Relação entre população e amostra

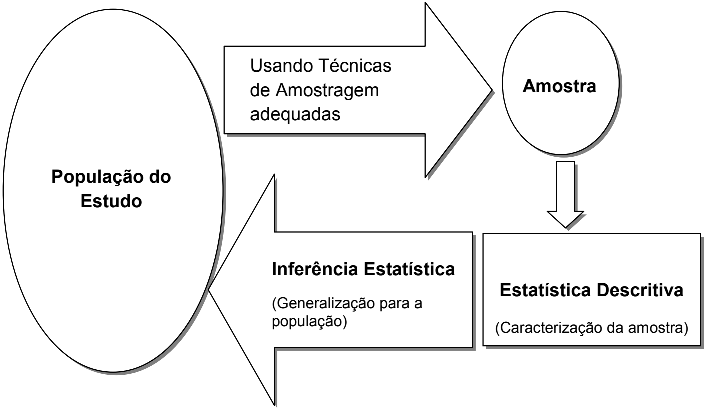
- População é o conjunto de todos os objetos cujas características pretendemos estudar e amostra é qualquer subconjunto finito da população.
- Usam-se medidas como a média e o desvio padrão para descrever amostras e populações.
- Quando as medidas se referem às características de uma amostra chamam-se estatísticas, e quando se referem às características da população chamam-se parâmetros.
- As estatísticas estimam o valor dos parâmetros que pretendemos determinar.

Amostra e valores observados

- Antes de uma amostra aleatória, de tamanho n, ser obtida, os seus elementos são considerados variáveis aleatórias,
$$X_1, X_2, ..., X_n,$$
que podem tomar qualquer um dos valores possíveis associados à característica em estudo, da população.
- Depois da amostra ser obtida, os valores observados das variáveis aleatórias $X_1, X_2, .., X_n$ (também chamados concretizações ou realizações da amostra) são designados por:
$$x_1, x_2, ..., x_n.$$

Uma amostra aleatória independente e identicamente distribuída (i.i.d.) é uma mostra em que as variáveis aletórias $X_1, X_2, ..., X_n$:

- são independentes;
- têm a mesma distribuição de probabilidade.


Usando Técnicas
de Amostragem
adequadas
Amostra
População do
Estudo
Inferência Estatística
(Generalização para a
população)
Estatística Descritiva
(Caracterização da amostra)

Estatísticas

- O ponto de partida de qualquer dos problemas de estimação que vamos estudar, é sempre uma amostra aleatória $X_1, X_2, ..., X_n$ da população.
- Dada uma amostra aleatória i.i.d., usamos funções da amostra, chamadas estatísticas, para fazer inferências sobre a população representada pela amostra, o que significa fazer inferências sobre o(s) parâmetro(s) da população em estudo.
- Como uma estatística é uma função de variáveis aleatórias, também é uma variável aleatória com uma certa distribuição de probabilidade.
- Sendo as estatísticas variáveis aleatórias, costumam representar-se por letras maiúsculas. Os valores que as estatísticas tomam são representados pelas correspondentes letras manísculas.

Distribuições amostrais

Média amostral

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n.

A média amostral é dada por:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

O valor calculado de $\overline{X}$, para os valores observados da amostra, representa-se por $\overline{x}$.

Variância amostral

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n. A variância amostral é dada por:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right].$$

O valor calculado de S^2, para os valores observados da amostra, representa-se por s^2.

Desvio padrão amostral

O desvio padrão amostral é a raíz quadrada positiva da variância amostral:

$$S = \sqrt{S^2}.$$

Proporção amostral

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de n observações associadas a um processo de Bernoulli, em que $X_i = 1$ (sucesso) ou $X_i = 0$ (insucesso), $i = 1, 2, ..., n$, consoante o elemento observado tem, ou não, a(s) característica(s) pretendida(s).

A proporção amostral é dada por:

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

e indica a proporção de sucessos da amostra.

O valor calculado de $\hat{P}$, para os valores observados da amostra, representa-se por $\hat{p}$.

Teorema do Limite Central (T.L.C.)

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória i.i.d. de tamanho n (isto é, n variáveis independentes e igualmente distribuídas), de uma população com média μ_X e variância σ_X^2 finitas. Então, se $\overline{X}$ é a média desta amostra, a função de distribuição da variável aleatória,

$$Z = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}.$$

tende, quando $n \to +\infty$, para a função de distribuição $N(0, 1)$.

$$Z \xrightarrow{\text{D}} N(0,1) \quad \text{ou} \quad \overline{X} \xrightarrow{\text{D}} N\left(\mu_X, \frac{\sigma_X^2}{n}\right).$$

(Convergência em distribuição).

- Obtém-se uma aproximação satisfatória se $n \geq 30$, considerando-se, neste caso, n suficientemente grande.

Se as observações são obtidas de uma população normal, então a distribuição da média amostral, $\overline{X}$, é exatamente normal, independentemente do tamanho da amostra.

Teorema

Seja $X_1, X_2, ..., X_n$ uma amostra de n observações independentes de uma população normal, com média μ_X e variância σ_X^2 e se $\overline{X}$ é a média desta amostra, então,

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

Exemplo 4.1: Considere que o tempo de viagem de autocarro entre o Porto e Coimbra segue uma distribuição Uniforme entre 100 a 120 minutos. Numa amostra aleatória de 30 viagens, qual a probabilidade de a média dos tempos de viagem ser inferior a 112 minutos?

Resolução: Seja X_i, $i = 1, 2, ..., 30$ a variável aleatória que representa o tempo da viagem i, de autocarro entre o Porto e Coimbra. Então,

$$X_i \sim U(100, 120).$$

Tem-se,

$$\mu_X = \frac{100 + 120}{2} = 110 \quad \text{e} \quad \sigma_X^2 = \frac{(120 - 100)^2}{12} = \frac{100}{3}.$$

A variável aleatória média amostral, $\overline{X}$, é dada por,

$$\overline{X} = \frac{X_1 + X_2 + ... + X_{30}}{30}.$$

Exemplo 4.1 (Cont.):

$\overline{X}$ representa a média dos 30 tempos de viagem. Pelo T.L.C., e como a população não é normal mas $n \geq 30$, tem-se

$$\overline{X} \xrightarrow{D} N\left(\mu_{\overline{X}} = \mu_X = 110, \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{30} = \frac{100}{3 \times 30}\right).$$

Logo, aproximadamente,

$$P(\overline{X} < 112) = 0.9711.$$

Distribuição da diferença entre médias amostrais de duas populações distintas

Sejam $\overline{X_1}$, $\overline{X_2}$ as médias de duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, obtidas de duas populações (discretas ou contínuas) com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2, respetivamente. Então, a função de distribuição da variável aleatória,

$$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tende, quando $n_1 \to +\infty$ e $n_2 \to +\infty$, para a função de distribuição normal estandardizada, $N(0, 1)$.

$$Z \xrightarrow{D} N(0,1) \quad \text{ou} \quad \overline{X_1} - \overline{X_2} \xrightarrow{D} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

(Convergência em distribuição)

Se as observações são obtidas de duas populações com distribuição normal, então a distribuição da diferença entre médias amostrais, $\overline{X_1} - \overline{X_2}$, é exatamente normal, independentemente do tamanho da amostra.

Teorema

Sejam $\overline{X_1}$, $\overline{X_2}$ as médias de duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, obtidas de duas populações com distribuição normal, com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2, respetivamente. Então,

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Exemplo 4.2: Considere o tempo (em horas) que uma pessoa passa, por dia, a ver televisão. Suponha que esse tempo é uma variável aleatória que, para um dado grupo etário, tem distribuição $N(\mu_1 = 3, \sigma_1^2 = 1)$, enquanto que para outro grupo etário tem distribuição $N(\mu_2 = 2, \sigma_2^2 = 1.5^2)$. Suponha ainda que se obteve uma amostra de cada grupo com tamanho $n_1 = 10$ e $n_2 = 20$, respetivamente. Calcule $P\left(\overline{X_1} - \overline{X_2} \geq 2\right)$.

Resolução:

Como ambas as populações são normais, tem-se,

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

ou seja, a diferença das médias é uma variável aleatória que segue uma distribuição normal.

Exemplo 4.2 (Cont.):

Sejam μ e σ^2 a média e a variância da distribuição que segue a variável aleatória $\overline{X_1} - \overline{X_2}$.

- $\mu = \mu_1 - \mu_2 = 3 - 2 = 1$;
- $\sigma^2 = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2} = \dfrac{1}{10} + \dfrac{1.5^2}{20} = 0.2125$

Então,

$$\overline{X_1} - \overline{X_2} \sim N(\mu = 1, \sigma^2 = 0.2125)$$

A probabilidade pedida é:

$$P\left(\overline{X_1} - \overline{X_2} \geq 2\right) = 1 - P\left(\overline{X_1} - \overline{X_2} < 2\right) = 0.015.$$

(Valor obtido por comando Python).

Distribuição da proporção amostral

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória i.i.d. de n observações de um processo de Bernoulli. Seja $\hat{P}$ a proporção amostral. A função de distribuição da variável aleatória,

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}},$$

tende, quando $n \to +\infty$, para a função de distribuição normal estandardizada, $N(0, 1)$.

$$Z \xrightarrow{D} N(0, 1), \quad \text{ou} \quad \hat{P} \xrightarrow{D} N\left(p, \frac{p(1-p)}{n}\right).$$

(Convergência em distribuição).

Quando o tamanho da amostra aumenta, a distribuição de $\hat{P}$ aproxima-se cada vez mais da distribuição normal.

Exemplo 4.3: Suponha que o Canal-Ideias reclama que 10% das residências o subscrevem, o que é verdade. No entanto, dada a sua reputação duvidosa, uma empresa de *marketing* resolveu estimar essa proporção, a partir de uma amostra de 100 residências, antes de renovar os seus contratos de publicidade com o Canal-Ideias.

- Determine a distribuição de probabilidade (aproximadamente) da proporção amostral $\hat{P}$.
- Supondo que os contratos só são renovados se a proporção amostral for superior a 8.5%, determine a probabilidade disso acontecer, usando o resultado aproximado da alínea anterior.

Resolução:

- Sejam $\mu = p = 0.1$ e $\sigma^2 = \dfrac{p(1-p)}{n} = 0.03^2$. A função de distribuição aproximada é:
$$\hat{P} \xrightarrow{\text{D}} N\left(\mu = 0.1, \sigma^2 = 0.03^2\right).$$
- $P(\hat{P} > 0.085) = 1 - P(\hat{P} \leq 0.085) = 0.691.$

Distribuição da diferença entre proporções amostrais de duas populações distintas

Considere duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2 (*suficientemente* grandes), obtidas de duas populações de Bernoulli.

Sejam X_1 e X_2, o número de sucessos em cada amostra (número de elementos com as características pretendidas). Seja

$$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2},$$

a diferença entre as proporções de sucesso (elementos com as características pretendidas) das duas amostras.

Distribuição da diferença entre proporções amostrais de duas populações distintas

Então, a função de distribuição da variável aleatória,

$$Z = \frac{\left(\hat{P}_1 - \hat{P}_2\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

tende, quando $n_1 \to +\infty$ e $n_2 \to +\infty$, para a função de distribuição normal estandardizada, $N(0,1)$.

$$Z \xrightarrow{\text{D}} N(0,1).$$

$$\hat{P}_1 - \hat{P}_2 \xrightarrow{\text{D}} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Exemplo 4.4: Em virtude de uma campanha publicitária, a proporção de consumidores que preferem uma determinada marca de café passou de $p_1 = 10\%$ para $p_2 = 12\%$. Suponha que se efetuaram duas sondagens, a primeira a 100 pessoas, antes da campanha se iniciar, e a segunda a 80 pessoas, depois da campanha terminar. Determine a probabilidade de $P\left(\hat{P}_2 - \hat{P}_1 > 0\right)$ e interprete o resultado obtido.

Resolução:

Considere-se,

- $\hat{P}_1$ v.a. que representa a proporção de clientes que preferem a marca do café antes da campanha.
- $\hat{P}_2$ v.a. que representa a proporção de clientes que preferem a marca do café depois da campanha.
- $\hat{P}_1$ e $\hat{P}_2$ resultam de amostras aleatórias retiradas de uma população de Bernoulli com parâmetros $p_1 = 0.10$ e $p_2 = 0.12$ e $n_1 = 100$ e $n_2 = 80$ suficientemente grandes.

Exemplo 4.4 (Cont.): Tem-se,

$$\hat{P}_1 \xrightarrow{D} N\left(\mu_1 = 0.10, \sigma_1^2 = \frac{0.10 \times 0.90}{100}\right).$$

$$\hat{P}_2 \xrightarrow{D} N\left(\mu_2 = 0.12, \sigma_2^2 = \frac{0.12 \times 0.88}{80}\right).$$

Então,

$$\hat{P}_2 - \hat{P}_1 \xrightarrow{D} N\left(\mu = 0.12 - 0.10, \sigma^2 = \frac{0.10 \times 0.90}{100} + \frac{0.12 \times 0.88}{80}\right)$$

$$\Longleftrightarrow$$

$$\hat{P}_2 - \hat{P}_1 \xrightarrow{D} N\left(\mu = 0.02, \sigma^2 = 0.047^2\right).$$

Assim,

$$P\left(\hat{P}_2 - \hat{P}_1 > 0\right) = 1 - P\left(\hat{P}_2 - \hat{P}_1 < 0\right) = 0.6644.$$

Estimação de parâmetros: intervalos de confiança

Parâmetro

Um parâmetro de uma população é uma constante θ, que é uma característica ou propriedade da população.

Estimador ou estatística

Um estimador ou estatística é uma função real das variáveis aleatórias que constituem a amostra, $\hat{\Theta} = G(X_1, ..., X_n)$, e, portanto, é também uma variável aleatória. As realizações desta v.a. fornecem aproximações para o parâmetro (desconhecido) da população, ou seja, é uma fórmula ou um processo que usa os valores da amostra para estimar um determinado parâmetro populacional.

Estimativa

Uma estimativa, $\hat{\theta}$ (valor calculado da estatística para uma dada amostra), é um valor específico, ou intervalo de valores, usado para aproximar o valor do parêmetro, θ, de uma população.

- Estimação pontual: produção de um valor, que se pretende que seja o melhor, para um determinado parâmetro da população, com base na informação amostral.
    - A estatística mais usada como medida de localização central é a média amostral $\overline{X}$ que estima a média populacional $E(X) = \mu$.
    - As estatísticas usadas para medir a variabilidade da amostra são a variância e o desvio padrão, S^2 e S, respetivamente, que são usadas para estimar a variância e o desvio-padrão populacional, $\text{var}(X) = \sigma^2$ e σ.
- Estimação intervalar: construção de um intervalo de confiança (IC) que, com certo grau de certeza previamente estipulado, contenha o verdadeiro valor do parâmetro da população. Em muitos casos, o intervalo é da forma $[\hat{\theta} - \epsilon, \hat{\theta} + \epsilon]$, sendo $\hat{\theta}$ uma estimativa para o parâmetro de interesse θ, e ϵ é considerado uma medida de precisão ou medida do erro inerente à estimativa $\hat{\theta}$. Usualmente, ϵ é designado por erro de estimativa ou margem de erro (absoluta). Desta forma, este método de estimação incorpora a confiança que se pode atribuir às estimativas.

Intervalo de confiança (IC)

Um intervalo de confiança (IC) de $(1-\alpha)\times 100\%$ para o parâmetro populacional θ (desconhecido) é um intervalo aleatório $[\hat{\Theta}_1, \hat{\Theta}_2]$, em que os limites de confiança $\hat{\Theta}_1$ e $\hat{\Theta}_2$ são duas estatísticas amostrais tais que:

$$P(\hat{\Theta}_1 \leq \theta \leq \hat{\Theta}_2) = 1-\alpha,$$

sendo:

- $1-\alpha$, o coeficiente (ou nível) de confiança;
- $\alpha \in]0,1[$, o nível de significância.

O coeficiente de confiança, $1-\alpha$, indica a proporção de vezes que os intervalos observados $[\hat{\theta}_1, \hat{\theta}_2]$ contêm o parâmetro θ.

Algumas considerações sobre intervalo de confiança

- Idealmente, um IC deverá ter amplitude pequena (grande precisão) e coeficiente de confiança elevado (probabilidade elevada de o IC conter o parâmetro desconhecido θ).
- Infelizmente, para um tamanho da amostra fixo, o coeficiente de confiança só pode aumentar, se a amplitude do intervalo também aumentar.
- Em geral, para valores do coeficiente de confiança elevados, a amplitude do IC aumenta rapidamente.
- Os valores mais típicos do coeficiente de confiança, $1-\alpha$, são, 0.99 ($\alpha = 0.01$), 0.95 ($\alpha = 0.05$), que é o valor mais comum, e 0.90 ($\alpha = 0.10$).

Intervalo de confiança para μ (σ^2 conhecida)

Pelo Teorema do Limite Central, sabemos que, para populações com variância σ^2 finita, quando as amostras aleatórias são independentes e identicamente distribuídas, a variável aleatória,

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

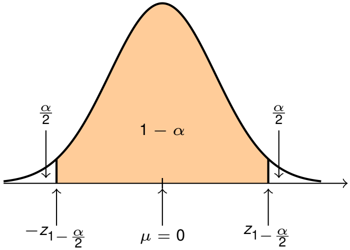
aproxima-se da distribuição normal estandardizada, quando o tamanho da amostra aumenta.

Vejamos como definir o intervalo de confiança de $(1 - \alpha) \times 100\%$ para a média μ, conhecendo σ^2.

Intervalo de confiança para μ (σ^2 conhecida)

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

- $100p\%$ das observações são menores que z_p,
- $P(Z < z_p) = p$.



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

$$\Leftrightarrow P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Intervalo de confiança para μ (σ^2 conhecida)

Seja $\overline{X}$ a média de uma amostra aleatória i.i.d., $X_1, X_2, ..., X_n$, de uma população com variância σ^2 conhecida. O intervalo aleatório,

$$IC_\mu = \left[\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right],$$

em que $z_{1-\frac{\alpha}{2}}$ é o percentil $100\left(1-\frac{\alpha}{2}\right)$ da distribuição normal $N(0,1)$, é um intervalo de confiança de $(1-\alpha)\times 100\%$ para a média populacional μ.

- $\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ são os limites de confiança;
- O erro de estimativa ou precisão é $z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$;
- A amplitude do IC é $2 \times z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

Intervalo de confiança para μ (σ^2 conhecida)

Depois da amostra ter sido realizada, substitui-se $\overline{X}$ por $\overline{x}$ e obtém-se o respetivo intervalo determinístico.

Para amostras relativamente pequenas de populações não normais (sobretudo as assimétricas) não podemos esperar que o grau de confiança seja próximo do indicado.

Para tamanhos de amostras aleatórias relativamente grandes (≥ 30), o Teorema do Limite Central sugere bons resultados.

Exemplo 4.5: O conteúdo médio de nicotina de uma amostra de dez cigarros de uma dada marca é 1.0 miligramas. O laboratório sabe, pela longa experiência neste tipo de análises, que o conteúdo de nicotina é uma variável aleatória aproximadamente normal com um desvio padrão de 0.15 miligramas. Determine intervalos de 90%, 95% e 99% de confiança para o conteúdo médio de nicotina.

Resolução:

Dados do problema:

- X é v. a. tal que, $X =$ quantidade de nicotina num cigarro;
- $X \sim N(\mu, \sigma_X^2 = 0.15^2)$;
- Numa amostra de $n = 10$ cigarros tem-se $\overline{x} = 1.0$g.

Pelo T.L.C., temos,

$$\overline{X} \sim N\left(\mu, \sigma_{\overline{X}}^2 = \frac{0.15^2}{10}\right).$$

Exemplo 4.5 (Cont.): Para os coeficientes de confiança, $(1-\alpha)\times 100\%$, de 90%, 95% e 99%, verifica-se:

- $1-\alpha = 0.90 \Leftrightarrow \alpha = 0.10 \Leftrightarrow 1-\frac{\alpha}{2} = 0.950.$
- $1-\alpha = 0.95 \Leftrightarrow \alpha = 0.05 \Leftrightarrow 1-\frac{\alpha}{2} = 0.975.$
- $1-\alpha = 0.99 \Leftrightarrow \alpha = 0.01 \Leftrightarrow 1-\frac{\alpha}{2} = 0.995.$

Assim, os intervalos de 90%, 95% e 99% de confiança são:

$$IC_\mu = \left[1.0 - z_{0.950}\frac{0.15}{\sqrt{10}}, 1.0 + z_{0.950}\frac{0.15}{\sqrt{10}}\right],$$

$$IC_\mu = \left[1.0 - z_{0.975}\frac{0.15}{\sqrt{10}}, 1.0 + z_{0.975}\frac{0.15}{\sqrt{10}}\right],$$

$$IC_\mu = \left[1.0 - z_{0.995}\frac{0.15}{\sqrt{10}}, 1.0 + z_{0.995}\frac{0.15}{\sqrt{10}}\right].$$

Falta calcular os valores de $z_{0.950}$, $z_{0.975}$ e $z_{0.995}$ e obter os intervalos de confiança.

Exemplo 4.5 (Cont.):

Comandos Python:

```
from scipy import stats
import numpy as np
conf = [0.90, 0.95, 0.99]
for value in conf:
    p =1 - (1-value)/2
    z =stats.norm.ppf(p, 0, 1)
    lim_inf = 1.0 − z ∗ 0.15/np.sqrt(10)
    lim_sup = 1.0 + z ∗ 0.15/np.sqrt(10)
    print(f'Para value∗100 : .0f%: z_p=z:.3f e o
            IC é [lim_inf:.3f, lim_sup:.3f]')
```

Output:

Para 90%: z_0.95 $= 1.645$ e o IC é $[0.922, 1.078]$

Para 95%: z_0.975 $= 1.960$ e o IC é $[0.907, 1.093]$

Para 99%: z_0.995 $= 2.576$ e o IC é $[0.878, 1.122]$

Intervalo de confiança para μ (σ^2 desconhecida)

No caso, mais comum, da variância da população σ^2 ser desconhecida, temos de recorrer à estatística S^2, variância amostral, e usar a variável aleatória,

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}.$$

Se a amostra aleatória é obtida de uma população normal, então,

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim T(n-1).$$

Vejamos como definir o intervalo de confiança de $(1-\alpha)\times 100\%$ para a média μ, desconhecendo σ^2.

Intervalo de confiança para μ (σ^2 desconhecida)

Sejam $\overline{X}$ e S^2 a média e a variância de uma amostra aleatória i.i.d., $X_1, X_2, ..., X_n$, de uma população normal com variância σ^2 desconhecida. O intervalo aleatório,

$$IC_\mu = \left[\overline{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right],$$

em que $t_{1-\frac{\alpha}{2}}$ é o percentil $100\left(1-\frac{\alpha}{2}\right)$ da distribuição $T(n-1)$, é um intervalo de confiança de $(1-\alpha) \times 100\%$ para a média populacional μ.

- $\overline{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ são os limites de confiança;
- O erro de estimativa ou precisão é $t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$;
- A amplitude do IC é $2 \times t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$.

Intervalo de confiança para μ (σ^2 desconhecida)

Se a população não for normal mas a amostra for suficientemente grande ($n \geq 30$):

$$IC_\mu = \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right].$$

Para ambos os casos, depois da amostra ter sido realizada, substitui-se $\overline{X}$ por $\overline{x}$ e S por s e obtém-se o respetivo intervalo determinístico.

IC a $(1-\alpha)\times 100\%$ para a média μ:

Sejam X_1, X_2,...,X_n uma amostra aleatória i.i.d. de uma população com média μ e variância σ^2. Sejam $\overline{X}$ e S^2 a média e a variância da amostra aleatória.

σ^2 conhecida:

$X \sim N(\mu, \sigma^2)$	$n \geq 30$	Estatística de teste	IC
Sim	Indiferente	$Z = \dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$	$\left[\overline{X} - z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right]$
Não	Sim	$Z = \dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0,1)$	$\left[\overline{X} - z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right]$

IC a $(1-\alpha)\times 100\%$ para a média μ:

Sejam X_1, X_2,...,X_n uma amostra aleatória i.i.d. de uma população com média μ e variância σ^2. Sejam $\overline{X}$ e S^2 a média e a variância da amostra aleatória.

σ^2 desconhecida:

$X \sim N(\mu, \sigma^2)$	$n \geq 30$	Estatística de teste	IC
Sim	Indiferente	$T = \dfrac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim T(n-1)$	$\left[\overline{X} - t_{1-\frac{\alpha}{2}}\dfrac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}}\dfrac{s}{\sqrt{n}}\right]$
Não	Sim	$Z = \dfrac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \xrightarrow{D} N(0,1)$	$\left[\overline{X} - z_{1-\frac{\alpha}{2}}\dfrac{S}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\dfrac{S}{\sqrt{n}}\right]$

Tamanho adequado para a amostra

- A precisão do intervalo de confiança de $(1-\alpha)\times 100\%$ para a média é metade da seu amplitude (semiamplitude do IC), ou seja, $z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, $t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$ ou $z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$.
- Para efetuar a amostragem, pode estimar-se, com um grau de confiança de $(1-\alpha)\times 100\%$ dado, o tamanho n da amostra que garante que o erro máximo cometido (precisão), não ultrapassa um valor ϵ desejado.
- Consoante o caso, resolvemos a inequação:
$$z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\leq\epsilon, \quad t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\leq\epsilon \quad \text{ou} \quad z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\leq\epsilon,$$
em ordem a n obtendo-se, respetivamente,
$$n\geq\left(\frac{z_{1-\frac{\alpha}{2}}\sigma}{\epsilon}\right)^2, \quad n\geq\left(\frac{t_{1-\frac{\alpha}{2}}s}{\epsilon}\right)^2 \quad \text{ou} \quad n\geq\left(\frac{z_{1-\frac{\alpha}{2}}s}{\epsilon}\right)^2,$$
pelo que basta tomar para n o menor inteiro que satisfaz a desigualdade.

Exemplo 4.6: Uma cordoaria, depois de efetuar alterações no processo de fabrico, testou uma amostra de 64 cordas, tendo obtido uma resistência média $\overline{x} = 300$N e um desvio padrão de $s = 24$N.

- Compare os intervalos de confiança a 95% para a resistência média μ desse tipo de cordas, quando se usa a distribuição t-Student ou a aproximação à normal, respetivamente.
- Qual deve ser o tamanho da amostra, se pretendermos estimar da resistência média com um erro de estimativa inferior a 3N e um nível de confiança de 95%?

Resolução: Sejam X, X_i e $\overline{X}$ v.a. tais que ($i = 1, ..., 64$):

- $X =$ resistência das cordas da cordoaria de média μ e desvio padrão s (distribuição e desvio padrão desconhecidos);
- $X_i =$ resistência da corda i de uma a.a. de $n = 64$;
- $\overline{X} = \dfrac{X_1 + ... + X_{64}}{64}$ representa a média amostral.

Exemplo 4.6 (Cont.): Verifica-se:

$$1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025 \Rightarrow 1-\frac{\alpha}{2} = 0.975$$

IC_μ a 95% para a resistência média:

Distribuição t-Student	Aproximação à normal
$[\overline{x} - t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{x} + t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}]$	$[\overline{x} - z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}]$
$[300 - t_{0.975}\frac{24}{\sqrt{65}}, 300 + t_{0.975}\frac{24}{\sqrt{64}}]$	$[300 - z_{0.975}\frac{24}{\sqrt{64}}, 300 + z_{0.975}\frac{24}{\sqrt{64}}]$
$\downarrow$	$\downarrow$
$[294.0, 306.0]$	$[294.1, 305.9]$

Exemplo 4.6 (Cont.):

Comandos Python:

```
from scipy import stats, import numpy as np
média_amostral = 300, desvio_padrão_amostral = 24
conf = 0.95, n= 64, alfa = 1−conf, z = 1 − (alfa)/2
t_a =stats.t.ppf(z, n - 1), z_a =stats.norm.ppf(z, 0, 1)
# t-Student
lim_inf = 300 − t_a ∗ 24/np.sqrt(64)
lim_sup = 300 + t_a ∗ 24/np.sqrt(64)
print(f'Para a distribuição t-Student IC é [lim_inf:.1f, lim_sup:.1f]')
# Aproximação à normal
lim_inf = 300 − z_a ∗ 24/np.sqrt(64)
lim_sup = 300 + z_a ∗ 24/np.sqrt(64)
print(f'Para a aproximação à normal IC é [lim_inf:.1f, lim_sup:.1f]')
```

Output:

Para a distribuição t-Student IC é $[294.0, 306.0]$,

Para a aproximação à normal IC é $[294.1, 305.9]$.

Exemplo 4.6 (Cont.):

Para estimar $\overline{X}$ com $\epsilon < 3$ e nível confiança de 95% $\Rightarrow \alpha = 0.05$.

$$\epsilon < 3 \Leftrightarrow z_{0.975} \times \frac{24}{\sqrt{n}} < 3 \Rightarrow$$

$$\Rightarrow n > \left(\frac{z_{0.975} \times 24}{3}\right)^2 \Leftrightarrow n > 245.8.$$

A amostra deve conter 246 peças.

Intervalo de confiança para $\mu_1 - \mu_2$ (σ_1^2 e σ_2^2 conhecidas)

Sejam $\overline{X_1}$ e $\overline{X_2}$ as médias de duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, de duas populações com médias desconhecidas μ_1 e μ_2 e variânias conhecidas σ_1^2 e σ_2^2, respetivamente.

A variável aleatória,

$$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \xrightarrow{D} N(0, 1),$$

aproxima-se da normal estandardizada, se o tamanho das amostras aumentam.

Intervalo de confiança para $\mu_1 - \mu_2$ (σ_1^2 e σ_2^2 conhecidas)

O intervalo de confiança de $(1-\alpha)\times 100\%$, para a diferença de médias $\mu_1 - \mu_2$, é dado por:

$$IC_{\mu_1-\mu_2} = \left[\left(\overline{X_1} - \overline{X_2}\right) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \left(\overline{X_1} - \overline{X_2}\right) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

em que $z_{1-\frac{\alpha}{2}}$ é o percentil $100\left(1-\frac{\alpha}{2}\right)$ da distribuição normal $N(0,1)$.

- $\left(\overline{X_1} - \overline{X_2}\right) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ são os limites de confiança;
- O erro de estimativa ou precisão é $z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$;
- A amplitude do IC é duas vezes o erro de estimativa.

Intervalo de confiança para $\mu_1 - \mu_2$ (σ_1^2 e σ_2^2 conhecidas)

Depois da amostra ter sido realizada, substitui-se $\overline{X_1}$ e $\overline{X_2}$ por $\overline{x_1}$ e $\overline{x_2}$, respetivamente, e obtém-se o corresondente intervalo determinístico.

O coeficiente de confiança é exato para populações normais, mas é aproximado para populações não normais.

Exemplo 4.7: A Eletricidade do Oriente, antes de fazer novos investimentos, resolveu estimar a evolução do consumo de eletricidade no último ano. Para isso, selecionou duas amostras aleatórias i.i.d. e mutuamente independentes de $n_1 = 120$ e $n_2 = 150$ consumidores domésticos, para estimar o consumo médio de eletricidade, por habitação, em janeiro do ano passado e do ano corrente, respetivamente. Os resultados obtidos, em quilowatt-hora, foram de $\overline{x_1} = 550$ e $\overline{x_2} = 567$. Supondo que o desvio padrão do consumo por habitação, em janeiro de ambos os anos, era conhecido, $\sigma_1 = \sigma_2 = 110$, determine o intervalo de 95% de confiança para a evolução (diferença) do consumo médio de eletricidade, $\mu_2 - \mu_1$.

Resolução: Sejam:

- $\overline{X_1}$ a variável aleatória que representa a média amostral dos consumos do último ano, com $\overline{x_1} = 550$.
- $\overline{X_2}$ a variável aleatória que representa a média amostral dos consumos do corrente ano, com $\overline{x_2} = 567$.

Exemplo 4.7 (Cont.): Tem-se:

$$\overline{X_1} \xrightarrow{D} N\left(\mu_1, \frac{110^2}{120}\right), \overline{X_2} \xrightarrow{D} N\left(\mu_2, \frac{110^2}{150}\right),$$

$$\overline{X_2} - \overline{X_1} \xrightarrow{D} N\left(\mu_2 - \mu_1, \frac{110^2}{120} + \frac{110^2}{150}\right)$$

$$IC_{\mu_2-\mu_1} = \left[17 - z_{0.975}\sqrt{\frac{110^2}{120} + \frac{110^2}{150}}, 17 + z_{0.975}\sqrt{\frac{110^2}{120} + \frac{110^2}{150}}\right] =$$

$$= [-9.4, 43.4].$$

Este resultado não garante (com 95% de confiança) que tenha havido uma evolução positiva do consumo, visto que admite valores negativos para a diferença $\mu_2 - \mu_1$. Assim, é aconselhável, antes de realizar novos investimentos, proceder a um estudo com amostras maiores, para reduzir o erro de amostragem.

Intervalo de confiança para $\mu_1 - \mu_2$ ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ desconhecidas)

Sejam $\overline{X_1}$ e $\overline{X_2}$ as médias de duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, de duas populações normais com médias desconhecidas μ_1 e μ_2 e variânias também desconhecidas $\sigma_1^2 = \sigma_2^2 = \sigma^2$, respetivamente.

A variável aleatória,

$$T = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}},$$

segue uma distribuição t-Student com $n_1 + n_2 - 2$ graus de liberdade, ou seja,

$$T \sim T(n_1 + n_2 - 2).$$

Intervalo de confiança para $\mu_1 - \mu_2$ ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ desconhecidas)

Sejam $\overline{X_1}$ e $\overline{X_2}$ as médias e S_1 e S_2 as variâncias de duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, de duas populações normais com médias desconhecidas μ_1 e μ_2 e variânias também desconhecidas $\sigma_1^2 = \sigma_2^2 = \sigma^2$, respetivamente. Após a realização da amostragem, o IC com $(1-\alpha) \times 100\%$ de confiança é

$$IC_{\mu_1-\mu_2} = \left[(\overline{x_1} - \overline{x_2}) - t_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times s_c, (\overline{x_1} - \overline{x_2}) + t_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times s_c \right]$$

em que $t_{1-\frac{\alpha}{2}}$ é o percentil $100(1-\frac{\alpha}{2})$ da distribuição $T(n_1+n_2-2)$ e s_c é o estimador combinado para σ dado por:

$$s_c = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

sendo s_1 e s_2 as estimativas de S_1 e S_2, respetivamente.

Para amostras suficientemente grandes, pode-se substituir a distribuição t-Student pela distribuição normal e estender a populações não normais.

Intervalo de confiança para $\mu_1 - \mu_2$ ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ desconhecidas e amostras grandes)

Sejam $\overline{X_1}$ e $\overline{X_2}$ as médias e S_1 e S_2 as variâncias (com estimativas $\overline{x_1}$, $\overline{x_2}$, s_1 e s_2) de duas a.a., i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, suficientemente grandes, de duas populações com médias e variâncias desconhecidas. O IC determinístico com $(1-\alpha)\times 100\%$ de confiança é:

$$IC_{\mu_1-\mu_2} = \left[(\overline{x_1} - \overline{x_2}) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times s_c, (\overline{x_1} - \overline{x_2}) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times s_c \right]$$

em que $z_{1-\frac{\alpha}{2}}$ é o percentil $100(1-\frac{\alpha}{2})$ da distribuição $N(0,1)$ e s_c é o estimador combinado para σ dado por:

$$s_c = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}.$$

Exemplo 4.8: Numa experiência para comparar dois novos analgésicos, 65 doentes voluntários, depois de submetidos ao mesmo tipo de cirurgia, foram divididos em dois grupos de $n_1 = 35$ e $n_2 = 30$ pessoas, a quem foi ministrada uma dose equivalente dos analgésicos 1 e 2, respetivamente. No primeiro grupo, a ausência de dor durou, em média, $\overline{x_1} = 6.3$ horas, com um desvio padrão de $s_1 = 1.2$ horas, enquanto que no segundo grupo, $\overline{x_2} = 5.2$ horas e $s_2 = 1.4$ horas. Supondo que as populações são aproximadamente normais e têm a mesma variância, determine um intervalo de 95% de confiança para a diferença da duração média do efeito dos analgésicos, $\mu_1 - \mu_2$.

Resolução: Sejam $\overline{X_1}$ e $\overline{X_2}$ duas v.a. tais que:

- $\overline{X_1}$ = duração média, em horas, de ausência de dor com o analgésico 1. Para uma amostra com $n_1 = 35$, observou-se, $\overline{x_1} = 6.3$h e $s_1 = 1.2$h;
- $\overline{X_2}$ = duração média, em horas, de ausência de dor com o analgésico 2. Para uma amostra com $n_2 = 30$, observou-se $\overline{x_2} = 5.2$h e $s_2 = 1.4$h.

Exemplo 4.8 (Cont.):

$$T = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{35} + \frac{1}{30}} \times \sqrt{\frac{34 \times S_1^2 + 29 \times S_2^2}{63}}} \sim T(63),$$

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow 1 - \frac{\alpha}{2} = 0.975 \Rightarrow t_{0.975} = 1.998.$

Temos $\overline{x_1} = 6.3$h, $\overline{x_2} = 5.2$h, $s_1 = 1.2$h e $s_2 = 1.4$h como estimativas de $\overline{X_1}$, $\overline{X_2}$, S_1 e S_2, respetivamente.

Os limites do IC de 95% de confiança para $\mu_1 - \mu_2$ são:

$$(6.3 - 5.2) \pm t_{0.975}\sqrt{\frac{1}{35} + \frac{1}{30}} \times \sqrt{\frac{34 \times 1.2^2 + 29 \times 1.4^2}{63}} = 1.1 \pm 0.644$$

e portanto, o IC de 95% é $[0.44, 1.76]$. Assim, com um nível de confiança de 95%, não temos evidência estatística para concluir que o analgésico 1 não tem um efeito mais duradouro do que o 2.

Intervalo de confiança para $\mu_1 - \mu_2$ ($\sigma_1^2 \neq \sigma_2^2$ desconhecidas e amostras grandes)

Sejam $\overline{X_1}$ e $\overline{X_2}$ as médias e S_1 e S_2 as variâncias (com estimativas $\overline{x_1}$, $\overline{x_2}$, s_1 e s_2, respetivamente) de duas a.a., i.i.d., mutuamente independentes, de tamanhos n_1 e n_2, suficientemente grandes, de duas populações com médias μ_1 e μ_2 desconhecidas e variâncias $\sigma_1^2 \neq \sigma_2^2$ também desconhecidas. O IC determinístico para $\mu_1 - \mu_2$ com $(1-\alpha) \times 100\%$ de confiança é:

$$IC_{\mu_1-\mu_2} = \left[(\overline{x_1} - \overline{x_2}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{x_1} - \overline{x_2}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

em que $z_{1-\frac{\alpha}{2}}$ é o percentil $100(1-\frac{\alpha}{2})$ da distribuição $N(0,1)$.

IC a $(1-\alpha)\times 100\%$ para $\mu_X - \mu_Y$: Sejam X e Y duas amostra aleatórias i.i.d. de duas populações com médias μ_X e μ_Y e variâncias σ_1^2 e σ_2^2, respetivamente. Sejam $\overline{X}$, $\overline{Y}$, S_X^2 e S_Y^2 as médias e as variâncias respetivas das amostras aleatórias.

σ_X^2 e σ_Y^2 conhecidas

Polu. normais	$n_X \geq 30$ $n_Y \geq 30$	Estatística de teste	Limites do IC
Sim	Indif.	$Z = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \sim N(0,1)$	$(\overline{X}-\overline{Y}) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$
Não	Sim	$Z = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \xrightarrow{D} N(0,1)$	$(\overline{X}-\overline{Y}) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$

IC a $(1-\alpha)\times 100\%$ para $\mu_X - \mu_Y$: Sejam X e Y duas amostra aleatórias i.i.d. de duas populações com médias μ_X e μ_Y e variâncias σ_1^2 e σ_2^2, respetivamente. Sejam $\overline{X}$, $\overline{Y}$, S_X^2 e S_Y^2 as médias e as variâncias respetivas das amostras aleatórias.

$\sigma_X^2 = \sigma_Y^2$ desconhecidas

Polu. normais	$n_X \geq 30$ $n_Y \geq 30$	Estatística de teste	Limites do IC
Sim	Indif.	$T = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim T(n_X+n_Y-2)$	$(\overline{X}-\overline{Y}) \pm t_{1-\frac{\alpha}{2}}\, S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}$
Não	Sim	$Z = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \xrightarrow{D} N(0,1)$	$(\overline{X}-\overline{Y}) \pm z_{1-\frac{\alpha}{2}}\, S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}$
			com $S = \sqrt{\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}}$

IC a $(1-\alpha)\times 100\%$ para $\mu_X - \mu_Y$: Sejam X e Y duas amostra aleatórias i.i.d. de duas populações com médias μ_X e μ_Y e variâncias σ_1^2 e σ_2^2, respetivamente. Sejam $\overline{X}$, $\overline{Y}$, S_X^2 e S_Y^2 as médias e as variâncias respetivas das amostras aleatórias.

$\sigma_X^2 \neq \sigma_Y^2$ desconhecidas

Polu. normais	$n_X \geq 30$ $n_Y \geq 30$	Estatística de teste	Limites do IC
Indif.	Sim	$Z = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{S_X^2}{n_X}+\dfrac{S_Y^2}{n_Y}}} \xrightarrow{D} N(0,1)$	$(\overline{X}-\overline{Y}) \pm z_{1-\frac{\alpha}{2}}\sqrt{\dfrac{S_X^2}{n_X}+\dfrac{S_Y^2}{n_Y}}$

Intervalo de confiança para uma proporção

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

$$\hat{P} = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{D} N\left(p, \frac{p(1-p)}{n}\right) \Leftrightarrow$$

$$\Leftrightarrow Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{D} N(0, 1)$$

Intervalo de confiança para uma proporção

O IC, de amplitude mínima, é obtido resolvendo:

$$P\left(-z_{1-\frac{\alpha}{2}} < \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(\hat{P} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha,$$

sendo $z_{1-\frac{\alpha}{2}}$ o percentil da distribuição $N(0,1)$.

Após realizada a amostragem, substitui-se $\hat{P}$ por $\hat{p}$ e obtém-se o IC, de amplitude mínima, para uma proporção.

Intervalo de confiança para uma proporção

O *IC* a $(1-\alpha)\times 100\%$ para a proporção p de uma população é:

$$IC_p = \left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

, sendo $z_{1-\frac{\alpha}{2}}$ o percentil $100\left(1-\frac{\alpha}{2}\right)$ da distribuição $N(0,1)$.

Exemplo 4.9: Realizou-se um inquérito telefónico a 50 pessoas para estimar a proporção da população de um país favorável a uma reforma fiscal, tendo 84% dessas pessoas manifestando-se favorável à reforma fiscal.

- Determine um intervalo de confiança a 95% para a proporção da população favorável a essa reforma fiscal.
- Supondo que o resultado obtido na sondagem seria o mesmo, determine o tamanho da amostra que garanta, com um grau de confiança de 95%, que o erro máximo cometido seja inferior a 0.05.

Resolução:

$\hat{P}$ é v.a. que representa a proporção de pessoas favoráveis à reforma fiscal.

$\hat{P} \xrightarrow{D} N\left(p, \dfrac{p(1-p)}{n}\right)$ e $n = 50$.

Uma estimativa de p associada ao estimador $\hat{P}$ é $\hat{p} = 0.84$.

Exemplo 4.9 (Cont.):

- O *IC* a 95% para a proporção da população é dado por:

$$IC_p = \left[0.84 - z_{0.975} \times \sqrt{\frac{0.84 \times 0.16}{50}}, 0.84 + z_{0.975} \times \sqrt{\frac{0.84 \times 0.16}{50}}\right] =$$
$$= [0.74, 0.94].$$

Pode afirmar-se que, com uma confiança de 95%, que não há evidência estatística para não afirmar que a maioria das pessoas é favorável à reforma fiscal.

- O tamanho da amostra deveria ser:

$$\epsilon < 0.05 \Leftrightarrow z_{0.975} \times \sqrt{\frac{0.84 \times 0.16}{n}} < 0.05 \Leftrightarrow$$

$$\Leftrightarrow n > \left(\frac{z_{0.975} \times \sqrt{0.84 \times 0.16}}{0.05}\right)^2 \Leftrightarrow n > 206.5.$$

A amostra teria de conter 207 inquiridos.

Intervalo de confiança para a diferença de proporções, $p_1 - p_2$

Se $\hat{P}_1$ e $\hat{P}_2$ são as proporções de sucessos de duas amostras aleatórias i.i.d., mutuamente independentes, de tamanhos n_1 e n_2 suficientemente grandes, de duas populações quaisquer, o intervalo de confiançade $(1-\alpha)\times 100\%$, para $p_1 - p_2$, após a amostragem, é o intervalo determinístico, dado por:

$$IC_{p_1-p_2} = \left[(\hat{p_1} - \hat{p_2})) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}, (\hat{p_1} - \hat{p_2})) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right]$$

,

sendo $z_{1-\frac{\alpha}{2}}$ o percentil $100\left(1-\frac{\alpha}{2}\right)$ da distribuição $N(0,1)$.

Observação: O coeficiente de confiança é aproximado.