Estatística Aplicada

Rita Gaio
Departamento de Matemática - FCUP
argaio@fc.up.pt

June 1, 2020

Índice Geral

Chapter 1

Intervalos de confiança

1.1. Intervalos de confiança

Seja X uma variável aleatória com média e variância designadas, respectivamente, por

$$E(X) = \mu \quad \text{e} \quad \mathrm{Var}(X) = \sigma^2.$$

Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de tamanho n da variável aleatória X e $x_1, \ldots, x_n$ é uma realização dessa amostra aleatória, definimos os valores numéricos da média e da variância amostrais associadas a esta realização por

$$\begin{aligned} \overline{x} &= \overline{X}(x_1, \ldots, x_n) \\ &= \frac{1}{n}(x_1 + \cdots + x_n) \\ s^2 &= S^2(x_1, \ldots, x_n) \\ &= \frac{1}{n-1}[(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2] \end{aligned}$$

respectivamente. Sempre que estiver implícito que estamos a trabalhar com a realização da amostra, usaremos a notação

$$\overline{x} \quad \text{e} \quad s^2$$

para designar $\overline{X}(x_1, \ldots, x_n)$ e $S^2(x_1, \ldots, x_n)$, respectivamente. Note-se que estas quantidades são *números* e não devem ser confundidas com as *variáveis aleatórias* $\overline{X}$ e S^2.

Suponhamos que se sabe que a função densidade de probabilidade f de X pertence a uma família $f(x; \theta)$ de densidades de probabilidade que dependem, para completa especificação, de um parâmetro $\theta \in \Theta$ (θ pode ser um único parâmetro ou um vector de parâmetros $\theta = (\theta_1, \ldots, \theta_k)$).
Podemos colocar o problema de usar os valores observados de uma amostra aleatória da população para:

- estimar o valor "mais provável" do parâmetro θ da distribuição que gerou as observações (teoria sobre **estimação pontual** ou sobre **testes de hipóteses**)

- estimar θ através de uma região que, de alguma forma, transmita informação sobre a provável magnitude de θ (teoria sobre **intervalos de confiança**).

De seguida, tomamos θ igual à média da população e deduzimos pormenorizadamente um intervalo *aleatório* que contenha essa média com um determinado nível de confiança pré-especificado. Só abordaremos intervalos *centrados*.
Intervalos de confiança para os restantes parâmetros populacionais serão apresentados de seguida. Não se inclui a dedução dessas fórmulas por resultarem das estatísticas amostrais correspondentes de forma absolutamente análoga ao que foi feito para a média.

- **Intervalo de confiança para a média - σ^2 conhecida**

  Como vimos, se $X \sim N(\mu, \sigma^2)$ a média amostral segue também uma distribuição normal, independentemente do tamanho da amostra; se X não segue uma distribuição normal mas o tamanho n das amostras é suficientemente grande ($n \geq 30$) então a distribuição da média amostral é assimptoticamente normal. Para cada uma das situações tem-se

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

  isto é

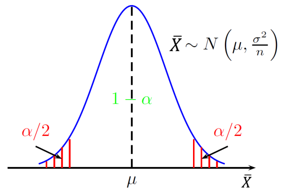
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{(resp. assimpt.).}$$

  Fixemos agora um número α com $0 < \alpha < 1$. O problema que se coloca é o da determinação de um intervalo que contenha a média da população com $100(1-\alpha)\%$ de confiança.

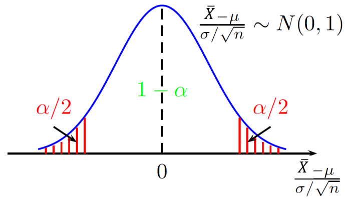
  Como a distribuição normal é simétrica, é usual (mas não obrigatório) pensar-se em intervalos centrados em μ. É dessa situação que vamos tratar em seguida.

  Começamos por dividir α de igual forma por entre as duas caudas da distribuição de $\overline{X}$, definindo um intervalo $[\mu_L, \mu_U]$ que contém $100(1-\alpha)\%$ de todos os possíveis valores de $\overline{X}$:

$$P(\mu_L < \overline{X} < \mu_U) = 1 - \alpha.$$



Equivalentemente,



Resolvendo a inequação

$$N_{\alpha/2}(0,1) \leq \frac{\overline{X} - \mu}{\sigma\sqrt{n}} \leq N_{1-\alpha/2}(0,1)$$

em ordem a μ, obtem-se

$$\frac{\sigma}{\sqrt{n}} N_{\alpha/2}(0,1) \leq \overline{X} - \mu \leq \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1)$$

donde

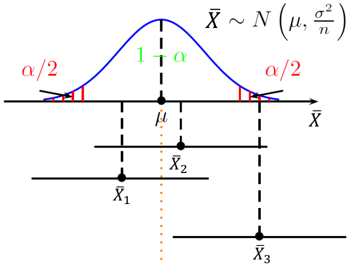
$$\overline{X} - \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1) \leq \mu \leq \overline{X} - \frac{\sigma}{\sqrt{n}} N_{\alpha/2}(0,1)$$

ou, dado que $N_{\alpha/2}(0,1) = -N_{1-\alpha/2}(0,1)$,

$$\overline{X} - \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}.^{a} \tag{1.1}$$

As desigualdades (1.1) representam um **intervalo aleatório** pois $\overline{X}$ é uma variável aleatória e como tal pode tomar uma infinidade de valores! A questão que se coloca é a de passar do intervalo aleatório para um intervalo concreto.

Suponhamos que temos uma amostra de tamanho n de X. Para essa amostra, calculamos (o escalar) $\overline{x}$ e o intervalo correspondente a (1.1). Em seguida, extraímos outra amostra e voltamos a calcular o intervalo correspondente a (1.1). E fazemos o mesmo noutra amostra. E noutra amostra...



O número de amostras que podemos extrair da população (com reposição) é infinito, podendo portanto encontrar um número infinito de concretizações de intervalos do tipo (1.1), todos com a mesma amplitude.

[a] Para simplificação da notação, representamos $N_q(0,1)$ apenas por N_q, para qualquer $q \in [0,1]$.

Dependendo do valor de $\overline{X}$ na amostra, alguns desses intervalos conterão a média μ da população, outros não.

Dada uma amostra, o intervalo (concreto)

$$IC_{100(1-\alpha)\%} = \left[\overline{x} - \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1), \overline{x} + \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1)\right]$$

diz-se o **intervalo de confiança a $100(1-\alpha)\%$ para a média**, e é representado por $IC_{100(1-\alpha)\%}$.

Tecem-se agora as seguintes considerações:

- a probabilidade de o intervalo aleatório (1.1) conter a média da população é igual a $1-\alpha$
- a média da população pode ou não estar contida num intervalo de confiança (concreto)
- não se pode afimar que um intervalo de confiança a $100(1-\alpha)\%$ contém a média com probabilidade $1-\alpha$
- de entre todos os intervalos de confiança possíveis de construir a partir de amostras aleatórias (com reposição) de tamanho n, exactamente $100(1-\alpha)\%$ desses intervalos contêm a média μ da população.
- utilizando um argumento frequencista, de entre N intervalos de confiança a 95% para a média, *aproximadamente* $N \times 0.95$ desses intervalos conterão a média da população

- qualquer intervalo a 99% de confiança merece mais **confiança**, no que diz respeito a cobrir μ, do que qualquer outro intervalo a 95% de confiança
- não se conhece a **probabilidade** com que um determinado intervalo de confiança para a média contém μ (apesar dessa probabilidade só poder ser 0 ou 1).
- em termos práticos, para facilitar a interpretação mas usando linguagem um pouco imprecisa, diz-se que dado um intervalo de confiança para a média a $100(1-\alpha)\%$, temos $100(1-\alpha)\%$ de ***confiança*** que a média da população esteja contida algures nesse intervalo.

- **Intervalo de confiança para a média - σ^2 desconhecida**

No caso em que a variância da população não é conhecida, tem de ser trabalhar com a variância amostral

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

Sabe-se já que, se $X \sim N(\mu, \sigma^2)$, então

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1). \tag{1.2}$$

Repetindo o argumento do caso anterior, resolvemos a inequação

$$t_{\alpha/2}(n-1) \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{1-\alpha/2}(n-1)$$

em ordem a μ para encontrar o intervalo aleatório que contém a média μ de X com uma probabilidade de $100(1-\alpha)\%$:

$$\overline{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) < \mu < \overline{X} + \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1).$$

Para uma amostra, o intervalo de confiança para a média na situação em que a variância é desconhecida é então

$$\begin{aligned} & IC_{100(1-\alpha)\%} \\ &= \left[\overline{x} + \frac{s}{\sqrt{n}} t_{\alpha/2}(n-1), \overline{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right] \\ &= \left[\overline{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1), \overline{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right] \end{aligned}$$

A igualdade entre os dois intervalos anteriores resulta da simetria da distribuição t.

Se a distribuição não for normal mas não apresentar assimetrias *exageradas* e n for suficientemente grande ($n \geq 30$ [a]), então $t(n-1)$ converge assimptótica para $N(0,1)$. O intervalo passa então a ser

$$\left[\overline{x} - \frac{s}{\sqrt{n}} N_{1-\alpha/2}(0,1), \overline{x} + \frac{s}{\sqrt{n}} N_{1-\alpha/2}(0,1)\right].$$

[a] Há autores que usam $n > 200$; por ex., Rosner, Fundamentals of Biostatistics. Essencialmente dependerá das características da distribuição empírica.

Mais uma vez, de entre todos os intervalos de confiança possíveis de construir a partir de amostras da população, com reposição e com o mesmo tamanho, $100(1-\alpha)\%$ deles conterão a média μ da população.

Observação: A amplitude do intervalo de confiança para a média reflecte a precisão com que $\overline{x}$ estima (pontualmente) μ e depende dos valores de n, s e α. Têm-se as seguintes considerações:

- quanto maior for o tamanho da amostra menor será a amplitude do intervalo (uma vez fixados s e α). De facto, quantos mais indivíduos conseguirmos amostrar, maior será a probabilidade de a amostra ser semelhante à população e consequentemente maior será a semelhança entre o parâmetro amostral e o populacional.
- quanto maior for s maior será a amplitude do intervalo de confiança. Isto relaciona-se com o facto de o desvio-padrão (amostral) reflectir a variabilidade da distribuição empírica.
- quanto maior for α (menor será $1-\alpha$), menor será a amplitude do intervalo de confiança.

- **Intervalo de confiança para a proporção**
Para $X_i \sim B(1,\pi)$, $i=1,\ldots,n$, seja

$$X=\sum_{i=1}^{n} X_i,$$

a v.a. que conta o número de sucessos em n experiências i.i.d.. Pelo teorema de De Moivre-Laplace, se n é suficientemente grande ($n \geq 30$ e $n\pi(1-\pi)>5$) então

$$X \sim B(n,\pi) \overset{a}{\sim} N\left(n\pi, n\pi(1-\pi)\right).$$

Para a proporção amostral

$$p=X/n$$

ter-se-á portanto

$$\frac{P-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \overset{a}{\sim} N(0,1).$$

Em particular, note-se também que $E(P)=\pi$. Pelo raciocínio seguido nos casos anteriores, o intervalo aleatório centrado no qual, com probabilidade $1-\alpha$, está a verdadeira proporção π da população é conseguido resolvendo a inequação em ordem a π

$$N_{\alpha/2}(0,1) \leq \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq N_{1-\alpha/2}(0,1)$$

e estimando π por p em $\sqrt{\frac{\pi(1-\pi)}{n}}$.

Para uma amostra concreta, obtem-se

$$
\begin{aligned}
&IC_{100(1-\alpha)\%}\\
&= \left[p - \sqrt{\frac{p(1-p)}{n}}N_{\alpha/2}, p + \sqrt{\frac{p(1-p)}{n}}N_{1-\alpha/2}\right]\\
&= \left[p - \sqrt{\frac{p(1-p)}{n}}N_{1-\alpha/2}, p + \sqrt{\frac{p(1-p)}{n}}N_{1-\alpha/2}\right]
\end{aligned}
$$

onde $N_{\alpha/2}$ e $N_{1-\alpha/2}$ representam, respectivamente, os $\alpha/2$ e $1-\alpha/2$ quantis da distribuição normal $N(0,1)$.

De entre todos os intervalos para a proporção com $100(1-\alpha)\%$ de confiança assim construídos, apenas 100α % desses intervalos não conterão o parâmetro π.

- **Intervalo de confiança para a variância**

Sabemos já que, se $X \sim N(\mu, \sigma^2)$, então

$$n\frac{\widehat{\sigma}}{\sigma^2} = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1).$$

Para qualquer $0 < \alpha < 1$, resulta que o intervalo aleatório (centrado) que com probabilidade $1-\alpha$, contém a variância σ^2 da população é definido por

$$\chi^2_{\alpha/2}(n-1) \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}(n-1)$$

isto é

$$\frac{1}{(n-1)S^2}\chi^2_{\alpha/2} \leq \frac{1}{\sigma^2} \leq \frac{1}{(n-1)S^2}\chi^2_{1-\alpha/2}$$

e portanto o $IC_{(1-\alpha)100\%}$ para uma amostra é

$$
\begin{aligned}
&IC_{100(1-\alpha)\%}\\
&= \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right].
\end{aligned}
$$

Repare-se que a distribuição χ^2 não é simétrica e portanto não existe nenhuma relação privilegiada entre $\chi^2_{\alpha/2}(n-1)$ e $\chi^2_{1-\alpha/2}(n-1)$.

Para n grande (para distribuições *razoáveis* pode pensar-se em $n-1 > 60$), pode-se usar o facto de uma distribuição $\chi^2(\nu)$ (com média ν e variância 2ν) poder ser aproximada por uma distribuição normal, isto é,

$$\chi^2(n-1) \approx N(n-1, 2(n-1)).$$

1.2. Intervalos de confiança - populações normais

- **Parâmetro**: Média μ (σ^2 conhecida)

  Condições de aplicação:

  $X \sim N(\mu, \sigma^2)$

  Variável fulcral:

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

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{x} - \frac{\sigma}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1), \overline{x} + \frac{\sigma}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1)\right)$$

- **Parâmetro**: Média μ (σ^2 desconhecida)

  Condições de aplicação:

  $X \sim N(\mu, \sigma^2)$

  Variável fulcral:

  $$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1), \overline{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)\right)$$ [a]

[a] Há autores que escrevem $t_{n-1,1-\frac{\alpha}{2}}$ para designar o $(1-\frac{\alpha}{2})$-quantil da distribuição $t(n-1)$, valor acima designado por $t_{1-\frac{\alpha}{2}}(n-1)$.

- **Parâmetro**: Variância σ^2

  Condições de aplicação:

  $X \sim N(\mu, \sigma^2)$

  Variável fulcral:

  $$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

  IC a $(1-\alpha)100\%$:

  $$\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right)$$

- **Parâmetro**: Diferença de médias $\mu_1 - \mu_2$ (pop. independentes; σ_1^2, σ_2^2 conhecidas)

  Condições de aplicação:

  $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$

  Variável fulcral:

  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{x}_1 - \overline{x}_2 - \sigma^* N_{1-\frac{\alpha}{2}}(0,1), \overline{x}_1 - \overline{x}_2 + \sigma^* N_{1-\frac{\alpha}{2}}(0,1)\right)$$

  onde $\sigma^* = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- **Parâmetro**: Diferença de médias $\mu_1 - \mu_2$ (pop. independentes; σ_1^2, σ_2^2 desconhecidas)

  Condições de aplicação:

  $X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$

  Variável fulcral:

  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{a}{\sim} t(\nu)$$

  onde ν é o maior inteiro que não excede

  $$\nu' = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}$$

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{x}_1 - \overline{x}_2 - s^* t_{1-\frac{\alpha}{2}}(\nu), \overline{x}_1 - \overline{x}_2 + s^* t_{1-\frac{\alpha}{2}}(\nu)\right)$$

  onde $s^* = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

- **Parâmetro**: Razão de variâncias $\frac{\sigma_2^2}{\sigma_1^2}$

  Condições de aplicação:

  $X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$

  Variável fulcral:

  $$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1)$$

  IC a $(1-\alpha)100\%$:

  $$\left(\frac{s_2^2}{s_1^2} F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1), \frac{s_2^2}{s_1^2} F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)\right)$$

- **Parâmetro**: Diferenças de médias $\mu_1 - \mu_2$ (amostras emparelhadas)

  Condições de aplicação:

  $D = X_1 - X_2 \sim N(\mu_D, \sigma_D^2)$

  Variável fulcral:

  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \sim t(n-1)$$

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{x}_1 - \overline{x}_2 - \frac{s_D}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1), \overline{x}_1 - \overline{x}_2 + \frac{s_D}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)\right)$$

1.3. Intervalos de confiança - amostras grandes

- **Parâmetro**: Média μ (σ^2 finita)

  Variável fulcral:

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{x}-\frac{s}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1), \overline{x}+\frac{s}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1)\right)$$

- **Parâmetro**: Proporção π

  Condições de aplicação:

  $n \geq 30$, $np > 5$ e $n(1-p) > 5$ [a]

  Variável fulcral:

$$\frac{P-\pi}{\sqrt{\frac{P(1-P)}{n}}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(p-\sqrt{\frac{p(1-p)}{n}}N_{1-\frac{\alpha}{2}}, p+\sqrt{\frac{p(1-p)}{n}}N_{1-\frac{\alpha}{2}}\right)$$

  onde $N_{1-\frac{\alpha}{2}}$ representa o $(1-\frac{\alpha}{2})$-quantil da distribuição $N(0,1)$.

[a] Caso contrário, usa-se a distribuição binomial.

- **Parâmetro**: Diferença de médias $\mu_1-\mu_2$ (pop. independentes; σ_1^2, σ_2^2 finitas)

  Variável fulcral:

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{x}_1-\overline{x}_2-s^*N_{1-\frac{\alpha}{2}}(0,1), \overline{x}_1-\overline{x}_2+s^*N_{1-\frac{\alpha}{2}}(0,1)\right)$$

  onde $s^*=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$.

- **Parâmetro**: Diferenças de médias $\mu_1-\mu_2$ (amostras emparelhadas)

  Variável fulcral:

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S_D/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{x}_1-\overline{x}_2-\frac{s_D}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1), \overline{x}_1-\overline{x}_2+\frac{s_D}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1)\right)$$

- **Parâmetro**: Diferença de proporções $\pi_1 - \pi_2$

  Condições de aplicação:

  $n_1, n_2 \geq 30$; $n_1 p_1, n_2 p_2 > 5$ e $n_1(1-p_1), n_2(1-p_2) > 5$ [a]

  Variável fulcral:

$$\frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}} \overset{a}{\sim} N(0,1)^{b}$$

  IC a $(1-\alpha)100\%$:

$$\left(p_1 - p_2 - p^* N_{1-\frac{\alpha}{2}}(0,1), p_1 - p_2 + p^* N_{1-\frac{\alpha}{2}}(0,1)\right)$$

  onde $p^* = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

[a] Caso contrário, usa-se o teste exacto de Fisher que usa a distribuição hipergeométrica.

[b] Com correcção de continuidade, a variável fulcral é

$$\frac{(P_1 - P_2) - (\pi_1 - \pi_2) - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}} \overset{a}{\sim} N(0,1)$$

Para amostras *bastante grandes*, a correcção de continuidade é usualmente ignorada (reparar que esta conduz a uma alteração da estatística quase insignificante).

Observação: Um intervalo de confiança **exacto** a $(1-\alpha)100\%$ para a proporção π é (π_1, π_2) onde π_1 e π_2 são tais que:

$$P(X \geq x | \pi = \pi_1) = \frac{\alpha}{2} = \sum_{i=x}^{n} \binom{n}{i} \pi_1^i (1-\pi_1)^{n-i}$$

$$P(X \leq x | \pi = \pi_2) = \frac{\alpha}{2} = \sum_{i=0}^{x} \binom{n}{i} \pi_2^i (1-\pi_2)^{n-i}$$

(aqui $X \sim B(n, \pi)$ é a v.a. que conta o número de sucessos na amostra de tamanho n.)

1.4. Intervalos de confiança: Exemplos

Exemplo 1: Suponhamos que um investigador está interessado em estimar o nível médio de uma determinada enzima numa população de indivíduos. Considera uma amostra aleatória $X_1, \ldots, X_{10}$ constituída por 10 pessoas, determina o nível da enzima em cada uma dessas pessoas e calcula a média amostral obtendo $\overline{x} = 22$. Suponhamos ainda que é razoável assumir que a variável "nível da enzima" segue uma distribuição normal com uma variância igual a 45. Pretende-se determinar um intervalo a 95% de confiança para a média (populacional) do nível de enzima.

Resolução: Temos $\overline{x} = 22$, $\sigma^2 = 45$ e $n = 10$.
O intervalo de confiança a 95% para a média da população é

$$\begin{aligned}&\left(22 - \frac{\sqrt{45}}{\sqrt{10}} N_{0.975}, 22 + \frac{\sqrt{45}}{\sqrt{10}} N_{0.975}\right)\\ &= (22 - 2.121(1.96), 22 + 2.121(1.96))\\ &= (17.8, 26.2).\end{aligned}$$

O valor do quantil-0.975 da distribuição normal reduzida foi obtido da forma usual.

Exemplo 2: Um grupo de investigadores recolheu os seguintes dados para o peso da glândula pituitária numa amostra constituída por vinte ratinhos:

média: 9.0 mg, desvio padrão: 0.6.

Pretende-se determinar um intervalo de confiança a 95% para o peso médio da glândula pituitária da população de ratinhos. Sob que hipóteses é que este intervalo é válido?

Resolução: Temos $n = 20$, $\overline{x} = 9.0$, $s^2 = 0.36$, $\alpha = 0.05$ e não se conhece a variância da população. Além disso não se sabe se o peso da glândula pituitária nos ratinhos segue uma distribuição normal e o tamanho da amostra não é *grande*.
Na hipótese de ser razoável assumir que a população segue uma distribuição normal, o intervalo de confiança a 95% será

$$\begin{aligned}&\left(9.0 - \frac{0.6}{\sqrt{20}} t_{0.975}(19), 9.0 + \frac{0.6}{\sqrt{20}} t_{0.975}(19)\right)\\ &= (9.0 - 0.134(2.093), 9.0 + 0.134(2.093))\\ &= (8.72, 9.28)\end{aligned}$$

Exemplo 3: Numa amostra aleatória constituída por 591 pacientes de um hospital psiquiátrico verificou-se que 204 desses pacientes admitiram ter consumido cannabis pelo menos uma vez na vida. Determinar o intervalo de confiança a 90% para a proporção de indivíduos que alguma vez consumiu cannabis considerando como população todos os indivíduos inscritos nesse hospital psiquiátrico.

Resolução: Temos $n = 591$, $p = 204/591 \approx 0.34$ e $\alpha = 0.1$. Sendo a amostra grande e p não muito próximo de 0 nem 1 (verifica-se que $n \geq 30$ e $np(1-p) > 5$), o intervalo de confiança a 90% para a proporção populacional π é

$$\left(0.34 - \sqrt{\frac{0.34(0.66)}{591}}N_{0.95}, 0.34 + \sqrt{\frac{0.34(0.66)}{591}}N_{0.95}\right)$$
$$= (0.34 - 0.02(1.645), 0.34 + 0.02(1.645)).$$
$$= (0.31, 0.38).$$

Exemplo 4: Uma equipa de investigação está interessada na diferença entre os níveis de ácido úrico em pacientes com e sem sindrome de Down. Uma amostra de 12 indivíduos com a doença foi seleccionada de um hospital psiquiátrico fornecendo uma média de 4.5 mg/100 ml e um desvio padrão de 1 mg/100 ml. Num outro hospital foi recolhida uma amostra de 15 indivíduos sem a doença, com as mesmas idades e o mesmo sexo, e a média e o devio padrão obtidos foram de 3.4 mg/100 ml e 1.5 mg/100 ml, respectivamente. Determine um intervalo de confiança a 99% para diferença entre os valores médios de ácido úrico entre doentes e não doentes.

Resolução: As amostras não são grandes portanto um intervalo de confiança para a diferença de médias só será conseguido se se assumir que as duas populações em causa (independentes) seguem distribuições normais. Dos dados, resulta que

$$\begin{array}{lll} n_1 = 12 & \overline{x}_1 = 4.5 & s_1 = 1 \\ n_2 = 15 & \overline{x}_2 = 3.4 & s_2 = 1.5. \end{array}$$

O intervalo de confiança a 99% para a diferença de médias é

$$\left(\overline{x}_1 - \overline{x}_2 - s^* t_{1-\frac{\alpha}{2}}(\nu), \overline{x}_1 - \overline{x}_2 + s^* t_{1-\frac{\alpha}{2}}(\nu)\right)$$

onde ν é o maior inteiro que não excede

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

Ora

$$\nu' = \frac{\left(\frac{1^2}{12} + \frac{(1.5)^2}{15}\right)^2}{\frac{1}{11}\left(\frac{1^2}{12}\right)^2 + \frac{1}{14}\left(\frac{(1.5)^2}{15}\right)^2} \approx 24.322$$

portanto $\nu = 24$, e

$$s^* = \sqrt{\frac{1^2}{12} + \frac{(1.5)^2}{15}} \approx 0.483.$$

Vem então que o intervalo pedido é

$$\begin{aligned}&(4.5 - 3.4 - 0.483\, t_{0.995}(24), 4.5 - 3.4 + 0.483\, t_{0.995}(24))\\ &= (-0.25, 2.45).\end{aligned}$$

Podemos dizer, em termos práticos, que temos 99% de *confiança* que este intervalo contenha a diferença de médias entre as duas populações. Ora isto implica que, para esse nível de confiança, essa diferença pode ser 0! portanto não se consegue afirmar que existem diferenças *significativas* entre os valores médios de ácido úrico nos doentes e nos não doentes.

Exemplo 5: Connor *et al* (2003) estudaram as diferenças de género entre agressões proactivas e reactivas numa amostra de 323 crianças e adolescentes (68 raparigas e 255 rapazes). Dos relatórios sobre os indivíduos amostrados, constatou-se que 31 das raparigas e 53 dos rapazes tinham sofrido abusos sexuais. Construa um intervalo de confiança a 95% para a diferença entre as proporções de abuso sexuais entre as duas populações consideradas. Interprete o resultado.

Resolução: Tem-se

$$\begin{aligned} n_1 &= 68, \quad p_1 = 31/68 \approx 0.456\\ n_2 &= 255, \quad p_2 = 53/255 \approx 0.208.\end{aligned}$$

Como as amostras são *grandes* e as proporções amostrais não são muito próximas dos extremos (0 ou 1), o intervalo de confiança a 95% para $\pi_1 - \pi_2$ é

$$\left(p_1 - p_2 - p^* N_{1-\frac{\alpha}{2}}(0,1), p_1 - p_2 + p^* N_{1-\frac{\alpha}{2}}(0,1)\right)$$

onde $p^* = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. No caso em questão, vem então

$$p^* = \sqrt{\frac{\frac{31}{68}\left(1 - \frac{31}{68}\right)}{68} + \frac{\frac{53}{255}\left(1 - \frac{53}{255}\right)}{255}} \approx 0.0655$$

e portanto o intervalo de confiança a 95% é

$$\begin{aligned}&\left(\frac{31}{68} - \frac{53}{255} - 0.0655\, N_{0.975}, \frac{31}{68} - \frac{53}{255} + 0.0655\, N_{0.975}\right)\\ &= \left(\frac{31}{68} - \frac{53}{255} - 0.0655 \times 1.96, \frac{31}{68} - \frac{53}{255} + 0.0655 \times 1.96\right)\\ &= (0.120, 0.376).\end{aligned}$$

Conclui-se que, com 95% de *confiança*, a diferença de proporções populacionais $\pi_1 - \pi_2$ situa-se entre 0.120 e 0.376. Dado que este intervalo não contém o 0, conseguimos afirmar, com esse nível de confiança, que a proporção de raparigas que sofrem abusos sexuais é diferente da proporção de rapazes que sofrem abusos sexuais.

Nota: As instruções anteriores aplicam-se à situação em que não conhecemos as observações da experiência mas apenas as medidas descritivas dessas observações. No caso geral, essa não é a situação. O que se tem é um conjunto de dados, resultantes de uma experiência, e procuram-se intervalos de confiança para medidas de localização ou escala da população de onde os dados foram retirados. A determinação desses intervalos usando um software de análise estatística adequado é, em geral, muito simples.