3. INTRODUÇÃO À INFERÊNCIA ESTATISTICA

Noções preliminares sobre estimação

Pode dizer-se que as probabilidades e a estatística têm objectivos diferentes: enquanto nas probabilidades se parte de um dado esquema ou modelo para calcular próbabilidades de certos resultados ou acontecimentos, na estatística parte-se de dados ou observações e procura saber-se algo sobre o modelo (Tiago de Oliveira (1990) e Murteira (1990)).

É a **inferência estatística** que tem como objectivo a construção e desenvolvimento de métodos que permitem a extensão do particular para o geral (chamada inferência indutiva), i.e., a partir de um conjunto de dados é possível fazer 'inferências' ou generalizações acerca de uma população da qual os dados foram extraídos.

A inferência estatística é então um método científico de tirar conclusões sobre os parâmetros da população a partir da recolha, tratamento e análise dos dados de uma amostra, recolhida dessa população.

O conjunto completo de todas as observações possíveis constitui a população, enquanto o conjunto dos valores efectivamente observados constitui a amostra. Chama-se população de amostras ao conjunto de todas as amostras observáveis.

Atenda-se a que **parâmetro de uma população** é uma **constante** desconhecida, cujo verdadeiro valor só se conseguiria saber, nalguns casos, após estudos exaustivos e noutros nem sequer é possível saber.

Num problema de inferência estatística ou se admite que a distribuição da população tem uma forma matemática conhecida, embora contendo um ou mais parâmetros desconhecidos, é o que se chama estatística paramétrica ou se pretende conhecer a forma da distribuição, é o domínio da estatística não paramétrica.

Os dois tipos mais importantes de inferência estatística são:

- estimação dos parâmetros e
- testes de hipóteses estatísticas.

A estimação permite-nos "adivinhar" ou melhor estimar o verdadeiro valor desconhecido do(s) parâmetro(s) da população, **estimação pontual**, ou obter um intervalo de valores plausíveis para esse parâmetro, com a indicação da confiança no procedimento, **estimação por intervalos**.

A outra grande área da inferência estatística, a dos testes de hipóteses, tem como objectivo decidir se o valor do parâmetro pertence ou não a um domínio de valores especificado pelo investigador.

A extensão do particular ao geral que temos estado a referir, chama-se inferência indutiva. É o caminho para a aquisição de novos conhecimentos. O grau de incerteza que acompanha as inferências indutivas pode ser medido rigorosamente em termos de probabilidade, se a experiência foi conduzida segundo determinados princípios (probabilísticos ou aleatórios).

Os procedimentos que levam à obtenção de amostras nestas condições são do domínio da **teoria da amostragem**. Vejamos alguns conceitos básicos, deixando como referências para um aprofundamento do assunto Murteira (1990), Barnett (1982) e Cochran (1977).

A amostra, que irá ser utilizada para tirar conclusões sobre parâmetros desconhecidos da população, deverá ser representativa dessa população. Para isso deverá obedecer a princípios de aleatoriedade, i.e., a selecção dos indivíduos a incluir na amostra é deixada completamente ao acaso.

Temos assim um dos conceitos fundamentais da teoria da amostragem, o de **amostra aleatória**.

Definição 3.1

Diz-se que $(X_1, X_2, ..., X_n)$ é uma **amostra aleatória** de dimensão n se as variáveis X_i $(i = 1, \cdots, n)$ são independentes e semelhantes, i.e., têm todas a mesma distribuição, que é a da população.

Se as variáveis $X_1, X_2, ..., X_n$ constituem uma amostra aleatória extraída de uma população com função densidade $f(x)$, então a função densidade conjunta daquele vector aleatório $\underline{X} = (X_1, X_2, ..., X_n)$ pode escrever-se (devido à independência entre das variáveis) como:

$$f_{\underline{X}}(x_1, x_2, ..., x_n) = f(x_1)f(x_2)...f(x_n)$$

É importante realçar que uma amostra aleatória $(X_1, X_2, ..., X_n)$ é um conjunto de n variáveis aleatórias. Antes da amostragem ser realizada as quantidades observáveis são variáveis aleatórias, depois de feitas as observações temos um conjunto de dados que constituem a amostra observada, que passaremos a representar por $(x_1, ..., x_n)$.

Note-se que nem todas as amostras conduzem a generalizações válidas para a população da qual foram extraídas. De facto os métodos de inferência que iremos utilizar baseiam-se na hipótese de que estamos a considerar **amostras aleatórias**.

Definição 3.2

Dada uma amostra aleatória, chama-se **estatística** a toda a função da amostra

aleatória, que não contenha parâmetros desconhecidos.

Uma estatística, $T = T(X_1, X_2, ..., X_n)$, função dos valores aleatórios, é portanto uma variável aleatória. A cada amostra observada $(x_1, ..., x_n)$, corresponde um valor numérico bem determinado para a estatística, $t(x_1, x_2, ..., x_n)$.

Para a amostra aleatória $(X_1, ...X_n)$, vejamos alguns exemplos de estatísticas importantes.

- **Média Amostral**

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{3.1}$$

- **Mediana Amostral**

$$\tilde{X} = \begin{cases} X_{(n+1)/2} & n \text{ ímpar} \\ \frac{X_{(n/2)} + X_{(n/2)+1}}{2} & n \text{ par} \end{cases}. \tag{3.2}$$

- **Variância Amostral**

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}. \tag{3.3}$$

Acabámos de referir alguns exemplos de estatísticas. Observe-se que **não são estatísticas**, por exemplo, as seguintes funções:

$$\sum \frac{X_i - \mu}{\sigma}; \qquad \sum \frac{X_i}{\sigma},$$

por conterem parâmetros desconhecidos.

Suponhamos que de uma população infinita qualquer, extraímos ao acaso uma primeira amostra de n observações:

$$x_1, ..., x_n \qquad \text{cuja média será então} \qquad \overline{x} = \frac{\sum x_i}{n}.$$

Se, nas mesmas condições, fosse outra pessoa a extrair uma amostra também de dimensão n, teríamos

$$x_1^{'}, ..., x_n^{'} \qquad \text{cuja média será, digamos,} \qquad \overline{x}^{'} = \frac{\sum x_i^{'}}{n},$$

que provavelmente será diferente da primeira, $\overline{x}$; e assim sucessivamente para outras amostras que possam ser extraídas, nas mesmas condições das anteriores.

Podiamos então considerar $\overline{x}$, $\overline{x}^{'}$, $\overline{x}^{''}$, $\overline{x}^{'''}$, ..., valores observados da variável aleatória $\overline{X}$.

Um processo de amostragem dá lugar, como vimos, a muitas amostras diferentes. Uma amostra particular é apenas uma das muitas amostras (em número infinito se a população for infinita), que é possível extrair da população.

As flutuações de amostragem só podem ser controladas quando a amostra é aleatória.

Observe-se que, o facto de se ter considerado a população infinita e no desenvolvimento que se segue, todo o estudo ter este pressuposto como base, não é efectivamente uma restrição. Muitas das populações que nós estudamos ou são infinitas, por exemplo, a população dos perfis pedológicos de determinada região, as realizações possíveis do mesmo jogo, etc, ou podem ser consideradas teoricamente infinitas, por a dimensão da população comparada com a dimensão da amostra ser de tal modo grande, que torna desprezível o erro cometido ao considerar-se infinita.

As estatísticas são, pela própria definição como vimos, variáveis aleatórias, tendo portanto distribuições a que é costume chamar **distribuições de amostragem**.

Se $(X_1, X_2, ..., X_n)$ é uma amostra aleatória extraída de uma população com função de probabilidade ou densidade $f(x|\theta)$, onde θ designa o(s) parâmetro(s) desconhecido(s), a distribuição de amostragem da estatística $T(X_1, X_2, ..., X_n)$ pode definir-se a partir da distribuição conjunta $\prod_{i=1}^{n} f(x_i|\theta)$.

Iremos considerar agora as distribuições por amostragem das estatísticas mais frequentemente usadas.

Distribuições de amostragem

Distribuição da Média Amostral

Consideremos uma amostra aleatória com n observações, retirada de uma população **normal**, com valor médio μ e variância σ^2. Sendo assim, por definição de amostra aleatória, cada observação $X_i,\ i = 1, ..., n$, tem distribuição normal, com valor médio e variância μ e σ^2, respectivamente.

Tendo em conta propriedades da distribuição normal, sabemos que

$$\overline{\boldsymbol{X}} \cap \boldsymbol{\mathcal{N}(\mu, \sigma/\sqrt{n})} \Longleftrightarrow \frac{\overline{\boldsymbol{X}} - \boldsymbol{\mu}}{\boldsymbol{\sigma/\sqrt{n}}} \cap \boldsymbol{\mathcal{N}(0,1)}. \tag{3.4}$$

Porém, se a amostragem foi feita numa população com distribuição desconhecida, a distribuição de amostragem de $\overline{X}$ é ainda aproximadamente normal com valor médio e variância μ e σ^2/n, respectivamente, desde que o tamanho da amostra seja grande (teorema limite central), i. e., quando $n \to \infty$

$$\frac{\overline{\boldsymbol{X}} - \boldsymbol{\mu}}{\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}} \sim \boldsymbol{\mathcal{N}(0,1)}. \tag{3.5}$$

Na prática, a aproximação anterior é regra geral boa quando $n > 30$ e a distribuição tenha apenas uma moda. No caso de $n < 30$ a aproximação ainda é razoável se a distribuição da população não diferir muito da normal. Se a população for normal, a distribuição de amostragem de $\overline{X}$ é exactamente normal qualquer que seja o tamanho da amostra.

As distribuições apresentadas em (3.4) e (3.5) são fundamentais para a realização de inferências sobre o parâmetro μ, **quando $\boldsymbol{\sigma}$ é conhecido**.

Veremos mais adiante a distribuição que surge quando σ não é conhecido. Antes disso, porém, necessitamos do conhecimento da distribuição da Variância Amostral.

Distribuição da Variância Amostral

Dada uma amostra aleatória de dimensão n extraída de uma população normal com valor médio μ e variância σ^2, pretendemos determinar a distribuição de amostragem da variância amostral S^2, definida em (3.3). Porém, a distribuição de S^2 tem pouca aplicação prática em estatística. Em vez daquela variável aleatória é costume considerar a distribuição de amostragem da variável aleatória

$$\boldsymbol{(n-1)S^2/\sigma^2}.$$

Teorema 3.1

Sendo X_i variáveis aleatórias normais independentes, $\mathcal{N}(\mu, \sigma)$, a variável aleatória $(n-1)S^2/\sigma^2$ tem distribuição χ^2 com $(n-1)$ g.l.

Dem:
Consideremos

$$\sum (X_i - \mu)^2 = \sum \left[(X_i - \overline{X}) + (\overline{X} - \mu)\right]^2 =$$

$$= \sum (X_i - \overline{X})^2 + \sum (\overline{X} - \mu)^2 + 2\sum (X_i - \overline{X})(\overline{X} - \mu) =$$

$$= \sum (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2$$

dado que a última parcela é nula.

Portanto podemos escrever

$$\frac{\sum (X_i - \mu)^2}{\sigma^2} = \frac{\sum (X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}$$

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2. \tag{3.6}$$

O primeiro membro da igualdade (3.6) tem, como se sabe, distribuição $\chi^2_{(n)}$, visto ser a soma dos quadrados de n v.a. independentes normais reduzidas (ver generalização do teorema 2.20, pág. 105) e o segundo termo do segundo membro tem distribuição $\chi^2_{(1)}$.

O resultado que se pretende provar resulta de dois teoremas fundamentais que iremos enunciar sem demonstração.

Teorema 3.2

Se X_i $(i = 1, \cdots, n)$ são n variáveis aleatórias normais independentes, as variáveis S^2 e $\overline{X}$ são independentes.

Teorema 3.3

Se X e Y são duas variáveis aleatórias independentes tais que $X + Y \cap \chi^2_{(n)}$ e $X \cap \chi^2_{(n_1)}$ com $n_1 < n$, então $Y \cap \chi^2_{(n-n_1)}$.

Como consequência destes dois teoremas e da igualdade (3.6) temos então

$$\boldsymbol{\frac{(n-1)S^2}{\sigma^2} \cap \chi^2_{(n-1)}}. \tag{3.7}$$

Exemplo 3.1

Um fabricante de baterias de automóveis garante que as suas baterias têm uma duração média de 3 anos com um desvio padrão de 1 ano. Se 5 dessas baterias tiverem durado

1.9 2.4 3.0 3.5 e 4.2 anos

continuará o fabricante convencido que as suas baterias têm um desvio padrão de 1 ano?

(Nota: admita-se que a duração das baterias se pode considerar uma v.a. normal).

Resolução:

Consideremos então $X \cap \mathcal{N}(3, 1)$, a v.a. que designa a duração de vida das referidas baterias, sendo o nosso objectivo estudar o seu desvio padrão.

Para isso considere-se a variável

$$(n-1)S^2/\sigma^2$$

que como sabemos tem distribuição $\chi^2_{(n-1)} = \chi^2_{(4)}$.

Vejamos qual o valor desta variável no caso da amostra observada:

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = 0.815$$

Portanto

$$\chi^2_{\text{calculado}} = \frac{(n-1)s^2}{\sigma^2} = 3.26$$

Ora sabemos que 95% dos valores de um χ^2 estão entre $\chi^2_{0.975}$ e $\chi^2_{0.025}$, que no nosso caso são $\chi^2_{0.975} = 0.484$ e $\chi^2_{0.025} = 11.143$. Como conclusão, podemos dizer que o valor da estatística calculado considerando $\sigma^2 = 1$ é um valor plausível, não tendo portanto o fabricante razões para suspeitar que o desvio padrão não seja de 1 ano.

Acabámos de estudar as distribuições de amostragem da média e da variância, porém a distribuição da média (3.4) ou mesmo a distribuição aproximada (3.5) **não têm interesse quando a variância é desconhecida**.

No caso de dispormos de **amostras de dimensão grande**, podemos substituir em (3.5) σ^2 pela variância observada na amostra, s^2, tendo-se então a v.a.

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{3.8}$$

Porém, se n **é pequeno**, os valores de s^2 estão sujeitos a grandes flutuações de amostra para amostra e a aproximação (3.8) já não é válida. Para dar resposta a esta situação, tem-se uma distribuição diferente da distribuição normal - a distribuição $\boldsymbol{t - Student}$[1]. Vejamos em primeiro lugar a definição da distribuição $t - Student$.

Definição 3.3

Sejam $Z \cap \mathcal{N}(0,1)$ e $Y \cap \chi^2_{(n)}$, variáveis aleatórias independentes. A v.a. assim definida

$$\boldsymbol{X = \frac{Z}{\sqrt{Y/n}}} \tag{3.9}$$

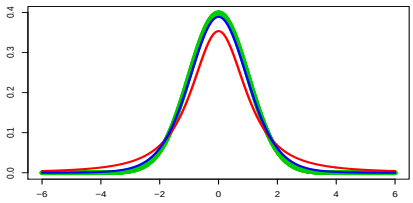
diz-se ter distribuição $t - Student$ com parâmetro n, ou com n graus de liberdade e representa-se por $\boldsymbol{X \cap t_{(n)}}$.

Propriedades da distribuição $\boldsymbol{t - Student}$

1. Trata-se de uma distribuição simétrica, unimodal.

2. A distribuição $\boldsymbol{t}$ é semelhante à distribuição normal reduzida (ver figura seguinte)

[1] A distribuição t de Student, foi introduzida em 1908 por William Gosset, que trabalhava então para uma fábrica de cervejas. Como esta não queria que os seus concorrentes soubessem do uso que os seus técnicos faziam de métodos estatísticos, Gosset teve de publicar o seu trabalho sob o pseudónimo de "Student".



Gráficos da função densidade de uma v.a. com distribuição $N(0,1)$ (a verde), $t_{(4)}$(a vermelho) e $t_{(10)}$(a azul).

Parâmetros da distribuição t

$$E[X] = 0 \quad (n > 1) \qquad Var[X] = \frac{n}{n-2} \quad (n > 2). \tag{3.10}$$

Existem também tabelas para esta distribuição. As utilizadas no nosso curso de Estatística dão-nos para cada par de valores n e α o ponto $t_{\alpha(n)}$ tal que, sendo $X \cap t_{(n)}$, $P[X > t_{\alpha(n)}] = \alpha$.

Atendendo ao facto de se tratar de uma distribuição simétrica, é fácil verificar a seguinte relação:

$$\boldsymbol{t_{1-\alpha} = -t_{\alpha}}. \tag{3.11}$$

Consideremos agora novamente o estudo da distribuição da variável $\overline{X}$, quando a amostra aleatória provém de uma população normal com variância desconhecida. Tem-se o seguinte teorema:

Teorema 3.4

Se $X_1, ..., X_n$ é uma amostra aleatória retirada de uma população normal de valor médio μ e variância σ^2, a v.a.

$$\boldsymbol{\frac{\overline{X} - \mu}{S/\sqrt{n}} \cap t_{(n-1)}} \tag{3.12}$$

Dem: Nas condições do teorema sabemos que

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \cap \mathcal{N}(0,1) \qquad \text{e} \qquad \frac{(n-1)S^2}{\sigma^2} \cap \chi^2_{(n-1)}.$$

Além disso, como no caso de variáveis aleatórias normais $\overline{X}$ e S^2 são independentes tem-se, aplicando a definição (3.3)

$$\frac{(\overline{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S^2/\sigma^2(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \cap t_{(n-1)}.$$

Distribuição da diferença entre duas médias amostrais

Consideremos duas amostras aleatórias $X_{11}, X_{12}, ..., X_{1n_1}$ e $X_{21}, X_{22}, ..., X_{2n_2}$ extraídas independentemente de duas populações normais $X_1 \cap \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \cap \mathcal{N}(\mu_2, \sigma_2)$.

As médias,

$$\overline{X_1} = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1} \qquad \text{e} \qquad \overline{X_2} = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}$$

são também normais independentes. Então aplicando o Exercício 2.15, pág. 97, que generaliza o teorema da estabilidade da soma de normais, temos

$$\overline{X_1} - \overline{X_2} \cap \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \tag{3.13}$$

Se as populações são não normais, a distribuição de amostragem de $\overline{X_1} - \overline{X_2}$ é aproximadamente normal, desde que as dimensões das amostras n_1 e n_2 sejam ambas grandes.

A distribuição de amostragem (3.13) só tem aplicação quando as variâncias σ_1^2 e σ_2^2 de cada uma das populações são conhecidas. No caso de n_1 e n_2 grandes pode substituir-se σ_1^2 por s_1^2 e σ_2^2 por s_2^2 sendo a aproximação à normal ainda bastante razoável.

Porém, se as dimensões são pequenas, a aproximação já não é válida. No caso de as variâncias das populações embora desconhecidas poderem ser supostas iguais (mais tarde veremos como se poderá verificar esta condição), é possível usar um procedimento que vai também conduzir à distribuição $t - Student$.

Teorema 3.5

Se $X_{11}, X_{12}, ..., X_{1n_1}$ e $X_{21}, X_{22}, ..., X_{2n_2}$ são amostras aleatórias extraídas independentemente de duas populações normais $X_1 \cap \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \cap \mathcal{N}(\mu_2, \sigma_2)$, respectivamente, com variâncias iguais, $\sigma_1^2 = \sigma_2^2$, então

$$\frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} \cap t_{(n_1+n_2-2)}$$

Dem:

Seja σ^2 a variância comum às duas populações normais, então

$$\frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}} \cap \mathcal{N}(0,1) \tag{3.14}$$

Sabemos ainda que as variáveis aleatórias

$$\frac{(n_1-1)S_1^2}{\sigma^2} \cap \chi^2_{(n_1-1)} \qquad \text{e} \qquad \frac{(n_2-1)S_1^2}{\sigma^2} \cap \chi^2_{(n_2-1)}$$

são independentes pois as amostras são seleccionadas independentemente de cada população. Pela estabilidade da soma da distribuição χ^2 tem-se

$$\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2} \cap \chi^2_{(n_1+n_2-2)}$$

Sendo assim, e utilizando de novo a definição (3.3), a variável aleatória definida como

$$\frac{\frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2\ (n_1+n_2-2)}}} =$$

$$\frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} \cap t_{(n_1+n_2-2)}$$

É costume designar por

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \quad \text{a que se chama variância "ponderada".}$$

No caso de as variâncias das populações não poderem ser consideradas iguais vários métodos têm sido sugeridos; referiremos aqui o procedimento devido a Welsh-Satterthwaite que estabelece que a variável aleatória

$$\frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \tag{3.15}$$

tem aproximadamente distribuição t em que o número de graus de liberdade γ, tem de ser estimado directamente a partir dos dados:

$$\gamma = \frac{(s_1^2/n_1+s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}. \tag{3.16}$$

Como o valor obtido para γ pela expressão anterior não é necessariamente um inteiro, deve arredondar-se para o inteiro imediatamente inferior.

Distribuição do quociente de variâncias amostrais

A estatística S_1^2/S_2^2, em que S_1^2 e S_2^2 são as variâncias amostrais de duas populações, tem um grande interesse em questões de inferência sobre a razão entre as variâncias das duas populações.

A distribuição de uma variável aleatória função do quociente entre duas variâncias amostrais, em populações normais com base em amostras independentes, foi estudada por Fisher e Snedecor. Vamos aqui apresentar a definição e algumas propriedades da distribuição $\boldsymbol{F - Snedecor}$.

Definição 3.4

Sejam $U \cap \chi^2_{(m)}$ e $V \cap \chi^2_{(n)}$ variáveis aleatórias independentes. A v.a. assim definida

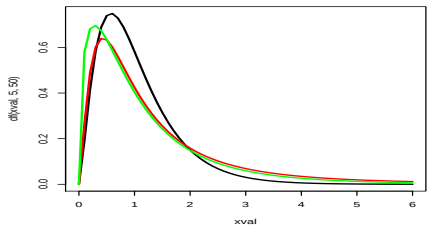
$$\boldsymbol{X = \frac{U/m}{V/n}} \tag{3.17}$$

tem distribuição F com (m, n) graus de liberdade e costuma representar-se por $\boldsymbol{X \cap F_{(m,n)}}$,

Esta distribuição é, juntamente com as distribuições normal, qui-quadrado e t, uma das mais importantes distribuições em estatística aplicada. Estas distribuições constituem um conjunto de instrumentos indispensáveis na resolução de problemas de inferência estatística.

A distribuição F encontra a sua grande aplicação em problemas de análise de variância (ANOVA), um dos métodos mais importantes da análise estatística.

Na figura seguinte pode ver-se algumas formas da função densidade para alguns valores dos parâmetros.



Gráficos da função densidade de uma v.a. com distribuição $F_{(5,50)}$ (a preto), $F_{(5,5)}$ (a vermelho) e $F_{(3,10)}$ (a verde).

Exercício 3.1

Seja $X \cap t_{(n)}$. Prove que $X^2 \cap F_{(1,n)}$.

Teorema 3.6

Se

$$X \cap F_{(m,n)} \Rightarrow Y = 1/X \ \cap F_{(n,m)}.$$

Dem: A demonstração é imediata tendo em conta a definição (3.4) de uma variável aleatória com distribuição F.

Consequência: Sendo $f_{\alpha(m,n)}$ o valor de uma distribuição F tal que $P[X > f_{\alpha(m,n)}] = \alpha$, tem-se

$$\boldsymbol{f_{1-\alpha(m,n)} = \frac{1}{f_{\alpha(n,m)}}} \tag{3.18}$$

Dem:
Se $X \cap F_{(m,n)}$ temos $f_{1-\alpha(m,n)}$ como o ponto tal que

$$P(X > f_{1-\alpha(m,n)}) = 1-\alpha \Longleftrightarrow P(1/X < 1/f_{1-\alpha(m,n)}) = 1-\alpha \Longleftrightarrow$$
$$\Longleftrightarrow P(Y > 1/f_{1-\alpha(m,n)}) = \alpha \qquad \text{com} \qquad Y = 1/X \cap F_{(n,m)}.$$

Portanto

$$1/f_{\alpha(m,n)} = f_{1-\alpha(n,m)}.$$

Vejamos agora a utilização desta distribuição na Inferência Estatística

Teorema 3.7

Suponhamos que temos amostras aleatórias independentes, de tamanhos n_1 e n_2 seleccionadas de duas populações normais com variâncias σ_1^2 e σ_2^2, respectivamente. Então a v.a.

$$\boldsymbol{\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \cap F_{(n_1-1,n_2-1)}.} \tag{3.19}$$

Dem: Como se trata de amostras aleatórias seleccionadas de populações normais, sabemos que

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \cap \chi^2_{(n_1-1)} \qquad \text{e} \qquad \frac{(n_2-1)S_2^2}{\sigma_2^2} \cap \chi^2_{(n_2-1)}$$

e como são variáveis independentes, tem-se atendendo à definição (3.4) da distribuição F,

$$\frac{(n_1-1)S_1^2/(n_1-1)\sigma_1^2}{(n_2-1)S_2^2/(n_2-1)\sigma_2^2} = \boldsymbol{\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \cap F_{(n_1-1,n_2-1)}.}$$

Teoria da Estimação

Consideremos uma população da qual é conhecida a forma da função de distribuição, mas tendo parâmetros desconhecidos, cujo valor pretendemos estimar. Tal como já foi referido atrás, o problema da estimação divide-se em duas grandes áreas:

- estimação pontual;
- estimação por intervalos.

Estimação pontual

Seja X uma população cuja distribuição iremos supor conhecida, ou pelo menos admitida, mas dependendo de um parâmetro θ desconhecido. Entre outras hipóteses, θ pode ser, um **valor médio**, uma **proporção**, uma **variância**, etc.

Pretendemos então estimar θ a partir da informação dada por uma amostra aleatória, $(X_1, X_2, ..., X_n)$, extraída da referida população.

Definição 3.5

Uma estatística definida sobre a amostra, que sirva para estimar o valor do parâmetro θ, diz-se que é um **estimador de** θ e costuma representar-se por $\hat{\boldsymbol{\Theta}}$. Um estimador é portanto uma v.a. com uma dada distribuição.

Definição 3.6

Chama-se **estimativa** ou **estimativa pontual de** θ e é costume representar-se por $\hat{\boldsymbol{\theta}}$, o valor numérico de $\hat{\Theta}$ para uma amostra concreta.

Exemplo 3.2

Se para uma dada população X pretendemos estimar

$$\boldsymbol{\theta = \mu = E(X)}$$

podemos usar como **estimador**

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

sendo a estimativa obtida a partir de uma amostra concreta $(x_1, ..., x_n)$,

$$\boldsymbol{\hat{\mu} = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}.$$

Nesta disciplina iremos apresentar os estimadores dos parâmetros mais comuns e fazer inferência sobre esses parâmetros. Na tabela seguinte encontram-se esses parâmetros, os estimadores que iremos considerar e estimativas associadas.

Parâmetro a estimar	**Estimador**	**Estimativa**
μ	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
σ^2	$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
p	$\widehat{P} = \frac{X^{(*)}}{n}$	$\hat{p} = \frac{x^{(**)}}{n}$
$\mu_1 - \mu_2$	$\overline{X_1} - \overline{X_2}$	$\overline{x_1} - \overline{x_2}$
σ_1^2/σ_2^2	S_1^2/S_2^2	s_1^2/s_2^2
$p_1 - p_2$	$\widehat{P_1} - \widehat{P_2}$	$\hat{p_1} - \hat{p_2}$

com $^{(*)}X$ - v.a. que conta o número de sucessos em n provas de Bernoulli e $^{(**)}x$ - número observado de sucessos na amostra de dimensão n (concretização daquelas provas de Bernoulli).

Para um dado parâmetro desconhecido é possível propor mais de um estimador, sendo assim, uma vez definido um estimador pontual, põe-se uma pergunta natural: "Quão bom é o estimador obtido?".

Obviamente, pretendemos que o estimador forneça estimativas que se espere estarem muito próximas do valor de θ.

É necessário então dispor de critérios ou propriedades que permitam escolher um estimador como o "melhor" de entre outros possíveis estimadores do parâmetro.

Propriedades de um estimador

- Estimador **centrado**.

Definição 3.7

O estimador $\hat{\Theta}$ do parâmetro θ, diz-se **centrado** ou **não enviesado** se e só se $E(\hat{\Theta}) = \theta$.

Exemplo 3.3

Seja $(X_1, ..., X_n)$ uma amostra aleatória de dimensão n extraída de uma população com média μ. A média amostral $\overline{X}$ é um estimador centrado para μ.

De facto

$$E(\overline{X}) = \frac{1}{n}E(X_1 + ... + X_n) = \frac{1}{n}(n\mu) = \mu$$

Exemplo 3.4

A variância amostral S^2 definida em (3.3), baseada numa amostra aleatória de dimensão n, retirada de uma população com valor médio μ e variância σ^2 é um estimador centrado de σ^2.

Pretendemos então provar que $E(S^2) = \sigma^2$. Ora

$$E(S^2) = E\left[\frac{\sum(X_i - \overline{X})^2}{n-1}\right] = E\left[\frac{\sum[(X_i - \mu) - (\overline{X} - \mu)]^2}{n-1}\right]$$

$$E\left[\frac{\sum(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum(X_i - \mu) + \sum(\overline{X} - \mu)^2}{n-1}\right]$$

$$E\left[\frac{\sum(X_i - \mu)^2 - n(\overline{X} - \mu)^2}{n-1}\right] = E\left[\frac{\sum(X_i - \mu)^2}{n-1}\right] - E\left[\frac{n(\overline{X} - \mu)^2}{n-1}\right]$$

Como X_i são elementos da amostra aleatória, são independentes e

$$E(X_i) = \mu \qquad \text{e} \qquad Var(X_i) = E[(X_i - \mu)^2] = \sigma^2,$$

então,

$$E\left[\frac{\sum(X_i - \mu)^2}{n-1}\right] = \frac{\sum E[(X_i - \mu)^2]}{n-1} = \frac{n\ \sigma^2}{n-1} \qquad \text{e}$$

$$E\left[\frac{n(\overline{X} - \mu)^2}{n-1}\right] = \frac{n}{n-1}E(\overline{X} - \mu)^2 = \frac{n}{n-1}\frac{\sigma^2}{n} = \frac{\sigma^2}{n-1}$$

pois $E(\overline{X} - \mu)^2 = Var(\overline{X}) = \sigma^2/n$

Tem-se então

$$E(S^2) = \frac{n}{n-1}\ \sigma^2 - \frac{\sigma^2}{n-1} = \sigma^2$$

logo S^2 é um estimador centrado de σ^2.

É importante perceber o que significa **centrado**: dizer que $\hat{\Theta}$ é um estimador centrado de θ, significa que, fazendo k repetições de uma experiência os valores das estimativas resultantes $\hat{\theta_1}, \hat{\theta_2}, ..., \hat{\theta_k}$, variarão em torno de θ e a média daquelas k estimativas está razoavelmente próxima de θ. É igualmente importante, compreender o que o termo centrado não implica. De facto o termo **centrado** não implica que uma estimativa esteja muito próxima do verdadeiro valor do parâmetro a estimar.

Porém, regra geral apenas uma amostra é obtida quando se estuda uma dada população, porque de facto os estudos estatísticos não se repetem muitas vezes, por forma a que das estimativas obtidas se possa calcular a média. Para ter alguma garantia de que a estimativa obtida está próxima do verdadeiro valor do parâmetro θ, o ideal será o estimador usado ser, não só centrado, como ter também variância pequena. Deste modo, mesmo que os valores estimados flutuem em torno de θ, a variabilidade é pequena. Aparece aqui um novo conceito, o conceito de eficiência de um estimador.

- Estimador **eficiente**.

  Uma maneira de definir a eficiência de um estimador é pelo valor de $E[(\hat{\Theta} - \theta)^2]$. A esta quantidade chama-se **erro quadrático médio** e o ideal é que seja mínimo.

  Se $\hat{\Theta}$ é um estimador centrado, $E[(\hat{\Theta} - \theta)^2] = var(\hat{\Theta})$.

Definição 3.8

Um estimador $\hat{\Theta}$ diz-se **eficiente** se é centrado e tem variância menor ou igual à variância de qualquer outro estimador do mesmo parâmetro.

Observe-se que só $\hat{\Theta}_1$ e $\hat{\Theta}_2$ são centrados, mas o estimador $\hat{\Theta}_1$ tem menor variância do que $\hat{\Theta}_2$ portanto é mais eficiente.

Existem outras propriedades que um estimador deve possuir, mas que não abordaremos neste curso (consultar Murteira – 2ºvolume, Tiago de Oliveira – 2ºvolume ou ainda Mood, Graybill e Boes).

Referimos até aqui alguns estimadores, algumas propriedades que eles deverão possuir, porém nunca falámos em métodos de estimação, i.e., procedimentos para obter estimadores de parâmetros. Não nos é possível estudar tais métodos no decorrer deste curso, ficando porém aqui a referência aos mais importantes, que se encontram expostos na bibliografia indicada:

- método dos momentos;
- método da máxima verosimilhança;
- método dos mínimos quadrados.

Estimação por intervalos

Até aqui considerámos o problema da **estimação pontual** e fizemos referência a métodos de determinar estimadores pontuais de parâmetros desconhecidos.

Porém, a indicação de um único valor como estimativa de um parâmetro não nos dá nenhuma informação sobre a precisão de tal valor. Por isso em muitas situações, interessa-nos dar uma medida de erro, ϵ, para indicar que o verdadeiro valor do parâmetro está muito provavelmente entre $\hat{\theta} - \epsilon$ e $\hat{\theta} + \epsilon$.

Definição 3.9

Um **intervalo de estimação** de um parâmetro θ é um intervalo da forma $\boldsymbol{\hat{\theta}_1 < \theta < \hat{\theta}_2}$, onde $\hat{\theta}_1$ e $\hat{\theta}_2$ são dois valores assumidos pelo estimador $\hat{\Theta}$, face a uma amostra concreta.

A medida da confiança com que aquele intervalo conterá o verdadeiro valor do parâmetro é feita em termos de probabilidades. Para isso é necessário conhecer, pelo menos aproximadamente, a distribuição de amostragem de $\hat{\Theta}$. Assim, fixada uma probabilidade que regra geral é elevada, interessa-nos determinar o intervalo aleatório que com a referida probabilidade contém o parâmetro que se pretende estimar.

A construção do intervalo baseia-se na determinação de duas variáveis aleatórias, $\hat{\Theta}_1$ e $\hat{\Theta}_2$ tais que fixado um valor α, com $0 < \alpha < 1$,

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

A $\boldsymbol{1 - \alpha}$ chama-se **coeficiente de confiança** e a $\boldsymbol{\alpha}$ chama-se **nível de significância**.

Definição 3.10

O intervalo $\boldsymbol{\hat{\theta}_1 < \theta < \hat{\theta}_2}$, calculado para uma amostra concreta chama-se **intervalo de confiança** a $(1 - \alpha) \times 100\%$.

Note-se que o intervalo $]\hat{\Theta}_1, \hat{\Theta}_2[$ é um intervalo aleatório e o que vamos determinar são $\hat{\theta}_1$ e $\hat{\theta}_2$, realizações de $\hat{\Theta}_1$ e $\hat{\Theta}_2$, respectivamente.

Quando temos $P(\hat{\Theta}_1 < \theta < \hat{\Theta}_2) = 0.95$, significa que temos uma confiança de 95% de que o nosso intervalo contenha o verdadeiro valor do parâmetro, ou ainda, significa que esperamos que em cerca de 95% dos intervalos $]\hat{\theta}_1, \hat{\theta}_2[$, obtidos a partir de amostras extraídas da população, o valor θ esteja lá incluído e nos restantes 5%, não esteja.

Diferentes amostras, conduzem a diferentes valores de $\hat{\theta}$, produzindo portanto diferentes intervalos de confiança para o parâmetro θ.

Quanto maior for o intervalo, maior é o grau de confiança que temos de que ele contenha o verdadeiro valor do parâmetro desconhecido, mas não há interesse em ter um intervalo muito largo. O ideal seria um intervalo curto com probabilidade elevada.

Como já dissémos, para construir um intervalo de confiança para um parâmetro θ, é necessário encontrar um variável aleatória cuja expressão contenha θ e cuja distribuição seja conhecida, pelo menos aproximadamente.

Iremos considerar intervalos de confiança para o valor médio, variância, diferença de valores médios, quociente de variâncias e para proporções.

Intervalos de confiança para o valor médio

a) Caso de uma população normal com σ conhecido

Seja $(X_1, ..., X_n)$ uma amostra aleatória retirada de uma população X, com distribuição normal com valor médio μ e variância σ^2 conhecida. Vimos já que um estimador para a média da população μ, é dado pela estatística $\overline{X}$, que no nosso caso tem distribuição normal com valor médio μ e variância σ^2/n.

Sendo assim tem-se

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \cap N(0,1) \tag{3.20}$$

É a v.a. Z, em cuja expressão intervém μ como único parâmetro desconhecido, que vai ser utilizada para a determinação do intervalo de confiança para μ.

Fixado o nível de significância α e designando por $z_{\alpha/2}$ o valor da v.a. Z tal que $P(Z > z_{\alpha/2}) = \alpha/2$, tem-se

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \Leftrightarrow P(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < z_{\alpha/2}) = 1 - \alpha$$

donde, após pequenos cálculos

$$P(\overline{X} - z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Para uma amostra concreta $(x_1, ..., x_n)$, seja $\overline{x}$ o valor da estatística $\overline{X}$, tem-se então

o intervalo a $(1-\alpha)\times 100\%$ de confiança para μ numa população normal com σ conhecido é

$$\boxed{\overline{\boldsymbol{x}} - \boldsymbol{z_{\alpha/2}}\ \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} < \boldsymbol{\mu} < \overline{\boldsymbol{x}} + \boldsymbol{z_{\alpha/2}}\ \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}} \tag{3.21}$$

b) Caso de uma população normal com σ desconhecido

Na verdade, na maioria dos casos não conhecemos a variância da população da qual pretendemos estimar a média. Se a amostra é grande, pode substituir-se σ por s em (3.21) e obter assim um intervalo de confiança para μ. Tal procedimento resulta da distribuição de amostragem (3.8).

Porém, se a amostra de que dispomos é pequena, a distribuição de $(\overline{X}-\mu)/(S/\sqrt{n})$, a variável que resulta de substituir σ por um seu estimador, já não é normal. Porém, vimos que se a distribuição da população é normal, a v.a. $T = (\overline{X}-\mu)/(S/\sqrt{n})$ tem distribuição t-Student com $(n-1)$ g.l.

A construção dos intervalos de confiança segue aqui um procedimento análogo ao considerado em a).

Sendo $t_{\alpha/2}$ o valor da v.a. T tal que $P(T > t_{\alpha/2}) = \alpha/2$, tem-se

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P(-t_{\alpha/2} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}) = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P(\overline{X} - t_{\alpha/2}\ S/\sqrt{n} < \mu < \overline{X} + t_{\alpha/2}\ S/\sqrt{n}) = 1-\alpha$$

.

Portanto,

o intervalo de confiança para μ a $(1-\alpha)\times 100\%$, no caso de σ^2 desconhecido e a população ser normal é

$$\boxed{\overline{\boldsymbol{x}} - \boldsymbol{t_{\alpha/2(n-1)}\frac{s}{\sqrt{n}}} < \boldsymbol{\mu} < \overline{\boldsymbol{x}} + \boldsymbol{t_{\alpha/2(n-1)}\frac{s}{\sqrt{n}}}} \tag{3.22}$$

Exemplo 3.5

Uma amostra de 10 medidas do diâmetro de uma esfera apresenta uma média $\overline{x} = 4.38$ e um desvio padrão $s = 0.06$. Determine um intervalo de confiança a 99% para o valor do diâmetro médio da esfera, supondo admissível que as medidas dos diâmetros seguem uma lei aproximadamente normal.

Resolução:

Como $n = 10$ é pequeno e não conhecemos o valor do desvio padrão da população das medidas dos diâmetros, o intervalo de confiança para o verdadeiro valor do diâmetro é dado por (3.22), portanto como $\alpha = 0.01$, tem-se $t_{\alpha/2(9)} = t_{0.005(9)} = 3.25$, donde o intervalo de confiança pedido é

$$]4.38 - 3.25(0.06/\sqrt{10}), 4.38 + 3.25(0.06/\sqrt{10})[=]4.318, 4.442[.$$

c) Caso de uma população não normal

Se a amostra de que dispomos tiver dimensão elevada, o Teorema Limite Central permite-nos resolver a questão da determinação dos intervalos de confiança.

De facto, se n grande, tem-se pelo Teorema Limite Central que, qualquer que seja a forma da distribuição da qual se retirou a amostra aleatória

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

então analogamente ao que foi feito na alínea anterior

o intervalo a $(1-\alpha)\times 100\%$ **de confiança para** μ **numa população qualquer com** σ **conhecido, desde que a dimensão da amostra,** n**, seja grande, é** dado por (3.21).

Se σ não é conhecido, mas a dimensão da amostra é grande pode ainda substituir-se na expressão do intervalo de confiança (3.21), σ por s (desvio padrão da amostra). Sendo assim temos

o intervalo a $(1-\alpha)\times 100\%$ **de confiança para** μ **numa população qualquer com** σ **desconhecido e com** n **grande, é**

$$\boxed{\overline{x} - z_{\alpha/2}\,\frac{s}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\,\frac{s}{\sqrt{n}}} \tag{3.23}$$

Exemplo 3.6

A média e o desvio padrão da classificação média de uma amostra aleatória de 100 alunos de uma dada escola são 2.6 e 0.3, respectivamente. Determine um intervalo de confiança a 95% para a classificação média de todos os alunos daquela escola.

Resolução:

Como $n = 100$ é grande, não se exige o conhecimento da distribuição subjacente às classificações; o desvio padrão também não é conhecido mas podemos usar s, vindo então o intervalo de confiança dado por (3.23) com $\overline{x} = 2.6$ e $s = 0.3$.

Para $\alpha = 0.05$, tem-se $z_{\alpha/2} = z_{0.025} = 1.96$; o intervalo de confiança a 95% para a classificação média é então

$$2.6 - 1.96\,\frac{0.3}{\sqrt{100}} < \mu < 2.6 + 1.96\,\frac{0.3}{\sqrt{100}}$$

$$2.54 < \mu < 2.66.$$

Um outro problema que se levanta frequentemente é o de saber que dimensão deverá ter a amostra para assegurar que o erro cometido ao estimar μ por $\overline{x}$ seja inferior a uma quantidade especificada ϵ.

Ora, considerando o intervalo de confiança para μ

$$\overline{x} - z_{\alpha/2}\sigma/\sqrt{n} < \mu < \overline{x} + z_{\alpha/2}\sigma/\sqrt{n}$$

se usarmos $\overline{x}$ como uma estimativa de μ, podemos estar $(1-\alpha)\times 100\%$ confiantes que o **erro cometido será menor que** $z_{\alpha/2}\sigma/\sqrt{n}$.

Sendo assim, para termos uma estimativa de μ com erro inferior a ϵ devemos escolher n tal que $z_{\alpha/2}\sigma/\sqrt{n} \leq \epsilon$, i.e.,

$$n \geq \left(\frac{z_{\alpha/2}\ \sigma}{\epsilon}\right)^2 \tag{3.24}$$

Porém esta igualdade só se pode aplicar se conhecermos a variância da população da qual se vai seleccionar a amostra. Mas, o que acontece na realidade é que na grande maioria dos casos não se conhece; então o que se deverá fazer é tomar uma amostra preliminar de $n > 30$, usá-la para obter uma estimativa s de σ e então entrar com este valor na fórmula (3.24).

Intervalos de confiança para a diferença entre duas médias populacionais

a) Caso de populações normais com variâncias conhecidas

Suponhamos duas populações normais $X_1 \cap \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \cap \mathcal{N}(\mu_2, \sigma_2)$.

Um estimador para a diferença das médias $\mu_1 - \mu_2$ é dado por $\overline{X_1} - \overline{X_2}$. Então, para ter uma estimativa pontual para $\mu_1 - \mu_2$ seleccionamos duas amostras aleatórias independentes uma de cada população com dimensões n_1 e n_2, respectivamente. Como vimos já,

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

Sendo assim, analogamente ao que fizemos atrás, fixado o nível de significância α,

o intervalo de confiança a $(1-\alpha)\times 100\%$ **para** $\mu_1 - \mu_2$ **no caso de populações normais com variâncias conhecidas das quais foram extraídas amostras independentes é**

$$\boxed{(\overline{x_1} - \overline{x_2}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.25}$$

b) Caso de populações normais, variâncias desconhecidas

O que acabámos de dizer sobre a diferença de duas médias só é aplicável no caso de σ_1^2 e σ_2^2 conhecidos ou estimados, no caso de grandes amostras. De facto se σ_1^2 e σ_2^2 são desconhecidas e n_1 e n_2 grandes, podemos substituir em (3.25) σ_1^2 por s_1^2 e σ_2^2 por s_2^2, sem haver alteração significativa na precisão do intervalo de confiança.

Se no entanto as amostras são pequenas, podemos contruir intervalos de confiança para a diferença dos valores médios se forem verificadas as seguintes hipóteses:

1. ambas as distribuições normais;

2. as variâncias das populações, embora desconhecidas possam ser supostas iguais (veremos mais adiante um procedimento que permite estudar a admissibilidade desta hipótese).

Como ficou dito atrás, teorema 3.5, se as variâncias embora desconhecidas, puderem ser supostas iguais, seja $\sigma_1^2 = \sigma_2^2 = \sigma^2$, a v.a.

$$\frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \cap t_{(n_1+n_2-2)} \tag{3.26}$$

com

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

A construção de um intervalo de confiança para $\mu_1 - \mu_2$ segue os procedimentos já atrás referidos: dado o nível de significância α, tem-se

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1-\alpha \;\Leftrightarrow$$

$$\Leftrightarrow P\left(-t_{\alpha/2} < \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{\alpha/2}\right) = 1-\alpha$$

donde, após pequenos cálculos se tem

o intervalo de confiança a $(1-\alpha)\times 100\%$ **para** $\mu_1 - \mu_2$ **no caso de amostras de pequena dimensão retiradas de populações normais com variâncias desconhecidas mas supostas iguais,** $\sigma_1^2 = \sigma_2^2 = \sigma^2$ **é**

$$\boxed{(\overline{\boldsymbol{x_1}} - \overline{\boldsymbol{x_2}}) - \boldsymbol{t_{\alpha/2}}\ \boldsymbol{s_p}\sqrt{\tfrac{1}{\boldsymbol{n_1}} + \tfrac{1}{\boldsymbol{n_2}}} < \boldsymbol{\mu_1 - \mu_2} < (\overline{\boldsymbol{x_1}} - \overline{\boldsymbol{x_2}}) + \boldsymbol{t_{\alpha/2}}\ \boldsymbol{s_p}\sqrt{\tfrac{1}{\boldsymbol{n_1}} + \tfrac{1}{\boldsymbol{n_2}}}} \tag{3.27}$$

Exemplo 3.7

Num processo químico pretende comparar-se o efeito de dois catalizadores sobre o produto obtido.

Utilizaram-se 12 quantidades iguais de produto nas quais se usou o catalizador 1 e 10 outras quantidades nas quais se usou o catalizador 2. Quando foi usado o catalizador 1, obteve-se uma média de 85 com desvio padrão 4, enquanto para o catalizador 2 se obteve média 81 com desvio padrão 5.

Determinar um intervalo de confiança a 90% para a diferença entre o efeito médio dos dois catalizadores, supondo que as populações são aproximadamente normais.

Resolução:

Para já admitamos poder aceitar-se a igualdade das variâncias (mais tarde teremos oportunidade de verificar se esta hipótese é admissível).

Como as amostras têm dimensão pequena e é suposto terem sido retiradas de populações aproximadamente normais, o intervalo de confiança é dado por (3.27), com

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 20.05$$

Como $t_{\alpha/2(20)} = t_{0.05(20)} = 1.725$, o intervalo de confiança a 90% para $\mu_1 - \mu_2$ é então

$$4 - (1.725)(4.478)\sqrt{1/12+1/10} < \mu_1 - \mu_2 < 4 + (1.725)(4.478)\sqrt{1/12+1/10}$$

ou seja,

$$0.69 < \mu_1 - \mu_2 < 7.31.$$

As hipóteses da normalidade das populações e da igualdade das variâncias são bastante restritivas e muitas vezes não são verificadas nas aplicações. Porém, vários autores mostraram que o grau de confiança nos intervalos não é seriamente afectado para pequenos afastamentos da normalidade. No caso de as populações serem normais, a hipótese da igualdade das variâncias pode ser não verificada continuando a ter-se bons resultados, desde que estejam a ser consideradas amostras do mesmo tamanho. Portanto, ao planear-se uma experiência, é de fazer um esforço para obter amostras com a mesma dimensão.

Se no entanto as amostras tiverem dimensões diferentes e não for admissiível supor as variâncias iguais é ainda possível contruir intervalos de confiança aproximados para a diferença das médias das duas populações considerando uma estatística T' aproximada, em que o número de graus de liberdade tem de ser estimado a partir dos dados. Esta variável foi já estudada atrás, (3.15), vindo-nos portanto

o intervalo de confiança a $(1-\alpha)\times 100\%$ **para** $\mu_1 - \mu_2$ **no caso de amostras de pequena dimensão retiradas de populações normais com variâncias diferentes**

$$\boxed{\overline{x_1} - \overline{x_2} - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{x_1} - \overline{x_2} + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \quad (3.28)$$

onde $t_{\alpha/2}$ é o valor de uma variável aleatória T', com distribuição t com γ g.l., definido em (3.16).

c) Caso de populações quaisquer e amostras de dimensão grande

Quando as populações X_1 e X_2 não são normais mas as dimensões das amostras retiradas independentemente de cada uma delas são grandes, (como regra prática n_1 e n_2 ambos maiores que 30) o intervalo de confiança (3.25) dá uma aproximação bastante boa quando as variâncias de cada uma das variáveis σ_1^2 e σ_2^2 são conhecidas.

Se σ_1^2 e σ_2^2 são desconhecidas e n_1 e n_2 grandes, podemos substituir em (3.25) σ_1^2 por s_1^2 e σ_2^2 por s_2^2, sendo razoável a precisão do intervalo de confiança.

Exemplo 3.8

Foi feito um teste a 50 raparigas e 75 rapazes. As raparigas tiveram uma pontuação média de 76 com desvio padrão 6 e os rapazes tiveram uma pontuação média de 82 com desvio padrão 8.

Construindo um intervalo de confiança a 95% para a diferença da pontuação média entre os rapazes e as raparigas, interprete o resultado obtido.

Resolução:

Designando por X_1 a população dos rapazes e X_2 a população das raparigas, temos $n_1 = 75$ e $n_2 = 50$. Como as amostras têm dimensões grandes não é necessário exigir a normalidade das populações a estudar, assim como o conhecimento das variâncias de cada uma delas.

Dado que $\overline{x_1} = 82$, $\overline{x_2} = 76$, $s_1 = 8$ e $s_2 = 6$, o intervalo de confiança a 95% para a diferença da pontuação média entre os rapazes e as raparigas é

$$(82-76) - 1.96\sqrt{64/75+36/50} < \mu_1 - \mu_2 < (82-76) + 1.96\sqrt{64/75+36/50}$$

donde o intervalo de confiança a 95% é

$$3.54 < \mu_1 - \mu_2 < 8.46.$$

Dado que o intervalo de confiança contém apenas valores positivos para $\mu_1 - \mu_2$, podemos dizer que, com uma confiança de 95%, as pontuações médias obtidas pelos rapazes são superiores às obtidas pelas raparigas.

e) Caso de amostras emparelhadas

Até aqui considerámos intervalos de confiança para a diferença das médias de duas populações, quando as amostras obtidas de cada uma dessas populações eram independentes.

Porém, em muitas situações tal não acontece. É o caso de observações de uma dada experiência ocorrendo aos pares, sendo o 1º elemento do par de uma amostra e o 2º elemento da outra. Por exemplo, ao ser feito um teste sobre uma nova dieta em n indivíduos, os pesos registados antes e após o tempo de teste constituem as observações de cada uma das nossas amostras. Como se vê, as observações em cada uma das amostras estão relacionadas (emparelhadas) pelo indivíduo.

Uma outra situação é a de haver factores estranhos ao fenómeno em estudo, que causam diferenças significativas nas médias. É o caso de se pretender fazer um ensaio de dois adubos na produção de trigo. Se para isso escolhermos vários talhões aplicando numa parte deles um dos adubos e na outra parte o outro adubo, pode acontecer que ao compararmos as produções médias, as diferenças verificadas se devam não aos adubos mas a condições diferentes de solo e clima a que os talhões estejam sujeitos.

A influência de factores estranhos pode reduzir-se considerando aquilo a que chamamos observações emparelhadas. De facto, ao compararmos dois tratamentos (o termo tratamentos usa-se regra geral para designar coisas que se pretendem comparar), é desejável que as unidades experimentais (são os elementos básicos a que são aplicados os tratamentos) sejam o mais homogéneas possível, por forma que, as diferenças que se verifique existirem entre os dois grupos em estudo possam ser atribuídas a diferenças nos tratamentos. Se existirem certos factores variando nas unidades experimentais e que possam influenciar a resposta , tais factores podem obscurecer as diferenças reais devidas ao efeito dos tratamentos. O requerimento da homogeneidade entre as duas amostras, que poderia resolver este problema, pode impor restrições grandes ao número de elementos a incluir nas amostras.

Por exemplo, para comparar dois analgésicos pode ser impraticável obter um número razoável de doentes que tenham as mesmas características: idade, sexo, condições físicas gerais e o mesmo grau de severidade de doença.

Além de ser muitas vezes impraticável obter amostras com mesmas características, também não tem interesse o estudo feito com um grupo tão restrito. Um estudo com muito mais interesse consistirá em considerar uma variedade grande de doentes, com idades, sexos, condições gerais diferentes e para aplicar os tratamentos tentar formar pares de indivíduos o mais semelhantes possível em cada par, mas variando de par para par. Em cada par escolhe-se aleatoriamente um elemento para receber um tratamento ficando o outro com outro tratamento.

Surge então um conceito muito importante, é o conceito de **bloco** que se mostra fundamental por permitir uma situação de compromisso entre o requesito da homogeneidade e da diversidade das unidades experimentais. O procedimento consiste então em formar grupos ou blocos por forma que em cada bloco as referidas unidades sejam homogéneas e nos diferentes blocos sejam bastante heterogéneas. Temos então amostras a

que chamamos **amostras emparelhadas**.

Depois em cada bloco é atribuído aleatoriamente o tratamento 1 a uma das unidades, ficando a outra para o tratamento 2. Este processo permite a comparação efectiva dentro de cada bloco e a diversidade de condições possíveis de existir entre os blocos.

No caso dos adubos, a experiência deve então ser planeada de modo a formar talhões emparelhados, i.e., o mais semelhantes possível quanto ao tipo de solo, humidade, exposição ao sol, etc, e num deles aplicar um adubo e no outro o outro adubo.

O emparelhamento serve portanto para remover fontes de variação que possam existir entre as duas populações em estudo e que obscureçam o efeito dos tratamentos. A diferença nas respostas dos dois grupos será assim atribuída às diferenças nos tratamentos e não aos grupos.

Consideremos a amostra emparelhada (X_i, Y_i) $(i = 1, ..., n)$, composta por pares de observações independentes, mas dentro de cada par X_i e Y_i são correlacionadas.

Designemos por $(D_1, D_2, ..., D_n)$, a amostra aleatória das diferenças, i.e.,

$$D_1 = X_1 - Y_1; \qquad D_2 = X_2 - Y_2; \qquad ... \qquad D_n = X_n - Y_n.$$

As variáveis aleatórias $D_1, D_2, ..., D_n$ são independentes.

Supondo que $D_1, D_2, ..., D_n$ são observações de uma população normal com valor médio μ_D e variância σ_D^2, desconhecida, pelas propriedades do valor médio vê-se que $\mu_X - \mu_Y = \mu_D$, por conseguinte a inferência sobre a diferença das médias entre as duas populações é o mesmo que fazer inferência sobre a média da diferença.

Naturalmente, como estimador pontual de σ_D^2 temos S_D^2 e como estimador pontual de μ_D temos $\overline{D}$ sendo

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \qquad \text{e} \qquad S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2.$$

Para determinarmos um intervalo de confiança para μ_D, uma vez fixado um nível α, temos de considerar a v.a.

$$\frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \cap t_{(n-1)} \tag{3.29}$$

donde

o intervalo de confiança a $(1 - \alpha) \times 100\%$ **para** μ_D**, onde as diferenças** $D_i = X_i - Y_i$ **são supostas normais e independentes é**

$$\boxed{\overline{\boldsymbol{d}} - \boldsymbol{t_{\alpha/2}\frac{s_D}{\sqrt{n}}} < \boldsymbol{\mu_D} < \overline{\boldsymbol{d}} + \boldsymbol{t_{\alpha/2}\frac{s_D}{\sqrt{n}}}} \tag{3.30}$$

Observação: Tal como tem vindo a ser referido se n grande não é necessária a hipótese da normalidade para a população das diferenças.

Exemplo 3.9

Numa investigação médica pretende-se estudar o efeito de determinado medicamento na redução da tensão arterial. Para isso registaram-se as tensões arteriais de 15 pessoas e posteriormente usaram o referido medicamento durante seis meses, no fim dos quais as suas tensões arteriais foram de novo registadas. Que inferências é possível efectuar a partir dos dados obtidos?

Indivíduo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Antes (x_i)	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
Depois (y_i)	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74
$d_i = x_i - y_i$	2	8	10	6	18	10	4	26	18	-8	0	32	0	-4	10

Resolução:

Estamos perante observações emparelhadas, por pessoa, feitas antes e após o uso do medicamento.

Tem-se

$$\overline{d} = \frac{1}{15}\sum d_i = 8.8\ , \qquad s_d = \sqrt{\frac{1}{14}\sum (d_i - \overline{d})^2} = 10.98.$$

Supondo que as diferenças são valores de uma amostra aleatória extraída de uma população normal, um intervalo de confiança a 95% para a diferença média, obtido a partir de (3.30) é

$$]8.8 - 2.145\ \frac{10.98}{\sqrt{15}}, 8.8 + 2.145\ \frac{10.98}{\sqrt{15}}[\Leftrightarrow]2.72, 14.88[$$

Sendo assim, podemos dizer que, com 95% de confiança, a redução média está entre 2.72 e 14.88 e como o intervalo de confiança inclui apenas valores positivos há fortes razões para acreditar que há redução da tensão arterial.

Observações:

Na altura de planear uma experiência para comparar dois tratamentos, por vezes teremos de decidir entre fazer amostras emparelhadas ou amostras independentes.

Vejamos algumas considerações:

- Ao considerarmos **amostras emparelhadas** e dispondo de n pares de observações, a variável a usar na construção do intervalo de confiança tem distribuição t com $(n-1)$ g.l.; se as **amostras** fossem consideradas **independentes** com n observações cada, a variável a usar teria distribuição t com $(2n-2)$ g.l.

  Vemos então que amostras emparelhadas resultam numa perda de graus de liberdade, o que acarreta um valor maior para $t_{\alpha/2}$ e consequentemente um intervalo de confiança maior.

Mas, se nas unidades experimentais existem condições que possam influenciar a resposta e que produzam uma grande variabilidade nos dois grupos em estudo, é pela amostragem emparelhada que devemos optar.

- Sempre que estejamos a trabalhar com unidades experimentais que sejam homogéneas ou cuja heterogeneidade não se pode relacionar a factores identificáveis é então preferível fazer amostras independentes.

Intervalo de confiança para a variância

Dada uma amostra aleatória $(X_1, X_2, ..., X_n)$, vimos já que um estimador da variância σ^2 da população era $S^2 = \sum(X_i - \overline{X})^2/(n-1)$. Por outro lado sabemos que se a amostra for extraída de uma população normal de parâmetros μ e σ,

$$\frac{(n-1)S^2}{\sigma^2} \cap \chi^2_{(n-1)}.$$

É esta variável aleatória que nós vamos utilizar para construir um intervalo de confiança para σ^2.

Fixado um nível α, e sendo $\chi^2_{\alpha/2}$ o valor de uma v.a. com distribuição $\chi^2_{(n-1)}$ tal que $P(\chi^2_{(n-1)} > \chi^2_{\alpha/2}) = \alpha/2$

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right] = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2(n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2(n-1)}}\right] = 1-\alpha$$

Então **o intervalo de confiança a** $(1-\alpha)\times 100\%$ **para a variância de uma população normal é**

$$\boxed{\frac{(n-1)s^2}{\chi^2_{\alpha/2(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2(n-1)}}} \tag{3.31}$$

Intervalo de confiança para o quociente entre duas variâncias

Suponhamos que temos duas populações normais $X \cap \mathcal{N}(\mu_1, \sigma_1)$ e $Y \cap \mathcal{N}(\mu_2, \sigma_2)$ das quais retiramos duas amostras independentes de dimensões n_1 e n_2, respectivamente.

Pretendemos comparar as variâncias das duas populações, determinando um intervalo de confiança para $\dfrac{\sigma_1^2}{\sigma_2^2}$.

Vimos, teorema 3.7, que a variável aleatória $\dfrac{\sigma_2^2\ S_1^2}{\sigma_1^2\ S_2^2}$ tem distribuição $F_{(n_1-1,n_2-1)}$

Para construir um intervalo de confiança, para um nível de significância α, basta então ter em conta que

$$P\left[f_{1-\alpha/2} < \frac{\sigma_2^2\ S_1^2}{\sigma_1^2\ S_2^2} < f_{\alpha/2}\right] = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left[\frac{S_1^2}{S_2^2\ f_{\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2\ f_{1-\alpha/2}}\right] = 1-\alpha.$$

Portanto, tendo ainda em conta o resultado (3.18) para a distribuição F,

o intervalo de confiança para $\frac{\sigma_1^2}{\sigma_2^2}$, razão das variâncias de duas populações normais das quais foram extraídas duas amostras independentes, é

$$\boxed{\frac{s_1^2}{s_2^2\ f_{\alpha/2;(n_1-1,n_2-1)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2\ f_{\alpha/2;(n_2-1,n_1-1)}}{s_2^2}} \tag{3.32}$$

Mais uma vez é de referir que estamos a supor as populações normais. A violação desta hipótese pode levar-nos a conclusões incorrectas. Porém, tem-se verificado que tal tal problema é minimizado se as amostras tiverem a mesma dimensão.

Intervalos de confiança para proporções.

Dada uma população X com distribuição binomial de parâmetros (n,p), regra geral p é o parâmetro desconhecido. Um estimador de p é

$$\hat{P} = \frac{X}{n}, \tag{3.33}$$

onde X, designando o número de sucessos em n provas, tem então distribuição $B(n,p)$

Como sabemos $E(X) = np$ e $Var(X) = npq$, portanto

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = p \qquad \hat{P} \text{ é então um estimador centrado de } p \text{ e}$$

$$Var(\hat{P}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Se estivermos nas condições de ser possível aproximar a v.a. binomial pela normal, teorema de De Moivre, sabemos que

$$X \sim \mathcal{N}(np, \sqrt{npq})$$

.

Portanto

$$\frac{X}{n} \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right) \Leftrightarrow Z = \sqrt{n}\,\frac{X/n - p}{\sqrt{pq}} \sim \mathcal{N}(0,1) \tag{3.34}$$

Usando esta variável aleatória podemos determinar um intervalo de confiança para a proporção p, atendendo a que

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

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

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

Observe-se porém que p é desconhecido, por isso os limites da expressão anterior não se conseguem obter. No entanto, como toda esta construção só é válida para n grande, nos limites da expressão anterior pode substituir-se p por $\hat{p} = x/n$. Temos então

um intervalo de confiança para p a $(1-\alpha)\times 100\%$ de confiança para amostras de dimensão grande é

$$\boxed{\hat{\boldsymbol{p}} - \boldsymbol{z_{\alpha/2}}\sqrt{\frac{\hat{\boldsymbol{p}}(\boldsymbol{1}-\hat{\boldsymbol{p}})}{\boldsymbol{n}}} < \boldsymbol{p} < \hat{\boldsymbol{p}} + \boldsymbol{z_{\alpha/2}}\sqrt{\frac{\hat{\boldsymbol{p}}(\boldsymbol{1}-\hat{\boldsymbol{p}})}{\boldsymbol{n}}}} \tag{3.35}$$

com $\hat{p} = x/n$.

Exemplo 3.10

Numa amostra de 400 famílias extraída de entre os habitantes de Lisboa, verificou-se que 140 tinham televisão via satélite. Determine um intervalo a 95% de confiança para a proporção de famílias possuindo TV satélite.

Resolução:

Designando por X a v.a. que conta o número de famílias possuindo TV satélite na amostra seleccionada, trata-se de uma variável com distribuição hipergeométrica. Mas $n = 400$ é muito pequeno relativamente à dimensão da população (todas as famílias de Lisboa) e por isso é possível fazer uma aproximação à binomial. Temos então a v.a. X aproximadamente binomial de parâmetros $(n = 400, p)$. O parâmetro p, desconhecido, tem como estimativa $\hat{p} = 140/400 = 0.35$.

O intervalo de confiança, para a proporção pedida, obtido a partir de (3.35) é então dado por

$$0.35 - 1.96\sqrt{\frac{(0.35)\ (0.65)}{400}} < p < 0.35 + 1.96\sqrt{\frac{(0.35)\ (0.65)}{400}}$$

ou seja

$$0.303 < p < 0.397.$$

O intervalo de confiança (3.29) pode escrever-se sob a forma

$$|p - \hat{p}| < z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

permitindo-nos dar uma ideia do erro cometido ao estimar p por $\hat{p}$.

Teorema 3.8

Se $\hat{p}$ é usado como uma estimativa de p, podemos ter $(1-\alpha)\times 100\%$ de confiança que o erro cometido ao usarmos $\hat{p}$ em vez de p é inferior a $z_{\alpha/2}\ \sqrt{\hat{p}\hat{q}/n}$.

Aplicando este teorema podemos determinar a dimensão da amostra que se deve tomar por forma que o erro cometido ao estimar p por $\hat{p}$ seja inferior a um certo valor especificado ϵ. Assim, basta fazer

$$z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq \epsilon \Leftrightarrow n \geq \frac{z^2_{\alpha/2}\hat{p}\hat{q}}{\epsilon^2}.$$

Podemos enunciar então o

Teorema 3.9

Se $\hat{p}$ é usado como estimativa de p, o erro cometido é inferior a ϵ com $(1-\alpha)\times 100\%$ de confiança, quando a dimensão da amostra fôr

$$n \geq \frac{z^2_{\alpha/2}\hat{p}\hat{q}}{\epsilon^2}. \tag{3.36}$$

Observe-se porém que, na determinação do tamanho da amostra necessária para ter uma estimativa de p com um erro inferior a uma quantidade especificada, usa-se $\hat{p}$, calculado a partir da amostra. Regra geral este problema é resolvido do seguinte modo:

– Se for possível ter uma estimativa grosseira de p sem necessidade de recolher uma amostra usa-se essa estimativa na expressão que nos dá o valor de n.

– na falta dessa estimativa, extrai-se uma primeira amostra de dimensão $n > 30$, estima-se $\hat{p}$ e depois este valor entra na expressão que permite determinar n.

Note-se ainda que, independentemente do grau de confiança é possível ter um limite inferior para n, tendo em conta que

$$\hat{p}\hat{q} = \hat{p}(1-\hat{p}) = \hat{p} - \hat{p}^2 = 1/4 - (\hat{p}^2 - \hat{p} + 1/4) = 1/4 - (\hat{p} - 1/2)^2 \leq 1/4$$

Sendo assim o máximo valor que $\hat{p}\hat{q}$ podem tomar é 1/4. Substituindo na expressão (3.36) obtemos um valor para a dimensão da amostra que se deve recolher para termos uma estimativa com um erro inferior a ϵ, independentemente do grau de confiança

$$n \geq \frac{z^2_{\alpha/2}}{4\epsilon^2}. \tag{3.37}$$

Intervalo de confiança para a diferença entre duas proporções

Consideremos duas populações binomiais das quais retiramos duas amostras aleatórias independentes, suficientemente grandes, de dimensões n_1 e n_2. Defina-se as variáveis aleatórias

$$X \cap B(n_1, p_1) \qquad \text{e} \qquad Y \cap B(n_2, p_2)$$

Pretendemos determinar um intervalo de confiança para $p_1 - p_2$. Sejam

$$\hat{P}_1 = X/n_1 \qquad \text{e} \qquad \hat{P}_2 = Y/n_2$$

estimadores de p_1 e p_2, respectivamente. Como estamos a considerar n_1 e n_2 grandes, temos

$$\hat{P}_1 \sim \mathcal{N}\left(p_1, \sqrt{\frac{p_1\ q_1}{n_1}}\right) \qquad \text{e} \qquad \hat{P}_2 \sim \mathcal{N}\left(p_2, \sqrt{\frac{p_2\ q_2}{n_2}}\right).$$

Como as amostras são independentes as v.a.s X e Y são independentes e portanto $\hat{P}_1$ e $\hat{P}_2$ também independentes , logo

$$\hat{P}_1 - \hat{P}_2 \sim \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1\ q_1}{n_1} + \frac{p_2\ q_2}{n_2}}\right).$$

Procedendo como temos feito até aqui temos

o intervalo de confiança a $(1-\alpha)\times 100\%$ **para a diferença** $p_1 - p_2$ **de duas populações binomiais das quais se retiraram amostras independentes e de dimensão elevada é**

$$\boxed{(\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2) - \boldsymbol{z}_{\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}_1\ \hat{\boldsymbol{q}}_1}{\boldsymbol{n}_1} + \frac{\hat{\boldsymbol{p}}_2\ \hat{\boldsymbol{q}}_2}{\boldsymbol{n}_2}} < \boldsymbol{p}_1 - \boldsymbol{p}_2 < (\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2) + \boldsymbol{z}_{\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}_1\ \hat{\boldsymbol{q}}_1}{\boldsymbol{n}_1} + \frac{\hat{\boldsymbol{p}}_2\ \hat{\boldsymbol{q}}_2}{\boldsymbol{n}_2}}} \tag{3.38}$$

Exemplo 3.11

Pretende-se saber se a proporção de ulmeiros afectados pela grafiose é idêntica em duas zonas A e B. Na zona A foi recolhida uma amostra de 150 ulmeiros e verificou-se que 107 estavam afectados e na zona B recolheu-se uma amostra de 100 havendo 63 afectados. Que conclusões se pode tirar ao nível de significância de 0.05?

Resolução:

Sejam p_1 e p_2 as proporções de ulmeiros afectados pela grafiose nas zonas A e B respectivamente. Em face dos resultados obtidos temos como estimativas

$$\hat{p_1} = \frac{107}{150} = 0.71 \qquad \hat{p_2} = \frac{63}{100} = 0.63.$$

Um intervalo de confiança para $p_1 - p_2$, aplicando directamente (3.38) é

$$(0.71 - 0.63) - 1.96 \times 0.06 < p_1 - p_2 < (0.71 - 0.63) + 1.96 \times 0.06 \Leftrightarrow$$

$$\Leftrightarrow -0.04 < p_1 - p_2 < 0.20$$

Atendendo a que o intervalo de confiança contém valores positivos e negativos, podemos dizer com 95% de confiança que não há diferença significativa entre as proporções de ulmeiros afectados em cada uma das zonas.

Testes de hipóteses

Vamos aqui introduzir uma outra metodologia usada em inferência estatística que já referimos atrás, chamada **Testes de Hipóteses Estatísticas**. Grosseiramente falando o objectivo dos testes de hipóteses estatísticas é determinar se certas afirmações sobre uma população são suportadas pelos dados da amostra. Portanto apresenta procedimentos adequados para pôr à prova ideias que formulamos sobre factos desconhecidos.

"Uma ideia, mesmo que pareça muito boa, deve ser pelo menos transitoriamente negada, e só se a evidência factual nos levar a rejeitar essa negação é que deve ser acolhida como promissora" ... Pestana e Velosa (2008), *Introdução à Probabilidade e à Estatística*

Vejamos em primeiro lugar algumas

Noções básicas

Definição 3.11

Uma **hipótese estatística** é uma conjectura sobre:
- um parâmetro desconhecido da população ou
- a forma da distribuição de uma característica em estudo na populaço

Pode ser então sobre a forma da distribuição ou sobre o valor de parâmetros desconhecidos da população, conhecida ou não a forma da distribuição.

Tal conjectura pode ou não ser verdadeira. A verdade ou falsidade nunca pode ser confirmada, a menos que observássemos toda a população, o que nalguns casos é impraticável (quando a população é muito grande) ou noutros mesmo impossível (caso de populações infinitas ou mesmo de a característica em estudo levar à destruição da população – duração de vida de um elemento, etc.)

É então através da informação fornecida por uma amostra que nós rejeitamos ou não a hipótese formulada.

Iremos tratar aqui hipóteses sobre o(s) valor(es) do(s) parâmetro(s) desconhecidos da população (**testes paramétricos**ou **testes de hipóteses paramétricas**).

Então, por exemplo, admitamos que temos uma população cuja lei é conhecida com excepção de um parâmetro (escalar ou vector) $\theta \in \Theta$ desconhecido.

A teoria de Neyman-Pearson sobre testes de hipóteses estabelece uma dicotomia no espaço, Θ, do parâmetro (conjunto de valores possíveis para o parâmetro desconhecido),i.e., $\Theta = \Theta_0 \cup \Theta_1$ e $\Theta_0 \cap \Theta_1 = \emptyset$.

Tal dicotomia consiste afinal na formulação de duas hipóteses alternativas, que é costume serem designadas por

$$\boldsymbol{H_0}\textbf{ – hipótese nula e }\boldsymbol{H_1}\textbf{ – hipótese alternativa}.$$

A formulação matemática das hipóteses é então a seguinte:

$$\boldsymbol{H_0 : \theta \in \Theta_0}$$

$$\boldsymbol{H_1 : \theta \in \Theta_1},$$

e dizemos que queremos testar H_0 contra H_1, escrevendo muitas vezes H_0 vs H_1.

Se Θ_0 (Θ_1) contém um só elemento, H_0 (H_1) diz-se hipótese simples, se contém mais de um elemento diz-se hipótese composta.

Escolha da hipótese nula e da hipótese alternativa

Ao testar uma hipótese nula contra uma hipótese alternativa a nossa atitude deverá ser admitir H_0 como verdadeira até que os dados testemunhem fortemente contra ela; nesse caso deverá ser rejeitada a favor de H_1.

Uma analogia à formulação das hipóteses nula e alternativa, pode ser a que se estabelece num julgamento, onde os jurados terão de decidir por considerar culpado ou inocente um réu. Deve considerar-se culpado apenas se houver provas verdadeiramente convincentes da culpa, portanto a hipótese nula deve ser aquela que deve ser olhada como verdadeira e só rejeitada quando os dados apresentarem forte evidência disso.

Portanto devem ser assim formuladas as hipóteses:

H_0 : não culpado
H_1 : culpado.

Regra geral toma-se como hipótese H_0 a hipótese a ensaiar, por exemplo, H_0 é o valor de um parâmetro cuja plausibilidade se pretende verificar, é a situação actual, etc.

Um teste de hipóteses é afinal uma regra que permite especificar um subconjunto do espaço-amostra (Chama-se espaço amostra $\mathcal{X}$ o conjunto de todas as amostras possíveis), $\boldsymbol{A \subset \mathcal{X}}$, que permita decidir se a amostra observada conduz à **rejeição ou não de $\boldsymbol{H_0}$**

Designe-se por A - **região de aceitação de** H_0 e R - **região de rejeição de** H_0, com $A \cap R = \emptyset$ e $A \cup R = \mathcal{X}$.

Para tomarmos uma decisão é necessário então estabelecer um conjunto de **regras**:

– Definir uma variável aleatória, função da amostra aleatória, cujo valor será calculado a partir de uma amostra observada:

– se o valor calculado $\in R$, rejeita-se H_0

– se o valor calculado $\in A$, aceita-se H_0

Podemos dizer que um teste de uma hipótese nula é o decurso de uma acção que especifica o conjunto de valores de uma v.a. X, para os quais H_0 é

rejeitada.

A v.a. cujos valores permitem decidir qual a atitude a tomar diz-se **Estatística do teste** e o conjunto de valores dessa variável que conduzem à rejeição de H_0 chama-se **Região de Rejeição** ou **Região Crítica**.

Os dois tipos de erros e a função potência de um teste

Sendo um teste de hipóteses um tipo particular de inferência estatística, é portanto um procedimento que leva do particular ao geral, podendo assim conduzir a erros.

Ao proceder-se a um teste de hipóteses podem cometer-se dois tipos de erros:

– rejeitar a hipótese nula, sendo ela verdadeira – chama-se a este **erro do tipo I** ou **erro de primeira espécie** e

– não rejeitar H_0 e ela ser falsa – chama-se a este **erro do tipo II** ou **erro de segunda espécie**.

Vejamos esquematizadas no seguinte quadro as situações que podem ocorrer

Conclusões do teste	Situações correctas H_0 verdadeira	desconhecidas H_0 falsa
Não rejeitar H_0	Correcto	Errado **Erro tipo II**
Rejeitar H_0	Errado **Erro tipo I**	Correcto

A decisão do teste é portanto errada nas duas situações seguintes:

– Rejeitar H_0 e H_0 ser verdadeira (erro de $1^{\underline{a}}$ espécie);

– Não rejeitar H_0 e H_0 ser falsa (erro de $2^{\underline{a}}$ espécie).

Às probabilidades de cometer cada um dos erros anteriores é costume designar por α e β:

$$\alpha = P(\text{erro tipo I}) = P(\text{Rejeitar } H_0 / \ H_0 \text{ verdadeiro})$$
é chamado **o nível de significância do teste**;
$$\beta = P(\text{erro tipo II}) = P(\text{não rejeitar } H_0 / \ H_0 \text{ falso}) \text{ e a}$$
$$1 - \beta = P(\text{rejeitar } H_0 / \ H_0 \text{ falso}) \text{ chama-se } \textbf{a potência do teste}.$$

O aspecto fundamental da teoria dos testes de hipóteses consiste na escolha da região crítica de modo a controlar cada um dos dois tipos de erros. A situação ideal seria aquela que minimizasse simultaneamente α e β. Mas, na realidade, mantendo fixa a dimensão da amostra, α e β variam em sentidos contrários. O que se faz é fixar o nível de significância α (o erro tipo I é considerado o mais gravoso e portanto o que deve ser controlado) e tomar o teste que minimize β, ou seja, que maximize a potência do

teste (na verdade a procura de um teste que minimize β só tem sentido quando H_0 e H_1 são hipóteses compostas, porque neste caso as probabilidades de cometer erros de 1ª e 2ª espécie dependem do valor particular do parâmetro em cada uma das regiões).

Vejamos esquematicamente os principais passos na formulação de um teste de hipóteses, respeitante a um parâmetro θ de uma população:

1. $H_0 : \theta \geq \theta_0$ ou $H_0 : \theta \leq \theta_0$ ou $H_0 : \theta = \theta_0$; e respectivamente
2. $H_1 : \theta < \theta_0$ ou $\theta > \theta_0$ ou $\theta \neq \theta_0$;
3. Escolher o nível de significância α;
4. Seleccionar a estatística do teste e definir a região de rejeição;
5. Calcular o valor da estatística para uma amostra observada, de dimensão n;
6. Decisão:
   – Rejeitar H_0, se o valor da estatística pertence à região critica
   – Não rejeitar H_0, caso contrário.

Podem se consultados os quadros com os diferentes testes que irão ser considerados neste curso, e que são testes ao **valor médio**, **variância**, **diferença de valores médios** no caso de amostras independentes e emparelhadas, **quociente de variâncias** e **proporções**.

Notas sobre Regras de decisão em Testes de Hipóteses

A indicação do valor observado da estatística do teste, seguido da consulta de uma tabela para a procura de um valor crítico de modo a tirar conclusões, tem sido recentemente "substituído" pelo cálculo

da probabilidade de se observar um valor igual ou mais extremo do que o observado, se a hipótese nula é verdadeira – chama-se a isto valor de prova; **valor-*p* (*p-value*)**

Nota: é esta quantidade que hoje em dia qualquer *software* está preparado para calcular e indicar quando se manda realizar um teste.

Podemos interpretar o **valor de prova**, **valor-*p*** ou ***p-value*** como a

medida do grau de concordância entre os dados e H_0. Assim

Quanto menor for o *p-value,* menor é a consistência entre os dados e a hipótese nula

Habitualmente adopta-se como regra de decisão:

$$\text{rejeitar } H_0 \text{ se } \boldsymbol{p\text{-}value} \leq \boldsymbol{\alpha} .$$

Testes de Normalidade

Vamos aqui considerar apenas **o caso** de se pretender testar a forma da distribuição. Estamos então no domínio dos testes não paramétricos que neste caso se designam por **testes de ajustamento**.

Comecemos com um teste muito importante nas nossas aplicações - **um teste de ajustamento** para averiguar se um dado conjunto de observações se pode considerar proveniente de uma população lue com distribuição normal – é um **teste de normalidade**, o **Teste de Shapiro Wilk**, que se tem revelado ser um dos mais potentes. Vejamos em síntese como se processa.

O teste de Shapiro-Wilk
Seja **X** a característica em estudo na população.

Formulam-se as hipóteses:
H_0 : X tem distribuição normal
$H_{1:}$ X não tem distribuição Normal

Calcula-se o valor da estatística do teste $W_{cal} = \dfrac{b^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ com b constante a determinar a partir dos dados e com recurso a uma tabela (que neste ano lectivo não será dada).

Valores pequenos de W_{cal} indicam não normalidade, i.e.

$$\textbf{RC:} \quad \boldsymbol{W_{cal} < W_{\alpha}}$$

Com W_α – **valor crítico** a consultar na tabela. Aqui iremos utilizar o teste de Shapiro-Wilk, apenas com recurso ao valor-$\boldsymbol{p}$, para a tomada de decisão.

- **Rejeita-se H_0** se **valor$-\boldsymbol{p} \leq \boldsymbol{\alpha}$**, significando que não se pode admitir que X tem distribuição normal;
- Se **valor$-\boldsymbol{p} > \boldsymbol{\alpha}$** não se rejeita $\mathbf{H}_0$, o que significa que a distribuição normal é uma distribuição possível para X.

Exercícios propostos

1. Suponhamos que se pretende comparar, numa mesma região, as produções leiteiras de duas raças bovinas e que se escolheu com esse objectivo, aleatória e independentemente, uma amostra de 50 animais de cada raça. Se obtivermos para cada uma das amostras rendimentos médios de $\overline{x}_1 = 4360$ kg e $\overline{x}_2 = 4190$ kg, deve-se admitir que existe realmente uma diferença entre as duas raças? Justifique a resposta.

2. Um fabricante de sacos de plástico vende os sacos, a peso, em lotes de 100. Cada lote deveria ter um peso de 500 gr. Recolhida uma amostra de 20 lotes de sacos, observaram-se os seguintes valores:

$$\overline{x} = 502.5 \qquad\qquad s = 4.31.$$

Diga se é de admitir que o peso médio dos sacos é efectivamente 500 gr.

3. Pretende-se comparar dois métodos para determinar a percentagem de magnésio existente num composto químico. Para isso foram submetidos àqueles métodos dez compostos diferentes, tendo-se obtido os seguintes resultados:

nºde composto	1	2	3	4	5	6	7	8	9	10
Método A	13.3	17.6	4.1	17.2	10.1	3.7	5.1	7.9	8.7	11.6
Método B	13.4	17.9	4.1	17.0	10.3	4.0	5.1	8.0	8.8	12.0

Haverá diferença significativa entre os dois métodos, a um nível de significância de 5%?

4. Para controlar a qualidade dos lotes de um dado produto que vai ser vendido a peso, produzidos numa linha de fabrico, decidiu-se usar o seguinte esquema:

– recolher uma amostra e dimensão n, de cada lote e calcular a média $\overline{x}$, dos pesos das embalagens;

– se $\overline{x} \leq c$ rejeitar o lote;

– se $\overline{x} > c$ aceitar o lote.

Decidiu-se ainda que, se a verdadeira média do peso das embalagens no lote (μ) for inferior ou igual a 5.3 a probabilidade de rejeitar o lote deve ser pelo menos 99% e, se μ for superior ou igual a 5.5 a probabilidade de aceitar o lote deve ser pelo menos 90%. Admita que o desvio padrão do peso, em cada lote, é de 0.2.

Calcule o valor de c e o menor valor de n requerido por este esquema de amostragem. Justifique.

Referências bibliográficas

Bhattacharyya, G.K. and Johnson R.A.(1977), *Statistical Concepts and Methods*, John Wiley & Sons Inc.

Dagnelie, P.(1973), *Estatística, Teoria e Métodos*, trad. do Prof. Doutor A. St.Aubyn, Europa América, vol I e II.

Walpole, R.E (1993), *Mathematical Statistics*, 3th edition, Englewood Cliffs, N.J., Prentice-Hall.

Milton, J.S. and Arnold, J.C. (1987), *Probability and Statistics in the Engineering and Computing Sciences*, Mc Graw Hill.

Mood, A. Graybill, F. and Boes D. (1985). *Introduction to the Theory of Statistics* , Mc Graw Hill.

Murteira, B. (1990), *Probabilidades e Estatística* (vol I), Mc Graw Hill.

Murteira, B., Ribeiro, C.S., Silva, J.A. e Pimenta C.(2002), *Introdução à Estatística*, Mc Graw Hill.

Pestana, D.D. e Velosa, S.F. (2008), *Introdução à Probabilidade e à Estatística* 3^a edição. Fundação Calouste Gulbenkian.