Capítulo III – Introdução à Inferência Estatística

Capítulo III-Introdução à Inferência Estatística

- A Probabilidade e a Inferência Estatística.
- A amostragem. Principais Distribuições por Amostragem
- Tópicos de Estimação e de Inferência
  - estimação pontual
  - estimação por intervalos
  - testes de hipóteses

A Probabilidade e a Inferência Estatística

Começámos esta unidade curricular pelo estudo descritivo de uma **amostra**. Passámos depois pela Introdução à Teoria da Probabilidade com o estudo de alguns dos modelos de probabilidade (discretos e contínuos).

Veremos agora como utilizar todo esse conhecimento na **Inferência Estatística**. A Inferência Estatística é *... aquilo que permite o salto que vai da amostra para a população* (Pestana e Velosa, 2008).

A Probabilidade e a Inferência Estatística

Mas ... sem conhecimento da probabilidade não é possível avançar na inferência estatística. Apenas para breve "orientação" podemos dizer que:

- na probabilidade temos o modelo da população que nos permite calcular a probabilidade de certos acontecimentos;
- na inferência estatística parte-se da amostra e pretende-se obter informação sobre o modelo ou sobre características da população.

Conceitos em Inferência Estatística

A inferência estatística tem como objectivos tirar conclusões sobre os parâmetros da população a partir da recolha, tratamento e análise dos dados de uma amostra, obtida nessa população.

Conceitos básicos:

População $\rightarrow$ conjunto completo de todas os objectos (elementos) com uma (ou mais) característica(s) comuns;

Unidade Estatística $\rightarrow$ cada elemento da população;

Amostra $\rightarrow$ conjunto dos valores efectivamente observados;

Parâmetro de uma população $\rightarrow$ **constante** desconhecida, cujo verdadeiro valor se pretende "estimar" ou "validar".

Introdução à Teoria da Amostragem

Amostragem $\rightarrow$ procedimento de recolha de elementos da população para obter uma amostra.

De uma dada **população** podemos retirar muitas amostras:

- Amostra 1
- Amostra 2
- ...
- Amostra k
- ...

Mas $\cdots$ quase sempre recolhemos só uma amostra para estudarmos uma característica $\boldsymbol{X}$ da população.

Seja $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}$ uma amostra de n observações da característica, obtidas após um processo de amostragem.

Introdução à Teoria da Amostragem

Vamos considerar a amostra $x_1, x_2, \cdots x_n$ como uma realização de n variáveis $X_1, X_2, \cdots X_n$ que são "cópias" da variável X

Recapitulando:

- Antes da amostragem ser realizada temos n variáveis aleatórias $X_1, X_2, ..., X_n$
- Depois de efectuada a amostragem temos um conjunto de dados que constituem a amostra observada $x_1, ..., x_n$.

Definição de amostra aleatória

Chama-se **amostra aleatória** de dimensão n a um conjunto de n variáveis aleatórias $X_1, X_2, ..., X_n$ independentes e semelhantes, i.e., tendo todas a mesma distribuição que é a da característica X em estudo na população.

Tópicos de Estimação

Quais os parâmetros desconhecidos da população que nos vão interessar e que procedimentos usar para **os estimar**?

Os procedimentos são métodos adequados (queremos "os melhores") para estimar os parâmetros desconhecidos.
Em Estatística, esses métodos consistem na obtenção de variáveis aleatórias "especiais", chamadas **estimadores**.

Definição de estimador

Chama-se **estimador de um parâmetro** a uma função da amostra aleatória $X_1, X_2, ..., X_n$, que "permite" obter um valor para o parâmetro desconhecido (esta função não possui quantidades desconhecidas).

NOTA: Não iremos referir procedimentos para obter estimadores

Tópicos de Estimação

Um estimador é então uma variável aleatória que terá uma dada distribuição (pelo menos aproximadamente).

O valor que o estimador (variável aleatória) toma quando se observa a amostra chama-se **estimativa**.

Definição de estimativa

Estimativa de um parâmetro é um valor concreto assumido pelo estimador.

Dada a amostra aleatória, $X_1, X_2, ..., X_n$, i.e., v.a. **i.i.d. a** $\boldsymbol{X}$, vamos ver os **parâmetros** que nos vão interessar, os seus estimadores e estimativas

Tópicos de Estimação

Parâmetro(s) a estimar	Estimador(es)	Estimativa(s)
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	$\widehat{P} = \frac{X}{n}$ [(a)]	$\hat{p} = \frac{x}{n}$ [(b)]
μ_1 ?? μ_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	$\widehat{P_1}$?? $\widehat{P_2}$	$\hat{p}_1$?? $\hat{p}_2$

[(a)] X - v.a. que conta o número de sucessos na amostra de dimensão n

[(b)] x - número observado de sucessos na amostra de dimensão n.

?? significa que teremos que escolher a melhor operação para fazer a comparação

Tópicos de Estimação

Temos uma amostra — como inferir para a população?

- Estimação dos parâmetros
  - Estimação pontual
  - Estimação intervalar $\longrightarrow$ intervalos de confiança
- Testes de hipóteses estatísticas.

Estimação pontual $\rightarrow$ indicar um único valor como estimativa.
Exemplos de estimativas: $\overline{x}$, s^2, $\hat{p}$, $\cdots$.

Estimação por intervalos $\rightarrow$ neste caso calcula-se um intervalo (aleatório) que com uma probabilidade elevada contenha o verdadeiro valor do parâmetro.

Estimação por intervalos

Para a construção do intervalo é necessário conhecer a distribuição - exacta ou aproximada - do **estimador** (ou qualquer expressão dele)

Definição de intervalo de confiança

Chama-se **intervalo de confiança** ao intervalo que resulta da concretização do intervalo (aleatório) e é portanto um intervalo (a, b), onde a e b são números reais e $a < b$.

Relembre-se que para obter o intervalo de confiança é então necessário conhecer **a lei do estimador** do parâmetro desconhecido.

Intervalo de confiança

Comecemos então por ver como se construiria **um intervalo de confiança para μ,** (valor médio de uma característica X.)

Precisamos de procurar a distribuição do estimador de $\mu \longrightarrow \overline{X}$

- Seja X v.a. c/ dist. Normal, i.e., $X \frown \mathcal{N}(\mu, \sigma)$
  Para uma amostra de dimensão n tem-se $\overline{X}_n \frown \mathcal{N}(\mu, \sigma/\sqrt{n})$

Então . . .

Intervalo de confiança

Intervalo de confiança a $(1-\alpha)\times 100\%$ para μ

- Seja X v.a. c/ dist. Normal , i.e., $X \frown \mathcal{N}(\mu, \sigma)$

  Se σ conhecido $\rightarrow$ considera-se $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \frown \mathcal{N}(0,1)$

  o intervalo a $(1-\alpha)\times 100\%$ de confiança para μ é

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

($z_{\alpha/2} \rightarrow$ valor da v.a. Z tal que $P(Z > z_{\alpha/2}) = \alpha/2$)

Observações: Chama-se **precisão da estimativa** à semi-amplitude do intervalo de confiança e **confiança** ou **grau de confiança** a $(1-\alpha)\times 100\%$

Quanto maior for o intervalo, maior é o grau de confiança, mas menor a precisão da estimativa.

Exercício 1

Um método utilizado na determinação do pH de uma dada solução fornece medições que se admite terem distribuição normal de valor médio igual ao verdadeiro valor do pH da solução e desvio padrão de 0.02.
Para avaliar o pH de uma solução, efectuaram-se 10 medições independentes tendo-se obtido os seguintes valores:

8.18 8.16 8.17 8.22 8.19 8.17 8.15 8.21 8.16 8.18

a) Indique uma estimativa do valor médio do pH da solução.
b) Com base nestas 10 medições, determine um intervalo a 95% de confiança para o valor médio do pH da solução.
c) Para um certo processo quimico é muito importante que uma dada solução tenha um pH de exactamente 8.20. Com base no resultado da alínea anterior, o que pode concluir relativamente à utilização desta solução no referido processo químico?

Distribuições por amostragem

A determinação de intervalos de confiança para os parâmetros μ, σ^2, p, $\mu_1 - \mu_2$, σ_1^2/σ_2^2 e $p_1 - p_2$, necessita do conhecimento da distribuição dos estimadores envolvidos $\longleftrightarrow$ **distribuições por amostragem**

$\Downarrow$

São distribuições de funções da amostra aleatória $(X_1, X_2, \cdots, X_n)$, que vamos usar para obter **Intervalos de Confiança**

Já vimos

Se $X \frown \mathcal{N}(\mu, \sigma)$ e σ conhecido usamos $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \frown \mathcal{N}(0,1)$ para construir um I.C. para o valor médio, μ.

Distribuições por amostragem

Mas na maior parte das situações não conhecemos a variância.
Então neste caso já não é possível obter um I.C. para μ, pois os limites do intervalo dependem de σ, desconhecido.

Vamos então referir brevemente alguns resultados importantes, que nos permitem definir distribuições que precisamos de utiizar agora

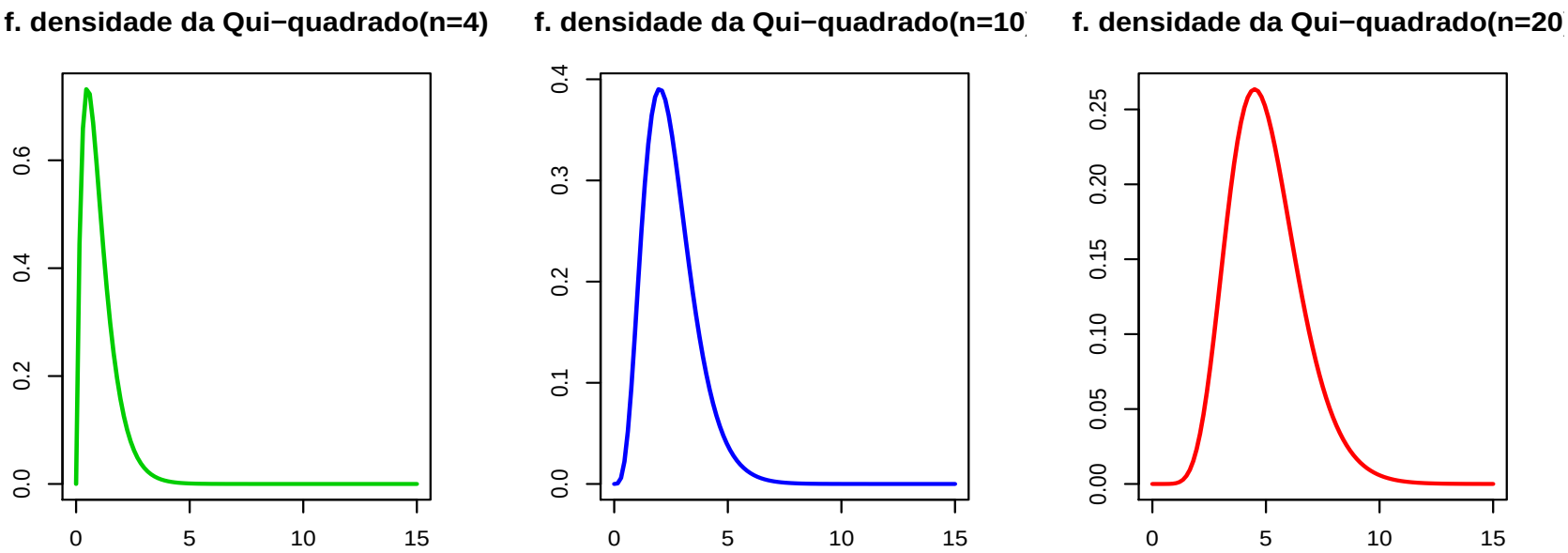
Resultados (distribuições) importantes

Teorema

Sejam X_i $(i = 1, \ldots, n)$ v.a. independentes com distribuição $\mathcal{N}(\mu, \sigma)$. Então:

- $\left(\frac{X_i - \mu}{\sigma}\right)^2 \frown \chi^2_{(1)}$, que se designa distribuição qui-quadrado com 1 grau de liberdade.
- $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \frown \chi^2_{(n)}$.

Gráficos da função densidade de uma v.a. com distribuição χ^2

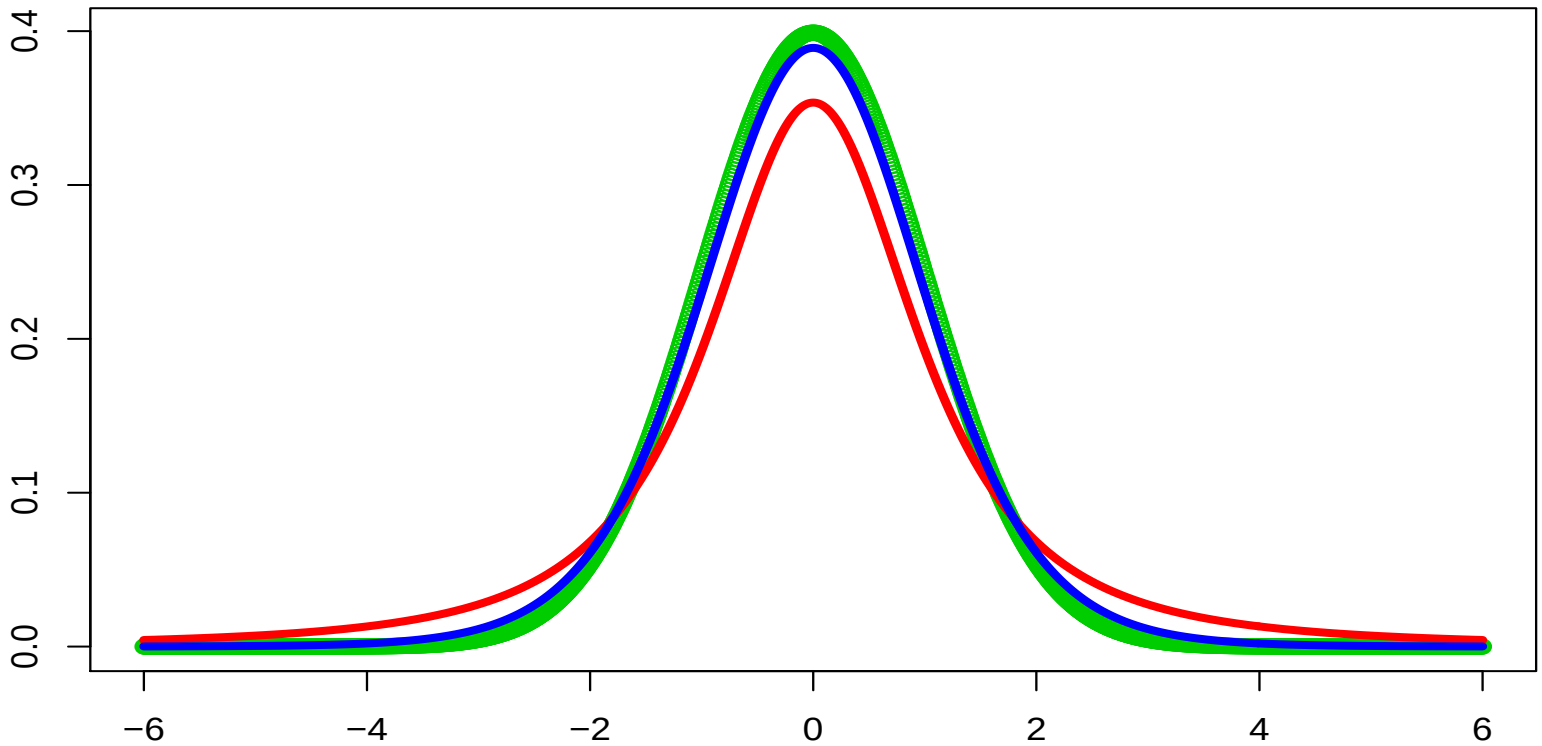


Gráficos da função densidade de uma v.a. com distribuição $\chi^2_{(4)}$, $\chi^2_{(10)}$ e $\chi^2_{(20)}$, da esquerda para a direita.

Resultados importantes (cont.)

- $\frac{(n-1)S^2}{\sigma^2} \frown \chi^2_{(n-1)}$, com $S^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$
- Sejam $Z \frown \mathcal{N}(0,1)$ e $Y \frown \chi^2_{(n)}$ v.a. independentes, então $\frac{Z}{\sqrt{Y/n}} \frown t_{(n)}$, $t_{(n)}$ diz-se distribuição $t-Student$ com n g.l.
  Para mais informações ver quadro das distribuições contínuas.
- $\frac{\overline{X}-\mu}{S/\sqrt{n}} \frown t_{(n-1)}$, com $S^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$

Gráficos



Gráficos da função densidade de uma v.a. com distribuição $\mathcal{N}(0,1)$, $t_{(2)}$ e $t_{(10)}$.

Intervalos de confiança (cont.)

O Intervalo a $(1-\alpha)\times 100\%$ de confiança para σ^2 numa população normal, constrói-se usando a v.a. $\frac{(n-1)S^2}{\sigma^2} \frown \chi^2_{(n-1)}$

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

O Intervalo a $(1-\alpha)\times 100\%$ de confiança para μ com σ desconhecido, numa população normal, constrói-se usando a v.a. $\frac{\overline{X}-\mu}{S/\sqrt{n}} \frown t_{(n-1)}$

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

Dada X com distribuição binomial de parâmetros (n, p)
(p parâmetro desconhecido).

Um **estimador** de p é $\hat{P} = \frac{X}{n}$ (X é o número de sucessos em n provas)
Sendo a **estimativa** $\hat{p} = \frac{x}{n}$ (x é o número observado de sucessos em n provas)

Se $\boldsymbol{X \frown B(n,p)}$ e n grande
$\boldsymbol{X \sim \mathcal{N}(np, \sqrt{npq})}$

Intervalo de confiança $(1-\alpha) \times 100\%$ para $\boldsymbol{p}$

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