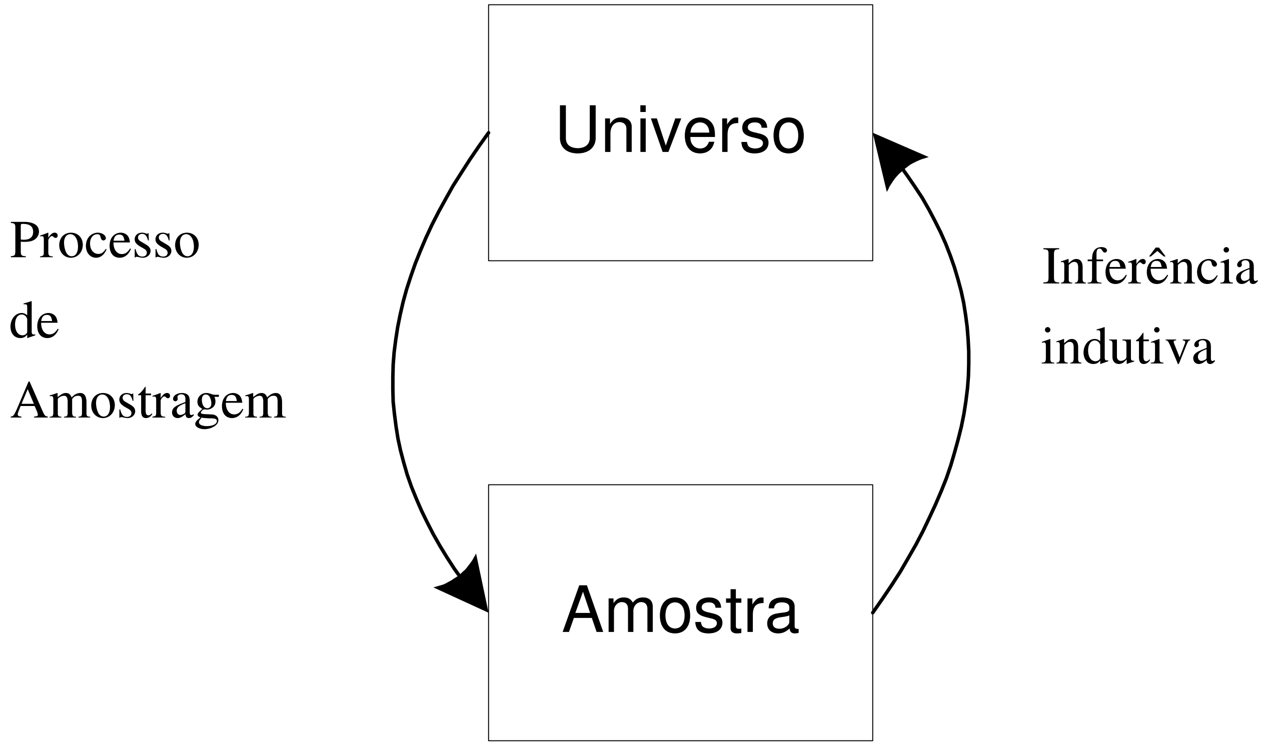
1- Probabilidades e inferência estatística

Procedimentos "complementares"

- **Teoria da probabilidade**: parte-se de determinado modelo e calcula-se a probabilidade de certos resultados ou acontecimentos;
- **Inferência estatística**: parte-se de observações e procura inferir-se alguma coisa sobre o modelo;
- **Estatística descritiva**: "Organização/arrumação" das observações/informação.


Universo
Amostra
Processo de Amostragem
Inferência indutiva

Exemplo 6.1 – Considere-se um grupo numeroso de pessoas (por exemplo, os estudantes matriculados no ISEG no ano lectivo de 2001-2002) entre os quais há uma proporção θ que pratica desporto. Escolhem-se ao acaso, com reposição, n pessoas, seja $n = 10$; se θ fosse conhecido, seja, $\theta = 0.3$, podia haver interesse em calcular a probabilidade de encontrar x praticantes, $0 \le x \le 10$, nesse grupo de 10 pessoas, probabilidade que se sabe ser determinada pela expressão,

$$\binom{10}{x}(0.3)^x(0.7)^{10-x}.$$

Trata-se de um problema de probabilidades.

Pode no entanto suceder – e na prática sucede quase sempre – que θ seja desconhecido; nesse caso interessa provavelmente ao observador utilizar o resultado da amostra, nomeadamente a proporção de praticantes de desporto na amostra, seja $x/10$ (ou, no caso geral, x/n), para tirar conclusões sobre a proporção de praticantes na população donde foi retirada a amostra.

Trata-se de um problema de inferência estatística.

2 – Especificação. Amostragem casual

- Especificação de um modelo (universo)

  Escolha de uma família de modelos probabilísticos que se supõe vigorar no universo. Esta escolha estará naturalmente sujeita a avaliação;

- Processo de amostragem / Amostragem casual

  O processo de recolha da amostra (**processo de amostragem**) deve depender do acaso. Apenas se vai ver um processo particular de amostragem aplicado a populações supostas infinitas.

- **Definição 6.1 – Amostragem casual -** Quando as n variáveis aleatórias observadas, componentes do vector $(X_1, X_2, \ldots, X_n)$, são **independentes e identicamente distribuídas** – simbolicamente **iid** – diz-se que se trata de amostragem casual.
  - Cada X_i, $i = 1,2,\ldots,n$, é uma **"cópia"** da variável aleatória X
  - Independência entre os X_i, $i = 1,2,\ldots,n$.

- Processo de amostragem aleatório $\rightarrow$ os dados observados formam apenas um dos muitos conjuntos de dados que poderiam ter sido obtidos operando nas mesmas circunstâncias;

  A amostra de n observações que se observou, $(x_1, x_2, \ldots, x_n)$, é uma realização da variável aleatória n-dimensional $(X_1, X_2, \ldots, X_n)$.

  - $(X_1, X_2, \ldots, X_n)$ Amostra aleatória
  - $(x_1, x_2, \ldots, x_n)$ Amostra observada

- O espaço-amostra, X, é o conjunto de todas as amostras passíveis de serem seleccionadas (subconjunto de $\Re^n$)

$$\text{População} \Rightarrow \begin{cases} \text{Amostra } 1 \\ \text{Amostra } 2 \\ \cdots \\ \text{Amostra } m \\ \cdots \end{cases}$$

Exemplo – Assuma-se que X (pratica ou não pratica desporto) é uma variável de Bernoulli de parâmetro θ, isto é

$$F_\theta = \left\{ f(x \mid \theta) = \theta^x (1-\theta)^{1-x} : x \in \{0,1\} \wedge \theta \in \Theta = (0,1) \right\} \rightarrow \text{Modelo}$$

Amostra casual $(X_1, X_2, ..., X_n)$, sendo $X_i = 1$ (i-ésimo indivíduo da amostra é praticante de desporto) ou $X_i = 0$ (caso contrário).

Os X_i, $i = 1,2,...,n$, são iid com distribuição de Bernoulli de parâmetro θ

Suponha-se que $n = 3$. O espaço amostra vem (8 elementos):

(0 ; 0; 0)	com probabilidade	$(1-\theta)^3$
(1 ; 0; 0) (0 ; 1; 0) (0 ; 0; 1)		$\theta \times (1-\theta)^2$
(1 ; 1; 0) (1 ; 0; 1) (0 ; 1; 1)		$\theta^2 \times (1-\theta)$
(1 ; 1; 1)		θ^3

Como é óbvio só se observa, habitualmente, uma das amostras.

3 – Estatísticas

- **Definição 6.2 – Estatística**

  Uma estatística é uma variável ou vector aleatório $T(X_1, X_2, ..., X_n)$, função da amostra aleatória $(X_1, X_2, ..., X_n)$, que não envolve qualquer parâmetro desconhecido.

- Comentários
  - A ideia é, sempre que possível, condensar a informação.
  - Depois de observar a amostra temos de estar em condições de atribuir um valor à estatística.

- **Exemplo 6.7 –** Se $(X_1, X_2, ..., X_n)$ é amostra casual de uma população de Bernoulli, a estatística $T_1(X_1, ..., X_n) = \sum_i X_i$, ou simplesmente $T_1 = \sum_i X_i$, representa o número de "sucessos" na amostra e a estatística $T_2 = \sum_i X_i / n$ indica a proporção de "sucessos" na amostra.

- **Exemplo 6.9 –** Se $(X_1, X_2, ..., X_n)$ é amostra casual de população normal $N(\mu, \sigma^2)$ com parâmetros μ e σ^2 desconhecidos, são exemplos de estatísticas unidimensionais,

$$\sum_i X_i,\ \bar{X} = \frac{1}{n}\sum_i X_i,\ \sum_i X_i^2,\ \frac{1}{n}\sum_i X_i^2,$$

e de estatísticas bidimensionais,

$$\left(\sum_i X_i, \sum_i X_i^2\right),\ \left(\bar{X},\ \sum_i (X_i - \bar{X})^2\right).$$

Não são estatísticas as funções,

$$\frac{1}{\sigma}\sum_i (X_i - \mu),\ \frac{1}{\sigma}\sum_i X_i,\ \frac{1}{\sigma^2}\sum_i X_i^2,$$

pois dependem de parâmetros desconhecidos.

4 - Distribuição da amostra e distribuição por amostragem da estatística

$$\text{População} \Rightarrow \begin{cases} \text{Amostra } 1 \rightarrow \text{valor } t_1 \text{ para a estatística } T(x_1, x_2, ..., x_n) \\ \text{Amostra } 2 \rightarrow \text{valor } t_2 \text{ para a estatística } T(x_1, x_2, ..., x_n) \\ \cdots \\ \text{Amostra } m \rightarrow \text{valor } t_m \text{ para a estatística } T(x_1, x_2, ..., x_n) \\ \cdots \end{cases}$$

- O comportamento probabilístico da estatística $T(X_1, X_2, ..., X_n)$ é dado pela respectiva função de distribuição (função densidade ou função probabilidade).
- Fala-se na **distribuição por amostragem** da estatística $T(X_1, X_2, ..., X_n)$.

- Distribuição da amostra **(função densidade ou função probabilidade conjunta):**

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i \mid \theta) \qquad \text{tira-se partido da amostra ser iid}$$

- **Distribuição por amostragem** da estatística $T(X_1, X_2, ..., X_n)$:

$$G(t \mid \theta) = P(T \le t) = \int \cdots \int_{A(t)} \left[\prod_{i=1}^{n} f(x_i \mid \theta) \right] dx_1 dx_2 ... dx_n,$$

no caso de T ser variável aleatória contínua, ou,

$$G(t \mid \theta) = P(T \le t) = \sum_{A(t)} \left[\prod_{i=1}^{n} f(x_i \mid \theta) \right],$$

no caso de T ser variável aleatória discreta. Em qualquer das hipóteses,

$$A(t) = \left\{ (x_1, x_2, ..., x_n) \in \Re^n : T(x_1, x_2, ..., x_n) \le t \right\},$$

- Para algumas situações existem formas mais simples de obter a distribuição por amostragem da estatística

Como obter a **distribuição por amostragem** de determinada estatística?

Pode-se utilizar:

a) A função geradora dos momentos relevante no estudo de T.

b) Distribuições aproximadas (Teorema do Limite Central)

c) O método de **Monte Carlo** (**simulação)** quando não se consegue chegar a uma solução analítica.

Exemplo 6.10 – Se $(X_1, X_2, ..., X_n)$ é uma amostra casual de uma população de Poisson, $X_i \sim \mathrm{Po}(\theta)$, então, pelo teorema 5.3, tem-se $T = \sum_i X_i \sim \mathrm{Po}(n\theta)$. Assim, a estatística T tem função probabilidade,

$$g(t \mid \theta) = \frac{e^{-n\theta}(n\theta)^t}{t!}, \quad t = 0,1,2,..., \quad \theta > 0.$$

Exemplo 6.11 – Se $(X_1, X_2, \ldots, X_n)$ é uma amostra casual de uma população exponencial, $X_i \sim \mathrm{Ex}(\theta)$, então, pelo teorema 5.8, $T = \sum_i X_i \sim \mathrm{G}(n, \theta)$. Assim, a estatística T tem função densidade,

$$g(t \mid \theta) = \frac{\theta^n \, e^{-\theta t} \, t^{n-1}}{\Gamma(n)}, \quad t > 0, \quad \theta > 0.$$

Distribuições por amostragem do **mínimo** e do **máximo** da amostra.

- Amostra $(X_1, X_2, ..., X_n)$ onde $X_i \sim F(x)$, f.d.p. ou f.p. $f(x)$.
- **Estatísticas de ordem**: obtêm-se ordenando a amostra: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.
- Estatísticas: $X_{(1)} = \min(X_i)$ e $X_{(n)} = \max(X_i)$.
- **Distribuição do mínimo**: Seja $G_1(x)$ a função de distribuição de $X_{(1)}$

$$G_1(x) = \Pr\left[X_{(1)} \le x\right] = 1 - [1 - F(x)]^n.$$

Se as v.a.'s são contínuas, $g_1(x) = n[1 - F(x)]^{n-1} f(x)$, $g_1(x)$ e $f(x)$ são as f.d.p.'s.

- **Distribuição do máximo**: Seja $G_n(x)$ a função de distribuição de $X_{(n)}$

$$G_n(x) = [F(x)]^n,$$

Caso as variáveis aleatórias sejam contínuas, $g_n(x) = n[F(x)]^{n-1} f(x)$, onde $g_n(x)$ e $f(x)$ são as respectivas funções densidade

Exemplo 6.15 – (adaptado) Seja X um universo com distribuição exponencial de parâmetro λ.

Distribuição do mínimo da amostra, $X_{(1)}$: Como se sabe, $X_{(1)} \sim \mathrm{Ex}(n\lambda)$.

Distribuição do máximo da amostra, $X_{(n)}$:

Amostra aleatória $(X_1, X_2, \ldots, X_n)$

$$G_n(x) = \Pr\left(X_{(n)} \le x\right) = \left(\Pr\left(X \le x\right)\right)^n = \left[1 - e^{-\lambda x}\right]^n,$$

que não é a função de distribuição de uma exponencial.

$$g_n(x) = n\lambda e^{-\lambda x}\left[1 - e^{-\lambda x}\right]^{n-1}.$$

5 - Primeiros resultados sobre a média e variância amostrais.

- **Média** e **variância** **amostrais**

$$\bar{X} = \frac{1}{n}\sum_i X_i \qquad\qquad S^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} - \bar{X}^2.$$

- **Teorema 6.1 –** Se $(X_1, X_2, ..., X_n)$ é uma amostra casual de população para a qual existem média $\mu = E(X_i)$ e variância $\sigma^2 = \mathrm{Var}(X_i)$ $(i = 1,2,...,n)$, tem-se,

$$E(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

(demonstrar)

- Comentários:
  - O teorema apenas exige a existência de μ e de σ^2 (no universo).
  - $E(\bar{X}) = \mu \rightarrow$ o VE da média da amostra é igual à média da população;
  - $\mathrm{Var}(\bar{X}) = \sigma^2/n \rightarrow$ quanto maior a dimensão da amostra, menor a variância de $\bar{X}$;

- **Teorema 6.2 –** Se $(X_1, X_2, ..., X_n)$ é amostra casual de população para a qual existem média $\mu = E(X_i)$ e variância $\sigma^2 = \mathrm{Var}(X_i)$ $(i = 1,2,...,n)$, tem-se,

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

- Os valores de S^2 são, em média, inferiores a σ^2. A variância amostral subavalia, em média, a variância da população.
- Correcção do problema → **variância corrigida** definida por,

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

Evidentemente que,

$$E(S'^2) = \sigma^2.$$

- Pode demonstrar-se que,

$$\mathrm{Var}(S^2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3},$$

$$\mathrm{Var}(S'^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right), \quad (n > 1).$$

Recorde-se que $\mu_r = E(X - \mu)^r$

6 – Distribuições por amostragem assintóticas

Em muitas situações não é possível obter **distribuições exactas** para as estatísticas $\sum_i X_i,\ \bar{X},\ S^2$ ou S'^2, mas podem obter-se **distribuições aproximadas.**

Distribuição assintótica da Média

Se $(X_1, X_2, \ldots, X_n)$ é uma amostra casual de população para a qual existem média $\mu = E(X_i)$ e variância $\sigma^2 = \mathrm{Var}(X_i)$, pelo **Teorema do Limite Central**

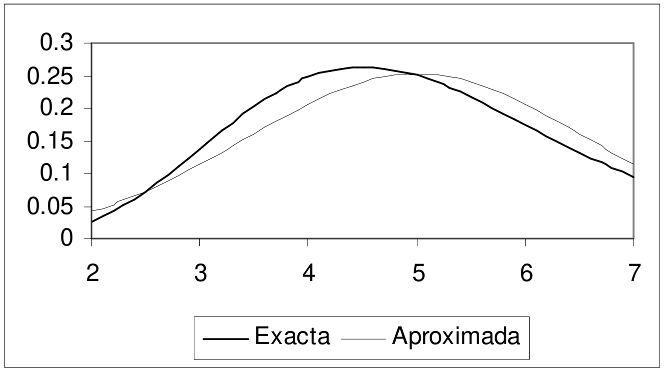
$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \overset{a}{\sim} N(0,1)$$

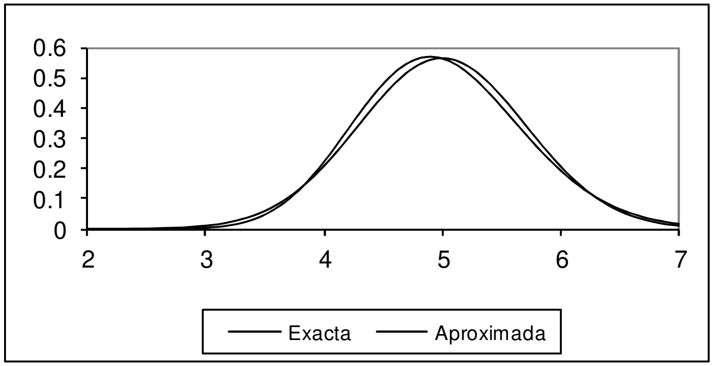
ou seja

$$\bar{X} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Exemplo 6.16 (alterado) – Considere uma população com distribuição $\text{Ex}(0.2)$ da qual se extraiu uma amostra de dimensão n. Compare a distribuição exacta com a distribuição aproximada de $\bar{X}$ para uma amostra de dimensão $n=10$ e para uma amostra de dimensão $n=50$.

Distribuição exacta: $\bar{X} \sim G(n, 0.2n)$. Distribuição aproximada: $\bar{X} \overset{a}{\sim} N(5; \frac{25}{n})$.





$n=10 \longrightarrow$ aproximação deficiente

$n=50 \longrightarrow$ aproximação aceitável

Exemplo 6.17 – Considerem-se as variáveis aleatórias *iid*, $X_1, X_2, \ldots, X_{30}$, com distribuição uniforme no intervalo (0,10). Pretende determinar-se $P(\bar{X} < 5.5)$.

Como o valor exacto é de difícil cálculo, recorre-se à distribuição assimptótica de $\bar{X}$. Tem-se $E(\bar{X}) = \mu = 5$ e $\sigma / \sqrt{30} = 0.53$.

Logo, $$P(\bar{X} < 5.5) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < \frac{5.5 - 5}{0.53}\right) \approx \Phi(0.94) = 0.8264.$$

7 – Amostragem de população de Bernoulli. Caso de uma proporção

- População é composta por elementos de dois tipos: os que possuem e os que não possuem determinado atributo ;
- Amostra casual $(X_1, X_2, ..., X_n)$: n variáveis aleatórias independentes e identicamente distribuídas, com função probabilidade individual da família

  $$\mathcal{F}_\theta = \{ f(x \mid \theta) = \theta^x (1-\theta)^{1-x} : x \in \{0,1\} \wedge 0 < \theta < 1 \}.$$

  e função probabilidade conjunta,

  $$\prod_{i=1}^{n} f(x_i \mid \theta) = \theta^{\sum_i x_i} (1-\theta)^{n-\sum_i x_i}, \quad 0 < \theta < 1, \quad x_i \in \{0,1\}, \quad i = 1,2,...,n.$$

- Interessa geralmente estabelecer a distribuição por amostragem de duas estatísticas: $Y = \sum_i X_i$ e $\overline{X} = \sum_i X_i / n$

- Solução:

1. $Y=\sum_i X_i \rightarrow$ soma de n variáveis aleatórias i.i.d. com distribuição de Bernoulli; logo $Y \sim B(n;\theta)$ e,

$$P(Y=y)=\binom{n}{y}\theta^{y}(1-\theta)^{n-y}, \quad y=0,1,\ldots,n,$$

$$P(\bar{X}=z)=P(Y=nz)=\binom{n}{nz}\theta^{nz}(1-\theta)^{n-nz}, \quad z=\frac{0}{n},\frac{1}{n},\ldots,\frac{n}{n}.$$

2. Quando a dimensão da amostra é razoavelmente grande, o teorema de De Moivre-Laplace permite estabelecer,

$$\frac{Y-n\theta}{\sqrt{n\theta(1-\theta)}} \overset{a}{\sim} N(0,1), \qquad \frac{\bar{X}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \overset{a}{\sim} N(0,1).$$

Utilizar a correcção de continuidade

3. Por vezes, torna-se aconselhável aproximar pela Poisson $\rightarrow Y=n\bar{X} \overset{a}{\sim} \mathrm{Po}(n\theta)$

Exemplo 6.19 – Admita que uma instituição bancária classifica os seus clientes possuidores de cartões de crédito em “maus” e “bons” riscos, conforme tenham ou não faltado a um pagamento nos últimos 2 anos. Suponha-se que a proporção de “maus” riscos (classificados por $X=1$) é de 0.05 para as agências da zona de Lisboa. Qual a probabilidade de se obter pelo menos 10% de maus riscos numa amostra de: (a) 10 clientes; (b) 50 clientes; (c) 400 clientes?

A resposta a qualquer das três alíneas é obtida calculando $P(\bar{X} \geq 0.1)$, sabendo-se que $X_i \sim B(1;0.05)$ para $i=1,\ldots,n$.

(a) pequena amostra → distribuição binomial

$$P(\bar{X} \geq 0.1) = P\left(\sum_{i=1}^{10} X_i \geq 10 \times 0.1\right) = P\left(\sum_{i=1}^{10} X_i \geq 1\right) = 0.4013$$

(b) $n = 50 > 20$, $\theta = 0.05 \leq 0.1$, $n\theta = 50 \times 0.05 = 2.5 < 5$ → aproximar pela Poisson

$$P(\bar{X} \geq 0.1) = P\left(\sum_{i=1}^{50} X_i \geq 50 \times 0.1\right) = P\left(\sum_{i=1}^{50} X_i \geq 5\right) = 1 - P\left(\sum_{i=1}^{50} X_i \leq 4\right) \approx 0.1088$$

valor “exacto” → 0.1036

(c) $n = 400 > 20$, $\theta = 0.05 \le 0.1$, $n\theta = 20 \ge 5$ $\rightarrow$ aproximar pela normal com correcção de continuidade

$$P(\bar{X} \ge 0.1) \approx 1 - \Phi\left(\frac{0.1 - (1/800) - 0.05}{\sqrt{0.05 \times 0.95 / 400}}\right) \approx 1 - \Phi\left(\frac{0.04875}{0.0109}\right) \approx 1 - \Phi(4.47) \approx 0.$$

Como é natural, a probabilidade de um acontecimento "invulgar" diminui com o crescimento da dimensão da amostra.

8 – Amostragem de população de Bernoulli. Caso de duas proporções

- 2 populações de Bernoulli com parâmetros θ_1 e θ_2 respectivamente.

  Habitualmente, quer-se comparar as duas proporções θ_1 e θ_2 (por exemplo, proporção de curas nos doentes tratados com o medicamento *A* e nos doentes tratados com o medicamento *B*).

  Nos estudos por amostragem esta diferença $(\theta_1 - \theta_2)$ nunca pode ser conhecida exactamente; A ideia será reolher 2 amostras independentes (uma de cada população) e utilizar a estatística $\bar{X}_1 - \bar{X}_2$ (a diferença entre proporções observadas) para inferir sobre $(\theta_1 - \theta_2)$.

- 2 amostras casuais independentes uma da outra:
  - $(X_{11}, X_{12}, \ldots, X_{1m}) \Rightarrow \bar{X}_1 = \sum_{i=1}^{m} X_{1i} / m$,
  - $(X_{21}, X_{22}, \ldots, X_{2n}) \Rightarrow \bar{X}_2 = \sum_{j=1}^{n} X_{2j} / n$,

- Distribuição por amostragem de $\bar{X}_1 - \bar{X}_2$
  - Pequenas amostra: Não existe resultado exacto que seja "simpático"
  - Distribuição assintótica (amostras razoavelmente grandes)

    Teorema de De Moivre-Laplace,

$$\bar{X}_1 \overset{a}{\sim} N\left(\theta_1, \frac{\theta_1(1-\theta_1)}{m}\right), \quad \bar{X}_2 \overset{a}{\sim} N\left(\theta_2, \frac{\theta_2(1-\theta_2)}{n}\right).$$

Logo

$$\frac{\bar{X}_1 - \bar{X}_2 - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{m} + \frac{\theta_2(1-\theta_2)}{n}}} \overset{a}{\sim} N(0,1).$$

Exemplo 6.20 – Retome-se o exemplo anterior e suponha-se que a percentagem de "maus" riscos na zona do Porto é de 0.06. Recolhidas amostras independentes nas zonas de Lisboa (índice 1) e Porto (índice 2) de dimensão 400 e 500 respectivamente, qual a probabilidade de se observar uma proporção maior de "maus" riscos em Lisboa do que no Porto?

$$P(\bar{X}_1 - \bar{X}_2 > 0) = P\left(\frac{\bar{X}_1 - \bar{X}_2 - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{m} + \frac{\theta_2(1-\theta_2)}{n}}} > \frac{0 - (0.05 - 0.06)}{\sqrt{\frac{0.05 \times 0.95}{400} + \frac{0.06 \times 0.94}{500}}} \right)$$

$$\approx 1 - \Phi(0.66) \approx 0.2546 \ .$$

Este valor evidencia os cuidados que se devem ter no processo de inferência das conclusões amostrais para a população. Com efeito, embora a proporção de "maus" riscos seja menor em Lisboa do que no Porto, a probabilidade da média da amostra de Lisboa ser superior à da média da amostra do Porto é aproximadamente 25%.

10 – População normal: distribuição da média

- $(X_1, X_2, ..., X_n)$ amostra casual da população normal, $N(\mu, \sigma^2)$.
- Recorde-se que $E(\bar{X}) = \mu$ e que $\mathrm{Var}(\bar{X}) = \sigma^2 / n$. À medida que a amostra cresce, a variância de $\bar{X}$ diminui.
- Assim $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ ou $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- **Exemplo 6.21 –** Suponha-se que a duração das chamadas telefónicas locais em determinada empresa pode ser bem aproximada por uma distribuição normal com média igual a 17 minutos e variância 25. Qual a probabilidade de, numa amostra aleatória de *n* chamadas, a duração média se situar entre (a) 16 e 18 minutos e (b) 14 e 16 minutos?

  Exemplificar para $n = 25$ e para $n = 100$.

  a) $P\left(16 < \bar{X}_{25} < 18\right) = 0.6826$ e $P\left(16 < \bar{X}_{100} < 18\right) = 0.9544$

  b) $P\left(14 < \bar{X}_{25} < 16\right) = 0.1573$ e $P\left(14 < \bar{X}_{100} < 16\right) = 0.02275$

10 – População normal: distribuição da variância

- $(X_1, X_2, ..., X_n)$ amostra casual da população normal, $N(\mu, \sigma^2)$.
- Sabe-se (Teorema 5.8) que $\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_{(n)}$
- Demonstra-se que (**Teorema 6.3**) – Se $(X_1, X_2, ..., X_n)$ é uma amostra casual de uma da população normal, $N(\mu, \sigma^2)$, então,

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)\, S'^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

***Dem.*:** Livro

- Ao comparar os 2 resultados vê-se que se perde um grau de liberdade por utilizar $\overline{X}$ em vez de μ

Exemplo 6.22 – Considere-se uma população normal da qual se extraiu uma amostra de dimensão 25. Calcule a probabilidade do quociente entre a variância corrigida da amostra e a variância da população se situar entre 0.79 e 1.18.

$$P\left(0.79 < \frac{S'^2}{\sigma^2} < 1.18\right) = P\left(18.96 < \frac{(n-1)S'^2}{\sigma^2} < 28.32\right) \approx 0.75 - 0.25 = 0.5 .$$

11 – População normal: rácio de "*Student*"

- $(X_1, X_2, ..., X_n)$ amostra casual da população normal, $N(\mu, \sigma^2)$.
- A variância σ^2 é desconhecida o que desaconselha a utilização de

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{ou} \qquad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0;1)$$

- Nesta situação, utiliza-se o **rácio de "*Student*"**,

$$\frac{\overline{X}-\mu}{S'/\sqrt{n}} = \frac{\overline{X}-\mu}{S/\sqrt{n-1}} \sim t(n-1)$$

- Este rácio tem uma distribuição designada por *t*-"*Student*" com (*n*-1) graus de liberdade (tabelas ou máquina)

- A distribuição *t-Student* (definição 6.3) pode ter origem num caso mais geral:

$$\left.\begin{array}{c} U \sim N(0,1) \\ V \sim \chi^2(n) \\ U \text{ e } V \text{ independentes} \end{array}\right\} \Rightarrow T = \frac{U}{\sqrt{V/n}} \sim t(n)$$

- função densidade de uma *t*-"*Student*" com *n* graus de liberdade:

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < +\infty.$$

- Simétrica em torno de $t=0$, abcissa que corresponde à moda (ordenada máxima);
- $E(T)=0;\ \ \mathrm{Var}(T)=\dfrac{n}{n-2}\ \ (n>2);\ \ \gamma_1=0;\ \ \gamma_2=\dfrac{3(n-2)}{n-4}\ \ (n>4).$
  - Tende para a $N(0;1)$ quando $n \to \infty$.

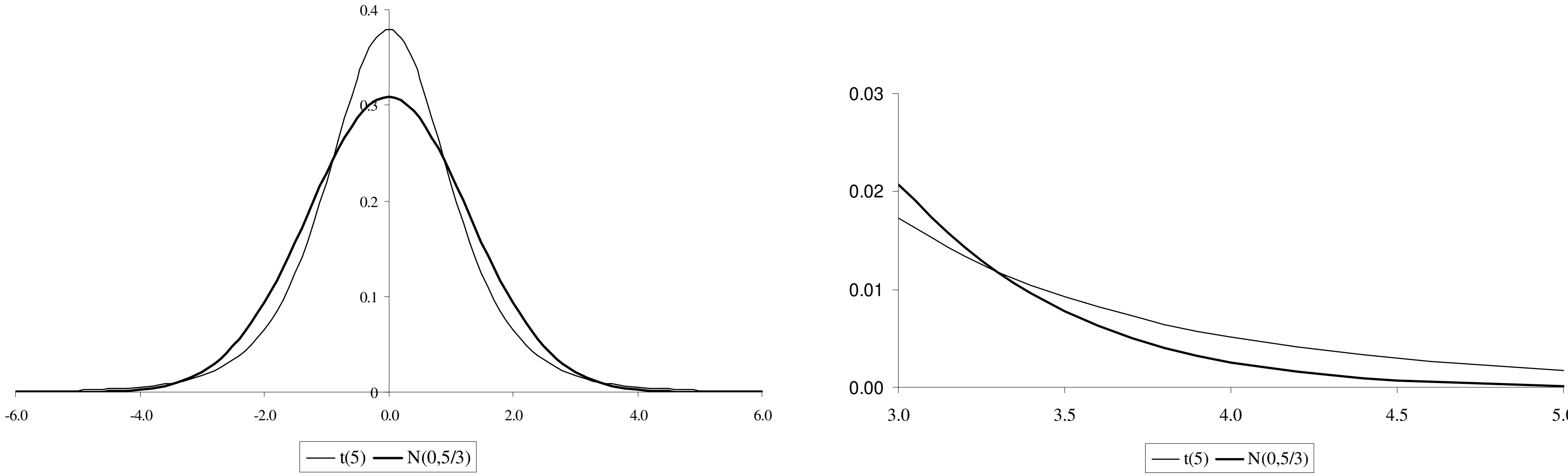
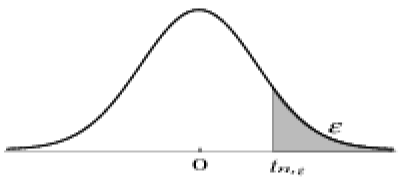


Fig. 6.8a e 6.8b – Comparação da $N(0,5/3)$ com a $t(5)$, densidade e cauda direita (as 2 distribuições têm a mesma média e a mesma variância)

TABELA 7 – DISTRIBUIÇÃO *t*-“*Student*”

$$t_{n,\varepsilon} : P(X > t_{n,\varepsilon}) = \varepsilon$$



$n \backslash \varepsilon$	**.400**	**.250**	**.100**	**.050**	**.025**	**.010**	**.005**	**.001**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.656	318.289
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.328
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173
…								
100	0.254	0.677	1.291	1.662	1.987	2.368	2.632	3.183
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090

12 – Populações normais: diferença entre duas médias

- 2 populações normais: $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$
- 2 amostras casuais independentes (dimensão *m* e *n* respectivamente)

$$(X_{11}, X_{12}, \ldots, X_{1m}) \quad \text{e} \quad (X_{21}, X_{22}, \ldots, X_{2n}),$$

- Estatísticas $\overline{X}_1 = \frac{1}{m}\sum_{i=1}^{m} X_{1i}$ e $\overline{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$
- Facilmente se conclui que,

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

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

- O resultado anterior só tem aplicação quando as variâncias das duas populações são conhecidas (problema semelhante ao que levou a introduzir do rácio de "*Student*")
- Quando as **variâncias**, embora **desconhecidas**, são **iguais**, pode recorrer-se a outro resultado para estabelecer inferências sobre $\mu_1 - \mu_2$.

Quando $\sigma_1^2 = \sigma_2^2 = \sigma^2$, tem-se

$$T = \frac{\dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}}{\sqrt{\dfrac{(m-1)S_1'^2 + (n-1)S_2'^2}{m+n-2}}} \sim t(m+n-2)$$

recorrendo à definição 6.3 com

$$U = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1),$$

$$V = \frac{(m-1)S_1'^2 + (n-1)S_2'^2}{\sigma^2} \sim \chi^2(m+n-2)$$

e mostrando que U e V são independentes

Tem-se $V \sim \chi^2(m+n-2)$ já que $\frac{(m-1)S_1'^2}{\sigma_1^2} \sim \chi^2(m-1)$ e $\frac{(n-1)S_2'^2}{\sigma_2^2} \sim \chi^2(n-1)$ e recordando que a qui-quadrado é aditiva (variáveis aleatórias independentes)

- Quando as **variâncias** das populações são **desconhecidas e diferentes**, as inferências sobre $\mu_1 - \mu_2$ tornam-se bem mais complexas.
  - Amostras grandes → distribuição assintótica Normal: substituir as variâncias da população pelas variâncias das amostras.
  - Amostras pequenas (particularmente se $m \neq n$) → **aproximação de Welch**:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1'^2}{m} + \frac{S_2'^2}{n}}} \overset{a}{\sim} t_{(r^*)},$$

sendo r^* dado pelo maior inteiro contido em,

$$r = \frac{\left(\frac{s_1'^2}{m} + \frac{s_2'^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1'^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{s_2'^2}{n}\right)^2}$$

13 – Populações normais: relação entre duas variâncias

- Para inferir sobre a relação entre as variâncias, σ_1^2 / σ_2^2, de duas populações normais **independentes** é natural pensar na estatística $S_1'^2 / S_2'^2$.
- Sendo as duas amostras independentes, torna-se fácil ver que esta estatística pode ser relacionada com quociente de duas variáveis independentes com distribuição do qui-quadrado, já que

$$U = \frac{(m-1)S_1'^2}{\sigma_1^2} \sim \chi^2(m-1) \;\text{ e }\; V = \frac{(n-1)S_2'^2}{\sigma_2^2} \sim \chi^2(n-1),$$

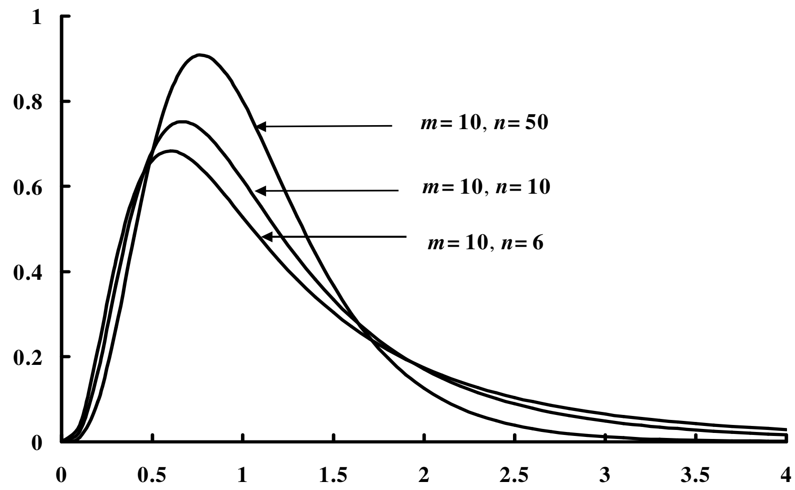
- Deste estudo resultou

$$F = \frac{U/(m-1)}{V/(n-1)} = \frac{S_1'^2}{S_2'^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(m-1, n-1)$$

Em que a variável F tem distribuição F-Snedecor com $m-1$ e $n-1$ graus de liberdade.

- Tal como a *t-Student* a *F-Snedecor* pode ser definida num quadro mais geral (definição 6.4)

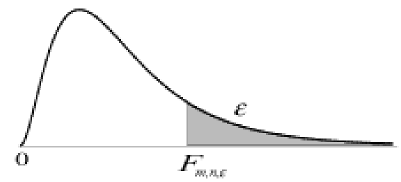
$$\left.\begin{array}{r} U \sim \chi^2(m) \\ V \sim \chi^2(n) \\ U \text{ e } V \text{ independentes} \end{array}\right\} \Rightarrow F = \frac{U/m}{V/n} \sim F(m,n)$$



Funções densidade de uma distribuição *F*-Snedcor

- $E(X)=\dfrac{n}{n-2}\quad (n>2),\quad \mathrm{Var}(X)=\dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}\quad (n>4).$
- Tabela 8: valores $F_{m,n,\varepsilon}$ para alguns pares (m,n) e para valores de ε de emprego frequente – 0.05, 0.025 e 0.01 – tais que $X \sim F(m,n) \Rightarrow P(X > F_{m,n,\varepsilon}) = \varepsilon$.

TABELA 8 – DISTRIBUIÇÃO *F*-SNEDCOR



$$F_{m,n,\varepsilon} : P(X > F_{m,n,\varepsilon}) = \varepsilon$$

n – graus de liberdade do denominador		m - graus de liberdade do numerador																		
	ε	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	**.100**	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.90	245.95	248.02	249.05	250.10	251.14	252.20	253.25	254.32
	.025	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	976.72	984.87	993.08	997.27	1001.40	1005.60	1009.79	1014.04	1018.26
	.010	4052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5928.33	5980.95	6022.40	6055.93	6106.68	6156.97	6208.66	6234.27	6260.35	6286.43	6312.97	6339.51	6365.59
2	**.100**	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	.010	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.50
3	**.100**	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	**.100**	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	**.100**	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	**.100**	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	**.100**	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65

- Os valores indicados pela tabela 8 situam-se na aba da direita da distribuição. Para obter valores na aba da esquerda, isto é, valores $F^{*}_{m,n,\varepsilon}$ tais que,

$$X \sim F(m,n) \Rightarrow P(X < F^{*}_{m,n,\varepsilon}) = \varepsilon,$$

tem de atender-se a uma propriedade da *F*-Snedcor que estabelece,

$$X \sim F(m,n) \Rightarrow Y = \frac{1}{X} \sim F(n,m).$$

Exemplo 6.23 – Suponha-se que os resultados do teste QI são bem modelados por distribuições normais de média 100 nos países *A* e *B* e que se recolheu uma amostra de dimensão 16 no país *A* e outra de dimensão 10 no país *B*. Admitindo que as variâncias nas duas populações são iguais, qual a probabilidade do quociente entre as variâncias corrigidas das duas amostras, $S_A'^2 / S_B'^2$, ser superior a 3.77?

$$P\left(\frac{S_A'^2}{S_B'^2} > 3.77\right) \approx 0.025, \text{ com } \frac{S_A'^2}{S_B'^2} \sim F(15;9).$$

Embora as variâncias sejam iguais nas duas populações, o facto de se estar em presença de amostras de dimensão diferente leva a que,

$$P\left(\frac{S_A'^2}{S_B'^2} > 3.77\right) \neq P\left(\frac{S_B'^2}{S_A'^2} > 3.77\right).$$

Suponha-se agora que se pretendia calcular a probabilidade de $S_A'^2 / S_B'^2 < 0.386$

$$P\left(\frac{S_A'^2}{S_B'^2} < 0.386\right) = P\left(\frac{S_B'^2}{S_A'^2} > \frac{1}{0.386}\right) = P\left(\frac{S_B'^2}{S_A'^2} > 2.59\right) \approx 0.05.$$