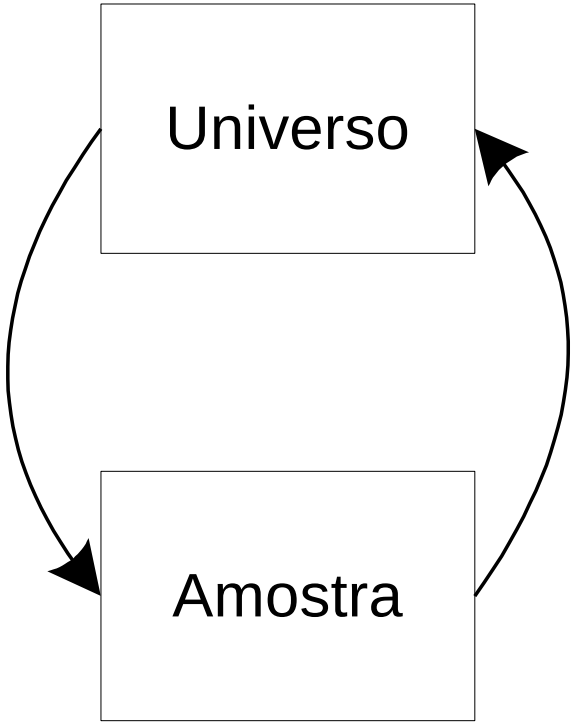
5. Amostragem. Distribuições por amostragem

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;



Processo de Amostragem

Inferência indutiva

Exemplo 5.1 – Considere-se um grupo numeroso de pessoas (por exemplo, os estudantes matriculados no ISEG no ano letivo 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, $0 \leq x \leq 10$, 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 – <u>Amostragem casual</u> -** Quando as *n* variáveis aleatórias observadas, componentes do vetor $(X_1, X_2, ..., X_n)$, são **<u>independentes</u> e <u>identicamente distribuídas</u>** – simbolicamente **<u>iid</u>** – diz-se que se trata de amostragem casual.
  - Cada X_i, $i = 1,2,...,n$, é uma **"cópia"** da variável aleatória X
  - Independência entre os X_i, $i = 1,2,...,n$.

- Processo de amostragem aleatório → 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, $(x_1, x_2, ..., x_n)$, é uma realização da variável aleatória *n*-dimensional $(X_1, X_2, ..., X_n)$.

  - $(X_1, X_2, ..., X_n)$ — Amostra aleatória
  - $(x_1, x_2, ..., x_n)$ — Amostra observada

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

$$\text{Popula}\tilde{c}ao \Rightarrow \begin{cases} Amostra\ \ 1 \\ Amostra\ \ 2 \\ \cdots \\ Amostra\ \ m \\ \cdots \end{cases}$$

Exemplo 5.2. – Assuma-se que X (pratica ou não pratica desporto) é uma variável de Bernoulli de parâmetro θ, isto é

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

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.

5.3. – Estatísticas

- **Definição – <u>Estatística</u>**

  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 5.3. –** 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 5.4** – Se $(X_1, X_2, \ldots, X_n)$ é amostra casual de população normal $N(\mu, \sigma^2)$ com parâmetros μ e σ^2 desconhecidos, são exemplos de estatísticas unidimensionais,

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

e de estatísticas bidimensionais,

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

Não são estatísticas as funções,

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

pois dependem de parâmetros desconhecidos.

5.4. Distribuição da amostra e distribuição por amostragem da estatística

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

- O comportamento probabilístico da estatística $T(X_1, X_2, ..., X_n)$ é dado pela respetiva 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):**

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

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

$$G(t|\theta) = P(T \leq t) = \int \cdots \int_{A(t)} [\prod_{i=1}^{n} f(x_i|\theta)]\, dx_1 dx_2 \ldots dx_n,$$

no caso de *T* ser variável aleatória contínua, ou,

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

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

$$A(t) = \{(x_1, x_2, \ldots, x_n) \in \Re^n : T(x_1, x_2, \ldots, x_n) \leq t\},$$

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

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

a) Assumindo que algumas condições de verificam é possível, por vezes, derivar a distribuição exata de $T(X_1, X_2, ..., X_n)$.

b) É geralmente possível obter distribuições aproximadas (Teorema do Limite Central)

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

Exemplo 5.5 – Se $(X_1, X_2, ..., X_n)$ é uma amostra casual de uma população de Poisson, $X_i \sim \text{Po}(\theta)$, então, pelo teorema 5.3, tem-se $T = \sum_i X_i \sim \text{Po}(n\theta)$. Assim, a estatística T tem função probabilidade,

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

Exemplo 5.6 – Se $(X_1, X_2, ..., X_n)$ é uma amostra casual de uma população exponencial, $X_i \sim \text{Ex}(\theta)$, então, pelo teorema 5.8, $T = \sum_i X_i \sim \text{G}(n, \theta)$. Assim, a estatística T tem função densidade,

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

5.5. Estatísticas de ordem. Distribuição por amostragem do máximo e do mínimo amostrais

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

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

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 respetivas funções densidade

Exemplo 5.7: Seja X um universo com distribuição exponencial de parâmetro λ.

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

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

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

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

que não é a função de distribuição de uma exponencial.

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

5.6. Primeiros resultados sobre a média e variância amostrais.

- **Média** e **variância amostrais**

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

- **Teorema** – 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 $\sigma^2 = \mathrm{Var}(X_i)$ $(i = 1,2,\ldots,n)$, tem-se,

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

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

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

.

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

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

5. 7. Distribuições por amostragem assintóticas

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

Distribuição assintótica da Média

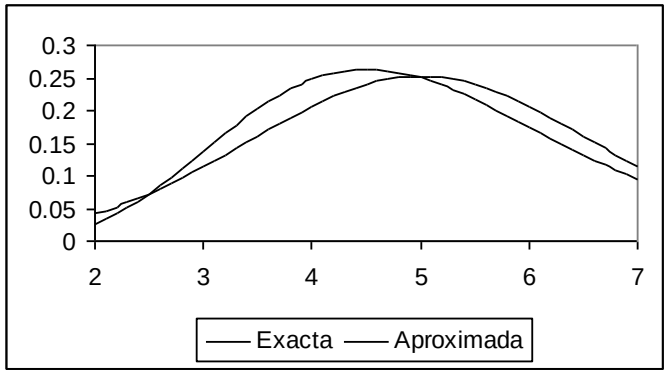
Se $(X_1, X_2, \ldots, X_n)$ é uma amostra casual de população para a qual existem média $\mu = E(X_i)$ e variância $\sigma^2 = \text{Var}(X_i)$, pelo **Teorema do Limite Central**

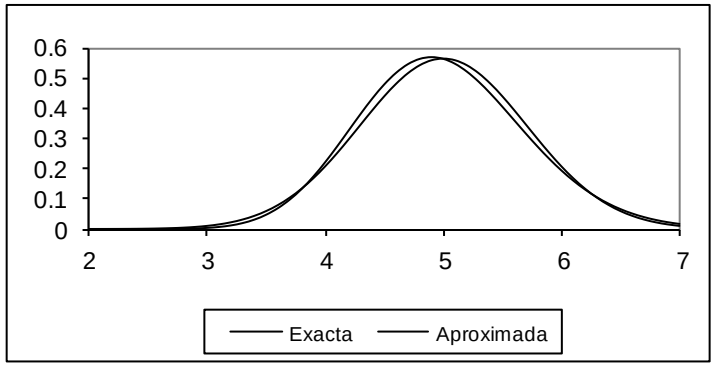
$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$$, ou seja

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

Exemplo 5.8 – Considere uma população com distribuição Ex(0.2) da qual se extraiu uma amostra de dimensão n. Compare a distribuição exata com a distribuição aproximada de $\bar{X}$ para uma amostra de dimensão $n = 10$ e para uma amostra de dimensão $n = 50$.

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





$n = 10 \rightarrow$ aproximação deficiente $\qquad$ $n = 50 \rightarrow$ aproximação aceitável

Exemplo 5.9 – Considerem-se as variáveis aleatórias *iid*, $X_1, X_2, \ldots, X_{30}$, com distribuição uniforme no intervalo (0,10). Calcule $P(\bar{X} < 5{,}5)$

5.8 – 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 $(X_1, X_2, \ldots, X_n)$: n variáveis aleatórias independentes e identicamente distribuídas, com função probabilidade individual da família

$$F_\theta = \{f(x|\theta) = \theta^x(1-\theta)^{1-x} : x \in \{0,1\},\ 0 < \theta < 1\}.$$

e função probabilidade conjunta,

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

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

- Solução:

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

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

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

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

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

(Utilizar a correção de continuidade)

3- Por vezes, torna-se aconselhável aproximar pela Poisson (lei dos acontecimentos raros)

$$Y = n\bar{X} \overset{a}{\sim} \text{Po}(n\theta)$$

__ *a*

Exemplo 5.10 – 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?

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

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

  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 exatamente; A ideia será recolher 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 exato 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 5.11 – 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 zo- nas de Lisboa (índice 1) e Porto (índice 2) de dimensão 400 e 500 respetivamente, qual a probabilidade de se observar uma proporção maior de "maus" riscos em Lisboa do que no Porto?

5.10. População normal: distribuição da média

- $(X_1, X_2, \ldots, X_n)$ amostra casual da população normal, $N(\mu, \sigma^2)$.
- Recorde-se que $E(\bar{X}) = \mu$ e que $\text{Var}(\bar{X}) = \sigma^2/n$. À medida que a amostra cresce, a variância de $\bar{X}$ diminui.
- Assim $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ ou $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- **Exemplo 5.12 –** 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 $n = 100$.

5.11. População normal: distribuição da variância

- $(X_1, X_2, \ldots, X_n)$ amostra casual da população normal, $N(\mu, \sigma^2)$.
- É possível provar que $\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)}$
- Demonstra-se que se $(X_1, X_2, \ldots, X_n)$ é uma amostra casual de uma da população normal, $N(\mu, \sigma^2)$, então,

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

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

Exemplo 5.13 – 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.

5. 12 – População normal: rácio de "*Student*"

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

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

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

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

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

- A distribuição *t-Student* 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);
- $E(T) = 0$; $\text{Var}(T) = \frac{n}{n-2}$ $(n > 2)$; $\gamma_1 = 0$; $\gamma_2 = \frac{3(n-2)}{n-4}$ $(n > 4)$.
- Tende para a $N(0;1)$ quando $n \to \infty$.

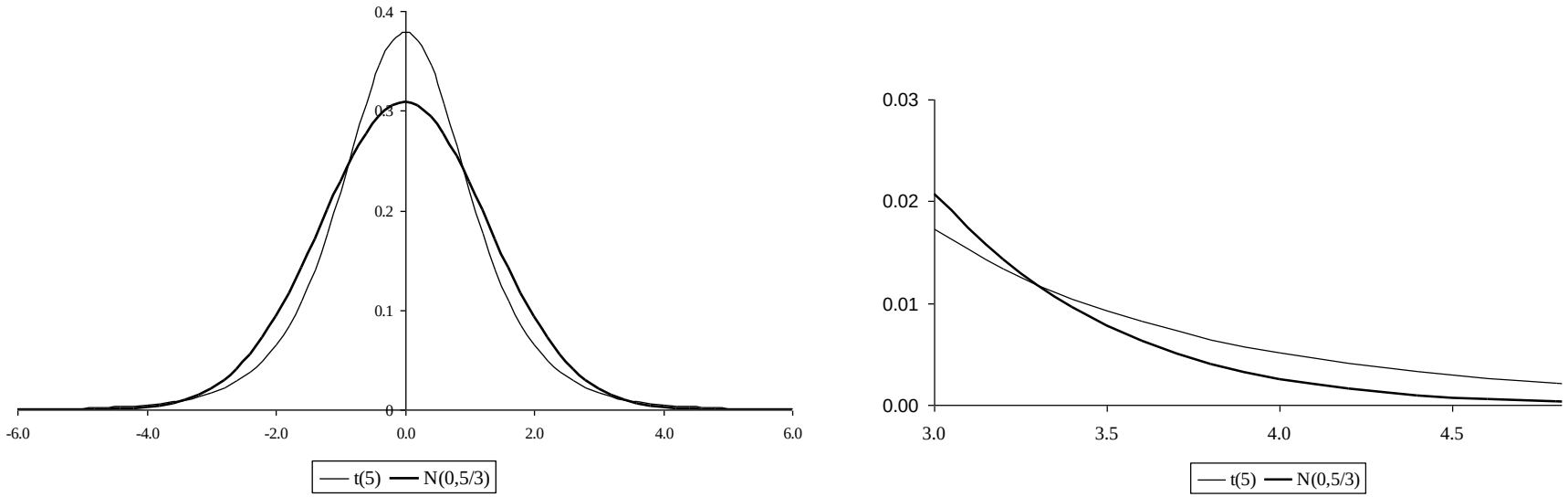


Fig. – Comparação da $N(0,5/3)$ com a $t(5)$, densidade e cauda direita

(as 2 distribuições têm a mesma média e a mesma variância)

5.13. 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 <u>independentes</u> (dimensão *m* e *n* respetivamente)

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

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

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

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

- 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

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

- 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**:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S'^2_1}{m} + \frac{S'^2_2}{n}}} \overset{a}{\sim} t(r^*),$$

sendo r^* dado pela parte inteira do número

$$r = \frac{\left(\frac{s'^2_1}{m} + \frac{s'^2_2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s'^2_1}{m}\right)^2 + \frac{1}{n-1}\left(\frac{s'^2_2}{n}\right)^2}$$

Exemplo: Parte inteira de 2,73 é 2. A parte inteira de 1,1 é 1.

5.14. 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'^2_1/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

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

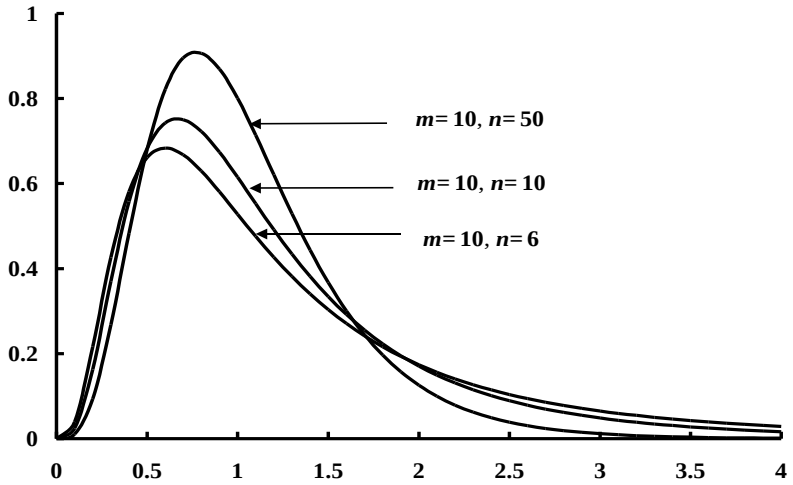
- Deste estudo resultou

$$F = \frac{U/(m-1)}{V/(n-1)} = \frac{S'^2_1}{S'^2_2}\,\frac{\sigma_2^2}{\sigma_1^2} \sim F(m-1, n-1)$$

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

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



Funções densidade de uma distribuição *F*-Snedcor

- $E(X) = \frac{n}{n-2} \quad (n > 2), \quad \text{Var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (n > 4).$
- 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 $X \sim F(m, n) \Rightarrow P(X > F_{m,n,\varepsilon}) = \varepsilon$.
- 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 5.14 – 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'^2_A/S'^2_B, ser superior a 3,77? Calcule a probabilidade de $S'^2_A/S'^2_B<0{,}386$.