1 Amostragem e Distribuições por Amostragem

1.1 Probabilidade e Inferência Estatística

Processos "complementares":

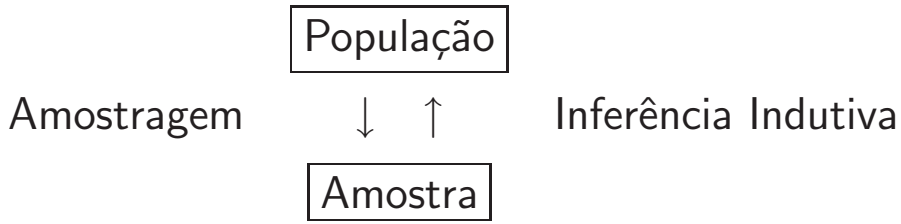
- Teoria da probabilidade: Parte-se de um modelo totalmente especificado, que se assume como correcto e calcula-se, e.g., a probabilidade de certos acontecimentos
- Inferência estatística: Observam-se certos acontecimentos, e procura-se inferir sobre o modelo probabilístico pelo qual se regirá a experiência aleatória.

Exemplo 1.1 *Considere-se um grupo numeroso de pessoas entre as quais há uma proporção θ de fumadores.*

- *Se θ conhecido e estivermos interessados em conhecer a probabilidade de encontrar x fumadores num grupo de 10 pessoas escolhidos ao acaso, temos um problema no domínio da Teoria da Probabilidade.*
- *Na prática, sucede quase sempre que θ é desconhecido. A partir da observação do número de fumadores na amostra de 10 pessoas, pretende-se tirar conclusões sobre a proporção de fumadores na população, θ. Trata-se então de um problema de Inferência Estatística*

■

- Dados estatísticos resultam de experiências conduzidas sobre um conjunto restrito — amostra — e procura-se alargar as conclusões à população — conjunto mais vasto
- Diagrama habitual:



1.2 Especificação. Amostragem casual.

Modelação matemática do processo de inferência

- Característica de interesse no universo é uma variável aleatória X com função de distribuição F que é um elemento de um conjunto $\mathcal{F}$ — o *modelo estatístico*
- Esse modelo tem que ser especificado:
  - modelos paramétricos — F é conhecida a menos de um parâmetro (de dimensão finita, k), eg, F é normal com média μ e variância σ^2 desconhecidas
  - modelos não-paramétricos — F especificada de forma não-paramétrica, eg, F é um elemento do conjunto das distribuições simétricas e (absolutamente) contínuas;

O caso que nos interessa aqui é o paramétrico. Modelo estatístico paramétrico:

$$\mathcal{F} = \{F(\cdot \mid \theta) : \theta \in \Theta\}$$

Ao conjunto Θ dá-se a designação de espaço paramétrico.

Exemplo 1.1 *(Continuação) A característica de interesse é* $X = 1$ *se o indivíduo for fumador e* $X = 0$ *caso contrário, logo o modelo estatístico consiste na família das distribuições de Bernoulli, isto é,* $\mathcal{F} = \{\theta^x(1-\theta)^{1-x},\ x = 0, 1 : \theta \in (0, 1), \}$. ■

Exemplo 1.2 *Para se estudar o retorno diário gerado por determinado activo financeiro (acções ou obrigações) utiliza-se frequentemente o seguinte modelo: seja* V_t *o preço de esse activo no instante* t. *Então, assume-se que*

$$R_t = \ln \frac{V_t}{V_{t-1}},\ t = 1, \ldots, T$$

são variáveis aleatórias independentes com distribuição log-normal. Assim, $\mathcal{F} = \{\mathrm{LN}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$,

Para estudar o número de sinistros num ano de uma apólice de seguro, recorrere-se frequentemente à distribuição de Poisson: $\mathcal{F} = \{\mathrm{Po}(\lambda) : \lambda > 0\}$. ■

A especificação é uma fase fundamental da inferência estatística. Resulta de uma combinação de vários factores, nomeadamente:

- conhecimento do fenómeno em estudo
- resultados de estudos anteriores
- conhecimentos de teoria da probabilidade para escolher um modelo probabilístico com as características desejadas

As consequências de uma má especificação serão sempre negativas mas em geral são tanto mais negativas quanto menor for a dimensão da amostra.

Processo de amostragem

- Existem naturalmente vários processos de amostragem, i.e., formas de recolher informação relativa à característica de interesse num conjunto restrito de elementos da população
- processo de amostragem aleatório: os dados observados são apenas um dos muitos conjuntos de dados que poderiam ter sido obtidos operando nas mesmas circunstâncias. O conjunto de n observações, $(x_1, \ldots, x_n)$, que se observou é uma realização da variável aleatória n-dimensional $(X_1, \ldots, X_n)$:

$$\begin{array}{ll} (X_1, \ldots, X_n) & \text{Amostra aleatória} \\ (x_1, \ldots, x_n) & \text{Amostra observada} \end{array}$$

- Espaço amostral: subconjunto de $\mathbb{R}^n$ que corresponde ao conjunto dos valores que a amostra $(x_1, \ldots, x_n)$ pode assumir. Designa-se por $\mathcal{X}$.
- Nesta disciplina, vamo-nos restringir quase exclusivamente a um processo de amostragem aleatório em particular:

Definição 1.1 Amostragem Casual: *Quando as n variáveis aleatórias que constituem a amostra aleatória são*

1. *independentes*
2. *identicamente distribuídas, com a mesma distribuição que X*

diz-se que $(X_1, \ldots, X_n)$ constitui uma amostra casual de tamanho n proveniente da população X. Simbolicamente, $X_1, \ldots, X_n \overset{iid}{\sim} X$. ■

Se $\mathcal{F} = \{F(\cdot \mid \theta) : \theta \in \Theta\}$ e $X_1, \ldots, X_n \overset{iid}{\sim} X$, então

$$\begin{aligned} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} F_{X_i}(x_i \mid \theta) \quad \text{por independência} \\ &= \prod_{i=1}^{n} F(x_i \mid \theta) \quad \text{pois } X_i \sim X \end{aligned}$$

e similarmente para a função (densidade de) probabilidade:

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \ .$$

Exemplo 1.3 *Se $X_1, \ldots, X_n$ é uma amostra casual proveniente de uma população $Po(\lambda)$, então*

$$P(X_1 = x_1, \ldots, X_n = x_n) = f(x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}, \ x_i \in \mathbb{N}_0$$

Se $X_1, \ldots, X_n$ é uma amostra casual proveniente de uma população $N(\mu, 1)$, então

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i - \mu)^2\right] = (2\pi)^{-n/2} \ \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right], \ x_i \in \mathbb{R}$$

■

1.3 Estatísticas

Definição 1.2 Estatística *Estatística é qualquer função da amostra casual que não dependa de parâmetros desconhecidos.* ■

Exemplo 1.4 *No contexto de uma população* $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ *e* $\sigma > 0$, *são exemplos de estatísticas unidimensionais*

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

e de estatísticas bidimensionais

$$\left(T, \sum_{i=1}^{n} X_i^2 \right), \quad \left(\bar{X}, S^2 \right) \ .$$

Não são estatísticas as funções

$$\sum_{i=1}^{n} (X_i - \mu)^2, \quad \frac{1}{\sigma^2} \sum_{i=1}^{n} X_i^2$$

pois dependem de parâmetros desconhecidos. Se σ^2 *for conhecido, já* $\frac{1}{\sigma^2} \sum_{i=1}^{n} X_i^2$ *é uma estatística.* ■

Exemplos de estatísticas

- A amostra casual $(X_1, \ldots, X_n)$ é uma estatística
- A média amostral

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

- A variância amostral

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

- A variância amostral corrigida

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

- o máximo da amostra, $\max\{X_1, \ldots, X_n\}$
- X_1 ou X_n

- Estatísticas são sumários da informação contida na amostra casual
- Estatísticas operam uma redução dos dados: claramente, observar $(X_1, \ldots, X_n)$ é pelo menos tão informativo quanto observar $\bar{X}$; observar $(\bar{X}, S^2)$ é pelo menos tão informativo quanto observar apenas $\bar{X}$
- Estatísticas são variáveis aleatórias. Há que distinguir a variável aleatória do seu valor observado

população	amostra casual	amostra observada
X	$(X_1, \ldots, X_n)$	$(x_1, \ldots, x_n)$
média da população	média amostral	média da amostra
$\mu = E[X]$	$\bar{X} = \frac{1}{n}\sum_i X_i$	$\bar{x} = \frac{1}{n}\sum_i x_i$
variância da população	variância amostral	variância da amostra
$\sigma^2 = \mathrm{Var}(X)$	$S^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$	$s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$

1.4 Distribuição por amostragem

- A distribuição por amostragem de uma estatística corresponde à sua distribuição de probabilidade: variando a amostra casual $(X_1, \ldots, X_n)$ de acordo com a sua distribuição de probabilidades, qual o comportamento probabilístico de $T(X_1, \ldots, X_n)$ daí resultante para cada valor de θ
- Vai revelar-se muito importante conhecer a distribuição por amostragem de estatísticas para, por exemplo, avaliar o desempenho de metodologias estatísticas nelas baseadas
- O objectivo do resto deste capítulo consiste em determinar (aspectos d)a distribuição por amostragem de uma estatística T, sabendo (aspectos d)a distribuição de probabilidades da população X.

Métodos para a obtenção da distribuição por amostragem de uma estatística:

- Método da mudança de variável: Se X é contínua,

$$F_T(t \mid \theta) = P(T \leq t \mid \theta) = \int_{A(t)} \prod_{i=1}^{n} f(x_i \mid \theta) dx_1 \ldots dx_n$$

  onde $A(t) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : T(x_1, \ldots, x_n) \leq t\}$. No caso de X ser discreta, substituir os integrais por somatórios. Laborioso, e, nos casos em que se consegue obter resultados explícitos, existem quase sempre soluções mais elegantes

- Recurso à função geradora de momentos de T
- Recurso a propriedades conhecidas da distribuição de X (relacionado com o ponto anterior)
- Aproximação assintótica da distribuição por amostragem de certas estatísticas recorrendo ao Teorema do Limite Central
- Por simulação: estratégia cada vez mais importante dado o aumento da capacidade de cálculo dos computadores pessoais, especialmente quando não existem soluções analíticas

Exemplo 1.5 *Seja* $T = \sum_{i=1}^{n} X_i$.

- *Se* $(X_1, \ldots, X_n)$ *é uma amostra casual de uma população* $Po(\lambda)$, *então pelo facto de a soma de Poisson independentes ser ainda Poisson, tem-se que* $T \sim Po(n\lambda)$, *logo*

$$f_T(t \mid \lambda) = e^{-n\lambda} \frac{(n\lambda)^t}{t!}, \quad t \in \mathbb{N}_0 \ .$$

- *Se* $(X_1, \ldots, X_n)$ *é uma amostra casual de uma população* $N(\mu, \sigma^2)$, *então* $T \sim N(n\mu, n\sigma^2)$
- *Se* $(X_1, \ldots, X_n)$ *é uma amostra casual de uma população* $B(1, \theta)$, *então* $T \sim B(n, \theta)$.

■

1.4.1 Simulação Monte Carlo

Teorema 1.1 Lei Forte dos Grandes Números *Se $X_1, \ldots, X_n$ constitui uma amostra casual proveniente de uma população X com valor esperado μ finito, então quando $n \to +\infty$*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{q.c.} \mu$$

■

- Aplicação frequente: aproximar $E[X]$ por $\bar{x} = \sum_{i=1}^{n} x_i/n$ no contexto de uma amostra observada
- Outra aplicação: representar aproximadamente a distribuição de probabilidades de X por uma amostra de dimensão suficientemente grande $x_1, \ldots, x_n$ extraída dessa população, mas gerada em computador.
- (Quase) todos os aspectos dessa distribuição podem ser arbitrariamente aproximados com recurso exclusivo a $x_1, \ldots, x_n$:

- Probabilidades: suponha-se que se pretende calcular $P(X > a)$. Note-se que se se definir

$$Y = \begin{cases} 1 & \text{se } X > a \\ 0 & \text{se } X \leq a \end{cases}$$

então $E[Y] = P(X > a)$. Logo, sabemos que para n suficientemente grande

$$P(X > a) = E[Y] \approx \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{\#\{i : x_i > a\}}{n}$$

- Função densidade: Para $\delta > 0$ suficientemente pequeno,

$$f(a) = \lim_{\delta \to 0^+} \frac{F(a+\delta) - F(a)}{\delta} \approx \frac{1}{\delta} \frac{\#\{i : a < x_i \leq a + \delta\}}{n}$$

ou seja, o histograma da amostra $x_1, \ldots, x_n$ é uma boa aproximação à densidade de X.

Como gerar uma amostra casual de dimensão n de uma dada distribuição de probabilidade F?

- depende de F!
- para as distribuições mais usuais na prática, existem algoritmos implementados em computador para o efeito
- no *software* `R`:

Distribuição	Comando
$\mathrm{N}(\mu, \sigma^2)$	`rnorm(n,mu,sigma)`
$\mathrm{B}(N, \theta)$	`rbinom(n,N,theta)`
$\mathrm{Po}(\lambda)$	`rpoi(n,lambda)`
$\chi^2(\nu)$	`rchisq(n,nu)`
$t(\nu)$	`rt(n,nu)`
...	...

Simulação Monte Carlo para obter amostra de dimensão N proveniente da distribuição por amostragem da estatística $T = T(X_1, \ldots, X_n)$:

$x_{11}, \ldots, x_{1n}$	$t_1 = T(x_{11}, \ldots, x_{1n})$
$x_{21}, \ldots, x_{2n}$	$t_2 = T(x_{21}, \ldots, x_{2n})$
$\ldots$	$\ldots$
$x_{N1}, \ldots, x_{Nn}$	$t_N = T(x_{N1}, \ldots, x_{Nn})$

- Gerar N amostras de tamanho n provenientes da distribuição de X;
- para cada uma dessas amostras, calcular o valor observado da estatística T;
- Os N números resultantes, $(t_1, \ldots, t_N)$, constituem uma amostra de dimensão N proveniente da distribuição por amostragem de T.

Exemplo 1.6 *Simulação Monte Carlo*

- Obter $N = 10000$ amostras de dimensão $n = 10$ provenientes de uma população $X \sim \mathrm{N}(0, 1)$
- Obter uma amostra de dimensão N proveniente da distribuição por amostragem de $T = \sum X_i$ e de $S = \min X_i$
- Comparar os histogramas com as densidades exactas (que são neste caso conhecidas; a distibuição por amostragem de S vai ser estudada em breve)
- Código em R:

```
n <- 10
N <- 10000
x <- matrix(rnorm(N*n),ncol=n)
t <- apply(x,1,sum)
s <- apply(x,1,min)

hist(t,probability=T)
curve(dnorm(x,0,sqrt(n)),add=T,col=2)

hist(s,probability=T)
curve(n*dnorm(x,0,1)*(1-pnorm(x))^(n-1),add=T,col=3)
```

Histogram of t
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
−10
−5
0
5
10
t

Histogram of s
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−5
−4
−3
−2
−1
0
s

Cálcular aspectos da distribuição por simulação:

- $P(T \leq 1) = \Phi(1/\sqrt{n}) = 0.6241$ enquanto que por simulação se obtém $P(T \leq 1) \approx 0.6162$
- $P(S \leq -1) = 1 - [1 - \Phi(-1)]^n = 0.8223$ enquanto que por simulação se obtém $P(S \leq -1) \approx 0.8214$
- O cálculo de $E[S]$ não parece ser possível analiticamente. Contudo, por simulação tem-se $E[S] \approx -1.5362$
- $\mathrm{Cov}(S, T) \approx 0.9935$
- ...

1.5 Primeiras propriedades dos momentos amostrais

Definição 1.3 Momentos amostrais *Seja* $(X_1, \ldots, X_n)$ *uma amostra casual de tamanho* n *proveniente de uma população* X*. Para* $k \in \mathbb{N}$ *define-se o momento ordinário (ou em relação à origem) amostral de ordem* k *por*

$$M'_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

e o momento central amostral de ordem k *pr*

$$M_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^k \ .$$

■

Observação importante: mais uma vez, não confundir momentos amostrais, M'_k e M_k, com momentos populacionais, $\mu'_k = E[X^k]$ e $\mu_k = E[(X - E[X])^k]$, e com os momentos da amostra, $m'_k = \sum_{i=1}^{n} x_i^k/n$ e $m_k = \sum_{i=1}^{n} (x_i - \bar{x})^k/n$.

Casos particulares importantes: $\bar{X} = M'_1$ e $S^2 = M_2$, média amostral e variância amostral (não-corrigida), respectivamente.

Teorema 1.2 Propriedades da média amostral: *Existindo os momentos envolvidos, tem-se que*

$$
\begin{aligned}
E[\bar{X}] &= E[X] = \mu \\
Var(\bar{X}) &= \frac{Var(X)}{n} = \frac{\sigma^2}{n} \\
\mu_3(\bar{X}) &= \frac{\mu_3}{n^2} \\
\mu_4(\bar{X}) &= \frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} \ .
\end{aligned}
$$

■

Observações:

- Estes resultados são válidos para qualquer distribuição da população (desde que existam os momentos)
- A distribuição de $\bar{X}$ está centrada em torno de μ
- $\lim_{n\to+\infty} \mathrm{Var}(\bar{X}) = 0$

Teorema 1.3 Propriedades da variância amostral: *Existindo os momentos envolvidos, tem-se que*

$$E[S^2] = \frac{n-1}{n}\sigma^2$$
$$Var(S^2) = \frac{\mu_4 - \mu_2^2}{n} - 2\frac{\mu_4 - 2\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$$

■

Observações:

- $E[S^2] < \sigma^2$
- Por esta razão, define-se a **variância amostral corrigida**

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

Teorema 1.4 Propriedades da variância amostral corrigida: *Existindo os momentos envolvidos,*

$$E[S'^2] = \sigma^2$$
$$Var(S'^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$$

■

Teorema 1.5 Propriedades dos momentos amostrais centrais: *Existindo os momentos envolvidos, tem-se que*

$$E[M_k] = \mu_k + \mathcal{O}\left(\frac{1}{n}\right)$$

$$Var(M_k) = \frac{c}{n} + \mathcal{O}\left(\frac{1}{n^2}\right)$$

onde c é uma constante que envolve momentos populacionais centrais de ordem $\leq 2k$. ■

Teorema 1.6 Distribuição assintótica de $\bar{X}$: *Desde que $Var(X)$ exista, tem-se como consequência imediata do Teorema do Limite Central que*

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} N(0,1)\ .$$

■

Observações:

- O resultado acima é na maioria dos casos utilizado na forma

$$P(\bar{X} \leq x) \approx \Phi\left(\sqrt{n}\,\frac{x-\mu}{\sigma}\right)$$

- a partir de que valor de n é que a aproximação é suficientemente boa?
- Depende. Em geral, unimodalidade e simetria da população têm um impacto positivo na taxa de convergência.
- Podem derivar-se resultados idênticos para outros momentos amostrais, mas o seu interesse prático é mais limitado.

1.6 Estatísticas de ordem

Definição 1.4 Estatísticas de ordem: *Seja* $(X_1, \ldots, X_n)$ *uma amostra casual. A* i*-ésima estatística de ordem é denotada por* $X_{(i)}$ *e satisfaz*

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

■

Observações:

- Designa-se também por estatística de ordem qualquer função dos $X_{(i)}$
- Estatísticas de ordem mais usuais: máximo da amostra, $X_{(n)}$, mínimo da amostra, $X_{(1)}$, mediana amostral, $M_e = X_{((n+1)/2)}$ se n ímpar; $M_e = [X_{(n/2)} + X_{(n/2+1)}]/2$ se n par, amplitude amostral, $R = X_{(n)} - X_{(1)}$.
- Vamos restringir-nos ao caso contínuo
- $(Y_1, \ldots, Y_n) \equiv (X_{(1)}, \ldots, X_{(n)})$ para simplificar notação.

Teorema 1.7 *As estatísticas de ordem têm densidade de probabilidade conjunta dada por*

$$g(y_1, y_2, \ldots, y_n) = n! \prod_{i=1}^{n} f(y_i), \quad se\ y_1 < y_2 < \ldots < y_n\ .$$

Se $u < v$, a densidade conjunta de (Y_u, Y_v) é

$$g_{u,v}(y,z) = \frac{n!}{(u-1)!(v-u-1)!(n-v)!} \times [F(y)]^{u-1}\ [F(z)-F(y)]^{v-u-1}\ [1-F(z)]^{n-v} f(y) f(z)\ , \quad se\ y < z.$$

A densidade e função distribuição de Y_v:

$$g_v(y) = \frac{n!}{(v-1)!(n-v)!} [F(y)]^{v-1} [1-F(y)]^{n-v} f(y)$$

$$G_v(y) = \sum_{j=v}^{n} \binom{n}{j} [F(y)]^j [1-F(y)]^{n-j}\ .$$

■

Teorema 1.8 Casos particulares importantes—o máximo e o mínimo:

$$\begin{aligned} G_1(y) &= 1 - [1 - F(y)]^n \\ G_n(y) &= [F(y)]^n \\ g_{1,n}(y,z) &= n(n-1)[F(z) - F(y)]^{n-2} f(y) f(z) \ , \quad y < z \ . \end{aligned}$$

■

Exemplo 1.7 *Se* $X \sim Pa(c, \theta)$, *i.e., com* $\theta > 0$ *e* $c > 0$,

$$f(x) = \frac{\theta}{c} \left(\frac{c}{x}\right)^{\theta+1} \ , \quad x > c \ ,$$

então $F(x) = 1 - (c/x)^\theta$, $x > c$. *Logo, para* $y, z > c$

$$\begin{aligned} g_1(y) &= n \frac{\theta}{y} \left(\frac{c}{y}\right)^{\theta n} \\ g_n(z) &= n \frac{\theta}{y} \left[1 - \left(\frac{c}{z}\right)^\theta\right]^{n-1} \left(\frac{c}{z}\right)^\theta \end{aligned}$$

Exemplo 1.8 *Recordar que se* $X \sim Ex(\lambda)$, *então* $X_{(1)} \sim Ex(n\lambda)$: *como* $F(x) = 1 - \exp(-\lambda x)$, $x > 0$,

$$G_1(x) = 1 - [1 - (1 - \exp(-\lambda x))]^n = 1 - \exp(-\lambda n x) \ , \quad x > 0 \ .$$

■

Teorema 1.9 Distribuição assintótica do quantil de ordem p: *Seja Z_p o quantil de ordem p de uma amostra casual simples de tamanho n, retirada de uma população contínua. Designe-se por ξ_p o quantil de ordem p da população. Sendo f contínua e positiva em ξ_p,*

$$\sqrt{n} f(\xi_p) \frac{Z_p - \xi_p}{\sqrt{p(1-p)}} \overset{a}{\sim} N(0,1) \ .$$

■

Casos particulares importantes: $p = 1/2$, mediana, e primeiro e terceiro quartis, $p = 1/4$ e $p = 3/4$.

Exemplo 1.9 *Se $X \sim N(\mu, \sigma^2)$, então $\xi_{1/2} = \mu$. Assim, $f(\xi_{1/2}) = (2\pi\sigma^2)^{-1/2}$. Logo,*

$$\sqrt{\frac{2n}{\pi\sigma^2}} (Z_{1/2} - \mu) \overset{a}{\sim} N(0,1) \ .$$

■

1.7 Função de distribuição empírica

- Forma de representar graficamente a amostra observada $(x_1, \ldots, x_n)$: Função de distribuição empírica da amostra observada

$$\hat{F}_n(x) = \frac{1}{n}\#\{i : x_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$$

- À semelhança de outras quantidades empíricas, define-se a mesma quantidade para a amostra casual de tamanho n proveniente de uma população X

$$F_n(x) = \frac{1}{n}\#\{i : X_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i), \qquad x \in \mathbb{R}$$

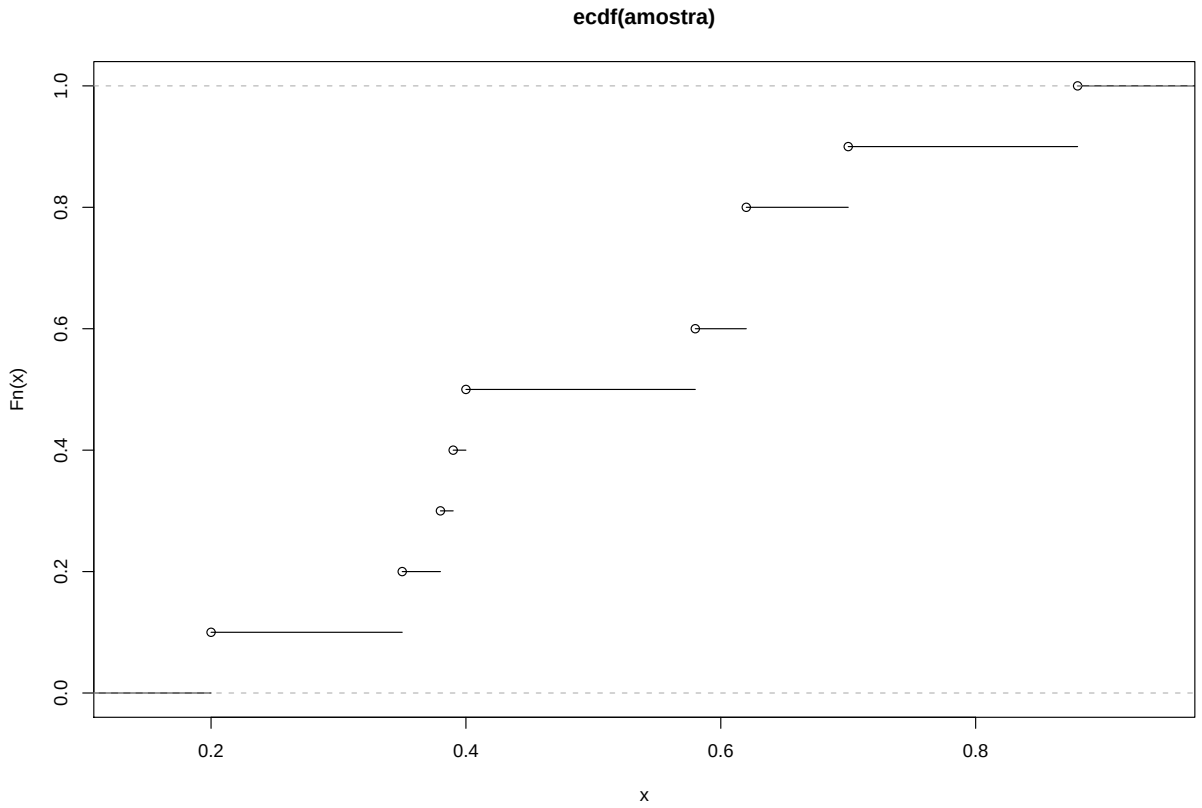
- Se X for contínua,

$$F_n(x) = \begin{cases} 0 & \text{se } x < X_{(1)} \\ \frac{i}{N} & \text{se } X_{(i)} \leq x < X_{(i+1)},\ i = 1, \ldots, n-1 \\ 1 & \text{se } x \geq X_{(n)}\ . \end{cases}$$

- $F_n(x)$ é, para cada x, uma estatística; $\hat{F}_n(x)$ é o correspondente valor observado

- Em R: comando ecdf

```
> amostra <- c(0.7,0.62,0.62,0.88,0.58,0.2,0.35,0.39,0.38,0.40)
> plot(ecdf(amostra))
> ecdf(amostra)(0.62)
[1] 0.8
```



Teorema 1.10 *Para cada $x \in \mathbb{R}$ tem-se que $nF_n(x) \sim B(n, F(x))$, ou seja,*

$$P\left(F_n(x) = i/n\right) = \binom{n}{i}[F(x)]^i[1-F(x)]^{n-i},\ i = 0, \ldots, n$$

logo

$$E[F_n(x)] = F(x)$$
$$Var[F_n(x)] = \frac{F(x)[1-F(x)]}{n}\ .$$

■

Teorema 1.11 *Pela Lei Forte dos Grandes Números, para cada $x \in \mathbb{R}$,*

$$F_n(x) \overset{q.c.}{\to} F(x)\ .$$

Pelo Teorema do Limite Central, para cada $x \in \mathbb{R}$,

$$\sqrt{n}\ \frac{F_n(x) - F(x)}{\sqrt{F(x)[1-F(x)]}} \overset{a}{\sim} N(0,1)\ .$$

■

Exemplo 1.10 *Se* $X \sim U(0,1)$, $F(x) = x$, $x \in (0,1)$, *logo,*

$$nF_n(x) \sim B(n,x)$$

$$P(F_n(x) = i/n) = \binom{n}{i} x^i (1-x)^{n-i}$$

para $0 < x < 1$ *e* $i = 0, \ldots, n$.

Se $X \sim Ex(1)$, $F(x) = 1 - e^{-x}$, $x > 0$, *logo*

$$nF_n(x) \sim B(n, 1-e^{-x})$$

$$P(F_n(x) = i/n) = \binom{n}{i} [1-e^{-x}]^i e^{x(n-i)}$$

para $x > 0$ *e* $i = 0, \ldots, n$. ■

1.8 Algumas distribuições por amostragem

1.8.1 População normal

No que se segue, seja $(X_1, \ldots, X_n)$ uma amostra casual de dimensão n proveniente de uma população $\mathrm{N}(\mu, \sigma^2)$.

Distribuição da média amostral, $\bar{X}$

- $\bar{X}$ é combinação linear de v.a. normais independentes, logo tem distribuição normal
- Sabemos que $E[\bar{X}] = \mu$ e $\mathrm{Var}(\bar{X}) = \sigma^2/n$
- logo

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

Exemplo 1.11 *Suponha-se que a duração, em minutos, 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 16 e 18 minutos?*

Com $\mu = 17$, $\sigma^2 = 25$, e $\bar{X}$ representando a média amostral, tem-se que

$$\begin{aligned} P(16 < \bar{X} < 18) &= P\left(\sqrt{n}\,\frac{16-\mu}{\sigma} < \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} < \sqrt{n}\,\frac{18-\mu}{\sigma}\right) \\ &= P\left(-0.2\sqrt{n} < \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} < 0.2\sqrt{n}\right) \\ &= 2\Phi(0.2\sqrt{n}) - 1\ . \end{aligned}$$

Esta probabilidade aumenta com o crescimento da dimensão da amostra e tende para 1 quando $n \to \infty$. O que acontece à medida que $n \to \infty$ com $P(14 < \bar{X} < 16)$? ■

Distribuição da variância amostral corrigida, S'^2

- Claramente, $\sum_{i=1}^{n}(X_i-\mu)^2/\sigma^2 \sim \chi^2(n)$
- Além disso,

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2$$

- Logo,

$$\sum_{i=1}^{n}(X_i-\mu)^2/\sigma^2 = (n-1)S'^2/\sigma^2 + n(\bar{X}-\mu)^2/\sigma^2$$

- Acabámos de ver que $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, logo $n(\bar{X}-\mu)^2/\sigma^2 \sim \chi^2(1)$
- Falta mostrar que no contexto de uma população normal $\bar{X}$ e S'^2 são independentes — ver Murteira (1990) — para concluir (como?) que

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

Exemplo 1.12 *Considere-se uma população normal da qual se extraiu uma amostra de dimensão 25. Supondo que se procura calcular a probabilidade de o quociente entre a variância amostral corrigida e a variância da população se situar entre 0.79 e 1.18, obtém-se*

$$\begin{aligned} P\left(0.79 < \frac{S'^2}{\sigma^2} < 1.18\right) &= P\left(18.96 < \frac{(n-1)S'^2}{\sigma^2} < 28.32\right) \\ &= \texttt{pchisq}(28.32, 24) - \texttt{pchisq}(18.96, 24) \\ &= 0.5073 \end{aligned}$$

Rácio de "Student"

- Quando a variância da população é desconhecida, torna-se problemático usar

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

- Nesta situação, usa-se o rácio de "Student":

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

- Sabemos que $\bar{X}$ e S'^2 são independentes; note-se que

$$\frac{\bar{X}-\mu}{S'/\sqrt{n}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S'^2}{\sigma^2}\dfrac{1}{n-1}}} = \frac{U}{\sqrt{V/(n-1)}}$$

  onde $U \sim \mathrm{N}(0,1)$ independente de $V \sim \chi^2(n-1)$. Segue-se de imediato o resultado.

- Note-se que $t(n) \to \mathrm{N}(0,1)$ quando $n \to \infty$: para grandes amostras, conhecer S'^2 é muito próximo de conhecer σ^2...

1.8.2 Duas populações normais

- $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$
- Duas amostras casuais, independentes entre si, de tamanhos m e n respectivamente: $(X_{11}, \ldots, X_{1m})$ e $(X_{21}, \ldots, X_{2n})$

Diferença entre duas médias

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

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

- 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 $\mu_1 - \mu_2$: se $\sigma_1^2 = \sigma_2^2 = \sigma^2$, então

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

já que neste caso

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

e

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

são independentes e $T = U/\sqrt{V/(m+n-2)}$.

- Quando as variâncias das populações são desconhecidas e diferentes, as inferências sobre $\mu_1 - \mu_2$ tornam-se mais complexas.
  - grandes amostras possibilitam substituir as variâncias populacionais pelas amostrais, estando garantida a distribuição normal assintoticamente
  - amostras pequenas: aproximação de Welch

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1'^2}{m} + \frac{S_2'^2}{n}}} \overset{a}{\sim} t(\nu)$$

sendo ν dado pelo maior inteiro que não excede

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

Relação entre duas variâncias

- Sendo as duas amostras independentes, as v.a.

$$U = \frac{(m-1)S_1'^2}{\sigma_1^2} \sim \chi^2(m-1)$$

$$V = \frac{(n-1)S_2'^2}{\sigma_2^2} \sim \chi^2(n-1)$$

são independentes, logo

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

- Em particular, quando $\sigma_1^2 = \sigma_2^2$, tem-se

$$\frac{S_1'^2}{S_2'^2} \sim F(m-1, n-1)$$

Exemplo 1.13 *Suponha-se que a distribuição do QI segue, em dois países, uma distribuição normal. Admita-se 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 de o quociente entre as variâncias corrigidas das duas amostras, S'^2_A/S'^2_B, ser superior a 3.77? A resposta é dada calculando*

$$P\left(\frac{S'^2_A}{S'^2_B} > 3.77\right) = 1 - \texttt{pf}(3.77, 15, 9) = 0.02499 \ .$$

■

1.8.3 População Bernoulli

- População é constituída por elementos de dois tipos: os que possuem e os que não possuem determinado atributo
- No que se segue, seja $(X_1, \ldots, X_n)$ uma amostra casual de dimensão n proveniente de uma população $\mathrm{B}(1, \theta)$.
- Interessa geralmente estabelecer a distribuição por amostragem de duas estatísticas: $T = \sum_{i=1}^{n} X_i$ e $\bar{X} = T/n$ — o número de indivíduos na amostra casual que apresenta o atributo, e a proporção de indivíduos na amostra casual que apresenta o atributo.
- Claramente, $T \sim \mathrm{B}(n, \theta)$, logo

$$P(T = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}, \ t = 0, \ldots, n$$

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

- Aproximações para grandes amostras: Teorema de De Moivre-Laplace e Lei dos Acontecimentos Raros
- De Moivre-Laplace:

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

- Regra empírica (há outras): usar quando $n > 20$, $n\theta \geq 5$ e $n\theta(1-\theta) \geq 5$, $0.1 < \theta < 0.9$, juntamente com a correcção de continuidade: com $a < b$, $a, b = 0, 1, \ldots, n$

$$P(a \leq T \leq b) \approx \Phi\left(\frac{b + 1/2 - n\theta}{\sqrt{n\theta(1-\theta)}}\right) - \Phi\left(\frac{a - 1/2 - n\theta}{\sqrt{n\theta(1-\theta)}}\right)$$

- O coeficiente de assimetria de uma $\mathrm{B}(1,\theta)$ é $\gamma_1 = (1-2\theta)/\sqrt{\theta(1-\theta)}$ logo quanto mais afastado de $1/2$ estiver θ maior deve ser n

- Lei dos acontecimentos raros:

$$T \overset{a}{\sim} \mathrm{Po}(n\theta)$$

- Regra empírica: usar para $n > 20$ quando $\theta \notin (0.1, 0.9)$ e $n\theta < 5$
- Aproximações úteis do ponto de vista analítico; do ponto de vista do cálculo de probabilidades são hoje em dia desnecessárias.

Exemplo 1.14 *Admita-se 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?* ■

Designando por $\bar{X}$ a proporção de maus clientes na amostra casual, a resposta a cada uma das alíneas é dada por $P(\bar{X} \geq 0.1)$

(a) Pequena amostra

$$P(\bar{X} \geq 0.1) = P(T \geq 10 \times 0.1) = 1 - P(T = 0) = 1 - (1 - 0.05)^{10} = 0.4013$$

(b) $n = 50 > 20$, $\theta = 0.05 < 0.1$, $n\theta = 2.5 < 5$: usar aproximação pela Poisson

$$P(\bar{X} \geq 0.1) = P(T \geq 5) = 1 - P(T \geq 4) \approx 1 - \texttt{ppois(4,50*0.05)} = 0.1088$$

Valor "exacto": $1 - \texttt{pbinom(4,50,0.05)} = 0.1036$

- $n = 400 > 20$, $\theta = 0.05 < 0.01$, $n\theta = 20 \geq 5$: usar a aproximação pela normal

  Sem correcção de continuidade:

$$P(\bar{X} \geq 0.1) \approx 1 - \Phi\left[(40 - 20)/\sqrt{400 \times 0.05 \times (1 - 0.05)}\right] = 2.23 \times 10^{-6}$$

  Com correcção de continuidade:

$$P(\bar{X} \geq 0.1) \approx 1 - \Phi\left[(40 - 1/2 - 20)/\sqrt{400 \times 0.05 \times (1 - 0.05)}\right] = 3.84 \times 10^{-6}$$

  Aproximação pela Poisson:

$$P(\bar{X} \geq 0.1) = 1 - \texttt{ppois(39,400*0.05)} = 5.32 \times 10^{-5}$$

  Valor "exacto":

$$P(\bar{X} \geq 0.1) = 1 - \texttt{pbinom(39,400,0.05)} = 3.15 \times 10^{-5}$$

1.8.4 Duas populações Bernoulli

- Duas populações Bernoulli com parâmetros θ_1 e θ_2.
- Quer-se comparar as duas proporções populacionais 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)
- $\theta_1 - \theta_2$ será em geral desconhecido; pretende-se estabelecer inferências acerca dessa quantidade através da estatística $\bar{X}_1 - \bar{X}_2$, a diferença entre as proporções observáveis nas amostras casuais extraídas de cada população
- Duas amostras casuais independentes uma da outra:
  - $(X_{11}, \ldots, X_{1m}) \Rightarrow \bar{X}_1 = \sum_{i=1}^{m} X_{1i}/m$
  - $(X_{21}, \ldots, X_{2n}) \Rightarrow \bar{X}_2 = \sum_{j=1}^{n} X_{2j}/n$
- Distribuição por amostragem de $\bar{X}_1 - \bar{X}_2$?

- Não existem resultados exactos simples
- Distribuição assintótica: pelo Teorema de De Moivre-Laplace, tem-se que

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

logo, por independência,

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

Exemplo 1.14 *(Continuação) 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?* ■

Com $\theta_1 = 0.05$, $\theta_2 = 0.06$, $m = 400$, $n = 500$,

$$P(\bar{X}_1 - \bar{X}_2 > 0) = P\left(\frac{\bar{X}_1 - \bar{X}_2 - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{m} + \frac{\theta_2(1-\theta_2)}{n}}} > \frac{0 - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{m} + \frac{\theta_2(1-\theta_2)}{n}}}\right)$$
$$\approx 1 - \Phi(0.66) \approx 0.2546$$

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, mesmo assim a probabilidade da média amostral de Lisboa ser superior da média amostral do Porto é aproximadamente 25%.

1.8.5 Outras populações: caso da gama

- Se $X \sim \mathrm{G}(\alpha, \lambda)$, então (para $\alpha, \lambda > 0$)

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x) \ , \quad x > 0$$

- Se $\alpha \in \mathbb{N}$, é também conhecida por distribuição de Erlang; $\alpha = 1$ cai-se no caso da $\mathrm{Ex}(\lambda)$; $\mathrm{G}(n/2, 1/2) = \chi^2(n)$

- α é parâmetro de forma; λ é a taxa. Muitas vezes esta distribuição é parametrizada em termos de $\beta = 1/\lambda$ — parâmetro de escala:

```
dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
```

- Se $X_1 \sim \mathrm{G}(\alpha_1, \lambda)$ for independente de $X_2 \sim \mathrm{G}(\alpha_2, \lambda)$ então

$$X_1 + X_2 \sim \mathrm{G}(\alpha_1 + \alpha_2, \lambda)$$

- Se $c > 0$ e $X \sim \mathrm{G}(\alpha, \lambda)$, então

$$cX \sim \mathrm{G}(\alpha, \lambda/c)$$

- Se $X \sim \mathrm{G}(\alpha, \lambda)$, então

$$2\lambda X \sim \mathrm{G}(\alpha, 1/2) = \chi^2(2\alpha)$$

- Seja $X_1, \ldots, X_n$ uma amostra casual proveniente de uma população $\mathrm{G}(\alpha, \lambda)$.
- Então

$$\sum_{i=1}^{n} X_1 \sim \mathrm{G}(n\alpha, \lambda) \Leftrightarrow \bar{X} \sim \mathrm{G}(n\alpha, n\lambda) \Leftrightarrow 2n\lambda\bar{X} \sim \chi^2(2n\alpha)$$

Fim do Capítulo 1