U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Capítulo 6

Amostragem Aleatória e Distribuições por Amostragem

AMG, JFO (v8 – 2017)
adaptado de: *Estatística*,
Rui Campos Guimarães,
José A. Sarsfield Cabral

Slide 6.-1

Conteúdo

Slide 6.0

Resultados de aprendizagem

- Reconhecer uma amostragem aleatória
- Distinguir uma amostragem aleatória de uma amostragem aleatória simples
- Obter uma distribuição por amostragem quer por enumeração de “todos os resultados possíveis” quer por “repetição de amostras”
- Calcular o valor esperado e a variância da média amostral
- Enunciar o Teorema do Limite Central e a sua versão relaxada
- Conhecer a regra prática de aplicação do Teorema do Limite Central
- Calcular probabilidades com base em distribuições obtidas pela aplicação do Teorema do Limite Central

Slide 6.1

6.1 Introdução

Amostragem

- Capítulo final do percurso dedutivo (população $\longrightarrow$ amostra)

- Estudo de como as estatísticas variam de amostra para amostra

⇒ Caracterizar as distribuições de certas estatísticas amostrais

Amostragem probabilística — necessária para se poder caracterizar as distribuições de certas estatísticas

⇒ estudaremos apenas a *amostragem aleatória*

Nota: ver Capítulo 1 para mais informações sobre outros tipos de amostragem

Slide 6.2

6.2 Amostragem aleatória

Amostra aleatória
Todos os elementos da população têm igual probabilidade de serem incluídos na amostra

$$\forall y : p_{Y_1}(y_1) = p_{Y_2}(y_2) = \ldots = p_{Y_N}(y_n) = p_Y(y)$$

Dimensão da amostra (N) — número de observações da população, ou número de réplicas da variável aleatória Y

Resultado de cada réplica — é também uma variável aleatória ($Y_1, Y_2, \ldots, Y_N$) com *distribuição igual* à da variável original Y

Amostra aleatória simples
Uma amostra aleatória diz-se *simples* quando as variáveis Y_1, Y_2, ..., Y_N forem também independentes

$$\forall y_1, y_2, \ldots, y_N : p_{Y_1 Y_2 \ldots Y_N}(y_1, y_2, \ldots, y_N) = p_Y(y_1) \cdot p_Y(y_1) \cdot \ldots \cdot p_Y(y_N)$$

Slide 6.3

Recolha de amostras aleatórias com e sem reposição

- Com reposição – $X_1, X_2, \ldots, X_n$ são v.a. independentes

$$\Downarrow$$
$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \cdots p_{X_n}(x_n)$$

- Sem reposição (ou em bloco) – $X_1, X_2, \ldots, X_n$ não são v.a. independentes

$$\Downarrow$$
$$\begin{aligned} p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) &= p_{X_n|X_1,X_2,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1}) \cdot \\ &\quad \cdot \ldots \cdot p_{X_2|X_1}(x_2|x_1) \cdot p_{X_1}(x_1) \end{aligned}$$

Amostragem aleatória simples ($X_1, X_2, \ldots, X_n$ são v.a. independentes)

- Quando a amostragem aleatória é realizada com reposição
- Quando a população muito maior que a amostra ($M >> N$) a dependência entre as variáveis por a amostragem ser feita sem reposição tende a desaparecer
- Em populações infinitas as amostragens aleatórias são sempre simples

Slide 6.4

6.3 Distribuições por amostragem

- Para uma dada população, e uma dada variável aleatória sobre ela definida, os parâmetros da distribuição correspondente (valor esperado, variância, ...) são *fixos*

- As estatísticas (média amostral, variância amostral, etc.) *variam* de amostra para amostra

$\Longrightarrow$ é possível, e desejável, estabelecer funções de probabilidade ou de densidade de probabilidade para o modo como as estatísticas variam

Objectivo
estabelecimento de *distribuições de estatísticas por amostragem*

Slide 6.5

Exemplo

Considere-se uma população com 4 elementos, que correspondem aos seguintes valores da variável aleatória $Y : \{2,4,6,6\}$, da qual se retiram, em bloco, 2 elementos. Determine a distribuição de $\bar{Y}$ (a média amostral de Y).

y	$p_Y(y)$
2	1/4
4	1/4
6	2/4

$$\mu_Y = \sum y_i \cdot p_Y(y_i) = 2 \times \tfrac{1}{4} + 4 \times \tfrac{1}{4} + 6 \times \tfrac{2}{4} = 4.5$$
$$\sigma_Y^2 = \sum (y_i - \bar{Y})^2 \cdot p_Y(y_i) = 2.5^2 \times \tfrac{1}{4} + 0.5^2 \times \tfrac{1}{4} + 1.5^2 \times \tfrac{2}{4} = 2.75$$

Conjunto de todas as amostras de dimensão 2 (sem reposição)

Amostra	$\bar{y}$	Prob. de ocorrência
2,4	3	1/4 x 1/3 = 1/12
2,6	4	1/4 x 2/3 = 2/12
4,2	3	1/4 x 1/3 = 1/12
4,6	5	1/4 x 2/3 = 2/12
6,2	4	2/4 x 1/3 = 2/12
6,4	5	2/4 x 1/3 = 2/12
6,6	6	2/4 x 1/3 = 2/12

$\bar{y}$	$p_{\bar{Y}}(\bar{y})$
3	1/6
4	2/6
5	2/6
6	1/6

$$\mu_{\bar{Y}} = \sum \bar{y}_i \cdot p_{\bar{Y}}(\bar{y}_i) = 3 \times \tfrac{1}{6} + 4 \times \tfrac{2}{6} + 5 \times \tfrac{2}{6} + 6 \times \tfrac{1}{6} = 4.5$$
$$\sigma_{\bar{Y}}^2 = \sum (\bar{y}_i - \bar{Y})^2 \cdot p_{\bar{Y}}(\bar{y}_i) = 1.5^2 \times \tfrac{1}{6} + 0.5^2 \times \tfrac{2}{6} + +0.5^2 \times \tfrac{2}{6} + 1.5^2 \times \tfrac{1}{6} = 0.917$$

Slide 6.6

- No exemplo anterior foi relativamente fácil obter a distribuição completa da média amostral devido ao reduzido tamanho da população, que permitiu que se gerassem todas as amostras possíveis
- E quando tal não for possível? Por exemplo quando as populações são muito grandes ...
- Recorre-se a métodos que permitem obter as distribuições por amostragem de uma forma geralmente aproximada ou parcial:

⇒ A via teórica permite-nos obter, sob certas condições, as distribuições de algumas estatísticas bem comportadas,

⇒ Quando a via teórica é inviável (populações não-normais, amostras de pequena dimensão ou estatísticas são mais complexas) recorre-se à geração artificial de amostras

(http://www.socr.ucla.edu/Applets.dir/SamplingDistributionApplet.html)

(http://onlinestatbook.com/chapter7/sampling_distributions.html)

Slide 6.7

6.3.1 Distribuição da média amostral

Média amostral: $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ (transformação linear)

- Valor esperado da média amostral: $\mu_{\bar{X}} = \mu_X$
- Variância da média amostral
  - amostragem aleatória simples: $\sigma_{\bar{X}}^2 = \frac{1}{N} \cdot \sigma_X^2$
  - população finita (M) e sem reposição: $\sigma_{\bar{X}}^2 = \left(\frac{M-N}{M-1}\right) \cdot \frac{1}{N} \cdot \sigma_X^2$

    $\left(\frac{M-N}{M-1}\right)$: factor de correcção para populações finitas ($N \le M$)

    $M \to \infty$**:** factor de redução $\to 1$ (se amostra finita)

    $N = 1$**:** factor de redução $= 1$ (reposição não faz diferença)

    $N = M$**:** factor de redução $= 0$ (amostra coincide com a pop.)

Demonstrações:

$$\mu_{\bar{X}} = E(\bar{X}) = E\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{1}{N}\sum_{i=1}^{N} E(X) = \frac{E(X)}{N} \cdot \sum_{i=1}^{N} 1 = E(X) = \mu_X$$

$$\sigma_{\bar{X}}^2 = Var(\bar{X}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) = \frac{1}{N^2}\sum_{i=1}^{N} Var(X) = \frac{N \cdot Var(X)}{N^2} = \frac{1}{N} \cdot \sigma_X^2$$

Slide 6.8

Forma da distribuição de $\bar{X}$ quando $X \rightsquigarrow N(\mu_X, \sigma_X^2)$

- X segue uma distribuição normal
$$X \rightsquigarrow N(\mu_X, \sigma_X^2)$$
- $\bar{X}$ é uma *combinação linear de v.a. normais e independentes*
$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$
- Pelo que $\bar{X}$ segue uma distribuição normal
$$\bar{X} \rightsquigarrow N\left(\mu_X, \frac{1}{N}\cdot\sigma_X^2\right) \qquad \text{(para amostras aleatórias simples)}$$

Notas

- o facto de X ser normalmente distribuído pressupõe que a população é infinita, ou seja que se trata de uma amostragem aleatória simples
- trata-se de um resultado limitado já que se refere apenas a uma estatística simples e obriga à normalidade da população

Slide 6.9

6.4 Teorema do limite central

6.4.1 Enunciado do teorema

- Sejam $X_1, \ldots, X_N$ um conjunto de v.a. *independentes* com a *mesma distribuição*, que se admite ter variância finita
$$E(X_i) = \mu_X \qquad Var(X_i) = \sigma_X^2$$
- Qualquer que seja a forma da distribuição destas variáveis, se o valor de N for suficientemente grande, então a variável soma
$$S = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{N} X_i$$
segue aproximadamente uma distribuição Normal, com parâmetros:
$$\mu_S = N\cdot\mu_X \qquad \sigma_S^2 = N\cdot\sigma_X^2$$

$$\Rightarrow \qquad S = X_1 + X_2 + \ldots + X_N \quad \overset{N\to\infty}{\longrightarrow} \quad N\left(N\cdot\mu_X, N\cdot\sigma_X^2\right)$$

Nota admitir variância finita é uma condição pouco restritiva, já que quase todas as distribuições com interesse prático têm variância finita

Slide 6.10

- A média amostral $\bar{X}$ calculada com base numa amostra aleatória simples tende para uma distribuição Normal

- O T.L.C. não impõe nenhuma condição relativamente à forma da distribuição original

- Quando é que N é suficientemente grande?
  - depende da forma da distribuição original
  - resposta pode ser obtida por via experimental
    (geração de amostras aleatórias pela técnica de Monte Carlo)

  $\Rightarrow$ Regra prática:
  $N \geq 10$ quando a distribuição original for simétrica
  $N \geq 50$ quando a distribuição original for muito assimétrica

Slide 6.11

O T.L.C. atrás apresentado obriga a que as variáveis X_i:

- sejam *independentes* e tenham *distribuições idênticas*

São, no entanto, *condições suficientes mas não necessárias*, pelo que *podem ser relaxadas*:

- As variáveis X_i podem ter distribuições distintas, desde que a contribuição da variância de cada uma delas para a variância de S seja pequena
- As variáveis X_i podem não ser independentes, desde que as correlações entre elas sejam fracas

$$S = X_1 + X_2 + \ldots + X_N \quad \overset{N\to\infty}{\longrightarrow} \quad N\left(\sum_{i=0}^{N} \mu_{X_i}, \sum_{i=0}^{N} \sigma^2_{X_i}\right)$$

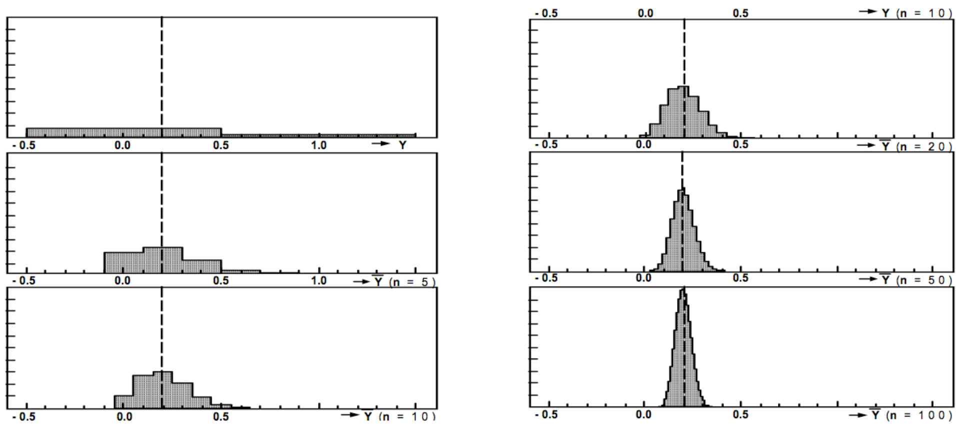
- Notar que são condições semelhantes às que estão na origem de uma distribuição Normal (ver slide 7.9)
- Para N grande, muitas estatísticas amostrais (para além da soma e da média amostral) têm uma distribuição muito próxima da distribuição Normal (p.e. a variância amostral)

Slide 6.12

Exemplo

A função de probabilidade de Y, $p(y)$, está definida na tabela ao lado. As figuras abaixo são as distribuições da média amostral ($\bar{Y}$), para amostras aleatórias simples de tamanho $N = 5$, 10, 20, 50 e 100

y	$p(y)$
0	0.80
1	0.20



(http://www.causeweb.org/repository/statjava/CLTApplet.html) Slide 6.13

6.4.2 Justificação da normalidade de certas variáveis

O T.L.C. permite ainda justificar *a priori* que certas variáveis sigam distribuições Normais:

> sempre que uma variável possa ser considerada, pelo menos aproximadamente, como sendo a soma de um conjunto significativo de variáveis independentes e identicamente distribuídas

Exemplos

- Para um sector de uma floresta no qual as árvores tenham sido plantadas há 50 anos, considere-se a variável diâmetro do tronco das árvores
  - resultado da soma de 50 crescimentos anuais
  - crescimentos anuais independentes e identicamente distribuídos
- E se há árvores de tipos diferentes, logo com crescimentos diferentes?

- Num aparelho de medida que tenha um número elevado de fontes potenciais de erro, razoavelmente independentes e sem que nenhuma delas seja dominante

Slide 6.14

6.4.3 Justificação de alguns resultados apresentados anteriormente

Justificação de alguns resultados apresentados anteriormente

- *Aproximação da Distribuição Binomial pela Distribuição Normal*

  Z_i são variáveis que tomam o valor 1 quando o resultado da experiência de Bernoulli for "sucesso" e o valor 0 no caso contrário.

  $Y \rightsquigarrow B(N,p)$ pode ser interpretada como $Y = \sum_{i=1}^{N} Z_i$

  Como Z_i são igualmente distribuídas (com variância finita) e independentes, do T.L.C. resulta que $Y \stackrel{N\to\infty}{\longrightarrow} N(N\cdot p,\ N\cdot p\cdot q)$

- *Aproximação da Distribuição Hipergeométrica pela Distribuição Normal*

  Justificação semelhante à anterior, relaxando a condição de independência entre as variáveis Z_i (dependência fraca, $M >> N$)

- *Aproximação da Distribuição χ^2_{GL} pela Distribuição Normal*

  A v.a. Qui-quadrado é dada pela soma $\chi^2_{GL} = \sum_{i=1}^{GL} Z_i^2$, em que Z_i são variáveis independentes, todas com distribuições $N(0,1)$. Para valores grandes de N resulta que χ^2_{GL} tende para a distribuição Normal

Slide 6.15

6.5 Exercícios

Exercícios

1. Um fabricante de produtos alimentares produz sacos de um dado tipo de arroz com a indicação de 1 Kg na embalagem. Sabe-se que o peso do conteúdo de um saco segue uma distribuição $U(950g,\ 1050g)$. Admita que o peso dos sacos de arroz são independentes entre si. Calcule ou indique como poderia calcular:

   a) A probabilidade de um lote de 500 sacos pesar mais de 501 Kg.

   b) A probabilidade de o peso de dois sacos seleccionados aleatoriamente diferir em mais de 20 gramas.

Resolução:

P: Peso de um saco de arroz [g] $\qquad P \rightsquigarrow U(950,\ 1050)$

L: Peso de um lote de 500 sacos de arroz [g] $\qquad L = \sum_{i=1}^{500} P_i$

Sabemos que:

$\mu_P = \frac{950+1050}{2} = 1000$ e $\sigma_P^2 = \frac{(1050-950)^2}{12} = 833.33$ (distr. uniforme)

$\mu_L = 500\cdot\mu_P = 500000$ e $\sigma_L^2 = 500\cdot\sigma_P^2 = 416666.7 = 645.5^2$ (transf. linear)

a) Como a amostra é de grande dimensão (500), os P_i são independentes e identicamente distribuídos, o T.L.C. diz-nos que $L \rightsquigarrow N(500000,\ 645.5^2)$

$$P(L > 501000) = P\left(Z > \frac{501000-500000}{645.5}\right) = P(Z > 1.55) = 0.06057$$

Slide 6.16

b) Queremos calcular $P(P_1 > P_2 + 20)$ ou seja $P(P_1 - P_2 > 20)$

   Onde:

   - $P_1, P_2 \rightsquigarrow U(950,\ 1050)$
   - $P_1 - P_2$ segue uma distribuição desconhecida

Teremos então recorrer à geração de uma amostra para estimar $P_1 - P_2$.

Procedimento:

(i) Gerar uma amostra de dimensão N de P_1

(ii) Gerar uma amostra de dimensão N de P_2

(iii) Definir uma amostra de dimensão N tal que $N_i = P_{1i} - P_{2i}$, com $(i = 1, \ldots, N)$

(iv) Contar na amostra definida em (iii) o número de elementos cuja diferença é superior a 20g (seja N_0)

(v) Estimar a probabilidade de acordo com

$$P(P_1 - P_2 > 20) \approx N_0/N$$

Recorrendo a uma folha de cálculo:

i	P_1	P_2	P_1 - P_2	N_0
1	1001	996	5	0
2	1002	1029	-27	0
3	1049	971	78	1
4	1012	994	18	0
...	...	...	...	...
499	1008	1047	-39	0
500	1035	954	81	1

$N = 500 \qquad N_0 = 171$

$$P(P_1 - P_2 > 20) \approx \frac{171}{500} = 0.342$$

Slide 6.17

2. Uma empresa de vendas por catálogo recebe encomendas de duas áreas geográficas: 30% das encomendas provêm do Norte e 70% do Sul.

   A distribuição do valor de cada encomenda varia consoante a região de acordo com as seguintes distribuições (expressas em euros):

$$\text{Norte: } X_N \rightsquigarrow N(10, 2^2) \qquad \text{e} \qquad \text{Sul: } X_S \rightsquigarrow U(5, 13)$$

   a) Sabendo que uma encomenda acabada de chegar tem um valor superior a 12 euros, calcule a probabilidade de tal encomenda provir do Norte.

   b) Numa determinada semana, a empresa recebeu 100 encomendas do Norte e 250 do Sul. Qual a probabilidade de o valor global destas encomendas ultrapassar 3180 euros?

Resolução:

a) $P(N|X>12) = \dfrac{P(N \cap X > 12)}{P(X>12)} = \dfrac{P(X>12|N)P(N)}{P(X>12|N)P(N) + P(X>12|S)P(S)} = 0.352$

em que:

- $P(N) = 0.3 \quad$ e $\quad P(S) = 0.7$
- $P(X>12|N) = P(X_N > 12) = P\left(Z > \dfrac{12-10}{2}\right) = P(Z>1) = 0.1587$
- $P(X>12|S) = P(X_S > 12) = \dfrac{13-12}{13-5} 0.125$

b) $G = \sum_{i=1}^{100} X_{Ni} + \sum_{j=1}^{250} X_{Sj}$ pelo T.L.C. sabemos que; $G \rightsquigarrow N(3250, 1733)$

Nota: trata-se de um amostra de grande dimensões, com v.a. independentes provenientes de distribuições diferentes ⇒ versão relaxada do T.L.C

com:

- $\mu_G = E(G) = 100 \cdot \mu_{X_N} + 250 \cdot \mu_{X_S} = 100 \cdot 10 + 250 \cdot \dfrac{5+13}{2} = 3250$
- $\sigma_G^2 = Var(G) = 100 \cdot \sigma_{X_N}^2 + 250 \cdot \sigma_{X_S}^2 = 100 \cdot 4 + 250 \cdot \dfrac{(13-5)^2}{12} = 1733$

$$P(G > 3180) = P\left(Z > \frac{3180 - 3250}{\sqrt{1733}}\right) = P(Z > -1.68) = 0.9535$$

Slide 6.18

3. Segundo um relatório sobre a situação mundial da pecuária do Departamento de Agricultura dos Estados Unidos, o país com maior consumo per capita de carne de porco é a Dinamarca.

   Em 1994, a quantidade de carne de porco consumida por uma pessoa a residir na Dinamarca tinha um valor médio de 147 kg com um desvio padrão de 62 kg, enquanto a quantidade de carne de porco consumida por um residente nos Estados Unidos tinha um valor médio foi 105 kg com um desvio padrão de 75 kg.

   a) Calcule a probabilidade de a quantidade média de carne de porco consumida pelos membros de uma amostra de aleatória constituída por 55 dinamarqueses em 1994 ultrapassar 155 kg. (Sol.: 16.93%)

b) Considere uma amostra aleatória de 55 dinamarqueses e uma amostra aleatória de 90 americanos. Calcule a probabilidade de a quantidade total de carne de porco consumida pelos 90 americanos ser inferior à quantidade total de carne de porco consumida pelos 55 dinamarqueses. (Sol.: 5.36%)

Slide 6.19

4. Um pequeno hotel tem 8 quartos que aluga a 55 € por noite. Cada quarto tem um custo fixo diário para o hotel de 44 €. Dada a sua localização, o hotel apenas aceita alugar quartos por reserva. Devido a esta estratégia, o hotel usa uma política de overbooking (aceitação de reservas acima da capacidade) para compensar clientes com reserva que acabem por não aparecer, aceitando 12 reservas por noite (nota: a realização de uma reserva não acarreta qualquer custo para quem a realiza). Se um cliente chegar ao hotel com uma reserva e não houver quarto disponível, o hotel paga ao cliente uma “multa” de valor igual ao aluguer de um quarto (55 €), como forma de compensar o incómodo causado. Dados históricos mostram que apenas 60% dos clientes com reserva acabam por aparecer. Considere que todas as noites existem 12 reservas e os clientes com reserva têm comportamentos independentes entre si.

   a) Calcule a probabilidade de, numa noite, não haver quartos suficientes para todos os clientes com reserva que aparecem no hotel. (Sol.: 22.53%)

   b) Calcule o valor esperado e o desvio padrão do lucro diário. (Sol.: $\mu_L = 7.64$€, $\sigma_L = 64.73$€)

   c) Devido à sua localização, o hotel funciona apenas durante aproximadamente 7 meses por ano (200 dias). Calcule a probabilidade de o lucro total ao fim de um ano ser negativo (i.e, o hotel gerar prejuízo ao fim de um ano). (Sol.: 4.76%)

Slide 6.20