Capítulo 6

Geração de amostras recorrendo à técnica de Monte Carlo

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

Slide 0.-1

Conteúdo

Slide 0.0

6.1 Enquadramento

Objectivo no início do capítulo – estudar como as estatísticas variam de amostra para amostra, conhecida a distribuição da população

Abordagens possíveis:

- Gerar todas as amostras possíveis e determinar as distribuições das estatísticas pretendidas a partir da função de probabilidades das amostras

  Inviável quando a população é muito grande ou infinita, pelo que . . .

- Para algumas estatísticas bem comportadas, e sob certas condições, a via teórica deu-nos essas distribuições das estatísticas

  Inviável quando as condições teóricas não se verificam (populações não-normais, amostras de pequena dimensão) ou as estatísticas são mais complexas (não-lineares) pelo que...

- Gerar artificialmente amostras, recorrendo à técnica de Monte-Carlo, e a partir das amostras geradas estudar experimentalmente o comportamento das estatísticas em causa

Geração de amostras – permite calcular valores das estatísticas em causa e fazer suposições quanto à forma da distribuição respectiva

Slide 0.1

Geração de amostras aleatórias – Procedimento experimental

Exemplo: Seja uma população infinita $X \rightsquigarrow EN(2)$. Para amostras de dimensão $N = 10$, pretende-se conhecer a distribuição do coeficiente de assimetria amostral:

$$g_1 = \frac{k_3}{s^3} = \frac{N^2}{(N-1)\times(N-2)} \times \frac{m_3}{s^3}$$

(m_3 é o momento amostral centrado de ordem 3 e s é o desvio padrão amostral)

Procedimento geral

1. Gerar uma amostra aleatória, constituída por N observações, de uma população com uma dada distribuição
2. Com base na amostra gerada, calcular o valor da estatística que se quer estudar e registá-lo (e_i)
3. Repetir K vezes os passos 1 e 2
4. Os valores registados $(e_1, e_2, \ldots, e_K)$ constituem uma amostra da estatística em causa, que nos permite caracterizar por via experimental a distribuição da estatística (e.g. histograma, média, desvio-padrão, etc.).

Slide 0.2

6.2 Geração de amostras aleatórias provenientes de uma população com uma distribuição $U(0,\ 1)$

Problema em aberto
Como gerar amostras aleatórias provenientes de populações com uma dada distribuição?

1. Gerar números aleatórios segundo uma dist. uniforme $U(0,1)$
2. Aplicar uma transformação a estes números para se obterem outros, também aleatórios, mas segundo a distribuição pretendida

Número aleatório: se a sua ocorrência numa sequência de números é imprevisível a partir de quaisquer ocorrências anteriores e se possui a mesma probabilidade de ocorrência ao longo de um intervalo predefinido de valores

Fenómenos verdadeiramente aleatórios: ruído radioactivo, ruído de fundo em componentes electrónicos, etc. (sistemas físicos)

Como gerar um número aleatório?

- recorrendo a tabelas de números aleatórios
- recorrendo a computadores (rotinas *standard* em linguagens de programação)

Slide 0.3

Geração de números aleatórios

Tabelas de números (dígitos) aleatórios:

- 763150 012500 268350 506190 953420 812760 ...
- agrupar dígitos conforme precisão desejada (0.76, 0.31, 0.50, 0.01, ...)
- como os dígitos são equiprováveis e independentes, também o são qualquer agrupamento desses dígitos

Geração de números aleatórios por computador:

- Apenas se conseguem gerar sequências *pseudo-aleatórias*
- Processo iterativo: $n_{i+1} = f(n_i)$
- Sequência repete-se quando surgir um n_j igual a um n_k anterior
- n_0 denomina-se de semente

Desvantagem: Sequência pode-se repetir

Vantagem: É possível duplicar exactamente uma sequência de números aleatórios (importante para a detecção e correcção de erros, verificação e validação de um modelo de simulação)

$\Rightarrow$ Necessário escolher uma função $f()$ adequada

Slide 0.4

Geração de números pseudo-aleatórios por computador
Método mais comum: *Método congruencial multiplicativo*

$$n_{i+1} = (b \times n_i + c) \bmod (m)$$

em que:

- *mod* representa o resto da divisão inteira
- m é o maior número que pode ser gerado (exclusive)
- b e c são parâmetros da fórmula
- ao primeiro valor n_0 dá-se o nome de semente do gerador

Valores concretos para geração de inteiros de 16-bits e 32-bits, respectivamente:

- $f(n_i) = (3993n_i + 1) \bmod (32767)$
- $f(n_i) = (16807n_i + 0) \bmod (2147483647)$

Para gerar um valor entre 0 e 1 basta dividir n_{i+1} por m:

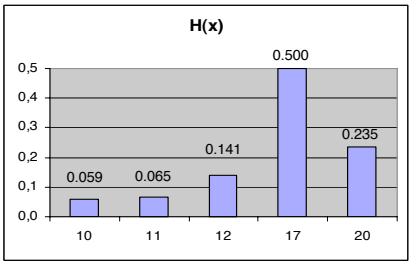
$$R_{i+1} = \frac{n_{i+1}}{m}$$

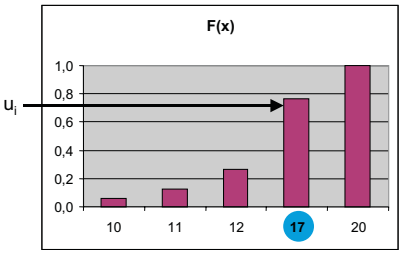
Slide 0.5

6.3 Geração de amostras aleatórias provenientes de uma população qualquer

6.3.1 Geração de amostras aleatórias – População discreta qualquer

Estratégia – Dividir o intervalo [0,1] em sub-intervalos de amplitudes proporcionais às probabilidades de ocorrência dos números que se pretendem gerar





1. Obter a distribuição de frequências acumuladas ($F(x)$) a partir da distribuição de frequências ou histograma ($H(x)$)
2. Gerar um número aleatório u_i entre $[0,1]$
3. Obter o valor x_i directamente da distribuição de frequências acumuladas
4. Repetir K vezes os passos 2 e 3 para se obter uma amostra de dimensão K

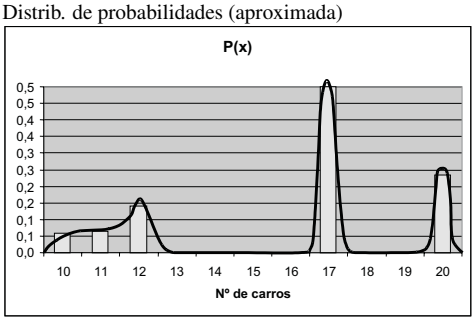
Pressuposto – a sequência de números (pseudo-)aleatórios reflecte uma distribuição de probabilidades constante

Slide 0.6

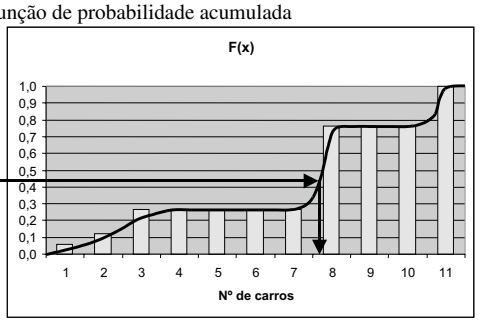
6.3.2 Geração de amostras aleatórias – População contínua qualquer

- Distribuição de probabilidades é dada de forma analítica $f(x)$ e x varia de forma contínua no intervalo dado
- Origem de $f(x)$:
  - obtida de forma aproximada por ajuste a valores experimentais
  - conhecido o comportamento *a priori* do modelo (p.e., X segue uma dist. normal), os parâmetros podem ser ajustados experimentalmente

Distrib. de probabilidades (aproximada)



Função de probabilidade acumulada

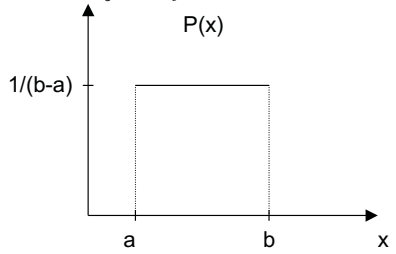


Conhecida a expressão analítica de $F(x)$, a determinação do valor de x para um dado $u_i \in [0,1]$ faz-se a partir da inversa de $F(x)$: $x = F^{-1}(u_i)$

Slide 0.7

6.3.3 Amostragem de distribuições contínuas típicas

1. Distribuição uniforme



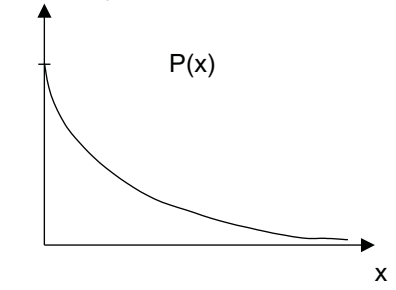
$$f(x) = \frac{1}{b-a},\ a \le x \le b$$

$$F(x) = \int_a^x \frac{1}{b-a}\,dx = \frac{1}{b-a}\cdot(x-a)$$
$$\Downarrow$$
$$u = F(x) = \frac{1}{b-a}\cdot(x-a)$$
$$\Downarrow$$
$$x = F^{-1}(u) = (b-a)\cdot u + a$$
$$\Downarrow$$
$$\boxed{X = (b-a)\cdot U + a}$$

(com *u* uniformemente distribuído sobre o intervalo [0,1])

Slide 0.8

2. Distribuição exponencial negativa



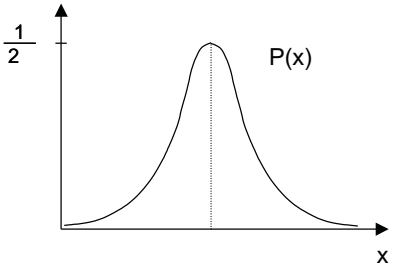
$$f(x) = \lambda\cdot e^{-\lambda x},\ x > 0$$

$$F(x) = \int_0^x \lambda\cdot e^{-\lambda x}dx = 1 - e^{-\lambda x}$$
$$\Downarrow$$
$$u = F(x) = 1 - e^{-\lambda x}$$
$$\Downarrow$$
$$x = F^{-1}(x) = -\frac{\ln(1-u)}{\lambda}$$
$$\Downarrow$$
$$\boxed{X = -\frac{\ln(1-U)}{\lambda}}$$

(com *u* uniformemente distribuído sobre o intervalo [0,1])

Slide 0.9

3. Distribuição normal padronizada $N(0,\ 1^2)$



$$f(x) = \frac{1}{2\pi}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F(x) = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

não existe forma de calcular analíticamente este integral

$$\Downarrow$$

Alternativa: *método de Box-Muller*

$$Z_1 = \sqrt{-2\ln U_1}\cdot\cos(2\pi U_2)$$
$$Z_2 = \sqrt{-2\ln U_1}\cdot\sin(2\pi U_2)$$

com

$$U_1, U_2 \rightsquigarrow U[0,1]$$
$$Z_1, Z_2 \rightsquigarrow N(0,1) \text{ e independentes}$$

3.a Distribuição normal genérica $N(\mu,\ \sigma^2)$

Para obter uma variável $X \rightsquigarrow N(\mu,\sigma^2)$ fazer: $\boxed{X = \mu + \sigma\cdot Z}$

Slide 0.10

6.4 Exemplo

Exemplo – Manutenção de uma linha de produção

Uma fábrica possui uma linha de produção que labora 24 horas por dia durante 360 dias por ano. Esta linha de produção gera um valor acrescentado de 500 euros por hora. Por vezes a linha apresenta avarias necessitando de reparação. Históricos detalhados permitiram concluir que o número de horas de funcionamento entre avarias apresenta a seguinte distribuição de probabilidades:

N^0 de horas	Probabilidade	N^0 de horas	Probabilidade
10	0.05	560	0.05
20	0.05	670	0.05
40	0.05	790	0.05
70	0.05	920	0.05
110	0.05	1060	0.05
160	0.05	1210	0.05
220	0.05	1370	0.05
290	0.05	1540	0.05
370	0.05	1720	0.05
460	0.05	1920	0.05

O tempo de funcionamento médio (entre avarias) é pois de 675.5 horas.

Slide 0.11

Após uma avaria o tempo necessário para a reparação é também variável. Se o tempo de reparação ultrapassar os 3 dias então uma unidade de substituição pode ser obtida, ao abrigo do contrato de garantia, para repor a fábrica em funcionamento. Como esta substituição leva um dia, no pior dos casos o tempo de reparação será de 4 dias. A distribuição de probabilidade para o tempo de reparação é dada pela seguinte tabela:

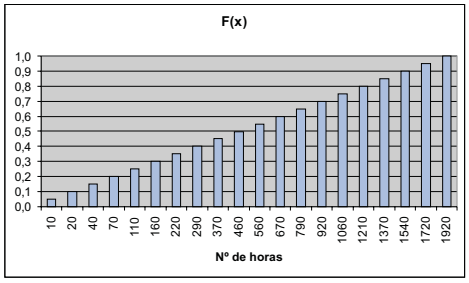
Tempo de reparação (horas)	Probabilidade
24	0.10
48	0.40
72	0.40
96	0.10

O tempo de reparação médio é de 60 horas. Está-se a assumir que o tempo de reparação assumirá sempre um valor múltiplo de 24 horas.

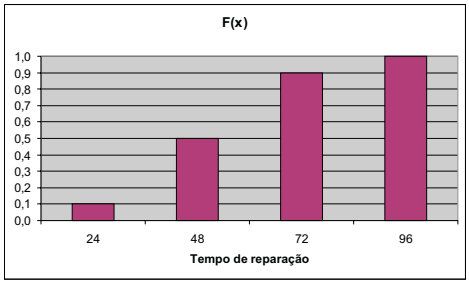
Slide 0.12

Histogramas de probabilidade acumulada

Tempo de funcionamento entre avarias



Tempo de reparação



Slide 0.13

Resultados de uma simulação de tamanho 10

#	F	TR	TT	TP	A	CA
1	1720	96	1816	5.29 %	4.758	228 370.04
2	1060	48	1108	4.33 %	7.798	187 148.01
3	1210	72	1282	5.62 %	6.739	242 620.90
4	670	96	766	12.53 %	11.279	541 409.92
5	1720	48	1768	2.71 %	4.887	117 285.07
6	220	24	244	9.84 %	35.410	424 918.03
7	40	48	88	54.55 %	98.182	2 356 363.64
8	920	72	992	7.26 %	8.710	313 548.39
9	110	72	182	39.56 %	47.473	1 709 010.99
10	1210	48	1258	3.82 %	6.868	164 833.07
			Médias	14.55 %	23.210	628 550.81

F – Número de horas em funcionamento [h]

TR – Tempo de reparação [h]

TT – Tempo total [h] (TT = F + TR)

TP – % de tempo perdido [%] (TP = TR / TT)

A – N^o de avarias por ano (A = 8640/ TT)

CA – Custo anual [€] ($500 \times 8640 \times TP$)

Slide 0.14

Resultados de uma simulação de tamanho 500 (folha de cálculo)

#	Aleatório 1	Aleatório 2	F	TR	TT	TP	A	CA
1	0.803912592	0.919235976	1466.28	60	1526.28	3.93%	5.661	169 825.20 €
2	0.678215165	0.372770009	1020.48	36	1056.48	3.41%	8.178	147 205.13 €
3	0.118657052	0.405819747	113.678	36	149.678	24.05%	57.72	1 039 033.14 €
4	0.112741828	0.649541753	107.657	48	155.657	30.84%	55.51	1 332 156.82 €
...	...	...	...	...	...	...	...	...
495	0.527863696	0.370136868	675.439	36	711.439	5.06%	12.14	218 599.27 €
496	0.109103853	0.818892176	103.975	48	151.975	31.58%	56.85	1 364 437.86 €
497	0.823176566	0.427847644	1559.34	36	1595.34	2.26%	5.416	97 483.72 €
498	0.232692884	0.195078969	238.381	36	274.381	13.12%	31.49	566 802.41 €
499	0.695068008	0.053693187	1068.9	24	1092.9	2.20%	7.906	94 866.88 €
500	0.347375127	0.038424919	384.078	12	396.078	3.03%	21.81	130 883.47 €
		Média	**900.38**	**42.2**	**942.62**	**12.96%**	**27.2**	**559758.2768**

F – Número de horas em funcionamento [h]

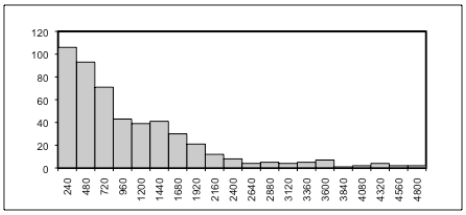
TR – Tempo de reparação [h]

TT – Tempo total [h] (TT = F + TR)

TP – % de tempo perdido [%] (TP = TR / TT)

A – N^o de avarias por ano (A = 8640/ TT)

CA – Custo anual [€] ($500 \times 8640 \times TP$)



Slide 0.15

6.5 Exercício

- Repetir o exemplo dos slides (Manutenção de uma linha de produção) considerando agora que o número de horas de funcionamento entre avarias segue uma distribuição exponencial negativa com valor esperado igual a 675.5 horas.

Resolução:

Como agora o tempo de funcionamento entre avarias (T_F) segue uma distribuição exponencial negativa, necessitámos de gerar números aleatórios com esta distribuição. Para tal teremos de inverter a respectiva função distribuição de probabilidade:

$$T_F \rightsquigarrow EN(\lambda = 1/675.5) \qquad \longrightarrow \qquad F(T_F) = 1 - e^{-\lambda T_F}$$

$$U = F(T_F) = 1 - e^{-\lambda T_F} \Leftrightarrow 1 - U = e^{-\lambda T_F} \Leftrightarrow T_F = -\frac{1-U}{\lambda}$$

As restantes colunas da tabela seguinte são idênticas à da situação anterior

Slide 0.16

Resultados de uma simulação de tamanho 500 (folha de cálculo)

#	Aleatório 1	Aleatório 2	F	TR	TT	TP	A	CA
1	0.803912592	0.919235976	1100.52	60	1160.52	5.17%	7.445	223 347.95 €
2	0.678215165	0.372770009	765.931	36	801.931	4.49%	10.77	193 931.98 €
3	0.118657052	0.405819747	85.3214	36	121.321	29.67%	71.22	1 281 884.71 €
4	0.112741828	0.649541753	80.8028	48	128.803	37.27%	67.08	1 609 902.62 €
...	...	...	...	...	...	...	...	...
494	0.713327427	0.413705633	843.98	36	879.98	4.09%	9.818	176 731.38 €
495	0.527863696	0.370136868	506.954	36	542.954	6.63%	15.91	286 432.92 €
496	0.109103853	0.818892176	78.0388	48	126.039	38.08%	68.55	1 645 208.06 €
497	0.823176566	0.427847644	1170.37	36	1206.37	2.98%	7.162	128 915.27 €
498	0.232692884	0.195078969	178.918	36	214.918	16.75%	40.2	723 623.37 €
499	0.695068008	0.053693187	802.269	24	826.269	2.90%	10.46	125 479.76 €
500	0.347375127	0.038424919	288.272	12	300.272	4.00%	28.77	172 643.75 €
		Média	**675.78**	**42.2**	**718.02**	**13.82%**	**32.7**	**672310.476**

F – Número de horas em funcionamento [h]

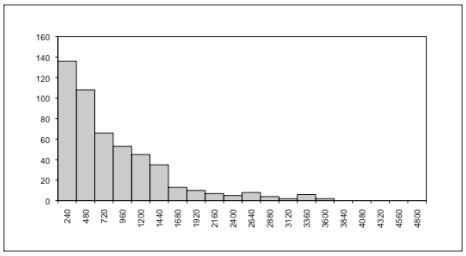
TR – Tempo de reparação [h]

TT – Tempo total [h] (TT = F + TR)

TP – % de tempo perdido [%] (TP = TR / TT)

A – N^o de avarias por ano (A = 8640/ TT)

CA – Custo anual [€] ($500 \times 8640 \times TP$)



Slide 0.17