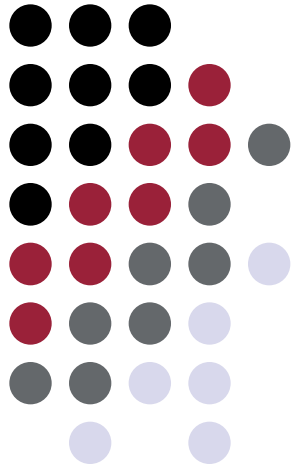
Projeto e Análise de Experimentos

CEA458 – Metodologia de Pesquisa Aplicada à Computação

Prof. MSc. George H. G. Fonseca
Universidade Federal de Ouro Preto



Sumário

- Sumarização de Dados
  - Métricas
  - Gráficos
- Projeto experimental
  - Projeto simples
  - Projeto c/ fatorial completo
  - Projeto c/ fatorial fracionado
  - Avaliação de impacto de fatores

Introdução

- A maioria das pesquisas desenvolvidas em computação é experimental
  - Avaliação de um algoritmo
  - Avaliação de desempenho um sistema
  - Comparação de configurações de sistema/algoritmo

Introdução

- Muitas vezes pode ser inviável avaliar todas as configurações possíveis de experimentos
- Também pode ser necessário sumarizar os dados, quando inviável ou improdutivo avaliar todos os valores
- Outro fator importante é a análise correta dos resultados

Sumarização de Dados

- Dada uma amostra $\{x_1, x_2, \ldots, x_n\}$ de observações:
  - Média: $\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$
  - Mediana: o $x_{n/2}$ elemento da lista ordenada
  - Moda: observação de maior frequência
  - Variância: $s^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}$
  - Desvio Padrão: $s = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \bar{x})}{n-1}}$, ou $s = \sqrt{s^2}$
  - Coeficiente de variação: $s/\bar{x}$

Sumarização de Dados

- Exemplo
  - Considere as seguintes notas {1, 3, 5, 6, 7, 8, 8, 9, 10}
  - Média: $\bar{x}$
  - Mediana:
  - Moda:
  - Variância: s^2
  - Desvio Padrão: s
  - Coeficiente de Variação: $s/\bar{x}$

Sumarização de Dados

Exemplo

- Considere as seguintes notas {1, 3, 5, 6, 7, 8, 8, 9, 10}
- Média: $\bar{x} = 6{,}33$
- Mediana: 7
- Moda: 8
- Variância:

$$s^2 = \frac{(1-6{,}33)^2+(3-6{,}33)^2+(5-6{,}33)^2+(6-6{,}33)^2+(7-6{,}33)^2+(8-6{,}33)^2+(8-6{,}33)^2+(9-6{,}33)^2+(10-6{,}33)^2}{8}$$

$$= 8{,}50$$

- Desvio Padrão: $s = \sqrt{8{,}50} = 2{,}91$
- Coeficiente de Variação: $s/\bar{x} = 0{,}46$

Sumarização de Dados

- Dada uma amostra $\{x_1, x_2, \ldots, x_n\}$ de observações:
  - Percentil: divide as observações em 100 partes, sendo que o x-ésimo percentil separa as x% maiores observações do resto
  - Quartil: casos especiais de percentil (25 (1º quartil), 50 (mediana), 75 (3º quartil))

Sumarização de Dados

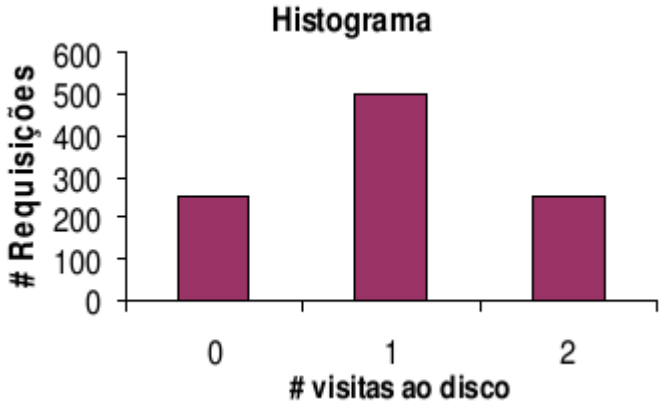
- Exemplo:

10	16	19	21	24	31	32	43	57	59	62	75	115	116	120	121	125	126	128	141
159	168	175	176	178	186	191	197	212	230	234	237	241	244	244	246	246	246	254	261
265	273	273	280	284	287	287	291	302	305	305	305	323	328	343	345	358	369	380	381
384	386	388	396	400	415	419	422	423	426	441	456	463	466	467	476	480	482	492	500
500	513	516	522	531	533	547	547	554	560	569	577	584	584	586	593	603	603	607	612
617	620	625	625	627	629	637	640	650	650	651	664	670	679	701	710	712	713	715	715
727	753	755	761	776	778	783	783	786	789	795	796	800	804	805	807	818	819	823	828
846	846	858	858	876	877	883	884	884	895	899	907	907	918	944	946	984	993	995	1000

- 25-ésimo percentil (1º quartil): 261
- 50-ésimo percentil (mediana, 2º quartil): 500
- 75-ésimo percentil (3º quartil): 715

Sumarização de Dados

- Histograma
  - Descreve o número de vezes em que o resultado de um experimento foi igual a cada ponto amostral
  - Ex.: Número de requisições que fizeram 0, 1 e 2 acessos ao disco



Sumarização de Dados

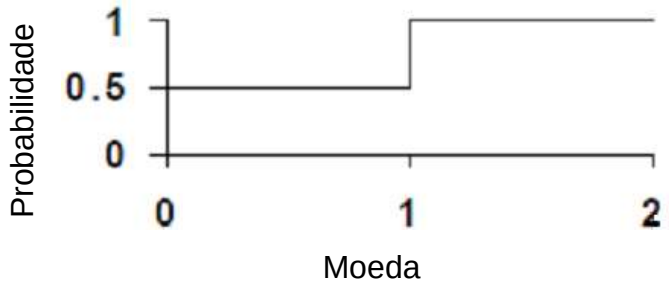
- Função de Distribuição Cumulativa
  - Ou CDF – *Cumulative Distribuition Function*
  - Mapeia um valor para uma probabilidade cujo resultado é menor ou igual a *a*

$$F_x(a) = P(x \leq a)$$

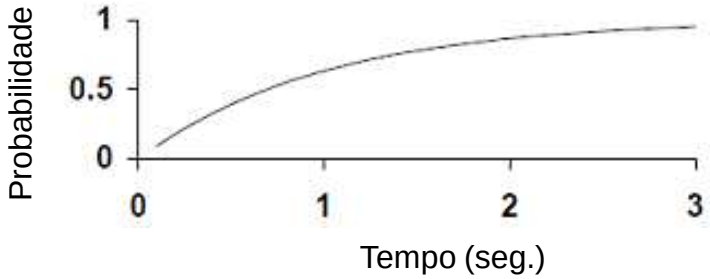
  - Permite visualizar diversos comportamentos nos dados

Sumarização de Dados

- Função de Distribuição Cumulativa
  - Ex. 1: Jogada de uma moeda (0 = cara; 1 = coroa)

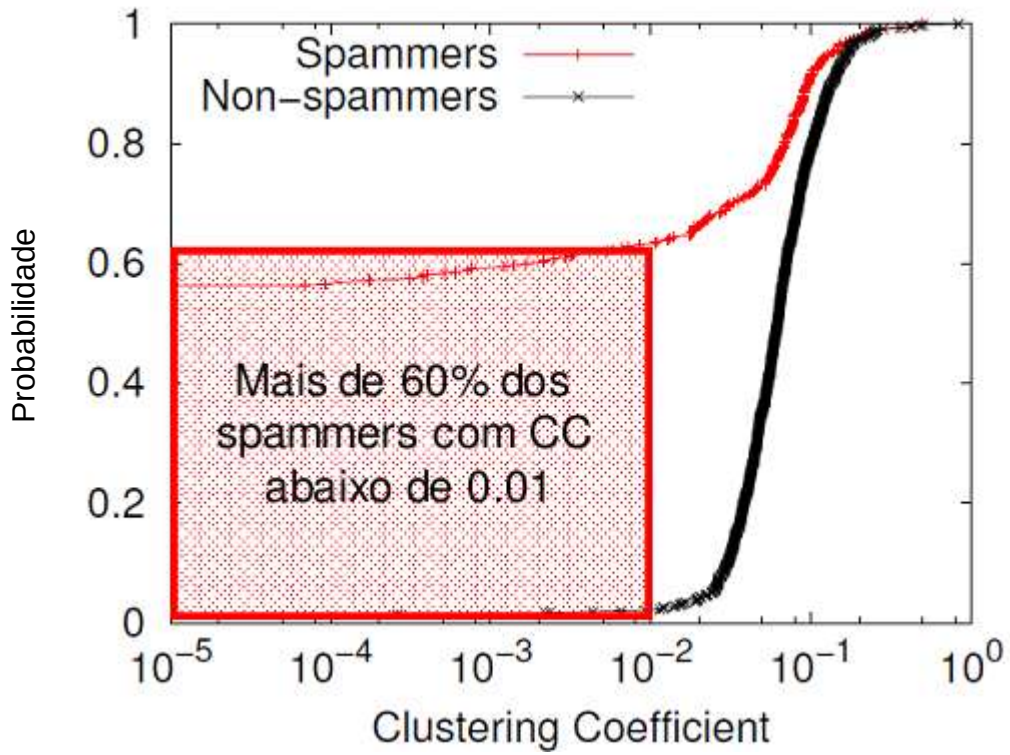


  - Ex. 2: Tempo entre chegas de pacotes



Sumarização de Dados

- Função de Distribuição Cumulativa
  - Ex. 3: Coeficiente de clusterização (0 = cara; 1 = coroa) (Probabilidade dos vizinhos de um nodo (indivíduo da rede social) estarem conectados)



Amigos de *spammers* não estão conectados entre si

Sumarização de Dados

- Função de Probabilidade Cumulativa
  - Ou PDF – *Probability Distribuition Function*
  - Derivada da CDF

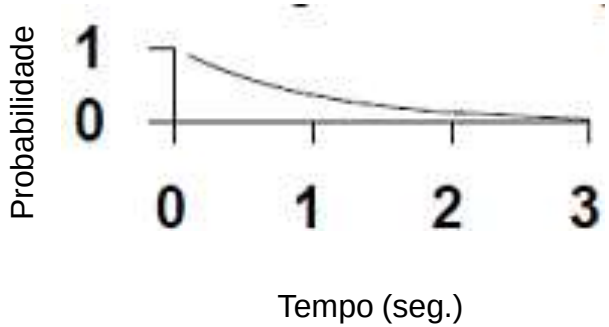
$$f(x) = \frac{dF(x)}{dx}$$

  - Útil para determinar intervalos de probabilidades

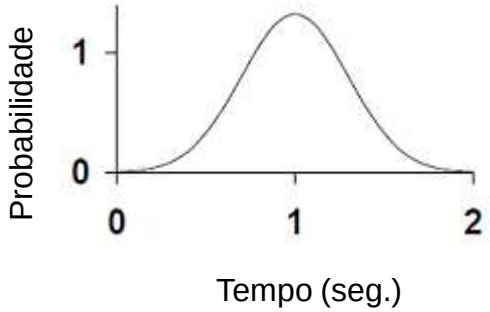
$$P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1)$$
$$= \int_{x_1}^{x_2} f(x)dx$$

Sumarização de Dados

- Função de Probabilidade Cumulativa
  - Ex. 1: Tempo entre chegadas de pacotes (protocolo 1)



  - Ex. 2: Tempo entre chegadas de pacotes (protocolo 2)



Projeto Experimental

- Objetiva a definir os experimentos a serem realizados em uma pesquisa

- Terminologia
  - Variável resposta
  - Fatores
  - Níveis
  - Replicação
  - Interação

Projeto Experimental

- Variável resposta
  - Valor obtido no experimento (tempo de resposta, custo de solução, utilização, etc)
- Fatores
  - Variáveis de entrada de um experimento (tamanho da mém. cache, tam. da mem. principal)
- Níveis
  - Valores específicos atribuídos a um fator (tam. mém cache: 8mb, 16mb, 32mb)

Projeto Experimental

- Replicação
  - Re-executar completamente um experimento com as mesmas entradas
  - Objetiva a determinar o impacto do erro experimental na variável resposta quando experimentos sujeitos a variações aleatórias

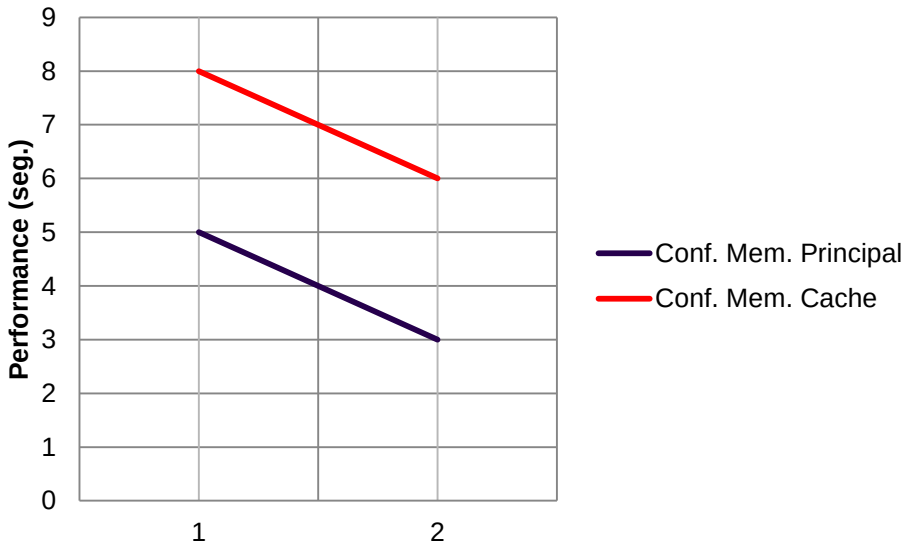
- Interação
  - Uma interação entre fatores ocorre quando o efeito de um fator depende do nível de outro fator

Projeto Experimental

- Interação

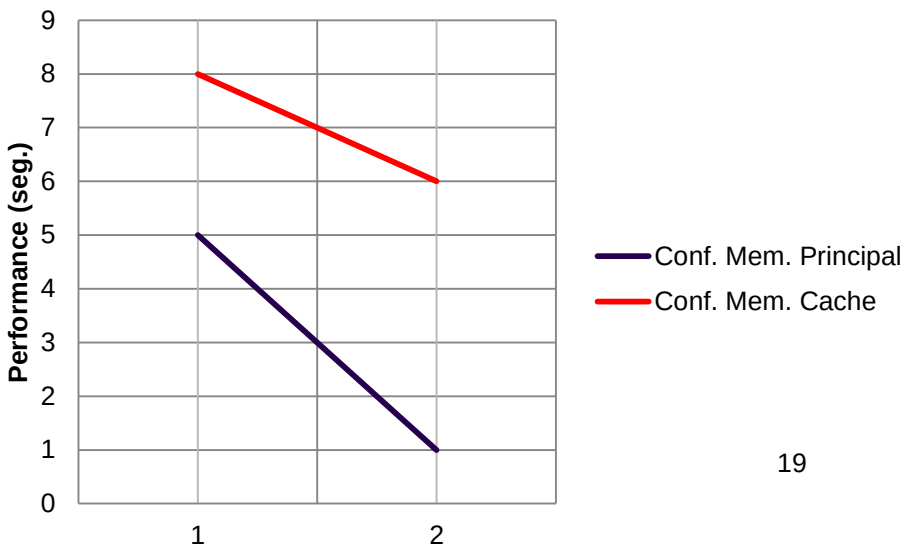
Fatores sem interação

Cache / Principal	1 gb. RAM	2 gb. RAM
8 mb. cache	5	3
16 mb. cache	8	6



Fatores com interação

Cache / Principal	1 gb. RAM	2 gb. RAM
8 mb. cache	5	1
16 mb. cache	8	6



Projeto Experimental

- Assumindo um experimento com 3 fatores, e.g.:

  Fator 1: (k_{11}, k_{12}, ..., k_{1n})

  Fator 2: (k_{21}, k_{22}, ..., k_{1m})

  Fator 3: (k_{31}, k_{32}, ..., k_{3p})

  - O estudo completo das configurações envolve n x m x p experimentos (fora as repetições)

- Há três formas de realizar os experimentos
  - Projeto Simples
  - Projeto com Fatorial Completo
  - Projeto com Fatorial Fracionado

Projeto Experimental

- Projeto Simples
  - Assume que um fator não interfere no outro (e.g. não há interação)
  - Varia um fator de cada vez:

Fixo: $k_{??}$
Varia: $k_{??}$
Melhor: $k_{??}$

Fator 1: (k_{11}, k_{12}, ..., k_{1n})
Fator 2: (k_{21}, k_{22}, ..., k_{1m})
Fator 3: (k_{31}, k_{32}, ..., k_{3p})

Fator 1: (k_{11}, k_{12}, ..., k_{1n})
Fator 2: (k_{21}, k_{22}, ..., k_{1m})
Fator 3: (k_{31}, k_{32}, ..., k_{3p})

Fator 1: (k_{11}, k_{12}, ..., k_{1n})
Fator 2: (k_{21}, k_{22}, ..., k_{1m})
Fator 3: (k_{31}, k_{32}, ..., k_{3p})

Projeto Experimental

- Projeto com Fatorial Completo
  - Teste de cada combinação possível dos fatores
  - Captura a informação completa da interação entre fatores
  - Trabalho enorme!!!
    - No exemplo abaixo, *n* x *m* x *p* execuções

Fixo: $k_{??}$
Varia: $k_{??}$
Melhor: $k_{??}$

Fator 1: (k_{11}, k_{12}, ..., k_{1n})

Fator 2: (k_{21}, k_{22}, ..., k_{1m})

Fator 3: (k_{31}, k_{32}, ..., k_{3p})

Projeto Experimental

- Projeto com Fatorial Fracionado
  - A medição é feita somente sobre alguns fatores
  - Se sabe *a priori* quais fatores têm interação
  - Ex.: Fator 1 e Fator 2 não têm interação

Fixo: $k_{??}$
Varia: $k_{??}$
Melhor: $k_{??}$

Fator 1: (k_{11}, k_{12}, ..., k_{1n})
Fator 2: (k_{21}, k_{22}, ..., k_{1m})
Fator 3: (k_{31}, k_{32}, ..., k_{3p})

Fator 1: (k_{11}, k_{12}, ..., k_{1n})
Fator 2: (k_{21}, k_{22}, ..., k_{1m})
Fator 3: (k_{31}, k_{32}, ..., k_{3p})

Projeto Experimental

- E como descobrir o nível de interação entre dois fatores e o impacto de cada fator no resultado???
  - Calculando a Soma dos Quadrados (SS) de cada fator / interação
  - Ex.: Considere os seguintes fatores e níveis

$$x_A = \begin{cases} -1, se\ 4\ núcleos \\ 1, se\ 16\ núcleos \end{cases}$$

$$x_B = \begin{cases} -1, se\ 8\ mb\ cache \\ 1, se\ 16\ mb\ cache \end{cases}$$

Fator B / Fator A	**8 núcleos**	**16 núcleos**
8 mb cache	820	217
16 mb cache	776	197

Projeto Experimental

- E como descobrir o nível de interação entre dois fatores e o impacto de cada fator no resultado???
  - Modelo de regressão não-linear

    As variáveis *y* representam o resultado de cada experimento

Experimento	**A**	**B**	**y**
1	-1	-1	$y_1 = 820$
2	1	-1	$y_2 = 217$
3	-1	1	$y_3 = 776$
4	1	1	$y_4 = 197$

Projeto Experimental

- E como descobrir o nível de interação entre dois fatores e o impacto de cada fator no resultado???
  - Solução para os *q*'s

$$q_0 = \frac{y_1 + y_2 + y_3 + y_4}{4} =$$ q_0 representa a simples média dos resultados

$$q_A = \frac{-y_1 - y_2 + y_3 + y_4}{4}$$ Os sinais de y representam o valor (+/-1) do fator A no resultado y

$$q_B = \frac{-y_1 + y_2 - y_3 + y_4}{4}$$ Os sinais de y representam o valor (+/-1) do fator B no resultado y

$$q_{AB} = \frac{y_1 - y_2 - y_3 + y_4}{4}$$ Os sinais de y representam o valor (+/-1) do interação AB no resultado y

Projeto Experimental

- E como descobrir o nível de interação entre dois fatores e o impacto de cada fator no resultado???
  - Solução para os *q*'s

$$q_0 = \frac{y_1 + y_2 + y_3 + y_4}{4} = \frac{802+217+776+197}{4} = 502{,}5$$

$$q_A = \frac{-y_1 + y_2 - y_3 + y_4}{4} = \frac{-802+217-776+197}{4} = -295{,}5$$

$$q_B = \frac{-y_1 - y_2 + y_3 + y_4}{4} = \frac{-802-217+776+197}{4} = -16$$

$$q_{AB} = \frac{y_1 - y_2 - y_3 + y_4}{4} = \frac{802 -217-776+197}{4} = 6$$

Projeto Experimental

- E como descobrir o nível de interação entre dois fatores e o impacto de cada fator no resultado???

  - Acabou? Ainda não!
  - Com esses *q*'s, procede-se ao cálculo da influência de cada fator no resultado

$$SSA = 2^2 \times q_A^2 = 349281 \quad (99{,}67\%)$$

$$SSB = 2^2 \times q_B^2 = 1024 \quad (0{,}29\%)$$

$$SSAB = 2^2 \times q_{AB}^2 = 144 \quad (0{,}04\%)$$

$$SST = SSA + SSB + SSAB = 350449 \quad (100{,}0\%)$$

Projeto Experimental

- E como descobrir o nível de interação entre dois fatores e o impacto de cada fator no resultado???
  - Interpretando o resultado

$SSA = 2^2 \times q_A^2 = 349281$ (99,67%)

$SSB = 2^2 \times q_B^2 = 1024$ (0,29%)

$SSAB = 2^2 \times q_{AB}^2 = 144$ (0,04%)

$SST = SSA + SSB + SSAB = 350449$ (100,0%)

O fator A explica 99,67% da variação no resultado

O fator B explica 0,29% da variação no resultado

A interação entre A e B explica 0,04% da variação no resultado

Exercícios

- Considerando os seguintes dados sobre tempos de execução de um programa, calcule a média, moda, mediana, desvio padrão, coeficiente de variação e 1º quartil

  {3, 4, 5, 5, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 11, 12, 15, 17, 20}

- Dados os seguintes resultados de experimentos, calcule a influência de cada um dos fatores no resultado e a interação entre os mesmos

Fator B / Fator A	**Temp. = 5**	**Temp. = 40**
Alpha = 0, 5	104	52
Alpha = 0,9	220	185

Bibliografia

- Jain, R. *The Art of Computer System Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling*. Wiley, 1992.

- Benevenuto, F. *Notas de aula. Metodologia de Pesquisa em Computação*. Universidade Federal de Ouro Preto, 2012.