Workload Characterization

Dr. John Mellor-Crummey

Department of Computer Science
Rice University

johnmc@cs.rice.edu

Goals for Today

Understand

- **Different approaches for characterizing workload**

Workload Characterization

Two key parts

- **Observe key performance characteristics of a workload**
- **Develop a model that can be used for further study**
- **Terms**
  - **workload unit/component: present service requests to SUT interface**
    - **examples: each application in a set, sites, user sessions**
    - **components should be homogeneous if possible, otherwise split**
- **Workload parameters:**
  - **measured quantities that depend on workload <u>not system</u>**
  - **types**
    - **service requests**
    - **resource demands**
  - **examples**
    - **transaction types, instructions, packet types & destinations**
    - **page reference patterns**

Techniques for Workload Characterization

- **Averaging**
- **Specifying Dispersion**
- **Single-parameter histograms**
- **Multiparameter histograms**
- **Principal components analysis**
- **Markov models**
- **Clustering**

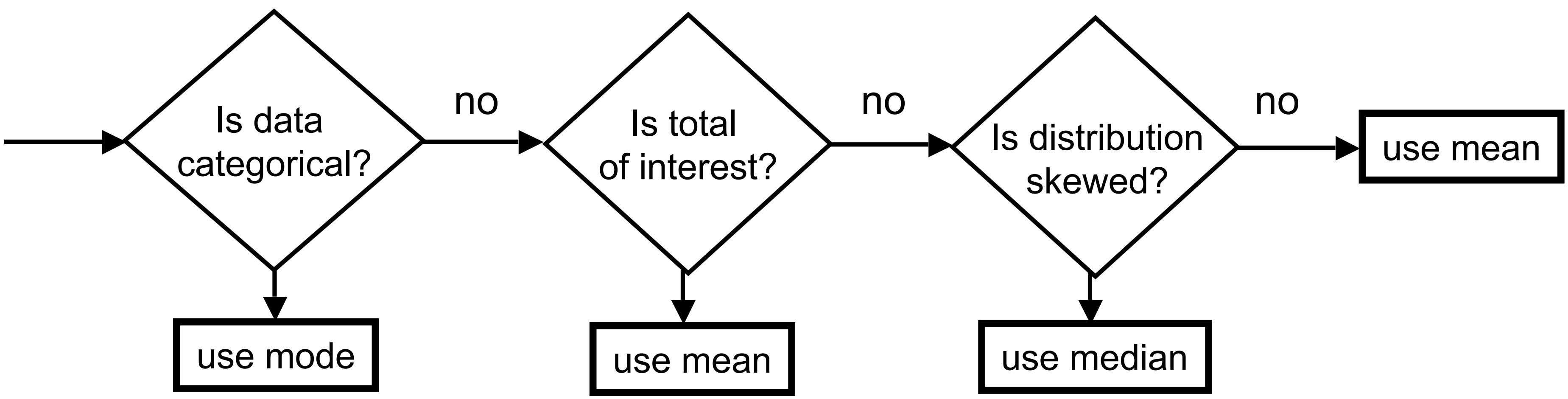
Note: the items marked in red will be discussed in the next lecture

Averaging

aka arithmetic mean of values $\{x_1,x_2,\ldots,x_n\}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Caution: arithmetic mean is not always appropriate "index of central tendency"**



- **Median = 50th percentile value**
- **Mode = most frequent**
  - **—e.g. most frequent destination for packets**

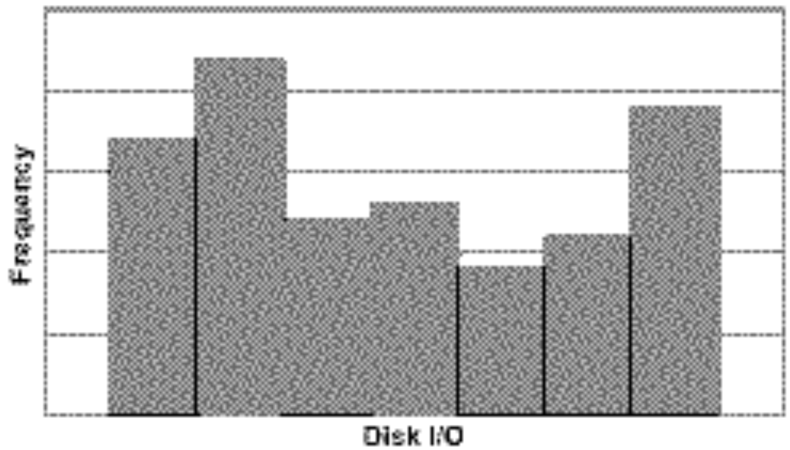
Specifying Dispersion

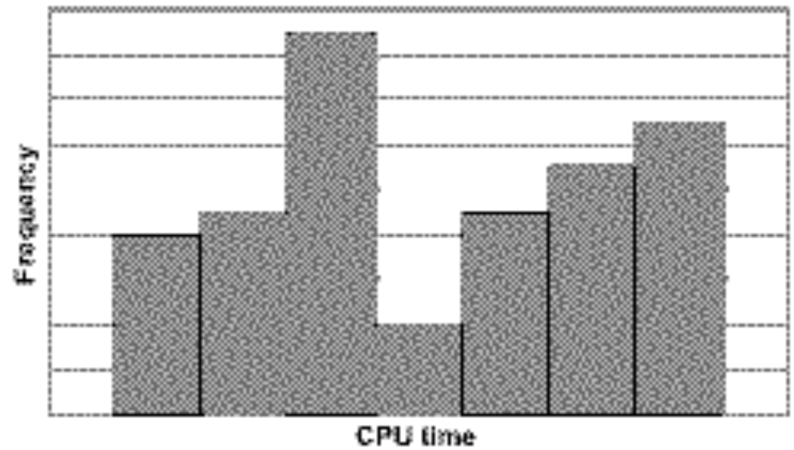
- Averaging is insufficient if ∃ large variability in data values
- Variability of $\{x_1, x_2, \ldots, x_n\}$ is commonly specified by variance

sample variance $$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Sample standard deviation, s = sqrt(sample variance)
  - —often more meaningful: same units as the mean
- Alternatives for summarizing variability
  - —range: maximum - minimum
  - —10 and 90 percentiles
  - —semi-interquartile range (SIQR) = $(x_{[.75(n-1)+1]} - x_{[.25(n-1)+1]})/2$
  - —mean absolute deviation $= \frac{1}{n} \sum_{i=1}^{n} | x_i - \bar{x} |$
- Coefficient of variation = ratio of std. dev to mean = $s / \bar{x}$
  - —if C.O.V. = 0, $\forall_i\ x_i = c$; high C.O.V. ⇒ mean is not sufficient

Single-parameter Histograms

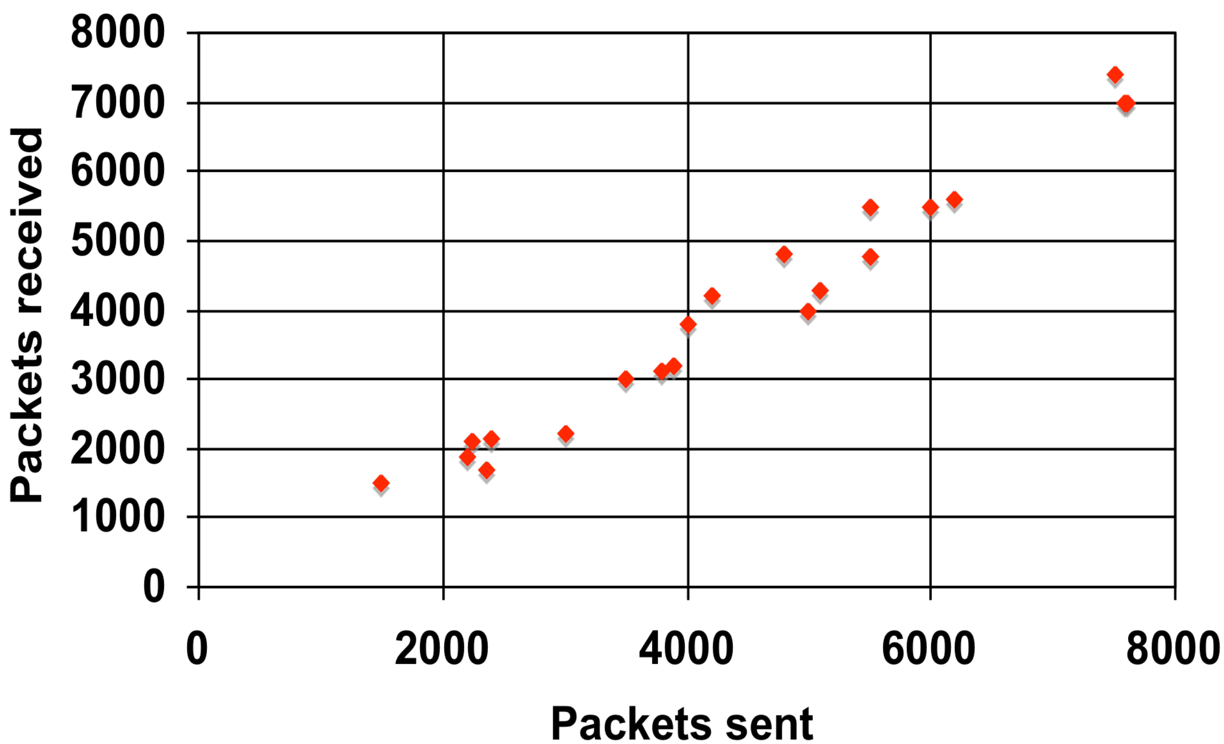




- **SPH: relative frequencies of various values of a parameter**
  - —divide complete range into buckets
  - —count observations that fall in each
- **Uses**
  - —simulation: generate test workload matching distribution
  - —analytical model: validate probability distribution used in model
- **Disadvantages**
  - —much data: n buckets, m parameters/ component, k components
    - – should only be used if variance is high and averages are inappropriate
  - —SPH ignore correlation among parameters

Multiparameter Histograms

- **MPH: use when significant correlation between parameters**



- **Disadvantages**
  - —rendering more than 2-parameter histograms is problematic
  - —much detail; uncommon to use them

Weighted Sum of Parameters

- **Classify workload components by weighted sum of their parameter values**

$$y = \sum_{j=1}^{n} a_j x_j$$

parameter

weight

- **Use y to classify components into categories, e.g. high, low**
- **Problem: choosing appropriate weights for parameters**
  - **—bad choice of weights may group dissimilar components**

Principal-Component Analysis

Choosing Good Weights

- **Problem**
  - —find weights so that weighted sums provide maximum discrimination among components

$$y_i = \sum_{j=1}^{n} a_{ij} x_j$$

- **For each component i,**
  - —y_i is a linear combination of parameter values x_j
  - —a_{ij} is loading of x_j on y_i
- **Choose weights so that y's form an orthogonal set, namely**

$$\langle y_i, y_j \rangle = \sum_k a_{ik} a_{kj} = 0$$

- **Properties: y's form an ordered set such that**
  - —y_1 explains highest percentage of variance in resource demands
  - —successive y_i explain increasingly lower percentages

Sample Problem

- **Given a set of n workstations**
  - x_{s_i} **number of packets sent by workstation i**
  - x_{r_i} **number of packets received by workstation i**
- **There is a considerable correlation between** x_{s_i} **and** x_{r_i}
- **Compute** y_{ki} **k=1,2 for each workstation i, such that successive** y_k **vectors provide next best discriminatory power**