# SUMMARIZING MEASURED DATA

Gaia Maselli

maselli@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

# Overview

- Basic concepts
- Summarizing measured data
- Summarizing data by a single number
- Summarizing variability
- Determining distribution of data

# Motivation

- A measurement project may result in several hundreds or millions of observations on a given variable.
- To present the measurements it is necessary to summarize data
- How to report the performance as a single number? Is specifying the mean the correct way?
- How to report the variability of measured quantities? What are the alternatives to variance and when are they appropriate?

# Basic concepts of probability and statistics

- **Independent Events:** Two events are called independent if the occurrence of one event does not in any way affect the probability of the other event.
  - ➢ Examples: Successive throws of a coin
- **Random Variable:** A variable is called a random variable if it takes one of a specified set of values with a specified probability.
- The outcome of a random event or experiment that yields a numeric value is a random variable
  - ➢ Examples: execution time

# CDF, PDF, and PMF

- **Cumulative Distribution Function (CDF)** of a random variable maps a given value *a* to the probability of the variable taking a value less than or equal to *a* (Starts at 0. Ends at 1)
  - For a random variable x,

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

- **Probability Density Function (PDF):** Starts at 0 and ends at 0

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

# CDF, PDF, and PMF (Cont)

- Given a probability density function f(x) the probability of x being in the interval ($x_1$,$x_2$) can be computed

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

- **Probability Mass Function (PMF):** For discrete random variables:

$$f(x_i) = p_i$$

$$\begin{aligned} P(x_1 < x \leq x_2) &= F(x_2) - F(x_1) \\ &= \textstyle\sum_i p_i \\ & \quad x_1 < x_i \leq x_2 \end{aligned}$$

# Mean, Variance, CoV

- **Mean** or **Expected Value**:

  Mean $\mu = E(x) = \sum_{i=1}^{n} p_i x_i = \int_{-\infty}^{+\infty} x f(x) dx$

- **Variance**: The expected value of the square of distance between x and its mean

$$\begin{aligned} Var(x) &= \sigma^2 = E[(x-\mu)^2] = \sum_{i=1}^{n} p_i (x_i - \mu)^2 \\ &= \int_{-\infty}^{+\infty} (x_i - \mu)^2 f(x) dx \end{aligned}$$

- **Coefficient of Variation**:

Standard deviation

$$\text{C.O.V.} = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{\sigma}{\mu}$$

SAPIENZA
UNIVERSITÀ DI ROMA

# Covariance and Correlation

- **Covariance**:

$$\begin{aligned} Cov(x,y) &= \sigma_{xy} = E[(x-\mu_x)(y-\mu_y)] \\ &= E(xy) - E(x)E(y) \end{aligned}$$

- For independent variables, the covariance is zero:

$$E(xy) = E(x)E(y)$$

- Although independence always implies zero covariance, the reverse is not true.
- **Correlation Coefficient**: normalized value of covariance

$$\text{Correlation}(x,y) = \rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

The correlation always lies between -1 and +1.

# Describing a set of measurements

- How can we summarize measured data?

***Numerical data (i.e., numbers)***

- Central tendency ➔ summarizing data by a single number
- Dispersion ➔ summarizing variability
- Histograms ➔ determining distribution of data
- Cumulative Distribution Function ➔ most detailed view

***Categorical (symbolic) data (e.g., symbols, names, etc.)***

- Histograms

# Summarizing data by a single number: central tendency

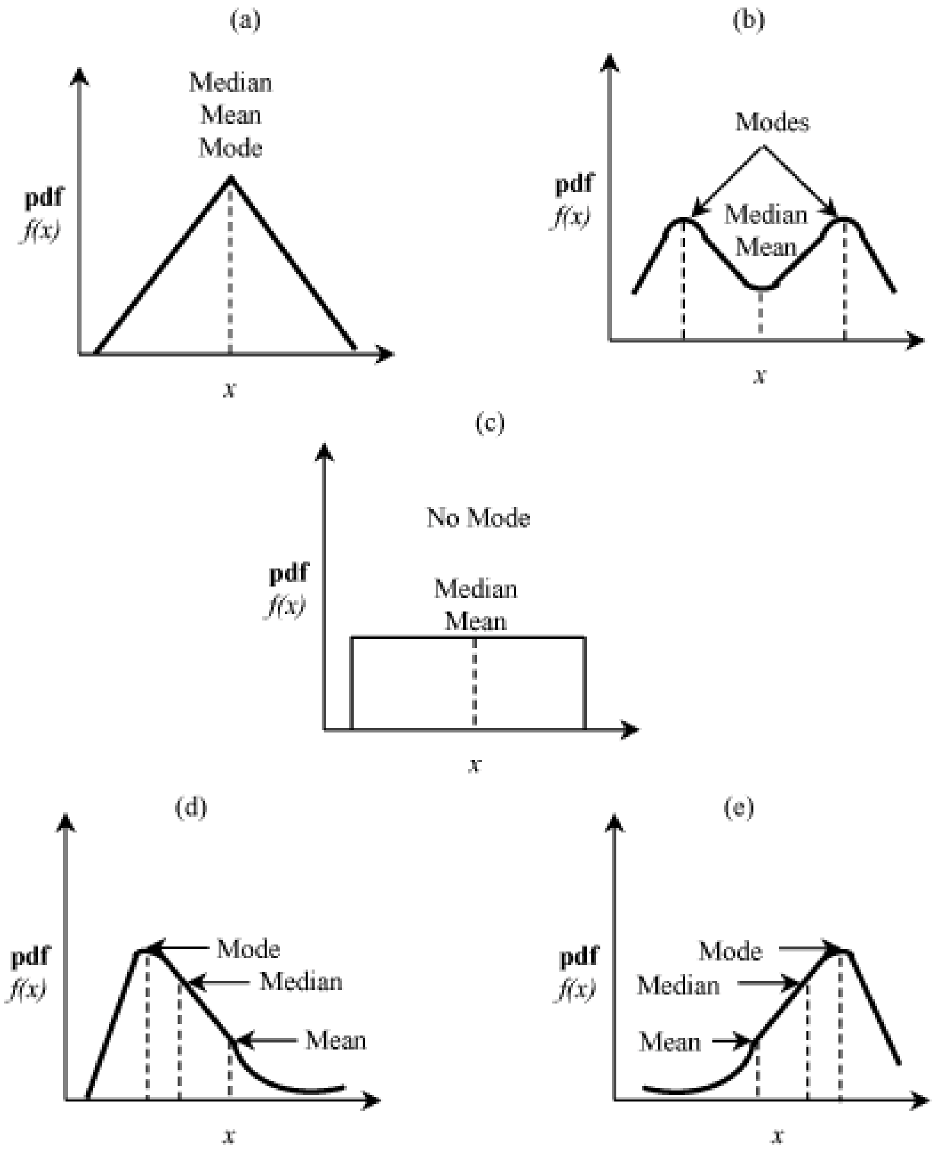
# Central tendency: empirical mean

- The simplest description of numerical data
- Given a dataset {$x_i$, i=1,…,N} central tendency tells us where on the number line the values tend to be located.
- ***Empirical mean (average)*** is the most widely used measure of central tendency.
- It is obtained by taking the sum of all observations and dividing this sum by the number of observations in the sample.

# Central tendency: median and mode

- Other very simple descriptions of numerical data
- ***Median***: (value which divides the sorted dataset into two equal parts) is obtained by sorting the observations in an increasing order and taking the observation that is in the middle of the series. If the number of observations is even, the mean of the middle two values is used as a median.
- ***Mode*** (most common/likely value) It is obtained by plotting a histogram and specifying the midpoint of the bucket where the histogram peaks.
- Mean and median always exist and are unique.
- Mode may not exist.

# Mean, median and mode: relationships



# Observations

- The main problem with the mean is that it is affected more by outliers than median and mode
- A single outlier can make a considerable change in the mean (this is particularly true for small samples)
- Median and mode are resistant to several outlying observations
- The mean gives equal weight to each observation and in this sense makes full use of the sample
- Median and mode ignore a lot of information
- The mean has an additivity or linearity property in that the mean of a sum is a sum of the means. This does not apply to the mode or median

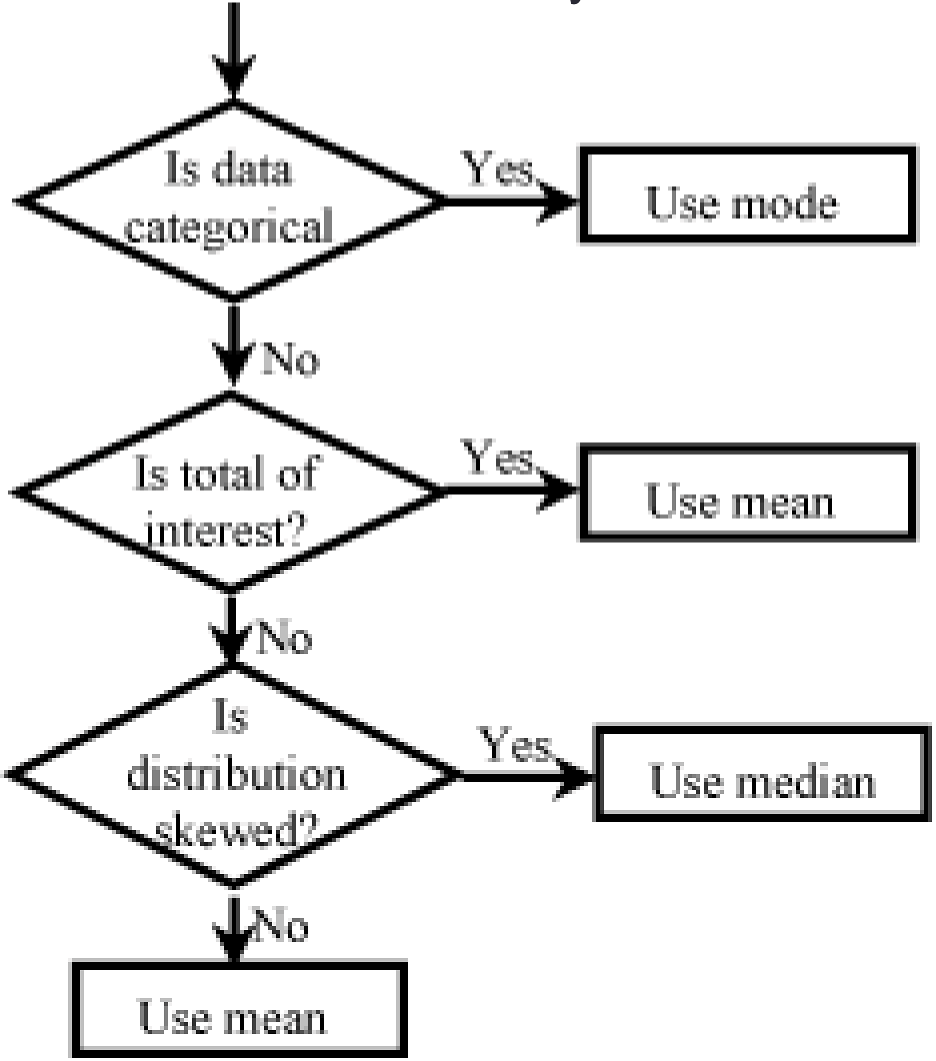
# Selecting mean, median, and mode

Guidelines to select a proper index of central tendency

Type of variable

Is the total of all observations of any interest?

In case of small samples if the ratio of the maximum and minimum of the observations is large, the data is skewed



# Examples

Examples of selections of indices of central tendency

- *Most used resource in a system*: Resources are categorical and hence the mode must be used
- *Inter arrival time*: total time is of interest and so the mean is the proper choice
- *Load on a computer*: the median is preferable due to highly skewed distribution

# Common misuses of means

**Using mean of significantly different values**

- When the mean is the correct index of central tendency for a variable, it does not automatically imply that a mean of any set of the variable will be useful
- *Usefulness* depends upon the number of values and the variance, not only on the type of the variable
- Example: it is not very useful to say that the mean packet latency is *2,501s* when the two measurements come out to be *2ms* and *5s*. An analysis based on 2,501s would lead nowhere close to the two possibilities (the mean is the correct index but is useless)

# Common misuses of means (Cont)

## Using mean without regard to skewness of distribution

| System A | System B |
|---:|---:|
| 10 | 5 |
| 9 | 5 |
| 11 | 5 |
| 10 | 4 |
| 10 | 31 |
| Sum=50 | Sum=50 |
| Mean=10 | Mean=10 |
| Typical=10 | Typical=5 |

- Both systems have mean response times of 10
- System A: it is **useful** to know the mean because the variance is low and 10 is the typical value
- System B: the typical value is 5; hence using 10 for the mean does not give any useful result. The variability is too large in this case

# Common misuses of means (Cont)

**Multiplying means to get the mean of a product**

- If the variable are correlated (not independent)

$$E(xy) \neq E(x)E(y)$$

**Taking a mean of a ratio with different bases**

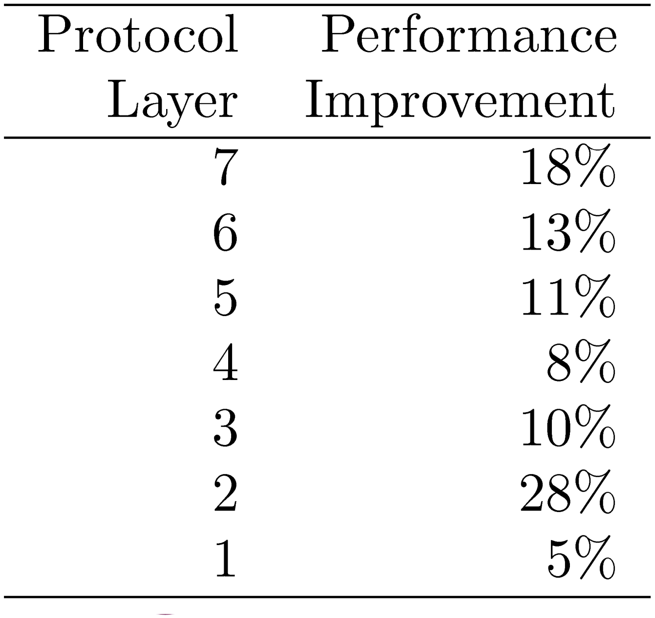
# Geometric mean

- The geometric mean of *n* values $x_1, x_2, \ldots, x_n$ is obtained by multiplying the values together and taking the *n*th root of the product

$$\dot{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

**Example**: The performance improvements in 7 layers.

What is the average improvement per layer?



| Protocol Layer | Performance Improvement |
|---:|---:|
| 7 | 18% |
| 6 | 13% |
| 5 | 11% |
| 4 | 8% |
| 3 | 10% |
| 2 | 28% |
| 1 | 5% |

# Geometric mean (Cont)

**Example**: The performance improvements in 7 layers. What is the average improvement per layer?

| Protocol Layer | Performance Improvement |
|---|---|
| 7 | 18% |
| 6 | 13% |
| 5 | 11% |
| 4 | 8% |
| 3 | 10% |
| 2 | 28% |
| 1 | 5% |

Average improvement per layer

$$= \{(1.18)(1.13)(1.11)(1.08)(1.10)(1.28)(1.05)\}^{\frac{1}{7}} - 1$$

$$= 0.13$$

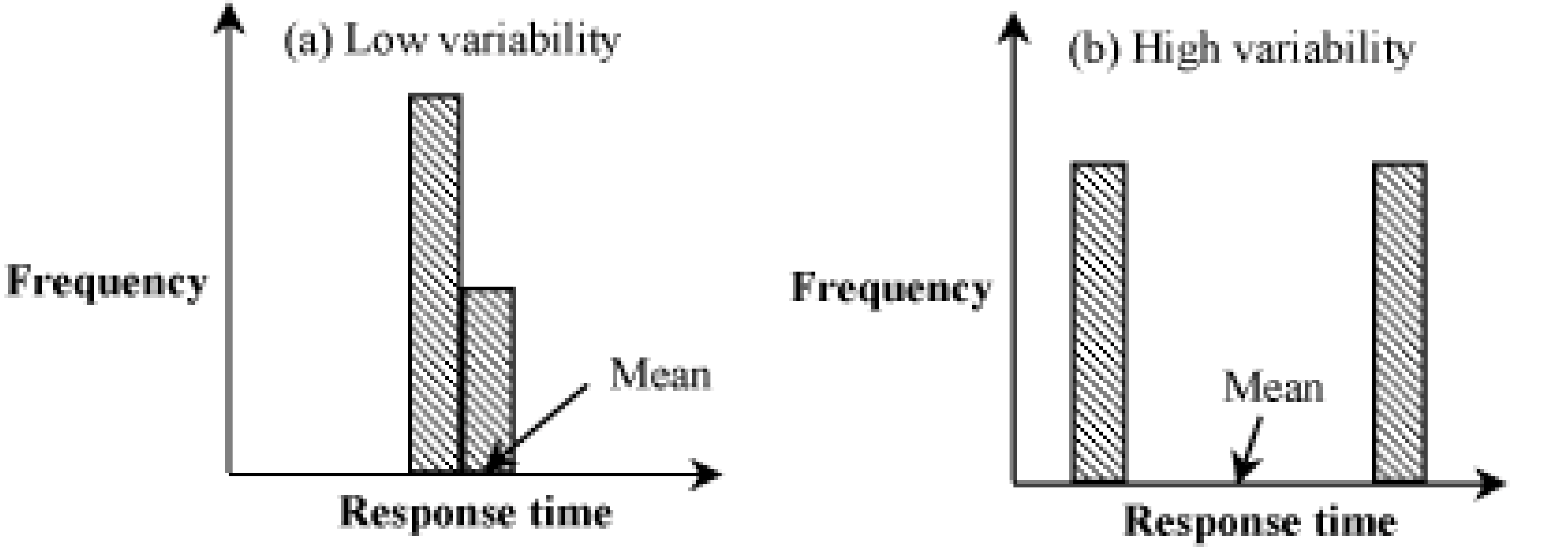
# Summarizing variability: dispersion

# Summarizing variability

- Given a data set, summarizing it by a single number is rarely enough
- It is important to include a statement about its variability in any summary of the data
- Motivation: given two systems with the same mean performance, one would generally prefer one whose performance does not vary much from the mean
- Systems with low variability are preferred
- Variability is specified using one of the following measures which are called *indices of dispersion*
  - Range-minimum and maximum of the values observed
  - Variance and standard deviation
  - Percentiles and quantiles
  - Mean absolute variation

# Example

- Histograms of response time of two systems



- Both systems have the same mean response time of 2 seconds but

a. The response time is close to the mean

b. The response time can be 1 ms and 1 minute

# Range

- The range of a stream of values can be easily calculated by keeping track of the minimum and the maximum
- Range = Max-Min
- Larger range => higher variability
- In most cases, range is not very useful.
- The minimum often comes out to be zero and the maximum comes out to be an ``outlier'' far from typical values.
- Unless the variable is bounded, the maximum goes on increasing with the number of observations, the minimum goes on decreasing with the number of observations, and there is no “stable'' point that gives a good indication of the actual range.
- Range is useful if, and only if, there is a reason to believe that the variable is bounded.

# Variance and coefficient of variation

- *Variance* is measured in squared units; the square root of variance is standard deviation, which is expressed in the same units as the data

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

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

- Variance is expressed in units that are the square of the units of the observations
- *Standard deviation s* is preferred
- *Coefficient of variation*: dimensionless measure of the dispersion of a dataset

# Quantile and percentile

- A more detailed description of dataset dispersion is in terms of quantiles and percentile
- The *pth quantile* is the value below which the *fraction p* of the values lie
- The median is the 0.5 quantile
- This can also be expressed as a *percentile*, e.g., the 90 th percentile is the value that is larger than 90% of the data

# Similar measures

- **Deciles**: percetiles at multiples of 10%
  - The first decile is the 10-percentile
- **Quartiles** divide the data into four parts at 25, 50 and 75%
  - Second quartile $Q_2$ is median
- The range between Q3 and Q1 is called **interquartile range** of the data.
- One half of this rage is called **Semi-Interquartile Range (SIQR)**

$$\mathrm{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{x_{0.75} - x_{0.25}}{2}$$

SAPIENZA
UNIVERSITÀ DI ROMA

# Mean absolute deviation

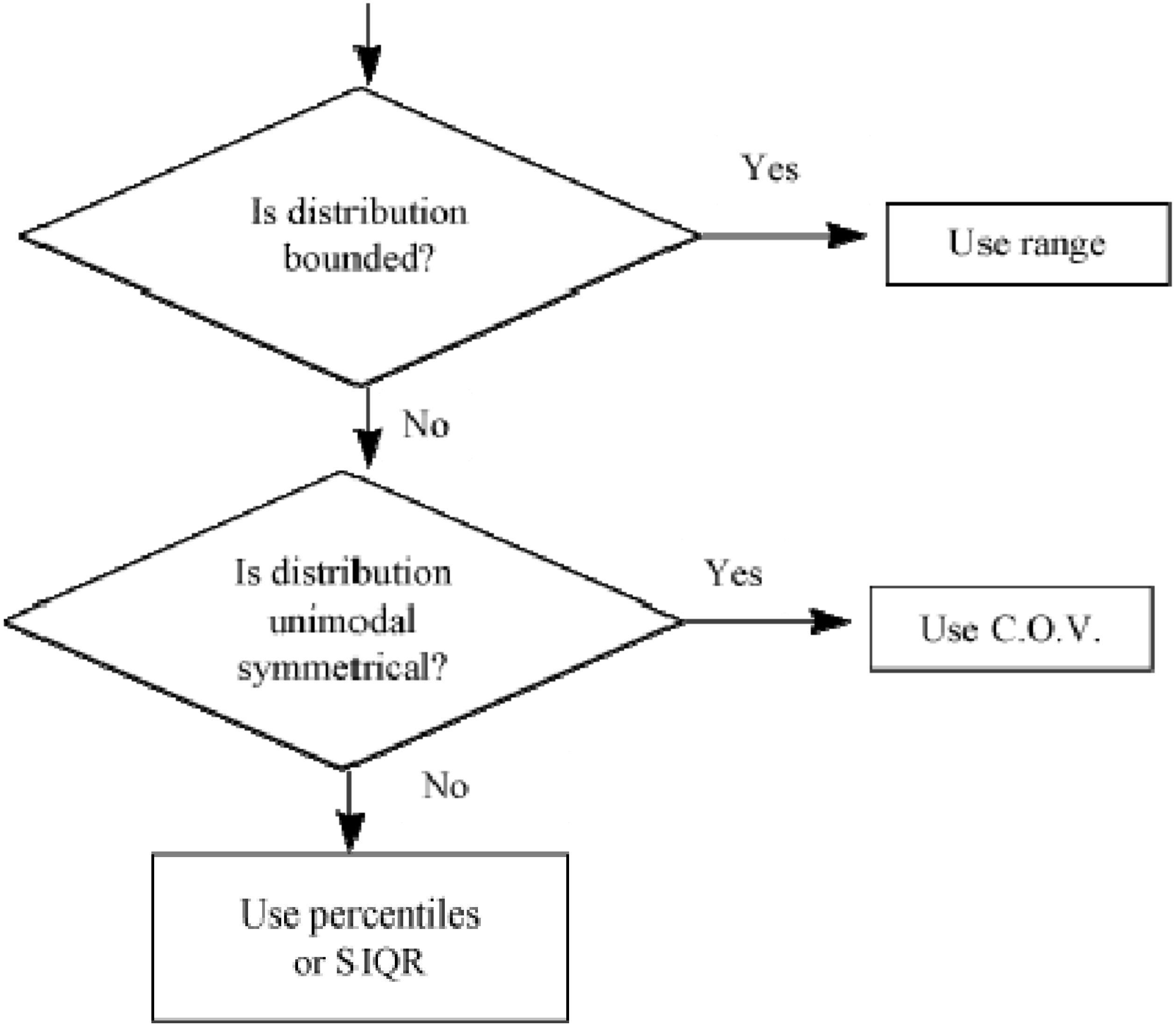
$$\text{Mean absolute deviation} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

The key advantage of the mean absolute deviation over the standard deviation is that no multiplication or square root is required

# Comparison of variation measures

- Range is affected considerably by outliers.
- Sample variance is also affected by outliers but the effect is less than that on the range
- Mean absolute deviation is next in resistance to outliers.
- Semi inter-quantile range is very resistant to outliers.
- If the distribution is highly skewed, outliers are highly likely and SIQR is preferred over standard deviation
- In general, SIQR is used as an index of dispersion whenever median is used as an index of central tendency.
- For qualitative (categorical) data, the dispersion can be specified by giving the number of most frequent categories that comprise the given percentile, for instance, top 90%.

# Selecting the index of dispersion



Is distribution bounded?

Yes

Use range

No

Is distribution unimodal symmetrical?

Yes

Use C.O.V.

No

Use percentiles or SIQR

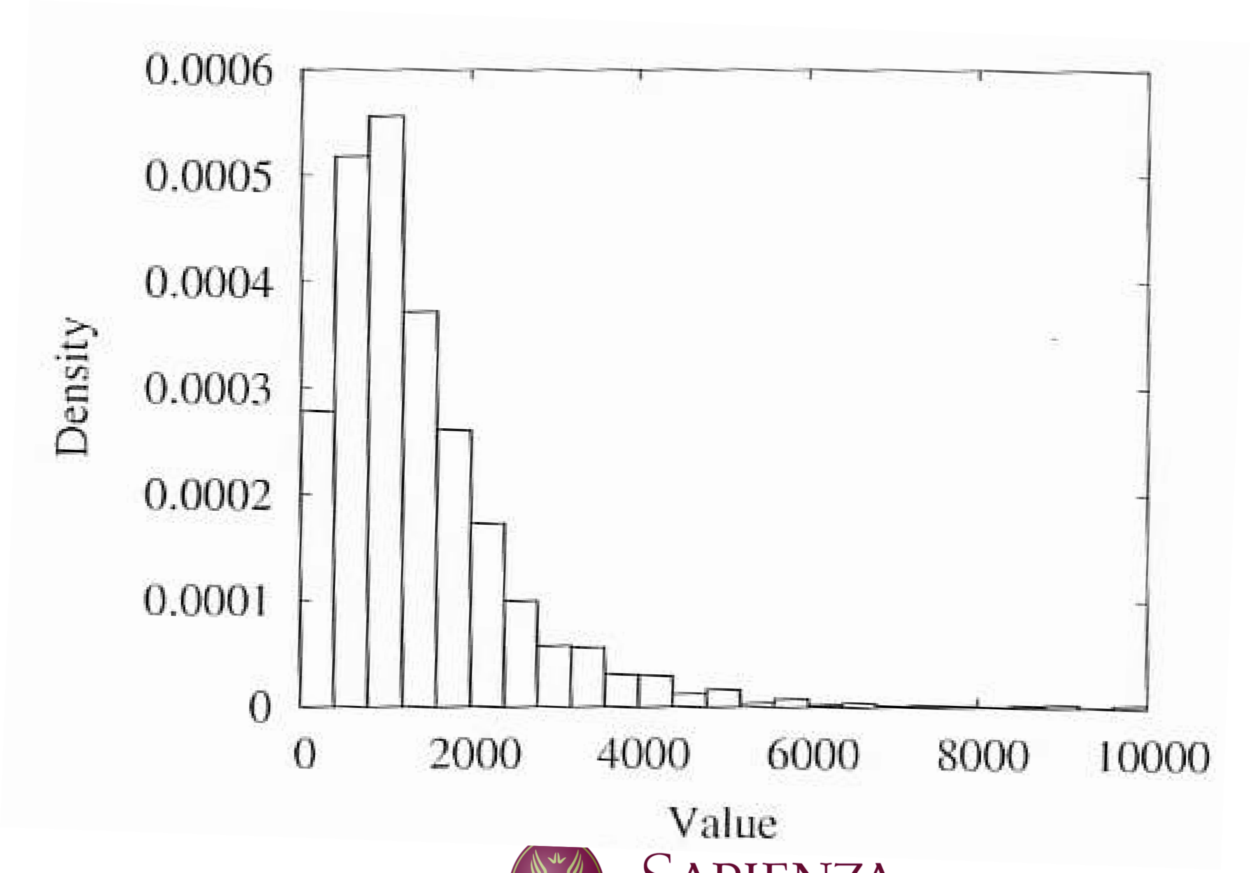
# Determining distribution of data: histograms and CDF

# Histograms

- Histograms give a more detailed description of measured data and allow to state the distribution the data follows
- The distribution information is also required if the summary has to be used later in simulation or analytical modeling
- The histogram is defined in terms of a set of bins (cells or buckets) that are a partition of the observed values
- To plot a histogram we need to determine the maximum and the minimum of the values and dividing the range into a number of sub-ranges (cells)
- The count of observations that fall into each cell is determined
- The count are normalized to cell frequency by dividing by the total number of observations
- The cell frequency are plotted as a column chart

# Histograms

- The histogram counts how many values fall in each bin



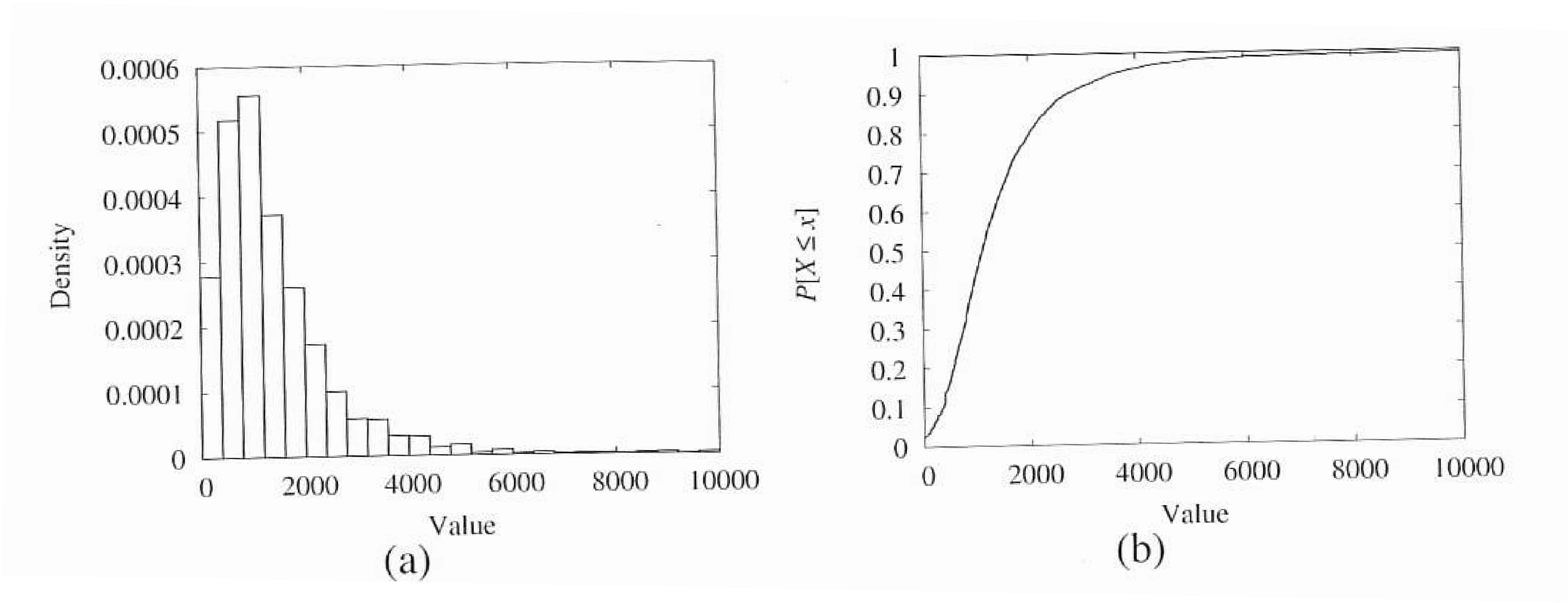
# Histograms

- A considerable problem in creating histograms is determining the cell sizes
- Small cells lead to very few observations per cell
- Large cells result in less variation but the details of the distribution are completely lost
- Given a set of data it is possible to reach very different conclusions about the distribution shape depending upon the cell size used
- One guideline is that if any cell has less than five observations, the cell size should be increased or a variable cell histogram should be used

# Cumulative distribution function (CDF)

- The most detailed view of a dataset is its empirical cumulative distribution function
- All the previous methods involve summarization of data in which some detail is lost
- The CDF potentially provides information about each value in the dataset
- The CDF is formed by plotting, for each unique value in the dataset, *the fraction of data items that are smaller than that value*
- In other words, one plots the quantile corresponding to each unique value in the dataset

# CDF: example



- The CDF involves no binning or averaging data values and so provides more information to the viewer than does the histogram

# Description of categorical (symbolic) data

- Most of data descriptions do not apply to categorical (symbolic) data
- The dispersion can be specified by giving the number of most frequent categories that comprise the given percentile, for instance, the top 90%
- Histograms can be used specifying categories
  - One can measure the empirical probability of each symbol in the dataset
  - The histogram can be plotted in order of decreasing empirical probability