Summarizing Measured Data

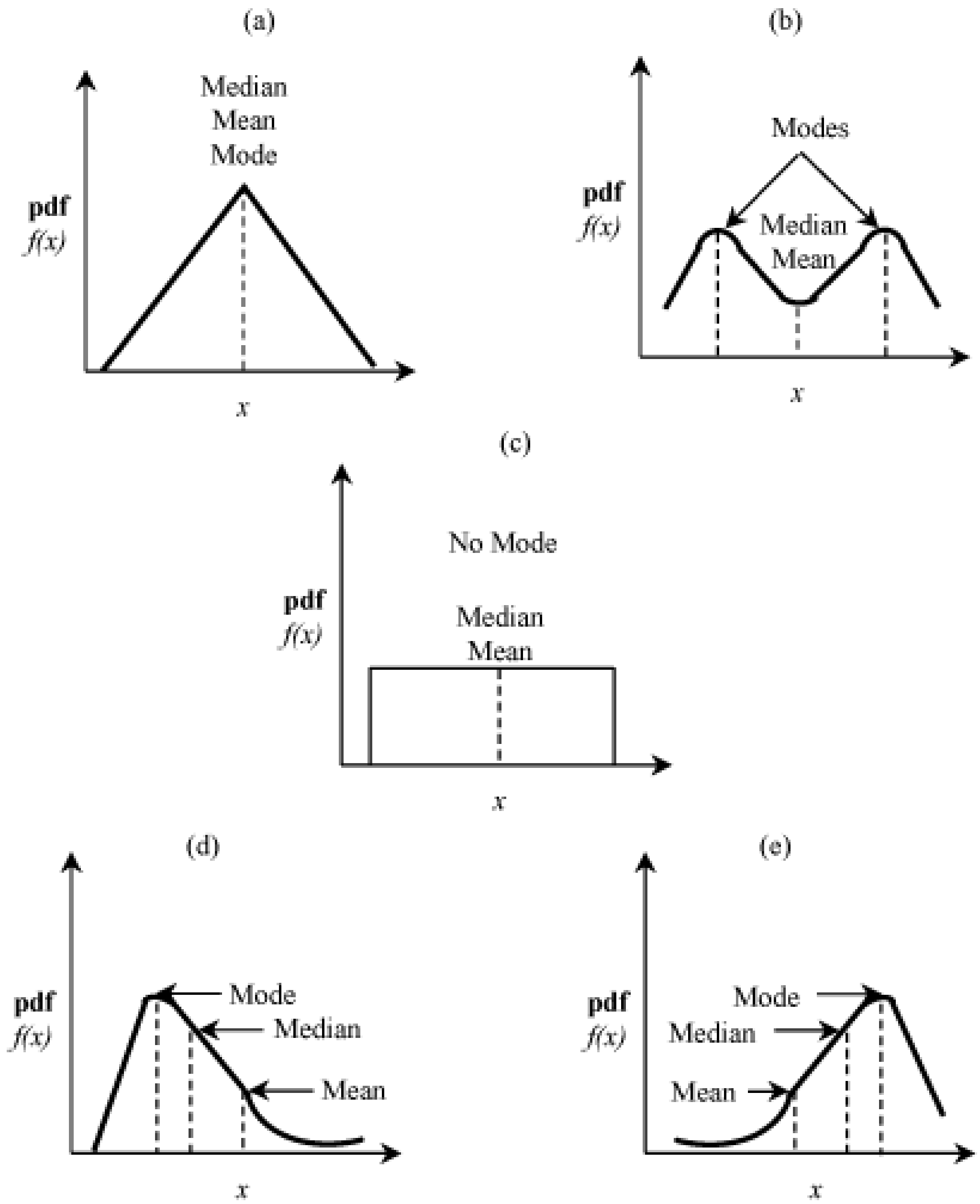
Adapted by Prof. Gheith Abandah

- Summarizing Data by a Single Number: Mean, Median, and Mode, Arithmetic, Geometric, Harmonic Means
- Mean of A Ratio
- Summarizing Variability: Range, Variance, percentiles, Quartiles

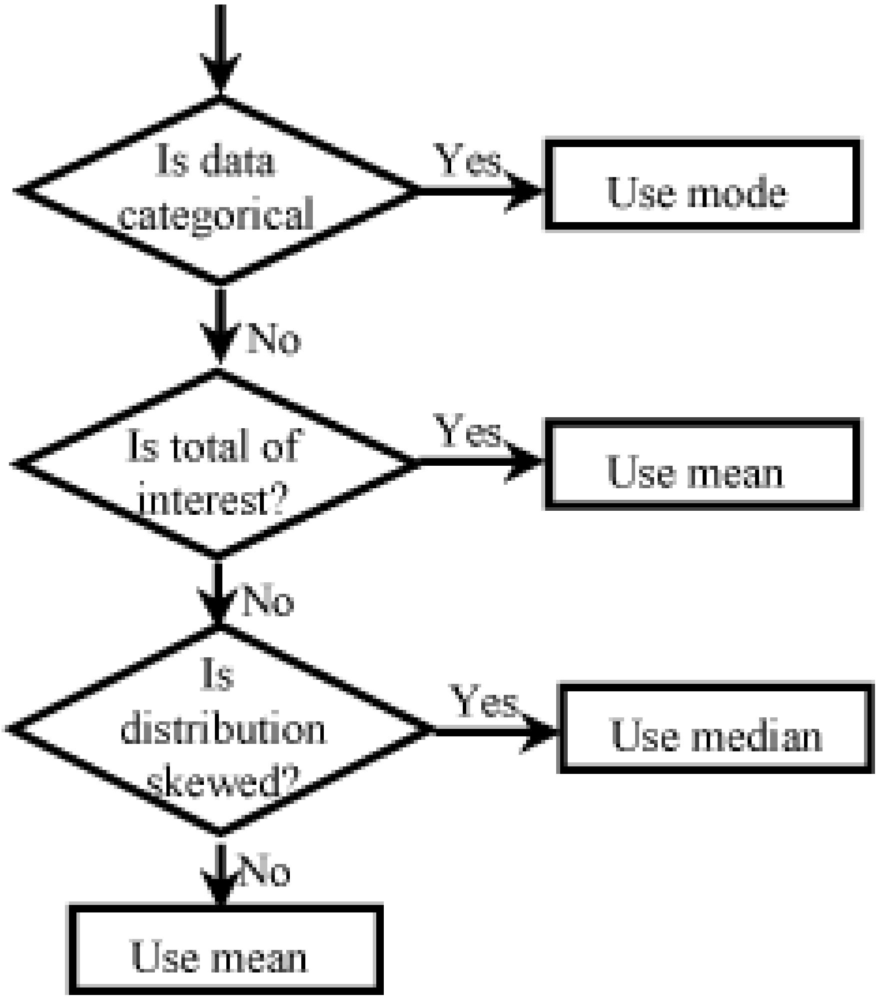
Summarizing Data by a Single Number

- **Indices of central tendencies**: Mean, Median, Mode
- **Sample Mean** is obtained by taking the sum of all observations and dividing this sum by the number of observations in the sample.
- **Sample Median** 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.
- **Sample Mode** is obtained by plotting a histogram and specifying the midpoint of the bucket where the histogram peaks. For categorical variables, mode is given by the category that occurs most frequently.
- Mean and median always exist and are unique. Mode, on the other hand, may not exist.

Mean, Median, and Mode: Relationships



Selecting Mean, Median, and Mode



Indices of Central Tendencies: Examples

- Most used resource in a system:
  - Resources are categorical and hence mode must be used.
- Interarrival time:
  - Total time is of interest and so mean is the proper choice.
- Load on a Computer:
  - Median is preferable due to a highly skewed distribution.
- Average Configuration:
  - Medians of number devices, memory sizes, number of processors are generally used to specify the configuration due to the skewness of the distribution.

Common Misuses of Means

- Using mean of significantly different values:
  - (10+1000)/2 = 505
- Using mean without regard to the 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

Misuses of Means (cont)

- Multiplying means to get the mean of a product

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

  - Example: On a timesharing system,
    Average number of users is 23
    Average number of sub-processes per user is 2
    What is the average number of sub-processes?
    Is it 46? No!
    The number of sub-processes a user spawns depends upon how much load there is on the system.
- Taking a mean of a ratio with different bases.
  - Already discussed in Chapter 11 on ratio games and is discussed further later

Geometric Mean

- Geometric mean is used if the product of the observations is a quantity of interest.

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

Geometric Mean: Example

- The performance improvements in 7 layers:

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

$$\begin{aligned}&\text{Average improvement per layer}\\&= \{(1.18)(1.13)(1.11)(1.08)(1.10)(1.28)(1.05)\}^{\frac{1}{7}} - 1\\&= 0.13\end{aligned}$$

Examples of Multiplicative Metrics

- Cache hit ratios over several levels of caches
- Cache miss ratios
- Percentage performance improvement between successive versions
- Average error rate per hop on a multi-hop path in a network.

Geometric Mean of Ratios

$$gm\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \ldots, \frac{x_n}{y_n}\right) = \frac{gm(x_1, x_2, \cdots, x_n)}{gm(y_1, y_2, \ldots, y_n)} = \frac{1}{gm\left(\frac{y_1}{x_1}, \frac{y_2}{x_2}, \ldots, \frac{y_n}{x_n}\right)}$$

- The geometric mean of a ratio is the ratio of the geometric means of the numerator and denominator
  - the choice of the base does not change the conclusion.
- It is because of this property that sometimes geometric mean is recommended for ratios.

Harmonic Mean

$$\ddot{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

- Used whenever an arithmetic mean can be justified for $1/x_i$
  - **Example**: Elapsed time of a benchmark on a processor
  - In the i^{th} repetition, the benchmark takes t_i seconds. Now suppose the benchmark has m million instructions, MIPS x_i computed from the i^{th} repetition is:

$$x_i = \frac{m}{t_i}$$

  - t_i's should be summarized using arithmetic mean since the sum of t_i has a physical meaning => x_i's should be summarized using harmonic mean since the sum of $1/x_i$'s has a physical meaning.

Mean of A Ratio

1. If the sum of numerators and the sum of denominators, both have a physical meaning, the average of the ratio is the ratio of the averages.
   - For example, if $x_i = a_i / b_i$, the average ratio is given by:

$$
\begin{aligned}
\text{Average}\left(\frac{a_1}{b_1}, \frac{a_2}{b_2}, \cdots, \frac{a_n}{b_n}\right) &= \frac{a_1 + a_2 + \cdots + a_n}{b_1 + b_2 + \cdots + b_n} \\
&= \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} a_i}{\frac{1}{n}\sum_{i=1}^{n} b_i} = \frac{\bar{a}}{\bar{b}}
\end{aligned}
$$

Mean of a Ratio: Example

- CPU utilization
  - Ratios cannot always be summarized by a geometric mean.
  - A geometric mean of utilizations is useless.

Measurement Duration	CPU Busy
1	45%
1	45%
1	45%
1	45%
100	20%
Sum	200
Mean	$\neq 200/5$ or 40%

$$\begin{aligned}
\text{Mean CPU utilization} &= \frac{\text{Sum of CPU busy times}}{\text{Sum of measurement durations}} \\
&= \frac{0.45 + 0.45 + 0.45 + 0.45 + 20}{1 + 1 + 1 + 1 + 100} \\
&= 21\%
\end{aligned}$$

Mean of a Ratio: Special Cases

1.1. If the denominator is a constant and the sum of numerator has a physical meaning, the arithmetic mean of the ratios can be used.

That is, if $b_i = b$ for all i's, then:

$$\text{Average}\left(\frac{a_1}{b}, \frac{a_2}{b}, \cdots, \frac{a_n}{b}\right)$$

$$= \frac{1}{n}\left(\frac{a_1}{b} + \frac{a_2}{b} + \cdots + \frac{a_n}{b}\right)$$

$$= \frac{\sum_{i=1}^{n} a_i}{nb}$$

- Example: mean resource utilization.

Mean of Ratio (Cont)

1.2. If the sum of the denominators has a physical meaning and the numerators are constant then a harmonic mean of the ratio should be used to summarize them.

That is, if $a_i = a$ for all i's, then:

$$\text{Average}\left(\frac{a}{b_1}, \frac{a}{b_2}, \cdots, \frac{a}{b_n}\right) = \frac{n}{\frac{b_1}{a} + \frac{b_2}{a} + \cdots + \frac{b_n}{a}} = \frac{na}{\sum_{i=1}^{n} b_i}$$

- Example: MIPS using the same benchmark

Mean of Ratios (Cont)

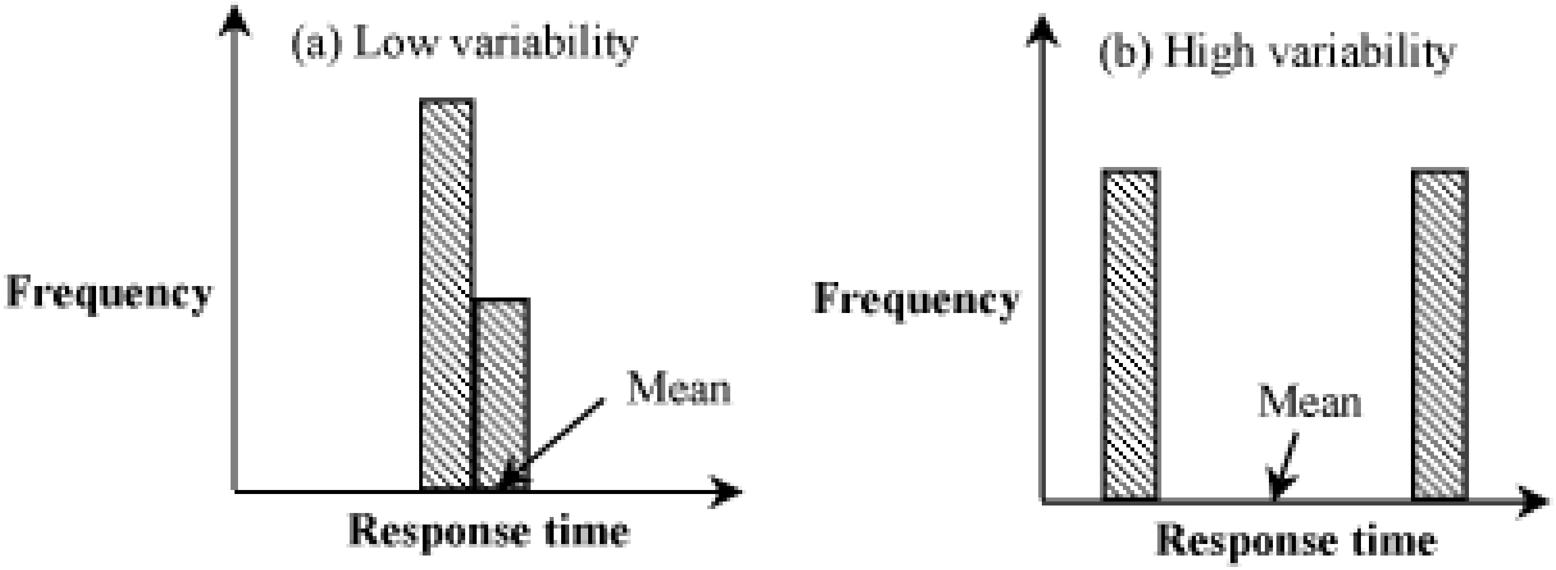
2. If the numerator and the denominator are expected to follow a multiplicative property such that $a_i = c\, b_i$, where c is approximately a constant that is being estimated, then c can be estimated by the geometric mean of a_i / b_i.
   - Example: Program Optimizer:
     $$a_i = cb_i$$
   - Where, b_i and a_i are the sizes before and after the program optimization and c is the effect of the optimization which is expected to be independent of the code size.

	Code Size		
Program	Before	After	Ratio
BubbleP	119	89	0.75
IntmmP	158	134	0.85
PermP	142	121	0.85
PuzzleP	8612	7579	0.88
QueenP	7133	7062	0.99
QuickP	184	112	0.61
SieveP	2908	2879	0.99
TowersP	433	307	0.71
Geometric Mean			0.79

Summarizing Variability

- "*Then there is the man who drowned crossing a stream with an average depth of six inches.*"

  - W. I. E. Gates



Indices of Dispersion

1. Range: Minimum and maximum of the values observed
2. Variance or standard deviation
3. 10- and 90- percentiles
4. Semi inter-quantile range
5. Mean absolute deviation

Range

- 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

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

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

- The divisor for s^2 is *n-1* and not *n*.
- This is because only *n-1* of the *n* differences $(x_i - \bar{x})$ are independent.
- Given *n-1* differences, n^{th} difference can be computed since the sum of all *n* differences must be zero.
- The number of independent terms in a sum is also called its **degrees of freedom**.

Variance (Cont)

- Variance is expressed in units which are square of the units of the observations.
  => It is preferable to use standard deviation.
- Ratio of standard deviation to the mean, or the **coefficient of variation** (COV), is even better because it takes the scale of measurement (unit of measurement) out of variability consideration.

Percentiles

- Specifying the 5-percentile and the 95-percentile of a variable has the same impact as specifying its minimum and maximum.
- It can be done for any variable, even for variables without bounds.
- When expressed as a fraction between 0 and 1 (instead of a percent), the percentiles are also called **quantiles**.
  => 0.9-quantile is the same as 90-percentile.
- **Fractile**= quantile
- The percentiles at multiples of 10% are called **deciles**. Thus, the first decile is 10-percentile, the second decile is 20-percentile, and so on.

Quartiles

- **Quartiles** divide the data into four parts at 25%, 50%, and 75%.
  => 25% of the observations are less than or equal to the first quartile Q_1,
  50% of the observations are less than or equal to the second quartile Q_2,
  and 75% are less than the third quartile Q_3.
- Notice that the second quartile Q_2 is also the median.
- The α-quantiles can be estimated by sorting the observations and taking the [(n-1)α+1]th element in the ordered set. Here, [.] is used to denote rounding to the nearest integer.
- For quantities exactly half way between two integers use the lower integer.

Semi Inter-Quartile Range

- **Inter-quartile range** = $Q_3 - Q_1$

- **Semi inter-quartile range** (SIQR)

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

Mean Absolute Deviation

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

- 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 affect is less
- 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%.

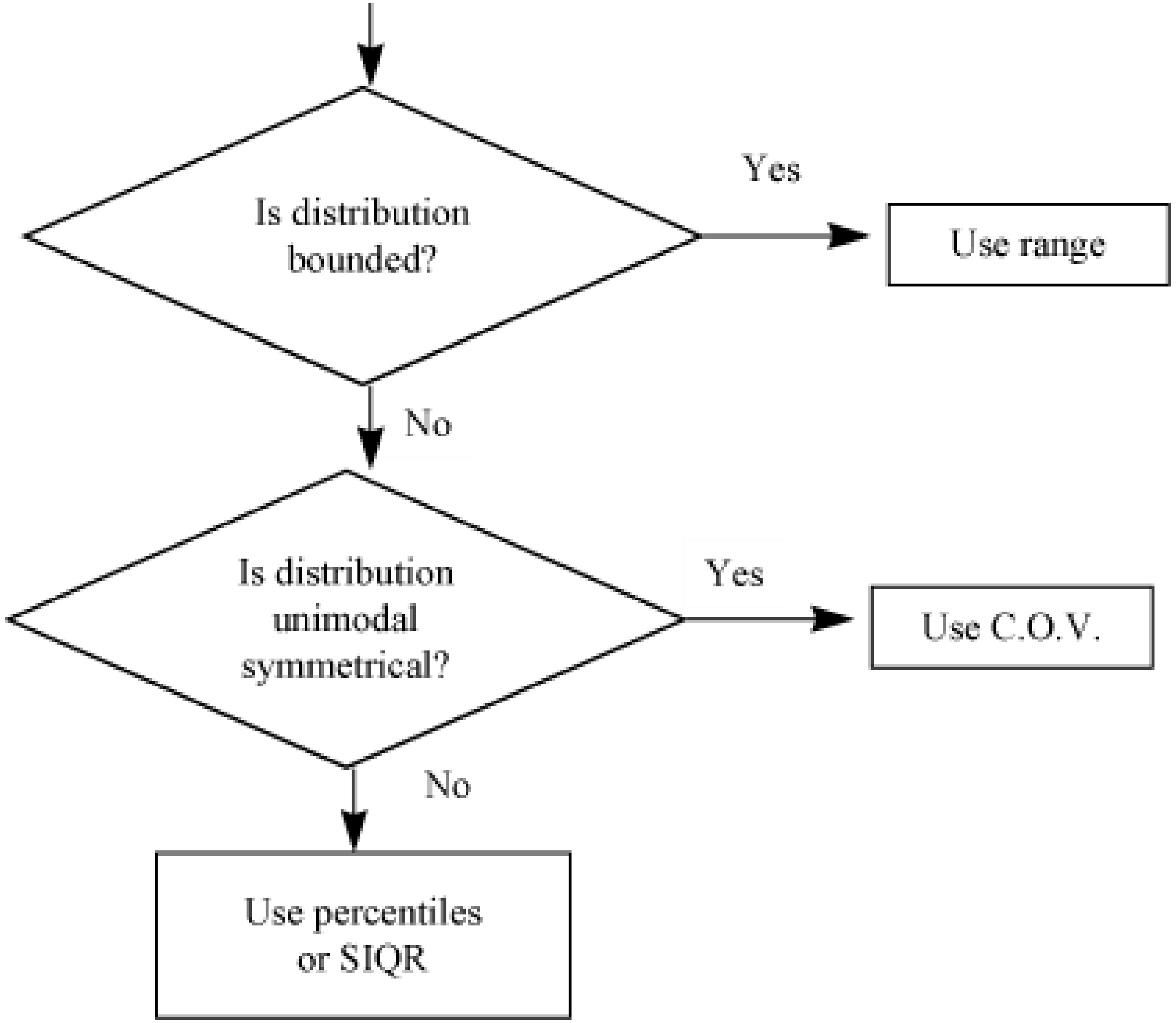
Measures of Variation: Example

In an experiment, which was repeated 32 times, the measured CPU time was found to be {3.1, 4.2, 2.8, 5.1, 2.8, 4.4, 5.6, 3.9, 3.9, 2.7, 4.1, 3.6, 3.1, 4.5, 3.8, 2.9, 3.4, 3.3, 2.8, 4.5, 4.9, 5.3, 1.9, 3.7, 3.2, 4.1, 5.1, 3.2, 3.9, 4.8, 5.9, 4.2}.

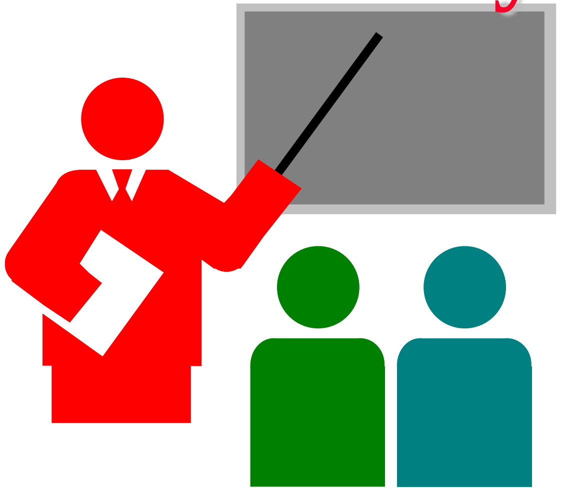
- The sorted set is {1.9, 2.7, 2.8, 2.8, 2.8, 2.9, 3.1, 3.1, 3.2, 3.2, 3.3, 3.4, 3.6, 3.7, 3.8, 3.9, 3.9, 3.9, 4.1, 4.1, 4.2, 4.2, 4.4, 4.5, 4.5, 4.8, 4.9, 5.1, 5.1, 5.3, 5.6, 5.9}.
- 10-percentile = [1+(31)(0.10) = 4th element = 2.8
- 90-percentile = [1+(31)(0.90)] = 29th element = 5.1
- First quartile Q_1 = [1+(31)(0.25)] = 9th element = 3.2
- Median Q_2 = [1+(31)(0.50)] = 16th element = 3.9
- Third quartile Q_3 = [1+(31)(0.75)] = 24th element = 4.5

$$\mathrm{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{4.5 - 3.2}{2} = 0.65$$

Selecting the Index of Dispersion



Summary



- Summarizing Data by a Single Number: Mean, Median, and Mode, Arithmetic, Geometric, Harmonic Means
- Mean of A Ratio
- Summarizing Variability: Range, Variance, percentiles, Quartiles