CS 147:
Computer Systems Performance Analysis
Summarizing Data

2015-06-15

CS147

CS 147:
Computer Systems Performance Analysis
Summarizing Data

Overview

Overview

"Standard" Indices of Central Tendency
Definitions
Characteristics
Selecting an Index
Other Indices
Geometric Mean
Harmonic Mean
Dealing with Ratios
Case 1: Two Physical Meanings
Case 1a: Constant Denominator
Case 1b: Constant Numerator
Case 2: Multiplicative Relationship

Summarizing Data With a Single Number

- Most condensed form of presentation of set of data
- Usually called the **average**
  - Average isn't necessarily the mean
- Must be representative of a major part of the data set

Summarizing Data With a Single Number

- Most condensed form of presentation of set of data
- Usually called the **average**
  - Average isn't necessarily the mean
- Must be representative of a major part of the data set

Indices of Central Tendency

- Mean
- Median
- Mode
- All specify *center of location* of distribution of observations in sample

Indices of Central Tendency

- Mean
- Median
- Mode
- All specify *center of location* of distribution of observations in sample

Sample Mean

- Take sum of all observations
- Divide by number of observations
- More affected by outliers than median or mode
- Mean is a linear property
  - Mean of sum is sum of means
  - Not true for median and mode

Sample Mean

- Take sum of all observations
- Divide by number of observations
- More affected by outliers than median or mode
- Mean is a linear property
  - Mean of sum is sum of means
  - Not true for median and mode

Sample Median

- Sort observations
- Take observation in middle of series
  - If even number, split the difference
- More resistant to outliers
  - But not all points given "equal weight"

Sample Median

- Sort observations
- Take observation in middle of series
  - If even number, split the difference
- More resistant to outliers
  - But not all points given "equal weight"

Sample Mode

- Plot histogram of observations
  - Using existing categories
  - Or dividing ranges into buckets
  - Or using kernel density estimation
- Choose midpoint of bucket where histogram peaks
  - For categorical variables, the most frequently occurring
- Effectively ignores much of the sample

Sample Mode

- Plot histogram of observations
  - Using existing categories
  - Or dividing ranges into buckets
  - Or using kernel density estimation
- Choose midpoint of bucket where histogram peaks
  - For categorical variables, the most frequently occurring
- Effectively ignores much of the sample

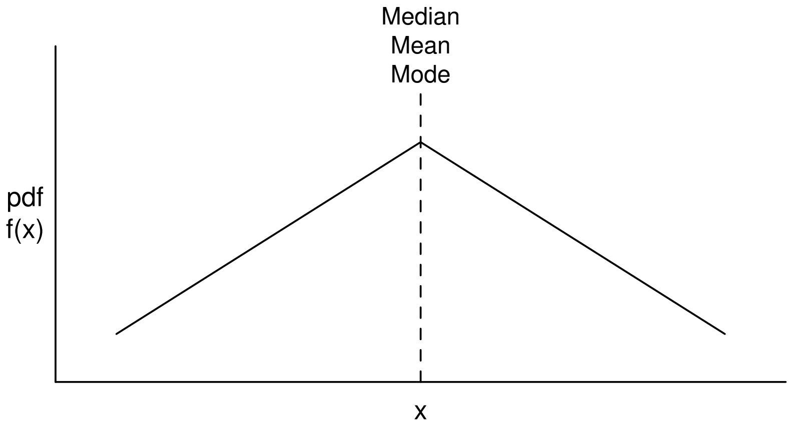
Characteristics of Mean, Median, and Mode

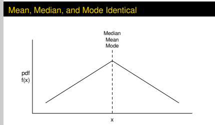
- Mean and median always exist and are unique
- Mode may or may not exist
  - If there is a mode, may be more than one
- Mean, median and mode may be identical
  - Or may all be different
  - Or some may be the same

Characteristics of Mean, Median, and Mode

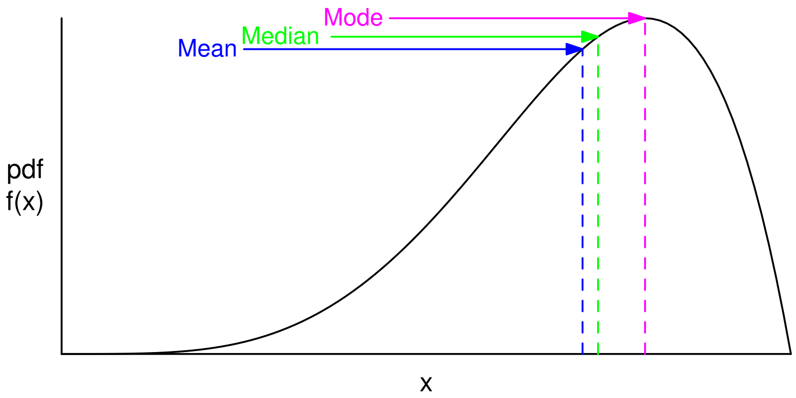
- Mean and median always exist and are unique
- Mode may or may not exist
  - If there is a mode, may be more than one
- Mean, median and mode may be identical
  - Or may all be different
  - Or some may be the same

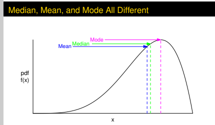
Mean, Median, and Mode Identical





Median, Mean, and Mode All Different





So, Which Should I Use?

- Depends on characteristics of the metric
- If data is categorical, use mode
- If a total of all observations makes sense, use mean
- If not (e.g., ratios), and distribution is skewed, use median
- Otherwise, use mean
  . . . but think about what you’re choosing

So, Which Should I Use?

- Depends on characteristics of the metric
- If data is categorical, use mode
- If a total of all observations makes sense, use mean
- If not (e.g., ratios), and distribution is skewed, use median
- Otherwise, use mean
  . . . but think about what you’re choosing

Some Examples

- Most-used resource in system

Some Examples

- Most-used resource in system

Some Examples

- Most-used resource in system
  - Mode
- Interarrival times

Some Examples

- Most-used resource in system
  - Mode
- Interarrival times

Some Examples

- Most-used resource in system
  - Mode
- Interarrival times
  - Mean
- Load

Some Examples

- Most-used resource in system
  - Mode
- Interarrival times
  - Mean
- Load

Some Examples

- Most-used resource in system
  - Mode
- Interarrival times
  - Mean
- Load
  - Median

Some Examples

- Most-used resource in system
  - Mode
- Interarrival times
  - Mean
- Load
  - Median

Don't Always Use the Mean

- Means are often overused and misused
  - Means of significantly different values
  - Means of highly skewed distributions
  - Multiplying means to get mean of a product
    - Only works for independent variables
  - Errors in taking ratios of means
  - Means of categorical variables

Don't Always Use the Mean

- Means are often overused and misused
  - Means of significantly different values
  - Means of highly skewed distributions
  - Multiplying means to get mean of a product
    - Only works for independent variables
  - Errors in taking ratios of means
  - Means of categorical variables

Geometric Means

- An alternative to the arithmetic mean

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

- Use geometric mean if product of observations makes sense

Geometric Means

- An alternative to the arithmetic mean

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

- Use geometric mean if product of observations makes sense

Good Places To Use Geometric Mean

- Layered architectures
- Performance improvements over successive versions
- Average error rate on multihop network path
- Year-to-year interest rates

2015-06-15

CS147
└Other Indices
  └Geometric Mean
    └Good Places To Use Geometric Mean

Good Places To Use Geometric Mean

- Layered architectures
- Performance improvements over successive versions
- Average error rate on multihop network path
- Year-to-year interest rates

Harmonic Mean

- Harmonic mean of sample $\{x_1, x_2, \ldots, x_n\}$ is

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

- Use when arithmetic mean of $1/x_i$ is sensible

Harmonic Mean

- Harmonic mean of sample $\{x_1, x_2, \ldots, x_n\}$ is

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

- Use when arithmetic mean of $1/x_i$ is sensible

Example of Using Harmonic Mean

- When working with MIPS numbers from a single benchmark
  - Since MIPS calculated by dividing constant number of instructions by elapsed time

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

- Not valid if different m's (e.g., different benchmarks for each observation)

Example of Using Harmonic Mean

- When working with MIPS numbers from a single benchmark
  - Since MIPS calculated by dividing constant number of instructions by elapsed time

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

- Not valid if different m's (e.g., different benchmarks for each observation)

Means of Ratios

- Given *n* ratios, how do you summarize them?
- Can't always just use harmonic mean
  - Or similar simple method
- Consider numerators and denominators

Means of Ratios

- Given *n* ratios, how do you summarize them?
- Can't always just use harmonic mean
  - Or similar simple method
- Consider numerators and denominators

Considering Mean of Ratios: Case 1

- Both numerator and denominator have physical meaning
- Then the average of the ratios is the ratio of the averages

Considering Mean of Ratios: Case 1

- Both numerator and denominator have physical meaning
- Then the average of the ratios is the ratio of the averages

Example: CPU Utilizations

Measurement Duration	CPU Busy (%)
1	40
1	50
1	40
1	50
100	20
Sum	200%
Mean?	

Example: CPU Utilizations

Measurement Duration	CPU Busy (%)
1	40
1	50
1	40
1	50
100	20
Sum	200%
Mean?	

Example: CPU Utilizations

Measurement Duration	CPU Busy (%)
1	40
1	50
1	40
1	50
100	20
Sum	200%
Mean?	Not 40%

Example: CPU Utilizations

Measurement Duration	CPU Busy (%)
1	40
1	50
1	40
1	50
100	20
Sum	200%
Mean?	Not 40%

Example: CPU Utilizations

Measurement Duration	CPU Busy (%)
1	40
1	50
1	40
1	50
100	20
Sum	200%
Mean?	Nor 1.92%!

Example: CPU Utilizations

Measurement Duration	CPU Busy (%)
1	40
1	50
1	40
1	50
100	20
Sum	200%
Mean?	Nor 1.92%!

Properly Calculating Mean For CPU Utilization

- Why not 40%?
- Because CPU-busy percentages are ratios
  - So their denominators aren't comparable
- The duration-100 observation must be weighted more heavily than the duration-1 ones

Properly Calculating Mean For CPU Utilization
- Why not 40%?
- Because CPU-busy percentages are ratios
  - So their denominators aren't comparable
- The duration-100 observation must be weighted more heavily than the duration-1 ones

So What Is the Proper Average?

- Go back to the original ratios:

$$\begin{aligned} \text{Mean CPU Utilization} &= \frac{0.40 + 0.50 + 0.40 + 0.50 + 20}{1 + 1 + 1 + 1 + 100} \\ &= 21\% \end{aligned}$$

So What Is the Proper Average?

- Go back to the original ratios:

$$\begin{aligned} \text{Mean CPU Utilization} &= \frac{0.40 + 0.50 + 0.40 + 0.50 + 20}{1 + 1 + 1 + 1 + 100} \\ &= 21\% \end{aligned}$$

Considering Mean of Ratios: Case 1a

- Sum of numerators has physical meaning
- Denominator is a constant
- Take arithmetic mean of the ratios to get overall mean

Considering Mean of Ratios: Case 1a

- Sum of numerators has physical meaning
- Denominator is a constant
- Take arithmetic mean of the ratios to get overall mean

For Example,

- What if we calculated CPU utilization from last example using only the four duration-1 measurements?
- Then the average is

$$\frac{1}{4}\left(\frac{.40}{1} + \frac{.50}{1} + \frac{.40}{1} + \frac{.50}{1}\right) = 0.45$$

For Example,

- What if we calculated CPU utilization from last example using only the four duration-1 measurements?
- Then the average is

$$\frac{1}{4}\left(\frac{.40}{1} + \frac{.50}{1} + \frac{.40}{1} + \frac{.50}{1}\right) = 0.45$$

Considering Mean of Ratios: Case 1b

- Sum of denominators has a physical meaning
- Numerator is a constant
- Take harmonic mean of the ratios

Considering Mean of Ratios: Case 1b

- Sum of denominators has a physical meaning
- Numerator is a constant
- Take harmonic mean of the ratios

Considering Mean of Ratios: Case 2

- Numerator and denominator are expected to have a multiplicative, near-constant property

$$a_i = cb_i$$

- Estimate c with geometric mean of a_i / b_i

Considering Mean of Ratios: Case 2

- Numerator and denominator are expected to have a multiplicative, near-constant property

$a_i = cb_i$

- Estimate c with geometric mean of a_i / b_i

Example for Case 2

- An optimizer reduces the size of code
- What is the average reduction in size, based on its observed performance on several different programs?
- Proper metric is percent reduction in size
- And we're looking for a constant *c* as the average reduction

Example for Case 2
- An optimizer reduces the size of code
- What is the average reduction in size, based on its observed performance on several different programs?
- Proper metric is percent reduction in size
- And we're looking for a constant *c* as the average reduction

Program Optimizer Example, Continued

	Code Size		
Program	Before	After	Ratio
BubbleP	119	89	.75
IntmmP	158	134	.85
PermP	142	121	.85
PuzzleP	8612	7579	.88
QueenP	7133	7062	.99
QuickP	184	112	.61
SieveP	2908	2879	.99
TowersP	433	307	.71

Program Optimizer Example, Continued

	Code Size		
Program	Before	After	Ratio
BubbleP	119	89	.75
IntmmP	158	134	.85
PermP	142	121	.85
PuzzleP	8612	7579	.88
QueenP	7133	7062	.99
QuickP	184	112	.61
SieveP	2908	2879	.99
TowersP	433	307	.71

Why Not Use Ratio of Sums?

- Why not add up pre- sizes and post-optimized sizes and take the ratio?
  - Benchmarks of non-comparable size
  - No indication of importance of each benchmark in overall code mix
  - When looking for constant factor, not the best method

Why Not Use Ratio of Sums?

- Why not add up pre- sizes and post-optimized sizes and take the ratio?
  - Benchmarks of non-comparable size
  - No indication of importance of each benchmark in overall code mix
  - When looking for constant factor, not the best method

So Use the Geometric Mean

- Multiply the ratios from the 8 benchmarks
- Then take the 1/8 power of the result

$$\begin{aligned} \ddot{x} &= (.75 \times .85 \times .85 \times .88 \times .99 \times .61 \times .99 \times .71)^{1/8} \\ &= .82 \end{aligned}$$

So Use the Geometric Mean
- Multiply the ratios from the 8 benchmarks
- Then take the 1/8 power of the result
$\ddot{x} = (.75 \times .85 \times .85 \times .88 \times .99 \times .61 \times .99 \times .71)^{1/8} = .82$