Summarizing Measured Data

Hongwei Zhang

http://www.cs.wayne.edu/~hzhang

The object of statistics is to discover methods of condensing information concerning large groups of allied facts into brief and compendious expressions sutable for discussions.

--- Francis Galton

Acknowledgement: this lecture is partially based on the slides of Dr. Raj Jain.

Outline

- Some Probability and Statistics Concepts
- Summarizing Data by a Single Number
- Summarizing Variability
- Determining Distribution of Data

Outline

- Some Probability and Statistics Concepts
- Summarizing Data by a Single Number
- Summarizing Variability
- Determining Distribution of Data

Some Probability and Statistics Concepts

- Probability model
- Random variable
- Expectation
- Quantiles, median, mode
- Covariance, correlation coefficient

Some Probability and Statistics Concepts

- Probability model
- Random variable
- Expectation
- Quantiles, median, mode
- Covariance, correlation coefficient

Probability model/experiment

- Sample space (S)
  - set of possible outcomes/sample-points
- Set of events (F)
  - An event is a subset of the sample space
- Rule of assigning probabilities to events (P)

- Q: what is the probability model of two independent flips of a fair coin?

Axioms of probability

- Normalization
  - $P(S) = 1$
- Monotonicity
  - $B \subset A \Rightarrow P(B) <= P(A)$
- Additivity (disjoint events)
  - If $A \cap B = \phi$, $P(A \cup B) = P(A) + P(B)$
- Note: the set of events should form a sigma(δ)-field so as to ensure the existence of a probability function P satisfying the above axioms

Field & sigma-field

- A set of sets is a “Field” if
  - it is closed under union, intersection, and complement; and
  - empty-set and universal set are elements of the set
- A field is a “δ-field” if
  - it is closed under any countable set of unions, intersections, and their combinations

Conditional probability

- For any two events A and B, with P(B) > 0, the *conditional probability* of A, conditional on B, is

$$P(A|B) = P(AB)/P(B)$$

- Two events A and B are *independent* if P(AB) = P(A)P(B)
  - For P(B) > 0, this is equivalent to P(A|B) = P(A)
- Two events A and B are *conditionally independent* given C if

$$P(AB|C) = P(A|C)*P(B|C)$$

or equivalently

$$P(A|B \cap C) = P(A|C)$$

Q: examples of conditional probability, independent events, conditionally independent events?

Some Probability and Statistics Concepts

- Probability model
- Random variable
- Expectation
- Quantiles, median, mode
- Covariance, correlation coefficient

Random variables

- A random variable X is a function: S -> R
- Given X and a real number x, there is an event $X \leq x$, and

$$P(X \leq x) = P(\{w \in S: X(w) \leq x\})$$

- $P(X \leq x)$ is the *distribution function* of X, and is usually denoted as $F_X(x)$
  - Monotonically nondecreasing
- Complex and vector random variables: map sample to a set of finite complex numbers or vectors in some finite dimensional vector space

Random variables (contd.)

- If $F_X(x)$ has a derivative $f_X(x)$, we call $f_X(x)$ the *probability density* of X, i.e.,

$$f_X(x) = d(F_X(x))/dx$$

- If $f_X(x)$ exists and is finite for all x, we say X is a *continuous* r.v.

- If X has only a countable number of possible outcomes, $x_1, x_2, \ldots,$
  - we say X is a *discrete* r.v.
  - the probability of each outcome x_i, $\{P_X(x_i): i>=1\}$, is called the *probability mass function* (PMF) of X

Random variables (contd.)

- *Joint distribution function* of r.v. X1, X2, ..., Xn

$$F_{X1,\ldots,Xn}(x1, x2, \ldots, xn) = P(X1 \leq x1, X2 \leq x2, \ldots, Xn \leq xn)$$

- Then
  - $F_{Xi}(xi) = F_{X1,\ldots,Xn}(\infty, \ldots, \infty, xi, \infty, \ldots, \infty)$
  - *Joint probability density* $f_{X1,\ldots,Xn}(x1, x2, \ldots, xn)$ is

$$\frac{\partial^n F(x1, x2, \ldots, xn)}{\partial x1 \partial x2 \ldots \partial xn}$$

Random variables (contd.)

- R.v. X1, X2, ..., Xn are *independent*, if, for all x1, ..., xn

$$F(x1,...,xn) = \prod_{i=1}^{n} P(Xi \leq xi)$$

- If the density or mass function exists, the above formula is equivalent to a product form for the density or mass function

- Note: pairwise independence does not imply that the entire set is independent (example?)

Example of “pairwise independence” vs. “set independence”

Suppose X, Y, and Z have the following joint probability distribution:

$$(X,Y,Z) = \begin{Bmatrix} (0,0,0) & \text{with probability } 1/4, \\ (0,1,1) & \text{with probability } 1/4, \\ (1,0,1) & \text{with probability } 1/4, \\ (1,1,0) & \text{with probability } 1/4. \end{Bmatrix}$$

Then

- X and Y are independent, and
- X and Z are independent, and
- Y and Z are independent, but
- X, Y, and Z are ***not*** independent (why?)

Some Probability and Statistics Concepts

- Probability model
- Random variable
- Expectation
- Quantiles, median, mode
- Covariance, correlation coefficient

Expectations

- The *expected value* (or the mean) of a r.v. X

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx; \quad or \quad E[X] = \sum_x x P_X(x)$$

For simplicity, write as

$$E[X] = \overline{X} = \int_{-\infty}^{\infty} x dF_X(x)$$

- For non-negative random variables,

$$E[X] = \int_0^{\infty} P(X \geq x)dx, \text{ if X is continuous}$$

or

$$E[X] = \sum_{x=1}^{\infty} P(X \geq x), \text{ if X is integer-valued and discrete}$$

Expectations (contd.)

- If Y = g(X), then

$$E[Y] = \int_{-\infty}^{\infty} y dF_Y(y) = \int_{-\infty}^{\infty} g(x) dF_X(x)$$

- Moments $E[X^n]$, central moments $E[(X-E[X])^n]$
  - VAR(X) (or σ_X^2) = $E[(X-E[X])^2] = E[X^2] - (E[x])^2$
  - Standard deviation σ_X

Z=X+Y …

- If X and Y are independent,

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(z-y)dF_Y(y) = \int_{-\infty}^{\infty} F_Y(z-x)dF_X(x) \quad \text{(convolution)}$$

  And if X and Y both have densities,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)\, f_Y(y)dy = \int_{-\infty}^{\infty} f_Y(z-x)\, f_X(x)dx$$

- Let Sn=X1+X2+…+Xn,
  - Whether or not X1, X2, …, Xn are independent

$$E[Sn] = E[X1] + E[X2] + \ldots + E[Xn] \quad (?)$$

  - If X1, X2, …, Xn are independent/uncorrelated

$$\sigma_{Sn}^2 = \sum_{i=1}^{n} \sigma_{Xi}^2 \quad (?)$$

- If X1, X2, …, Xn are independent

$$E[\prod_{i=1}^{n} Xi] = \prod_{i=1}^{n} E[Xi] \quad (?)$$

Example: overhead in bit-stuffing

- Bit-stuffing in data-link framing: 011111 -> 0111110
- Q: expected number of inserted bits in a string of length n?

Solution

- Xi = 1, if the insertion occurred after the i-th bit

  0 , otherwise

- Note: E[Xi] = 0 for $i \leq 5$, and E[Xi] = 2^{-6} otherwise
- Number of bits inserted (Sn) = X1+x2+...+X6+...+Xn
- Thus, E[Sn] = (n-5) 2^{-6}

Some Probability and Statistics Concepts

- Probability model
- Random variable
- Expectation
- Quantiles, median, mode
- Covariance, correlation coefficient

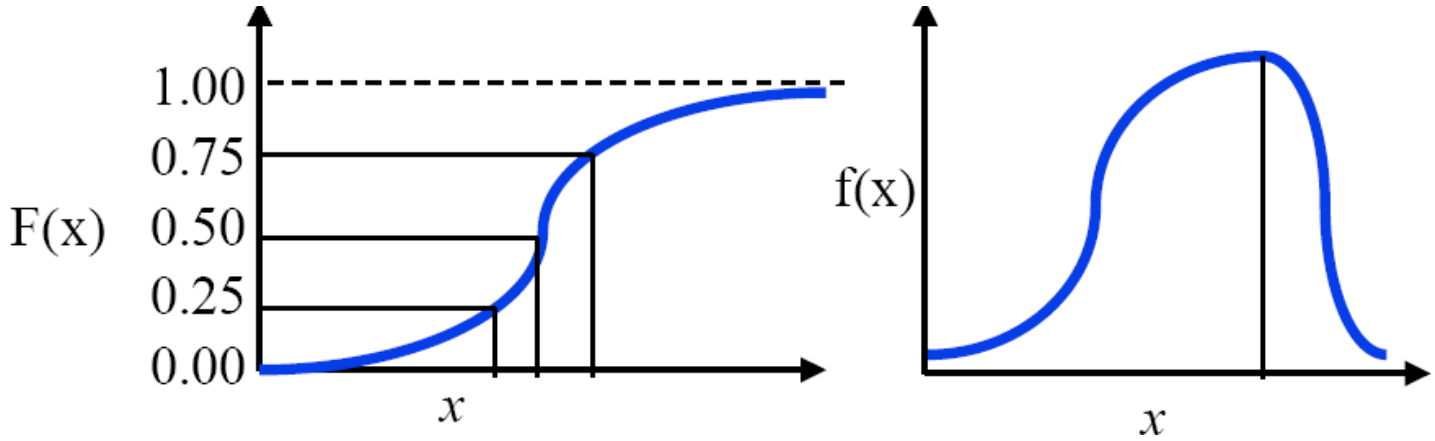
Quantiles, median, mode

- **Quantile**: The *x* value at which the CDF takes a value α is called the α-quantile or 100α-percentile. It is denoted by x_{α}.

$$P(x \leq x_\alpha) = F(x_\alpha) = \alpha$$

- **Median**: the 50-percentile or (0.5-quantile) of a random variable

- **Mode**: The most likely value, that is, x_i that has the highest probability p_i, or the *x* at which pdf is maximum, is called mode of *x*



Some Probability and Statistics Concepts

- Probability model
- Random variable
- Expectation
- Quantiles, median, mode
- Covariance, correlation coefficient

Covariance and correlation

- **Covariance**:

$$\begin{aligned} Cov(x,y) &= \sigma_{xy}^2 = E[(x-\mu_x)(y-\mu_y)] \\ &= E(xy) - E(x)E(y) \end{aligned}$$

  - For independent variables, the covariance is zero because

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

  - 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

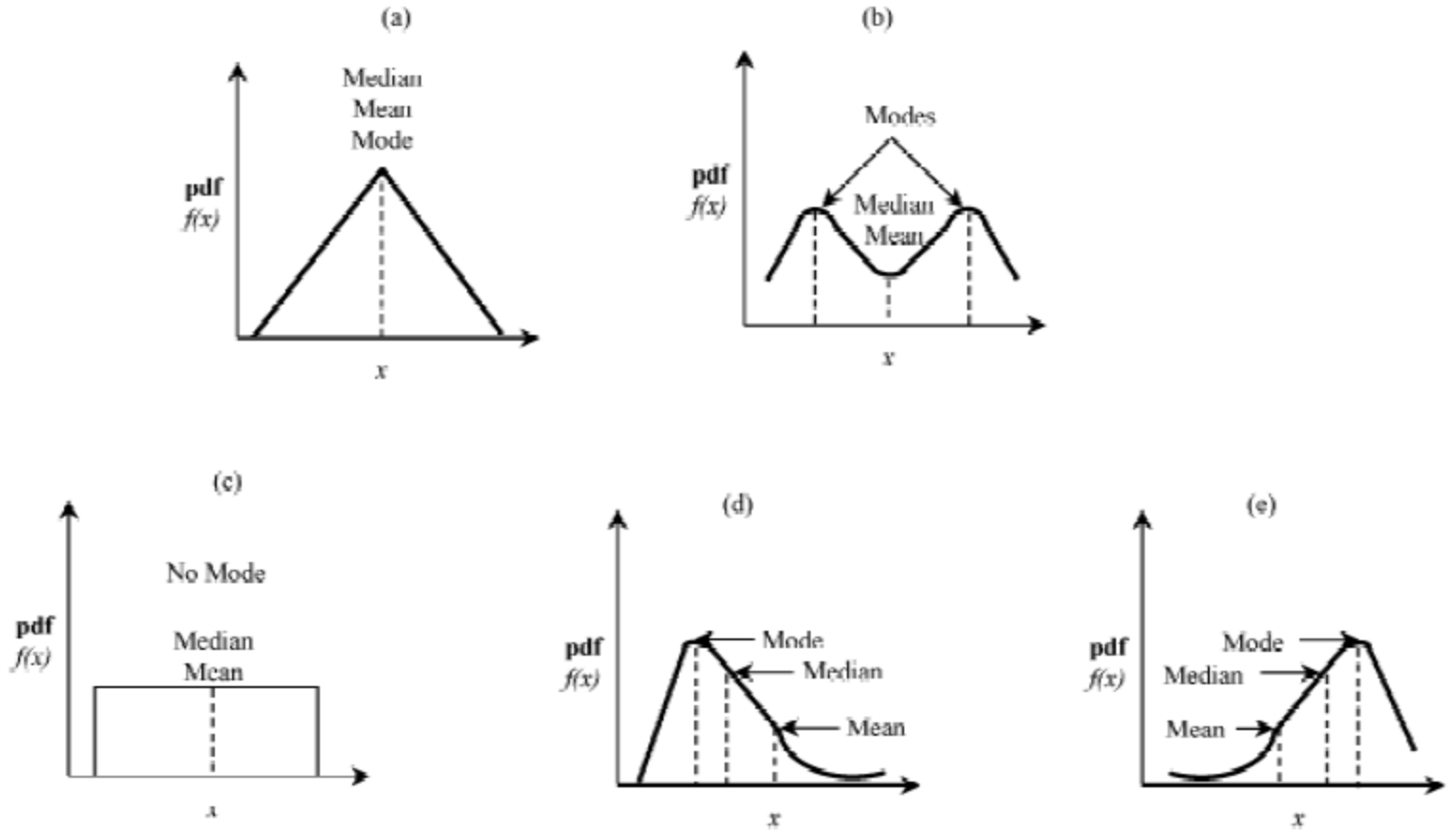
Outline

- Some Probability and Statistics Concepts
- Summarizing Data by a Single Number
- Summarizing Variability
- Determining Distribution of Data

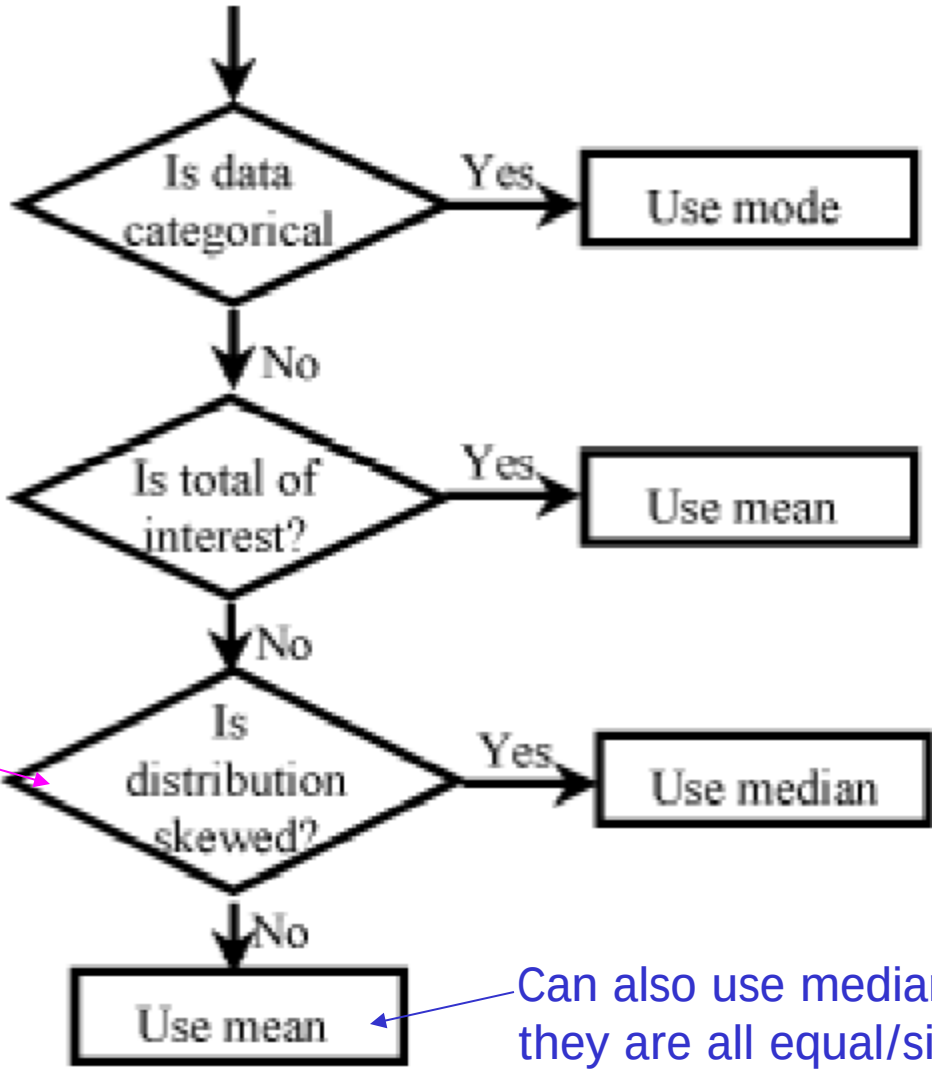
Summarizing data by a single number?

- **Indices of central tendencies**: Mean, Median, Mode
- **Sample Mean:** taking the sum of all observations and dividing this sum by the number of observations in the sample
- **Sample Median:** sorting the observations in an increasing order and taking the observation that is in the middle of the series
  - If # of observations is even, the mean of the middle two values is used as a median
- **Sample Mode:** 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 may not exist and may not be unique

Mean, median, mode: relationships



Mean, median, mode: which one to use?



A simple way of judging skewness: ratio of max. to min.

Can also use median, or mode: they are all equal/similar

Examples

- Most used resource in a system
  - Resources are categorical and hence mode must be used
- Interarrival time
  - Total time is of interest, and mean is the proper choice
- Load on a Computer
  - Median is preferable due to a highly skewed distribution
- Average Configuration
  - Medians of number of 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
  - E.g., packet processing time: (10+1000)/2 = 505 => correct operation, but *useless* (i.e., does not convey useful information)
- Using mean without regard to the skewness of distribution
  - E.g., system response time for 5 days

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

Check whether it is an outlier

Misuses of means (contd.)

- Multiplying means to get the mean of a product
  - $E[XY] \neq E[X]E[Y]$
  - E.g., 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.

Misuses of means (contd.)

- Taking a mean of a ratio with different bases

System A:

Test	Total	Pass	% Pass
1	300	60	20%
2	50	2	4%

System B:

Test	Total	Pass	% Pass
1	32	8	25%
2	500	40	8%

  - Had we taken mean of ratios, B is “better” than A
  - But in fact: A is better than B

Geometric mean

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

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

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

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

Geometric means 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
  - => choice of the base does not change the conclusion
  - It is because of this property that sometimes geometric mean is recommended for ratios
  - However, if the geometric mean of the numerator or denominator do not have any physical meaning, the geometric mean of their ratio is meaningless as well

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 $1/x_i$ (e.g., elapsed time of a benchmark on a processor)
- Example
  - 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

Harmonic mean (contd.)

- The average MIPS rate for the processor is:

$$\ddot{x} = \frac{n}{\frac{1}{m/t_1} + \frac{1}{m/t_2} + \cdots + \frac{1}{m/t_n}} = \frac{m}{\frac{1}{n}(t_1 + t_2 + \cdots + t_n)}$$

- However, if x_is represent the MIPS rate for n different benchmarks so that i-th benchmark has m_i million instructions, then harmonic mean of n ratios m_i/t_i cannot be used since the sum of the t_i/m_i does not have any physical meaning

- Instead, as shown later, the quantity $\Sigma m_i/\Sigma t_i$ is a preferred average MIPS rate

Weighted harmonic mean (contd.)

- The weighted harmonic mean is defined as follows:

$$\ddot{x} = \frac{1}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_n}{x_n}}$$

where, w_i's are weights which add up to one:

$$w_1 + w_2 + \cdots + w_n = 1$$

  - All weights equal => Harmonic, I.e., $w_i = 1/n$.

- In case of MIPS rate, if the weights are proportional to the size of the benchmark: $w_i = \frac{m_i}{m_1 + m_2 + \cdots + m_n}$
- Weighted harmonic mean would be:

$$\ddot{x} = \frac{m_1 + m_2 + \cdots + m_n}{t_1 + t_2 + \cdots + t_n}$$

Mean of a ratio

- Rule #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
  - E.g., if xi=ai/bi, the average ratio is given by

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

Example

- CPU utilization

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

Example (contd.)

$$\text{Mean CPU utilization} = \frac{\text{Sum of CPU busy times}}{\text{Sum of measurement durations}}$$

$$= \frac{0.45 + 0.45 + 0.45 + 0.45 + 20}{1 + 1 + 1 + 1 + 100}$$

$$= 21\%$$

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

Mean of a Ratio: Special Cases

- Case #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. I.e., if bi=b for all i's, then:

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

- Example: mean resource utilization.

Special cases (contd.)

- Case #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 ai=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 a ratio (contd.)

- Rule #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
  - 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
  - Summation of code size may not make sense

	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

Outline

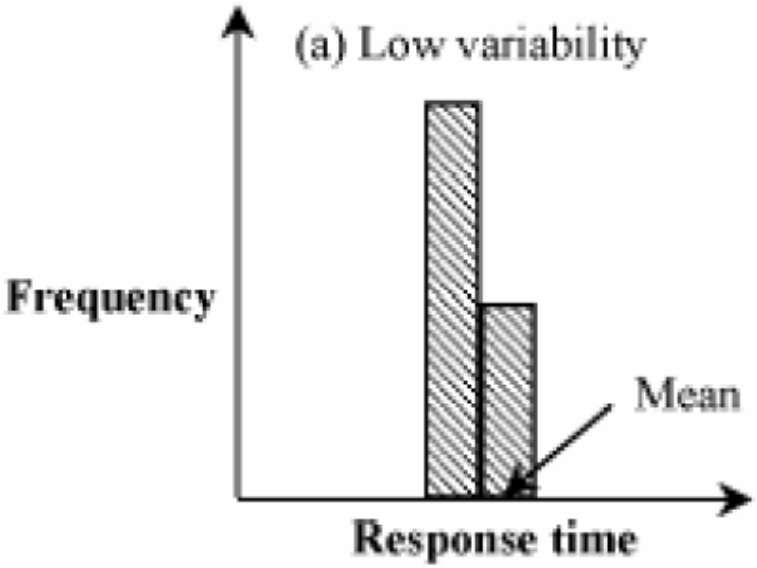
- Some Probability and Statistics Concepts
- Summarizing Data by a Single Number
- Summarizing Variability
- Determining Distribution of Data

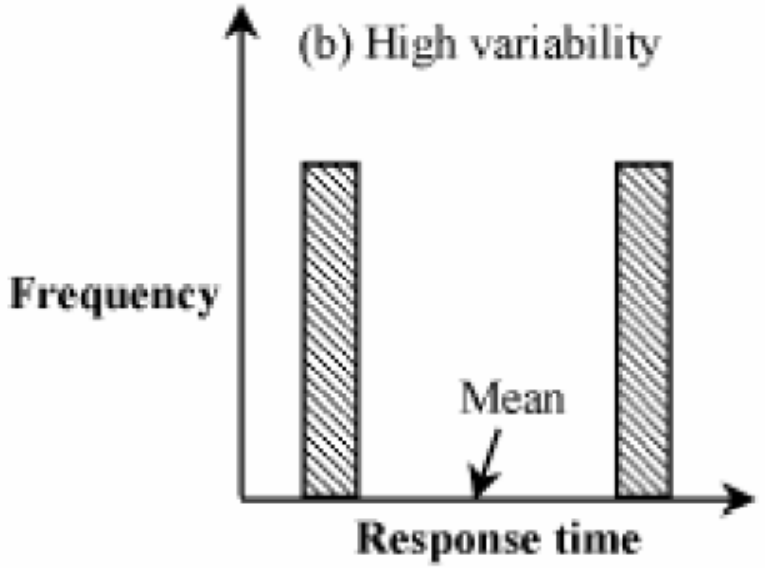
Summarizing variability

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

?

- W. I. E. Gates





Indices of dispersion?

- Range
  - Minimum and maximum of the values observed
- 10- and 90- percentiles
- Semi inter-quantile range (SIQR)
- Variance, standard deviation, or coefficient of variation
- 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*

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**
  - E.g., 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

Percentiles (contd.)

- **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 *Q1*, 50% of the observations are less than or equal to the second quartile *Q2*, and 75% are less than the third quartile *Q3*
  - Notice that the second quartile *Q2* is also the median
- The α-quantiles can be estimated by sorting the observations and taking the $[(n-1)\alpha+1]$-th element in the ordered set
  - [.] is used to denote rounding to the *nearest* integer
  - For quantities exactly half way between two integers use the lower integer

Semi-interquartile range

- **Interquartile range** = Q_3 - Q_1
- **Semi interquartile range** (SIQR)

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

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 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 (contd.)

- 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 (or unit) of measurement out of variability consideration

Mean absolute deviation

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

- Compared with variance/standard deviation/C.O.V., no multiplication or square root is required

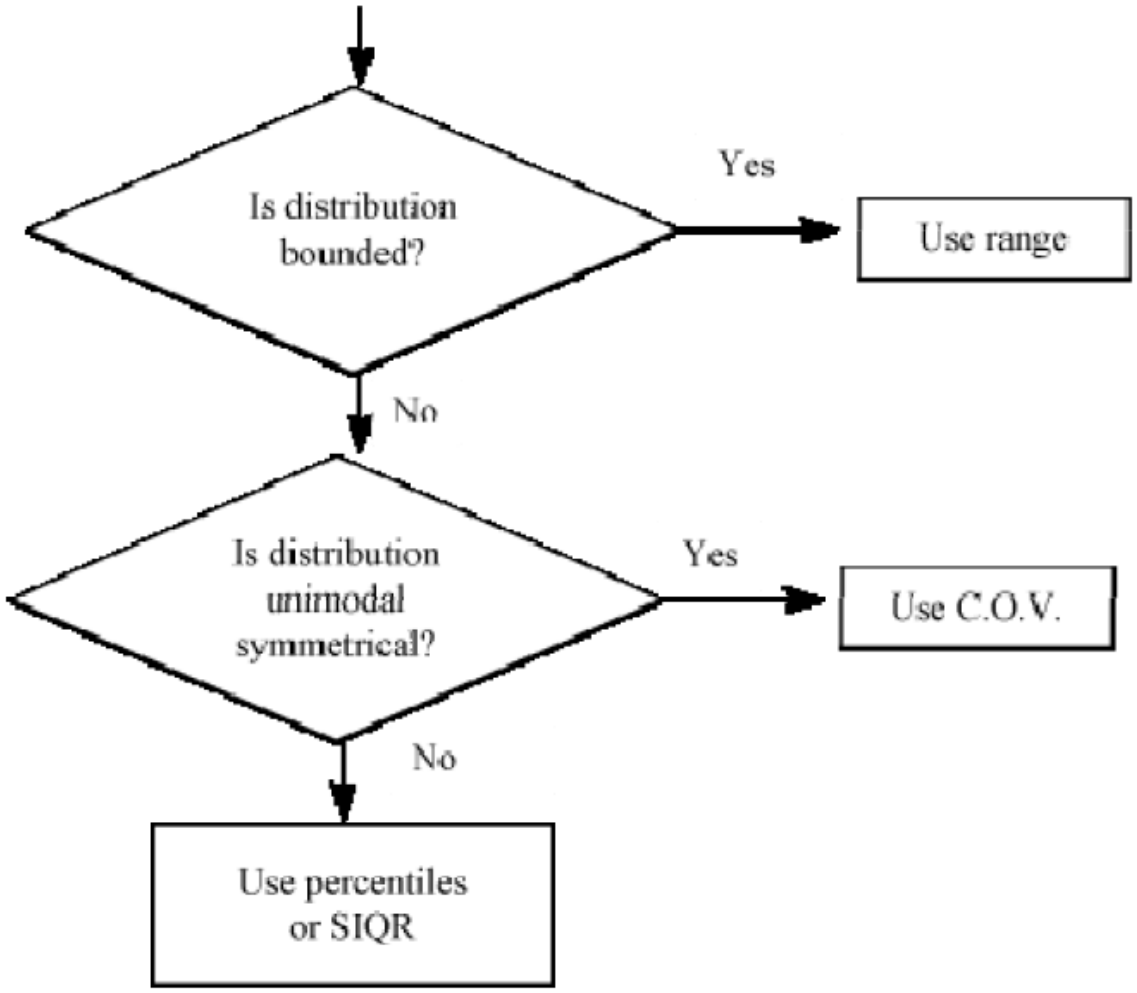
Comparison of variation measures

- Resistance to outliers
  - 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%

Example

- In an experiment, which was repeated 32 times, the measured packet-processing-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 *Q1* = [1+(31)(0.25)] = 9th element = 3.2
- Median *Q2* = [1+(31)(0.50)] = 16th element = 3.9
- Third quartile *Q1* = [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



Selecting the index of dispersion (contd.)

- The decision rules given above are not hard and fast
- E.g.,
  - Network designed for average traffic is grossly under-designed. The network load is highly skewed => Networks are designed to carry 95 to 99-percentile of the observed load levels =>Dispersion of the load should be specified via range or percentiles
  - Power supplies are similarly designed to sustain peak demand rather than average demand

Outline

- Some Probability and Statistics Concepts
- Summarizing Data by a Single Number
- Summarizing Variability
- Determining Distribution of Data

Determining distribution of data?

- The simplest way to determine the distribution is to plot a *histogram*
- Count observations that fall into each cell or bucket
- The key problem is determining the cell size
  - Small cells => large variation in the number of observations per cell
  - Large cells => details of the distribution are completely lost
  - It is possible to reach very different conclusions about the distribution shape
- *One guideline*: if any cell has less than five observations, the cell size should be increased or a variable cell histogram should be used

Determining distribution of data (contd.)

- A better technique (especially for *small samples*) is to use a quantile-quantile plot (Q-Q plot)
- y_i is the observed q_ith quantile

  x_i = theoretical q_ith quantile

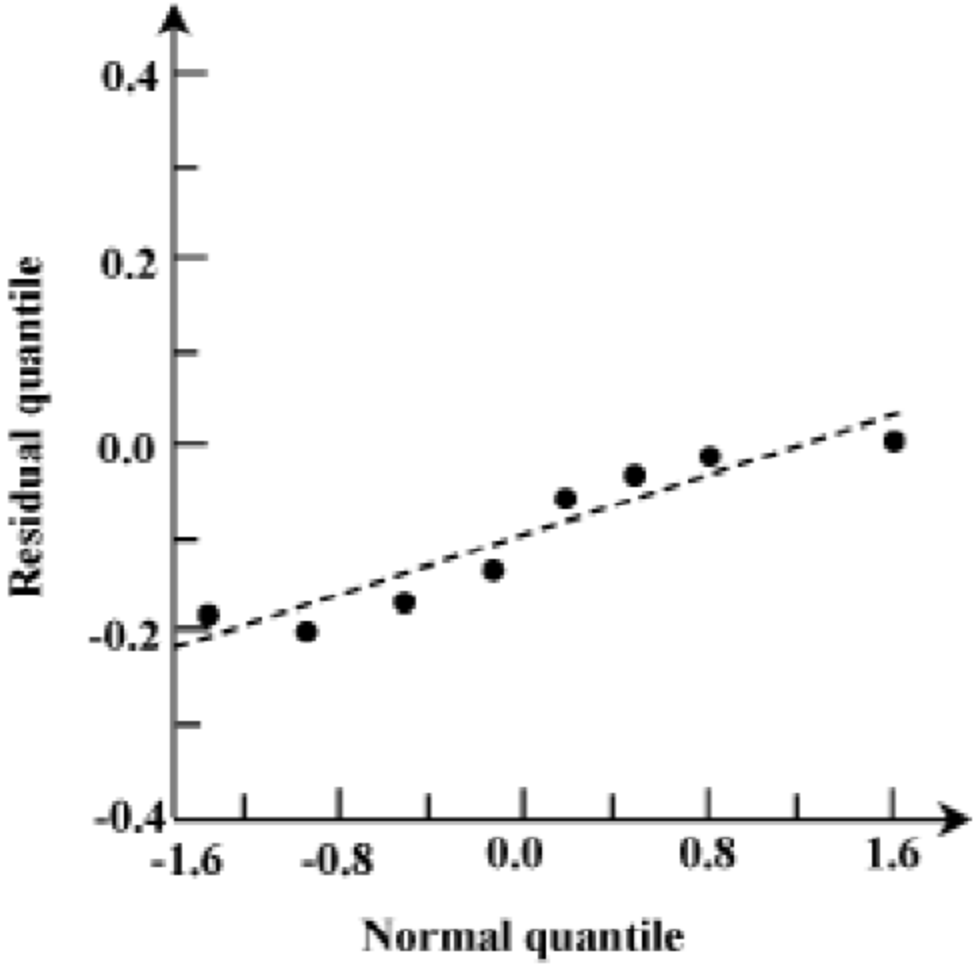
  => (x_i, y_i) plot should be a straight line
- To determine the q_ith quantile x_i, need to invert the cumulative distribution function

$$q_i = F(x_i)$$

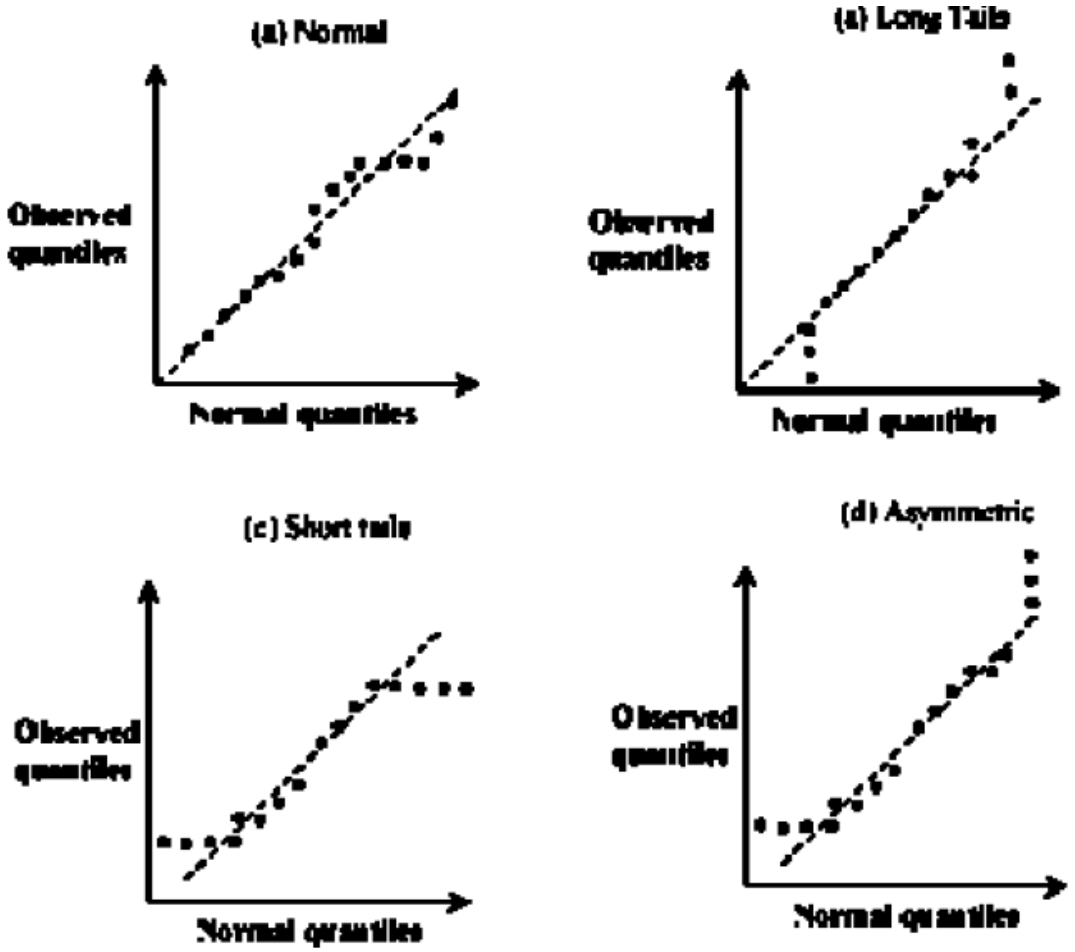
or

$$x_i = F^{-1}(q_i)$$

Example


0.4
0.2
0.0
-0.2
-0.4
Residual quantile
-1.6
-0.8
0.0
0.8
1.6
Normal quantile

Interpretation of quantile-quantile data



Summary

- Indices of Central Tendencies
  - Mean, Median, Mode
  - Arithmetic, Geometric, Harmonic means
- Indices of Dispersion
  - Range, percentiles/Quartiles/SIQR, Variance/Standard-devication/C.O.V., mean absolute deviation
- Determining Distribution of Data
  - Histogram, Quantile-Quantile plot