Summarizing Data

Daniel A. Menascé, Ph.D.
Dept of Computer Science
George Mason University

Major Properties of Numerical Data

- Central Tendency: arithmetic mean, geometric mean, median, mode.
- Variability: range, interquartile range, variance, standard deviation, coefficient of variation, mean absolute deviation.
- Skewness: coefficient of skewness.
- Kurtosis

Measures of Central Tendency

- Arithmetic Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Based on all observations ➡ greatly affected by extreme values.

Effect of Outliers on Average

	1.1	1.1
	1.4	1.4
	1.8	1.8
	1.9	1.9
	2.3	2.3
	2.4	2.4
	2.8	2.8
	3.1	3.1
	3.4	3.4
	3.8	3.8
	10.3	3.5
Average	**3.1**	**2.5**

Geometric Mean

- Geometric Mean: $\left(\prod_{i=1}^{n} X_i\right)^{1/n}$
- Used when the product of the observations is of interest.
- Important when multiplicative effects are at play:
  - Cache hit ratios at several levels of cache
  - Percentage performance improvements between successive versions.
  - Performance improvements across protocol layers.

Example of Geometric Mean

	Performance Improvement			
Test Number	Operating System	Middleware	Application	Avg. Performance Improvement per Layer
1	1.18	1.23	1.10	1.17
2	1.25	1.19	1.25	1.23
3	1.20	1.12	1.20	1.17
4	1.21	1.18	1.12	1.17
5	1.30	1.23	1.15	1.23
6	1.24	1.17	1.21	1.21
7	1.22	1.18	1.14	1.18
8	1.29	1.19	1.13	1.20
9	1.30	1.21	1.15	1.22
10	1.22	1.15	1.18	1.18
Average Performance Improvement per Layer				1.20

Properties of the Geometric Mean

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

- The choice of the base does not change the conclusion.
- Useful for benchmarks
  - x: throughput on target system.
  - y: throughput on base system.

Median

- Middle Value in an Ordered Set of Data.
- If there are no ties, 50% of the values are smaller than the median and 50% are larger.

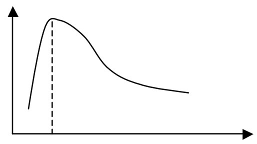
	1.1	1.1
	1.4	1.4
	1.8	1.8
	1.9	1.9
	2.3	2.3
	2.4	**2.4**
	2.8	2.8
	3.1	3.1
	3.4	3.4
	3.8	3.8
	10.3	3.5
Median	**2.4**	**2.4**

Median

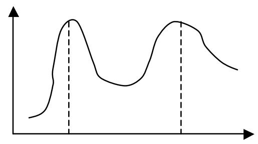
- The median is unaffected by extreme values.
- Obtaining the median:
  - – Odd-sized samples: $X_{(n+1)/2}$
  - – Even-sized samples: $\frac{X_{n/2} + X_{(n/2)+1}}{2}$

Mode

- Most frequently occurring value.
- Mode may not exist.
- Single mode distributions: unimodal.
- Distributions with two modes: bimodal.



unimodal

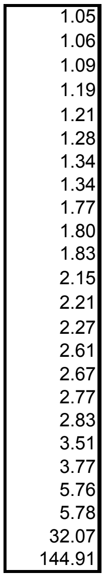


bimodal

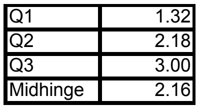
Quantiles (quartiles, percentiles) and midhinge

- Quartiles: split the data into quarters.
  - First quartile (Q1): value of Xi such that 25% of the observations are smaller than Xi.
  - Second quartile (Q2): value of Xi such that 50% of the observations are smaller than Xi.
  - Third quartile (Q3): value of Xi such that 75% of the observations are smaller than Xi.
- Percentiles: split the data into hundredths.
- Midhinge: $Midhinge = \frac{Q_3 + Q_1}{2}$

Example of Quartiles



1.05
1.06
1.09
1.19
1.21
1.28
1.34
1.34
1.77
1.80
1.83
2.15
2.21
2.27
2.61
2.67
2.77
2.83
3.51
3.77
5.76
5.78
32.07
144.91



Q1	1.32
Q2	2.18
Q3	3.00
Midhinge	2.16

In Excel:
Q1=PERCENTILE(<array>,0.25)
Q2=PERCENTILE(<array>,0.5)
Q3=PERCENTILE(<array>,0.75)

Example of Percentile

1.05
1.06
1.09
1.19
1.21
1.28
1.34
1.34
1.77
1.80
1.83
2.15
2.21
2.27
2.61
2.67
2.77
2.83
3.51
3.77
5.76
5.78
32.07
144.91

80-percentile	3.613002

In Excel:
p-th percentile=PERCENTILE(<array>,p)
($0 \leq p \leq 1$)

Range, Interquartile Range, Variance, and Standard Deviation

- Range: $X_{\max} - X_{\min}$
- Interquartile Range: $Q_3 - Q_1$
  - not affected by extreme values.
- Variance:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

In Excel:
s^2=VAR(<array>)

- Standard Deviation:

In Excel:
s=STDEV(<array>)

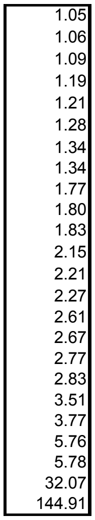
$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

Meanings of the Variance and Standard Deviation

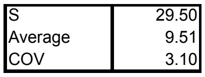
- The larger the spread of the data around the mean, the larger the variance and standard deviation.
- If all observations are the same, the variance and standard deviation are zero.
- The variance and standard deviation cannot be negative.
- Variance is measured in the square of the units of the data.
- Standard deviation is measured in the same units as the data.

Coefficient of Variation

- Coefficient of variation (COV) : $s / \overline{X}$
  - no units



1.05
1.06
1.09
1.19
1.21
1.28
1.34
1.34
1.77
1.80
1.83
2.15
2.21
2.27
2.61
2.67
2.77
2.83
3.51
3.77
5.76
5.78
32.07
144.91



S	29.50
Average	9.51
COV	3.10

Coefficient of Skewness

- Coefficient of skewness: $\frac{1}{ns^3}\sum_{i=1}^{n}(X_i - \overline{X})^3$

	(X-Xi)^3
1.05	-606.1
1.06	-602.9
1.09	-596.1
1.19	-575.2
1.21	-571.8
1.28	-557.9
1.34	-546.4
1.34	-544.8
1.77	-464.5
1.80	-458.1
1.83	-453.1
2.15	-398.9
2.21	-388.8
2.27	-379.0
2.61	-328.5
2.67	-320.5
2.77	-306.6
2.83	-298.7
3.51	-215.9
3.77	-189.6
5.76	-52.9
5.78	-52.1
32.07	11476.6
144.91	2482007.1
	4.033

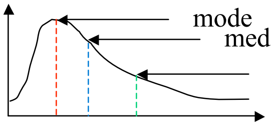
Mean Absolute Deviation

- Mean absolute deviation: $\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$

	abs(Xi-Xbar)
1.05	8.46
1.06	8.45
1.09	8.42
1.19	8.32
1.21	8.30
1.28	8.23
1.34	8.18
1.34	8.17
1.77	7.74
1.80	7.71
1.83	7.68
2.15	7.36
2.21	7.30
2.27	7.24
2.61	6.90
2.67	6.84
2.77	6.74
2.83	6.68
3.51	6.00
3.77	5.74
5.76	3.75
5.78	3.73
32.07	22.56
144.91	135.39
	315.90

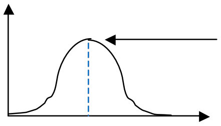
Average	9.51
Mean absolute deviation	13.16

Shapes of Distributions



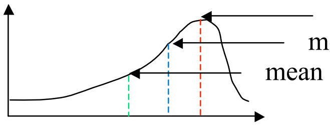
mode
median
mean

Right-skewed distribution



Mode, median, mean

Symmetric distribution



mode
median
mean

Left-skewed distribution

Confidence Interval for the Mean

- The sample mean is an estimate of the population mean.
- Problem: given *k* samples of the population (with *k* sample means), get a single estimate of the population mean.
- Only probabilistic statements can be made:

Confidence Interval for the Mean

$$\Pr[c_1 \leq \mu \leq c_2] = 1 - \alpha$$

where,

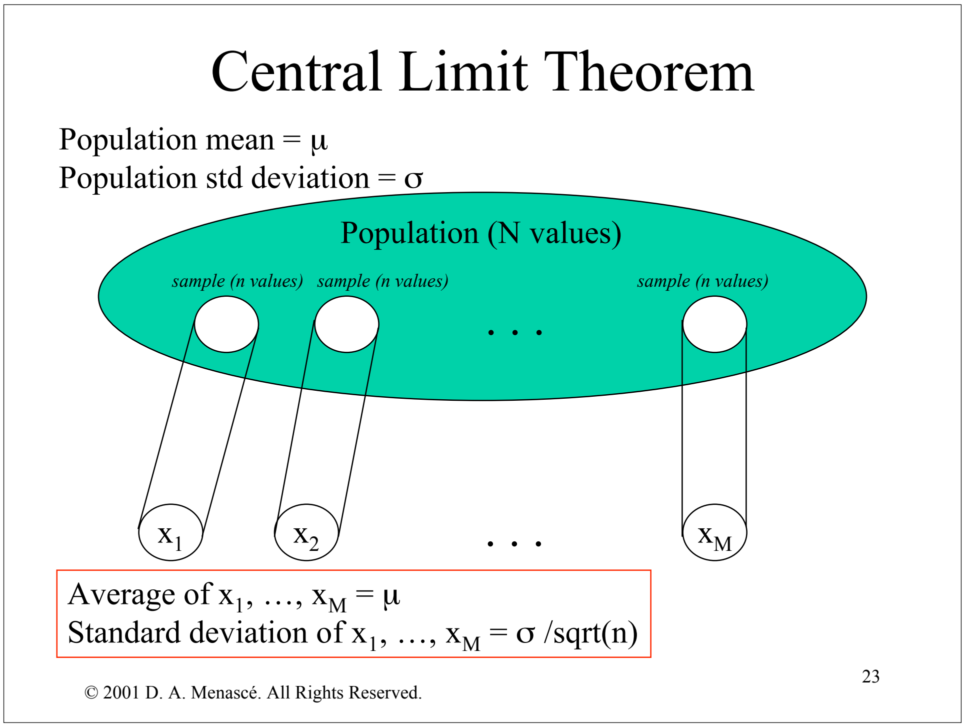
(c_1, c_2): confidence interval

$100(1-\alpha)$: confidence level (usually 90 or 95%)

$1-\alpha$: confidence coefficient.

Central Limit Theorem

- If the observations in a sample are independent and come from the same population that has mean μ and standard deviation σ then the sample mean for large samples has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
- The standard deviation of the sample mean is called the *standard error.*



Central Limit Theorem

Population mean = μ
Population std deviation = σ

Population (N values)

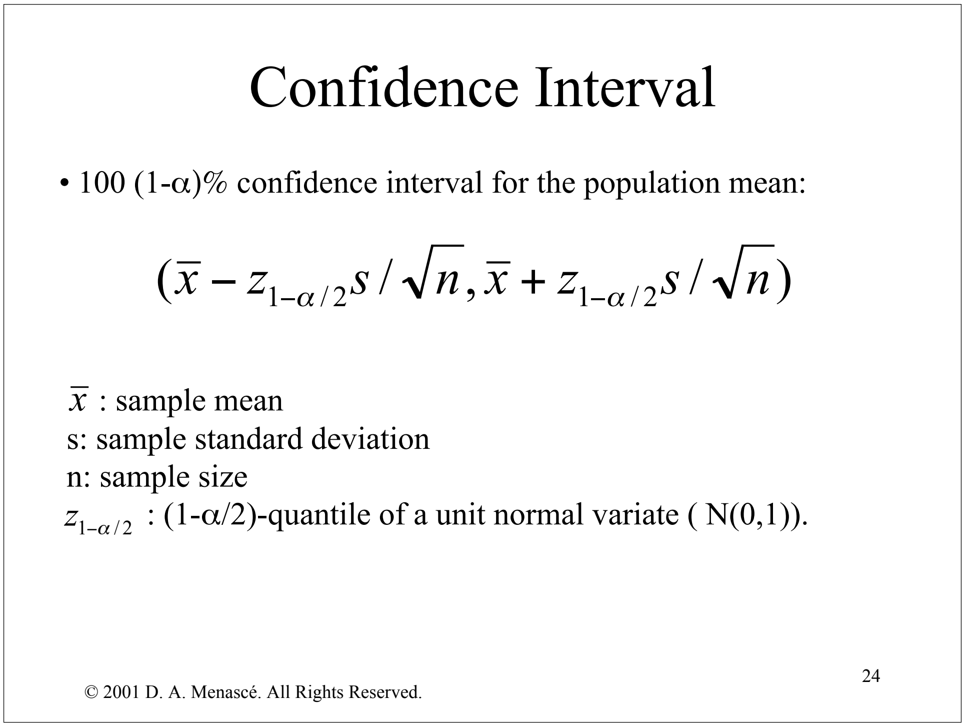
sample (n values) *sample (n values)* *sample (n values)*

. . .

x_1 x_2 . . . x_M

Average of $x_1, \ldots, x_M = \mu$
Standard deviation of $x_1, \ldots, x_M = \sigma$ /sqrt(n)

23



Confidence Interval

- 100 (1-α)% confidence interval for the population mean:

$$(\bar{x} - z_{1-\alpha/2} s / \sqrt{n}, \bar{x} + z_{1-\alpha/2} s / \sqrt{n})$$

$\bar{x}$: sample mean
s: sample standard deviation
n: sample size
$z_{1-\alpha/2}$: (1-α/2)-quantile of a unit normal variate (N(0,1)).

24

Example of Confidence Interval Computation

CPU Time (msec)
5.76
2.67
3.77
2.27
2.83
1.05
2.61
1.06
5.78
3.51
2.77
1.83
1.77
1.19
2.21
24.80
1.80
1.34
1.28
1.21
2.15
1.09
1.34
32.07

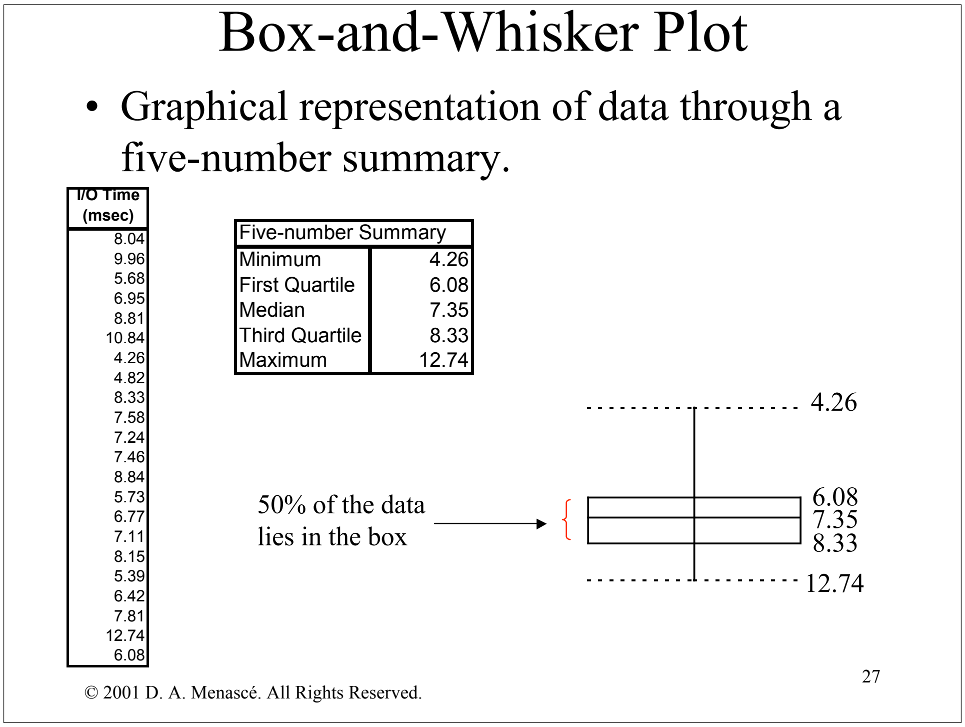
n	24	
sample mean	4.51	
sample std	7.56	
alpha	0.1	
conf level	90	
1-(alpha/2)	0.95	
z0.95	1.645	from a Normal Table
c1	1.97	
c2	7.04	

With 90% confidence the population mean		
is in the interval	**1.97**	**7.04**

From Excel: Tools > Data Analysis > Descriptive Statistics

Descriptive Statistics (from Excel Analysis Pack)	
Mean	9.510589
Standard Error	6.021322
Median	2.180555
Mode	#N/A
Standard Deviation	29.49833
Sample Variance	870.1515
Kurtosis	21.65021
Skewness	4.59114
Range	143.8572
Minimum	1.047923
Maximum	144.9051
Sum	228.2541
Count	24
Confidence Level(95.0%)	12.45604

$$\frac{s}{\sqrt{n}}$$



Box-and-Whisker Plot

- Graphical representation of data through a five-number summary.

I/O Time (msec)
8.04
9.96
5.68
6.95
8.81
10.84
4.26
4.82
8.33
7.58
7.24
7.46
8.84
5.73
6.77
7.11
8.15
5.39
6.42
7.81
12.74
6.08

Five-number Summary	
Minimum	4.26
First Quartile	6.08
Median	7.35
Third Quartile	8.33
Maximum	12.74

50% of the data lies in the box

27