Quatitative Methods and Experimental Design

CS 700

Jana Kosecka, 4444 Research II
kosecka@gmu.edu , 3-1876

Logistics

- **Prerequisites:** at least two 600 level CS courses
- **Course web page** cs.gmu.edu/~kosecka/cs700/
- **Course newsgroup**

- Homeworks 30%
- Midterm 25%
- Final 20%
- Project 25%
- Late policy: semester budget of 3 late days

Readings

- Textbook
  - David Lilja, "Measuring Computer Performance: A Practitioner's Guide"
    - Alternative Text: Raj Jain, "Art of Computer Systems Performance Analysis"
    - Cohen "Empirical techniques in AI"
- Online resources
- Class notes, slides
- Relevant research articles (links on class web site)

3

Software

- Required Software MATLAB + one language of your choice for homeworks and project
- Project – apply techniques covered in the class to the problem of your choice
- Focus on quantitative analysis or simulation
- Project proposal due early November

Course Topics

- Basic techniques in "experimental" computer science
  - measurement tools and techniques
  - Quantitative characterizations of measurement
  - Simulation
  - Design of experiments
- Quantitative Methods
  - Use of statistical techniques in design of experiments
  - Use of statistical techniques in comparing alternatives
  - Characterizing and interpreting measured data
- Simple analytical modeling
  - Initial examples from performance measurement of computer systems and networks, but techniques are applicable in all fields of CS
- Methods used in applied science in general
  - interdisciplinary nature of computer science

The Role of Experimentation in CS

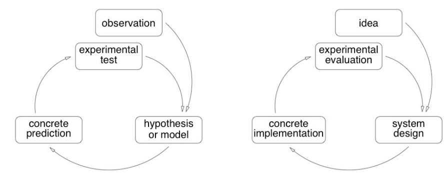


Figure 1: A comparison of the scientific method (on the left) with the role of experimentation in system design (right).

Schedule

- Introduction
- Performance Metrics (time, rate, size)
- Summarizing Measured Data
- Comparing Alternatives, hypothesis testing
- Simulation, design of experiments
- Analytical Modeling
- Linear Regression Models
- Basic optimization

- Statistical Analysis of multidimensional data
- Interpreting & characterizing measured data

Course Goals

- Understand the inherent trade-offs involved in using simulation, measurement, and analytical modeling.
- Rigorously compare computer systems/networks/ software/artifacts/… often in the presence of measurement noise
- Usually compare/measure performance in many fields of CS
- Many times "quality" of the output is more important than raw performance, e.g. face recognition
- Study variability
- Determine whether results are statistically significant impact (related to the amount of evidence)

Course Goals

- Provide intuitive conceptual background for some standard statistical tools
- Draw meaningful conclusions in presence of noisy measurements
- Allow you to correctly and intelligently apply techniques in new situations.
- Present techniques for aggregating and interpreting large quantities of data.
  - Obtain a big-picture view of your results.
  - Obtain new insights from complex measurement and simulation results.

→ E.g. How does a new feature impact the overall system?

Course Goals

- Traditional measurements one dimensional
- Study of analysis of multidimensional data
- Analysis of real and categorical data

Summarizing measured data
means, variability, distributions

Goals in Studying Statistics

- Analyze, present, and describe numerical information properly.
- Draw conclusions about the properties of large populations from sample information (inference)
- Descriptive statistics – characterize sample of populations
- Inferential statistics – draw conclusions about whole population
- Design experiments to learn about real-world situations.
- To forecast or predict not-measured values from a set of measurements.

Population and Sample

- Population (or universe): all N members of a class or group (people, objects, items of interest)
  - E.g., all files retrieved from a Web site since the site went into operation.
- Census: gather data about the whole population
- Sample: portion of the population. Its size is denoted by *n*.
  - E.g., the set of files retrieved from a Web site from 10:00 AM to 2:00 PM on January 03, 2001.

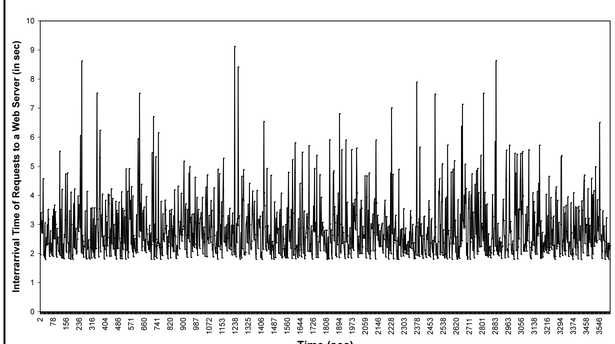
Census, Parameter, Statistic

- Parameter: summary measure of the individual observations made in a census of an entire population.
  - E.g., average size of all files ever retrieved from the Web site.
- Statistic: summary measure obtained from a sample.
  - E.g., average size of all files retrieved from the Web site from 10:00 AM to 2:00 PM on January 03, 2001.

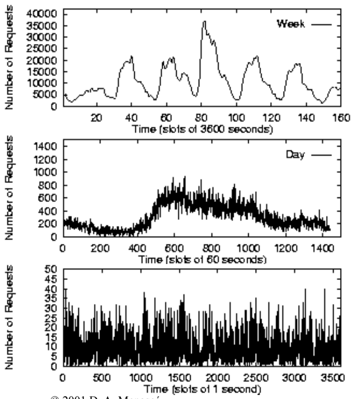
Visualizing Numerical Data

- Type of Plots:
  - Time ordered plots: the time scale is time.
    - Time-scale analysis: time is slotted into fixed time intervals. The y-axis displays a statistics over the time slot (e.g., sum, average).
    - Changing the time scale may reveal interesting properties about the variable being plotted (e.g., strong correlations between adjacent time intervals).
    - Percent frequency histograms: show the percentage of occurrences of values in a bin (range of values).
    - Cumulative frequency histograms.

Example of a Time Plot



Time-scale Analysis



From "In Search of Invariants in E-commerce Workloads," Menascé et al, Proc. ACM Conference on E-commerce, Minneapolis, MN, October 17-20, 2000.

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
- Kurtosis
- Type of distribution

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
- In the absence of other information about data
- Desire to reduce performance to a single number
  - Makes comparisons easy
    - Mine Apple is faster than your Cray!
  - People like a measure of "typical" performance

Mean

- For discrete random variable
- Expected value of X = E[X]
  - "First moment" of X
  - x_i = values measured
  - Sample mean
  - p_i = Pr(X = x_i) = Pr(we measure x_i)

$$E[X] = \sum_{i=1}^{n} x_i p_i$$

For continuous random variable (more details later)

$$\mu = \int x f(x) dx$$

The Problem

- Performance is multidimensional
  - CPU time
  - I/O time
  - Network time
  - Interactions of various components
  - Etc, etc

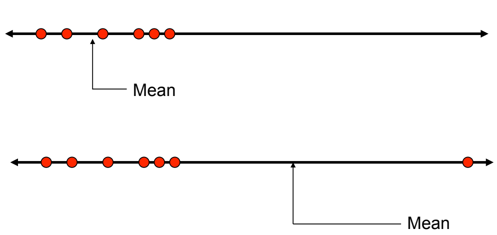
You will be pressured to provide mean values

- Understand how to choose the best type for the circumstance
- Be able to detect bad results from others

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**

Potential Problem with Means



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.

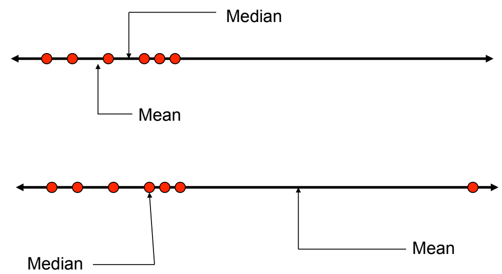
	1.1	1.1
	1.4	1.4
	1.8	1.8
	1.9	1.9
	2.3	2.3
	2.4	**2.4**
	2.8	2.8
	3.1	3.1
	3.4	3.4
	3.8	3.8
	10.3	3.5
Median	**2.4**	**2.4**

Median

- The median is unaffected by extreme values.
- Obtaining the median:
  - Odd-sized samples: $X_{(n+1)/2}$
  - Even-sized samples: $\frac{X_{n/2} + X_{(n/2)+1}}{2}$
- Measured values: 10, 20, 15, 18, 16
  - Mean = 15.8
  - Median = 16
- Obtain one more measurement: 200
  - Mean = 46.5
  - Median = ½ (16 + 18) = 17

25

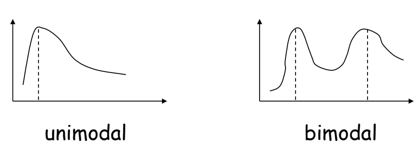
Potential Problem with Means



26

Mode

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



27

Mean, Median, or Mode ?

- Mean
  - If the sum of all values is meaningful
  - Incorporates all available information
- Median
  - Intuitive sense of central tendency with outliers
  - What is "typical" of a set of values?
- Mode
  - When data can be grouped into distinct types, categories (*categorical data*)

- Size of messages sent on a network, Number of cache hits
- Execution time, Bandwidth, Speedup, Cost
- Categorical data type of operating system, name of school

28

Yet Even More Means!

- Arithmetic
- Harmonic?
- Geometric?
- Which one should be used when?

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.
  - Time performance index example

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

Harmonic mean

$$\overline{x_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

What makes a good mean?

- *Time*-based mean (e.g. seconds)
  - Should be *directly proportional* to total weighted time
  - *Time doubles, mean value doubles*
- *Rate*-based mean (e.g. operations/sec)
  - Should be *inversely proportional* to total weighted time
  - *Time doubles, mean value reduced by half*
- Which means satisfy these criteria?

33

Assumptions

- Measured execution times of *n* benchmark programs
  - T_i, i = 1, 2, …, *n*
- Total work performed by each benchmark is constant
  - F = # operations performed
  - Relax this assumption later
- Execution rate = $M_i = F / T_i$

34

Arithmetic mean for times

- Produces a mean value that is *directly proportional to total time*

→ Correct mean to summarize *execution time*

$$\overline{T_A} = \frac{1}{n}\sum_{i=1}^{n} T_i$$

35

Arithmetic mean for rates

- Produces a mean value that is proportional to *sum of inverse of times*
- But we want *inversely proportional to sum of times*

$$\overline{M_A} = \frac{1}{n}\sum_{i=1}^{n} M_i = \sum_{i=1}^{n} \frac{F / T_i}{n} = \frac{F}{n}\sum_{i=1}^{n} \frac{1}{T_i}$$

36

Arithmetic mean for rates

- Produces a mean value that is proportional to *sum of inverse of times*
- But we want *inversely proportional to sum of times*

→ Arithmetic mean is ***not*** appropriate for summarizing rates

$$\overline{M_A} = \frac{1}{n}\sum_{i=1}^{n} M_i = \sum_{i=1}^{n} \frac{F / T_i}{n} = \frac{F}{n}\sum_{i=1}^{n}\frac{1}{T_i}$$

37

Harmonic mean for times

- Not directly proportional to *sum of times*

$$\overline{T_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{T_i}}$$

38

Harmonic mean for times

- Not directly proportional to *sum of times*

→ Harmonic mean is ***not*** appropriate for summarizing times

$$\overline{T_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{T_i}}$$

39

Harmonic mean for rates

- Produces
  (total number of ops)
  ÷ (sum execution times)
- Inversely proportional to total execution time

→ Harmonic mean is appropriate to summarize rates

$$\overline{M_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{M_i}} = \frac{n}{\sum_{i=1}^{n} \frac{T_i}{F}} = \frac{Fn}{\sum_{i=1}^{n} T_i}$$

40

Harmonic mean for rates

Sec	10^9 FLOPs	MFLOPS
321	130	405
436	160	367
284	115	405
601	252	419
482	187	388

$$\overline{M_H} = \frac{5}{\left(\frac{1}{405}+\frac{1}{367}+\frac{1}{405}+\frac{1}{419}+\frac{1}{388}\right)} = 396$$

$$\overline{M_H} = \frac{844 \times 10^9}{2124} = 396$$

Geometric mean

- Claim: Correct mean for averaging normalized values
  - Used to compute SPECmark
- Claim: Good when averaging measurements with wide range of values
- Maintains consistent relationships when comparing normalized values
  - Independent of basis used to normalize

Geometric mean with times

	System 1	System 2	System 3
	417	244	134
	83	70	70
	66	153	135
	39,449	33,527	66,000
	772	368	369
Geo mean	587	503	499
Rank	3	2	1

Geometric mean normalized to System 1

	System 1	System 2	System 3
	1.0	0.59	0.32
	1.0	0.84	0.85
	1.0	2.32	2.05
	1.0	0.85	1.67
	1.0	0.48	0.45
Geo mean	1.0	0.86	0.84
Rank	3	2	1

Geometric mean normalized to System 2

	System 1	System 2	System 3
	1.71	1.0	0.55
	1.19	1.0	1.0
	0.43	1.0	0.88
	1.18	1.0	1.97
	2.10	1.0	1.0
Geo mean	1.17	1.0	0.99
Rank	3	2	1

Total execution times

	System 1	System 2	System 3
	417	244	134
	83	70	70
	66	153	135
	39,449	33,527	66,000
	772	368	369
Total	40,787	34,362	66,798
Arith mean	8157	6872	13,342
Rank	2	1	3

What's going on here?!

	System 1	System 2	System 3
Geo mean wrt 1	1.0	0.86	0.84
Rank	3	2	1
Geo mean wrt 2	1.17	1.0	0.99
Rank	3	2	1
Arith mean	8157	6872	13,342
Rank	2	1	3

Geometric mean for times

- Not directly proportional to *sum of times*

$$\overline{T_G} = \left(\prod_{i=1}^{n} T_i \right)^{1/n}$$

Geometric mean for times

- Not directly proportional to *sum of times*

→ Geometric mean is ***not*** appropriate for summarizing times

$$\overline{T_G} = \left(\prod_{i=1}^{n} T_i\right)^{1/n}$$

49

Geometric mean for rates

- Not inversely proportional to *sum of times*

$$\overline{T_G} = \left(\prod_{i=1}^{n} M_i\right)^{1/n}$$

$$= \left(\prod_{i=1}^{n} \frac{F}{T_i}\right)^{1/n}$$

50

Geometric mean for rates

- Not inversely proportional to *sum of times*

→ Geometric mean is ***not*** appropriate for summarizing rates

$$\overline{T_G} = \left(\prod_{i=1}^{n} M_i\right)^{1/n}$$

$$= \left(\prod_{i=1}^{n} \frac{F}{T_i}\right)^{1/n}$$

51

Geometric mean

- Does provide consistent rankings
  - Independent of basis for normalization
- But can be consistently wrong!
- Value can be computed
  - But has no physical meaning

52

Other uses of Geometric Mean

- 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

Summary of Means

- Avoid means if possible
  - Loses information
- Arithmetic
  - When sum of raw values has physical meaning
  - Use for summarizing **times** (not rates)
- Harmonic
  - Use for summarizing **rates** (not times)
- Geometric mean
  - Not useful when *time* is best measure of perf
  - Useful when multiplicative effects are in play

Normalization

- Averaging normalized values doesn't make sense mathematically
  - Gives a number
  - But the number has no physical meaning
- First compute the mean
  - Then normalize

Weighted means

$$\sum_{i=1}^{n} w_i = 1$$

$$\bar{x}_A = \sum_{i=1}^{n} w_i x_i$$

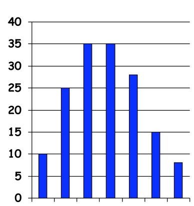
$$\bar{x}_H = \frac{1}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

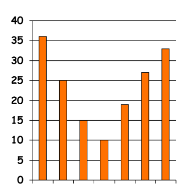
- Standard definition of mean assumes all measurements are equally important
- Instead, choose weights to represent relative importance of measurement *i*

Quantifying variability

- Mean hides information about variability
- How spread are the values
- What is the shape of distributions
- Indices of dispersion
  - Range
  - Variance or standard deviation
  - 10- and 90- percentiles
  - Semi-interquartile range
  - Mean absolute deviation

Histograms





- Similar mean values
- Widely different distributions
- How to capture this variability in one number?

Index of Dispersion

- Quantifies how "spread out" measurements are
- Range
  - (max value) – (min value)
- Maximum distance from the mean
  - Max of | x_i – mean |
- Neither efficiently incorporates all available information

Sample Variance

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

$$= \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$$

- *Second moment of random variable X*
- Second form good for calculating "on-the-fly"
  - One pass through data
- (n-1) *degrees of freedom*

Sample Variance

- Gives "units-squared"
- Hard to compare to mean
- Use *standard deviation, s*
  - s = square root of variance
  - Units = same as mean

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) : $s / \bar{X}$
- Ratio of standard deviation to mean
  - 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:
- Measure of assymetry of distribution

- Used for measuring deviation from normal Gaussian distribution

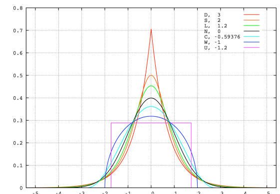
$$\frac{1}{ns^3}\sum_{i=1}^{n}(X_i - \overline{X})^3$$

	(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

65

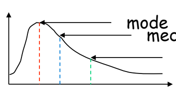
Coefficient of Kurtosis

- Coefficient of kurtosis: $\frac{1}{ns^4}\sum_{i=1}^{n}(X_i - \overline{X})^4$
- Measure of peakedness of distribution
- High kurtosis – variance is due to many infrequent observations
- Used another 'feature' of the distribution
- Kurtosis of common distributions

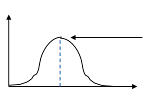


66

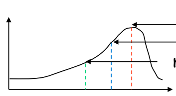
Shapes of Distributions



positive skew distribution
Right skew



Symmetric distribution



negative skew distribution
Left skew

67

Mean Absolute Deviation

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

	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

68

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:

$$Midhinge = \frac{Q_3 + Q_1}{2}$$

69

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)

70

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≤p≤1)

71

Range, Interquartile Range, Variance, and Standard Deviation

- Interquartile Range: $Q_3 - Q_1$
  - not affected by extreme values.
- Semi-interquartile Range (SIQR): $(Q_3 - Q_1)/2$
- Variance:

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

- Standard Deviation:

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

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

- If the distribution is highly skewed, SIQR is Preferred to the standard deviation for the same reason that median is preferred to mean

72

Selecting the index of dispersion

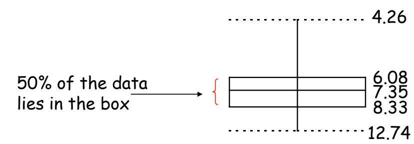
- Numerical data
  - If the distribution is bounded, use the range
  - For unbounded distributions that are unimodal and symmetric, use C.O.V.
  - O/w use percentiles or SIQR

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



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:
- E.g. we want mean of the population but can get only mean of the sample
- k samples, k estimates of the mean
- Finite size samples, we cannot get the true mean
- We can get probabilistic bounds

Determining the Distributions of a Data Set

- A measured data set can be summarized by stating its average and variability
- If we can say something about the distribution of the data, that would provide all the information about the data
  - Distribution information is required if the summarized mean and variability have to be used in simulations or analytical models
- To determine the distribution of a data set, we compare the data set to a theoretical distribution
  - Heuristic techniques (Graphical/Visual): Histograms, Q-Q plots
  - Statistical goodness-of-fit tests: Chi-square test, Kolmogrov-Smirnov test
    - Will discuss this topic in detail later this semester

Comparing Data Sets

- Problem: given two data sets D1 and D2 determine if the data points come from the same distribution.
- Simple approach: draw a histogram for each data set and visually compare them.
- To study relationships between two variables use a scatter plot.
- To compare two distributions use a quantile-quantile (Q-Q) plot.

Histogram

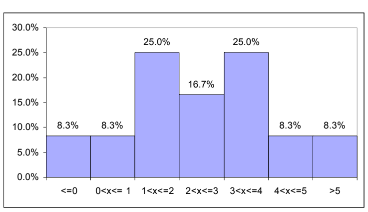
- Divide the range (max value – min value) into equal-sized cells or bins.
- Count the number of data points that fall in each cell.
- Plot on the y-axis the relative frequency, i.e., number of point in each cell divided by the total number of points and the cells on the x-axis.
- Cell size is critical!
  - Sturge's rule of thumb
    Given n data points, number of bins $k = \lfloor 1 + \log_2 n \rfloor$

Histogram

Data
-3.0
0.8
1.2
1.5
2.0
2.3
2.4
3.3
3.5
4.0
4.5
5.5

Bin	Frequency	Relative Frequency
<=0	1	8.3%
0<x<= 1	1	8.3%
1<x<=2	3	25.0%
2<x<=3	2	16.7%
3<x<=4	3	25.0%
4<x<=5	1	8.3%
>5	1	8.3%

In Excel:
Tools -> Data Analysis -> Histogram

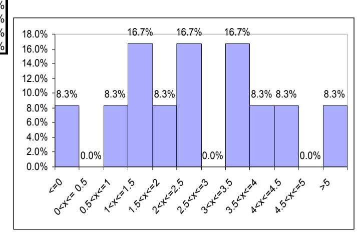


Histogram

Data
-3.0
0.8
1.2
1.5
2.0
2.3
2.4
3.3
3.5
4.0
4.5
5.5

Bin	Frequency	Relative Frequency
<=0	1	8.3%
0<x<= 0.5	0	0.0%
0.5<x<=1	1	8.3%
1<x<=1.5	2	16.7%
1.5<x<=2	1	8.3%
2<x<=2.5	2	16.7%
2.5<x<=3	0	0.0%
3<x<=3.5	2	16.7%
3.5<x<=4	1	8.3%
4<x<=4.5	1	8.3%
4.5<x<=5	0	0.0%
>5	1	8.3%

Same data, different cell size, different shape for the histograms !



Example System- Robotic Navigation

- Stanford Stanley Grand Challenge
- Outdoors unstructured env., single vehicle

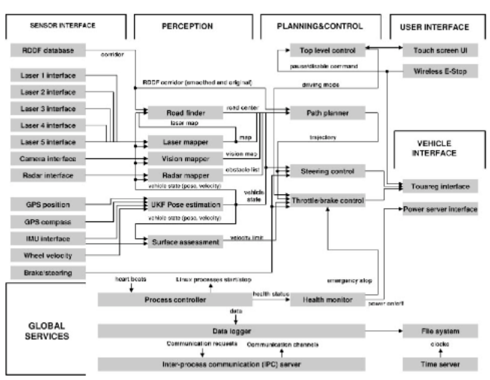
- Urban Challenge
- Outdoors structured env., mixed traffic, traffic rules

Robot Components (Stanley)

- Sensors
- Actuators-Effectors
- Locomotion System
- Computer system – Architectures

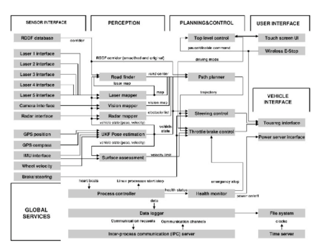
- Lasers, camera, radar, GPS, compass, antenna, IMU,
- Steer by wire system
- Rack of PC's with Ethernet for processing information from sensors

Example System

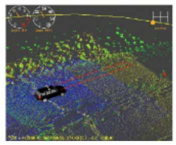


System performance

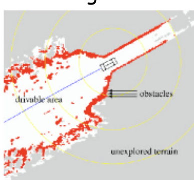
- Performance can be analyzed an many levels
- Sensors – speed, accuracy, noise characterization
- Design of algorithms for sensing and control
- Characterizing throughput and delays in the system
- Accuracy of the classification algorithms
- Complexity and accuracy of the planning algorithms

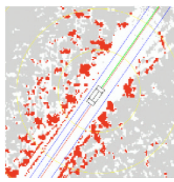


- Terrain mapping using lasers

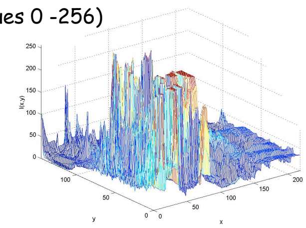


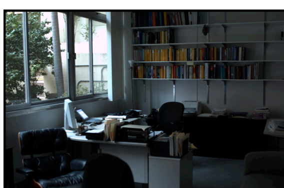
- Determining obstacle course



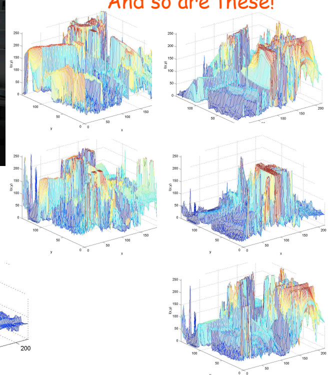


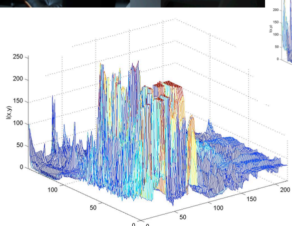
This is how a computer represents it
(gray level values 0 -256)





And so are these!





We need to extract some "invariant", i.e. what is common to all these images (they are all images of an office)

Face/object detection

Face detection
Car detection
Pedestrian detection

- Performance of the algorithm
- Classification accuracy
- Precision/recall curves

- True positives
- True negatives
- False positives
- False negatives

$$precision = \frac{tp}{tp + fp} \qquad recall = \frac{tp}{tp + fn}$$

Document retrieval applications

Information retrieval context
- set of retrieved documents
- set of relevant documents

$$precision = \frac{relevant \cap retrieved}{retrieved}$$

$$recall = \frac{relevant \cap retrieved}{relevant}$$