Summarizing Measured Data - Means, Variability, Distributions

Major Properties of Numerical Data

- Central Tendency: arithmetic mean, geometric mean, harmonic mean, median, mode.
- Variability: range, inter-quartile range, variance, standard deviation, coefficient of variation, mean absolute deviation
- Distribution: type of distribution

Why mean values?

- 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
- Leads to all sorts of crazy ways for summarizing data
  - X = *f*(10 parts A, 25 parts B, 13 parts C, …)
  - X then represents "typical" performance?!

3

The Problem

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

4

The Problem

- Systems are often specialized
  - Performs great on application type X
  - Performs lousy on anything else
- Potentially a wide range of execution times on one system using different benchmark programs

5

The Problem

- Nevertheless, people still want a single number answer!
- *How to (correctly) summarize a wide range of measurements with a single value?*

6

Index of Central Tendency

- Tries to capture "center" of a distribution of values
- Use this "center" to summarize overall behavior
- Not recommended for real information, but …
  - 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

7

Indices of Central Tendency

- Sample mean
  - Common "average"
- Sample median
  - ½ of the values are above, ½ below
- Mode
  - Most common

8

Indices of Central Tendency

- "Sample" implies that
  - Values are measured from a random process on discrete random variable X
- Value computed is only an approximation of true mean value of underlying process
- True mean value cannot actually be known
  - Would require infinite number of measurements

9

Sample mean

- Expected value of X = E[X]
  - "First moment" of X
  - x_i = values measured
  - $p_i = Pr(X = x_i) = Pr(\text{we measure } x_i)$

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

10

Sample mean

- Without additional information, assume
  - p_i = constant = 1/n
  - n = number of measurements
- ***Arithmetic mean***
  - Common "average"

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

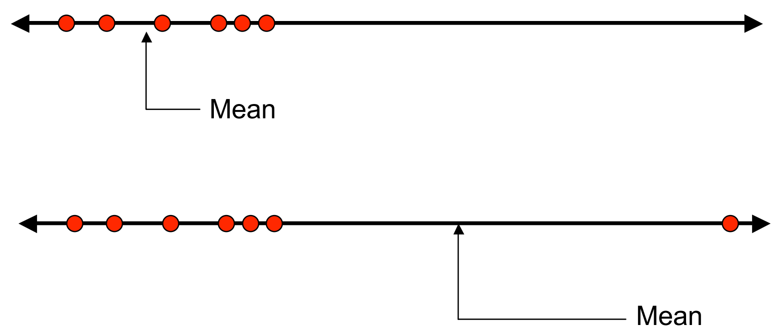
11

Potential Problem with Means

- Sample mean gives equal weight to all measurements
- *Outliers* can have a large influence on the computed mean value
- Distorts our intuition about the *central tendency* of the measured values

12

Potential Problem with Means



13

Median

- Index of central tendency with
  - ½ of the values larger, ½ smaller
- Sort n measurements
- If n is odd
  - Median = middle value
  - Else, median = mean of two middle values
- Reduces skewing effect of outliers on the value of the index

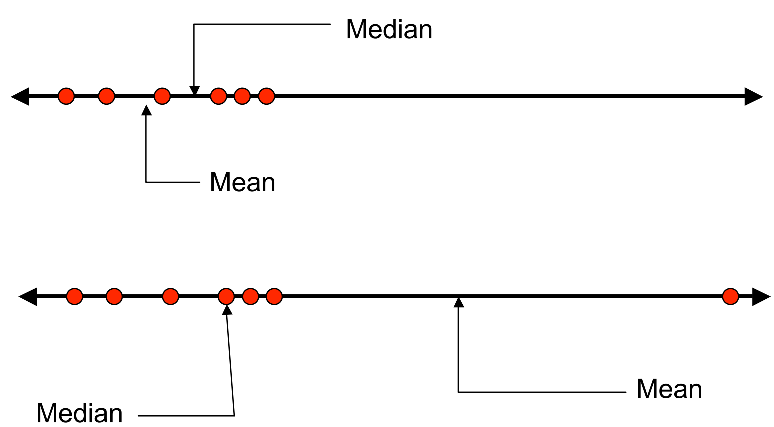
14

Example

- 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
- Median give more intuitive sense of central tendency

15

Potential Problem with Means



16

Mode

- Value that occurs most often
- May not exist
- May not be unique
  - E.g. “bi-modal” distribution
    - Two values occur with same frequency

17

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*)

18

Mean, Median, or Mode?

- Size of messages sent on a network
- Number of cache hits
- Execution time
- MFLOPS, MIPS
- Bandwidth
- Speedup
- Cost

19

Yet Even More Means!

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

20

Arithmetic mean

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

21

Harmonic mean

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

22

Geometric mean

$$\overline{x_G} = \sqrt[n]{x_1 x_2 \cdots x_i \cdots x_n}$$

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

23

Which mean to use?

- Mean value must still conform to characteristics of a *good* performance metric
  - Linear
  - Reliable
  - Repeatable
  - Easy to use
  - Consistent
  - Independent
- Best measure of performance still is *execution time*

24

What makes a good mean?

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

25

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 = $M_i = F / T_i$

26

Arithmetic mean for times

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

→ Correct mean to summarize *execution time*

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

27

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*

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

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

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

28

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

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

29

Harmonic mean for times

- Not directly proportional to *sum of times*

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

30

Harmonic mean for times

- Not directly proportional to *sum of times*

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

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

31

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

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

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

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

32

Harmonic mean for rates

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

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

$$\overline{M_H} = \frac{844\times10^9}{2124} = 396$$

33

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

34

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

35

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

36

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

37

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

38

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

39

Geometric mean for times

- Not directly proportional to *sum of times*

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

40

Geometric mean for times

- Not directly proportional to *sum of times*

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

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

41

Geometric mean for rates

- Not inversely proportional to *sum of times*

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

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

42

Geometric mean for rates

- Not inversely proportional to *sum of times*

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

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

43

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

44

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.

45

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

46

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

47

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

48

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*

49

Summarizing Variability

Quantifying variability

- Means hide information about variability
- How "spread out" are the values?
- How much spread relative to the mean?
- What is the shape of the distribution of values?

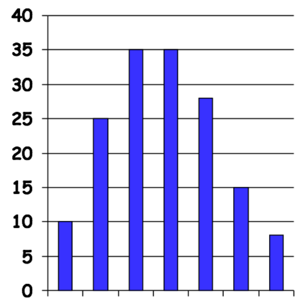
51

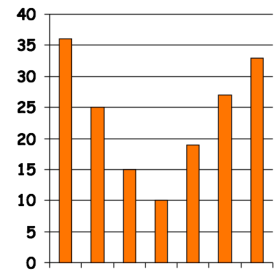
Quantifying variability

- Indices of dispersion
  - Range
  - Variance or standard deviation
  - 10- and 90- percentiles
  - Semi-interquartile range
  - Mean absolute deviation

52

Histograms





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

53

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

54

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*

55

Sample Variance

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

56

Coefficient of Variation (COV)

$$COV = \frac{s}{\bar{x}}$$

- Dimensionless
- Compares relative size of variation to mean value
- Not meaningful for distributions with negative or zero mean

57

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:

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

58

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)

59

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)
($0 \leq p \leq 1$)

60

Interquartile Range

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

  SIQR = $(Q_3 - Q_1)/2$
- If the distribution is highly skewed, SIQR is preferred to the standard deviation for the same reason that median is preferred to mean

61

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

62

Mean Absolute Deviation

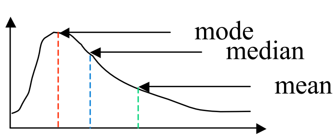
- Mean absolute deviation: $\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$

	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

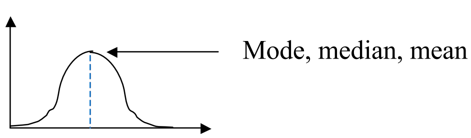
Average	9.51
Mean absolute deviation	13.16

63

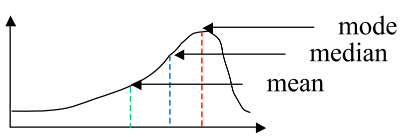
Shapes of Distributions



Right-skewed distribution



Symmetric distribution



Left-skewed distribution

64

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

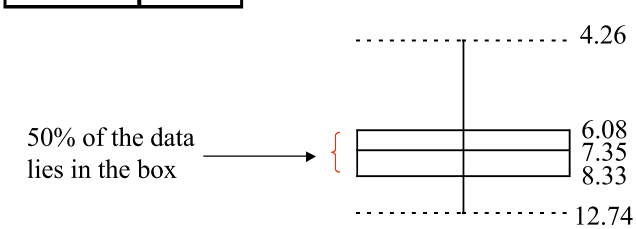
65

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



66

Determining Distributions

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.

69

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 $k = \lfloor 1 + \log_2 n \rfloor$

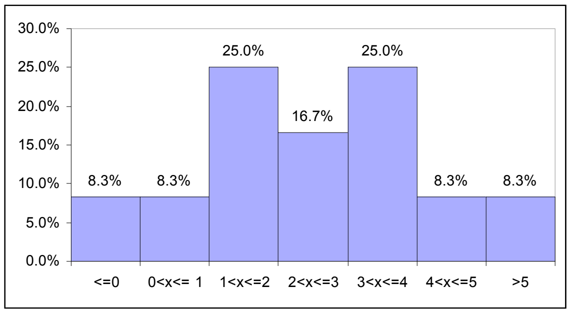
70

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



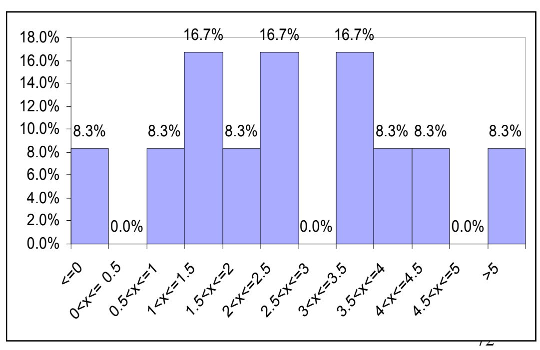
71

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!



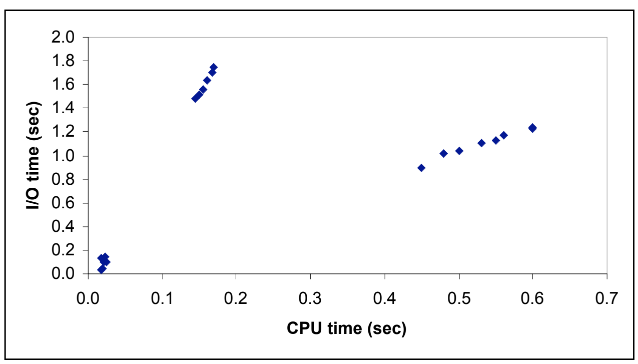
Scatter Plot

- Plot a data set against each other to visualize potential relationships between the data sets.
- Example: CPU time vs. I/O Time
- In Excel: XY (Scatter) Chart Type.

73

Scatter Plot

CPU Time (sec)	I/O Time (sec)
0.020	0.043
0.150	1.516
0.500	1.037
0.023	0.141
0.160	1.635
0.450	0.900
0.170	1.744
0.550	1.132
0.018	0.037
0.600	1.229
0.145	1.479
0.530	1.102
0.021	0.094
0.480	1.019
0.155	1.563
0.560	1.171
0.018	0.131
0.600	1.236
0.167	1.703
0.025	0.103



74

Plots Based on Quantiles

- Consider an ordered data set with n values $x_1, \ldots, x_n$.
- If $p = (i\text{-}0.5)/n$ for $i \leq n$, then the p quantile $Q(p)$ of the data set is defined as

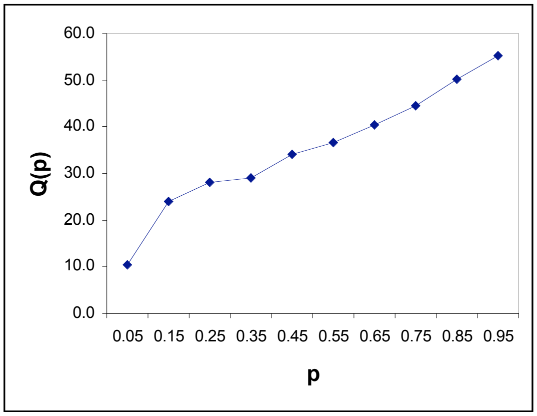
$$Q(p)= Q([i\text{-}0.5]/n)=x_i$$

- $Q(p)$ for other values of p is computed by linear interpolation.
- A quantile plot is a plot of $Q(p)$ vs. p.

75

Example of a Quantile Plot

i	*p=(i-0.5)/n*	*xi = Q(p)*
1	0.05	10.5
2	0.15	24.0
3	0.25	28.0
4	0.35	29.0
5	0.45	34.0
6	0.55	36.5
7	0.65	40.3
8	0.75	44.5
9	0.85	50.3
10	0.95	55.3



76

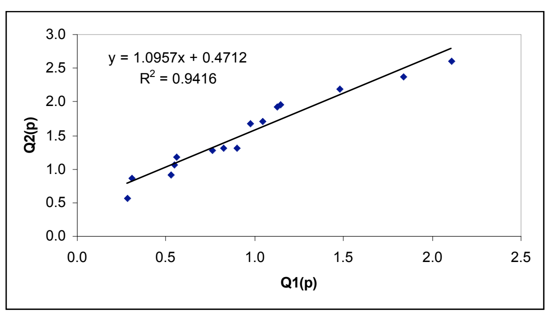
Quantile-Quantile (Q-Q plots)

- Used to compare distributions.
- "Equal shape" is equivalent to "linearly related quantile functions."
- A Q-Q plot is a plot of the type $(Q_1(p),Q_2(p))$ where $Q_1(p)$ is the quantile function of data set 1 and $Q_2(p)$ is the quantile function of data set 2. The values of p are $(i\text{-}0.5)/n$ where n is the size of the smaller data set.

77

Q-Q Plot Example

i	p=(i-0.5)/n	Data 1	Data 2
1	0.033	0.2861	0.5640
2	0.100	0.3056	0.8657
3	0.167	0.5315	0.9120
4	0.233	0.5465	1.0539
5	0.300	0.5584	1.1729
6	0.367	0.7613	1.2753
7	0.433	0.8251	1.3033
8	0.500	0.9014	1.3102
9	0.567	0.9740	1.6678
10	0.633	1.0436	1.7126
11	0.700	1.1250	1.9289
12	0.767	1.1437	1.9495
13	0.833	1.4778	2.1845
14	0.900	1.8377	2.3623
15	0.967	2.1074	2.6104



A Q-Q plot that is reasonably linear indicates that the two data sets have distributions with similar shapes.

78

Theoretical Q-Q Plot

- Compare one empirical data set with a theoretical distribution.
- Plot (x_i, $Q_2([i\text{-}0.5]/n)$) where x_i is the $[i\text{-}0.5]/n$ quantile of a theoretical distribution ($F^{-1}([i\text{-}0.5]/n)$) and $Q_2([i\text{-}0.5]/n)$ is the i-th ordered data point.
- If the Q-Q plot is reasonably linear the data set is distributed as the theoretical distribution.

79

Examples of CDFs and Their Inverse Functions

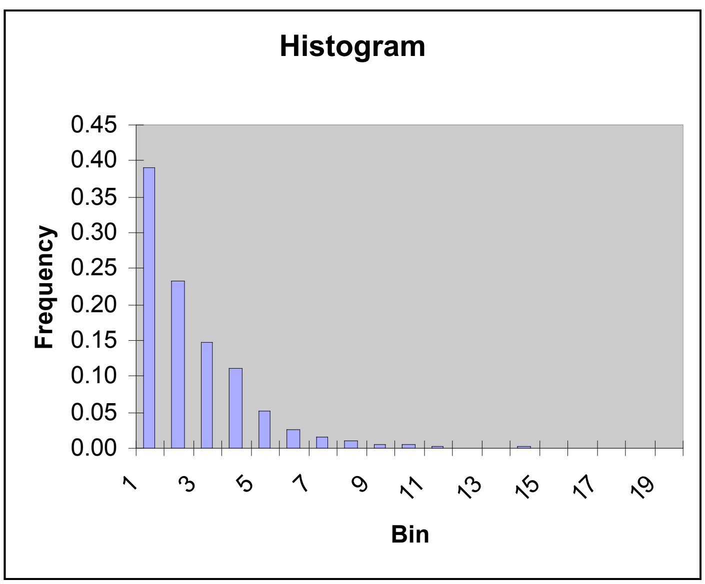
Exponential	$F(x) = 1 - e^{-x/a}$	$-a\mathrm{Ln}(1-u)$
Pareto	$F(x) = 1 - x^{-a}$	$\dfrac{1}{(1-u)^{1/a}}$
Geometric	$F(x) = 1 - (1-p)^x$	$\left\lceil \dfrac{\mathrm{Ln}(u)}{\mathrm{Ln}(1-p)} \right\rceil$

80

Example of a Quantile-Quantile Plot

- One thousand values are suspected of coming from an exponential distribution (see histogram in the next slide). The quantile-quantile plot is pretty much linear, which confirms the conjecture.

81

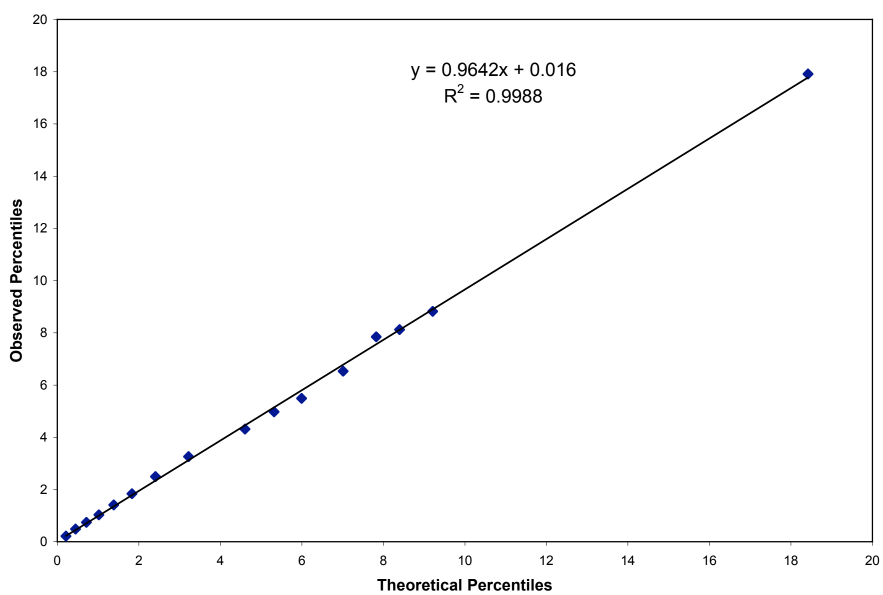


82

Data for Quantile-Quantile Plot

qi	yi	xi
0.100	0.22	0.21
0.200	0.49	0.45
0.300	0.74	0.71
0.400	1.03	1.02
0.500	1.41	1.39
0.600	1.84	1.83
0.700	2.49	2.41
0.800	3.26	3.22
0.900	4.31	4.61
0.930	4.98	5.32
0.950	5.49	5.99
0.970	6.53	7.01
0.980	7.84	7.82
0.985	8.12	8.40
0.990	8.82	9.21
1.000	17.91	18.42

83



84

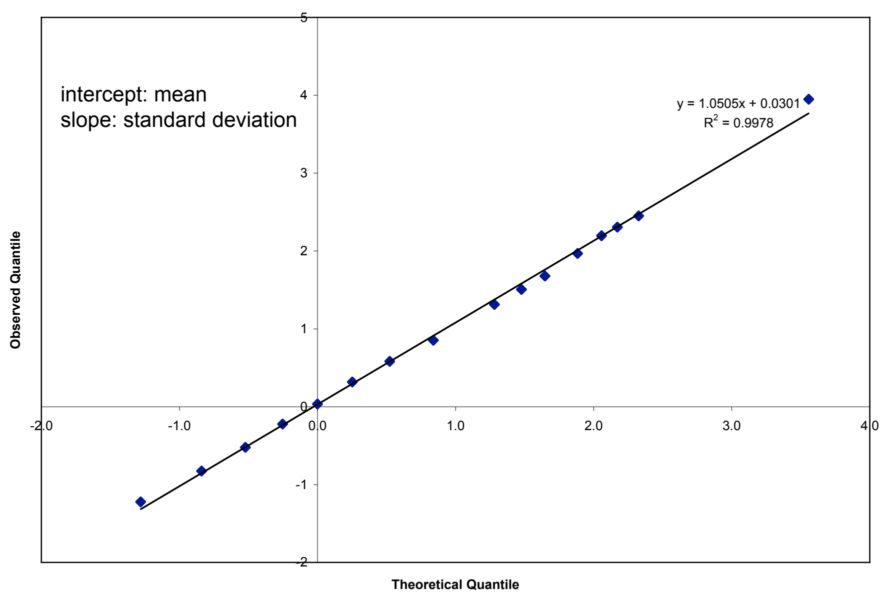
What if the Inverse of the CDF Cannot be Found?

- Use approximations or use statistical tables
  - Quantile tables have been computed and published for many important distributions
- For example, approximation for N(0,1):

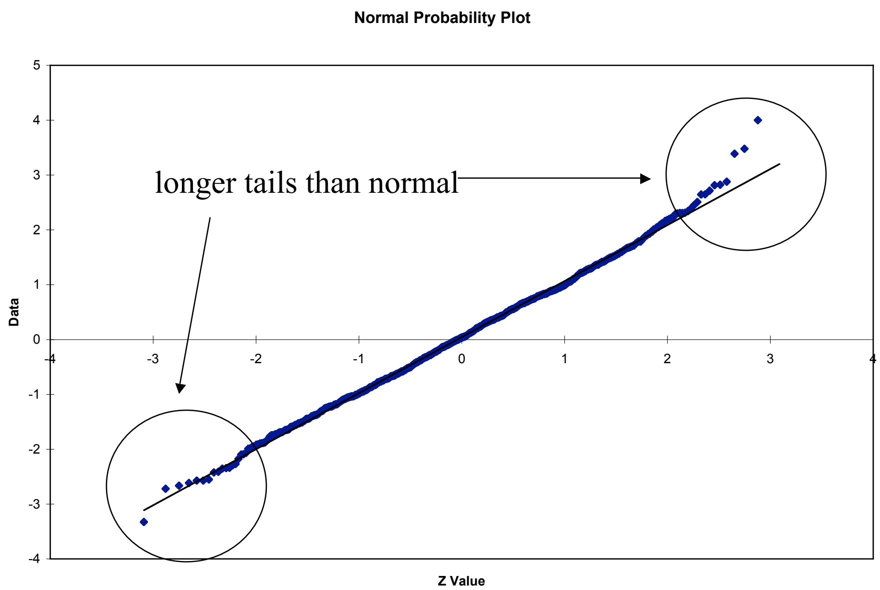
$$x_i = 4.91[q_i^{0.14} - (1-q_i)^{0.14}]$$

- For N(μ,σ) the x_i values are scaled as $\mu + \sigma x_i$ before plotting.

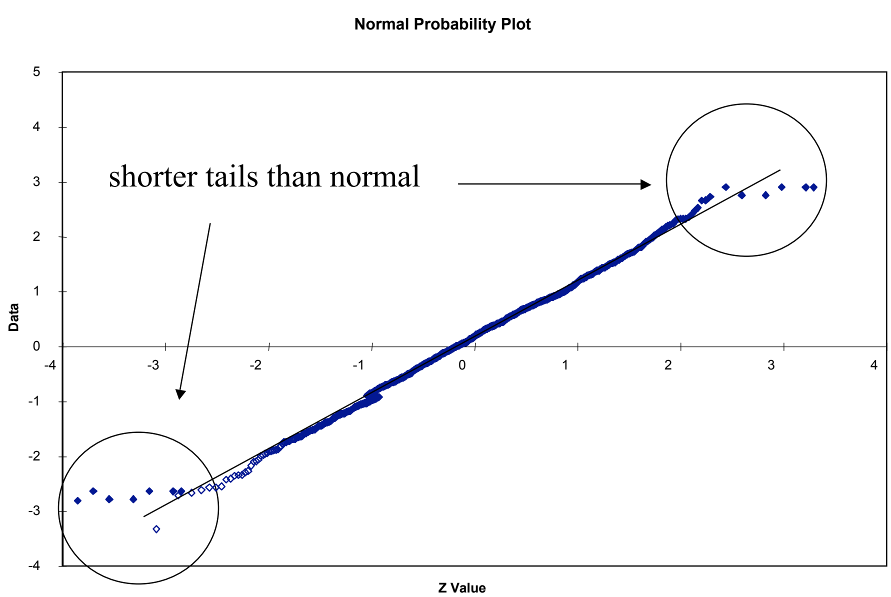
85



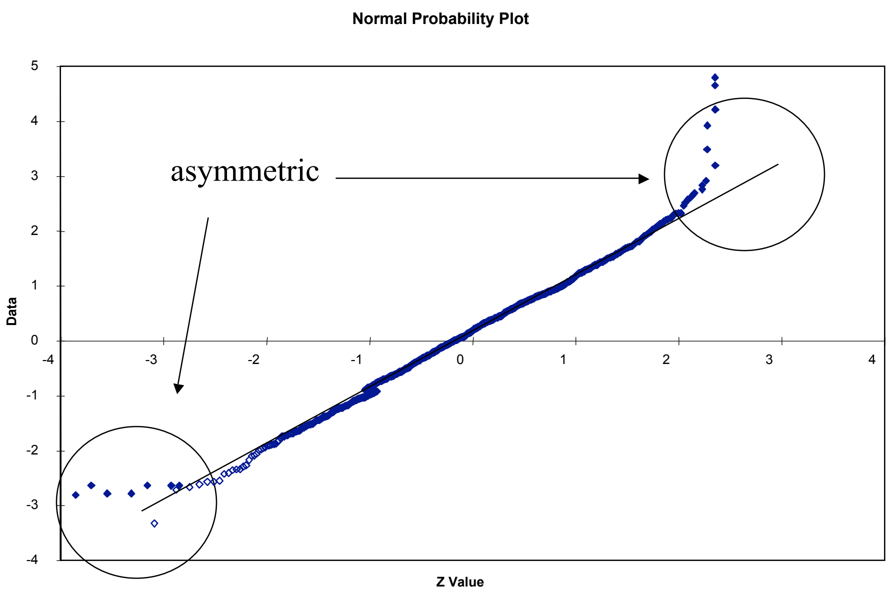
86


Normal Probability Plot
longer tails than normal
Data
Z Value
5
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4

87


Normal Probability Plot
shorter tails than normal
Data
Z Value
5
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4

88


Normal Probability Plot
5
4
3
2
1
0
-1
-2
-3
-4
Data
-4
-3
-2
-1
0
1
2
3
4
Z Value
asymmetric

89