Summarizing Measured Data

CS 700

1

Acknowledgement

These slides are based on presentations created and copyrighted by Prof. Daniel Menasce (GMU)

2

Types of Data

- Qualitative (also called categorical)
  - Data has states, categories, or levels that are mutually exclusive and exhaustive
    - E.g. computers can be classified as laptops, handheld (PDAs), desktops, servers
  - Categories can be ordered or unordered
- Quantitative (also called numerical)
  - Discrete variables
  - Continuous variables

3

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

4

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.

5

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**

6

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.

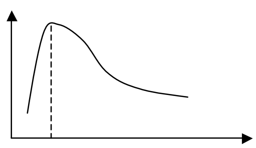
	1.1	1.1
	1.4	1.4
	1.8	1.8
	1.9	1.9
	2.3	2.3
	2.4	**2.4**
	2.8	2.8
	3.1	3.1
	3.4	3.4
	3.8	3.8
	10.3	3.5
Median	**2.4**	**2.4**

Median

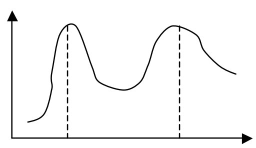
- The median is unaffected by extreme values.
- Obtaining the median:
  - Odd-sized samples: $X_{(n+1)/2}$
  - Even-sized samples: $\dfrac{X_{n/2} + X_{(n/2)+1}}{2}$

Mode

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



unimodal



bimodal

9

Selecting between the mean, mode, and median

- Categorical data
  - Use mode
- Numerical data
  - If the total of all observations is meaningful, use mean
    - E.g. total execution time for five different queries
  - If total not of interest, select median if distribution is skewed, o/w select mean

10

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

Harmonic Mean

- The harmonic mean of a sample $\{x_1, x_2, \ldots, x_n\}$ is defined as

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

- Weighted harmonic mean

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

  where w_i's are weights that add up to 1
- A harmonic mean or weighted harmonic mean should be used whenever an arithmetic mean can be justified for $1/x_i$ (or w_i/x_i)

13

Selecting between arithmetic, geometric and harmonic means

- Controversy (in late 1980s) over which mean to use to characterize the results of benchmarks consisting of a suite of programs
  - See link to article on class home page
- Basic idea: should be guided by physical interpretation of number produced by benchmark
  - Can be confusing if benchmark reports a ratio of two numbers, e.g. floating pt operations and execution time

14

Selecting between means (cont'd)

- If number produced by individual programs in the benchmark is proportional to execution time, then arithmetic mean makes sense to characterize the benchmark suite
- If the inverse of the number produced by individual benchmarks has a physical interpretation, then harmonic mean is appropriate for characterizing the performance of the benchmark suite
  - E.g. if benchmark reports MFLOPs rating of a program, i.e. number of floating pt ops divided by execution time

15

Summarizing variability

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

16

Range, Variance, and Standard Deviation

- Range: $X_{max} - X_{min}$
- Variance: $$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

In Excel:
s²=VAR(<array>)

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

In Excel:
s=STDEV(<array>)

17

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.

18

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

COV not very meaningful if the random variable has a negative or zero mean

19

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:

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

20

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)

21

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)
($0 \le p \le 1$)

22

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

23

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

24

Mean Absolute Deviation

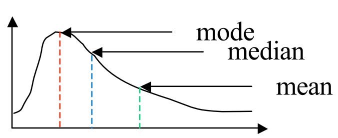
❑ Mean absolute deviation: $\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$

	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

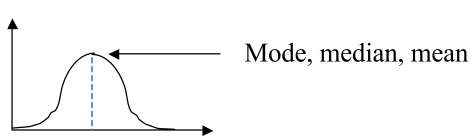
Average	9.51
Mean absolute deviation	13.16

25

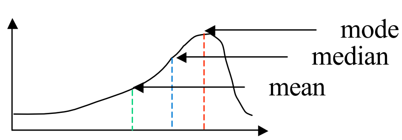
Shapes of Distributions



Right-skewed distribution



Symmetric distribution



Left-skewed distribution

26

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

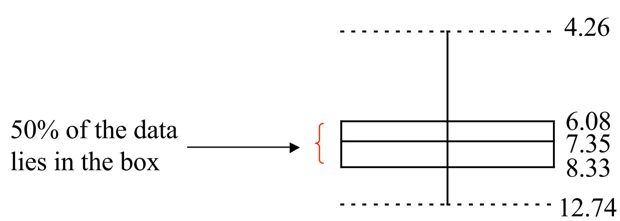
27

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



28

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

29

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.

30

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

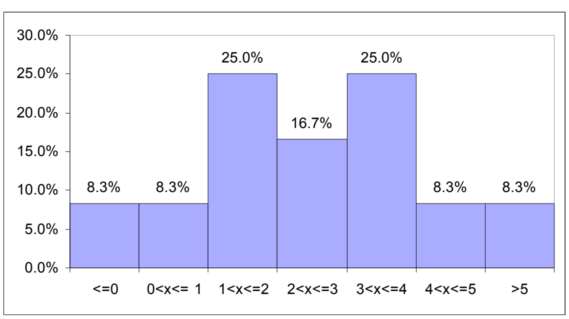
31

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



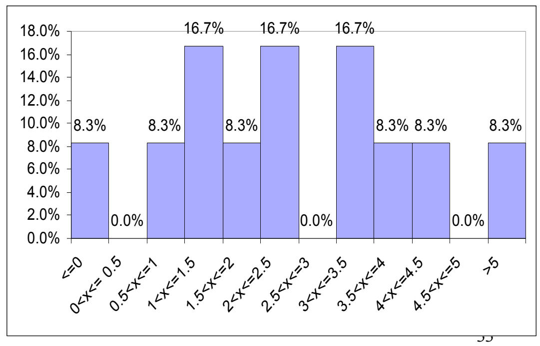
32

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!



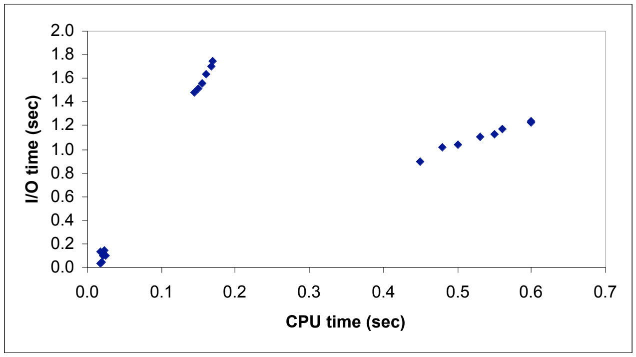
Scatter Plot

- 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.

34

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



35

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

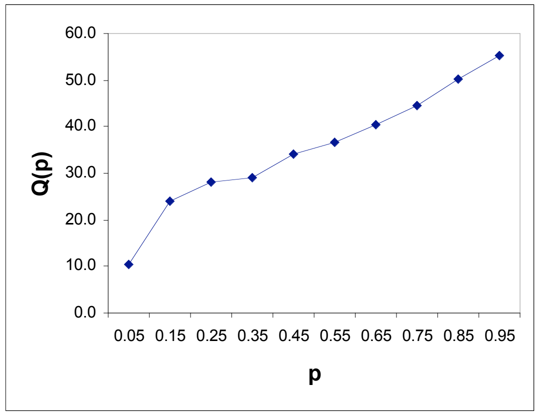
  $$Q(p)= Q([i\text{-}0.5]/n)=x_i$$

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

36

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



37

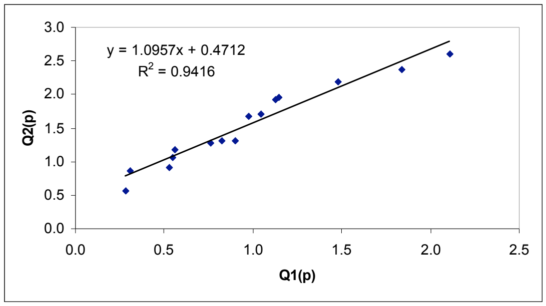
Quantile-Quantile (Q-Q plots)

- 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.

38

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.

39

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.

40

Examples of CDFs and Their Inverse Functions

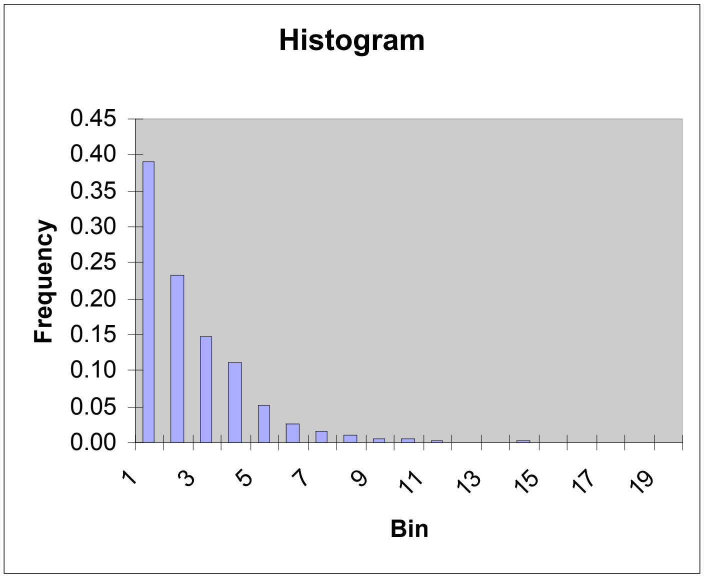
Exponential	$F(x) = 1 - e^{-x/a}$	$-a\mathrm{Ln}(1-u)$
Pareto	$F(x) = 1 - x^{-a}$	$\dfrac{1}{(1-u)^{1/a}}$
Geometric	$F(x) = 1 - (1-p)^{x}$	$\left\lceil \dfrac{\mathrm{Ln}(u)}{\mathrm{Ln}(1-p)} \right\rceil$

41

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.

42

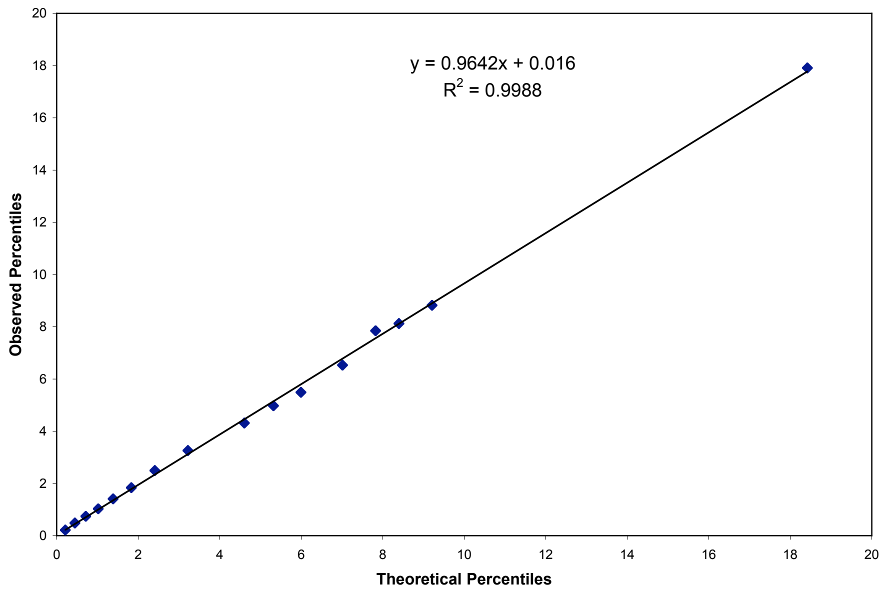


43

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

44



45

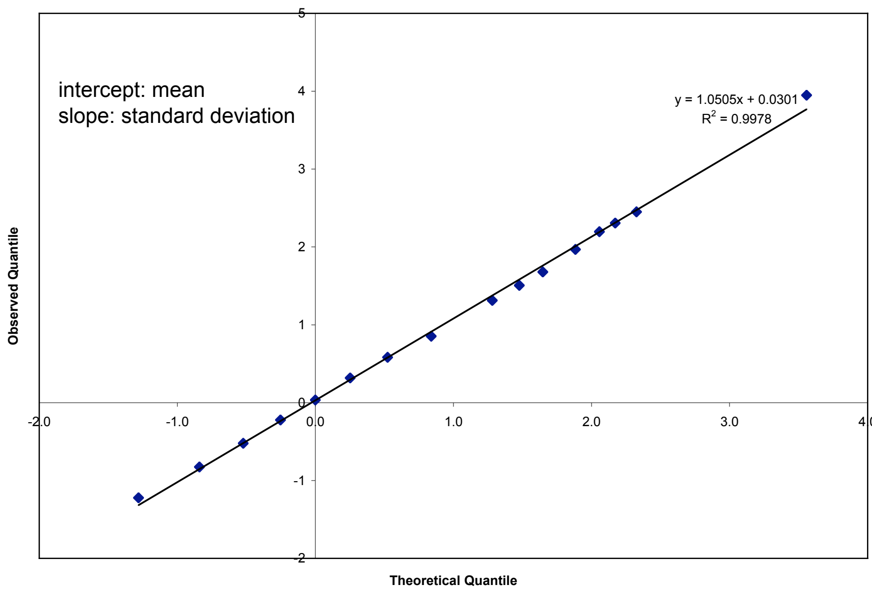
What if the Inverse of the CDF Cannot be Found?

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

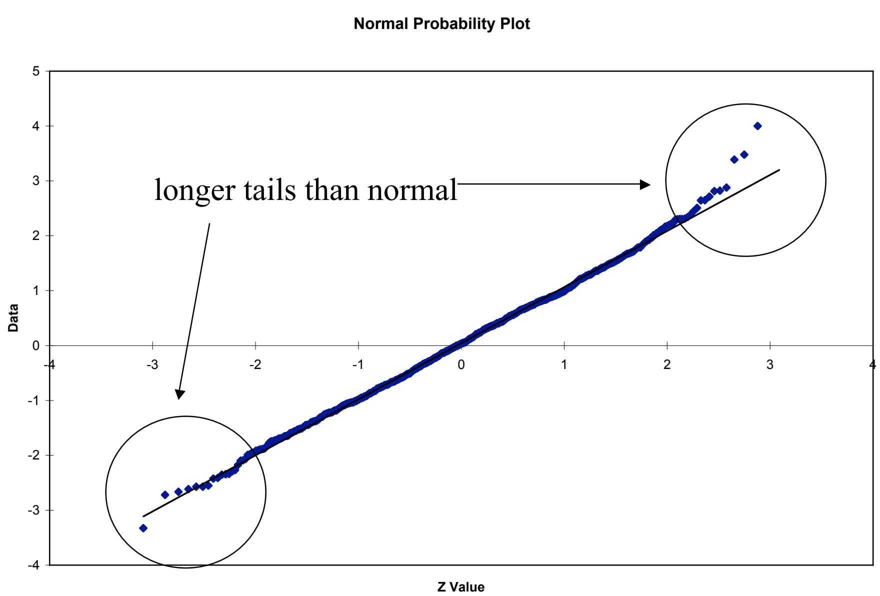
$$x_i = 4.91[q_i^{0.14} - (1-q_i)^{0.14}]$$

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

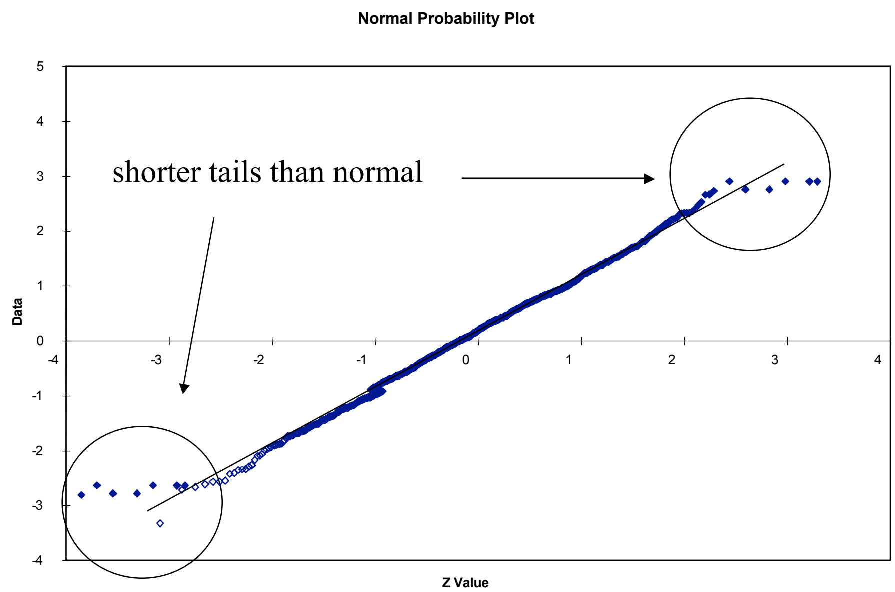
46


intercept: mean
slope: standard deviation
y = 1.0505x + 0.0301
R2 = 0.9978
Observed Quantile
Theoretical Quantile

47

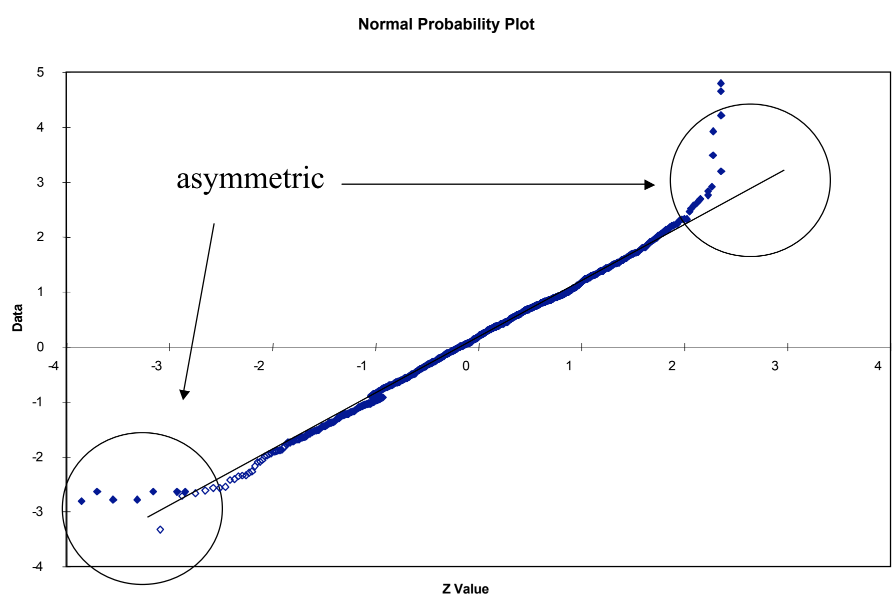

Normal Probability Plot
longer tails than normal
Data
Z Value

48



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

49



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

50