Fitting Distributions and Comparing Data Sets

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

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

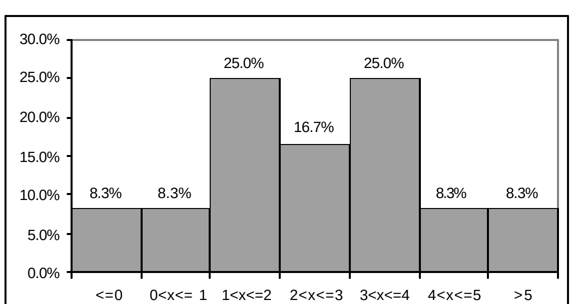
- Divide the range (max value – min value) into equal-sized cells or bins.
- Count the number of data points that fall in each cell.
- Plot on the y-axis the relative frequency, i.e., number of point in each cell divided by the total number of points and the cells on the x-axis.
- Cell size is critical!

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

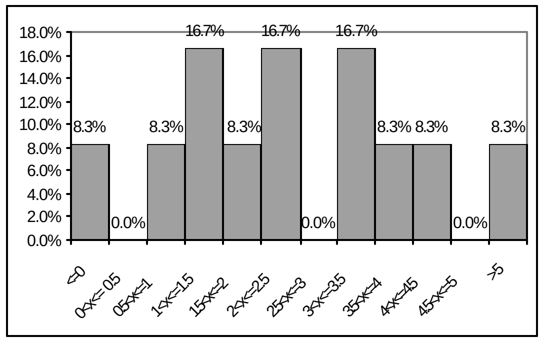


Histogram

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!



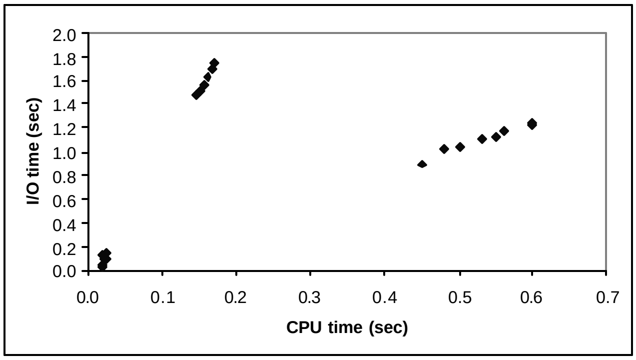
Scatter Plot

- 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.

6

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



Plots Based on Quantiles

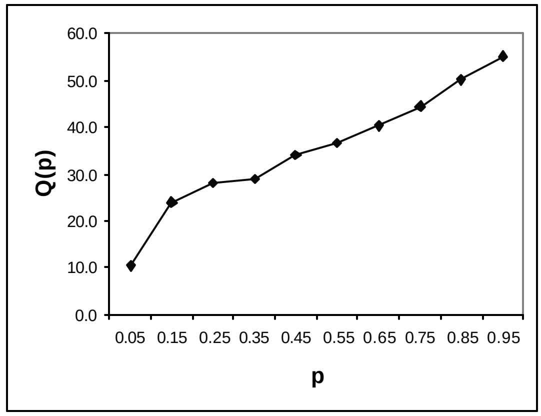
- Consider an ordered data set with n values $x_1, \ldots, x_n$.
- If $p = (i\text{-}0.5)/n$ for $i \leq n$, then the p quantile Q(p) of the data set is defined as

$$Q(p) = Q([i-0.5]/n) = x_i$$

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

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



9

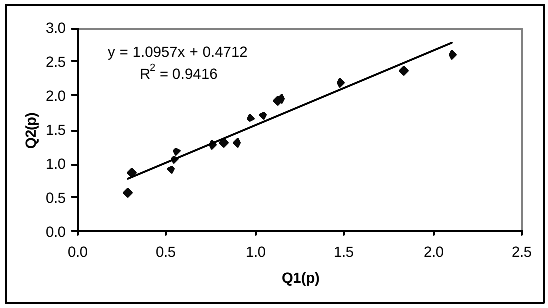
Quantile-Quantile (Q-Q plots)

- 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 $(i-0.5)/n$ where n is the size of the smaller data set.

10

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.

11

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.

12

Examples of CDFs and Their Inverse Functions

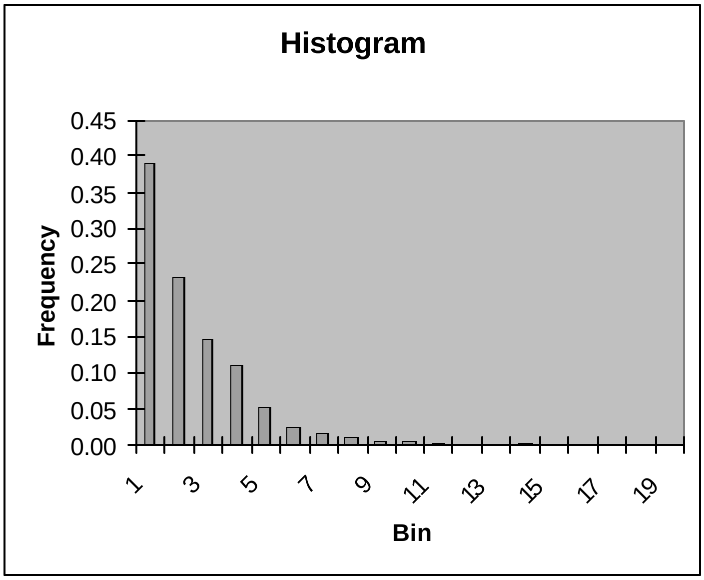
Exponential	$F(x) = 1 - e^{-x/a}$	$-a\mathrm{Ln}(1-u)$
Pareto	$F(x) = 1 - x^{-a}$	$\dfrac{1}{(1-u)^{1/a}}$
Geometric	$F(x) = 1 - (1-p)^{x}$	$\left\lceil \dfrac{\mathrm{Ln}(u)}{\mathrm{Ln}(1-p)} \right\rceil$

13

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.

14

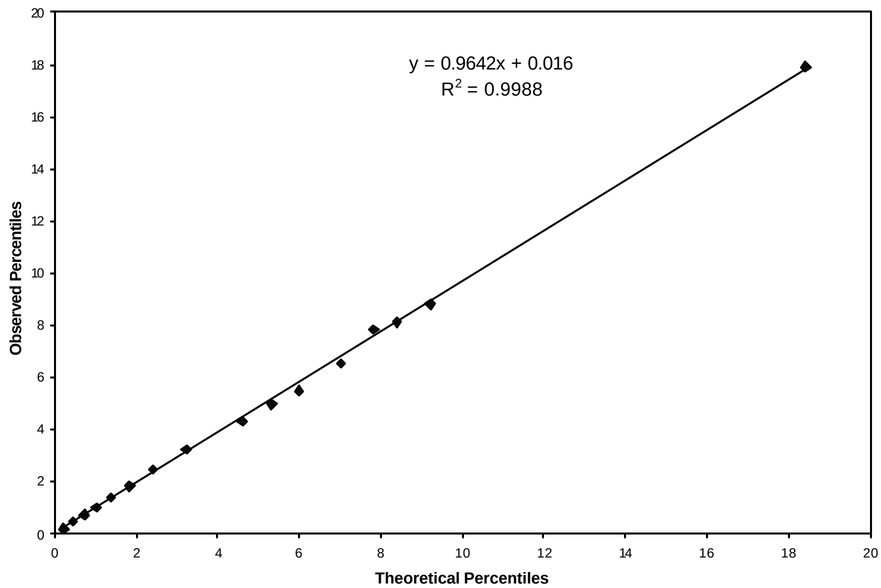


15

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

16



17

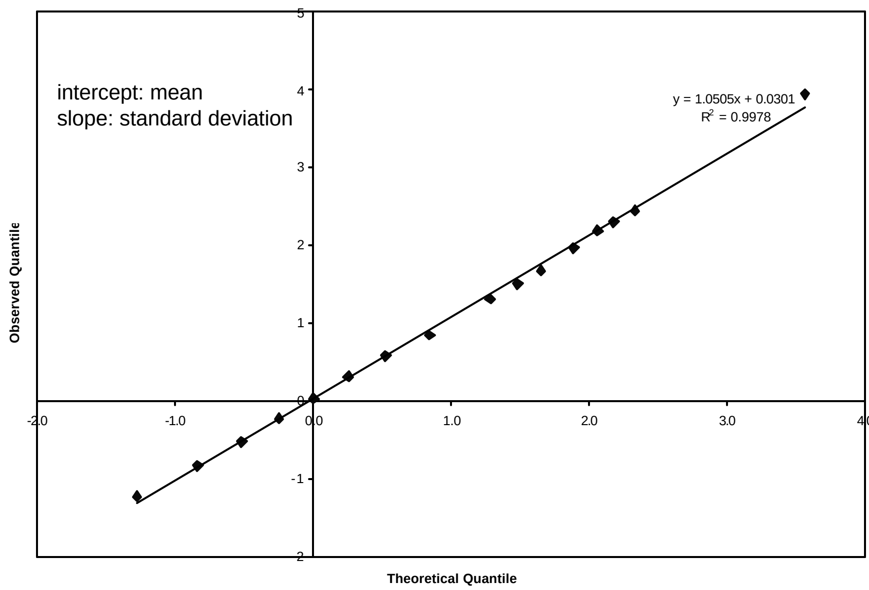
What if the Inverse of the CDF Cannot be Found?

- Use tables and interpolate.
- Approximation for N(0,1):

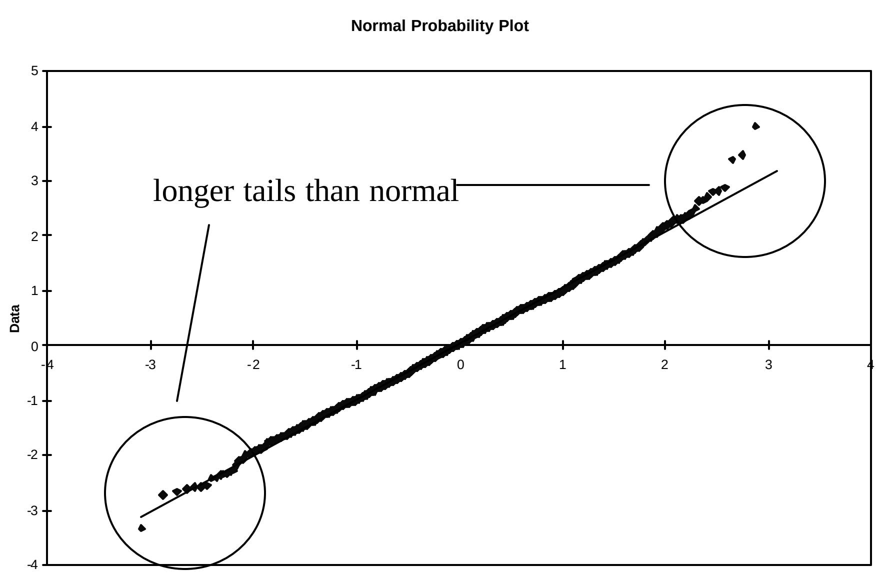
$$x_i = 4.91[q_i^{0.14} - (1 - q_i)^{0.14}]$$

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

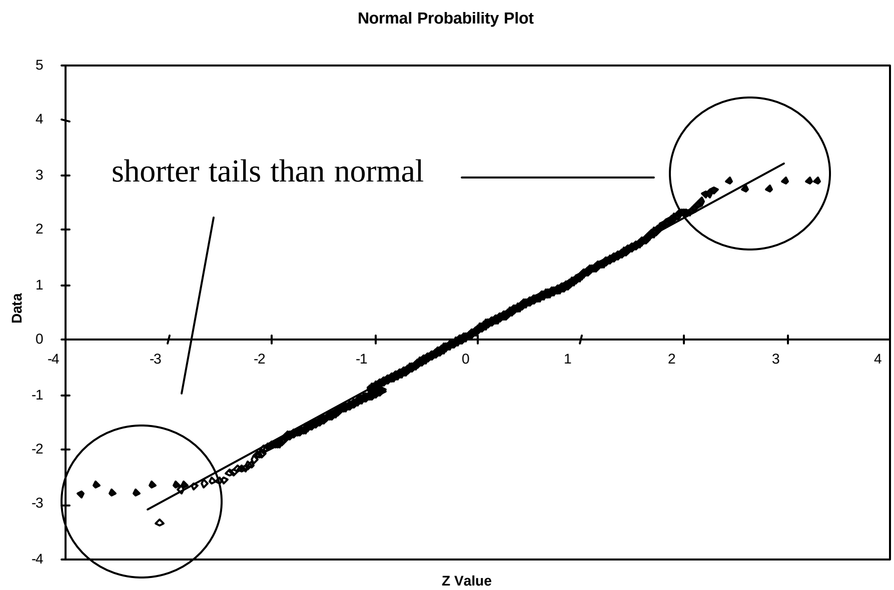
18


intercept: mean
slope: standard deviation
y = 1.0505x + 0.0301
R^2 = 0.9978
Observed Quantile
Theoretical Quantile

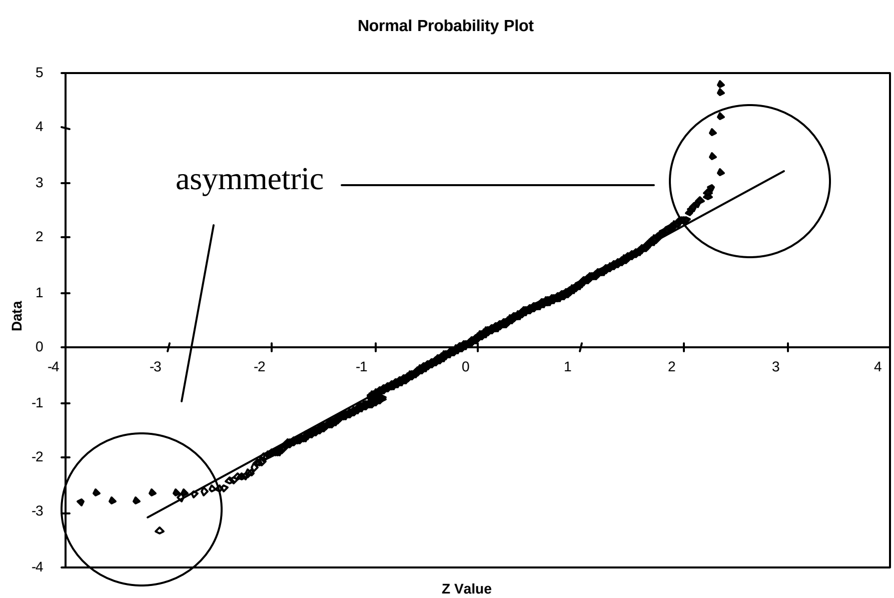
19


Normal Probability Plot
longer tails than normal
Data
Z Value

20


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

21


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

22