Statistical Learning

Measures of central tendency, dispersion and correlation

Outline

1. Raw Data
2. Frequency Distribution - Histograms
3. Cumulative Frequency Distribution
4. Measures of Central Tendency
5. Mean, Median, Mode
6. Measures of Dispersion
7. Range, IQR, Standard Deviation,coefficient of variation
8. Normal distribution, Chebyshev Rule.
9. Five number summary, boxplots, QQ plots, Quantile plot, scatter plot.
10. Visualization: scatter plot matrix, parallel coordinates.
11. Correlation analysis

Data versus Information

When managers are bewildered by plethora of data, which do not make any sense on the surface of it, they are looking for methods to classify data that would convey meaning. The idea here is to help them draw the right conclusion. This session provides the nitty-gritty of arranging data into information.

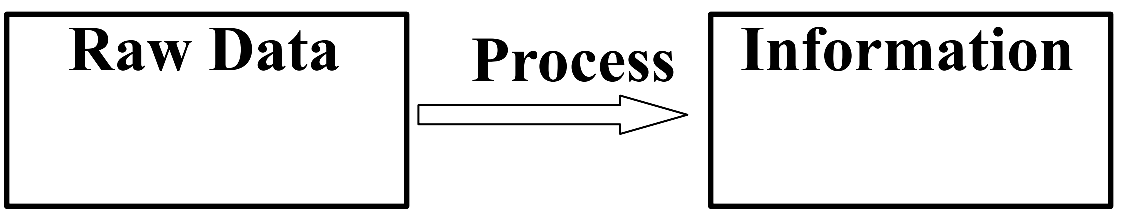
Raw Data

Meaning of Raw Data:

Raw Data represent numbers and facts in the original format in which the data have been Collected. You need to convert the raw data into information for managerial decision Making.

Information is Key

Large and massive raw data tend to bewilder you so much that the overall patterns are obscured. You cannot see the wood for the trees. This implies that the raw data must be processed to give you useful information.

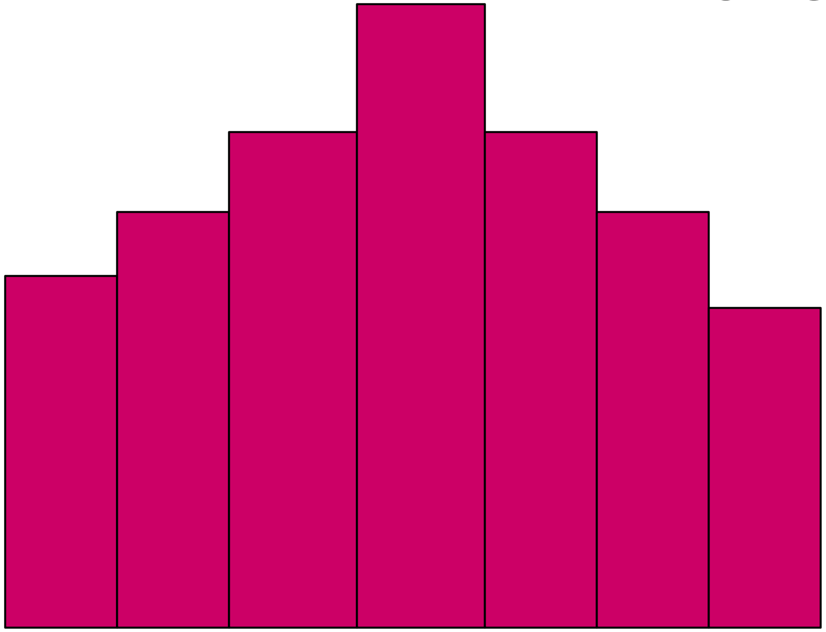


Frequency Distribution

In simple terms, frequency distribution is a summarized table in which raw data are arranged into classes and frequencies.

Frequency distribution focuses on classifying raw data into information. It is the most widely used data reduction technique in descriptive statistics.

HISTOGRAM



Histogram (also known as frequency histogram) is a snap shot of the frequency distribution.

Histogram is a graphical representation of the frequency distribution in which the X-axis represents the classes and the Y-axis represents the frequencies in bars

Histogram depicts the pattern of the distribution emerging from the characteristic being measured.

Histogram- Example

The inspection records of a hose assembly operation revealed a high level of rejection. An analysis of the records showed that the "leaks" were a major contributing factor to the problem. It was decided to investigate the hose clamping operation. The hose clamping force (torque) was measured on twenty five assemblies. (Figures in foot-pounds). The data are given below: Draw the frequency histogram and comment.

8	13	15	10	16
11	14	11	14	20
15	16	12	15	13
12	13	16	17	17
14	14	14	18	15

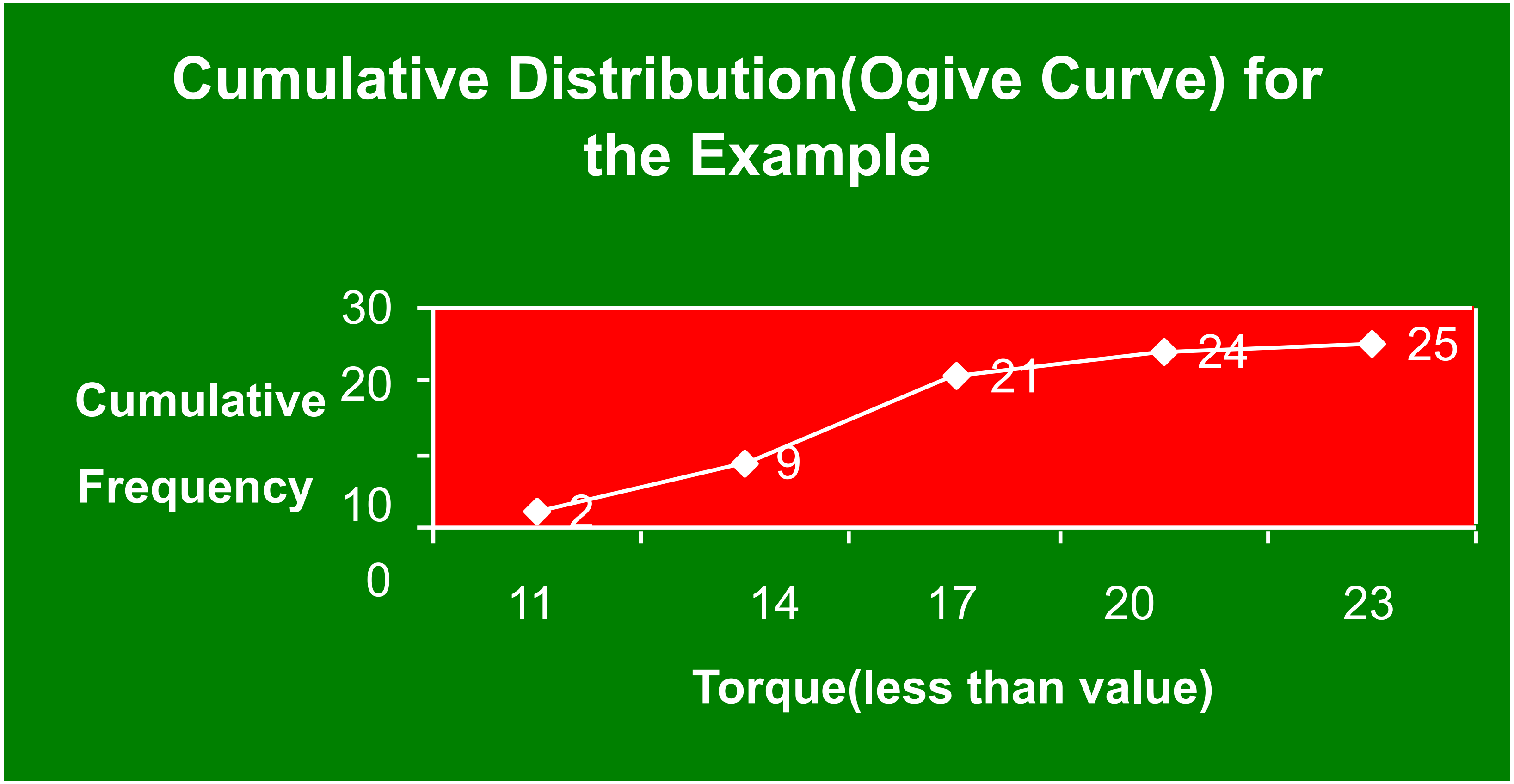
Histogram Example Solution

Cumulative Frequency Distribution

A type of frequency distribution that shows how many observations are above or below the lower boundaries of the classes. You can formulate the following from the previous example of hose clamping force(torque)

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
8-11	2	0.08	2	0.08
11-14	7	0.28	9	0.36
14-17	12	0.48	21	0.84
17-20	3	0.12	24	0.96
20-23	1	0.04	25	1.00
Total	**25**	**1.00**		

Ogive (Cumulative Frequency Distribution



What is Central Tendency?

Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. Such a value is called a measure of "Central Tendency". The other terms that are used synonymously are "Measures of Location", or "Statistical Averages".

Measures of Central Tendency

As a manager, You need the summary measures of central tendency to draw meaningful conclusions in your functional area of operation. The most widely used measures of central tendency are **Arithmetic Mean**, **Median**, and **Mode**.

Arithmetic Mean

Arithmetic Mean (called mean) is defined as the sum of all observations in a data set divided by the total number of observations. For example, consider a data set containing the following observations:

In symbolic form mean is given by $\overline{X} = \frac{\sum X}{n}$

$\overline{X}$ = Arithmetic Mean

$\sum X$ = Indicates sum all X values in the data set

n = Total number of observations(Sample Size)

Arithmetic Mean -Example

The inner diameter of a particular grade of tire based on 5 sample measurements are as follows:(figures in millimeters)

565, 570, 572, 568, 585

Applying the formula $\overline{X} = \frac{\sum X}{n}$

We get mean = (565+570+572+568+585)/5 =572

Caution: Arithmetic Mean is affected by extreme values or fluctuations in sampling. It is not the best average to use when the data set contains extreme values (Very high or very low values).

Mean Calculation - Problem 1

Let's say you own a piece of art(worth 90,000$) that increases in value by 50% the first year after you buy it, 20% the second year, and 90% the third year. Calculate the mean increase in the value of the art piece.

Solution

What these numbers tell you is that at the end of the first year the value was multiplied by 150% or 1.5, the second year the value at the end of year 1 was multiplied by 120% or 1.2 and at the end of the third year the value at the end of year 2 was multiplied by 190% or 1.9. As these are *multiplied*, what you are looking for is the geometric mean which can be calculated in the following way:
(1.5*1.2*1.9)^(1/3) = 1.50663725458… or about 1.51

What the answer of 1.51 is telling you is that if you multiplied your initial investment by 1.51 each year, you would get the same amount as if you had multiplied it by 1.5, 1.2 and 1.9.

- Art work value year 0: $90,000.
- Art work value year 1: $90,000 * 1.5 = $135,000
- Art work value year 2: $135,000 * 1.2 = $162,000
- Art work value year 3: $90,000 * 1.9 = $307,000

or, using the geometric mean:
$90,000 * 1.506637254583 = $307,000*

Geometric Mean

The geometric mean is a type of average , usually used for growth rates, like population growth or interest rates. While the arithmetic mean **adds** items, the geometric mean **multiplies** items. Also, you can only get the geometric mean for positive numbers.

Like most things in math, there's an easy explanation, and there's a more, *mathematical* way of stating the same thing. Formally, the geometric mean is defined as "…the nth root of the product of n numbers."

Mean Calculation - Problem 2

The distance from my house to town is 40 km. I drove to town at a speed of 40 km per hour and returned home at a speed of 80 km per hour. What was my average speed for the whole trip?

Solution

Solution: If we just took the arithmetic mean of the two speeds I drove at, we would get 60 km per hour. This isn't the correct average speed, however: it ignores the fact that I drove at 40 km per hour for twice as long as I drove at 80 km per hour. To find the correct average speed, we must instead calculate the harmonic mean.

Solution

For two quantities A and B, the harmonic mean is given by: $\frac{2}{\frac{1}{A}+\frac{1}{B}}$

This can be simplified by adding in the denominator and multiplying by the reciprocal: $\frac{2}{\frac{1}{A}+\frac{1}{B}} = \frac{2}{\frac{B+A}{AB}} = \frac{2AB}{A+B}$

For N quantities: A, B, C......

Harmonic mean $= \frac{N}{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\dots}$

Let us try out the formula above on our example:

Harmonic mean $= \frac{2AB}{A+B}$

Our values are A = 40, B = 80. Therefore, harmonic mean $= \frac{2\times40\times80}{40+80} = \frac{6400}{120} \approx 53.333$

Is this result correct? We can verify it. In the example above, the distance between the two towns is 40 km. So the trip from A to B at a speed of 40 km will take 1 hour. The trip

from B to A at a speed to 80 km will take 0.5 hours. The total time taken for the round distance (80 km) will be 1.5 hours. The average speed will then be $\frac{80}{1.5} \approx 53.33$ km/hour.

The harmonic mean also has physical significance.

What is the Harmonic Mean?

The harmonic mean is a very specific type of average. It's generally used when dealing with averages of units, like speed or other rates and ratios.

The formula is:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Median

Median is the middle most observation when you arrange data in ascending order of magnitude. Median is such that 50% of the observations are above the median and 50% of the observations are below the median.

Median is a very useful measure for ranked data in the context of consumer preferences and rating. It is not affected by extreme values (greater resistance to outliers)

$$\text{Median} = \frac{n+1}{2} \text{ th value of ranked data}$$

n = Number of observations in the sample

Median - Example

Marks obtained by 7 students in Computer Science Exam are given below: Compute the median.

45 40 60 80 90 65 55

Arranging the data after ranking gives

90 80 65 60 55 45 40

Median = (n+1)/2 th value in this set = (7+1)/2 th observation= 4^{th} observation=60

Hence Median = 60 for this problem.

Mode

Mode is that value which occurs most often. It has the maximum frequency of occurrence. Mode also has resistance to outliers.

Mode is a very useful measure when you want to keep in the inventory, the most popular shirt in terms of collar size during festival season.

Mode -Example

The life in number of hours of 10 flashlight batteries are as follows: Find the mode.

340 350 340 340 320 340 330 330 340 350

340 occurs five times. Hence, mode=340.

Comparison of Mean, Median, Mode

Mean	Median	Mode
Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.
Requires measurement on all observations.	Does not require measurement on all observations	Does not require measurement on all observations
Uniquely and comprehensively defined.	Cannot be determined under all conditions.	Not uniquely defined for multi-modal situations.

Comparison of Mean, Median, Mode Cont.

Mean	Median	Mode
Affected by extreme values. Can be treated algebraically. That is, Means of several groups can be combined.	Not affected by extreme values. Cannot be treated algebraically. That is, Medians of several groups cannot be combined.	Not affected by extreme values. Cannot be treated algebraically. That is, Modes of several groups cannot be combined.

Measures of Dispersion

In simple terms, measures of dispersion indicate how large the spread of the distribution is around the central tendency. It answers unambiguously the question " What is the magnitude of departure from the average value for different groups having identical averages?".

Range

Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in the data set.

$$\text{Range} = X_{\text{Maximum}} - X_{\text{Minimum}}$$

Range-Example

Example for Computing Range

The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate Range.

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

Range = $X_{Maximum} - X_{Minimum}$ = 18-9=9

Caution: If one of the components of range namely the maximum value or minimum value becomes an extreme value, then range should not be used.

Inter-Quartile Range(IQR)

IQR= Range computed on middle 50% of the observations after eliminating the highest and lowest 25% of observations in a data set that is arranged in ascending order. IQR is less affected by outliers.

$IQR = Q_3 - Q_1$

Interquartile Range-Example

The following data represent the return on percentage investment for 9 mutual funds per annum. Calculate interquartile range.

Data Set: 12, 14, 11, 18, 10.5, 12, 14, 11, 9

Arranging in ascending order, the data set becomes

9, 10.5, 11, 11, 12, 12, 14, 14, 18

$IQR = Q_3 - Q_1 = 14 - 10.75 = 3.25$

Standard Deviation

Standard deviation forms the cornerstone For Inferential Statistics.

To define standard deviation, you need to define another term called variance. In simple terms, standard deviation is the square root of variance.

Key Formulas

Important Terms with Notations

Sample Variance

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1}$$

Sample Standard Deviation

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Population Variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Population Standard Deviation $\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$

Where $\overline{X} = \frac{\sum X}{n}$ (Sample Mean) and

$\mu = \frac{\sum X}{N}$ (Population Mean)

n =Number of observations in the sample(Sample size)
N =Number of observations in the Population (Population Size)

Remarks

1. $S^2 = \frac{\sum (X - \overline{X})^2}{n-1}$ is an unbiased estimator of $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
2. $\overline{X} = \frac{\sum X}{n}$ is an unbiased estimator of $\mu = \frac{\sum X}{N}$
3. The divisor n-1 is always used while calculating sample variance for ensuring property of being unbiased
4. Standard deviation is always the square root of variance

Example for Standard Deviation

The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate the sample standard deviation.

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

Solution for the Example

A	B	C	D
1			
2	X	$X - \overline{X}$	$(X - \overline{X})^2$
3	12	-0.28	0.08
4	14	1.72	2.96
5	11	-1.28	1.64
6	18	5.72	32.72
7	10.5	-1.78	3.17
8	11.3	-0.98	0.96
9	12	-0.28	0.08
10	14	1.72	2.96
11	11	-1.28	1.64
12	9	-3.28	10.76
13	Mean =		56.96
14	12.28	Variance=	6.33
15		Standard Deviation=	2.52

Coefficient of Variation (Relative Dispersion)

Coefficient of Variation (CV) is defined as the ratio of Standard Deviation to Mean.

In symbolic form

$CV = \frac{S}{\bar{X}}$ for the sample data and $= \frac{\sigma}{\mu}$ for the population data.

Coefficient of Variation Example

Consider two Sales Persons working in the same territory. The sales performance of these two in the context of selling PCs are given below. Comment on the results.

Sales Person 1	Sales Person 2
Mean Sales (One year average) 50 units	Mean Sales (One year average)75 units
Standard Deviation 5 units	Standard deviation 25 units

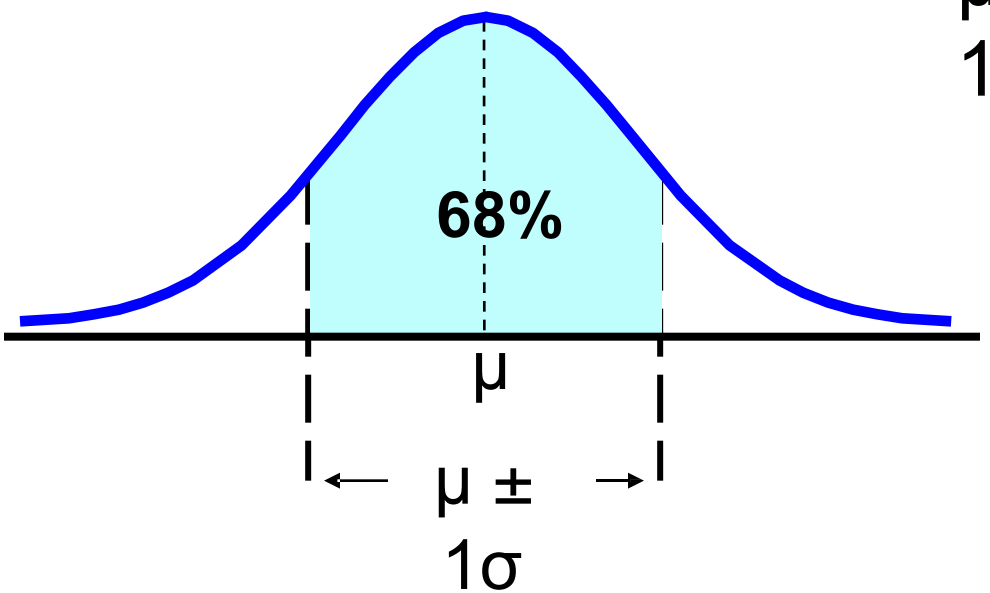
Interpretation for the Example

The CV is 5/50 =0.10 or 10% for the Sales Person1 and 25/75=0.33 or 33% for sales Person2.

The moral of the story is "don't get carried away by absolute number". Look at the scatter. Even though, Sales Person2 has achieved a higher average, his performance is not consistent and seems erratic.

The Empirical Rule

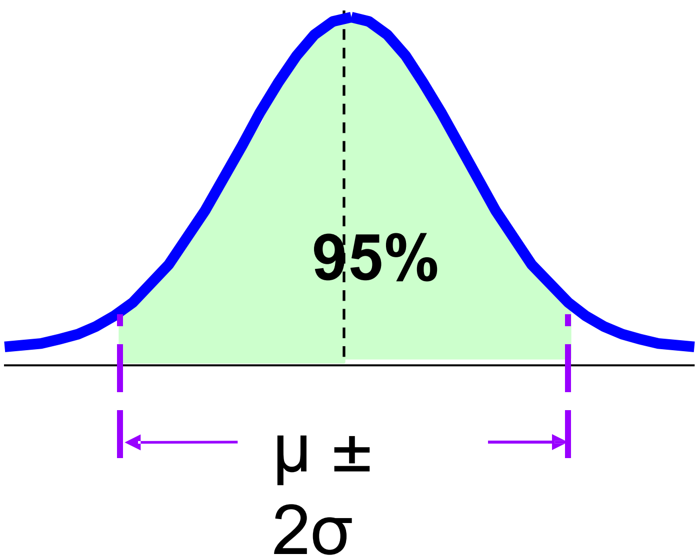
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$

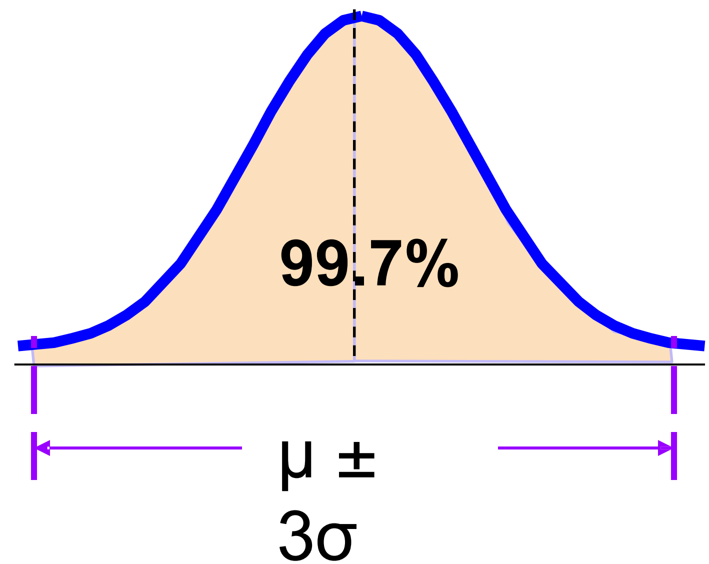


The Empirical Rule

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$

- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$





Chebyshev Rule

- Regardless of how the data are distributed, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for $k > 1$)

- For Example, when $k=2$, at least 75% of the values of any data set will be within $\mu \pm 2\sigma$

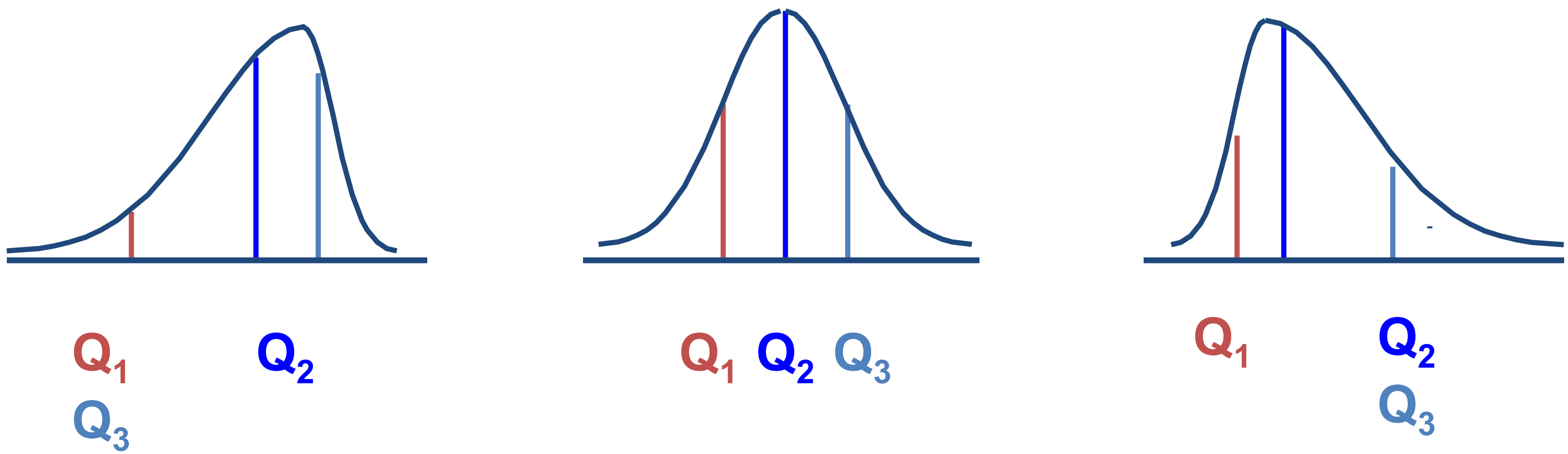
The Five Number Summary

The five numbers that help describe the center, spread and shape of data are:

- $X_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $X_{largest}$

Distribution Shape

Left-Skewed Symmetric Right-Skewed



Relationships among the five-number summary and distribution shape

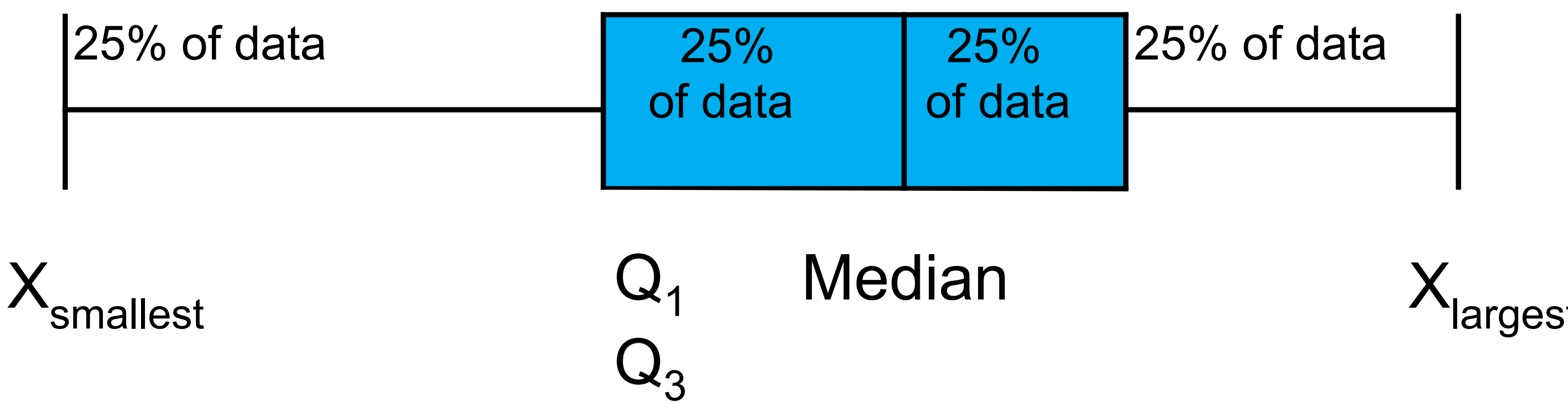
Left-Skewed	Symmetric	Right-Skewed
Median – $X_{smallest}$ > $X_{largest}$ – Median	Median – $X_{smallest}$ ≈ $X_{largest}$ – Median	Median – $X_{smallest}$ < $X_{largest}$ – Median
Q_1 – $X_{smallest}$ > $X_{largest}$ – Q_3	Q_1 – $X_{smallest}$ ≈ $X_{largest}$ – Q_3	Q_1 – $X_{smallest}$ < $X_{largest}$ – Q_3
Median – Q_1 > Q_3 – Median	Median – Q_1 ≈ Q_3 – Median	Median – Q_1 < Q_3 – Median

Five Number Summary and The Boxplot

- The Boxplot: A Graphical display of the data based on the five-number summary:

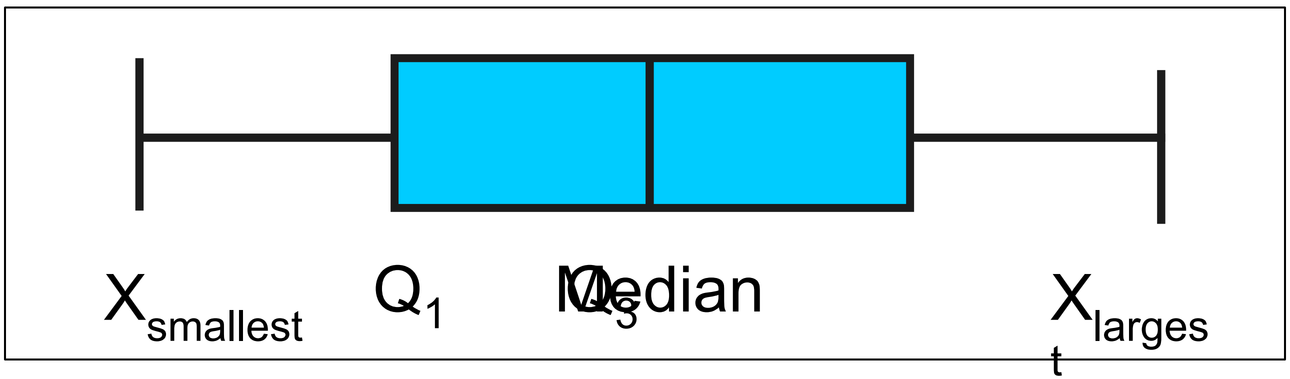
$X_{smallest}$ -- Q_1 -- Median $X_{largest}$

Example: -- Q_3 --



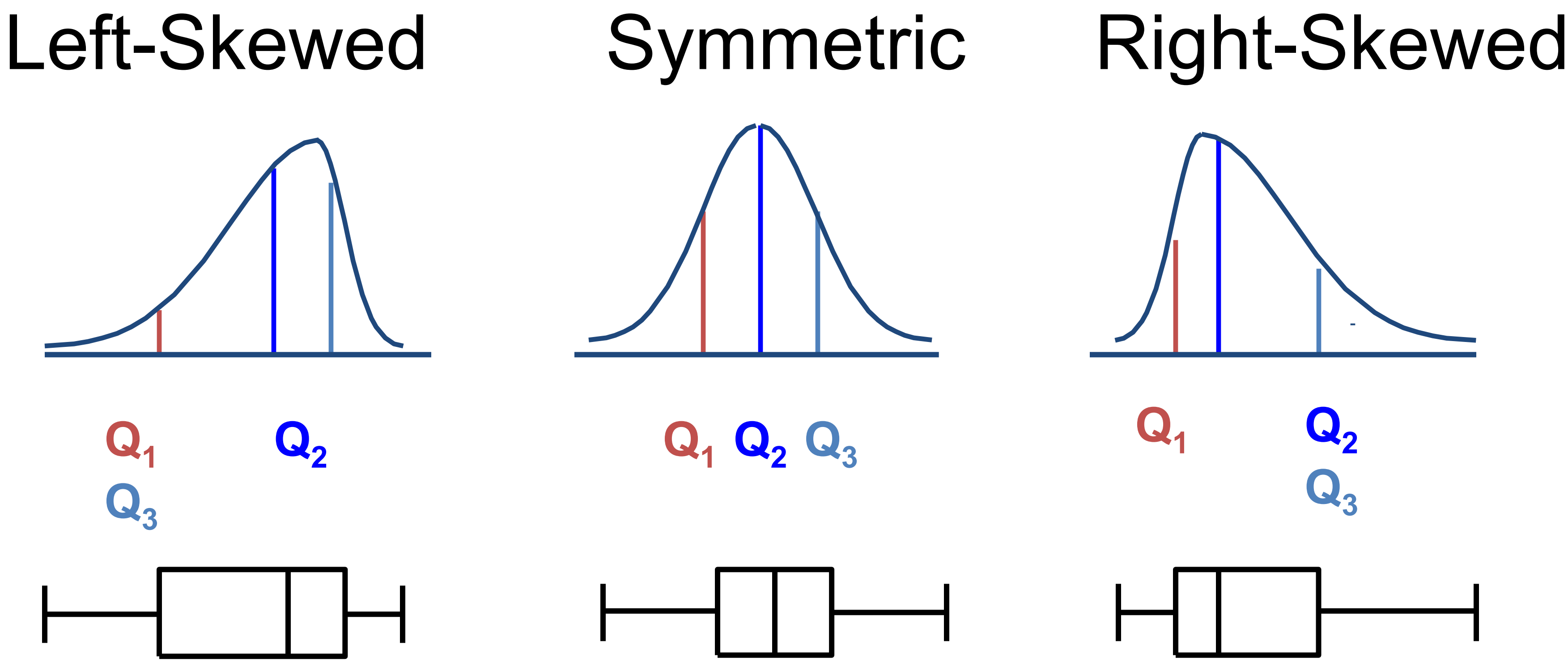
Five Number Summary: Shape of Boxplots

- If data are symmetric around the median then the box and central line are centered between the endpoints



- A Boxplot can be shown in either a vertical or horizontal orientation

Distribution Shape and The Boxplot



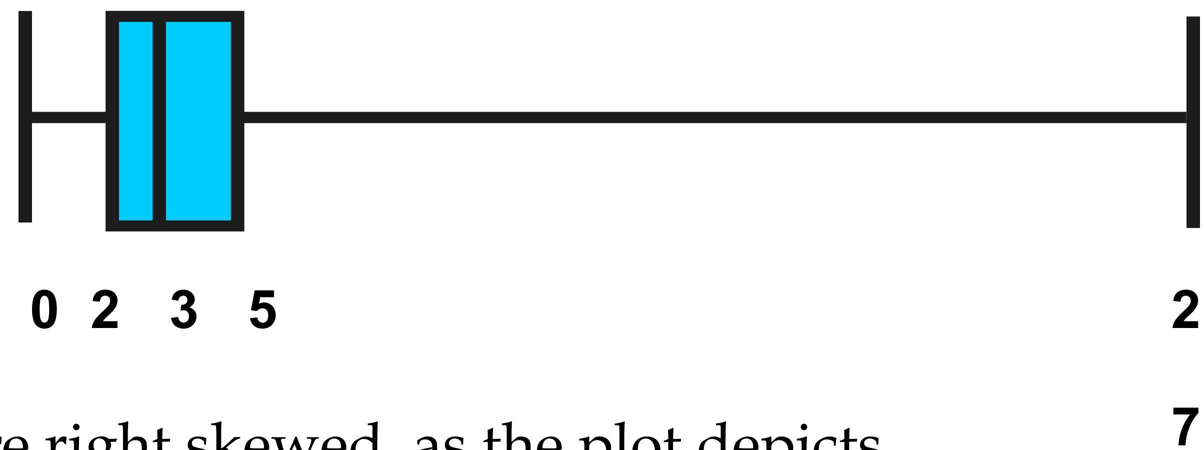
Boxplot Example

Below is a Boxplot for the following data:

$X_{smallest}$ Q_1 Q_2 Q_3 $X_{largest}$

0 2 2 3 3 27

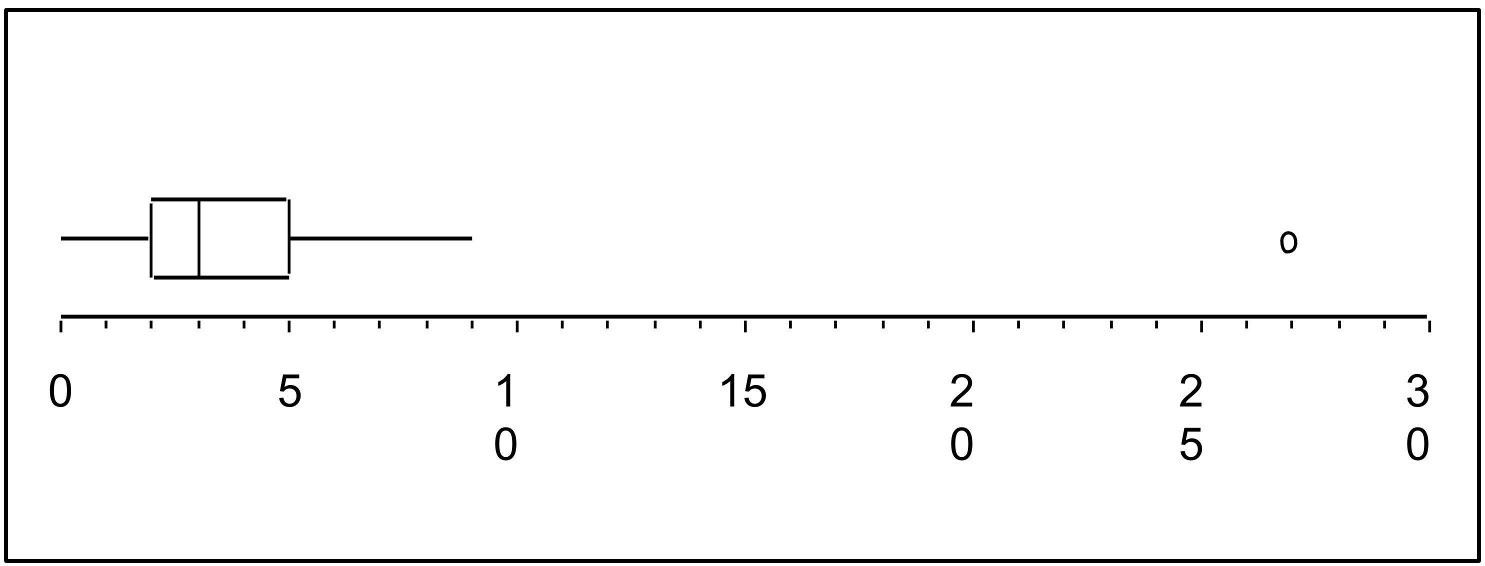
2 4 5 5 9



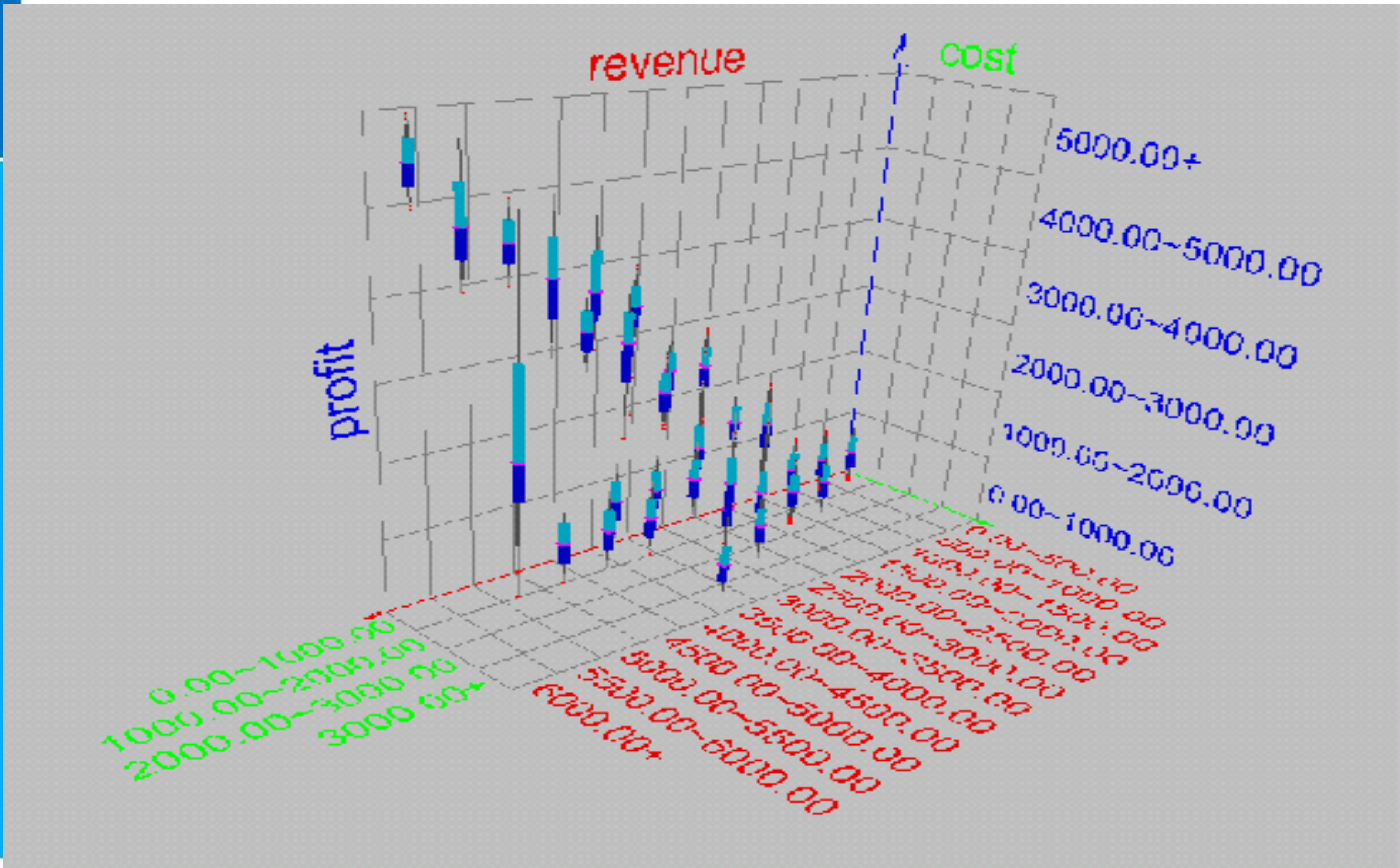
The data are right skewed, as the plot depicts

Box plot example showing an outlier

- The boxplot below of the same data shows the outlier value of 27 plotted separately
- A value is considered an outlier if it is more than 1.5 times the interquartile range below Q_1 or above Q_3



Visualization of Data Dispersion: 3-D Boxplots



Graphic Displays of Basic Statistical Descriptions

Boxplot: graphic display of five-number summary

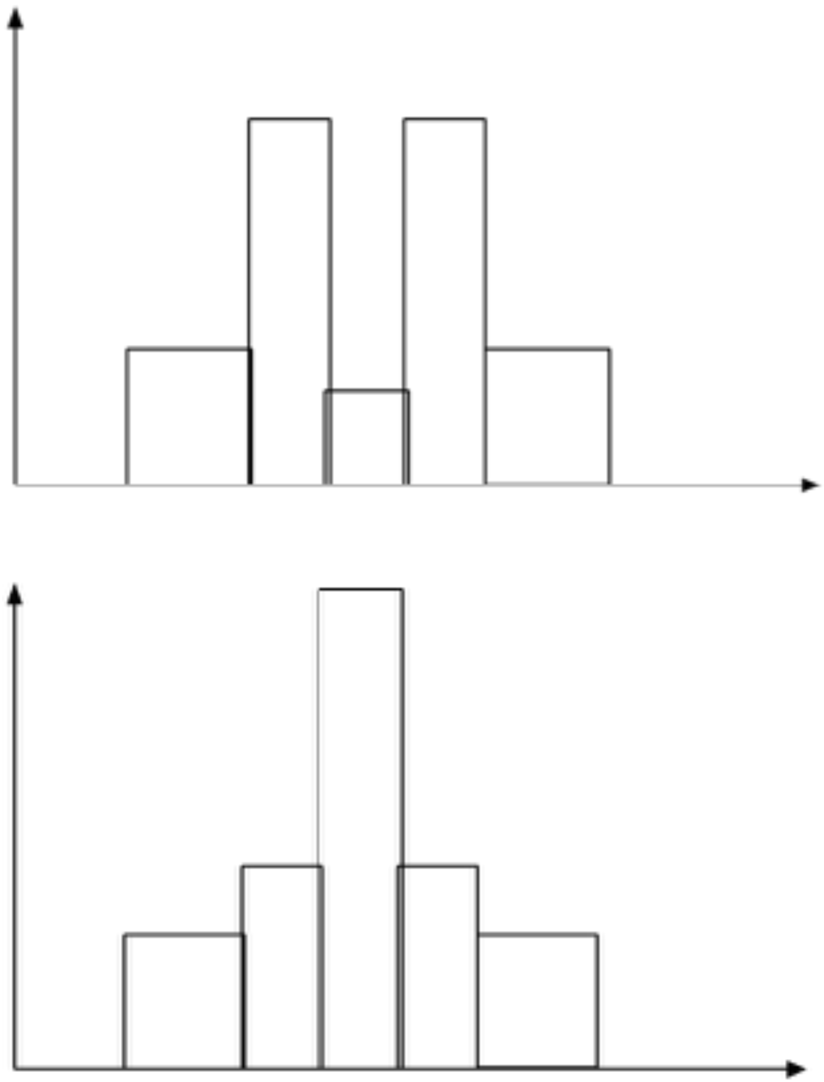
Histogram: x-axis are values, y-axis repres. frequencies

Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are ≤ x_i

Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another

Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

Histograms Often Tell More than Boxplots



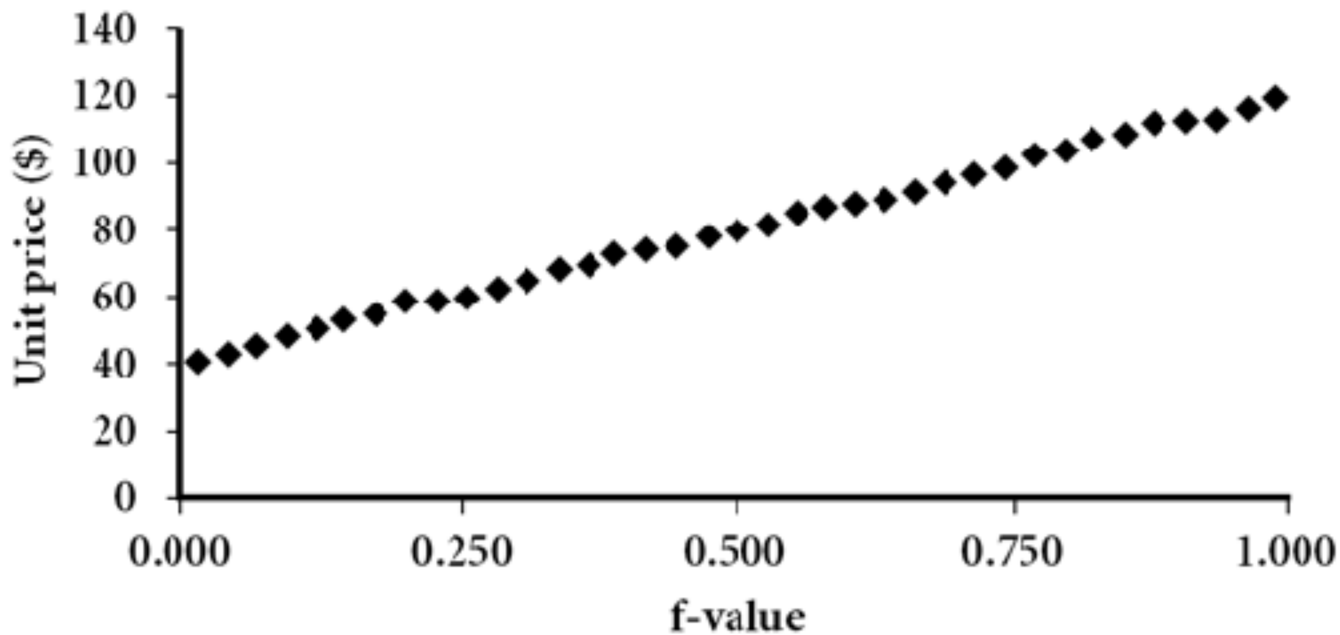
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Quantile Plot

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

Plots **quantile** information

- For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to t

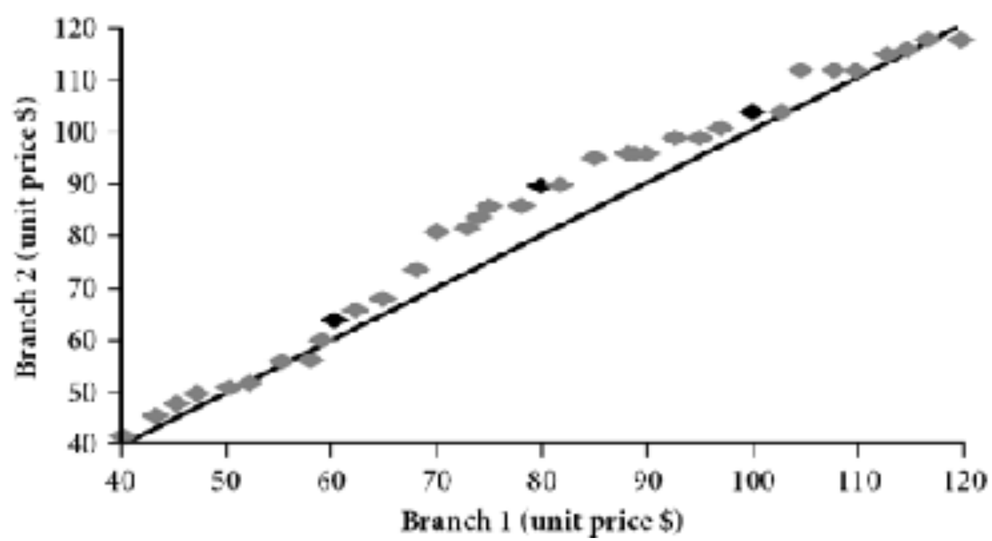


Quantile-Quantile (Q-Q) Plot

Graphs the quantiles of one univariate distribution against the corresponding quantiles of another

View: Is there is a shift in going from one distribution to another?

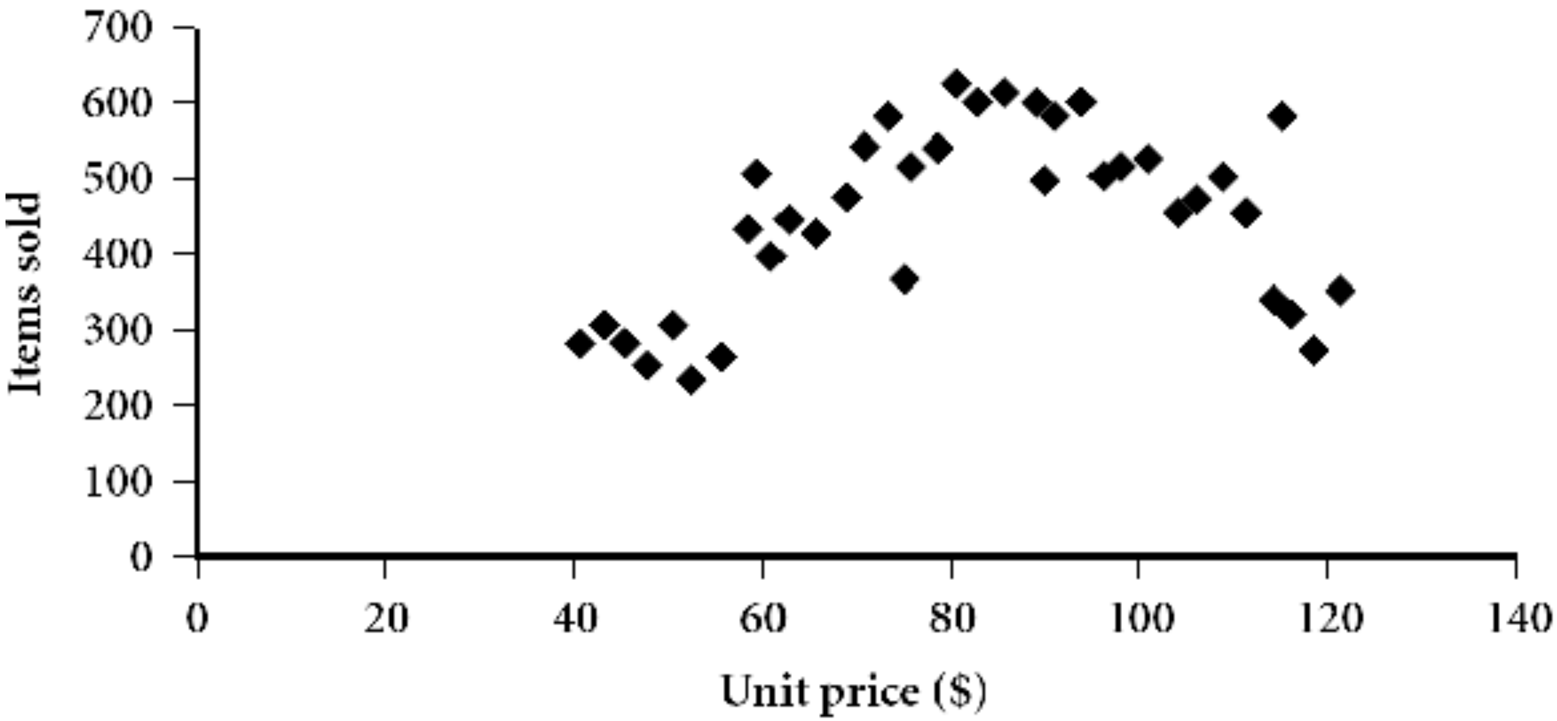
Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



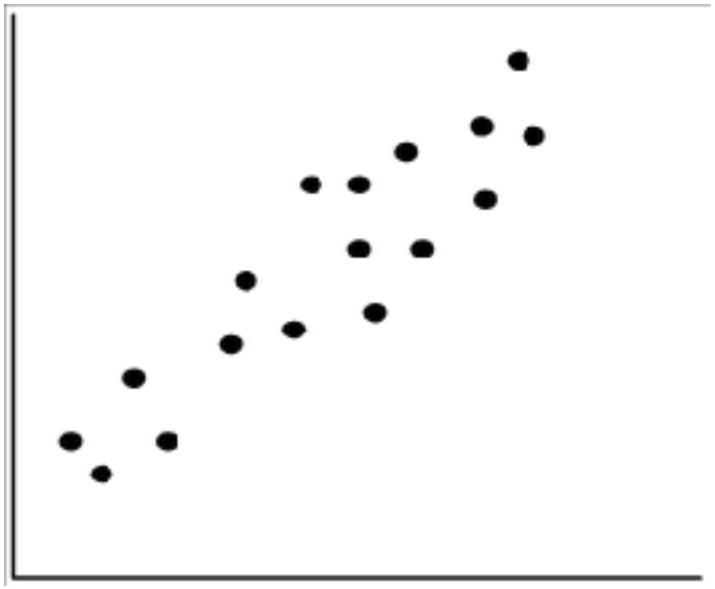
Scatter plot

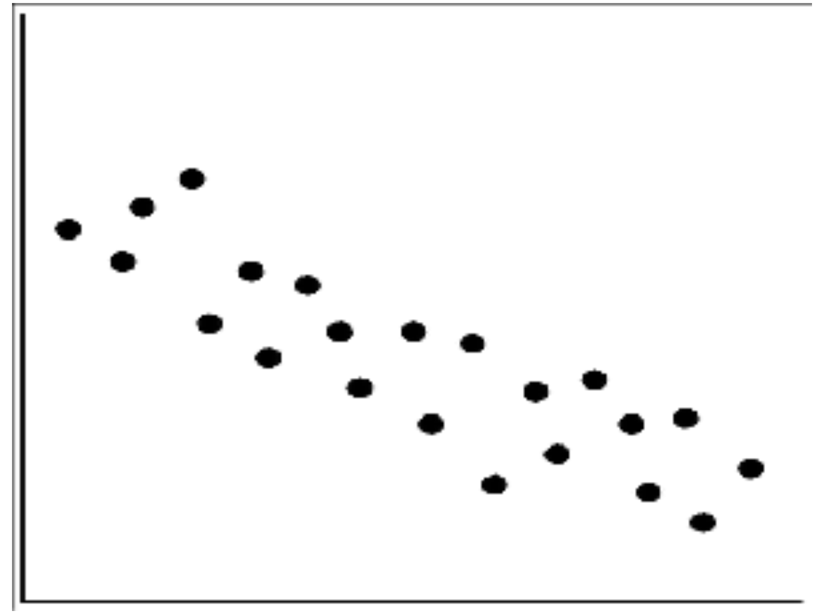
Provides a first look at bivariate data to see clusters of points, outliers, etc

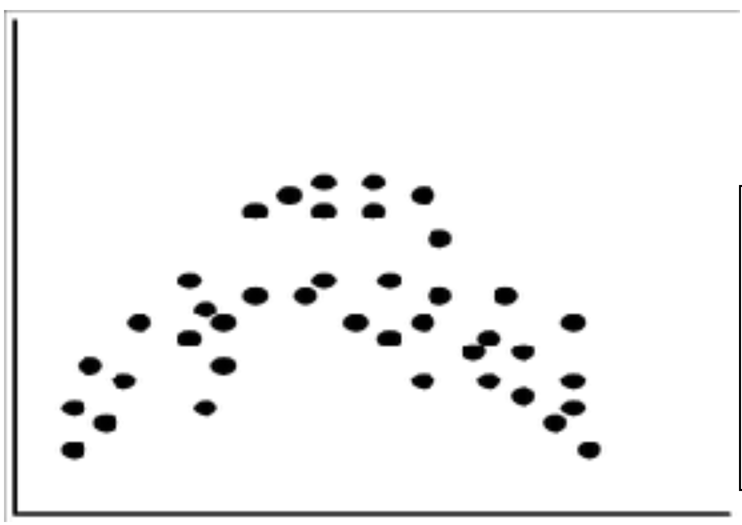
Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data







The left half fragment is positively correlated

The right half is negative correlated

Uncorrelated Data

Data Visualization

Why data visualization?

- Gain insight into an information space by mapping data onto graphical primitives
- Provide qualitative overview of large data sets
- Search for patterns, trends, structure, irregularities, relationships among data
- Help find interesting regions and suitable parameters for further quantitative analysis
- Provide a visual proof of computer representations derived

Categorization of visualization methods:

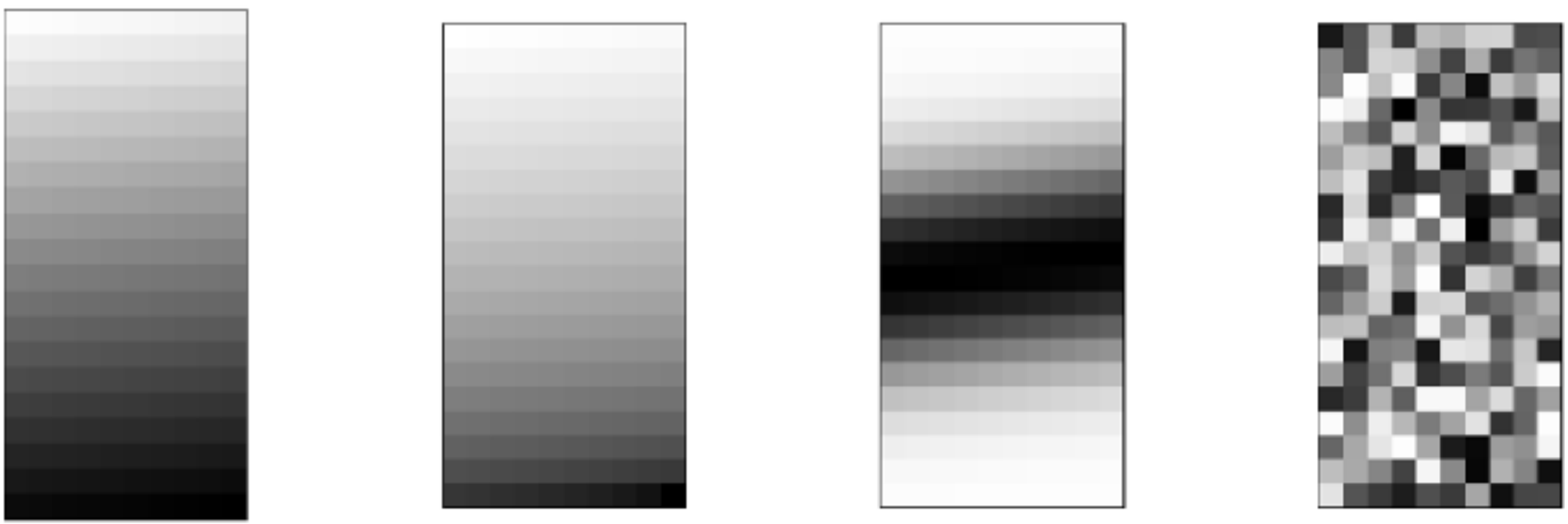
- Pixel-oriented visualization techniques
- Geometric projection visualization techniques
- Icon-based visualization techniques
- Hierarchical visualization techniques
- Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

For a data set of m dimensions, create m windows on the screen, one for each dimension

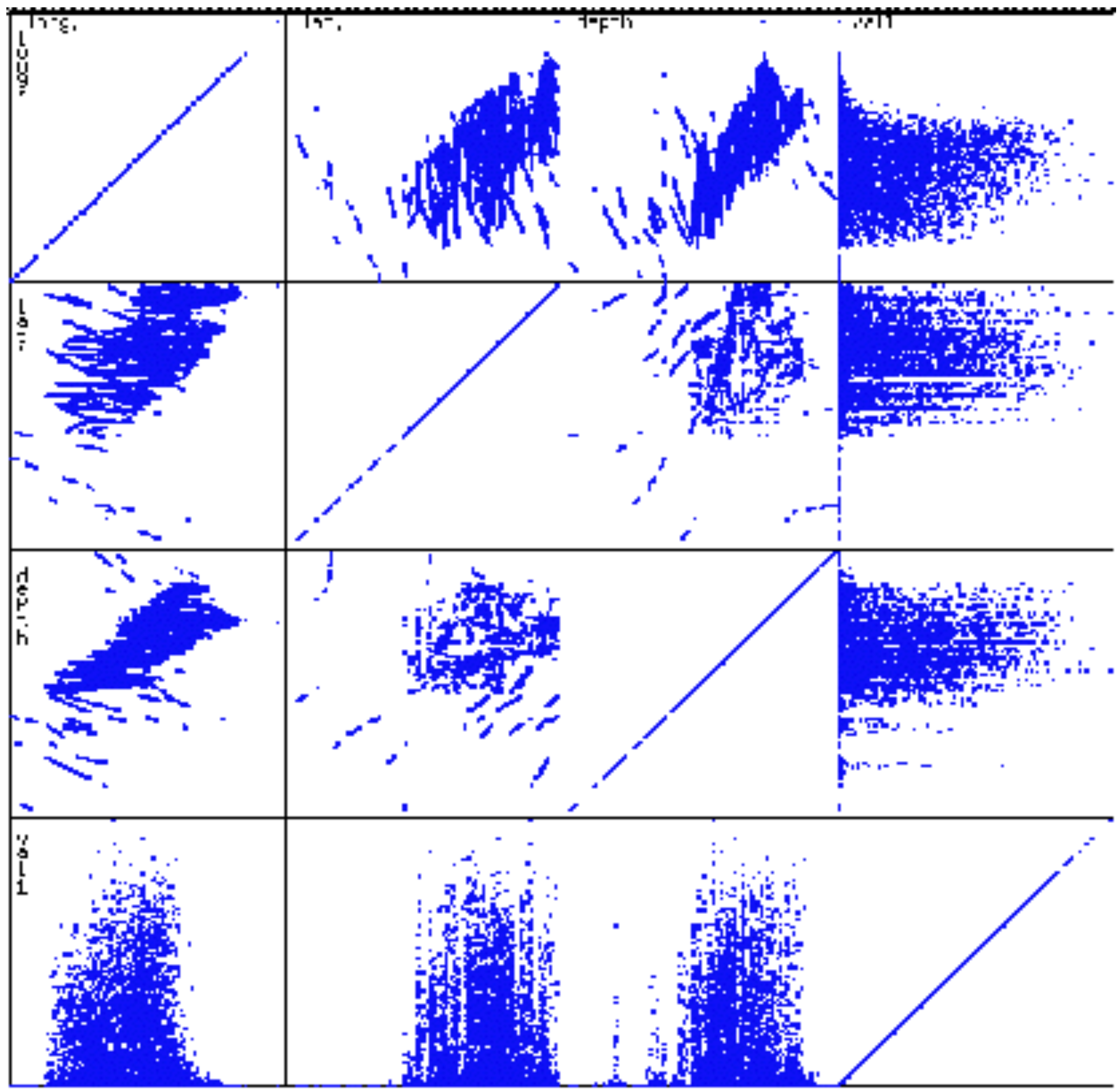
The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows

The colors of the pixels reflect the corresponding values



(a) Income (b) Credit Limit (c) transaction volume (d) age

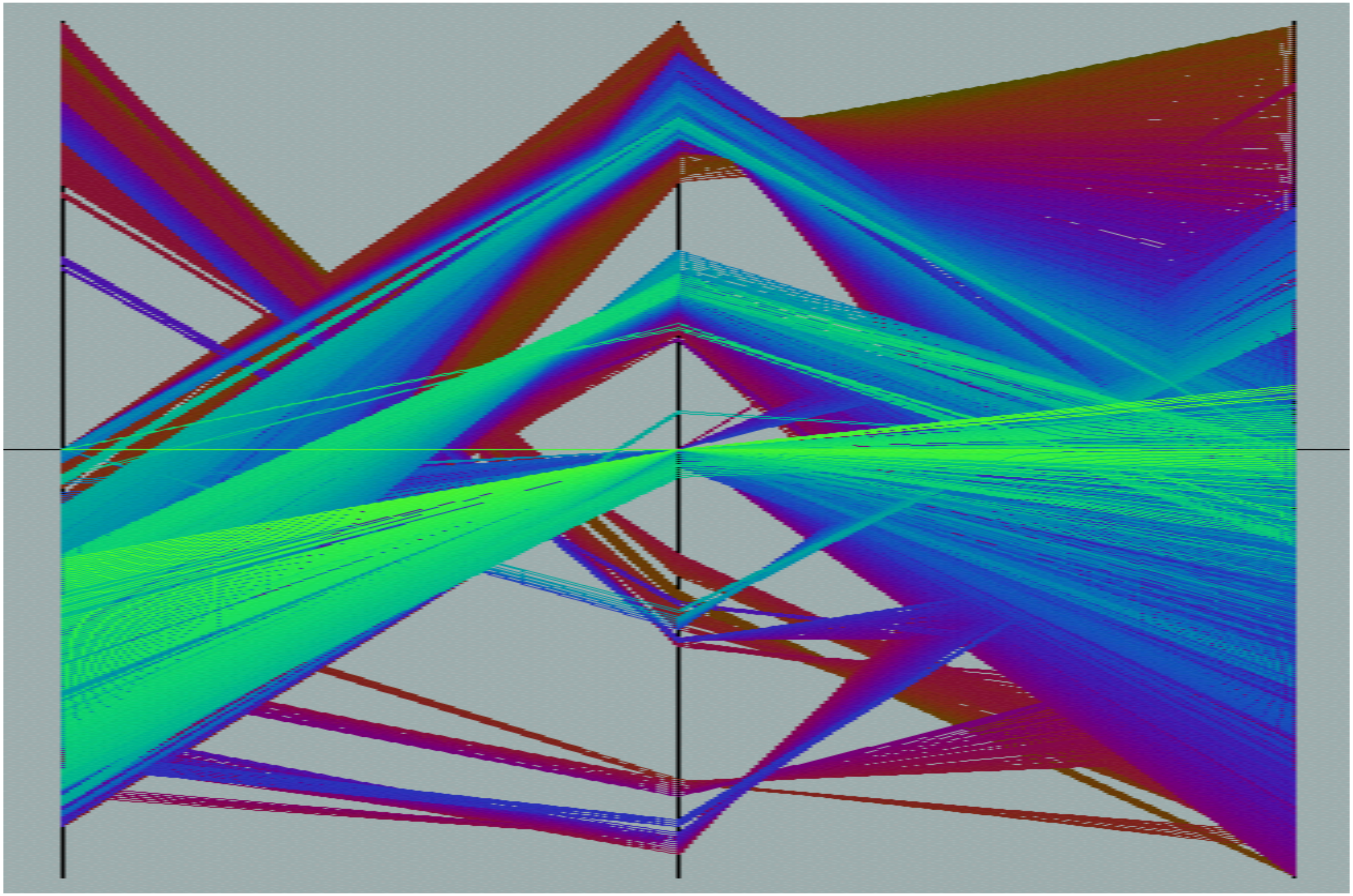
Scatterplot Matrices



Used by_ermission of M. Ward, Worcester Polytechnic Institute

Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of (k2/2-k) scatterplots]

Parallel Coordinates of a Data Set



Correlation Analysis (Nominal Data): Chi-Square Test

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

$$e_{ij} = \frac{count(A - a_i) \times count(B - b_j)}{n}$$

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90,$$

For this 2 × 2 table, the degrees of freedom are (2 − 1)(2 − 1) = 1. For 1 degree of freedom, the 2 value needed to reject the hypothesis at the 0.001 significance level is 10.828

X2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

507.93>10.828 shows that like_science_fiction and play_chess are correlated in the group

Correlation Analysis (Numeric Data)

Correlation coefficient (also called Pearson's product moment coefficient)

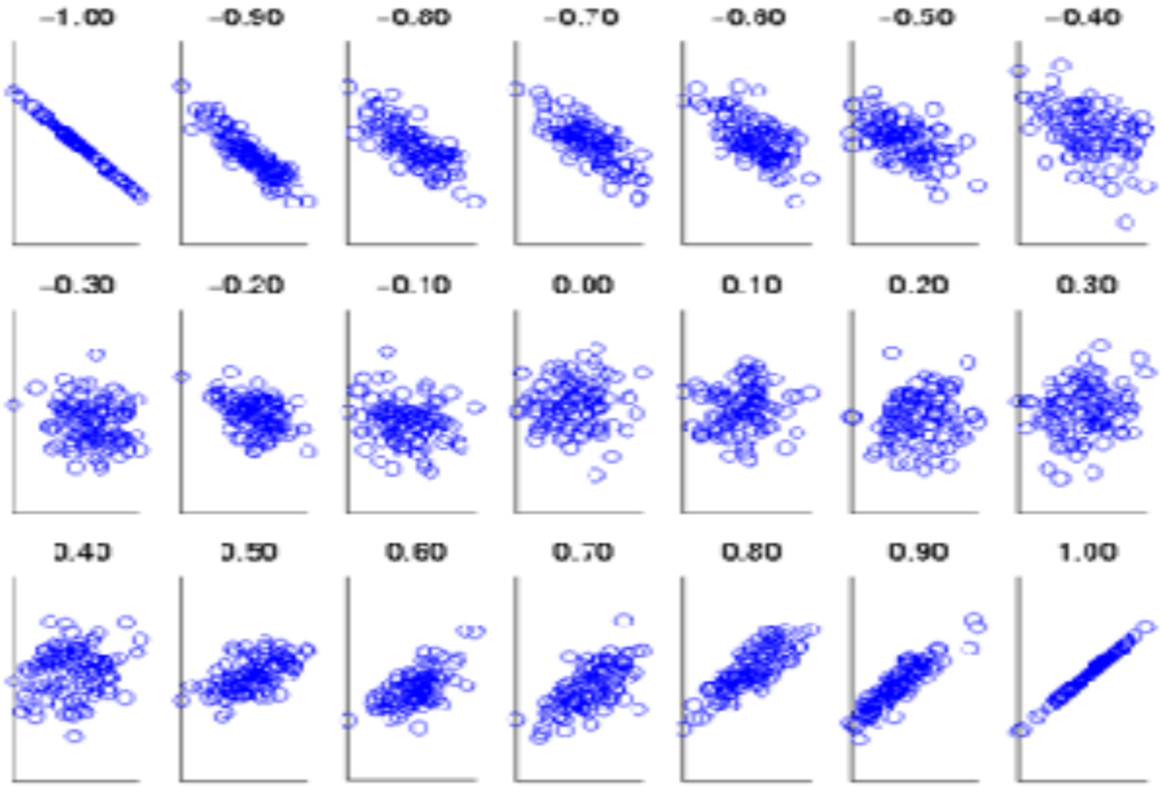
$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.

$r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Summary

- Histograms
- Measures of central tendency: mean, mode, median
- Measures of dispersion: range, IQR, variance, std deviation, coefficient of variation.
- Normal distribution, Chebyshev Rule.
- Five number summary, boxplots, QQ plots, Quantile plot, scatter plot.
- Visualization: scatter plot matrix, parallel coordinates.
- Correlation analysis.