

MATES ED2MIT  
Education and Training for Data Driven Maritime Industry

Tutorial D01

Statistical Data Analysis Basics: Data Structures, Statistical  
Characteristics

Yuri Demchenko MATES Project  
University of Amsterdam

**Maritime Alliance for fostering the  
European Blue economy through  
a Marine Technology Skilling  
Strategy**



Co-funded by the  
Erasmus+ Programme  
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



# Outline

- Types of data
  - Quantitative data
  - Qualitative data
- Statistical characteristics
- Distributions
  - Normal distribution
- Measures of data dissimilarity
- Summary and takeaway

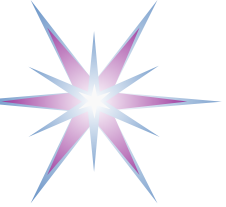


This work is licensed under the Creative Commons Attribution 4.0 International License.  
To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>



Co-funded by the  
Erasmus+ Programme  
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



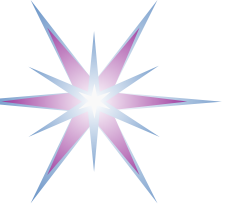
# Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

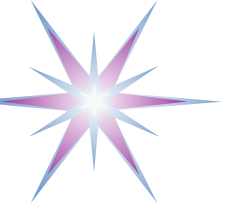
Cases/row/  
events



# Concepts and Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

- **Population**: the whole set of a “universe”
- **Sample**: a sub-set of a population
- **Parameter**: an unknown “fixed” value of population characteristic
- **Statistic**: a known/calculable value of sample characteristic representing that of the population. E.g.  
 $\mu$  = mean of population,  $\bar{X}$  = mean of sample



# Data Objects

- Data sets are made up of data objects
  - In a simple form data records
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples*, *instances*, *data points*, *objects*, *tuples*
- **Data objects are described by attributes**
- Database rows -> data objects; columns -> attributes



# Attributes

- **Attribute (or dimensions, features, variables)**: a data field, representing a characteristic or feature of a data object.
  - *E.g., customer\_ID, name, address*
- **Types**:
  - Nominal
  - Binary
  - Numeric: quantitative, e.g. numerical/numbers
    - Interval-scaled
    - Ratio-scaled
  - Categorical e.g. Yes or No



# Attribute Types

- **Nominal:** categories, states, or “names of things”
  - *Hair\_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary (aka dummy)**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings



# Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C° or F°, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*





# Discrete vs. Continuous Attributes

- **Discrete Attribute**

- Has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

- **Continuous Attribute**

- Has real numbers as attribute values
  - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables



# Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube



# Central Tendency: Mean, Median, Mode

Measure	Advantages	Disadvantages
<b>Mean</b> (Sum of all values ÷ no. of values)	<ul style="list-style-type: none"><li>* Best known average</li><li>* Exactly calculable</li><li>* Make use of all data</li><li>* Useful for statistical analysis</li></ul>	<ul style="list-style-type: none"><li>* Affected by extreme values</li><li>• Can be absurd for discrete data (e.g. Family size = 4.5 person)</li><li>* Cannot be obtained graphically</li></ul>
<b>Median</b> (middle value)	<ul style="list-style-type: none"><li>• Not influenced by extreme values</li><li>• Obtainable even if data distribution unknown (e.g. group/aggregate data)</li><li>• Unaffected by irregular class width</li><li>* Unaffected by open-ended class</li></ul>	<ul style="list-style-type: none"><li>• Needs interpolation for group/aggregate data (cumulative frequency curve)</li><li>• May not be characteristic of group when: (1) items are only few; (2) distribution irregular</li><li>* Very limited statistical use</li></ul>
<b>Mode</b> (most frequent value)	<ul style="list-style-type: none"><li>* Unaffected by extreme values</li><li>* Easy to obtain from histogram</li><li>* Determinable from only values near the modal class</li></ul>	<ul style="list-style-type: none"><li>• Cannot be determined exactly in group data</li><li>* Very limited statistical use</li></ul>



# Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

Note:  $n$  is sample size and  $N$  is population size.

- Weighted arithmetic mean:
- Trimmed mean: chopping extreme values

- Median:

- Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left( \frac{n/2 - (\sum freq)l}{freq_{median}} \right) width$$

- Mode

- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

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

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44



# Central Tendency – “Mean”, $\bar{x}$ Example

- For individual observations  
 $X = \{3, 5, 7, 7, 8, 8, 8, 9, 9, 10, 10, 12\}$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\sum fx = 96; n = 12$$

$$\text{Thus, } \bar{x} = \frac{\sum x_i}{n} = 96/12 = 8$$

- The above observations can be organised into a frequency table and mean calculated on the basis of frequencies

x	3	5	7	8	9	10	12
f	1	1	2	3	2	2	1
$\Sigma f$	3	5	14	24	18	20	12

$$\bar{x} = \frac{\sum fx}{\sum f}$$

$$\sum fx = 96; \sum f = 12$$

$$\text{Thus, } \bar{x} = \frac{\sum fx}{\sum f} = 96/12 = 8$$



# Variability (1)

- Indicates dispersion, spread, variation, deviation
- For single population or sample data:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

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

where  $\sigma^2$  and  $s^2$  = population and sample variance respectively,  $x_i$  = individual observations,  $\mu$  = population mean,  $\bar{x}$  = sample mean, and  $n$  = total number of individual observations.

- The standard deviation is the square roots

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

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



## Variability (2)

- Why “measure of dispersion” important?
- Consider returns from two categories of shares:

\* Shares A (%) = {1.8, 1.9, 2.0, 2.1, 3.6}

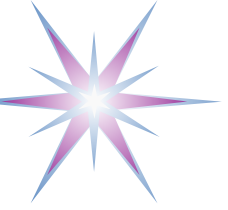
\* Shares B (%) = {1.0, 1.5, 2.0, 3.0, 3.9}

Mean A = mean B = 2.28%

But different variability!

$\text{Var}(A) = 0.557$ ,  $\text{Var}(B) = 1.367$

- \* Would you invest in category A shares or category B shares?



## Variability (3)

- Coefficient of variation – COV – std. deviation as % of the mean:

$$\text{COV} = \frac{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}}{\frac{\sum x_i}{n}} \times 100$$

- Could be a better measure compared to std. dev.  
COV(A) = 32.73%, COV(B) = 51.28%





# Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles:**  $Q_1$  (25<sup>th</sup> percentile),  $Q_3$  (75<sup>th</sup> percentile)
  - **Inter-quartile range:**  $IQR = Q_3 - Q_1$
  - **Five number summary:** min,  $Q_1$ , median,  $Q_3$ , max
  - **Boxplot:** ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier:** usually, a value higher/lower than  $1.5 \times IQR$
- Variance and standard deviation (*sample:  $s$ , population:  $\sigma$* )
  - **Variance:** (algebraic, scalable computation)

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

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

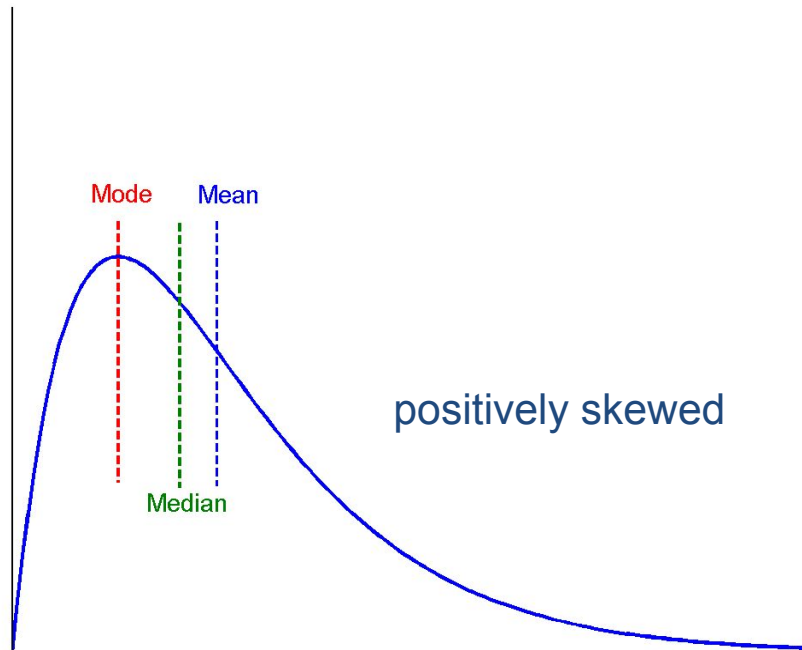
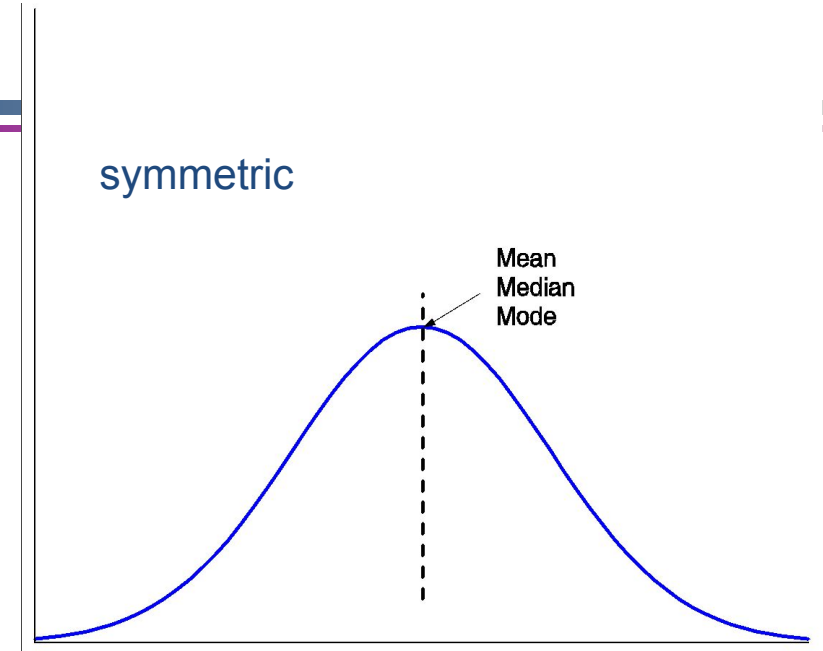
- **Standard deviation  $s$  (or  $\sigma$ )** is the square root of variance  $s^2$  (or  $\sigma^2$ )



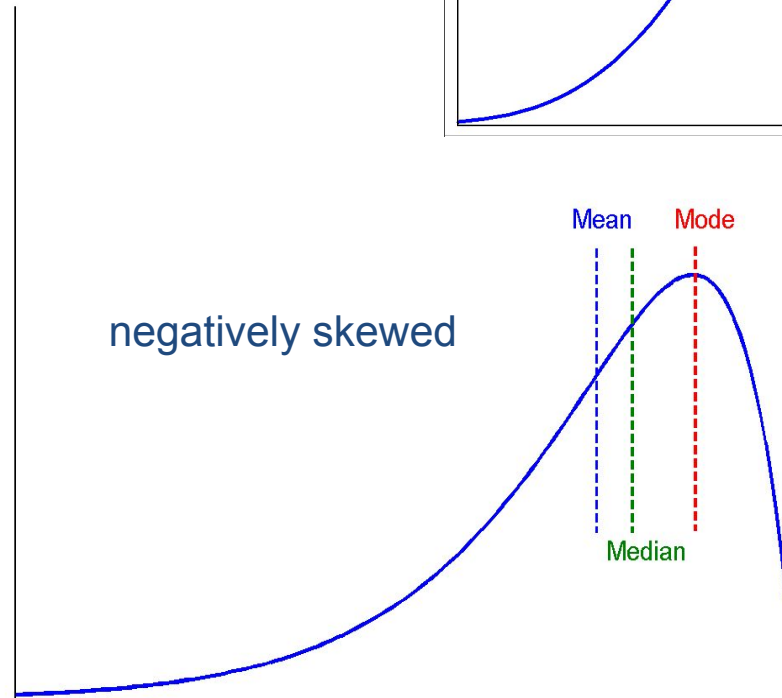
# Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

symmetric



negatively skewed





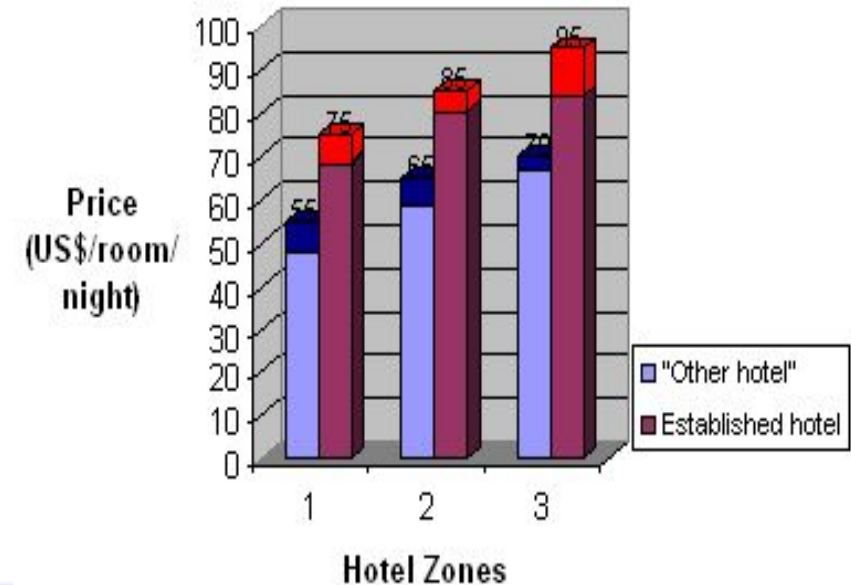
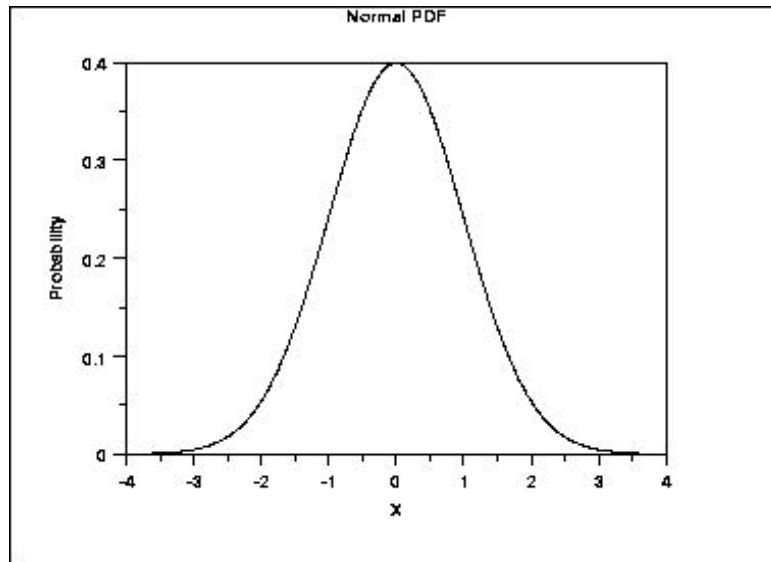
# Probability Distribution

- Defined as of probability density function (pdf).
- Many types: Z, t, F, gamma, etc.
- “God-given” nature of the real world event.
- General form:

$$\int_a^b f(x)dx = Pr[a \leq X \leq b] \quad (\text{continuous})$$

$$f(x) = Pr[X = x]$$

(discrete)





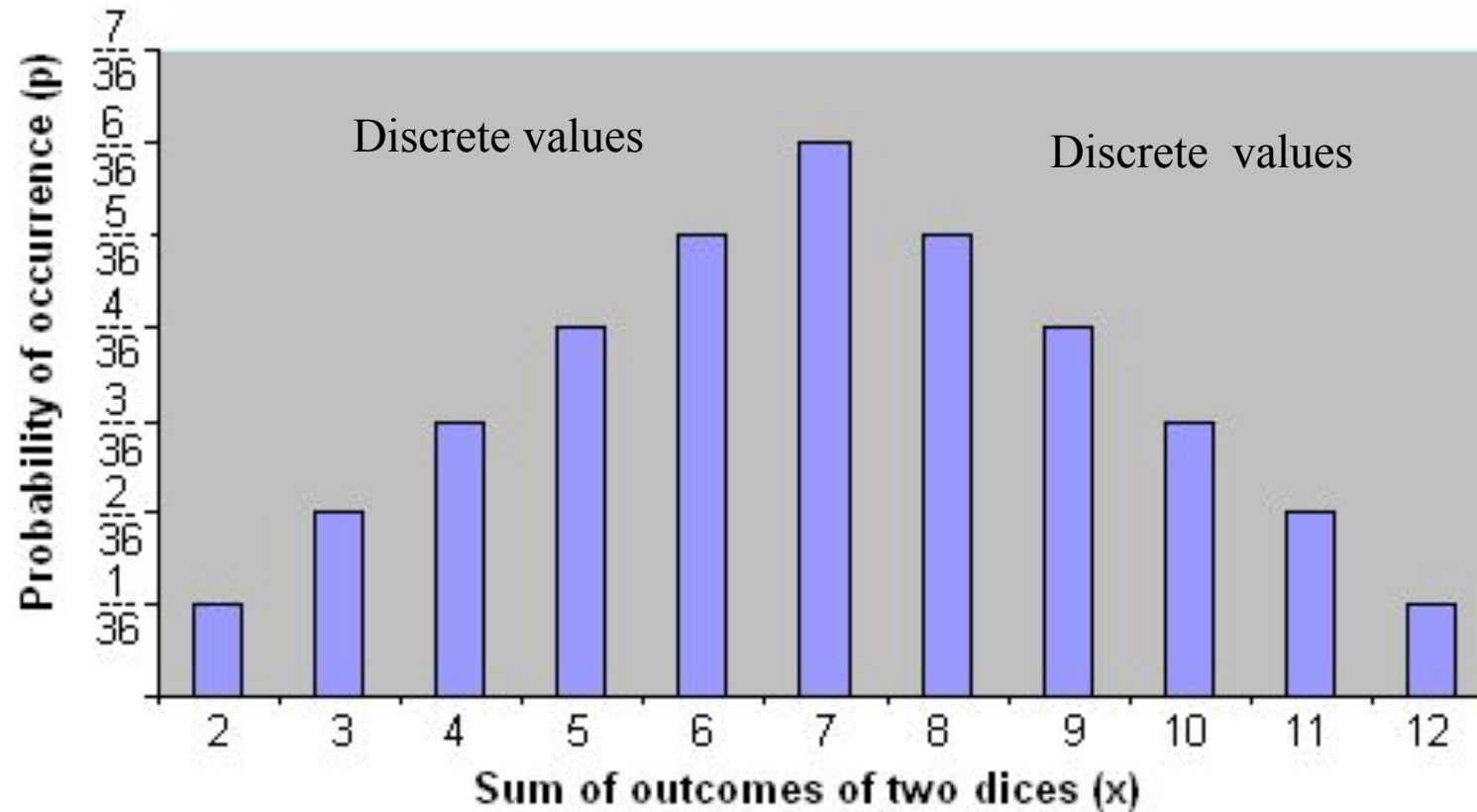
# Probability Distribution – Dice example



Dice1 Dice2		1	2	3	4	5	6
1	2	3	4	5	6	7	
2	3	4	5	6	7	8	
3	4	5	6	7	8	9	
4	5	6	7	8	9	10	
5	6	7	8	9	10	11	
6	7	8	9	10	11	12	



# Probability Distribution – Dice bets probability

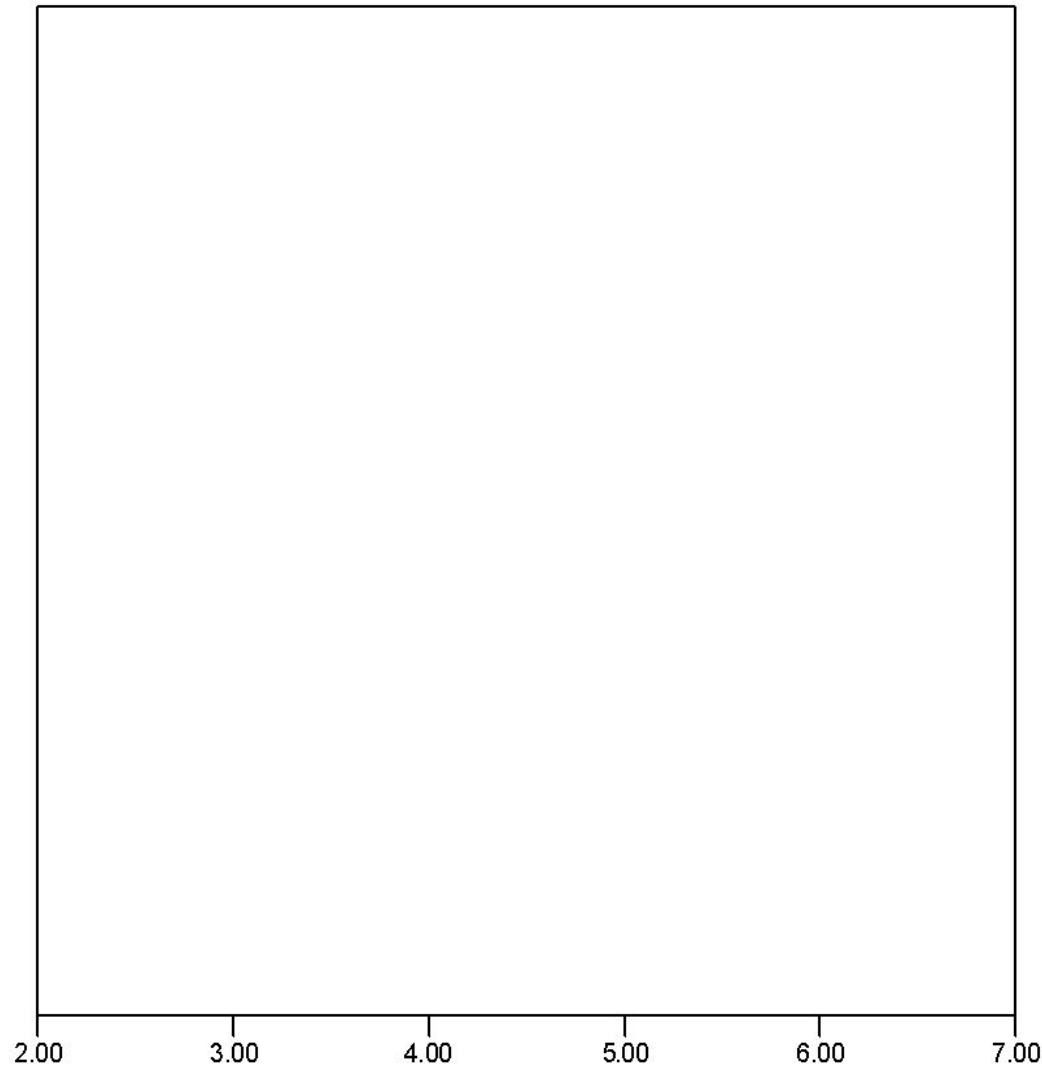


Values of  $x$  are discrete (discontinuous)

Sum of lengths of vertical bars  $\sum_{\text{all } x} p(X=x) = 1$



# Probability Distribution – Discrete Distribution

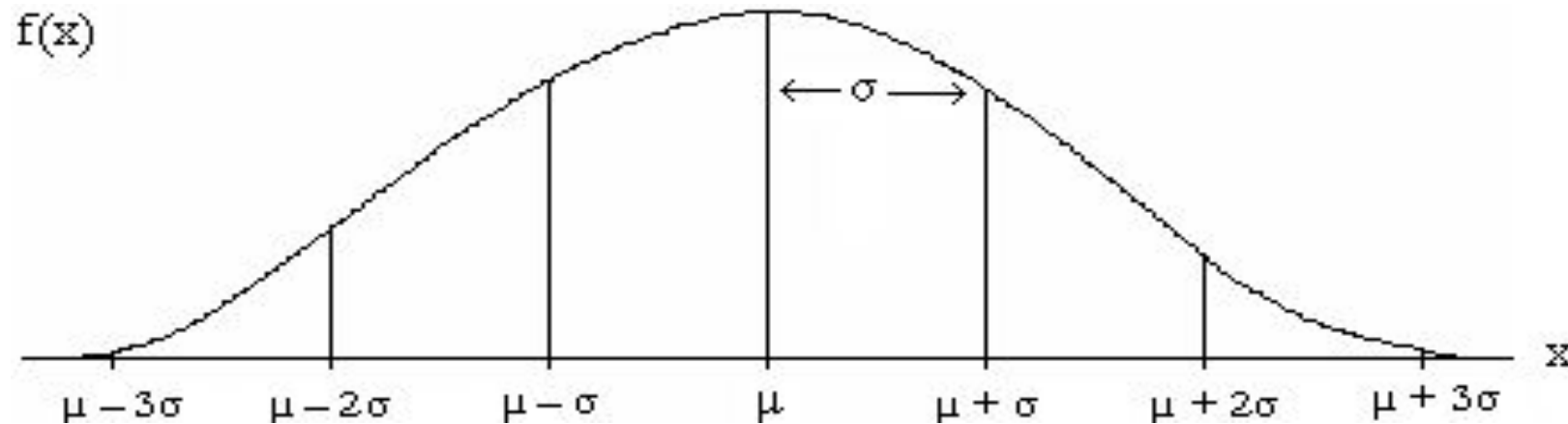


- Many real world phenomena take a form of continuous random variable
- Can take any values between two limits (e.g. income, age, weight, price, rental, etc.)



# Probability Distribution – Bell shaped

- Ideal distribution of a phenomena:



- Bell-shaped, symmetrical

- Has a function of

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu$  = mean of variable  $x$

$\sigma$  = std. dev. Of  $x$

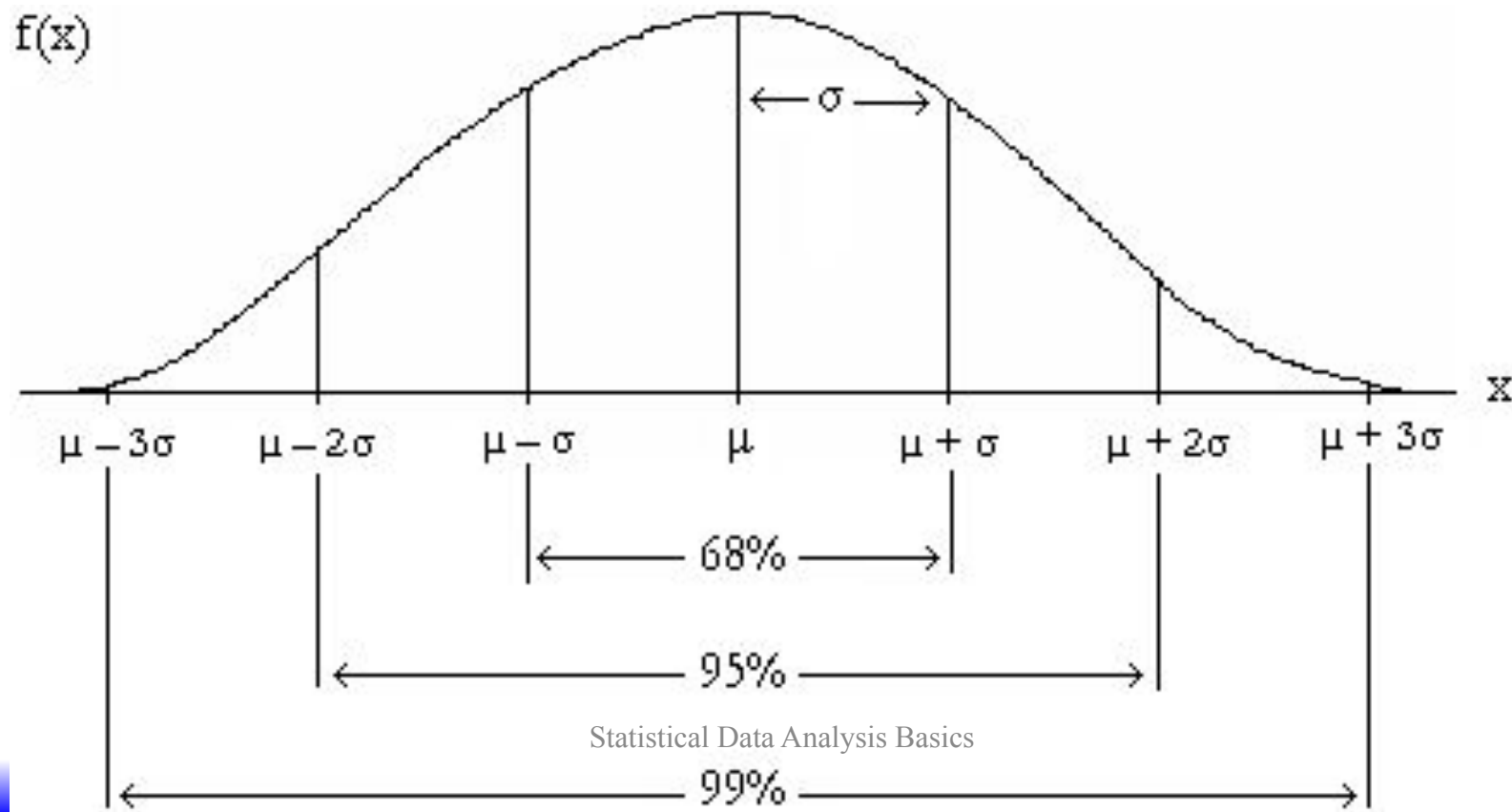
$\pi$  = ratio of circumference of a circle to its diameter = 3.14

$e$  = base of natural log = 2.71828



# Properties of Normal Distribution Curve

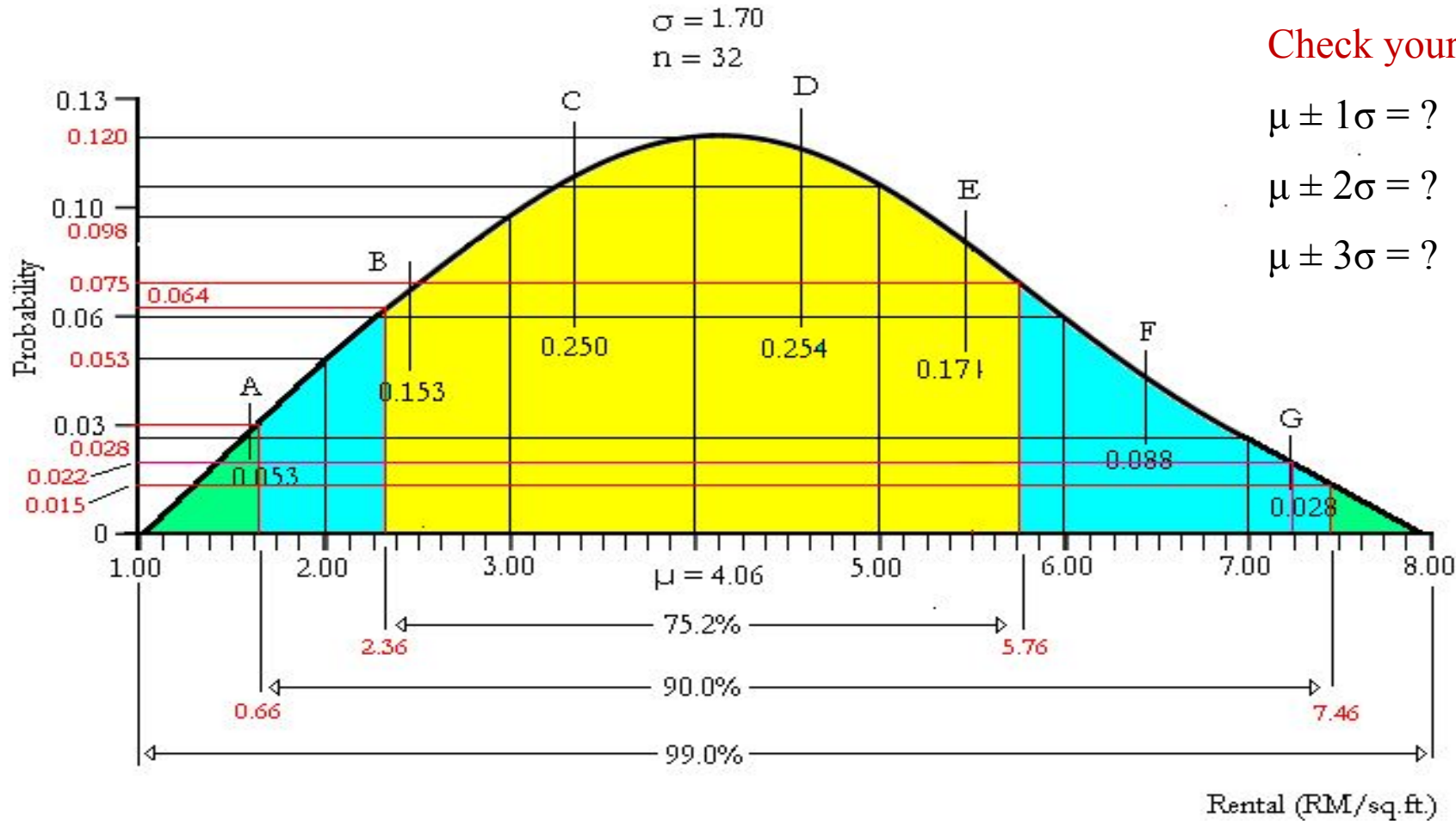
- The normal (distribution) curve
  - From  $\mu - \sigma$  to  $\mu + \sigma$ : contains about 68% of the measurements ( $\mu$ : mean,  $\sigma$ : standard deviation)
  - From  $\mu - 2\sigma$  to  $\mu + 2\sigma$ : contains about 95% of it
  - From  $\mu - 3\sigma$  to  $\mu + 3\sigma$ : contains about 99.7% of it







# Probability distribution – Standard Deviation



Check yourself

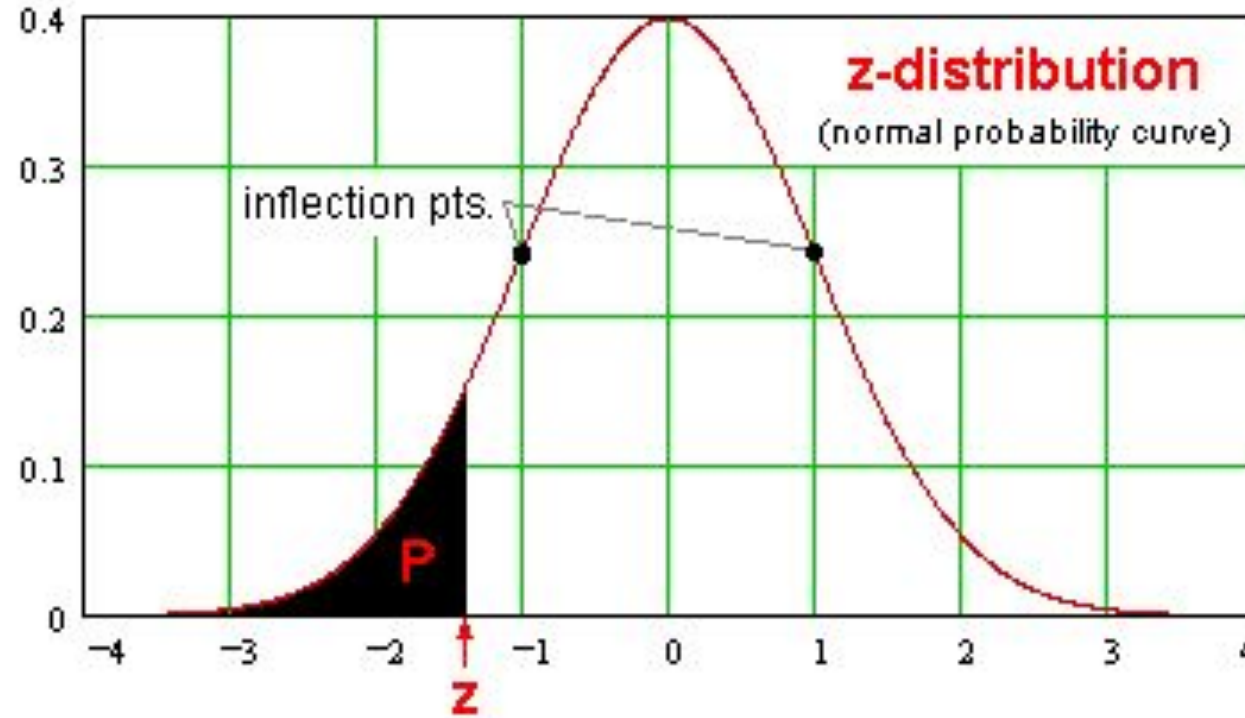
$\mu \pm 1\sigma = ?$  = \_\_\_\_% from total observation

$\mu \pm 2\sigma = ?$  = \_\_\_\_% from total observation

$\mu \pm 3\sigma = ?$  = \_\_\_\_% from total observation



# Z-distribution

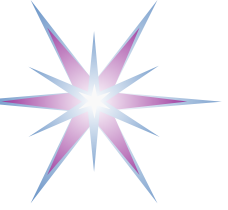


The z- is a  $N(0, 1)$  distribution, given by the equation:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

The area within an interval  $(a,b) = \text{normalcdf}(a,b) =$   
*(It is not integratable algebraically.)*

$$\int_a^b e^{-\frac{z^2}{2}} dz$$



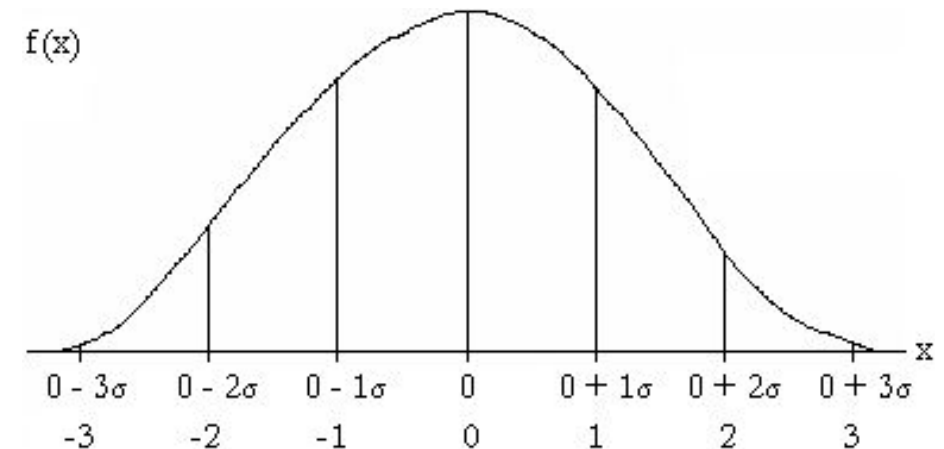
# Z-Distribution

- $\phi(X=x)$  is given by area under curve
- Has no standard algebraic method of integration  $\rightarrow Z \sim N(0,1)$
- It is called “normal distribution” (ND)
- Standard reference/approximation of other distributions. Since there are various  $f(x)$  forming NDs, SND is needed
- To transform  $f(x)$  into  $f(z)$ :

$$Z = \frac{x - \mu}{\sigma} \sim N(0, 1)$$

$$\text{E.g. } Z = \frac{160 - 155}{5.4} = 0.926$$

- Probability is such a way that:
  - \* Approx. 68%  $-1 < z < 1$
  - \* Approx. 95%  $-1.96 < z < 1.96$
  - \* Approx. 99%  $-2.58 < z < 2.58$



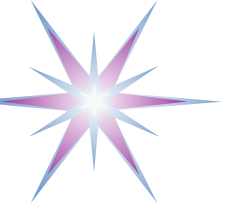


## Z-distribution (2)

- When  $X = \mu$ ,  $Z = 0$ , i.e.

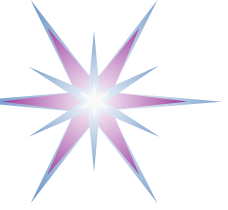
$$Z = \frac{X - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

- When  $X = \mu + \sigma$ ,  $Z = 1$
- When  $X = \mu + 2\sigma$ ,  $Z = 2$
- When  $X = \mu + 3\sigma$ ,  $Z = 3$  and so on.
- It can be proven that  $P(X_1 < X < X_k) = P(Z_1 < Z < Z_k)$
- SND (standard normal distribution) shows the *probability to the right* of any particular value of  $Z$ .



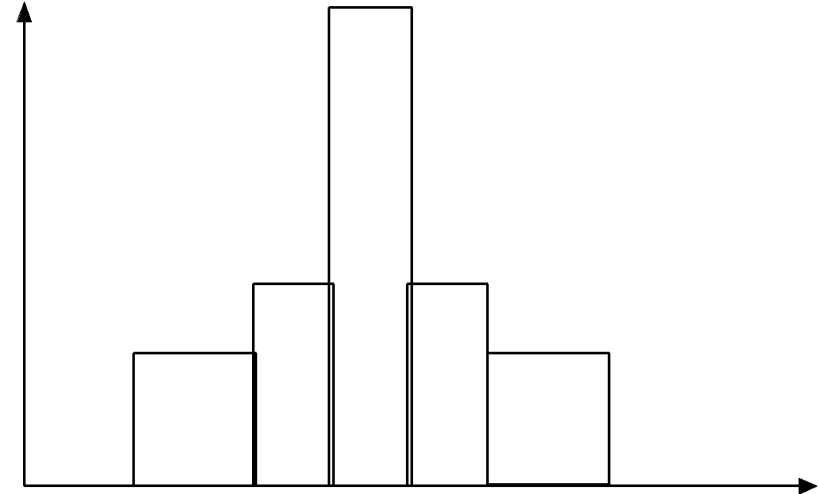
# Graphic Displays of Basic Statistical Descriptions

- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Boxplot:** graphic display of five-number summary
- **Quantile plot:** each value  $x_i$  is paired with  $f_i$  indicating that approximately 100  $f_i$  % of data are  $\leq x_i$
- **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



# Histogram Analysis

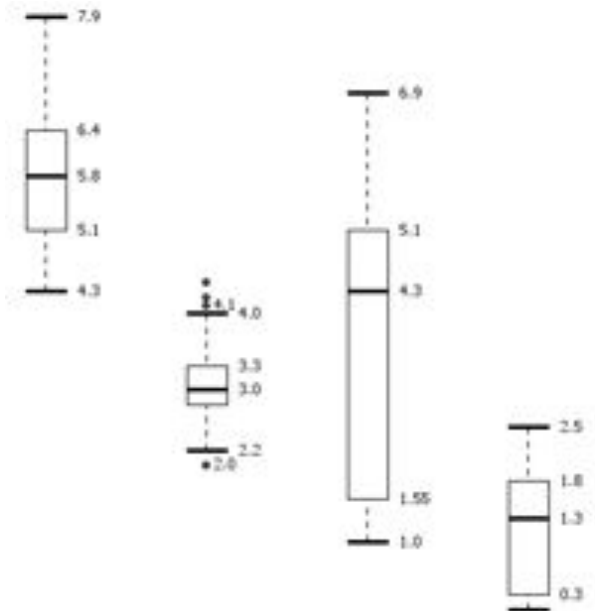
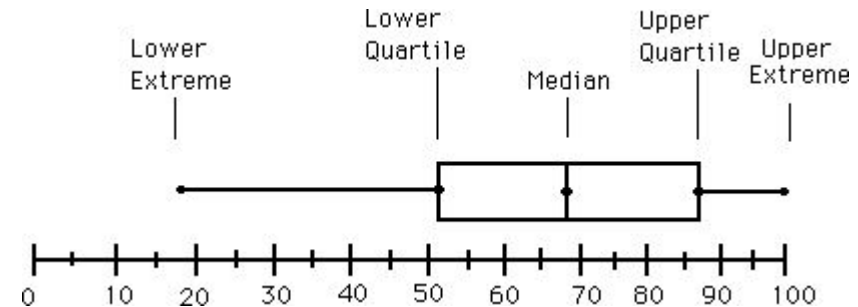
- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent





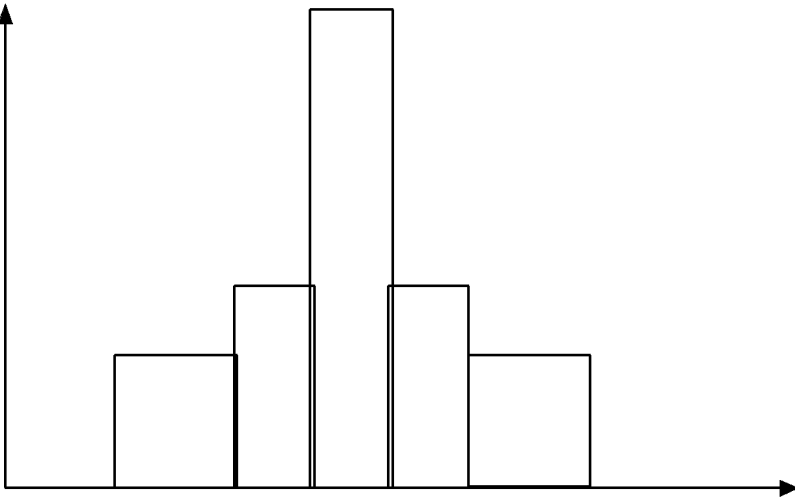
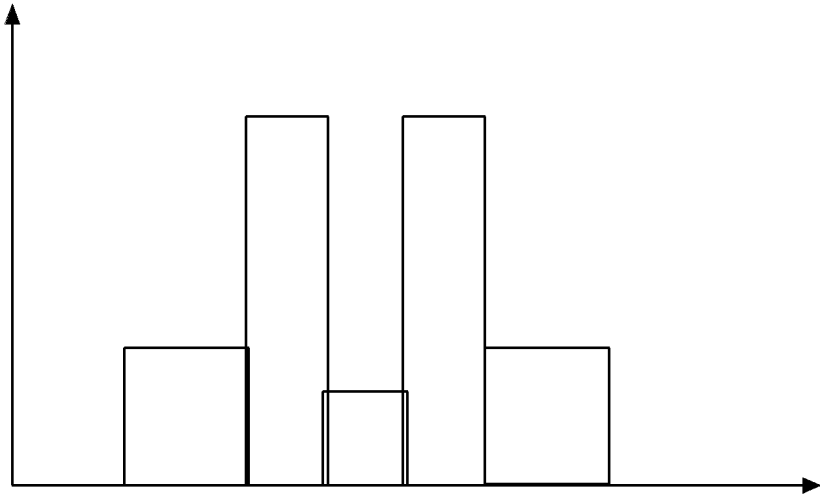
# Boxplot Analysis

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



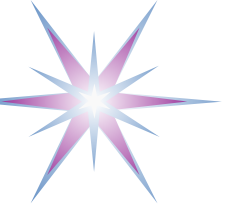


# Histograms Often Tell More than Boxplots



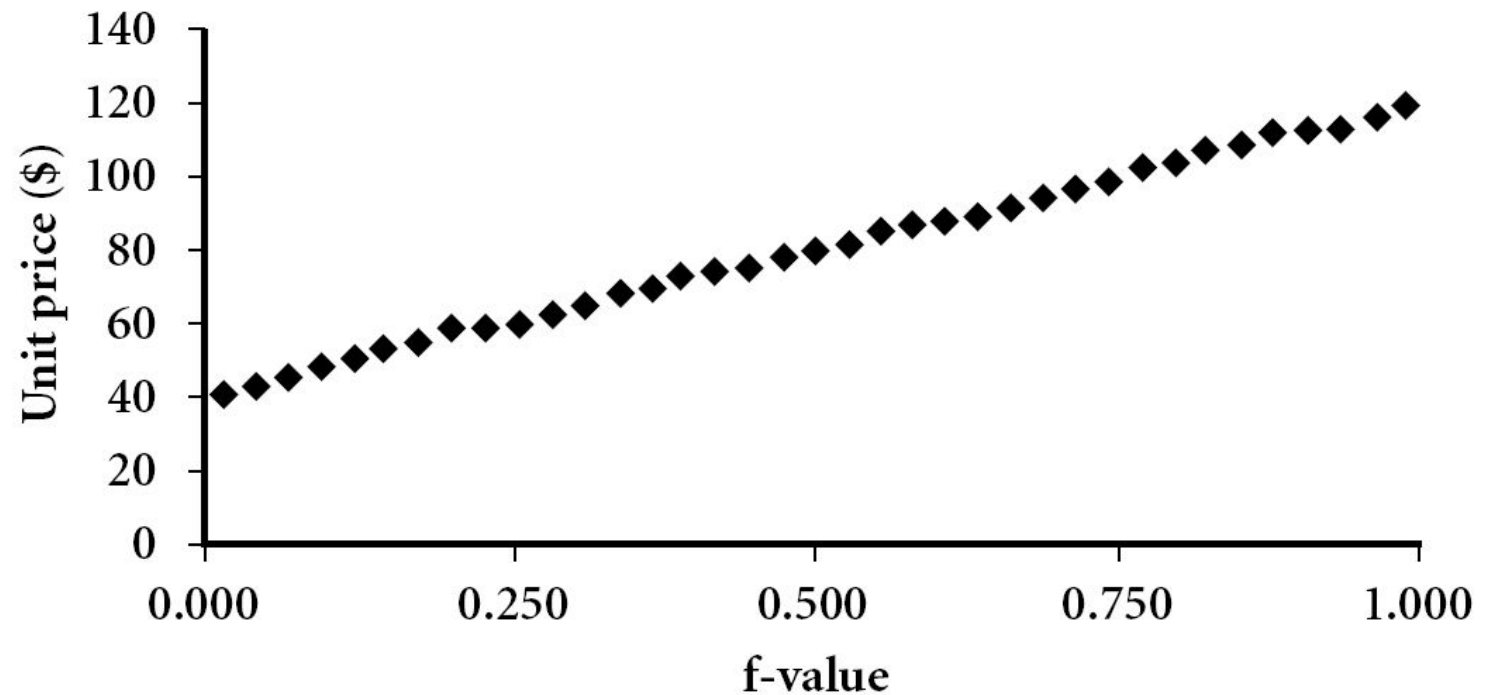
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions





# Quantile Plot

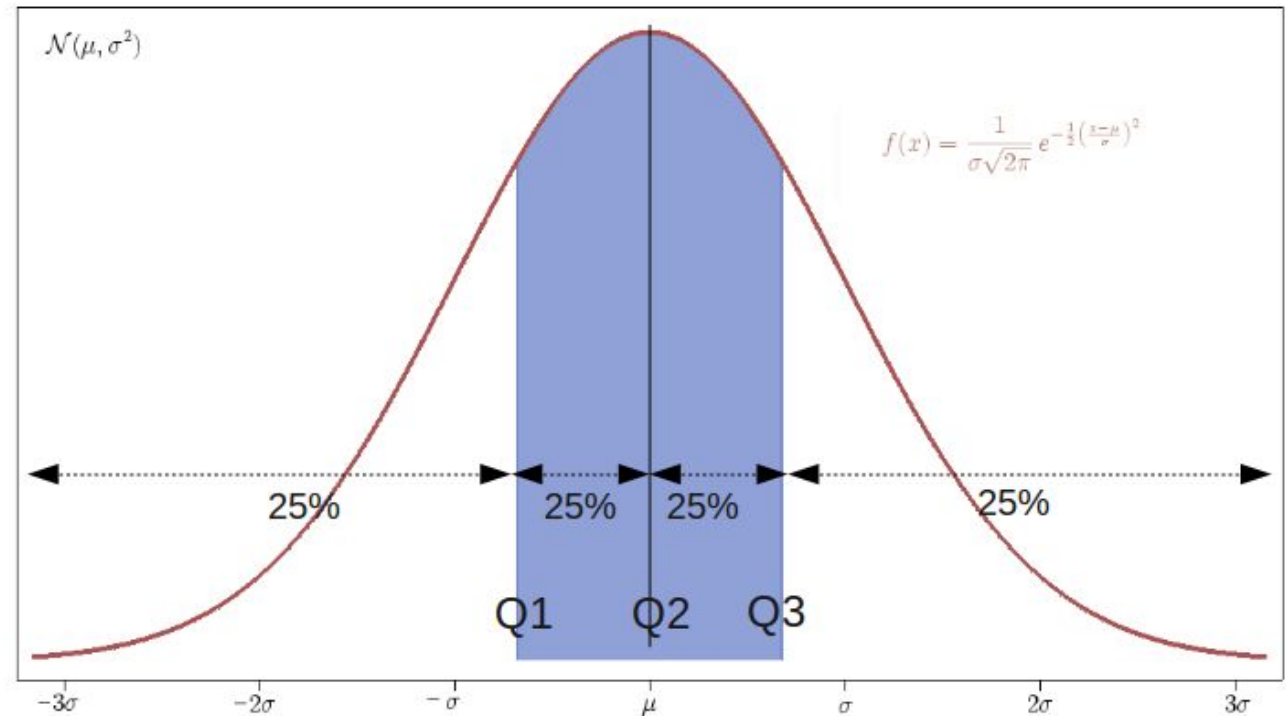
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data  $x_i$ , data sorted in increasing order,  $f_i$  indicates that approximately 100  $f_i$ % of the data are below or equal to the value  $x_i$





# Quantile Plot – What is quartile?

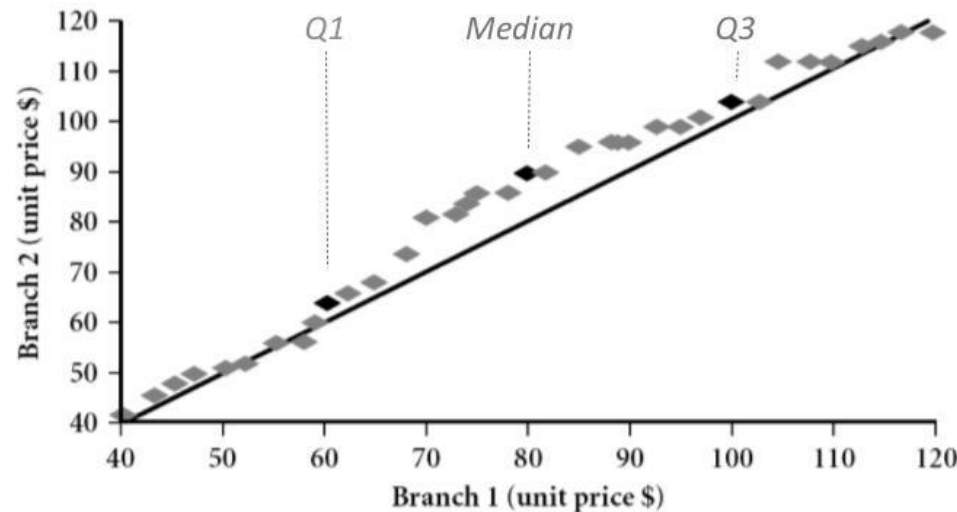
- Probability density of a normal distribution, with quartiles shown.
- The area below the red curve is the same in the intervals  $(-\infty, Q1)$ ,  $(Q1, Q2)$ ,  $(Q2, Q3)$ , and  $(Q3, +\infty)$ .
- Quartiles are used in boxplot
- The only 2-quantile is called the median





# Quantile-Quantile (Q-Q) Plot

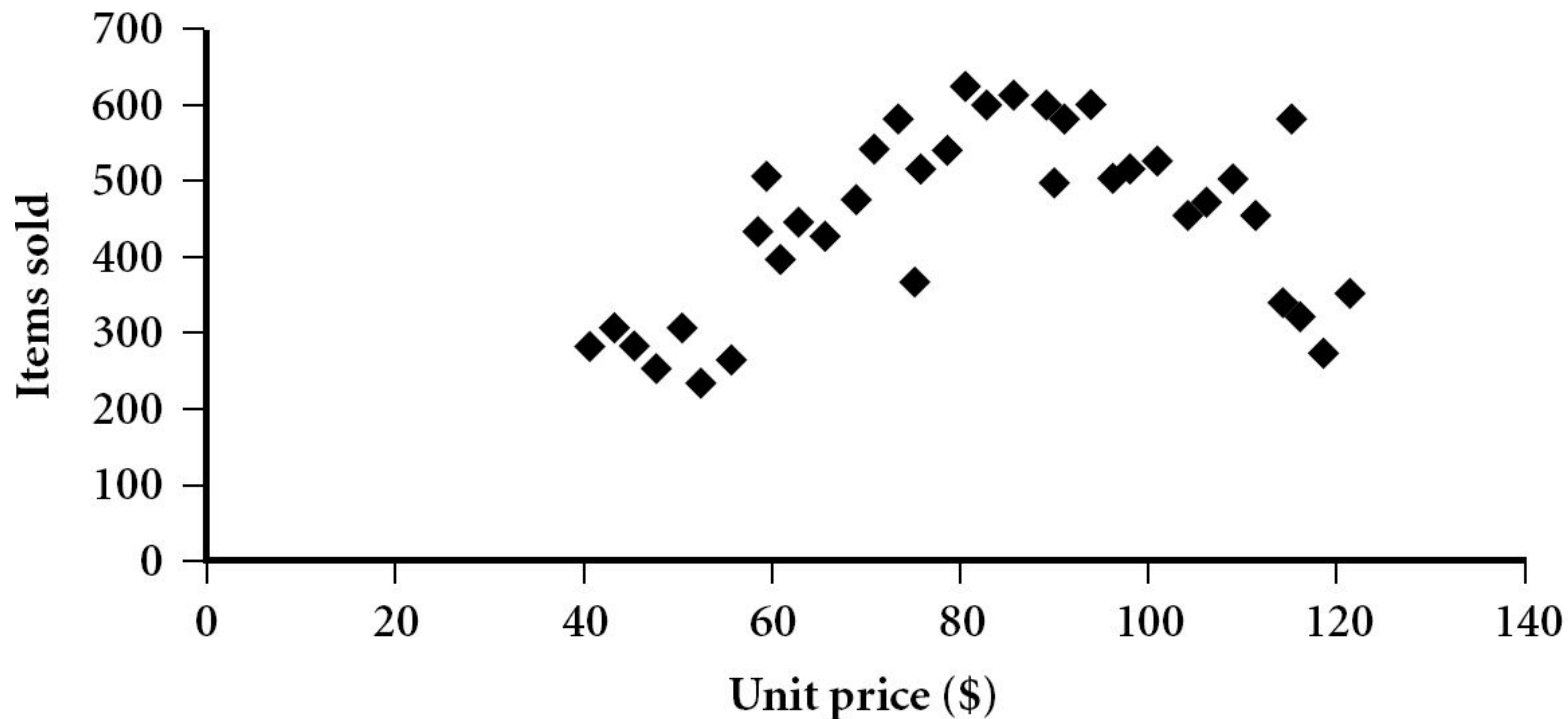
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

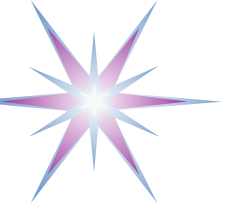




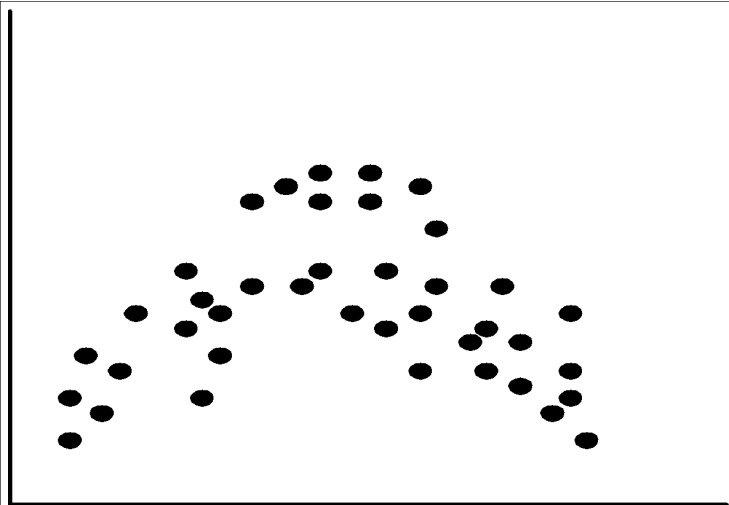
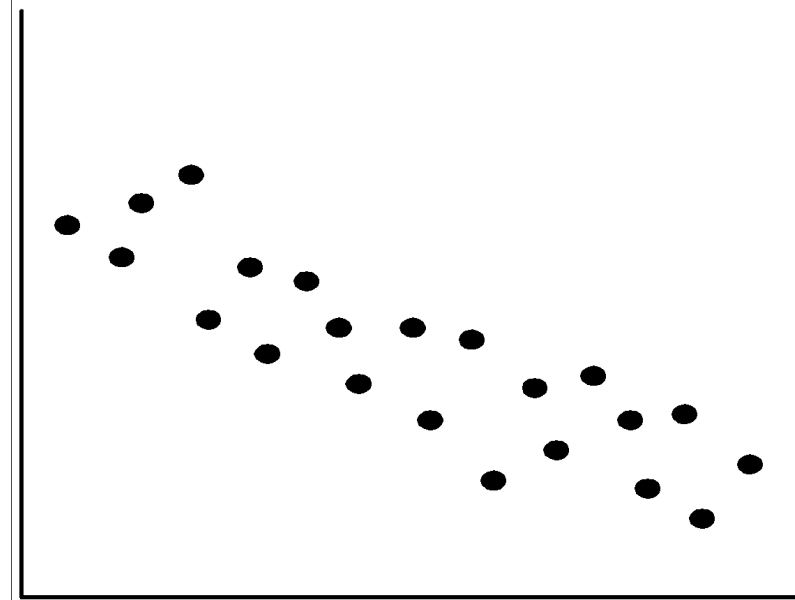
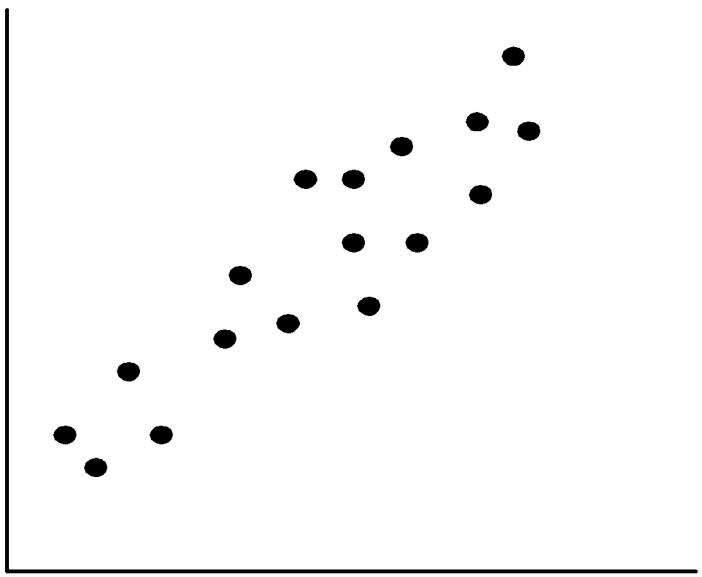
# Scatter plot

- 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





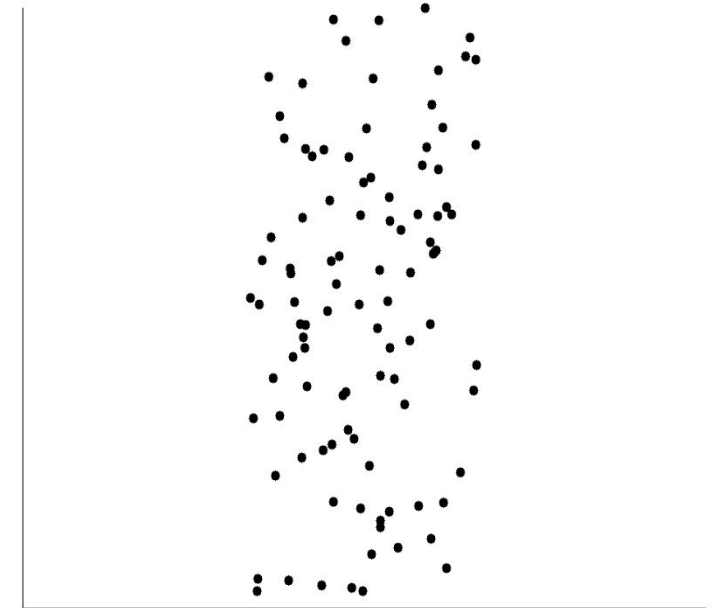
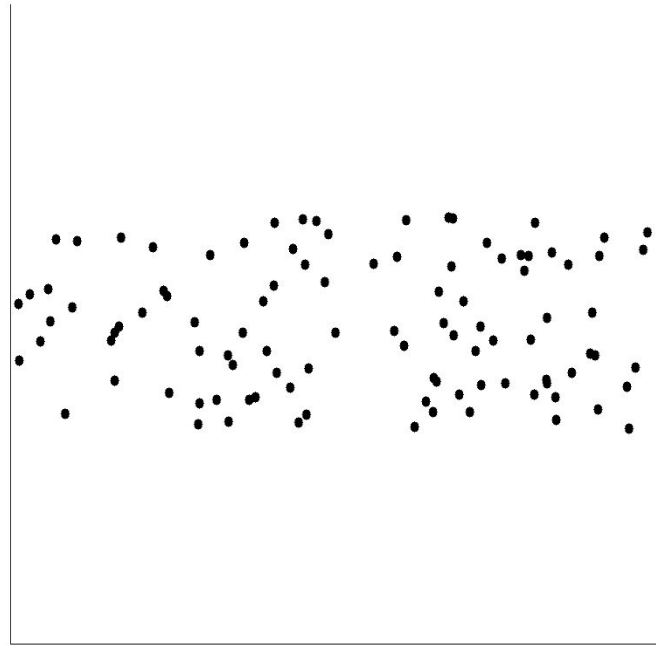
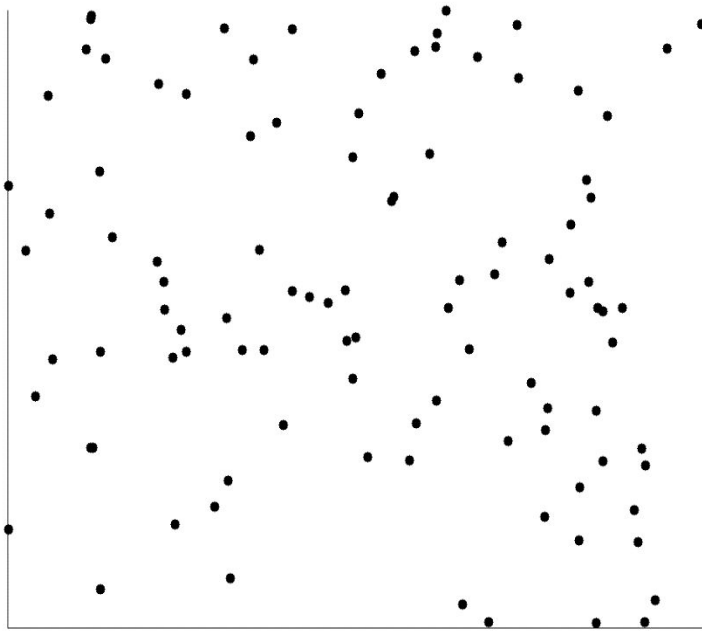
# Positively and Negatively Correlated Data

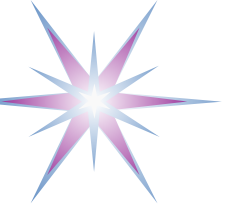


- The left half fragment is positively correlated
- The right half is negative correlated



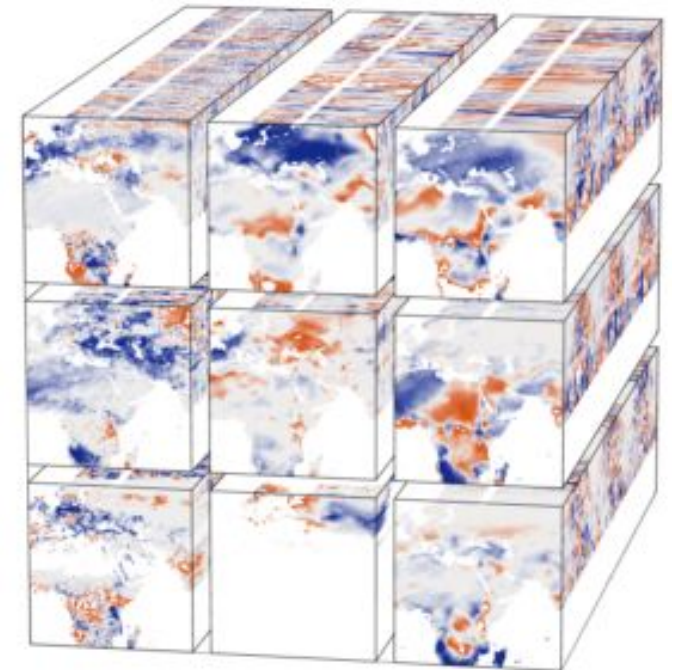
# Uncorrelated Data





# Data Cube

- A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on
- Multi-dimensional data cubes



[ref] Earth Data Cube <https://deepcube-h2020.eu/technology/earth-system-data-cube/>



# Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range  $[0,1]$
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity





# Data Matrix and Dissimilarity Matrix

- Data matrix

- n data points with p dimensions
- Two modes

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix
- Single mode

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$



# Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
  - $m$ : # of matches,  $p$ : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
  - creating a new binary attribute for each of the  $M$  nominal states



# Proximity Measure for Binary Attributes

- A contingency table for binary data
- Distance measure for symmetric binary variables:
- Distance measure for asymmetric binary variables:
- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

		Object $j$		
		1	0	sum
Object $i$	1	$q$	$r$	$q + r$
	0	$s$	$t$	$s + t$
sum		$q + s$	$r + t$	$p$

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

$$d(i, j) = \frac{r + s}{q + r + s}$$

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

- Note: Jaccard coefficient is the same as “coherence”:

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$



# Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$



# Standardizing Numeric Data

- Z-score: 
$$z = \frac{x - \mu}{\sigma}$$
  - X: raw score to be standardized,  $\mu$ : mean of the population,  $\sigma$ : standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, “+” when above
- An alternative way: Calculate the mean absolute deviation

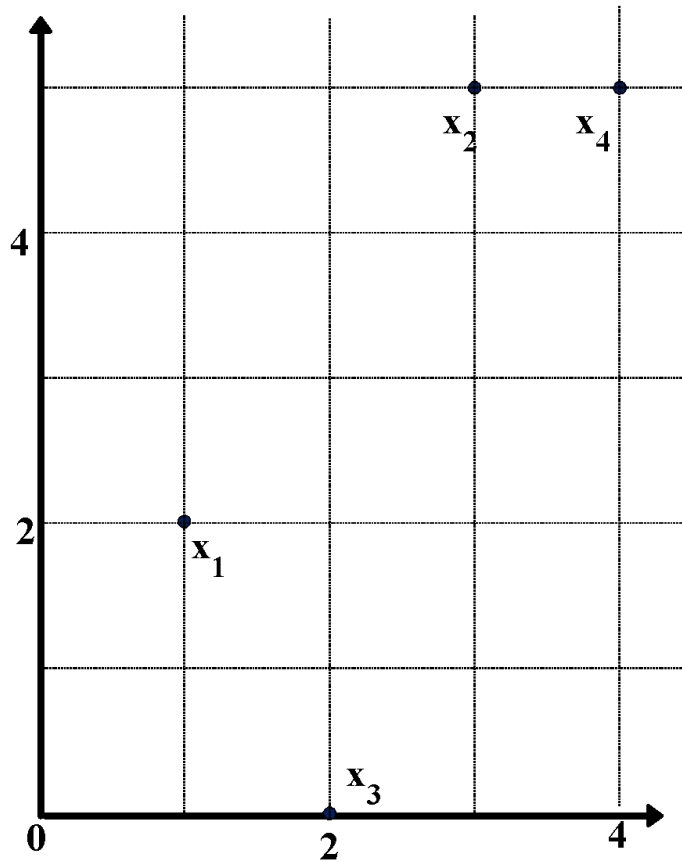
where

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$
$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf}).$$

- standardized measure (z-score): 
$$z_{if} = \frac{x_{if} - m_f}{s_f}$$
- Using mean absolute deviation is more robust than using standard deviation



# Example: Matrix and Dissimilarity Matrix



**Data Matrix**

point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

**Dissimilarity Matrix**  
(with **Euclidean Distance**)

	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$	3.61	0		
$x3$	5.1	5.1	0	
$x4$	4.24	1	5.39	0



# Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A general definition for distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  and  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  are two  $p$ -dimensional data objects, and  $h$  is the order (the distance so defined is also called L- $h$  norm)

- Properties
  - $d(i, j) > 0$  if  $i \neq j$ , and  $d(i, i) = 0$  (Positive definiteness)
  - $d(i, j) = d(j, i)$  (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$  (Triangle Inequality)
- A distance that satisfies these properties is a metric



# Special Cases of Minkowski Distance

- $h = 1$ : **Manhattan** (city block,  $L_1$  norm) **distance**
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

- $h = 2$ : ( $L_2$  norm) **Euclidean** distance

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \rightarrow \infty$ . **“supremum”** ( $L_{\max}$  norm,  $L_\infty$  norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \rightarrow \infty} \left( \sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f^p |x_{if} - x_{jf}|$$

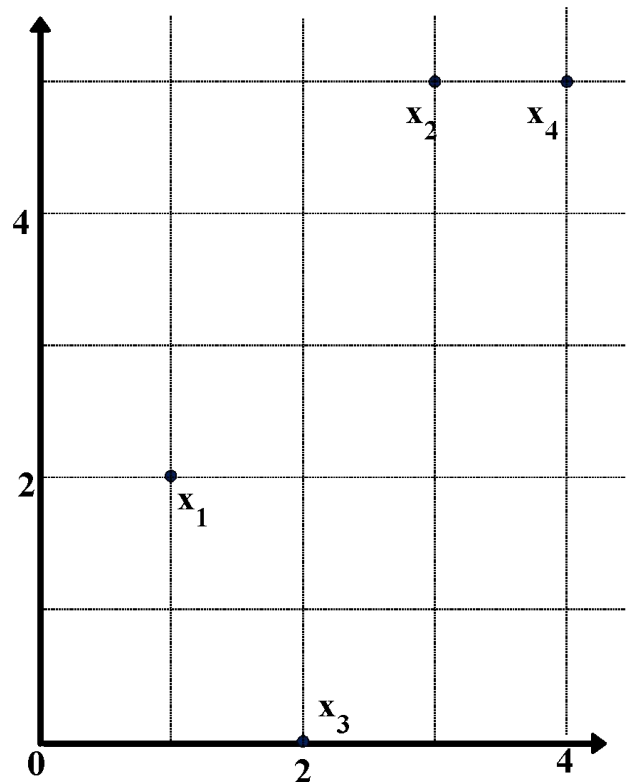




# Example: Minkowski Distance

## Dissimilarity Matrices

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



### Manhattan ( $L_1$ )

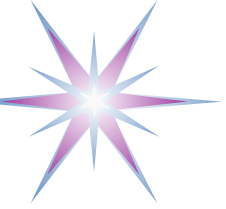
L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

### Euclidean ( $L_2$ )

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

### Supremum

$L_\infty$	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0



# Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace  $x_{if}$  by their rank  $r_{if} \in \{1, \dots, M_f\}$
  - map the range of each variable onto  $[0, 1]$  by replacing  $i$ -th object in the  $f$ -th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- compute the dissimilarity using methods for interval-scaled variables



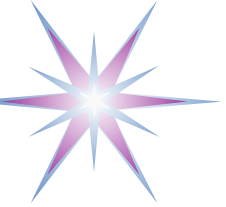
# Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- $f$  is binary or nominal:  
 $d_{ij}^{(f)} = 0$  if  $x_{if} = x_{jf}$ , or  $d_{ij}^{(f)} = 1$  otherwise
- $f$  is numeric: use the normalized distance
- $f$  is ordinal
  - Compute ranks  $r_{if}$  and
  - Treat  $z_{if}$  as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$



# Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

<i>Document</i>	<i>team</i>	<i>coach</i>	<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

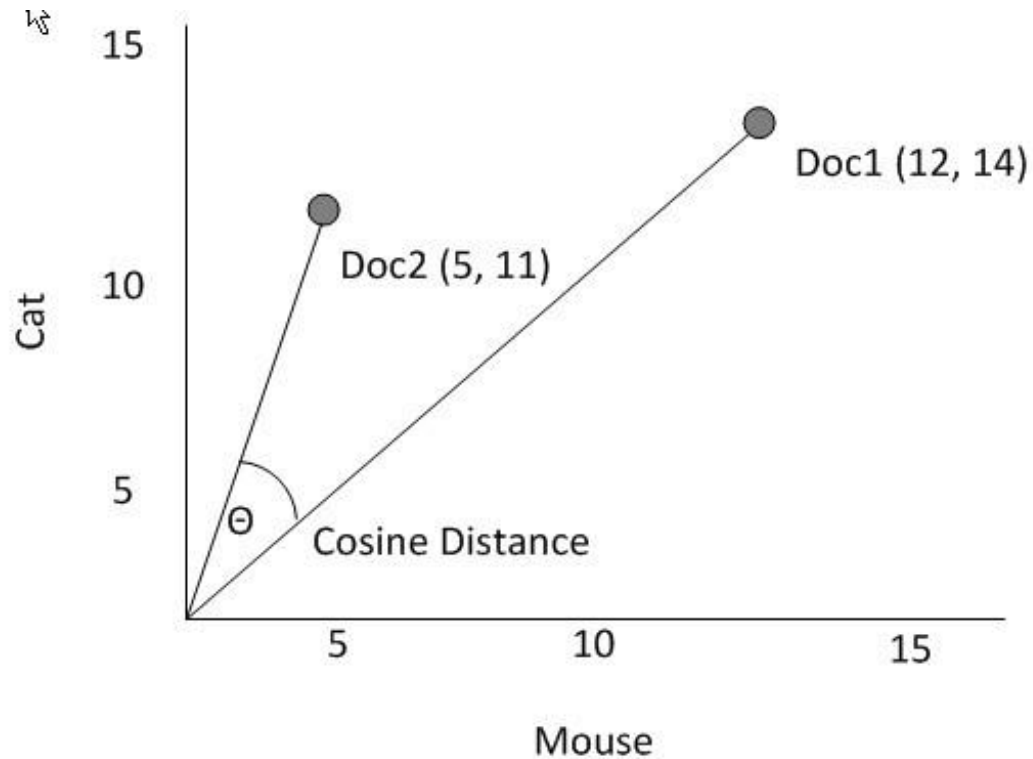
- Other vector objects: gene features in micro-arrays, ...
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If  $d_1$  and  $d_2$  are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / (\|d_1\| \|d_2\|),$$

where  $\cdot$  indicates vector dot product,  $\|d\|$ : the length of vector  $d$



# Cosine Distance and Similarity



- Cosine distance is the angle subtended at the origin between the two documents. A value of 0 degrees represents identical documents and 90 degrees dissimilar documents.
  - Note that this distance is based on the relative frequency of words in a document. A document with, say, twice as many occurrences of all words compared to another document will be regarded as identical.
  - (example code <https://nickgrattan.wordpress.com/>)



# Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\|$  ,  
where  $\cdot$  indicates vector dot product,  $\|d\|$ : the length of vector  $d$
- Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

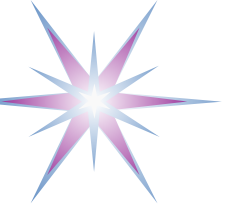
Solution:

$$d_1 \cdot d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$$

$$\|d_1\| = (5*5 + 0*0 + 3*3 + 0*0 + 2*2 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (3*3 + 0*0 + 2*2 + 0*0 + 1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d_1, d_2) = 0.94$$



# Summary and Takeaway

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Measure data similarity
- Statistical data analysis is a first step of data preprocessing.
- Many methods have been developed but still an active area of research.



## Practice Part – Investigating Selected Dataset

- Use self-study exercises provided for this course both in Python and RapidMiner
- Investigate and visualize dataset characteristics

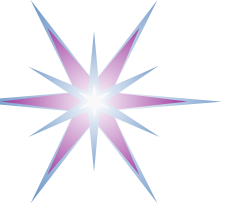




# Test yourself – Data Statistics

---

Test yourself with the exercises in the next slides



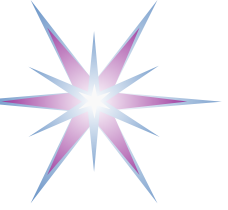
# Test yourself

Q1: Calculate the min and std. variance of the following data:

PRICE - RM '000	130	137	128	390	140	241	342	143
SQ. M OF FLOOR	135	140	100	360	175	270	200	170

Q2: Calculate the mean price of the following low-cost houses, in various localities across the country:

PRICE - RM '000 (x)	36	37	38	39	40	41	42	43
NO. OF LOCALITIES (f)	3	14	10	36	73	27	20	17



# Test yourself

Q3: From a sample information, a population of housing estate is believed have a “normal” distribution of  $X \sim (155, 45)$ .

What is the general adjustment to obtain a Standard Normal Distribution of this population?

Q4: Consider the following ROI for two types of investment:

A: 3.6, 4.6, 4.6, 5.2, 4.2, 6.5

B: 3.3, 3.4, 4.2, 5.5, 5.8, 6.8

Decide which investment you would choose.



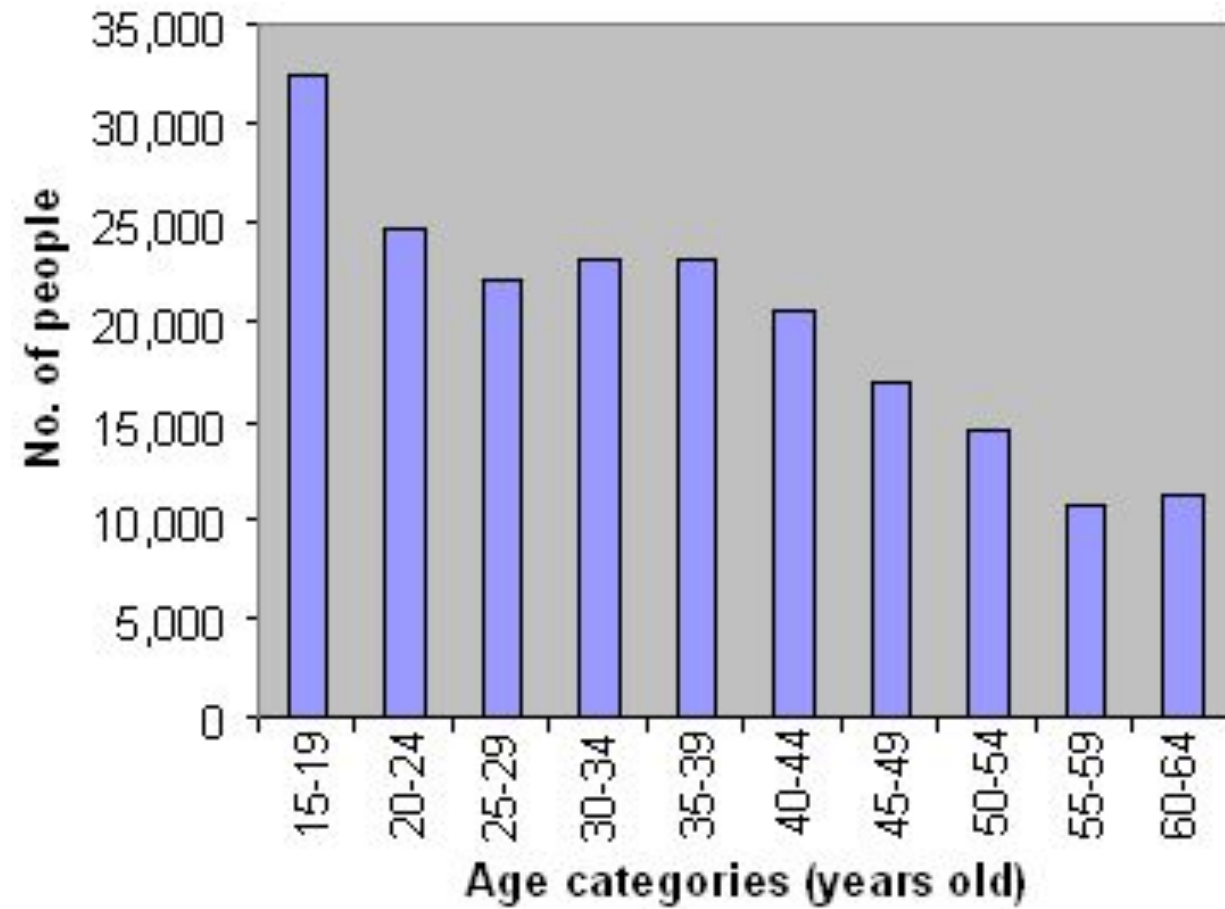
# Test yourselves!

Q5: Find:

$\phi(\text{AGE} > \text{"30-34"})$

$\phi(\text{AGE} \leq \text{"20-24"})$

$\phi(\text{"35-39"} \leq \text{AGE} < \text{"50-54"})$





# Test yourself

Q6: You are asked by a property marketing manager to ascertain whether or not *distance to work* and *distance to the city* are “equally” important factors influencing people’s choice of house location.

You are given the following data for the purpose of testing:

Explore the data as follows:

- Create histograms for both *distances*. Comment on the shape of the histograms. What is your conclusion?
- Construct scatter diagram of both *distances*. Comment on the output.
- Explore the data and give some analysis.
- Set a hypothesis that means of both distances are the same. Make your conclusion.



# Questions: Normal distribution (1)

Your sample found that the mean price of “affordable” homes in Johor Bahru, Y, is RM 155,000 with a variance of  $RM\ 3.8 \times 10^7$ . On the basis of a normality assumption, how sure are you that:

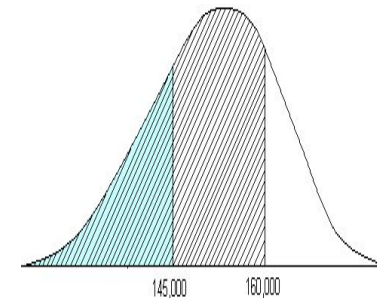
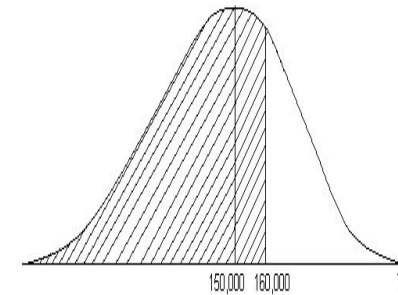
- (a) The mean price is really  $\leq RM\ 160,000$
- (b) The mean price is between RM 145,000 and 160,000

Answer (a):

$$\begin{aligned} P(Y \leq 160,000) &= P\left(Z \leq \frac{160,000 - 155,000}{\sqrt{3.8 \times 10^7}}\right) \\ &= P(Z \leq 0.811) \\ &= 0.1867 \end{aligned}$$

Using Z-table, the required probability is:

$$1 - 0.1867 = 0.8133$$



Always remember: to convert to SND, subtract the mean and divide by the std. dev.



## Questions: Normal distribution (2)

Answer (b):

$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{145,000 - 155,000}{\sqrt{3.8 \times 10^7}} = -1.622$$

$$Z_2 = \frac{X_2 - \mu}{\sigma} = \frac{160,000 - 155,000}{\sqrt{3.8 \times 10^7}} = 0.811$$

$$P(Z_1 < -1.622) = 0.0455; P(Z_2 > 0.811) = 0.1867$$

$$\begin{aligned} \therefore P(145,000 < Z < 160,000) \\ &= P(1 - (0.0455 + 0.1867)) \\ &= 0.7678 \end{aligned}$$



## Questions: Normal distribution (3)

You are told by a property consultant that the average rental for a shop house in Johor Bahru is RM 3.20 per sq. After searching, you discovered the following rental data:

2.20, 3.00, 2.00, 2.50, 3.50, 3.20, 2.60, 2.00,  
3.10, 2.70

What is the probability that the rental is greater than RM 3.00?