

## Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

### Introduction to Statistics

[naydin@yildiz.edu.tr](mailto:naydin@yildiz.edu.tr)  
<http://www3.yildiz.edu.tr/~naydin>

## Data analysis – “The Concept”

- Approach to de-synthesizing **data**, **informational**, and/or **factual** elements to answer research questions
- Method of putting together **facts** and **figures** to solve research problems
- Systematic process of utilizing **data** to address research questions
- Breaking down research issues through utilizing **controlled data** and **factual information**

## Categories of data analysis

- Narrative (e.g. laws, arts)
- Descriptive (e.g. social sciences)
- Statistical/mathematical (pure/applied sciences)
- Audio-Optical (e.g. telecommunication)
- Others
- Most research analyses adopt the first three
- The second and third are most popular in pure, applied, and social sciences


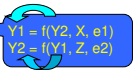
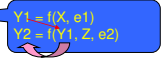
## Statistical Methods

- Something to do with “**statistics**”
  - **Statistics**
    - meaningful quantities about a sample of objects, things, persons, events, phenomena, etc.
    - Widely used in many fields (social sciences, engineering, etc.)
    - Simple to complex issues. E.g.
      - correlation
      - anova
      - manova
      - regression
      - econometric modelling
- Two main categories:
  - **Descriptive statistics**
  - **Inferential statistics**

## Descriptive statistics

- Use sample information to explain/make abstraction of population “**phenomena**”
- Common “**phenomena**”:
  - Association
  - Tendency
  - Causal relationship
  - Trend,
  - Pattern,
  - Dispersion,
  - Range
- Used in non-parametric analysis
  - e.g. chi-square, t-test, 2-way anova

## Inferential statistics

- Using sample statistics to infer some **phenomena** of population parameters
- Common **phenomena**: cause-and-effect
  - **One-way relationship** 
  - **Multi-directional relationship** 
  - **Recursive** 
- Use parametric analysis

## Which one to use?

- Nature of research
  - Descriptive in nature?
  - Attempts to infer, predict, find cause-and-effect, influence, relationship?
  - Is it both?
- Research design (including variables involved)
  - E.g. outputs/results expected
    - research issue
    - research questions
    - research hypotheses

7

## How to avoid mistakes - Useful tips

- Crystalize the research problem
  - operability of it!
- Read literature on data analysis techniques
- Evaluate various techniques that can do similar things with respect to research problem
- Know what a technique does and what it doesn't
- Consult people, esp. supervisor
- Pilot-run the data and evaluate results

8

## Principles of analysis...

- Goal of an analysis is to...
  - explain cause-and-effect phenomena
  - relate research with real-world event
  - predict/forecast the real-world phenomena based on research
  - find answers to a particular problem
  - make conclusions about real-world event based on the problem
  - learn a lesson from the problem

9

## ...Principles of analysis...

- Data cannot talk
- An analysis contains some aspects of scientific reasoning/argument:
  - Define
  - Interpret
  - Evaluate
  - Illustrate
  - Discuss
  - Explain
  - Clarify
  - Compare
  - Contrast

10

## ...Principles of analysis...

- An analysis must have four elements:
  - Data/information (what)
  - Scientific reasoning/argument
    - what? who? where? how? what happens?
  - Finding
    - what results?
  - Lesson/conclusion
    - so what? so how? therefore, ...

11

## ...Principles of data analysis...

- Basic guide to data analysis:
  - Analyze, not narrate
  - Go back to research flowchart
  - Break down into research objectives and research questions
  - Identify phenomena to be investigated
  - Visualize the expected answers
  - Validate the answers with data
  - Do not tell something not supported by data

12

## ...Principles of data analysis

- When analyzing:
  - Be objective
  - Be accurate
  - Be true
- Separate facts and opinion
- Avoid “wrong” reasoning/argument.
  - E.g. mistakes in interpretation.

13

## Description of samples and populations

- Statistics is about making statements about a population from data observed from a representative sample of the population.
- A population
  - a collection of subjects whose properties are to be analyzed.
  - contains all subjects of interest.
- A sample
  - a part of the population of interest
  - a subset selected by some means from the population.

14

## Description of samples and populations

- A parameter
  - a numerical value that describes a characteristic of a population
- A statistic
  - a numerical measurement that describes a characteristic of a sample
- We use a statistic to infer something about a parameter.

15

## Data exploration

- After collecting data, the next step towards statistical inference and decision making is to perform data exploration,
  - which involves visualizing and summarizing the data.
    - The objective of data visualization is to obtain a high level understanding of the sample and their observed (measured) characteristics.
- To make the data more manageable, we need to further reduce the amount of information in some meaningful ways so that we can focus on the key aspects of the data.
  - Summary statistics are used for this purpose.

16

## Data exploration

- Using data exploration techniques, we can learn about the distribution of a variable.
  - The distribution of a variable tells us
    - the possible values it can take,
    - the chance of observing those values,
    - how often we expect to see them in a random sample from the population.
- Through data exploration, we might detect previously unknown patterns and relationships that are worth further investigation.
  - We can also identify possible data issues, such as unexpected or unusual measurements, known as outliers.

17

## Statistical inference

- We collect data on a sample from the population in order to learn about the whole population.
  - {For example, Mackowiak, et al. (1992) measure the normal body temperature for 148 people to learn about the normal body temperature for the entire population.
    - In this case, we say we are estimating the unknown population average.
      - However, the characteristics and relationships in the whole population remain unknown.
    - Therefore, there is always some uncertainty associated with our estimations.}

18

## Statistical inference

- The mathematical tool to address uncertainty in Statistics
  - probability.
- The process of using the data to draw conclusions about the whole population, while acknowledging the extent of our uncertainty about our findings, is called **statistical inference**.
  - The knowledge we acquire from data through statistical inference allows us to make decisions with respect to the scientific problem that motivated our study and our data analysis.

19

## Data types

- The type(s) of data collected in a study determine
  - the type of statistical analysis that can be used
  - which hypotheses can be tested
  - which model we can use for prediction.
- Broadly speaking, data can be classified into two major types:
  - categorical
  - quantitative

20

## Categorical data

- **Categorical data** can be grouped into categories based on some **qualitative** trait.
- The resulting data are merely **labels** or **categories**,
  - {examples include
    - gender (male and female)
    - ethnicity (e.g., Caucasian, Asian, African)}
- We can further sub-classify categorical data into two types:
  - nominal
  - ordinal

21

## Categorical data

- **Nominal data**
  - When there is no natural ordering of the categories we call the data **nominal**.
    - {Hair color is an example of nominal data}
- **Ordinal data**
  - When the categories may be ordered, the data are called **ordinal variables**.
    - {Categorical variables that judge pain (e.g., none, little, heavy) or income (low-level income, middle-level income, or high-level income) are examples of ordinal variables.}

22

## Quantitative data

- **Quantitative data** are numerical measurements where
  - the numbers are associated with a scale measure rather than just being simple labels.
- **Quantitative data** fall in two categories:
  - discrete
  - continuous

23

## Quantitative data

- **Discrete quantitative data**
  - numeric data variables that have a finite or countable number of possible values.
    - When data represent counts, they are discrete.
      - {Examples include household size or the number of kittens in a litter.}
- **Continuous quantitative data**
  - The real numbers are continuous with no gaps;
    - physically measurable quantities like length, volume, time, mass, etc., are generally considered continuous.

24

## Categorical vs Quantitative data

- **Categorical data** are typically summarized using **frequencies** or **proportions** of **observations** in each category
- **Quantitative data** typically are summarized using **averages** or **means**.

25

## Describing Data

- Once data are collected, the next step is to summarize it all to get a handle on the big picture.
- Statisticians describe data in two major ways:
  - with **pictures**
    - that is, charts and graphs
  - with **numbers**,
    - called descriptive statistics.

26

## Charts and graphs

- Data are summarized in a visual way using charts and/or graphs
  - Some of the basic graphs used include **pie charts** and **bar charts**
  - Some data are numerical
  - Data representing counts or measurements need a different type of graph that either keeps track of the numbers themselves or groups them into numerical groupings.
    - One major type of graph that is used to graph numerical data is a **histogram**.

27

## Descriptive statistics

- Numbers that describe a data set in terms of its important features
  - **Categorical data** are typically summarized using
    - the number of individuals in each group (the **frequency**)
    - the percentage of individuals in each group (the **relative frequency**)
  - **Numerical data** represent measurements or counts, where the actual numbers have meaning
    - more features can be summarized
      - measures of center
      - measures of spread
      - measures of the relationship between two variables
    - Some descriptive statistics are better than others,
    - some are more appropriate than others

28

## Data Visualization and Summary Statistics

- Preliminary steps before analysis:
  - defining the scientific question we try to answer,
  - selecting a set of representative members from the population of interest
  - collecting data (either through observational studies or randomized experiments),
- Analysis usually begins with **data exploration**.
  - We start by focusing on data exploration techniques for one variable at a time.

29

## Data Visualization and Summary Statistics

- Objective is
  - to develop a high-level understanding of the data,
  - learn about the possible values for each characteristic,
  - find out how a characteristic varies among individuals in our sample.
- Basically, we want to learn about the **distribution** of variables.
  - Recall that for a variable, the **distribution** shows
    - the possible values,
    - the chance of observing those values,
    - how often we expect to see them in a random sample from the population.

30

## Data Visualization and Summary Statistics

- The data exploration methods allow us to reduce the amount of information so that we can focus on the key aspects of the data.
- We do this by using **data visualization** techniques and **summary statistics**.
- The visualization techniques and summary statistics we use for a variable depend on its type
  - Recall that we can classify them into two general groups:
    - Numerical (quantitative) variables
      - discrete, continuous
    - Categorical variables
      - nominal, ordinal

31

## Graphical summarization of data

- Before blindly applying the statistical analysis, it is always good to look at the raw data,
  - usually in a graphical form,and then use graphical methods to summarize the data in an easy to interpret format.
  - A Picture is worth a thousand word
- The types of graphical displays that are most frequently used by engineers
  - scatterplots, time series, box-and-whisker plots, and histograms.

32

## Exploring Categorical Variables

- A simple way for summarizing the data is to create a table that shows the number of times each category has been observed.
- **The number of times a specific category is observed is called frequency.**
  - We denote the frequency for category  $c$  by  $n_c$ .
- The sum of the frequencies for all categories is equal to the total sample size

$$\sum_c n_c = n$$

33

## Relative Frequency and Percentage

- The **relative frequency** is the sample proportion for each possible category.
- It is obtained by dividing the frequencies  $n_c$  by the total number of observations  $n$ :
$$p_c = \frac{n_c}{n}$$
- Relative frequencies are sometimes presented as **percentages** after multiplying proportions  $p_c$  by 100.
- Since the relative frequencies are proportions of the sample size, their sum is 1,

$$\sum_c p_c = 1$$

where  $p_c$  is the relative frequency of category  $c$ .

34

## Exploring Numerical Variables

- For numerical variables, we are especially interested in two key aspects of the distribution:
  - its **location**
    - refers to the **central tendency** of values, that is, the point around which most values are gathered.
  - its **spread**
    - refers to the **dispersion** of possible values, that is, how scattered the values are around the location.

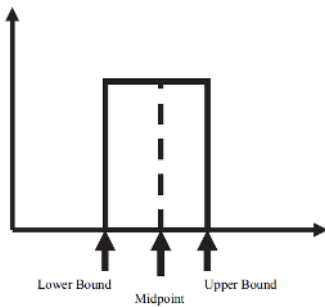
35

## Histograms

- defined as a frequency distribution commonly used to visualize **numerical variables**.
- A **histogram** is similar to a bar graph after the values of the variable are grouped (**binned**) into a finite number of nonoverlapping intervals (**bins**), usually of equal width.
- Given  $N$  samples or measurements,  $x_i$  ranging from  $X_{min}$  to  $X_{max}$ , the samples are binned into bins
- Typically, the number of bins is on the order of 7–14, depending on the nature of the data.
  - In addition, we typically expect to have at least 3 samples per bin.
    - Sturges' rule may also be used to estimate the number of bins and is given by  $k = 1 + 3.3 \log(n)$ .
      - where  $k$  is the number of bins and  $n$  is the number of samples.

36

## Histograms



- One bin of a histogram plot
- The bin is defined by
  - a lower bound,
  - a midpoint,
  - an upper bound

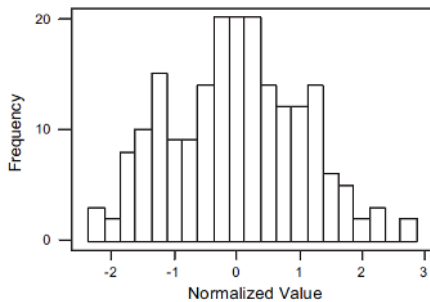
37

## Histograms

- constructed by plotting the number of samples in each bin.
  - horizontal axis,
    - the sample value,
  - the vertical axis,
    - the number of occurrences of samples falling within a bin
- Next slide illustrates a histogram for 1000 samples drawn from a normal distribution with mean ( $\mu$ ) = 0 and standard deviation ( $\sigma$ ) = 1.0.

38

## Histograms



39

## Histograms

- Two useful measures in describing a histogram:
  - the absolute frequency in one or more bins
    - $f_i$  = absolute frequency in  $i$ th bin
  - the relative frequency in one or more bins
    - $f_i/n$  = relative frequency in  $i$ th bin,
      - where  $n$  is the total number of samples being summarized in the histogram
- The histogram can exhibit several shapes
  - symmetric, skewed, or bimodal.

40

## Histograms

- The bar height for each interval could be set to its relative frequency  $p_c = n_c/n$ , or the percentage  $p_c \times 100$ , of observations that fall into that interval.
- For histograms, however, it is more common to use the density instead of the relative frequency or percentage.
  - The density is the relative frequency for a unit interval.
    - It is obtained by dividing the relative frequency by the interval width:
 
$$f_c = p_c / w_c$$
      - Here,  $p_c = n_c/n$  is the relative frequency with  $n_c$  as the frequency of interval  $c$  and  $n$  as the total sample size.
      - The width of interval  $c$  is denoted  $w_c$ .

41

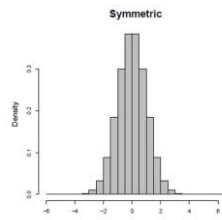
## Histograms

- The height of each bar in density histogram shows the density of the corresponding interval (as opposed to its frequency).
- For each interval  $c$ , the area of the corresponding bar in the density histogram is calculated as follows (height  $\times$  width):
 
$$a_c = f_c \times w_c = (p_c / w_c) \times w_c = p_c$$
- Therefore, the area of each bar (rectangle) is the relative frequency for the corresponding interval.
  - Since the sum of relative frequencies is 1, the total area of bars in a density histogram is 1.

42

## Histograms

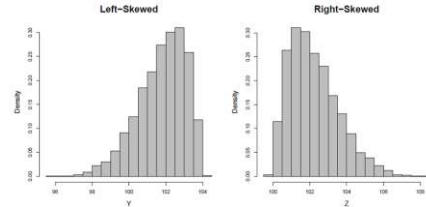
- When creating a histogram, it is important to choose an appropriate value for  $w$  (Number of Bins).
- Besides the location and spread of a distribution, the shape of a histogram also shows us how the observed values spread around the location.
- We say the following histogram is symmetric around its location (here, zero) since the densities are the [almost] same for any two intervals that are equally distant from the center.



43

## Histograms

- In many situations, we find that a histogram is stretched to the left or right.
- We call such histograms skewed.
  - More specifically, we call them left-skewed if they are stretched to the left, or right-skewed if they are stretched to the right.

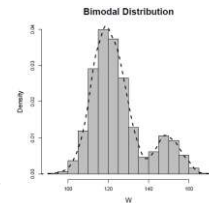
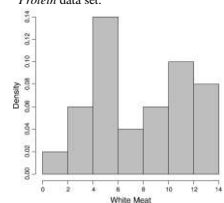


44

## Histograms

- The histograms in previous slides, whether symmetric or skewed, have one thing in common – they all have one peak (or mode).
- We call such histograms (and their corresponding distributions) unimodal.
- Sometimes histograms have multiple modes.
  - The bimodal histogram appears to be a combination of two unimodal histograms.
    - Indeed, in many situations bimodal histograms (and multimodal histograms in general) indicate that the underlying population is not homogeneous and may include two (or more in case of multimodal histograms) subpopulations.

## Histograms

- Histogram of a bimodal distribution
 
- Histogram of protein consumption in 25 European countries for white meat in Protein data set.
 
  - The histogram is bimodal, which indicates that the sample might be comprised of two subgroups

45

46

## Histograms

- The histogram is important because it serves as
  - a rough estimate of the true probability density function or
  - probability distribution of the underlying random process from which the samples are being collected.
- The probability density function or probability distribution is a function that quantifies the probability of a random event,  $x$ , occurring.
  - When the underlying random event is discrete in nature, we refer to the probability density function as the probability mass function

47

## Measures of Central Tendency

- Histograms are useful for visualizing numerical data and identifying their location and spread.
- However, we typically use descriptive or summary statistics for more precise specification of the
  - central tendency
  - dispersion
 of observed values.

48



## Measures of Central Tendency

- A **central tendency** is a central or typical value for a probability distribution.
  - also called a **center** or **location of the distribution**.
- Measures of central tendency are often called **averages**.
- There are several measures that reflect the central tendency
  - sample mean,
  - sample median,
  - sample mode.

49

## Mean

- In mathematics, mean has several different definitions depending on context.
- In probability and statistics
  - **mean** and **expected value** are synonymous
- In case of a discrete probability distribution of random variable  $x$ ,
  - the **mean** is equal to the sum over every possible value weighted by the probability of that value

$$\mu = \sum xP(x)$$

50

## Mean

- For a data set, the terms
  - arithmetic mean,
  - mathematical expectation,
  - sometimes average
 are used synonymously to refer to a central value of a discrete set of numbers
  - specifically, the sum of the values divided by the number of values.
- If the data set were based on a series of observations obtained by sampling from a statistical population,
  - the **arithmetic mean** is termed as the **sample mean** to distinguish it from the **population mean**

51

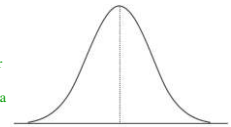
## Mean

- **Arithmetic mean** (or simply **mean**) of a sample  $x_1, x_2, \dots, x_n$ , usually denoted by  $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- It is used when the spread of the data is fairly similar on each side of the mid point

- when the data are “normally distributed”.
  - If a value is a lot smaller or larger than the others, “skewing” the data, the mean will then not give a good picture of the typical value.



52

## Mean

- **Geometric mean** is an average that is useful for sets of positive numbers that are interpreted according to their product, e.g. rates of growth

$$\bar{x} = \sqrt[n]{x_1 \times x_2 \times \dots \times x_n}$$

- **Harmonic mean** is an average which is useful for sets of numbers that are defined in relation to some unit, for example speed

$$\bar{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

53

## Mean

- The relationship between Arithmetic mean, Geometric mean, and Harmonic mean:
  - Arithmetic mean  $\times$  Harmonic mean = Geometric mean<sup>2</sup>
- Arithmetic mean, Geometric mean, and Harmonic mean satisfy the following inequalities:
  - Arithmetic mean  $\geq$  Geometric mean  $\geq$  Harmonic mean
  - Equality holds if and only if all the elements of the given sample are equal
- The **arithmetic mean** is best used in situations where:
  - the data are not skewed (no extreme outliers)
  - the individual data points are not dependent on each other
- The **geometric mean** should be used whenever the data are inter-related
- The **harmonic mean** is best to use when there is:
  - A large population where the majority of the values are distributed uniformly but where there are a few outliers with significantly higher values

54

## Mean

- Weighted arithmetic mean is used if one wants to combine average values from samples of the same population with different sample sizes

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

- The weights  $w_i$  represent the sizes of the different samples.
- In other applications, they represent a measure for the reliability of the influence upon the mean by respective values.

55

## Mean

- A power mean is a mean of the form

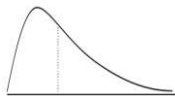
$$M_p = \left( \frac{1}{n} \sum_{k=1}^n x_k^p \right)^{1/p}$$

$M_{-\infty}$	minimum
$M_{-1}$	harmonic mean
$M_0$	geometric mean
$M_1$	arithmetic mean
$M_2$	root-mean-square
$M_{\infty}$	maximum

56

## Median

- Sometimes known as the mid-point.
  - It is used to represent the average when the data are not symmetrical (skewed distribution)



- The median value of a group of observations or samples,  $x_p$ , is the middle observation when samples,  $x_p$ , are listed in descending order.

- Note that if the number of samples,  $n$ , is odd, the median will be the middle observation.
- If the sample size,  $n$ , is even, then the median equals the average of two middle observations.
- Compared with the sample mean, the sample median is less susceptible to outliers.

57

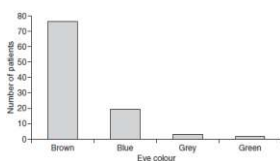
## Median

- The median may be given with its inter-quartile range (IQR).
- The 1st quartile point has the 1/4 of the data below it
- The 3rd quartile point has the 3/4 of the sample below it
- The IQR contains the middle 1/2 of the sample
- This can be shown in a “box and whisker” plot.

58

## Mode

- the most common of a set of events
  - used when we need a label for the most frequently occurring event
  - Example: An eye clinic sister noted the eye colour of 100 consecutive patients. The results are shown below

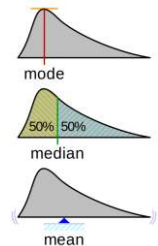
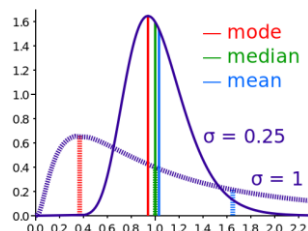


- Graph of eye colour of patients attending an eye clinic.
- In this case the mode is brown, the commonest eye colour.

59

## Mean, Median, Mode

- Comparison of the arithmetic mean, median and mode of two skewed (log-normal) distributions.
- Geometric visualisation of the mode, median and mean of an arbitrary probability density function.



60

## Measures of Variability

- When summarizing the variability of a population or process, we typically ask,
  - “How far from the center (sample mean) do the samples (data) lie?”
- To answer this question, we typically use the following estimates that represent the spread of the sample data:
  - sample variance,
  - sample standard deviation.
  - interquartile ranges,

61

## Variance and standard deviation

- Two common summary statistics for measuring dispersion are the **sample variance** and **sample standard deviation**.
- These two summary statistics are based on the **deviation** of observed values from the mean as the center of the distribution.
- For each observation, the deviation from the mean is calculated as

$$x_i - \bar{x}$$

62

## Variance and standard deviation

- The sample **variance** is a common measure of dispersion based on the squared deviations.

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

- The square root of the variance is called the sample **standard deviation**.

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

63

## Measures of Variability

- **Standard deviation (SD)** is used for data which are “normally distributed”,
  - to provide information on how much the data vary around their mean.
- **SD** indicates how much a set of values is spread around the average.
  - A range of one **SD** above and below the mean (abbreviated to  $\pm 1$  **SD**) includes 68.2% of the values.
  - $\pm 2$  **SD** includes 95.4% of the data.
  - $\pm 3$  **SD** includes 99.7%.

64

## Variance and standard deviation

- Some properties that can help you when interpreting a standard deviation:
  - The standard deviation can never be a negative number.
  - The smallest possible value for the standard deviation is 0
    - (when every number in the data set is exactly the same).
  - Standard deviation is affected by outliers, as it's based on distance from the mean, which is affected by outliers.
  - The standard deviation has the same units as the original data, while variance is in square units.

65

## Measures of Variability

- It is important to note that for normal distributions (symmetrical histograms),
  - sample mean and sample deviation are the only parameters needed to describe the statistics of the underlying phenomenon.
- Thus, if one were to compare two or more normally distributed populations,
  - one only needs to test the equivalence of the means and variances of those populations.

66

## Quantile

- comes from the word **quantity**
- A **quantile** is where a sample is divided into equal-sized, adjacent, subgroups
  - (quantile is also called a **fractile**)
- It can also refer to dividing a probability distribution into areas of equal probability
- **Quartiles** are also quantiles;
  - they divide the distribution into 4 equal parts.
- **Percentiles** are quantiles;
  - they divide a distribution into 100 equal parts
- **Deciles** are quantiles;
  - they divide a distribution into 10 equal parts.

67

## Percentiles

- the most common way to report relative standing of a number within a data set
- A percentile is the percentage of individuals in the data set who are below where your particular number is located.
  - For example,
  - if your exam score is at the 90th percentile, that means
    - 90% of the people taking the exam with you scored lower than you did
    - 10 percent scored higher than you did

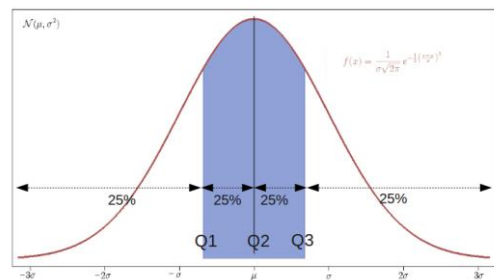
68

## Quantile

- For sampled data, the **median** is also known as
  - the **2nd quartile, Q2**.
- Given Q2, we can find the 1st quartile, Q1,
  - by simply taking the median value of those samples that lie below the 2nd quartile.
- We can find the 3d quartile, Q3,
  - by taking the median value of those samples that lie above the 2nd quartile.
- Quartiles can also be found in terms of percentiles:
  - 1st quartile is 25th percentile
  - 2nd quartile is 50th percentile
  - 3rd quartile is 75th percentile

69

## Measures of Variability



70

## Five-number summary

- The **minimum (min)**, which is the smallest value of the variable in our sample, is in fact the **0 quantile**.
- On the other hand, the **maximum (max)**, which is the largest value of the variable in our sample, is the **1 quantile**.
- The minimum and maximum along with quartiles (Q1, Q2, and Q3) are known as **five-number summary**.
- These are usually presented in the increasing order:
  - min, 1st quartile, median, 3rd quartile, max
  - min, 25th percentile, median, 75th percentile, max
- This way, the **five-number summary** provides
  - 0, 0.25, 0.50, 0.75, and 1 quantiles

71

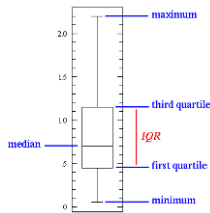
## Five-number summary

- The five-number summary can be used to derive two measures of dispersion:
  - the **range**
    - the difference between the maximum observed value and the minimum observed value.
  - the **interquartile range (IQR)**
    - the difference between the third quartile (Q3) and the first quartile (Q1).

$$IQR = Q3 - Q1$$

72

## Boxplot



- This simplest possible box plot displays the full range of variation (from **min** to **max**), the likely range of variation (the **IQR**), and a typical value (the **median**).
- Not uncommonly real datasets will display surprisingly high maximums or surprisingly low minimums called **outliers**.

– John Tukey has provided a precise definition for two types of outliers:

- $3 \times \text{IQR}$  or more above the  $Q3 \Rightarrow (Q3 + 3 \times \text{IQR})$
- $3 \times \text{IQR}$  or more below the  $Q1 \Rightarrow (Q1 - 3 \times \text{IQR})$

– Suspected outliers are slightly more central versions of outliers:

- $1.5 \times \text{IQR}$  or more above the  $Q3 \Rightarrow (Q3 + 1.5 \times \text{IQR})$
- $1.5 \times \text{IQR}$  or more below the  $Q1 \Rightarrow (Q1 - 1.5 \times \text{IQR})$

73

## Data Transformation

- We rely on data transformation techniques
  - to reduce the influence of extreme values in our analysis.
- The reasons for data transformation:
  - to make the distribution of the data **normal**,
  - to create more informative graphs of the data,
  - better outlier identification
  - increasing the sensitivity of statistical tests
- Two of the most common transformation functions for this purpose are
  - logarithm
  - square root.

74

## Coefficient of Variation

- In general, the **coefficient of variation** is used to compare variables in terms of their dispersion when the means are substantially different
  - possibly as the result of having different measurement units.
- To quantify dispersion independently from units, we use the **coefficient of variation**,
  - which is the **standard deviation** divided by the **sample mean**
    - assuming that the mean is a positive number:

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

75

## Scaling and Shifting Variables

- In general, when we multiply the observed values of a variable by a constant  $a$ , its mean, standard deviation, and variance are multiplied by  $a$ ,  $|a|$ , and  $a^2$ , respectively.
  - That is, if  $y = ax$ , then
    - $\bar{y} = a\bar{x}$ ,  $s_y = |a|s_x$ ,  $s_y^2 = a^2s_x^2$
- The **coefficient of variation** is not affected.

$$CV_y = \frac{s_y}{\bar{y}} = \frac{as_x}{a\bar{x}} = \frac{s_x}{\bar{x}} = CV_x$$

76

## Scaling and Shifting Variables

- If we shift the observed values by  $b$ , i.e.,  $y = x + b$ , then
  - $\bar{y} = \bar{x} + b$ ,  $s_y = s_x$ ,  $s_y^2 = s_x^2$
- If we multiply the observed values by the constant  $a$  and then add the constant  $b$  to the result, i.e.,  $y = ax + b$ , then
  - $\bar{y} = a\bar{x} + b$ ,  $s_y = |a|s_x$ ,  $s_y^2 = a^2s_x^2$
- the coefficient of variation will change.
- If  $y = ax + b$  (assuming  $a > 0$  and  $b = 0$ ), then

$$CV_y = \frac{s_y}{\bar{y}} = \frac{as_x}{a\bar{x} + b} \neq \frac{s_x}{\bar{x}}$$

77

## Variable Standardization

- Variable standardization** is a common **linear transformation**,
  - where we subtract the sample mean  $\bar{x}$  from the observed values and divide the result by the sample standard deviation  $s$ ,
    - in order to shift the mean to zero and make the standard deviation 1:
- Using such transformation is especially common in regression analysis and clustering.
- Subtracting  $\bar{x}$  from the observations shifts the sample mean to **zero**.
  - This, however, does not change the standard deviation.
    - Dividing by  $s$ , on the other hand, changes the sample standard deviation to 1

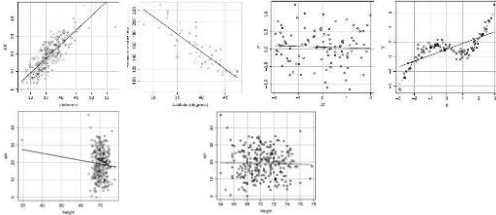
78

## Exploring Relationships

- Two numerical variables

- Scatterplot Matrix

- The relationship is simply an association and should not be regarded as causation since the data come from an observational study



79

## Exploring Relationships

- Two numerical variables

- Correlation

- To quantify the strength and direction of a linear relationship between two numerical variables
  - Considering a set of observed pairs of values,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , for a sample of  $n$  observations Pearson's correlation coefficient,  $r$ , can be used as a summary statistic

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

- $s_x$  and  $s_y$  denote the sample standard deviations
- The values of  $r$  are always between -1 and +1.
- The relationship is strong when  $r$  approaches -1 or +1.
- The sign of  $r$  shows the direction (negative or positive) of the linear relationship.

80

## Exploring Relationships

- Two numerical variables

- Cross-correlation

- Is a measure of similarities of two signals

$$r_{xy}(k) = \sum_{n=0}^{N-1} x(n)y(k+n)$$

- Auto-correlation

- when  $x(n) = y(n)$

- Algorithm for Cross/Auto-correlation

```
for k=1:K+N-1
    for n=1:N
        r(k)=r(k)+x(n)*y(k+n-1);
    end
end
```

81

## Exploring Relationships

- Two numerical variables

- Sample Covariance

- If the standard deviations are removed from the denominator in Pearson's correlation coefficient, the statistic is called the sample covariance,

$$v_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Therefore

$$r_{xy} = \frac{v_{xy}}{s_x s_y}$$

82

## Exploring Relationships

- Two categorical variables

- contingency tables

- used to summarize the relationship between several categorical variables.
- a special type of frequency distribution table, where two variables are shown simultaneously.

	Heart attack	No heart attack	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

- sample proportion can be calculated

- proportion of people suffered from a heart attack in the placebo group  
 $p_1 = 189/11034 = 0.0171$
- proportion of people suffered from heart attack in the aspirin group  
 $p_2 = 104/11037 = 0.0094$

83

## Exploring Relationships

- Two categorical variables

- One way of measuring the strength of the relationship is to calculate the difference of proportions,  $p_2 - p_1$ .

- In the example,  $p_2 - p_1 = -0.0077$

- proportion of people suffered from heart attack reduces by 0.0077 in the aspirin group compared to the placebo group, or

$$\frac{p_2 - p_1}{p_1} \times 100\% = \frac{-0.0077}{0.0171} \times 100\% = -45\%$$

- Another common summary statistic for comparing sample proportions is the relative proportion,  $p_2/p_1$

- If  $p_2 = p_1$ ,  $p_2/p_1 = 1$ , which is interpreted as no relationship between the two categorical variables

84

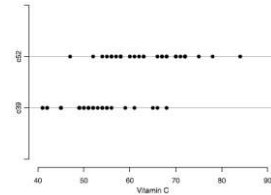
## Exploring Relationships

- Two categorical variables
  - It is more common to compare the sample odds
 
$$o = \frac{p}{1-p}$$
    - where  $p$  is the sample proportion for the event of interest
  - The odds of a heart attack in the placebo group,  $o_1$ , and in the aspirin group,  $o_2$ , are
 
$$o_1 = \frac{0.0171}{(1-0.0171)} = 0.0174, \quad o_2 = \frac{0.0094}{(1-0.0094)} = 0.0095.$$
  - We usually compare the sample odds using the sample odds ratio
 
$$OR_{21} = \frac{o_2}{o_1} = \frac{0.0095}{0.0174} = 0.54.$$
    - If  $OR = 1$ , no relationship between
    - Values of  $OR$  away from 1, indicate strong relationship

85

## Exploring Relationships

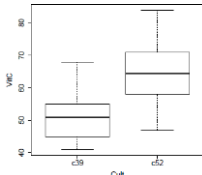
- Numerical and Categorical Variables
  - dot plots (a.k.a. strip chart)
    - The dot plots of *ascorbic acid* (one form of vitamin C) content (numerical) by *cultivar* (categorical).



86

## Exploring Relationships

- Numerical and Categorical Variables
  - boxplots of the numerical variable for different values of the categorical variable
    - This plot suggests that vitamin C content tends to be higher in the c52 group compared to the c39 group.
      - This is indicative of a possible relationship between these two variables.
        - Summary statistics of *vitamin C* content by *cultivar* from the cabbages data set



	min	Q1	Median	Q3	Max	Mean	SD	N
c39	45.0	50.0	52.0	55.0	65.0	52.5	10.0	10
c52	55.0	60.0	62.0	65.0	75.0	62.5	10.0	10

87

## Probability as a Measure of Uncertainty

- Plots and summary statistics are used to learn about the distribution of variables and to investigate their relationships.
  - However, we always remain uncertain about the true distributions and relationships in the population since we almost never have access to all of its members.
  - Furthermore, our findings based on the observed sample can change if different samples from the population were obtained.
- Therefore, when we generalize our findings from a sample to the whole population, we should explicitly specify the extent of our uncertainty.
  - We use probability as a measure of uncertainty.

88

## Probability as a Measure of Uncertainty

- A phenomenon is called **random** if its outcome (value) cannot be determined with certainty before it occurs.
- The collection of all possible outcomes  $S$  is called the **sample space**.
- To each possible outcome in the sample space, we assign a probability  $P$ ,
  - which represents how certain we are about the occurrence of the corresponding outcome.
- For an outcome  $o$ , we denote the probability as  $P(o)$ , where  $0 \leq P(o) \leq 1$ .
- The total probability of all outcomes in the sample space is always 1

89

## Probability as a Measure of Uncertainty

- An **event** is a subset of the sample space  $S$ .
- We denote the **probability of event  $E$**  as  $P(E)$ .
  - The probability of an event is the sum of the probabilities for all individual outcomes included in that event.
- For any event  $E$ , we define its complement,  $E^c$ , as the set of all outcomes that are in the sample space  $S$  but not in  $E$ .
- The probability of the complement event is
 
$$P(E^c) = 1 - P(E)$$

90

## Probability as a Measure of Uncertainty

- The **odds** of an event shows how much more certain we are that the event occurs than we are that it does not occur.
  - For event  $E$ , we calculate the odds as follows:
 
$$\frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$
- For two events  $E_1$  and  $E_2$  in a sample space  $S$ , we define their union  $E_1 \cup E_2$  as the set of all outcomes that are **at least in one of the events**.
- For two events  $E_1$  and  $E_2$  in a sample space  $S$ , we define their intersection  $E_1 \cap E_2$  as the set of outcomes that are **in both events**.

91

## Probability as a Measure of Uncertainty

- We refer to the probability of the intersection of two events,  $P(E_1 \cap E_2)$ , as their **joint probability**.
- In contrast, we refer to probabilities  $P(E_1)$  and  $P(E_2)$  as the **marginal probabilities** of events  $E_1$  and  $E_2$ .
- For any two events  $E_1$  and  $E_2$ , we have  $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
- Two events are called disjoint or mutually exclusive if they never occur together

92

## Probability as a Measure of Uncertainty

- The **conditional probability**, denoted  $P(E_1|E_2)$ , is
  - The probability of event  $E_1$  given that another event  $E_2$  has occurred. ( $P(E_2) \neq 0$ )
 
$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$
- Two events  $E_1$  and  $E_2$  are **independent** if our knowledge of the occurrence of one event does not change the probability of occurrence of the other event.
- If events  $E_1, E_2, \dots, E_n$  are independent:
 
$$P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1) \times P(E_2) \times \dots \times P(E_n)$$

93

## Probability as a Measure of Uncertainty

- According to **Bayes' theorem** or **Bayes' rule**
  - for two events  $E_1$  and  $E_2$ , the following equation shows the relationship between  $P(E_2|E_1)$  and  $P(E_1|E_2)$ :
 
$$P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)}$$
- The general form of Bayes' theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^K P(A|B_k)P(B_k)}$$

94

## Interpretation of Probability as the Relative Frequency

- The random phenomena can be observed repeatedly.
  - These repeated experiments or observations are called **trials**.
- For such random phenomena, the probability of an event can be interpreted in terms of the **relative frequency**.
- The above interpretation of probability requires two important assumptions.
  - We assume that the probability of events does not change from one trial to another.
  - We also assume that the outcome of one trial does not affect the outcome of another trial.

95

## Gaussian Distribution...

- The spread (distribution) of data may be rectangular, skewed, Gaussian, or other.
- The Gaussian distribution is given by

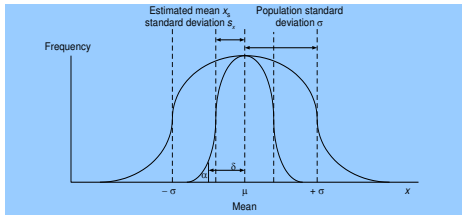
$$f(X) = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}$$

where  $\mu$  is the true mean and  $\sigma$  is the true standard deviation of a very large number of measurements.

96



## ...Gaussian Distribution



- For the **normal distribution**, **68%** of the data lies within  **$\pm 1$  SD**.
- By measuring samples and averaging, we obtain the estimated mean  $\bar{x}$ , which has a smaller standard deviation  $s_x$ .
- $\alpha$  is the tail probability that  $\bar{x}_s$  does not differ from  $\mu$  by more than  $\delta$ .

97

## Poisson Probability...

- The **Poisson probability** density function is another type of distribution.
  - It can describe, among other things, the probability of radioactive decay events, cells flowing through a counter, or the incidence of light photons.
- The probability that a particular number of events  $K$  will occur in a measurement (or during a time) having an average number of events  $m$  is

$$p(K, m) = \frac{e^{-m} m^K}{K!}$$

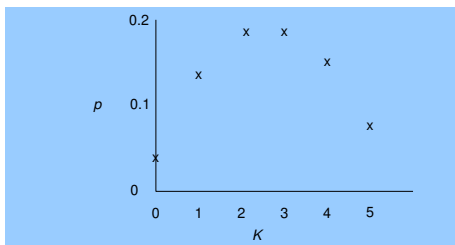
- The standard deviation of the Poisson distribution is

$$\sqrt{m}$$

98

## ...Poisson Probability

- A typical Poisson distribution for  $m = 3$ .



99

## Parameter Estimation

- The objective of statistics is to make **inferences** about a population based on information contained in a sample.
- Populations are characterized by numerical descriptive measures called **parameters**.
- Typical population parameters are the **mean  $\mu$** , the **median  $M$** , the **standard deviation  $\sigma$** , and a **proportion  $\pi$** .
- Most inferential problems can be formulated as an inference about one or more parameters of a population.

100

## Parameter Estimation

- Methods for making inferences about parameters fall into one of two categories:
  - **estimate the value of the population parameter of interest**
  - **test a hypothesis about the value of the parameter**
- These two methods of statistical inference involve different procedures, and they answer two different questions about the parameter.
  - In estimating a population parameter, we are answering the question
    - “What is the value of the population parameter?”
  - In testing a hypothesis, we are seeking an answer to the question
    - “Does the population parameter satisfy a specified condition?”

101

## Parameter Estimation

- Estimation refers to the process of guessing the unknown value of a parameter (e.g., population mean) using the observed data.
- For this, an **estimator**, which is a **statistic**, is used.
  - A **statistic** is a function of the observed data only.
- Sometimes we only provide a single value as our estimate.
  - This is called **point estimation**.
    - Point estimates do not reflect our uncertainty when estimating a parameter.
    - We always remain uncertain regarding the true value of the parameter when we estimate it using a sample from the population.
- To address this issue, we can present our estimates in terms of a range of possible values.
  - This is called **interval estimation**.

102

## Hypothesis Testing

- A **hypothesis** (plural: **hypotheses**),
  - a testable statement about the relationship between two or more variables
  - a proposed explanation for some observed phenomenon.
- In a scientific experiment or study, the hypothesis is
  - a brief summation of the researcher's prediction of the study's findings, which may be supported or not by the outcome.
- Hypothesis testing is the core of the **scientific method**.

103

## Scientific method

- an approach to seeking knowledge that involves forming and testing a **hypothesis**.
- used to answer questions in a wide variety of disciplines outside of science, including business.
- provides a logical, systematic way to answer questions and removes subjectivity by requiring each answer to be authenticated with objective evidence that can be reproduced.
- Goal of scientific method is to gather data that will validate or invalidate a cause and effect relationship.
  - often carried out in a linear manner, but the approach can also be cyclical, because once a conclusion has been reached, it often raises more questions.

104

## Scientific method



105

## Hypothesis

- In general, many scientific investigations start by expressing a **hypothesis**.
- To evaluate **hypotheses**, we rely on
  - estimators,
  - their sampling distributions,
  - their specific values
 from observed data.
- For example,
  - Mackowiak et al.\* hypothesized that the average normal (i.e., for healthy people) body temperature is less than the widely accepted value of 98.6°F.
  - If we denote the population mean of normal body temperature as  $\mu$ , then we can express this hypothesis as  $\mu < 98.6$ .

\*Mackowiak P.A., Wasserman S.S., Levine M.M.: A critical appraisal of 98.6°F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. JAMA 268, 1578-1580 (1992)

106

## Null and Alternative hypotheses

- The **null hypothesis** usually reflects the “status quo” or “nothing of interest”.
- In contrast, we refer to our hypothesis (i.e., the hypothesis we are investigating through a scientific study) as the **alternative hypothesis** and denote it as  $H_A$ .
- The procedure for evaluating a hypothesis is called **hypothesis testing**, and it rises in many scientific problems.
- For hypothesis testing, we focus on the **null hypothesis** since it tends to be simpler.
- To this end, we examine the evidence that the observed data provide against the null hypothesis  $H_0$ .
  - If the evidence against  $H_0$  is strong, we reject  $H_0$ .
  - If not, we state that the evidence provided by the data is not strong enough to reject  $H_0$ , and we fail to reject it.

107

## Null and Alternative hypotheses

- With respect to our decision regarding the null hypothesis  $H_0$ , we might make two types of errors:
  - **Type I error**:
    - we reject  $H_0$  when it is true and should not be rejected.
  - **Type II error**:
    - we fail to reject  $H_0$  when it is false and should be rejected.
- We denote the probability of making **type I error** as  $\alpha$  and the probability of making **type II error** as  $\beta$ .

Decision Made	Actual Validity of $H_0$	
	$H_0$ is true	$H_0$ is false
	Accept $H_0$ True Negative	False Negative (Type II Error)
Reject $H_0$	False Positive (Type I Error)	True Positive

108

## Null and alternative hypotheses

- Now suppose that we have a hypothesis testing procedure that fails to reject the null hypothesis when it should be rejected with probability  $\beta$ .
  - This means that our test correctly rejects the null hypothesis with probability  $1 - \beta$ .
    - Note that the two events are complementary.
  - We refer to this probability (i.e.,  $1 - \beta$ ) as the **power** of the test.
- In practice, it is common to first agree on a tolerable type I error rate  $\alpha$ , such as 0.01, 0.05, and 0.1.
- Then try to find a test procedure with the highest power among all reasonable testing procedures.

109

## Hypothesis testing for the population mean

- To decide whether we should reject the null hypothesis, we quantify the empirical support (provided by the observed data) against the null hypothesis using some statistics.
- We use statistics to evaluate our hypotheses.
  - We refer to them as **test statistics**.
    - To evaluate hypotheses regarding the population mean, we use the sample mean  $\bar{X}$  as the test statistic
- For a statistic to be considered as a test statistic, its sampling distribution must be fully known (exactly or approximately) under the null hypothesis.
  - We refer to the distribution of test statistics under the null hypothesis as the **null distribution**.
    - For the sample mean, the CLT states that the sampling distribution is approximately normal when the sample size is large.

110

## Regression Analysis

- The modeling of the relationship between a response variable and a set of explanatory variables is one of the most widely used of all statistical techniques.
  - We refer to this type of modeling as **regression analysis**.
- A **regression model** provides the user with a functional relationship between the **response variable** and **explanatory variables** that allows the user to determine which of the explanatory variables have an effect on the response.
  - The **regression model** allows the user to explore what happens to the response variable for specified changes in the explanatory variables.

111

## Regression Analysis

- The basic idea of regression analysis is to obtain a model for the functional relationship between a **response variable** (often referred to as the **dependent variable**) and one or more **explanatory variables** (often referred to as the **independent variables**).
- **Regression models have a number of uses:**
  - The model provides a description of the major features of the data set.
    - In some cases, a subset of the explanatory variables will not affect the response variable, and, hence, the researcher will not have to measure or control any of these variables in future studies.
      - This may result in significant savings in future studies or experiments.

112

## Regression Analysis

- The equation relating the response variable to the explanatory variables produced from the regression analysis provides estimates of the response variable for values of the explanatory variables not observed in the study.
  - For example, a clinical trial is designed to study the response of a subject to various dose levels of a new drug.
  - Because of time and budgetary constraints, only a limited number of dose levels are used in the study.
    - The regression equation will provide estimates of the subjects' response for dose levels not included in the study.
- In business applications, the prediction of future sales of a product is crucial to production planning.
  - If the data provide a model that has a good fit in relating current sales to sales in previous months, prediction of sales in future months is possible.

113

## The linear relationship

- The linear relationship between  $Y$  and  $X$  in the entire population can be presented in a similar form,
$$Y = \alpha + \beta X + \varepsilon$$
- where  $\alpha$  is the intercept, and  $\beta$  is the slope of the regression line,  $\varepsilon$  is called the **error term**, representing the difference between the estimated and the actual values of  $Y$  in the population.
- We refer to the above equation as the **linear regression model**.
  - We refer to  $\alpha$  and  $\beta$  as the **regression parameters**.
  - More specifically,  $\beta$  is called the **regression coefficient** for the explanatory variable.
  - The process of finding the regression parameters is called **fitting** a regression model to the data.

114

## Supervised learning

- **Linear regression models** are used to predict the unknown values of the response variable.
  - In these models, the response variable has a central role;
    - the model building process is guided by explaining the variation of the response variable or predicting its values.
  - Therefore, building regression models is known as supervised learning.

115

## Unsupervised learning

- Building statistical models to identify the underlying structure of data is known as **unsupervised learning**.
  - An important class of unsupervised learning is **clustering**,
    - which is commonly used to identify subgroups within a population.
- In general, **cluster analysis** refers to the methods that attempt to divide the data into subgroups such that
  - the observations within the same group are more similar compared to the observations in different groups.

116

## Distance Measure

- The core concept in any cluster analysis is the notion of **similarity** and **dissimilarity**.
  - It is common to quantify the degree of dissimilarity based on a **distance measure**,
    - which is usually defined for a pair of observations.
- The most commonly used distance measure is the **squared distance**,
 
$$d_{ij} = (x_i - x_j)^2,$$
 where  $d_{ij}$  refers to the distance between observations  $i$  and  $j$ ,  $x_i$  is the value of random variable  $X$  for observation  $i$ , and  $x_j$  is the value for observation  $j$ .

117

## Similarity and Dissimilarity

- **Similarity**
  - is a numerical measure of how alike two data objects are
  - is higher when objects are more alike
  - often falls in the range [0,1]
- **Dissimilarity**
  - is a numerical measure of how two data objects are different
  - is lower when objects are more alike
    - Minimum dissimilarity is often 0
    - Upper limit varies
- Proximity refers to a similarity or dissimilarity

118

## Distance

- **Euclidean Distance**

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

where  $n$  is the number of dimensions (attributes) and  $p_k$  and  $q_k$  are, respectively, the  $k$ th attributes (components) of data objects  $p$  and  $q$ .

- **Minkowski Distance** is a generalization of Euclidean Distance

$$dist = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

where  $r$  is a parameter,  $n$  is the number of dimensions (attributes) and  $p_k$  and  $q_k$  are, respectively, the  $k$ th attributes (components) of data objects  $p$  and  $q$ .

119

## Distance

- In **Minkowski Distance**,
  - if  $r = 1$   $dist$  is City block (Manhattan, taxicab, L1 norm) distance
  - if  $r = 2$   $dist$  is Euclidean distance
  - if  $r = \infty$   $dist$  is “supremum” (Lmax norm, L $\infty$  norm) distance
- In general, if we measure  $p$  random variables  $X_1, \dots, X_p$ , the squared distance between two observations  $i$  and  $j$  in our sample is
 
$$d_{ij} = (x_{i1} - x_{j1})^2 + \dots + (x_{ip} - x_{jp})^2$$
- This measure of dissimilarity is called the **squared Euclidean distance**

120

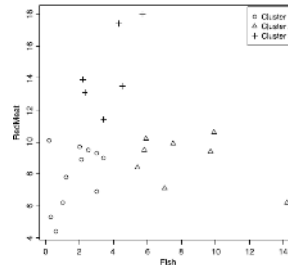
## K-means Clustering

- **K-means clustering** is a simple algorithm that uses the **squared Euclidean distance** as its measure of dissimilarity.
- After randomly partitioning the observations into  $K$  **groups** and finding the **center** or **centroid** of each cluster, the **K-means** algorithm finds the best clusters by iteratively repeating the following steps
  - For each observation, find its **squared Euclidean distance** to all  $K$  **centers**, and assign it to the cluster with the smallest distance.
  - After regrouping all the observations into  $K$  clusters, recalculate the  $K$  centers.
- These steps are applied until the clusters do not change
  - i.e., the centers remain the same after each iteration.

121

## K-means Clustering

- An example of visualizing the results of  $K$ -means clustering with a scatterplot



- The three clusters are represented by circles, triangles, and crosses.

122

## Hierarchical Clustering

- There are two potential problems with the  $K$ -means clustering algorithm.
  - It is a **flat clustering method**.
  - We need to specify the number of clusters  $K$  a priori.
- An alternative approach that avoids these issues is **hierarchical clustering**.
- The result of this method is a **dendrogram** (a tree).
  - The **root** of the dendrogram is its highest level and contains all  $n$  observations.
  - The **leaves** of the tree are its lowest level and are each a unique observation.

123

## Hierarchical Clustering

- There are two general algorithms for hierarchical clustering:
  - **Divisive (top-down)**:
    - We start at the top of the tree, where all observations are grouped in a single cluster.
    - Then we divide the cluster into two new clusters that are most dissimilar.
      - Now we have two clusters.
    - We continue splitting existing clusters until every observation is its own cluster.

124

## Hierarchical Clustering

- **Agglomerative (bottom-up)**:
  - We start at the bottom of the tree, where every observation is a cluster
    - i.e., there are  $n$  clusters.
  - Then we merge two of the clusters with the smallest degree of dissimilarity
    - i.e., the two most similar clusters.
    - Now we have  $n - 1$  clusters.
  - We continue merging clusters until we have only one cluster (the root) that includes all observations.

125

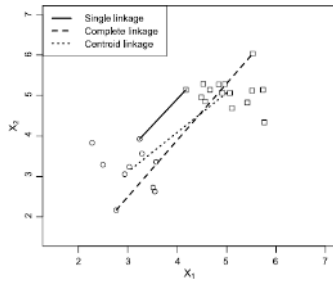
## Hierarchical Clustering

- We can use one of the following methods to calculate the overall distance between two clusters
  - **Single linkage clustering** uses the minimum  $d_{ij}$  among all possible pairs as the distance between the two clusters.
  - **Complete linkage clustering** uses the maximum  $d_{ij}$  as the distance between the two clusters.
  - **Average linkage clustering** uses the average  $d_{ij}$  over all possible pairs as the distance between the two clusters.
  - **Centroid linkage clustering** finds the centroids of the two clusters and uses the distance between the centroids as the distance between the two clusters.

126

## Hierarchical Clustering

- The following figure illustrates the difference between the single linkage method, the complete linkage method, and the centroid linkage method to determine the distance  $d_{ij}$  between the two clusters shown as circles and squares.



- Note that the dotted line connects the centers (as opposed to observations) of the two clusters.
- There are of course other ways for defining the distance between two clusters.
- However, the above measures are the most commonly used.

127