Statistic Quality & Engineering
Universitas Indonesia

01. Introduction to Statistics and Data Analysis

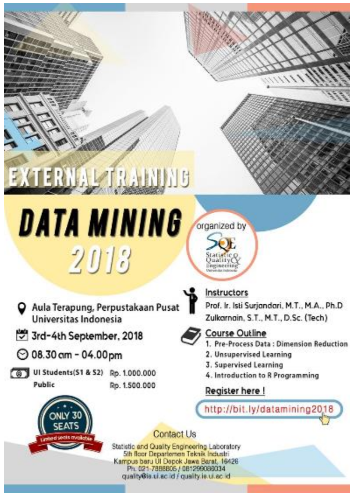
1. The Role of Statistics in Engineering
2. The Role of Probability in Statistics
3. Measures of Location and Variability: Data Summary Using Descriptive Measures

Part 1

The Role of Statistics in Engineering

Adapted From :
Probability & Statistics for Engineers & Scientists, 9th Ed.
Walpole/Myers/Myers/Ye (c)2012
Applied Statistics and Probablity for Engineers, 6rd Ed.
Montgomery/Runger © *2014*
Probability & Statistics for Engineering & The Science, 9th Ed.
J.L.Devore © *2014*

Laboratorium Rekayasa Kualitas dan Statistika



EXTERNAL TRAINING

DATA MINING 2018

organized by
Statistic Quality & Engineering Universitas Indonesia

Aula Terapung, Perpustakaan Pusat Universitas Indonesia

3rd-4th September, 2018

08.30 am - 04.00pm

UI Students(S1 & S2)	Rp. 1.000.000
Public	Rp. 1.500.000

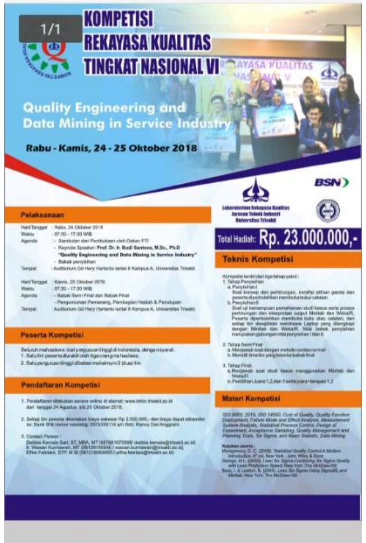
ONLY 30 SEATS
Limited seats available

Instructors
Prof. Ir. Isti Surjandari, M.T., M.A., Ph.D
Zulkarnain, S.T., M.T., D.Sc. (Tech)

Course Outline
1. Pre-Process Data : Dimension Reduction
2. Unsupervised Learning
3. Supervised Learning
4. Introduction to R Programming

Register here !
http://bit.ly/datamining2018

Contact Us
Statistic and Quality Engineering Laboratory
5th floor Departemen Teknik Industri
Kampus baru UI Depok Jawa Barat, 16426
Ph. 021-7888805 / 081299088034
quality@ie.ui.ac.id / quality.ie.ui.ac.id



1/1

KOMPETISI REKAYASA KUALITAS TINGKAT NASIONAL VI

Quality Engineering and Data Mining in Service Industry

Rabu - Kamis, 24 - 25 Oktober 2018

BSN

Laboratorium Rekayasa Kualitas
Jurusan Teknik Industri
Universitas Trisakti

Total Hadiah: Rp. 23.000.000,-

Pelaksanaan

Hari/Tanggal : Rabu, 24 Oktober 2018
Waktu : 07.30 - 17.30 WIB
Agenda : - Sambutan dan Pembukaan oleh Dekan FTI
- Keynote Speaker: Prof. Dr. Ir. Budi Santosa, M.Sc., Ph.D "Quality Engineering and Data Mining in Service Industry"
- Babak penyisihan
Tempat : Auditorium Gd Hery Hartanto lantai 8 Kampus A, Universitas Trisakti

Hari/Tanggal : Kamis, 25 Oktober 2018
Waktu : 07.30 - 17.30 WIB
Agenda : - Babak Semi Final dan Babak Final
- Pengumuman Pemenang, Pembagian Hadiah & Penutupan
Tempat : Auditorium Gd Hery Hartanto lantai 8 Kampus A, Universitas Trisakti

Peserta Kompetisi

Seluruh mahasiswa dari perguruan tinggi di Indonesia, dengan syarat:
1. Satu tim peserta diwakili oleh tiga orang mahasiswa.
2. Satu perguruan tinggi dibatasi maksimum 2 (dua) tim

Pendaftaran Kompetisi

1. Pendaftaran dilakukan secara online di alamat: www.tekin.trisakti.ac.id dari tanggal 24 Agustus s/d 20 Oktober 2018.
2. Setiap tim peserta dikenakan biaya sebesar Rp 2.000.000,- dan biaya dapat ditransfer ke: Bank BNI nomor rekening: 0574398114 a/n Sdri. Ranny Dwi Anggraini
3. Contact Person :
Debbie Kemala Sari, ST, MBA, MT (08788 1070999/ debbie.kemala@trisakti.ac.id)
Ir. Wawan Kurniawan, MT (08129155946 / wawan.kurniawan@trisakti.ac.id)
Elfira Febriani, STP, M.Si (081218964950 / elfira.febriani@trisakti.ac.id)

Teknis Kompetisi

Kompetisi terdiri dari tiga tahap yakni :
1. Tahap Penyisihan
   a. Penyisihan I
   Soal konsep dan perhitungan, bersifat pilihan ganda dan peserta diperbolehkan membuka buku/ catatan.
   b. Penyisihan II
   Soal uji kemampuan pemahaman studi kasus serta proses perhitungan dan interpretasi output Minitab dan Weka/R. Peserta diperbolehkan membuka buku atau catatan, dan setiap tim diwajibkan membawa Laptop yang dilengkapi dengan Minitab dan Weka/R. Nilai babak penyisihan merupakan gabungan nilai penyisihan I dan II.
2. Tahap Semi Final
   a. Menjawab soal dengan metode cerdas cermat
   b. Memilih finalis tim yang lolos ke babak final
3. Tahap Final
   a. Menjawab soal studi kasus menggunakan Minitab dan Weka/R
   b. Pemilihan Juara 1,2,dan 3 serta juara Harapan 1,2

Materi Kompetisi

ISO 9001: 2015, ISO 14000, Cost of Quality, Quality Function Deployment, Failure Mode and Effect Analysis, Measurement System Analysis, Statistical Process Control, Design of Experiment, Acceptance Sampling, Quality Management and Planning Tools, Six Sigma, and Basic Statistic, Data Mining

Rujukan utama :
Montgomery, D. C. (2009). Statistical Quality Control A Modern Introduction, 6th ed. New York : John Wiley & Sons
George, M.L. (2002). Lean Six Sigma-Combining Six Sigma Quality with Lean Production Speed. New York: The McGraw-Hill
Bass, I. & Lawton, B. (2009). Lean Six Sigma Using SigmaXL and Minitab. New York: The McGraw-Hill

Learning Objectives

- The Role of Statistics in Engineering
- Define Statistics
- Define Variability
- Distinguish descriptive & inferential statistics
- The Role of Probability
- Summarize the sources of data
- Describe the types of data & scales
- Explain the types of samples
- Describe survey process & errors

Statistics Intro

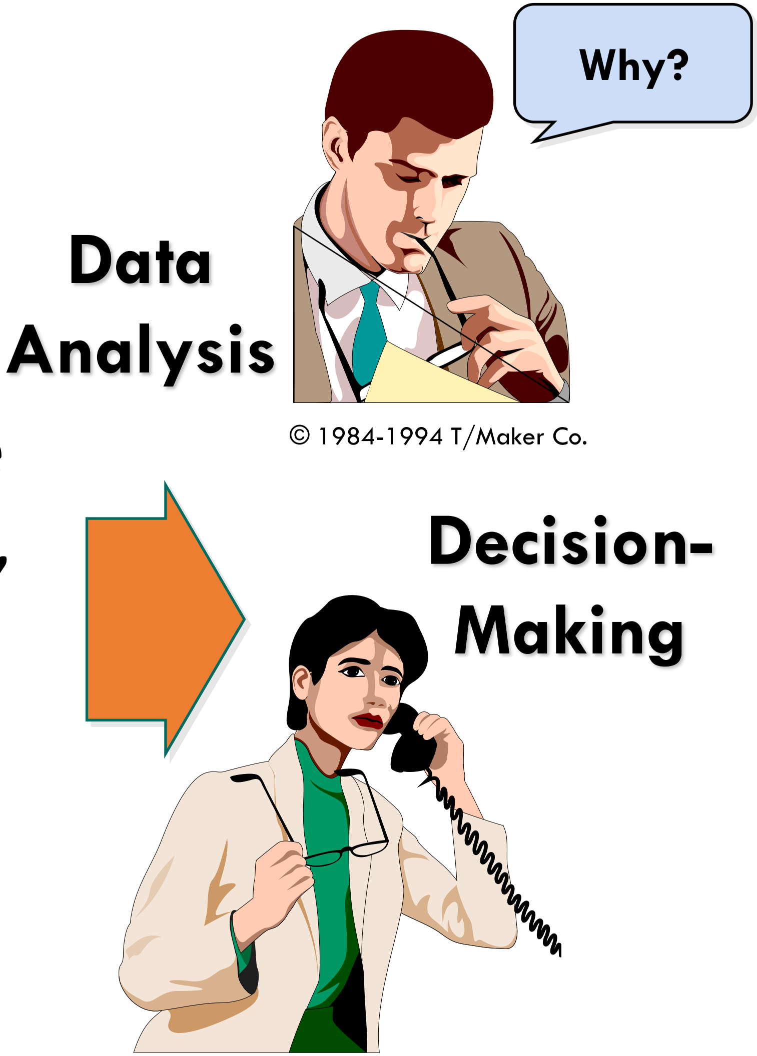
- **“There are three types of lies, lies, damn lies, and statistics”, a quote attributed to Mark Twain among others.**
    - In academia, we hopefully pursue the truth when analyzing data. Companies, unfortunately, don’t always have this goal.
    - We can use statistics to verify the validity of published results and find the truth, if raw data can be obtained.
- **Statistics, used properly, is a tool for analyzing data and discovering and proving it’s true meaning.**

What is Statistics ?

Statistics is a science that helps us make decisions and draw conclusions in the presence of variability

The field of **statistics** deals with the collection, presentation, analysis, and **use of data** to make decisions, solve problems, and design products and processes

Statistics is the science of data



Why?

Data Analysis

© 1984-1994 T/Maker Co.

Decision-Making

Why We Need Data

- To provide input to study a situation
- To measure performance of service or production process
- To evaluate conformance to standards
- To assist in formulating alternative courses of action
- To satisfy curiosity

The Engineering Method

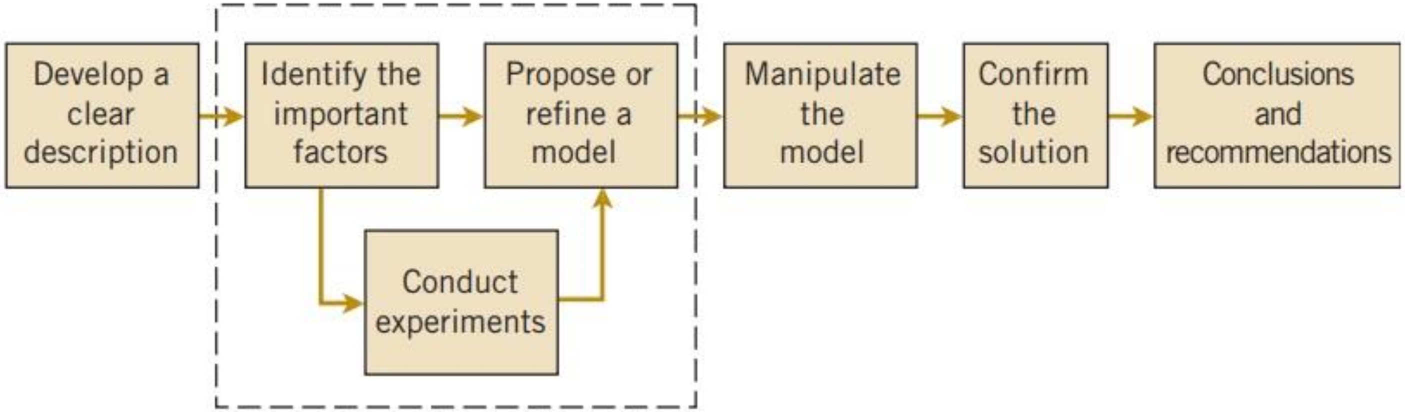

Develop a clear description
Identify the important factors
Propose or refine a model
Manipulate the model
Confirm the solution
Conclusions and recommendations
Conduct experiments

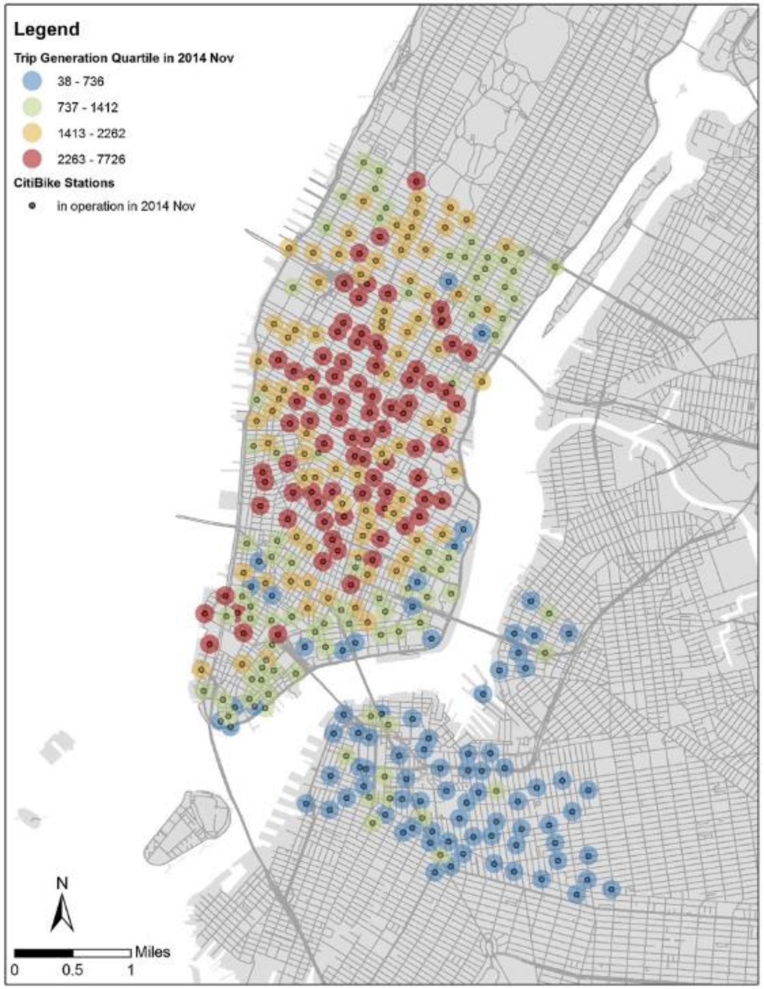
FIGURE 1-1 The engineering method.

Because many aspects of engineering practice involve working with data, obviously knowledge of **Statistics is just as important to an engineer as are the other engineering sciences.**

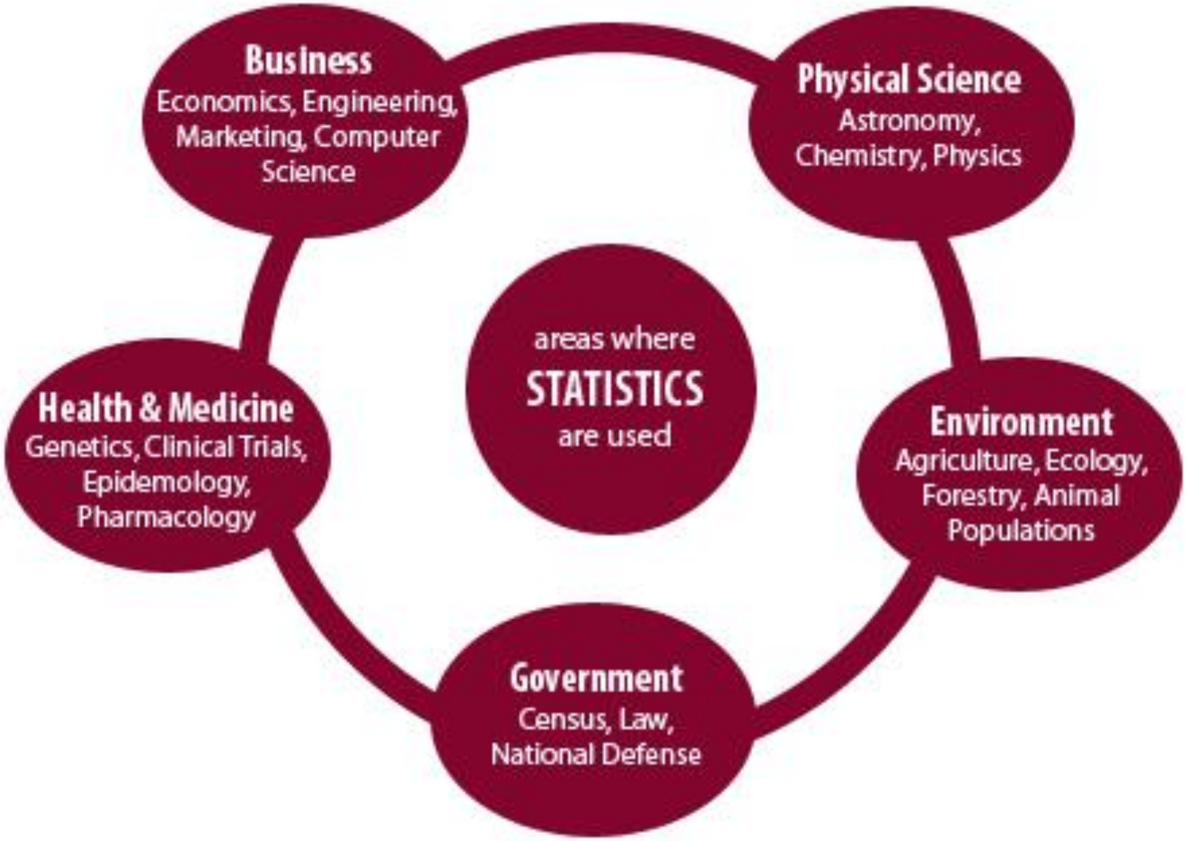
Specifically, **statistical techniques can be powerful aids in designing new products and systems, improving existing designs, and designing, developing, and improving production processes.**

Cases in Engineering

- Quality Engineering
- Ergonomics
- Transportation Engineering
  - Trip Generation Model
- Healthcare Service
- Queuing System
  - Toll both



Areas where Statistics is used



areas where **STATISTICS** are used

Business
Economics, Engineering, Marketing, Computer Science

Physical Science
Astronomy, Chemistry, Physics

Environment
Agriculture, Ecology, Forestry, Animal Populations

Government
Census, Law, National Defense

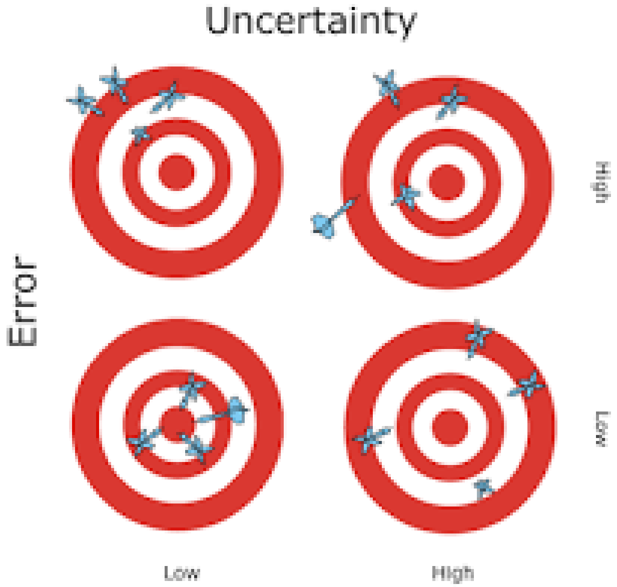
Health & Medicine
Genetics, Clinical Trials, Epidemology, Pharmacology

Areas where an Industrial Engineers will contribute using Statistics

- Quality Improvement
- Ergonomics
- Logistics
- Production system
- Maintenance system
- Human Capital
- Marketing
- Finance
- Etc.

Use of Statistics Methods

Statistical Methods are designed to contribute to the process of making scientific judgments in the face of **uncertainty** and **variation.**



Variability

- Suppose I’m trying to measure the performance of two alternative hardware/software combinations. Setup A takes 9.8 seconds to execute a given user task, and Setup B takes 10.8 seconds.
  - What can we infer from this experiment?
  - The answer is, without knowing the inherent variation in the system being measured and the measurement procedure, not much.
- Every experiment and data set has some inherent variation that must be understood before inferences about data can be made.
  - No two measurements are ever exactly the same, due to both process and measurement variability.
  - We must always gather a sample of several data points in order to make valid inferences.

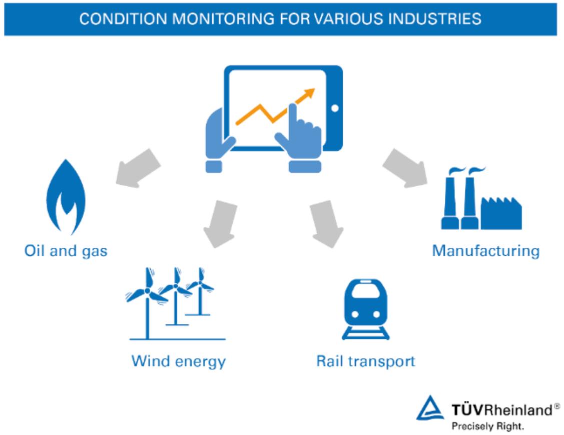
Variability (2)

- By variability, we mean that successive observations of a system or phenomenon do not produce exactly the same result.
- We all encounter variability in our everyday lives, and **statistical thinking** can give us a useful way to incorporate this variability into our decision-making processes.
- For example, consider the gasoline mileage performance of your car. Do you always get exactly the same mileage performance on every tank of fuel?
- Think about the **sources of variability** in the system.
- **Statistics provides a framework for describing this variability and for learning about which potential sources of variability are the most important or which have the greatest impact on the gasoline mileage performance.**

Methods in Data Collection

Three basic methods of collecting data are

- A **retrospective study** using historical data
  - A retrospective study would use either all or a sample of the historical process data archived over some period of time.
- An **observational study**
  - In an observational study, the engineer observes the process or population, disturbing it as little as possible, and records the quantities of interest.
- A **designed experiment**
  - The engineer makes *deliberate* or *purposeful changes* in the controllable variables of the system or process, observes the resulting system output data, and then makes an inference or decision about which variables are responsible for the observed changes in output performance.



Part 2

The Role of Probability in Statistics

Adapted From :

Probability & Statistics for Engineers & Scientists, 9th Ed.

Walpole/Myers/Myers/Ye (c)2012

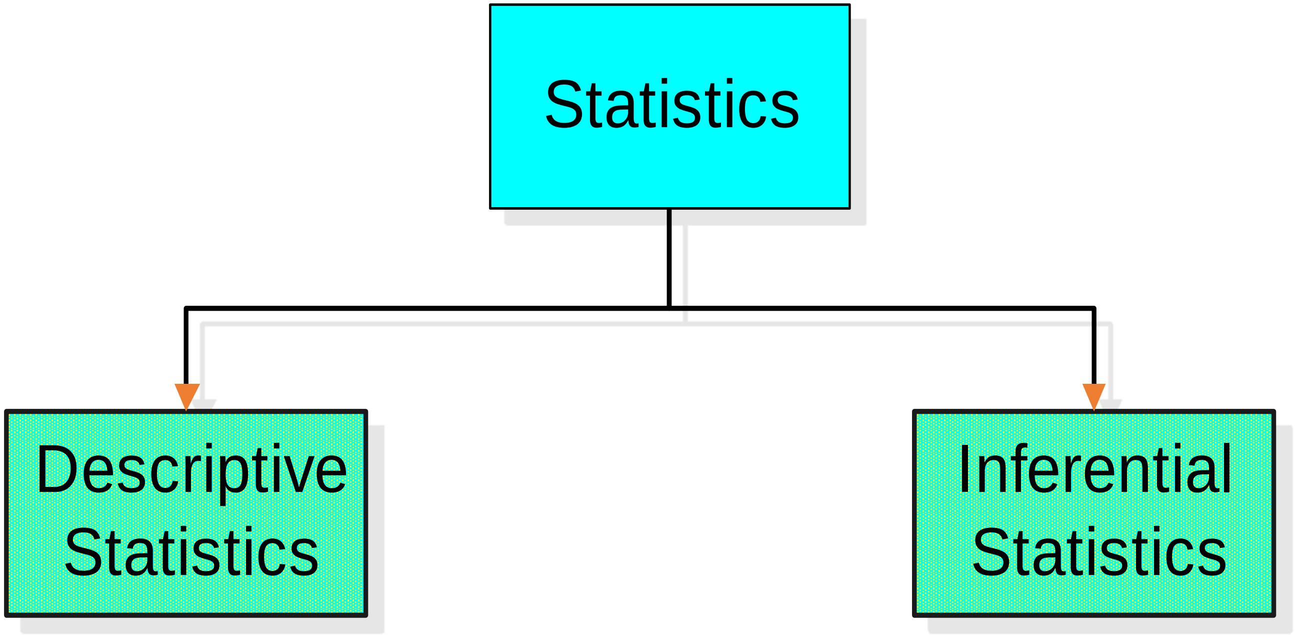
Applied Statistics and Probablity for Engineers, 6rd Ed.

Montgomery/Runger © 2014

Probability & Statistics for Engineering & The Science, 9th Ed.

J.L.Devore © 2014

Areas of Statistics



Statistics

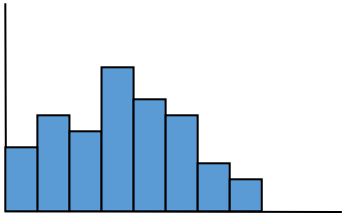
Descriptive Statistics

Inferential Statistics

Descriptive Statistics

- Collect data
  - e.g. Survey

- Present data
  - e.g. Tables and graphs



- Characterize data
  - e.g. Sample mean = $\dfrac{\sum X_i}{n}$

Describe Data

Inferential Statistics

- Estimation
  - e.g.: Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g.: Test the claim that the population mean weight is over 120 pounds

Drawing conclusions and/or making decisions concerning a population based on sample results.

Probability

- Can anybody tell me what probability is?
  - When we know the underlying model that governs an experiment, we use probability to figure out the chance that different outcomes will occur.
  - For example, if we flip a fair coin 3 times, what is the probability of obtaining 3 heads?
- By definition, probability values are between 0 and 1.
- What does it mean if Outcome A of an experiment has a probability of 1/3rd of occurring?
  - If the experiment is repeated a large number of times, Outcome A will occur 1/3rd of the time.

Probability and Statistics

- In Probability, we use our knowledge of the underlying model to determine the probability that different outcomes will occur.
- How does statistics compare to probability?
  - In statistics, we don't know the underlying model governing an experiment.
  - All we get to see is a sample of some outcomes of the experiment.
  - We use that sample to try to make inferences about the underlying model governing the experiment.
- So a thorough understanding of probability is essential to understanding statistics.

Probability and Statistics

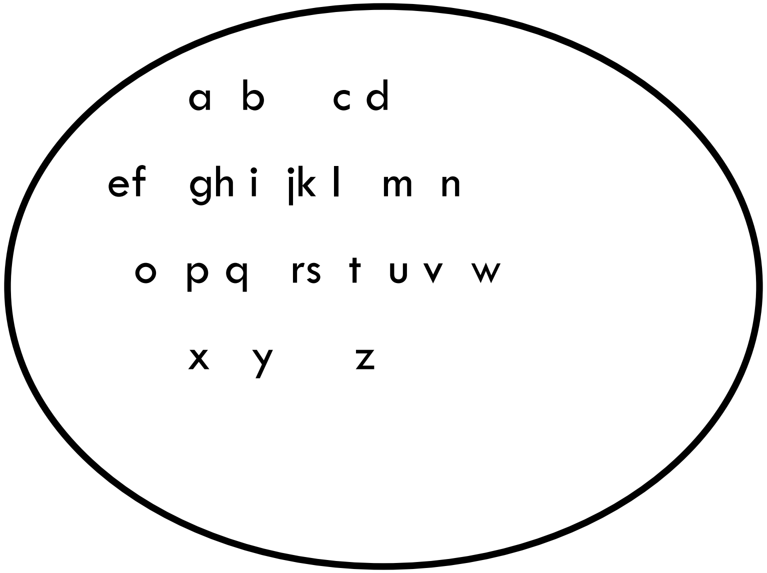
- Example: Suppose for a manufacturing process we have an upper limit of 5% defective items produced for the process to be “in control”.
- We take a sample of 100 items produced and find 10 defective items. Is the manufacturing process in control?
    - One way to do look at this is to say “if the process has 5% defective items, what is the probability that there will be 10 or more defective items in a sample of size 100?
    - The probability of this outcome is called a **<u>P-value</u>**.
    - In this case the P-value is only .0282. What does this mean?
    - That this outcome would occur by chance only 2.82% of the time. So what is the definition of a P-value?
    - The **<u>P-value</u>** is the probability of getting the measured outcome if the assumed underlying model were true.

Samples and Populations

- In statistics, a population is the set of all possible outcomes of an experiment (may be infinite).
    - In probability, we will call the set of all possible outcomes the sample space.
- A sample is a set of observations taken from a population.
- A random sample is selected so that every element in the population has an equal chance of being selected.
- Often in statistics, we compare samples from two different populations and try to determine statistically if the populations are significantly different.

Population vs. Sample

Population



a b c d

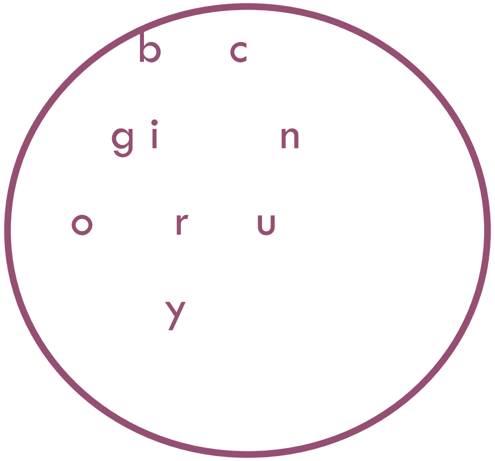
ef gh i jk l m n

o p q rs t u v w

x y z

Measures used to describe the population are called **parameters**

Sample



b c

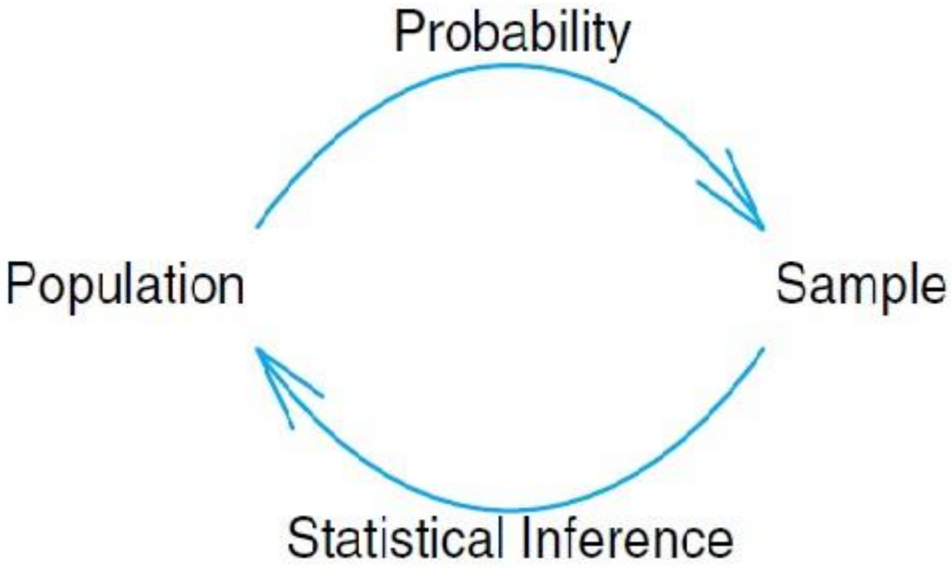
g i n

o r u

y

Measures computed from sample data are called **statistics**

How do Probability and Statistical Inference Work Together ?



Nothing can be learned about a population from a sample until the analyst learns **the rudiments of uncertainty** in that sample.

The Role Probability in Statistics

- **Elements of probability** allow us to **quantify the strength** or **“confidence”** in our conclusions.
- In this sense, **concepts in probability form a major component that supplements statistical methods and helps us gauge the strength of the statistical inference.**
- The discipline of probability, then, provides the transition between **descriptive statistics and inferential methods**.
- The notion of **a *P*-value** provides the “bottom line” in the interpretation of results from the use of statistical methods.

Mechanistic and Empirical Model

Example of mechanistic model:
Current =Voltage/Resistance
or
$I = E / R$

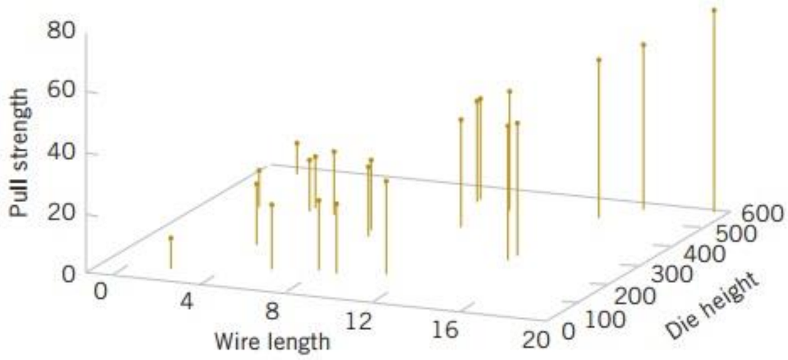


FIGURE 1-15 Three-dimensional plot of the wire bond pull strength data.

Example of empirical model: regression model

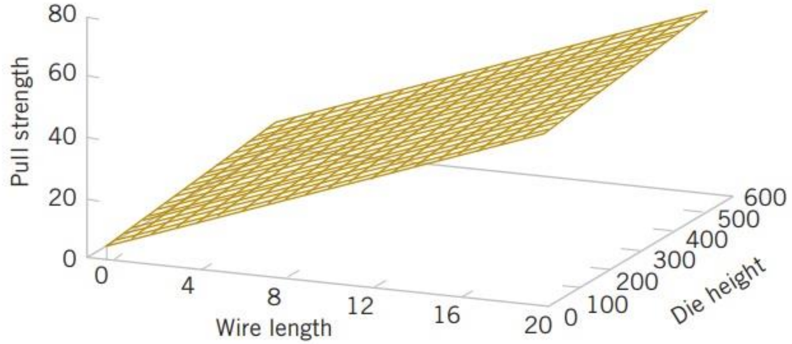


FIGURE 1-16 Plot of predicted values of pull strength from the empirical model.

Probabilistic Model

- To make good decisions, **an analysis of how well a sample represents a population is clearly necessary.**
- If the lot contains defective wafers, **how well will the sample detect these defective items? How can we quantify the criterion to "detect well?"**
- Basically, **how can we quantify the risks of decisions based on samples?** Furthermore, **how should samples be selected to provide good decisions—ones with acceptable risks?**
- **Probability models** help quantify the risks involved in statistical inference, that is, the risks involved in decisions made every day
- **A probability model** is used to calculate this proportion under reasonable assumptions for the manner in which the sample is selected.

Simple Random Sampling

- Any particular sample of a specified *sample size* has the same chance of being selected as any other sample of the same size.
- **sample size** means the number of elements in the sample.
- A table of random numbers can be utilized in sample selection in many instances.

Example of **Biased Sample**

- A sample is to be chosen to answer certain questions regarding political preferences in a certain state in the United States.
- The sample involves the choice of, say, 1000 families, and a survey is to be conducted. Now, suppose it turns out that random sampling is not used. Rather, all or nearly all of the 1000 families chosen live in an urban setting. **It is believed that political preferences in rural areas differ from those in urban areas.**
- The inferences need to be made about the state as a whole, the sample of size 1000 described here is often referred to as a **biased sample**.

Stratification in Sampling

- **Simple random sampling** is not always appropriate. Which alternative approach is used **depends on the complexity of the problem**. For example, the sampling units are not homogeneous and naturally divide themselves into non overlapping groups that are homogeneous.
- These groups are called *strata*, and a procedure called ***stratified random sampling*** involves random selection of a sample *within* each stratum.
- The purpose is to be sure that each of the strata is neither over- nor underrepresented.
  - For example, suppose a sample survey is conducted in order to gather preliminary opinions regarding a bond referendum that is being considered in a certain city. **The city is subdivided into several ethnic groups which represent natural strata**. In order not to disregard or over represent any group, separate random samples of families could be chosen from each group.

Part 3

Measures of Location and Variability: Data Summary Using Descriptive Measures

Adapted From :

Probability & Statistics for Engineers & Scientists, 9th Ed.

Walpole/Myers/Myers/Ye (c)2012

Applied Statistics and Probablity for Engineers, 6rd Ed.

Montgomery/Runger © 2014

Probability & Statistics for Engineering & The Science, 9th Ed.

J.L.Devore © 2014

Statistics for Managers, 5th Ed

Levine/Stephan/Krehbiel/Berenson © 2008

Data Sources

Primary

Data Collection

Observation

Survey

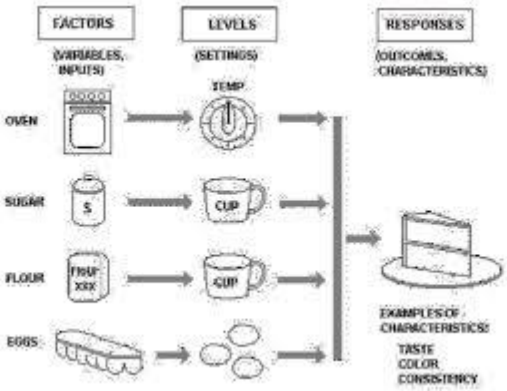
Experiment

Secondary

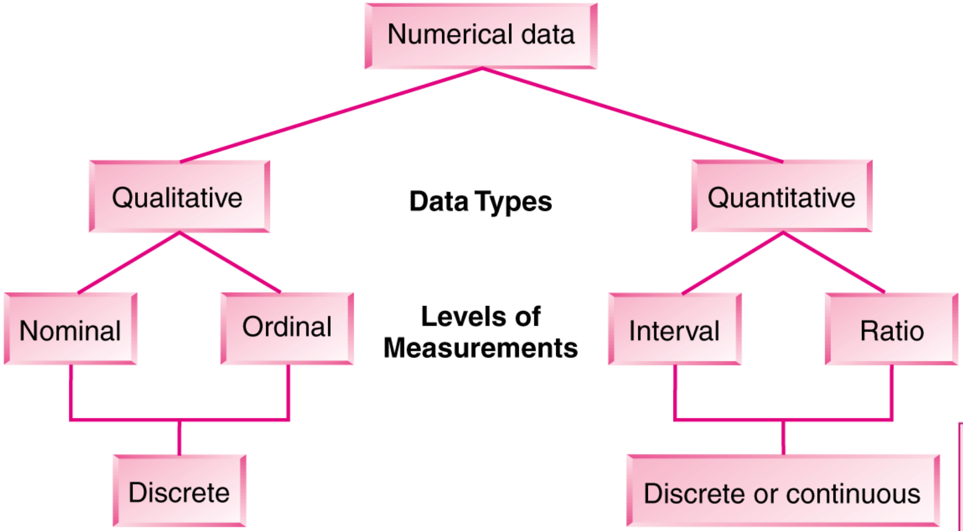
Data Compilation

Retrospective Study

Link: https://www.kdnuggets.com/2011/02/free-public-datasets.html
https://archive.ics.uci.edu/ml/index.php



Types of Data



Numerical data

Data Types

Qualitative

Quantitative

Levels of Measurements

Nominal

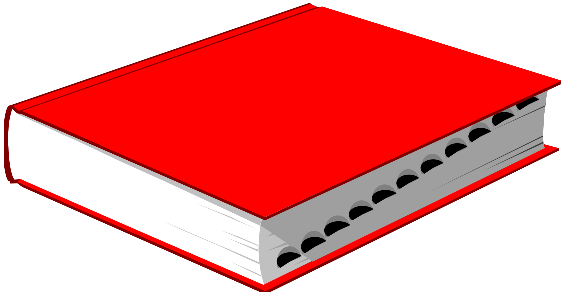
Ordinal

Interval

Ratio

Discrete

Discrete or continuous



EXAMPLES OF DISCRETE DATA

1. Nominal: Ownership status of resident dweller (1 = own, 2 = rent)
2. Ordinal: Level of customer satisfaction (1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = somewhat satisfied, 4 = very satisfied)
3. Interval: Person's score on IQ test
4. Ratio: Number of defective lightbulbs in a carton

EXAMPLES OF CONTINUOUS DATA

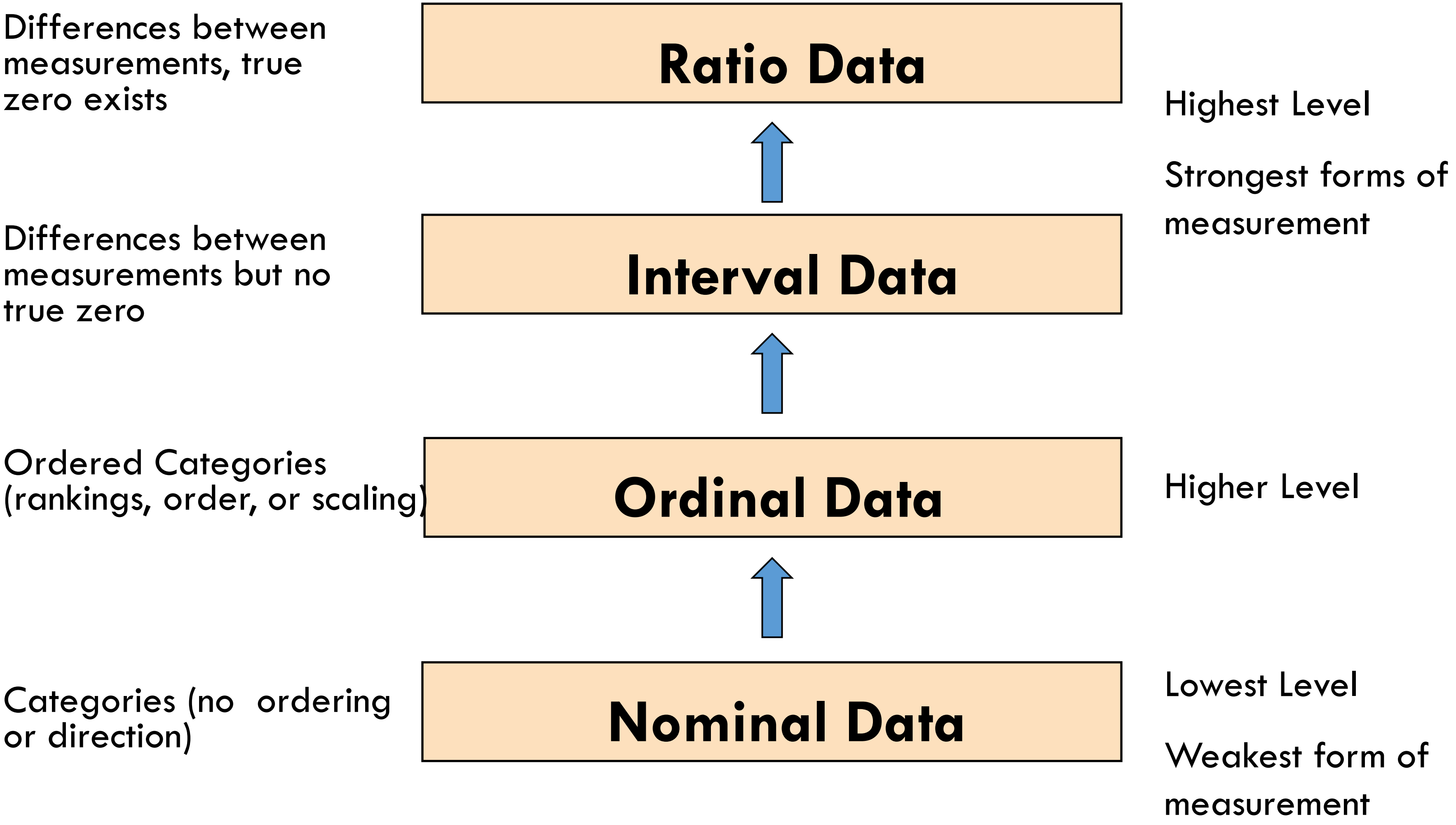
1. Interval: Actual temperature, °F
2. Ratio: Weight of packaged dog food

How are data measured ?

- Nominal scale
  - Categories
    - e.g., Male-female
  - Count
- Ordinal scale
  - Categories
  - Ordering implied
    - e.g., High-low
  - Count

- Interval scale
  - Equal intervals
  - No true 0
    - e.g., Degrees Celsius
  - Measurement
- Ratio scale
  - Equal intervals
  - True 0
  - Meaningful ratios
    - e.g., Height in inches
  - Measurement

Levels of Measurement and Measurement Scales



Differences between measurements, true zero exists

Ratio Data

Highest Level

Strongest forms of measurement

Differences between measurements but no true zero

Interval Data

Ordered Categories (rankings, order, or scaling)

Ordinal Data

Higher Level

Categories (no ordering or direction)

Nominal Data

Lowest Level

Weakest form of measurement

Example of Data

Subject	Name	Height	Income	Gender	Eye color
1	Mary	62	10,350	Female	Blue
2	John	72	30,500	Male	Brown
3	Jill	64	35,600	Female	Green
4	Donna	59	20,700	Female	Brown
5	Sam	73	15,300	Male	Blue
6	Bill	70	52,800	Male	Black
7	Mario	71	19,400	Male	Blue
8	Carol	73	12,500	Female	Brown
9	Betty	70	30,200	Female	Brown
10	Linda	68	22,700	Female	Brown

Thinking Challenge

Are they categorical? Numerical (discrete or continuous)? What scale?

- Gender
  - Male, female
- Weight
  - 123, 140.2 etc.
- Auto speed
  - 78, 64, 45 etc.

- Temperature
  - 78, 64, 85 etc.
- # Siblings
  - 0-2, 3-5, 6+
- Letter grade
  - A, B, C etc.

Measures of Location

- Suppose, then, that our data set is of the form x_1, x_2,. . ., x_n, where each x_i is a number. What features of such a set of numbers are of most interest and deserve emphasis? One important characteristic of a set of numbers is its location, and in particular its center.
- This section presents methods for describing the location of a data set.

The Mean

- The sample mean is the most important single statistic measuring the location of a sample.
  - What is the common term for the sample mean?
  - The numerical average of the sample observations. How is this calculated?
  - The sum of the observations divided by the sample size n.
  - The sample mean is an estimate of the population mean.
  - For a set of n observations, x_1, x_2, …, x_n , the sample mean is calculated as follows:

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

The Median

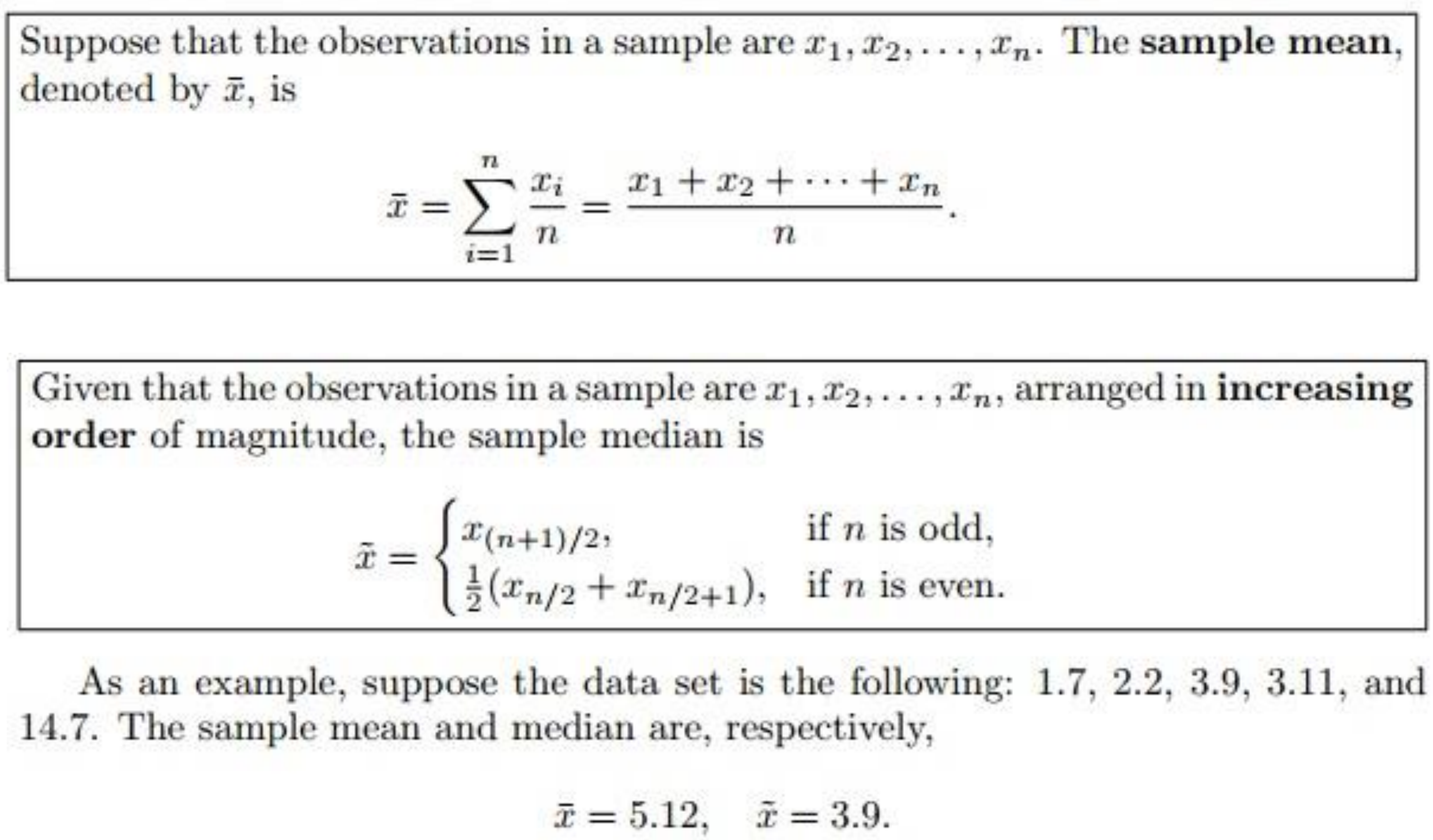
- The word *median* is synonymous with “middle,” and the sample median is indeed the middle value once the observations are ordered from smallest to largest.

- When the observations are denoted by $x_1,\ldots, x_n$, we will use the symbol $\tilde{x}$ to represent the sample median.

The Median

- The sample median is another measure of location.
  - What is the median of a sample?
  - The observation separating the upper and lower halves of the sample.
  - The middle observation if n is odd. What if n is even?
  - The average of the two middle observations.
- As an example, the median is always used for housing prices. Why use the median rather than the mean?
  - Outliers (multi-million dollar houses) skew the results.

Measures of Location



Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

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

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

As an example, suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

Data of An Experiment

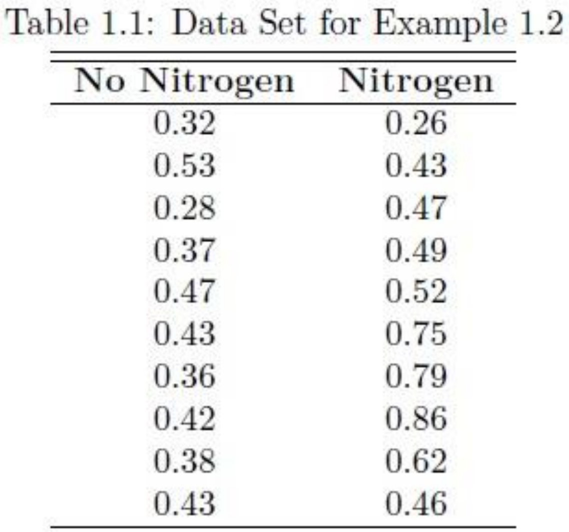


Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

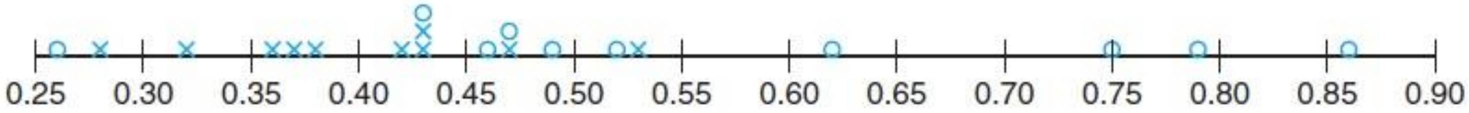
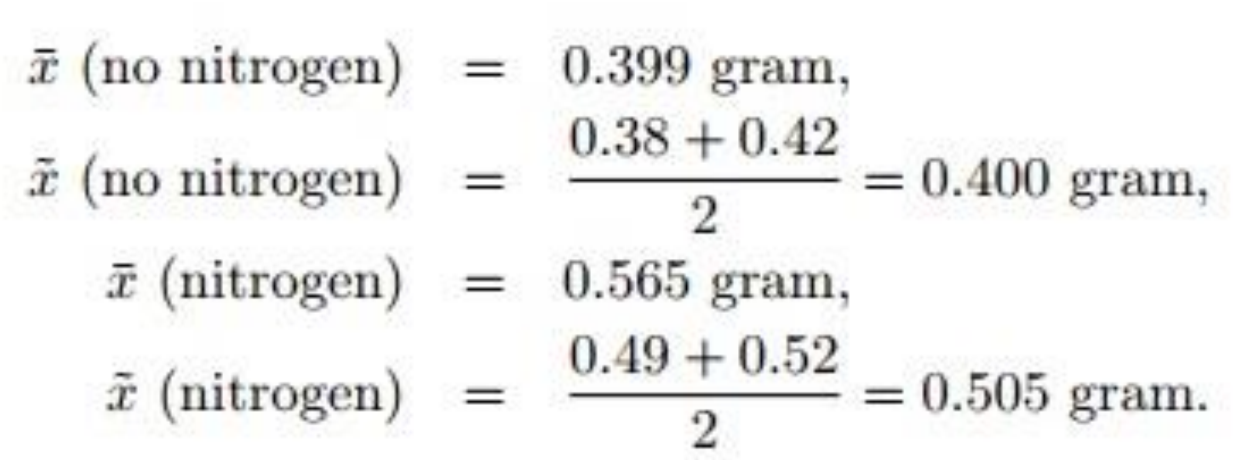


Figure 1.1: A dot plot of stem weight data.

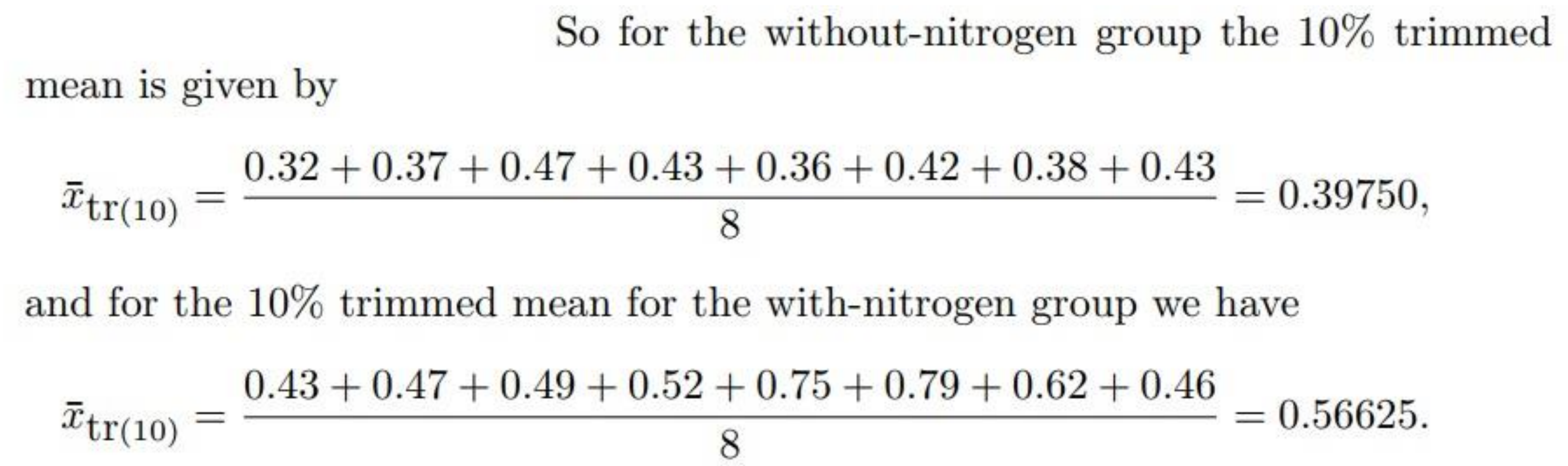
Mean & Median



$$
\begin{aligned}
\bar{x}\ (\text{no nitrogen}) &= 0.399 \text{ gram},\\
\tilde{x}\ (\text{no nitrogen}) &= \frac{0.38+0.42}{2} = 0.400 \text{ gram},\\
\bar{x}\ (\text{nitrogen}) &= 0.565 \text{ gram},\\
\tilde{x}\ (\text{nitrogen}) &= \frac{0.49+0.52}{2} = 0.505 \text{ gram}.
\end{aligned}
$$

Trimmed Means

- The trimmed mean is calculated by eliminating the highest and lowest values in the sample and taking the mean of the remaining values.
  - A trimmed mean is computed by “trimming away” a certain percent of both the largest and the smallest set of values.
  - For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.



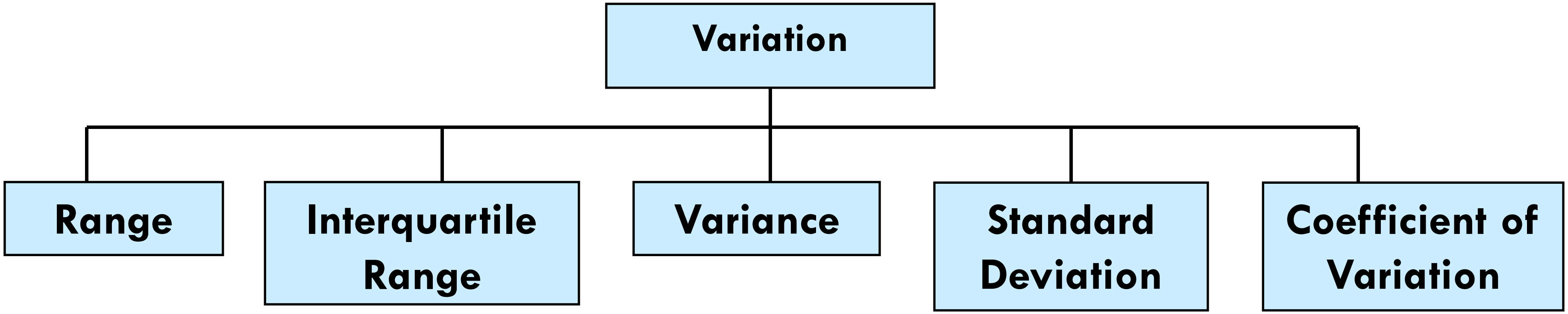
So for the without-nitrogen group the 10% trimmed mean is given by

$$\bar{x}_{\text{tr}(10)} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

and for the 10% trimmed mean for the with-nitrogen group we have

$$\bar{x}_{\text{tr}(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Measures of Variation



Variation

Range

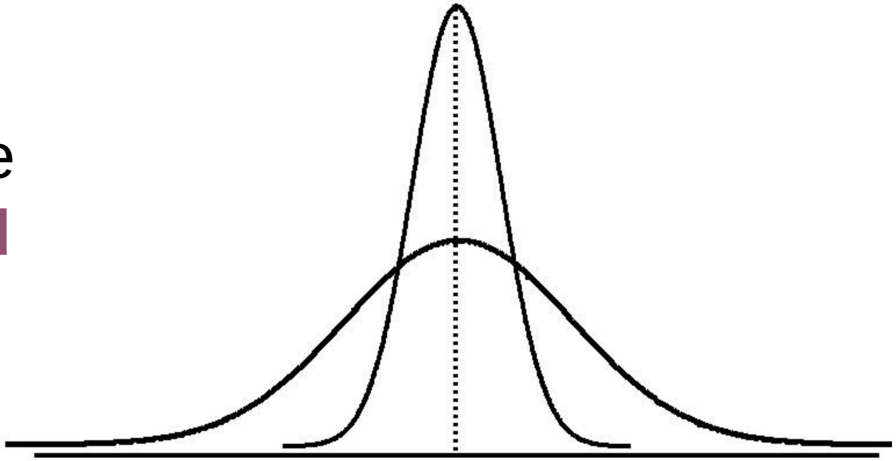
Interquartile Range

Variance

Standard Deviation

Coefficient of Variation

- Measures of variation give information on the **spread** or **variability** of the data values.



Same center, different variation

Variability Measures: Variance

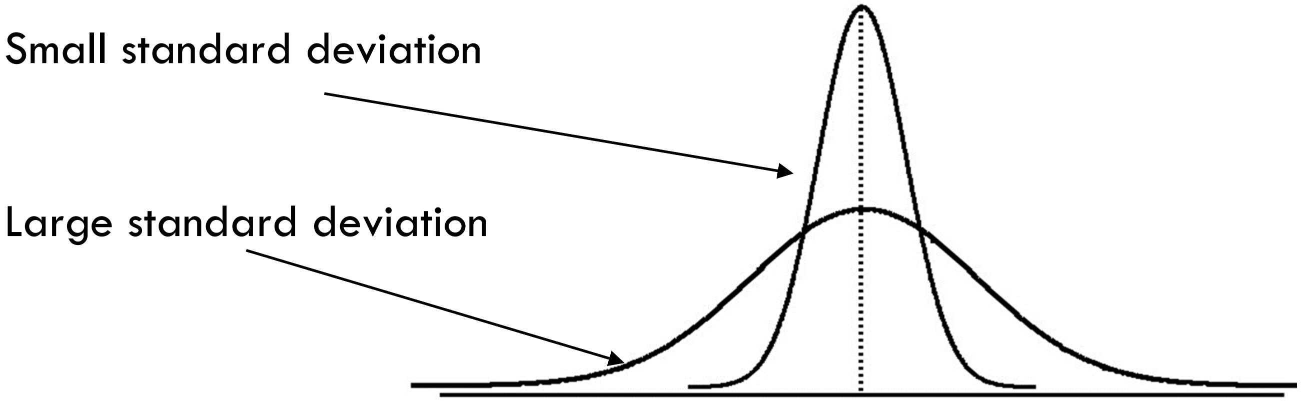
- Sample variability is critical to statistical calculations.
- Sample variance and standard deviation are the most important measures of variability.
  - Does anyone know how to calculate variance?
  - For a set of n observations, x_1, x_2, …, x_n , the sample variance, s^2, is calculated as follows:

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

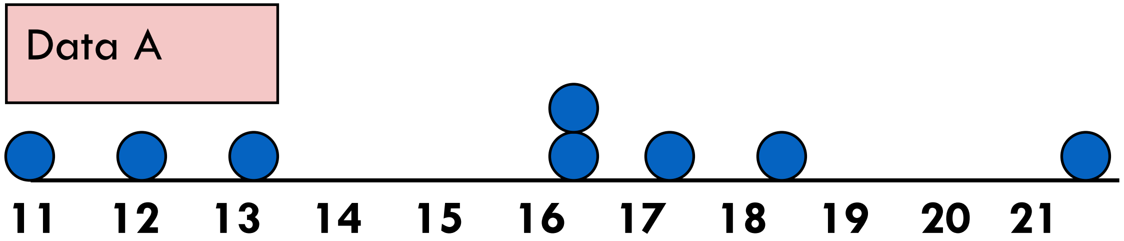
  - n –1 is called the degrees of freedom associated with the variance. This is the number independent squared deviations, or pieces of information that make up s^2.

Measures of Sample Variability

- The standard deviation, s, is the square root of the variance.
  - What are the units of the standard deviation?
  - What does it mean if the variance (and thus the standard deviation) are large?

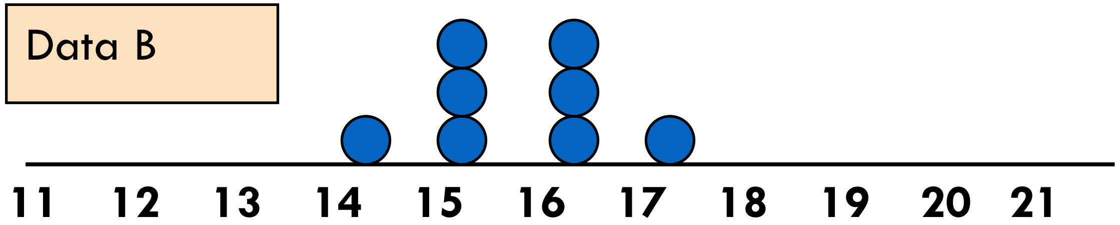


Comparing Standard Deviations



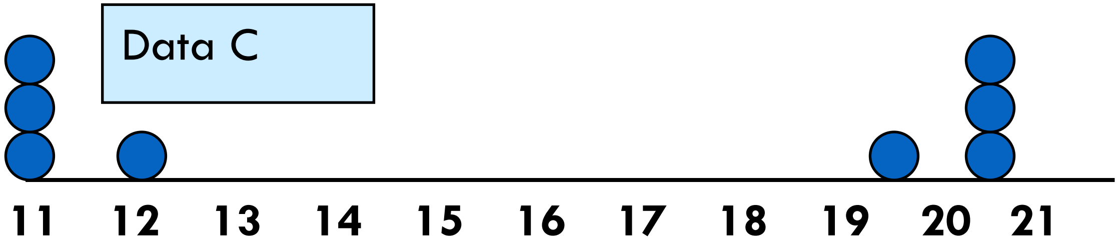
Mean = 15.5

S = 3.338



Mean = 15.5

S = .9258



Mean = 15.5

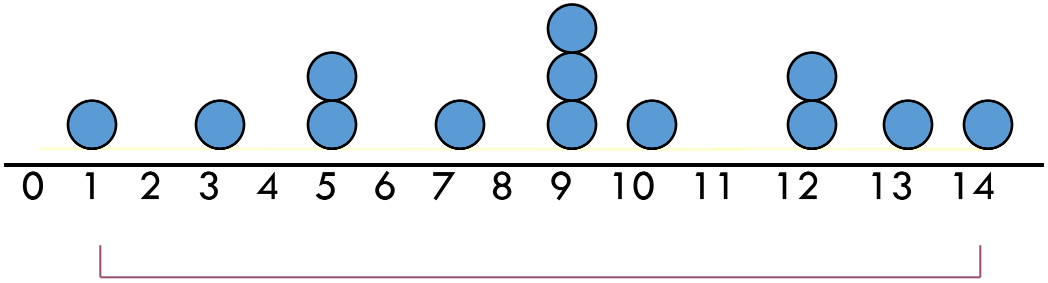
S = 4.57

Range

- The simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

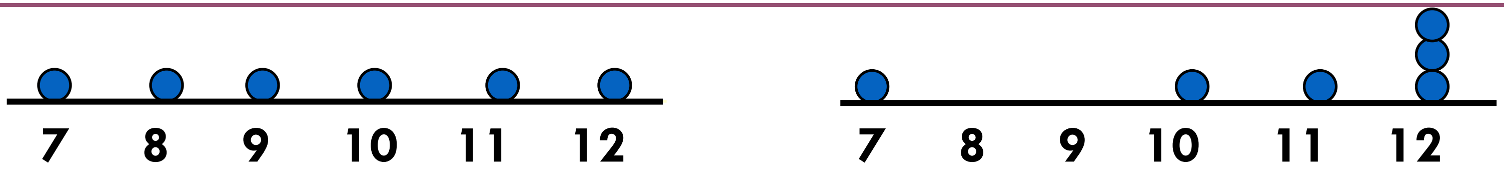
Example:



Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



Range = 12 - 7 = 5

Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

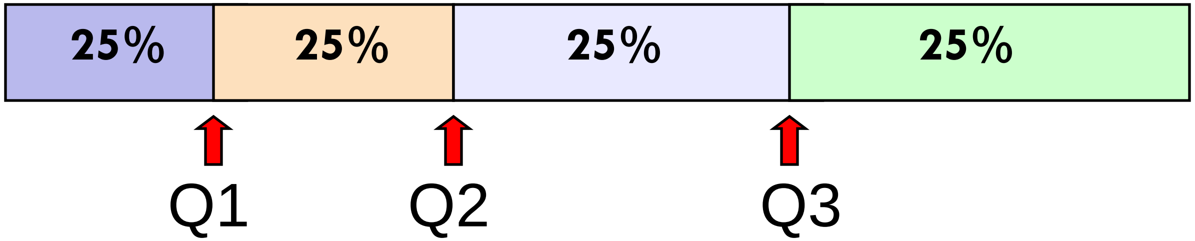
Quartiles

- The median (population or sample) divides the data set into two parts of equal size. To obtain finer measures of location, we could divide the data into more than two such parts.

- Roughly speaking, quartiles divide the data set into four equal parts, with the observations above the third quartile constituting the upper quarter of the data set, the second quartile being identical to the median, and the first quartile separating the lower quarter from the upper three-quarters.

Quartiles

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartile Formulas

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $Q_1 = (n+1)/4$

Second quartile position: $Q_2 = (n+1)/2$ (the median position)

Third quartile position: $Q_3 = 3(n+1)/4$

where n is the number of observed values

Quartiles

- Example: Find the first quartile

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 = is in the (9+1)/4 = 2.5 position of the ranked data

so use the value half way between the 2nd and 3rd values,

so **$Q_1 = 12.5$**

Q_1 and Q_3 are measures of noncentral location
Q_2 = median, a measure of central tendency

Interquartile Range

- You can eliminate some outlier problems by using the **interquartile range**

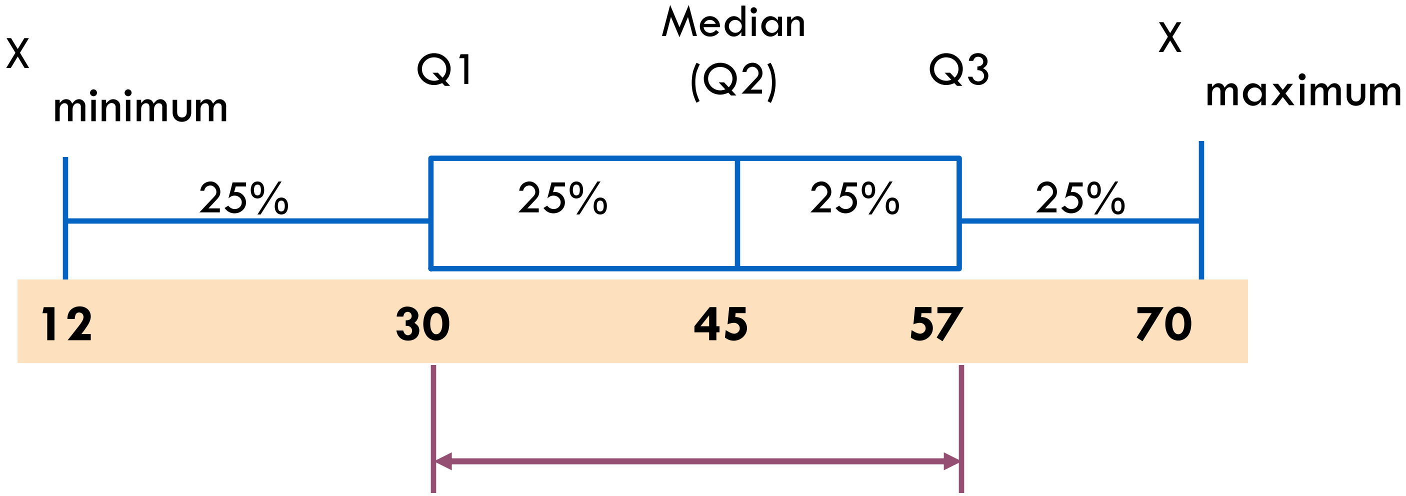
- Difference between the first and third quartiles

- Interquartile range = 3rd quartile – 1st quartile

  $$= Q_3 - Q_1$$

Interquartile Range

Example:



Interquartile range
= 57 – 30 = 27

Percentile

- Similarly, a data set (sample or population) can be even more finely divided using percentiles; the 99th percentile separates the highest 1% from the bottom 99%, and so on.

- Unless the number of observations is a multiple of 100, care must be exercised in obtaining percentiles.

Coefficient of Variation

- Measures relative variation
- Always a percentage (%)
- Shows variation relative to mean
- Is used to compare two or more sets of data measured in different units

$$\boxed{CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%}$$

Comparing Coefficients of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = \boxed{10\%}$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = \boxed{5\%}$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Population Summary Measures

- The population mean is the sum of the values in the population divided by the population size, N

$$\boxed{\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}}$$

Where

μ = population mean

N = population size

X_i = ith value of the variable X

Population Variance

- Average of squared deviations of values from the mean

  - Population variance:

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

Where

μ = population mean

N = population size

X_i = ith value of the variable X

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

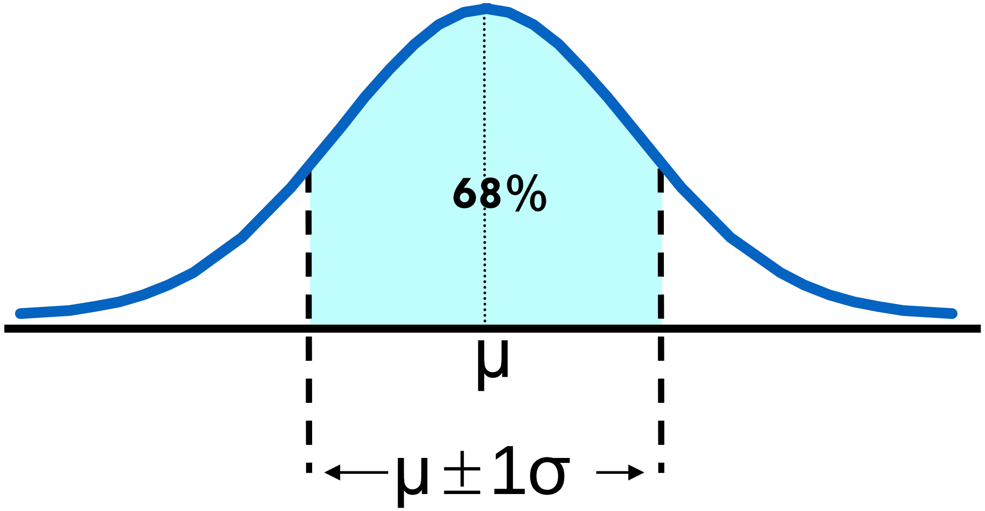
- Population standard deviation:

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}}$$

The Empirical Rule

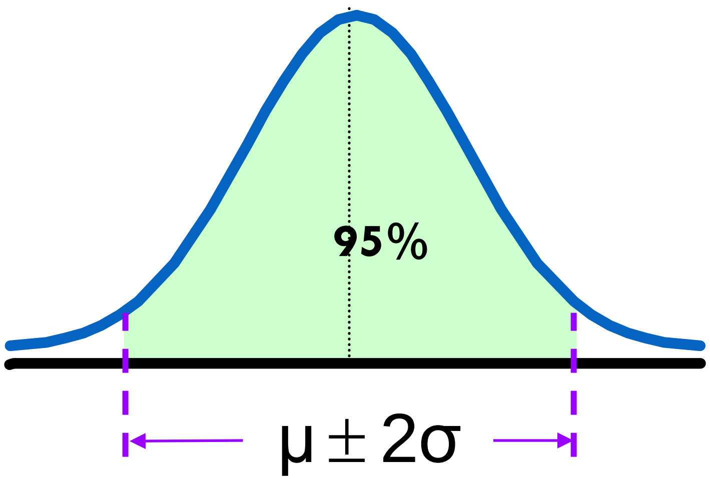
- If the data distribution is bell-shaped, then the interval: $\mu \pm 1\sigma$

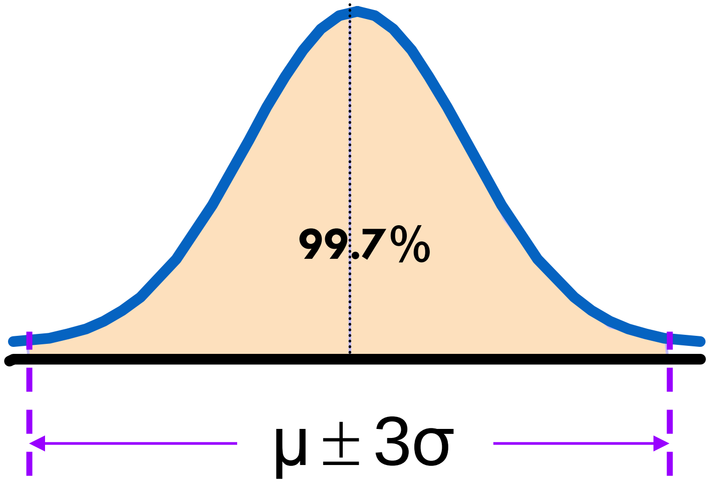
contains about 68% of the values in the population or the sample



The Empirical Rule

- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
- $\mu \pm 3\sigma$ contains about 99.7% of the values in the population or the sample





Bienaymé-Chebyshev Rule

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

  - Examples:

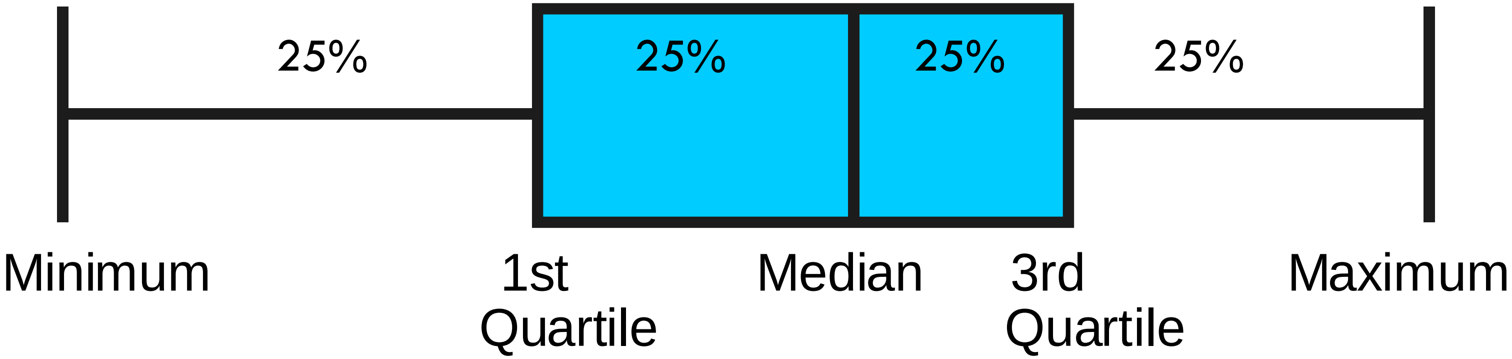
At least	within
$(1 - 1/1^2) =$ 0%	k=1 ($\mu \pm 1\sigma$)
$(1 - 1/2^2) =$ 75%	k=2 ($\mu \pm 2\sigma$)
$(1 - 1/3^2) =$ 89%	k=3 ($\mu \pm 3\sigma$)

Exploratory Data Analysis

- Box-and-Whisker Plot: A Graphical display of data using 5-number summary:

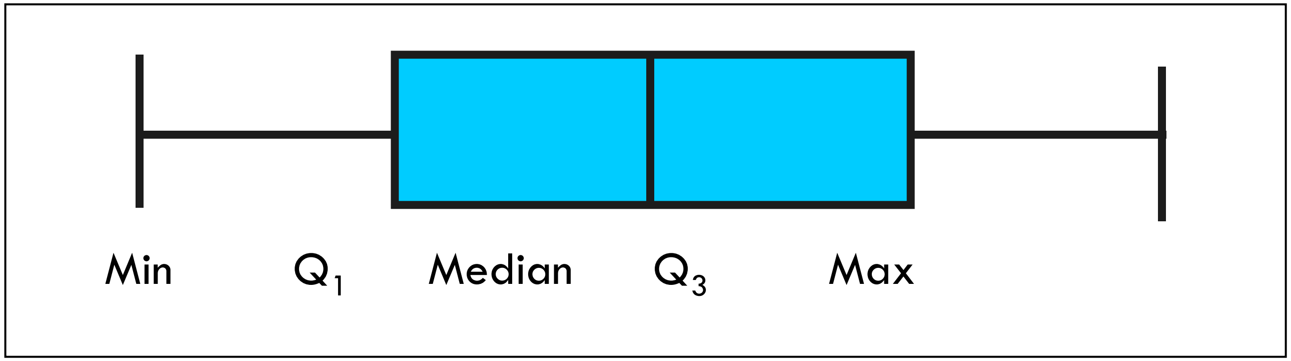
Minimum -- Q1 -- Median -- Q3 -- Maximum

Example:



Shape of Box and Whisker Plots

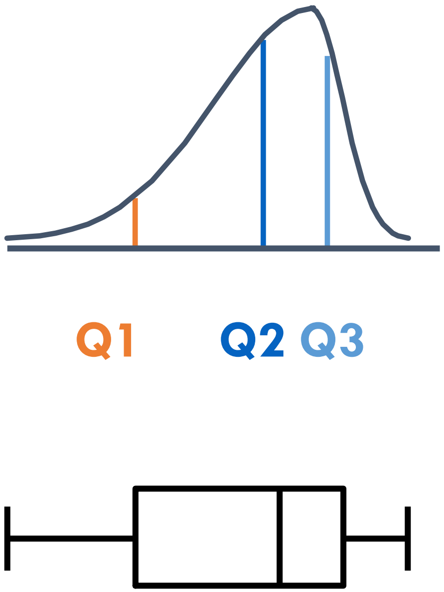
- The Box and central line are centered between the endpoints if data is symmetric around the median



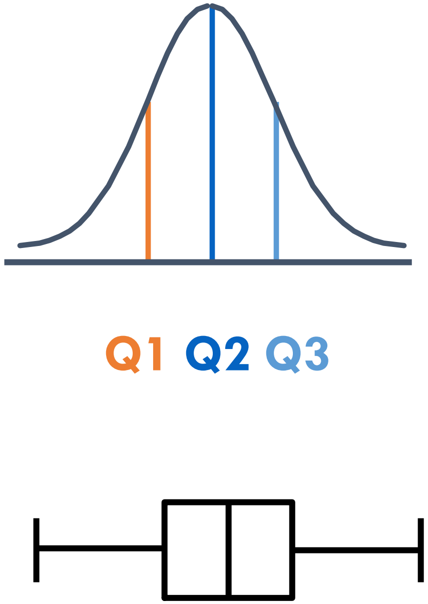
- A Box and Whisker plot can be shown in either vertical or horizontal format

Distribution Shape and Box and Whisker Plot

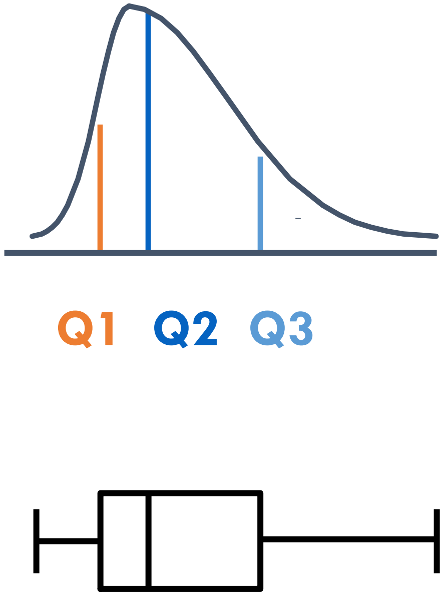
Left-Skewed



Symmetric



Right-Skewed

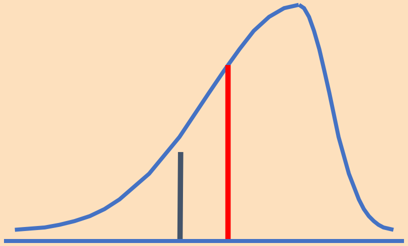


Shape of a Distribution

- Describes how data is distributed
- Measures of shape
    - Symmetric or skewed

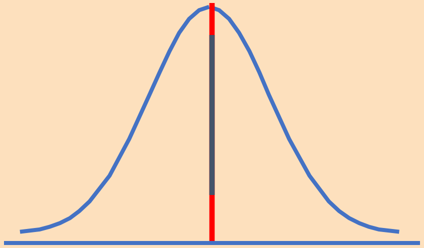
Left-Skewed

Mean < Median



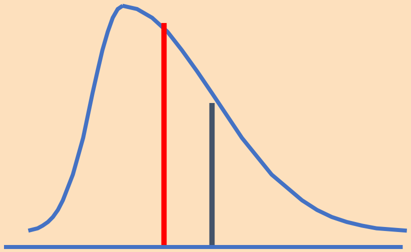
Symmetric

Mean = Median



Right-Skewed

Median < Mean

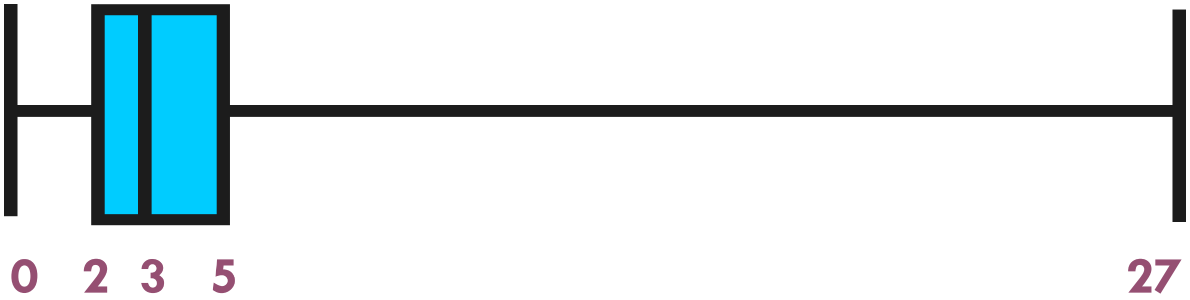


Box-and-Whisker Plot Example

- Below is a Box-and-Whisker plot for the following data:

-

Min		Q1		Q2			Q3			Max
0	2	2	2	3	3	4	5	5	10	27



- This data is right skewed, as the plot depicts

Organizing and Presenting Data Graphically

- Data in raw form are usually not easy to use for decision making
  - Some type of organization is needed
    - Table
    - Graph
- Techniques reviewed here:
  - Scatter Plot
  - Stem-and-Leaf Display
  - Frequency Distributions and Histograms
  - Bar charts

Notation

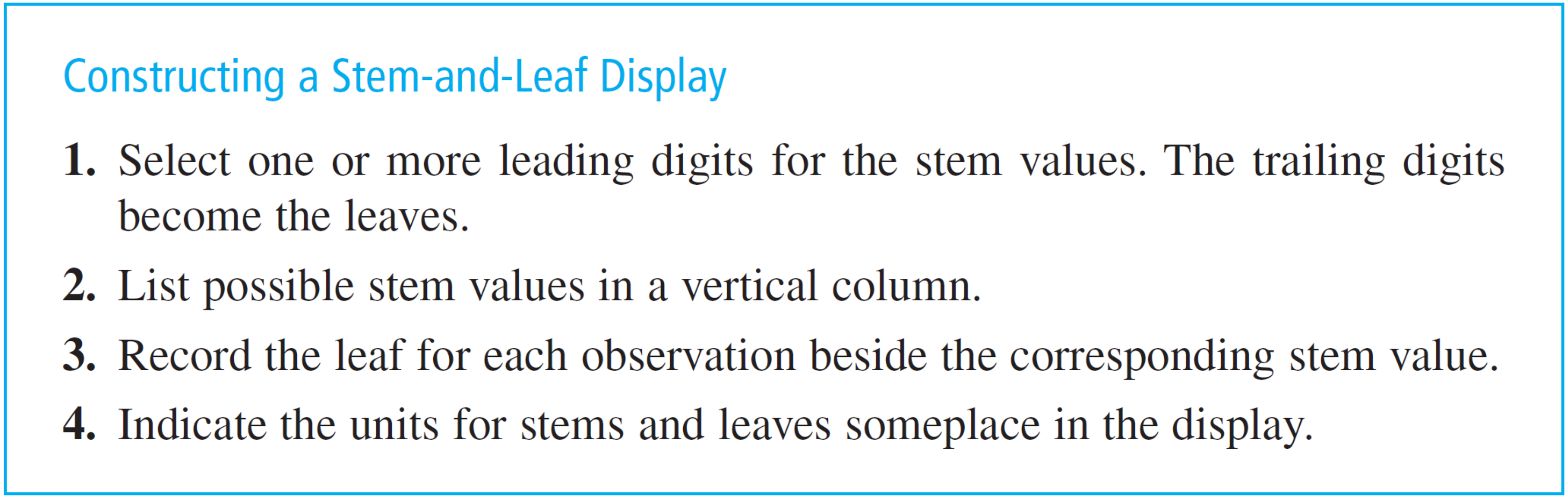
- Some general notation will make it easier to apply our methods and formulas to a wide variety of practical problems.
- The number of observations in a single sample, that is, the *sample size,* will often be denoted by *n,* so that $n = 4$ for the sample of universities {Stanford, Iowa State, Wyoming, Rochester} and also for the sample of pH measurements {6.3, 6.2, 5.9, 6.5}.
- If two samples are simultaneously under consideration, either m and n or n_1 and n_2 can be used to denote the numbers of observations.

Notation

- An experiment to compare thermal efficiencies for two different types of diesel engines might result in samples {29.7, 31.6, 30.9} and {28.7, 29.5, 29.4, 30.3}, in which case m= 3 and n= 4.
- Given a data set consisting of n observations on some variable x, the individual observations will be denoted by x_1, x_2, x_3,…, x_n. The subscript bears no relation to the magnitude of a particular observation.
- Thus x_1 will not in general be the smallest observation in the set, nor will x_n typically be the largest.
- In many applications, x_1 will be the first observation gathered by the experimenter, x_2 the second, and so on. The ith observation in the data set will be denoted by x_i.

Stem-and-Leaf Displays

- Consider a numerical data set x_1, x_2, x_3,…, x_n for which each x_i consists of at least two digits. A quick way to obtain an informative visual representation of the data set is to construct a *stem-and-leaf display.*



Constructing a Stem-and-Leaf Display

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for each observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

Stem-and-Leaf Displays

- If the data set consists of exam scores, each between 0 and 100, the score of 83 would have a stem of 8 and a leaf of 3.

- If all exam scores are in the 90s, 80s, and 70s use of the tens digit as the stem would give a display with three rows. In this case, it is desirable to stretch the display by repeating each stem value twice—9H, 9L, 8H, . . . ,7L—once for high leaves 9, . . . , 5 and again for low leaves 4, ... , 0. Then a score of 93 would have a stem of 9L and leaf of 3.

Example 1.6

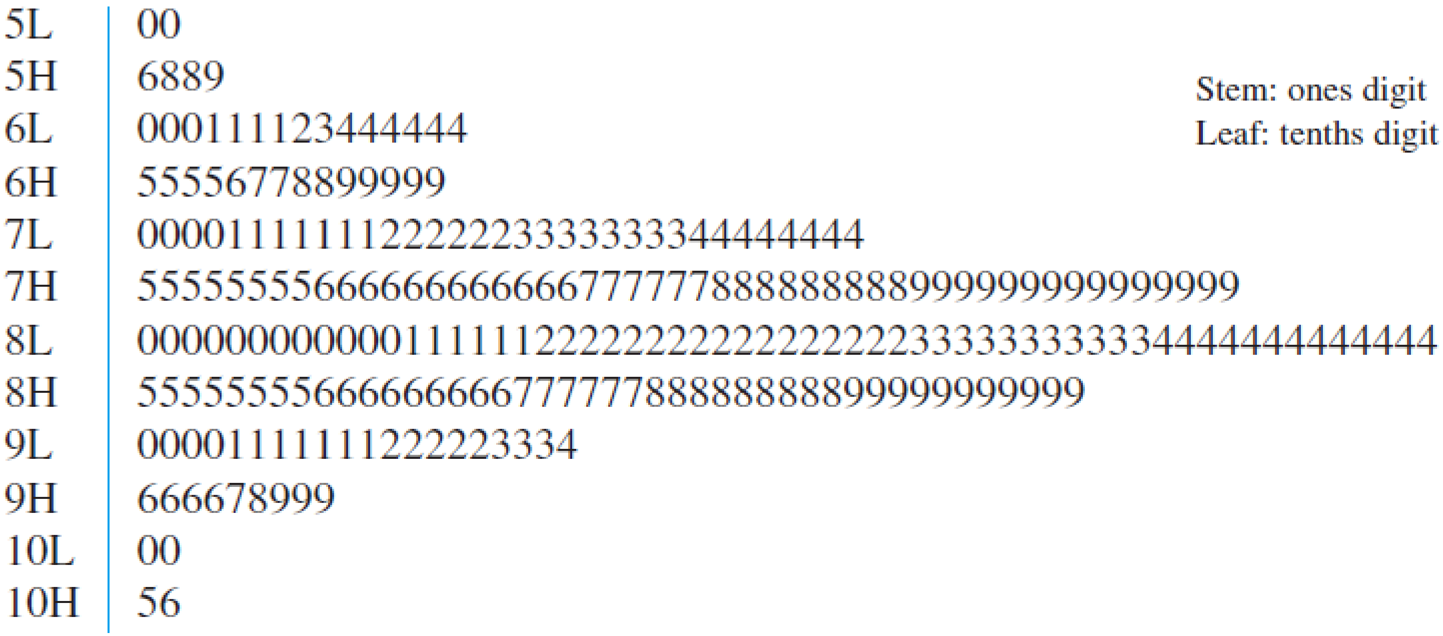
- A common complaint among college students is that they are getting less sleep than they need.

- The article “Class Start Times, Sleep, and Academic Performance in College: A Path Analysis” *Chronobiology Intl*., 2012: 318–335) investigated factors that impact sleep time.

cont'd

Example 1.6

- The stem-and-leaf display in Figure 1.4 shows the average number of hours of sleep per day over a



5L	00
5H	6889
6L	000111123444444
6H	55556778899999
7L	000011111112222223333333344444444
7H	55555555666666666666777777888888888999999999999999
8L	00000000000011111122222222222222222333333333334444444444444
8H	5555555566666666677777788888888899999999999
9L	00001111111222223334
9H	666678999
10L	00
10H	56

Stem: ones digit
Leaf: tenths digit

Figure 1.4 Stem-and-leaf display for average sleep time per day

cont'd

Example 1.6

- The first observation in the top row of the display is 5.0, corresponding to a stem of 5 and leaf of 0, and the last observation at the bottom of the display is 10.6. Note that in the absence of a context, without the identification of stem and leaf digits in the display, we wouldn't know whether the observation with stem 7 and leaf 9 was .79, 7.9, or 79.
- The leaves in each row are ordered from smallest to largest; this is commonly done by software packages but is not necessary if a display is created by hand.

cont'd

Example 1.6

- The display suggests that a typical or representative sleep time is in the stem 8L row, perhaps 8.1 or 8.2. The data is not highly concentrated about this typical value as would be the case if almost all students were getting between 7.5 and 9.5 hours of sleep on average.
- The display appears to rise rather smoothly to a peak in
- the 8L row and then decline smoothly (we conjecture that the minor peak in the 6L row would disappear if more data was available).

cont’d

Example 1.6

- The general shape of the display is rather symmetric, bearing strong resemblance to a bell-shaped curve; it does not stretch out more in one direction than the other.

- The two smallest and two largest values seem a bit separated from the remainder of the data—perhaps they are very mild, but certainly not extreme,“outliers”.

cont’d

Example 1.6

- A reference in the cited article suggests that individuals in this age group need about 8.4 hours of sleep per day. So it appears that a substantial percentage of students in the sample are sleep deprived.

Example 1:

Data in ordered array:

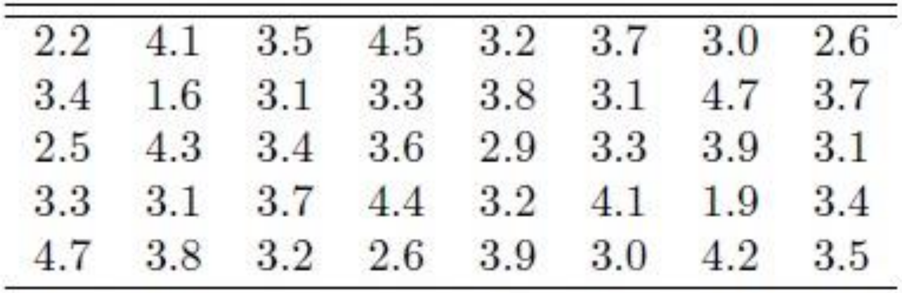
21, 24, 24, 26, 27, 27, 30, 32, 38, 41

- Completed Stem-and-leaf diagram:

Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

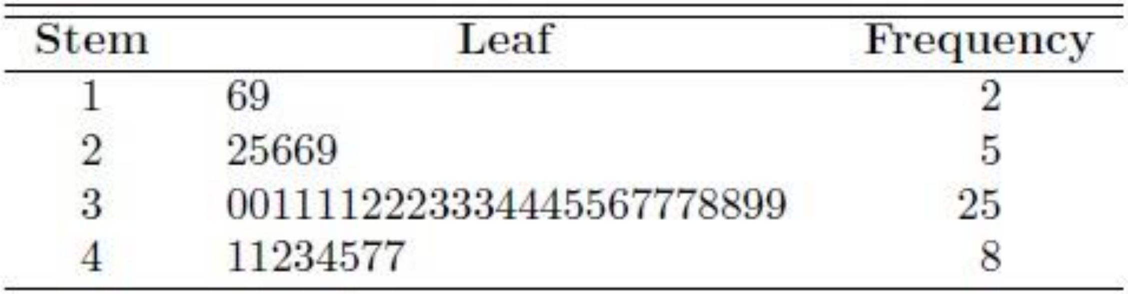
Example 2:

Table 1.4: Car Battery Life



2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Table 1.5: Stem-and-Leaf Plot of Battery Life



Stem	Leaf	Frequency
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

Using other stem units

- Using the 100's digit as the stem:
  - Round off the 10's digit to form the leaves

- 613 would become →
- 776 would become →
- . . .
- 1224 becomes →

Stem	Leaf
6	1
7	8
12	2

Using other stem units

(continued)

- Using the 100’s digit as the stem:
  - The completed stem-and-leaf display:

Data:

613, 632, 658, 717, 722, 750, 776, 827, 841, 859, 863, 891, 894, 906, 928, 933, 955, 982, 1034, 1047,1056, 1140, 1169, 1224

Stem	Leaves
6	1 3 6
7	2 2 5 8
8	3 4 6 6 9 9
9	1 3 3 6 8
10	3 5 6
11	4 7
12	2

Stem-and-Leaf Displays

- A stem-and-leaf display conveys information about the following aspects of the data:
  - identification of a typical or representative value
  - extent of spread about the typical value
  - presence of any gaps in the data
  - extent of symmetry in the distribution of values
  - number and location of peaks
  - presence of any outlying values

Dotplots

- A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale.

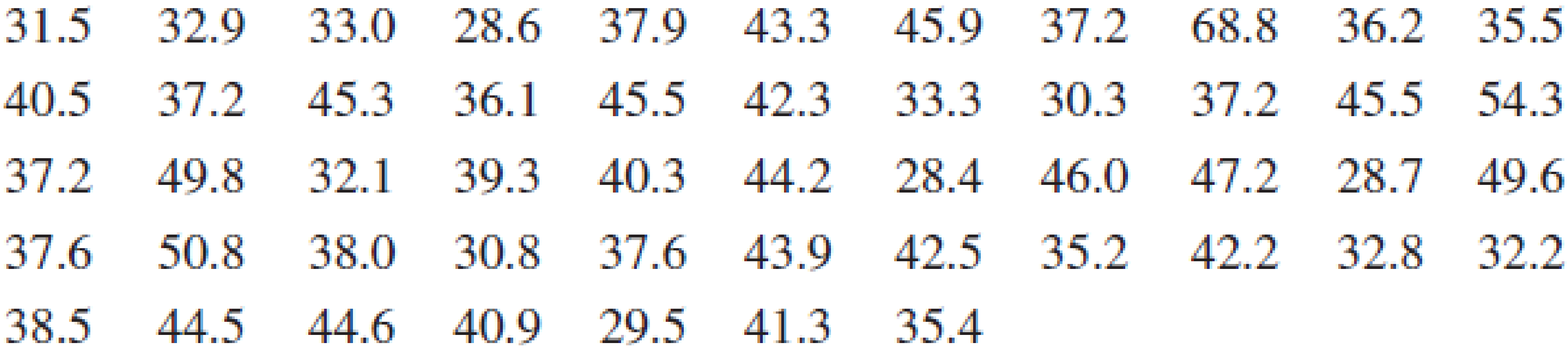
- When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Example 1.8

- There is growing concern in the U.S. that not enough students are graduating from college. America used to be number 1 in the world for the percentage of adults with college degrees, but it has recently dropped to 16th. Here is data on the percentage of 25- to 34-year-olds in each state who had some type of postsecondary degree as of 2010 (listed in alphabetical order, with the District of Columbia included):
    - 31.5 32.9 33.0 28.6 37.9 43.3 45.9 37.2 68.8 36.2 35.5
    - 40.5 37.2 45.3 36.1 45.5 42.3 33.3 30.3 37.2 45.5 54.3
    - 37.2 49.8 32.1 39.3 40.3 44.2 28.4 46.0 47.2 28.7 49.6
    - 37.6 50.8 38.0 30.8 37.6 43.9 42.5 35.2 42.2 32.8 32.2
    - 38.5 44.5 44.6 40.9 29.5 41.3 35.4

Example 1.8

- Here is data on the percentage of 25- to 34-year-olds in each state who had some type of postsecondary degree as of 2010 (listed in alphabetical order, with the District of Columbia included):



31.5	32.9	33.0	28.6	37.9	43.3	45.9	37.2	68.8	36.2	35.5
40.5	37.2	45.3	36.1	45.5	42.3	33.3	30.3	37.2	45.5	54.3
37.2	49.8	32.1	39.3	40.3	44.2	28.4	46.0	47.2	28.7	49.6
37.6	50.8	38.0	30.8	37.6	43.9	42.5	35.2	42.2	32.8	32.2
38.5	44.5	44.6	40.9	29.5	41.3	35.4				

cont’d

Example 1.8

- Figure 1.6 shows a dotplot of the data. There is clearly a

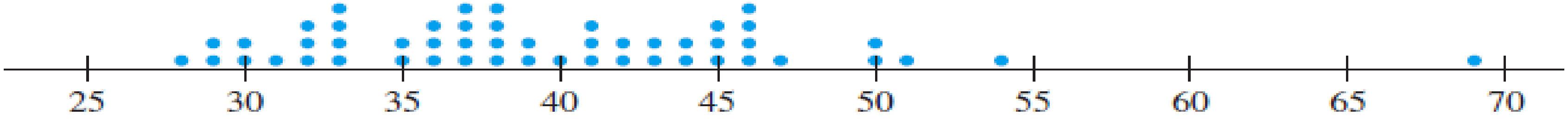


Figure 1.6 A dotplot of the data from Example 1.8

- The largest value, for D.C., is obviously an extreme
- outlier, and four other values on the upper end of the data are candidates for mild outliers (MA, MN, NY, and ND). There is also a cluster of states at the low end, primarily located in the South and Southwest.

Dotplots

- The overall percentage for the entire country is 39.3%; this
- is not a simple average of the 51 numbers but an average weighted by population sizes.
- A dotplot can be quite cumbersome to construct and look crowded when the number of observations is large. Our next technique is well suited to such situations.

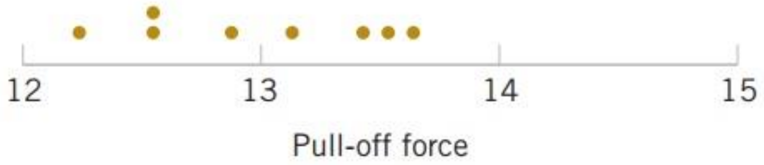


FIGURE 1-2 Dot diagram of the pull-off force data when wall thickness is 3/32 inch.

$$X = \mu + \in,$$

Where X= random variable, μ is constant and $\in$ is random disturbance

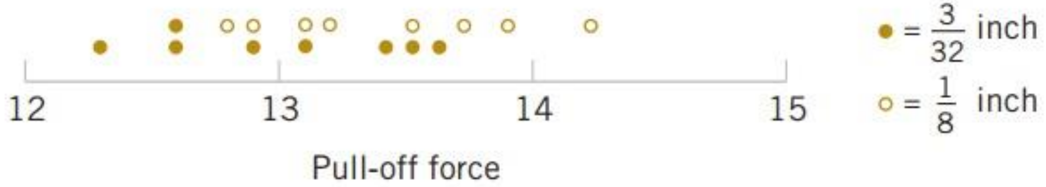


FIGURE 1-3 Dot diagram of pull-off force for two wall thicknesses.

- Data for Figure 1-2 : 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6, and 13 1. The average is 13.0
- Data for Figure 1-3: 12.9, 13.7, 12.8, 13.9, 14.2, 13.2, 13.5, and 13.1. The average is 13.4

Observing Processes overtime: Time series Plot

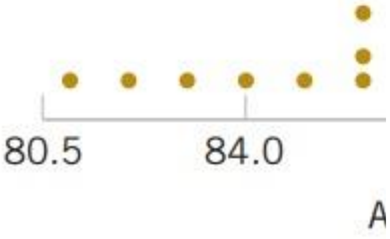


FIGURE 1-8 The
but does not ident

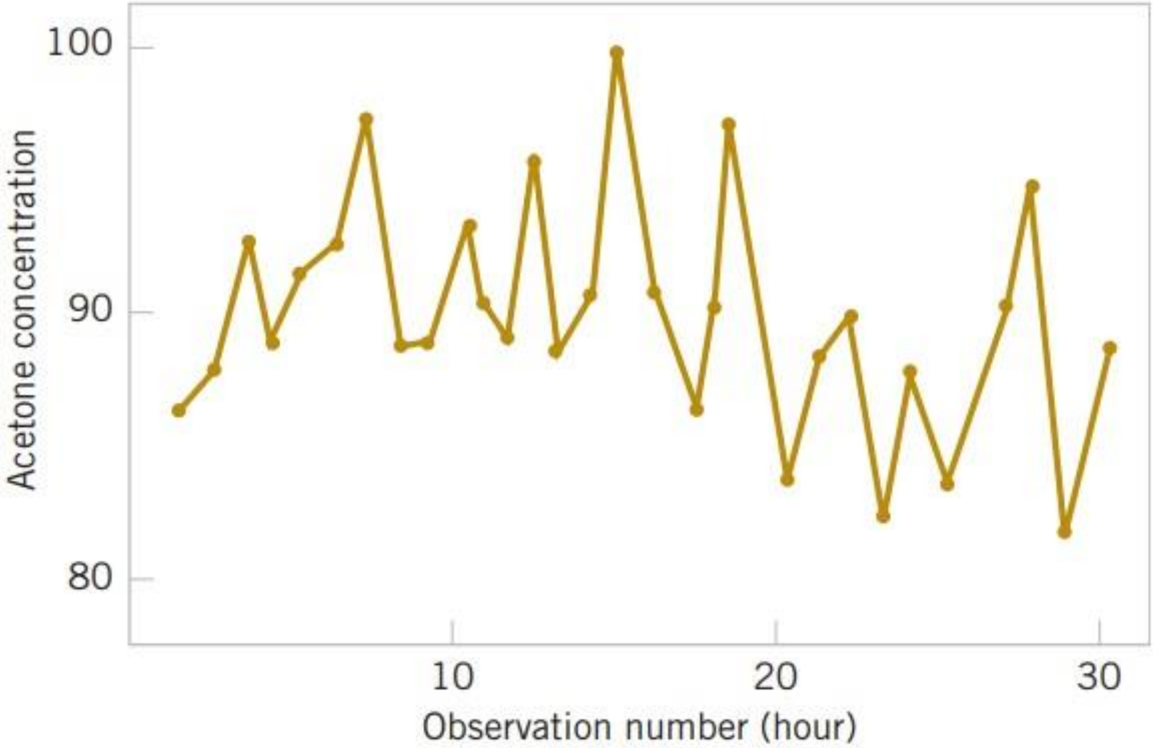


FIGURE 1-9 A time series plot of concentration provides more information than the dot diagram.

Histograms

- Some numerical data is obtained by **counting to determine the value of a variable** (the number of traffic citations a person received during the last year, the number of customers arriving for service during a particular period), whereas other data is obtained by **taking measurements** (weight of an individual, reaction time to a particular stimulus).

- The prescription for drawing a histogram is generally different for these two cases.

Histograms

- A numerical variable is **discrete** if its set of possible values either is finite or else can be listed in an infinite sequence (one in which there is a first number, a second number, and so on). A numerical variable is **continuous** if its possible values consist of an entire interval on the number line.
- A **discrete variable *x* almost always results from counting**, in which case possible values are 0, 1, 2, 3, ... or some subset of these integers. **Continuous variables arise from making measurements.** For example, if *x* is the pH of a chemical substance, then in theory *x* could be any number between 0 and 14: 7.0, 7.03, 7.032, and so on.

Histograms

- Of course, in practice **there are limitations on the degree of accuracy of any measuring instrument**, so we may not be able to determine pH, reaction time, height, and concentration to an arbitrarily large number of decimal places.
- However, from the point of view of creating mathematical models for distributions of data, **it is helpful to imagine an entire continuum of possible values.**
- Consider data consisting of observations on a discrete variable *x.* The **frequency** of any particular *x* value is the **number of times that value occurs in the data set.**

Histograms

- The **relative frequency** of a value is the fraction or proportion of times the value occurs:

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

- Suppose, for example, that our data set consists of 200 observations on *x* = the number of courses a college student is taking this term. If 70 of these *x* values are 3, then
- **frequency** of the *x* value 3: 70
- **relative frequency** of the *x* value 3:

$$\frac{70}{200} = .35$$

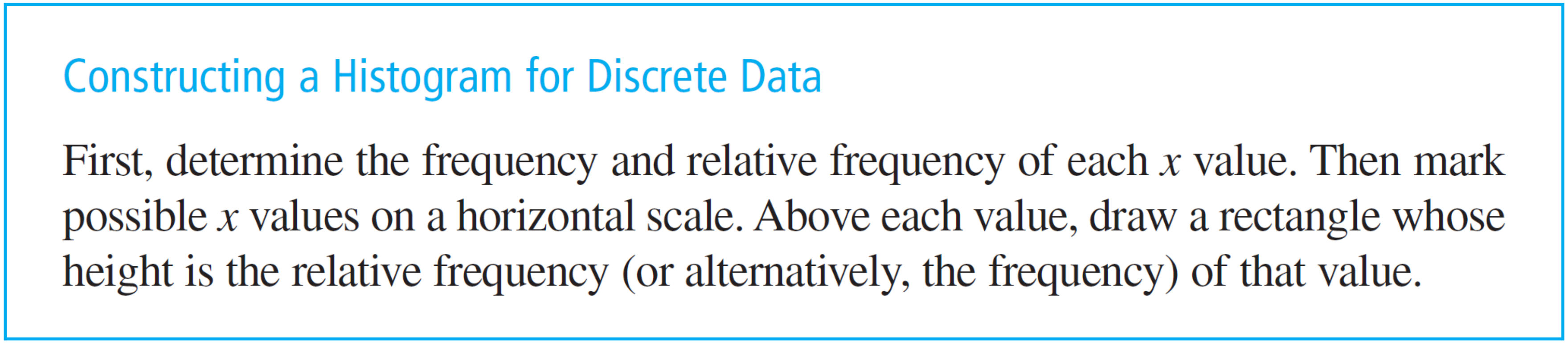
Histograms

- Multiplying a relative frequency by 100 gives a percentage; in the college-course example, **35% of the students in the sample are taking three courses.**

- **The relative frequencies, or percentages, are usually of more interest than the frequencies themselves.** In theory, the relative frequencies should sum to 1, but in practice the sum may differ slightly from 1 because of rounding.

- A **frequency distribution** is **a tabulation of the frequencies and/or relative frequencies.**

Histograms



Constructing a Histogram for Discrete Data

First, determine the frequency and relative frequency of each *x* value. Then mark possible *x* values on a horizontal scale. Above each value, draw a rectangle whose height is the relative frequency (or alternatively, the frequency) of that value.

- **This construction ensures that the *area* of each rectangle is proportional to the relative frequency of the value.** Thus if the **relative frequencies of *x* = 1 and *x* = 5 are .35 and .07, respectively, then the area of the rectangle above 1 is five times the area of the rectangle above 5.**

Frequency Histogram

- Frequency histogram:
  - Given a sample of data points, we divide data into equally-spaced intervals, and count the number of data points that fall into each interval.
  - A histogram is a bar chart with the length of each bar proportional to the number of observations in that interval.
  - A histogram for a sample will be an approximation of the probability distribution of the population.
- Probability distributions:
  - Show much more about a population than just the mean and standard deviation.
  - A distribution may be symmetric, or may be skewed to the right or the left.
  - The tail of a distribution shows the distance from the mean of the outlying points (for example, the 95th percentile point).

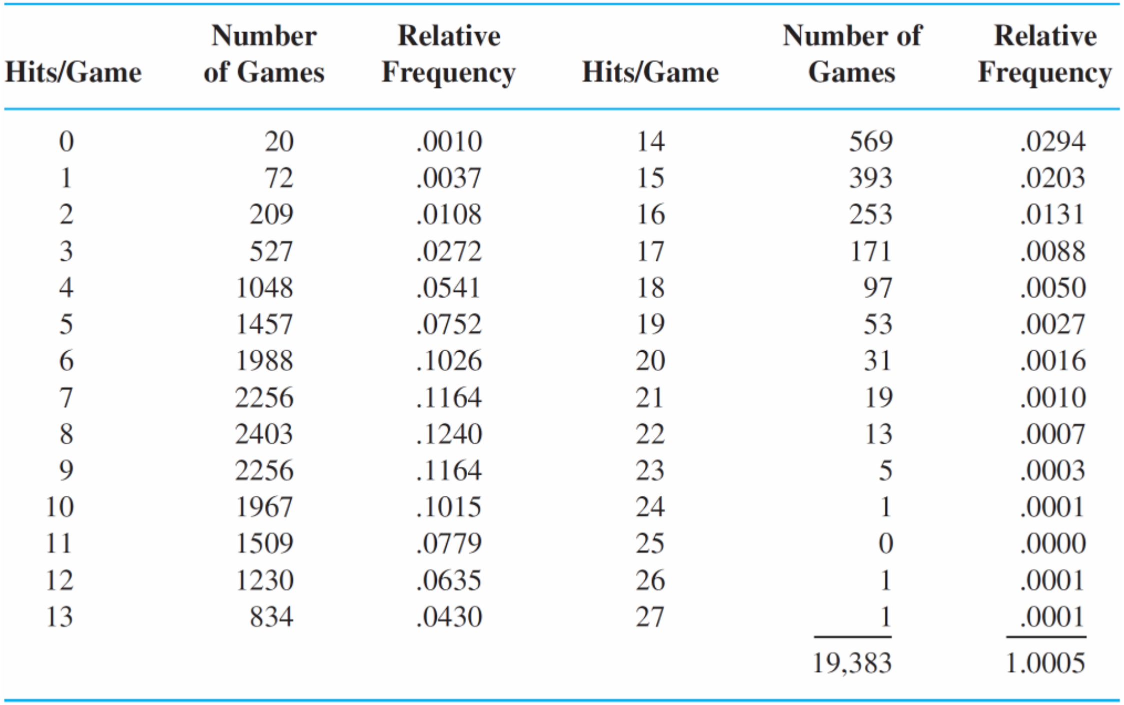
Example 1.9

- How unusual is a no-hitter or a one-hitter in a major league baseball game, and how frequently does a team get more than 10, 15, or even 20 hits?

Example 1.9

cont'd

- **Table 1.1 is a frequency distribution for the number of hits per team per game for all nine-inning games that were played between 1989 and 1993.**



Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

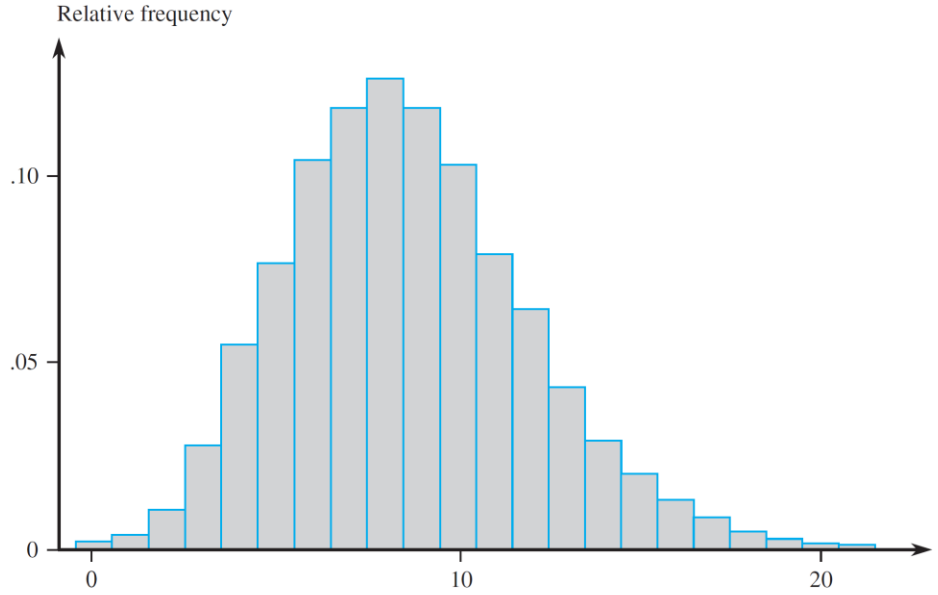
Frequency Distribution for Hits in Nine-Inning Games

Table 1.1

Example 1.9

cont'd

- The corresponding histogram in Figure 1.7 rises rather smoothly to a single peak and then declines. The histogram extends a bit more on the right (toward large values) than it does on the left—a slight "positive skew."



Histogram of number of hits per nine-inning game

Figure 1.7

cont’d

Example 1.9

- Either from the tabulated information or from the histogram itself, we can determine the following:

$$\begin{array}{l} \text{proportion of games with} \\ \text{at most two hits} \end{array} = \begin{array}{c} \text{relative} \\ \text{frequency} \\ \text{for } x = 0 \end{array} + \begin{array}{c} \text{relative} \\ \text{frequency} \\ \text{for } x = 1 \end{array} + \begin{array}{c} \text{relative} \\ \text{frequency} \\ \text{for } x = 2 \end{array}$$

$$= .0010 + .0037 + .0108$$

$$= .0155$$

cont’d

Example 1.9

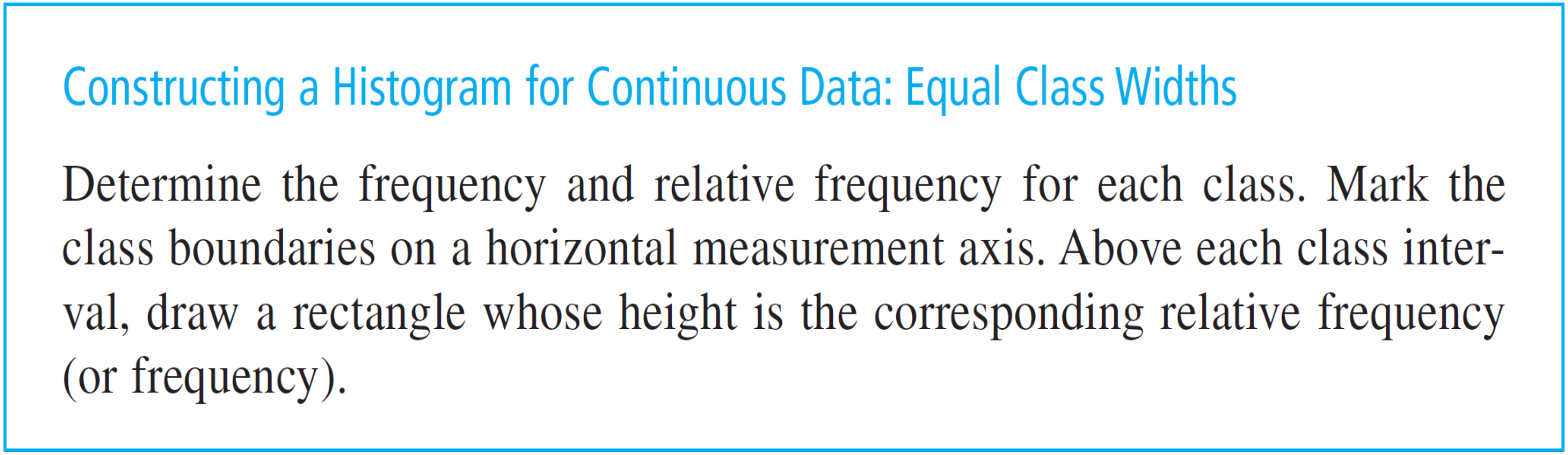
- Similarly,

$$\begin{array}{l}\text{proportion of games with} \\ \text{between 5 and 10 hits (inclusive)}\end{array} = .0752 + .1026 + \cdots + .1015$$

$$= .6361$$

- That is, **roughly 64% of all these games resulted in between 5 and 10 (inclusive) hits.**

Histograms



Constructing a Histogram for Continuous Data: Equal Class Widths

Determine the frequency and relative frequency for each class. Mark the class boundaries on a horizontal measurement axis. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

- Constructing a histogram for continuous data (measurements) entails subdividing the measurement axis into a suitable number of **class intervals** or **classes,** such that each observation is contained in exactly one class.

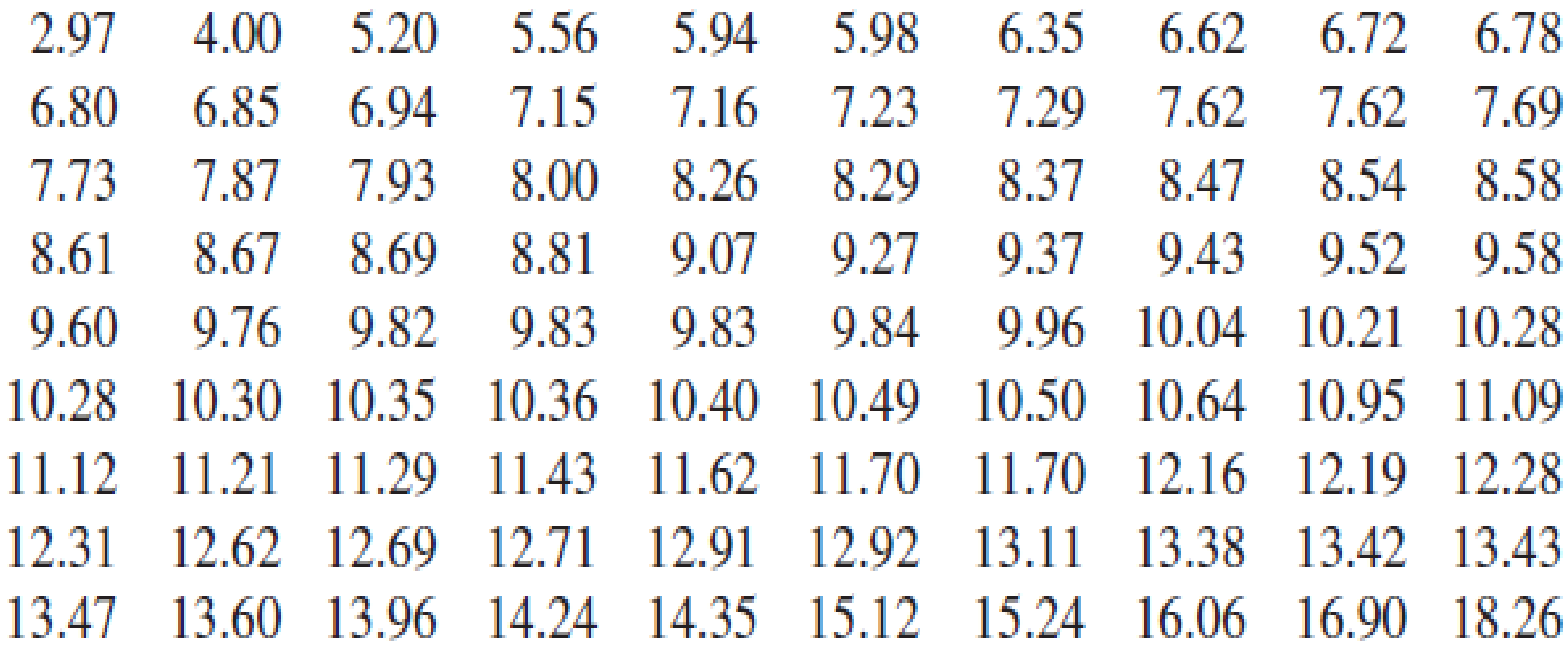
Example 1.10

- Power companies need information about customer usage to obtain accurate forecasts of demands. Investigators from Wisconsin Power and Light determined energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes. An adjusted consumption value was calculated as follows:

$$\text{adjusted consumption} = \frac{\text{consumption}}{(\text{weather, in degree days})(\text{house area})}$$

Example 1.10

- This resulted in the accompanying data (part of the stored data set FURNACE.MTW available in Minitab), which we have ordered from smallest to largest.



2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

Example 1.10

- The most striking feature of the histogram in Figure 1.8 is its resemblance to a bell-shaped curve, with the point of symmetry roughly at 10.

Class	1–<3	3–<5	5–<7	7–<9	9–<11	11–<13	13–<15	15–<17	17–<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011

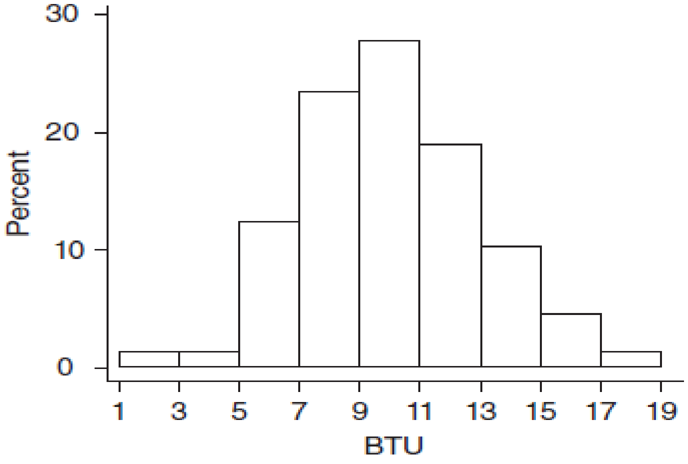
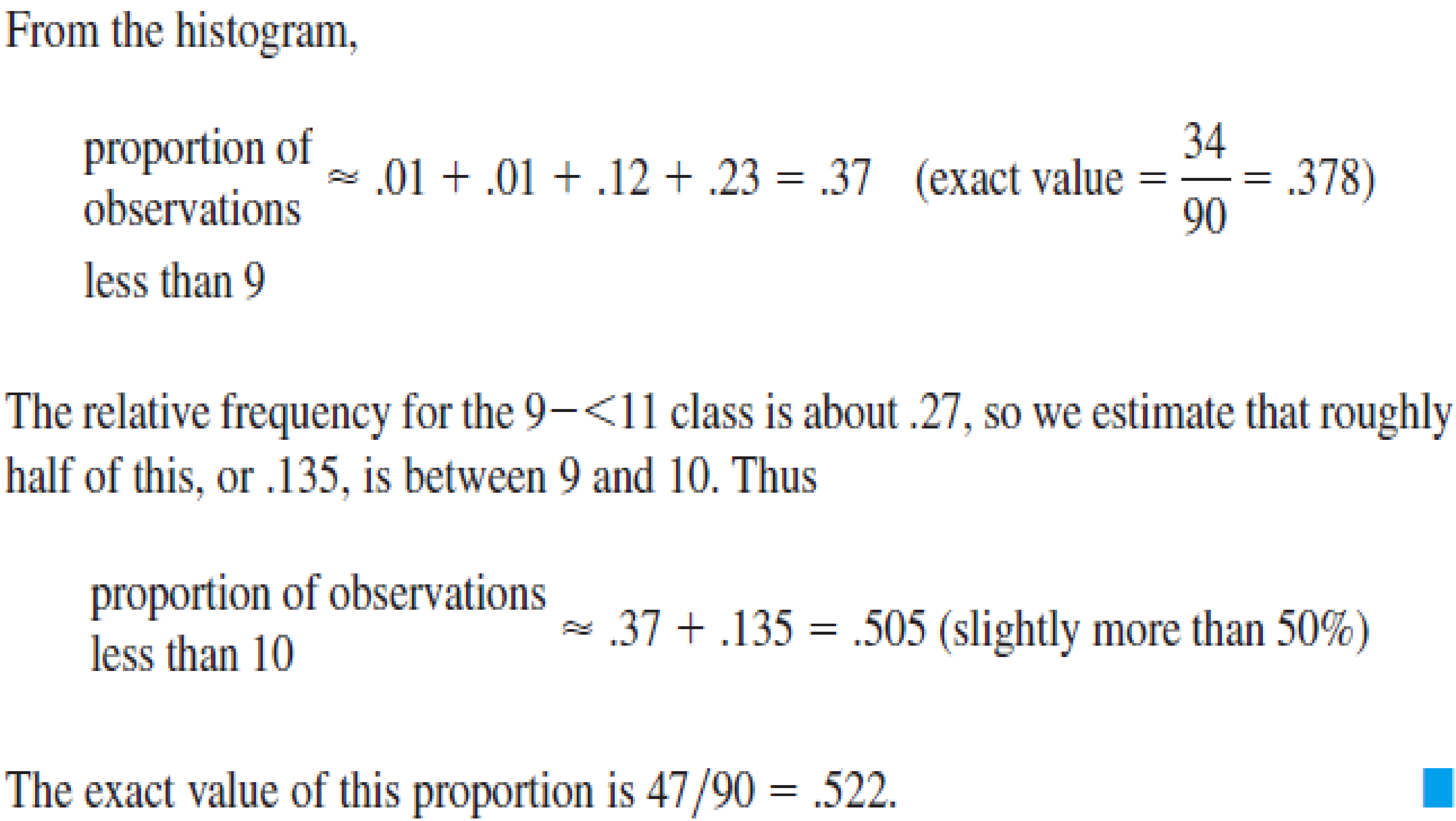


Figure 1.8 Histogram of the energy consumption data from Example 1.10

Example 1.10



From the histogram,

$$\begin{array}{l}\text{proportion of}\\ \text{observations}\\ \text{less than 9}\end{array} \approx .01 + .01 + .12 + .23 = .37 \quad \left(\text{exact value} = \frac{34}{90} = .378\right)$$

The relative frequency for the 9−<11 class is about .27, so we estimate that roughly half of this, or .135, is between 9 and 10. Thus

$$\begin{array}{l}\text{proportion of observations}\\ \text{less than 10}\end{array} \approx .37 + .135 = .505 \text{ (slightly more than 50\%)}$$

The exact value of this proportion is $47/90 = .522$. ■

Histograms

- **Equal-width classes may not be a sensible choice if there are some regions of the measurement scale that have a high concentration of data values and other parts where data is quite sparse.**

- Figure 1.9 shows a dotplot of such a data set; there is high concentration in the middle, and relatively few observations stretched out to either side. Using a small number of equal-width classes results in almost all observations falling in just one or two of the classes.

Histograms

- If a large number of equal-width classes are used, many classes will have zero frequency. **A sound choice is to use a few wider intervals near extreme observations and narrower intervals in the region of high concentration.**

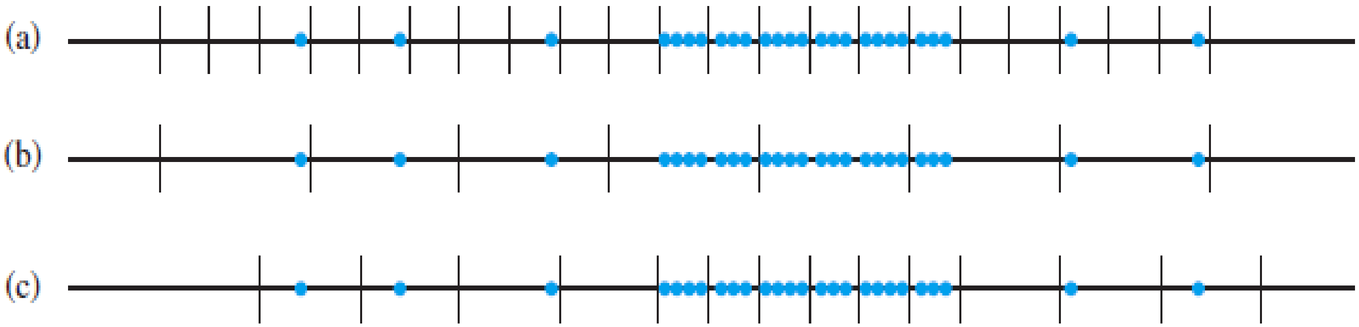
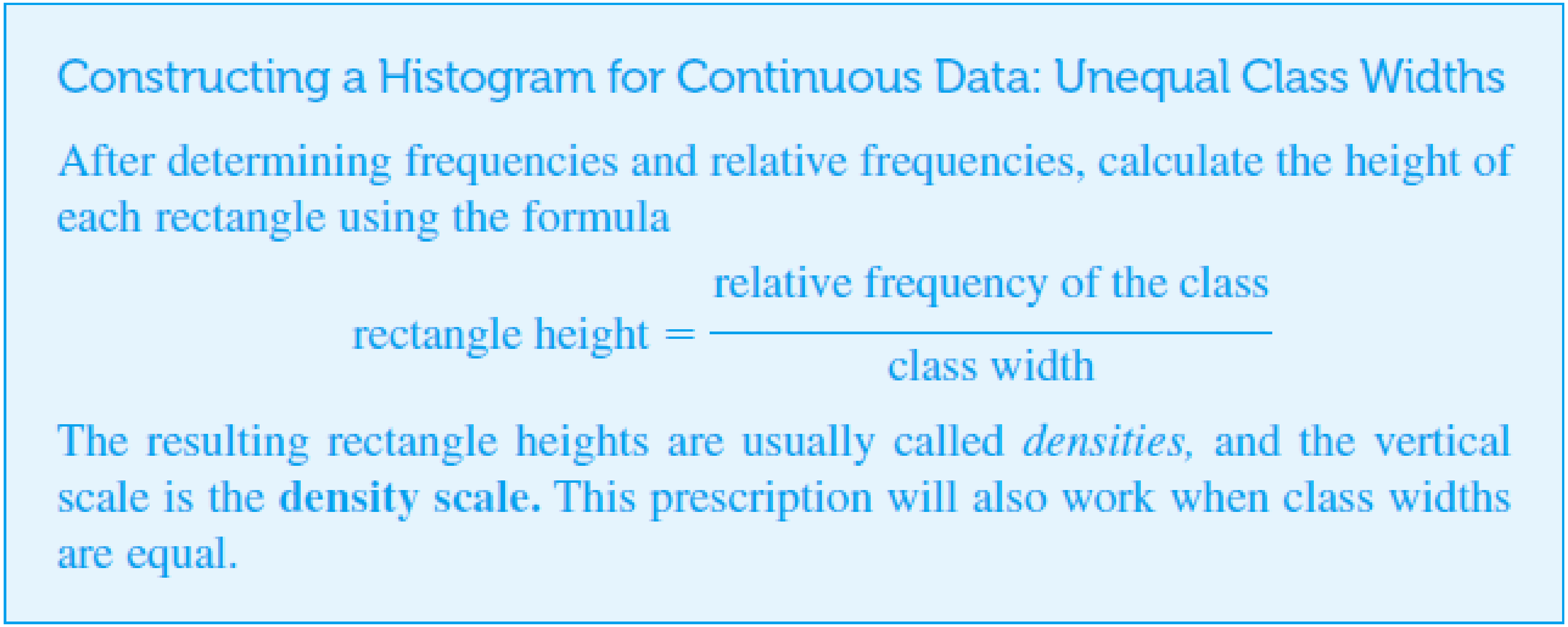


Figure 1.9 Selecting class intervals for "varying density" data: (a) many short equal-width intervals; (b) a few wide equal-width intervals; (c) unequal-width intervals

Histograms



Constructing a Histogram for Continuous Data: Unequal Class Widths

After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

The resulting rectangle heights are usually called *densities*, and the vertical scale is the **density scale.** This prescription will also work when class widths are equal.

Example 1.11

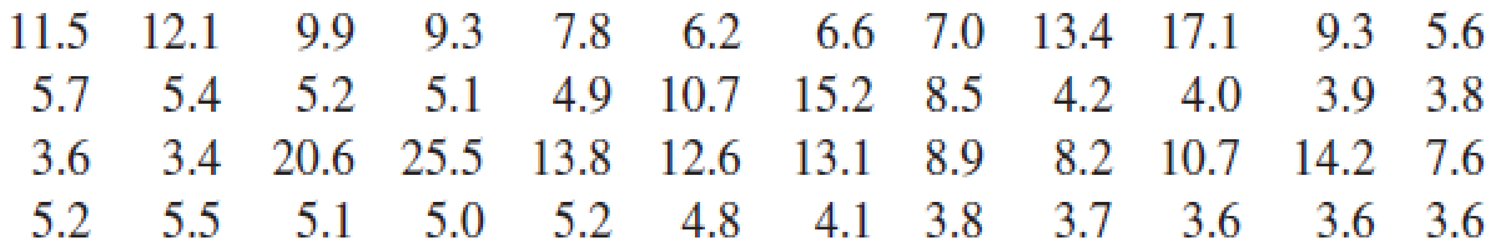
- Corrosion of reinforcing steel is a serious problem in concrete structures located in environments affected by severe weather conditions.

- For this reason, researchers have been investigating the use of reinforcing bars made of composite material.

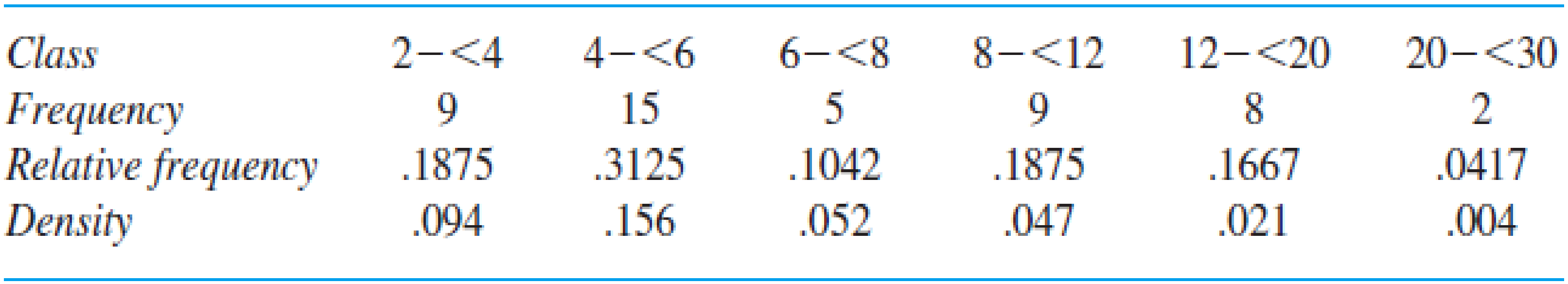
- One study was carried out to develop guidelines for bonding glass-fiber-reinforced plastic rebars to concrete ("Design Recommendations for Bond of GFRP Rebars to Concrete," *J. of Structural Engr.*, 1996: 247–254).

Example 1.11

- Consider the following 48 observations on measured bond strength:



11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6



Class	2−<4	4−<6	6−<8	8−<12	12−<20	20−<30
Frequency	9	15	5	9	8	2
Relative frequency	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

Example 1.11

- The resulting histogram appears in Figure 1.10. The right or upper tail stretches out much farther than does the left or lower tail—a substantial departure from symmetry.

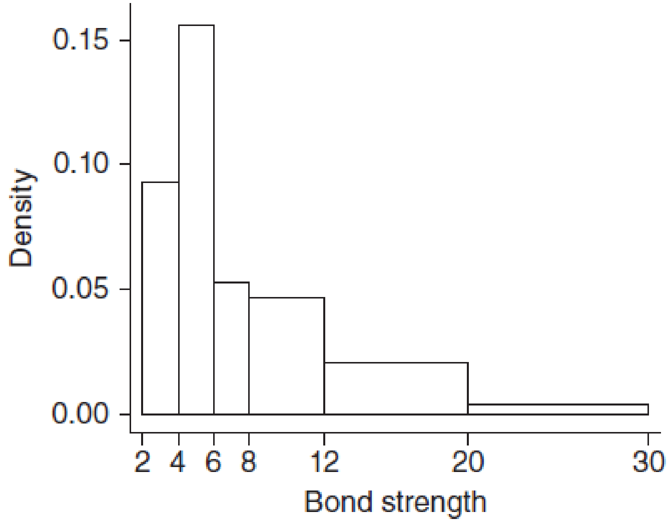


Figure 1.10 A Minitab density histogram for the bond strength data of Example 1.11 ■

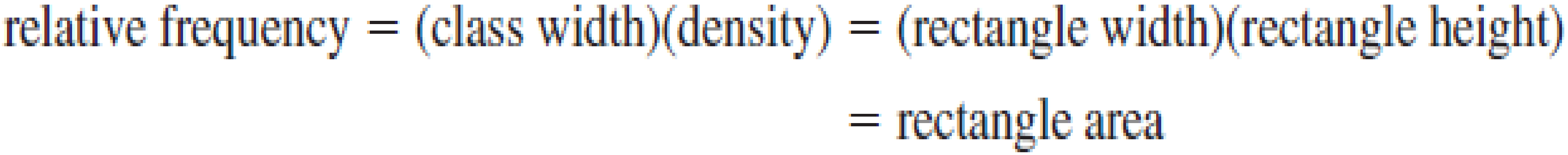
Histograms

- When class widths are unequal, not using a density scale will give a picture with distorted areas.

- For equal-class widths, the divisor is the same in each density calculation, and the extra arithmetic simply results in a rescaling of the vertical axis (i.e., the histogram using relative frequency and the one using density will have exactly the same appearance).

Histograms

- Multiplying both sides of the formula for density by the class width gives



$$\text{relative frequency} = (\text{class width})(\text{density}) = (\text{rectangle width})(\text{rectangle height})$$
$$= \text{rectangle area}$$

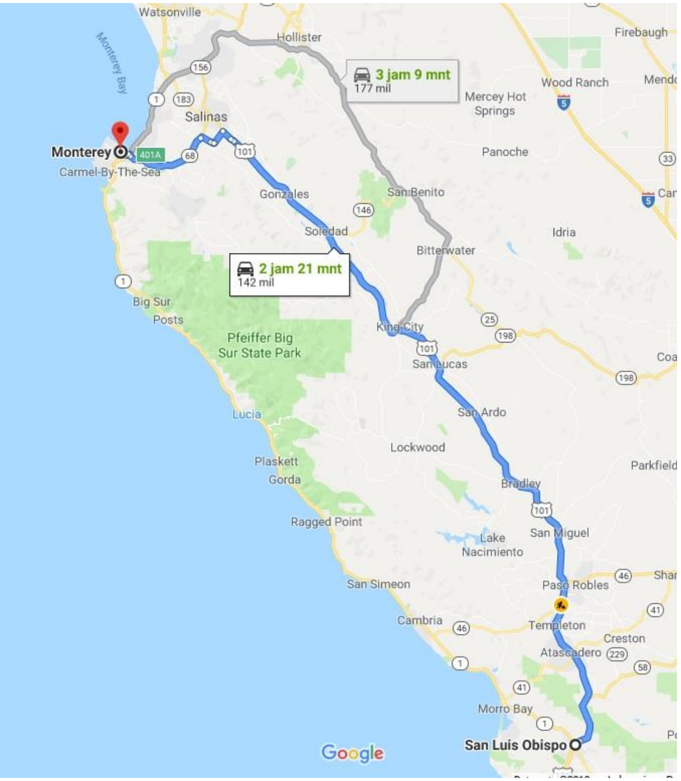
- **That is, *the area of each rectangle is the relative frequency of the corresponding class.*** Furthermore, **since the sum of relative frequencies should be 1, *the total area of all rectangles in a density histogram is* l.**

Histogram Shapes

- Histograms come in a variety of shapes. A **unimodal** histogram is one that rises to a single peak and then declines. A **bimodal** histogram has two different peaks.
- **Bimodality can occur when the data set consists of observations on two quite different kinds of individuals or objects.**

Histogram Shapes

- For example, consider a large data set consisting of driving times for automobiles traveling between **San Luis Obispo, California, and Monterey, California** (exclusive of stopping time for sightseeing, eating, etc.).
- This histogram would show two peaks: one for those cars that took the inland route (roughly 2.5 hours) and another for those cars traveling up the coast (3.5–4 hours).
- However, **bimodality does not automatically follow in such situations. Only if the two separate histograms are "far apart" relative to their spreads will bimodality occur in the histogram of combined data.**
- Thus a large data set consisting of heights of college students should not result in a bimodal histogram because the typical male height of about 69 inches is not far enough above the typical female height of about 64–65 inches.

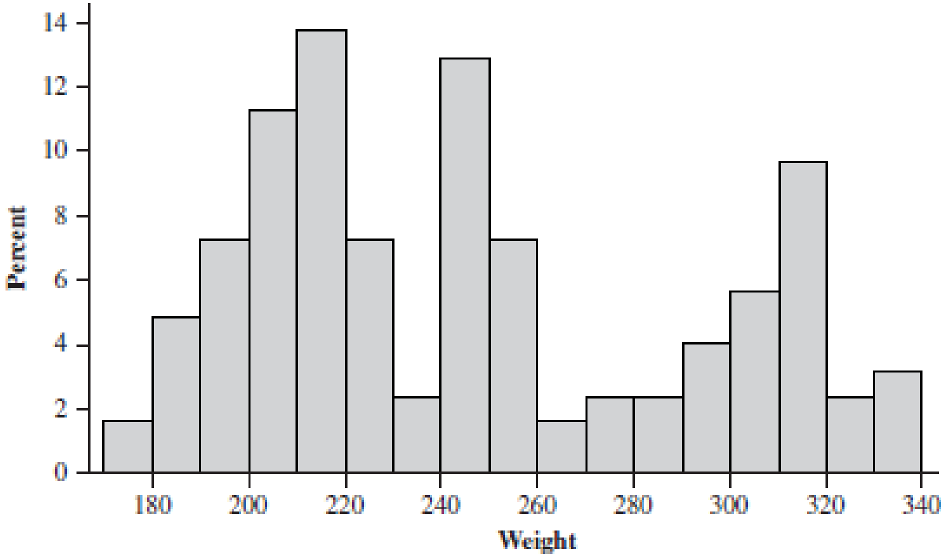


Histogram Shapes

- A histogram with more than two peaks is said to be **multimodal.** Of course, the number of peaks may well depend on the choice of class intervals, particularly with a small number of observations. The larger the number of classes, the more likely it is that bimodality or multimodality will manifest itself.

Example 1.12

- Figure 1.11(a) shows a Minitab histogram of the weights (lb) of the 124 players listed on the rosters of the San Francisco 49ers and the New England Patriots (teams the author would like to see meet in the Super Bowl) as of Nov. 20, 2009.



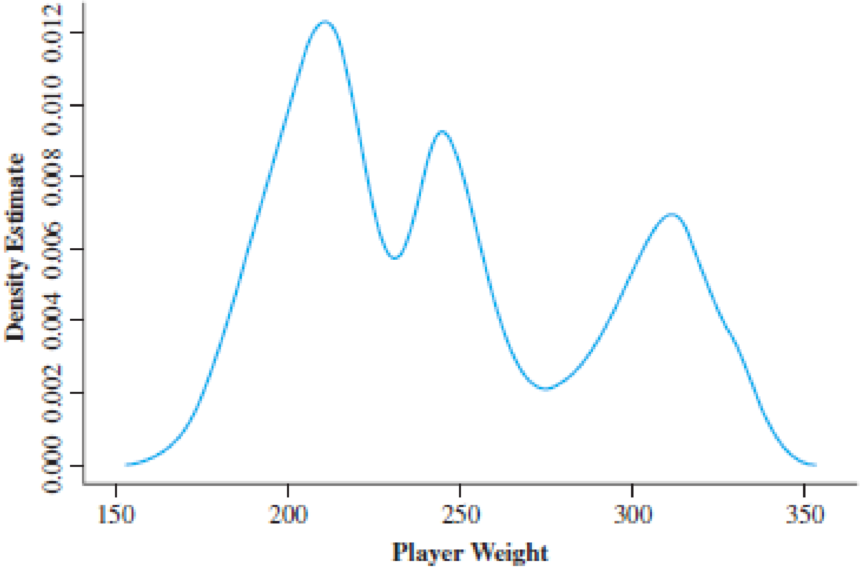
NFL player weights Histogram

Figure 1.11(a)

cont'd

Example 12

- Figure 1.11(b) is a smoothed histogram (actually what is called a *density estimate*) of the data from the R software package.



NFL player weights Smoothed histogram

Figure 1.11(b)

Example 1.12

- Both the histogram and the smoothed histogram show three distinct peaks; **the one on the right is for linemen**, **the middle peak corresponds to linebacker weights**, and **the peak on the left is for all other players** (wide receivers, quarterbacks, etc.).

- A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail and **negatively skewed** if the stretching is to the left.

cont’d

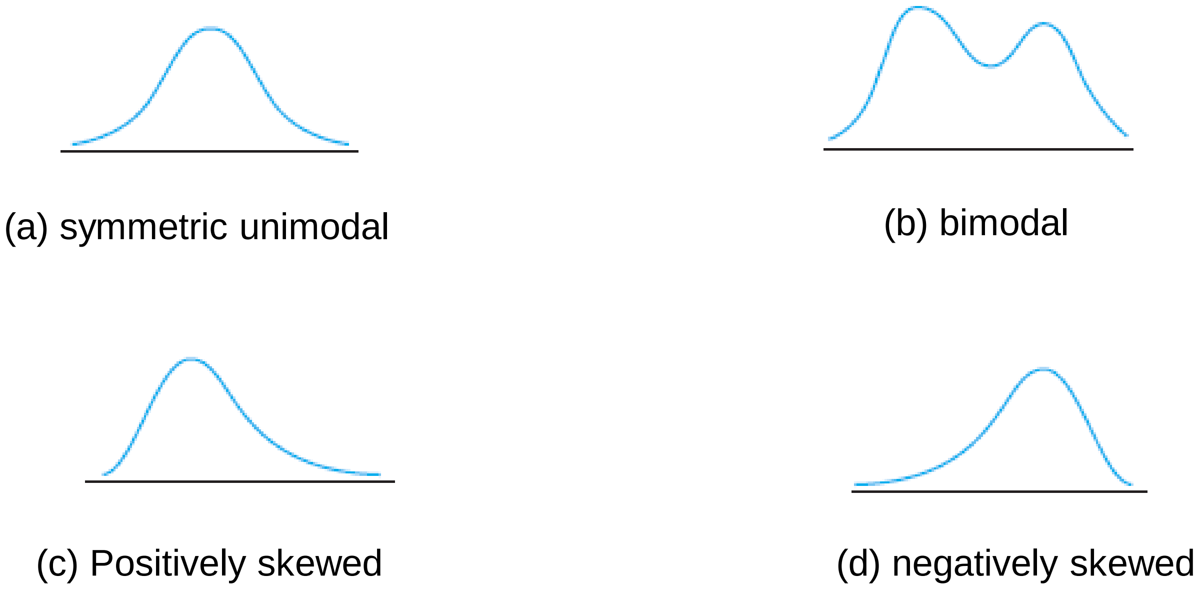
Example 1.12

- A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail and **negatively skewed** if the stretching is to the left.

cont’d

Example 1.12

- Figure 1.12 shows “smoothed” histograms, obtained by superimposing a smooth curve on the rectangles, that illustrate the various possibilities.



(a) symmetric unimodal

(b) bimodal

(c) Positively skewed

(d) negatively skewed

Smoothed histograms

Figure 1.12

Qualitative Data

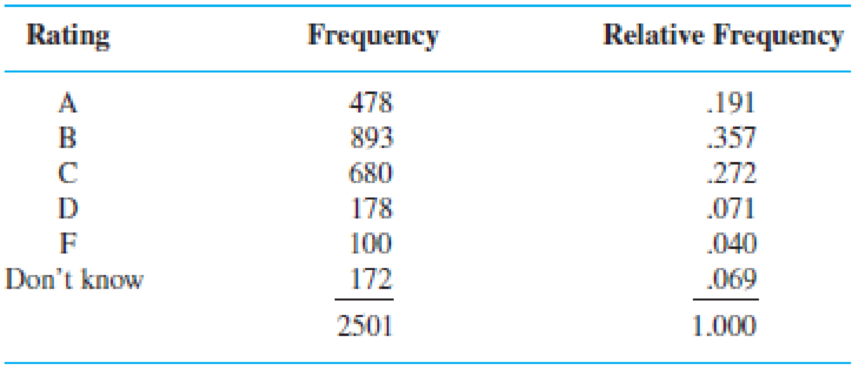
- Bar chart can be constructed when the data set is *qualitative* (categorical) in nature.
- In some cases, there will be a natural ordering of classes—for example, freshmen, sophomores, juniors, seniors, graduate students—whereas in other cases the order will be arbitrary—for example, Catholic, Jewish, Protestant, and the like.
- With such categorical data, the intervals above which rectangles are constructed should have equal width.

Example 1.13

- The Public Policy Institute of California carried out a telephone survey of 2501 California adult residents during April 2006 to ascertain how they felt about various aspects of K-12 public education. One question asked was “Overall, how would you rate the quality of public schools in your neighborhood today?”

Example 1.13

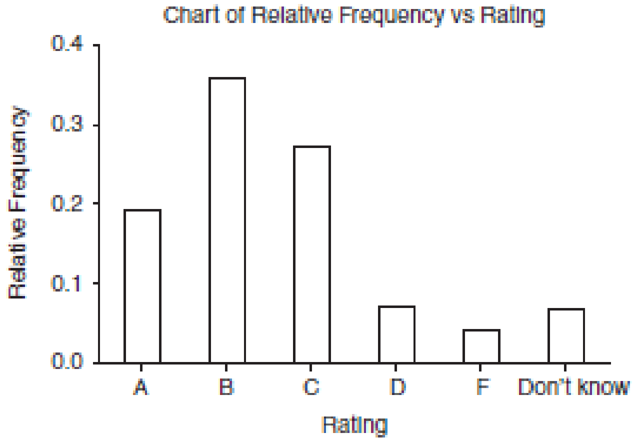
- Table 1.2 displays the frequencies and relative frequencies, and Figure 1.13 shows the corresponding bar chart.



Rating	Frequency	Relative Frequency
A	478	.191
B	893	.357
C	680	.272
D	178	.071
F	100	.040
Don't know	172	.069
	2501	1.000

Frequency Distribution for the School Rating Data

Table 1.2



Histogram of the school rating data from Minitab

Figure 1.13

cont'd

Example 1.13

- More than half the respondents gave an A or B rating, and only slightly more than 10% gave a D or F rating. The percentages for parents of public school children were somewhat more favorable to schools: 24%, 40%, 24%, 6%, 4%, and 2%.

Multivariate Data

- Multivariate data is generally rather difficult to describe visually. Several methods for doing so appear later in the book, notably scatter plots for bivariate numerical data.

Principles of Graphical Excellence

- Present data in a way that provides substance, statistics and design
- Communicate complex ideas with clarity, precision and efficiency
- Give the largest number of ideas in the most efficient manner
- Excellence almost always involves several dimensions
- Tells the truth about the data

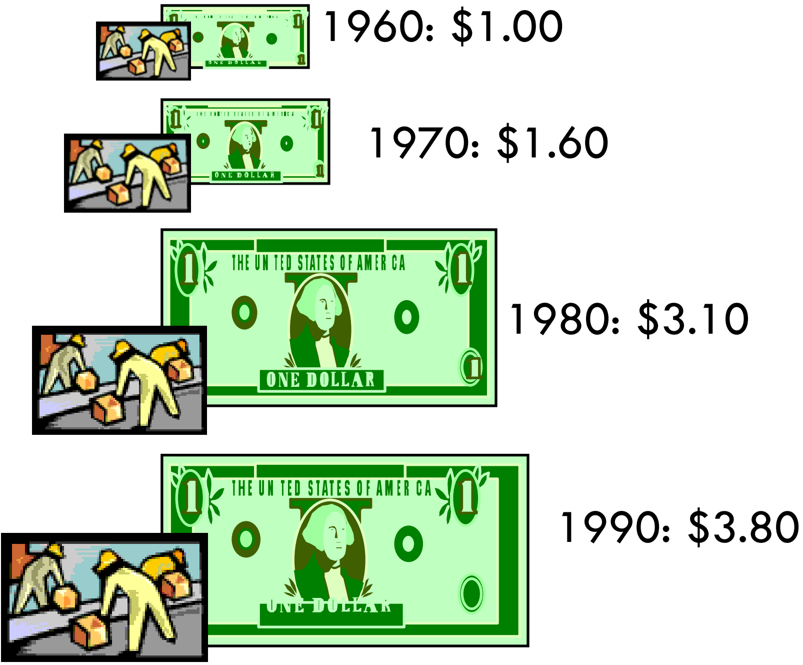
Errors in Presenting Data

- Using “chart junk”
- Failing to provide a relative basis in comparing data between groups
- Compressing or distorting the vertical axis
- Providing no zero point on the vertical axis

“Chart Junk” and “Lie Factor”

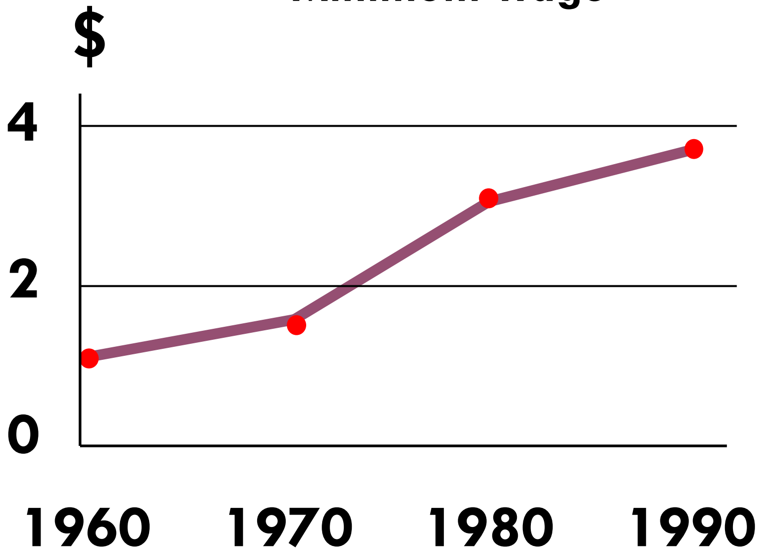
🚫 Bad Presentation

Minimum Wage



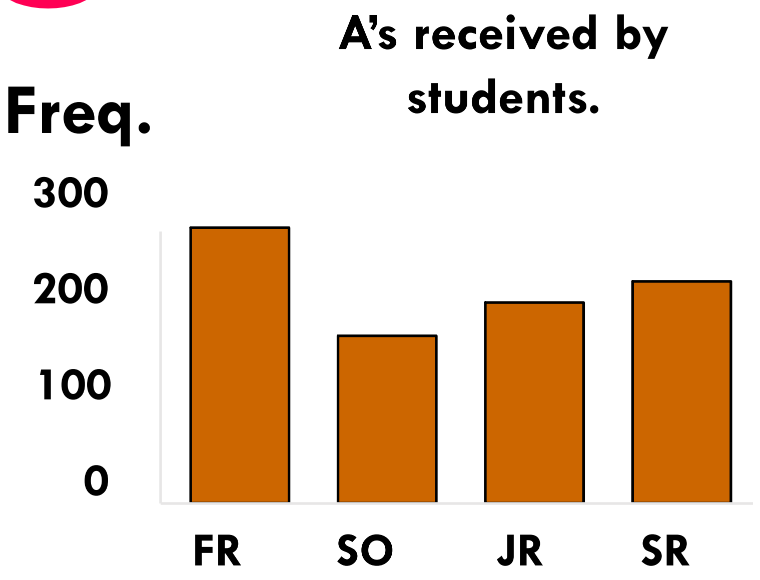
✓ Good Presentation

Minimum Wage

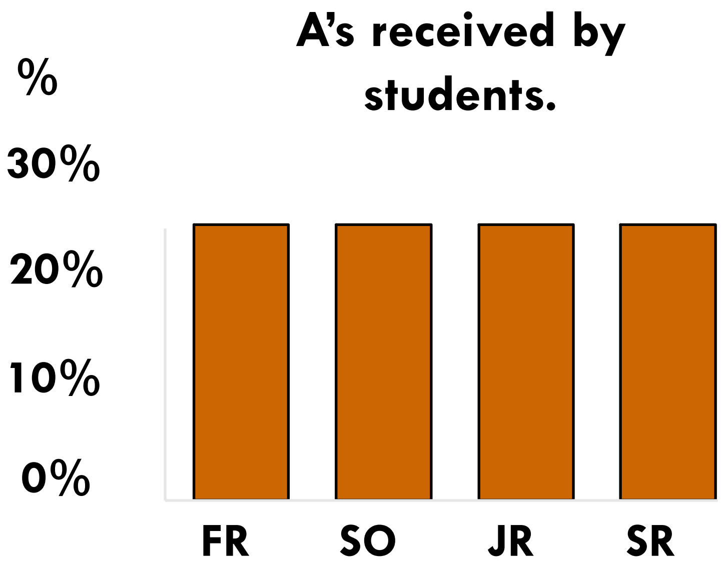


No Relative Basis

Bad Presentation


A’s received by students.

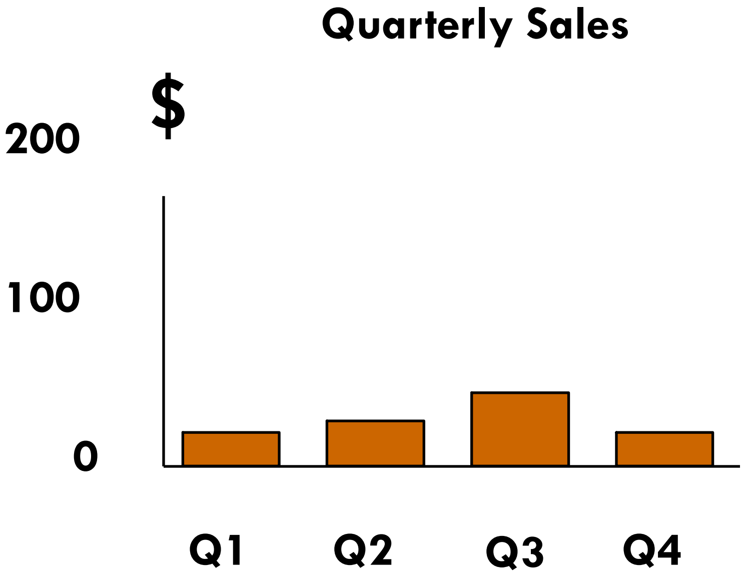
✓ Good Presentation


A’s received by students.

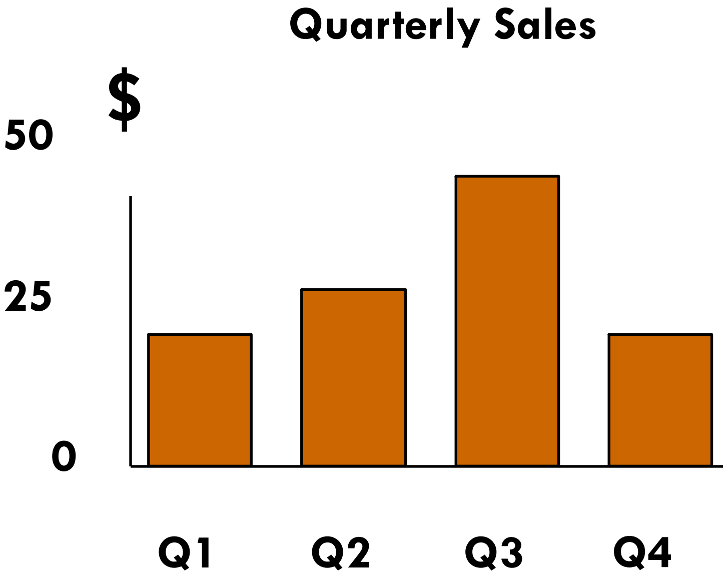
FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

Compressing Vertical Axis

Bad Presentation

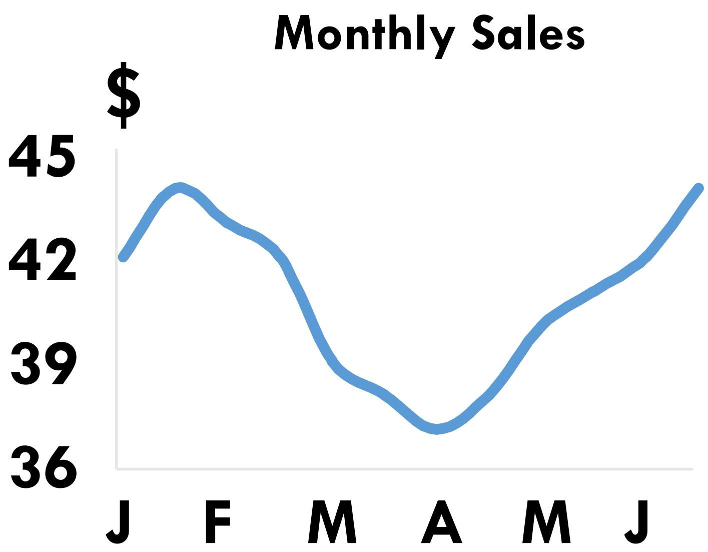


✓ Good Presentation



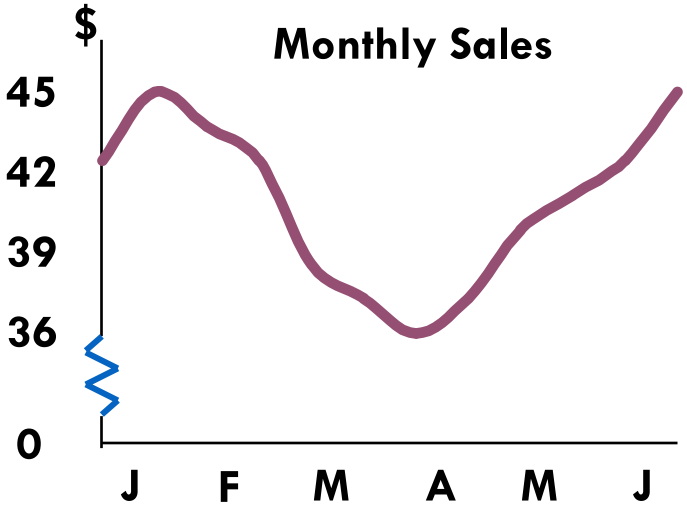
No Zero Point On Vertical Axis

Bad Presentation

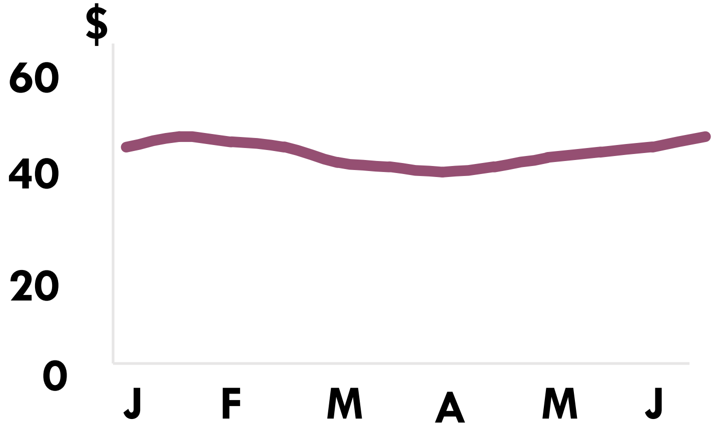


Graphing the first six months of sales

✓ Good Presentations



or



Ethical Considerations

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts

Thank You

“We trust in GOD, all others must bring data”