Lecture 01. Describing Data (Chapters 1 and 2)

Ping Yu

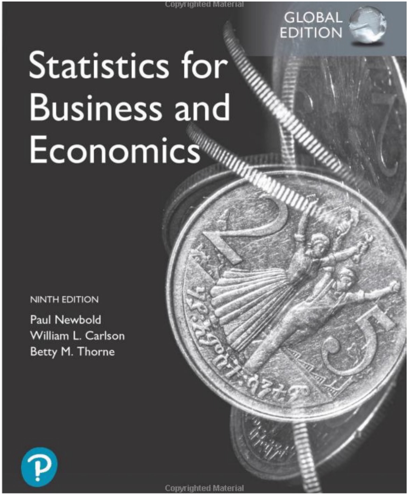
HKU Business School
The University of Hong Kong

Course Information

- Instructor: Yu, Ping
- Email: pingyu@hku.hk
- Teaching Time: 9:30am-10:45am, and 10:55am-12:10noon, Friday
- Teaching Location: Virtuall by Zoom for the first two weeks, and then f2f at LE1 (?, the lectures will be recorded and uploaded on moodle anyway)
- Office Hour: 11:00am-12:00noon, Thursday, f2f at KKL1108/Virtual by Zoom
  - Due to the recent COVID-19 situation in HK, I will provide only online OHs before the study break; please make an appointment beforehand so that I need not wait online lonely.
  - I will NOT answer questions in email if the answer is long or is not easy to explain exactly by words. Please stop by during my office hour.

- Tutor: Chou, Mike
- Email: choukp@hku.hk
- Teaching Time: TBA
- Teaching Location: TBA
- Office Hour: TBA
  - Any issues on administration (e.g., enrollment, Moodle, time clash, lab entrance, absence from the exams, etc.) and HWs (e.g., clarification of problems) should contact the tutor.

The Textbook: SBE Hereafter



- *Statistics for Business and Economics* (Global Edition, 9th edition), by Paul Newbold, William Carlson and Betty Thorne, Pearson, 2019.

Software

- We will use ℝ and RStudio as the statistical software in this course; RStudio is an Integrated Development Environment (IDE) for ℝ.

- Different from STATA or other softwares, both of them are free to download and install.
- Website for ℝ: https://www.r-project.org/
- Website for RStudio: https://www.rstudio.com/
- Wiki for ℝ: https://en.wikipedia.org/wiki/R_(programming_language)
- Wiki for RStudio: https://en.wikipedia.org/wiki/RStudio

Evaluation

- Evaluation: 4 HWs (10%), Midterm Test (40%), Final Exam (50%)
- HW: Evenly distributed over the 12 weeks. Must be typed (e.g., by LaTex or Word). ℝ commands need not be submitted. Turn in your HW on moodle on the due day (usually before midnight, 11:59pm, of some Sunday).
  - Late HWs are not acceptable for whatever reasons. To avoid any risk, start your HW early (the HWs indicate clearly which problems can be solved after each lecture; usually, one problem is assigned to each section).
- Tutorial: The answer key to the HWs and midterm would not be posted on moodle and will be discussed by the tutor. The tutorial class starts from week three (the week starting from Feb. 6). Tutorial questions will be posted on moodle one week in advance. Tutorial questions are not HWs; there is no need to turn them in.
- Examination: Mimic HWs and tutorials. Closed book and closed note. A formula sheet would be provided for the midterm and final and posted on moodle before the midterm and final. ℝ commands are not tested. No past exams are provided (due to the university policy).
  - You must take the final to pass this course; if you cannot take the midterm due to sickness, then the weight of midterm would be automatically shifted to the final.
  - Midterm: March 13, Sunday, 10:00am-12:00noon, ??.
  - Final: TBA.
- Suggestion: Preview slides before the class.

UG Econometric Courses at HKU Business School

- ECON1280. Analysis of Economic Data (both fall and spring): introduction to statistics, prerequisite of the other courses especially ECON2280.
  - You'd better enroll in ECON1280 in the fall if you want to enroll in ECON2280 in the spring, and vice versa.
- ECON2280. Introductory Econometrics (both fall and spring): linear regression.
- ECON3225. Big Data Economics (only spring): machine learning.
- ECON3283. Economic Forecasting (only spring): time series.
- ECON3284. Introduction to Causal Inference and Statistical Learning (only spring): treatment effects evaluation.

- I ever taught ECON2280, but will teach ECON1280 and ECON3225 in this semester.
- In ECON1280, I will emphasize concepts understanding and their empirical applications.
- To avoid repetition with ECON2280, I will not cover linear regression (Chapters 11-13 of SBE).
- To avoid repetition with ECON3283, I will not cover time series analysis and forecasting (Chapter 16 of SBE) and related materials in other chapters.
- I plan to cover all the other chapters of SBE (depending on whether time allows), roughly following the notations of the textbook.

Course Policy

- In Class: (i) turn off your cell phone and keep quiet; (ii) come to class and return from the break on time; (iii) you can ask me freely in class, but if your question is far out of the course or will take a long time to answer, I will answer you after class; (iv) speak English in class!
- Policy on Plagiarism: If judged as "plagiarism", you are in serious trouble. If a few students are judged to copy each other, each gets zero mark. I will not judge who copied whom. So **DO NOT** copy others and **DO NOT** be copied by others.

  - You may discuss with your classmates about the HWs, but **DO NOT** copy each other.

  - This policy applies to HW, midterm and final.
- Feedback: Any feedback to my teaching (e.g., the lecturer's English is hard to follow, technicalities are too hard to understand, the teaching should slow down, more interactions are required, there are some typos in the slides, etc.) is very welcome. I would incorporate your feedbacks in my future teaching during the semester. You can also give your feedbacks (e.g., some difficult points in the lectures) to the tutor so that the tutor can discuss them in tutorial classes.
- Guest Account (cannot receive announcements):

  - Website: http://hkuportal.hku.hk/moodle/guest

  - Guest Username: econ1280_2b_2021_guest

  - Password: ECON1280@ping

Course Outline

- Lecture 01: Describing Data (Chapters 1 and 2)
- Lecture 02: Probability (Chapter 3)
- Lecture 03: Discrete Random Variables (Chapter 4)
- Lecture 04: Continuous Random Variables (Chapter 5)
- Lecture 05: Sampling Distribution Theory (Chapter 6)

  Midterm: usually during the first week after the break and cover Lectures **1-4** (**Note**: one lecture need not be finished in one week.).

- Lecture 06: Hypothesis Testing (Chapters 9 and 10)
- Lecture 07: Confidence Interval Estimation (Chapters 7 and 8)
- Lecture 08: Analysis of Variance (Chapter 15)
- Lecture 09: Nonparametric Statistics (Chapter 14)
- Lecture 10: Sampling (Chapter 17)

  - The first eight lectures will definitely be covered, and whether or which of the remaining three are covered depends on how fast I will teach.

  - The final will concentrate on the materials that are not covered by the midterm.

- Slides indexed by (*): covered in the lecture or by the tutor, maybe related to the assignments, but not tested in the midterm or final.
- Slides indexed by (**): not covered in the lecture, only for after-class reading.
- I won't cite (in my slides) the section numbers in the textbook unless necessary.

Plan of This Lecture

- Statistics can help us process, summarize, analyze, and interpret data to make better decisions in uncertain environment (although usually loses some information of the raw data). It permits us to make sense of all the data.
- Data in raw form are usually not easy to use for decision making. I will introduce tables and graphs in the first half of this lecture to provide visual support for improved decision making, and introduce numerical measures in the second half for more rigorous analysis.
  - Pay special attentions to the differences in describing categorical and numerical variables both graphically and numerically.

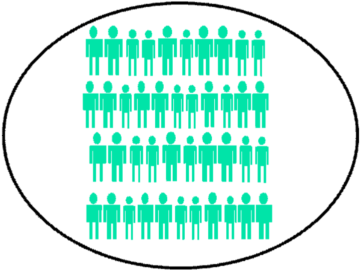
- Describing Data: Graphical
  - Decision Making in an Uncertain Environment
  - Classification of Variables
  - Graphs to Describe Categorical Variables
  - Graphs to Describe Numerical Variables
- Describing Data: Numerical
  - Measures of Central Tendency and Location
  - Measures of Variability
  - Weighted Mean and Measures of Grouped Data
  - Measures of Relationships Between Variables

Describing Data: Graphical

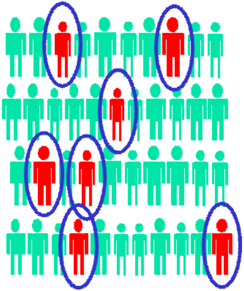
Decision Making in an Uncertain Environment

- Decisions are often made based on limited information – data (or samples).
  - This may be due to the cost constraints or time constraints.
- A population is the complete set of all items of interest. Population size, N, can be very large or even infinite.
  - e.g., all potential buyers of a new product.
  - e.g., all stocks traded on the NYSE.
- A sample is an observed subset (or portion) of a population with sample size given by n. [figure here]
- We hope the sample can represent the population, since our decision is made on the population.

Population vs. Sample



Population



Sample

Random and (**) Systematic Sampling

- (Simple) random sampling is a sampling scheme in that ① each member of the population has the same probability of being selected, ② the selection of one member is independent of the selection of any other member, and ③ every possible sample of a given size, n, has the same probability of selection.
  - Although random sampling is too ideal in practice (due to the cost issue), it serves as a benchmark for other sampling schemes discussed in Lecture 10.
- (**) Suppose that the population list is arranged in some fashion unconnected with the subject of interest (i.e., in random order). Systematic sampling involves the selection of every jth item in the population, where $j = N/n$, and the first item is randomly selected from 1 to j.
  - Suppose $n = 100$ samples are desired, $N = 5000$, then $j = 50$. If your first item is numbered 20, then the 20th, 70th, 120th,$\cdots$ items are sampled.
  - Systematic samples provide a good representation of the population if there is no cyclical variation in the population.

Parameter and Statistic

- A parameter is a numerical measure that describes a specific characteristic of a population.
- A statistic is a numerical measure that describes a specific characteristic of a sample.
- In other words, a parameter is a function of the population and a statistic is a function of the sample.
- Descriptive statistics focus on graphical and numerical procedures that are used to summarize and process data.
- Inferential statistics focus on using the data to make predictions, forecasts, and estimates to make better decisions.
- Usually, descriptive statistics are elementary and intuitive, and inferential statistics are more advanced and more powerful.

Sampling and Nonsampling Errors

- The target of statistics is to make decisions on a population parameter based on a sample statistic.
- Because $n < N$, there must be some uncertainty in the decision making based on the statistic (about the parameter); the resulting error is called sampling error.
- Even the whole population were collected, there are still some errors called nonsampling error.

  - The population actually sampled is not the relevant one, e.g., the voting opinion on Franklin Roosevelt.

  - Survey subjects may give inaccurate or dishonest answers, e.g., the voting opinion on Donald Trump.

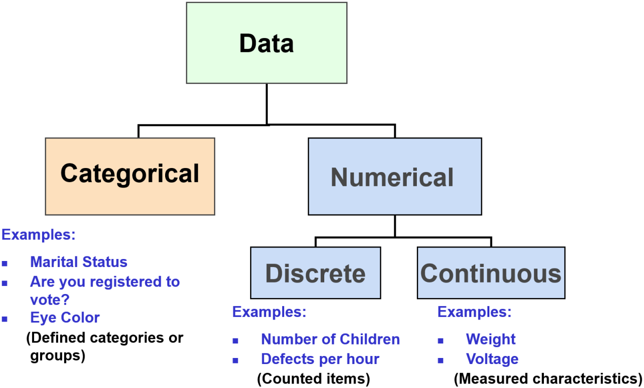
  - There may be no response to survey questions, e.g., income level of the rich.
- Read the textbook (Page 28) for other examples of nonsampling errors, but we will focus on sampling errors in this course.

Classification of Variables: Categorical and Numerical Variables

- A variable is a specific characteristic (such as age or weight) of an individual or object.
  - A variable is any property or descriptor that can take multiple values.
  - A variable can be though of as a question, to which the value is the answer. E.g., "How old are you?", "42 years old". Here, "age" is the variable, and "42" is its value.
- Based on the type and amount of information contained in the data, we classify variables into categorical and numerical variables.
- Based on the levels of measurement, we classify variables into qualitative and quantitative variables.
- Categorical variables produce responses that belong to groups or categories.
  - e.g., responses to yes/no questions.
  - e.g., choices from "strongly disagree" to "strongly agree".
- Numerical variables includes both discrete and continuous variables.

Discrete and Continuous Variables

- A discrete numerical variable may (but does not necessarily) have a finite number of values.
  - The most common type of discrete variable produces a response that comes from a counting process, i.e., takes values from infinite numbers, $0, 1, 2, 3, \cdots$, e.g., the number of customers.
- A continuous numerical variable may take on any value within a given range of real numbers.
  - The continuous variable usually arises from a measurement (not a counting) process, e.g., the salary of a worker.
  - In daily life, we tend to truncate continuous variables as if they were discrete ones due to the precision of measurement instruments or convenience.



Data

Categorical

Examples:

- Marital Status
- Are you registered to vote?
- Eye Color

(Defined categories or groups)

Numerical

Discrete

Examples:

- Number of Children
- Defects per hour

(Counted items)

Continuous

Examples:

- Weight
- Voltage

(Measured characteristics)

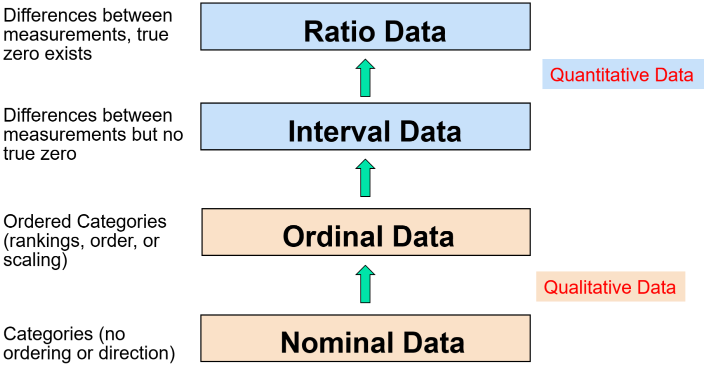
Qualitative and Quantitative Variables

- Qualitative data do not assign measurable meaning to the "difference" in numbers.
  - e.g., the numbers assigned to the football players – number 10 does not play twice as number 5.
- Quantitative data assign measurable meaning to the "difference" in numbers.
  - e.g., the exam score 90 is twice of 45.
- Qualitative data include nominal data and ordinal data.
  - Nominal data are considered the lowest or weakest type of data, e.g., gender, country of citizenship, phone number, etc., where numerical identification is chosen only for convenience and does not imply ranking of responses.
  - Ordinal data indicate the rank of ordering, e.g., product quality rating (1: poor; 2: average; 3: good), but the difference in numbers is meaningless.

continue

- Quantitative data include interval data and ratio data.
  - Interval data indicate rank and distance from an arbitrarily determined benchmark or zero, e.g., Celsius and Fahrenheit degrees of temperature or the year based on the Gregorian calendar, where the difference makes sense but ratio is meaningless.
  - Ratio data indicate both rank and distance from a natural zero, e.g., age and weight, where the ratios of two measures have meaning.
- From nominal, to ordinal, to interval, and to ratio, more and more information is contained in the data.

Measurement Levels



Differences between measurements, true zero exists

Ratio Data

Quantitative Data

Differences between measurements but no true zero

Interval Data

Ordered Categories (rankings, order, or scaling)

Ordinal Data

Qualitative Data

Categories (no ordering or direction)

Nominal Data

Graphs to Describe Categorical Variables: (Relative) Frequency Distributions

- A frequency distribution is a table used to organize data.

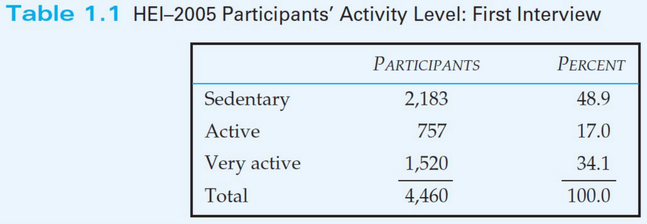


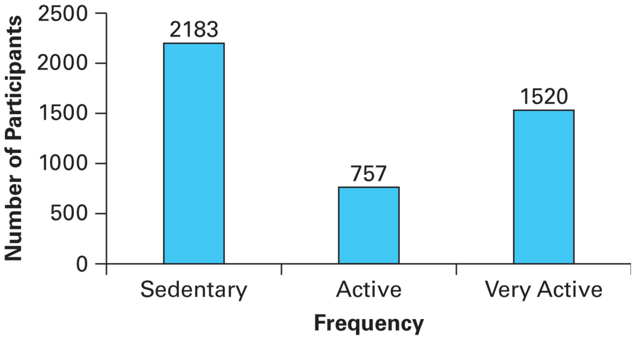
Table 1.1 HEI–2005 Participants' Activity Level: First Interview

	PARTICIPANTS	PERCENT
Sedentary	2,183	48.9
Active	757	17.0
Very active	1,520	34.1
Total	4,460	100.0

Figure: HEI: Healthy Eating Index

- The left column (called classes or groups) includes all possible responses on a variable under study.
- The right column is a list of the frequencies, or number of observations, for each class.
- A relative frequency distribution: $\frac{\text{frequency}}{n} \times 100\%$.

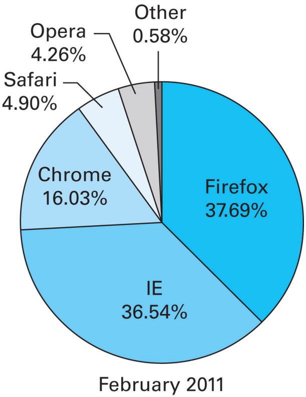
Bar Charts

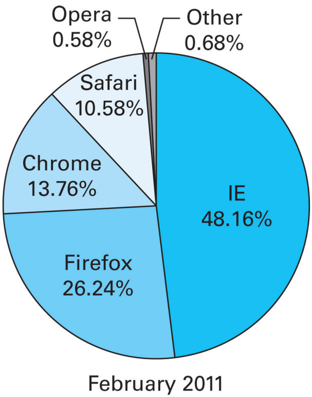


- Bar charts draw attention to the frequency itself (not proportion of frequencies) of each category.
- The height of bars represents frequency, and bars need not touch.

Pie Charts

- If the focus is the proportion of frequencies, then pie charts are appropriate.

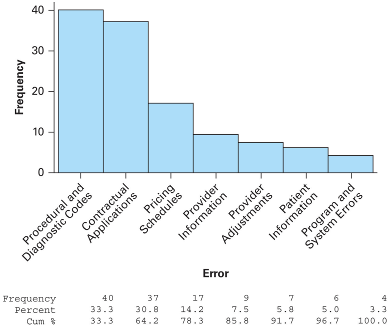




Browser Wars: European Market Share North America Market Share

Pareto Diagrams

- A Pareto diagram is a bar chart that displays the frequency of defect causes. It is used to separate the "vital few" from the "trivial many".



Frequency	40	37	17	9	7	6	4
Percent	33.3	30.8	14.2	7.5	5.8	5.0	3.3
Cum %	33.3	64.2	78.3	85.8	91.7	96.7	100.0

Copyright ©2013 Pearson Education, publishing as Prentice Hall

Figure: Errors in Health Care Claims Processing

- The bars are arranged in the descending order of frequencies.
- The first three causes contribute about 80% of errors.

The Pareto Principle or "80-20 Rule"

- The Pareto principle or "80-20 rule" states that 80% of outcomes are due to 20% of causes. [figure here]
- (**) The Pareto density function: $f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} 1(x \geq x_m)$.

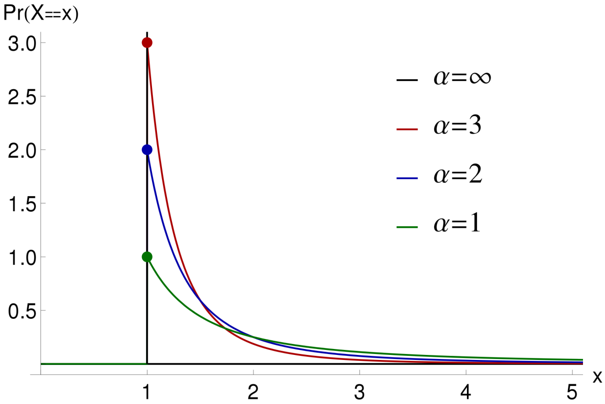
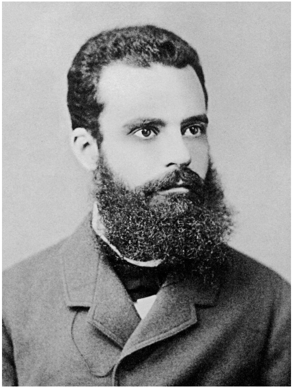


Figure: Pareto Density Functions for Various α's with $x_m = 1$

History of the Pareto Principle



Vilfredo F. D. Pareto (1848-1923),
Italian, University of Lausanne

Cross Tables

- Cross tables (or crosstabs/contingency tables) are used to describe relationships between categorical or ordinal variables.

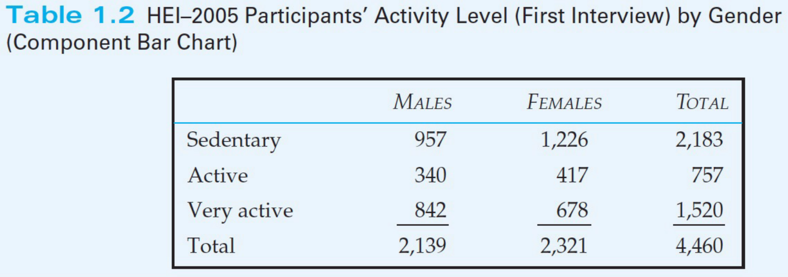


Table 1.2 HEI–2005 Participants' Activity Level (First Interview) by Gender (Component Bar Chart)

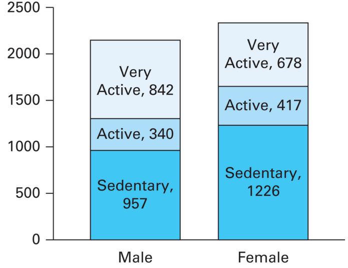
	MALES	FEMALES	TOTAL
Sedentary	957	1,226	2,183
Active	340	417	757
Very active	842	678	1,520
Total	2,139	2,321	4,460

Figure: A 3 × 2 Cross Table

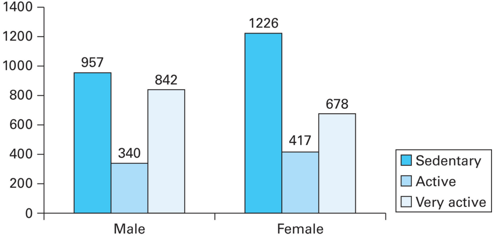
- It lists the frequencies of all combinations of values for the two variables.

Component or Cluster Bar Charts

- A component (or stacked) bar chart and cluster (or side-by-side) bar chart are used to picture the information in cross tables, and are extensions of the bar chart above.



Copyright ©2013 Pearson Education, publishing as Prentice Hall



Copyright ©2013 Pearson Education, publishing as Prentice Hall

Graphs to Describe Numerical Variables: Frequency Distributions

- Different from categorical data, we must construct the classes by ourselves.
- Three Rules:
- Rule 1: Determine k, the number of classes.
- Rule 2: Classes should be the same width, w, which is determined by

$$w = \frac{\text{Largest Observation} - \text{Smallest Observation}}{\text{Number of Classes}}.$$

  - w should always be rounded upward to include all observations in the frequency distribution table.

- Rule 3: Classes must be inclusive and nonoverlapping: each observation must belong to one and only one class.
  - Make sure the boundary values are clearly classified.

Number of Classes

Sample Size n	Number of Classes k
$n < 50$	$5 - 7$
$50 \leq n \leq 100$	$7 - 8$
$101 \leq n \leq 500$	$8 - 10$
$501 \leq n \leq 1000$	$10 - 11$
$1001 \leq n \leq 5000$	$11 - 14$
$n > 5000$	$14 - 20$

Table: Rule of Thumb in Determining k

- Intuitions are used in practice to guarantee that each class includes "not too few" or "not too many" observations.

(Relative) Cumulative Frequency Distributions

- A cumulative frequency distribution contains the total number of observations whose values are less than the upper limit for each class, which can be constructed by adding the frequencies of all classes up to and including the present class.
- A relative cumulative frequency distribution: $\frac{\text{cumulative frequency}}{n} \times 100\%$.

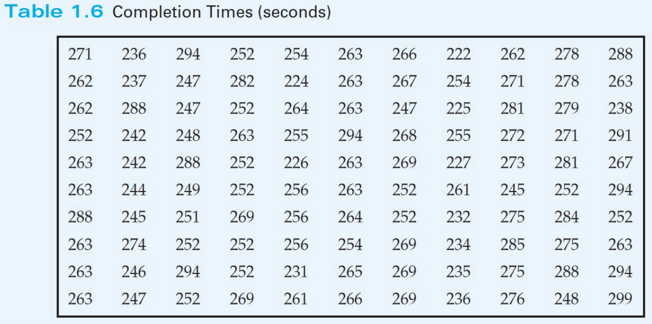


Table 1.6 Completion Times (seconds)

271	236	294	252	254	263	266	222	262	278	288
262	237	247	282	224	263	267	254	271	278	263
262	288	247	252	264	263	247	225	281	279	238
252	242	248	263	255	294	268	255	272	271	291
263	242	288	252	226	263	269	227	273	281	267
263	244	249	252	256	263	252	261	245	252	294
288	245	251	269	256	264	252	232	275	284	252
263	274	252	252	256	254	269	234	285	275	263
263	246	294	252	231	265	269	235	275	288	294
263	247	252	269	261	266	269	236	276	248	299

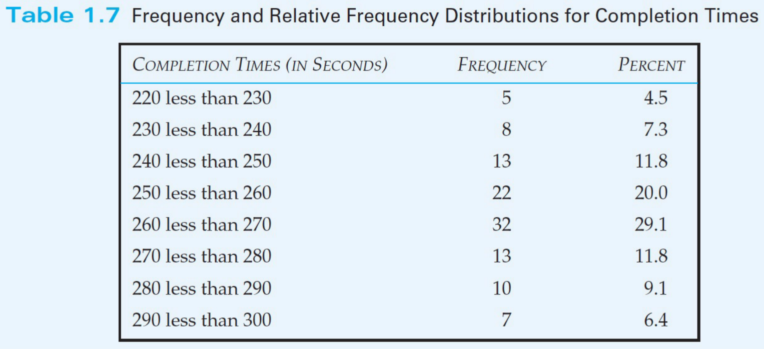


Table 1.7 Frequency and Relative Frequency Distributions for Completion Times

COMPLETION TIMES (IN SECONDS)	FREQUENCY	PERCENT
220 less than 230	5	4.5
230 less than 240	8	7.3
240 less than 250	13	11.8
250 less than 260	22	20.0
260 less than 270	32	29.1
270 less than 280	13	11.8
280 less than 290	10	9.1
290 less than 300	7	6.4

- $n = 110$, so set $k = 8$ and $w = \frac{299-222}{8} = 10$ (rounded up).

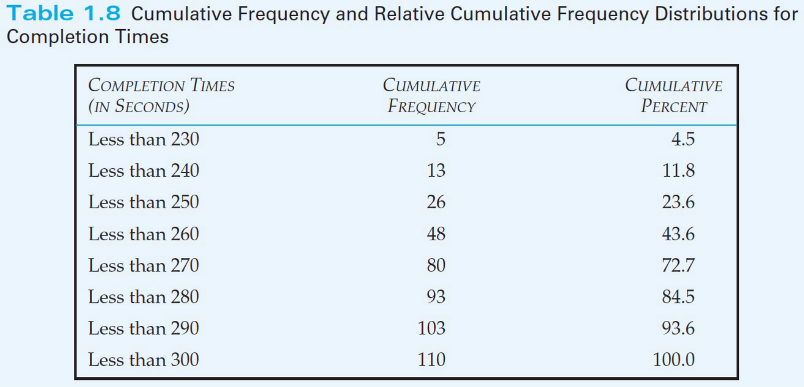


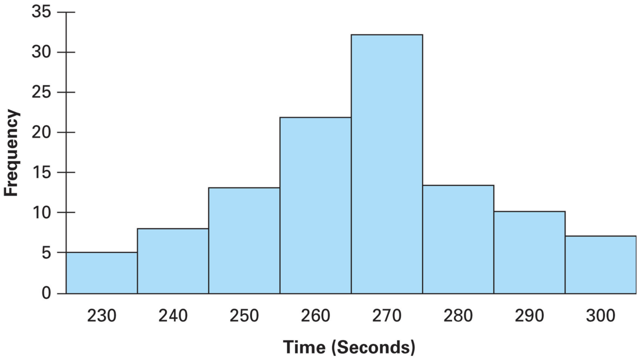
Table 1.8 Cumulative Frequency and Relative Cumulative Frequency Distributions for Completion Times

Completion Times (in Seconds)	Cumulative Frequency	Cumulative Percent
Less than 230	5	4.5
Less than 240	13	11.8
Less than 250	26	23.6
Less than 260	48	43.6
Less than 270	80	72.7
Less than 280	93	84.5
Less than 290	103	93.6
Less than 300	110	100.0

- Suppose the goal is 4.5 minutes; then we can tell from Table 1.8 that less than $3/4$ (72.7%) employees can achieve the goal.

Histograms

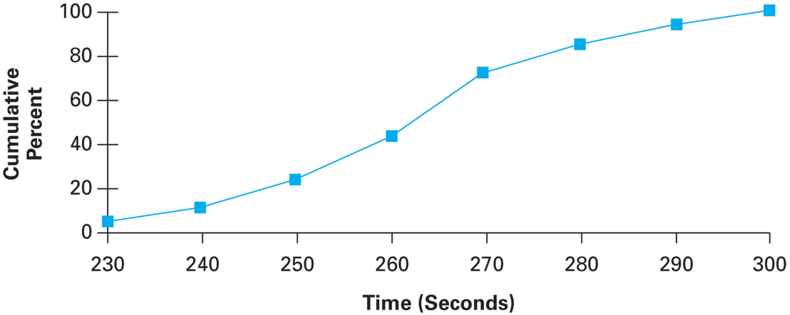
- A histogram is a counterpart of a bar chart for numerical variables.



- Read Section 1.6 for some popular mistakes in presenting histograms. These mistakes can be easily avoided by using statistical softwares properly.

Ogives

- A ogive (or cumulative line graph) is a line connecting points that are the cumulative percent of observations below the upper limit of each interval in a cumulative frequency distribution.



Shape of a Distribution: Symmetricity

- A distribution is symmetric if the observations are balanced, or approximately evenly distributed, about its center.

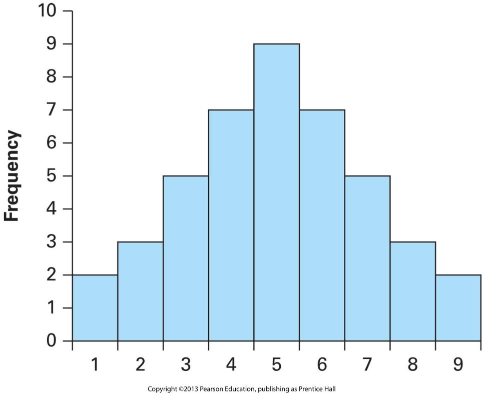
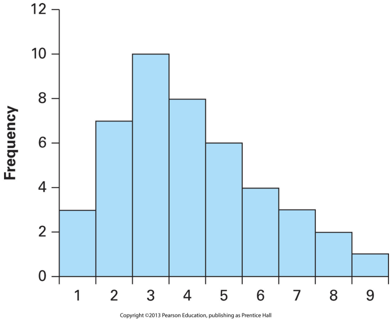


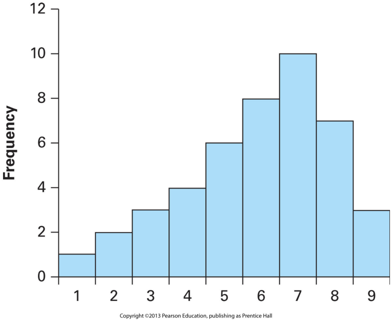
Figure: Symmetric Distribution

Shape of a Distribution: Skewedness

- A distribution is skewed (or asymmetric) if the observations are not symmetrically distributed on either side of the center.
  - A skewed-right (or positively skewed) distribution has a tail that extends farther to the right, e.g., the income distribution.
  - A skewed-left (or negatively skewed) distribution has a tail that extends farther to the left, e.g., the GPAs in Example 1.10.



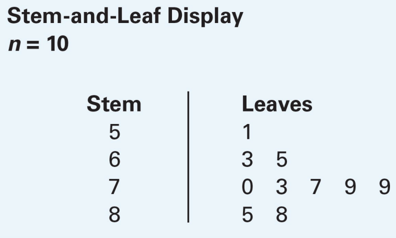
Skewed-Right Distribution



Skewed-Left Distribution

(**) Stem-and-Leaf Displays

- A stem-and-leaf display is an exploratory data analysis (EDA) graph that is an alternative to the histogram.
  - Data are grouped according to their leading digits (called stems), and the final digits (called leaves) are listed separately for each member of a class.
  - The leaves are displayed individually in ascending order after each of the stems.
- Accounting Final-Exam Grades: 88, 51, 63, 85, 79, 65, 79, 70, 73, 77



Stem-and-Leaf Display
***n* = 10**

Stem	Leaves
5	1
6	3 5
7	0 3 7 9 9
8	5 8

Copyright ©2013 Pearson Education, publishing as Prentice Hall

- The stem-and-leaf display is a quick way to identify possible patterns for a small data set.

Scatter Plots

- A scatter plot locates one point for each observation of two variables. It can provide a picture of the data, including (i) the range of each variable, (ii) the pattern of values over the range, (iii) a suggestion as to a possible relationship between the two variables, and (iv) an indication of outliers (extreme points, i.e., data values that are much larger or smaller than other values).

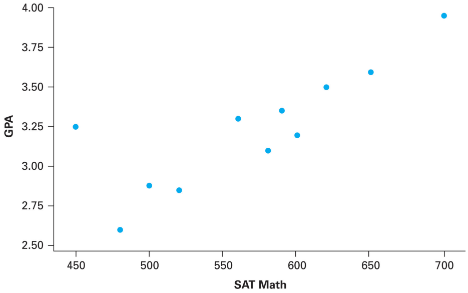


Figure: GPA at College Graduation vs. Entrance SAT Math Scores

Summary of Techniques

	Categorical Variables		Numerical Variables	
	One Variable	Two Variables	One Variable	Two Variables
Tables	Frequency distribution	Cross Table	(Cumulative) Frequency distribution	
Graphs	Bar chart	Component Bar Chart Cluster Bar Chart	Histogram	
	Pie chart Pareto diagram		(Ogive) Stem-and-leaf display	Scatter plot

Describing Data: Numerical

Measures of Central Tendency: Mean, Median, and Mode

- Measures of central tendency provide information about a "typical" observation in the data.
- They are usually computed from sample data rather than from population data.
- The (arithmetic) mean of a set of data is the sum of the data values divided by the number of observations.

  - The population mean is a parameter given by

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

  - The sample mean is a statistic given by

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

  - The mean is appropriate for numerical data.

continue

- The median is the middle observation of a set of observations that are arranged in increasing (or decreasing) order.
  - If n is odd, the median is the middle observation.
  - If n is even, the median is the average of the two middle observations.
  - The median will be the number located in the $0.5(n+1)$th ordered position.
  - The median is more robust to outliers than the mean. [why?]
- The mode, if one exists, is the most frequently occurring value.
  - A distribution with one mode is called unimodal; with two (local) modes, it is called bimodal; and with more than two (local) modes, it is said to be multimodal.
  - The mode is most commonly used with categorical data. [see more discussions below]
- The most appropriate measure of central tendency is context specific.
  - e.g., for clothing retailers, the mode is more informative than the mean for inventory decisions. [why?]
- For categorical data, median and mode are appropriate, but mean is not.
  - e.g., what is the mean of "male" (coded 1) and "female" (coded 0)?
- For numerical data (the most popular data type in business applications), mean and median (esp. outliers exist) are more appropriate than the mode (maybe each value occurs only once, which one is the center?).

Example 2.1: Demand for Bottled Water

- The number of bottled water sold in $n = 12$ hours at one store during hurricane season is

$$60, 84, 65, 67, 75, 72, 80, 85, 63, 82, 70, 75.$$

- The mean is

$$\bar{x} = \frac{60 + 84 + 65 + 67 + 75 + 72 + 80 + 85 + 63 + 82 + 70 + 75}{12} = 73.17.$$

- Arrange the sales from least to greatest:

$$60, 63, 65, 67, 70, 72, 75, 75, 80, 82, 84, 85,$$

so the median is

$$x_{.5} = \frac{72 + 75}{2} = 73.5.$$

- The mode is clearly 75 bottles.

Percentiles and Quartiles

- Percentiles and quartiles are measures that indicate the location, or position, of a value relative to the entire set of data.
- They are generally used to describe large data sets, e.g., sales data, survey data, or even the weights of newborn babies.
- Arranging the data in order from the smallest to the largest, the pth percentile is a value such that approximately p% of the observations are at or below that number.
  - Percentiles separate large ordered data sets into 100ths.
  - The 50th percentile is the median.
  - pth percentile = value located in the $\frac{p}{100}(n+1)$th order position.
- Quartiles are descriptive measures that separate large data sets into four quarters. The first quartile, Q_1, (or 25th percentile) separates approximately the smallest 25% of the data from the remainder of the data. The second quartile, Q_2, (or 50th percentile) is the median. The third quartile, Q_3, (or 75th percentile) separates approximately the smallest 75% of the data from the remainder of the data.
  - $Q_1 =$ the value in the $0.25(n+1)$th ordered position.
  - $Q_2 =$ the value in the $0.50(n+1)$th ordered position.
  - $Q_3 =$ the value in the $0.75(n+1)$th ordered position.

Five-Number Summary

- The five-number summary: minimum, first quartile, median, third quartile, and maximum, in ascending order.
- Example 2.5: Demand for Bottled Water Ascendingly ordered sales:

$$60, 63, 65, 67, 70, 72, 75, 75, 80, 82, 84, 85,$$

- We use this sample for illustration although n is small here.
- $Q_1 =$ the value located in the $0.25\,(12+1) = 3.25$th ordered position.
- The value in the third position is 65 and in the fourth position is 67, so

$$Q_1 = 65 + 0.25\,(67 - 65) = 65.5.$$

- $Q_3 =$ the value located in the $0.75\,(12+1) = 9.75$th ordered position.
- The value in the 9th position is 80 and in the 10th position is 82, so

$$Q_3 = 80 + 0.75\,(82 - 80) = 81.5.$$

- The five-number summary is

$$60 < 65.5 < 73.5 < 81.5 < 85.$$

Measures of Variability: Range and Interquartile Range

- Two data sets can have the same mean but the observations in one set could vary more from the mean than do those in the other set:

  Sample A: 1, 2, 1, 36,
  Sample B: 8, 9, 10, 13,

  both of which have mean 10, but the spread of Sample A is obviously larger than that of Sample B.
  - Intuition: gunfire.
- The range is the difference between the largest and smallest observations.
  - The greater the spread of the data from the center of the distribution, the larger the range will be.
  - Although the range measures the total spread, it is sensitive to outliers.
  - One solution is to discard a few of the highest and a few of the lowest numbers, such as the IQR below.
- The interquartile range (IQR) measures the spread in the middle 50% of the data:

$$\text{IQR} = Q_3 - Q_1.$$

Box-and-Whisker Plots

- A box-and-whisker plot is a graph that describes the shape of the distribution in terms of the five-number summary. The inner box shows the numbers that span the range from the first to the third quartile. A line is drawn through the box at the median. There are two "whiskers": one from the 25th percentile to the minimum, and the other from the 75th percentile to the maximum.
- Read from the following Figure 2.3 from the textbook about the median and spread.
- In the boxplots of $\mathbb{R}$, the upper whisker ends at

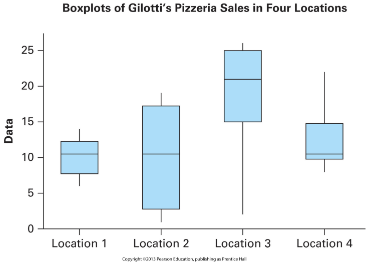
$$\min\left(Q_3 + 1.5 \times \text{IQR}, \text{maximum}\right),$$

the lower ends at

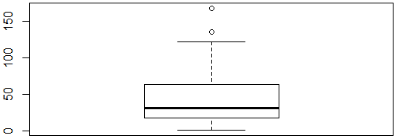
$$\max\left(Q_1 - 1.5 \times \text{IQR}, \text{minimum}\right),$$

and the points beyond that are marked as extreme points. [figure here]

- Both the stem-and-leaf display and the box-and-whisker plot were invented by John Tukey. [figure here]



box-and-whisker plots



a boxplot in $\mathbb{R}$

- (**) For a standard normal distribution, $Q_3 = 0.6745$, so

$$P\left(X > 2 \times \text{IQR}\right) = P\left(X > 4 \times 0.6745\right) \approx 0.35\%.$$

History of Exploratory Data Analysis (EDA)

John W. Tukey (1915-2000), Princeton[1]

[1] He ever coined two popular terms in computer science: bit (binary digit) and software.

Variance and Standard Deviation

- Both range and IQR use only two of the data values. Variance uses the distances of all observations from the mean.
- The population variance is the sum of the squared differences between each observation and the population mean divided by the population size:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \overset{[\text{Exercise}]}{=} \frac{\sum_{i=1}^N x_i^2}{N} - \mu^2. \qquad (1)$$

- Population A: $\{-2, 2\}$ and Population B: $\{-1, 1\}$:

$$\sigma_A^2 = \frac{(-2)^2 + 2^2}{2} = 4 > 1 = \frac{(-1)^2 + 1^2}{2} = \sigma_B^2$$

  matches the intuition that Population A is more spreading (or more risky) than Population B, where $\mu_A = \mu_B = 0$.
- The sample variance is the sum of the squared differences between each observation and the sample mean divided by the sample size minus one:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \overset{[\text{Exercise}]}{=} \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}.$$

  - The reason for dividing by $n-1$ rather than n will be explained in Lecture 5.

continue

- The population standard deviation, σ, is the (positive) square root of the population variance:
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}.$$
- The sample standard deviation, s, is the (positive) square root of the sample variance:
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$
- The name "standard deviation" came from Karl Pearson [figure below].
  - Variance measures the average squared "deviation" from the mean and has the unit of the squared unit of x_i.
  - By taking $\sqrt{\cdot}$, we get back to the "standard" (original) unit of x_i.
- (**) In the definition of s^2, why use $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ rather than $\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{n-1}$ or $\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n-1}$?

Example 2.9: Gilotti's Pizzeria Sales At Location 1

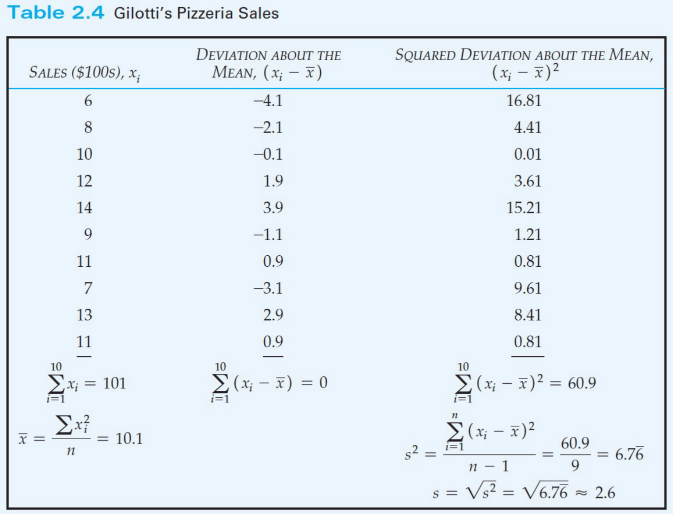


Table 2.4 Gilotti's Pizzeria Sales

SALES ($100s), x_i	DEVIATION ABOUT THE MEAN, $(x_i - \bar{x})$	SQUARED DEVIATION ABOUT THE MEAN, $(x_i - \bar{x})^2$
6	–4.1	16.81
8	–2.1	4.41
10	–0.1	0.01
12	1.9	3.61
14	3.9	15.21
9	–1.1	1.21
11	0.9	0.81
7	–3.1	9.61
13	2.9	8.41
11	0.9	0.81
$\sum_{i=1}^{10} x_i = 101$	$\sum_{i=1}^{10} (x_i - \bar{x}) = 0$	$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 60.9$
$\bar{x} = \frac{\sum x_i^2}{n} = 10.1$		$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} = \frac{60.9}{9} = 6.7\overline{6}$
		$s = \sqrt{s^2} = \sqrt{6.7\overline{6}} \approx 2.6$

- Typo: $\bar{x} = \frac{\sum x_i}{n}$.

Coefficient of Variation

- The coefficient of variation (CV), is a measure of relative dispersion that expresses the standard deviation as a percentage of the mean (provided the mean is positive).
  - The population CV is
  $$\text{CV} = \frac{\sigma}{\mu} \times 100\%.$$
  - The sample CV is
  $$\text{CV} = \frac{s}{\bar{x}} \times 100\%.$$
- When the means of two objects are different, it is better to compare them using CV rather than σ^2 or s^2.
  - e.g., a large store can be treated as a sum of many small stores, so both s^2 and $\bar{x}$ of its sales are larger than those of a small store.

Chebyshev's Theorem

- Chebyshev's Theorem: For any population with mean μ, standard deviation σ, and $k > 1$, the percent of observations that lie within the interval $[\mu \pm k\sigma]$ is at least $100\left[1 - \left(1/k^2\right)\right]\%$, where k is the number of σ. [figure here]
- (**) Why? For a random variable X, $P\left(|X-\mu| \le k\sigma\right) = 1 - P\left(|X-\mu|^2 > k^2\sigma^2\right)$, while

$$E\left[|X-\mu|^2\right] \ge E\left[|X-\mu|^2 1\left(|X-\mu|^2 > k^2\sigma^2\right)\right] \ge k^2\sigma^2 P\left(|X-\mu|^2 > k^2\sigma^2\right),$$

so

$$P\left(|X-\mu| \le k\sigma\right) \ge 1 - \frac{E\left[|X-\mu|^2\right]}{k^2\sigma^2} = 1 - \frac{\sigma^2}{k^2\sigma^2} = 1 - \frac{1}{k^2}. \quad \square$$

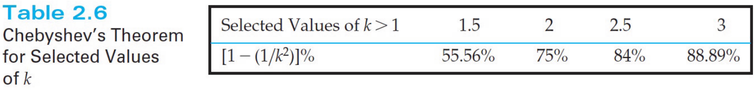
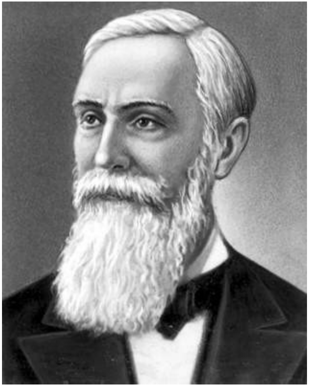


Table 2.6 Chebyshev's Theorem for Selected Values of k

Selected Values of $k > 1$	1.5	2	2.5	3
$[1-(1/k^2)]\%$	55.56%	75%	84%	88.89%

- Chebyshev's theorem can be applied to any distribution, but it is often too conservative. [check $k = 1$ for the extreme case]

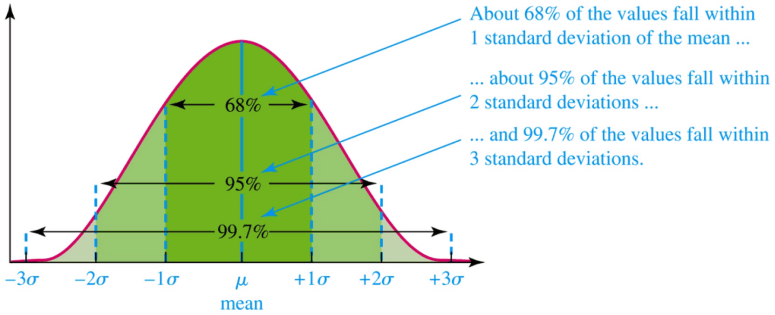
History of Chebyshev's Theorem



Pafnuty L. Chebyshev (1821-1894),
St. Petersburg University

Empirical Rule

- An empirical rule, called the 68-95-99.7 rule, gives more precise guidelines for the percentage of data values that lie within 1, 2, and 3 standard deviations (σ) of the mean (μ) for many large populations (mounded, bell-shaped).



About 68% of the values fall within 1 standard deviation of the mean ...

... about 95% of the values fall within 2 standard deviations ...

... and 99.7% of the values fall within 3 standard deviations.

- This empirical rule actually applies to the normal distribution which will be discussed in Lecture 4.

z-Score

- Percentiles and quartiles are measures that indicate the location or position of a value relative to the entire set of data, while a z-score measures the location or position of a value relative to the mean of the distribution: it is a standardized value that indicates the number of standard deviations a value is from the mean.
- For the population, the z-score of each value x_i is

$$z_i = \frac{x_i - \mu}{\sigma},$$

  which is positive if $x_i > \mu$, negative if $x_i < \mu$, and zero if $x_i = \mu$.
- For the sample, the z-score of each value x_i is

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Shape of a Distribution

- Skewness is defined as

$$\text{skewness} = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{s^3}.$$

  - The numerator is the key, and the denominator serves the purpose of standardization (free of units of x_i).

- Skewness is positive if a distribution is skewed to the right, negative if skewed to the left, and zero if bell-shaped that are mounded and symmetric about its mean [why? refere to the figures in slides 38 and 50].

- For continuous numerical unimodal data, the mean is usually less than the median in a skewed-left distribution, and vice versa.
  - e.g., the distribution of income is usually right skewed, so the median is more appropriate than the mean since the latter is too optimatic to the economic well-being of the community.
  - For a symmetric distribution, the mean and median are equal, but the converse is not true.
  - Mean is more popular than median in practice because the former is more straightforward and better understood than the latter.

Weighted Mean

- The weighted mean of a set of data is

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

where w_i is the weight of the ith observation, and $n = \sum_{i=1}^{n} w_i$.
- This concept would be used in Lectures 6 and 9.

- Example 2.17: Stock Recommendation:

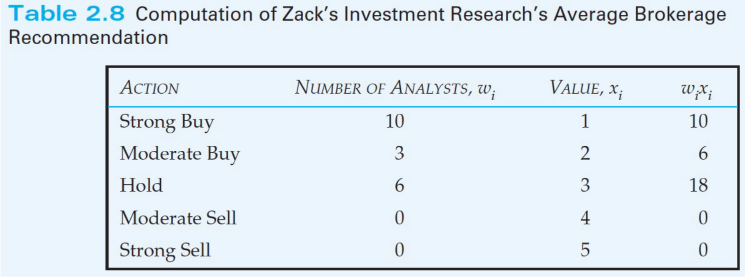


Table 2.8 Computation of Zack's Investment Research's Average Brokerage Recommendation

Action	Number of Analysts, w_i	Value, x_i	$w_i x_i$
Strong Buy	10	1	10
Moderate Buy	3	2	6
Hold	6	3	18
Moderate Sell	0	4	0
Strong Sell	0	5	0

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{n} = \frac{10+6+18+0+0}{19} = 1.79.$$

Measures of Grouped Data

- If the data are intervals rather than specific values, e.g., age intervals, wage intervals, etc., then we cannot calculate the exact mean and variance, but we can approximate them.
- Suppose that data are grouped into K classes, with frequencies $f_1, f_2, \cdots, f_K$. If the midpoints of these classes are $m_1, m_2, \cdots, m_K$, then the sample mean and sample variance can be approximated as

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^K f_i m_i}{n}, \\ s^2 &= \frac{\sum_{i=1}^K f_i \left(m_i - \bar{x}\right)^2}{n-1}, \end{aligned}$$

where $n = \sum_{i=1}^K f_i$.
- These concepts would be used in Lecture 9.

Measures of Relationships Between Variables: Covariance

- Covariance and correlation are numerical measures of the linear relationship between two variables as intuitively indicated in a scatter plot. A positive value indicates a direct or increasing linear relationship, and a negative value indicates a decreasing linear relationship.
- A population covariance is

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}.$$ [figure here]

- A sample covariance is

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

- It is easy to check that for any constants a_1, b_1, a_2 and b_2,

$$Cov(a_1 + b_1 x, a_2 + b_2 y) = b_1 b_2 Cov(x,y).$$ [Exercise]

- From this property, the covariance depends on units of measurement (i.e., not invariant to the scaling of x and y); its unit is the product of the units of x and y. In other words, it measures the direction, but not strength, of the linear relationship between x and y.

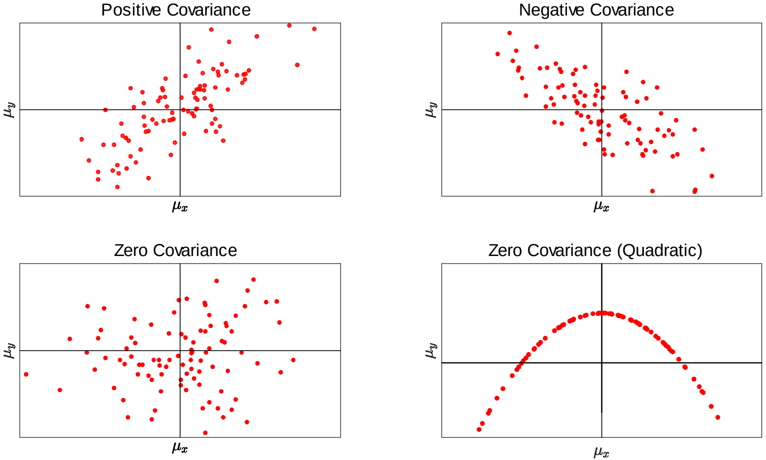


Figure: Positive, Negative an Zero Covariance

Measures of Relationships Between Variables: Correlation

- The correlation (coefficient), also called Pearson's product-moment correlation coefficient or Pearson's r, was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s. It gives a standardized measure of the linear relationship between two variables. [figure here]
  - It is more useful than covariance because it is free of units and provides both the direction and strength of a relationship.
- The correlation is computed by dividing the covariance by the product of the standard deviations of the two variables.
- A population correlation is

$$Corr\left(x,y\right)=\rho_{xy}=\frac{Cov\left(x,y\right)}{\sigma_x\sigma_y}.$$

- A sample correlation is

$$r_{xy}=\frac{s_{xy}}{s_x s_y}.$$

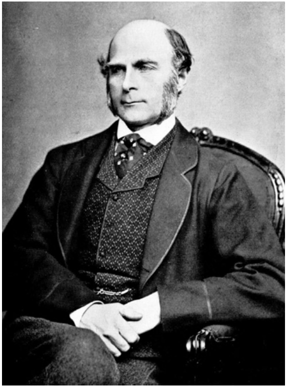
- A useful rule to determine a relationship exists is

$$\left|r_{xy}\right|\geq\frac{2}{\sqrt{n}},$$

which will be explained in Lecture 8.

History of Correlation

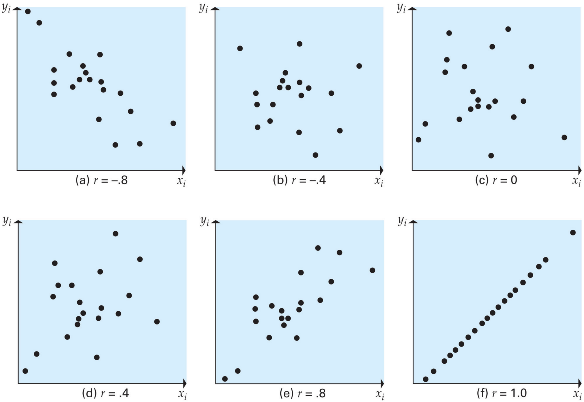
Karl Pearson (1857-1936), UCL



Sir Francis Galton (1822-1911), English[2]

[2] Galton was Charles Darwin (1809-1882)'s half-cousin, sharing the common grandparent. He was also the advisor of Karl Pearson, (South West) African explorer, and inventor of fingerprinting.

Figure 2.4: Scatter Plots and Correlation



- Because σ_x and σ_y are positive, σ_{xy} and ρ_{xy} always have the same sign, and $\rho_{xy} = 0$ if and only if (iff) $\sigma_{xy} = 0$. This is also true for r_{xy}.
- Both ρ_{xy} and $r_{xy} \in [-1, 1]$ [proof not required]. What does $r_{xy} = \pm 1$ mean?[3]

[3] This is why we know covariance measures the linear relationship between x and y.

Correlation Does NOT Imply Causation: Polio and Ice-cream

- "By 1910, frequent epidemics became regular events throughout the developed world, primarily in cities during the summer months. At its peak in the 1940s and 1950s, polio would paralyze or kill over half a million people worldwide every year."
  - From Wiki

- Folk legends: (i) A pretty woman causes death? (inauspicious or unlucky?)
  (ii) Old age causes death of my children? (bad omen or natural phenomenon?)
- Donald Trump: More tests, more infections.
- Warmonger: (larger) war induces (longer) peace.

Summary of Measures

Central Tendency	Location	Variation	Shape	Relationship
Mean	Minimum	Range	Skewness	Covariance
Median	Maximum	Interquartile Range		Correlation
Mode	Percentiles	Variance		
	Quartiles	Standard Deviation		
	z-Score	Coefficient of Variation		