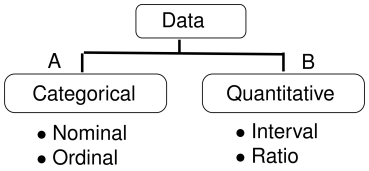
CH 1: Data and Statistics

1. Key Definitions:
   a. Elements: The entities on which data are collected.
   b. Population (universe): The set of all elements of interest in a particular study. (size N)
   c. Sample: A subset of the population selected for analysis. (size n)
   EX 1: ID the population and the sample for the Gallup Poll (estimation of the percentage of popular vote for each Candidate based on interviews of 1500 adults)

   d. Census: A survey to collect data on the entire population.
   e. Sample survey: A survey to collect data on a sample.
   f. Variable: A characteristic of interest for the elements.

2. Scales of Measurement
   a. Nominal: classification data (e.g. m/f); no ordering (e.g. it makes no sense to state that $m > f$; arbitrary labels (e.g., m/f, 0/1, etc).
   b. Ordinal: ordered but differences between values are not important (e.g., political parties on left to right spectrum given lables 1,1,2; restaurant ratings)
   c. Interval: ordered, constant scale, but no natural zero; differences make sense, but ratios do not (e.g., temperature: $30^0 - 20^0 = 20^0 - 10^0$, but $\frac{20^0}{10^0}$ is not twice as hot).
   d. Ratio: ordered, constant scale, natural zero (e.g., height, weight, age, length)

3. Data Structure



   A Data classified in categories. e.g. Gender, Hair Color, Bond Rating.
   B Data measured on numerical scale. e.g. Age, temperature.

EX2 Table 1.6 shows the fuel efficiency ratings for 10 cars.

Car	Class	Cylinders	City MPG	Highway MPG	Fuel Type
Audi A8	Large	12	13	19	Premium
BMW 328Xi	Compact	6	17	25	Premium
Cadillac CTS	Midsize	6	16	25	Regular
Chevrolet Malibu	Midsize	6	17	26	Regular
Chrysler 300	Large	8	13	18	Premium
Ford Focus	Compact	4	24	33	Regular
Hyundai Elantra	Midsize	4	25	33	Regular
Pontiac G6	Compact	6	15	22	Regular
Toyota Camry	Midsize	4	21	31	Regular
Volkswagen Jetta	Compact	5	21	29	Regular

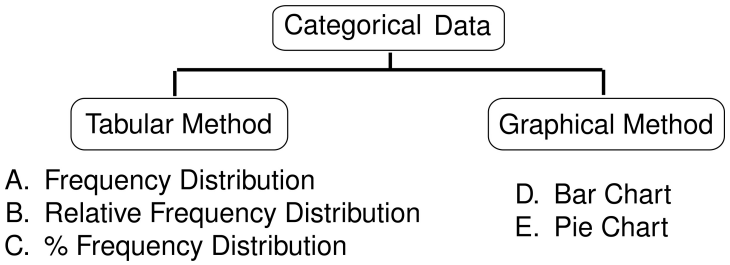
   a How many elements are in this data set?
   b How many variables are in this data set?
   c ID the data structure of the variables.

   d ID the measurement scale of the variables.

CH 2: Descriptive Statistics: Tabular and Graphical Presentations

Part 1 Organizing Categorical Data

1. Case 1: One Variable Categorical Data (Sec 2.1)



(A) **Frequency Distribution**: A tabular summary of data showing the number (frequency) of data values in each of several nonoverlapping classes.

(B) **Relative Frequency Distribution**: A tabular summary of data showing the fraction or proportion of data values in each of several nonoverlapping classes.

Note: Relative frequency of a class $= \dfrac{\text{Frequency (in the interval)}}{\text{Total number of observations}}$

(C) **Percent Frequency Distribution**: A tabular summary of data showing the percentage of data values in each of several nonoverlapping classes.

Note: Percentage frequency of a class = relative frequency $\times 100\%$

(D) **Bar chart**: A graphical device for displaying categorical data that have been summarized in a frequency, relative frequency, or percent frequency distribution.

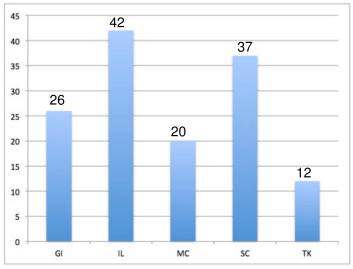
(E) **Pie chart**: Use pie slices to display the percent frequency distribution of each category.

EX1 Given a summary table of 137 mutual funds:

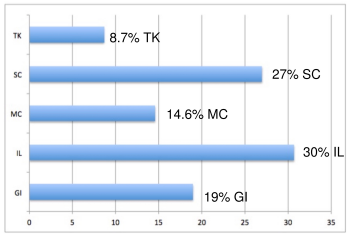
Fund Objective	Number of Funds	Rel. Frequency	% Frequency
Growth and Income (GI)	26		
International (IL)	42		
Midcap (MC)	20		
SmallCap (SC)	37		
Technology(TK)	12		

(a) Provide the relative frequency and the percent frequency distributions.

(b) Construct a bar chart

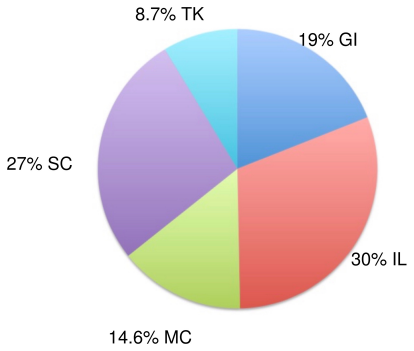


Frequency Bar Chart



Percent Frequency Bar Chart

(c) Construct a pie chart



(d) Based on the pie chart, what percentage of the fund is from $MC + TK$?

2. Case 2: Bivariate Categorical Data (Sec 2.4 Crosstabulations)

   Crosstabulation: A tabular summary of data for two variables. The classes for one variable are represented by the rows; the classes for the other variable are presented by the columns.

EX2 Crosstabulation (contingency table) of whether the fund has a sales charge vs. mutual funds

Sales Charge	GI	IL	MC	SC	TK	Total
Y	17	25	6	15	9	72
N	9	17	14	22	3	65
Total	26	42	20	37	12	137

3. **Side-by-side bar chart**: Bar charts arranged side-by-side according to different categories. Useful when looking for patterns or relationships.

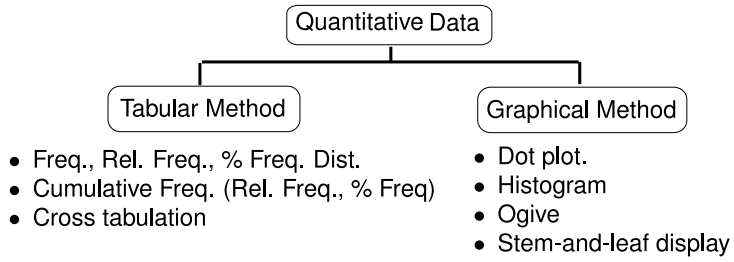
EX 2 (cont.) Construct the side-by-side bar chart.

4. Note: If we divided each cell of Table 1 by the grand total, we obtain a % based overall table.

Sales Charge	GI	IL	MC	SC	TK	Total
Y	17/137=12.4%	18.3%	4.4%	10.9%	6.6%	52.6%
N	9/137=6.6%	12.4%	10.2%	16.1%	2.2%	47.4%
Total	26/137=19%	30.7%	14.6%	27%	8.8%	100%

CH 2: Descriptive Statistics: Tabular and Graphical Presentations

Part 2 Organizing Quantitative Data



1. Dot Plot: A graphical device that summarizes data by the number of dots above each data value on the horizontal axis.

EX3 Given a set of data: 3, 5, 8, 9, 10, 1. Construct a dot plot.

2. Histogram:

   (A) **Histogram**: A graphical presentation of a frequency distribution, relative frequency distribution, or percent frequency distribution of quantitative data constructed by placing the class intervals on the horizontal axis and the frequencies, relative frequencies, or percent frequencies on the vertical axis.

   (B) Width of the classes:

$$\text{Approximate class width} = \frac{\text{Largest data value - Smallest data value}}{\text{Number of classes}}$$

EX3 (Cont.) Construct the distributions (with 2 classes).

EX4 The data set of the monthly finance charges of 50 customers from a store's record is given: \$20, \$5, \$13, ... The following table is the frequency distribution of the data set.

Class interval	number of customers (Freq)	RelativeFrequency
\$0-5	15	
\$ 5-10	20	
\$10-15	10	
\$15-20	5	

a. Construct a histogram (with Relative Frequency)

b. Referring to the histogram, what % of the finance charges is between \$5 to \$15?

c. Referring to the histogram, what % of the finance charges is below \$10?

d. Referring to the histogram, 70% of the finance charges is above what amount?

3. Cumulative distribution and Ogive

(A) **Cumulative frequency distribution**: A tabular summary of quantitative data showing the number of data values that are less than or equal to the upper class limit of each class.

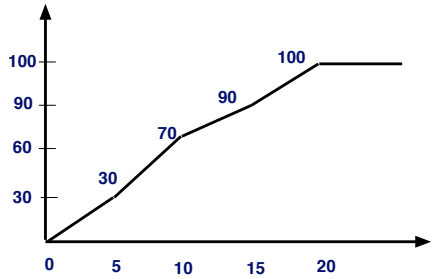
(B) **Cumulative relative frequency distribution**: A tabular summary of quantitative data showing the fraction or proportion of data values that are less than or equal to the upper class limit of each class.

(C) **Cumulative percent frequency distribution**: A tabular summary of quantitative data showing the percentage of data values that are less than or equal to the upper class limit of each class.

(D) **Ogive**: A graph of a cumulative distribution.

EX4 (cont.) Construct a Ogive (cumulative % distribution).

\$	Cumulative %
0	0 %
5	30 %
10	70 %
15	90 %
20	100 %



4. Stem and Leaf Display:

(A) **Ordered array**: Sorting the observations in rank order from the smallest to the largest.

(B) **Stem and Leaf**: An exploratory data analysis technique that simultaneously rank orders quantitative data and provides insight about the shape of the distribution. In this case, data are separated into leading digits (stems) and the remaining digits (leaves).

EX5 Given a set of data: 269, 272, 305, 283, 438, and 343. Construct a stem-and-leaf display.

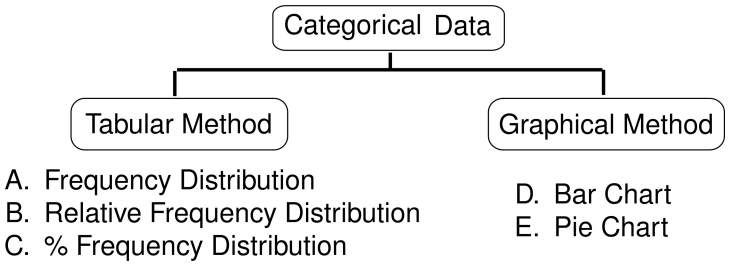
EX6 Given a stem-and-leaf display of the exam scores (0-100). What % of the exam scores is higher than 80?

Stem	Leaf
5	5 7
6	2 3
7	1 5 5
8	4
9	0 8

CH 2: Descriptive Statistics: Tabular and Graphical Presentations

Part 1 Organizing Categorical Data

1. Case 1: One Variable Categorical Data (Sec 2.1)



(A) **Frequency Distribution**: A tabular summary of data showing the number (frequency) of data values in each of several nonoverlapping classes.

(B) **Relative Frequency Distribution**: A tabular summary of data showing the fraction or proportion of data values in each of several nonoverlapping classes.

Note: Relative frequency of a class $= \dfrac{\text{Frequency (in the interval)}}{\text{Total number of observations}}$

(C) **Percent Frequency Distribution**: A tabular summary of data showing the percentage of data values in each of several nonoverlapping classes.

Note: Percentage frequency of a class = relative frequency $\times 100\%$

(D) **Bar chart**: A graphical device for displaying categorical data that have been summarized in a frequency, relative frequency, or percent frequency distribution.

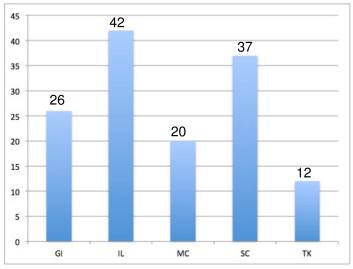
(E) **Pie chart**: Use pie slices to display the percent frequency distribution of each category.

EX1 Given a summary table of 137 mutual funds:

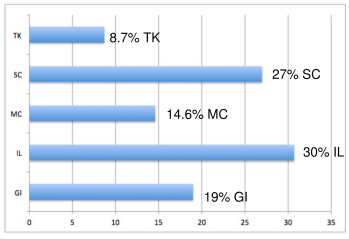
Fund Objective	Number of Funds	Rel. Frequency	% Frequency
Growth and Income (GI)	26		
International (IL)	42		
Midcap (MC)	20		
SmallCap (SC)	37		
Technology(TK)	12		

(a) Provide the relative frequency and the percent frequency distributions.

(b) Construct a bar chart

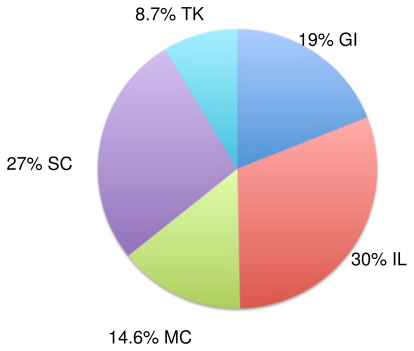


Frequency Bar Chart



Percent Frequency Bar Chart

(c) Construct a pie chart



(d) Based on the pie chart, what percentage of the fund is from $MC + TK$?

2. Case 2: Bivariate Categorical Data (Sec 2.4 Crosstabulations)

   Crosstabulation: A tabular summary of data for two variables. The classes for one variable are represented by the rows; the classes for the other variable are presented by the columns.

EX2 Crosstabulation (contingency table) of whether the fund has a sales charge vs. mutual funds

Sales Charge	GI	IL	MC	SC	TK	Total
Y	17	25	6	15	9	72
N	9	17	14	22	3	65
Total	26	42	20	37	12	137

3. **Side-by-side bar chart**: Bar charts arranged side-by-side according to different categories. Useful when looking for patterns or relationships.

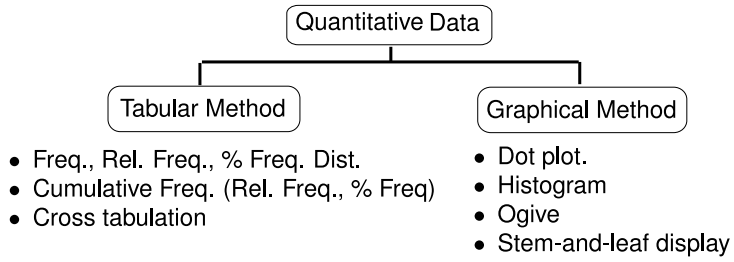
EX 2 (cont.) Construct the side-by-side bar chart.

4. Note: If we divided each cell of Table 1 by the grand total, we obtain a % based overall table.

Sales Charge	GI	IL	MC	SC	TK	Total
Y	17/137=12.4%	18.3%	4.4%	10.9%	6.6%	52.6%
N	9/137=6.6%	12.4%	10.2%	16.1%	2.2%	47.4%
Total	26/137=19%	30.7%	14.6%	27%	8.8%	100%

CH 2: Descriptive Statistics: Tabular and Graphical Presentations

Part 2 Organizing Quantitative Data



1. Dot Plot: A graphical device that summarizes data by the number of dots above each data value on the horizontal axis.

EX3 Given a set of data: 3, 5, 8, 9, 10, 1. Construct a dot plot.

2. Histogram:

   (A) **Histogram**: A graphical presentation of a frequency distribution, relative frequency distribution, or percent frequency distribution of quantitative data constructed by placing the class intervals on the horizontal axis and the frequencies, relative frequencies, or percent frequencies on the vertical axis.

   (B) Width of the classes:

$$\text{Approximate class width} = \frac{\text{Largest data value - Smallest data value}}{\text{Number of classes}}$$

EX3 (Cont.) Construct the distributions (with 2 classes).

EX4 The data set of the monthly finance charges of 50 customers from a store's record is given: \$20, \$5, \$13, ... The following table is the frequency distribution of the data set.

Class interval	number of customers (Freq)	RelativeFrequency
\$0-5	15	
\$ 5-10	20	
\$10-15	10	
\$15-20	5	

a. Construct a histogram (with Relative Frequency)

b. Referring to the histogram, what % of the finance charges is between \$5 to \$15?

c. Referring to the histogram, what % of the finance charges is below \$10?

d. Referring to the histogram, 70% of the finance charges is above what amount?

3. Cumulative distribution and Ogive

(A) **Cumulative frequency distribution**: A tabular summary of quantitative data showing the number of data values that are less than or equal to the upper class limit of each class.

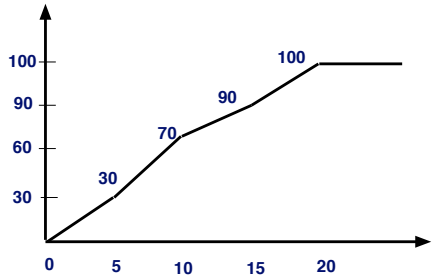
(B) **Cumulative relative frequency distribution**: A tabular summary of quantitative data showing the fraction or proportion of data values that are less than or equal to the upper class limit of each class.

(C) **Cumulative percent frequency distribution**: A tabular summary of quantitative data showing the percentage of data values that are less than or equal to the upper class limit of each class.

(D) **Ogive**: A graph of a cumulative distribution.

EX4 (cont.) Construct a Ogive (cumulative % distribution).

\$	Cumulative %
0	0 %
5	30 %
10	70 %
15	90 %
20	100 %



4. Stem and Leaf Display:

(A) **Ordered array**: Sorting the observations in rank order from the smallest to the largest.

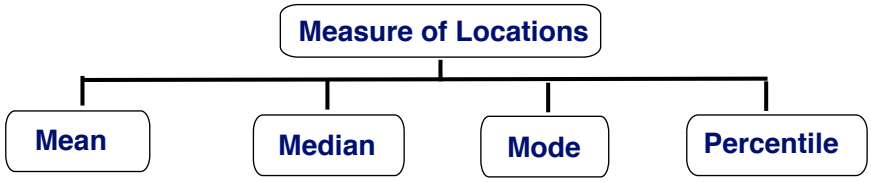
(B) **Stem and Leaf**: An exploratory data analysis technique that simultaneously rank orders quantitative data and provides insight about the shape of the distribution. In this case, data are separated into leading digits (stems) and the remaining digits (leaves).

EX5 Given a set of data: 269, 272, 305, 283, 438, and 343. Construct a stem-and-leaf display.

EX6 Given a stem-and-leaf display of the exam scores (0-100). What % of the exam scores is higher than 80?

Stem	Leaf
5	5 7
6	2 3
7	1 5 5
8	4
9	0 8

CH 3: Descriptive Statistics: Numerical Measures Part 1



1. Measure of Locations

(A) Observation Notation x_i: the ith observation in the list of observations.

(B) Summation Notation Σ ("Sigma"–Computing the sum):

We write $\Sigma_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$

(C) Sample Mean (Notation: $\bar{x}$)

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad \text{(eq3.1)}$$

EX 1 Given a set of data with $n = 5$ (the birth weights): 9.2, 6.4,10.5, 8.1,7.8. Find the mean.

(D) The Population Mean (Notation: μ)

$$\mu = \frac{\Sigma x_i}{N} \qquad \text{(eq3.2)}$$

(E) Median: the middle value when the observations are arranged in ascending order (smallest value to largest value).

Note 1: For an odd number of observations, the median is the middle value; for an even number of observations, the median is the average of the two middle values.

EX 1 (cont.) Find the median.

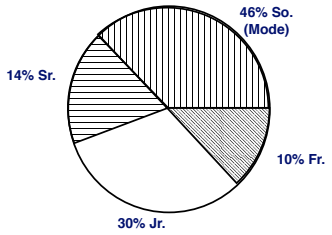
EX 2 Find the mean and median of the data set: ($n = 6$) 15, 3, 46, 623, 126, 64, Find the mean and the median.

Note 2: In some cases, median is a more sensible measure of center than the mean, for example, government uses median income.

(F) Mode: The mode is the value that occurs with greatest frequency.

EX 3 Find the mode for the following ordered array: 0, 0, 1, 2, 2, 3, 3, 3, 3, 3, 4, 5, 6, 26.

EX 4 Find the mode for the pie chart.



(G) Percentile: The pth percentile is a value such that at least p percent of the observation are less than or equal to this value and at least $(100 - p)$ percent of the observations are greater than or equal to the value. To find the percentile, the following procedure can be used:

(1) Order the data from the smallest to the largest.

(2) Find the location of the pth percentile

$$L_p = \frac{p}{100}(n+1) \tag{eq3.5}$$

(3) Rules to follow: ithe rank is split into integer component k and decimal component d, such that $L_p = k + d$. The value (the pth percentile) is calculated as

$$r_k + d(r_{k+1} - r_k)$$

EX 5 Given a set of data: 15, 20, 25, 25, 27, 28, 30, 34. Find the 20th percentile and the 75th percentile.

2. Measures of Variability

(A) Variance (Notation: Sample Variance S^2, Population Variance σ^2)

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \tag{eq3.7}$$

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{eq3.8}$$

(B) Standard Deviation ((Notation: Sample Variance s, Population Variance σ)

$$s = \sqrt{s^2} \tag{eq3.9}$$

$$\sigma = \sqrt{\sigma^2} \tag{eq3.10}$$

EX 6 Given a set of data: $n = 5$: 3, 7, 5, 8, 7. Find the variance and the standard deviation.

Step 1: Find

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{3+7+5+8+7}{5} = 6$$

Step 2: Set up a table to find $(x - \bar{x})^2$

obs.	$(x_i - 6)^2$
3	$(3-6)^2 = 9$
7	$(7-6)^2 = 1$
5	$(5-6)^2 = 1$
8	$(8-6)^2 = 4$
7	$(7-6)^2 = 1$

$\Rightarrow \Sigma = 16$

Step 3: Sample Variance

$$S^2 = \frac{16}{5-1} = 4$$

Step 4: Standard Deviation

$$S = \sqrt{4} = 2$$

CH 3: Descriptive Statistics: Numerical Measures Part 2

(C) Range

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

EX 6 cont. Find the range of the data set: ($n = 5$: 3, 7, 5, 8, 7).

(D) Interquartile Range

$$\text{Interquartile Range} = Q_3 - Q_1 \qquad \text{(eq3.6)}$$

(1) Quartiles: dividing the ordered data into four portions.
(2) Q_1: the first quartile (25th percentile).
(3) Q_2: the second quartile (the median, 50th percentile).
(4) Q_3: the third quartile (the 75th percentile).

EX 5 (cont.) Given a set of data: 15, 20, 25, 25, 27, 28, 30, 34. Find Q_1, median(Q_2), and Q_3 and find the interquartile range.

(E) Coefficient of Variation

$$\left(\frac{\text{Standard deviation}}{Mean} \times 100\right)\% = \frac{s}{\bar{x}} \times 100\% \qquad \text{(eq3.11)}$$

CV is used in comparing two or more sets of data measured in different units

3. Five Number Summary and the Boxplot

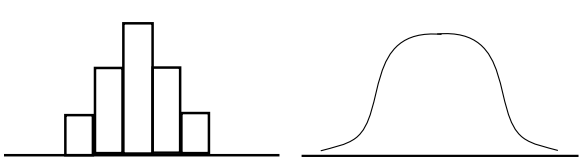
(A) The five-number summary: smallest value, Q_1, Q_2(median), Q_3, largest value

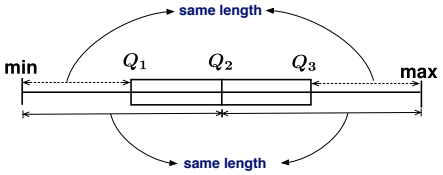
(B) Boxplot: A graphic display of the Five-Number Summary

EX 5 (cont.) Construct the Boxplot of the given data set.

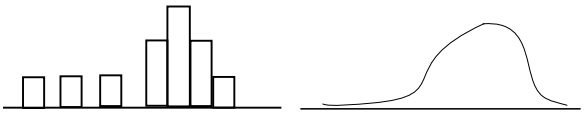
(C) Distribution Shape based on Boxplot:

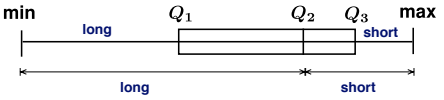
Case 1: Symmetric



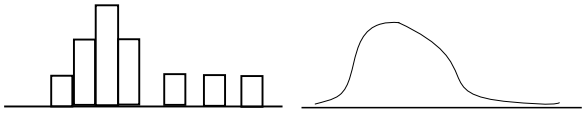


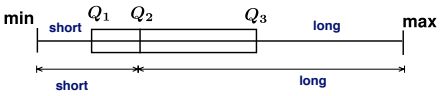
Case 2: Left-skewed





Case 3: Right-skewed





EX 5 (cont.) Find the distribution shape of the data set.

Note: An important numerical measure of the shape of a distribution is called Skewness. Case 1 symmetric ($skewness = 0$);

Case 2 Left-skewed ($skewness < 0$);

Case 3 Right-skewness ($skewness > 0$)

4. z Scores

   (1) z-Score

$$z_i = \frac{x_i - \bar{x}}{s} \qquad \text{(eq3.12)}$$

   (2) z-score is often called the standardized value.

   (3) A z-score reflects how many standard deviations above or below the population mean an observation is. For instance, on a scale that has a mean of 500 and a standard deviation of 100, a value of 450 would equal a z score of (450-500)/100 = -50/100 = -0.50, which indicates that the value is half a standard deviation below the mean.

5. The Empirical Rule:

   For a "Bell-Shaped" normal distribution. About 68% (2/3 of the data) lie within one standard deviation of the mean; about 95% of the data lie within two standard deviation of the mean; Almost all (about 99.7%) of the data lie within three standard deviation of the mean.

CH 3: Descriptive Statistics: Numerical Measures Part 3

5. The Empirical Rule:

For a "Bell-Shaped" normal distribution. About 68% (2/3 of the data) lie within one standard deviation of the mean; about 95% of the data lie within two standard deviation of the mean; Almost all (about 99.7%) of the data lie within three standard deviation of the mean.

6. Measures of Association Between Two Variables

(A) Scatter diagram: Given paired observations (x_i, y_i) (i.e. data set that is concerning with two measurement variables x and y), a scatter diagram uses the x and y axis to represent the data.

(B) The Covariance:

(1) The covariance measures the strength of the linear relationship between two numerical variables (x and y).

(2) The sample covariance is computed from the following equation:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

(C) The correlation Coefficient

(1) The correlation coefficient measures the strength of the linear relationship between two numerical variables (x and y).

(2) The sample correlation coefficient is computed from the following equation:

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

where s_x is the sample standard deviation of x and s_y is the sample standard deviation of y.

(3) In particular, $-1 \leq r_{xy} \leq 1$.

EX 7 Given a set of paired observations with $n = 4$: (2 , 5), (1, 3), (5, 6), (0, 2)

(1) Obtain the scatter diagram.

(2) Compute the covariance s_{xy}.

(3) Compute the sample standard deviations s_x and s_y.

(4) compute the correlation coefficient r_{xy}.

(5) Interpret the result.

CH 4: Basic Probability

1. Basic Concepts

(A) **Sample Space**: The collection of all possible outcomes.

(B) **An Event**: An event is a subset (part) of the sample space in which you are interested.

(C) **Combination:**

$$C_n^N = \frac{N!}{n!(N-n)!} \qquad \text{(eq4.1)}$$

$N! = N(N-1)(N-2)\cdots(2)(1)$, $n! = n(n-1)(n-2)\cdots(2)(1)$, and $1! = 1$, $0! = 1$

EX 1. From a committee of 10 people, in how many ways can we choose a subcommittee of 3 people?

(D) **Probability:** A numerical measure of the likelihood that an event will occur.

EX 2. Two coins are tossed, find the probability of getting at least one head.

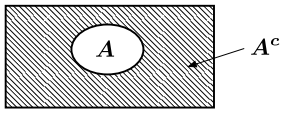
Note 1: The probability of an event is a number between 0 and 1.

Note 2: The probability of the sample space is 1.

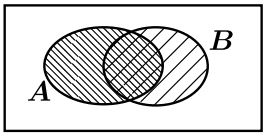
(E) Some basic set notations and formulas.

(1) The complement of an event A, denoted by A^c (the set of all outcomes that are not in A). The equation for computing probability using the complement is given by

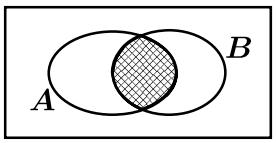
$$P(A) = 1 - P(A^c) \qquad \text{(eq4.5)}$$



(2) The union of two events A, B, denoted by $(A \cup B)$, is the set of all outcomes that are in A, B or both.



(3) The intersection of two events A, B, denoted by $(A \cap B)$, is the set of all outcomes that are in A and B. Note: $P(A \cap B)$ is the joint probability.



(4) If the intersection $(A \cap B)$ is empty (i.e. $P(A \cap B) = 0$), then the two events A, B are called mutually exclusive (disjoint).

(5) Addition law

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad \text{(eq4.6)}$$

(F) Conditional Probability:

The probability of the occurrence of an event A, given the occurrence of another event B, denoted by $P(A|B)$ is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad \text{(eq4.7)}$$

Similarly, the conditional probability of event B given that event A has occurred as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Now, based on the conditional probability, we can write $P(A \cap B) = P(A|B)P(B)$ or $P(A \cap B) = P(B|A)P(A)$ (called the multiplication law). If $P(A|B) = P(A)$ and $P(B|A) = P(B)$, then A and B are said to be independent.

(G) Independent Events: Two events A and B are said to be statistically independent if and only if

$$P(A \cap B) = P(A)P(B). \quad \text{(eq4.13)}$$

EX 3. Suppose that an employment agency has found that 40% of the applicants are college grads and 30% of the applicants have had computer skills. The two characteristics are independent. Find the probability that the applicants are college grads and have had computer skills.

2. Application: Computing Probability from crosstabulations.

	A	B	C	Total
D				
E				
Total				

EX 4. The manager of a shirt manufacture wants to study the connection between shifts and shirt quality. 600 shirts are randomly selected. The results are shown below:

Shirt quality	Shift 1	Shift2	Shift3	Total
Perfect	240	191	139	570
Flawed	10	9	11	30
Total	250	200	150	600

(1) What proportion (%, probability) of the shirts were perfect or made by shift 2?

(2) Given that the shirts were made by shift 2, what proportion (%, probability) of the shirts were perfect?

CH 5: Discrete Probability Distributions

Part 1: Discrete Probability Distribution

1. Basic Concepts

(A) **Random Variable** (x): is a numerical description of the outcome of an experiment.

EX 1 Tossing a fair coin twice. Let x be the random variable associated with the number of heads of the experiment. List all possible outcomes for x.

(B) **Discrete Random Variable**: A random variable that may assume either a finite number of values or an infinite sequence of values.

(C) **Continuous Random Variable**: A random variable that may assume any numerical value in an interval or collection of intervals.

EX 2 Determine if the following random variable is discrete or continuous.

(1) Number of cars arriving at a tollbooth in two-hour period.

(2) Amount of time spent trying to find a parking spot on campus.

(D) **Probability distribution**: A description of how the probabilities are distributed over the values of the random variable.

(E) We usually use a table or a chart to represent the discrete probability distributions.

All Possible Variables

x	$f(x)$
x_1	$f(x_1) = P(x = x_1)$
x_2	$f(x_2) = P(x = x_2)$
$\vdots$	$\vdots$
x_N	$f(x_N) = P(x = x_N)$

Associated probabilities
$\sum f(x) = 1$
$f(x) \geq 0$

EX 1 (cont) Construct the probability distribution for the experiment with random variable x (# of heads).

(F) We can use the probability distribution table to calculate some given probabilities.

Step 1: Write the probability statement.

Step 2: Find the probability.

EX 3. Probability distribution for the number of automobiles sold during a day at a car dealer is given. Find the following probabilities:

x	$f(x)$
0	0.1
1	0.2
2	0.4
3	0.2
4	0.1

(1) x is exactly 1.

(2) x is at most 2.

(3) x is between 2 and 3 (the end points are included).

(4) x is at least 1.

2. Given the probability distribution table, compute the expectation (mean, expected value), variance and standard deviation.

$$\text{Mean:} E(x) = \mu = \sum xf(x) \qquad \text{(eq5.4)}$$

$$\text{Variance:} \sigma^2 = \sum (x-\mu)^2 f(x) \qquad \text{(eq5.5)}$$

$$\text{Standard Deviation:} \sigma = \sqrt{\sigma^2} = \sqrt{\sum (x-\mu)^2 f(x)}$$

EX 3 (Cont). Compute the mean (expected value), the variance, and the standard deviation of random variable x.

EX 4 A trip Insurance policy pays \$1000 to the customer in case of a loss due to theft. If the risk of such a loss is assured to be 1 in 200. What is a fair premium?

CH 5: Discrete Probability Distributions

Part 2: Binomial Distribution

1. Characteristics of a Binomial Distribution:

   (A) The experiment consists of a sequence of n identical trials.

   (B) Each trial is classified into one of the two outcomes (Success/Failure).

   (C) The probability of a success p is the same for each trial. The probability of a failure for each trial is $1 - p$.

   (D) The trials are independent.

   EX 5.Tossing a coin 3 times. Let us assume that getting a head is a success. This experiment is a binomial distribution.

   EX 6.Selecting random multiple choice with 10 questions, each question has 4 possible answer. This is also a binomial distribution.

2. Binomial Distribution Formula

   Given a binomial distribution with n trials and success probability p, then the probability of x successes is (called binomial probability function)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{(eq5.12)}$$

   Where: $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ (Note: $n! = n * (n-1) * (n-2) * \ldots 2 * 1$; $0! = 1$; $1! = 1$).

   $x =$ the number of successes in the sample ($x = 0, 1, \ldots, n$).

   $n =$ the number of trials.

   $p =$ Probability of success, $\quad 1 - p =$ Probability of failure

EX 5(cont.) Compute the probability of all possible outcomes using Eq.5.12

EX 6(cont.) Find the probability of getting exactly 6 questions right.

EX 7. A roofing contractor estimates that after the "quick fix" job on leaking roofs is done, 15% of the roofs will still leak. He fixed eight roofs, find the probability that at least two of these roofs will still leak.

3. Binomial Mean, Variance, and Standard Deviation

$$\text{Mean: } E(x) = \mu = np \tag{eq5.13}$$

$$\text{Variance: } Var(x) = \sigma^2 = np(1-p) \tag{eq5.14}$$

$$\text{Standard Deviation: } \sigma = \sqrt{np(1-p)}$$

EX 8 Suppose that past history shows that 7% of the production is defective, 200 samples are selected, find the mean, the variance, and the standard deviation of the problem.

CH 6: Continuous Probability Distribution–Normal Distribution

1. Characteristics of a Normal Distribution:

   (A) The distribution is bell-shaped with mean μ and standard deviation σ.

   (B) The total area under the curve is 1.

   (C) The empirical rule holds.

   (D) The distribution function: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ for $-\infty < x < \infty$

   (E) The probability of observing the value x between a and b:

   $$P(a \leq x \leq b) = \text{Area under the curve between a and b}$$

   (F) Notation $N(\mu, \sigma)$

2. Standard Normal Distribution

   (A) The standard normal distribution had a bell-shaped distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. It is denoted by $N(0, 1)$, and the random variable is denoted by Z (instead of X).

   (B) We use the z-table to find the probability (area under the curve).

   (C) The z-table represents the area under the curve to the left of a given value (left-tail).

   Case 1 (left-tail): To find the probability $P(Z \leq a)$, we use the table directly.

   EX 1 Draw the graph of standard normal and use the z-table to find $P(z \leq 1.10) =$

   EX 2 Draw the graph of standard normal and use the z-table to find $P(z \leq -2.22) =$

   Case 2 (right-tail): To find the probability $P(Z \geq a)$, we use the complement probability

   $$P(Z \geq a) = 1 - P(Z < a)$$

   EX 3 Draw the graph of standard normal and use the z-table to find $P(z \geq -0.91) =$

Case 3 (In between): To find the probability of an interval $[a, b]$, we use the formula

$$P(a \leq Z \leq b) = \text{(Area to the left of b)-(Area to the left of a)}$$

EX 4 Draw the graph of standard normal and use the z-table to find $P(-1.37 \leq z \leq 1.82)$

(D) Using the table to find a z-value if the probability is given

EX 5 Find the value of z such that the probability of being less than that value is 1.5%

EX 6 Find a value of z such that the probability of being more than that value is 30.5%

3. Application 1: Finding the probability of a random variable X that is normally distributed

   Step 1: Write down the probability statement (say: $P(x < a)$, $P(x > a)$, $P(a < x < b)$.)

   Step 2: Use $Z = \frac{X-\mu}{\sigma}$ (eq6.3) to cover the random variable into standard normal z:

   Step 3: Look up the standard normal table to find the probability.

EX 7 Certain costs x is assumed to follow a normal distribution with a mean \$35,000 and a standard deviation of \$10,000.

(1) What is the probability that the costs will be less than \$40,000?

(2) What is the probability that the costs will be between \$45,000 and \$50,000?

4. Application 2: Find the value of x for a given probability.

   Step 1: Find the z-value from the standard normal table for the given probability (left-tail or right-tail).

   Step 2: Solve for x using equation $Z = \frac{X-\mu}{\sigma}$ (eq 6.3).

EX 8 Assumed that american family spends an average of \$75 with a standard deviation of \$5 on food per week (it's normally distributed). If 10.03% of the American families spend x or more on food per week. What would be the value of x?

CH 7: Sampling Distributions

1. Basic Concepts

   (A) Parameter: A numerical characteristic of a population, such as a population mean μ, a population standard standard deviation σ, a population proportion p.

   (B) Sample statistic: A sample characteristic, such as sample mean $\overline{x}$, sample standard deviation s, a sample proportion $\overline{p}$.

   (C) Our goal in this chapter is to use sample statistics to estimate certain parameters, such as point estimator $\bar{x}$ for μ, point estimator $\bar{p}$ for p.

   (D) Any sample statistic will have a probability distribution called the **sampling distribution** of the statistic.

2. Sample Distribution of $\bar{x}$.

   (A) The expected value of $\bar{x}$ (or the population mean of the sample mean, denoted by $E(\bar{x})$)

   $$E(\bar{x}) = \mu \qquad \text{(eq7.1)}$$

   where μ is the population mean.

   (B) The standard deviation the sample mean (or the standard error of the sample mean, denoted by $\sigma_{\overline{x}}$)

   $$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \qquad \text{(eq7.3)}$$

   where σ is the population standard deviation.

   EX 1 A population has a mean of 99 and standard deviation of 7. Compute the expected value of the sample mean and the standard error of the sample mean for
   (1). $n = 4$

   (2). $n = 25$

   (C) If a random variable x is from a normal distribution, i.e., $N(\mu, \sigma)$, then the random variable sample mean $\overline{x}$ would have a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, i.e., $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
   Application 1: Finding the probability of the sample mean $\overline{x}$:
   Step 1: Write down the probability statement (say: $P(\bar{x} < a)$. $P(\bar{x}) > a$, $P(a < \bar{x} < b)$)
   Step 2: Use $Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ to standardize the value of $\overline{x}$ into Z
   Step 3: Look in the standard normal table (z-table) to find the probability.

Application 2: Recovering the $\overline{x}$ value for a given probability p.

Step 1: Find the Z-value from the standard normal table for the given probability.

Step 2: Use the formula $Z = \dfrac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ to solve for $\overline{x}$

EX 2 Apples have a mean weight of 7 ounces and a standard deviation of 2 ounces (they are normally distributed) and they are chosen at random and put in a box of 30.

(1) Find the probability that the average weight of the apples in a box is greater than 6.5 ounces.

(2) Below what value do 12.1% of the average weight of the apples fall?

(D) Question: what if the sampling is from a nonnormal population, do we have similar result? Answer: yes! if the sample size n is large enough (say, at least 30). This result is called the Central Limit Theorem: Whatever the population, the distribution of $\overline{x}$ is approximately normal with mean μ and standard deviation $\dfrac{\sigma}{\sqrt{n}}$ if n is large.

EX 3 Consider a population with mean $\mu = 82$ and standard deviation $\sigma = 12$. If a random sample of size of 64 is selected. What is the probability that the sample mean will lie between 80.8 and 83.2?

3. Sample distribution of $\bar{p}$

(A) The sample proportion $\bar{p}$ can be computed use the equation $\bar{p} = \dfrac{x}{n}$ where x is the number of elements in the sample that possess the characteristic of interest and n is the sample size.

(B) Expected value of $\bar{p}$

$$E(\bar{p}) = p \qquad \text{(eq7.4)}$$

where p is the population proportion

(C) The standard deviation of $\bar{p}$ (or called the standard error, denoted by $\sigma_{\bar{p}}$)

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \text{(eq7.6)}$$

EX 4 A simple random sample of size 100 is selected from a population with $p = 0.40$. What is the expected value of $\bar{p}$? What is the standard error of $\bar{p}$?

CH 8: Interval Estimation (Part 1: for mean)

1. Basic Concepts

   (A) Interval estimate (Confidence Interval): An estimate of a population parameter that provides an interval believed to contain the value of the parameter. Note: the interval estimate has the form: point estimate ± margin of error.

   (B) Confidence level: The probability of including the population parameter within the confidence interval at $100(1-\alpha)\%$. Say 95%, 99%, etc.

   (C) α is called the level of significance. In this case, it is the probability that the interval estimation procedure will generate an interval that does not contain the parameter.

   (D) Why does it work?

2. Case I: $100(1-\alpha)\%$ confidence interval estimation of the mean μ (σ known).

   (A) formula:

$$\overline{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{(eq8.1)}$$

   Note: We assume: (a) The population is normally distributed or n is large; (b) The population standard deviation σ is known. (c) $Z_{\alpha/2}$ is called the critical value and $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the margin of error for the estimation.

   (B) $Z_{\alpha/2}$ notation:

   $Z_{\alpha/2}$ = the right-tail (upper tail) probability $\alpha/2$ point of the standard normal; i.e., the area to the right of $Z_{\alpha/2}$ is $\alpha/2$.

   EX 1 Find the values of $Z_{\alpha/2}$ for 90%, 95% and 99%

   (1) 90%

   (2) 95%

   (3) 99%

(C) Using the formula

EX 2 The computer paper is expected to have a standard deviation of 0.02inch. 100 sheets are selected and the mean is 10.998 inches. Set up a 95% confidence interval estimates of the population mean paper length.

3. Case II: $100(1-\alpha)\%$ confidence interval estimation of the mean μ (σ unknown).

(A) Formula:

$$\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \qquad \text{(eq8.2)}$$

where s is the sample standard deviation.

(B) Student's t distribution: Let $x_1, x_2, \ldots\ x_n$ be a random sample from a normal population with mean μ and standard deviation σ, then $t = \dfrac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ is called the t-distribution with $(n-1)$ degrees of freedom.

(C) $t_{\alpha/2}$ notation

(D) How to read the t-table:

CH 8: Interval Estimation

Part 2: Confidence Interval (CI) (cont.)

(E) How to use formula (eq 8.2)

$$\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \qquad \text{(eq8.2)}$$

EX 3 Suppose that a sample of 100 sales invoices is selected from the population of sales invoices during the month and the sample mean is 110.27 and the sample variance is 838.10. Set up a 95% confidence intervals for the mean μ.

4. Case III: $100(1-\alpha)\%$ confidence interval estimation for the proportion p.

(A) We use the sample proportion $\bar{p}$ to estimate the population proportion p combined with the margin of error term.

(B) The sample proportion is defined as $\bar{p} = \frac{x}{n}$,where x is the number of elements in the sample that possess the characteristic of interest and n is the sample size.

(C) In Chapter 7 we indicated that the sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$. We use $Z_{\alpha/2}$ for the critical value.

(D) Formula for the $100(1-\alpha)\%$ confidence interval for the population proportion p:

$$\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad \text{(eq8.6)}$$

(E) How to use (eq8.6)

Step 1: Find the sample proportion $\bar{p} = \dfrac{x}{n}$.

Step 2: Find the critical value $Z_{\alpha/2}$.

Step 3: Compute the confidence interval.

EX 4 A company wants to determine the frequency of occurrence of invoices error. Suppose that in a sample of 100 sales invoices, 10 contain errors. Construct a 90% confidence interval for the true proportion of error.

EX 5 Out of 268 interviewed, 83 people said that they would buy a certain product. Use a 95% confidence interval to estimate the true proportion of the customer who would buy the product.

CH 9: Hypothesis Testing Part 1

1. Introduction

   (A) The aim of testing statistical hypotheses is to determine whether a claim or conjecture about some feature of the population parameter (say, the mean μ, or the proportion p) is strongly supported by the information obtained from the sample data.

   (B) Some basic concepts in Hypothesis Testing

      (a) A set of hypotheses:
      H_1 or H_a : the claim or the research hypothesis that we wish to establish is called the alternative hypothesis.

      H_0 (Null hypothesis): Refers to a specified value of the population parameter.

      (b) There are three forms of Hypotheses in this chapter:

      Form 1: Two-tailed test
      H_0 : Parameter = reference value
      H_1 : Parameter $\neq$ reference value

      Form 2: Upper, one-tailed test
      H_0 : Parameter $\leq$ reference value
      H_1 : Parameter $>$ reference value

      Form 3: lower, one-tailed test
      H_0 : Parameter $\geq$ reference value
      H_1 : Parameter $<$ reference value

   (C) Type I and Type II error of the test

   (D) The probability of making a type I error= α: level of significance.

   (E) Test Statistic: A statistic whose value helps determine whether a null hypothesis should be rejected.

   (F) p-value: A probability that provides a measure of the evidence against the null hypothesis provided by the sample. If the p-value is less than α, we reject H_0; If the p-value is more than α, we fail to reject H_0.

2. Application: Hypothesis Testing

   Step 1: State H_0 vs. H_1.

   Step 2: Compute the test statistic

   Step 3: Compute the p-value based on the test statistic and making a decision:

   if the p-value is less than α, we reject H_0, otherwise, we fail to reject H_0.

(A) Case I: Z-test for the population mean μ (σ known)

$$Z_{cal} = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad \text{(eq9.1)}$$

Note: two-tailed test: $p - value = 2P(Z > |Z_{cal}|)$
upper, one-tail test: $p - value = P(Z > Z_{cal})$
lower, one-tail test: $p - value = P(Z < Z_{cal})$

EX 1. A manager wants to know if the amount of paint in 1-gallon cans is indeed 1-gallon. Given that the population standard deviation is 0.02 gallon. A random sample of 50 cans is selected and the sample mean is 0.995 gallon. Is there evidence that the mean amount is different from 1 gallon ($\alpha = 0.01$)?

(a) State H_0 and H_1

(b) Compute the test statistic

(c) Find the p-value and make a decision.

(B) Case II: t-test for the population mean μ (σ unknown)

$$t_{cal} = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \qquad \text{(eq9.2)}$$

Note: use the t-table (with $n - 1$ degrees of freedom) to obtain the range of the p-value and then make a decision.

EX 2. 100 candy bars are random selected with a mean of 1.466 and standard deviation of 0.132. For $\alpha = 0.05$, is there evidence that the average weight of the candy bars is less than 1.5 ounces?

(a) State H_0 and H_1

(b) Compute the test statistic

(c) Guessing the range of the p-value and make a decision.

CH 9: Hypothesis Testing Part 2

EX 2 (Cont) Is there evidence that the average weight of the candy bars is different from 1.5 ounces ($\alpha = 0.05$)?

(C) Case III: Z-test for the population proportion p

$$Z_{cal} = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} \qquad \text{(eq9.4)}$$

Note: two-tailed test: $p - value = 2P(Z > |Z_{cal}|)$
upper, one-tail test: $p - value = P(Z > Z_{cal})$
lower, one-tail test: $p - value = P(Z < Z_{cal})$

EX 3 It's claim that the usual percentage of overdraw is more than 10% on checking account(CA) at a bank. To test this claim, a random sample of 50 CA is examined and six out of 50 were found to be overdraw. What conclusion can you make at $\alpha = 0.05$?

(a) State H_0 and H_1

(b) Compute the test statistic

(c) Find the p-value and make a decision.

3. Making a decision based on the Critical Value

Step 1: State H_0 vs. H_1.

Step 2: Compute the test statistic and find the critical value.

Step 3: Make a decision based on the critical value.

EX 1 (cont) For the two-tail test, make a decision using the critical approach.

Step 1: State H_0 and H_1

Step 2: Compute the test statistic and find the critical value

Step 3: Make a decision.

EX 2 (cont) Use the critical value approach to test if the average weight of the candy bars is less than 1.5 ounces ($\alpha = 0.05$).

EX 3 (cont) Use the critical value approach to test the hypothesis.

CH 10: Hypothesis Testing for Data from Two or More Samples Part 1

1. Case 4: Z-test for difference in Means $(\mu_1 - \mu_2)$ with both σ_1 and σ_2 known.

   (A) Concepts

   (B) The Test Statistic

   eq 10.5 : Test statistic for mean difference $\mu_1 - \mu_2$ (σ_1, σ_2 known): $Z_{cal} = \dfrac{(\overline{X_1} - \overline{X_2}) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

   (C) Assumptions for using this formula: the populations are normally distributed or the samples are large; the two samples are randomly and independently drawn.

   (D) We can draw our conclusion either based on the critical value approach or the p-value approach.

   (E) For two-tailed test: $p - value = 2 * P(Z > |Z_{cal}|)$;
   upper, one-tail test: $p - value = P(Z > Z_{cal})$,
   lower, one-tail test: $p - value = P(Z < Z_{cal})$

   EX 1 Given two independent samples, a sample of size $n_1 = 40$ from a population 1 with known standard deviation $\sigma_1 = 20$ is selected and resulting in a sample mean of $\overline{X_1} = 72$; another sample of size $n_2 = 50$ from population 2 with known standard deviation $\sigma_2 = 10$ is also selected and the sample mean $\overline{X_2} = 66$. Test if the average for population 1 is more than the average for population 2 ($\alpha = 0.025$).

   Step 1: State H_0 vs. H_1.

   Step 2: Compute the test statistic

   Step 3: Make a decision using either the p-value approach or the critical value approach.

2. Case 5: t-test for difference in Means $(\mu_1 - \mu_2)$ with both σ_1 and σ_2 unknown.

   (A) Concepts

   (B) Compute the test statistic

   eq10.8: Test statistic for Mean difference $\mu_1 - \mu_2$ (σ_1, σ_2 unknown): $t_{cal} = \dfrac{(\overline{X_1} - \overline{X_2}) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

   with eq10.7: $df = \dfrac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{s_2^2}{n_2})^2}$

   Note: we round the noninterger degrees of freedom down.

   EX 2 Comparing the lifetimes of two brands of batteries, a researcher has randomly selected 20 batteries of brand A with $\overline{X}_A = 22.5$ months and $S_A = 2.5$ months and 30 batteries from brand B with $\overline{X}_B = 20.1$ months and $S_B = 4.8$ months . Test if the means are different ($\alpha = 0.05$)

   Step 1: State H_0 vs. H_1.

   Step 2: Compute the test statistic and *df*

   Step 3: Make a decision using either the p value approach or the critical value approach.

CH 10: Hypothesis Testing for Data from Two or More Samples Part 2

3. Case 6: t-test for difference in two related samples μ_d

   (A) Basic Concept and Data Structure

   (B) Test Statistic

   eq 10.9 Test statistic for mean difference (related samples): $t_{cal} = \dfrac{\overline{d} - \mu_d}{\frac{S_d}{\sqrt{n}}}$

   (with $(n-1)$ degrees of freedom)

   (C) Hypothesis Testing

   Step 1: State H_0 vs. H_1.

   Step 2: Compute the test statistic

   Step 3: Make a decision using either the p-value approach or the critical value approach.

EX 3 Given a set of matched pair of data, test if the mean has been changed (use $\alpha = 0.05$).

Step 1: State H_0 vs. H_1.

Step 2: Compute the test statistic

Step 3: Make a decision using either the p-value approach or the critical value approach.

4. Case 7: Z-test for the Difference Between Two Proportions $p_1 - p_2$

(A) Basic Concepts

(B) The Test Statistic

eq10.16: Test statistic for the difference between two proportions $Z_{cal} = \dfrac{(\overline{p}_1 - \overline{p}_2)}{\sqrt{\overline{p}(1-\overline{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$,

where eq 10.15: $\overline{p} = \dfrac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2}$

EX 4 Auto company suspects that singles have more claims than married policyholders. Let the single policyholder be population 1 and married policyholder be population2. If a random survey indicates that 76 out of 400 single and 90 out of 900 married policyholders did auto claim last year, test the theory with $\alpha = 0.05$.

Step 1: State H_0 vs. H_1.

Step 2: Compute the test statistic

Step 3: Make a decision using either the p-value approach or the critical value approach.

CH 12: Testing the Equality of Population Proportions for Three or more Population Proportions

(A) The pair of hypothesis:

H_0: $p_1 = p_2 = p_3 = ... = p_k$.

H_1: Not all population proportions are equal.

(B) Test Statistic (χ^2-test):

eq12.5: The test stat: $\chi^2 = \sum_i \sum_j \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$

(where f_{ij} is the observed frequency for the cell in row i and column$j$, e_{ij} expected frequency for the cell in row i and column j under the assumption H_0 is true.)

Note1: Select a random sample from each of the populations and recored the observed frequencies, f_{ij} in a table with 2 rows and k columns.

Note2: The expected frequencies $e_{ij} = \dfrac{(Row\ i\ Total)(Column\ j\ Total)}{Total\ sample\ size}$

Note 3: We set up a table to compute the test statistic

(C) How to use the χ^2 table:

(1) The test statistic has chi-square distribution with $k-1$ degrees of freedom.

(2) α is the level of significance (upper tail).

(D) We can use either the p-value approach or the critical value approach to make a decision.

EX Suppose that in a particular study we want to compare the customer loyalty for three automobiles. Chevrolet Impala, Ford Fusion, and Honda Accord. The Hypotheses are stated as follows:

H_0: $p_1 = p_2 = p_3$

H_1: Not all population proportions are equal

Sample results of likely to repurchase for three populations of automobile owners are given from the following table:

Find the test statistic and use $\alpha = 0.01$ to make a decision.

(1) Set up a table to find the test statistic: eq12.5: $\chi^2 = \sum_i \sum_j \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$

(2) Use the p-value approach to make a decision.

(3) Use the critical value approach to make a decision.

CH 14: Simple Linear Regression Model Part 1

1. The Simple Linear Regression Model (Population):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where β_0 and β_1 are the population parameters. Moreover, β_0 is the y-intercept; β_1 is the slope; and ε is the random error in y.

2. The Simple Linear Regression Model (Sample):

(A) Scatter diagram:
Given paired observations (x_i, y_i), a scatter diagram uses the x and y-axies to represent the data

(B) We use r (correlation coefficient) to measure the strength of the linear relation between the x variable and y variable:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

(C) We want to find the relationship between x and y by fitting a line to the data set.

eq14.4: Estimated Simple Regression Equation: $\hat{y} = b_0 + b_1 x$

(D) Linear regression equation:
At each observation, the predicted value of y is given by: $\widehat{y_i} = b_0 + b_1 x_i$
where b_0 and b_1 are regression coefficients.
Moreover,
b_0 is the y-intercept: the average value of y when $x = 0$.
b_1 is the slope.
$\widehat{y_i}$ is the predicted value of y for observation i
x_i is the value of x for observation i

(E) We use the least squares method to compute b_0 and b_1:
(a) In this case, we minimize $\sum(y_i - \widehat{y_i})^2$.

(b) Using differential calculus, we can obtain the following results:

eq14.6: The Slope $b_1 = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$

eq14.7: The Y-intercept: $b_0 = \overline{y} - b_1 \overline{x}$

EX1 Given are five observations for two variables x and y.

(a) Develop a scatter diagram and approximate the relationship between x and y by drawing a straight line through the data.

(b) Compute b_0 and b_1.

(c) Intercept the regression equation and predict the average value of y when $x = 5$.

3. Three important measures of variation

   (1) Sum of squares Due to Error (SSE): measure of the error in using the estimated regression equation to estimate the values of the dependent variable in the sample.

   eq14.8: Sum of squares Due to Error: $SSE = \sum(y_i - \hat{y}_i)^2$

   (2) Total sum of squares (SST): Measure of variation of the y_i values around their mean $\overline{y}$.

   eq14.9: Total sum of squares: $SST = \sum(y_i - \bar{y})^2$

   (3) Sum of squares Due to Regression (SSR): measure of variation attributable to the relationship between X and Y.

   eq14.10: Sum of Squares Due to Regression: $SSR = \sum(\hat{y}_i - \bar{y})^2$

   eq14.11: Relationship Among SST,SSR,and SSE: $SST = SSR + SSE$

4. Coefficient of Determination (Notation: r^2): A measure of the goodness of fit of the estimated regression equation. It can be interpreted as the proportion of the variability in the dependent variable y that is explained by x in the estimated regression equation.

   eq14.12: Coefficient of determination: $r^2 = \frac{SSR}{SST}$

5. Standard error of estimate (Notation: s): Measures how much the data vary around the regression line. Its the square root of the mean square error (MSE).

   eq14.16: Standard error of the estimate: $s = \sqrt{\frac{SSE}{n-2}}$

EX 2 Given $SSR = 66$, $SST = 88$ and $n = 22$, (a) compute the coefficient of determination and interpret its meaning. (b) Find the standard error of estimate s.

CH 14: Simple Linear Regression Model Part 2: Hypotheses Test and Confidence Intervals:

6. Population Model Assumptions:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

where β_0 and β_1 are the population parameters. Moreover, β_0 is the y-intercept; β_1 is the slope; and ε is the random error in y (assumed to be normally distributed with $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2$).

Note:

7. Testing for Significance: We use t-test for the slope β_1 to determine the existence of a significant linear relationship between the x and y variables.

Step1: State H_0 vs. H_1.

Step2: Compute the test statistic and critical value.

eq14.19: t Test Statistic $t_{cal} = \dfrac{b_1}{S_{b_1}}$ with $(n-2)$ degrees of freedom

where eq14.18: Estimated Standard Deviation of b_1: $s_{b_1} = \dfrac{s}{\sqrt{\sum(x_i - \overline{x})^2}}$

Step3: Make a decision using either p-value approach or the critical value approach.

EX 3 Given $SSR = 27.51$, $SST = 41.27$, $\sum(x_i - \overline{x})^2 = 18.38$, $\hat{y}_i = 3.0 + 0.5x_i$, and $n = 20$. Use the t test to test the existence of a linear relationship between x and y ($\alpha = 0.05$).

Step 1: State H_0 and H_1

Step 2 Compute the test statistic

Step 3 Make a decision

8. F test for significance in simple linear regression

   Step 1: State H_0 and H_1

   Step 2 Compute the test statistic

   eq14.20: Mean Square Regression: $MSR = \dfrac{SSR}{\#indvar}$

   eq14.21: F Test Statistic: $F = \frac{MSR}{MSE}$

   Step 3 Make a decision

9. The $100(1-\alpha)\%$ confidence interval for the slope β_1:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

EX 3 (cont.) Find the 95% confidence interval for the true slope β_1

10. $100(1-\alpha)\%$ CI for the mean value of y ($E(y^*)$) for a given value of x^*:

$$\text{eq14.24: } \hat{y^*} \pm t_{\alpha/2}\ s_{\hat{y}^*}$$

with eq12.23: $s_{\hat{y}^*} = s\sqrt{\dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

EX 4: To study the relationship between the size of a store (x, in 1000 square feet) and its annual sales (y, in \$1,000,000). We randomly selected 14 store and obtained the following data: $\hat{y} = 0.964 + 1.670x$, $SST = 116.95$, $SSR = 105.75$, and $\sum(x_i - \bar{x})^2 = 37.924$. Obtain a 95% confidence interval of the average annual sales for a store that is 4000 square feet (with $\bar{x} = 2.921$).

Step 1 : find $\hat{y}^*$

Step 2: find the standard error of the estimate s:

Step 3 find the critical value $t_{\alpha/2}$

Step 4: Find the confidence interval

11. How to real Excel computer output for the simple regression model.

CH 15: Multiple Regression: Part 1 The Model

1. Review of the simple linear regression model

   (A) The population

   (B) The prediction equation (regression equation)

   In this chapter, we are interested in developing a model with more than one independent variable (multiple regression).

2. Multiple regression model: describes the relationship between one dependent variable (y) and two or more independent variables ($x_1, x_2, \ldots, x_p$) in a linear function. Note: p is the number of independent variables.

   (A) The population model:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

   (B) The multiple regression equation:

   $$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

   (C) The prediction equation (estimated multiple linear regression equation)

   $$\text{eq15.3: } \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

   Where $b_0, b_1, \ldots, b_p$ are the regression coefficients: (b_0 is the y intercept and $b_1, \ldots, b_p$ are the slopes.)

   (1) $b_0, b_1, \ldots, b_p$ are the estimates of $\beta_0, \beta_1, \ldots . \beta_p$.

   (2) The least squares method is used to minimize $\sum (y_i - \hat{y}_i)^2$ (for the ith observation) to provide the values of $b_0, b_1, \ldots, b_p$.

   (D) The interpretation of the regression coefficients:

   (1) The y intercept (b_0): The estimated average value of y when all the independent variables satisfy $x_1 = x_2 = \cdots = x_p = 0$.

   (2) The slope (b_i and $i = 1, 2, \ldots, p$): Estimate the average of y changes by b_i for each one-unit increase in x_i holding constant the effect of all other independent variables.

EX1 To study the relationship amount the number of Omni-Powerbars sold in a month (y), the price of the Omni-Powerbar (x_1, in cents), and the monthly budget of promotion (x_2, in \$), thirty-four stores were selected and resulting in the following computer output of the multiple regressions model:

a). What is the value of p (the number of independent variable)?

b). What is the prediction equation?

c). Interpret of meaning of b_1.

d). Interpret the meaning of b_2.

e). Predict the average number of bars sold for a store that has a sales price of \$.79 and the promotion expenditures of \$400.

3. Multiple Coefficient of Determination:

Multiple coefficient of determination measures the proportion of total variation in y explained by all independent variables $x_1, x_2, \ldots, x_p$.

$$\text{eq15.8: Multiple Coefficient of Determination: } R^2 = \frac{SSR}{SST}$$

$$\text{eq15.9: Adjusted Multiple Coefficient of Determination: } R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

EX 1 (cont). From a computer output we find out that $R^2 = 0.7577$, interpret this result.

CH 15: Multiple Regression: Part 2 Hypotheses Test and Confidence Intervals

1. The multiple linear regression model and equation (Population):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the population parameters. Moreover, β_0 is the y-intercept; $\beta_j, j = 1, \ldots, k$ is the slope; and ε is the random error in y (assumed to be normally distributed with $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2$).

2. t-test for the slope β_i:

   To determine the existence of a significant linear relationship between the x and y variables. In this case, a hypothesis test of whether β_j is equal to zero or not.

   Step 1: State H_0 vs. H_1.

   Step 2: Compute the test statistic

   The test statistic:

$$\text{Eq 13.15:} \quad t_{cal} = \frac{b_i}{S_{b_i}}$$

   with $(n - p - 1)$ degrees of freedom.

   Note:

   p is the number of independent variables;

   b_i is the slope of variable x_i, holding constant the effects of all other independent variables;

   S_{b_i} is the standard error of the slope b_i.

   Step 3: Make a decision using p-value approach or the critical value approach.

$$p\text{-value approach: Reject } H_0 \text{ if } p\text{-value} \leq \alpha$$

$$\text{Critical value approach } (\ CV = \pm t_{\alpha/2}): \text{ Reject } H_0 \text{ if } t_{cal} \leq -t_{\alpha/2} \text{ or if } t_{cal} \geq t_{\alpha/2}$$

   Note: If we reject H_0, the corresponding independent is significant in explaining y, and should be included in the model. Otherwise, it should not be included in the model.

3. The $100(1 - \alpha)\%$ confidence interval for the true slope β_i

$$b_i \pm t_{\alpha/2} S_{b_i}$$

   with $n - p - 1$ degrees of freedom

EX 2 The firm wants predict the sales (y, in \$1,000's) using the market value (x_1, in \$1,000's), the total assets (x_2, in \$1,000's), and the number of employees (x_3). To do so, thirty-four firms were selected and the following Excel Output was obtained:

(a) If the firm wants to test whether the coefficient on Market value is significant, what is the relevant test statistic? What decision should be made? (Use the critical value approach with $\alpha = 0.05$).

(b) If the firm wants to test whether the coefficient on total assets is significant, what is the relevant p-value? What decision should be made? (Use the p-value approach with $\alpha = 0.05$).

(c) Find the 95% confidence interval for the true slope of the number of employees (β_3).

4. Which multiple regression model to choose?

   (a) The multiple of coefficient determination

   (b) The standard error of estimate

EX 2 (cont.)

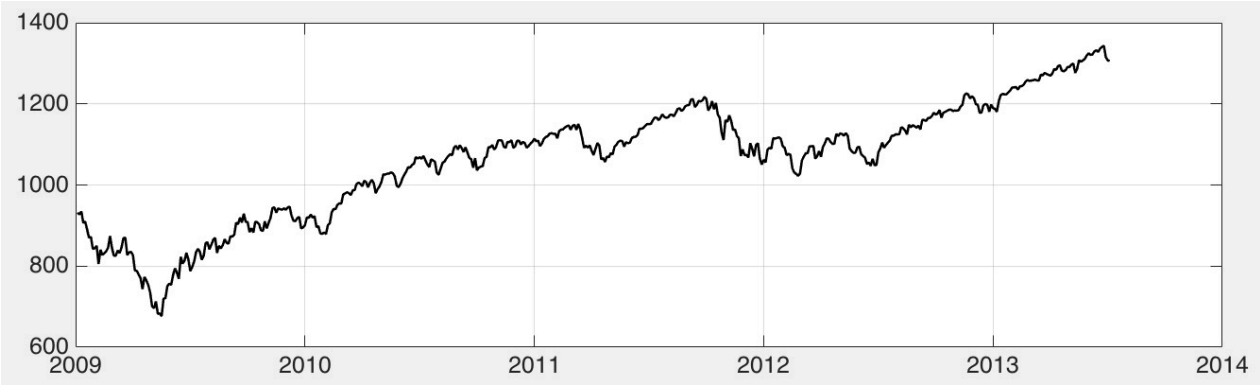
CH 17: Time Series Analysis and Forecasting

1. Basic Concepts:

   (A) Definition: A time series is a set of observations (responses), each one being recorded at a specified time.

   (B) Plotting time-series data:
   Y-axis: Response measured at the corresponding time
   X-axis: A time indicator (in years, months and so on)

   EX 1: The S&P 500 Index for year 2009 to 2013

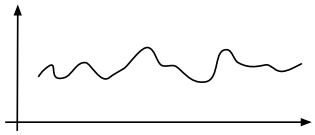


   (C) Time series analysis is used for Business Forecasting (to predict the future behavior of the estimated model). Say, you are a financial analyst and you need to forecast revenues of some companies in order to better evaluate investment opportunities for your clients.

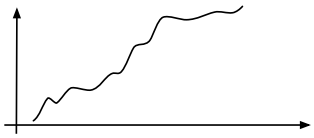
   (D) How do we do forecasting? Through identifying and isolating influence patterns of the time series.

2. Time Series Patterns

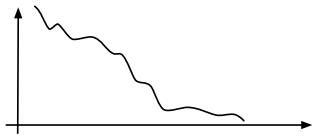
   (A) Horizontal Pattern: A horizontal pattern exists when the data fluctuate around a constant mean.



   (B) Trend Pattern: A trend is an overall or persistent, long-term upward or downward pattern of movement.

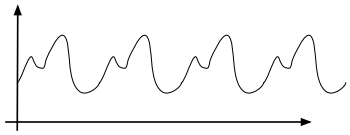


Increasing data

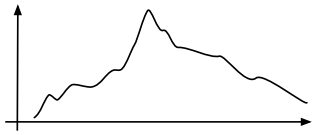


Decreasing data

   (C) Seasonal Pattern: Fairly regular periodic fluctuations that occur within some period, year after year. (Repeating patterns)

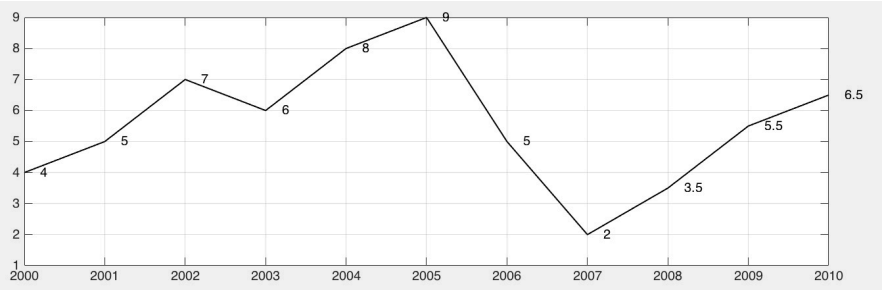


   (D) Cyclical Pattern: A wavelike pattern describing a gradual ups and downs movement that is generally apparent over a year.



3. Time-series Notation: Response: Y_t, Time $t = 1, 2, \cdots n$. Thus, at time point 1, Response is Y_1; at time point 2, Response is Y_2;..., at time point n, Response is Y_n.

EX2: The following represent total revenues (in millions) of a car rental agency over the 11-year periods from 2000 to 2010: 4.0, 5.0, 7.0, 6.0, 8.0, 9.0, 5.0, 2.0, 3.5, 5.5, and 6.5. Obtain the time-series plot.



4. Introduction to two data smoothing techniques-Moving averages and Exponential Smoothing for Forecasting.

(A) Moving Average: The moving averages method uses the average of the most recent k data values in the time series as the forecast for the next period.

eq17.1: Moving Average Forecast of order k:

$$F_{t+1} = \frac{\sum(\text{most recent } k \text{ data values})}{k} = \frac{y_t + y_{t-1} + ... + y_{t-k+1}}{k}$$

where F_{t+1} is the value of the time series being computed in time period $t + 1$.

EX2 (Cont) Find the forecasting values for year 2005, 2006, and 2007 with $k = 5$.

$$F_{2005} = \tfrac{1}{5}(Y_{2004} + Y_{2003} + Y_{2002} + Y_{2001} + Y_{2000}) = \frac{8+6+7+5+4}{5} = 6$$
$$F_{2006} = \tfrac{1}{5}(Y_{2005} + Y_{2004} + Y_{2003} + Y_{2002} + Y_{2001}) = \frac{9+8+6+7+5}{5} = 7$$
$$F_{2007} = \tfrac{1}{5}(Y_{2006} + Y_{2005} + Y_{2004} + Y_{2003} + Y_{2002}) = \frac{5+9+8+6+7}{5} = 7$$

(B) eq17.2 Exponential Smoothing Forecast

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t$$

(A recursive equation with $F_1 = y_1$)

where F_{t+1} is the value of the time series being computed in time period $t + 1$, F_t is the value of the time series being computed in time period t, and α is the subjectively assigned weight or smoothing coefficient ($0 < \alpha < 1$).

EX2 (Cont) Find the forecasting values for year 2003 and 2004 using exponential smoothing technique (use $\alpha = 0.4$, $1 - \alpha = 1 - 0.4 = 0.6$)

t	y_t	Exponential Smoothing Forecasting
2000	$y_1 = 4$	$F_1 = y_1 = 4$
2001	$y_2 = 5$	$F_2 = \alpha y_1 + (1 - \alpha)F_1 = \alpha y_1 + F_1 - \alpha y_1 = F_1 = 4.000$
2002	$y_3 = 7$	$F_3 = \alpha y_2 + (1 - \alpha)F_2 = 0.4 \times 5 + 0.6 \times 4 = 4.400$
2003	$y_4 = 6$	$F_4 = \alpha y_3 + (1 - \alpha)F_3 = 0.4 \times 7 + 0.6 \times 4.4 = 5.440$
2004	$y_5 = 8$	$F_5 = \alpha y_4 + (1 - \alpha)F_4 = 0.4 \times 6 + 0.6 \times 5.44 = 5.664$

CH 19: Statistical Method for Quality Control Part 1

1. Basic Concepts

   (A) A Process is any business activity that takes imputs and transforms them into outputs.

   (B) Quality control: A series of inspections and measurements used to determine whether quality standards are being met. i.e., we use statistical methods to detect and fix problems in a process.

   (C) We will use control chart to achieve this goal.

2. Control chart: A graphical tool used to help determine whether a process is in control or out of control. In our study, the control chart displays successive measurement of a process together with a upper control limit (UCL) and lower control limit (LCL).

3. Control Chart for the mean: the $\bar{x}$ chart.

   (A) The $\bar{x}$ is used to monitor the process average.

   (B) The UCL and LCL for $\bar{x}$ chart with standard deviation (error) known:

   eq19.1: Standard Error of the Mean: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

   eq19.2: $UCL = \mu + 3\sigma_{\bar{x}}$ eq 19.3: $LCL = \mu - 3\sigma_{\bar{x}}$

   EX1 Temperature is used to measure the output of a production process. When the process is in control, the mean of the process is $\mu = 128.5$ and the standard deviation is $\sigma = 0.4$.
   a). Compute the UCL and LCL for an $\bar{x}$ chart if sample of size 6 is provided.

   b). If the first two measurements are $\bar{x}_1 = 129.25$ and $\bar{x}_2 = 128.2$, what can you say about the process?

   (C) The UCL and LCL for $\bar{x}$ chart with standard deviation (error) unknown:

   eq19.4: Overall Sample Mean: $\overline{\bar{x}} = \dfrac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k}$

   eq19.5: Average Range: $\overline{R} = \dfrac{R_1 + R_2 + \cdots + R_k}{k}$

   eq19.8: Control limits (UCL, LCL) For an $\bar{x}$ standard deviation unknown:$\overline{\bar{x}} \pm A_2\bar{R}$

4. Control limits for an R chart:

   (A) The R chart is a control chart for the range (the largest value minus the smallest value). It's used to monitor the variation in the process.

   (B) The UCL and LCL for an R chart:

$$\text{eq19.14: } UCL = \overline{R}D_4$$

$$\text{eq19.15: } LCL = \overline{R}D_3$$

   (C) A_2, D_3, and D_4 is the control factors obtained from a table based on the sample size n.

EX 2 A toothpaste manufacturer monitors the amount of active ingredients found in a tube. Five samples are drawn each day for eight days with the following data.

(a) Find UCL and LCL for the $\bar{x}$ chart and the R chart,

(b) Is the process in control?

CH 19: Statistical Method for Quality Control Part 2

5. Review for the Control Chart: A graphical tool used to help determine whether a process is in control or out of control. In our study, the control chart displays successive measurement of a process together with a upper control limit (UCL) and lower control limit (LCL).

6. Control chart: p chart: The control chart used to monitor the proportion of defective items. Again, the control chart consists of the upper control limit (UCL), the lower control limit (LCL), the center line is also considered.

   (A) The data structure:

   (B) Find the proportion p (the center line).

   (C) Find the standard error of the proportion.

   $$\text{eq19.16: } \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

   Where n is the sample size.

   (D) Find the Control Limits for a p Chart:

   $$\text{eq19.17: } UCL = p + 3\sigma_{\bar{p}}$$

   $$\text{eq19.18: } LCL = p - 3\sigma_{\bar{p}}$$

EX3 Given the following data for a period of 10 days of a manufacture process. a sample of 100 items were randomly selected and the number of defective items were recored.

(a) Find the proportion p (the center line)

(b) Find the standard error of the proportion $\sigma_{\bar{p}}$

(c) Compute the control limits

(d) Is the process in control or out of control?

(e) Graph the p chart