Chapter 1

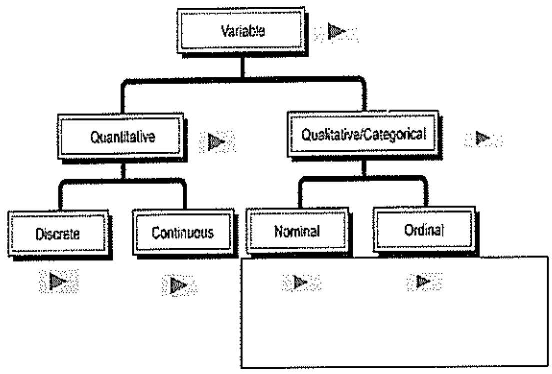
Data Description and Numerical Measures

Chapter 1 : Data Description and Numerical Measures

INTRODUCTION

1.1 WHAT IS STATISTICS

Statistics is a field of study which implies collecting, presenting, analyzing and interpreting data as a basis for explanation, description and comparison.



1.2 TYPES OF STATISTICS

Statistics can be divided into two

(i) Descriptive Statistics

(ii) Inferential Statistics

Descriptive Statistics is a field of study which involves organizing, displaying and describing data by using tables, graphs and summary measures.

Inferential Statistics consists of generalizing from samples to populations, performing hypothesis tests, determining relationships among variables, and making predictions.

1.3 POPULATION VERSUS SAMPLE

Population refers to every element in an observation which are of interest for data collection.
For example, If data on final exam results at UTeM is needed, every student in UTeM forms the population.

Sample refers to a certain number of elements that have been chosen from a population for observation. Sample in subset to population.
For example, choose any 100 students in UTeM for interviews. The sample size is 100.

ORGANIZING DATA

2.1 RAW DATA

Once data has been collected, it is crucial that the data be well presented for analysis and interpretation.

Once data has been collected, before they are processed or ranked we called ***raw data***.
Raw data also called as individual data.

2.2 ORGANIZING AND GRAPHING QUALITATIVE DATA

FREQUENCY DISTRIBUTION
Numerical data can be presented in form of a table. The data would have to be classified.

Frequency distribution is the **lists all** categories or classes and the number of elements or values that belong to each of the categories or classes.

Two types :
(i) category type - ***ungrouped frequency***

(ii) interval type - ***grouped frequency***

2.3 ORGANIZING AND GRAPHING QUANTITATIVE DATA

GROUPED FREQUENCY DISTRIBUTION

A ***class interval*** is a range of values defined by the lower class limit and upper class limit.

Class boundary is the midpoint of the upper limit of one class and the lower limit of the next class. Formulas for finding the class boundaries are as follows :

(lower class limit) - 0.5 = (lower class boundary)
(upper class limit) - 0.5 = (upper class boundary)
OR
(lower class limit) - 0.05 = (lower class boundary)
(upper class limit) - 0.05 = (upper class boundary)

Class midpoint or ***class mark*** is a average of lower class limit and upper class limit.
Formula:

$$\text{Class midpoint} = \frac{\text{upper class limit} + \text{lower class limit}}{2}$$

Range is equal to highest value minus lowest value.

Number of classes can be obtained by using Sturge's formula.

Number of classes = 1 + 3.3 log n ;
n =the number of observation in data set

Class size (Class width) can be obtained by dividing the range with a number of classes.

$$\text{Class size} = \frac{\text{Range}}{\text{number of classes}}$$

Tally marks used to count class frequency by marking strokes against each class for each data that falls in that class.

IMPORTANT NOTES

Relative frequency of a class is just the ratio of its frequency to the total frequency. Each relative frequency has value between 0 and 1, and the total of all relative frequencies would then be equal to 1.

Table 2.7: Relative Frequency Distribution for the Books on Weekly Sales

Class	34 - 43	44 - 53	54 - 63	64 - 73	74 - 83	84 - 93	94 - 103	Sum
Frequency (f)	2	5	12	18	10	2	1	50
Relative Frequency	0.04	0.1	0.24	0.36	0.20	0.04	0.02	1.00
Relative Frequency (%)	4	10	24	36	20	4	2	100

2.5 CUMULATIVE FREQUENCY DISTRIBUTIONS

Cumulative frequencies are obtained by finding the total number of values or frequency that fall below the upper class boundary of each class.

Formula:

$$\text{Cummulative relative frequency} = \frac{\text{cummulative frequency of each class}}{\text{sum of all frequencies}}$$

We can use **cummulative frequency** or **cumulative relative frequency** to represent the vertical axis.

CUMULATIVE FREQUENCY DISTRIBUTION

The total frequency of all values less than the upper class boundary of a given class is called a cumulative frequency up to and including the upper limit of that class.

Table 2.9: The "Less-than or Equal" Cumulative Distribution for the Books on Weekly Sales

Upper Boundary	Cumulative Frequency	Cumulative Frequency (%)
≤ 33.5	0	0
≤ 43.5	2	4
≤ 53.5	7	14
≤ 63.5	19	38
≤ 73.5	37	74
≤ 83.5	47	94
≤ 93.5	49	98
≤ 103.5	50	100

GRAPHING GROUPED DATA

After the data have been organized into a frequency distribution, they can be represented in graphic forms such as histograms, frequency polygon, and ogives

BAR CHART

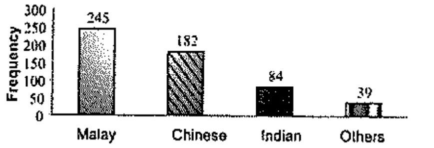


Figure 3.1: Bar chart for the number of students by their ethnic background in School J.

Pie Chart

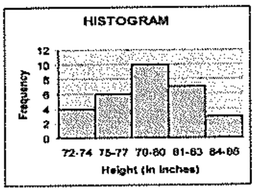
Figure 3.3(a): Using frequency to develop the Pie Chart

Figure 3.3(b): Using Relative Frequency to develop the Pie Chart

HISTOGRAMS

A graphical representation of a grouped frequency distribution.
class intervals - horizontal axis
frequency - vertical axis.

It is obtained by adjoining rectangles, the width of each rectangle is the size of each class and the height of each rectangle is the frequency of the class interval. The area of each rectangle is important.

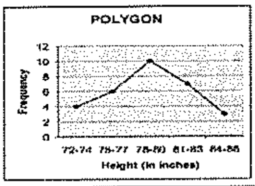


FREQUENCY POLYGONS AND CURVE

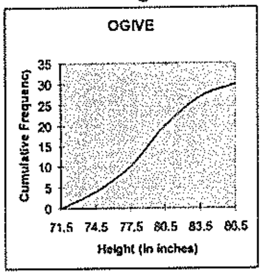
It is obtained by connecting with straight lines the midpoints of adjacent class intervals of histogram
A frequency curve is obtained by smoothing the corners of a frequency polygon.

Relative frequency = frequency of each class / Sum of all frequencies

We can use **frequency** or **relative frequency** to represent the vertical axis.



OGIVES
These are the graphical representations of a cumulative frequency distribution. Ogive can be drawn by joining with straight lines the dots marked above the upper boundaries of classes at heights.



Example
Below is the distance in km of a random sample of 50 employees in Z company who traveled to work each day.

1	2	6	7	12	13	2	6	9	5
18	7	3	15	15	4	17	1	14	5
4	16	4	5	8	6	5	18	5	2
9	11	12	1	9	2	10	11	4	10
9	18	8	8	4	14	7	3	2	6

i) Construct a frequency distribution table.
ii) Construct a histogram.
iii) Construct a frequency polygon
iv) Construct a relative frequency polygon.
v) Construct an ogive.
vi) Find the mean, variance and standard deviation for this data set. (give the answer in 4 decimal places).

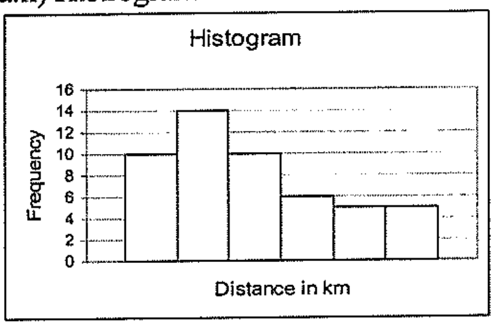
Solution
a.i) Determine the number of classes and class width using Sturge's formula.

Highest value = 18
Lowest value = 1
Number of classes = 1 + 3.3 log n n=the number of observation in data set
= 1 + 3.3 log 50
= 6.61
≈ 6 classes

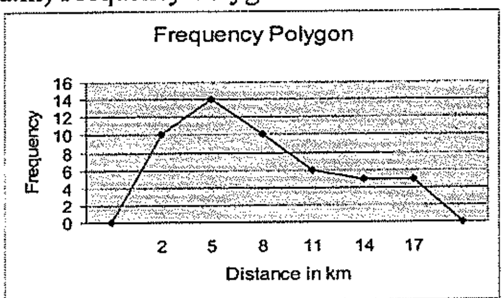
Class width = $\frac{\text{Highest value - Lowest value}}{\text{Numbers of classes}}$
$= \frac{18-1}{6}$
= 2.8
≈ 3

Class limit	Tally	Frequency, f	Cumulative Frequency, cf	Class boundaries	Midpoints, m	fm	fm^2
1-3	卌 卌	10	10	0.5-3.5	2	20	40
4-6	卌 卌 ||||	14	24	3.5-6.5	5	70	350
7-9	卌 卌	10	34	6.5-9.5	8	80	640
10-12	卌 |	6	40	9.5-12.5	11	66	726
13-15	卌	5	45	12.5-15.5	14	70	980
16-18	卌	5	50	15.5-18.5	17	85	1445
		$\Sigma f =$				$\sum fm =$	$\sum fm^2 =$

a.ii) Histrogram

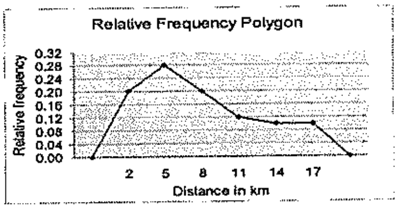


a.iii) Frequency Polygon



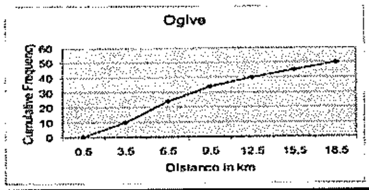
a. iv)

Class boundary	Midpoint	Frequency	Relative Frequency
0.5-3.5	2	10	0.20
3.5-6.5	5	14	0.28
6.5-9.5	8	10	0.20
9.5-12.5	11	6	0.12
12.5-15.5	14	5	0.10
15.5-18.5	17	5	0.10



a.v)

Class boundary	Midpoint	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5-3.5	2	10	0.20	0.20
3.5-6.5	5	14	0.28	0.48
6.5-9.5	8	10	0.20	0.68
9.5-12.5	11	6	0.12	0.80
12.5-15.5	14	5	0.10	0.90
15.5-18.5	17	5	0.10	1.00



a.vi)

Sample mean, $\bar{x} = \frac{\sum fm}{\sum f}$

$= \frac{391}{50}$

$= 7.8200$

Variance,

$$s^2 = \frac{\sum fm^2}{\sum f - 1} - \frac{(\sum fm)^2}{\sum f(\sum f - 1)}$$

$= \frac{4181}{50-1} - \frac{(391)^2}{50(49)}$

$= 22.9261$

Standard deviation, $s = \sqrt{22.9261}$

$= 4.7881$

NUMERICAL DESCRIPTIVE MEASURES

3.1 MEASURES OF CENTRAL TENDENCY

The three common measures of central tendency are mean, median and mode.

MEAN

The mean is the average

The mean from sample is denoted by $\bar{x}$

The mean from population is denoted by .

Calculation of mean

(i) Individual data

25, 40 ,15, 60 , 50 , 50, 60 , 40 , 35 ,

Formula:

$\bar{x} = \frac{\sum x}{n}$ or $\mu = \frac{\sum x}{N}$

(ii) Ungrouped frequency

Formula:

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

Data , X	Frequency ,f
45	2
50	4
60	1
85	5

(iii) Grouped frequency

Formula:

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

Data	Midpoint, m	Frequency, f
20 - 30	25	2
30 - 40	35	5
40 - 50	45	3
50 - 60	55	1

MEDIAN

The median is the value of the item which is located at the center of the distribution.

Calculation of the median.

(i) Individual data

$$\text{Location} = \frac{n+1}{2} \text{ th term}$$

(ii) Ungrouped data

$$\text{Location of median} = \frac{n+1}{2} \text{ th term}$$

Example

Ten customers purchased the following number of magazines: 1,7,5,3,6,2,3,1,5,8. Find the median.

Solution

1,1,2,3,3,5,5,6,7,8

↑

Median

Hence, the median $= \frac{3+5}{2}$

$= 4$

MODE

The mode is the value, which occurs most frequently in a distribution.

(i) Individual data

Identify the data with the highest occurrence.

Note:

In any set of data may be there is no mode, or one or more than one mode.

(ii) Ungrouped frequency

Identify the data with the highest occurrence.

Example

Find the mode for below numbers:

110 , 731 , 1031 , 84 , 20 , 118 , 1162 , 1977 , 103 , 752

Solution:

Since each value occurs only once, there is no mode.

Note: Do not say that the mode is zero. That would be incorrect, because in some data , zero can be an actual value.

3.2 MEASURES OF DISPERSION

RANGE

The range is the difference between highest and lowest value in the distribution.

Formula:

Range = highest value - lowest value

VARIANCE AND STANDARD DEVIATION

The standard deviation measures the spread of the data as compared to the mean.

(i) Individual data

Formula:

$$\sigma^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 \qquad s^2 = \frac{\sum x^2}{n-1} - \frac{(\sum x)^2}{n(n-1)}$$

(ii) Ungrouped frequency

Formula:

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 \qquad s^2 = \frac{\sum fx^2}{\sum f - 1} - \frac{(\sum fx)^2}{\sum f(\sum f - 1)}$$

(iii) Grouped frequency

Formula:

$$\sigma^2 = \frac{\sum fm^2}{\sum f} - \left(\frac{\sum fm}{\sum f}\right)^2 \qquad s^2 = \frac{\sum fm^2}{\sum f - 1} - \frac{(\sum fm)^2}{\sum f(\sum f - 1)}$$

where the m = Midpoint

Example

The following exam score frequency distribution was obtained from all the students in ABC college.

Class limits	Frequency, f	Cumulative frequency	Midpoint, m	fm	fm^2
90-98	6	6	94	564	53 016
99-107	22	28	103	2266	233 398
108-116	43	71	112	4816	539 392
117-125	28	99	121	3388	409 948
126-134	9	108	130	1170	152 100
	$\sum f = 108$			$\sum fm = 12204$	$\sum fm^2 = 1387854$

Find the (a) mean (b) median (c) mode (d) standard deviation

Solution

(a) mean, $\mu = \frac{\sum fm}{\sum f}$

$= \frac{12204}{108}$

$= 113$

(b) Location of median $= \frac{n+1}{2} th$ term

$= \frac{108+1}{2}$

$= 54.4$

The median class is 107.5-116.5. Sometimes, the class limits is used. Hence, the median class could also given as 108-116.

(c) The modal class is 107.5-116.5 since it has the largest frequency.

Note: Sometimes the midpoint of the class is used rather than the boundaries; hence, the mode could also be given as 112.0.

(d) standard deviation $= \sqrt{\frac{\sum fm^2}{\sum f} - \left(\frac{\sum fm}{\sum f}\right)^2}$

$= \sqrt{\frac{1387854}{108} - \left(\frac{12204}{108}\right)^2}$

$= \sqrt{81.5} = 9.03$

Statistics for Managers Using Microsoft® Excel 5th Edition

Chapter 2
Presenting Data in Tables and Charts

Learning Objectives

In this chapter, you will learn:

- To develop tables and charts for categorical data
- To develop tables and charts for numerical data
- The principles of properly presenting graphs

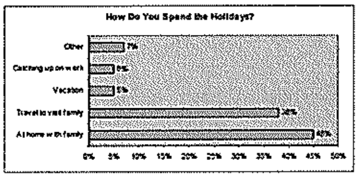
Organizing Categorical Data: Summary Table

- A **summary table** indicates the frequency, amount, or percentage of items in a set of categories so that you can see differences between categories.

How do you spend the holidays?	Percent
At home with family	45%
Travel to visit family	38%
Vacation	5%
Catching up on work	5%
Other	7%

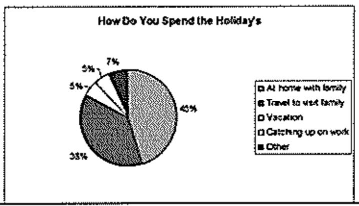
Organizing Categorical Data: Bar Chart

- In a **bar chart**, a bar shows each category, the length of which represents the amount, frequency or percentage of values falling into a category.



Organizing Categorical Data: Pie Chart

- The pie chart is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.

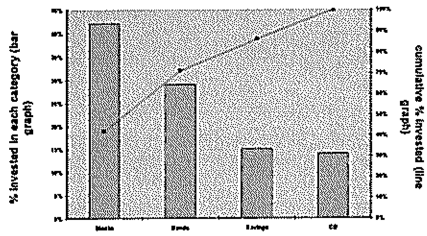


Organizing Categorical Data: Pareto Diagram

- Used to portray categorical data
- A bar chart, where categories are shown in descending order of frequency
- A cumulative polygon is shown in the same graph
- Used to separate the "vital few" from the "trivial many"

Organizing Categorical Data: Pareto Diagram

Current Investment Portfolio



Organizing Numerical Data: Ordered Array

- An **ordered array** is a sequence of data, in rank order, from the smallest value to the largest value.

Age of Surveyed College Students						
	Day Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Organizing Numerical Data: Stem and Leaf Display

- A **stem-and-leaf display** organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.

Age of College Students

Day Students

Stem	Leaf
1	67788899
2	0012257
3	28
4	2

Night Students

Stem	Leaf
1	8899
2	0138
3	23
4	15

Organizing Numerical Data: Frequency Distribution

- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered class groupings.
- You must give attention to selecting the appropriate *number* of class groupings for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.
- To determine the **width of a class interval**, you divide the range (Highest value–Lowest value) of the data by the number of class groupings desired.

Organizing Numerical Data: Frequency Distribution Example

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Organizing Numerical Data: Frequency Distribution Example

- Sort raw data in ascending order:
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: **58 - 12 = 46**
- Select number of classes: **5 (usually between 5 and 15)**
- Compute class interval (width): **10 (46/5 then round up)**
- Determine class boundaries (limits): **10, 20, 30, 40, 50, 60**
- Compute class midpoints: **15, 25, 35, 45, 55**
- Count observations & assign to classes

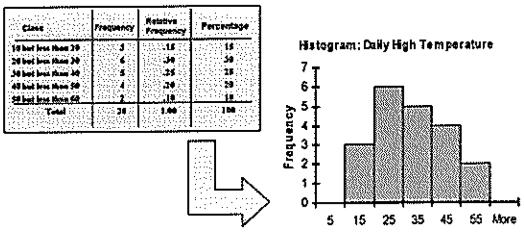
Organizing Numerical Data: Frequency Distribution Example

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

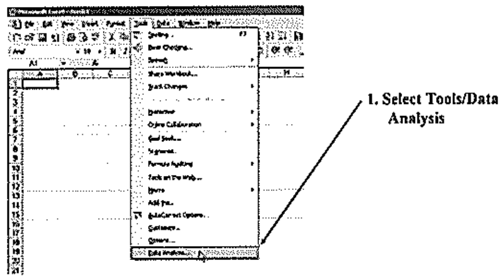
Organizing Numerical Data: The Histogram

- A graph of the data in a frequency distribution is called a histogram.
- The class boundaries (or class midpoints) are shown on the horizontal axis.
- The vertical axis is either frequency, relative frequency, or percentage.
- Bars of the appropriate heights are used to represent the number of observations within each class.

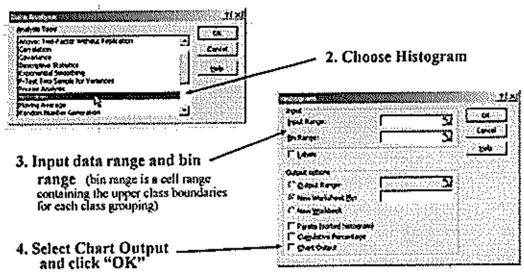
Organizing Numerical Data: The Histogram



Organizing Numerical Data: The Histogram in Excel



Organizing Numerical Data: The Histogram in Excel



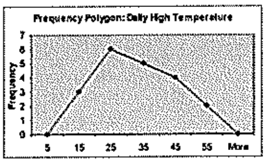
Organizing Numerical Data: The Polygon

- A percentage polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages.
- The cumulative percentage polygon, or ogive, displays the variable of interest along the X axis, and the cumulative percentages along the Y axis.

Organizing Numerical Data: The Polygon

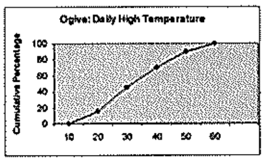
Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)



Organizing Numerical Data: The Cumulative Percentage Polygon

Class	Lower Boundary	% Less Than Lower Boundary
10<20	10	0
20<30	20	15
30<40	30	45
40<50	40	70
50<60	50	90
	60	100



Cross Tabulations: The Contingency Table

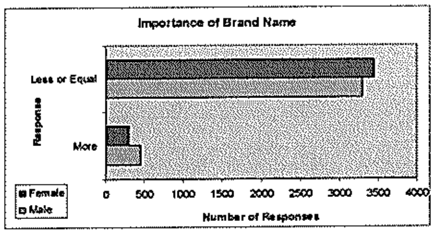
- A **cross-classification (or contingency) table** presents the results of two categorical variables. The joint responses are classified so that the categories of one variable are located in the rows and the categories of the other variable are located in the columns.
- The cell is the intersection of the row and column and the value in the cell represents the data corresponding to that specific pairing of row and column categories.
- A useful way to visually display the results of cross-classification data is by constructing a **side-by-side bar chart.**

Cross Tabulations: The Contingency Table

A survey was conducted to study the importance of brand name to consumers as compared to a few years ago. The results, classified by gender, were as follows:

Importance of Brand Name	Male	Female	Total
More	450	300	750
Equal or Less	3300	3450	6750
Total	3750	3750	7500

Cross Tabulations: Side-By-Side Bar Charts

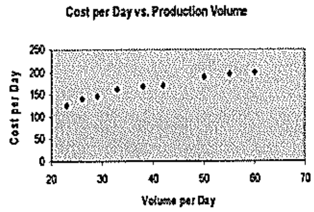


Scatter Plots

- Scatter plots are used for numerical data consisting of paired observations taken from two numerical variables
- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis

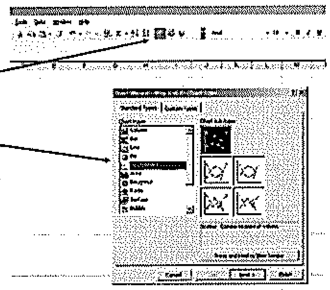
Scatter Plot Example

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



Scatter Plot in Excel (97-2003)

1. Select the chart wizard
2. Select XY(Scatter) option, then click "Next"
3. When prompted, enter the data range, then click "Next".
4. Enter Title, Axis Labels, and Legend and click "Finish"



Time Series

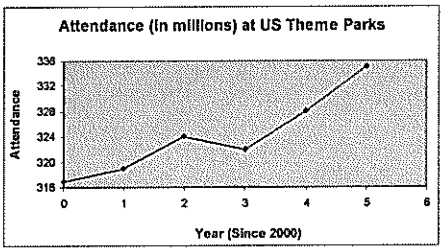
- A time-series plot is used to study patterns in the values of a numerical variable over time. Each value is plotted as a point in two dimensions with the time period on the horizontal *X* axis and the variable of interest on the *Y* axis.

Time Series Example

Attendance (in millions) at USA amusement/theme parks from 2000-2005

Year	Year Number	Attendance
2000	0	317
2001	1	319
2002	2	324
2003	3	322
2004	4	328
2005	5	335

Time Series Example

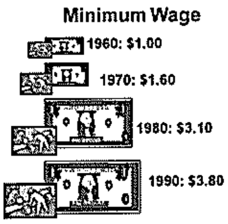


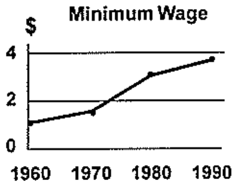
Principles of Excellent Graphs

- The graph should not distort the data.
- The graph should not contain unnecessary adornments (sometimes referred to as chart junk).
- The scale on the vertical axis should begin at zero.
- All axes should be properly labeled.
- The graph should contain a title.
- The simplest possible graph should be used for a given set of data.

Graphical Errors: Chart Junk

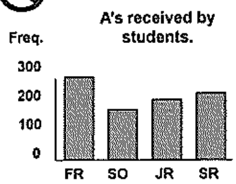
Bad Presentation | **✓ Good Presentation**

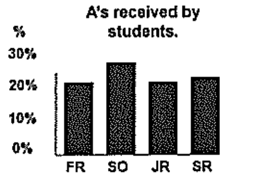




Graphical Errors: No Relative Basis

Bad Presentation | **✓ Good Presentation**

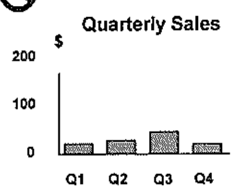




FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

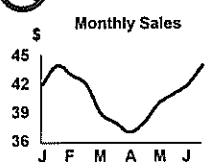
Graphical Errors: Compressing the Vertical Axis

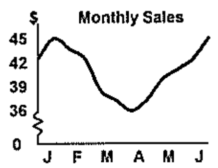
Bad Presentation | **✓ Good Presentation**



Graphical Errors: No Zero Point on the Vertical Axis

Bad Presentation | **✓ Good Presentations**





Graphing the first six months of sales

Chapter Summary

In this chapter, we have

- Organized categorical data using the summary table, bar chart, pie chart, and Pareto diagram.
- Organized numerical data using the ordered array, stem and leaf display, frequency distribution, histogram, polygon, and ogive.
- Examined cross tabulated data using the contingency table and side-by-side bar chart.
- Developed scatter plots and time series graphs.
- Examined the do's and don'ts of graphically displaying data.

Fundamentals of Hypothesis Testing:

One Sample Tests
Population Mean

The Hypothesis

- A hypothesis is a claim (assumption) about a population parameter:
  - population mean

    Example: The mean monthly cell phone bill of this city is $\mu = \$52$

The Null Hypothesis, H_0

- States the assumption (numerical) to be tested

  Example: The mean number of TV sets in U.S. Homes is equal to three.

  $$H_0 : \mu = 3$$

- Is always about a population parameter, not about a sample statistic.

The Null Hypothesis, H_0

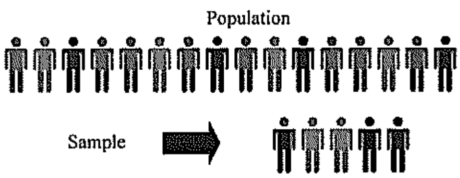
- Begin with the assumption that the null hypothesis is true
  - Similar to the notion of innocent until proven guilty
- It refers to the status quo
- Always contains "=" , "≤" or "≥" sign
- May or may not be rejected

The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis
  - e.g., The mean number of TV sets in U.S. homes is not equal to 3 ($H_1: \mu \neq 3$)
- Contains the " ≠" , "<" or ">" sign
- May or may not be proven

The Hypothesis Testing Process

- Claim: The population mean age is 50.
  - $H_0: \mu = 50$, $H_1: \mu \neq 50$
- Sample the population and find sample mean.



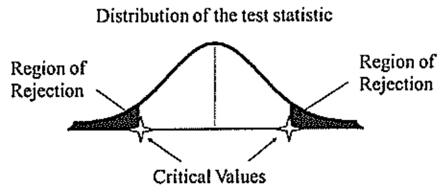
The Hypothesis Testing Process

- Suppose the sample mean age was $\overline{X} = 20$.
- This is significantly lower than the claimed mean population age of 50.
- If the null hypothesis were true, the probability of getting such a different sample mean would be very small, so you reject the null hypothesis.
- In other words, getting a sample mean of 20 is so unlikely if the population mean was 50, you conclude that the population mean must not be 50.

The Test Statistic and Critical Values

- If the sample mean is close to the assumed population mean, the null hypothesis is not rejected.
- If the sample mean is far from the assumed population mean, the null hypothesis is rejected.
- How far is "far enough" to reject H_0?
- The critical value of a test statistic creates a "line in the sand" for decision making.

The Test Statistic and Critical Values



Errors in Decision Making

- **Type I Error**
  - Reject a true null hypothesis
  - Considered a serious type of error
  - The probability of a Type I Error is α
    - Called level of significance of the test
    - Set by researcher in advance
- **Type II Error**
  - Failure to reject false null hypothesis
  - The probability of a Type II Error is β

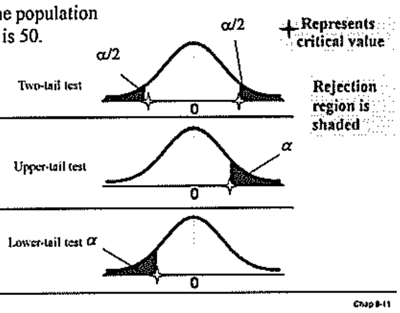
Level of Significance, α

Claim: The population mean age is 50.

H_0: $\mu = 50$
H_1: $\mu \neq 50$

H_0: $\mu \leq 50$
H_1: $\mu > 50$

H_0: $\mu \geq 50$
H_1: $\mu < 50$



Hypothesis Testing: σ Known

For two tail test for the mean, σ known:

- Convert sample statistic ($\overline{X}$) to test statistic

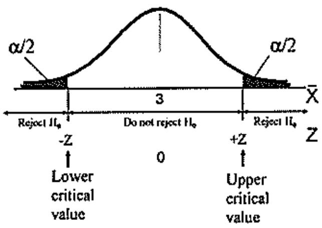
$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Determine the critical Z values for a specified level of significance α from a table or by using Excel
- Decision Rule: If the test statistic falls in the rejection region, reject H_0 ; otherwise do not reject H_0

Hypothesis Testing: σ Known

$H_0: \mu = 3$
$H_1: \mu \neq 3$

- There are two cutoff values (critical values), defining the regions of rejection



Hypothesis Testing: σ Known

Example: Test the claim that the true mean weight of chocolate bars manufactured in a factory is 3 ounces.

- State the appropriate null and alternative hypotheses
  - $H_0: \mu = 3$ $H_1: \mu \neq 3$ (This is a two tailed test)
- Specify the desired level of significance
  - Suppose that $\alpha = .05$ is chosen for this test
- Choose a sample size
  - Suppose a sample of size n = 100 is selected

Hypothesis Testing: σ Known

- Determine the appropriate technique
  - σ is known so this is a Z test
- Set up the critical values
  - For $\alpha = .05$ the critical Z values are ±1.96
- Collect the data and compute the test statistic
  - Suppose the sample results are

    $n = 100$, $\bar{X} = 2.84$

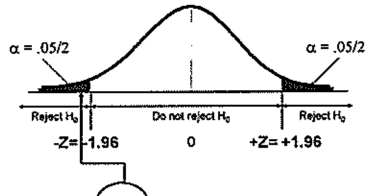
    ($\sigma = 0.8$ is assumed known from past company records)

  So the test statistic is: $$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0$$

Hypothesis Testing: σ Known

- Is the test statistic in the rejection region?

Reject H_0 if $Z < -1.96$ or $Z > 1.96$; otherwise do not reject H_0



Here, Z = -2.0 < -1.96, so the test statistic is in the rejection region

Hypothesis Testing: σ Known

- Reach a decision and interpret the result
  - Since $Z = -2.0 < -1.96$, you reject the null hypothesis and conclude that there is sufficient evidence that the mean weight of chocolate bars is not equal to 3.

Hypothesis Testing: σ Known

6 Steps of Hypothesis Testing:

1. State the null hypothesis, H_0 and state the alternative hypotheses, H_1
2. Choose the level of significance, α, and the sample size n.
3. Determine the appropriate statistical technique and the test statistic to use
4. Find the critical values and determine the rejection region(s)

Hypothesis Testing: σ Known

5. Collect data and compute the test statistic from the sample result
6. Compare the test statistic to the critical value to determine whether the test statistic falls in the region of rejection. Make the statistical decision: Reject H_0 if the test statistic falls in the rejection region. Express the decision in the context of the problem

Chap 9-19

Hypothesis Testing: σ Known Confidence Interval Connections

- For $\overline{X} = 2.84$, $\sigma = 0.8$ and $n = 100$, the 95% confidence interval is:

$$2.84 - (1.96)\frac{0.8}{\sqrt{100}} \text{ to } 2.84 + (1.96)\frac{0.8}{\sqrt{100}}$$

$$2.6832 \leq \mu \leq 2.9968$$

- Since this interval does not contain the hypothesized mean (3.0), you reject the null hypothesis at $\alpha = .05$

Chap 9-20

Hypothesis Testing: σ Known One Tail Tests

- In many cases, the alternative hypothesis focuses on a particular direction

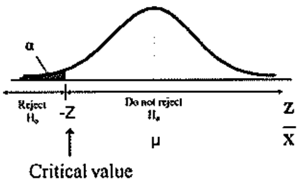
$H_0: \mu \geq 3$
$H_1: \mu < 3$
→ This is a lower-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

$H_0: \mu \leq 3$
$H_1: \mu > 3$
→ This is an upper-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3

Chap 9-21

Hypothesis Testing: σ Known Lower Tail Tests

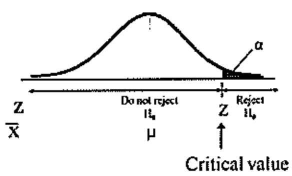
- There is only one critical value, since the rejection area is in only one tail.



Chap 9-22

Hypothesis Testing: σ Known Upper Tail Tests

- There is only one critical value, since the rejection area is in only one tail.



Chap 9-23

Hypothesis Testing: σ Known Upper Tail Test Example

A phone industry manager thinks that customer monthly cell phone bills have increased, and now average more than $52 per month. The company wishes to test this claim. Past company records indicate that the standard deviation is about $10.

Form hypothesis test:

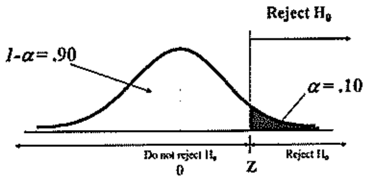
$H_0: \mu \leq 52$ the mean is less than or equal to than $52 per month

$H_1: \mu > 52$ the mean is greater than $52 per month (i.e., sufficient evidence exists to support the manager's claim)

Chap 9-24

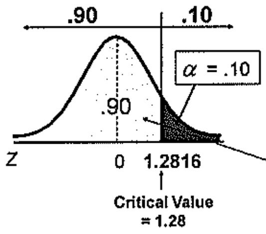
Hypothesis Testing: σ Known Upper Tail Test Example

- Suppose that $\alpha = .10$ is chosen for this test
- Find the rejection region:



Hypothesis Testing: σ Known Upper Tail Test Example

What is Z given $\alpha = 0.10$?



Standard Normal Distribution Table 4 (Portion)

x	$\Phi(x)$
1.20	0.8849
.21	.8869
.22	.8888
.23	.8907
.24	.8925
1.25	0.8944
.26	.8962
.27	.8980
.28	.8997
.29	.9015

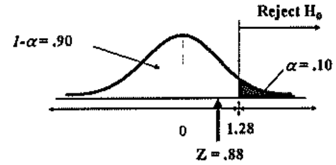
Hypothesis Testing: σ Known Upper Tail Test Example

- Obtain sample and compute the test statistic.
- Suppose a sample is taken with the following results: $n = 64$, $\bar{X} = 53.1$ ($\sigma = 10$ was assumed known from past company records)
  - Then the test statistic is:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{53.1 - 52}{\frac{10}{\sqrt{64}}} = 0.88$$

Hypothesis Testing: σ Known Upper Tail Test Example

- Reach a decision and interpret the result:



Do not reject H_0 since $Z = 0.88 \le 1.28$

i.e.: there is not sufficient evidence that the mean bill is greater than $52

Hypothesis Testing: σ Unknown

- If the population standard deviation is unknown, you instead use the sample standard deviation S.
- Because of this change, you use the t distribution instead of the Z distribution to test the null hypothesis about the mean.
- All other steps, concepts, and conclusions are the same.

Hypothesis Testing: σ Unknown

- Recall that the t test statistic with n-1 degrees of freedom is:

$$t_{n-1} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Hypothesis Testing: σ Unknown Example

The mean cost of a hotel room in New York is said to be $168 per night. A random sample of 25 hotels resulted in $\bar{X}$ = $172.50 and S = 15.40. Test at the $\alpha = 0.05$ level.

(A stem-and-leaf display and a normal probability plot indicate the data are approximately normally distributed)

H_0: μ = 168
H_1: μ ≠ 168

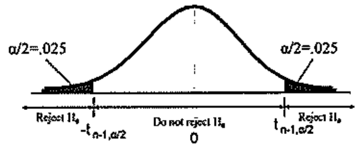
Chap 8-31

Hypothesis Testing: σ Unknown Example

H_0: μ = 168
H_1: μ ≠ 168

- α = 0.05
- n = 25
- σ is unknown, so use a t statistic
- Critical Value: $t_{24} = \pm 2.064$

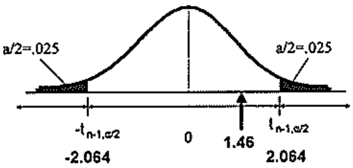
Determine the regions of rejection



Chap 8-32

Hypothesis Testing: σ Unknown Example

$$t_{n-1} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{172.50-168}{\frac{15.40}{\sqrt{25}}} = 1.46$$



Do not reject H_0: not sufficient evidence that true mean cost is different from $168

Chap 8-33

Chap 8-34

Two-Sample Tests for Population Mean

Two Sample Tests Overview

Two Sample Tests

Independent Population Means

Examples:

Group 1 vs. Group 2

Two Sample Tests

Independent Population Means *

- σ_1 and σ_2 known
- σ_1 and σ_2 unknown

Goal: Test hypothesis or form a confidence interval for the difference between two population means, $\mu_1 - \mu_2$

The point estimate for the difference is

$$\overline{X}_1 - \overline{X}_2$$

Two-Sample Tests Independent Populations

Independent Population Means *

- σ_1 and σ_2 known → Use a Z test statistic
- σ_1 and σ_2 unknown → Use S to estimate unknown σ, use a t test statistic

σ_1 and σ_2 Known

Independent Population means

- σ_1 and σ_2 known *
- σ_1 and σ_2 unknown

Assumptions:

- Samples are randomly and independently drawn
- Population distributions are normal

σ_1 and σ_2 Known

Independent Population means

- σ_1 and σ_2 known *
- σ_1 and σ_2 unknown

When σ_1 and σ_2 are known and both populations are normal or both sample sizes are at least 30, the test statistic is a Z-value...

...and the standard error of $\overline{X}_1 - \overline{X}_2$ is

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

σ_1 and σ_2 Known

Independent Population Means

- σ_1 and σ_2 known *
- σ_1 and σ_2 unknown

The test statistic for $\mu_1 - \mu_2$ is:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Two-Sample Tests Independent Populations

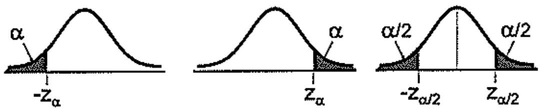
Two Independent Population, Comparing Means

Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \mu_1 \geq \mu_2$	$H_0: \mu_1 \leq \mu_2$	$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 < \mu_2$	$H_1: \mu_1 > \mu_2$	$H_1: \mu_1 \neq \mu_2$
i.e.,	i.e.,	i.e.,
$H_0: \mu_1 - \mu_2 \geq 0$	$H_0: \mu_1 - \mu_2 \leq 0$	$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 < 0$	$H_1: \mu_1 - \mu_2 > 0$	$H_1: \mu_1 - \mu_2 \neq 0$

Two-Sample Tests Independent Populations

Two Independent Population, Comparing Means

Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \mu_1 - \mu_2 \geq 0$	$H_0: \mu_1 - \mu_2 \leq 0$	$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 < 0$	$H_1: \mu_1 - \mu_2 > 0$	$H_1: \mu_1 - \mu_2 \neq 0$



Reject H_0 if $Z < -Z_\alpha$ | Reject H_0 if $Z > Z_\alpha$ | Reject H_0 if $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$

Worked Example 1

Two types of chocolate chips are suitable for use in decorating cakes. The melting points of these chocolate chips are important. It is known that $\sigma_1 = \sigma_2 = 1.0°C$. From a random sample of size $n_1 = 12$ and $n_2 = 15$, we obtain $\bar{x}_1 = 50°C$ and $\bar{x}_2 = 40°C$. The bakery will use chocolate chip 1 if its mean melting point exceeds that of chocolate chip 2 by at least $12°C$. Based on the sample information, should the bakery use chocolate chip 1? Use $\alpha = 0.05$ to make the decision.

Solution:

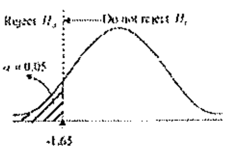
$\sigma_1 = \sigma_2 = 1.0°C$

$n_1 = 12$ $\quad n_2 = 15$

$\bar{x}_1 = 50$ $\quad \bar{x}_2 = 40$

$H_0: \mu_1 - \mu_2 \geq 12$

$H_1: \mu_1 - \mu_2 < 12$



Since σ_1 and σ_2 are known, we use -

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1^2}{12} + \frac{1^2}{15}} = 0.387298$$

From the normal distribution table, at $\alpha = 0.05$, $z = -1.65$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(50 - 40) - 12}{0.387298} = -5.16$$

The value of the test statistic $z = -5.16$ is smaller than the critical value of $z = -1.65$ and it falls in the rejection region. As a result, we reject H_0 at $\alpha = 0.05$. Therefore, the mean melting point of chocolate chip 1 does not exceed chocolate chip 2 by at least $12°C$. The bakery should not use chocolate chip 1.

σ_1 and σ_2 Unknown

Independent Population Means

- σ_1 and σ_2 known
- σ_1 and σ_2 unknown *

Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed or both sample sizes are at least 30
- Population variances are unknown but assumed equal

σ_1 and σ_2 Unknown

Independent Population Means
- σ_1 and σ_2 known
- σ_1 and σ_2 unknown *

Forming interval estimates:

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate σ
- the test statistic is a t value with $(n_1 + n_2 - 2)$ degrees of freedom

σ_1 and σ_2 Unknown

Independent Population Means
- σ_1 and σ_2 known
- σ_1 and σ_2 unknown *

The pooled standard deviation is

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

σ_1 and σ_2 Unknown

(continued)

Independent Population Means
- σ_1 and σ_2 known
- σ_1 and σ_2 unknown *

The test statistic for $\mu_1 - \mu_2$ is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

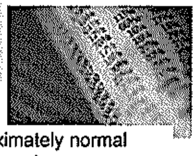
Where t has $(n_1 + n_2 - 2)$ d.f., and

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Pooled Variance t Test: Example

You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number	21	25
Sample mean	3.27	2.53
Sample std dev	1.30	1.16



Assuming both populations are approximately normal with equal variances, is there a difference in average yield ($\alpha = 0.05$)?

Calculating the Test Statistic

The test statistic is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(3.27 - 2.53) - 0}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = \boxed{2.040}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(21 - 1)1.30^2 + (25 - 1)1.16^2}{(21 - 1) + (25 - 1)} = 1.5021$$

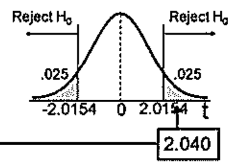
Solution

H_0: $\mu_1 - \mu_2 = 0$ i.e. $(\mu_1 = \mu_2)$
H_1: $\mu_1 - \mu_2 \neq 0$ i.e. $(\mu_1 \neq \mu_2)$

$\alpha = 0.05$

df = 21 + 25 - 2 = 44

Critical Values: t = ± 2.0154



Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = \boxed{2.040}$$

Decision:
Reject H_0 at $\alpha = 0.05$

Conclusion:
There is evidence of a difference in means.

13-1

Chapter 9
Simple Linear Regression

13-2

Simple Linear Regression

GOALS

When you have completed this topic, you will be able

ONE

Draw a scatter diagram.

TWO

Understand and interpret the terms *dependent* variable and *independent* variable.

THREE

Calculate and interpret the coefficient of correlation and the coefficient of determination.

FOUR

Calculate the least squares regression line and interpret the slope and intercept values.

Goals

Correlation Analysis

is a group of statistical techniques to measure the association between two variables.

A Scatter Diagram is a chart that portrays the relationship between two variables.

The Independent Variable provides the basis for estimation.

The Dependent Variable is the variable being predicted or estimated.

Correlation Analysis

Correlation Coefficient (r)

is to measure the strength of the linear relationship between two variables.

Also called the Pearson's Product Moment Correlation Coefficient

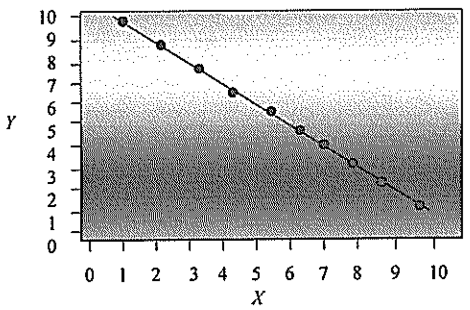
It can range from -1.00 to 1.00.

Values close to 0.0 indicate weak correlation.

r = 1	indicates a perfect positive relationship
r = 0	no relationship/independent
r = -1	indicates a perfect negative relationship

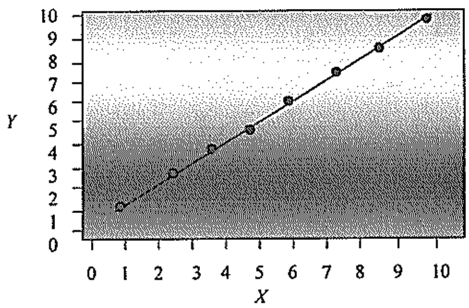
The Coefficient of Correlation, *r*

13-5



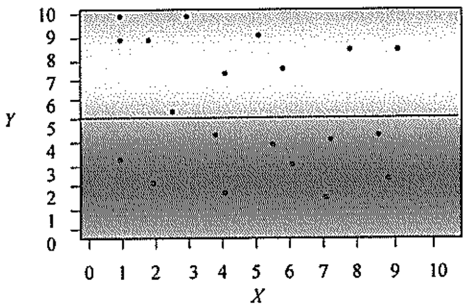
Perfect Negative Correlation

13-6



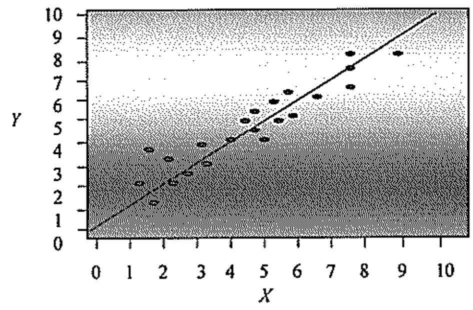
Perfect Positive Correlation

13-7



Zero Correlation

13-8



Strong Positive Correlation

13-9

Formula to Calculate the Coefficient of Correlation

$$r = \frac{\sum XY - \left[\frac{\sum X \sum Y}{n}\right]}{\sqrt{\left[\sum X^2 - \left(\frac{(\sum X)^2}{n}\right)\right]\left[\sum Y^2 - \left(\frac{(\sum Y)^2}{n}\right)\right]}}$$

Use calculator to obtain this value.

Formula for *r*

Coefficient of Determination (r^2) is the proportion of the total variation in the dependent variable (*Y*) that is explained by the variation in the independent variable (*X*).

It is the square of the coefficient of correlation.
It ranges from 0 to 1.
It does not give any information on the direction of the relationship between the variables.

Coefficient of Determination

13-11

Penerbit Universiti from UTeM is concerned about the cost to students of textbooks. He believes there is a relationship between the number of pages in the text and the selling price of the book. To provide insight into the problem he selects a sample of 8 textbooks currently on sale in the bookstore. Draw a scatter diagram, compute the Correlation Coefficient & the Coefficient of Determination.

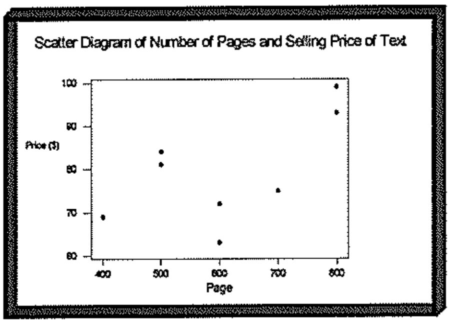
Example 1

12

Book	Page	Price(RM)
Introduction to Statistics	500	84
Basic Algebra	700	75
Introduction to Psychology	800	99
Introduction to Sociology	600	72
Business Management	400	69
Introduction to Biology	500	81
Fundamentals of Finance	600	63
Principles of Marketing	800	93

Example 1 continued

13-13



Example 1 continued

Formula to Calculate the Correlation Coefficient

$n = 8$	$\sum XY = 397{,}200$
$\sum X = 4900$	$\sum X^2 = 3{,}150{,}000$
$\sum Y = 636$	$\sum Y^2 = 51{,}606$

$$r = \frac{\sum XY - \left[\frac{\sum X \sum Y}{n}\right]}{\sqrt{\left[\sum X^2 - \left(\frac{(\sum X)^2}{n}\right)\right]\left[\sum Y^2 - \left(\frac{(\sum Y)^2}{n}\right)\right]}}$$

Correlation Coefficient (r)

$$r = \frac{\sum XY - \left[\frac{\sum X \sum Y}{n}\right]}{\sqrt{\left[\sum X^2 - \left(\frac{(\sum X)^2}{n}\right)\right]\left[\sum Y^2 - \left(\frac{(\sum Y)^2}{n}\right)\right]}}$$

$$= \frac{397200 - \left[\frac{4900(636)}{8}\right]}{\sqrt{\left[3150000 - \left(\frac{(4900)^2}{8}\right)\right]\left[51606 - \left(\frac{(636)^2}{8}\right)\right]}} = 0.614$$

It indicates a moderate positive relationship between the number of pages & the selling price of the book.

Coefficient of Determination (r^2)

$r = 0.614 \qquad r^2 = 0.377$

That is 37.7% of the total variation in the selling price of the book (Y) is explained by the variation in the number of pages of the book (X).

The balance of 62.3% is the unexplained variation.

In Regression Analysis, the independent variable (X) is used to estimate the dependent variable (Y).

The relationship between the variables is linear.

The least squares principle is used to determine the regression equation.

The Regression Equation expresses the linear relationship between two variables.

Regression Analysis

13-18

The regression equation is $Y' = \beta_0 + \beta_1 X$

where

Y' is the average predicted value of Y for any X.

β_0 is the Y-intercept.
It is the estimated Y value when $X = 0$

β_1 is the slope of the line, or the average change in Y' for each change of one unit in X

Regression Analysis

The least squares principle is used to obtain β_0 and β_1.

Y-intercept

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

slope

$$\beta_1 = \frac{\sum XY - \left[\frac{\sum X \sum Y}{n}\right]}{\left[\sum X^2 - \left(\frac{(\sum X)^2}{n}\right)\right]}$$

Regression Analysis

13-20

Develop a Regression Equation for the information given in Example 1 that can be used to estimate the selling price (Y) based on the number of pages (X).

$n = 8$	$\sum XY = 397{,}200$
$\sum X = 4900$	$\sum X^2 = 3{,}150{,}000$
$\sum Y = 636$	$\sum Y^2 = 51{,}606$

Example 1 revisited

13-21

$$\beta_1 = \frac{\sum XY - \left[\frac{\sum X \sum Y}{n}\right]}{\left[\sum X^2 - \left(\frac{(\sum X)^2}{n}\right)\right]}$$

$$= \frac{397200 - \left[\frac{4900\,(636)}{8}\right]}{\left[3150000 - \left(\frac{(4900)^2}{8}\right)\right]} = 0.05143$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X} = 79.5 - 0.05143\,(612.5) = 48$$

Use calculator to obtain these values.

Example 1 revisited

13-22

The regression equation is:
$Y' = 48 + 0.05143X$

The slope of the line is 0.05143. It means that each addition page costs about 5 cents.

The equation crosses the Y-axis at RM48 or when X = 0.
So, that means a book with no pages would cost RM48.

Example 1 revisited

13-23

We can use the regression equation to estimate values of Y.

What is the estimated selling price for a book that has 800 pages?

Price = RM48 + 0.05143 (Number of Pages)
= RM48 + 0.05143 (800)
= RM89.14

Example 1 revisited

13-24

Types of Estimation

<u>Interpolated Estimate</u> is an estimation made within the given data range.

<u>Extrapolated Estimate</u> is an estimation made outside the given data range.

<u>Interpolated Estimate</u> is always <u>more reliable</u> than the Extrapolated Estimate.

Types of Estimation

Using the Regression Equation in Example 1, compute Y when X = 1100 & X = 550

The regression equation is:
$Y' = 48 + 0.05143X$

Price = RM48 + 0.05143 (Number of Pages)
= RM48 + 0.05143 (1100)
= RM104.57

Price = RM48 + 0.05143 (Number of Pages)
= RM48 + 0.05143 (550)
= RM76.29

Example 1 revisited

13-26

Below is the Extrapolated Estimate.

Price = RM48 + 0.05143 (Number of Pages)
= RM48 + 0.05143 (1100)
= RM104.57

Below is the Interpolated Estimate.

Price = RM48 + 0.05143 (Number of Pages)
= RM48 + 0.05143 (550)
= RM76.29

Example 1 revisited

13-27

Why X = 1100 is not a reliable estimate whereas X = 550 is more reliable?

The minimum value of X is 400 pages & the maximum value of X is 800 pages.

Because, X = 1100 is outside the given data range while X = 550 is within the given data range.

Example 1 revisited

13-28

Book	Page	Price(RM)
Introduction to Statistics	500	84
Basic Algebra	700	75
Introduction to Psychology	800	99
Introduction to Sociology	600	72
Business Management	400	69
Introduction to Biology	500	81
Fundamentals of Finance	600	63
Principles of Marketing	800	93

Example 1 continued

13-29

EXERCISE 1

A company manufacturing machine parts would like to develop a model to estimate the number of worker hours required for production runs of varying lot sizes. A random sample of 14 production runs is selected with the following results.

a) Calculate the correlation coefficient & coefficient of determination.
b) Determine the least square regression line.
c) Estimate the worker hours for these lot sizes; 35 units & 100 units.
d) Which of the two estimates that is more reliable?

13-30

Lot Size	Worker Hours
20	50
20	55
30	73
30	67
40	87
40	95
50	108
50	112
60	128
60	135
70	148
70	160
80	170
80	162

EXERCISE 2

13-31

Sunflowers, a chain of women's clothing stores, has improved its market share over the past 25 years by increasing the number of stores in the chain. As the director of special projects and planning, you need to develop a strategic plan for opening several new stores. This plan must be able to forecast annual sales for all potential stores under consideration. You believe that the size of the store is significantly related to its success and want to incorporate this information in the decision-making process. To estimate the relationship between the store size (sq. ft) and its annual sales, a sample of 14 stores was selected.

a) Calculate and interpret the correlation coefficient & coefficient of determination.

b) Determine the least square regression line.

c) Estimate the annual sales for these store sizes;

1.4 sq. ft. & 6 sq. ft

d) Which of the two estimates that is more reliable?

13-32

Store	Sq. Ft. ('000)	Annual Sales ($'000)
1	1.7	3.7
2	1.6	3.9
3	2.8	6.7
4	5.6	9.5
5	1.3	3.4
6	2.2	5.6
7	1.3	3.7
8	1.1	2.7
9	3.2	5.5
10	1.5	2.9
11	5.2	10.7
12	4.6	7.6
13	5.8	11.8
14	3.0	4.1