Data Visualization

Dr. P. Kalpana, M.E., PhD.
Faculty of Mechanical Engineering
IIITDM Kancheepuram

Representation of Categorical variables

- Describe Data
  - frequency distribution tables
  - graphs
  - bar charts
  - pie charts
  - Pareto diagrams.

- used by managers and marketing researchers
- to describe data collected from surveys and questionnaires.

Frequency Distribution

- A frequency distribution is a table used to organize data.
- A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non overlapping classes.
- The left column (called classes or groups) includes all possible responses on a variable being studied.
- The right column is a list of the frequencies, or number of observations, for each class.
- A relative frequency distribution is obtained by dividing each frequency by the number of observations and multiplying the resulting proportion by 100%.

DATA FROM A SAMPLE OF 50 SOFT DRINK PURCHASES

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

Soft Drink	Frequency
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Soft Drink	Relative Frequency	Percent Frequency
Coke Classic	.38	38
Diet Coke	.16	16
Dr. Pepper	.10	10
Pepsi	.26	26
Sprite	.10	10
Total	1.00	100

Bar Chart

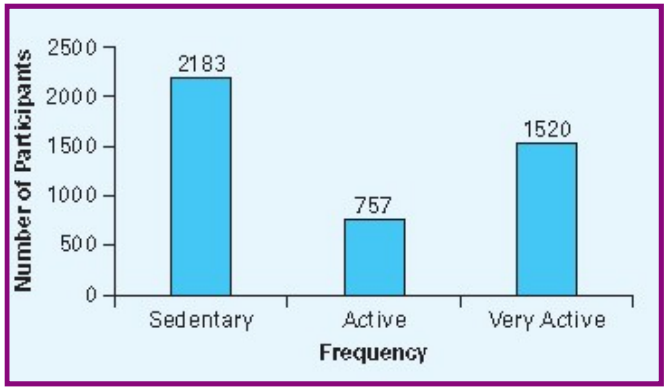
- A graphical device for depicting categorical data summarized in a frequency, relative frequency, or percent frequency distribution
- *Frequency of each category is drawn as a bar chart*
- *In a bar chart the* height of a rectangle represents each frequency
- **Example: Healthy Eating Index 2005 (HEI–2005):**
- Activity Level (Frequency Distribution and Bar Chart)
- One variable in the HEI–2005 study is a participant's activity level coded as
  - 1 = sedentary, 2 = active, and 3 = very active

Frequency distribution Table 1

HEI–2005 Participants' Activity Level: First Interview

	Participants	*Percent*
Sedentary	2,183	48.9
Active	757	17.0
Very active	1,520	34.1
Total	4,460	100.0

Bar Chart



Cross Tables

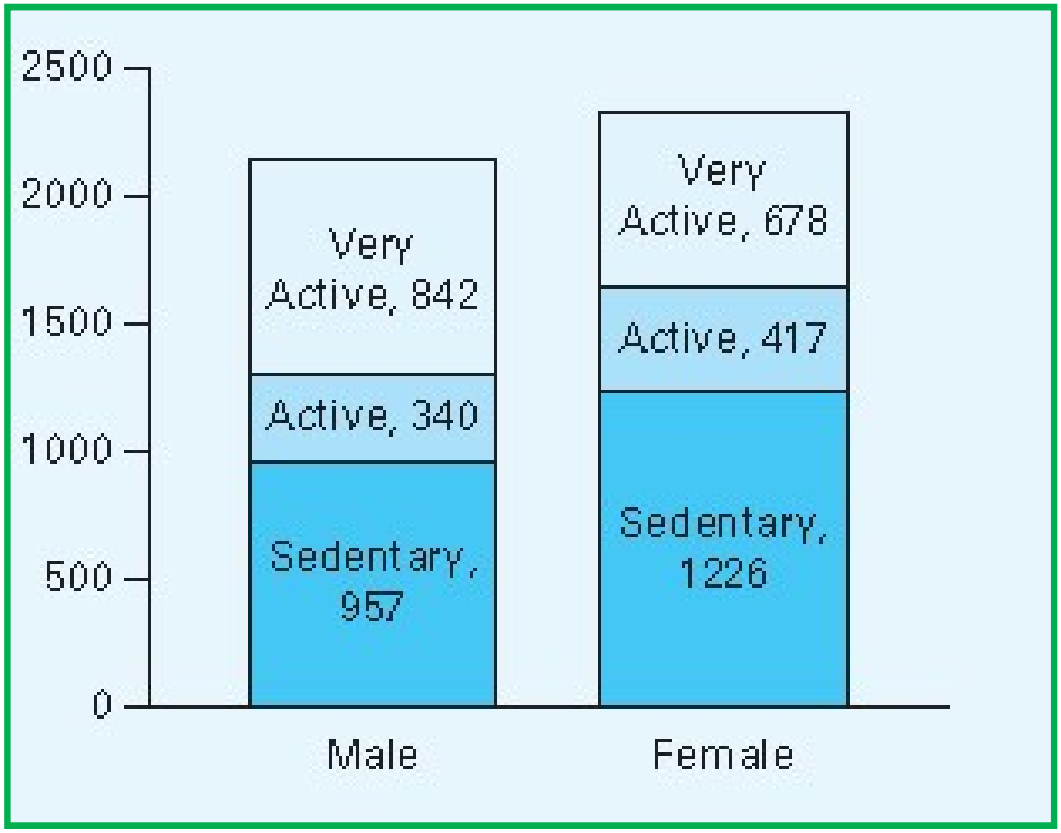
- A **cross table, sometimes called a crosstab or a contingency table, lists the** number of observations for every combination of values for two categorical or ordinal variables. The combination of all possible intervals for the two variables defines the cells in a table. A cross table with *r rows and c columns is* referred to as an *r * c cross table.*
- **Example 2 HEI–2005: Activity Level and**
- **Gender (Component and Cluster Bar Charts)**
- Consider again the data in Table1. Sometimes a comparison of one variable (activity level) with another variable (such as gender) is of interest. Construct component and cluster bar charts that compare activity level and gender.
- cross table of activity levels (1 = sedentary; 2 = active; and 3 = very active) and gender (0 = male; 1 = female)

Cross tabulation is a method to quantitatively analyze the relationship between multiple variables.

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

Component Bar Chart or stacked bar chart

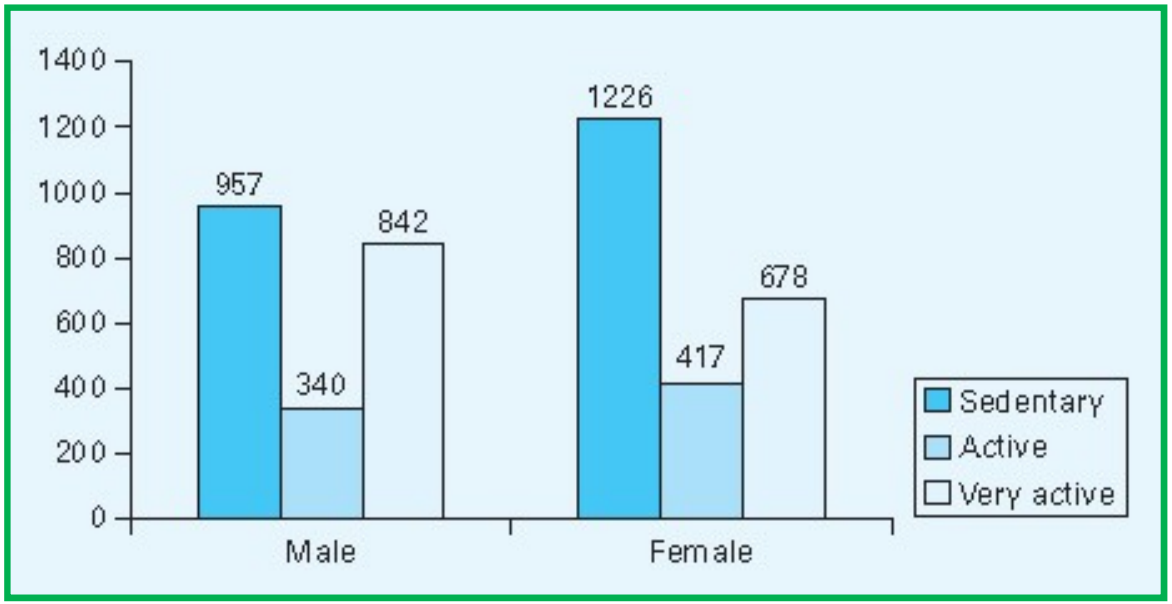
Used to represent data in which the total magnitude is divided into different or components



by Gender

Cluster Bar Chart

- A cluster chart is like a bar chart except that it clusters several bars into a category and displays each cluster separately from the rest.
- Each gender has three bars



by Gender

- The greater the number of categories a chart contains, the harder it is to compare between them.

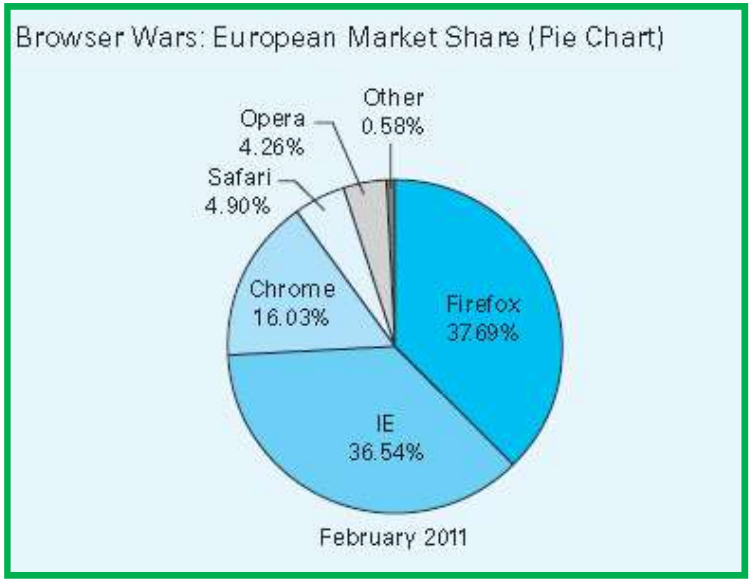
Pie Charts

- To draw attention to the proportion of frequencies in each category use a pie chart to depict the division of a whole into its constituent parts.
- The circle (or "pie") represents the total, and the segments (or "pieces of the pie") cut from its center depict shares of that total
- The pie chart is constructed so that the area of each segment is proportional to the corresponding frequency
- Table lists the market shares for various browsers in both Europe and North America during the month of February 2011

Market Shares (Pie Chart)

	European Market	North American Market
Firefox	37.69	26.24
Internet Explorer	36.54	48.16
Google Chrome	16.03	13.76
Safari	4.90	10.58
Opera	4.26	0.58
Others	0.58	0.68

source: http://gs.statcounter.com

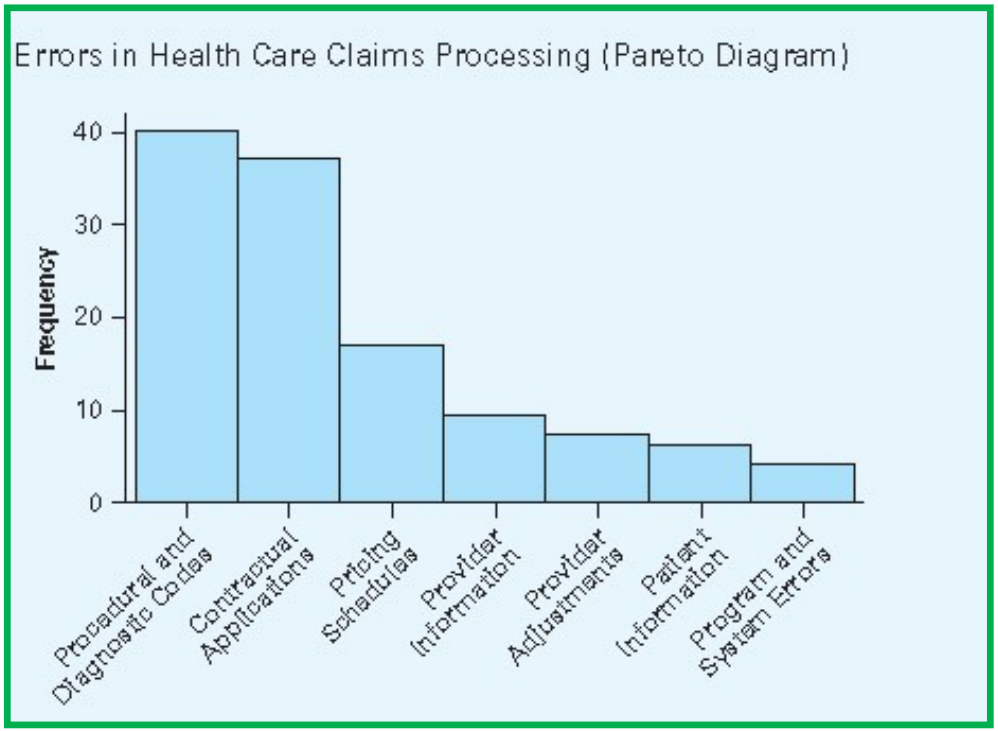


Pareto Diagrams

- A **Pareto diagram is a bar chart that displays the frequency of defect causes.**
- The bar at the left indicates the most frequent cause and the bars to the right indicate causes with decreasing frequencies.
- A Pareto diagram is used to separate the “vital few” from the “trivial many.”
- It is sometimes referred to as the 80–20 rule.
- The use of a Pareto diagram can also improve communication with employees or management and within production teams.
- Managers who need to identify major causes of problems and attempt to correct them quickly with a minimum cost frequently use a special bar chart known as a Pareto diagram.

Pareto Diagrams

CATEGORY	ERROR TYPE	FREQUENCY
1	Procedural and Diagnostic Codes	40
2	Provider Information	9
3	Patient Information	6
4	Pricing Schedules	17
5	Contractual Applications	37
6	Provider Adjustments	7
7	Program and System Errors	4



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

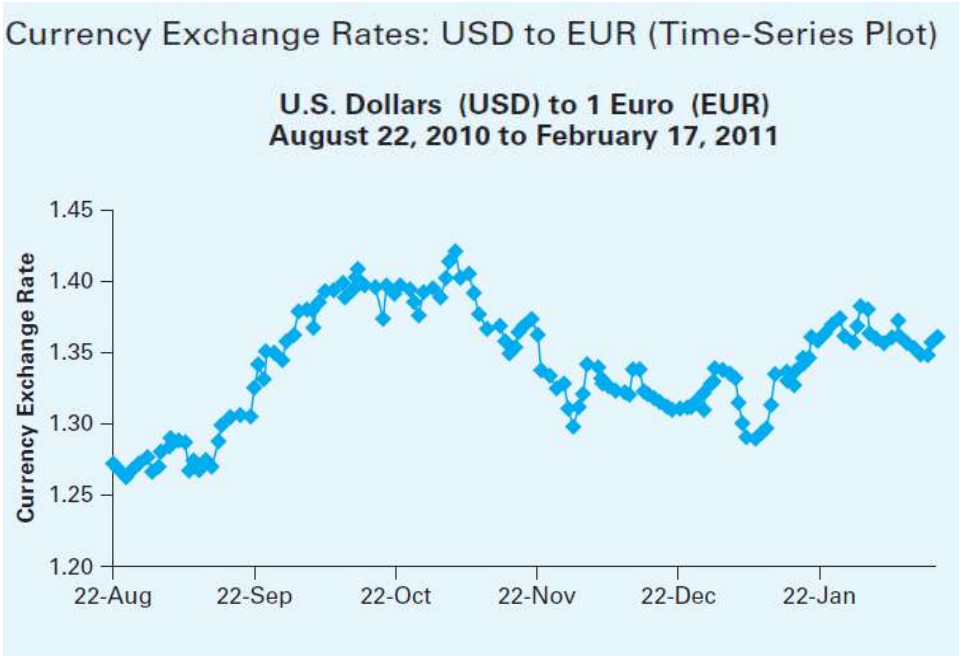
Pareto analysis separated
the vital few causes from the trivial many

Representation of Time-series data

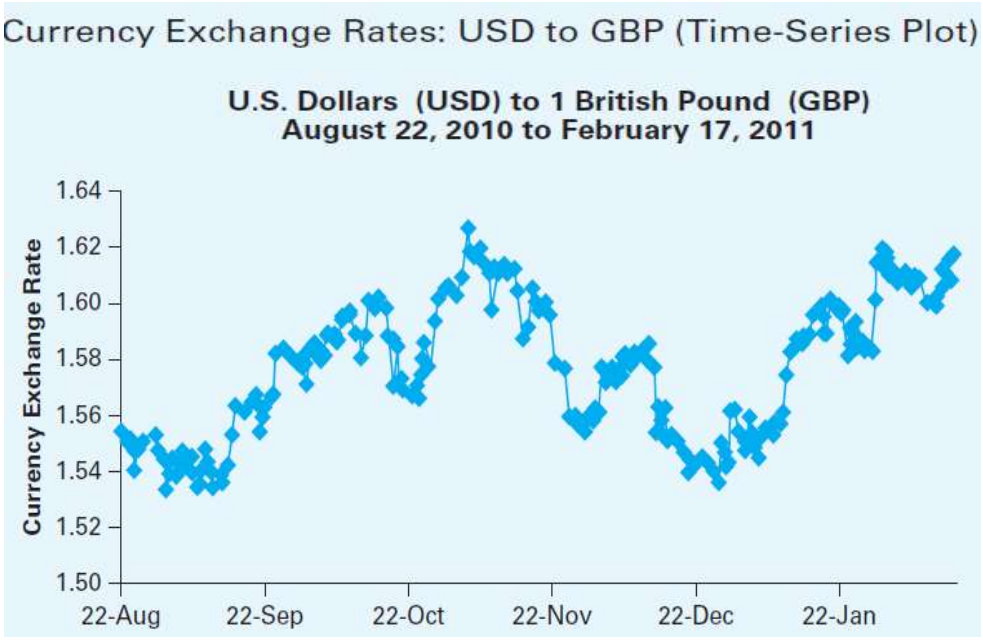
- **Line Chart (Time-Series Plot)**
  - a series of data plotted at various time intervals.
  - A graph of time-series data is called a line chart or time-series plot
  - Measurements of sample at one point in time obtained are known as cross-sectional data.
  - Data on a particular quantity of interest measured at successive points in time are called time-series data.
  - Measuring time along the horizontal axis and the numerical quantity of interest along the vertical axis yields a point on the graph for each observation.
  - Joining points adjacent in time by straight lines produces a time-series plot.
- Ex:
  - annual university enrollment
  - annual interest rates
  - the gross domestic product over a period of years
  - monthly product
  - sales, quarterly corporate earnings

Representation of Time-series data

Currency Exchange Rates: USD to EUR (Time-Series Plot)



Currency Exchange Rates: USD to GBP (Time-Series Plot)

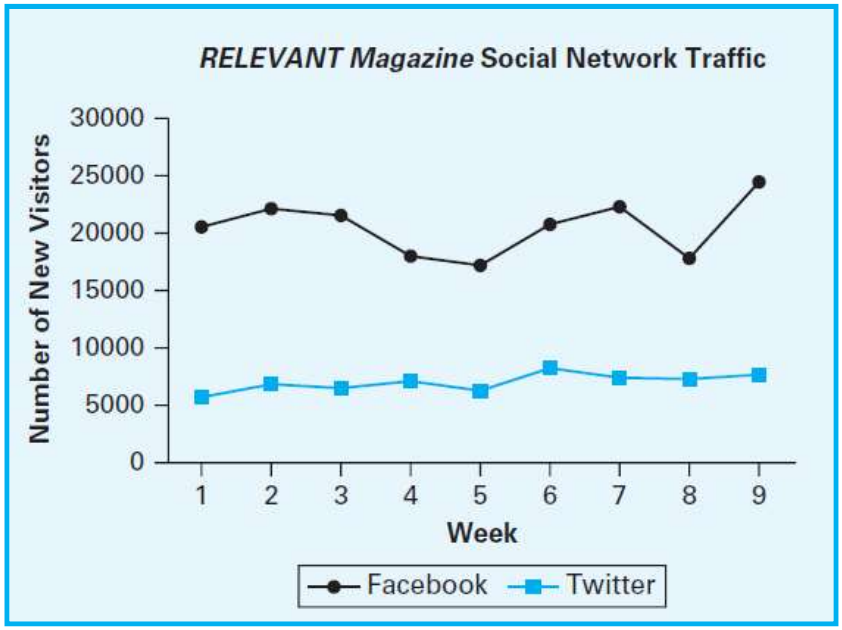


Representation of Time-series data

Sometimes a time-series plot is used to compare more than one variable over time

Social Network Traffic: Weekly New Visitors to *RELEVANT Magazine* Web Site

WEEK	TWITTER	FACEBOOK
1	5,611	20,499
2	6,799	22,060
3	6,391	21,365
4	6,966	17,905
5	6,111	17,022
6	8,101	20,572
7	7,370	22,201
8	7,097	17,628
9	7,531	24,256



Construct two time-series plots

Representation Numerical variables

- **Frequency Distributions:**
  - Similar to a frequency distribution for categorical data
  - Lists the classes in the left column and the number of observations in each class in the right column
  - Classes of a frequency distribution
    - How many classes should be used?
    - How wide should each class be?
  - Construction of a Frequency Distribution
    - Rule 1: Determine *k, the number of classes.*
    - Rule 2: Classes should be the same width, *w ; the width is determined by* the following:

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

Always round class width, w, upward

    - Rule 3: Classes must be inclusive and non overlapping

Frequency Distributions

- ***Rule 1: Number of Classes***

➢The number of classes used in a frequency distribution is decided in a somewhat arbitrary manner.

➢Too few classes, the patterns and various characteristics of the data may be hidden

➢Too many classes, we will discover that some of our intervals may contain no observations or have a very small frequency

Quick Guide to Approximate Number of Classes for a Frequency Distribution

Sample Size	Number of Classes
Fewer than 50	5–7
50 to 100	7–8
101 to 500	8–10
501 to 1,000	10–11
1,001 to 5,000	11–14
More than 5,000	14–20

Frequency Distributions

- **Cumulative Frequency Distributions**
  - Contains the total number of observations whose values are less than the upper limit for each class
  - We construct a cumulative frequency distribution by adding the frequencies of all frequency distribution classes up to and including the present class
- **Relative Cumulative Frequency Distributions**
  - Cumulative frequencies can be expressed as cumulative proportions or percents
- **Class Midpoint**
  - The *midpoint of each class interval*
  - Calculated as *the average of the two class endpoints*

Frequency Distributions

- **Employee Completion Time (in Seconds)**
- The goal is to complete this task in less than 4.5 minutes.

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

$$w = \frac{299 - 222}{8} = 10 \ (rounded\ up)$$

Frequency Distributions

Frequency and Relative Frequency Distributions for Completion Times

Completion Times (in Seconds)	Frequency	Percent
220 less than 230	5	4.5
230 less than 240	8	7.3
240 less than 250	13	11.8
250 less than 260	22	20.0
260 less than 270	32	29.1
270 less than 280	13	11.8
280 less than 290	10	9.1
290 less than 300	7	6.4

Cumulative Frequency and Relative Cumulative Frequency Distributions for Completion Times

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

Less than three-fourths (72.7%) of the employees sampled completed the task within the desired goal
Initiate an extra training session for the employees who failed to meet the time constraint

Exercise 1

- The following data represent the afternoon high temperatures for 50 construction days during a year in St. Louis.

42	70	64	47	66	69	73	38	48	25
55	85	10	24	45	31	62	47	63	84
16	40	81	15	35	17	40	36	44	17
38	79	35	36	23	64	75	53	31	60
31	38	52	16	81	12	61	43	30	33

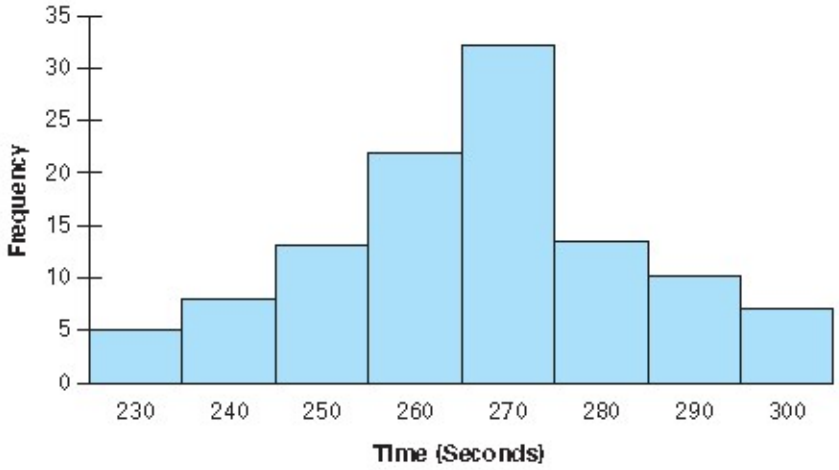
a. Construct a frequency distribution for the data using five class intervals.

b. Construct a frequency distribution for the data using 10 class intervals.

c. Examine the results of (a) and (b) and comment on the usefulness of the frequency distribution in terms of temperature summarization capability.

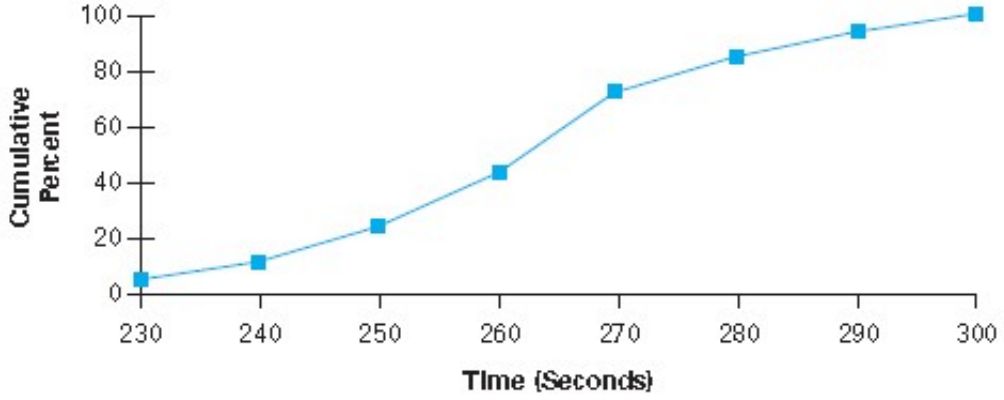
Histogram

- Consists of vertical bars constructed on a horizontal line that is marked off with intervals for the variable being displayed.
- The intervals correspond to the classes in a frequency distribution table.
- The height of each bar is proportional to the number of observations in that interval.
- The number of observations can be displayed above the bars.
- Yields information about
  - the shape of the distribution of a large database
  - the variability of the data
  - the central location of the data
  - and outlier data.



Ogive or cumulative frequency polygon

- An **ogive, sometimes called a *cumulative line graph, is a line that connects*** points that are the cumulative percent of observations below the upper limit of each interval in a cumulative frequency distribution.

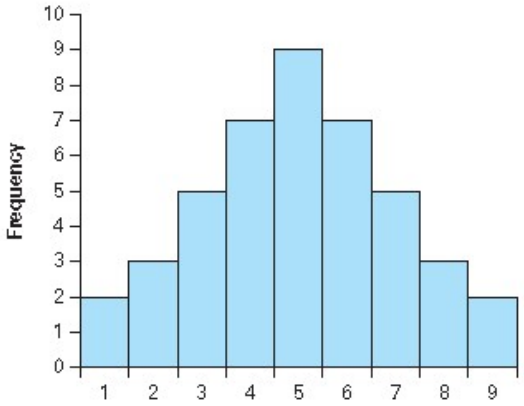


- Ogives are most useful when the decision maker wants to see *running totals.*
- For example, if a comptroller is interested in controlling costs, an ogive could depict cumulative costs over a fiscal year.

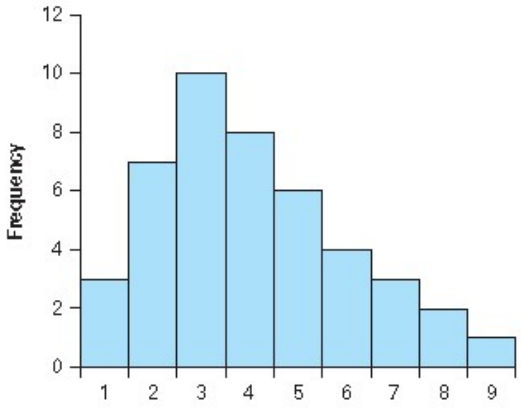
Shape of a Distribution

describe graphically the shape of the distribution by a histogram

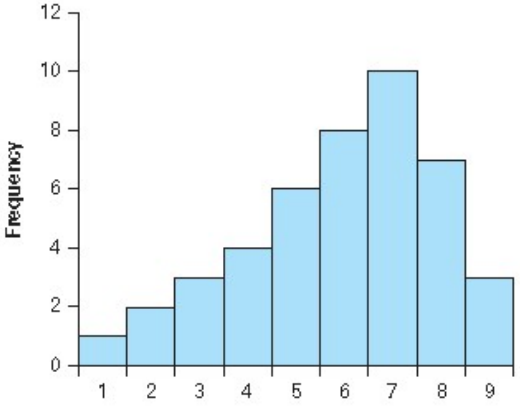
- Symmetry
  - If the observations are balanced, or approximately evenly distributed, about its center.
- Skewness (asymmetric)
  - If the observations are not symmetrically distributed on either side of the center.
  - A *skewed-right distribution* (sometimes called *positively skewed) has a tail that extends farther to the* right.
  - A *skewed-left distribution (sometimes called negatively skewed) has a* tail that extends farther to the left.



Symmetric Distribution



Skewed-right Distribution



Skewed-left Distribution

Stem-and-Leaf Displays

- A quick way to identify possible patterns when you have a small data set
- A stem-and-leaf display is an Exploratory data analysis (EDA) graph
- An alternative to the histogram.
- Data are grouped according to their leading digits (called stems),
- The final digits (called leaves) are listed separately for each member of a class.
- The leaves are displayed individually in ascending order after each of the stems.
- all stems are included, even if there are no observations (leaves) in the corresponding subset

- Ex: Random sample of 10 final exam grades

88 51 63 85 79 65 79 70 73 77

51 63 65 70 73 77 79 79 85 88

Stem-and-Leaf Display
$n = 10$

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

Exercise

The following data represent the costs (in dollars) of a sample of 30 postal mailings by a company.

3.67	2.75	9.15	5.11	3.32	2.09
1.83	10.94	1.93	3.89	7.20	2.78
6.72	7.80	5.47	4.15	3.55	3.53
3.34	4.95	5.42	8.64	4.84	4.10
5.10	6.45	4.65	1.97	2.84	3.21

- Using dollars as a stem and cents as a leaf, construct a stem-and-leaf plot of the data.

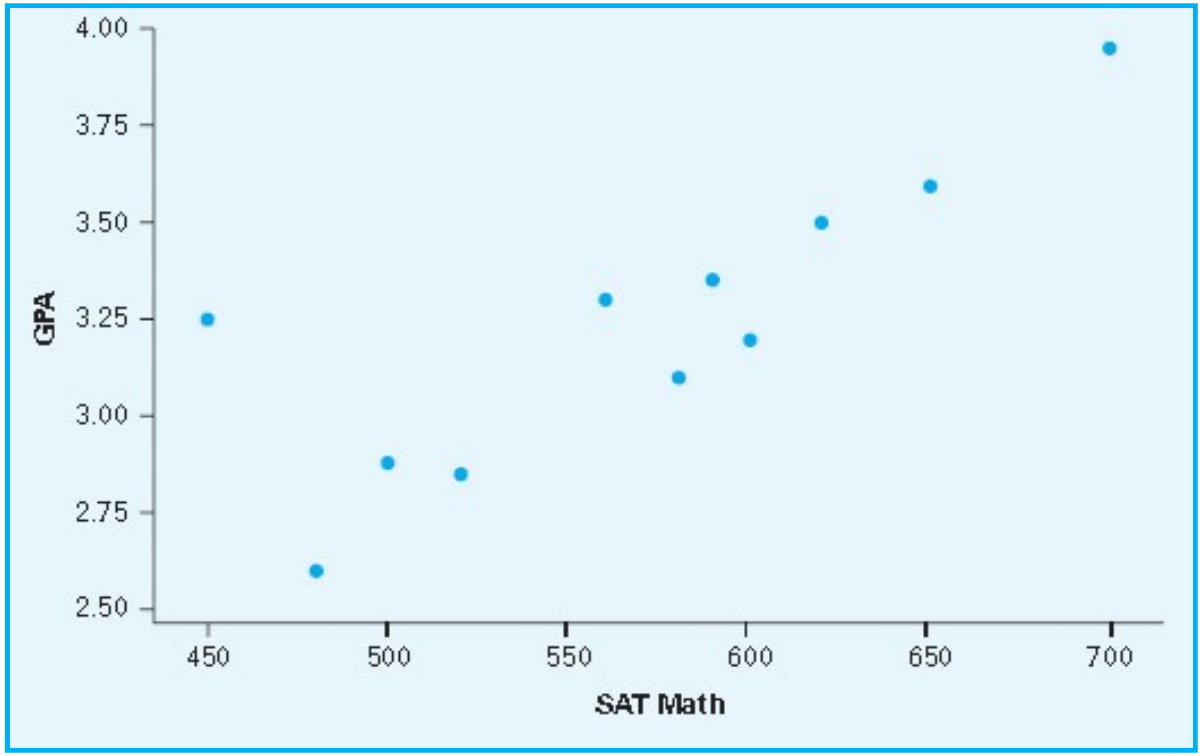
Scatter Plots

- Used to investigate possible relationships between two numerical variables
- Business and economic analyses are often concerned about relationships between variables.
  - What is the effect of advertising on total profits?
  - What is the change in quantity sold as the result of a change in price?
- One variable may depend to a certain extent on the other variable
  - Dependent variable and label it Y
  - Independent variable and label it X
- scatter plot by locating one point for each pair of two variables that represent an observation in the data set.
- The scatter plot provides a picture of the data, including the following:
  - The range of each variable
  - The pattern of values over the range
  - A suggestion as to a possible relationship between the two variables
  - An indication of outliers (extreme points)

Scatter Plots

- GPA vs. SAT Math Scores (Scatter Plot)

SAT MATH	GPA
450	3.25
480	2.60
500	2.88
520	2.85
560	3.30
580	3.10
590	3.35
600	3.20
620	3.50
650	3.59
700	3.95



Data Presentation Errors

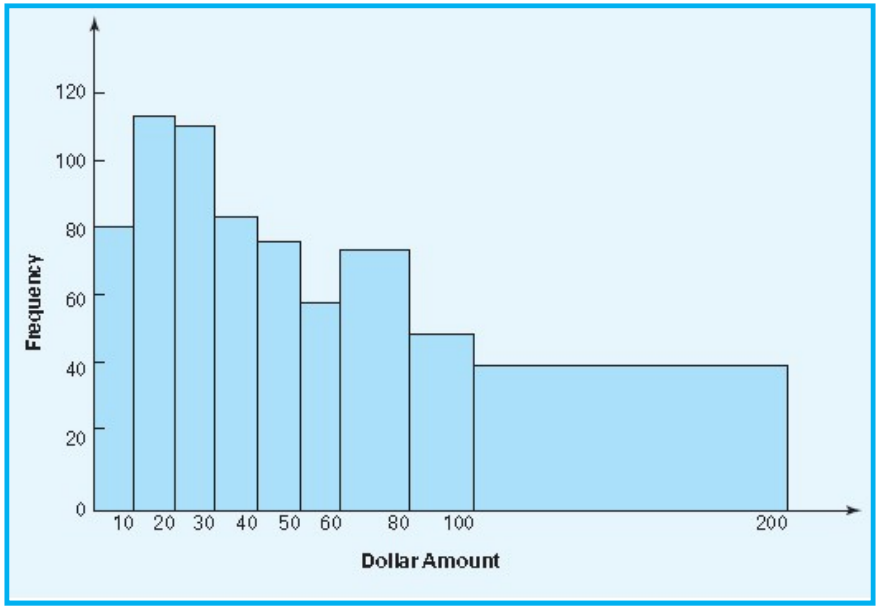
- Improper graphs can produce a distorted picture, yielding a false impression. It is possible to convey the wrong message without being deliberately dishonest
- graphs must be persuasive, clear, and truthful

Grocery Receipts (Unequal Interval Widths)

Dollar Amount	Number of Receipts	Proportions
$0 < $10	84	84/692
$10 < $20	113	113/692
$20 < $30	112	112/692
$30 < $40	85	85/692
$40 < $50	77	77/692
$50 < $60	58	58/692
$60 < $80	75	75/692
$80 < $100	48	48/692
$100 < $200	40	40/692

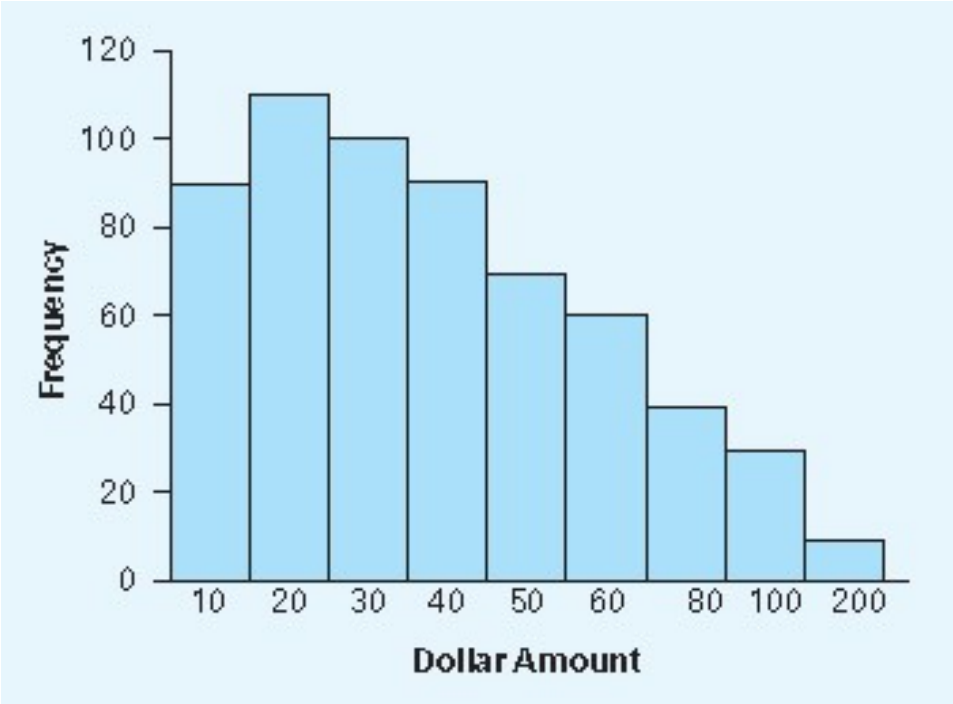
Data Presentation Errors

- Misleading Histogram of Grocery Receipts **(Error: Heights Proportional to Frequencies for Distribution with Varying Interval Widths)**



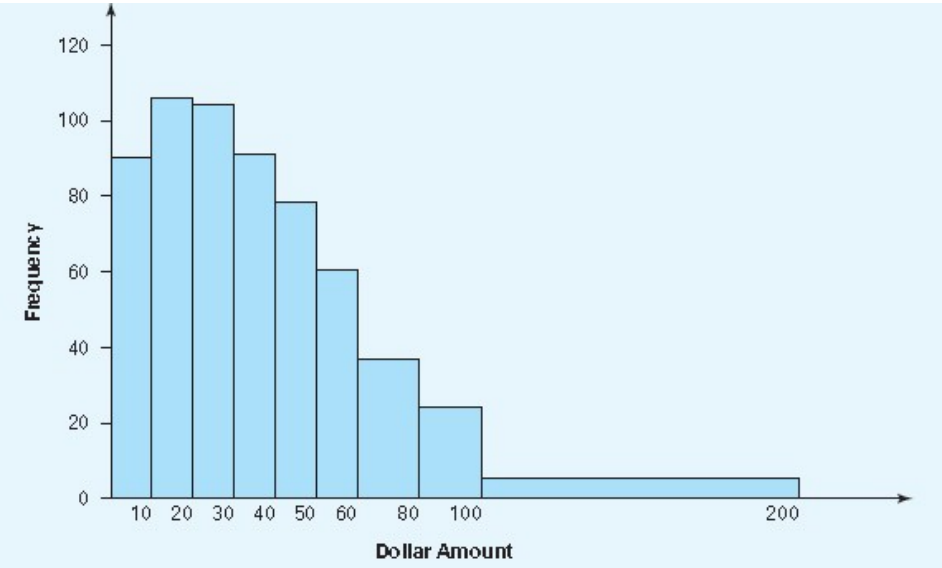
Data Presentation Errors

- Misleading Histogram of Grocery Receipts (Error: Bars of Equal Width for Distribution with Varying Interval Widths)



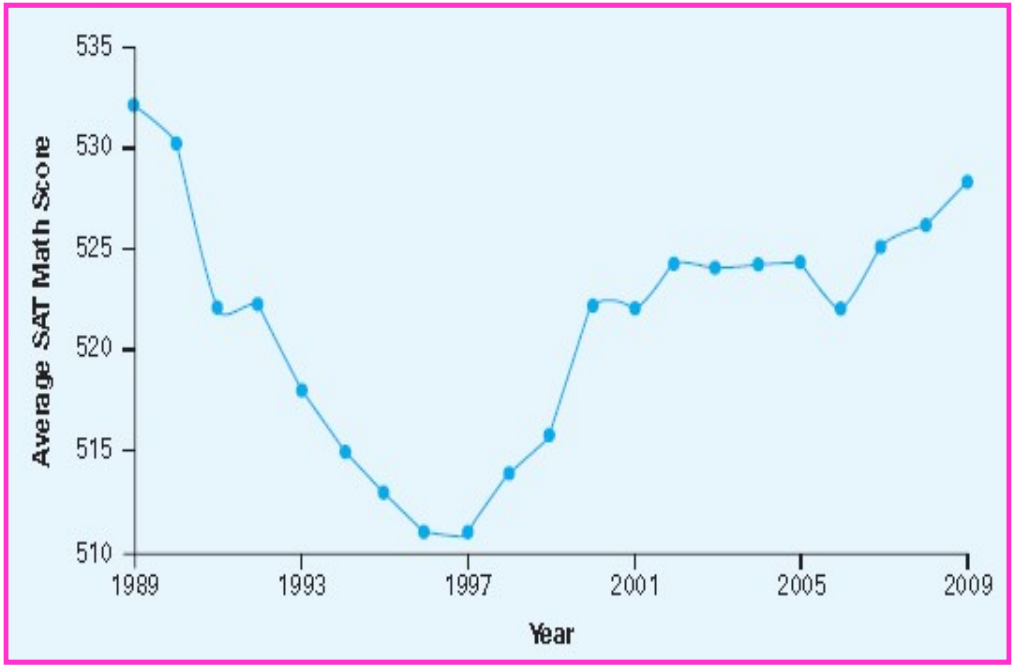
Data Presentation Errors

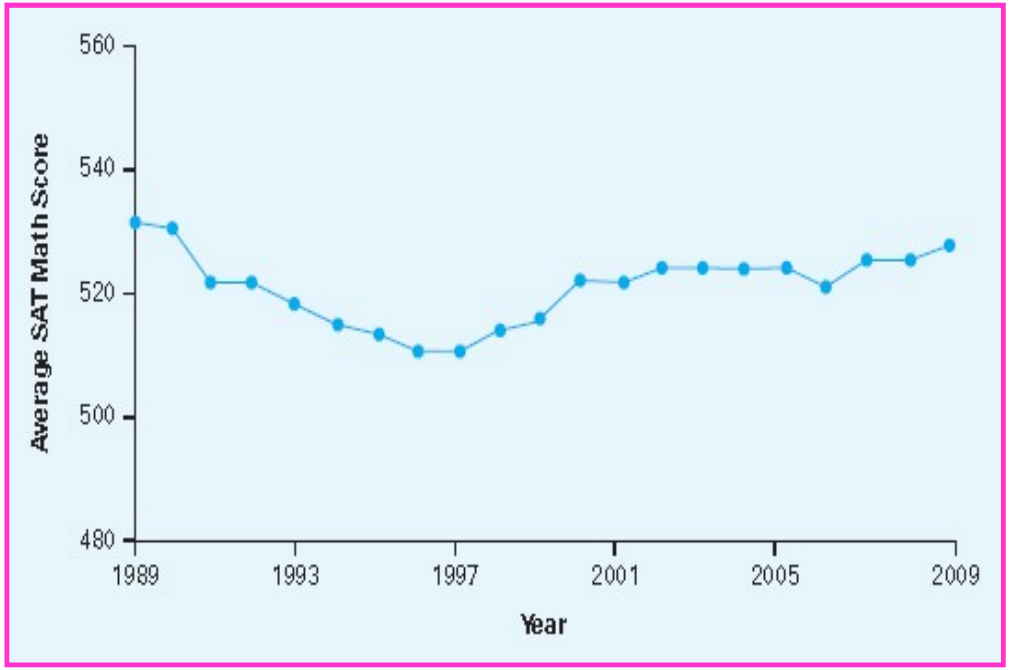
- Grocery Receipts (Histogram)



Choice of Scale for Time-Series Plot

SAT Math Scores 1989–2009





The shape of the plot alone is inadequate for obtaining a clear picture of the data
It is also necessary to keep in mind the scale on which the measurements are made

Reference

- Statistics for Business and Economics, Pearson edition (2013), William L Carlson, Paul Newbold, Betty M.Thorne
- statistics-for-business-and-economics-8th-edition, Paul newbold, William L.Carlson,Betty M.Thome
- Business statistics for contemporary Decision making , 6th edition by Ken Black
- https://medium.com/budding-data-scientist
- statistics-for-business-and-economics-11th edition, David R. Anderson,Dennis J. Sweeney,Thomas A. Williams


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
gracias
tapadh leat
dank je
bedankt
dziękuję
obrigado
sukriya
kop khun krap
go raibh maith agat
arigatō
takk
dakujem
grazie
terima kasih
merci
мерси
감사합니다
ευχαριστώ
நன்றி

- Steps to construct Frequency Distribution:
  - Determine the range of the raw data
    - the difference between the largest and smallest numbers.
  - determine how many classes it will contain
    - One rule of thumb is to select between 5 and 15 classes
    - too few classes, the data summary may be too general to be useful
    - Too many classes may result in a frequency distribution that does not aggregate the data enough to be helpful
  - determine the width of the class interval
  - Class mark

Frequency Distribution

- **Class Midpoint**
  - The *midpoint of each class interval*
  - Calculated as *the average of the two class endpoints*
- **Relative Frequency**
  - *the proportion of the total frequency that is in any given class interval in a frequency distribution*
- **Cumulative Frequency**
  - *A running total of frequencies through the classes of a frequency distribution*

Interval	Frequency	Class Midpoint	Relative Frequency	Cumulative Frequency
1–under 3	4	2	.0667	4
3–under 5	12	4	.2000	16
5–under 7	13	6	.2167	29
7–under 9	19	8	.3167	48
9–under 11	7	10	.1167	55
11–under 13	5	12	.0833	60
Total	60			