Co-financed by the Connecting Europe
Facility of the European Union

xAIM

Describing Data: Graphical

Data Driven Healthcare
Module B
Prof. Paola Cerchiello – University of Pavia

Lecture Goals (1 of 3)

After completing this lecture, you should be able to:

- Explain how decisions are often based on incomplete information
- Explain key definitions:
  - Population vs. Sample
  - Parameter vs. Statistic
  - Descriptive vs. Inferential Statistics
- Describe random sampling and systematic sampling
- Explain the difference between Descriptive and Inferential statistics

Chapter Goals (2 of 3)

After completing this lecture, you should be able to:

- Identify types of data and levels of measurement
- Create and interpret graphs to describe categorical variables:
  - frequency distribution, bar chart, pie chart, Pareto diagram
- Create a line chart to describe time-series data
- Create and interpret graphs to describe numerical variables:
  - frequency distribution, histogram, ogive, stem-and-leaf display

Chapter Goals (3 of 3)

After completing this lecture, you should be able to:

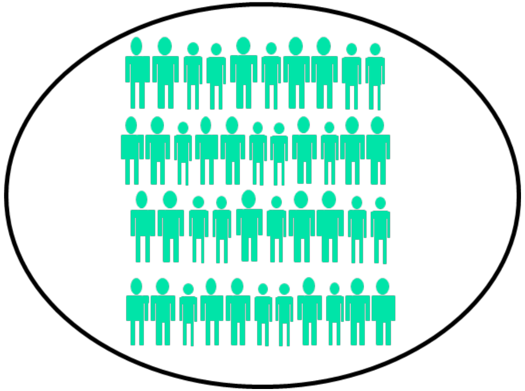
- Construct and interpret graphs to describe relationships between variables:
  - Scatter plot, cross table
- Describe appropriate and inappropriate ways to display data graphically

Key Definitions

- A population is the collection of all items of interest or under investigation
    - *N* represents the population size
- A sample is an observed subset of the population
    - *n* represents the sample size
- A parameter is a specific characteristic of a population
- A statistic is a specific characteristic of a sample

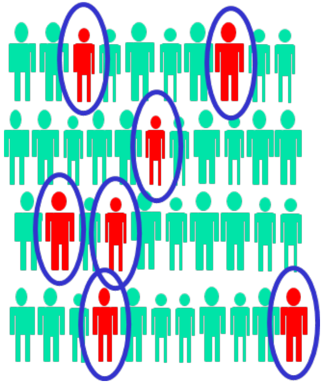
Population vs. Sample

Population



Values calculated using population data are called parameters

Sample



Values computed from sample data are called statistics

Examples of Populations

- Names of all registered voters in the United States
- Incomes of all families living in Daytona Beach
- Annual returns of all stocks traded on the New York Stock Exchange
- Grade point averages of all the students in your university

Random Sampling

Simple random sampling is a procedure in which

- each member of the population is chosen strictly by chance,
- each member of the population is equally likely to be chosen,
- every possible sample of n objects is equally likely to be chosen

The resulting sample is called a random sample

Descriptive and Inferential Statistics

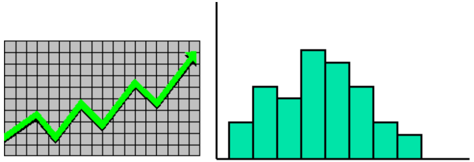
Two branches of statistics:

- Descriptive statistics
  - Graphical and numerical procedures to summarize and process data
- Inferential statistics
  - Using data to make predictions, forecasts, and estimates to assist decision making

Descriptive Statistics

- Collect data
  - e.g., Medical Reports

Present data
e.g., Tables and graphs



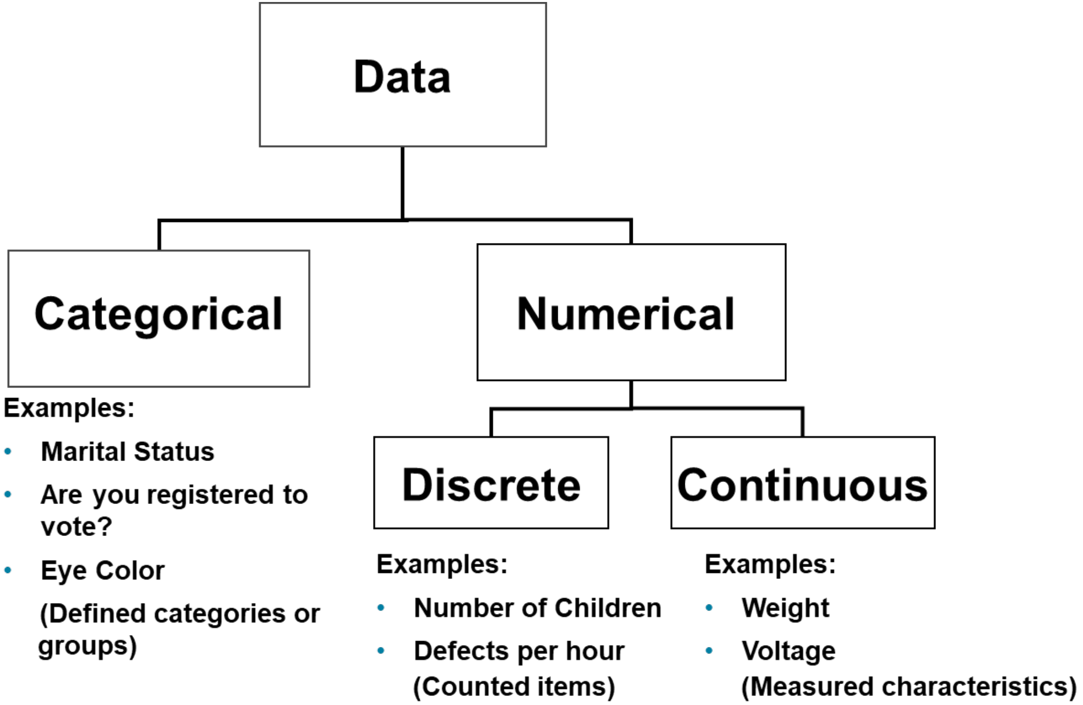
- Summarize data
  - e.g., $\text{Sample mean} = \frac{\sum X_i}{n}$

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 140 pounds

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Classification of Variables



Data

Categorical

Examples:

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

(Defined categories or groups)

Numerical

Discrete

Examples:

- Number of Children
- Defects per hour

(Counted items)

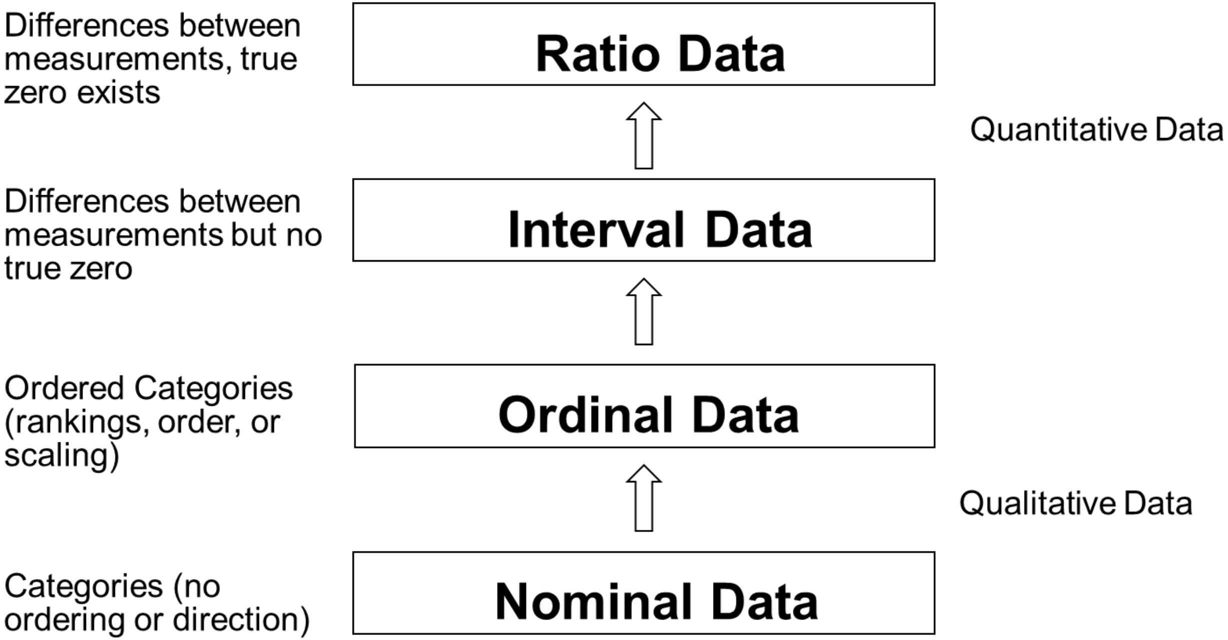
Continuous

Examples:

- Weight
- Voltage

(Measured characteristics)

Measurement Levels



Differences between measurements, true zero exists

Ratio Data

Quantitative Data

Differences between measurements but no true zero

Interval Data

Ordered Categories (rankings, order, or scaling)

Ordinal Data

Qualitative Data

Categories (no ordering or direction)

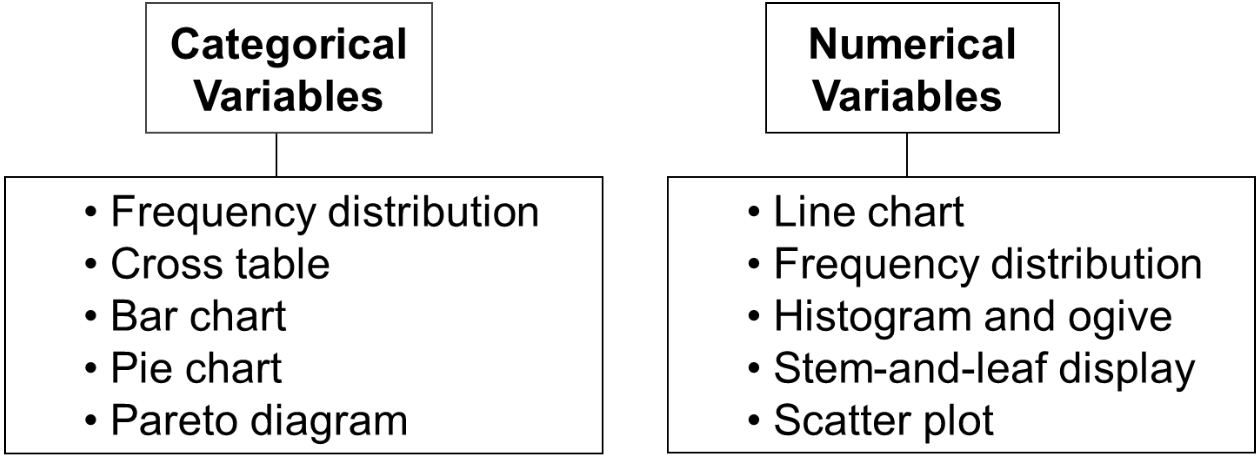
Nominal Data

Graphical Presentation of Data (1 of 2)

- Data in raw form are usually not easy to use for decision making
- Some type of organization is needed
  - Table
  - Graph
- The type of graph to use depends on the variable being summarized

Graphical Presentation of Data (2 of 2)

- Techniques reviewed in this chapter:



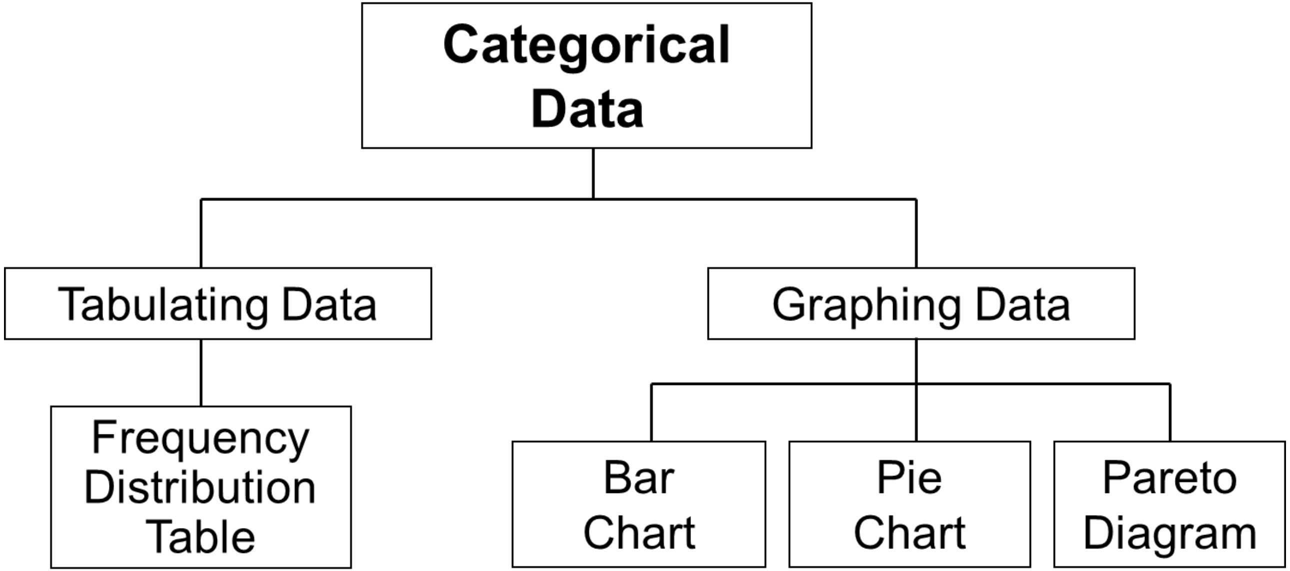
Categorical Variables

- Frequency distribution
- Cross table
- Bar chart
- Pie chart
- Pareto diagram

Numerical Variables

- Line chart
- Frequency distribution
- Histogram and ogive
- Stem-and-leaf display
- Scatter plot

Tables and Graphs for Categorical Variables



Categorical Data

Tabulating Data

Frequency Distribution Table

Graphing Data

Bar Chart

Pie Chart

Pareto Diagram

The Frequency Distribution Table

Summarize data by category

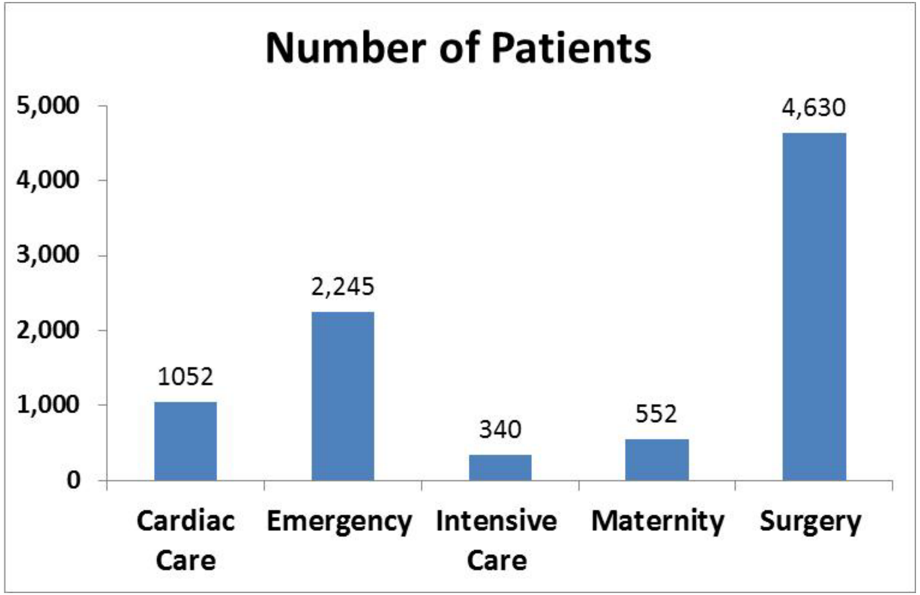
Example: Hospital Patients by Unit

Hospital Unit	Number of Patients	Percent (rounded)
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50
Total:	8,819	100.0

(Variables are categorical)

Graph of Frequency Distribution

- Bar chart of patient data



Cross Tables

- Cross Tables (or contingency tables) list the number of observations for every combination of values for two categorical or ordinal variables

If there are *r* categories for the first variable (rows) and *c* categories for the second variable (columns),

the table is called an $r \times c$ cross table

Cross Table Example

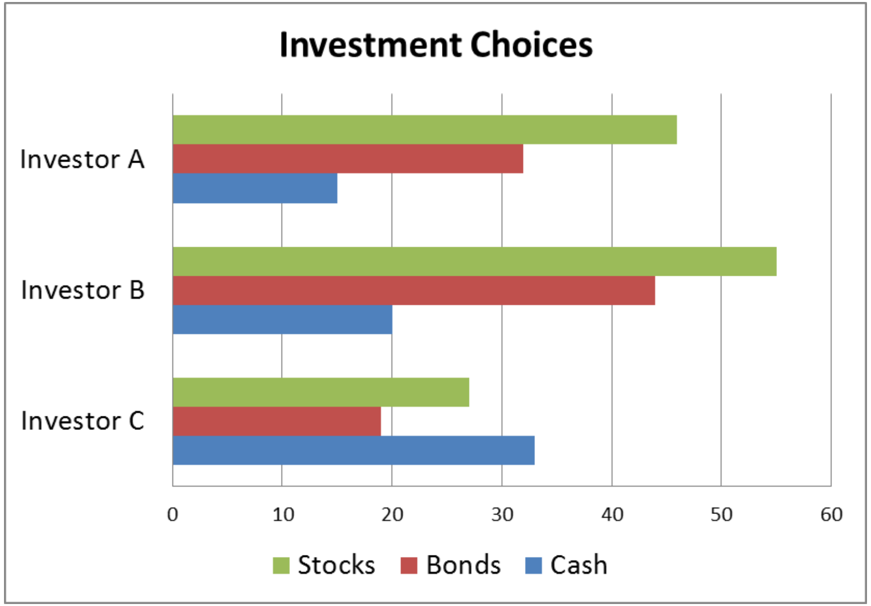
3×3 Cross Table for Investment Choices by Investor

(values in $1000’s)

Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46	55	27	**128**
Bonds	32	44	19	**95**
Cash	15	20	33	**68**
Total	**93**	**119**	**79**	**291**

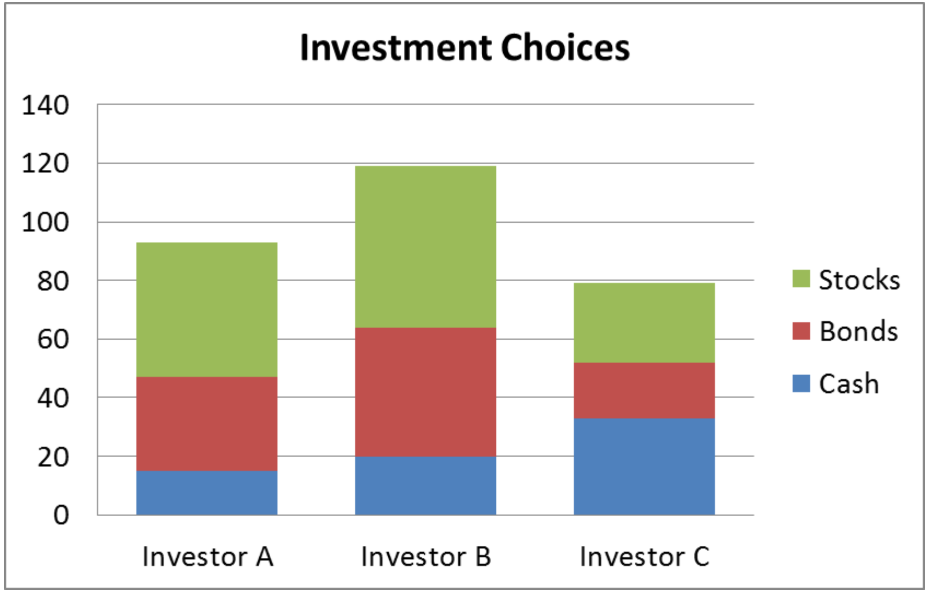
Graphing Multivariate Categorical Data (1 of 2)

- Side by side horizontal bar chart



Graphing Multivariate Categorical Data (2 of 2)

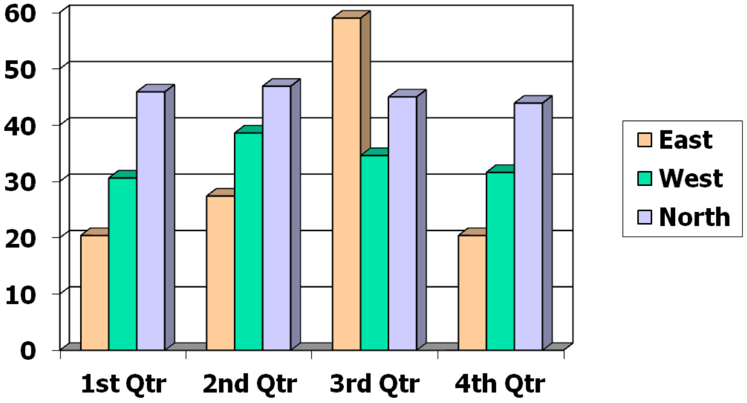
- Stacked bar chart



Vertical Side-by-Side Chart Example

- Sales by quarter for three sales territories:

	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr
East	20.4	27.4	59	20.4
West	30.6	38.6	34.6	31.6
North	45.9	46.9	45	43.9

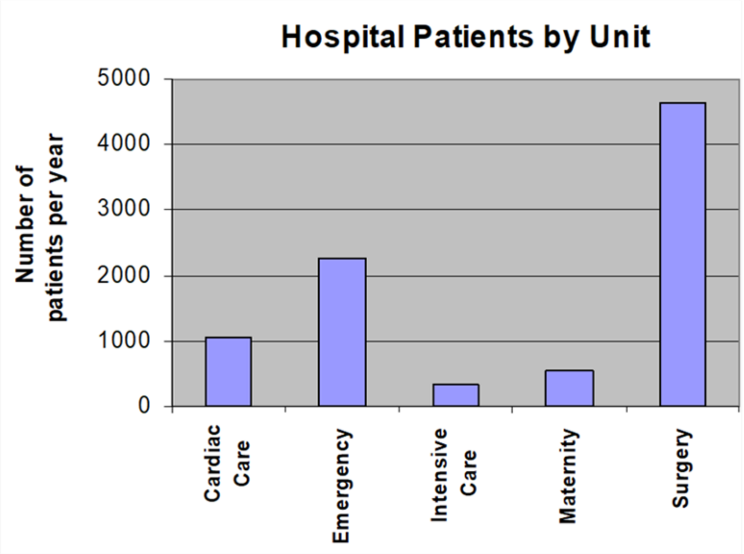


Bar and Pie Charts

- Bar charts and Pie charts are often used for qualitative (categorical) data
- Height of bar or size of pie slice shows the frequency or percentage for each category

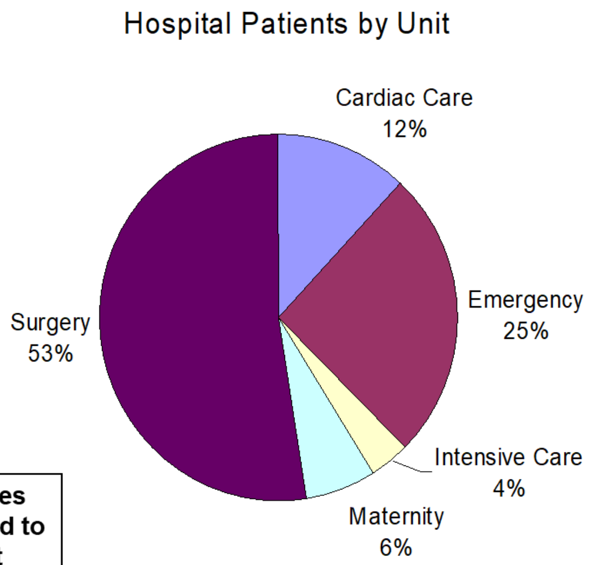
Bar Chart Example

Hospital Unit	Number of Patients
Cardiac Care	1,052
Emergency	2,245
Intensive Care	340
Maternity	552
Surgery	4,630



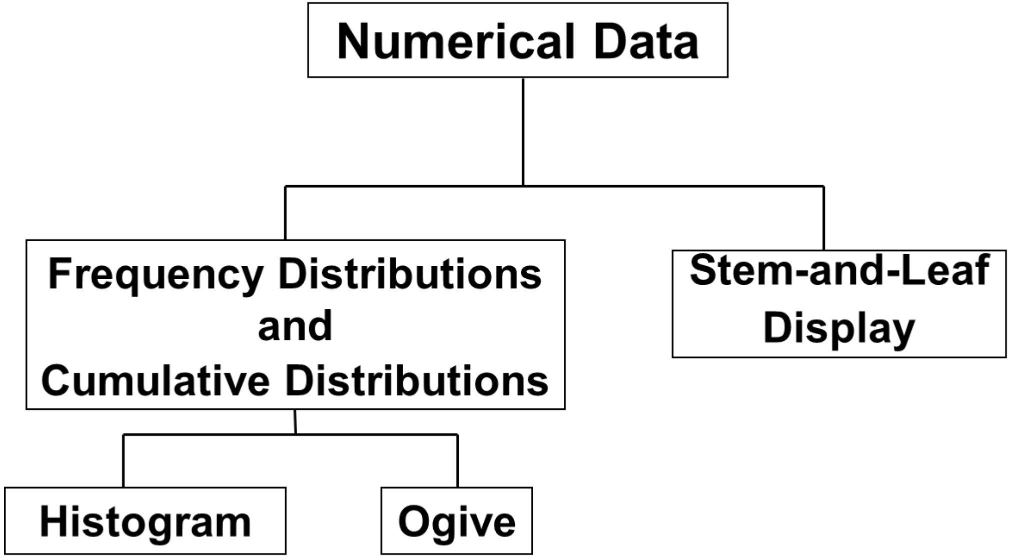
Pie Chart Example

Hospital Unit	Number of Patients	% of Total
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50



(Percentages are rounded to the nearest percent)

Graphs to Describe Numerical Variables



Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a list or a table…
- containing class groupings (categories or ranges within which the data fall)…
- and the corresponding frequencies with which data fall within each class or category

Why Use Frequency Distributions?

- A frequency distribution is a way to summarize data
- The distribution condenses the raw data into a more useful form...
- and allows for a quick visual interpretation of the data

Class Intervals and Class Boundaries

- Each class grouping has the same width
- Determine the width of each interval by

$$w = \text{interval width} = \frac{\text{largest number} - \text{smallest number}}{\text{number of desired intervals}}$$

Use at least 5 but no more than 15-20 intervals
Intervals never overlap
Round up the interval width to get desirable interval endpoints

Frequency Distribution Example (1 of 3)

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

data:

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

Frequency Distribution Example (2 of 3)

- 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: $\mathbf{58 - 12 = 46}$

Select number of classes: **5 (usually between 5 and 15)**

Compute interval width: $\mathbf{10 \left(\frac{46}{5} \textbf{ then round up} \right)}$

Determine interval boundaries: **10 but less than 20, 20 but**

less than $\mathbf{30, \ldots, 60}$ **but less than 70**

Count observations & assign to classes

Frequency Distribution Example (3 of 3)

Data in ordered array:

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

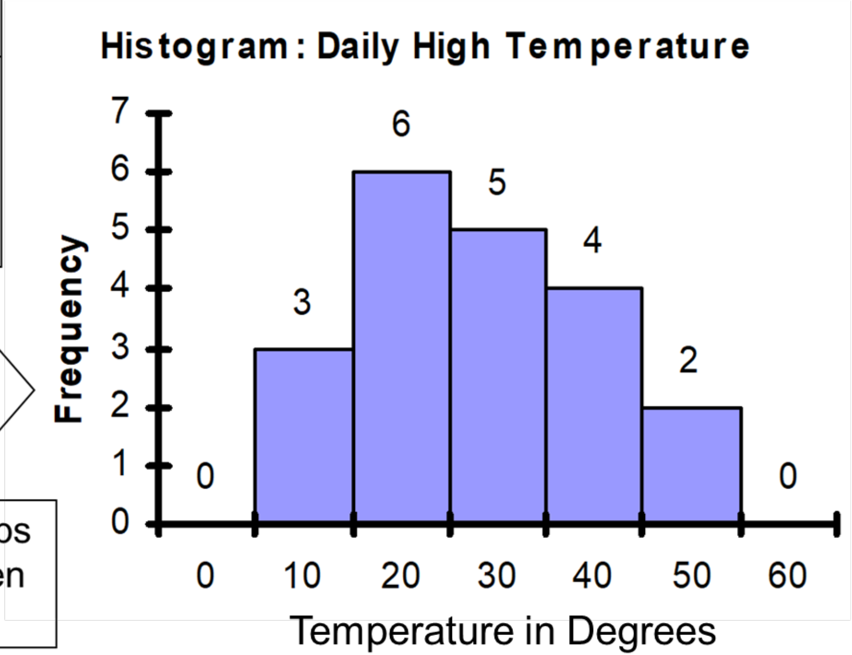
Interval	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Histogram

- A graph of the data in a frequency distribution is called a **histogram**
- The **interval endpoints** 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

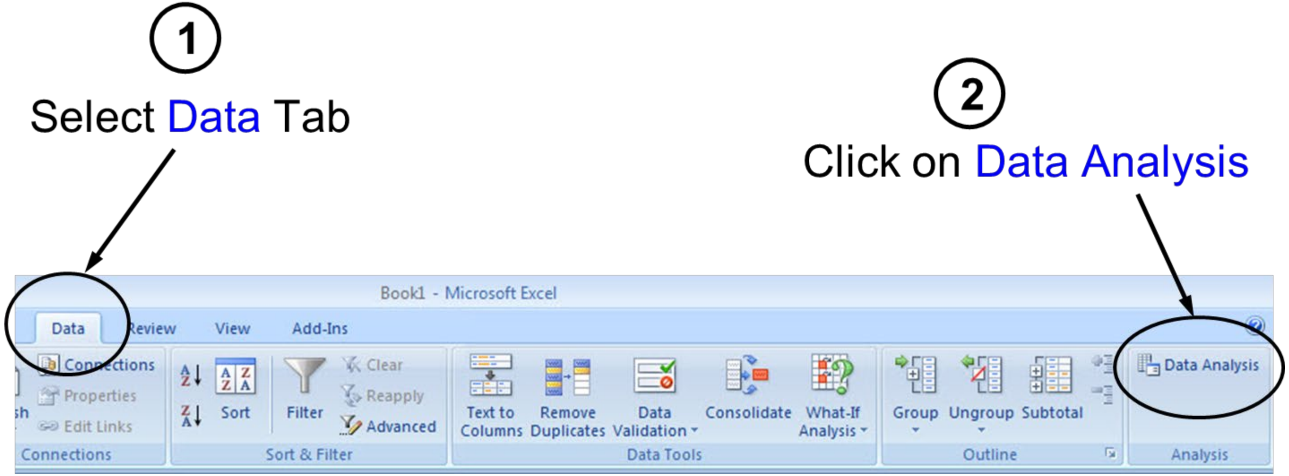
Histogram Example

Interval	Frequency
10 but less than 20	3
20 but less than 30	6
30 but less than 40	5
40 but less than 50	4
50 but less than 60	2



(No gaps between bars)

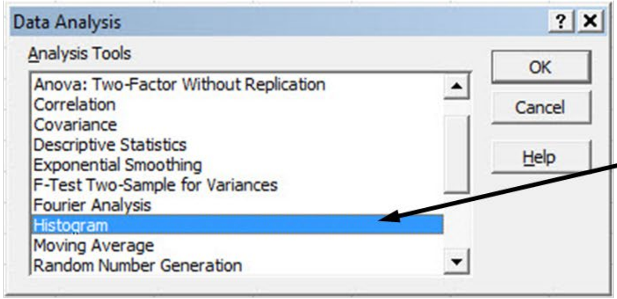
Histograms in Excel (1 of 2)



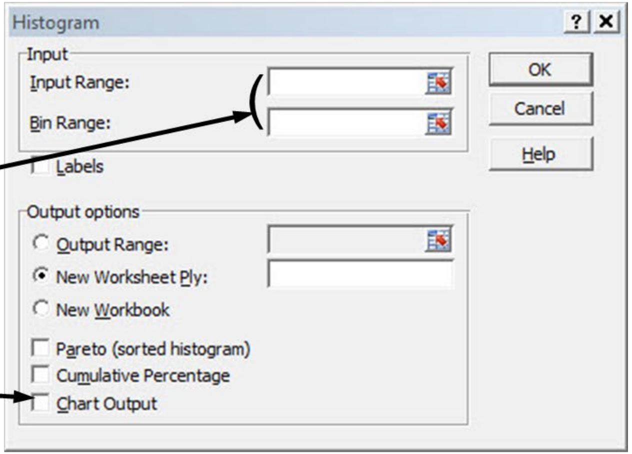
1 Select Data Tab

2 Click on Data Analysis

Histograms in Excel (2 of 2)



3 Choose Histogram



Input

Input Range:

Bin Range:

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

□ Chart Output

4 Input data range and bin range (bin range is a cell range containing the upper interval endpoints for each class grouping)

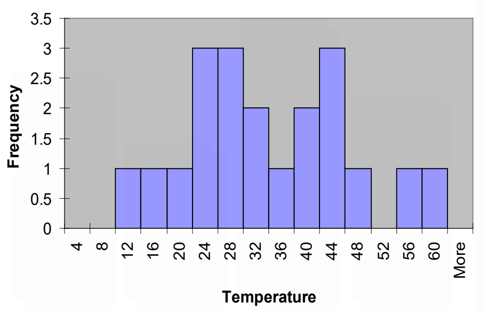
Select Chart Output and click "OK"

Questions for Grouping Data into Intervals

- How wide should each interval be?
  (How many classes should be used?)
- How should the endpoints of the intervals be determined?
  - Often answered by trial and error, subject to user judgment
  - The goal is to create a distribution that is neither too "jagged" nor too "blocky”
  - Goal is to appropriately show the pattern of variation in the data

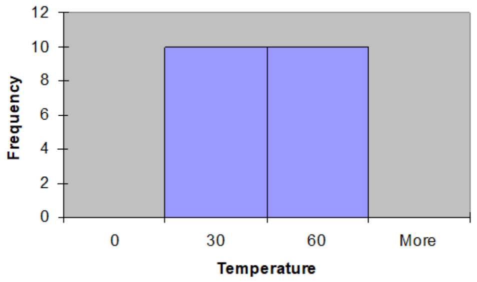
How Many Class Intervals?

- **Many (Narrow class intervals)**
  - may yield a very jagged distribution with gaps from empty classes
  - Can give a poor indication of how frequency varies across classes



Few (Wide class intervals)

- may compress variation too much and yield a blocky distribution
- can obscure important patterns of variation.



(*X* axis labels are upper class endpoints)

The Cumulative Frequency Distribution

Data in ordered array:

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

Class	Frequenc y	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

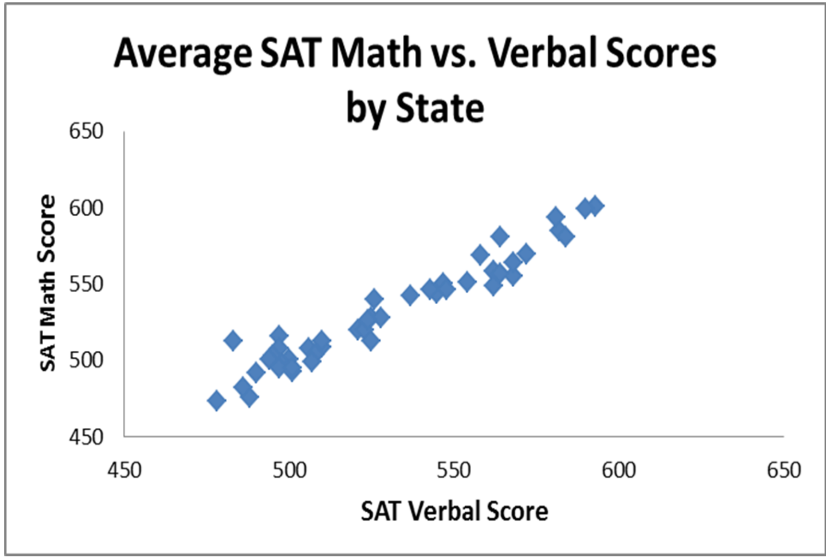
Scatter Diagrams

- Scatter Diagrams are used for paired observations taken from two numerical variables
- The Scatter Diagram:
  - one variable is measured on the vertical axis and the other variable is measured on the horizontal axis

Scatter Diagram Example

Average SAT scores by state: 1998

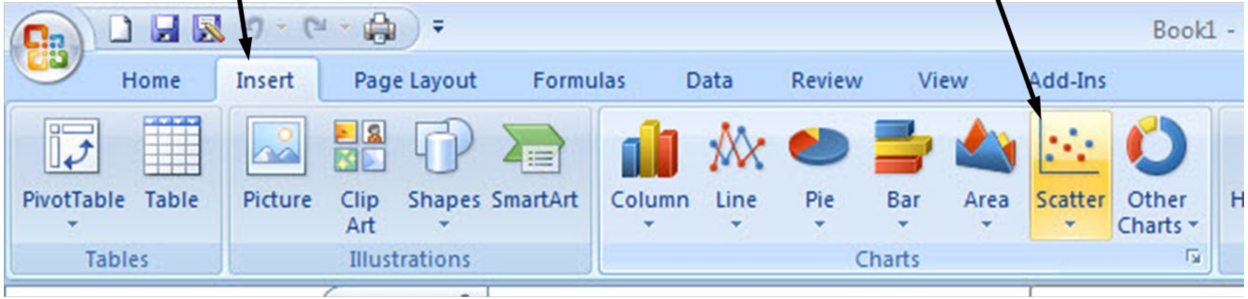
	Verbal	Math
Alabama	562	558
Alaska	521	520
Arizona	525	528
Arkansas	568	555
California	497	516
Colorado	537	542
Connecticut	510	509
Delaware	501	493
D.C.	488	476
Florida	500	501
Georgia	486	482
Hawaii	483	513
. . .		
W.Va.	525	513
Wis.	581	594
Wyo.	548	546



Scatter Diagrams in Excel

1 Select the Insert tab

2 Select Scatter type from the Charts section



3 When prompted, enter the data range, desired legend, and desired destination to complete the scatter diagram

Data Presentation Errors (1 of 2)

Goals for effective data presentation:

- Present data to display essential information
- Communicate complex ideas clearly and accurately
- Avoid distortion that might convey the wrong message

Data Presentation Errors (2 of 2)

- Unequal histogram interval widths
- Compressing or distorting the vertical axis
- Providing no zero point on the vertical axis
- Failing to provide a relative basis in comparing data between groups