Chapter 1- Describing Data: Graphical

Ercan Karadas

New York University

ercan@nyu.edu

Spring, 2016

Overview

Overview of Statistics

Descriptive Statistics

Describing Categorical Data

Describing Numerical Data

Data Presentation Errors

Decision Making in an Uncertain Environment

Everyday decisions are based on incomplete information! Here is some examples:

- Will the job market be strong when I graduate?
- Will the price of Yahoo stock be higher in six months than it is now?
- Will interest rates remain low for the rest of the year if the federal budget deficit is as high as predicted?

How to deal with these uncertain situations?

- Collect data to assist decision making under uncertainty.

Statistics is a tool to help process, summarize, analyze, and interpret data.

Some 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 or targeted subset of the population.

- n represents the sample size

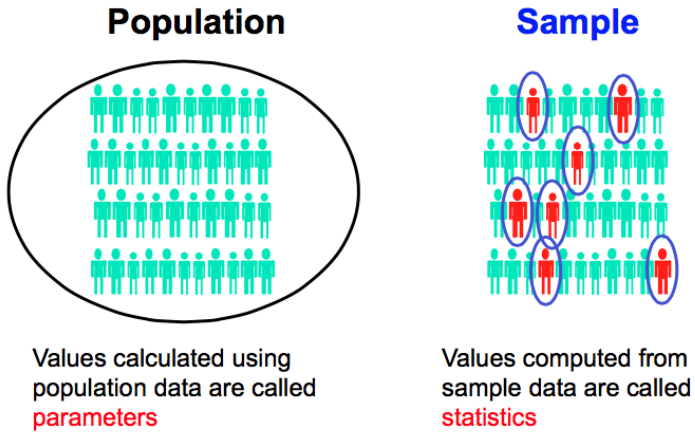
A **parameter** is a specific characteristic of a population.
A **statistic** is a specific characteristic of a sample.

Basic Idea: Use a specific characteristic of the sample (*statistic*) to make inference about a specific characteristic of the population (*parameter*).

Example: Predicting population height

Goal: Suppose we would like to estimate average height of NYU students.

- Parameter: average height of all NYU students. Usually denoted by μ.
- Statistic: average height of all NYU students in the sample. Usually denoted by $\bar{X}$.



- It is considered that population parameter μ is infeasible to obtain or calculate (otherwise working with the sample would be meaningless), therefore from the sample data we calculate the value of $\bar{X}$ and then based on this value make some prediction about μ. Broadly,

$$\text{(inferential) statistics is going from } \quad \overline{X} \quad \text{ to } \quad \mu.$$

Population vs Sample

Some more 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 NYU students

How do we choose the sample?

- Use simple random sampling in which,
  - each member of the population is chosen strictly by chance,
  - each member of the population is equally likely to be chosen,
  - Therefore: every possible sample of n members of the population is equally likely to be chosen as the sample.

Descriptive and Inferential Statistics

There are two branches/stages of statistics:

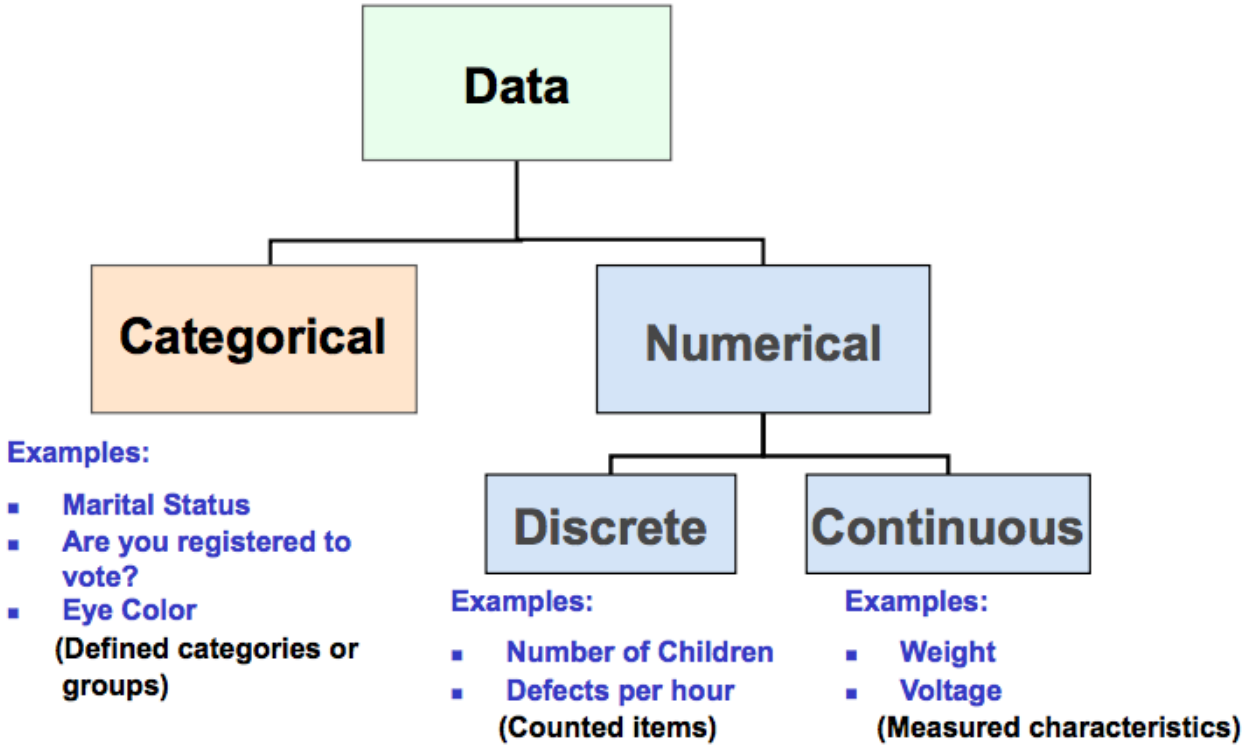
Descriptive statistics: Graphical and numerical procedures to summarize and process data. Main steps:

- Collect data, e.g., survey or national statistical data resources
- Present data, e.g., tables and graphs
- Summarize data, e.g., sample mean

Inferential statistics: Using sampled data to make predictions, forecasts, and estimates to assist decision making about some population characteristics.

- Broadly, it consists of estimation and hypothesis testing.
- In the previous example
  - Estimation: estimate the population mean height using the sample mean height
  - Hypothesis testing: test the claim that the population mean height is 5.7" .

Data Types



- Data in raw form are usually not easy to use for decision making
- Some type of organization is needed:
  - Table
  - Graph

Data Types cont'd

The type of graph to use depends on the variable being summarized. For categorical and numerical data, some of the most frequently used graphs are:

Categorical Variables:

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

Numerical Variables:

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

Goal: effective data presentation

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

Categorical Data

1. Frequency Distribution Table

Summarize data by category

Example: Hospital Patients by Unit

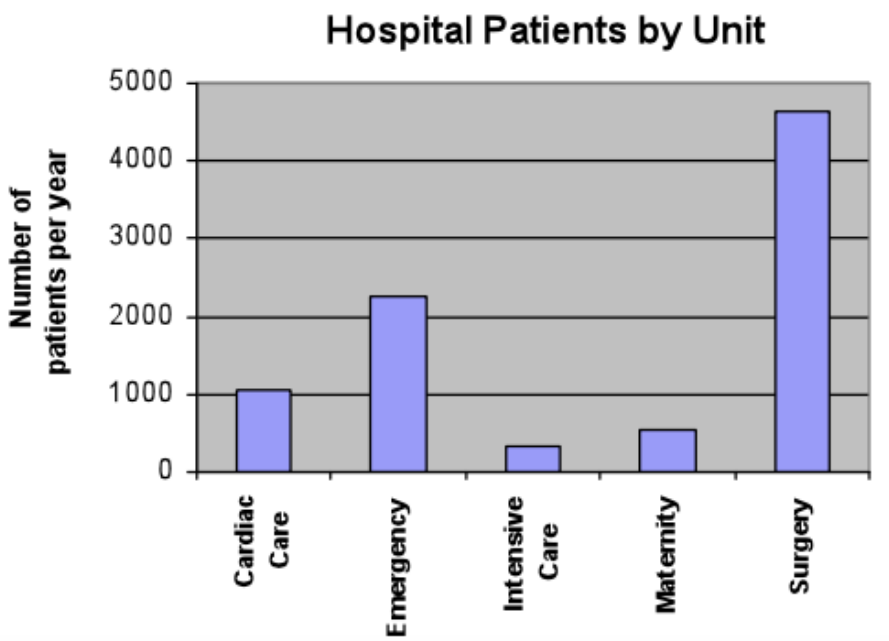
Hospital Unit	Number of Patients	Percent (rounded)
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50
Total:	8,819	100.0

(Variables are categorical)

Categorical Data

2. Bar Chart

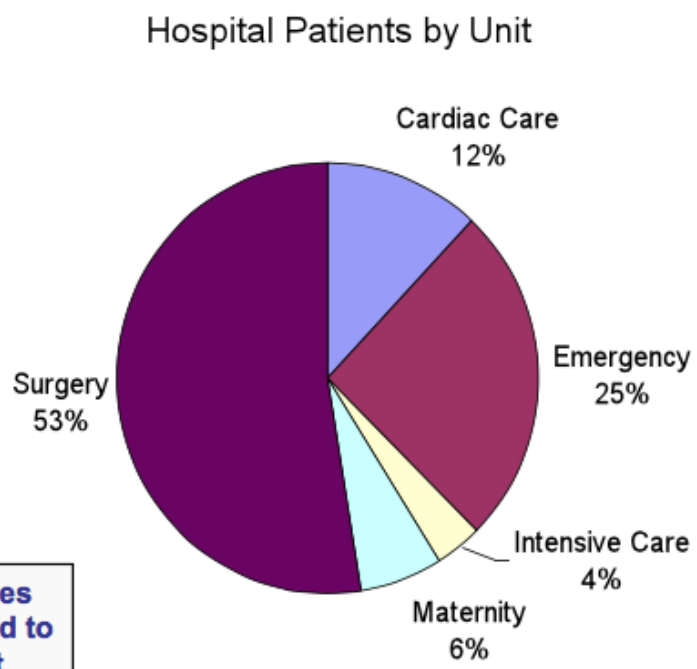
Hospital Unit	Number of Patients
Cardiac Care	1,052
Emergency	2,245
Intensive Care	340
Maternity	552
Surgery	4,630



Categorical Data

3. Pie Chart

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)

Numerical Data

1. Frequency Distribution Table

- A frequency distribution is a list or a table just like in the categorical data case above
- However, here *we need to determine* the categories or ranges within which the data fall
- and the corresponding frequencies with which data fall within each class or category

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

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

- 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 the range: $58 - 12 = 46$
- Select number of classes: 5 (usually between 5 and 15)
- Compute interval width: 10 (46/10 then round up)
- Determine interval boundaries: 10 but less than 20, 20 but less than 30, . . . , 60 but less than 70, ...
- Count observations & assign to classes

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

Numerical Data

2. Cumulative Frequency Distribution Table

We can expand the frequency distribution table by adding cumulative frequency and cumulative percentages:

Data in ordered array:

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

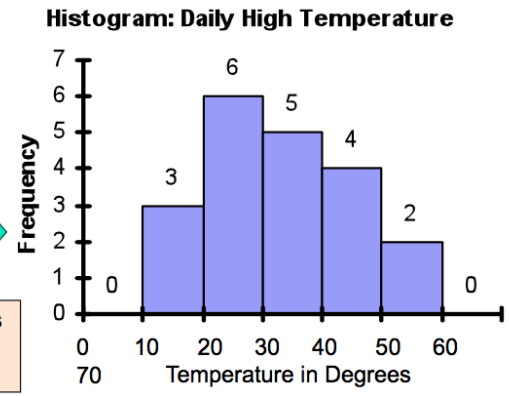
Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

Numerical Data

3. 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

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)

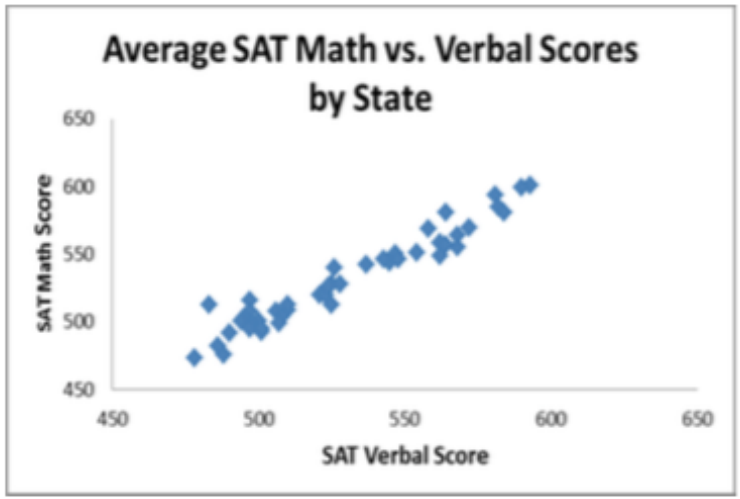
Numerical Data

4. Scatter Diagrams

- Scatter Diagrams are used for 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

Average SAT scores by state: 1998

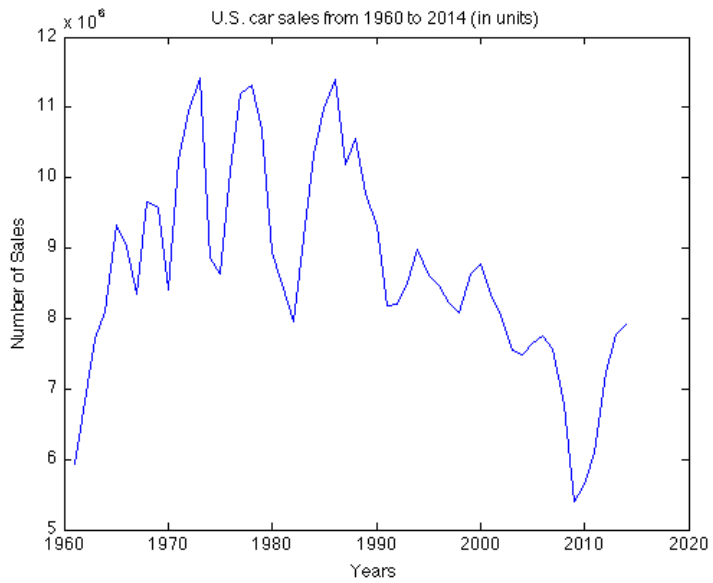
	Verbal	Math
Alabama	562	558
Alaska	521	520
Arizona	525	528
Arkansas	568	555
California	497	516
Colorado	537	542
Connecticut	510	509
Delaware	501	493
D.C.	488	476
Florida	500	501
Georgia	486	482
Hawaii	483	513
...		
W.Va.	525	513
Wis.	581	594
Wyo.	548	546



Numerical Data

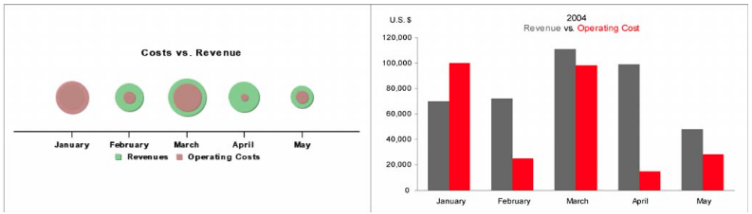
5. Line Chart

- A line chart (time-series plot) is used to show the values of a variable over time
- Time is measured on the horizontal axis, and the variable of interest on the vertical axis
- For example, following line chart depicts how total number of car sales has evolved since 1960's. Note that on top of the vertical axis, there is '$\times 10^6$', which means all the values on the vertical axis should be multiplied by that number. For instance, car sales in 1960 were approximately 6×10^6 (6 million).



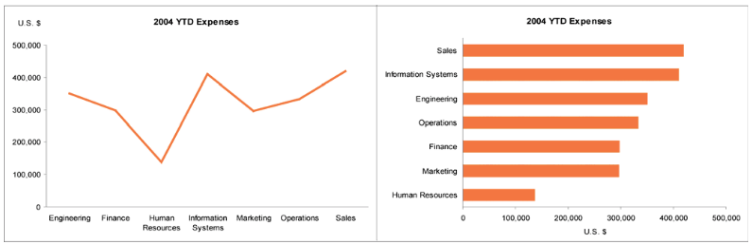
Data Presentation Errors

- Use the Right Graph Type
- In the following graph, presenter has used an exotic graph (on the left) instead of simply using a more appropriate chart type, like bar chart (on the right)



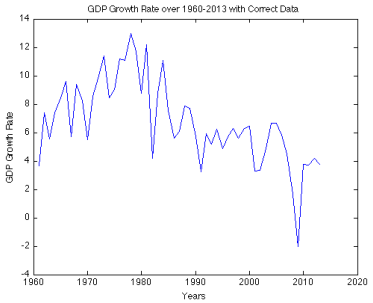
or even worse

- In the following example, using line chart gives the wrong impression that the graph exhibits some sort of change between two points on the graph because of the slope (on the left). Again, in this case simply bar chart would be more direct and expressive.

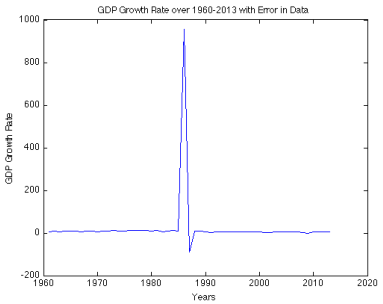


Data Presentation Errors cont'd

- Always check extreme values, outliers and likely data entrance problems.
  - For example, the following chart depicts US GDP growth rate from 1960 to 2013



  - But you would get the following picture if you mistakenly added just one extra zero in 1986 GDP data!



  - As it is clear from the above figure that all the fluctuations in the GDP growth rate disappears because scale of the vertical axis jumps to 1000 and therefore dominates the figure.
- Unequal histogram interval widths
- Compressing or distorting the vertical axis