Chapter 1

Summary and Display of Univariate Data

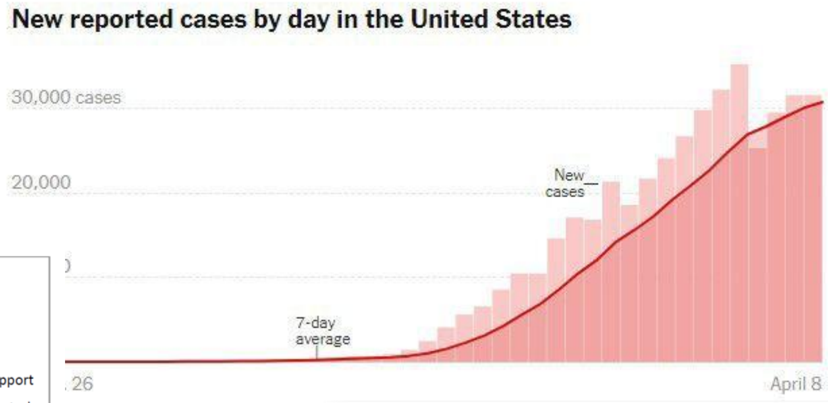
Lecture 2

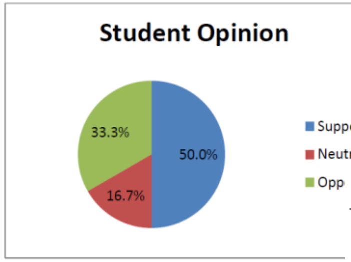
Some Key Statistical Concepts

Classification of Variables

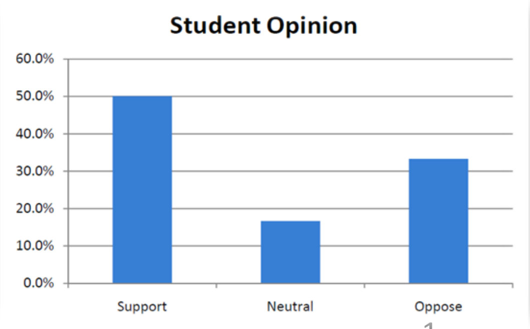
Summarizing data using tables and graphs

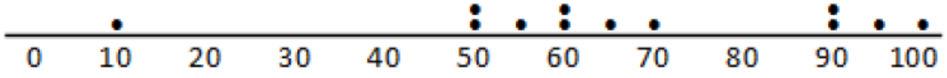
Dr. Lasantha Premarathna





Stem	Leaf
2	5
3	
4	1
5	057
6	2359
7	0255
8	0125
9	025





Chapter 1

Learning Outcomes

Demonstrate the ability to apply fundamental concepts in exploratory data analysis.

- Distinguish between different types of data.
- Interpret examples of methods for summarizing data sets, including common graphical tools (such as boxplots and histograms) and summary statistics (such as mean, median, variance and IQR).
- Assess which methods for summarizing a data set are most appropriate given data.
- Identify the features that describe a data distribution.
- Use an appropriate software tool for data summary and exploratory data analysis.

What is Statistics?

Statistics is a science involving the design of studies, data collection, summarize and analyse data, interpreting results and drawing conclusions.

Statistics is the science of learning from data, and of measuring, controlling and communicating uncertainty

Statistics is a branch of applied mathematics dealing with data collection, organization, analysis, interpretation and presentation.

Some Key Statistical Concepts...

Population and Samples

- Population: all subjects of interest in a particular study
- Sample: subset of the population

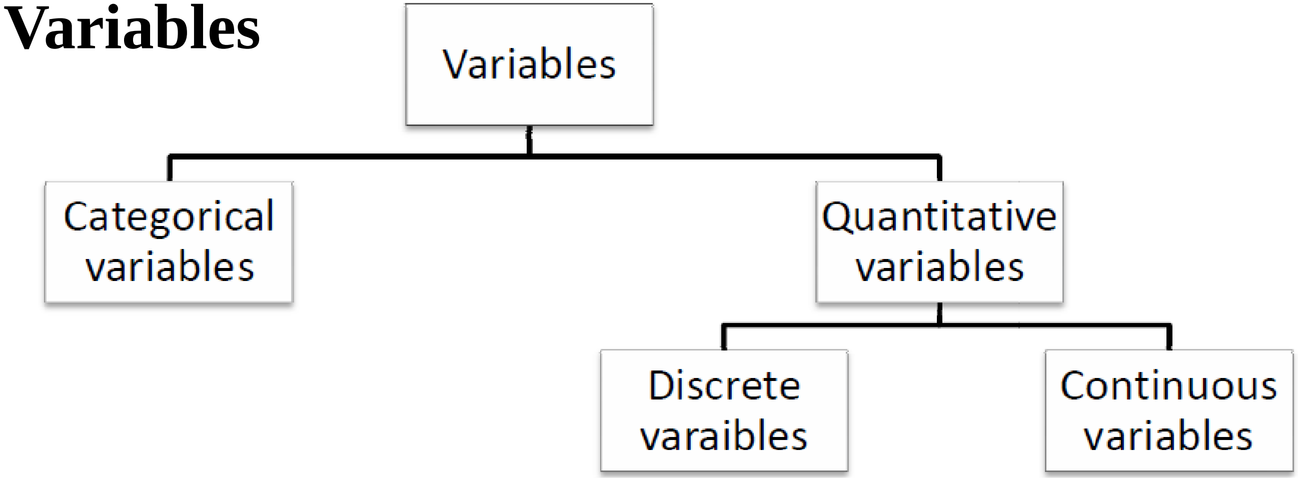
Parameter and Statistic

- Parameter : a descriptive measure of a *population.*
- Statistic : a descriptive measure of a *sample***.**

Census and Sample Survey

- Census : collecting data for the entire population
- Sample survey : collecting data for a sample

Classification of Variables



Variables

Categorical variables

Quantitative variables

Discrete varaibles

Continuous variables

- A variable can be classified as **categorical** if each observation belongs to one of a set of categories

- A variable is called **quantitative** if observations on it take numerical values that represent different magnitudes of the variable

Ex: Categorical(A) or quantitative (B) ?

1. Number of siblings in a family
2. County of residence
3. Distance (in km) of commute to school
4. Blood type

- A **quantitative variable is discrete** if its possible values form a set of separate numbers, such as 0,1,2,3,…. Also can 1, 1.5, 2, 2.5

  e.g. shoe sizes → distinct val.

- A **quantitative variable is continuous** if its possible values form an interval

↳ for any given 2 val, any val in between is possible, e.g Weight

↳ 48, 48.12, 57.375, 63

- Any is possible if accurately measured

✓ Dog weight

X speeding tickets

X People waiting in line.

Descriptive vs. Inferential Statistics

→ Describes

- Descriptive Statistics refers to methods for summarizing the data. Summaries consist of graphs and numbers → e.g. Avg. grade.

  ↳ mean, sd.

  ↳ pie, bar (sum up)

- Inferential statistics refers to methods of making decisions or predictions about a population based on data obtained from a sample of that population.

  - cannot collect data from pop.
  - After Sample $\xrightarrow[\text{Analyze}]{}$ Decision about pop.

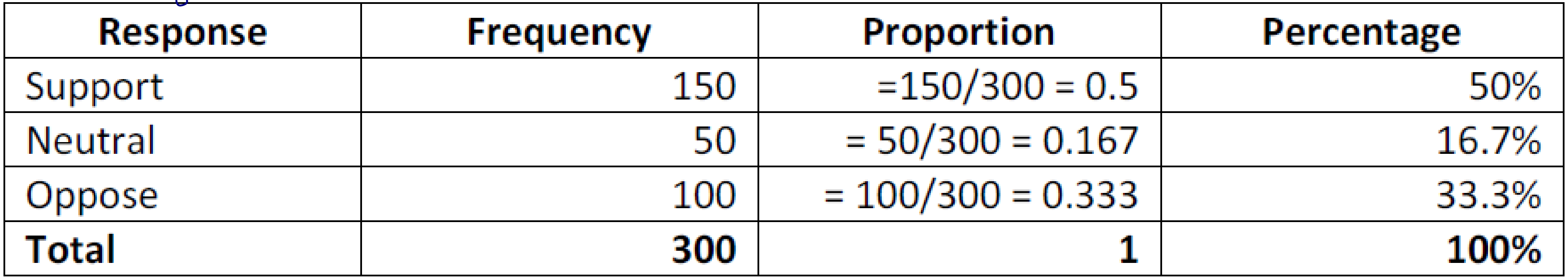
→ for ease of interpretation

Summarizing data using tables and graphs

Frequency Table : A frequency table is a listing of possible values for a variable , together with the number of observations and/ or relative frequencies for each value

e.g. A campus press polled a sample of 300 undergrads in order to study the attitude towards a proposed change in on campus housing regulations. Summary of results of an opinion poll is as follows.

Categorical Count $\frac{\text{Value}}{\text{total}}$ Proportion x 100%



Response	Frequency	Proportion	Percentage
Support	150	=150/300 = 0.5	50%
Neutral	50	= 50/300 = 0.167	16.7%
Oppose	100	= 100/300 = 0.333	33.3%
Total	**300**	**1**	**100%**

↳ rel. freq. table

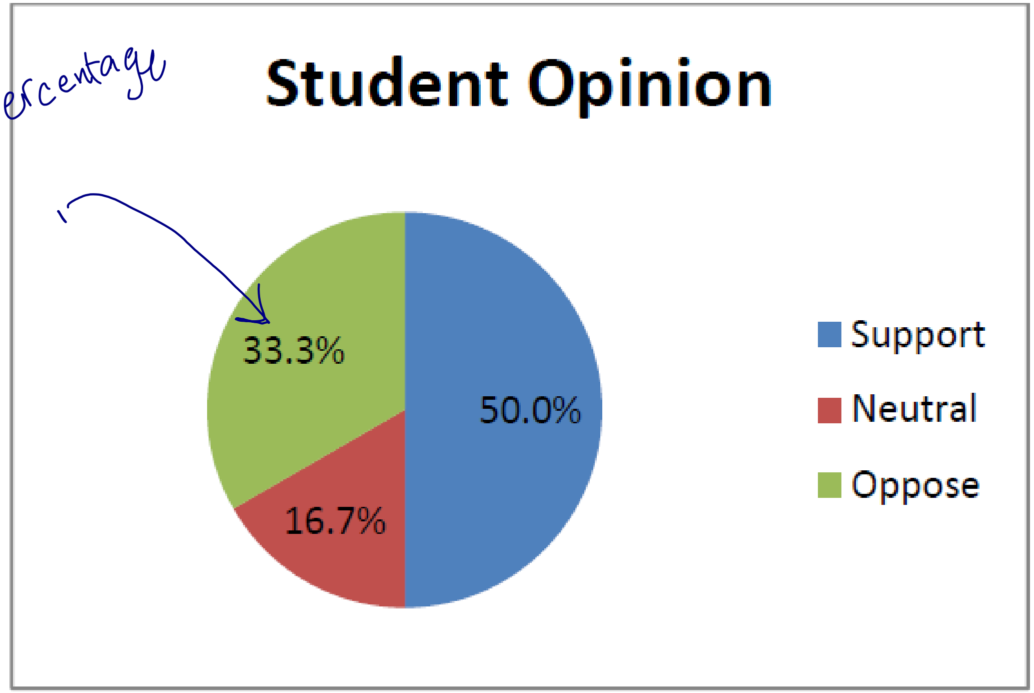
Pie Chart

→ easy to convey ideas.

- used for summarizing a categorical variable
- Drawn as a circle where each category is represented as a "slice of the pie"
- The size of each pie slice is proportional to the percentage of observations falling in that category

Count v.s. percentage

- it is true
- easier to interpret

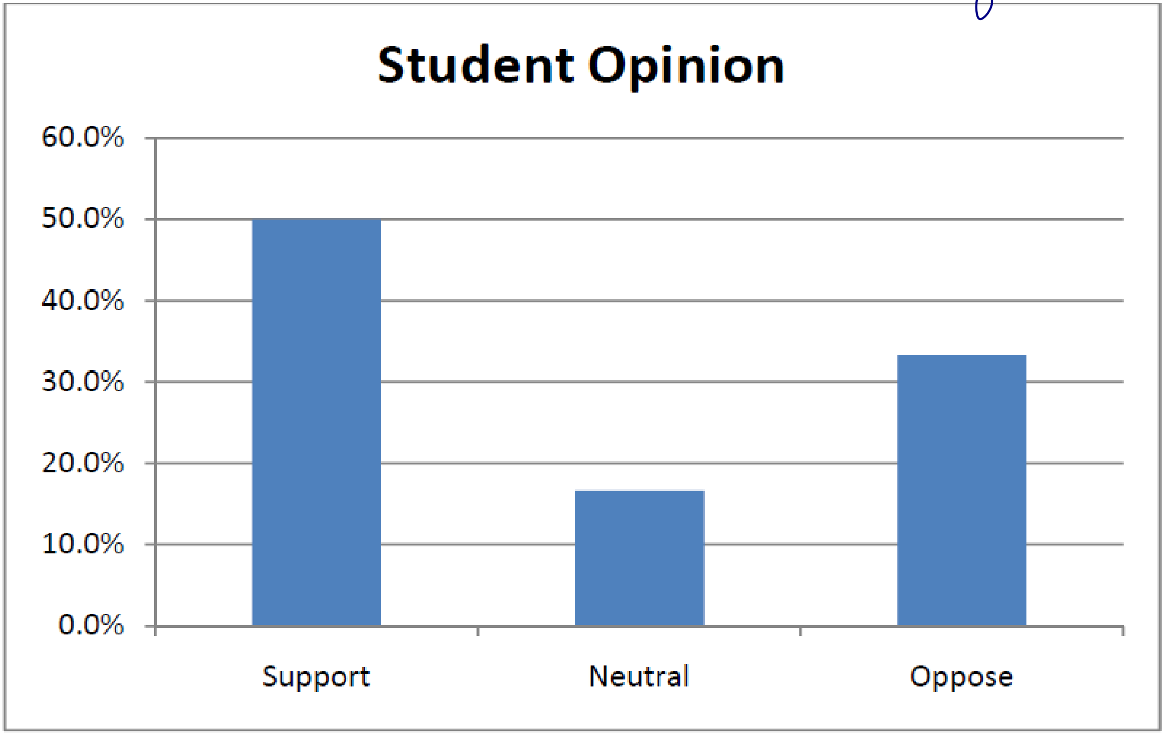


Bar Graphs

→ better for more categories

- used for summarizing a categorical variable
- Bar Graphs display a vertical bar for each category
- The height of each bar represents either counts ("frequencies") or percentages ("relative frequencies") for that category
- Usually easier to compare categories with a bar graph than with a pie chart ← Sort bars from hi to lo for better visibility

- Count o.k.
- easy to compare
- can see height easily



Graphs for Quantitative variables

Dot plot

- Draw a horizontal line and Label it with the name of the variable
- Mark regular values of the variable on it
- For each observation, place a dot above its value on the number line

- good for discrete data.

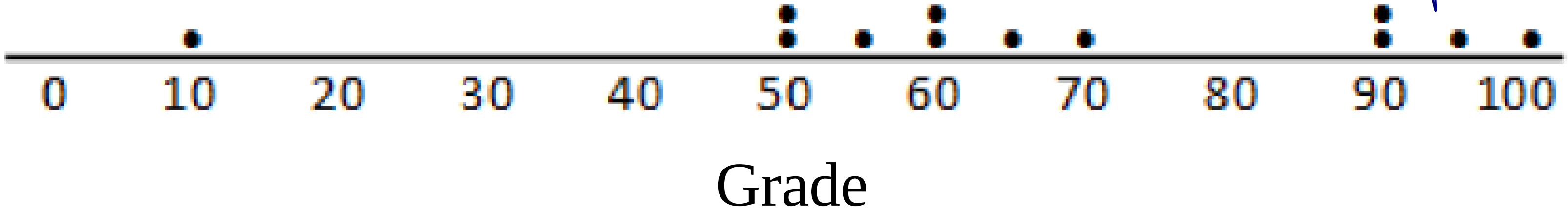
X continuous . data =) ......

e.g. The following set of data is the scores obtained for midterm test on a 0-100 scale. Construct a dot plot.

10, 90, 95, 100, 65, 50, 60, 50, 90, 55, 60, 70

- easily see min / max

- see pts. of conc.



Stem-and-leaf plots

x Contimous

- Separate each observation into a **stem** (first part of the number) and a **leaf** (typically the last digit of the number)
- Write the stems in a vertical column ordered from smallest to largest, including empty stems; draw a vertical line to the right of the stems
- Write each leaf in the row to the right of its stem; order leaves if desired

- if leaf is x10, and there are ones, cannot represent

Example: Consider the following data

80	85	75	90	62
50	55	65	75	82
70	25	92	57	63
72	81	95	41	69

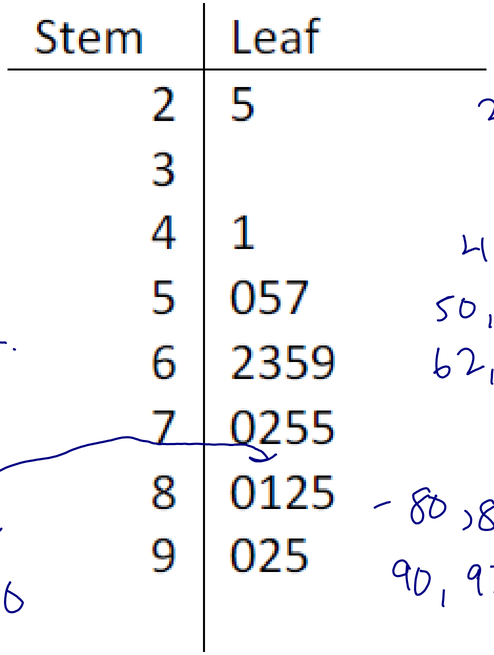
1. See from 20s - 90s

leaf = 0.1 -ok.

leaf = 1

leaf = 10 e.g. 810

<Multiply whole>



Stem	Leaf
2	5
3	
4	1
5	057
6	2359
7	0255
8	0125
9	025

25

41

50, 55, 57

62, 63, 65, 69

- 80, 81, 82, 85

90, 92, 95 etc.

Histograms

Example: Here are the data (number of hours worked) for 25 students in a particular semester. Construct a histogram from the following data.

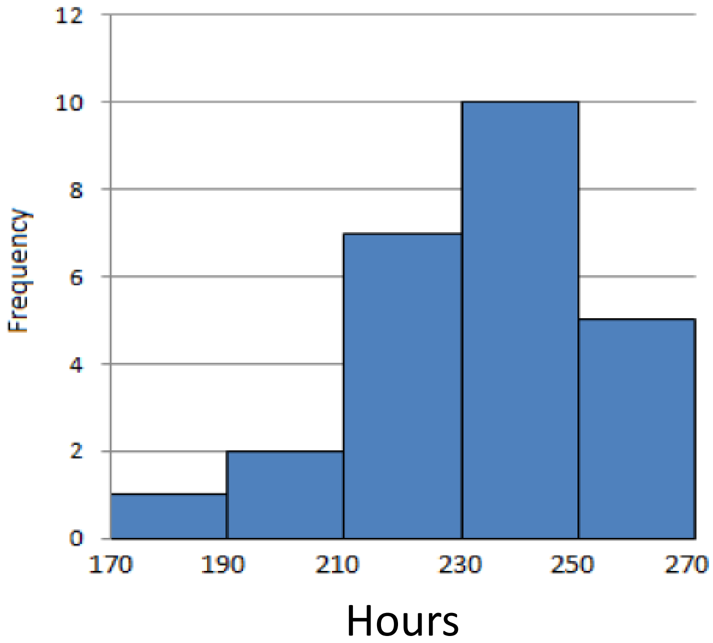
175, 192, 207, 212, 213, 214, 218, 225, 229, 230,
231, 235, 235, 237, 240, 240, 240, 242, 248, 250,
253, 257, 260, 265, 265

Range = 265-175 = 90
Number of intervals needed = 5
Width of an interval = 90/5 = 18

Create the frequency table

Interval(hours)	Frequency
170-190	1
190-210	2
210-230	7
230-250	10
250-270	5

Histogram of Hours worked



Constructing a Histogram

- Divide the range of the data into intervals of equal width
- Count the number of observations in each interval, creating a frequency table
- On the horizontal axis, label the values or the endpoints of the intervals.
- Draw a bar over each value or interval with height equal to its frequency (or percentage), values of which are marked on the vertical axis.
- Label axes and provide proper headings

Summary

- Some Key Statistical Concepts
- Classification of Variables
- Summarizing data using tables and graphs
- Graphs for Categorical variables
- Graphs for Quantitative variables

Before the next class

- Read course guide and assessment regulations
- Review the lecture 2 and related sections in the text book
- Register to iClicker Cloud, if not done already

Next Class:

- Chapter 1 : Summary and Display of Univariate Data (contd)