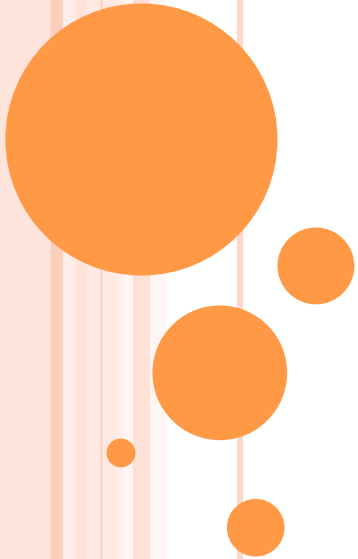


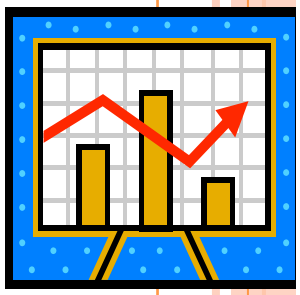
Descriptive statistics

Luis Eduardo Mujica

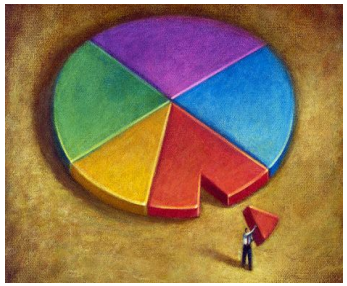
Magda Ruiz

INTRODUCTION

Descriptive statistics



- It is the discipline of summarizing information quantitatively to describe the main features of a collection of data.
  - Tables & Graphs
  - Measures of Central Tendency
  - Measures of Variability
- Examples:
  - Average rainfall in Barcelona last year
  - Number of car thefts in last year
  - Your test results
  - Percentage of males in our class



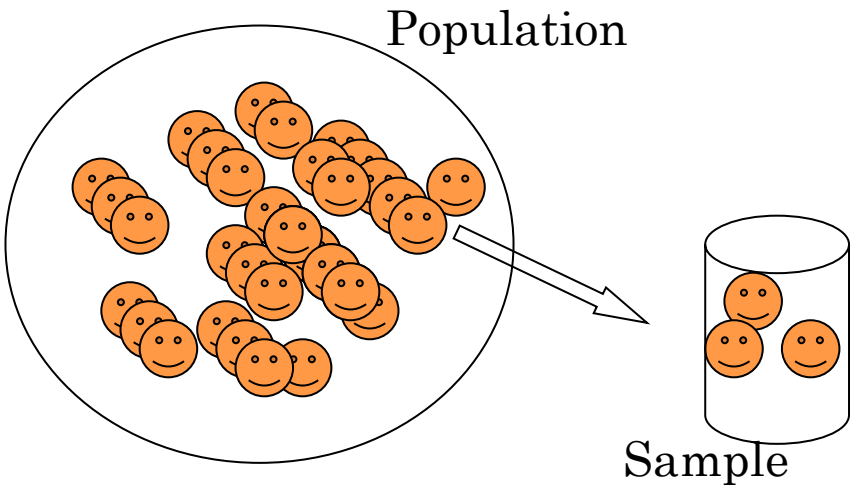
STATISTICAL TERMS (I)

Population

- Complete set of individuals, objects or measurements

Sample

- A sub-set of a population



Statistical terms (II)

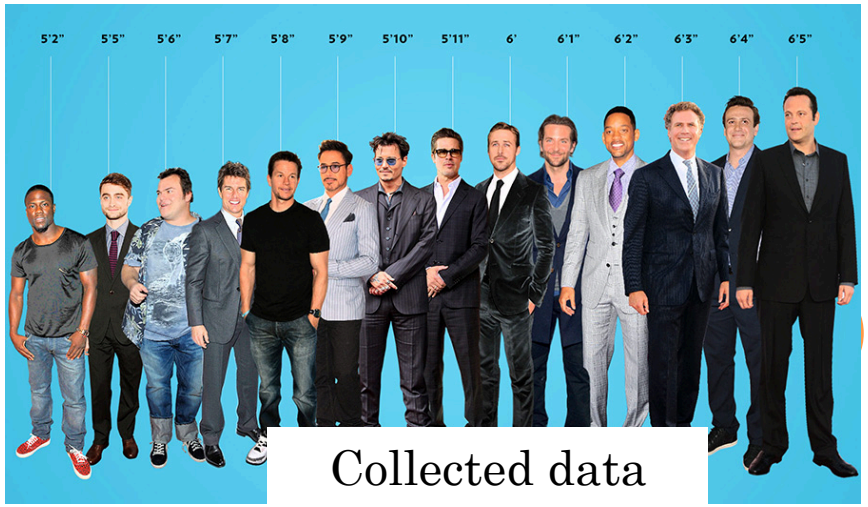
Variable

- A characteristic which may take on different values

Data

- Numbers or measurements collected

Variable: Height



Collected data

STATISTICAL TERMS (III)

Parameter

- A characteristic of a population
  - e.g., the *average* height of all Hollywood men

Statistic

- A characteristic of a sample
  - e.g., the *average* height of a sample of Hollywood men

TYPES OF DESCRIPTIVE STATISTICS

Organize Data

- Tables
  - Frequency distributions
- Graphs
  - Stem and leaf plot – Bar plot - Histogram - Frequency polygon - Pie chart -

Summarize Data

- Central Tendency
  - Mean – Median - Mode
- Variability / Dispersion
  - Standard Deviation- Variance - Range – Quartile – Range Interquartile
- Graphs
  - Box plots

Types of descriptive statistics

Organize Data

Frequency distributions

Absolute frequency

- Number of times that a certain value of variable appears in the study

$$n_i \leq n, \quad \sum_{i=1}^{k} n_i = n$$

Relative frequency

- Number of times that a certain value of variable appears divided by all outcomes

$$f_i = \frac{n_i}{n}$$

Cummulative frequency

- Is the sum of the all frequencies that lie below a particular value
  - Absolute cummulative frequency N_i
  - Relative cummulative frequency F_i

Types of descriptive statistics

Organize Data

Frequency distributions

Example I

- The data set for the quality control of the water from different reactors is as follows, where each number represents the reactor that was chosen as the best:

  1, 5, 3, 1, 2, 3, 4, 5, 1, 4, 2, 4, 4, 5, 1, 4, 2, 4, 2, 2

Reactor	Choosen-Tally Frequency	Absolute Frequency	Relative Frequency	Cumm. Absolute Frequency	Cumm. Relative Frequency
1	xxxx	4	4/20 = 0.2	4	0.2
2	xxxxx	5	5/20 = 0.25	9	0.45
3	xx	2	2/20 = 0.1	11	0.55
4	xxxxx x	6	6/20 = 0.3	17	0.85
5	xxx	3	3/20 = 0.15	20	1

Types of descriptive statistics

Organize Data

Frequency distributions

Example II

- The alloy compressive strengths in pounds per square inch (psi) of 80 specimens of a new aluminum-lithium alloy undergoing evaluation as a possible material for aircraft structural elements.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Types of descriptive statistics

Organize Data

Frequency distributions

Example II (cont)

- Frequency distribution should be grouped
- Divide the range of the data into intervals (class intervals, cells, or bins)
- If possible, the bins should be of equal width.
- The number of bins depends on the number of observations and the amount of scatter or dispersion in the data (usually 5 - 20 bins).

Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

Types of descriptive statistics

Organize Data

Representations of Frequency distributions

Stem-and-Leaf Diagrams

- It is a device for presenting quantitative data in a graphical format to assist in visualizing the shape of a distribution.
- It is a special table where each data value is split into a "leaf" (usually the last digit) and a "stem" (the other digits).
- Gives information about location, spread, extremes, and gaps.

Example II

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem: Tens and hundreds digits (psi); Leaf: Ones digits (psi).

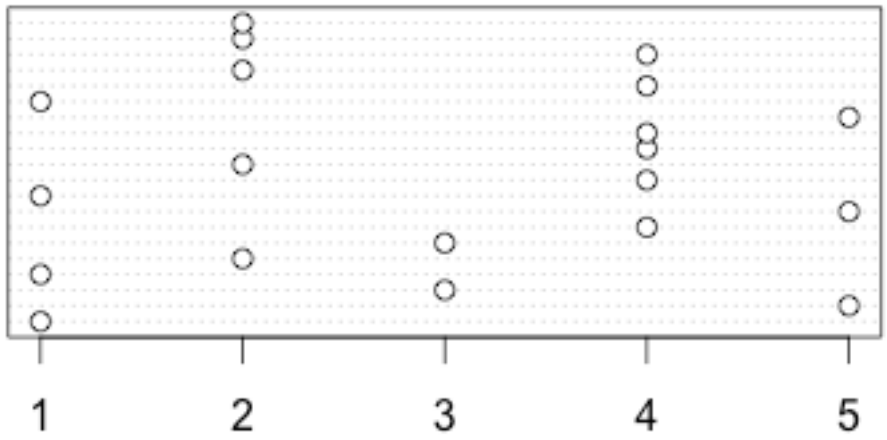
Types of descriptive statistics

Organize Data

Representations of Frequency distributions

Dotplots

- An attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values.
- Each observation is represented by a dot above the corresponding location on a horizontal measurement scale.
- When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically



Example I

Reactor	Absolute Frequency
1	4
2	5
3	2
4	6
5	3

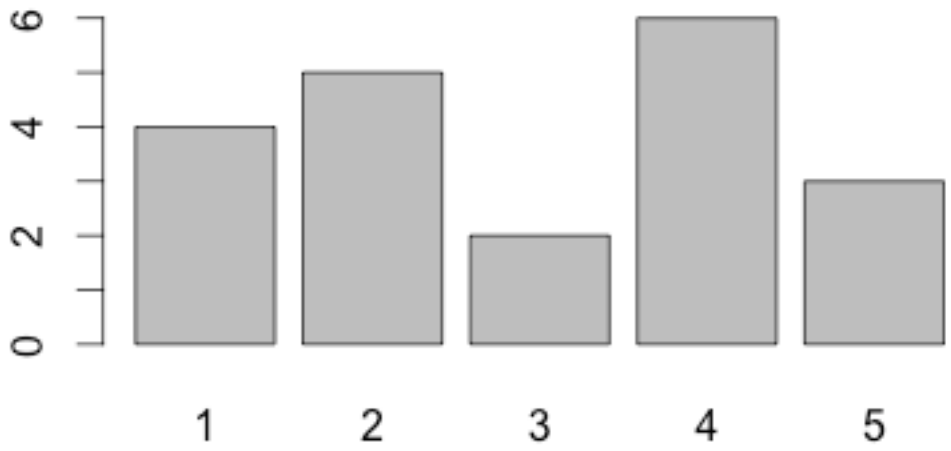
Types of descriptive statistics

Organize Data

Representations of Frequency distributions

Bar graph (Bar chart)

- It is a chart with rectangular bars with lengths proportional to the frequency of each value.



Example I

Reactor	Absolute Frequency
1	4
2	5
3	2
4	6
5	3

Types of descriptive statistics

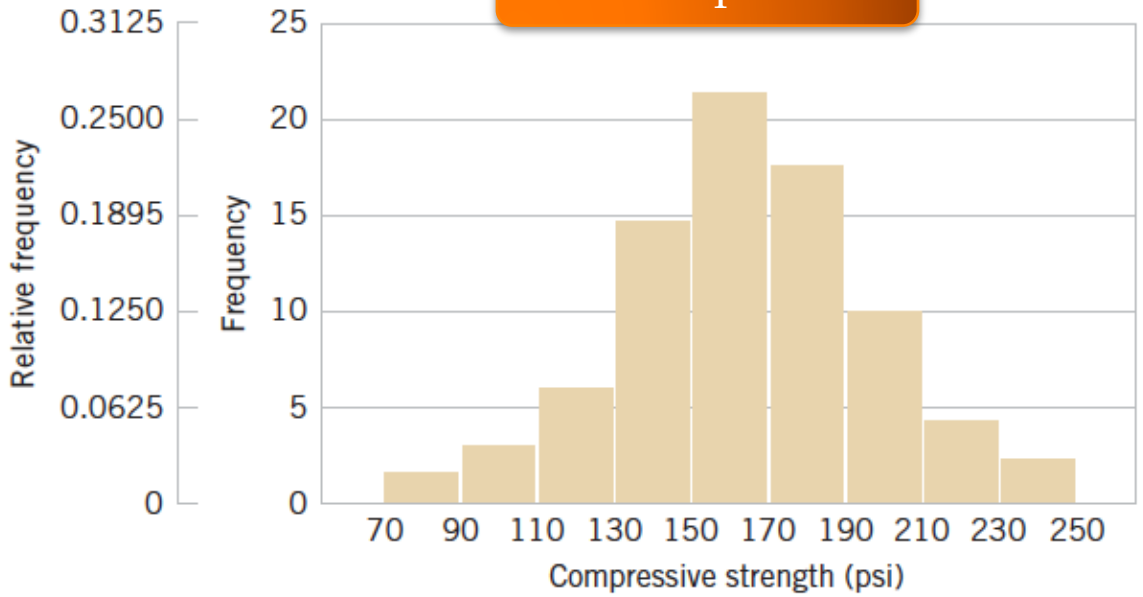
Organize Data

Representations of Frequency distributions

Histogram

- It is a visual display of the frequency distribution
  - X values (or midpoints of class intervals) on x axis
  - Plot each *f(x)* with a bar, equal size, touching
  - No gaps between bars

Example II

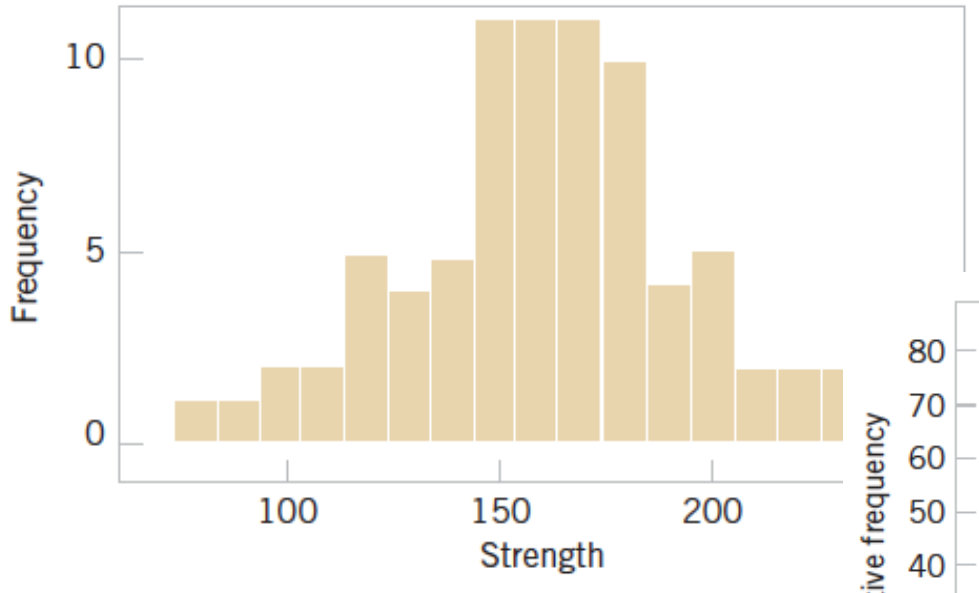


Types of descriptive statistics

Organize Data

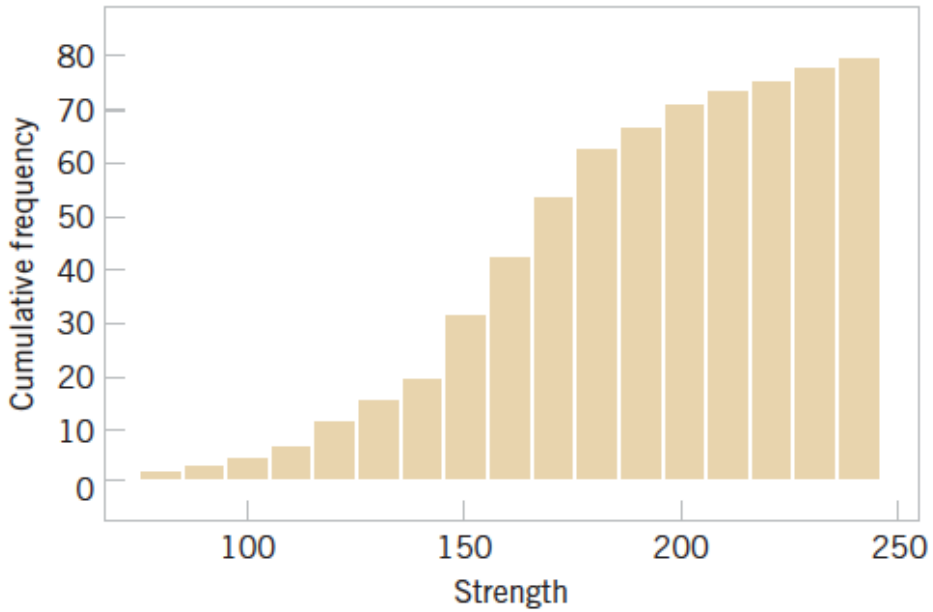
Representations of Frequency distributions

Histogram



Example II

Cumulative histogram



Types of descriptive statistics

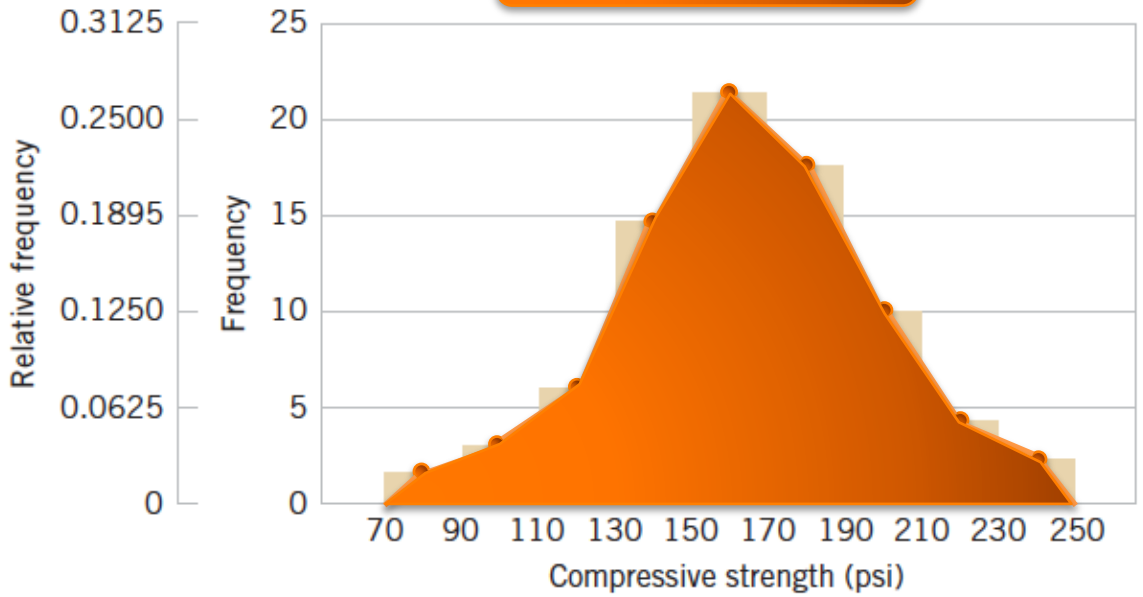
Organize Data

Representations of Frequency distributions

Frequency Polygons

- Depicts information from a frequency table or a grouped frequency table as a **line graph**

Example II



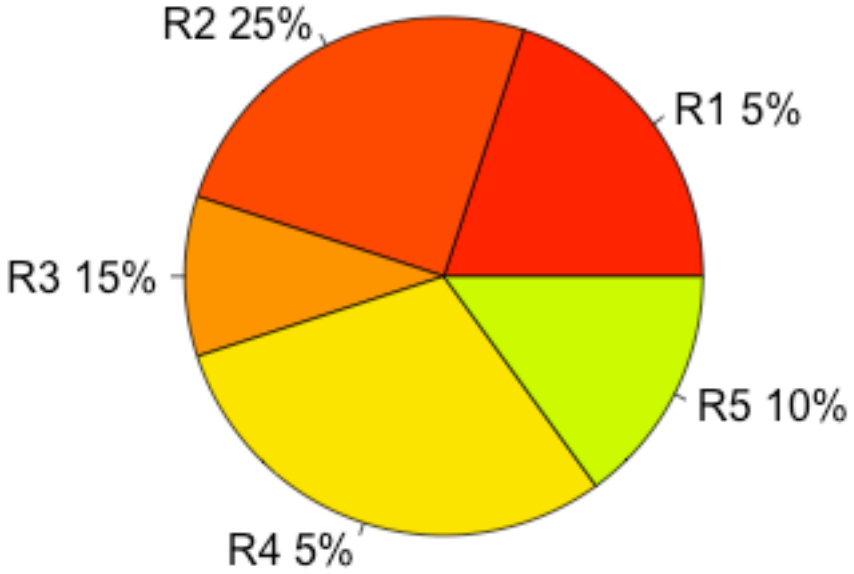
Types of descriptive statistics

Organize Data

Representations of Frequency distributions

Pie Charts

- It is a pie divided into sectors, illustrating numerical proportion
- The arc length of each sector (and consequently its central angle and area), is proportional the frequency of each value.



Example I

Types of descriptive statistics
Summarize Data

Central Tendency measures

- They are computed to give a "center" around which the measurements in the data are distributed

Variation or Variability measures.

- They describe "data spread" or how far away the measurements are from the center.

Relative Standing measures

- They describe the relative position of specific measurements in the data.

Types of descriptive statistics

Summarize Data

Central Tendency

Mean

- Most commonly called the "average."
- It is the "balance point."
- Add up the values for each case and divide by the total number of cases.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Crucial for inferential statistics
- It is not very resistant to outliers
- A "trimmed mean" may be better for descriptive purposes

Types of descriptive statistics

Summarize Data

Central Tendency

Mean (Example)

rim diameter (cm)

	unit 1	unit 2
	12.6	16.2
	11.6	16.4
	16.3	13.8
	13.1	13.2
	12.1	11.3
	26.9	14.0
	9.7	9.0
	11.5	12.5
	14.8	15.6
	13.5	11.2
	12.4	12.2
	13.6	15.5
		11.7
n	12	13
total	168.1	172.6
total/n	14.0	13.3

rim diameter (cm)

	unit 1	unit 2
	~~9.7~~	~~9.0~~
	11.5	11.2
	11.6	11.3
	12.1	11.7
	12.4	12.2
	12.6	12.5
	13.1	13.2
	13.5	13.8
	13.6	14.0
	14.8	15.5
	16.3	15.6
	~~26.9~~	16.2
		~~16.4~~
n	10	11
total	131.5	147.2
total/n	13.2	13.4

Types of descriptive statistics

Summarize Data

Central Tendency

Mean

- If data are grouped in a frequency table, so

$$\bar{x} = \frac{x_1 . n_1 + x_2 . n_2 + x_3 . n_3 + \ldots + x_N . n_N}{N} = \frac{1}{N} \sum_{i=1}^{k} x_i . n_i$$

where $n_1 + n_2 + n_3 + \ldots + n_k = N$

Types of descriptive statistics
Summarize Data
Central Tendency

Median

- Middlemost or most central item in the set of ordered numbers; it separates the distribution into two equal halves

- If odd n, middle value of sequence
  - if X = [1,2,4,6,**9**,10,12,14,17]
  - then **9** is the median
- If even n, average of 2 middle values
  - if X = [1,2,4,6,**9**,**10**,11,12,14,17]
  - then **9.5** is the median; i.e., (9+10)/2

- Median is not affected by extreme values

Types of descriptive statistics

Summarize Data

Central Tendency

Median (Example)

rim diameter (cm)

unit 1	unit 2	
9.7	9.0	
11.5	11.2	
11.6	11.3	
12.1	11.7	
12.4	12.2	
12.6	12.5	
12.9 <--	13.2	13.2
13.1	13.8	
13.5	14.0	
13.6	15.5	
14.8	15.6	
16.3	16.2	
26.9	16.4	

Types of descriptive statistics

Summarize Data

Central Tendency

Mode

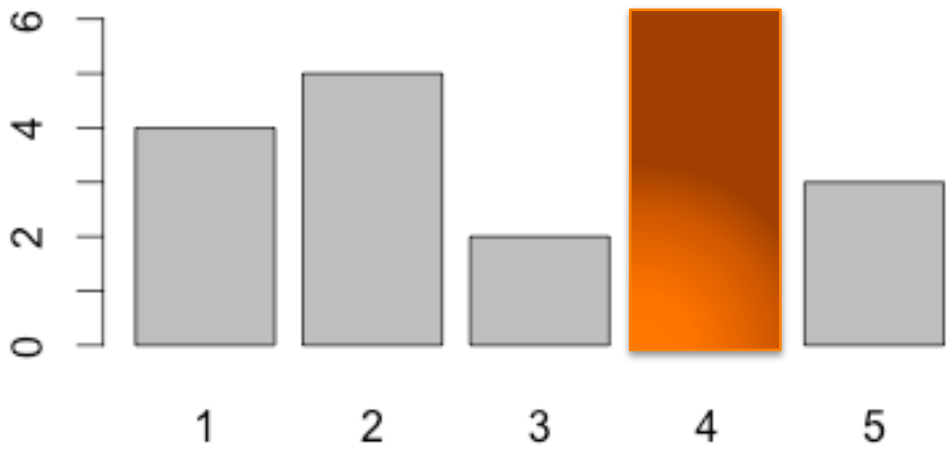
- The mode is the most frequently occurring number in a distribution
  - if X = [1,2,4,**7,7,7**,8,10,12,14,17]
  - then 7 is the mode
- Easy to see in a simple frequency distribution
- Possible to have no modes or more than one mode
  - bimodal and multimodal
- Don't have to be exactly equal frequency
  - major mode, minor mode
- Mode is not affected by extreme values

Types of descriptive statistics

Summarize Data

Central Tendency

Mode (Example)



Example I

Reactor	Absolute Frequency
1	4
2	5
3	2
4	**6**
5	3

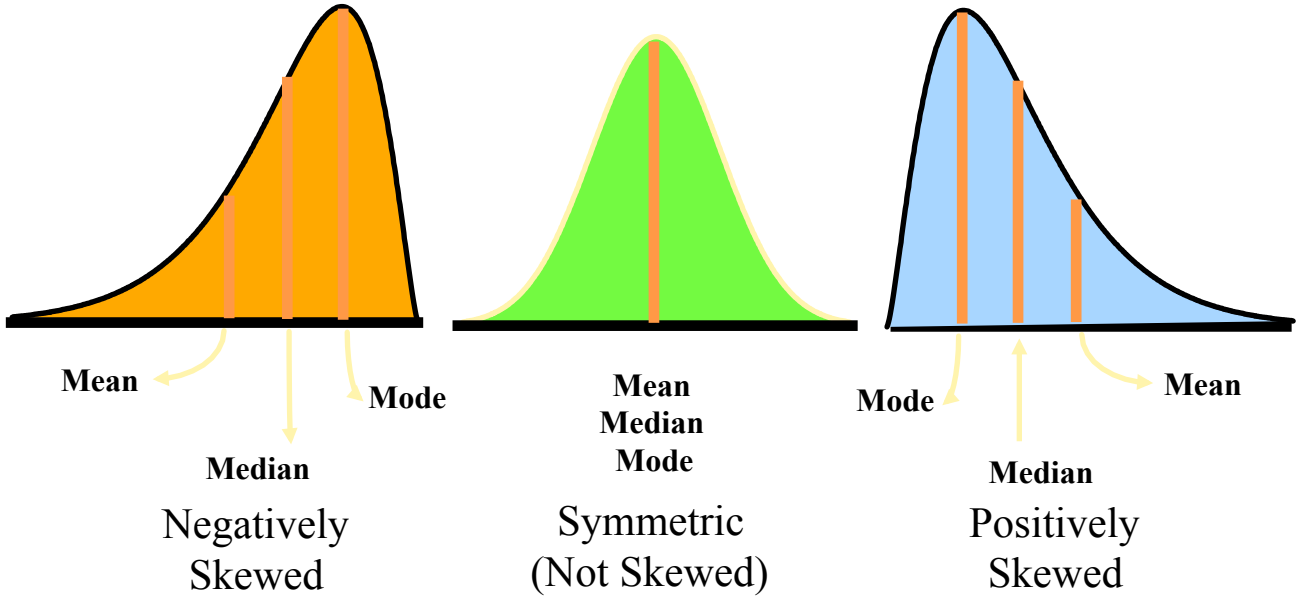
Types of descriptive statistics

Summarize Data

Central Tendency

When to use what?

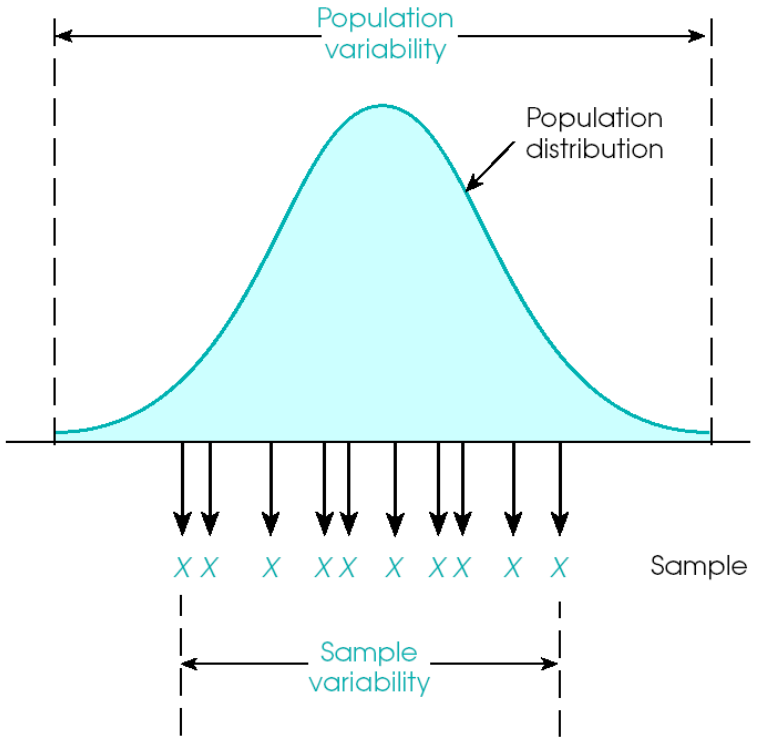
- Mean is a great measure. But, there are time when its usage is inappropriate or impossible.
  - Nominal data: Mode
  - The distribution is bimodal: Mode
  - You have ordinal data: Median or mode
  - Are a few extreme scores: Median



Types of descriptive statistics
Summarize Data
Variability

Dispersion

- How tightly clustered or how variable the values are in a data set.
- Example
  - Data set 1: [0,25,50,75,100]
  - Data set 2: [48,49,50,51,52]
  - Both have a mean of 50, but data set 1 clearly has greater Variability than data set 2.
- Range – Variance - Standard deviation



Types of descriptive statistics

Summarize Data

Variability

Range

- The spread between the lowest and highest values of a variable.
- Highly sensitive to outliers, insensitive to shape.
- It ignores how data are distributed and only takes the extreme scores into account

unit 1	unit 2
9.7	9.0
11.5	11.2
11.6	11.3
12.1	11.7
12.4	12.2
12.6	12.5
13.1	13.2
13.5	13.8
13.6	14.0
14.8	15.5
16.3	15.6
26.9	16.2
	16.4

$$Range\_1 = 26.9 - 9.7 = 17.2$$

$$Range\_2 = 16.4 - 9 = 7.4$$

Types of descriptive statistics

Summarize Data

Variability

Variance

- It measures how far each number in the set is from the Mean
- The average of the squared differences from the Mean

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

- Note: units of variance are squared, and makes variance hard to interpret.
- Ex.: projectile point sample:
  - Mean = 22.6 mm
  - Variance = 38 mm^2
- **What does this mean???**

Types of descriptive statistics
Summarize Data
Variability

Standard deviation

- Square root of variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$$

- Units are in same units as base measurements

- Ex.: projectile point sample:
  - mean = 22.6 mm
  - standard deviation = 6.2 mm

- Mean +/- sd (16.4—28.8 mm)
  - should give at least some intuitive sense of where most of the cases lie, barring major effects of outliers

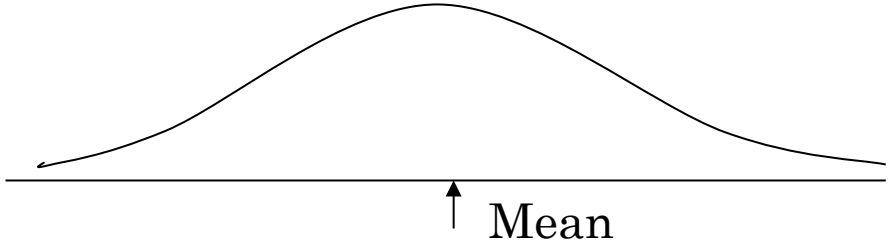
Types of descriptive statistics

Summarize Data

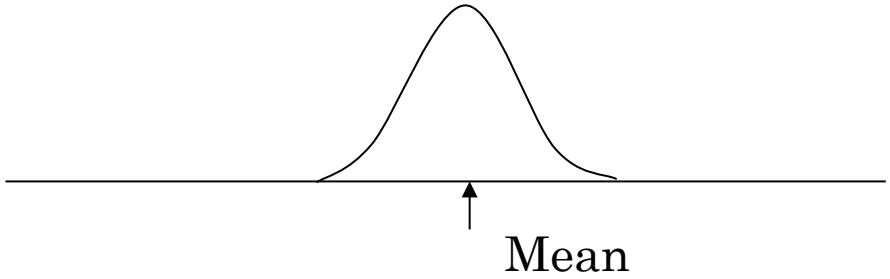
Variability

Variance and Standard deviation

- The larger the variance, the further the individual cases are from the mean.



- The smaller the variance, the closer the individual scores are to the mean.



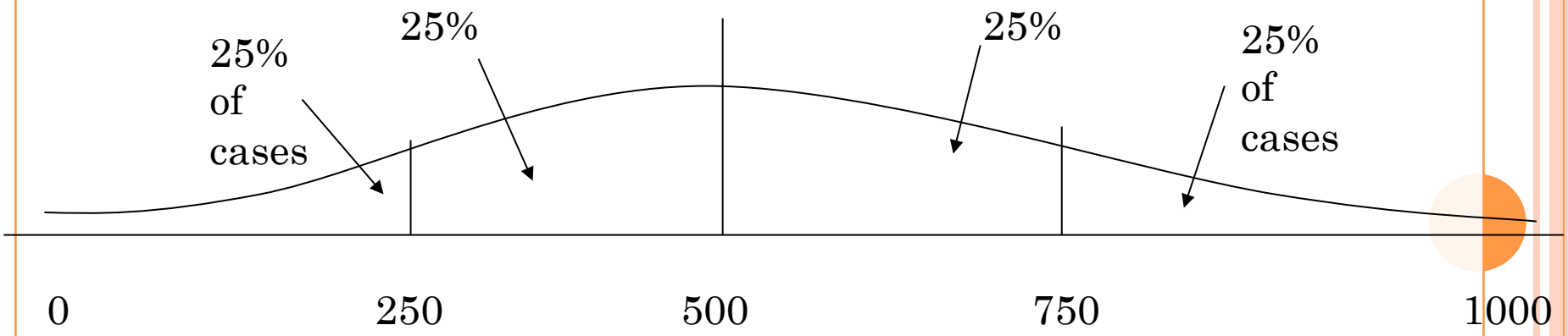
Types of descriptive statistics

Summarize Data

Relative Standing

Percentiles

- The p-th percentile is a number such that at most p% of the measurements are below it and at most 100 – p percent of the data are above it.
  - $P_{25} = Q_1$ => First quartil
  - $P_{50} = Q_2$ => Second quartil => Median
  - $P_{75} = Q_3$ => Third quartil



Types of descriptive statistics

Summarize Data

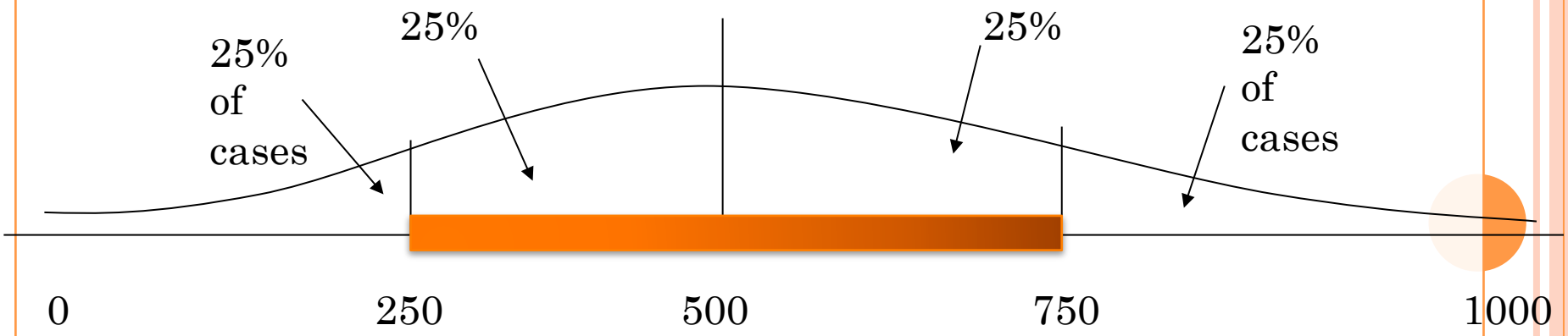
Relative Standing

Interquartile range

- Difference between third & first quartiles

$$IQR = Q_3 - Q_1$$

- Spread in middle 50%
- Not affected by extreme values



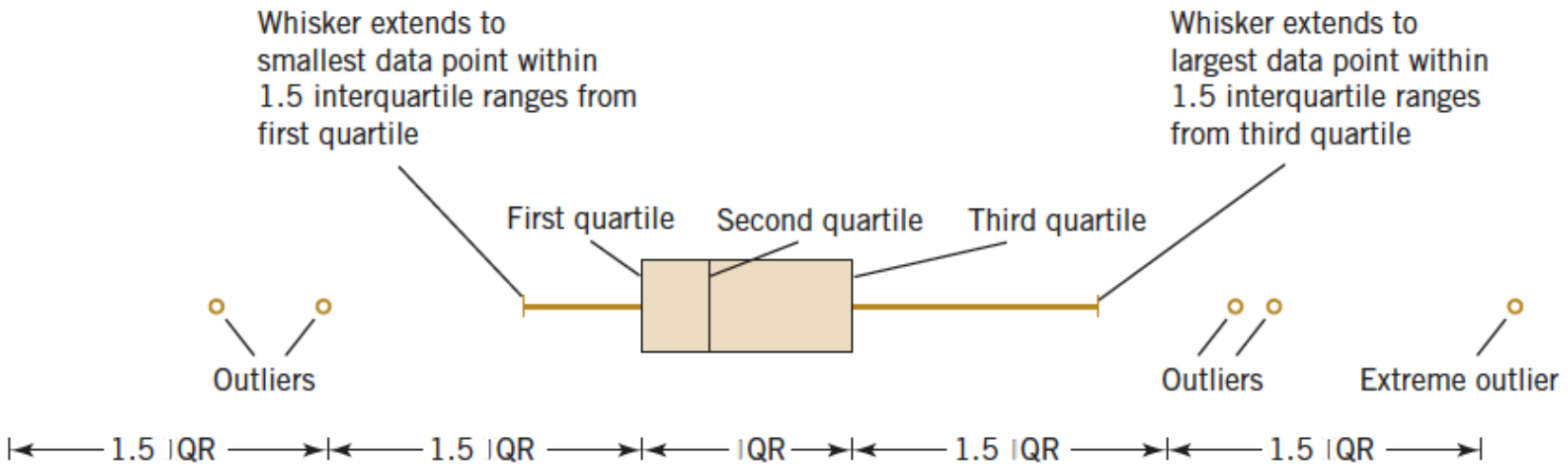
Types of descriptive statistics

Summarize Data

Representation of variability and standing

Box plots

- A plot that have a box from Q_1 to Q_3, and contains also some number to summarize the data: Minimum, Q_1, Median, Q_3, Maximum
- Also indicates outliers identified separately
- Outlier = observation falling
- below LQ – 1.5(IQR)
- or above UQ + 1.5(IQR)



Types of descriptive statistics

Summarize Data

Representation of variability and standing

Box plots