GMBA 7098: Statistics and Data Analysis (Fall 2014)

Descriptive Statistics

Ling-Chieh Kung

Department of Information Management
National Taiwan University

September 22, 2014

Visualizing the data

- We will introduce some common ways to **summarize** a set of data.
  - By **graphs**.
  - By **statistics**.
- This is always the **first step** of any data analysis project: To get intuitions that guide our directions.
- We will introduce more than concepts.
  - We will review basic algebra and mathematical notations.
  - We will practice how to use R to generate them (on Monday).

Road map

- **Visualizing data with graphs**.
- Describing central tendency.
- Describing variability.
- Describing correlation.

Frequency distributions

- Suppose that these are your midterm grades:[1]

42	26	32	34	57	30	58	37	50	30
53	40	30	47	49	50	40	32	31	40
52	28	23	35	25	30	36	32	26	50
55	30	58	64	52	49	33	43	46	32
61	31	30	40	60	74	37	29	43	54

- When data are **ungrouped**, visualizing them is hard.
- We start by **grouping** them into a **frequency distribution**.
  - Grouped data presented in the form of class intervals and frequencies.
- Let's create an intuitive frequency distribution.

[1] Don't worry, there is no midterm at all!

Frequency distributions: an example

42	26	32	34	57	30	58	37	50	30
53	40	30	47	49	50	40	32	31	40
52	28	**23**	35	25	30	36	32	26	50
55	30	58	64	52	49	33	43	46	32
61	31	30	40	60	**74**	37	29	43	54

- Step 1: **Range** $= 74 - 23 = 51$.
- Step 2: Let's divide the range into six **classes**.
- Step 3: **Class width** $\geq \lceil \frac{51}{6} \rceil = 9$.[2] Let's try 10.
  - Why ceiling? Why not floor ($\lfloor \frac{51}{6} \rfloor = 8$)?

[2] In general, $\lceil x \rceil$ is the smallest integer that is no less than x.

Frequency distributions: an example

- The resulting classes:

Class	Class interval	(Which means)
1	$[20, 30)$	$20 \leq x < 30$
2	$[30, 40)$	$30 \leq x < 40$
3	$[40, 50)$	$40 \leq x < 50$
4	$[50, 60)$	$50 \leq x < 60$
5	$[60, 70)$	$60 \leq x < 70$
6	$[70, 80)$	$70 \leq x < 80$

  - How about $[20, 29]$, $[30, 39]$, etc.?
  - How about $(20, 30]$, $(30, 40]$, etc.?

Frequency distributions: an example

- Then we count:

Class interval	Frequency
$[20, 30)$	6
$[30, 40)$	18
$[40, 50)$	11
$[50, 60)$	11
$[60, 70)$	3
$[70, 80)$	1

- This is a complete frequency distribution. It is a set of **grouped data**.
- Some remarks:
  - If there are **outliers**, they should be removed first.
  - A rule of thumb: **5 to 15 classes**.
  - Typically all classes have the same width.
  - Be aware of class endpoints! Classes should NOT overlap with each other.

Something more

- We may add **class midpoints**, **relative frequencies**, and **cumulative frequencies** into a frequency table:

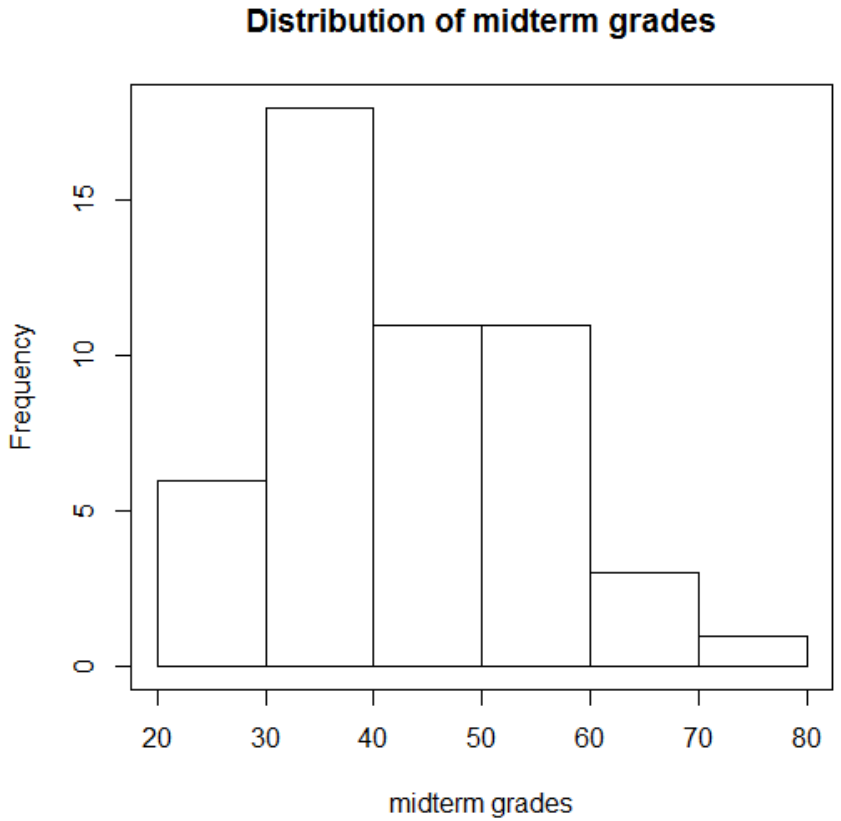
Class interval	Frequency	Class midpoint	Relative frequency	Cumulative frequency
$[20, 30)$	6	25	0.12	6
$[30, 40)$	18	35	0.36	24
$[40, 50)$	11	45	0.22	35
$[50, 60)$	11	55	0.22	46
$[60, 70)$	3	65	0.06	49
$[70, 80)$	1	75	0.02	50

- How about **cumulative relative frequencies**?

Histograms

- A frequency distribution may be depicted as a **histogram**.

Interval	Freq.
$[20, 30)$	6
$[30, 40)$	18
$[40, 50)$	11
$[50, 60)$	11
$[60, 70)$	3
$[70, 80)$	1



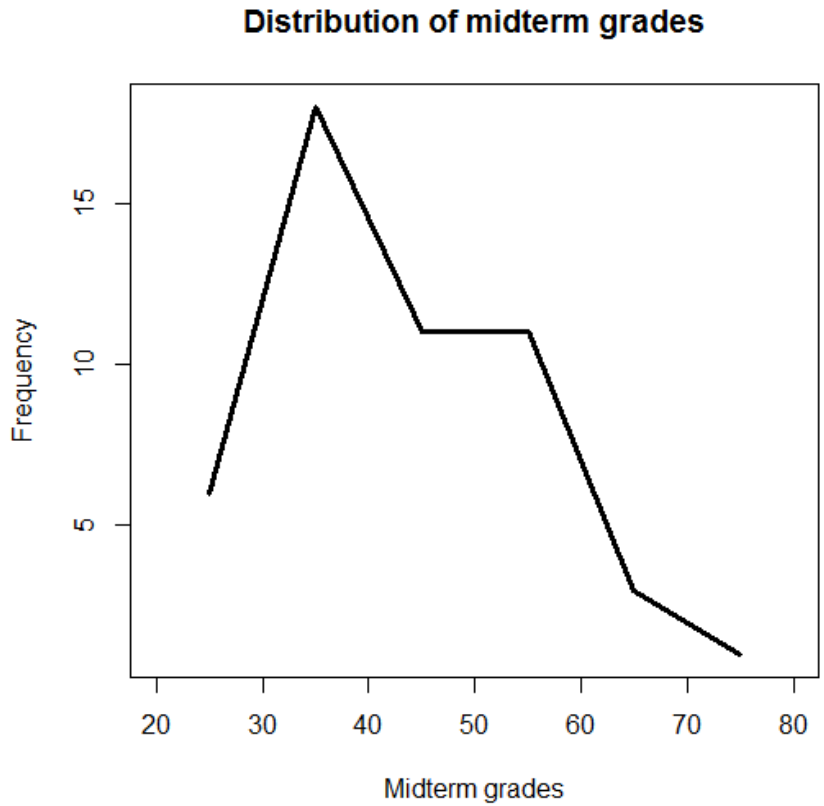
- It consists of a series of **contiguous** rectangles, each representing the frequency in a class.

Histograms

- Histograms may be the most important type of data graphs.
- One particular reason to draw histograms is to get some ideas about the **distribution**.
  - We will discuss distributions in more details.
  - Any outlier?

Frequency polygons

- Alternatively, we may draw a **frequency polygon** by using **line segments** connecting dots plotted at class **midpoints**.
- It is more convenient to use a frequency polygon to **compare multiple** frequency distributions.
- People may **misinterpret** a frequency polygon as a **line chart** (for data with a time sequence).



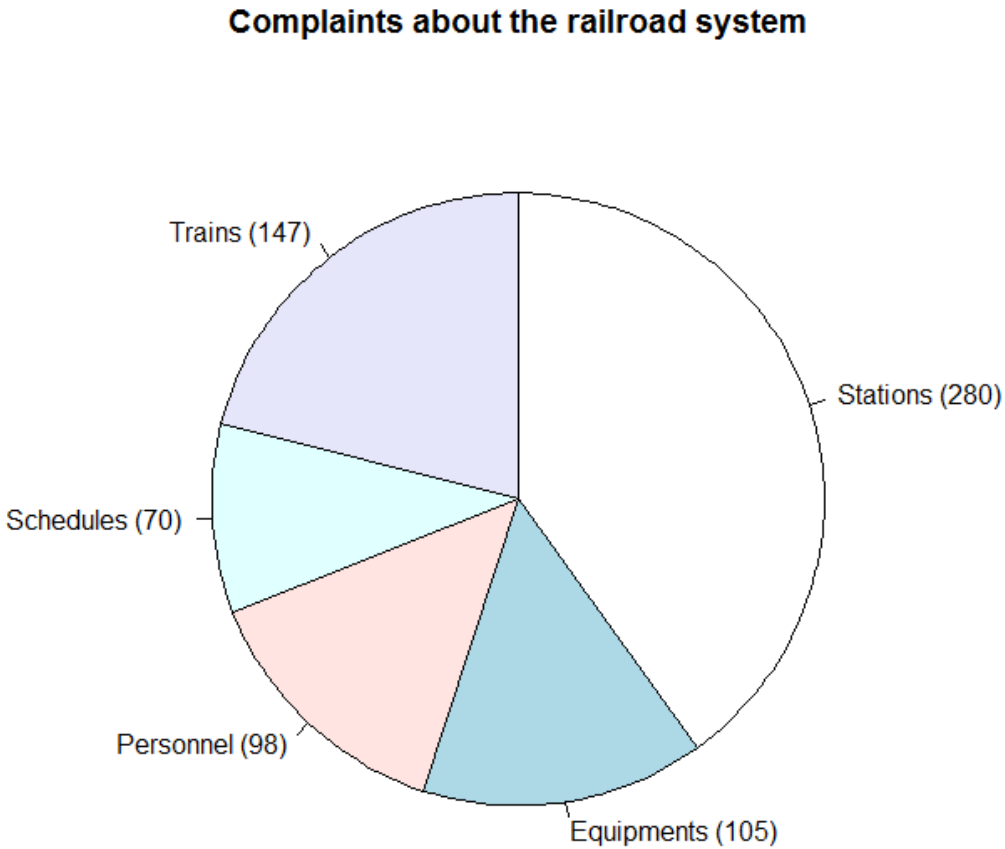
Pie charts

- A **pie chart** is a **circular** depiction of data where each slice represents the percentage of the corresponding category.
- It visualizes **relative frequency distributions** well.
- Consider a survey in a city on what do passengers complain about the railroad system:

Complaint	Number	**Proportion**	**Degrees**
Stations	280	0.58	208.7
Equipment	105	0.22	78.3
Personnel	98	0.20	73.0
Schedules	70	0.14	52.2
Trains	147	0.30	109.6
Total	483	1.00	360.0

Pie charts

Complaint	Number
Stations	280
Equipment	105
Personnel	98
Schedules	70
Train	147

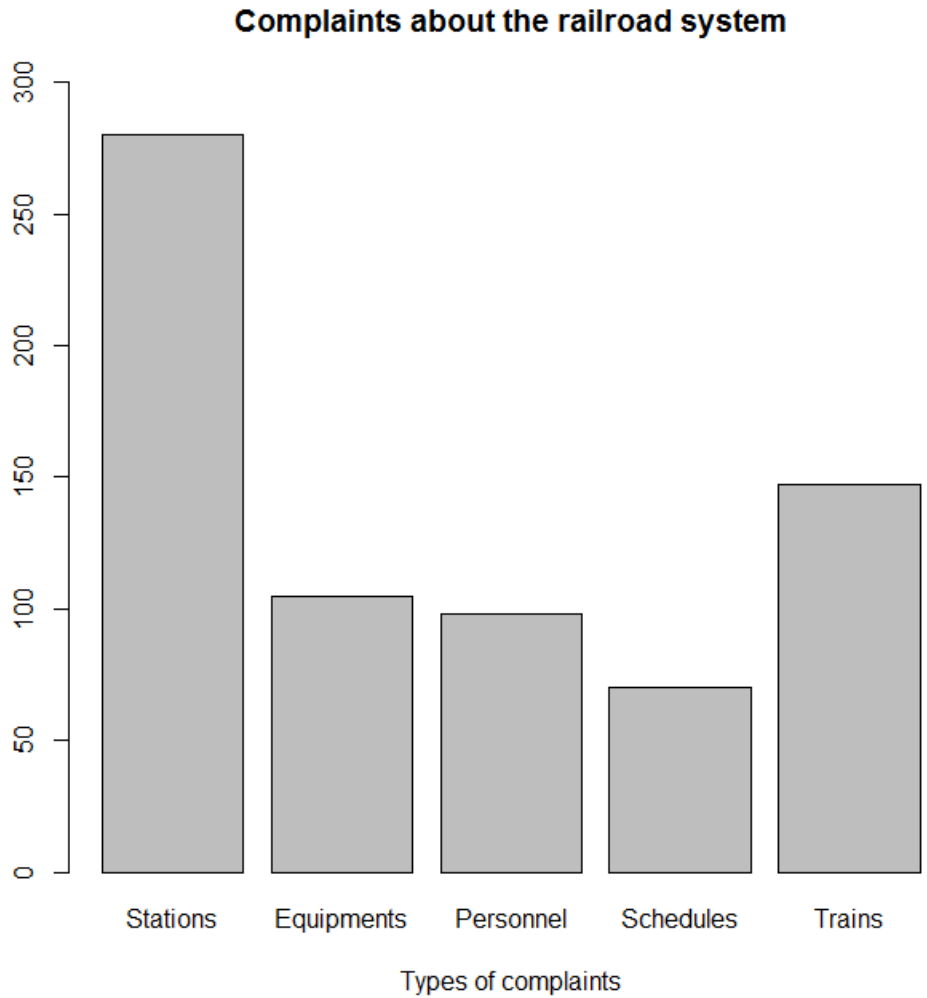


Bar charts

- Pie charts are useful in visualizing the **proportions** of each categories.
- In demonstrating the **differences** among categories, a **bar chart** is a better choice.
  - The larger the category, the longer the bar.
  - Some people draw bars vertically; some horizontally.

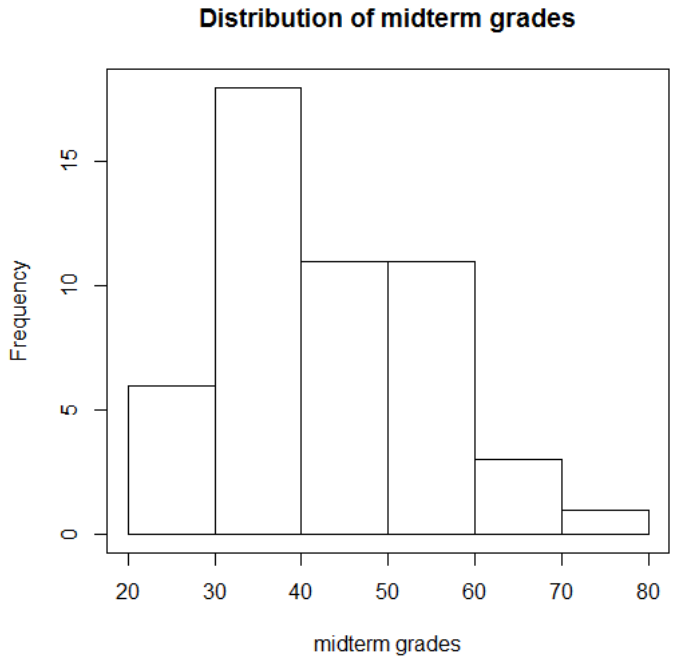
Bar charts

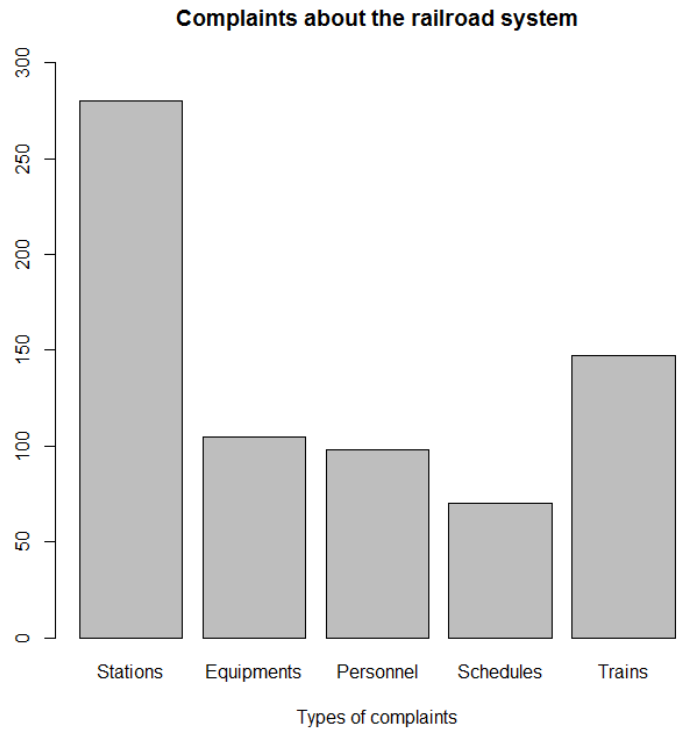
Complaint	Number
Stations	280
Equipment	105
Personnel	98
Schedules	70
Train	147



Bar charts v.s. histograms

- What are differences that distinguish a bar chart from a histogram?





- A bar chart uses **noncontiguous** bars to visualize **categorical** data.
- A histogram uses **contiguous** bars to visualize **quantitative** data.

Visualizing two variables

- When we have data for two variables, typically we want to identify whether there is any **relationship** between them.
- Visualizing the data in a two-dimensional manner helps.

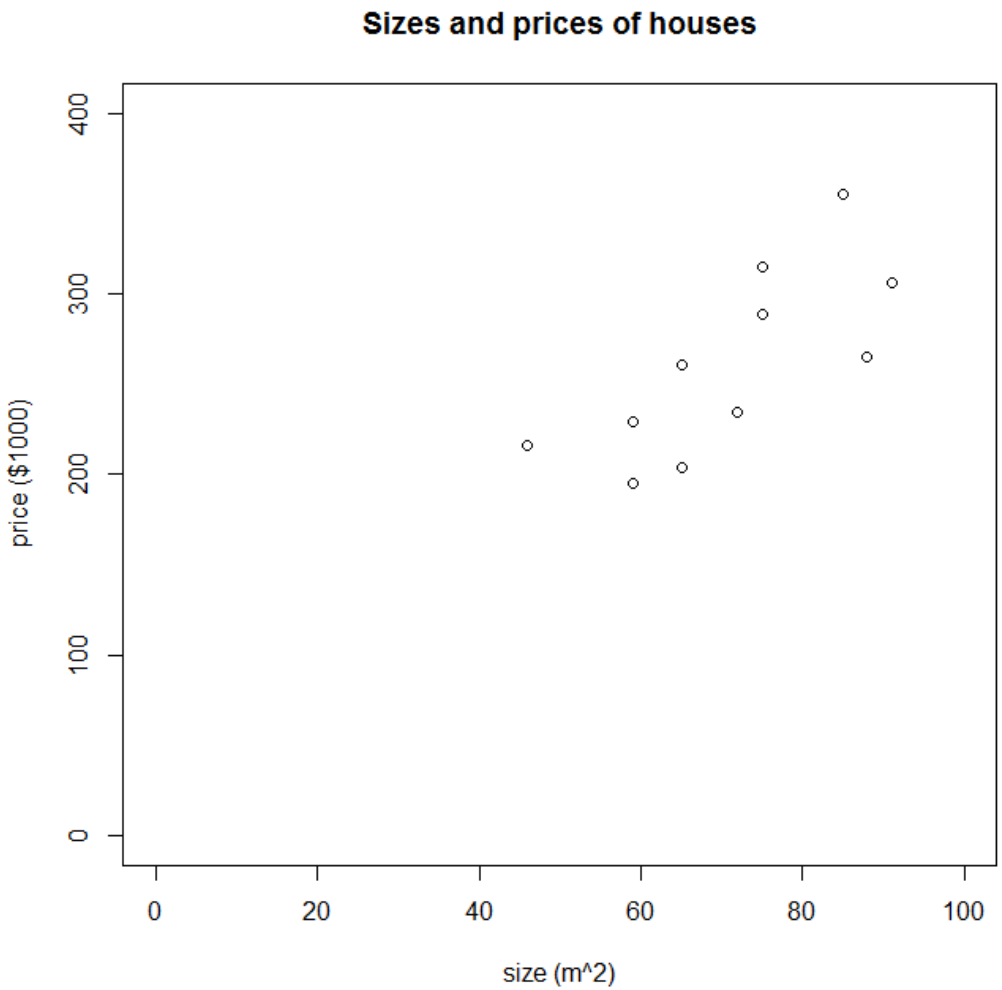
Scatter plots

- Sometimes in an observation there are **two** values recorded.
- When the two vales are both measured in quantitative scales, we may depict each observation as a point on a plane to create a **scatter plot**.
- Consider the size of a house and its price in a city:

House	1	2	3	4	5	6
Size (m^2)	75	59	85	65	72	46
Price ($1000)	315	229	355	261	234	216
House	7	8	9	10	11	12
Size (m^2)	107	91	75	65	88	59
Price ($1000)	308	306	289	204	265	195

Scatter plots

- Is there any relationship?
- In general, relationships may also be **nonlinear**.
- **Regression** will be introduced in this semester for discovering patterns from multi-dimensional data.



Road map

- Visualizing data with graphs.
- **Describing central tendency**.
- Describing variability.
- Describing correlation.

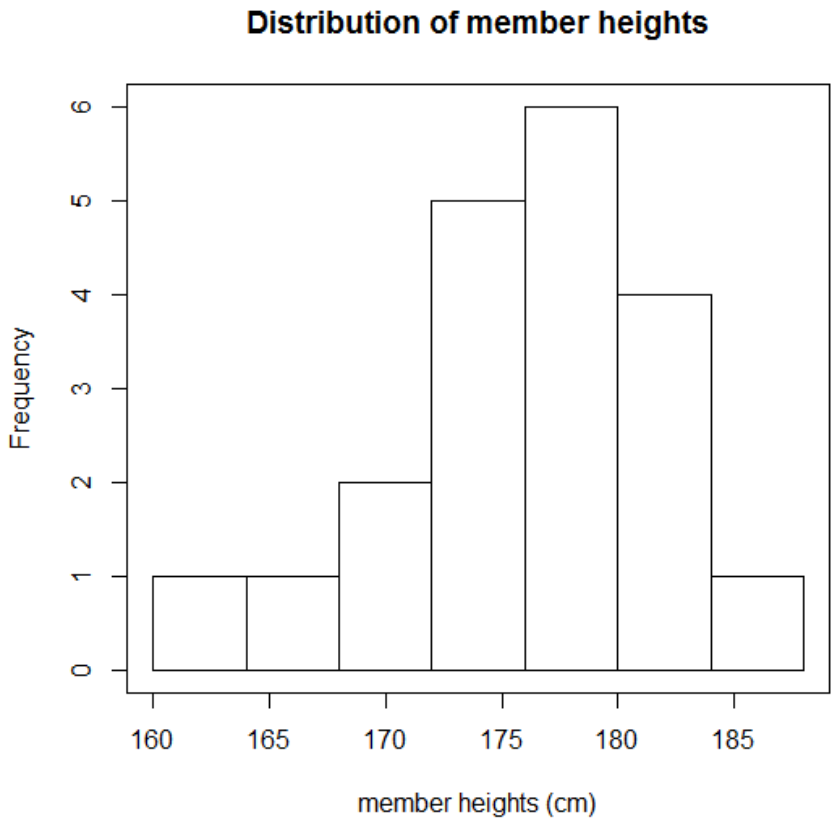
Using numbers to describe data

- Besides using graphs to visualize data, we may also use **numbers** to describe/summarize data.
  - These are either **parameters** for populations or **statistics** for samples.
- Measures of **central tendency** provides information about the center or middle part of a group of numbers.
- Measures of **variability** provides information about how variable the data are.
- Measures of **correlation** provides information about the relationship between two variables.

Central tendency

- In a baseball team, players' heights (in cm) are:

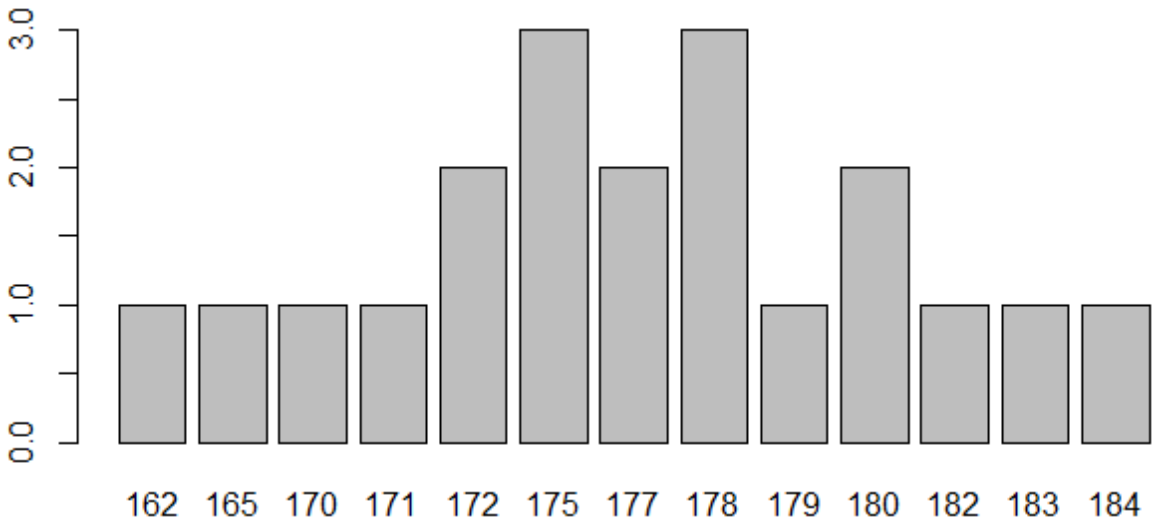
178	172	175	184
172	175	165	178
177	175	180	182
177	183	180	178
179	162	170	171



- Let's try to describe the central tendency of this set of data.

Modes

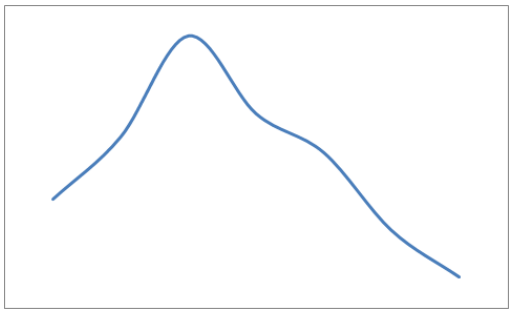
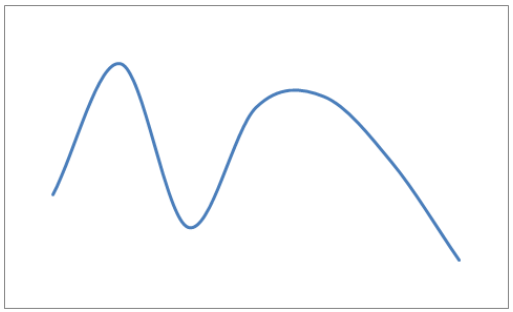
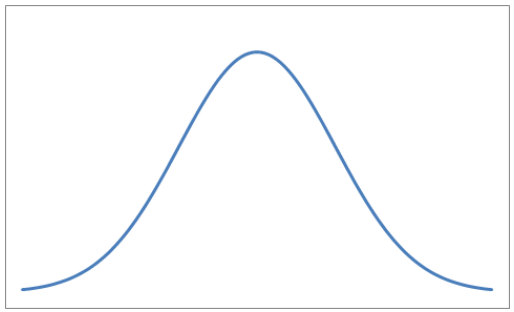
- The **mode**(s) is (are) the **most frequently** occurring value(s) in a set of data.
  - In the team, the modes are 175 and 178.



- It is better to look for a mode in a set of **qualitative** data.
  - Otherwise, maybe all values are modes!

The number of modes

- The data of the IM team is **bimodal**.
- In general, data may be **unimodal**, bimodal, or **multimodal**.
  - When the mode is unique, the data is unimodal.
  - When there are two modes or two values of similar frequencies that are more dominant than others, the data is bimodal.

- An important type of unimodal data is the **bell shape**.

Medians

- The **median** is the **middle** value in an ordered set of numbers.
  - Roughly speaking, **half** of the numbers are below and **half** are above it.
- Suppose there are N numbers:
  - If N is odd, the median is the $\frac{N+1}{2}$th large number.
  - If N is even, the median is the **average** of the $\frac{N}{2}$th and the $(\frac{N}{2}+1)$th large number.
- For example:
  - The median of $\{1, 2, 4, 5, 6, 8, 9\}$ is 5.
  - The median of $\{1, 2, 4, 5, 6, 8\}$ is $\frac{4+5}{2} = 4.5$.

Medians

- A median is unaffected by the magnitude of extreme values:
  - The median of $\{1, 2, 4, 5, 6, 8, 9\}$ is 5.
  - The median of $\{1, 2, 4, 5, 6, 8, 900\}$ is still 5.
- Medians may be calculated from **quantitative** or **ordinal** data.
- Unfortunately, a median uses only **part** of the information contained in these numbers.
  - For quantitative data, a median only treats them as ordinal.

Means

- The **mean** is the **average** of a set of data.
  - Can be calculated only from quantitative data.
  - The mean of $\{1, 2, 4, 5, 6, 8, 9\}$ is

$$\frac{1+2+4+5+6+8+9}{7} = 5.$$

- A mean uses **all** the information contained in the numbers.
- Unfortunately, a mean will be affected by extreme values.
  - The mean of $\{1, 2, 4, 5, 6, 8, 900\}$ is $\frac{1+2+4+5+6+8+900}{7} \approx 132.28$!
  - Using the mean and median **simultaneously** can be a good idea.
  - We should try to identify **outliers** (extreme values that seem to be "strange") before calculating a mean (or any statistics).

Population means v.s. sample means

- Let $\{x_i\}_{i=1,\ldots,N}$ be a population with N as the **population size**. The **population mean** is
$$\mu \equiv \frac{\sum_{i=1}^{N} x_i}{N}.$$
- Let $\{x_i\}_{i=1,\ldots,n}$ be a sample with $n < N$ as the **sample size**. The **sample mean** is
$$\bar{x} \equiv \frac{\sum_{i=1}^{n} x_i}{n}.$$
- People use μ and $\bar{x}$ in almost the whole statistics world.

Population means v.s. sample means

$$\mu \equiv \frac{\sum_{i=1}^{N} x_i}{N} \qquad \bar{x} \equiv \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Isn't these two means the same?
  - From the perspective of calculation, yes.
  - From the perspective of statistical inference, **no**.
- Typically the population mean is **fixed but unknown**.
  - The sample mean is **random**: We may get different values of $\bar{x}$ today and tomorrow.
  - To start from $\bar{x}$ and use **inferential statistics** to estimate or test μ, we need to apply **probability**.

Quartiles and percentiles

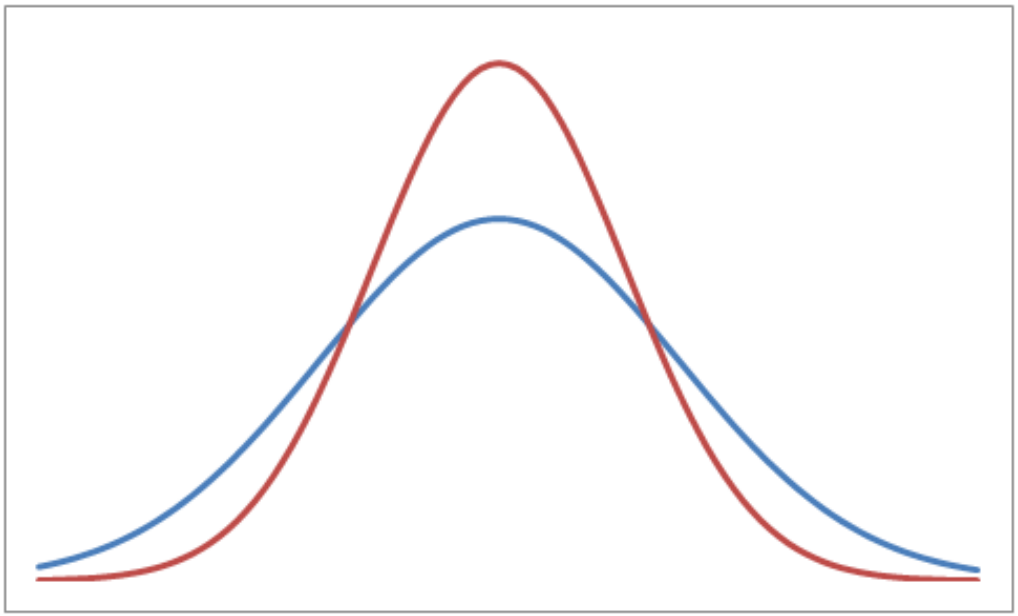
- The median lies at the middle of the data.
- The **first quartile** lies at the middle of the **first half** of the data.
- The **third quartile** lies at the middle of the **second half** of the data.
- For the pth **percentile**:
  - $\frac{p}{100}$ of the values are below it.
  - $1 - \frac{p}{100}$ of the values are above it.
- Median, quartiles, and percentiles:
  - The 25th percentile is the first quartile.
  - The 50th percentile is the median (and the second quartile).
  - The 75th percentile is the third quartile.

Road map

- Visualizing data with graphs.
- Describing central tendency.
- **Describing variability**.
- Describing correlation.

Variability

- **Measures of variability** describe the **spread** or **dispersion** of a set of data.
- Especially useful when two sets of data have the same center.



Ranges and Interquartile ranges

- The **range** of a set of data $\{x_i\}_{i=1,\ldots,N}$ is the difference between the maximum and minimum numbers, i.e.,

$$\max_{i=1,\ldots,N}\{x_i\} - \min_{i=1,\ldots,N}\{x_i\}.$$

- The **interquartile range** of a set of data is the difference of the first and third quartile.
  - It is the range of the middle 50 of data.
  - It excludes the effects of extreme values.

Deviations from the mean

- Consider a set of population data $\{x_i\}_{i=1,\ldots,N}$ with mean μ.
- Intuitively, a way to measure the dispersion is to examine how each number **deviates from the mean**.
- For x_i, the deviation from the population mean is defined as

$$x_i - \mu.$$

- For a **sample**, the deviation from the sample mean of x_i is

$$x_i - \bar{x}.$$

i	x_i	deviation
1	1	$1-5=-4$
2	2	$2-5=-3$
3	4	$4-5=-1$
4	5	$1-5=0$
5	6	$6-5=1$
6	8	$8-5=3$
7	9	$9-5=4$
Mean	15	

Mean deviations

- May we summarize the N deviations into a single number to summarize the aggregate deviation?
- Intuitively, we may sum them up and then calculate the **mean deviation**:

$$\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}.$$

- Is it always 0?

i	x_i	deviation
1	1	$1-5=-4$
2	2	$2-5=-3$
3	4	$4-5=-1$
4	5	$1-5=0$
5	6	$6-5=1$
6	8	$8-5=3$
7	9	$9-5=4$
Mean	15	0

Adjusting mean deviations

- People use two ways to adjust it:
  - Mean **absolute** deviations (MAD):

$$\frac{\sum_{i=1}^{N} |x_i - \mu|}{N}.$$

  - Mean **squared** deviations (variance):

$$\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}.$$

i	x_i	deviation d_i	$\lvert d_i \rvert$	d_i^2
1	1	$1-5=-4$	4	16
2	2	$2-5=-3$	3	9
3	4	$4-5=-1$	1	1
4	5	$1-5=0$	0	0
5	6	$6-5=1$	1	1
6	8	$8-5=3$	3	9
7	9	$9-5=4$	4	16
Mean	5	0	2.29	7.43

Measuring variability

- **Larger** MADs and variances means the data are **more disperse**:

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	1	-4	4	16
2	2	-3	3	9
3	4	-1	1	1
4	5	0	0	0
5	6	1	1	1
6	8	3	3	9
7	9	4	4	16
Mean	5	0	2.29	7.43

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	0	-5	5	25
2	2	-3	3	9
3	4	-1	1	1
4	5	0	0	0
5	6	1	1	1
6	8	3	3	9
7	10	5	5	25
Mean	5	0	2.57	10

MADs vs. variances

- The main difference:
  - An MAD puts the same weight on all values.
  - A variance puts more weights on **extreme values**.

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	0	−5	5	25
2	4	−1	1	1
3	5	0	0	0
4	6	1	1	1
5	10	5	5	25
Mean	5	0	2.4	10.4

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	1	4	4	16
2	2	3	3	9
3	5	0	0	0
4	8	3	3	9
5	9	4	4	16
Mean	15	0	2.8	10

- In general, people use variances more than MADs.
  - But MADs are still popular in some areas, e.g., demand forecasting.
  - It is the analyst's discretion to choose the appropriate one.

Standard deviations

- One drawback of using variances is that the unit of measurement is the **square** of the original one.
- For the baseball team, the variance of member heights is 34.05 **cm^2**. What is it?!
- People take the square root of a variance to generate a **standard deviation**.
- The standard deviation of member heights is

$$\sqrt{34.05} \approx 5.85 \text{ cm}.$$

178	172	175	184
172	175	165	178
177	175	180	182
177	183	180	178
179	162	170	171

- A standard deviation is typically of more managerial implications.

Population v.s. sample variances

- Recall that the formulas for population and sample means are

$$\mu \equiv \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{and} \quad \bar{x} \equiv \frac{\sum_{i=1}^{n} x_i}{n}, \text{ respectively.}$$

  - Formula-wise there is no difference.
- However, **population** and **sample variances** are

$$\sigma^2 \equiv \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \quad \text{and} \quad s^2 \equiv \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \text{ respectively.}$$

  - Note the difference between N and $n-1$!
  - Population and sample standard deviations are $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$ and $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$, respectively.
- People use σ^2, σ, s^2, and s in almost the whole statistics world.

Coefficient of variation

- The **coefficient of variation** is the **ratio** of the standard deviation to the mean:

$$\text{Coefficient of variation} = \frac{\sigma}{\mu}.$$

- When will you use coefficients of variation?

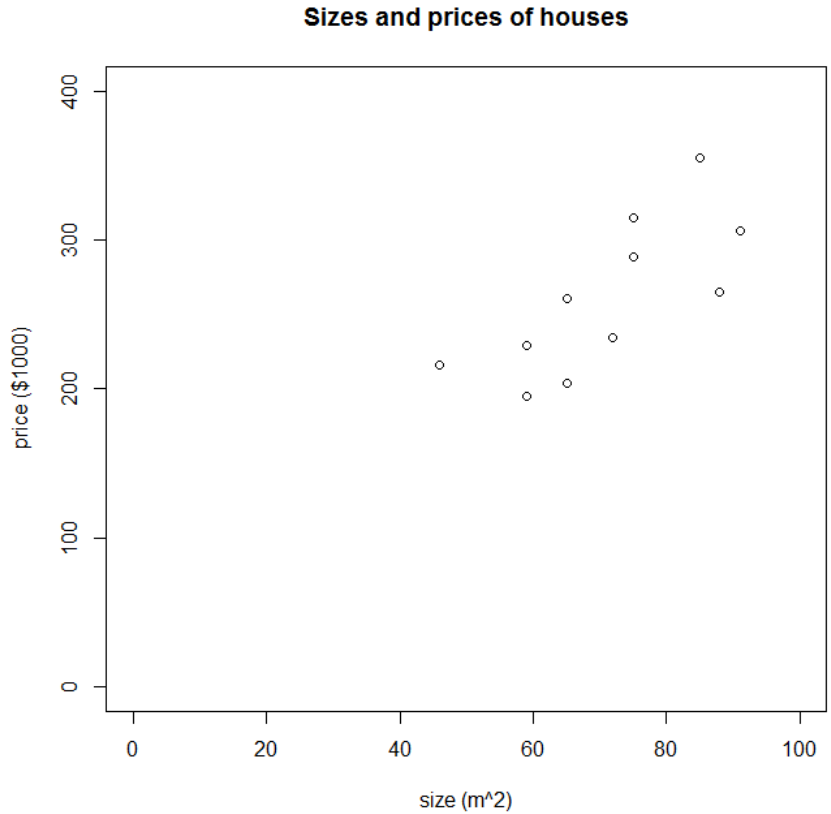
Road map

- Visualizing data with graphs.
- Describing central tendency.
- Describing variability.
- **Describing correlation**.

Introduction

- Consider the size of a house and its price in a city:

Size (in m^2)	Price (in \$1000)
75	315
59	229
85	355
65	261
72	234
46	216
107	308
91	306
75	289
65	204
88	265
59	195



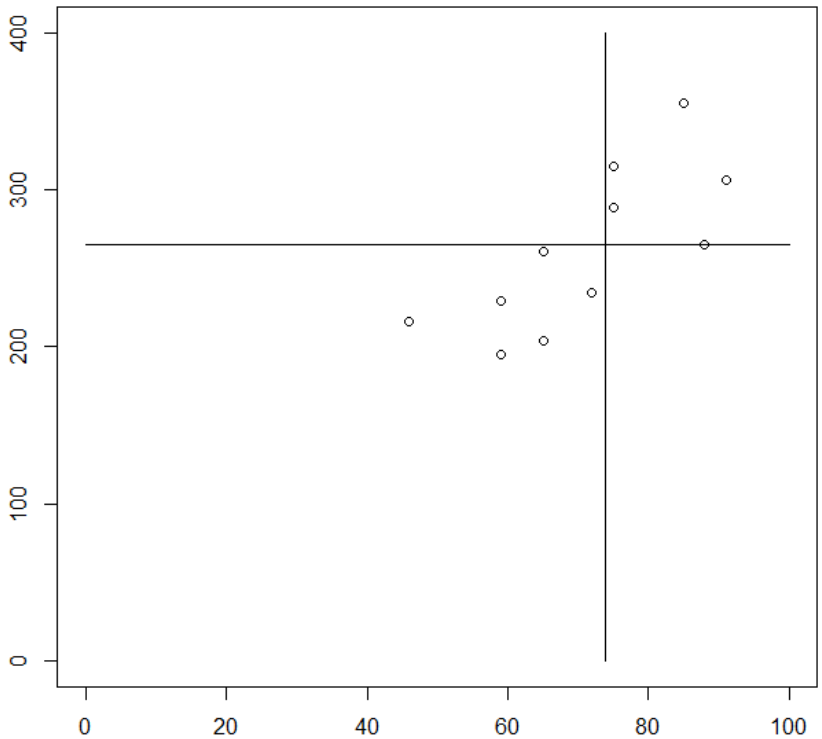
- How do we measure/describe the **correlation** (linear relationship) between the two variables?

Intuition

- When one variable goes up, does the other one **tend to** go up or down?
- More precisely, if x_i is larger than μ_x (the mean of the x_is), is it more likely to see $y_i > \mu_y$ or $y_i < \mu_y$?
- Let's highlight the two means on the scatter plot.

Intuition

- The scatter plot with the two means:



- We say that the two variables have a **positive** correlation.
  - If one goes up when the other goes down, there is a **negative** correlation.

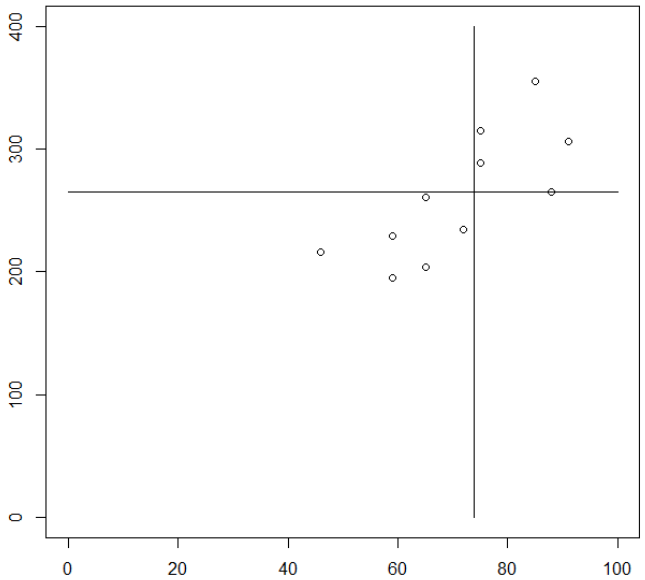
Covariances

- We define the **covariance** of a set of two-dimensional **population** data as

$$\sigma_{xy} \equiv \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}.$$

  - If most points fall in the first and third quadrants, most $(x_i - \mu_x)(y - \mu_y)$ will be positive and σ_{xy} tends to be positive.
  - Otherwise, σ_{xy} tends to be negative.
- The **sample covariance** is

$$s_{xy} \equiv \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$



Example: house sizes and prices

- For our example:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
75	315	1.08	50.25	54.44
59	229	−14.92	−35.75	533.27
85	355	11.08	90.25	1000.27
65	261	−8.92	−3.75	33.44
72	234	−1.92	−30.75	58.94
46	216	−27.92	−48.75	1360.94
107	308	33.08	43.25	1430.85
91	306	17.08	41.25	704.69
75	289	1.08	24.25	26.27
65	204	−8.92	−60.75	541.69
88	265	14.08	0.25	3.52
59	195	−14.92	−69.75	1040.44
$\bar{x} = 73.92$	$\bar{y} = 264.75$	–	–	$s_{xy} = 617.16$

- So the covariance of house size and price is 617.16.
- Is it large or small?

Correlation coefficients

- To take away the auto-variability of each variable itself, we define the population and sample **correlation coefficients** as

$$\rho \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{and} \quad r \equiv \frac{s_{xy}}{s_x s_y},$$

  - σ_x and σ_y are the population standard deviations of x_is and y_is.
  - s_x and s_y are the sample standard deviations of x_is and y_is.
  - In our example, we have $r = \frac{617.16}{16.78 \times 50.45} \approx 0.729$.
- It can be shown that we always have

$$-1 \leq \rho \leq 1 \quad \text{and} \quad -1 \leq r \leq 1.$$

  - $\rho > 0$ ($s > 0$): Positive correlation.
  - $\rho = 0$ ($s = 0$): No correlation.
  - $\rho < 0$ ($s < 0$): Negative correlation.

Magnitude of correlation

- In practice, people often determine the degree of correlation based on $|\rho|$ or $|s|$:
  - $0 \leq |\rho| < 0.25$ or $0 \leq |s| < 0.25$: A weak correlation.
  - $0.25 \leq |\rho| < 0.5$ or $0.25 \leq |s| < 0.5$: A moderately weak correlation.
  - $0.5 \leq |\rho| < 0.75$ or $0.5 \leq |s| < 0.75$: A moderately strong correlation.
  - $0.75 \leq |\rho| \leq 1$ or $0.75 \leq |s| \leq 1$: A strong correlation.
- The last remark:
  - Correlation coefficients only measure **linear** relationships.