Statistics I – Chapter 2
Visualizing the Data

Ling-Chieh Kung

Department of Information Management
National Taiwan University

September 12, 2012

Visualizing the data

- In this chapter, we introduce some commonly adopted techniques for visualizing data.
- Raw data, or data that have not been summarized in any way, are called **ungrouped data**.
- We will learn how to generate and present **grouped data**, either in tables or in figures.

Road map

- **Frequency distributions**.
- Quantitative data graphs.
- Qualitative data graphs.
- Visualizing two variables.

Frequency distributions

- A **frequency distribution** is a summary of data presented in the form of class intervals and frequencies.
- Three steps to construct a frequency distribution from ungrouped data:
  - Determine the **range**, the difference between the largest and the smallest numbers.
  - Determine the **number of classes**.
    - A rule of thumb: **5 to 15 classes**.
  - Determine the **width** of each class; then count!
    - Typically all classes have the same width.
    - Be aware of class endpoints! Classes should NOT overlap with each other.

Frequency distributions: an example

- A sample: ages of managers from urban child care centers in the IM city.
- Ungrouped data:

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

- Let's summarize this sample by a frequency distribution.

Frequency distributions: an example

- Step 1: Range $= 74 - 23 = 51$.
- Step 2: As we only have 50 numbers, it is not very good to have many classes. Let's try 6.
- Step 3: Class width $\geq \lceil \frac{51}{6} \rceil = 9$. But widths like 5 or 10 are always preferred. So let's try 10.
  - Why ceiling? Why not floor?

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	$[70, 80)$	$70 \leq x < 80$

  - Why not $[21, 31)$, $[31, 41)$, ...?
  - Why not $(20, 30]$, $(30, 40]$, ...?
  - How about $[20, 29]$, $[30, 39]$, ...?

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 grouped data. It is a **description (summary)** of the sample.

Some remarks

- You may also call them frequency tables.
- It general, deciding the number of classes, the class width, and the starting point is an **art**. It requires experiences and domain knowledge to make a good choice.
- There is NO best choice. There is NO standard answer.

Something more on frequency tables

- We may add class midpoints, relative frequencies, and cumulative frequencies into a frequency table.
  - A **class midpoint** (or a class mark) is the midpoint of the class interval.
  - A **relative frequency** is the proportion of the total frequency in a given class.
  - A **cumulative frequency** is the sum of all frequencies up to a given class.

Something more

- The extended our 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**?

Road map

- Frequency distributions.
- **Quantitative data graphs**.
- Qualitative data graphs.
- Visualizing two variables.

Quantitative data graphs

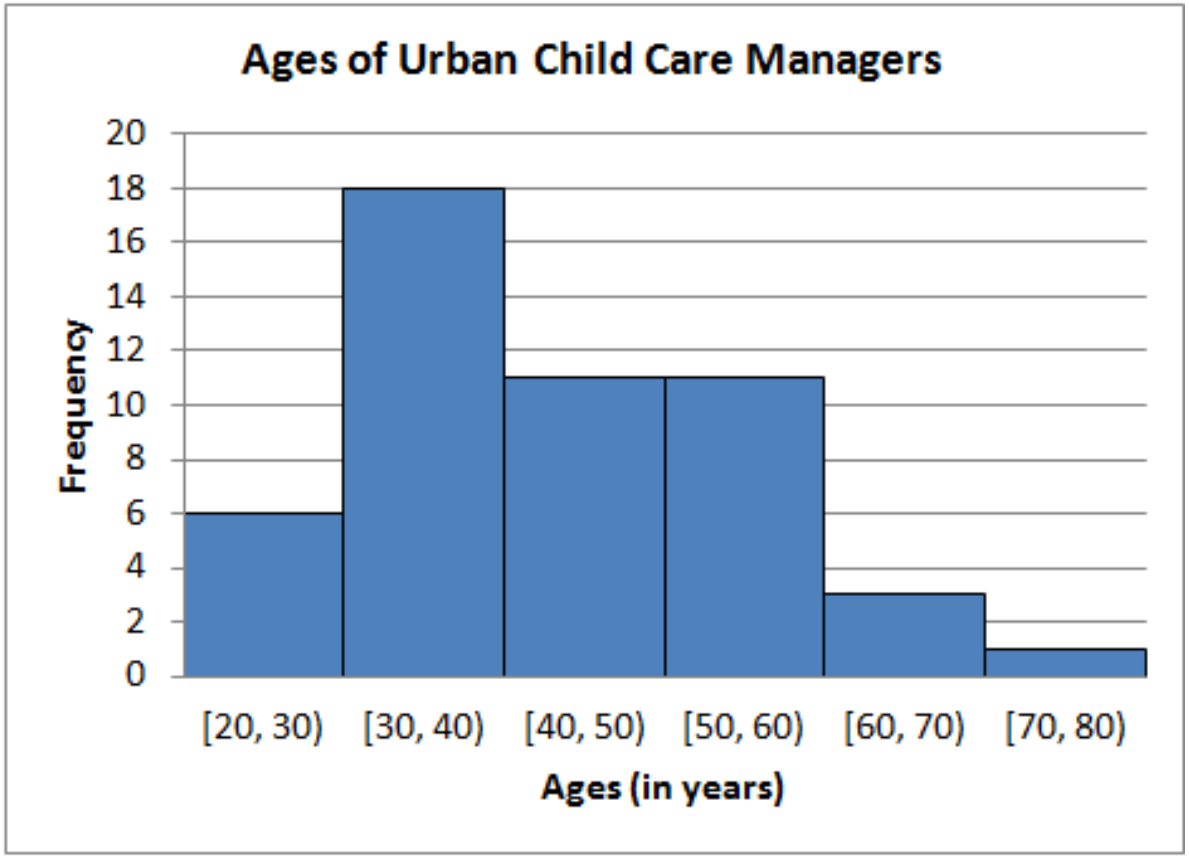
- "A picture is worth a thousand words."
  - Graphs are intuitive to interpret.
  - Graphs are helpful for determining the shape of a distribution.
- Typically we draw graphs to get some rough ideas before conducting rigorous statistical studies.
- Moreover, (probably) your boss can read nothing but graphs... orz

Histograms

- A **histogram** is a graphical representation of a frequency distribution.
- It consists of a series of **contiguous** rectangles, each representing the frequency in a class.

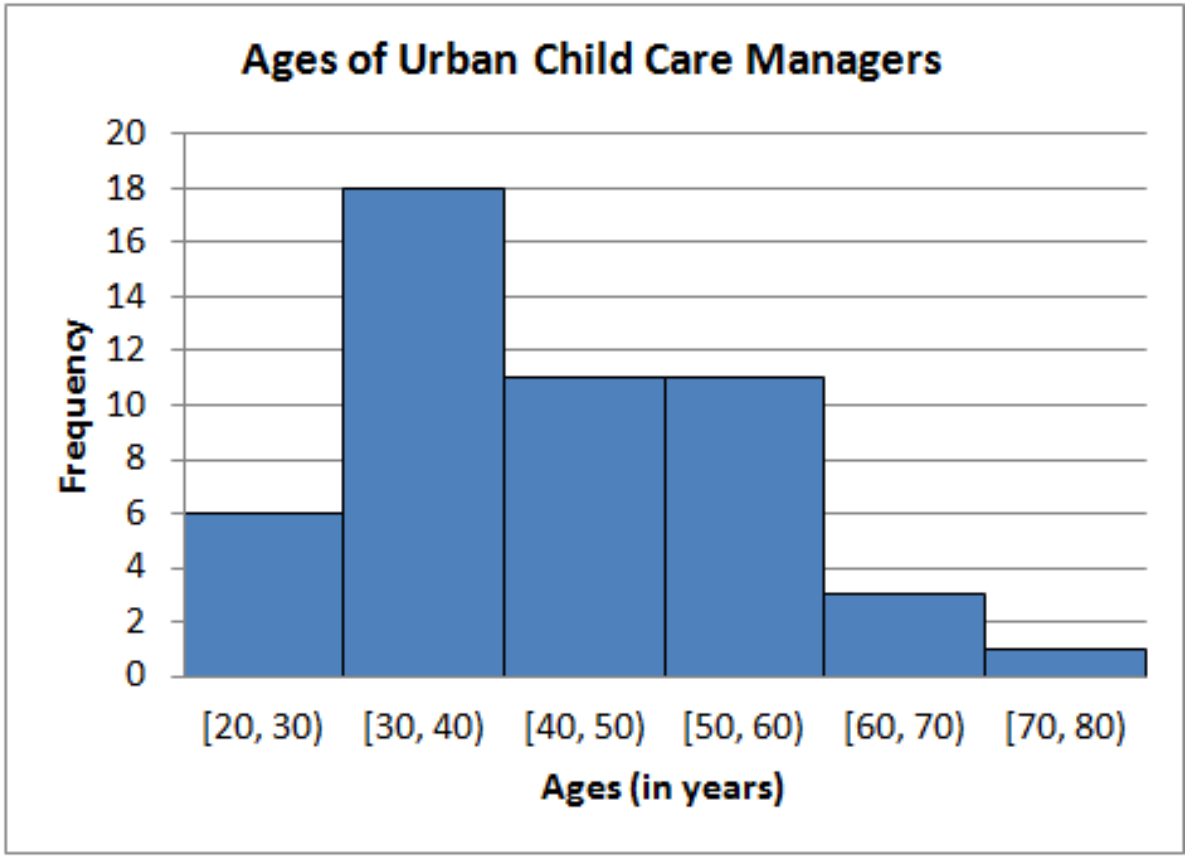
Histograms

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



Histograms

- Never forget:
  - Caption.
  - Captions and labels for the x- and y-axes.
  - Unit of measurement.
  - Contiguous rectangles.



Histograms

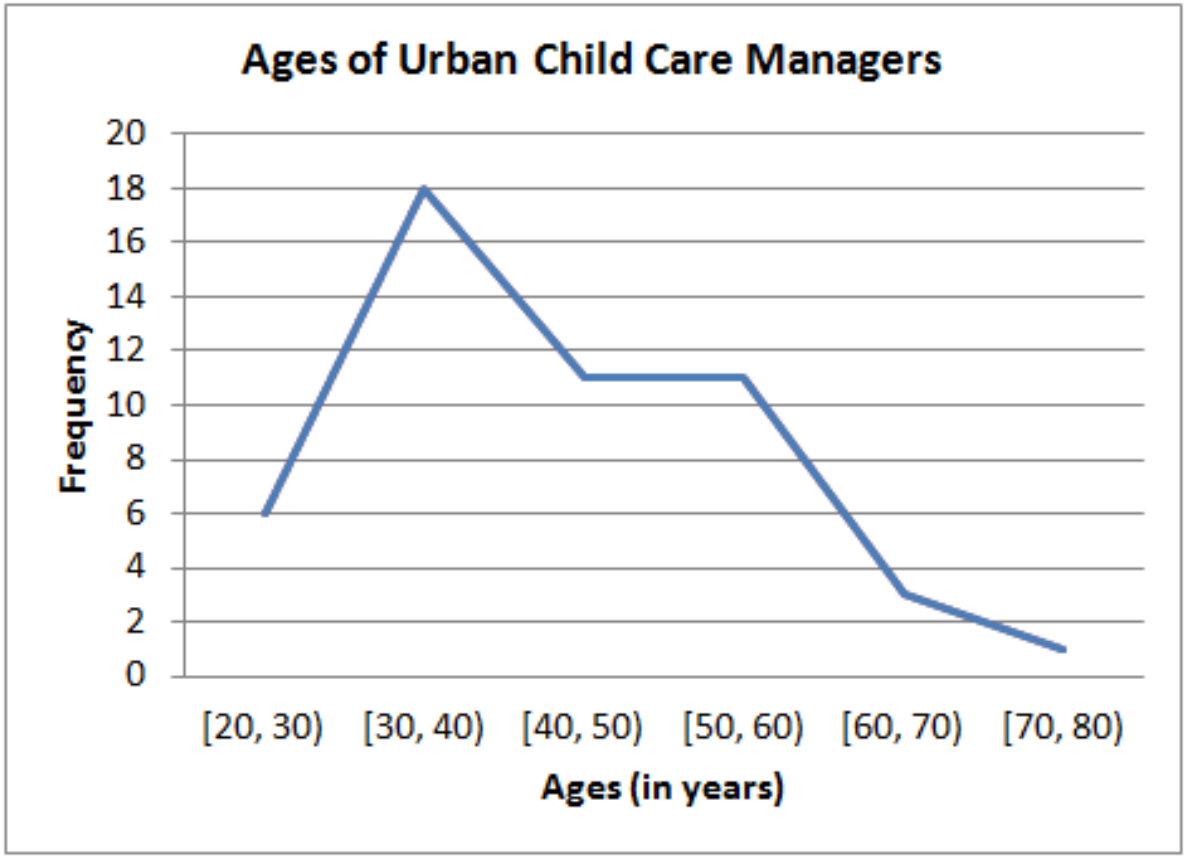
- Histograms are one of the most important types of quantitative graph.
- One particular reason to draw histograms is to get some ideas about the **distribution**.
  - Bell shape? M shape? Skewed?
  - Any outlier?
  - Uniformly distributed? Normally distributed?

Frequency polygons

- A **frequency polygon** also graphically visualizes a frequency distribution.
- Instead of using rectangles, it uses **line segments** connecting dots plotting at class **midpoints**, where dots represents frequencies.
- The information contained in a frequency polygon is quite similar to that contained in a histogram.

Frequency polygons

- Never forget:
  - Plot dots at class midpoints.



Frequency polygons

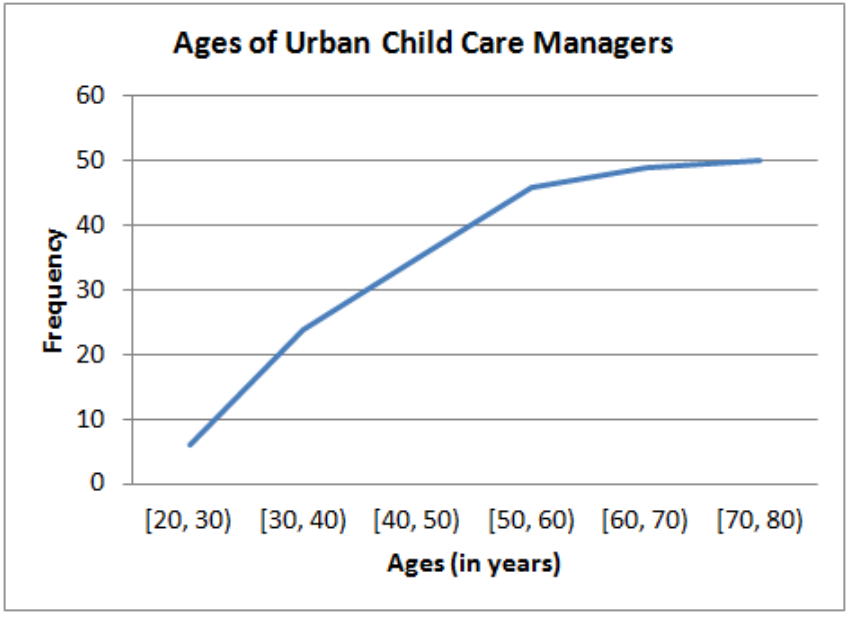
- It is more convenient to use a frequency polygon to **compare multiple** frequency distributions.
- However, people may **misunderstand** a frequency polygon by feeling that there are some connections between consecutive classes.

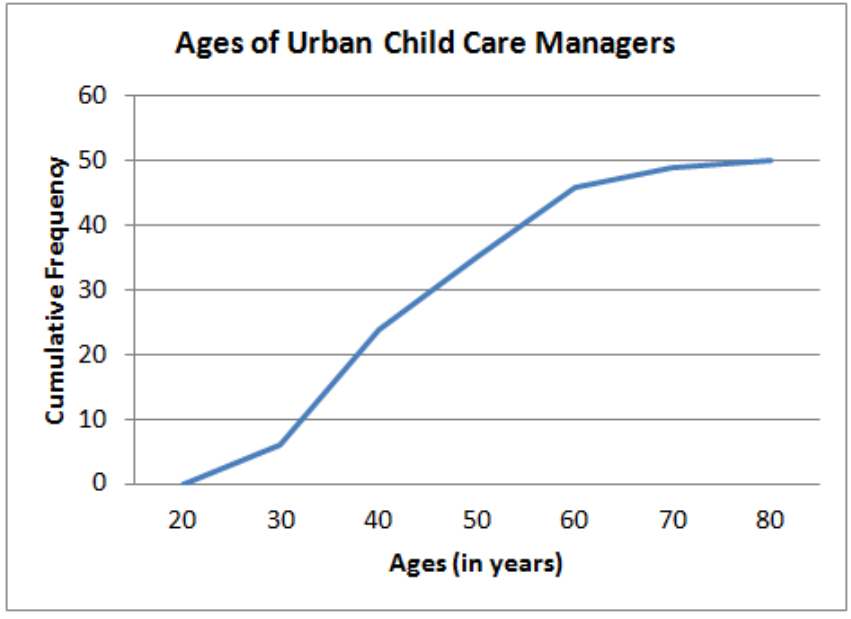
Ogives

- An ogive is a **cumulative** frequency polygon.
  - A dot of zero frequency is plotted at the **beginning** of the first class.
  - Dots of cumulative frequencies are plotted at the **end** of all classes.
- Useful for seeing **running totals**.
  - How many classes, from bottom to top, do we need to achieve 30 people?

Ogives

- Which one is a correct ogive?





Stem-and-leaf plots

- An **stem-and-leaf plot** separates the digits for each number into two groups, a **stem** and a **leaf**.
  - The leftmost digits form the stem.
  - The other digits form the leave.
- The stems will be treated as categories (like those classes in a histogram). The leaves are to distinguish numbers.
- In our example, the tens are stems and the units are leaves.
  - E.g., 42: Stem is 4 and leaf is 2.
  - E.g., 26: Stem is 2 and leaf is 6.

Stem-and-leaf plots

- In a column at left, one ranks stems in an ascending order **from top to bottom**. A stem may have no leaf if there is no corresponding number.
- For each stem, one ranks leaves in an ascending order **from left to right**. Repeated leaves are all listed.
- The stem-and-leaf plot for our example:

2	3	5	6	6	8	9												
3	0	0	0	0	0	0	1	1	2	2	2	2	3	4	5	6	7	7
4	0	0	0	0	2	3	3	6	7	9	9							
5	0	0	0	2	2	3	4	5	7	8	8							
6	0	1	4															
7	4																	

Stem-and-leaf plots

- The main advantage of a stem-and-leaf plot is that it does NOT **conceal any information**.
- The main disadvantage is the **table size**, especially when the data size is large.
- Good for small-size data but impossible for large-size data.
- In general, how to divide a number into a stem and a leaf is the plot drawer's discretion.
- Personally, I don't think stem-and-leaf plots are widely used ...

Road map

- Frequency distributions.
- Quantitative data graphs.
- **Qualitative data graphs**.
- Visualizing two variables.

Qualitative data graphs

- Qualitative data graphs are for qualitative data... XD
  - Which two data scales belong to qualitative data?
- Qualitative data graphs are also for grouped quantitative data.

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.

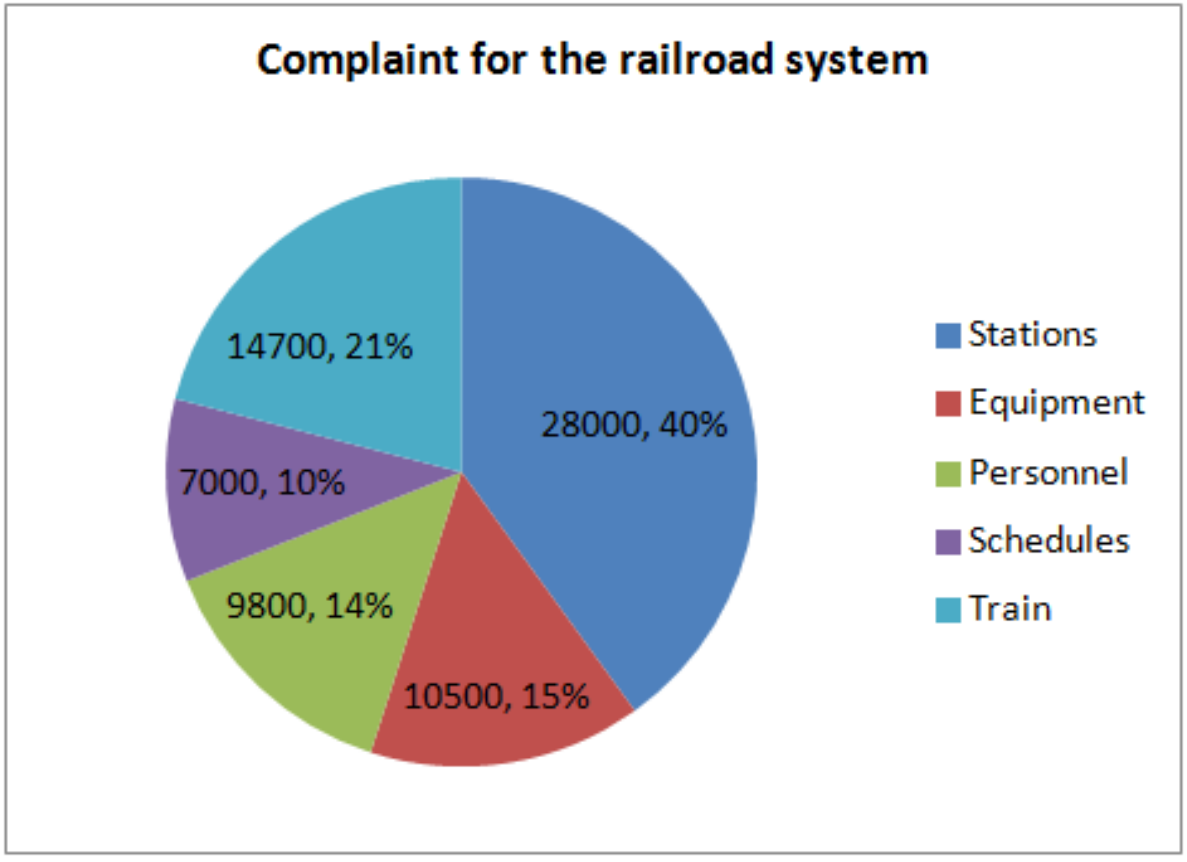
Pie charts

- Consider a survey in the IM city on what do passengers complain about the railroad system:

Complaint	Number	**Proportion**	**Degrees**
Stations	28000	0.40	144.0
Equipment	10500	0.15	54.0
Personnel	9800	0.14	50.4
Schedules	7000	0.10	36.0
Train	14700	0.21	75.6
Total	70000	1.00	360.0

Pie charts

Complaint	Number
Stations	28000
Equipment	10500
Personnel	9800
Schedules	7000
Train	14700



Pie charts

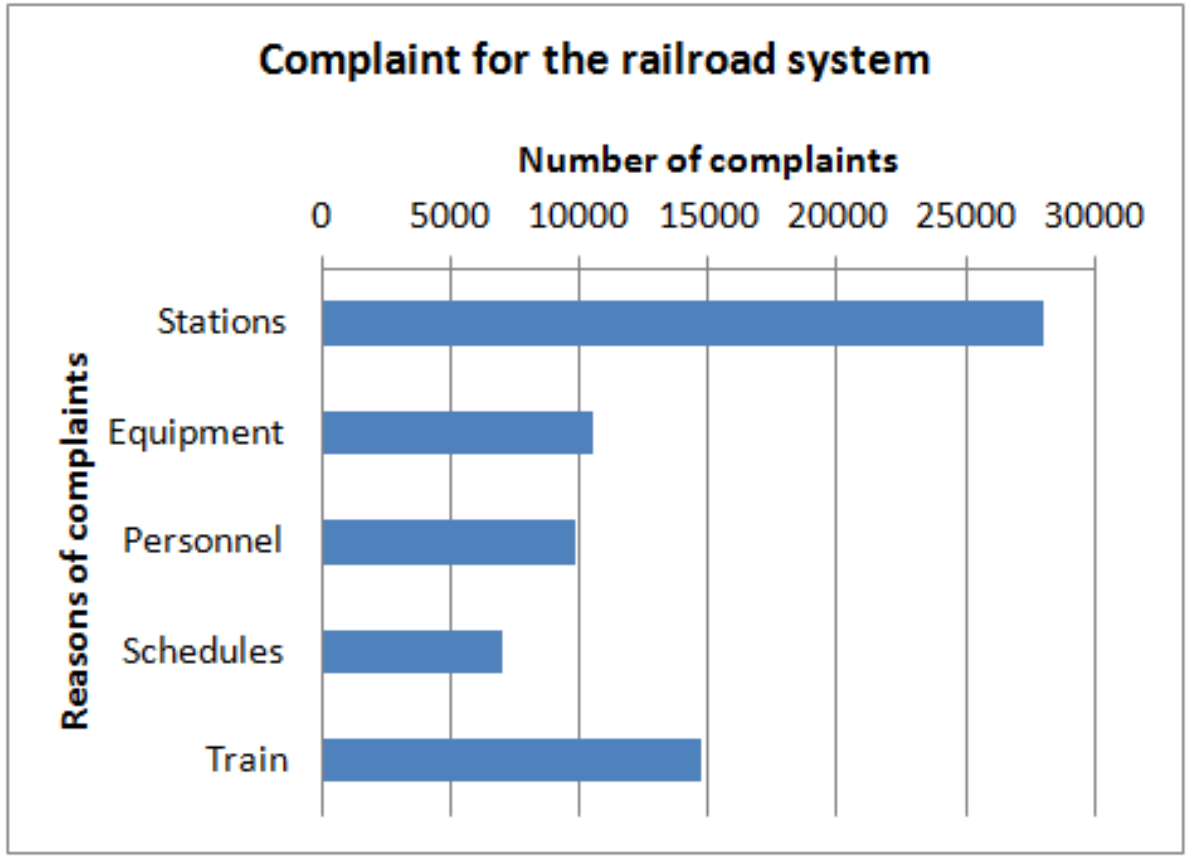
- No one says those slices must be sorted by their sizes. But you may do it if you want.
- Pie charts are useful in visualizing the **proportions** of each categories.
- However, determining the **relative size** of slides in a pie char may be hard.
- In demonstrating the differences among categories, a bar chart is a better choice.

Bar charts

- A **bar chart** (or bar graph) depicts each category by a bar. The larger the category, the longer the bar.
  - It does not matter to draw bars vertically or horizontally.
- No one says those bars must be sorted by their lengths. But you may do it if you want.

Bar charts

Complaint	Number
Stations	28000
Equipment	10500
Personnel	9800
Schedules	7000
Train	14700



Bar charts

- A bar chart is different from a histogram!!

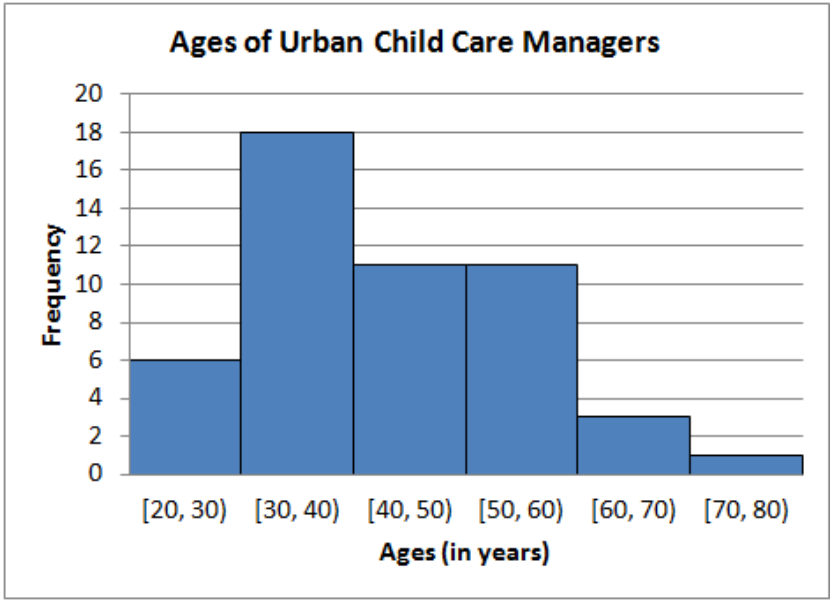
	Data type	Bars are ...
Histograms	Quantitative	Contiguous
Bar charts	Qualitative	Noncontiguous[1]

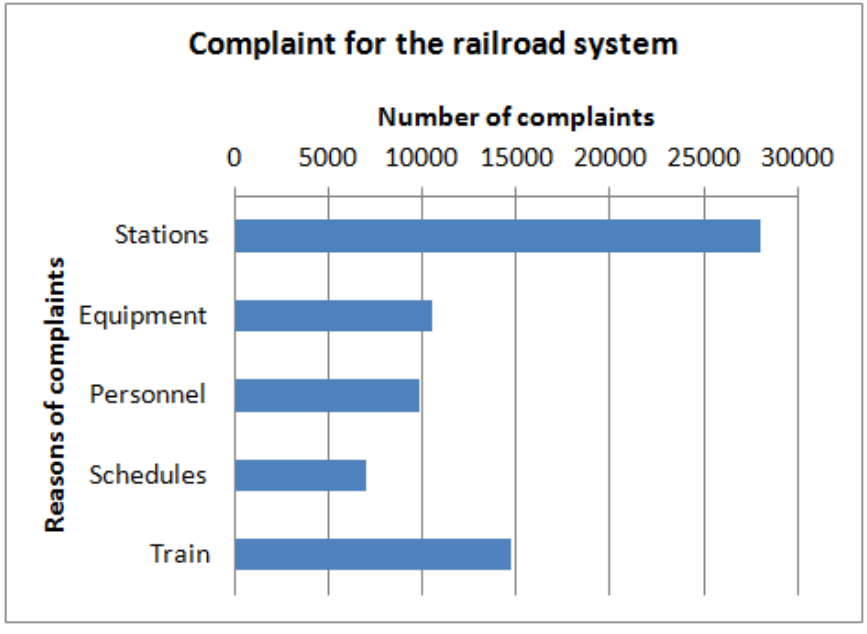
- A bar chart is better for comparing difference categories; a pie chart is better for presenting the proportion of a single category.

[1] While it is still allowed for bars in a bar chart to be contiguous, I suggest you to make them noncontiguous. For histograms, however, bars MUST be contiguous.

Bar charts v.s. histograms

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



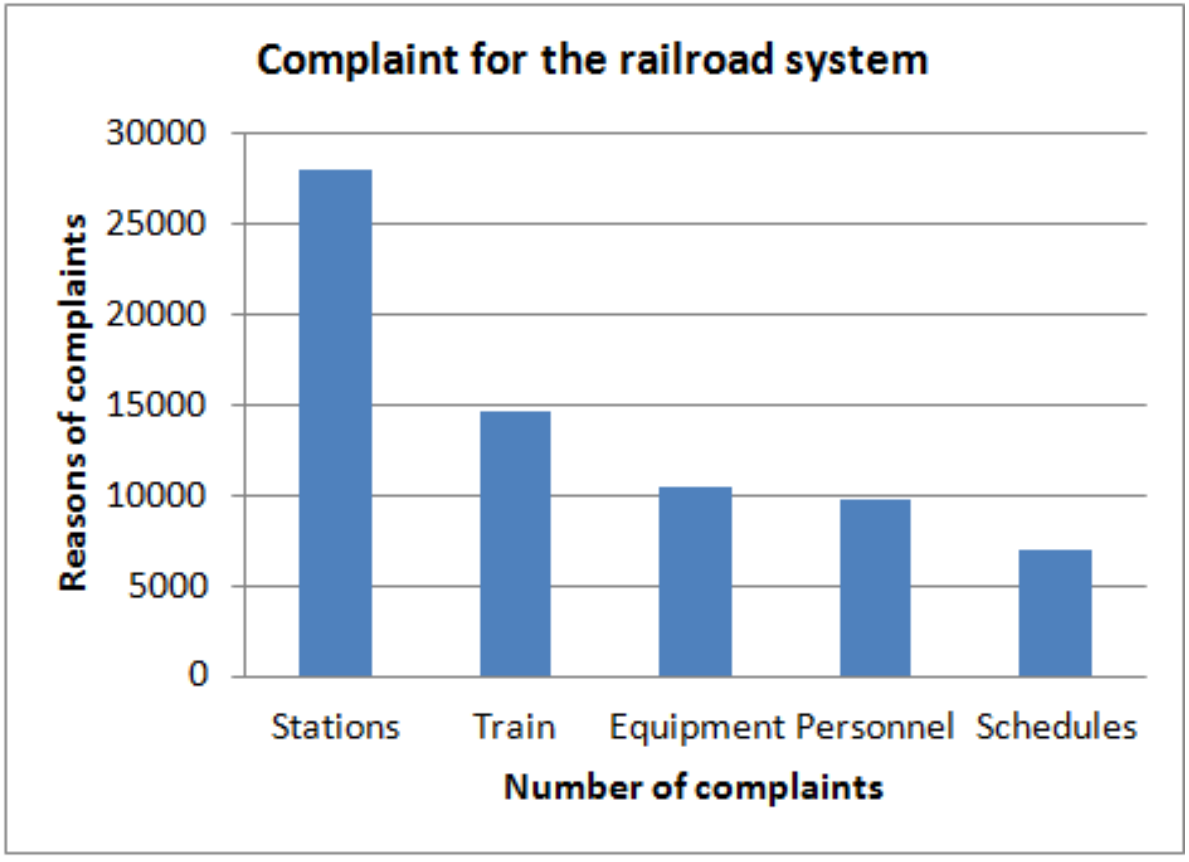


Pareto charts

- A **Pareto chart** is a bar chart in which bars are **sorted** according to their lengths.
  - Pareto is not Plato!! He is Vilfredo Pareto, an Italian economist.
- Typically, bars in a Pareto chart are vertically depicted. The longest bar are put at the leftmost position.

Pareto charts

Complaint	Number
Stations	28000
Equipment	10500
Personnel	9800
Schedules	7000
Train	14700



Pareto charts

- A Pareto chart is good for identifying those most critical categories.
- Some people add a cumulative frequency distribution on a Pareto chart.

Road map

- Frequency distributions.
- Quantitative data graphs.
- Qualitative data graphs.
- **Visualizing two variables**.

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.

Cross tabulation

- Cross tabulation produces a two-dimensional table that displays the frequency counts for two variables simultaneously.
- Consider how people in three occupations select one out of four brands of newspaper.
  - Labels occupations as 1, 2, and 3.
  - Labels newspaper as 1, 2, 3, and 4.
  - Data:

Person	1	2	3	4	5	...	354
Occupation	2	1	2	3	1	...	1
Newspaper	2	3	2	2	1	...	2

Cross tabulation

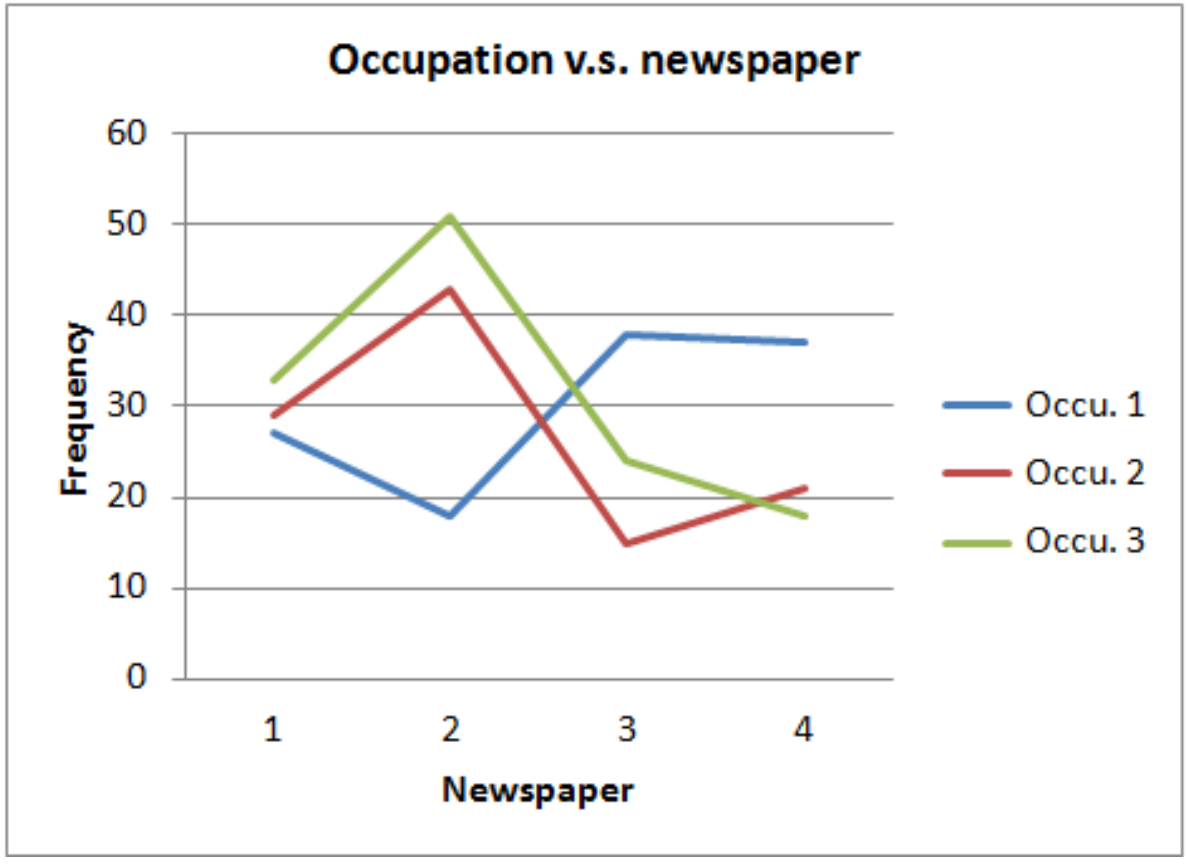
- The data can be organized into a **contingency table**:

Occupation	Newspaper				
	1	2	3	4	Total
1	27	18	38	37	120
2	29	43	15	21	108
3	33	51	24	18	126
Total	89	112	77	76	354

- Do people in different occupation prefer different newspaper?

Depicting a contingency tables

- What do you think?



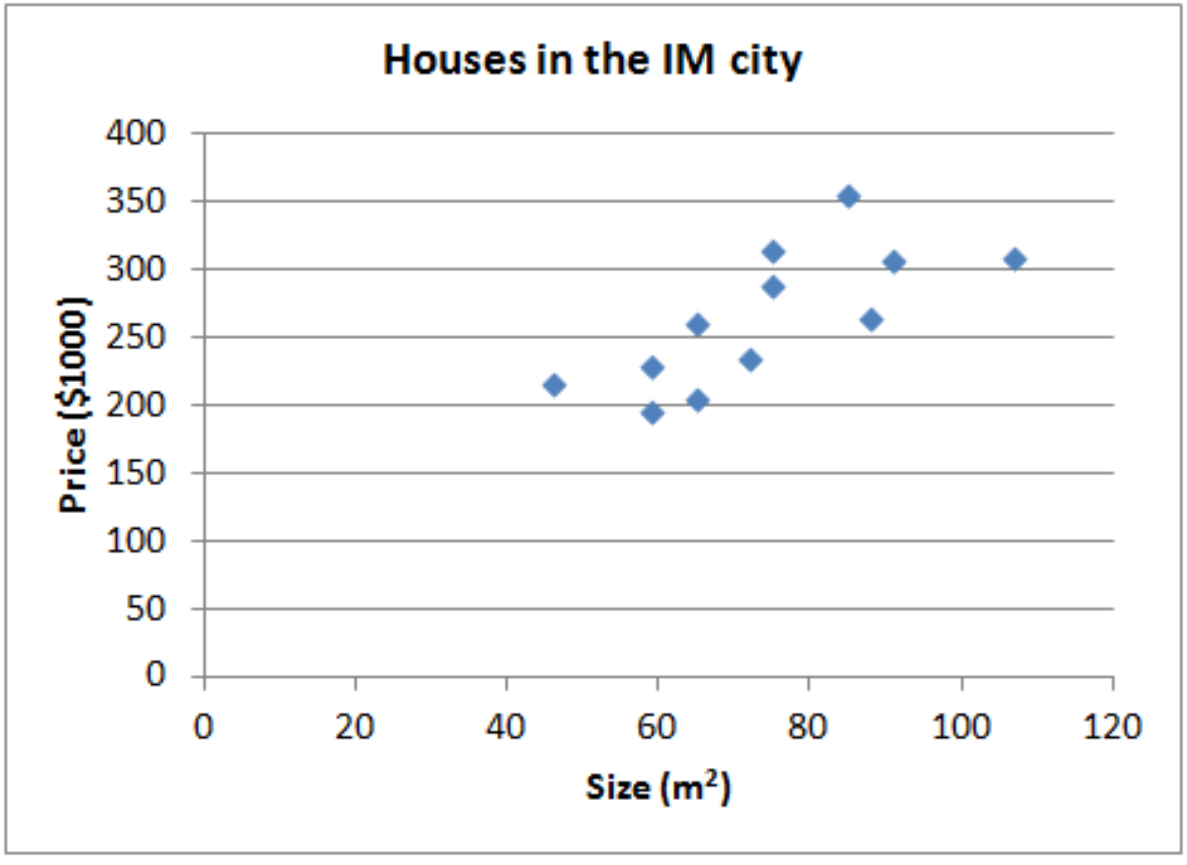
Scatter Plots

- When the two variables are both measured in quantitative scales, we may depict each point on a two-dimensional Cartesian coordinate system and create a **scatter plot**.
- Consider the size of a house and its price in the IM 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

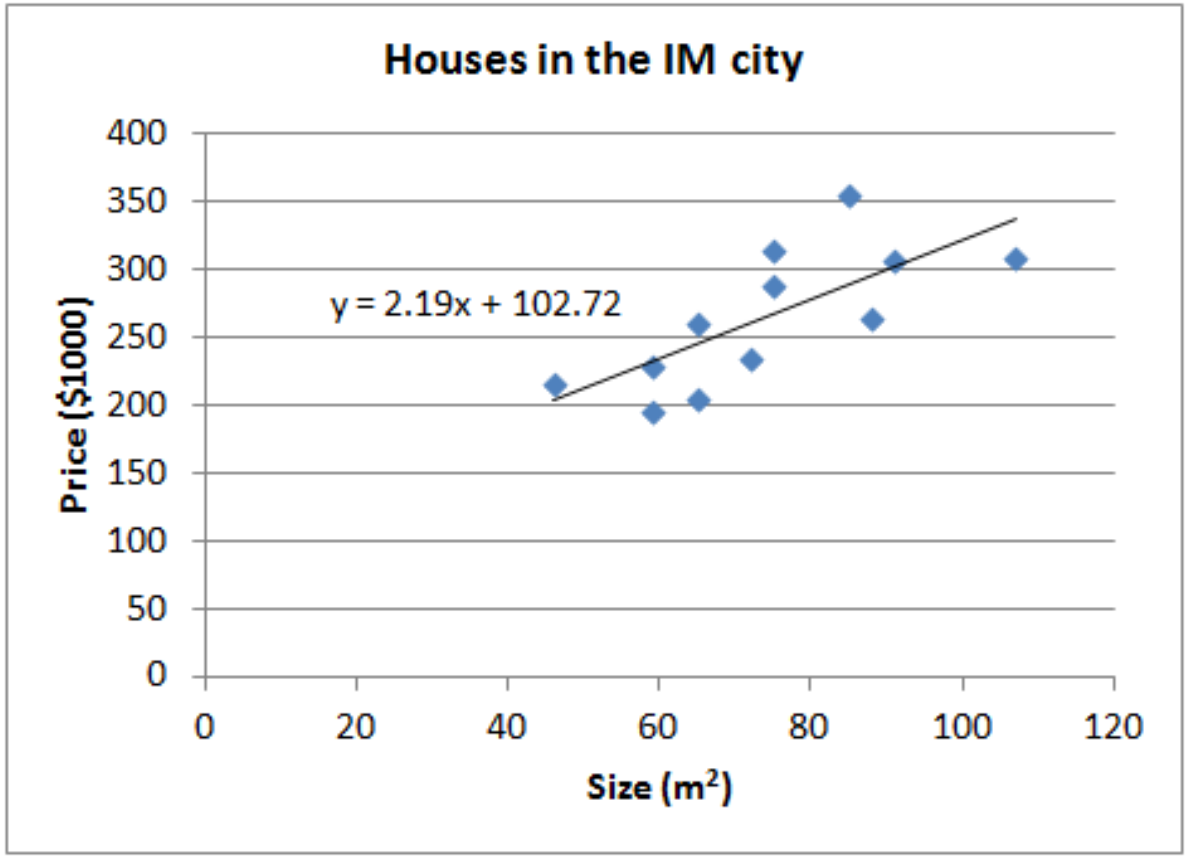
Scatter Plots

- We may switch the two axes.
- Is there any relationship?



Scatter Plots

- Does the line fit our data?



Scatter Plots

- Whether there exists a "significant" relationship between two (or more) variables?
  - Relationships may also be **nonlinear**.
  - A scientific way, **regression**, will be introduced in the Spring semester.
  - At this moment, judge a scatter plot by intuitions.
- Scatter plots are typically for two quantitative variables.
  - Scatter plots can be drawn when one variable is qualitative.
  - What if both variables are qualitative?

Some final remarks

- There is NO standard way of making frequency distributions and drawing graphs. It requires experiences and domain knowledge.
- In drawing a graph, never forget:
  - Caption.
  - Captions and labels for the x- and y-axes.
  - Unit of measurement.