MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Last time: Statistical concepts

- Today, we continue to go over the notes, and introduce more concepts.

More concepts.

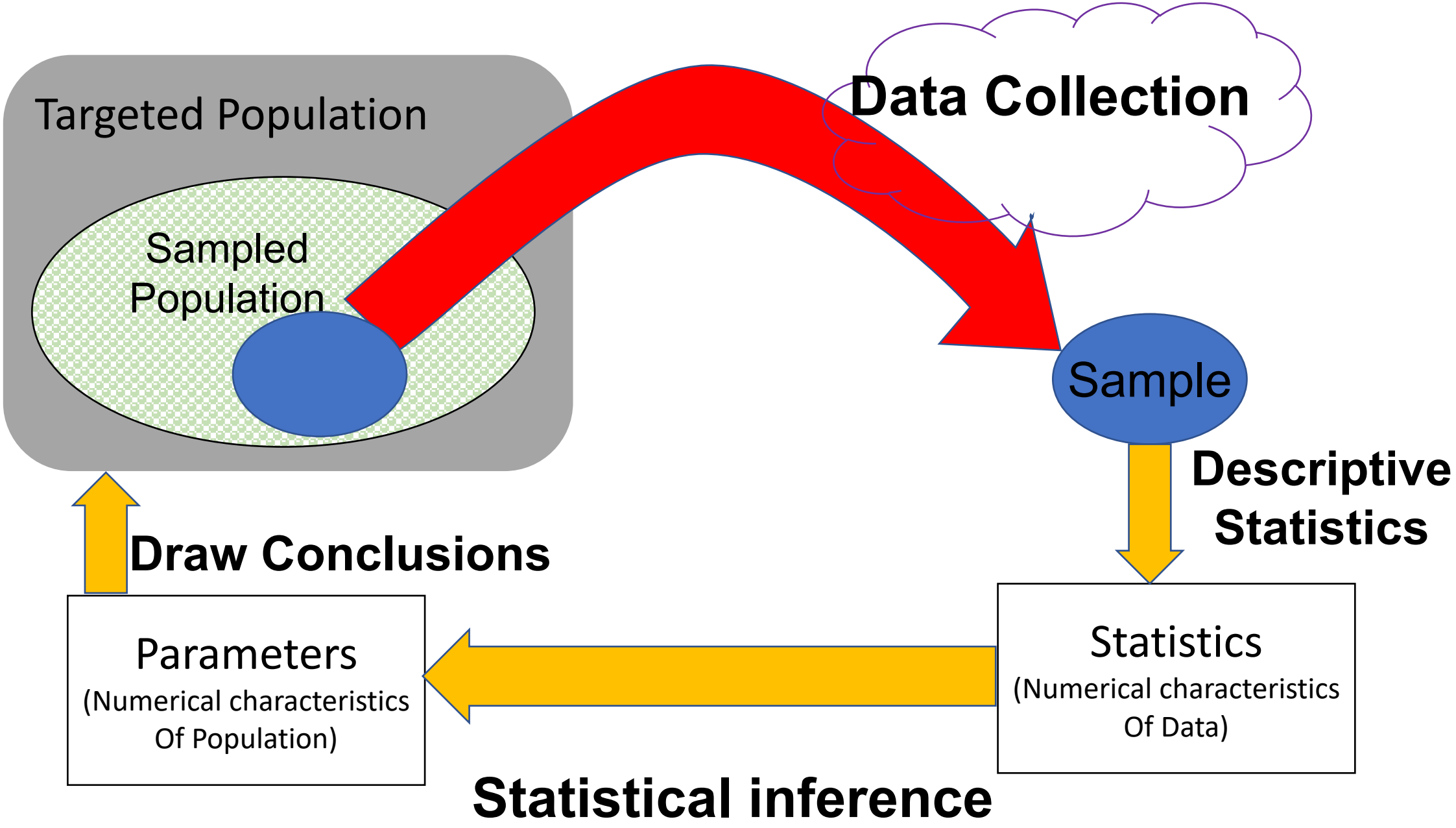
- Random sampling: A method for selecting a sample from a population such that every member of the population has an equal chance of being selected.

- A sample is a random sample form a population if
- (1) Each data value is selected in a manner totally unrelated to the other data values in the sample (independence).
- (2) Each data value is selected from a set of values that is the same as those in the population (identically distributed).

Concepts Example 1. *Political polling. (Cont'd)*

- To ensure the sample is a random sample, a table of random digit is used. (Or more likely, a computer algorithm generating random numbers is used.) Could see chapter 22 if you are interested in more sampling theory.

- An example of Non-random sample could be that the first 376 names in the phone directory being chosen. We would then be stuck with all persons with the last name starts with an "A". This could leads to many people from the same racial group or even the same family and rendering any conclusion from the sample useless for the general population.


Targeted Population
Sampled Population
Data Collection
Sample
Descriptive Statistics
Draw Conclusions
Parameters
(Numerical characteristics Of Population)
Statistics
(Numerical characteristics Of Data)
Statistical inference

Association is NOT Causation

- **Association**: There is a relationship between two random variables X and Y.

For example, “owning an iPhone in 2016” is associated with “having higher income” according to this paper.

==========

Researchers find that owning an iPhone or iPad is the number-one way to guess if you’re rich or not

Kif Leswing

Jul. 8, 2018, 11:00 AM AP

==========

- **Causation**: Changes in one variable X directly causes changes in another variable Y.

Would buying an iPhone in 2016 cause you to have higher income?

Association is NOT Causation

- You may also read the following paper "Association, correlation and causation" (https://www.nature.com/articles/nmeth.3587)

- *Do we need a causation result? Is associative result ok?*
- (Answer: depends on the study objective: marketing research? Scientific study? ….)

- **Can we use statistical studies to infer if X is associated Y or X causes Y?**
- Association can be judged statistically if the study is properly designed (random sampling).
- To infer causation, we need controlled study design.

Controlled study versus Observational study

- In a controlled experiment the allocation of "treatment" to subjects is done *at random*. Randomization eliminates biases and *enables the inference of causation*.

- In an observational study the "treatment" is a characteristic of the subject which we merely observe. It is usually not possible to eliminate biases and we can *only infer association*.

Example 4. Controlled vs. Observational

- Question: What is the effect of forest fires on the density of oak seedlings in a national forest?
- Experiment: Count the number of oak seedlings in a 10 square meters plot in the forest.

Controlled experiment

- 1. Select sites to study.
- 2. Choose half of sites at random and burn them.
- 3. Return later to count the seedlings at the sites.

Observational Study

1. Select the sites to study such that some are burned and some are not.
2. Count the seedlings at the selected sites.

Example 5. Video display terminals and miscarriage.

- Question: Does exposure to video display terminals increase risk of miscarriage?

- Target population: All pregnant women in the U.S. over next several years.
- Sample: 1583 pregnant women who attended one of the three Boston area hospitals.
- Statistics: Difference in miscarriage rates between women exposed to VDTs and women not exposed.

- Conclusion:

Example 6. Polio Vaccine effectiveness.

- In 1950s, a Polio vaccine was tested on all elementary school children whose parent gave consent.

- Target Population: All children in the US over next several years.
- Sample: The 400,000 plus children in the trial.
- Statistics:

Vaccine group: 33 of 200,745 developed paralysis. Rate 16/100000

Placebo group: 115 of 201,229 developed paralysis. Rate 57/100000

- Conclusion:

What is the relevant sample size?

- Experimental Unit: The largest unit to which a single treatment is applied.

- Sampling Unit: The largest unit on which measurements are made.

(Variance among treatments can only be reflected through variance among different experimental units. If sampling units are different from experimental units, measurement variance and treatment variance need careful distinction.)

Example 7 popcorn taste rating

The Professor pops two bags of regular and two bags of gourmet popcorn. One bag of each type has been stored at room temperature, while the other has been frozen. He passes the popped corn around the class, and each student takes a handful and consumes them.

- Experimental unit: a bag of popcorn.
- Sampling unit: a handful of popcorn.
- How many experimental units here?
- How many sampling units here?
- Is it possible to distinguish the treatment effect, with large number of sampling units in each experimental units?

Example 8 fertilizer testing

- A farmer puts five experimental fertilizers (including one control) on ten different plots in a field. Each fertilizer is placed on two plots. He then plants corn on each of the plots.
- (a) At the end of the growing season, he selects ten plants from each plot and weighs them individually.

  experimental unit: a plot of corn.

  sampling unit: an individual corn plant.
- (b) At the end of the growing season, he harvests and weighs all the corn on each plot.

  experimental unit: a plot of corn.

  sampling unit: a plot of corn.

Type of Data

- Quantitative data are measurements or counts that have meaningful numerical values.
- Qualitative data are attributes such as gender, occupation or the responses to a yes/no question that do not have inherent numerical values. Usually represented as nominal, ordinal or ranked data.

- Continuous data are measurements that could in principle be made arbitrarily precise. For example, weight or temperature.
- Discrete data belong to distinct classes and are usually counts or qualitative.

Basic statistical concepts. End of Module 1

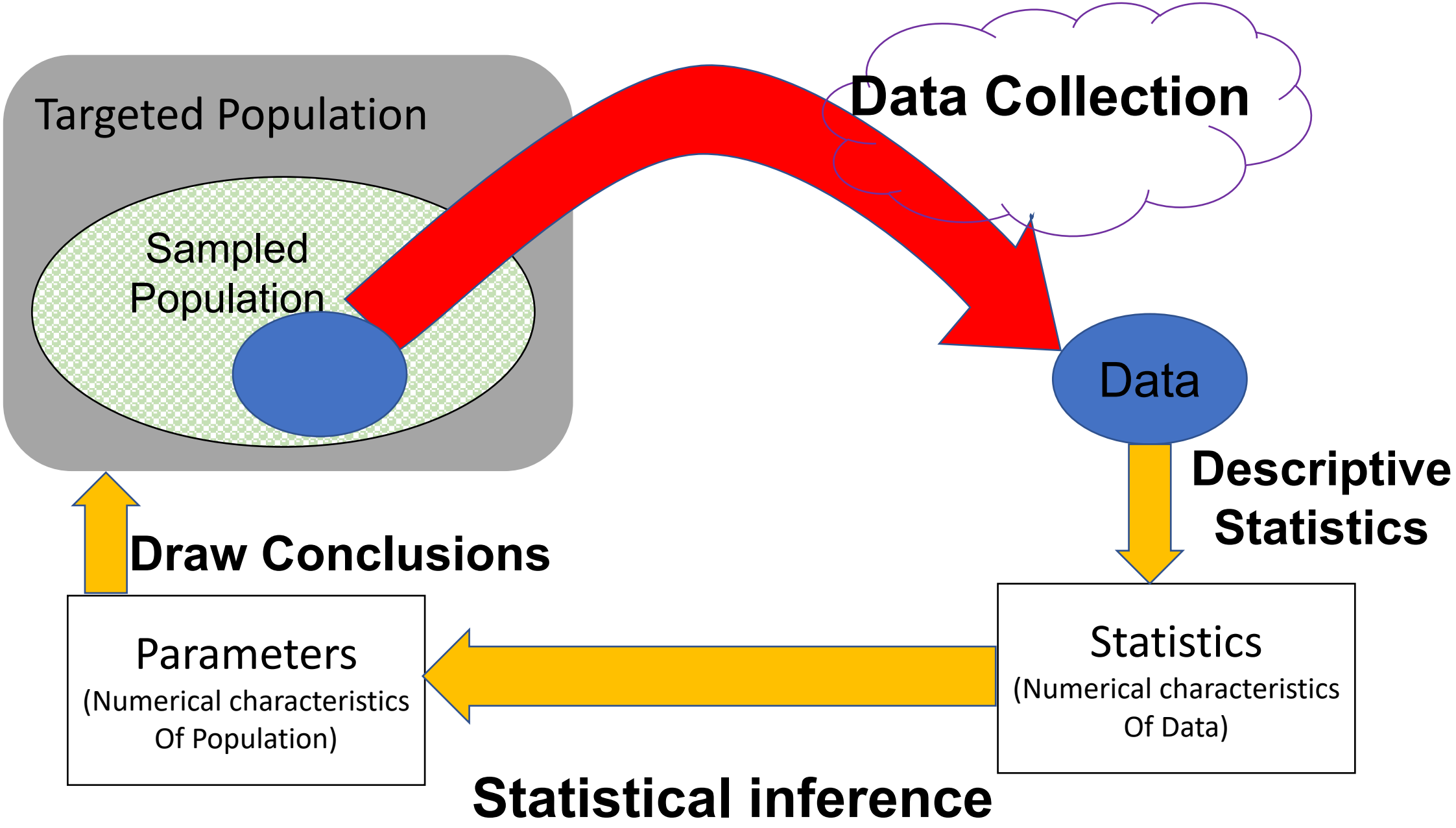
Statistical inferences is about drawing conclusions of the parameter (of population) using the statistics (of sample). Random sampling is important to get a representative sample of the population.

Inferential statistics is only a part of the process of statistical analysis.

Association is not causation. To infer causation, we need controlled study design.

Next topic: Descriptive Statistics

How to summarize data, as a few numbers or graphs.


Targeted Population
Sampled Population
Data Collection
Data
Descriptive Statistics
Draw Conclusions
Parameters
(Numerical characteristics Of Population)
Statistics
(Numerical characteristics Of Data)
Statistical inference

Descriptive Statistics

Ex 1. A marketing consultant observed 50 consecutive shoppers at a grocery store. Here are the amounts that each shopper spent (in dollars).

18.71	32.82	37.52	33.26	6.90
31.99	39.28	69.49	19.55	12.66
27.07	63.85	34.76	20.89	16.55
23.85	30.54	40.80	52.36	15.01
14.35	14.52	20.58	33.80	13.72
36.22	29.15	43.97	45.58	15.33
21.13	14.55	13.67	61.57	18.30
20.91	64.30	11.34	18.22	17.15
2.32	26.04	28.76	8.04	9.45
19.54	11.63	6.61	12.95	10.26

- It is really hard to get any information staring at a data set like this. We have to summary the data somehow.

Descriptive Statistics

(1) Frequency table. We could separate the range of the data into several intervals and count how many cases in each interval.

Category	(Absolute) Frequency	Relative Frequency
0<=x<10	5	0.100
10<=x<20	19	0.380
20<=x<30	9	0.180
30<=x<40	9	0.180
40<=x<50	3	0.060
50<=x<60	1	0.020
60<=x<70	4	0.090

Descriptive Statistics

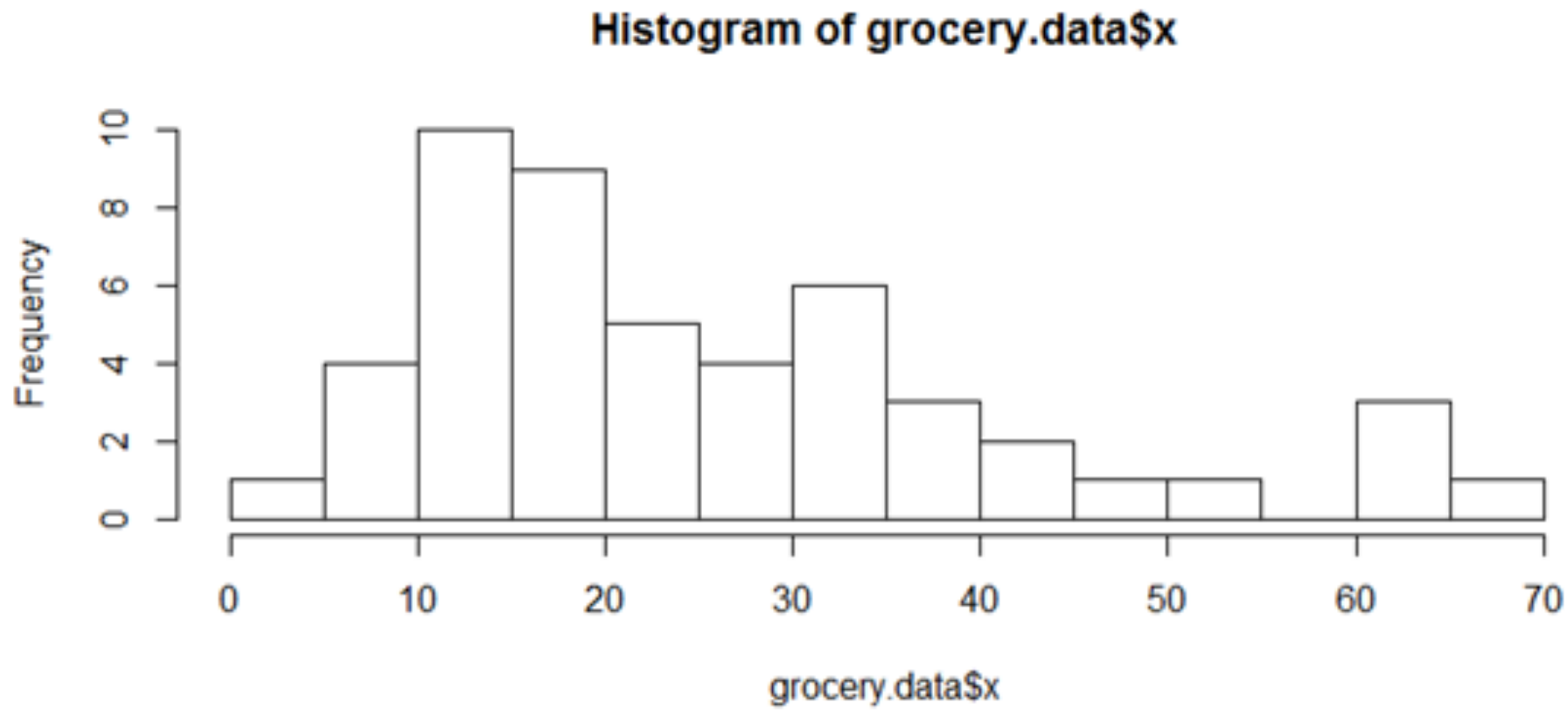
(2) Stem-and-leaf plot. Table served as a graph also.

Category	Stem & leaf
0<=x<10	26689
10<=x<20	01122334445567888 99
20<=x<30	136789
30<=x<40	012334679
40<=x<50	035
50<=x<60	2
60<=x<70	1449

18.71	…
31.99	…
27.07	…
23.85	…
14.35	…
36.22	…
21.13	…
20.91	…
2.32	…
19.54	…

Descriptive Statistics

(3) Histogram.



Descriptive Statistics

(4) Frequency Polygons

Descriptive Statistics

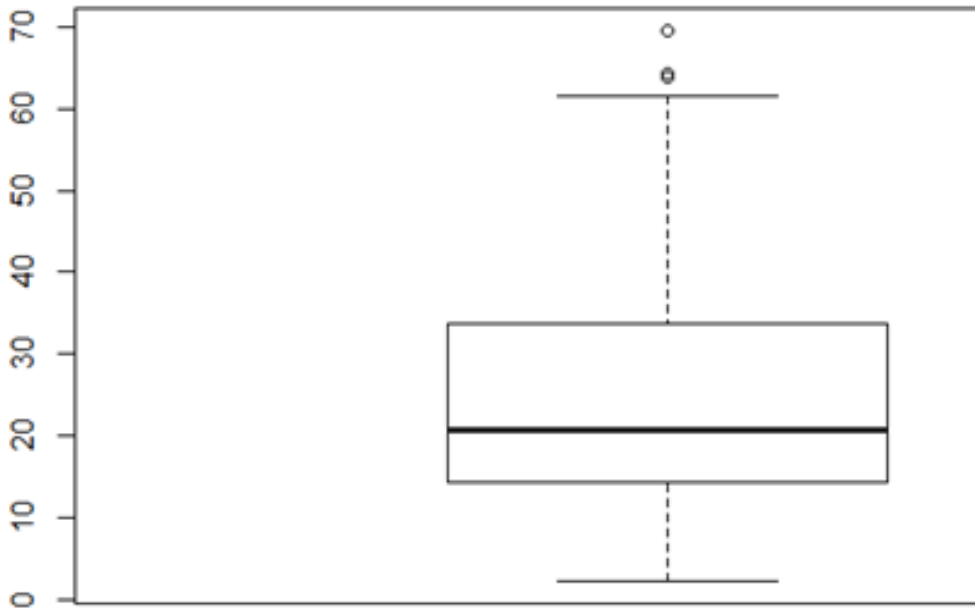
(5) Box Plot.

Lines in middle: Mean & Median

Box height: Interquartile range (IQR)

i.e. 25^{th} to 75^{th} percentile

Circles: Outliers, i.e. data points 1.5 IQR away from the box.



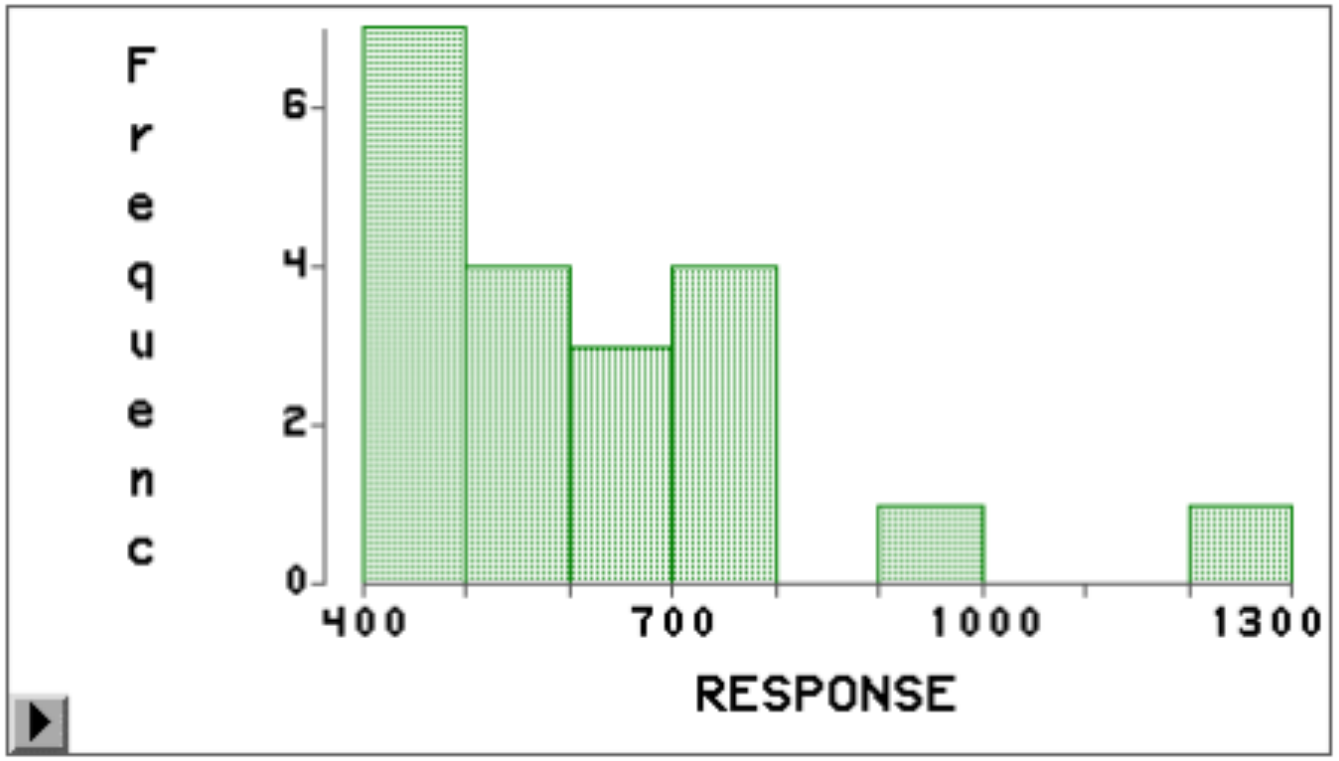
The line down below is the minimum value or 1.5IQR below box, the line above is the maximum or 1.5IQR above.

An Example: Response times

observations	response time	observations	response time
1	1270	11	660
2	600	12	500
3	710	13	440
4	600	14	490
5	720	15	490
6	930	16	490
7	770	17	550
8	720	18	490
9	490	19	550
10	440	20	500

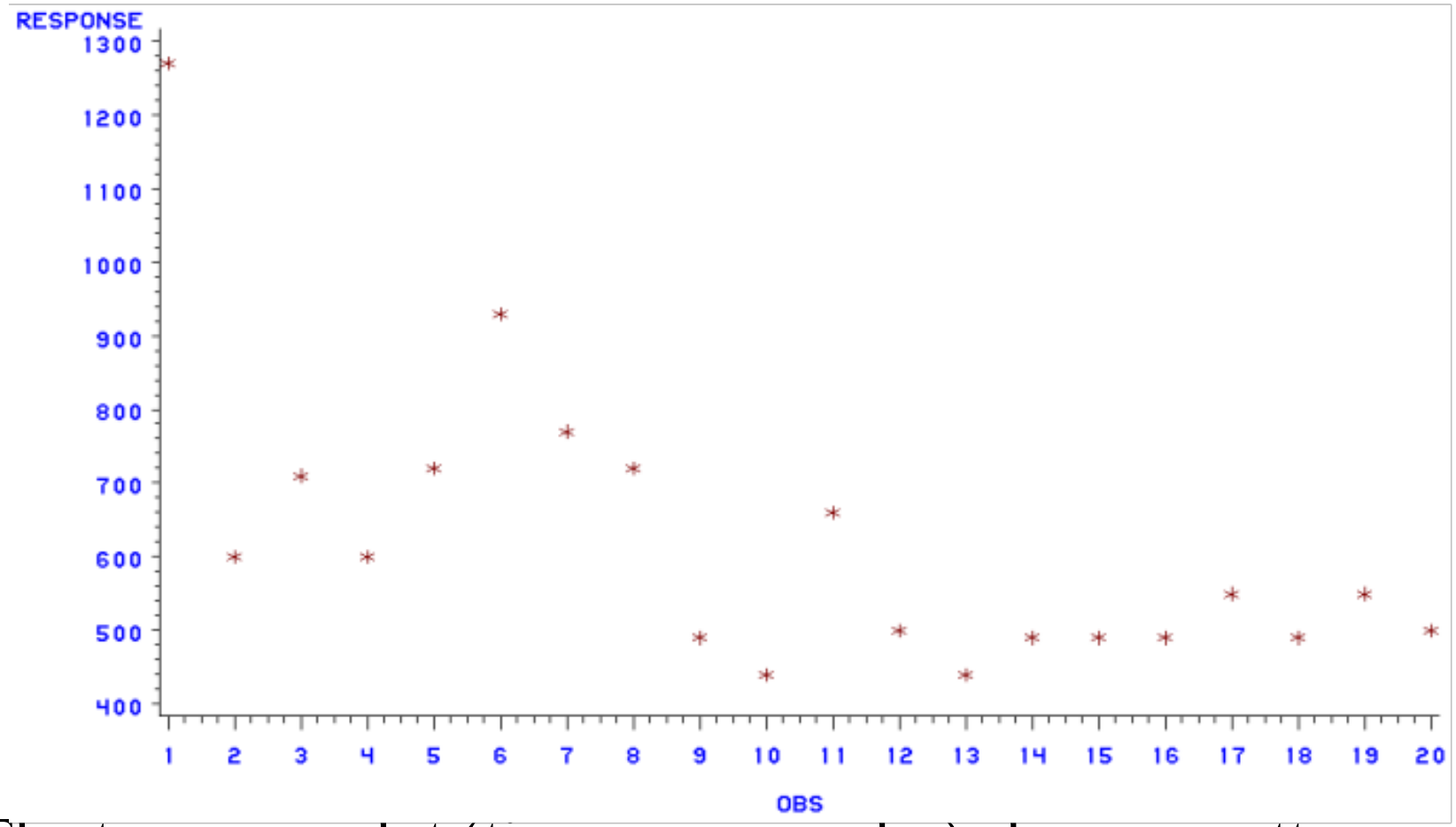
From A webpage that measure your response time

Histogram of Response times



- Judging from the data set, would the following response time be considered ordinary, too small or too large?
- 580 Ok
- 490 Ok
- 20 Too small
- 2000 Too big
- 750 Ok
- 1060 A little big, Ok.

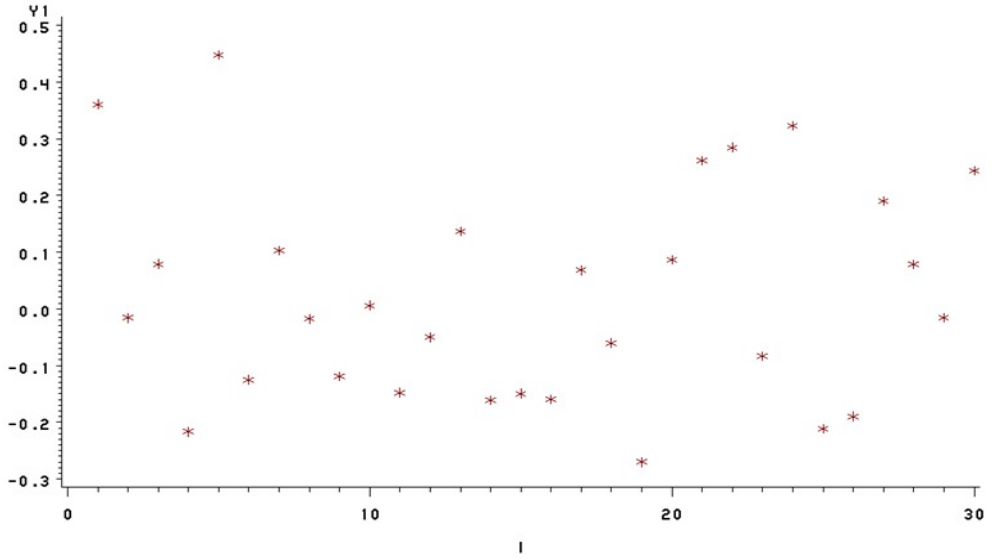
Scatter plot of Response times



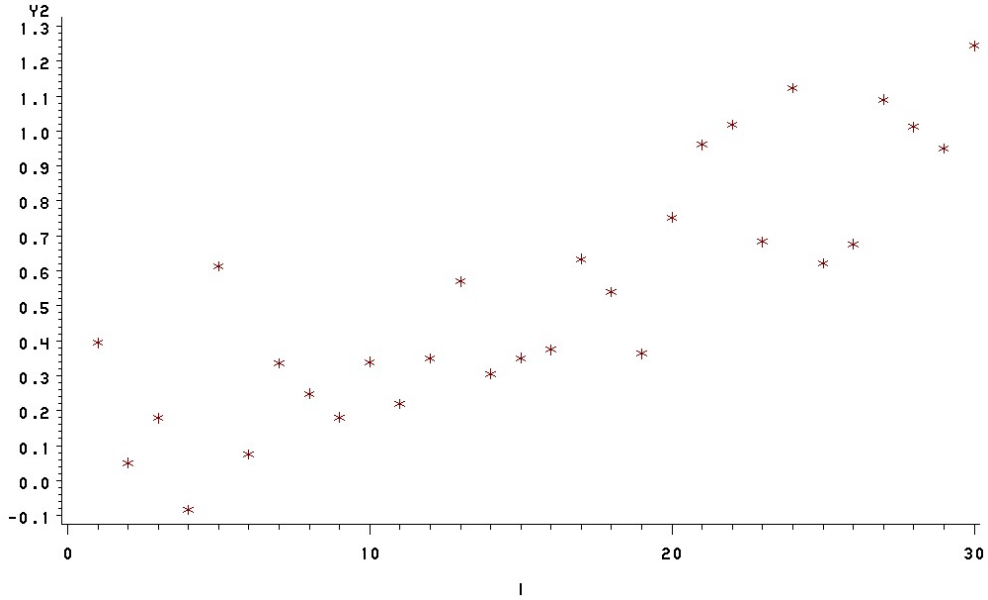
The two-way plot (time versus order) shows a pattern not observable in histogram: first try is very slow, then response gets faster and stabilizes.

Some patterns from scatter plots

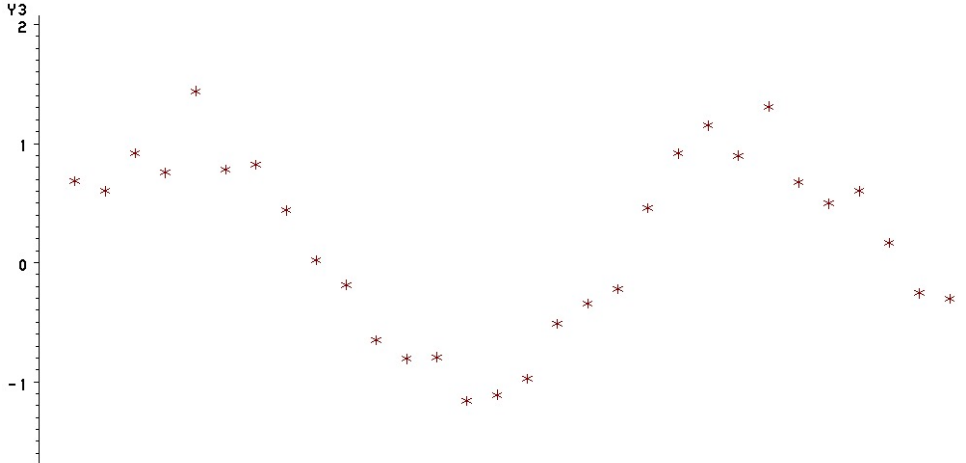
Random stationary



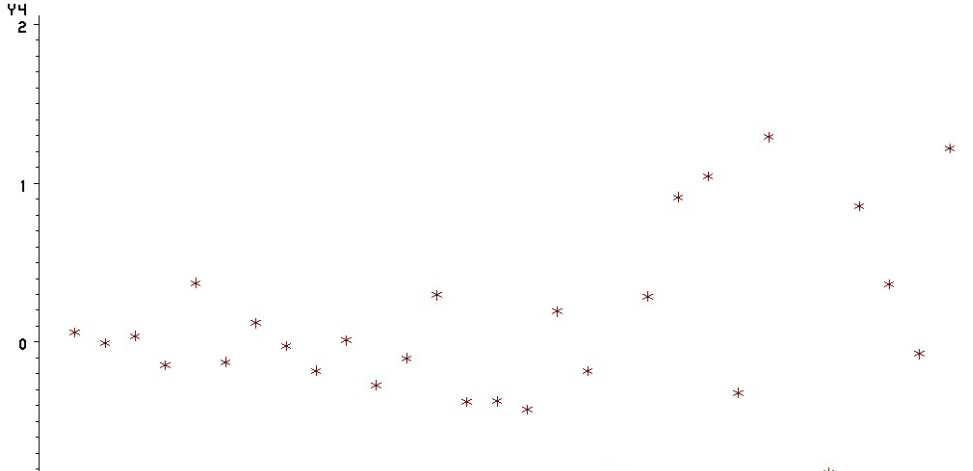
linear increasing



Periodical



increasing variance



Data visualization

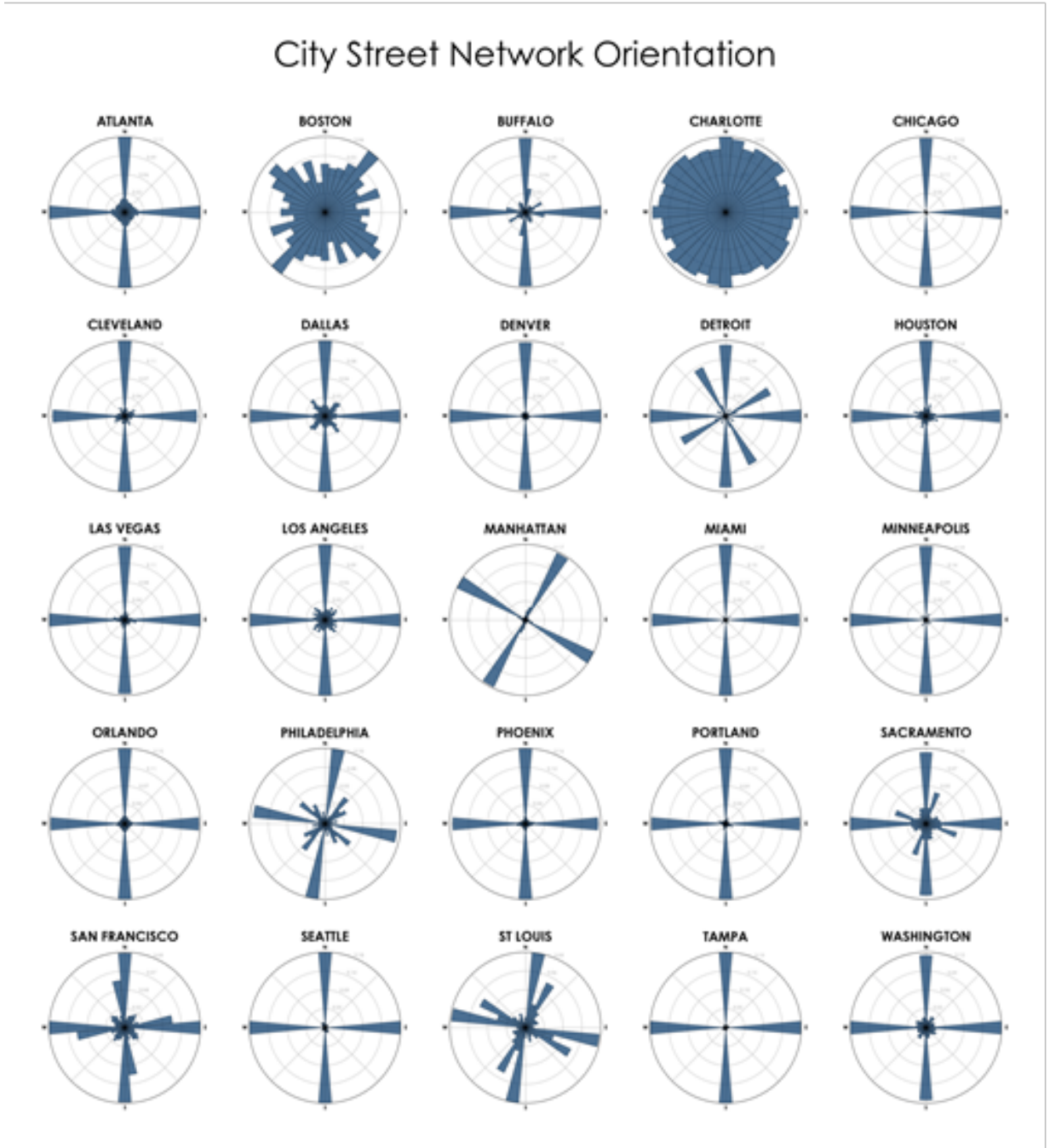
• Above are some very basic common plots to display data. Nowadays, data visualization itself is an important evolving research topic. People are inventing new ways to display high-dimensional data so that we can visualize patterns. The techniques (e.g. heatmap) are continuously invented and then programmed into common statistical software such as R. You are encouraged to read and learn about those techniques on your own. The idea is to display data to provide intuition with the available computational capability.

• Following is an example by Professor Boeing to visualize the distribution of city street orientations.
(https://geoffboeing.com/2018/07/comparing-city-street-orientations/)

Polar histogram

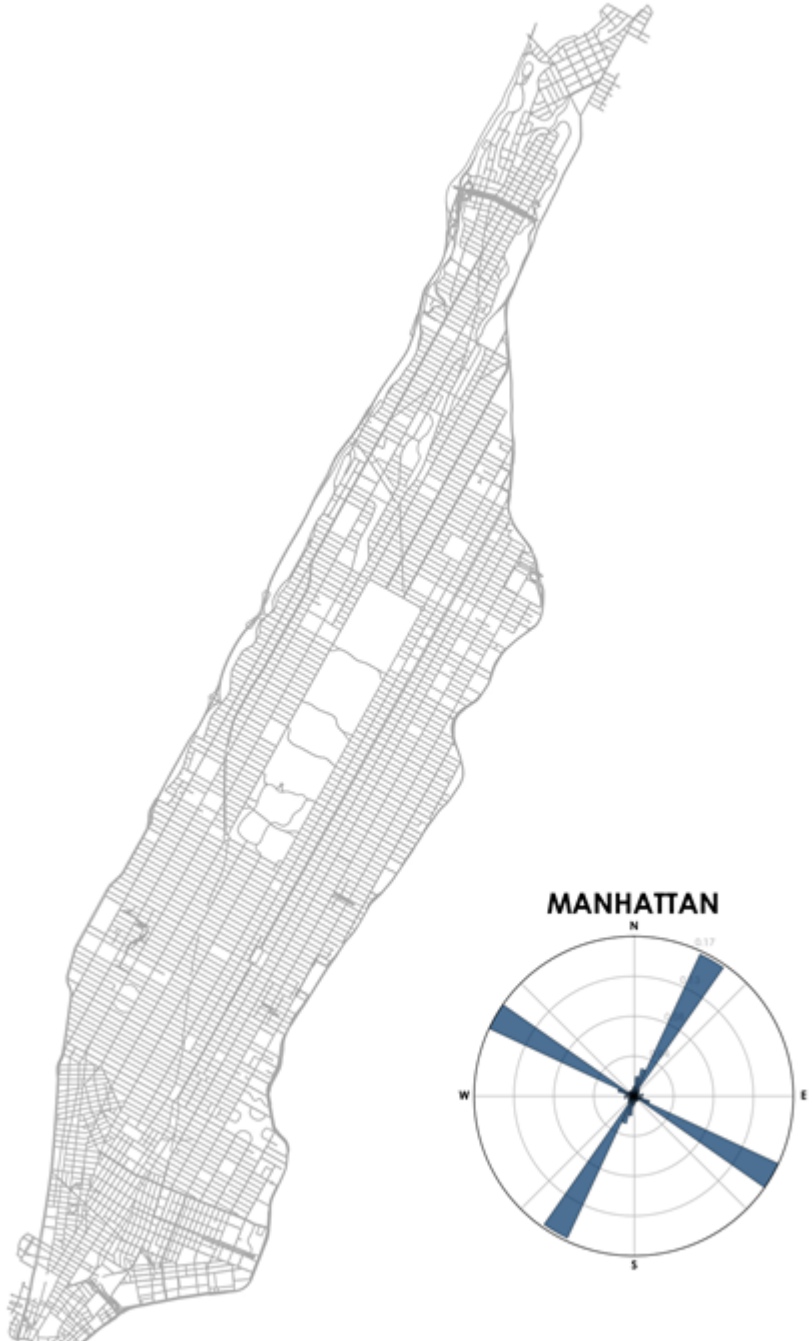
Most cities' streets are organized in grids, concentrate in two directions.

Boston and Charlotte are two exceptions.



Polar histogram

Most cities' streets are organized in grids, concentrate in two directions.
Boston and Charlotte are two exceptions.



Descriptive Statistics: Numerical summary

- Example: Ages at death for 8 (randomly sampled) women who divorced within 5 years of their first marriage:

  32, 83, 71, 75, 45, 68, 56, 57

- How to summarize?
- (1) Mean -- a measure for the center of the data

Definition: $\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$

In the example, $\bar{X} = \frac{1}{8}(32+83+71+75+45+68+56+57) = 60.875$

Descriptive Statistics: Numerical summary

• (2) Median -- another measure for the center of the data.

Definition: A value m such that at least half of the data ≤m and at least half of the data ≥m. This m is also the 50^{th} percentile.

In the example, sort the data from smallest to biggest,

32, 45, 56, 57, 68, 71, 75, 83

Any number m between 57 and 68 can be the median by the definition. In practice, we make this number unique by rule:

Odd number of data points: m is already unique, the $\frac{n+1}{2}$-th.

Even number of data points: average the middle two numbers.

In the example, m=(57+68)/2 = 62.5

Descriptive Statistics: Numerical summary

- (2) Median m is also the 50^{th} percentile.

Generally to find the p-th percentile (at least p percent data $\leq$m and at least 100-p percent data data $\geq$m) :

1. Sort the data from smallest to biggest.
2. Calculate k= $\frac{np}{100}$.
3. If k is an integer, average the k-th and (k+1)-th data.
   If k is not an integer, round up and use the k-th data.

In the example, m=(57+68)/2 = 62.5

Descriptive Statistics: Numerical summary

Generally to find the p-th percentile :

1. Sort the data from smallest to biggest.
2. Calculate k= $\frac{np}{100}$.
3. If k is an integer, average the k-th and (k+1)-th data.
   If k is not an integer, round up and use the k-th data.

In the example, 32, 45, 56, 57, 68, 71, 75, 83

50th percentile: k= $\frac{8*50}{100} = 4$. average the 4th and 5th data.

10th percentile: k= $\frac{8*10}{100} = 0.8$. Use the 1st data: $p_{10} = 32$

80th percentile: k= $\frac{8*80}{100} = 6.4$. Use the 7th data: $p_{80} = 75$

Descriptive Statistics: Numerical summary

- (3) Statistics measuring the spread of the data

  Range = Max- Min,

  Interquartile range (IQR) = 75^{th} percentile – 25^{th} percentile

In the example, 32, 45, 56, 57, 68, 71, 75, 83

Range = 83-32= 51,

IQR = 73-50.5= 22.5

75^{th} percentile: $k= \frac{8*75}{100} = 6$. Average : $\frac{71+75}{2} = 73$

25^{th} percentile: $k= \frac{8*25}{100} = 2$. Average : $\frac{45+56}{2} = 50.5$

Descriptive Statistics: Numerical summary

- (3*) Statistics measuring the spread of the data

(Sample) Variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$= \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$

(Sample) Standard deviation $S = \sqrt{S^2}$

In the example, 32, 45, 56, 57, 68, 71, 75, 83

Recall $\bar{X}$ =60.875.

$\sum_{i=1}^{n} X_i^2 = 32^2 + 45^2 + 56^2 + 57^2 + 68^2 + 71^2 + 75^2 + 83^2$

=31613

$S^2 = \frac{1}{8-1}(31613 - 8 * 60.875^2 60.875) = 280.98$

$S = \sqrt{280.98}$ =16.8

Descriptive Statistics

- Summary the data graphically or numerically.
- Technically easy: Histogram, scatter plot, boxplot, range, IQR, mean, median, variance, etc.
- Still need to consider: are we using statistics properly? We will continue next lecture.