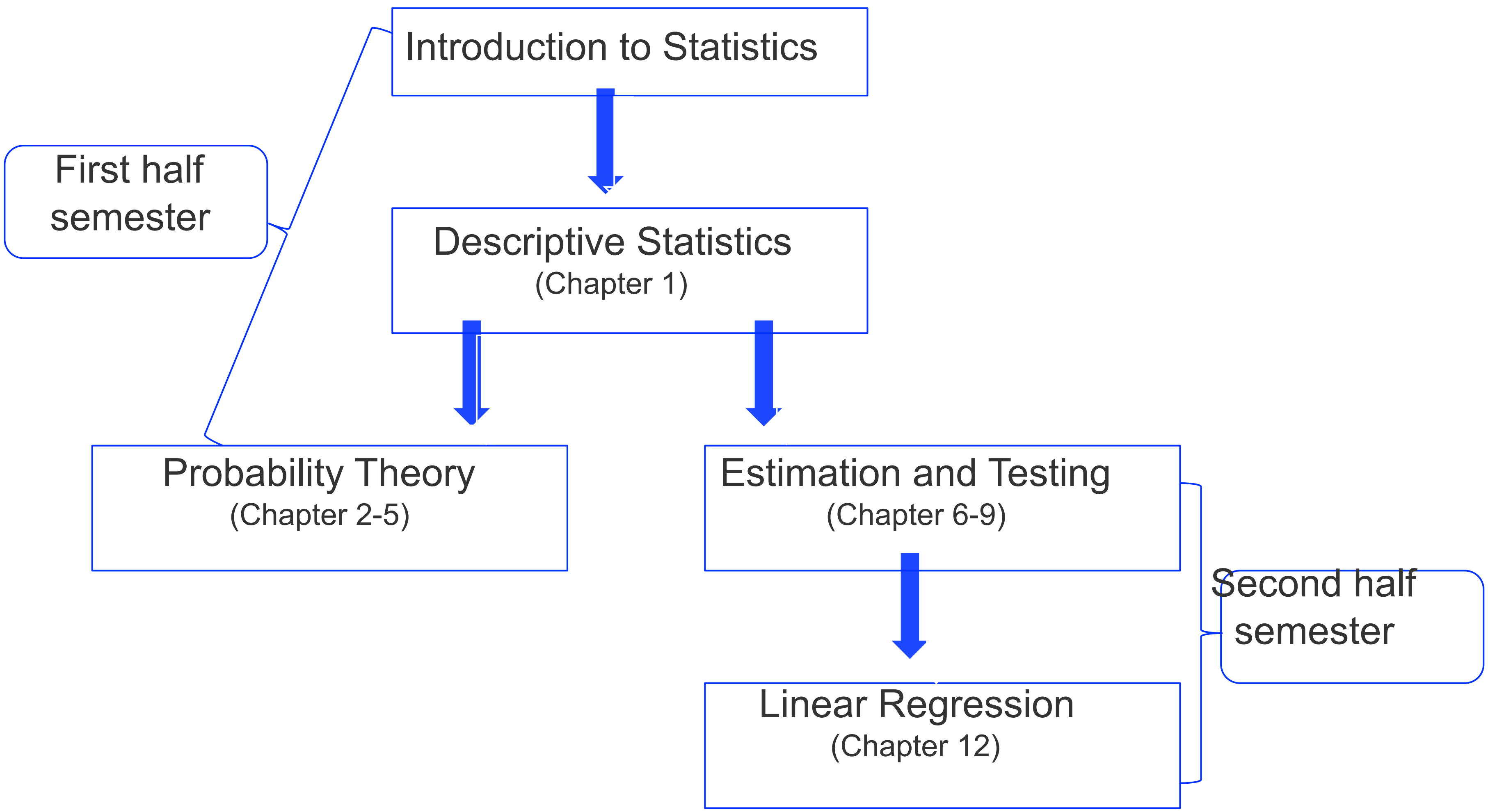
# S1211Q Introduction to Statistics
# Lecture 2

Wei Wang

Sep 10, 2012

# Overview of the course



# Basic concepts

- **Population**: the whole class of individuals which an investigator is interested in.
- **Census**: the desired information is available for all objects in the population.
- **Sample**: a subset (part) of the population which is examined or observed.
- **Sample Size**: the number of observations in a single sample.
- **Variable**: any characteristic whose value may change from one object to another in the population, including *univariate*, *bivariate*, *multivariate*.

# Probability

- What are random variables? Example: coin tosses.

- To describe random variables: *distribution*. This course will cover a variety of commonly used probability distributions.
  - Discrete distributions: Binomial, Poisson, etc.
  - Continuous distributions: Exponential, Normal (Gaussian), etc.

- Conditional probability.

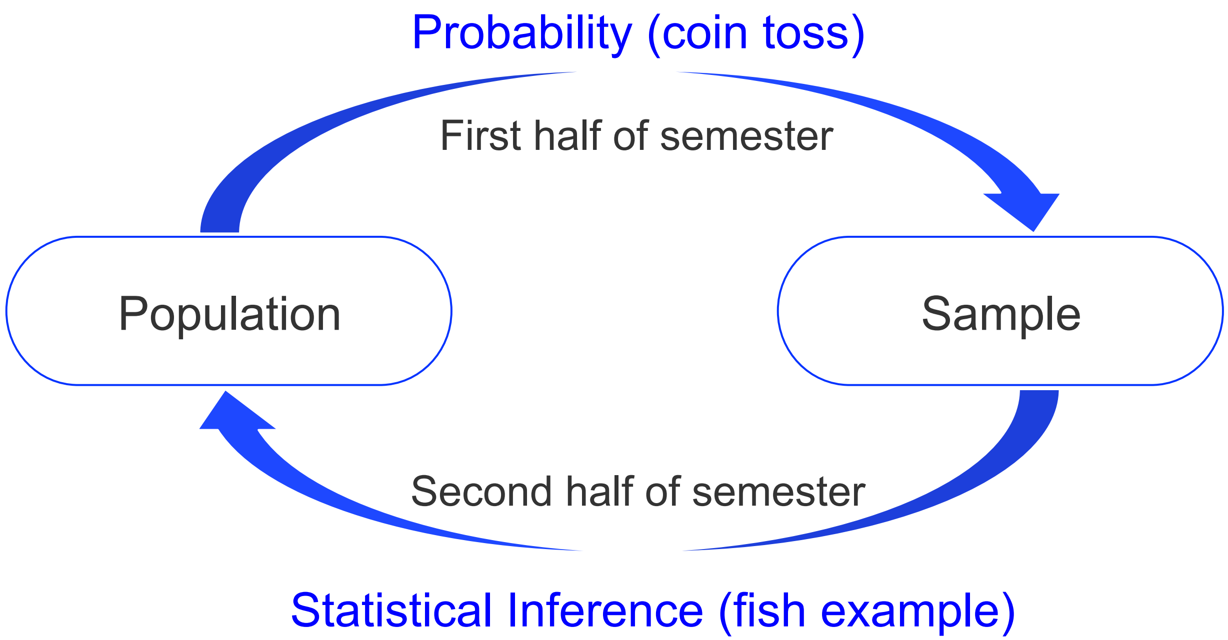
# Statistical Inference

- Estimation:
  - Point estimation. Example: What is the total number of fish in a lake?
  - Interval estimation.

- Hypothesis testing:
  - One sample testing.
  - Two sample testing. Example: Is there a significant improvement in the new drug?

- Estimation and hypothesis testing are just two different ways of looking at the same problem.

# Probability and Inference



# Examining a new data set

1. Examine each variable by itself.
2. Study the relationship between variables.

For both steps 1 and 2 we want to:

- Display the data graphically.
- Summarize the data numerically (Statistics).
- Construct a mathematical model.

# Descriptive Statistics

- Pictorial methods:
  - Stem-and-Leaf Displays.
  - Dotplots.
  - Histograms.

- All these methods convey information about the following aspects of the data:
  - Identification of a typical or representative value
  - Extent of spread about the typical value
  - Presence of any gaps in the data
  - Extent of symmetry in the distribution of values
  - Number and location of peaks
  - Presence of any outlying values

# Stem-and-Leaf displays

- Steps for constructing a Stem-and-Leaf Display:
  1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
  2. List possible stem values in a vertical column.
  3. Record the leaf for every observation beside the corresponding stem value.
  4. Indicate the units for stems and leaves someplace in the display.

- R demo for Stem-and-Leaf:
  - Command: `>stem(x)`
  - Option: `scale=`…, scale has to be a positive number. It controls the plot length. A value of scale=2 will cause the plot to be roughly twice as long as the default (=1).

# More basic concepts

- **Discrete Variable**: Its set of possible values is either finite or else can be listed in an infinite sequence. (Gender, Age, etc.)

- **Continuous Variable**: Its possible values consist of an entire interval on the real number line. (Height, Weight, etc.)

- **Frequency**: Number of times a value occurs in the data set.

- **Relative Frequency**: Frequency/(Sample size).

# Histogram

- Most commonly used tool in descriptive statistics.

- Histogram for discrete data:
  - Determine the frequency and relative frequency of each *x* value.
  - Mark possible x values on a horizontal scale.
  - Above each value, draw a rectangle whose height is the relative frequency (or the frequency) of that value.

- Histogram for continuous data:
  - Divide the range of the data into classes (5-10) of *equal width*. (It can also be unequal.)
  - Determine the frequency and relative frequency for each class.
  - Mark the class boundaries on a horizontal measurement axis.
  - Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

# Constructing histogram

- Example: The maximum daily temperature in degrees Fahrenheit measured from May to September 1973 at La Guardia Airport. (154 observations)

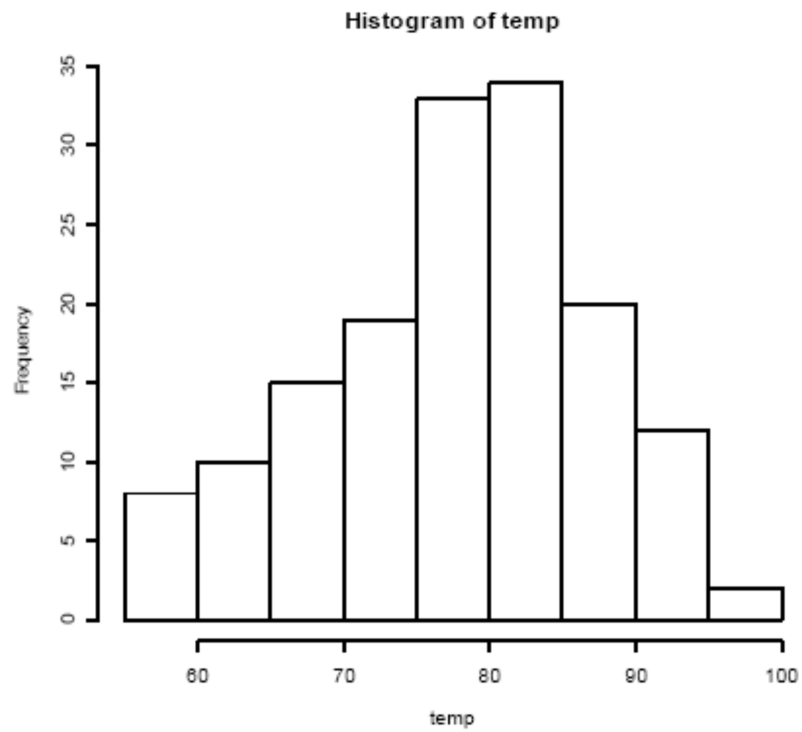
  Data

  {67 72 74 62 56 66 65 59 61 69 74 69 66 68 58 64 66 57 68 62 59 73 61 61 57 58 57 67 81 79 76 78 74 67 84 85 79 82 87 90 87 93 92 82 80 79 77 72 65 73 76 77 76 76 76 75 78 73 80 77 83 84 85 81 84 83 83 88 92 92 89 82 73 81 91 80 81 82 84 87 85 74 81 82 86 85 82 86 88 86 83 81 81 81 82 86 85 87 89 90 90 92 86 86 82 80 79 77 79 76 78 78 77 72 75 79 81 86 88 97 94 96 94 91 92 93 93 87 84 80 78 75 73 81 76 77 71 71 78 67 76 68 82 64 71 81 69 63 70 77 75 76 68}

  Draw a histogram.

# Example cont.

| Class | Count | Percent |
|---|---|---|
| 55-59.9 | 8 | 5.2 |
| 60-64.9 | 10 | 6.5 |
| 65-69.9 | 15 | 9.8 |
| 65-74.9 | 19 | 12.4 |
| 75-79.9 | 33 | 21.6 |
| 80-84.9 | 34 | 22.2 |
| 85-89.9 | 20 | 13.1 |
| 90-94.9 | 12 | 7.9 |
| 95-99.9 | 2 | 1.3 |




- R demo. `>hist(x)` (option: `breaks=`…)

# Density Histogram

- Besides Frequency/Relative Frequency histogram, there is another type of histogram: Density Histogram.
- The only difference is that in Density Histogram

$$\text{rectangle height} = \frac{\text{relative frequency}}{\text{class width}}$$

- Then the areas of all the rectangles add up to 1.
- We can also allow the class width to be unequal. Why?
- In `R` function `stem`, set argument `freq=FALSE` produces density histograms.

# Examining distributions

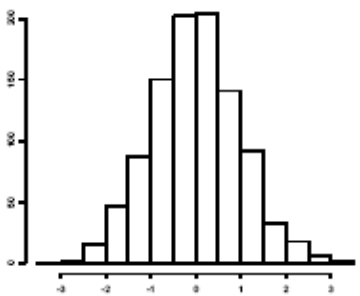
- When examining a distribution, look at its shape, center and spread. Look for clear deviations from the overall shape.

- We are interested in whether it is symmetric or skewed, as well as the number of modes.

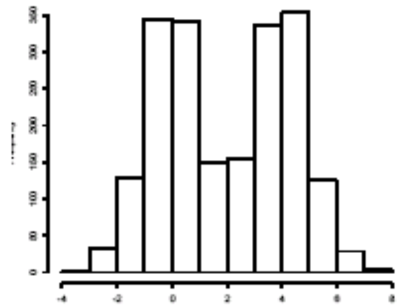
- Outliers are observations that lie outside of the overall pattern of a distribution.

# Examining distributions

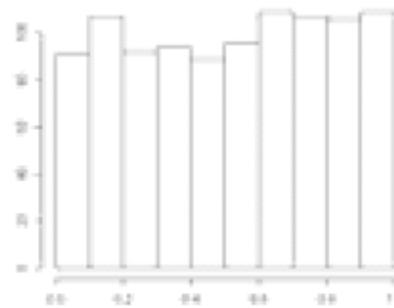
(a) Symmetric, unimodal



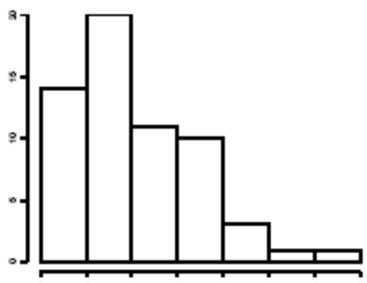
(b) bimodal



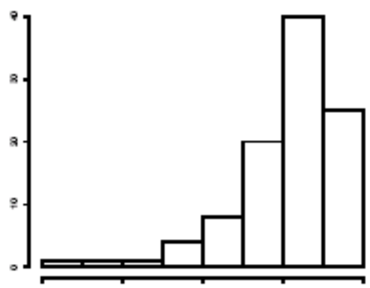
(c) Uniform



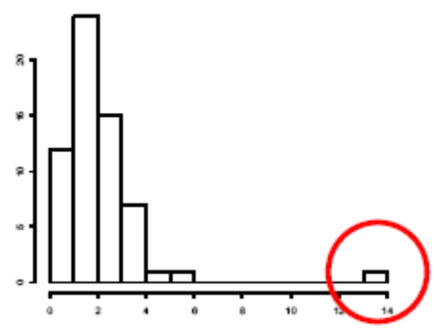
(d) right skewed



(e) left skewed



(f) Outlier



# Describing distributions numerically

- For single variables, We are interested in summaries that provide information about the center and spread of the distribution.

- A statistic is a numerical summary of data.

- The two most common measures of center are the mean and median.

- “generous” vs. “selfish”.

# Mean

- If we have *n* ,observations, their mean is defined by,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

or

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ex. Calculate the mean of the data set: {1,2,3,4,5}.

$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

Ex. Calculate the mean of the data set: {1,2,3,4,30}.

$$\bar{x} = \frac{1+2+3+4+30}{5} = \frac{40}{5} = 8$$

# Mean cont.

- The mean is non-resistant, meaning that it is influenced by very large or very small data points that are extreme values for the data set.

# Median

The median, written as M, is defined as the middle value of a data set.

1. List all *n* observations in order of size.
2. If *n* is odd, the median is the center value of the ordered list.
3. If *n* is even, the median is the average of the two center observations.

# Median Cont.

Ex. Calculate the median of {6,2,5,19,12,10}.

2 5 6 | 10 12 19

M is the average of 6 and 10, hence M=8.

Ex. Calculate the median of {1,2,3,4,5} and {1,2,3,4,30}.

1 2 3 4 5 M=3.

1 2 3 4 30 M=3.

# Median cont.

- The median is resistant (robust) to the extremes in the data set. Extremely large or small values do NOT influence the median.

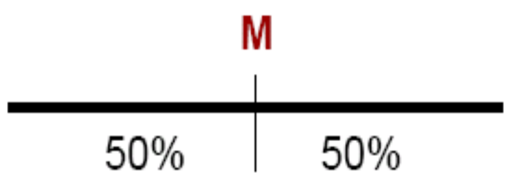
# Trimmed Mean

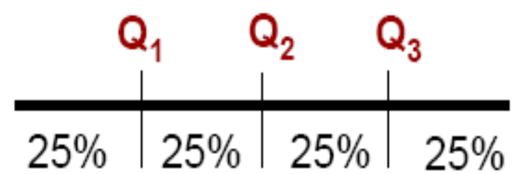
- To explicitly remove the impact of outliers, Trimmed Means can be used. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and largest 10% of the sample and then averaging what remains.
- In sports like Gymnastics and Diving, trimmed means are used to calculate the scores of athletes.

# Quartiles

- The median divides the data into two groups of equal size.

- The quartiles divide the data into four groups of equal size.



# Quartiles cont.

To find the quartiles:

1. Find the median.
2. Find the first quartile (Q1, or the *lower fourth*) by finding the median of the lower half of the data.
3. Find the third quartile (Q3, or the *upper fourth*) by finding the median of the upper half of the data.

   (*When n is odd include the median in both halves in steps 2 and 3.*)

Ex. Find the quartiles for the data set {2,4,6,8,12,14,18,19,41}.

2 4 6 8 12 14 18 19 41

2 4 6 8 12

12 14 18 19 41

M = 12   Q1 = 6   Q3=18

# Percentiles (Quantiles)

- First Quartile and Third Quartile are two particular examples of Percentiles, specifically, 25% percentile and 75% percentile.
- For any $\alpha \in (0, 1)$, we can define a $(100 \times \alpha)\%$ percentile.