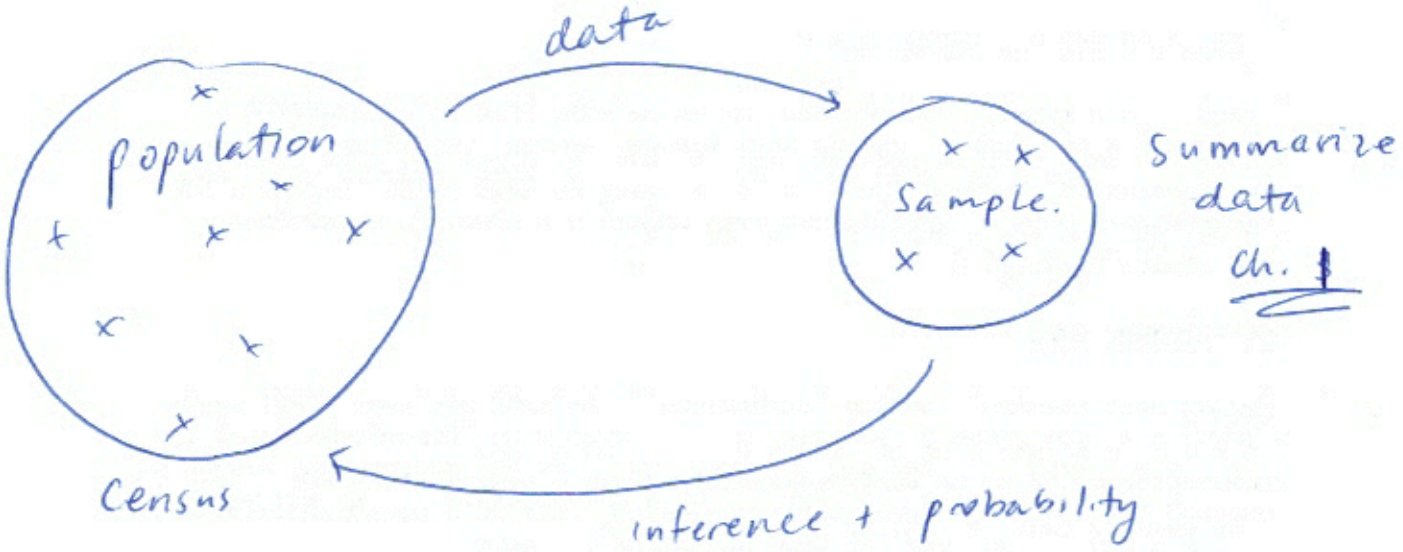

data
population
Sample.
Summarize data
Ch. 1
Census
inference + probability

Ch. 1 - Summary and Display of Univariate Data

Statistics is a science involving the design of studies, data collection, summarizing data, interpreting resulting and drawing conclusions.

Often in statistics, we want to know something about a certain **population** (entire class of individuals which we want to generalize about). So we collect a **sample** (subset of the population) in order to infer something about that population.

Suppose some students at UBC are interviewed about their study habits and the results are shown in the data table below.

ordinal

Subject	Age (years)	Gender	Hours spent studying/week	Stress Level
1	19	M	2	Low
2	20	F	4	Medium
3	20	M	19	Medium
4	28	M	8	Low
5	21	F	11	Low
6	22	F	7	High
7	21	M	6	Medium
8	20	M	10	High

Table rows: individual **cases** about whom we record characteristics, called **variables** (characteristic of the measured object or individual).
e.g. columns in the table above are each variables

1 Types of Variables

1. Categorical

- categories that cases fall into
- summarized by counts/frequencies
- subdivided into 3 categories:
  - binary (2 named categories)
  - nominal (> 2 unordered categories)
  - ordinal (> 2 ordered categories)

gender (M/F)
political view
ex. A, B, C, D, F grade

2. Quantitative

- measured numeric values
- underlying scale
- usually has units

ex. percent grade.

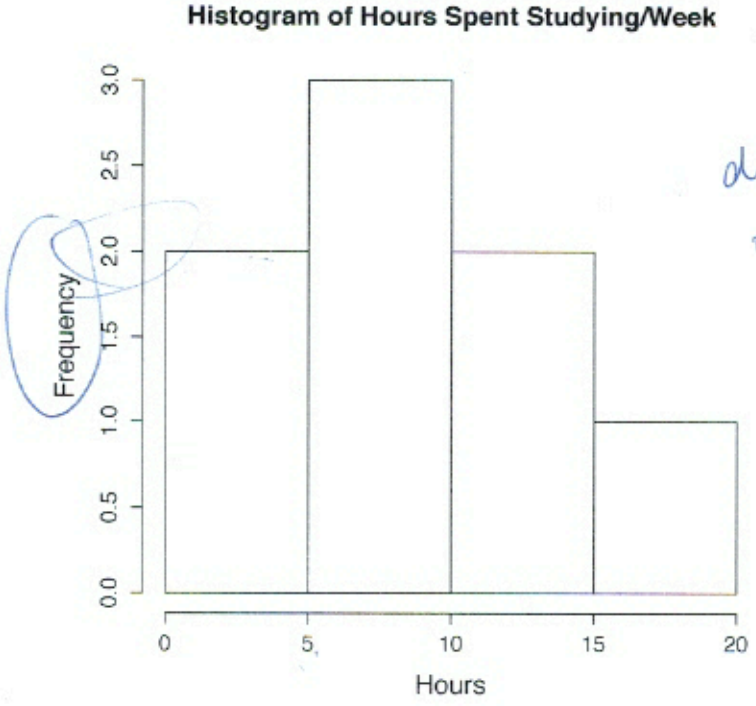
2 Displaying and Summarizing Quantitative Variables

Histogram: breaks the range of values of a variable into "bins" and displays the count or percent of the observations that fall into each bin.
e.g. below is a histogram of number of hours spent studying/week

Frequency Table

Bin	Center of Bin (c_i)	Frequency (n_i)
0 - 5	2.5	2
5 - 10	7.5	3
10 - 15	12.5	2
15 - 20	17.5	1

" border obs. " bin to right.



distribution of # hours studied.

Describing a distribution

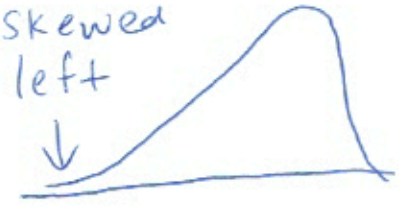
1. Shape
   (a) Number of peaks (unimodal, bimodal, multimodal)
   (b) Symmetry (symmetric, right skewed, left skewed)
   (c) Outliers - unusually large or small observations
2. Center - where do the observations cluster about?
3. Spread - how spread out are the observations?

Shape

a) # peaks

b) Symmetry

ex. grades



ex. age of death from natural causes.

age school.

Measures of Center

1. **Mean**: average of the observations

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Example 1. Find the mean number of hours spent studying/week.
Solution:

$$\bar{x} = \frac{2+4+19+8+11+7+6+10}{8} = 8.375 \text{ hours}$$

The sample mean can also be approximately calculated from a frequency table using the formula:

$$\bar{x} \approx \frac{\sum_{i=1}^{k} c_i n_i}{\sum_{i=1}^{k} n_i}$$

Example 2. Find the approximate mean number of hours spent studying/week from the frequency table above.
Solution:

$$\bar{x} \approx \frac{2.5 \times 2 + 7.5 \times 3 + 12.5 \times 2 + 17.5 \times 1}{8} = 8.75 \text{ hours}$$

2. **Median**: middle number of the observations. The number such that half of the observations are smaller and half are larger.

To find the median:

(a) Arrange the data in ascending order

(b) If the number of observations is...

- Odd: median = $(\frac{n+1}{2})th$ observation (median is the middle value)
- Even: median = average of $(\frac{n}{2})th$ and $(\frac{n+1}{2})th$ observations (median is the average of the 2 middle values)

n = 5

o o ● o o

5+1 / 2 = 3rd.

n/2 + 1

n = 4

o ● | ● o

Example 3. Find the median number of hours spent studying/week.
Solution:
First we arrange the numbers in order: 2, 4, 6, 7, 8, 10, 11, 19

$$\begin{aligned}\text{median} &= \text{average of } 4th \text{ and } 5th \text{ values} \\ &= \frac{7+8}{2} \\ &= 7.5 \text{ hours}\end{aligned}$$

Measures of Spread

1. **Variance and Standard Deviation:**

$$\text{Variance:} \quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

$$\text{Standard deviation:} \quad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Example 4. Find the standard deviation of number of hours spent studying/week. Solution:

$$s^2 = \frac{(2-8.375)^2 + (4-8.375)^2 + \ldots + (10-8.375)^2}{8-1} = 27.125$$

$$s = \sqrt{27.125} = 5.21 \text{ hours}$$

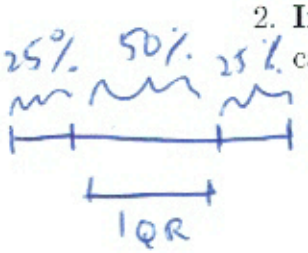
The sample standard deviation can also be approximately calculated from a frequency table using the formula:

$$s \approx \sqrt{\frac{\sum_{i=1}^{k}(c_i - \bar{x})^2 n_i}{n-1}}$$

Example 5. Find the approximate standard deviation of the number of hours spent studying/week from the frequency table above.
Solution:

$$s \approx \sqrt{\frac{(2.5-8.75)^2 \times 2 + (7.5-8.75)^2 \times 3 + (12.5-8.75)^2 \times 2 + (17.5-8.75)^2 \times 1}{8-1}}$$

$$\approx 5.18 \text{ hours}$$



2. **Interquartile Range (IQR)**: length of the range of an interval that captures the middle 50% of the data

- **First quartile** (Q1) is the value in the sample that has 25% of the data at or below it (it is the median of the lower half of the sorted data, excluding M).
- **Third quartile** (Q3) is the value in the sample that has 75% of the data at or below it (it is the median of the upper half of the sorted data, excluding M).

$$IQR = Q_3 - Q_1$$

Example 6. Find the IQR of hours spent studying/week.

Observation	Hours Spent Studying/Week
1	2
2	4
3	6
4	7
5	8
6	10
7	11
8	19

$Q_1 = \frac{4+6}{2} = 5$

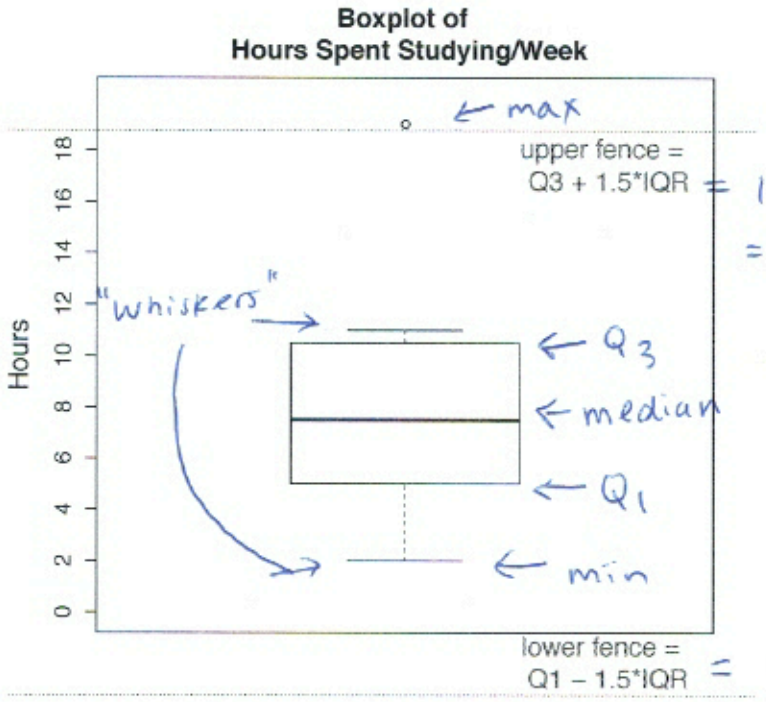
$Q_3 = 10.5$

$IQR = 10.5 - 5 = 5.5$

5-Number Summary and Boxplots

The 5-number summary includes the: Minimum, Q_1, Median, Q_3, Maximum

The boxplot provides the graphical display of the 5-number summary. It shows the distribution of the variable much like a histogram. It is very useful for comparing a quantitative variable over different levels of a categorical variable (eg. blood pressure over 4 different drugs) to see distributions of data over these categories.

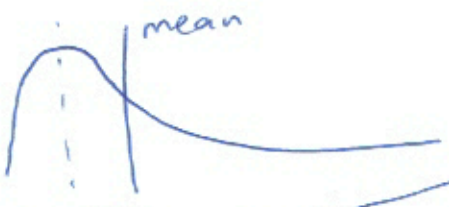


When to use which summary statistic?

- Distribution is roughly symmetric and no huge outliers - mean and standard deviation are good measures of center and spread.
- Distribution is skewed or has outliers - median and IQR will be better measures.

report mean + SD.

mean ≈ median



report median + IQR

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$