Lecture 2

Descriptive Statistics

Today's Updates / Reminders

- Get registered with WebAssign if you haven't already. There is an orientation assignment that is for a grade, due at the start of next week at 11:59pm
  - Technically, all homework really closes at 4am. I give a 4-hour grace period in case you're last-minute and the computer glitches.
- Due to popular request, I have started the process to create a course pack of these notes. Fill out the survey in Moodle to indicate your interest.
- Your first homework is open now. It will be due a week from today since we will finish chapter 1 today.
- Lab 1 opens today after class. Due a week from Friday.
- If this is your first day, please come see me after class to get "first day" information.

How to avoid/reduce some of the above?
An experiment should be able to show causation.

Causation: the changes in the factors are the only cause of the changes in the response variables (or outcomes). A good experiment should be able to minimize the effect of lurking variables.

There are **three principles of experimental design** (to show causation)

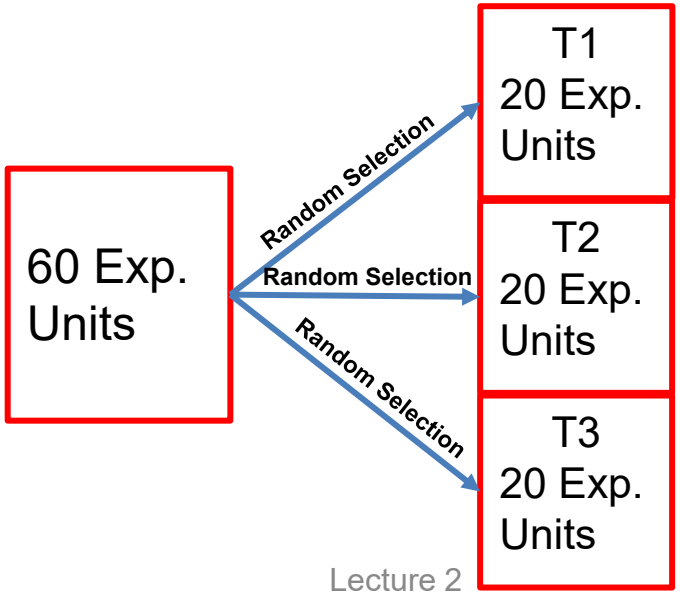
1. **Compare:** An experiment should have two or more treatments to control the effects of lurking variables on the response. One group should receive no treatment; that is the **control group**.

2. **Randomize:** Use impersonal chance to assign experimental units to the different treatments. This avoids bias. Also, ideally, an experiment should be **double-blind** (neither the subjects nor the experimenters know which treatment the subject received).

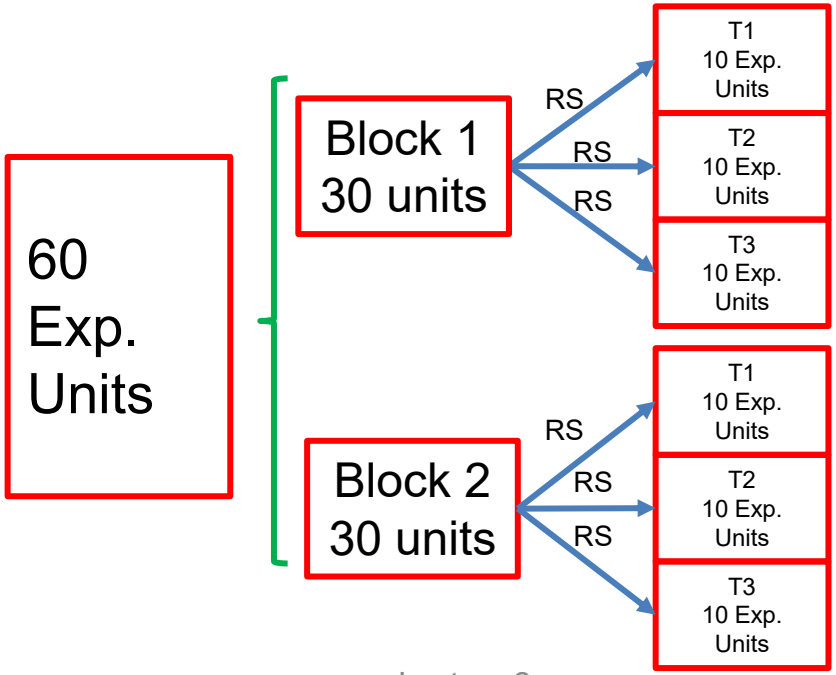
3. **Repeat:** Each treatment should be applied to many units to reduce chance of variation in the results.

Types of Experimental Design

1. **Completely randomized design (CRD):** all experimental units are allocated at random among the different treatments.



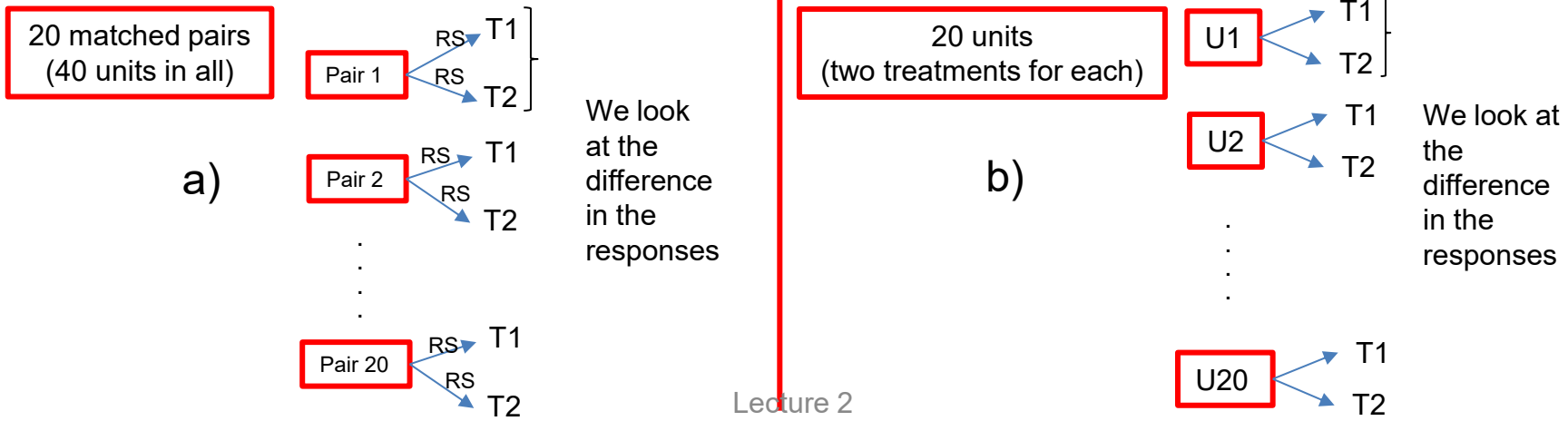
2. **Randomized block design:** the units are separated into blocks based on similarities before performing the random assignment to all the treatments.



3. Matched pairs design: Two possibilities:

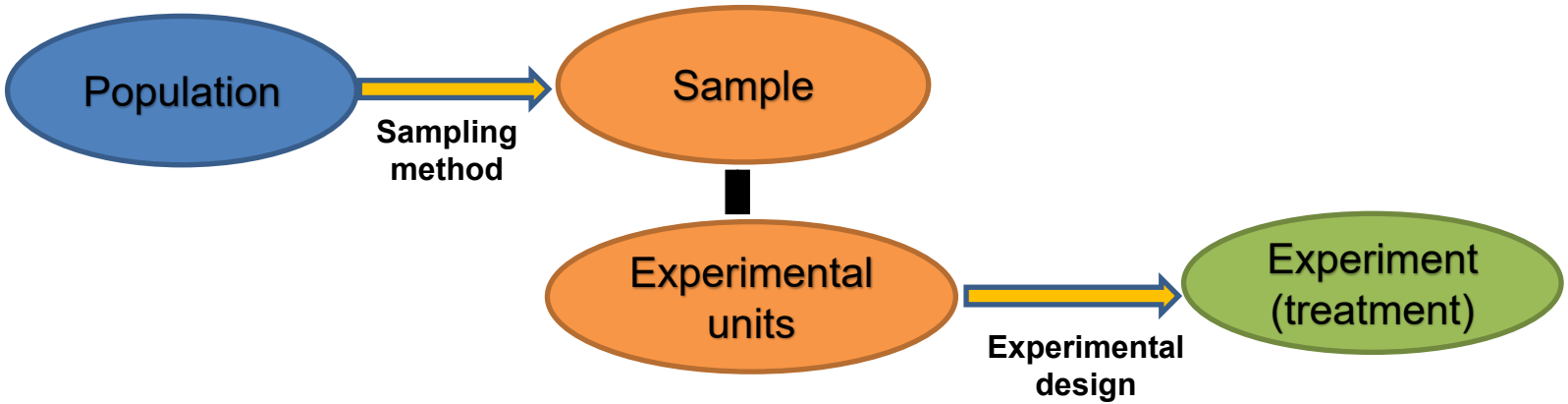
a) It compares 2 treatments; the experimental units are matched in pairs based on similar characteristics; the units in each pair are randomly allocated between the 2 treatments.

b) It compares 2 treatments; each experimental unit receives both treatments. Example: 2 results on each experimental unit, one result previous to a treatment and the other obtained after a treatment.



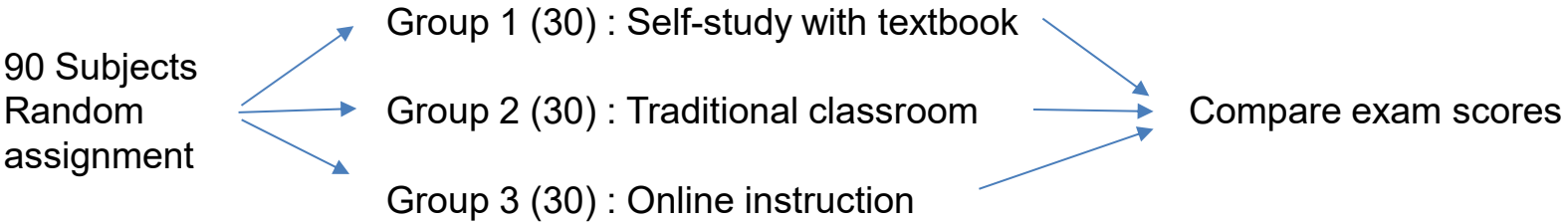
Do not confuse sampling method with experimental design!!

The sampling procedure comes before the assignment of the experimental units to the treatments. Then the sample will form or become all the experimental units to which the experiment is applied.



Example of Completely Randomized Design CRD:

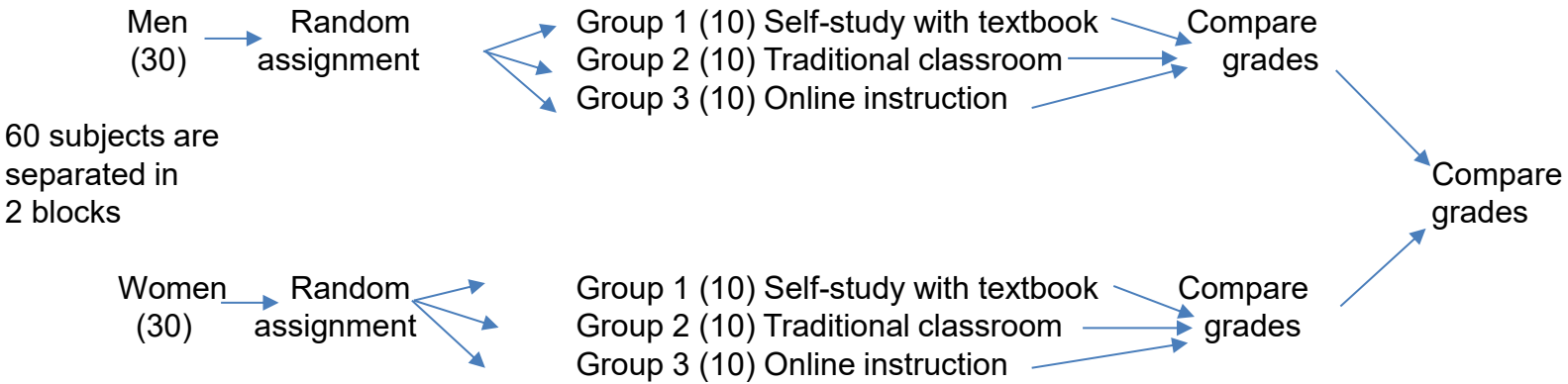
What instructional method helps students learn new material better: self-study from the textbook, traditional classroom, online instruction. A sample of 90 students are randomly assigned to one of these three groups (treatments). Scores on an exam are measured to compare these methods of teaching.



- What are the experimental units?
- Treatment?
- Factors?
- Levels? How many?
- Response variable?

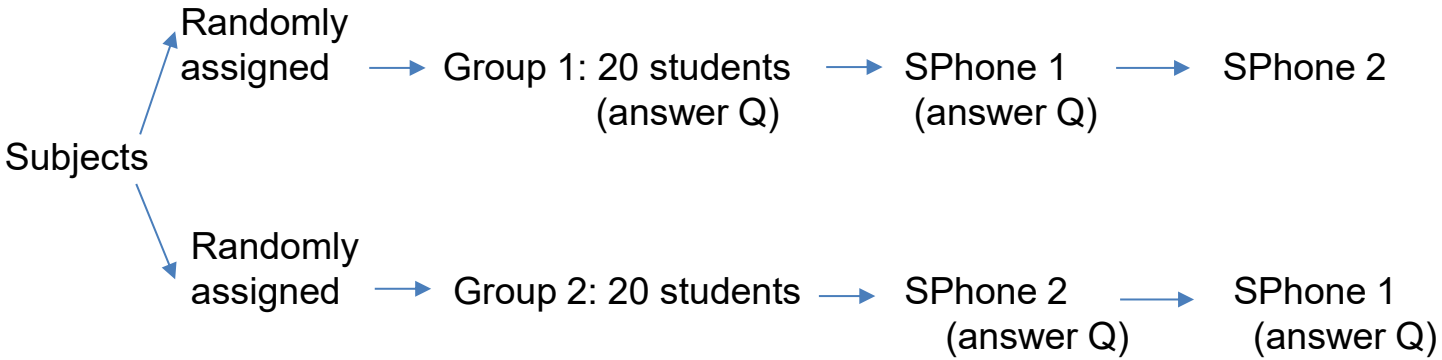
Example of Randomized Block Design:

Men and women may have different learning styles. Does this affect the instructional method that is most effective? A sample of 30 men and 30 women are assigned randomly to the three instructional methods from the previous example. An exam is given to determine the amount of material retained.



Example of Matched Pairs Design:

Two teams have each prepared a prototype for a new smartphone. Before deciding which one will be marketed , the smartphones will be evaluated by college students. Forty students will try the smartphones. Each student will use each phone for two weeks, and then answer some questions about how did they like the phone. Twenty students will be chosen randomly to evaluate Phone 1 first and Phone 2 second. The other 20 will evaluate Phone 2 first and Phone 1 second.



At the end we compared the answers to see which smartphone should be marketed.

A study is conducted to see if reading before bedtime decreases the time it takes to fall asleep for college students in the US. To select the participants, researchers first select 10 US colleges randomly and then, from each college, they randomly select 10 students. After that, from the 100 participants, 50 students are randomly chosen to sleep without any reading while the other 50 students read 30 min. before sleep. The time to fall asleep is measured for each student.

1. The population is:
    a. The 100 students that participate in the study.
    b. All college students in the US.
    c. All the books read by the students in the study.
    d. All young adults in the US.
2. The sampling method is:
    a. Voluntary
    b. SRS
    c. Stratified
    d. Multistage
3. The experimental design is:
    a. Completely Randomized
    b. Randomized Block
    c. Matched Pairs
    d. The study is not an experiment
4. The response variable is:
    a. The minutes the students read before bedtime.
    b. The number of students participating in the study.
    c. The time to fall asleep.
5. The researchers let the students choose any type of book they wanted to read. This could be labeled as:

a. Response bias
b. Lurking variable
c. Lack of realism
d. Non response

Lab 1

- First download R at https://www.r-project.org/
- Second, download RStudio at https://posit.co/download/rstudio-desktop/
- Set working directory in RStudio
- Move your gdp.csv file to the working directory
- Read the data in using **gdp<-read.csv("gdp.csv")**
- Let's go into RStudio and I'll show you how to get started.
- And your TA (Makenna Meyer) used R for a living before coming back to grad school! 😉

Chapter 1.2

Pictorial and Tabular Methods in Descriptive Statistics

A statistical study starts with the collection of data. After we decide what objects or **cases** we want to study, **we record the information** in that we call variables (**response variables**).

Two kinds of variables:

Qualitative: It is a word. It places the objects we study (cases) into one of several categories (eye color, nationality, college major).

Quantitative: It is a number. It takes numerical values for which calculations like adding or averaging make sense (age, height, annual salary, etc.). It has a unit of measurement associated to it (sec., $, ft., km.).

Examples:

- What letter grade did a student get into a math class?
- How many votes were cast in the last presidential election?
- How many songs were released last year?
- Which song was the most downloaded last year?

After collecting the data we examine them → Plot, Graph

For each response variable we have values or categories. We can look at each value or category and count how many times it appears (**frequency**).

Distribution

Distribution of a qualitative variable: lists the categories and gives the frequency (as the count or as the percent) of the cases that fall in each of them.

Two types of graphs are useful to display the distribution of a **categorical variable**: **Bar graph & Pie chart**

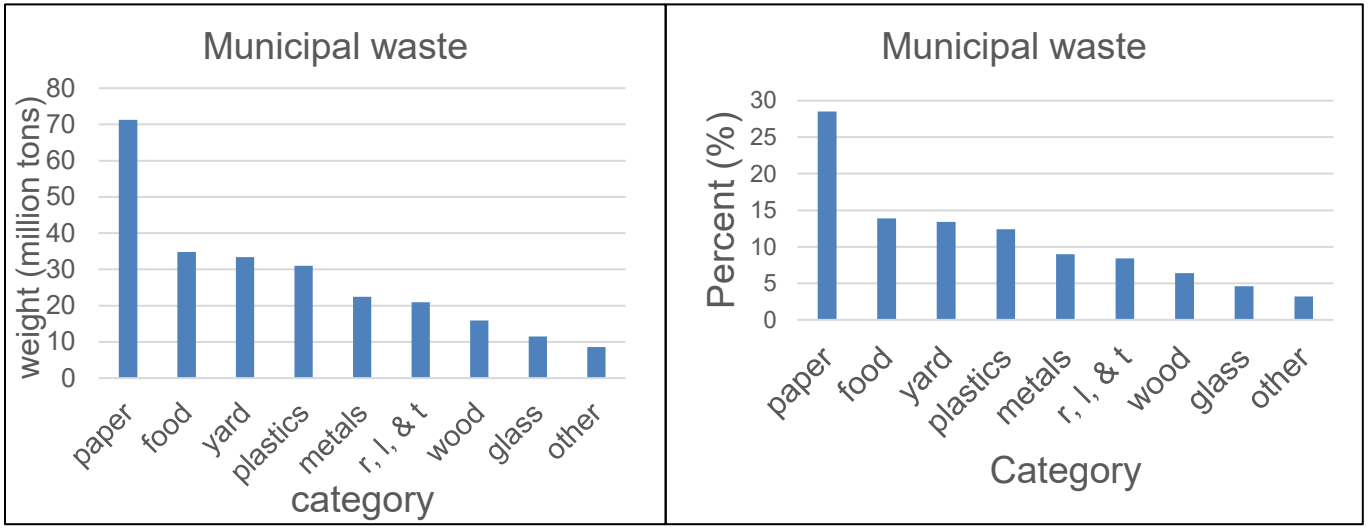
Example:

1.34 The formal name for garbage is “municipal solid waste”. Here is a table that shows the breakdown of the materials that make up American solid waste:

Material	Weight (million tons)	Percent of total (%)
Paper	71.3	28.6
Food scraps	34.8	13.9
Yard trimmings	33.4	13.4
Plastics	31.0	12.4
Metals	22.4	9.0
Rubber, leather, textiles	20.9	8.4
Wood	15.9	6.4
Glass	11.5	4.6
Other	8.6	3.2
Total	249.6	100.0

$$\frac{71.3}{249.6} \times 100$$

Bar Graph:



- Horizontal axis – list the categories (values of the variable)
- Vertical axis – frequency (count or %)
- Height of bars = frequency
- Bars do not need to touch each other
- Variables do not need to appear in any given order

Quantitative Variables

- Useful graphs to represent the distribution of a quantitative variable are:
  - Stem-and-Leaf Displays
  - Dotplots
  - Histograms
  - Boxplots (see section 1.4)

Stem-and-Leaf Displays (or stemplots): Separates each observation (or number) into a stem consisting of all but the rightmost digit of the number, and a leaf with the rightmost digit. It is useful for small data sets.

Example: HW time

During a school week, 30 2nd grade students were asked to record the time spent doing HW each day. At the end of the school week an average daily time (in minutes) spent doing HW was calculated. Results are shown below.

10	15	15	21	21	21	21	21	23	23
25	25	25	25	22	22	24	20	20	20
20	27	27	28	29	30	30	35	40	42

Stemplot of HW time

1. Order your numbers from least to greatest

10	15	15	20	20	20	20	21	21	21
21	21	22	22	23	23	24	25	25	25
25	27	27	28	29	30	30	35	40	42

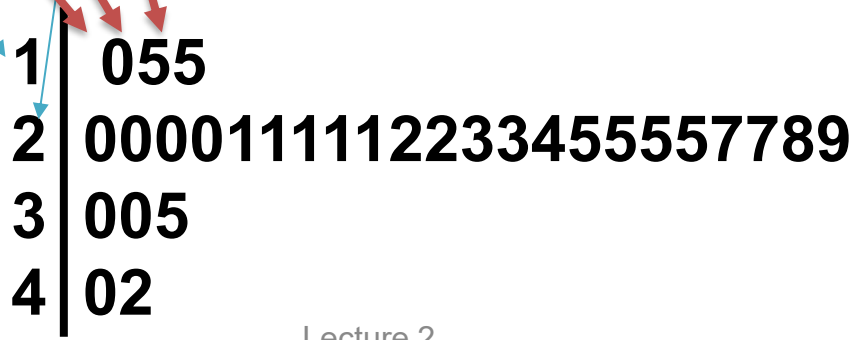
2. Separate each number into the stem (all but the ones digit) and the leaf (ones digit) 10: 1-0; 27:2-7; 42:4-2

How about 1235?

How about -257?

3. Build the stemplot as follows:

10	15	15	20	20	20	20	21	21	21
21	21	22	22	23	23	24	25	25	25
25	27	27	28	29	30	30	35	40	42



Stemplots. ***In Exercises 13 and 14, construct the stemplot.***

13. Car Crash Tests Refer to Data Set 13 in Appendix B and use the 21 pelvis (PLVS) deceleration measurements from the car crash tests. Is there strong evidence suggesting that the data are *not* from a population having a normal distribution?

CAR	PLVS
Mitsubishi Lancer	45
Subaru Impreza	53
Nissan Altima	53
Dodge Charger	53
Honda Accord	55
Toyota Avalon	55
Toyota Camry	57
Nissan Maxima	59
VW Passat	61
Lincoln Town	61
Merc Gr Marq	61
VW Jetta	66
Volvo S60	67
Chev Aveo	71
Honda Civic	71
Hyundai Elantra	71
Hyundai Azera	75
Cadillac DTS	75
Buick Lucerne	76
Ford Fusion	78
Kia Rio	84

Dotplots

- A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale.

- When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Dotplot Example

- There is growing concern in the U.S. that not enough students are graduating from college. America used to be number 1 in the world for the percentage of adults with college degrees, but it has recently dropped to 16th. Here is data on the percentage of 25- to 34-year-olds in each state who had some type of postsecondary degree as of 2010 (listed in alphabetical order, with the District of Columbia included):

31.5 32.9 33.0 28.6 37.9 43.3 45.9 37.2 68.8 36.2 35.5
40.5 37.2 45.3 36.1 45.5 42.3 33.3 30.3 37.2 45.5 54.3
37.2 49.8 32.1 39.3 40.3 44.2 28.4 46.0 47.2 28.7 49.6
37.6 50.8 38.0 30.8 37.6 43.9 42.5 35.2 42.2 32.8 32.2
38.5 44.5 44.6 40.9 29.5 41.3 35.4

A dotplot can be quite cumbersome to construct and look crowded when the number of observations is large. Our next technique is well suited to such situations.

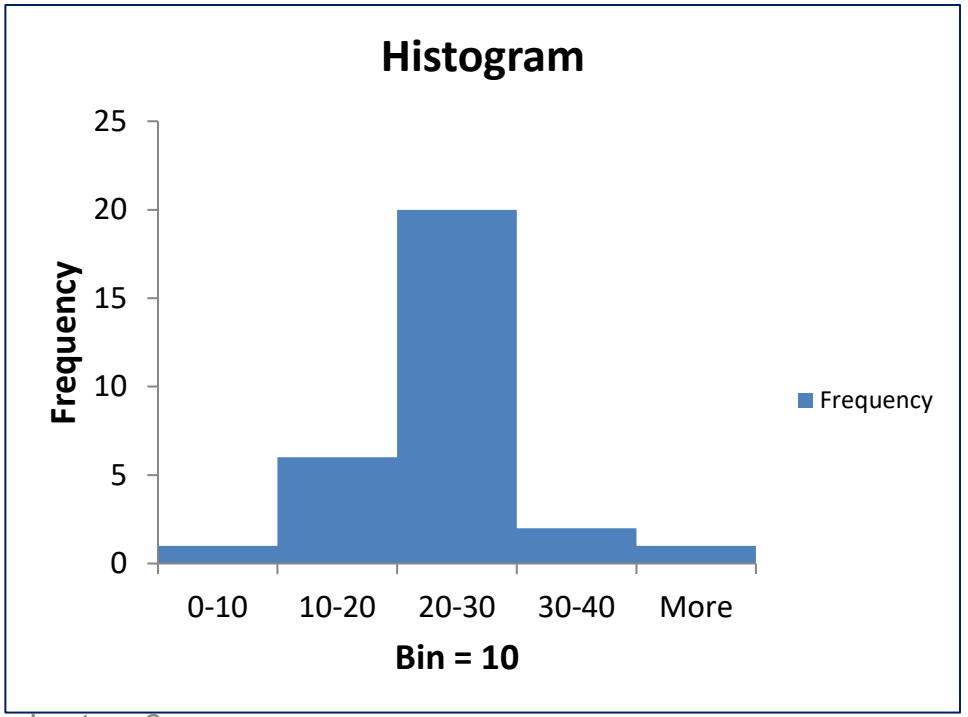
Figure 1.6 A dotplot of the data from Example 1.8

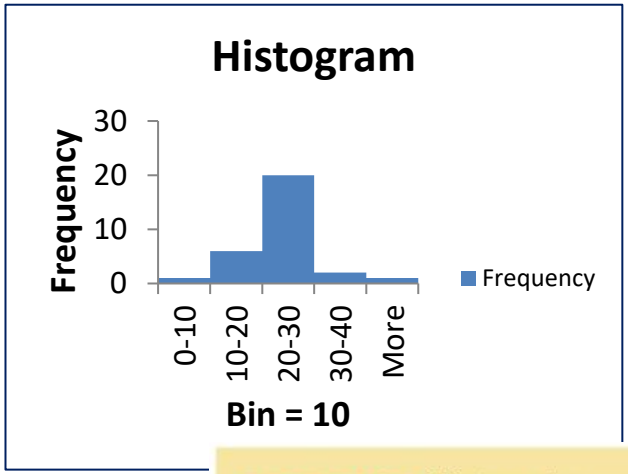
Histogram of HW time

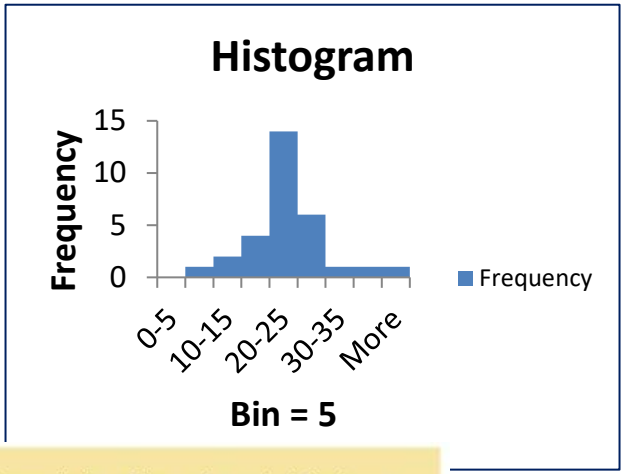
Common practice:
- left-inclusive
- right-exclusive

10	15	15	20	20	20	20	21	21	21
21	21	22	22	23	23	24	25	25	25
25	27	27	28	29	30	30	35	40	42

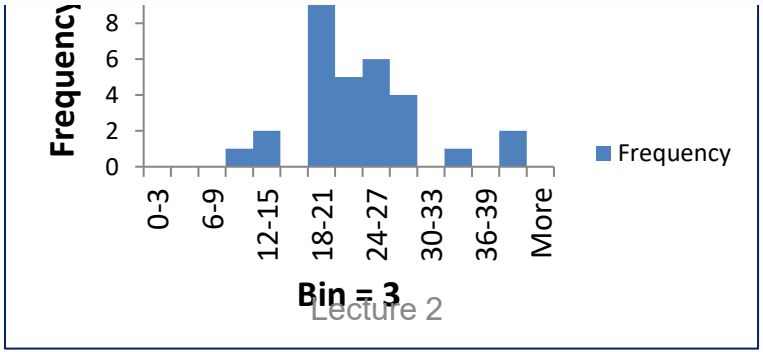
- **Horizontal axis:** continuous range of values for the variable broken into bins
- **Vertical axis:** frequency (count or %) corresponding to different bins
- **Display:** vertical bars for each bin, touching each other
- **Bins:** choose any convenient number of bins but all of equal width

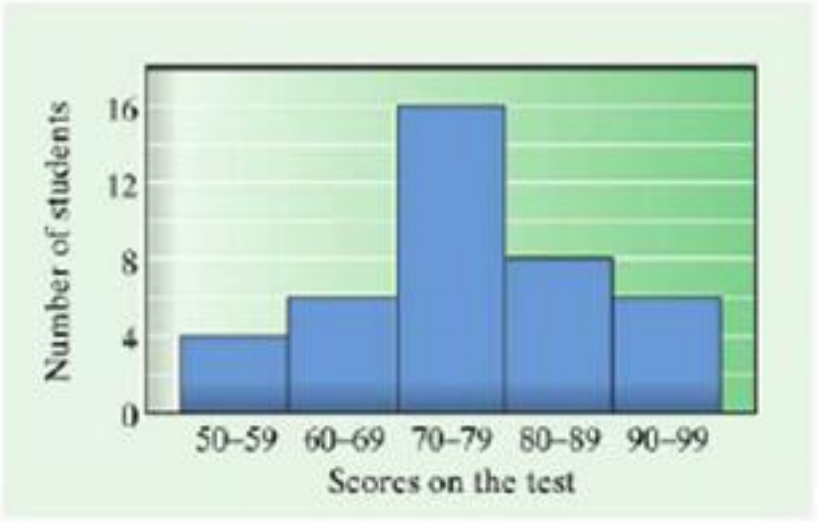






DEFINITION A **histogram** is a graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data). The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to the frequency values.





1) How many people made a 70 or higher on this test?
2) How many people took the test?
3) What percentage of students made an A?

Constructing Histograms

- Constructing a histogram for continuous data (measurements) entails subdividing the measurement axis into a suitable number of class intervals or classes, such that each observation is contained in exactly one class.

Constructing a Histogram for Continuous Data: Equal Class Widths

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).

- Let's do one together using SALT (HW 1, #6)

Interpreting Histograms

The overall pattern is determined by:

I. **Shape** → number of peaks, symmetry

II. **Center** → mean, median, mode

III. **Spread** → range, interquartile range, P^{th} percentile, 5 number summary (boxplots), variance, standard deviation

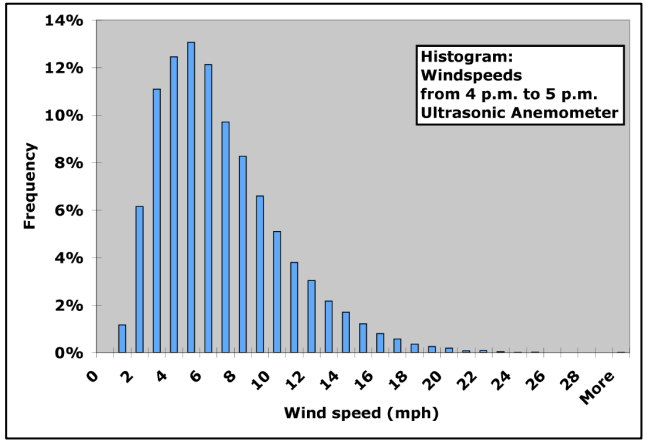
What is an outlier?

Outlier: **It is an observation that lies outside the overall pattern.**

An outlier can be dropped from the data if you discover it is an error. Otherwise, you keep it and use resistant measures in your analysis so that the outliers don't influence too much.

Shape

- Look at the number of major peaks (or modes) in the distribution: 1 peak = unimodal; 2 peaks = bimodal; more than 2 peaks = multimodal.



http://patarnott.com/

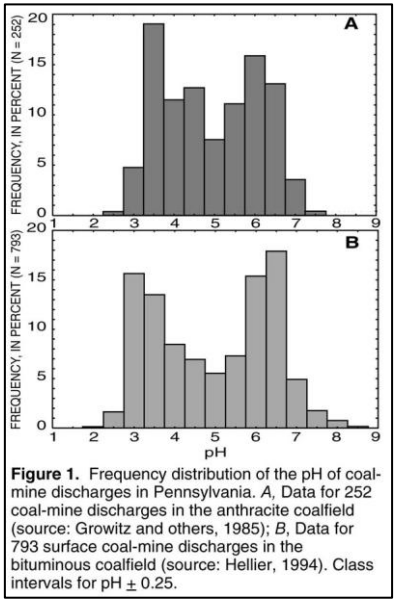
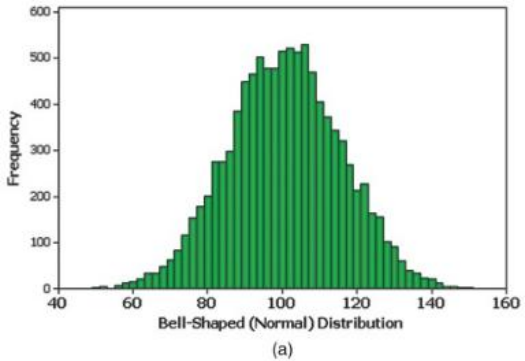


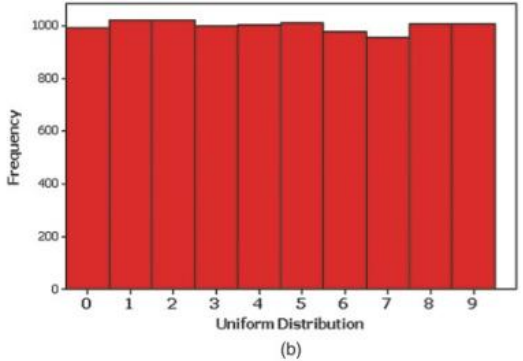
Figure 1. Frequency distribution of the pH of coal-mine discharges in Pennsylvania. *A*, Data for 252 coal-mine discharges in the anthracite coalfield (source: Growitz and others, 1985); *B*, Data for 793 surface coal-mine discharges in the bituminous coalfield (source: Hellier, 1994). Class intervals for pH ± 0.25.

maps.unomaha.edu

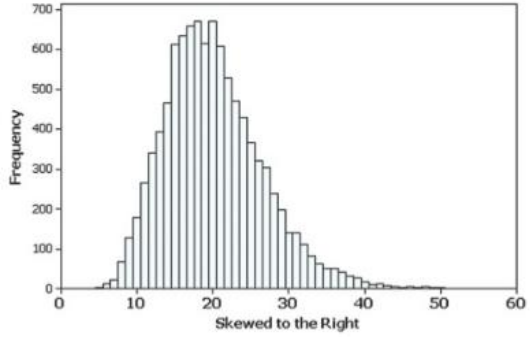
- Look at the symmetry. A distribution can be **symmetric**, **right** (positively) **skewed** (long tail to the right), or **left** (negatively) **skewed** (long tail to the left).



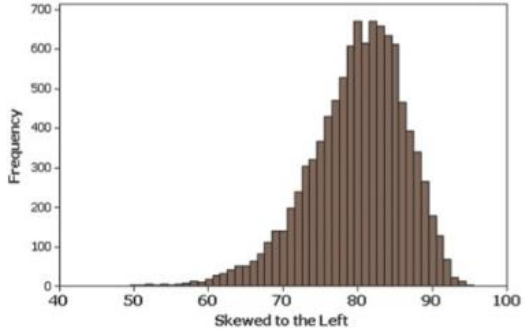
(a)



(b)



(c)



(d)

Chapter 1.3

Measures of Location

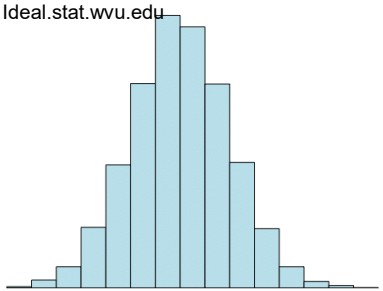
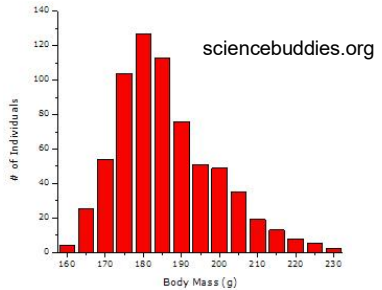
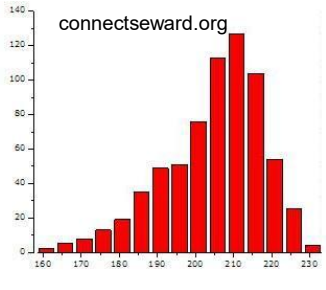
Center

- To begin the numerical description of a distribution we need a measure of its center → mean, median, and mode (the mean and the median are the two most common)
  - **Mean (average, $\bar{x}$):** **It is the sum of all the observations divided by the number of observations.**

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

  - **Median ($\tilde{x}$):** **It is the midpoint of the distribution. If n is the number of data→(n+1)/2 : center. If n is odd, $\tilde{x}$ is the observation at the center; If n even, $\tilde{x}$ is the average of two center observations.**
  - **Mode:** **The highest peak of the distribution, the value of the response variable with the highest frequency (the most common number).**

How does the shape of the distribution relate to the values of the mean and the median?

Symmetric	Positively skewed (long tail on right)	Negatively skewed (long tail on left)
		
Median ≈ Mean **Mode ≈ Mean**	**Median < Mean** **Mode < Mean**	**Median > Mean** **Mode > Mean**

Important: **When the distribution is skewed, the median is a better measure of the center of the distribution. It is more resistant to the influence of outliers!**

Trimmed Mean

- A trimmed mean is a compromise between $\bar{x}$ and $\tilde{x}$. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.
- If the desired trimming percentage is 100α% and nα is not an integer, the trimmed mean must be calculated by interpolation.

Example

- The total number of visits to Bojangles in a year for a sample of 10 people is given by:

  0, 3, 5, 7, 8, 10, 10, 11, 26, 67

- Mean, $\bar{x}$ = (0 + 3 + 5 + 7 + 8 + 10 + 10 + 11 + 26 + 67)/10 = 14.7
- Median, $\tilde{x}$ = (8 + 10)/2 = 9
- Notice how the 67 skews the measurement for mean…
  - mean > median
  - What kind of skew is this?

Example (continued)

- The total number of visits to Bojangles in a year for a sample of 10 people is given by:

0, 3, 5, 7, 8, 10, 10, 11, 26, 67

- 20% trimmed mean = 100*(2/10) → remove the two smallest and two largest observations ➔

- 12% trimmed mean = use 20% trimmed mean and 10% trimmed mean ($\bar{x}_{tr(10)}$=10) and interpolate.

$$\frac{20-10}{8.5-10}=\frac{20-12}{8.5-x}$$

Chapter 1.4

Measures of Variability

Spread

The measure of center alone is misleading. Example: Two countries might have the same median family income, but one could have extreme wealth and poverty.

Or simply consider the data set 20, 30, 40 versus the data set 29, 30, 31. They both have an average of 30 but their spread is very different.

We need to have measures for the spread.

- **Range:** **It is the difference between the maximum observation and the minimum observation.**

  Set of data: 12, -5, 6, 2, -8, 7, 5, -1, 1, 9

  We order them from least to greatest: -8, -5, -1, 1, 2, 5, 6, 7, 9, 12

- **Interquartile range:** **IQR = Q3 - Q1**

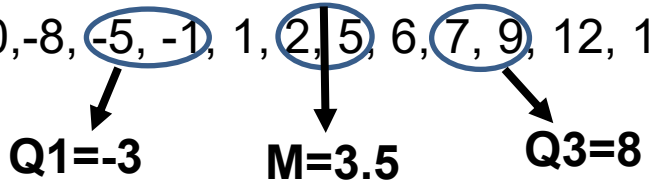
  Q1: **first quartile; median of the lower half of the data set**

  Q3: **third quartile; median of the upper half of the data set**

  -8, -5, -1, 1, 2, 5, 6, 7, 9, 12

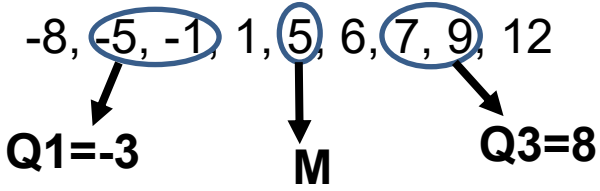
Be careful when finding Q1 and Q3!

If the number of data is multiple of 4: -10,-8, -5, -1, 1, 2, 5, 6, 7, 9, 12, 15 (even)

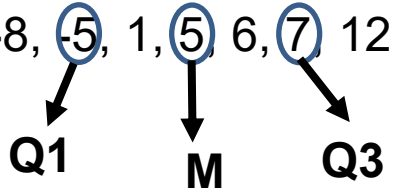


With an odd number of data: -8, -5, -1, 1, 5, 6, 7, 9, 12

Keep $\tilde{x}$ aside to count the observations in the lower and upper half!



or: -8, -5, 1, 5, 6, 7, 12



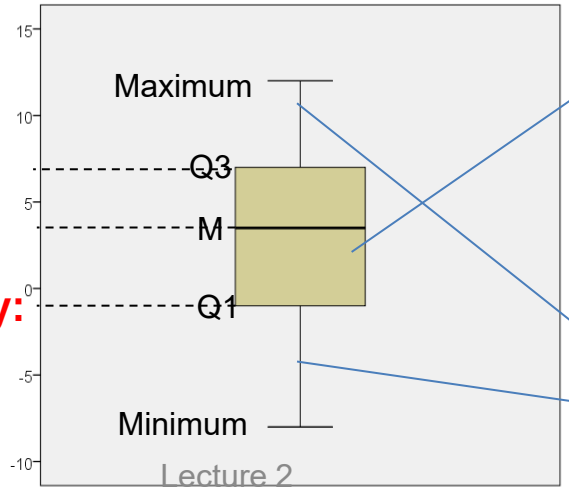
Five Number Summary:

Minimum Q1 M Q3 Maximum

We can build a **Boxplot** with these numbers:

-8, -5, -1, 1, 2, 5, 6, 7, 9, 12

Minimum = -8
Q1 = -1
M= 3.5
Q3 = 7
Maximum = 12
Five number summary:
-8, -1, 3.5, 7, 12



How do we know if an observation is an outlier?

1.5 x IQR Rule for Outliers: Call an observation an outlier if it falls more than 1.5 x IQR above Q3 or below Q1.

S: observation if S
- **< Q1 - 1.5 IQR** → S is a low outlier
- **or**
- **> Q3 + 1.5 IQR** → S is a high outlier

Example: -8, -5, -1, 1, 2, 5, 6, 7, 9, 12

Q1 = -1 Q3 = 7 IQR = Q3 - Q1 = 8 1.5(IQR) = 1.5 · 8 = 12

Outliers are observations that are less than Q1 - 1.5 IQR = -1 - 12 = -13; or greater than Q3 + 1.5 IQR = 7 +12 = 19 → **no outliers in this data set!**

In the same way as the median is more resistant to outliers than the mean, The interquartile range (IQR) is more resistant to outliers than the range.

Example: -8, -5, -1, 1, 2, 5, 6, 7, 9, 12
M = 3.5 ; Range = 12- (-8) = 12+8 = 20
IQR = Q3 – Q1 = 7 – (-1) = 7+1 = 8

Suppose we add one clear outlier: -8, -5, -1, 1, 2, 5, 6, 7, 9, 12, 70
Careful! Now we have 11 data, so: M=5; Q1 = -1; Q3 = 9

Range = 70 – (-8) = 78
IQR = 9 – (-1) = 10

70 is a clear outlier: 1.5 IQR = 1.5 · 10 = 15; Q3 + 1.5 IQR = 9+15 = 24
Anything greater than 24 is classified as high outlier!

The range changed by a factor of 3.9 (78/20=3.9) while the IQR changed by a factor 1.25 (10/8=1.25) → **the IQR is considerably better!**

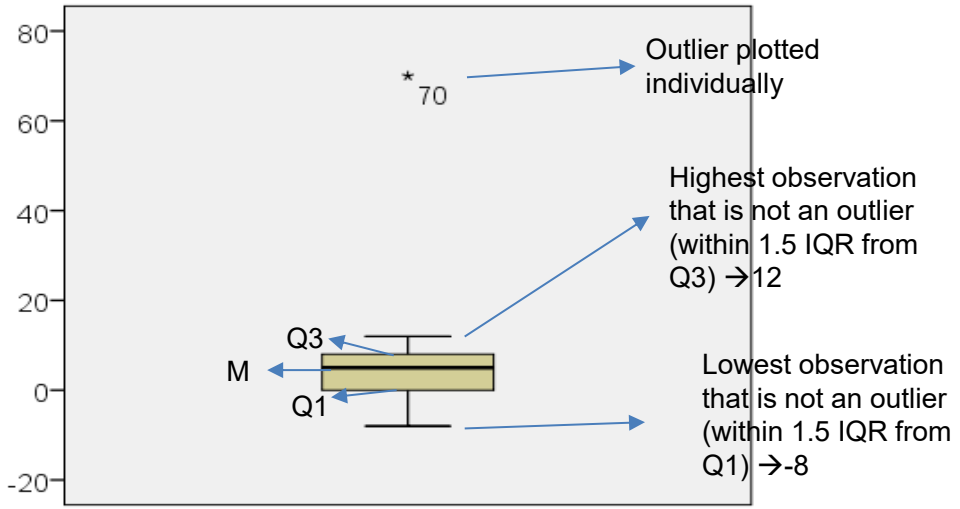
To see outliers clearly → Modified Boxplot

Modified Boxplot: Instead of having whiskers that extend to the Maximum and Minimum values, we have whiskers that extend up to the observations that are within 1.5 IQR from the quartiles. The outliers are then plotted outside individually.

Example:

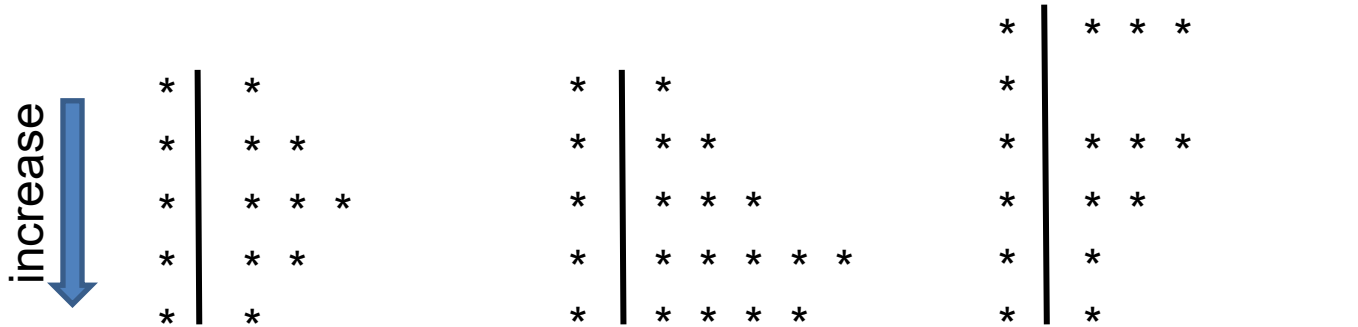
-8, -5, -1, 1, 2, 5, 6, 7, 9, 12, 70

M = 5
Q1 = -1
Q3 = 9
IQR = 9-(-1) = 10
1.5 IQR = 1.5 · 10 = 15
Q1-1.5IQR = -1 -15 = -16
Q3+1.5 IQR = 9+15 = 24
Whiskers: from -8 to -1, and
from 9 to 12

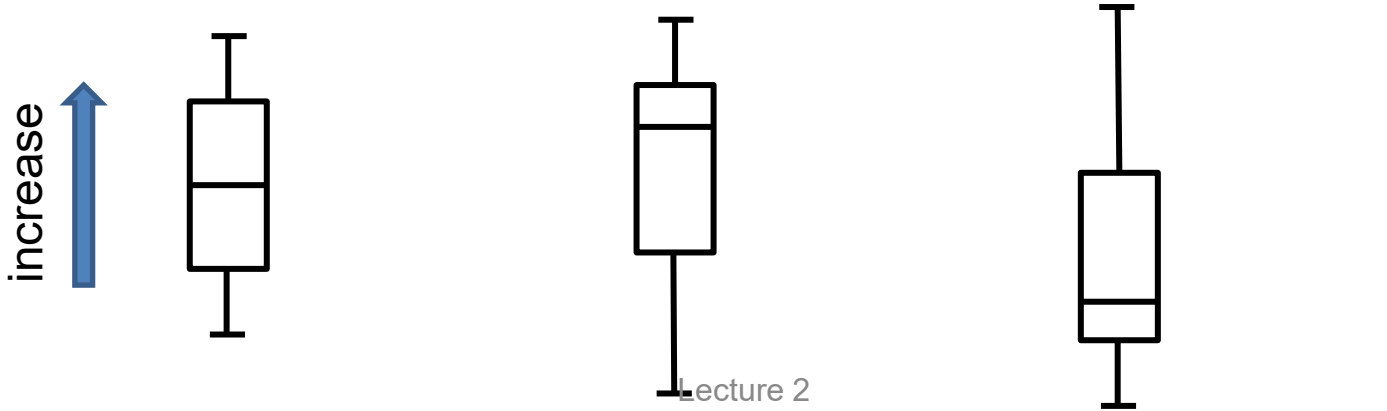


Just look at the box and whiskers to define symmetry!

Looking at symmetry with stemplots and boxplots

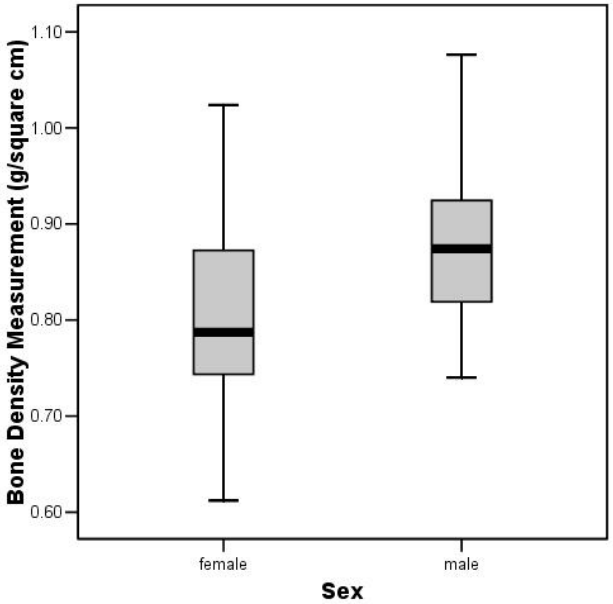


Symmetric **Negatively Skewed** **Positively Skewed**



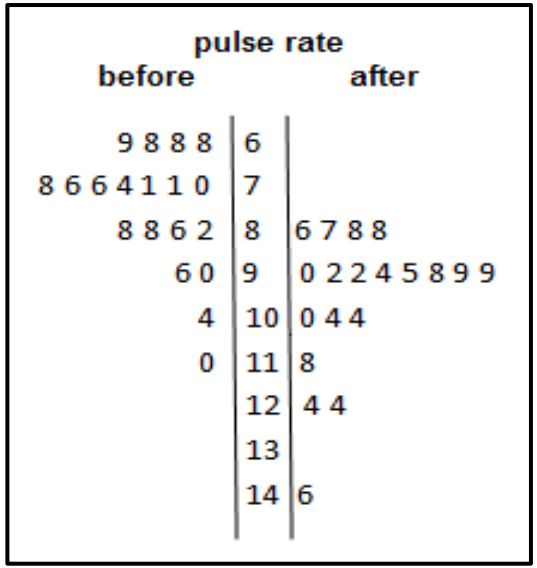
Comparing 2 or more groups

Side-by-side Boxplot



Web.anglia.ac.uk

Back-to-back



www.australiancurriculum.edu.au

Another measure of the spread: **Standard deviation**

Standard deviation: Measures the spread by looking at how far the observations are from the mean.

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

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

If the observations are widely spread from the mean → standard deviation is large

Variance = (standard deviation)^2
OR
Standard deviation = $\sqrt{variance}$

Properties of the Standard Deviation

- S measures the spread about the mean and should be used only when we choose the mean as a measure of center of the distribution.
- S = 0 only when all the observations have the same value. Otherwise, s > 0.
- The greater the spread, the larger the standard deviation.
- S, like the mean, is not resistant. A few outliers can make s very large.

How do we choose the measures of center and spread?

For skewed distributions or with strong outliers → **Five number summary: Min, Q1, $\tilde{x}$, Q3, Max**

Resistant $\tilde{x}$, IQR

For symmetric distributions without outliers → $\bar{x}$, S

Not Resistant