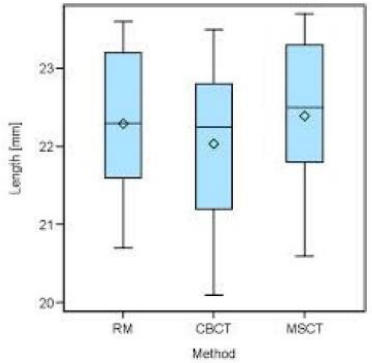
L02: Descriptive Statistics

Prof. Yiming QIN
Assistant Professor
School of Energy and Environment
City University of Hong Kong

Two Branches of Statistics

1. **Descriptive statistics**
   - Data collection, data summarization, and data presentation.



2. **Inferential statistics**
   - Using sample data to draw conclusions about a population.

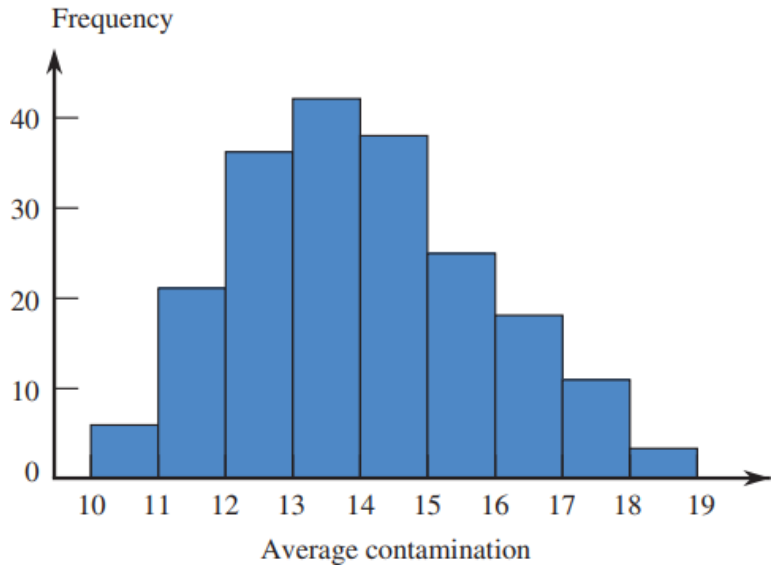
Descriptive Procedure

- **Description of the basic features of the data in our study:**
  - Provides a summary about the sample.
  - Use different sorts of tools such as the sample mean, median, quartiles, variance, etc.

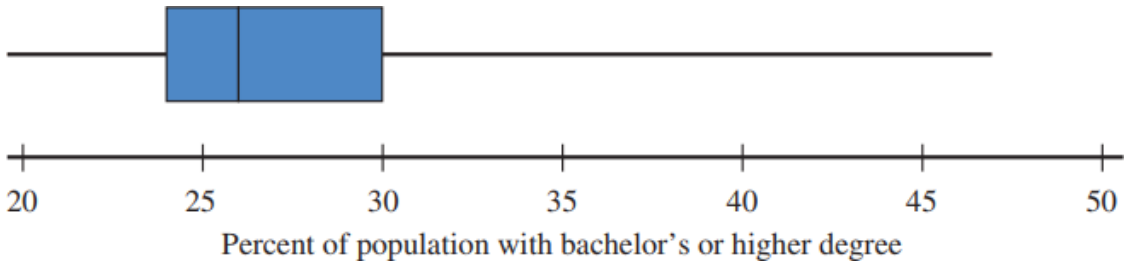
- **Methodologies used:**
  - Graphical: use graphs like histogram and boxplot to summarize data.
  - Tabular: use tables like frequency table to summarize data.
  - Numerical: use certain values to summarize data.

- **Steps:**
  - Collect data, e.g. surveying and sampling.
  - Classify data, e.g. grouping.
  - Characterize data, e.g. sample mean.
  - Present data, e.g. table and boxplot.

Descriptive Statistics Examples



Histogram



Boxplot

Big Mac Prices for 7 Countries

Country	Big Mac Price in U.S. Dollars
Argentina	3.02
Brazil	4.67
Chile	3.28
Colombia	3.51
Costa Rica	3.42
Peru	2.76
Uruguay	2.87

Descriptive Statistics: Big Mac Price in U.S. Dollars

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median
Big Mac Price	7	3.361	0.242	0.641	2.760	2.870	3.280

Variable	Q3	Maximum
Big Mac Price	3.510	4.670

What are Data?

Data (its singular form is datum) in science are a collection of measurements and identifiers. These can take many forms like numerical, character, or any other form of output.

For example:

Respondent	Age	Gender
2411	30	Male
2421	43	Female

numerical

identifier

character

What are Data?

Data summarization:

Consider the following data:

35, 42, 21, 59, 47, 55, 55, 38, 50, 41, 51, 44, 31, 42, 30, 32, 40, Yes, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, No, Yes, Yes, Yes

Note that without summarizing them, data are completely meaningless and non-informative.

What are Data?

Consider the following data extracted from medical records of 50 patients with low back pain:

Subject	Age	Gender	In employment?	Duration of pain	Severity of pain	First episode of pain
1	35	F	No	3 weeks	Mild	< 1 year
2	42	F	Yes	13 weeks	Severe	1-6 years
3	21	M	Yes	4 weeks	Moderate	< 1 year
4	59	F	No	72 weeks	Moderate	≥ 11 years
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	40	M	Yes	30 weeks	severe	6-11 years

- Note that in this table each row represents all measurements for a ***SINGLE case/patient***.
- Each column refers to a characteristic of our interest which we often call a ***variable***.
- Each datum (the value of a single variable for a single case) is known as an ***observation***.

What is Variability?

Suppose you went into a convenience store to purchase a soft drink. Does every can on the shelf contain exactly 12 ounces?

What is Variability?

Suppose you went into a convenience store to purchase a soft drink. Does every can on the shelf contain exactly 12 ounces?

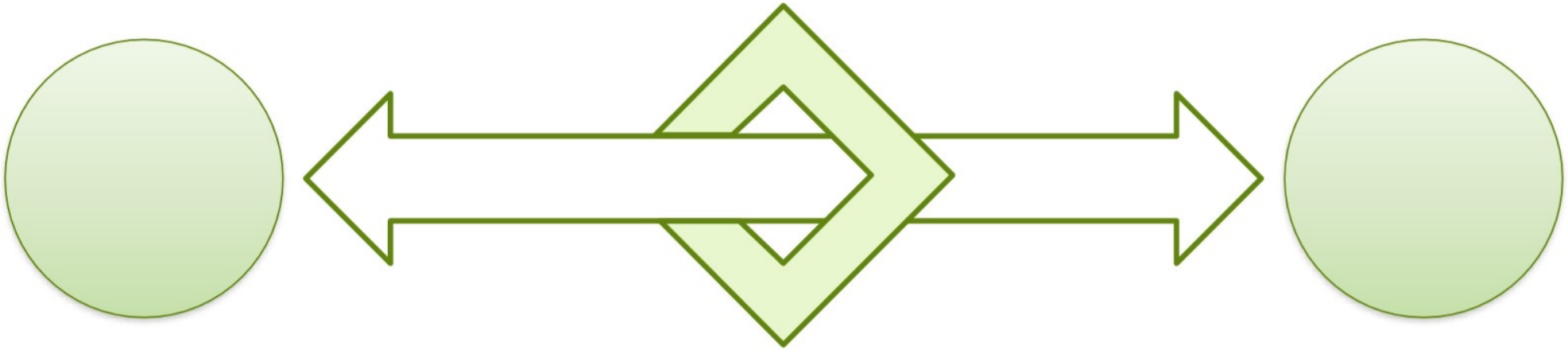
NO :

There may be a little more or less in the various cans due to the variability that is inherent in the filling process

It is variability that makes life interesting!!

CityU

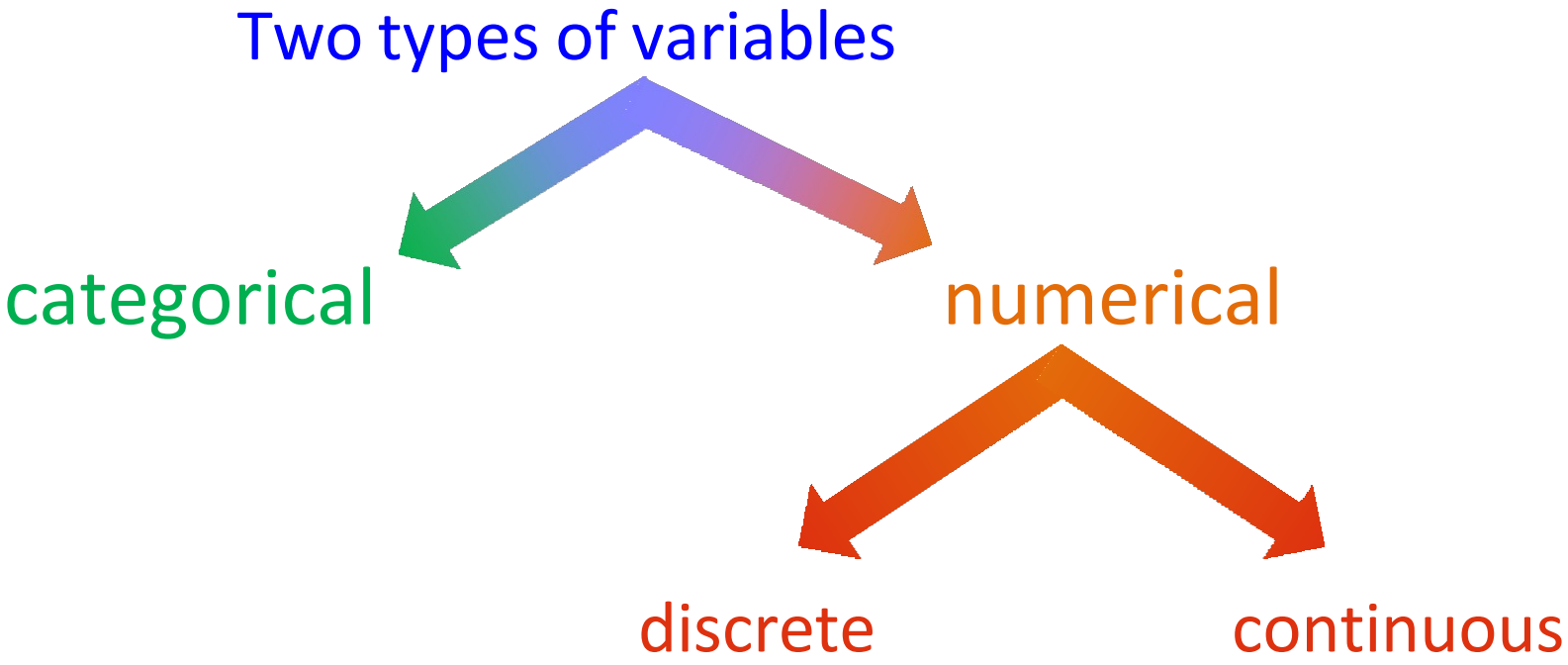
The Role of Statistics



The Data Analysis Process

1. Understand the nature of the problem
2. Decide what to measure and how to measure it
3. Collect data
4. Summarize data and perform preliminary analysis
5. Perform formal analysis
6. Interpret results

Types of Variables



Types of Variables

- **Categorical/ Qualitative**

  Questions that yield distinct categorical/label response.

  - **Nominal**: The distinct categories cannot be ranked.

  For instance, "Do you currently own any notebook computer" ➔ Yes/No

  - **Ordinal**: The distinct categories can be ranked.

  For instance, "Please rate the instructor overall" ➔ Very Bad, Bad, Satisfactory, Good, Very Good.

- **Quantitative**

  Questions that produce numerical response.

  - **Discrete:** The numerical responses arise from a counting process.

  For instance, "How many courses did you enroll?" ➔ 0, 1, 2, 3, …

  - **Continuous:** The numerical responses arise from a measuring process.

  For instance, "What is your height?" ➔ 168.5cm, 172.0cm, 185.4cm, …

CityU

Types of Variables

WHY SHOULD WE CARE ABOUT THE TYPE OF VARIABLE?

Note that in practice categorical variables are often recorded using numbers (e.g. yes = 1, no = 0). Please don't mistake these for quantitative variables.

Thus, knowing the type of variables can help

- Decide how to **interpret** the data from variable.
  For instance, if a measure is nominal, then the numerical values are just short codes for the longer names.

- Decide what **statistical analysis** is appropriate on the values that were assigned.
  For instance, if a measure is ordinal, averaging the data values is INCORRECT/ INVALID.

Classifying variables by the number of variables

Suppose that the PE coach records the height of each student in his class.

This is an example of a univariate data

Univariate - data that describes a single characteristic of the population

Classifying variables by the number of variables

Suppose that the PE coach records the height and weight of each student in his class.

This is an example of a bivariate data

Bivariate - data that describes two characteristics of the population

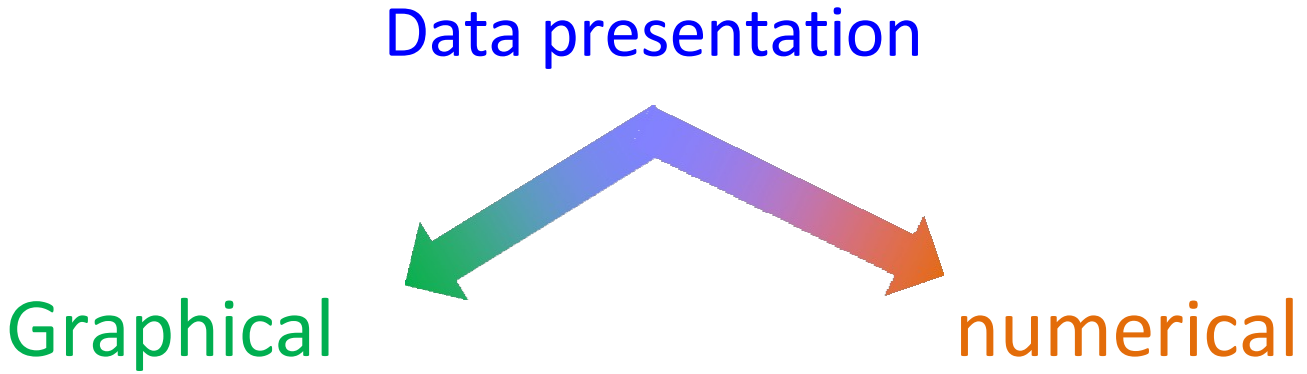
Classifying variables by the number of variables

Suppose that the PE coach records the height, weight, number of sit-ups, and number of push-ups for each student in his class.

This is an example of a multivariate data

Multivariate - data that describes more than two characteristics

Data Presentation



Data Presentation (Graphical)

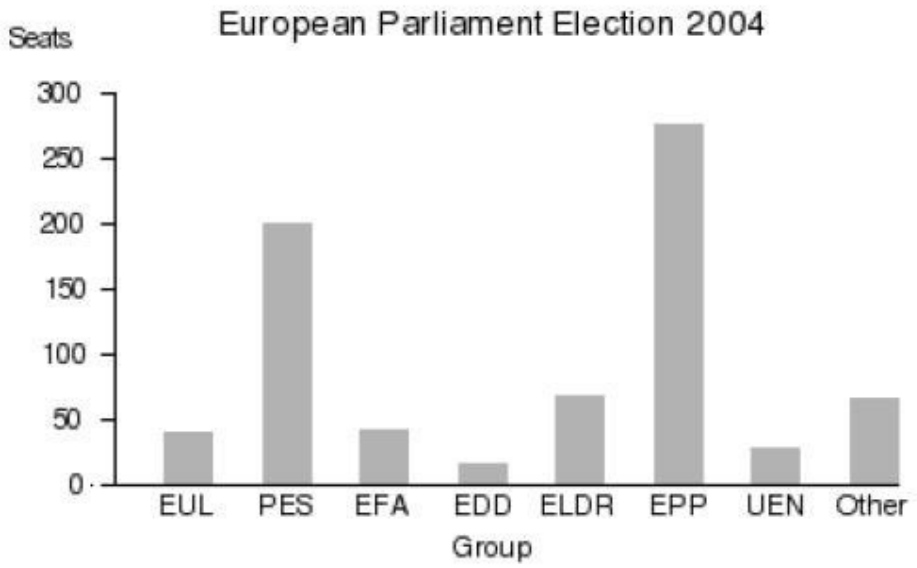
Frequency table

Mark	Tally	Frequency
4	II	2
5	II	2
6	IIII	4
7	𝍸	5
8	IIII	4
9	II	2
10	I	1

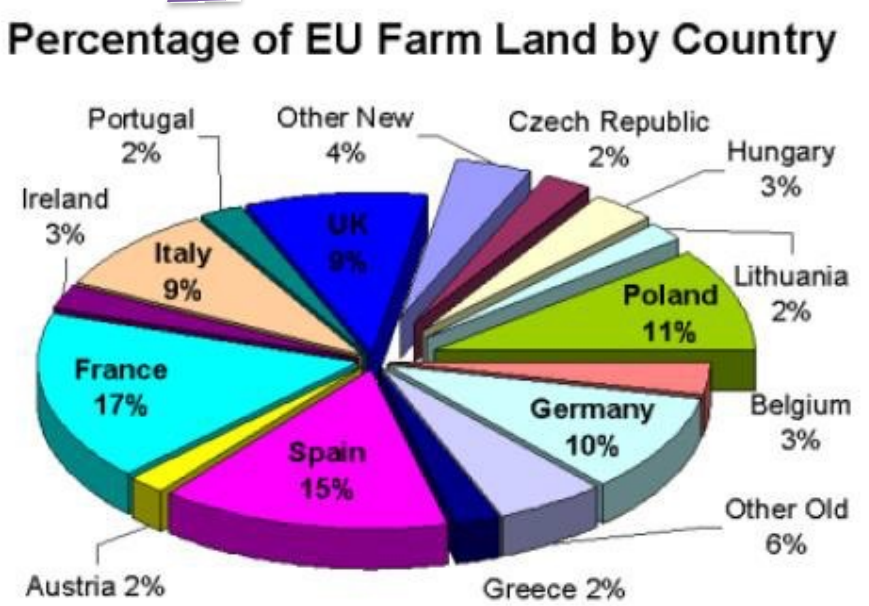
Presenting categorical data:

- **Frequency table**
- **Bar chart** (Similar to histogram)
- **Pie chart**: A pie is divided into slices according to the percentage in each category

Bar Chart



Pie Chart



Data Presentation (Tabular) - Frequency Table

A summary table in which the collected data can be arranged into numerically ordered and non- overlapping categories or class intervals so that we can condense the data into a more useful form and then allow for a quick interpretation of the data.

- The frequency of a particular DATA value is the number of times the data value occurs.
- The mode of the data is the data values with the highest frequency.
- The frequency of the CLASS interval is the number of data values in that class.

Procedure:

i. Selects an appropriate number of class intervals. Normally use 5 - 20 classes.
ii. Obtains a suitable class intervals by dividing the range of the data by the number decided at the previous step.
iii. Establishes the boundaries of each class to avoid overlapping.
iv. Construct a table with columns of the class interval, class midpoint, and class frequency.

Example

The marks awarded for an assignment set for a Year 8 class of 20 students were as follows:

6	7	5	7	7	8	7	6	9	7
4	10	6	8	8	9	5	6	4	8

Present this information in a frequency table.

Mark	Tally	Frequency
4	||	2
5	||	2
6	||||	4
7	$\cancel{\text{IIII}}$	5
8	||||	4
9	||	2
10	|	1

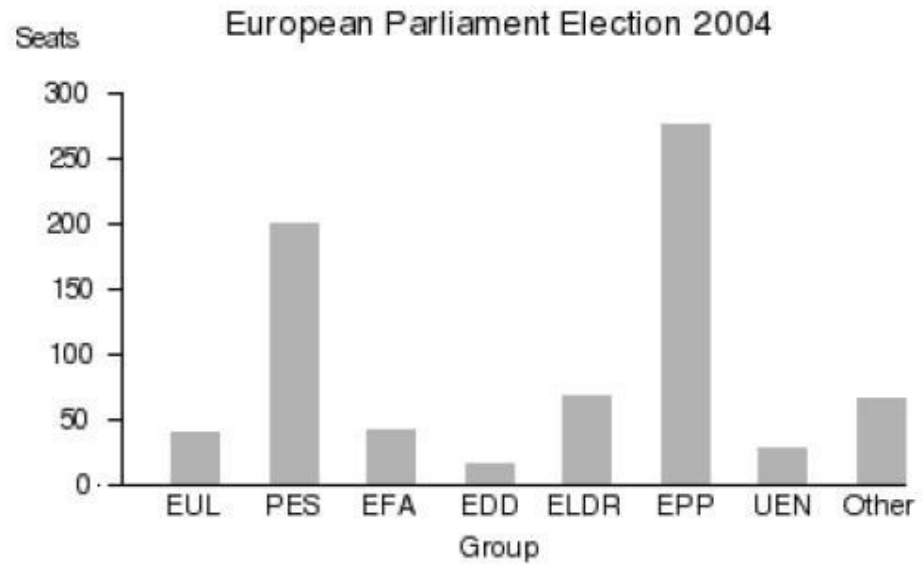
Data Presentation (Graphical) - Bar Chart

When to Use:

- Categorical data

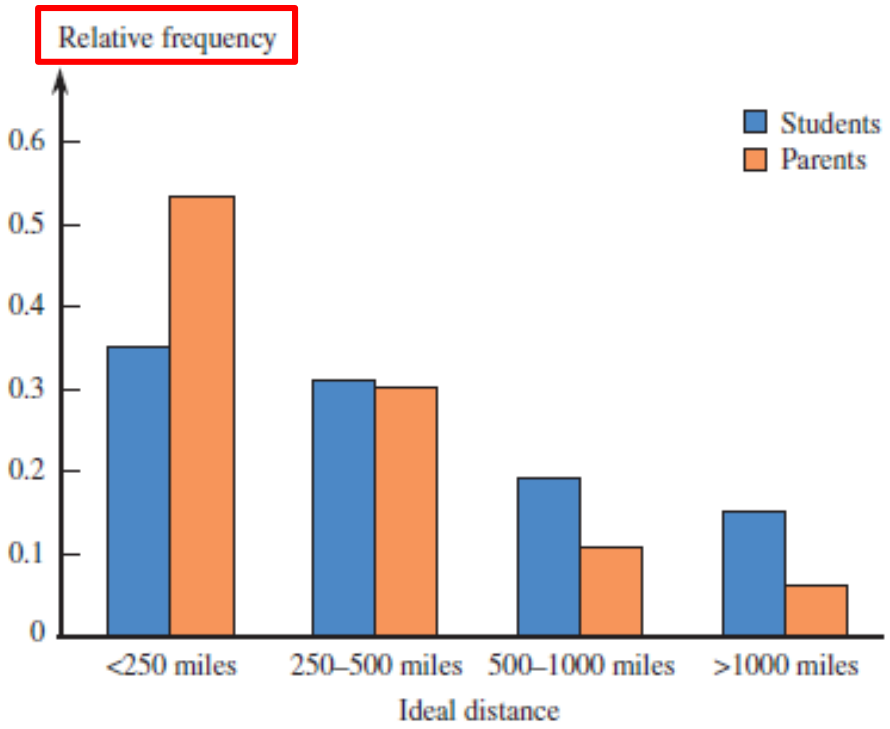
How to construct:

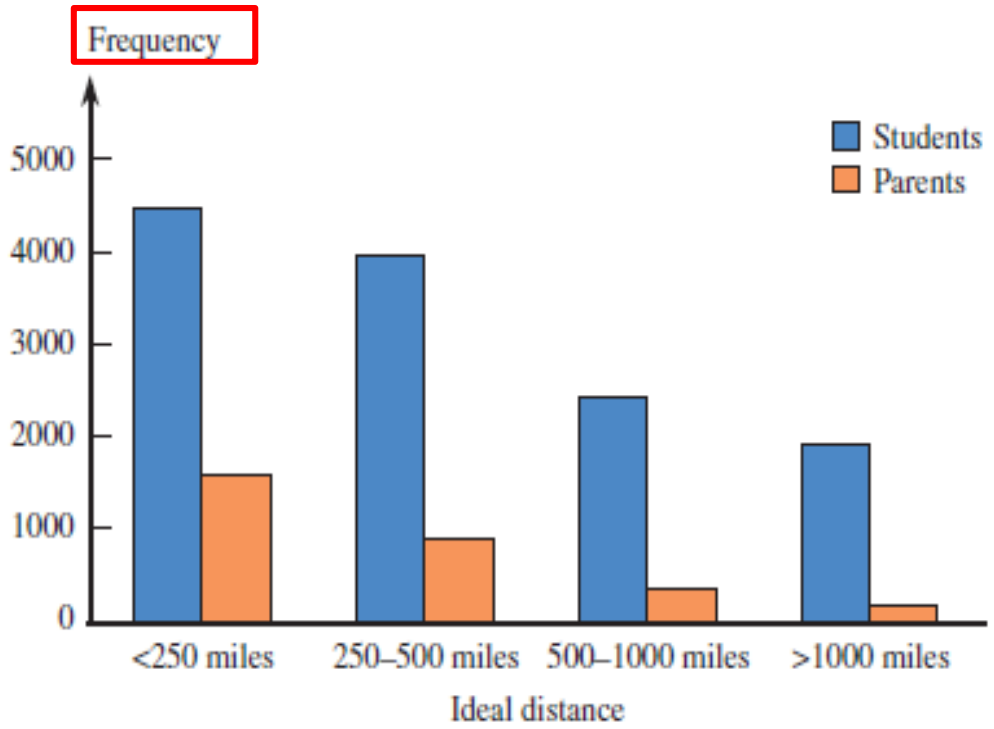
- Constructed like bar charts, but with two (or more) groups being compared
- MUST use relative frequencies on the vertical axis
- MUST include a key to denote the different bars



Example : Comparative Bar Chart

When constructing a comparative bar chart we use the relative frequency rather than the frequency to construct the scale on the vertical axis so that we can make meaningful comparisons even if the sample sizes are not the same.





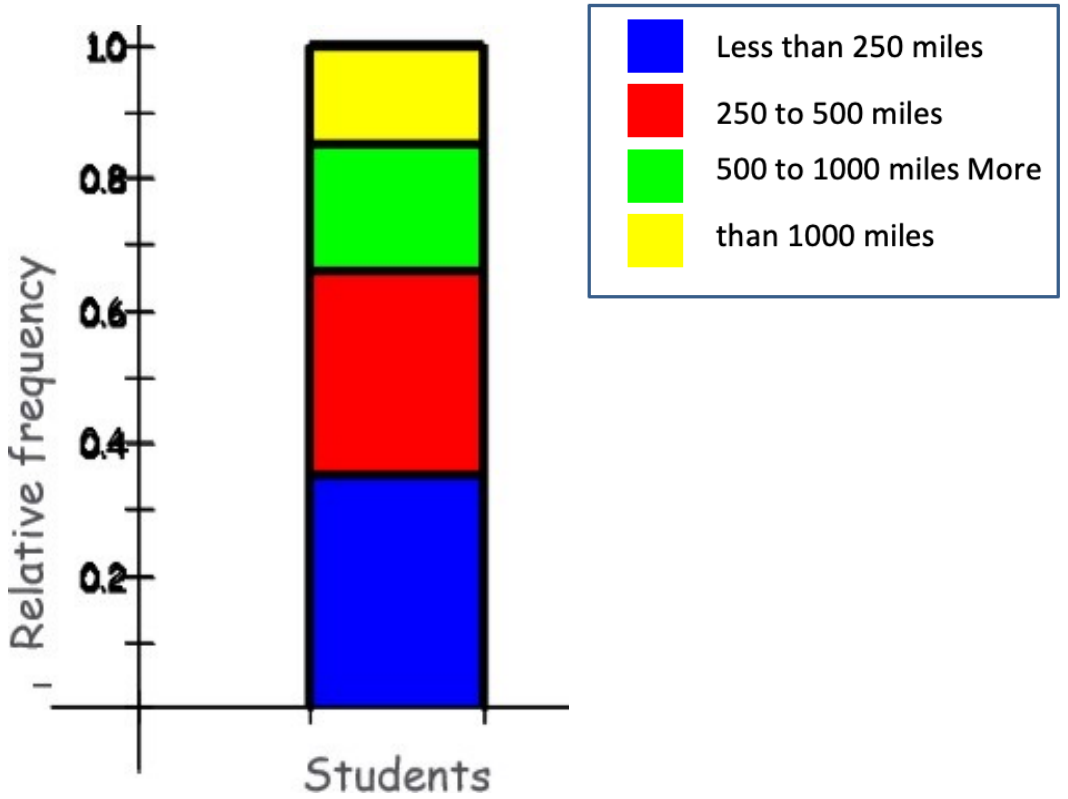
Data Presentation (Graphical) - Segmented (or Stacked) Bar Chart

When to Use:

- Categorical data

How to construct:

- MUST first calculate relative frequencies
- Draw a bar representing 100% of the group
- Divide the bar into segments corresponding to the relative frequencies of the categories



Data Presentation (Graphical) - Pie Chart

When to Use:

- Categorical data

How to construct:

- Draw a circle to represent the entire data set
- Calculate the size of each "slice":

$$\text{Relative frequency} \times 360^{\circ}$$

- Using a protractor, mark off each slice

To describe

– comment on which category had the largest proportion or smallest proportion

Example : Pie Chart

Typos on a résumé do not make a very good impression when applying for a job. Senior executives were asked how many typos in a résumé would make them not consider a job candidate **("Job Seekers Need a Keen Eye," *USA Today,* August 3, 2009)**. The resulting data are summarized in the accompanying relative frequency distribution.

Number of Typos	Frequency	Relative Frequency
1	60	.40
2	54	.36
3	21	.14
4 or more	10	.07
Don't know	5	.03

Total Frequency: 150

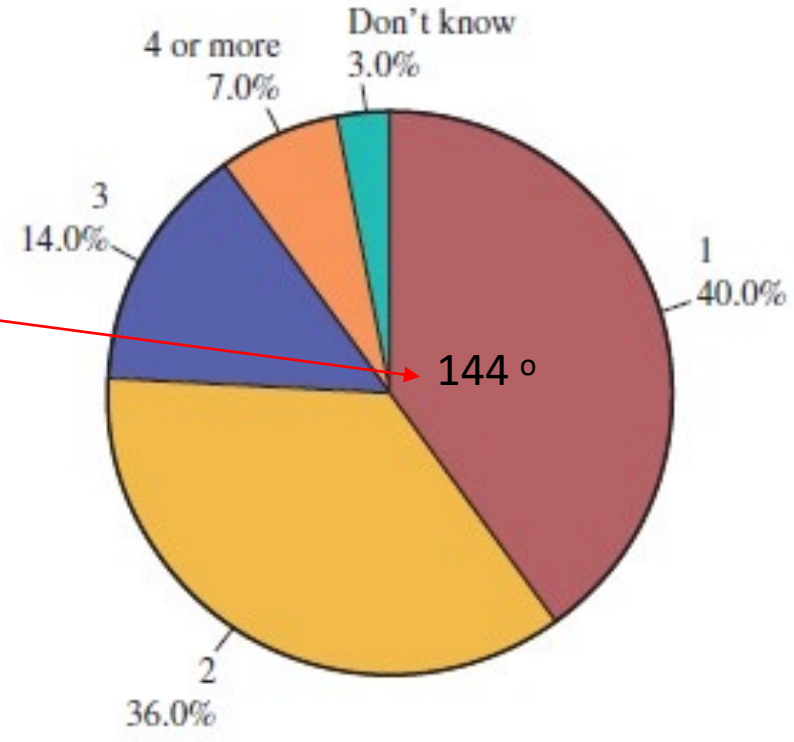
Let's draw a pie chart together!

Example : Pie Chart

Typos on a résumé do not make a very good impression when applying for a job. Senior executives were asked how many typos in a résumé would make them not consider a job candidate **("Job Seekers Need a Keen Eye," *USA Today,* August 3, 2009)**. The resulting data are summarized in the accompanying relative frequency distribution.

Number of Typos	Frequency	Relative Frequency
1	60	.40
2	54	.36
3	21	.14
4 or more	10	.07
Don't know	5	.03

Total Frequency: 150

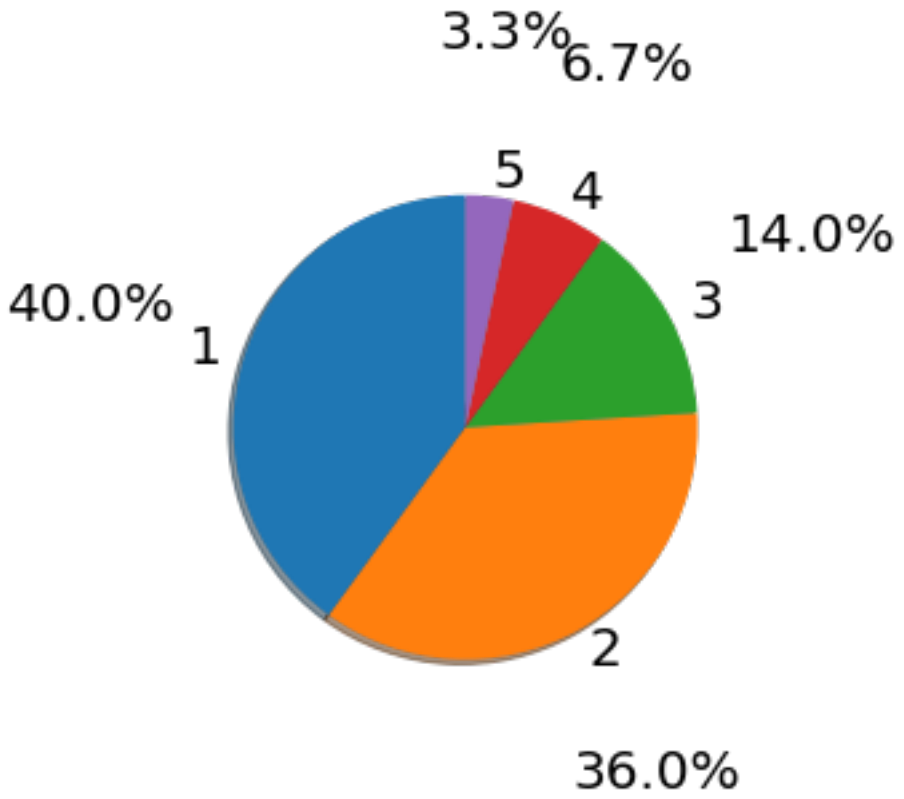


slice size = (.40)(360) = 144 degrees

Example : Pie Chart

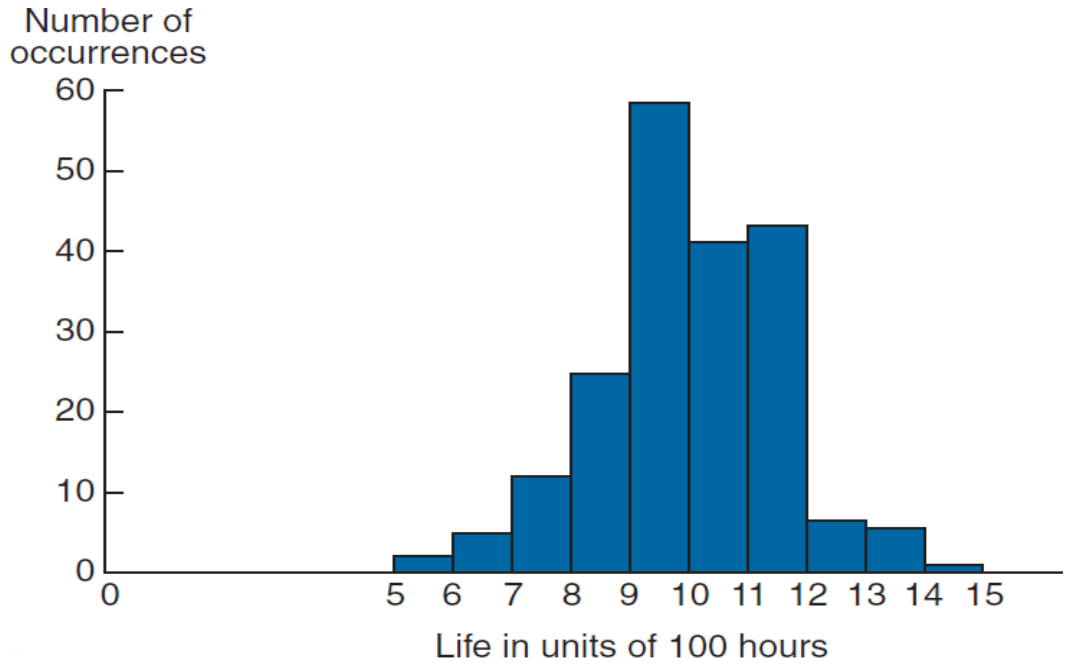
Let's do it in Python with matplotlib

```
# Frequencies
sizes = [60, 54, 21, 10, 5]

# Labels for each section
labels = ['1', '2', '3', '4', '5']

# Plot the pie chart
plt.pie(
    sizes,
    labels=labels,
    autopct='%1.1f%%',
    shadow=True,
    startangle=90,
    pctdistance=1.7,
    labeldistance=1.1,
    textprops={'fontsize': 20}
)


# Show the plot
plt.show()
```



Data Presentation (Graphical) - Histogram

A graphical version of the frequency table so that we can allow for a quick visual interpretation of the data. The class frequency is on the y-axis and the class intervals are placed on the x-axis.



Data Presentation (Graphical) - Histogram

When to Use:

- Univariate numerical data; Discrete Data

How to construct:

- Draw a horizontal scale and mark it with the possible values for the variable
- Draw a vertical scale and mark it with frequency or relative frequency
- Above each possible value, draw a rectangle centered at that value with a height corresponding to its frequency or relative frequency

To describe

– comment on the center, spread, and shape of the distribution and if there are any unusual features

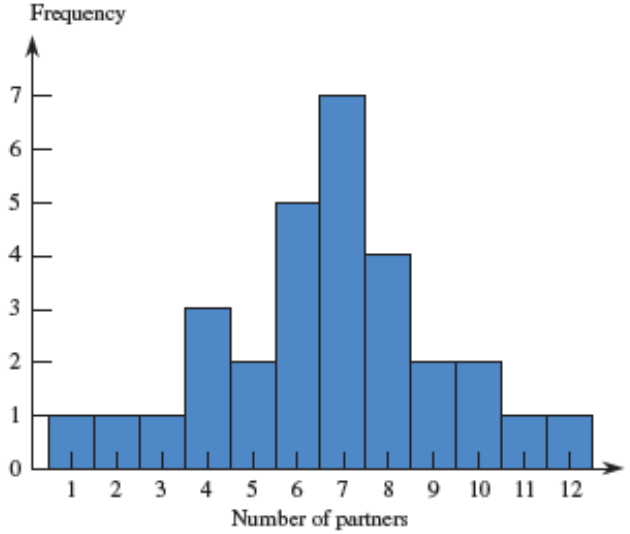
Example : Histogram

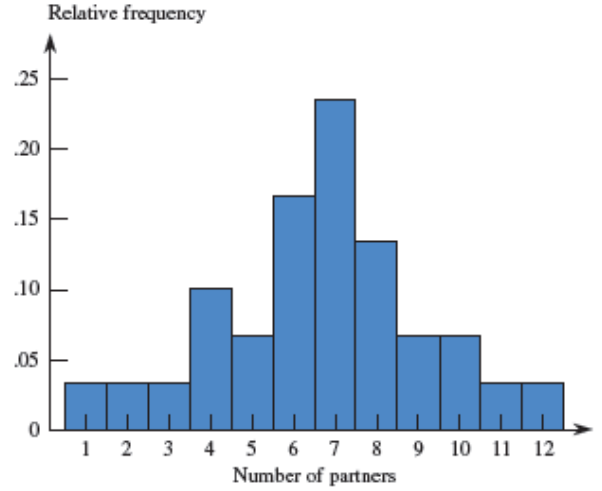
A study on reading hour per week for 30 university students.

12	2	4	6	6	7
8	7	8	11	8	3
5	6	7	10	1	9
7	6	9	7	5	4
7	4	6	7	8	10

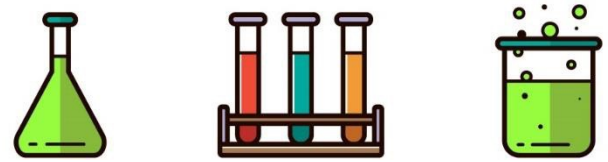
A frequency is the number of times a value of the data occurs

Relative frequency = $\frac{\text{frequency}}{\text{total}}$





Example : Histogram

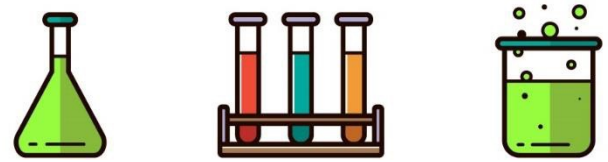


A laboratory's measurement process was assessed by randomly inserting 27 specimens having a known concentration of 8.0 mg/L into the normal flow of work over a period of 2 weeks.

The results in order of observation were 6.9, 7.8, 8.9, 5.2, 7.7, 9.6, 8.7, 6.7, 4.8, 8.0, 10.1, 8.5, 6.5, 9.2, 7.4, 6.3, 5.6, 7.3, 8.3, 7.2, 7.5, 6.1, 9.4, 5.4, 7.6, 8.1, and 7.9 mg/L.

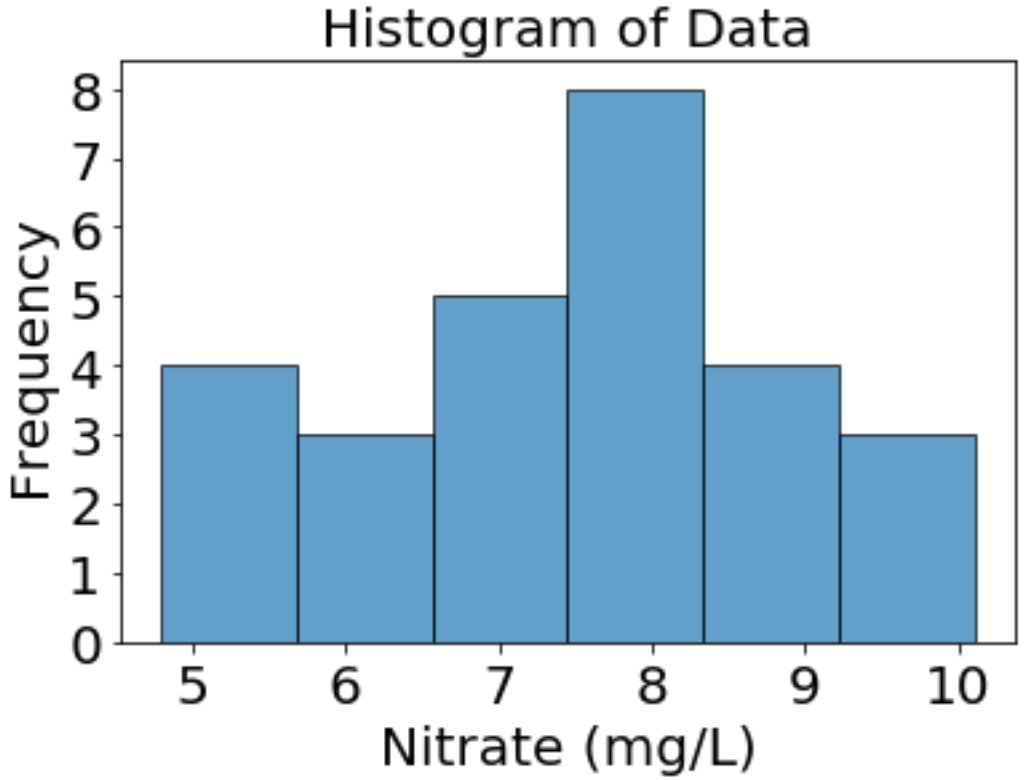
The *population* is all specimens having a known concentration of 8.0 mg/L. The *sample* is the 27 observations (measurements). The *sample size* is $n = 27$.

Example : Histogram



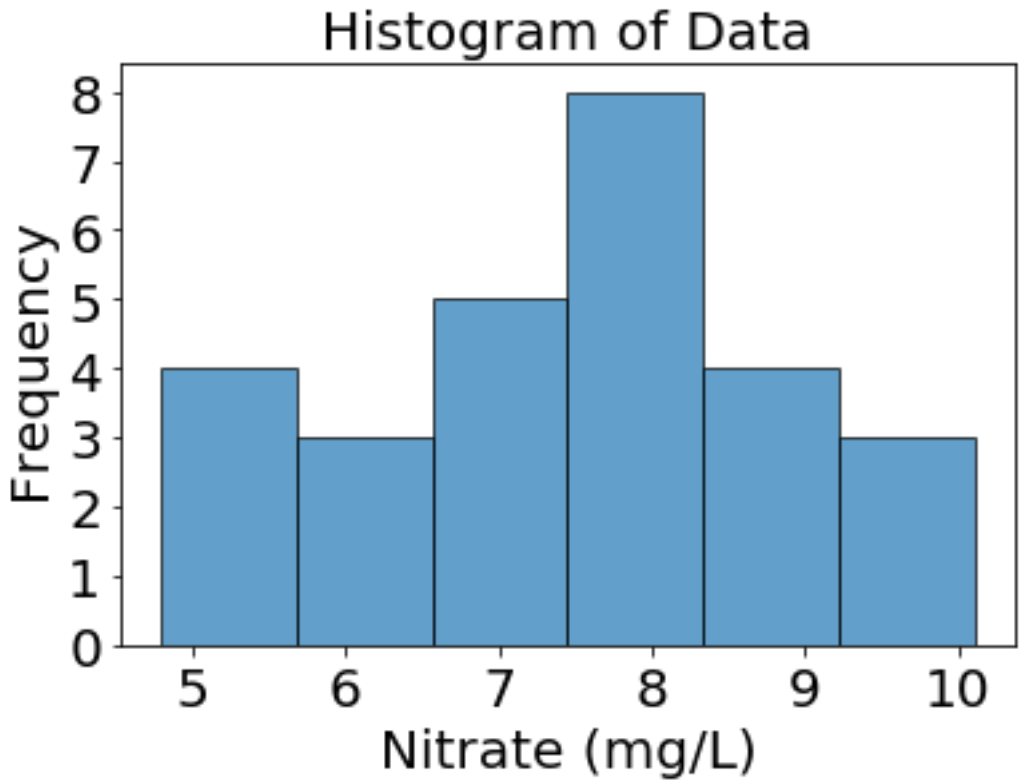
Data: 6.9, 7.8, 8.9, 5.2, 7.7, 9.6, 8.7, 6.7, 4.8, 8.0, 10.1, 8.5, 6.5, 9.2, 7.4, 6.3, 5.6, 7.3, 8.3, 7.2, 7.5, 6.1, 9.4, 5.4, 7.6, 8.1, and 7.9 mg/L

Concentration (mg/L)	Frequency
4-5	1
5-6	3
6-7	5
7-8	8
8-9	5
9-10	3
10-11	1



Example : Histogram Let's do it in Python with matplotlib

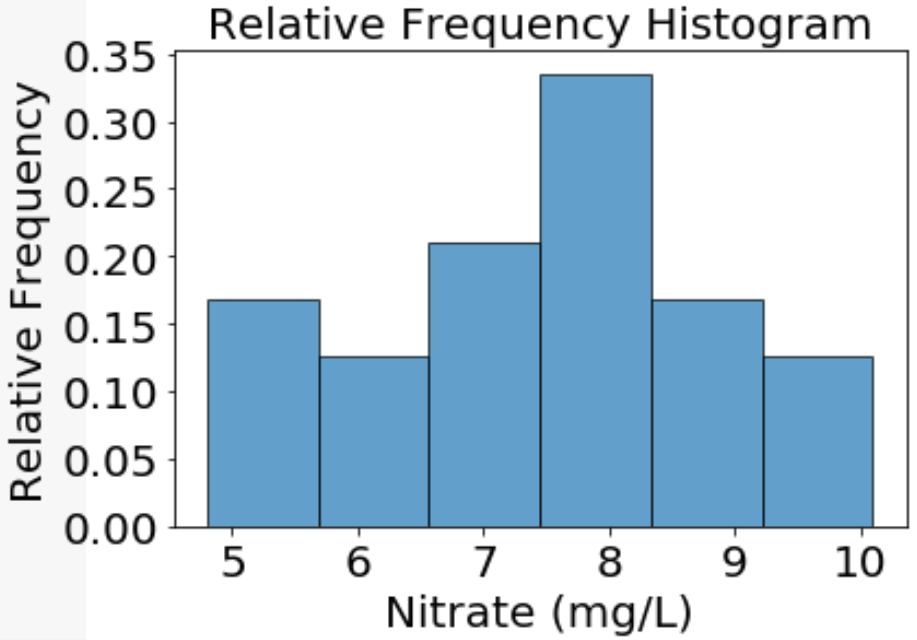
```
: # Given data set
data = [
    6.9, 7.8, 8.9, 5.2, 7.7, 9.6, 8.7,
    6.7, 4.8, 8.0, 10.1, 8.5, 6.5,
    9.2, 7.4, 6.3, 5.6, 7.3, 8.3, 7.2,
    7.5, 6.1, 9.4, 5.4, 7.6, 8.1, 7.9
]

# Create the histogram
plt.hist(data, bins='auto', edgecolor='black', alpha=0.7)

plt.title('Histogram of Data', fontsize=20)
plt.xlabel('Nitrate (mg/L)', fontsize=20)
plt.ylabel('Frequency', fontsize=20)
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)

# Show the histogram
plt.show()
```



Example : Histogram Let's do it in Python with matplotlib

With relative frequency

```
# Create the histogram with relative frequencies
plt.hist(data, bins='auto', edgecolor='black',
         alpha=0.7, density=True)

plt.title('Relative Frequency Histogram', fontsize=20)
plt.xlabel('Value', fontsize=20)
plt.ylabel('Relative Frequency', fontsize=20)

plt.xticks(fontsize=20)
plt.yticks(fontsize=20)

# Show the histogram
plt.show()
```



Data Presentation (Graphical)

– Histogram with unequal intervals

When to Use:

- when you have a concentration of data in the middle with some extreme values

How to construct:

- construct similar to histograms with continuous data, but with density on the vertical axis

$$\text{density} = \frac{\text{relative frequency for interval}}{\text{width of interval}}$$

Data Presentation (Graphical)

- Cumulative Relative Frequency Plot

When to Use:

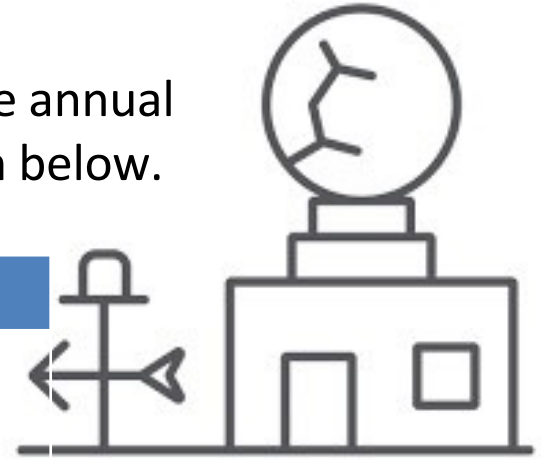
- Used to answer questions about percentiles (a value with a given percent of observations at or below that value)

How to construct:

- Mark the boundaries of the intervals on the horizontal axis
- Draw a vertical scale and mark it with relative frequency
- Plot the point corresponding to the upper end of each interval with its cumulative relative frequency, including the beginning point
- Connect the points.

Example : Cumulative Relative Frequency Plot

The National Climatic Center has been collecting weather data for many years. The annual rainfall amounts from 1950 to 2008 were used to create the frequency distribution below.

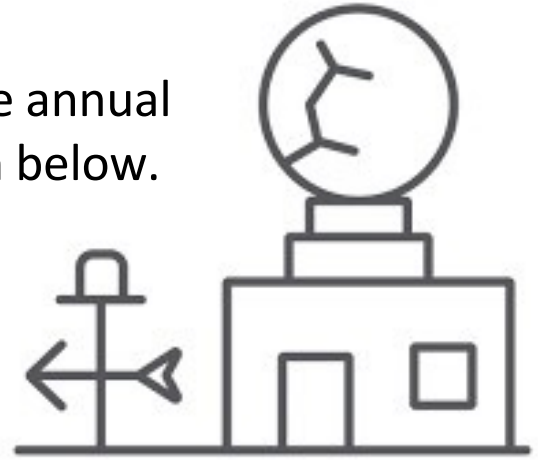


Annual Rainfall (in inches)	Relative frequency	Cumulative relative frequency
4 to <5	0.052	**0.052**
5 to <6	0.103	**0.155**
6 to <7	0.086	**0.241**
7 to <8	0.103	
8 to <9	0.172	
9 to <10	0.069	
10 to <11	0.207	
11 to <12	0.103	
12 to <13	0.052	
13 to <14	0.052	

Continue this pattern to complete the table

Example : Cumulative Relative Frequency Plot

The National Climatic Center has been collecting weather data for many years. The annual rainfall amounts from 1950 to 2008 were used to create the frequency distribution below.



Annual Rainfall (in inches)	Relative frequency	Cumulative relative frequency
4 to <5	0.052	**0.052**
5 to <6	0.103	**0.155**
6 to <7	0.086	**0.241**
7 to <8	0.103	**0.344**
8 to <9	0.172	**0.516**
9 to <10	0.069	**0.585**
10 to <11	0.207	**0.792**
11 to <12	0.103	**0.895**
12 to <13	0.052	**0.947**
13 to <14	0.052	**0.999**

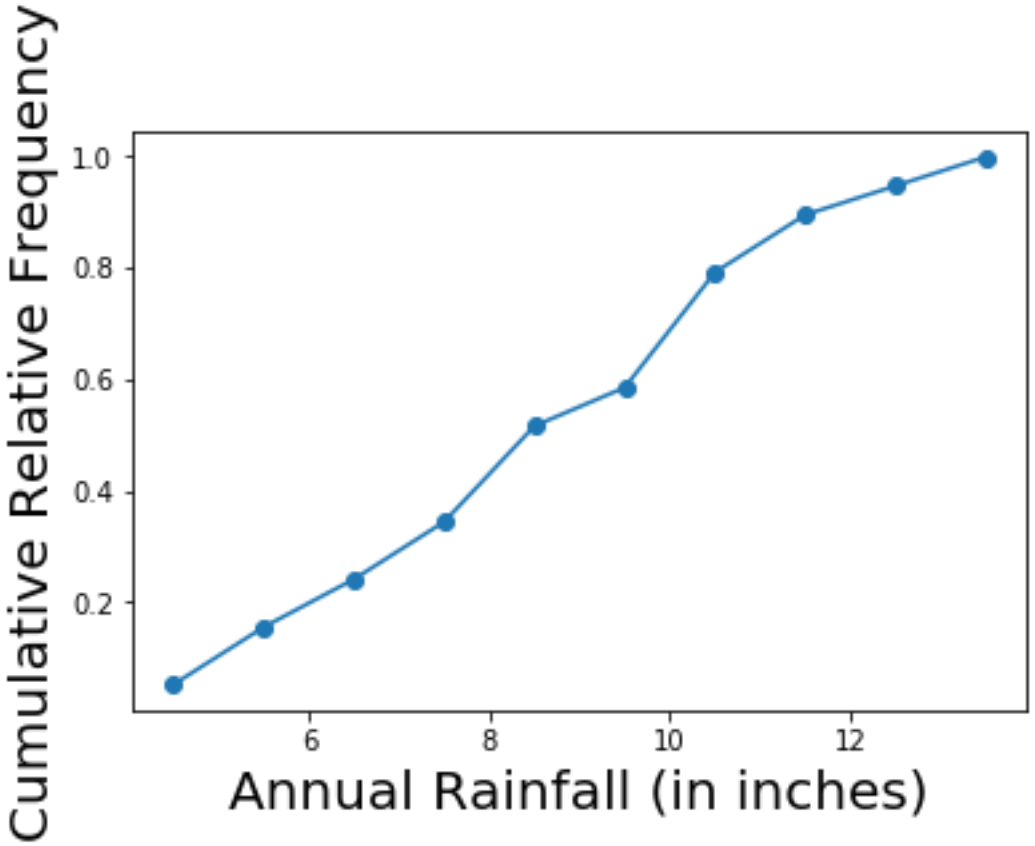
Example : Cumulative Relative Frequency Plot

 # Calculate the cumulative relative frequency

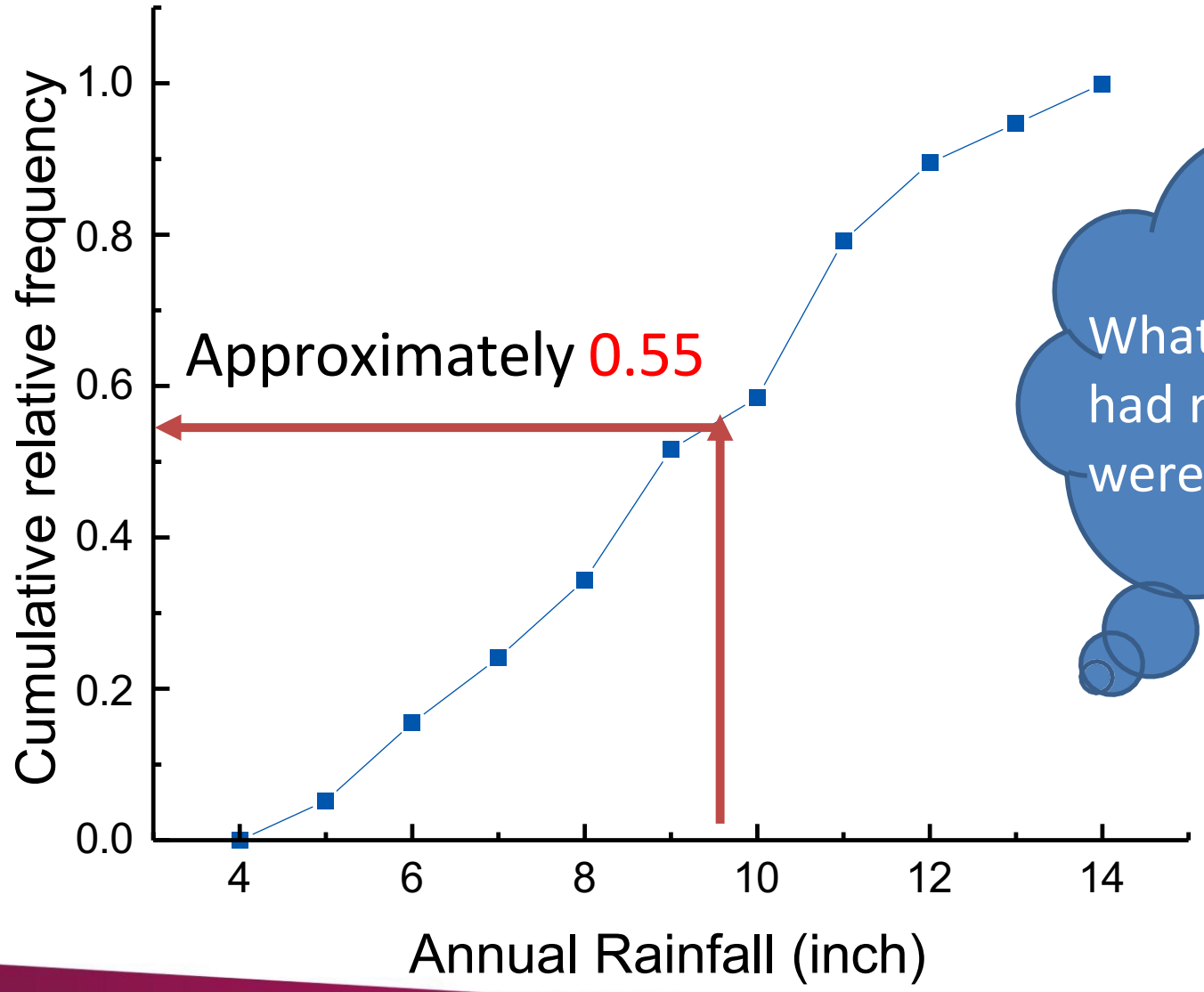
```
cumulative_relative_frequencies = [sum(relative_frequencies[:i+1]) for i in range(len(relative_frequencies))]
```

 # Create a dot plot for the cumulative relative frequency

```
plt.plot(midpoints, cumulative_relative_frequencies, 'o-')
```

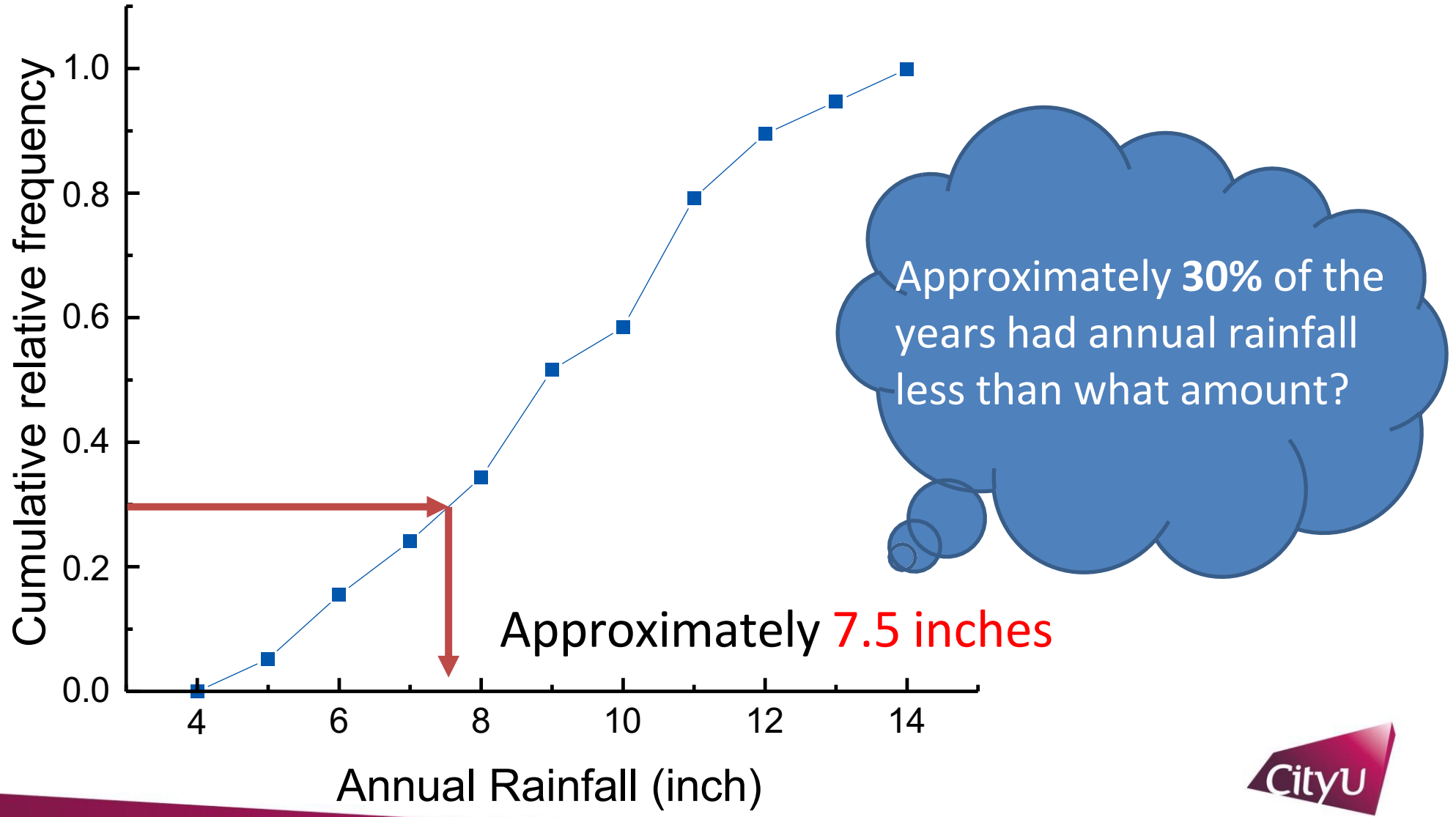


Example : Cumulative Relative Frequency Plot

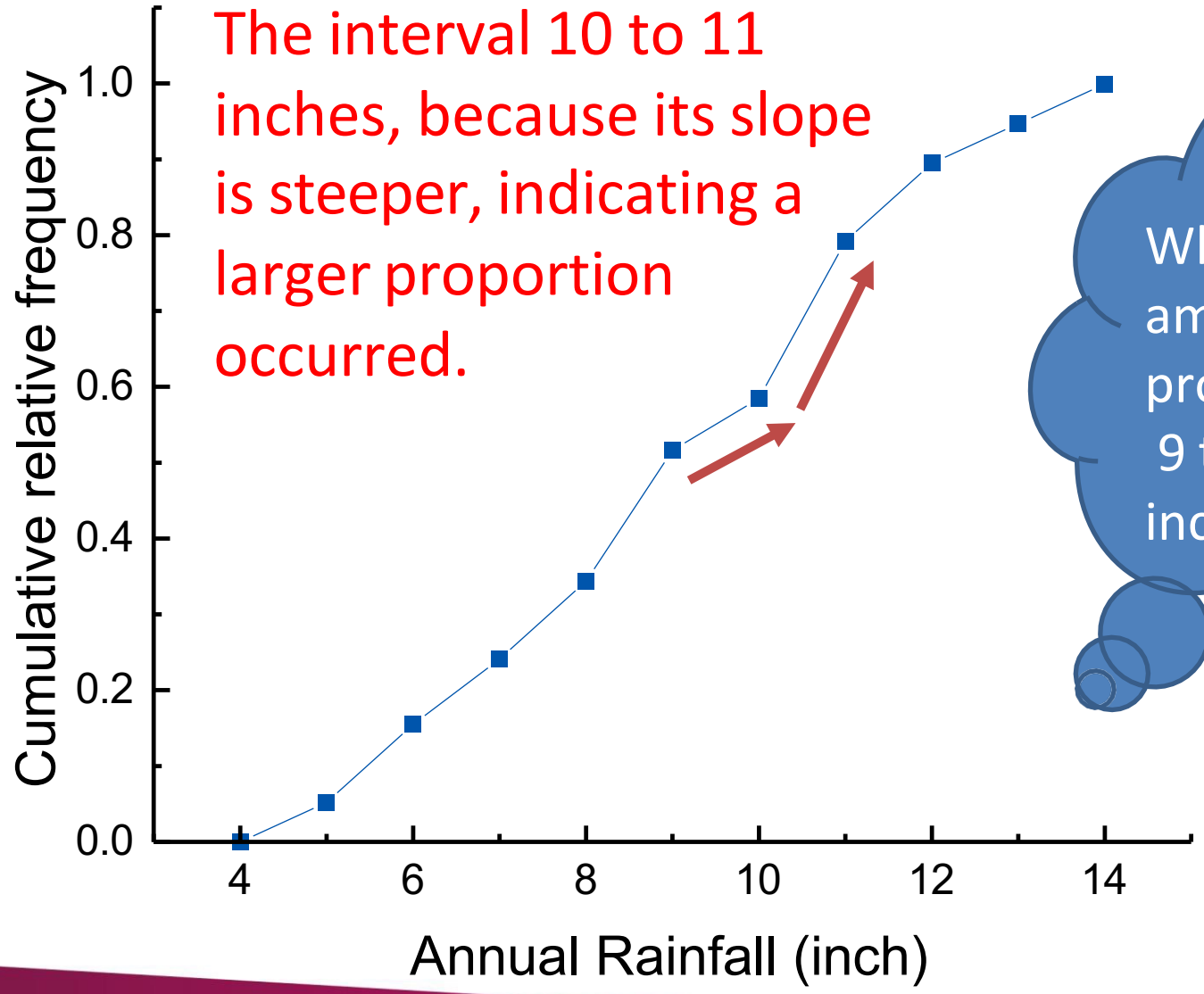


What proportion of years had rainfall amounts that were **9.5 inches or less**?

Example : Cumulative Relative Frequency Plot



Example : Cumulative Relative Frequency Plot



The interval 10 to 11 inches, because its slope is steeper, indicating a larger proportion occurred.

Which interval of rainfall amounts had a larger proportion of years:
9 to 10 inches or 10 to 11 inches? Explain

Data Presentation (Graphical) - Scatterplot

When to Use:

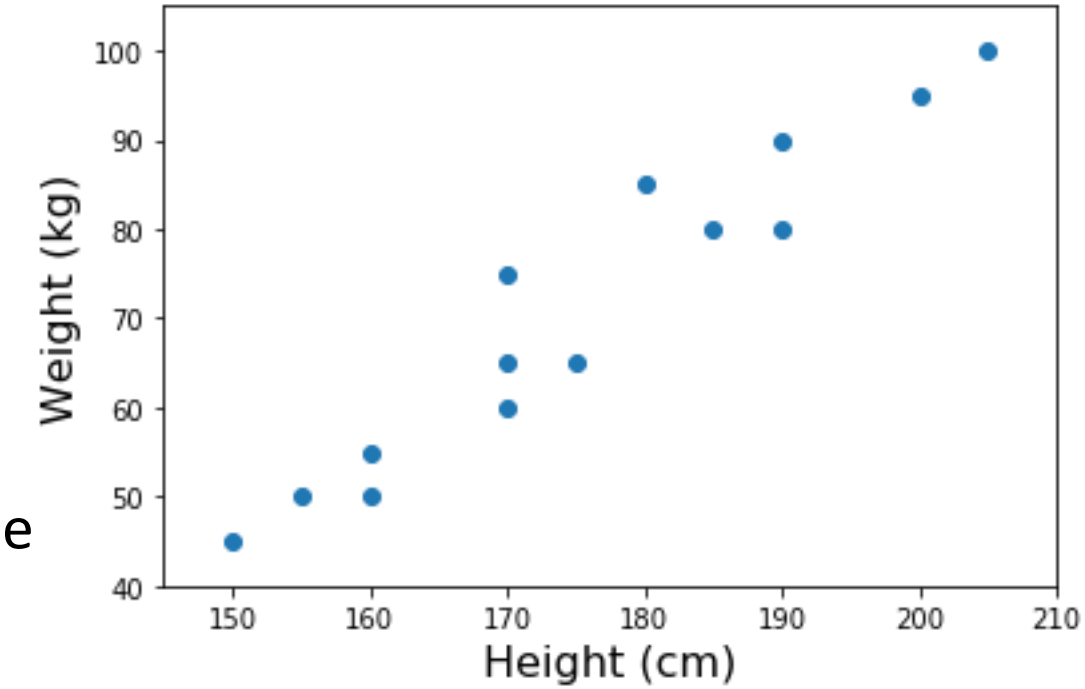
- Bivariate numerical data

How to construct:

- Draw a horizontal scale and mark it with appropriate values of the independent variable
- Draw a vertical scale and mark it appropriate values of the dependent variable
- Plot each point corresponding to the observations

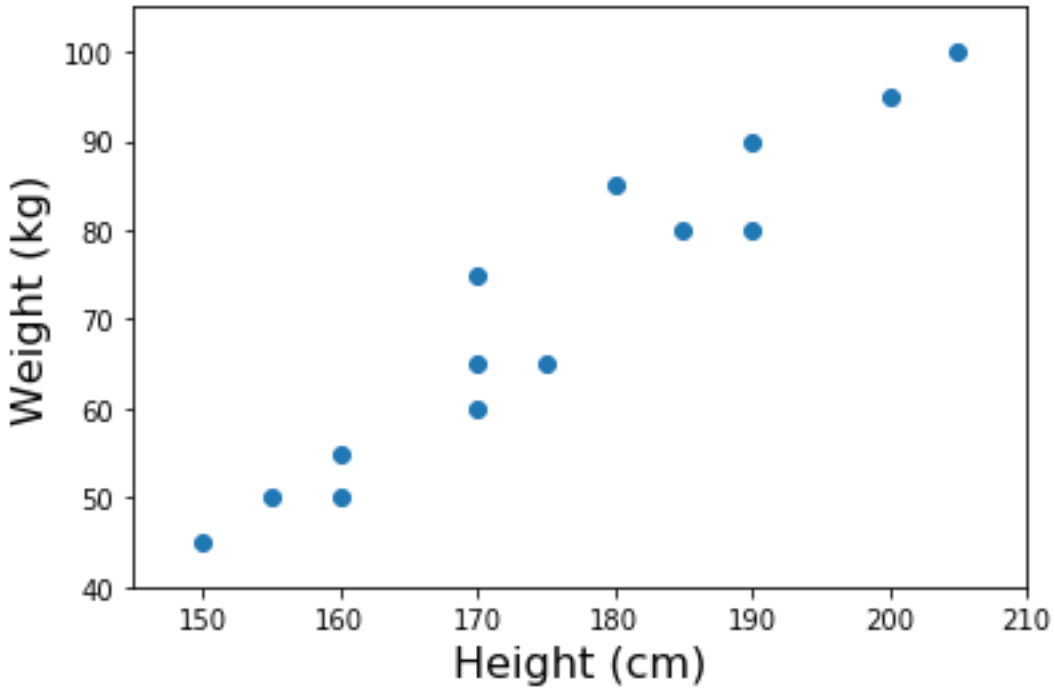
To describe

– comment the relationship between the variables

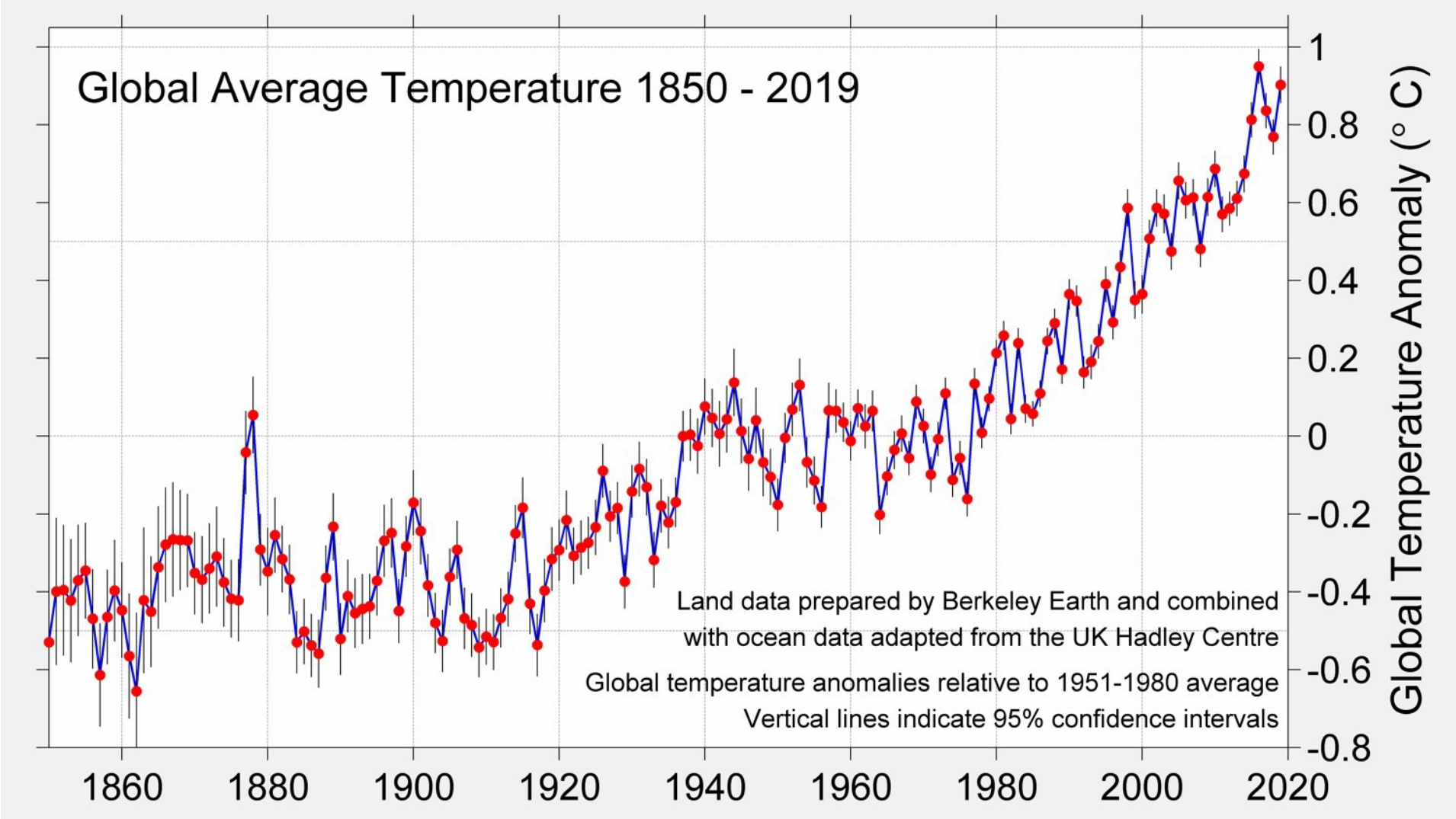


Data Presentation (Graphical) - Scatterplot

```
# Create a scatter plot
plt.scatter(heights, weights, marker='o')

plt.xlabel('Height (cm)')
plt.ylabel('Weight (kg)')

plt.show()
```



Data Presentation (Graphical) - Time Series Plots



Data Presentation (Graphical) - Time Series Plots

When to Use:

- measurements collected over time at regular intervals

How to construct:

- Draw a horizontal scale and mark it with appropriate values of time
- Draw a vertical scale and mark it appropriate values of the observed variable
- Plot each point corresponding to the observations and connect

To describe

- comment on any trends or patterns over time

Data Presentation (Graphical) – Stem-and-Leaf Displays

A stem-and-leaf display or stem-and-leaf plot is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in **visualizing the shape of a distribution**.

Unlike histograms, stem-and-leaf displays retain the original data to at least two significant digits, and put the data in order.

Stem	Leaf
4	4 6 7 9
5	
6	3 4 6 8 8
7	2 2 5 6
8	1 4 8
9	
10	6

Data Presentation (Graphical) – Stem-and-Leaf Displays

When to Use:

- Univariate numerical data

How to construct:

- Select one or more of the leading digits for the stem
- List the possible stem values in a vertical column
- Record the leaf for each observation beside each corresponding stem value
- Indicate the units for stems and leaves in a key or legend

To describe

– comment on the center, spread, and shape of the distribution and if there are any unusual features

Example : Stem-and-Leaf Displays

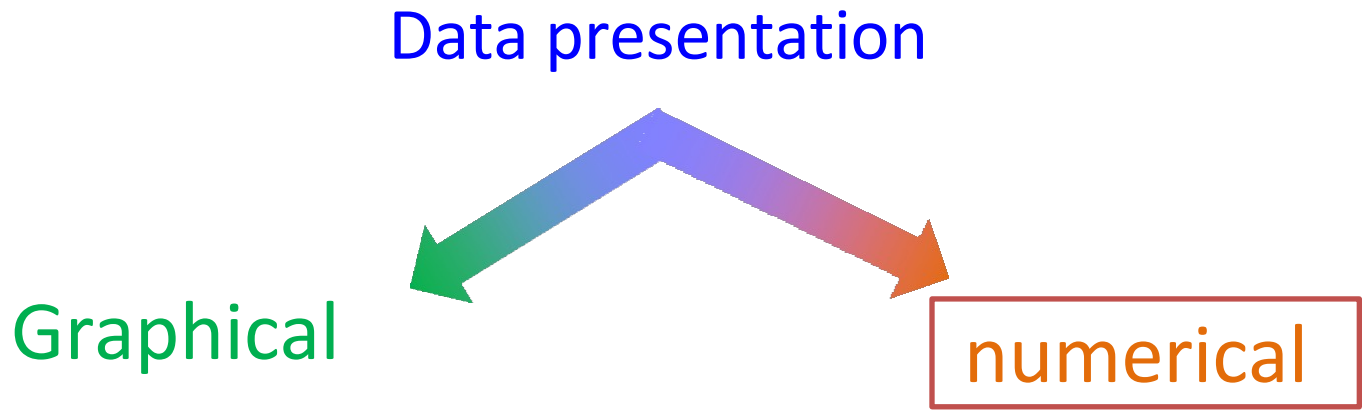
44, 46, 47, 49, 63, 64, 66, 68, 68, 72, 72, 75, 76, 81, 84, 88, 106

Stem	Leaf
4	4 6 7 9
5	
6	3 4 6 8 8
7	2 2 5 6
8	1 4 8
9	
10	6

Leaf unit: 1.0

Stem unit: 10.0

Data Presentation



Data Presentation (Numerical)

US Medicare introduced a new prescription drug program. The article **"Those Most in Need May Miss Drug Benefit Sign-Up"** notes that only 24% of those eligible for low-income subsidies under this program. The article also gave the percentage of those eligible who had signed up in each of 49 states and the District of Columbia:

24	27	12	38	21	26	23	33	19	19	26	28
16	21	28	20	21	41	22	16	29	26	22	16
27	22	19	22	22	22	30	20	21	34	26	20
25	19	17	21	27	19	27	34	20	30	20	21
14	18										

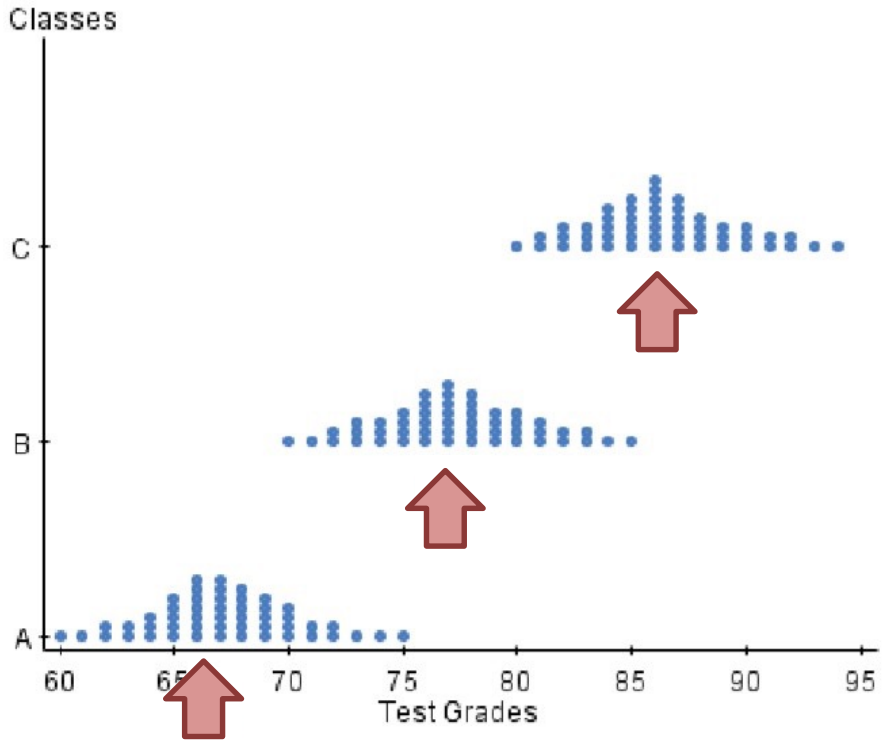
Data Presentation (Numerical)

- What is a typical value for this data set?
- The enrollment percentages differ widely from state to state.
- How might we summarize this variability numerically?

24	27	12	38	21	26	23	33	19	19	26	28
16	21	28	20	21	41	22	16	29	26	22	16
27	22	19	22	22	22	30	20	21	34	26	20
25	19	17	21	27	19	27	34	20	30	20	21
14	18										

Data Presentation (Numerical) - Center

The exam scores in classes A, B, & C



What strikes you as the most distinctive difference among the distributions?

mean, median, & mode?

Data Presentation (Numerical) - Center

Sample mean

The mean of a numerical data set is just the familiar arithmetic average: the sum of the observations divided by the number of observations.

DEFINITION

The **sample mean** of a sample consisting of numerical observations $x_1, x_2, \ldots, x_n$, denoted by $\bar{x}$, is

$$\bar{x} = \frac{\text{sum of all observations in the sample}}{\text{number of observations in the sample}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

Data Presentation (Numerical) - Center

Population mean

DEFINITION

The **population mean**, denoted by μ, is the average of all x values in the entire population.

Data Presentation (Numerical) - Center

Sample median

Once the data values have been listed in order from smallest to largest, the **median** is the middle value in the list, and it divides the list into two equal parts.

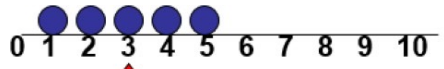
DEFINITION

The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included, so that every sample observation appears in the ordered list). Then

$$\text{sample median} = \begin{cases} \text{the single middle value if } n \text{ is odd} \\ \text{the average of the middle two values if } n \text{ is even} \end{cases}$$

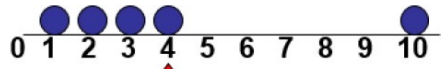
Data Presentation (Numerical) - Center

Comparison between the mean and the median



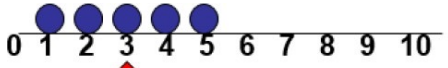
Mean = 3

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

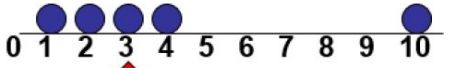


Mean = 4

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$



Median = 3

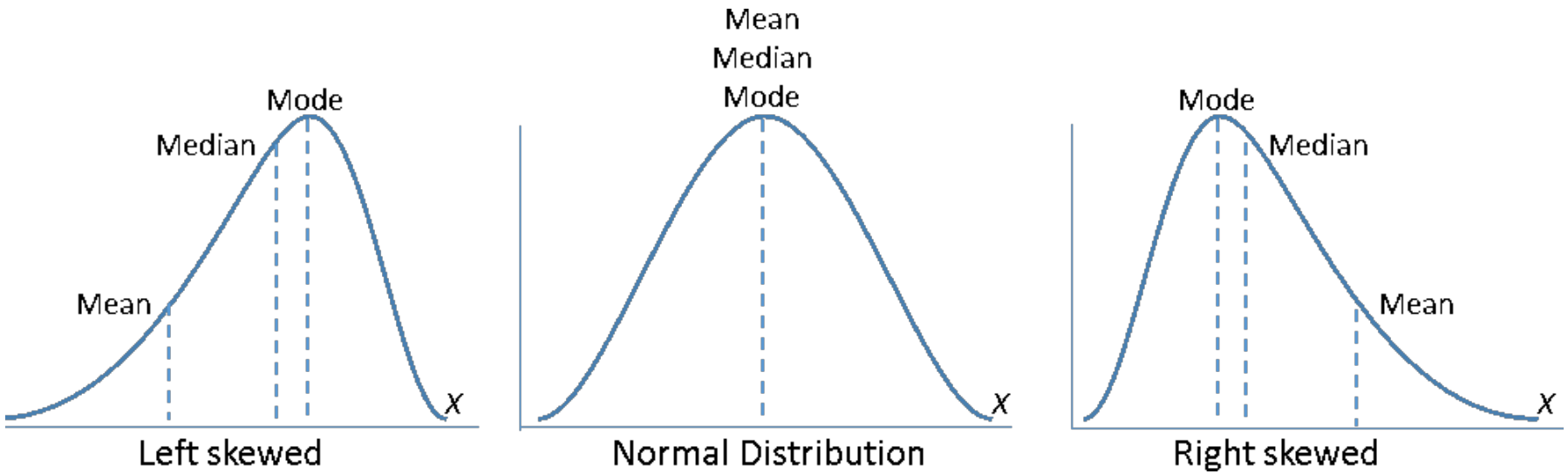


Median = 3

- ✓ The sample mean summarizes the center well if the distribution is symmetric, unimodal and there are no outliers. Otherwise, the sample median is a better alternative.
- ✓ The sample mean is sensitive to outliers, but the sample median is **not**.

Data Presentation (Numerical) - Center

- **Sample Mode :** the observation that occurs the most often
  - Can be more than one mode
  - If all values occur only once – there is no mode
  - Not used as often as mean & median



Example : Center

Data was collected on each of ten consecutive days:

44	50	38	96	42	47	40	39	46	50

Find the sample **mean** and **median**

Example : Center

Data was collected on each of ten consecutive days:

44	50	38	96	42	47	40	39	46	50

Find the sample **mean** and **median**

To find the sample mean, add them and divide by 10:

(44 + 50 + 38 + 96 + 42 + 47 + 40 + 39 + 46 + 50) / 10 = 49.2

To find the median, first sort the data:

38, 39, 40, 42, 44, 46, 47, 50, 50, 96

Notice that there are two middle numbers 44 and 46. To find the median we take the average of the two.

Median =(44 + 46) / 2= 45

Example: Median

Suppose we catch a sample of fish from the lake. The lengths of the fish (in inches) are listed below. Find the **median** length of fish.

Case 1

3	4	**5**	8	10

Median = 5

The numbers are in order & n is odd – so find the middle observation.

The numbers are in order & n is even – so find the middle two observations.

Case 2

3	4	**5**	**6**	8	10

Median = 5.5

Example : Mean & Deviation

Now find how each observation deviates from the mean. $\bar{X} = 6$

x	3	4	5	6	8	10	Sum
$x - \bar{x}$	-3	-2	-1	0	2	4	**0**

This is the **deviation** from mean

The mean is considered the balance point of the distribution because it "balances" the positive and negative deviations.

Example : Mean & Deviation

Now find how each observation deviates from the mean. $\bar{X} = 6$

x	3	4	5	6	8	10	Sum
$x - \bar{x}$	-3	-2	-1	0	2	4	**0**

This is the **deviation** from mean

The mean is considered the balance point of the distribution because it "balances" the positive and negative deviations.

Will this sum always equal zero?

Example : Mean & Deviation

Now find how each observation deviates from the mean. $\bar{X} = 6$

x	3	4	5	6	8	10	Sum
$x - \bar{x}$	-3	-2	-1	0	2	4	**0**

This is the **deviation** from mean

The mean is considered the balance point of the distribution because it "balances" the positive and negative deviations.

Will this sum always equal zero?

Yes

CityU

Example : Mean & Median

What happens to the median & mean if the length of 10 inches was 15 inches or 20 inches ?

3	4	5	6	8	10

median	mean
5.5	6

3	4	5	6	8	15

5.5	6.833

3	4	5	6	8	20

5.5	7.667

Is the median resistant?

YES

Is the mean affected by extreme values?

YES

Data Presentation (Numerical) - Center

The extreme sensitivity of the mean to even a single outlier and the extreme insensitivity of the median to a substantial proportion of outliers can sometimes make both of them suspect as a measure of center.

A *trimmed mean* is a compromise between these two extremes.

DEFINITION

A **trimmed mean** is computed by first ordering the data values from smallest to largest, deleting a selected number of values from each end of the ordered list, and finally averaging the remaining values.

The **trimming percentage** is the percentage of values deleted from *each* end of the ordered list.

Data Presentation (Numerical) - Center

Sometimes, the number of observations to be deleted from each end of the dataset is specified. Then, the corresponding trimming percentage is calculated as

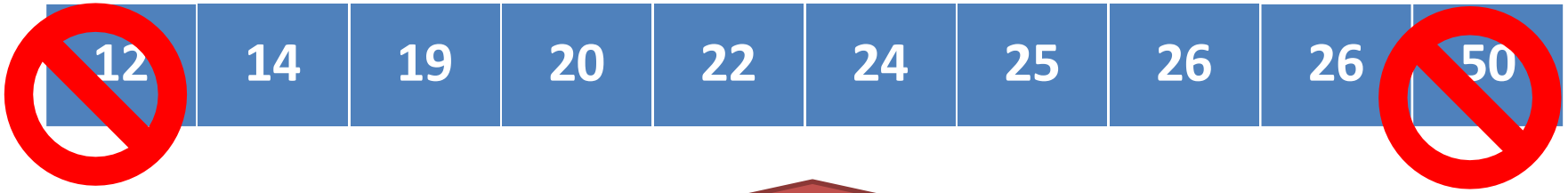
$$\text{trimming percentage} = \left(\frac{\text{number deleted from each end}}{n} \right) \cdot 100$$

In other cases, the trimming percentage is specified and then used to determine how many observations to delete from each end, with

$$\text{number deleted from each end} = \left(\frac{\text{trimming percentage}}{100} \right) \cdot n$$

Example : Trimmed Center

Find the mean of the following set of lab data



Mean = 23.8

Find a 10% trimmed.

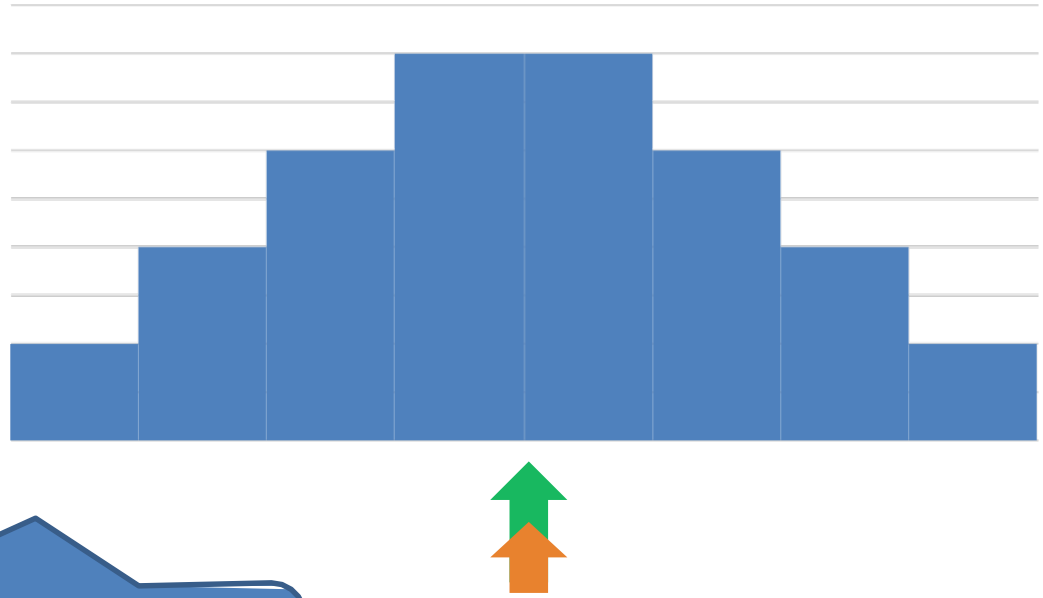
10%(10) = 1 ; So remove one observation from each side!

Trimmed Center : $\bar{x}_T = \frac{14+19+20+22+24+25+26+26}{8} = 22$

Example : Center

Suppose we caught a sample of 20 fish with the following lengths.
Create a histogram for the lengths of fish.

3	5	6	10	6
7	7	8	4	5
6	4	7	5	9
9	8	7	6	8



Look at the placement of the mean and median in this **symmetrical distribution**.

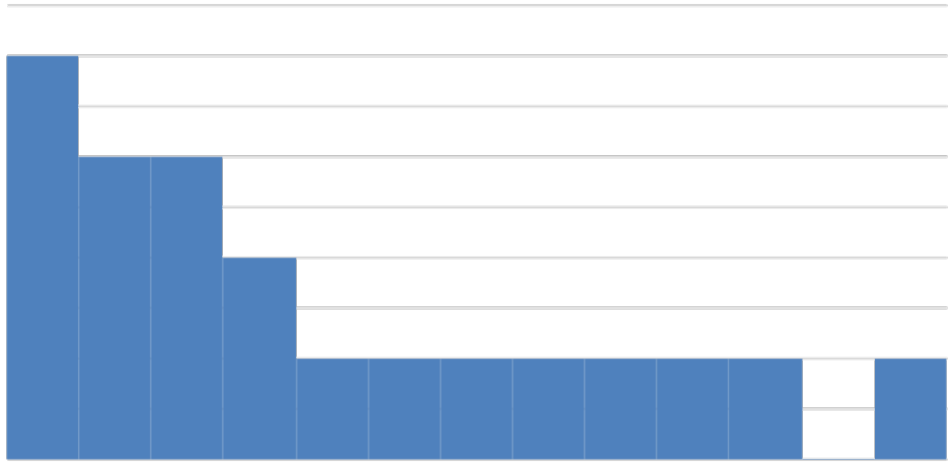
Mean =6.5

Median =6.5

Example : Center

Suppose we caught a sample of 20 fish with the following lengths.
Create a histogram for the lengths of fish.

3	5	6	10	15
7	3	3	4	5
6	4	12	5	3
4	8	13	11	9



Look at the placement of the mean and median in this **skewed distribution**.

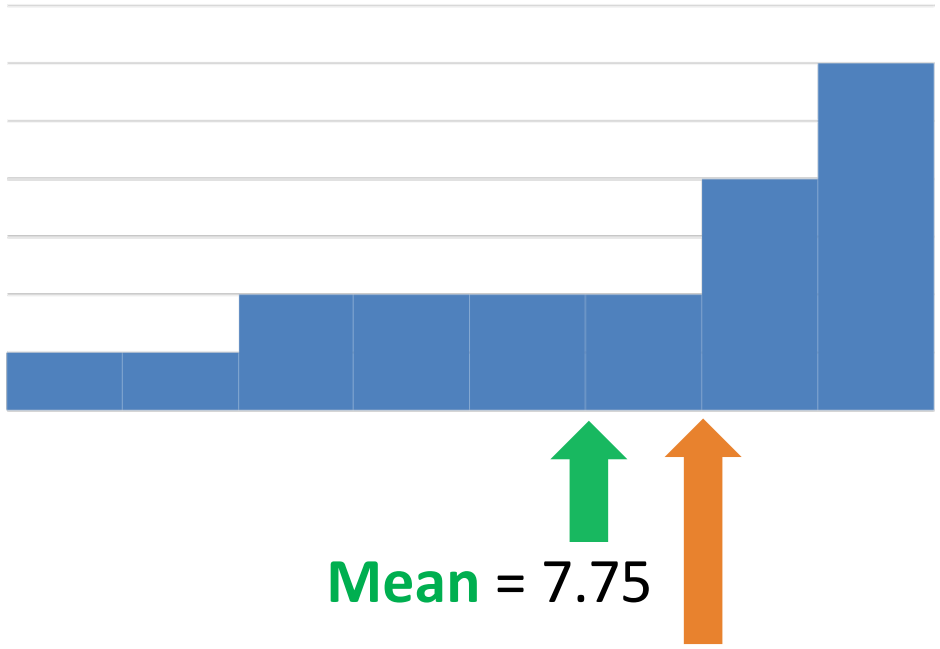
Mean =6.8

Median =5.5

Example : Center

Suppose we caught a sample of 20 fish with the following lengths. Create a histogram for the lengths of fish.

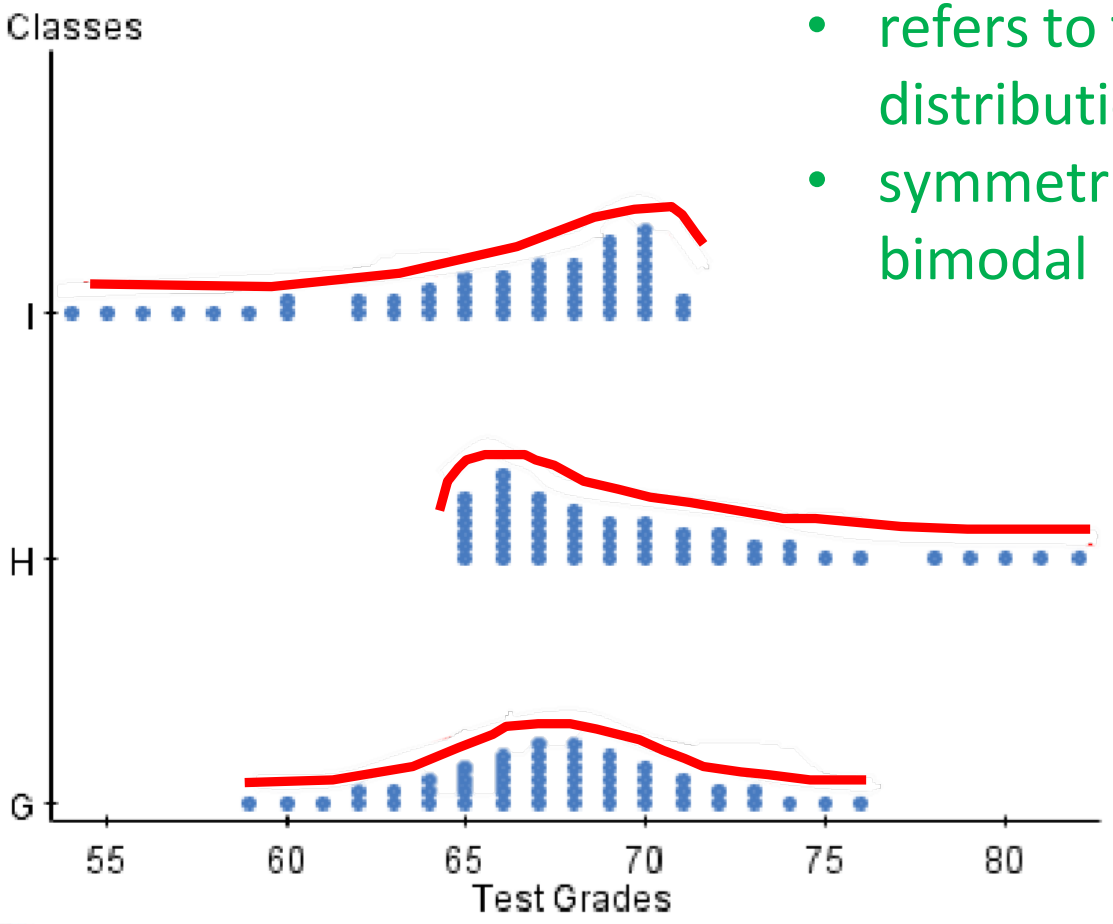
3	5	6	10	10
7	10	8	9	5
6	4	9	10	9
9	10	7	10	8



Look at the placement of the mean and median in this **skewed distribution**.

Data Presentation (Numerical) - Shape

What strikes you as the most distinctive difference among the distributions of exam scores in classes G, H, & I

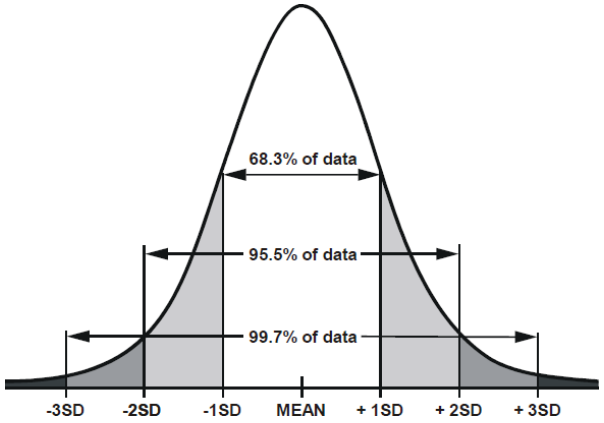


- refers to the overall shape of the distribution
- symmetrical, uniform, skewed, or bimodal

Data Presentation (Numerical) - Shape

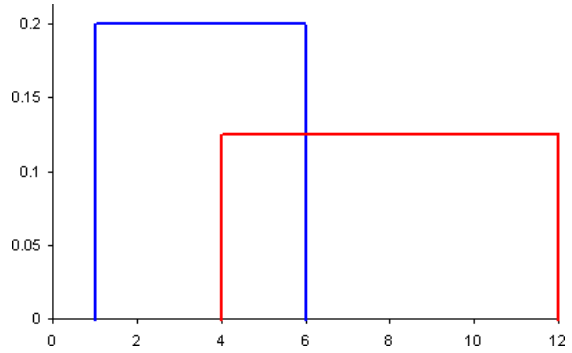
Symmetrical

- refers to data in which both sides are (more or less) the same when the graph is folded vertically down the middle
- bell-shaped is a special type
  has a center mound with two sloping tails



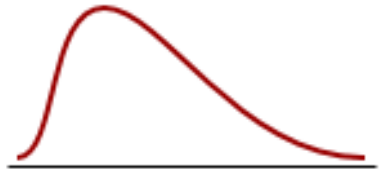
Uniform

- refers to data in which every class has equal or approximately equal frequency



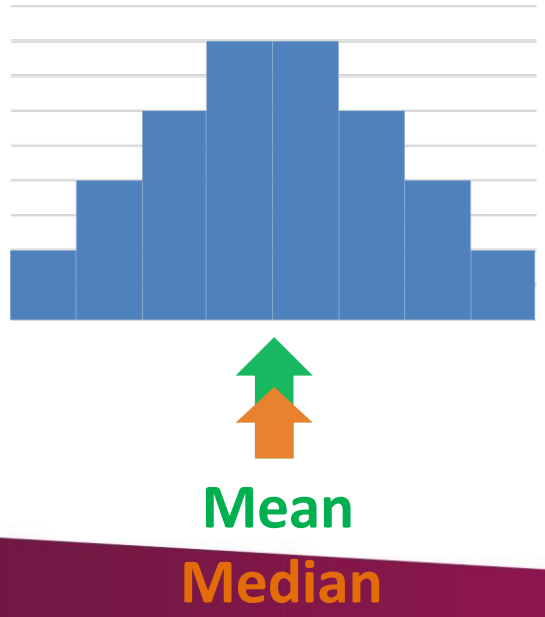
Skewed

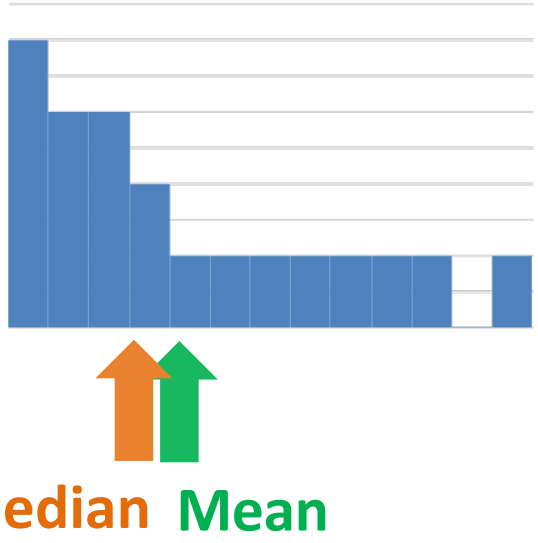
- refers to data in which one side (tail) is longer than the other side
- the direction of skewness is on the side of the longer tail

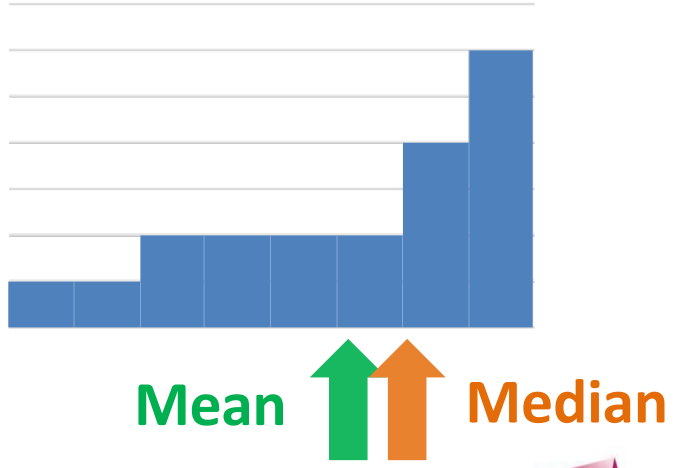


Data Presentation (Numerical) - Shape

- In a symmetrical distribution, the mean and median are equal.
- In a skewed distribution, the mean is pulled in the direction of the skewness.
- In a symmetrical distribution, you should report the mean!
- In a skewed distribution, the median should be reported as the measure of center!



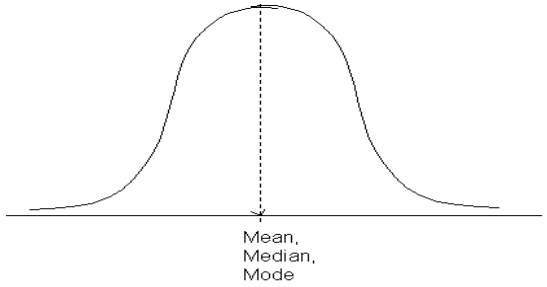


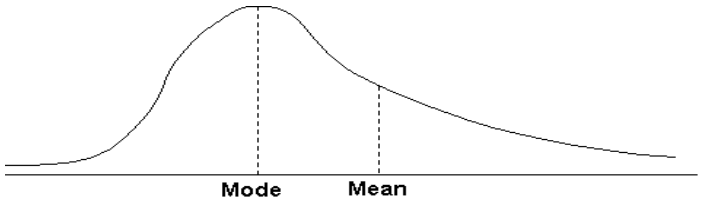


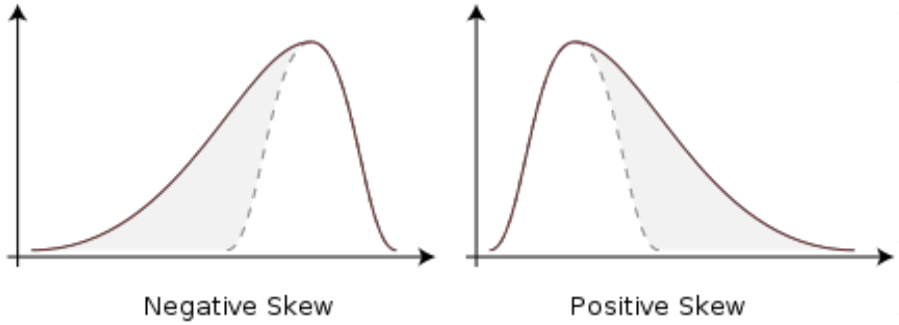
Data Presentation (Numerical) - Shape

A normal distribution is a bell-shaped distribution of data where the mean, median and mode all coincide.

If there are extreme values towards the positive end of a distribution, the distribution is said to be positively skewed. A negatively skewed distribution, on the other hand, has a mean which is less than the mode because of the presence of extreme values at the negative end of the distribution.







- **Positive Skewness**:
- distribution has a long right tail

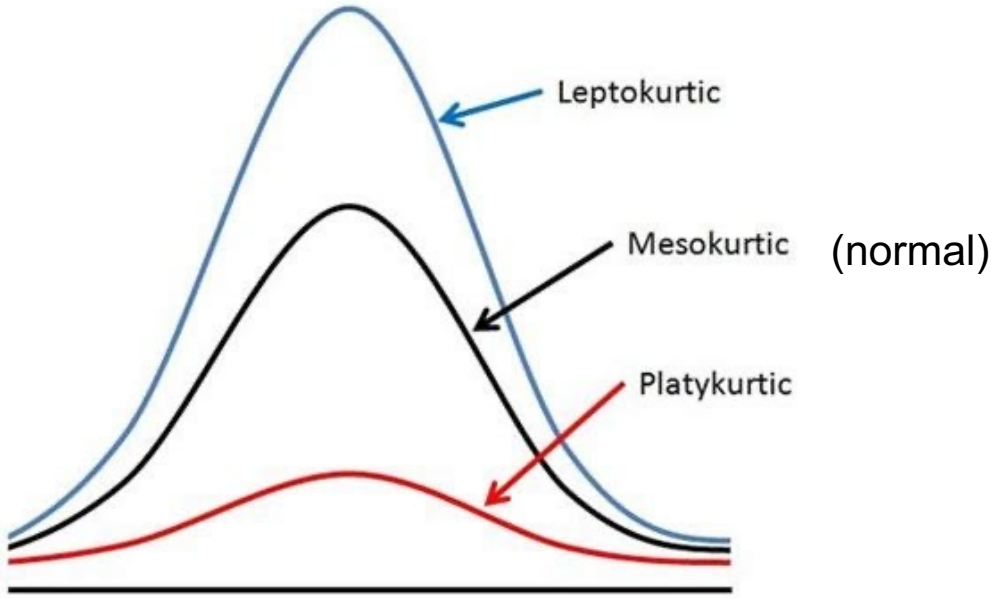
- **Negative Skewness:**
- distribution has a long left tail

Data Presentation (Numerical) - Shape

Kurtosis: Describe the distribution of observed data around the mean.

- **Positive Kurtosis** (leptokurtic)
- Distribution has a thin tall peak.

- **Negative Kurtosis** (platykurtic)
- distribution has a flat low peak.



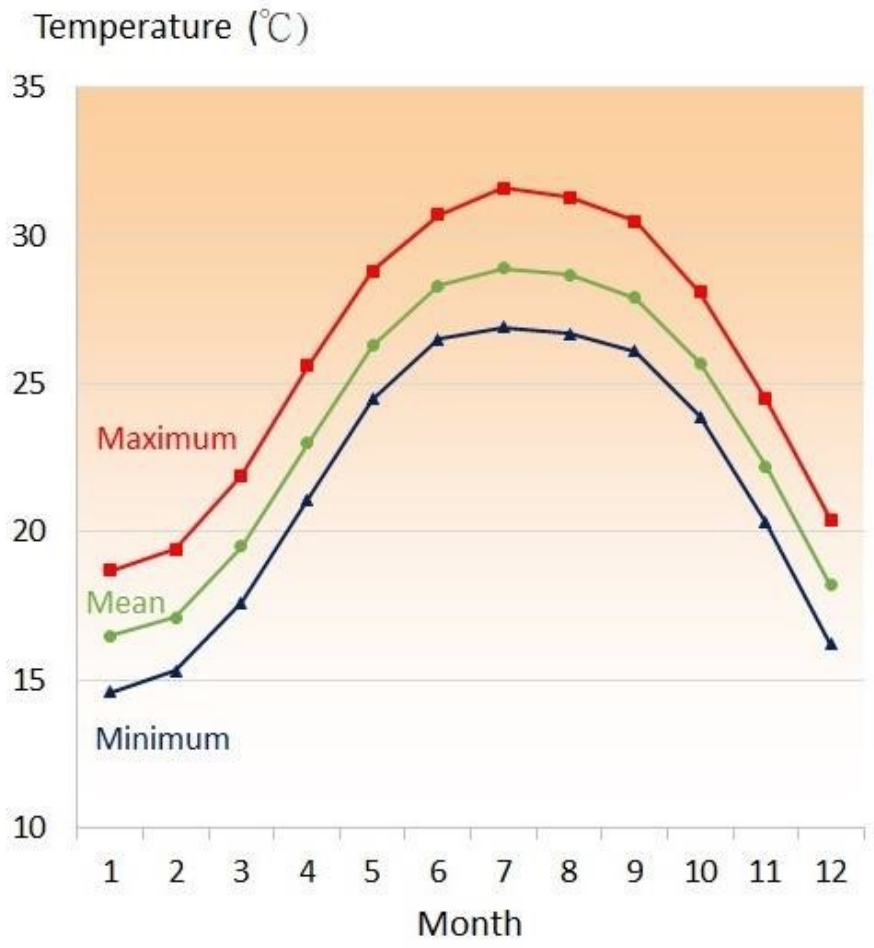
Data Presentation (Numerical) - Variability

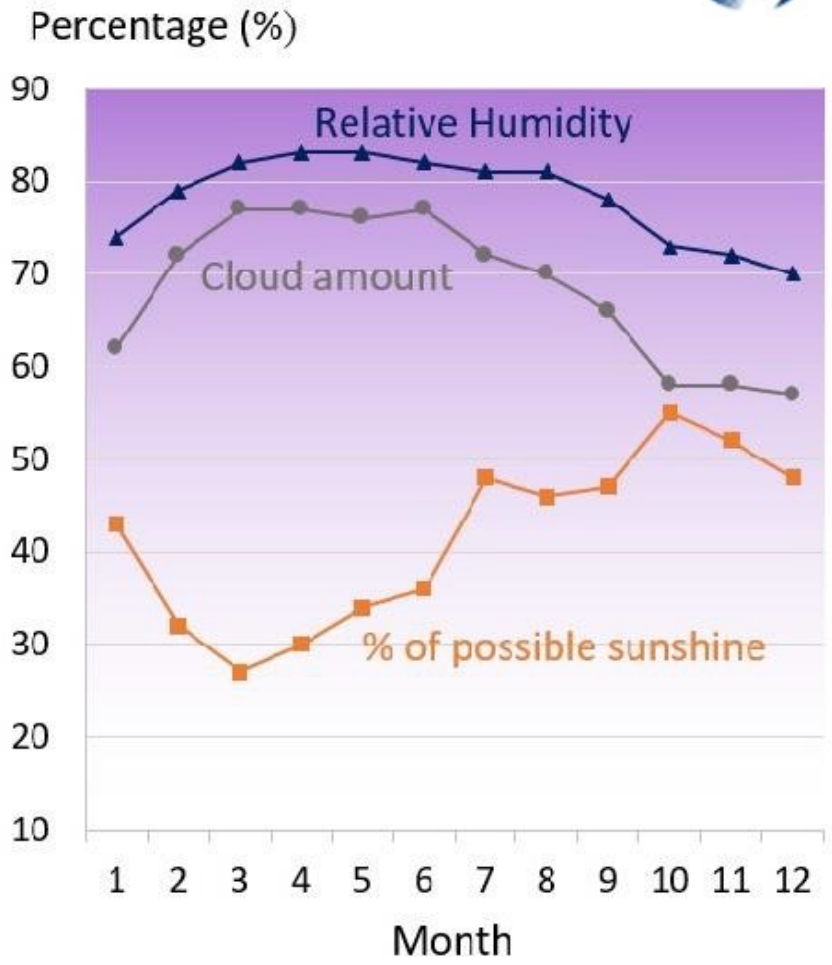
Why is the study of variability important?

- There is variability in virtually everything
- Allows us to distinguish between usual & unusual values
- Reporting only a measure of center doesn't provide a complete picture of the distribution.

Example : Variability

香港天文台
HONG KONG OBSERVATORY



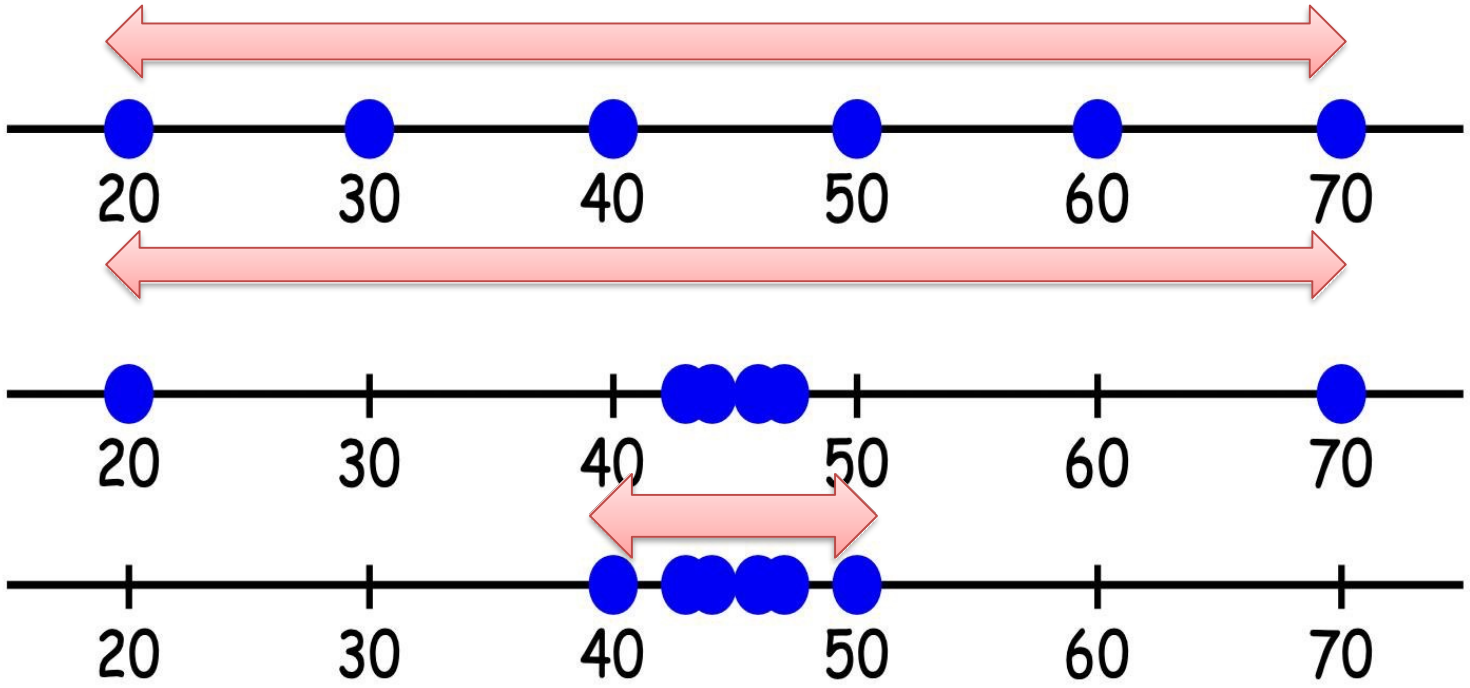


Monthly means of daily maximum, mean and minimum temperature(left), relative humidity, cloud amount recorded at the Hong Kong Observatory and percentage of possible sunshine at King's Park (right) between 1981-2010

CityU

Data Presentation (Numerical) – Variability - Range

What strikes you as the most distinctive difference among the distributions

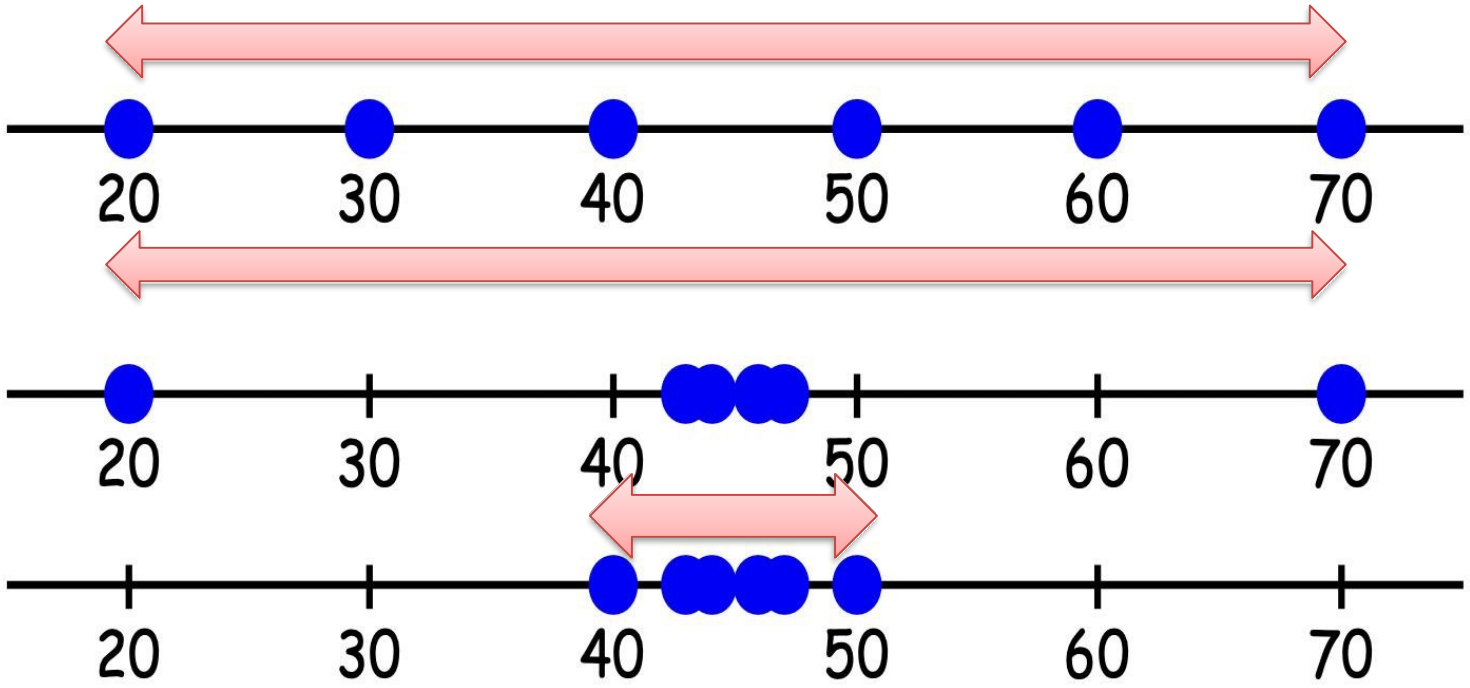


Notice that these three data sets all have the similar mean and median, but they have very different amounts of variability.

Data Presentation (Numerical) – Variability - Range

The simplest numeric measure of variability is range.

Range = largest observation – smallest observation



The first two data sets have a range of 50 (70-20) but the third data set has a much smaller range of 10.

Data Presentation (Numerical) – Variability -Deviation

Another measure of the variability in a data set uses the deviations from the mean $(x - \overline{x})$.

Remember the sample of 6 fish that we caught from the lake . . .
They were the following lengths:

3”, 4”, 5”, 6”, 8”, 10”

The mean length was 6 inches. Recall that we calculated the deviations from the mean.
What was the sum of these deviations?

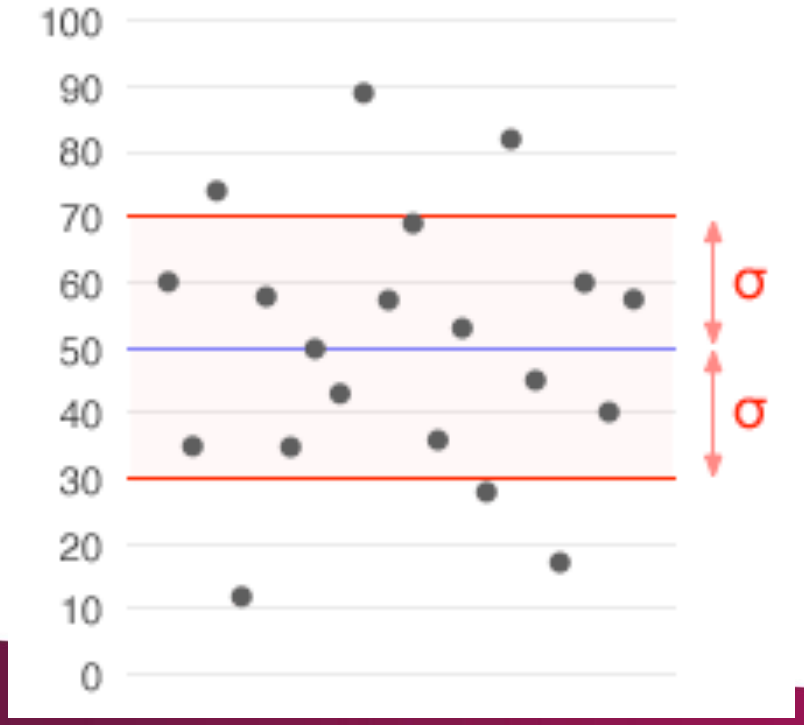
The estimated **average of the deviations squared** is called the **variance**.

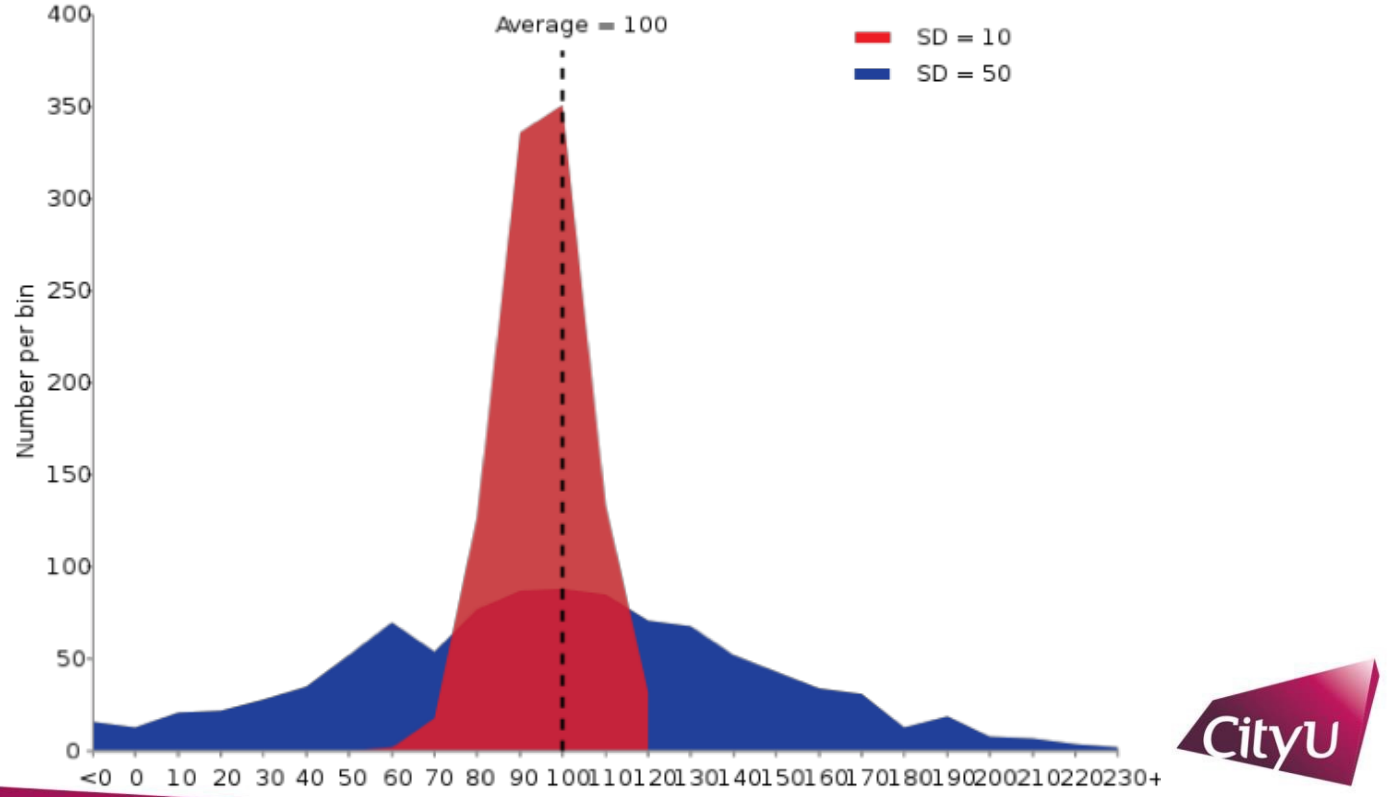
$$s^2 = \frac{\sum (x - \overline{x})^2}{n - 1}$$

When calculating sample variance, we use degrees of freedom (n – 1) in the denominator instead of n because this tends to produce better estimates.

Data Presentation (Numerical) – Variability -Deviation

- Standard deviation is a widely used measurement of variability or diversity.
- A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.





Data Presentation (Numerical) – Variability -Deviation

Z-score

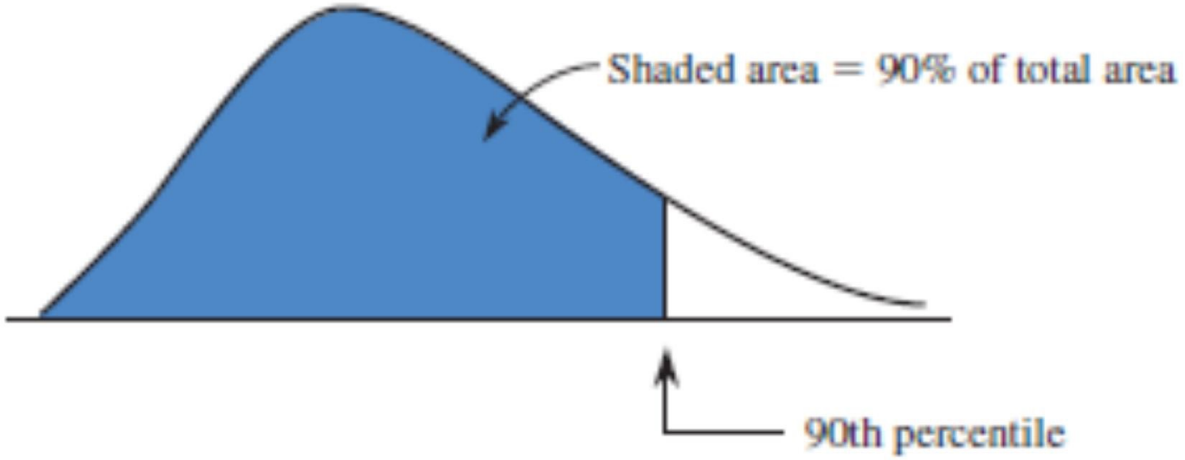
A z-score tells us how many standard deviations the value is from the mean.

$$z\text{-}score = \frac{value\ -mean}{standard\ deviation}$$

Data Presentation (Numerical) – Variability -Deviation

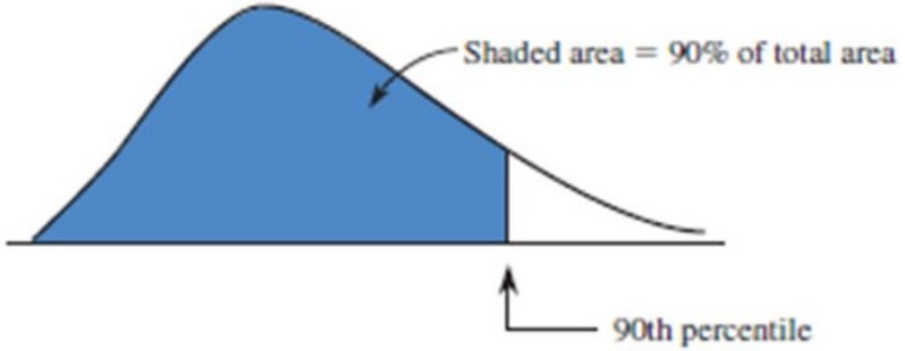
Percentiles

A percentile is a value in the data set where r percent of the observations fall AT or BELOW that value



Example : Z-score

The following summary values for annual rainfall (in cm) in a city



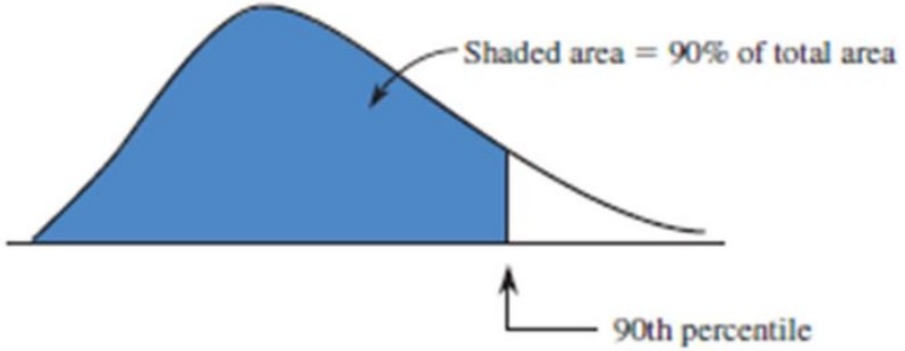
data	180	232	275	318	370	422	586
Percentile	5	10	25	50	75	90	95

1. What percent of annual rainfall greater than 370 cm?

2. 10% of annual rainfall bigger than what value?

Example : Z-score

The following summary values for annual rainfall (in cm) in a city



data	180	232	275	318	370	422	586
Percentile	5	10	25	50	75	90	95

1. What percent of annual rainfall greater than 370 cm?

25%

2. 10% of annual rainfall bigger than what value?

422cm

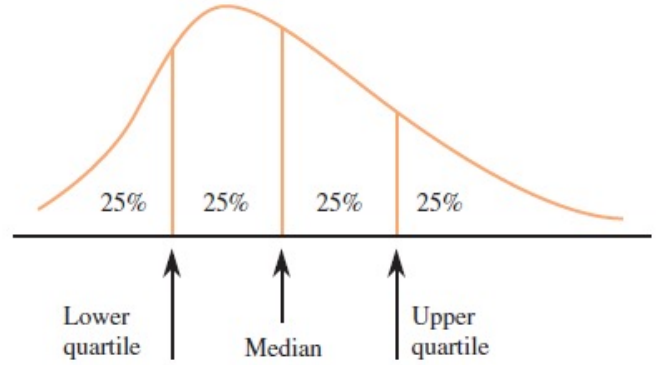
Data Presentation (Numerical) – Variability - iqr

The third measurement method of Variability:

Interquartile range (iqr) is the range of the middle half of the data.

Lower quartile (Q_1) is the median of the lower half of the data

Upper quartile (Q_3) is the median of the upper half of the data



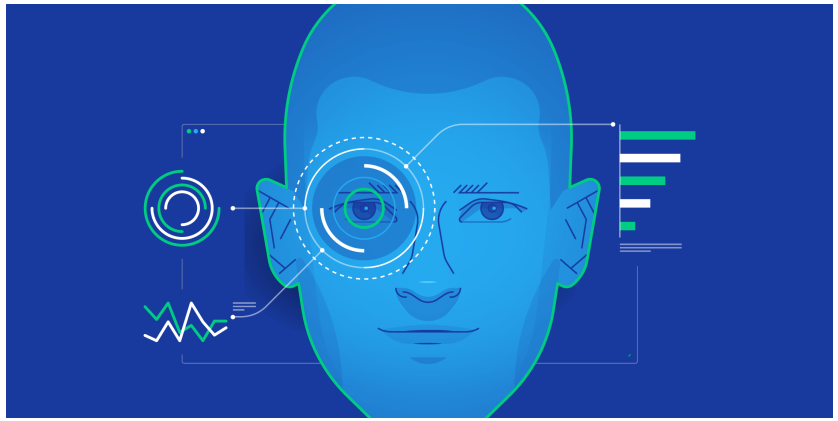
$$iqr = Q_3 - Q_1$$

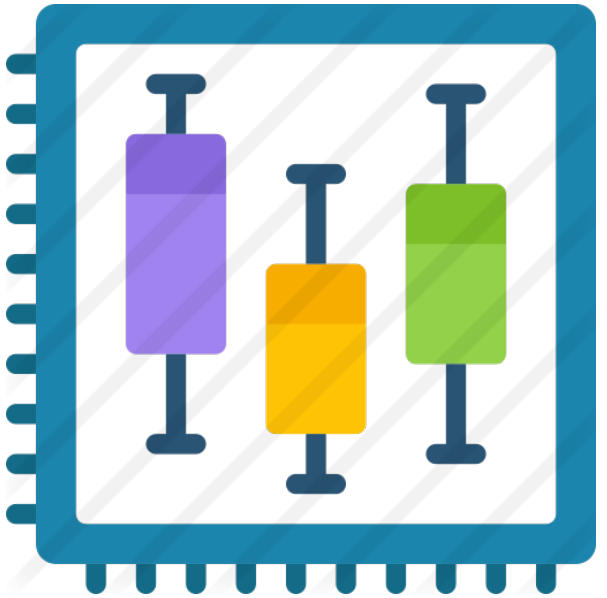
What advantage does the interquartile range have over the standard deviation?

-The iqr is resistant to extreme values

Data Presentation (Graphical) - Boxplot

We looked at ways of describing the center and variability of a data set using numerical measures. It would be nice to have a method of summarizing data that gives more detail than just a measure of center and spread and yet less detail than a stem-and-leaf display or histogram.

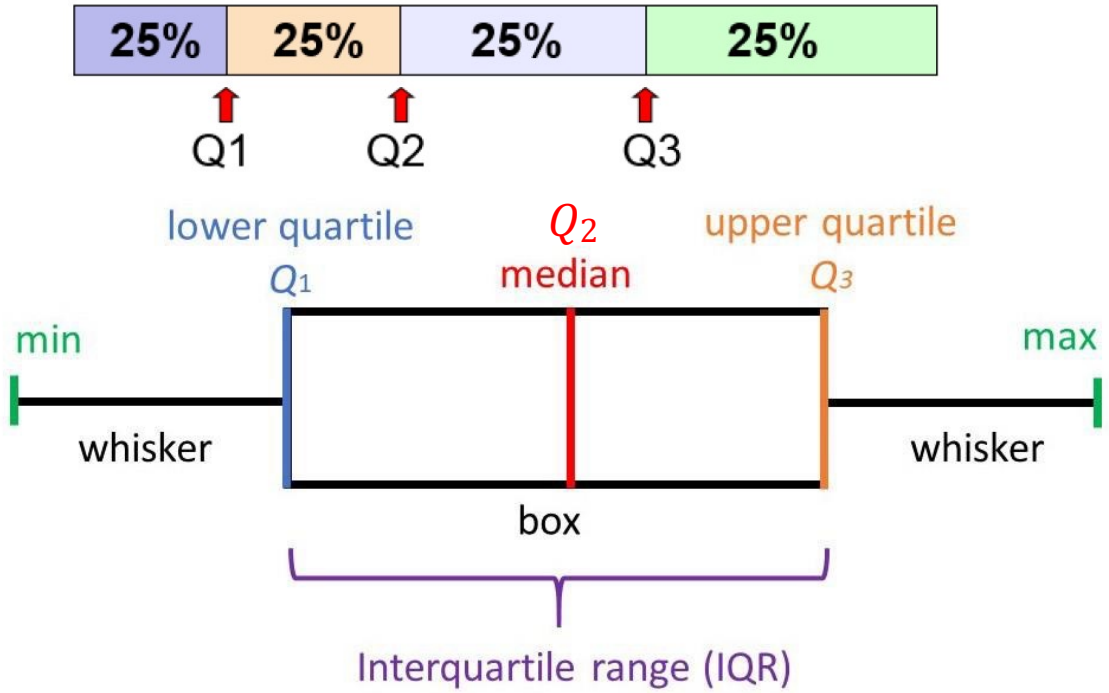




Data Presentation (Graphical) - Boxplot

A graph of showing five summary statistics which are the minimum, lower quartile, median, upper quartile and the maximum of the data points.

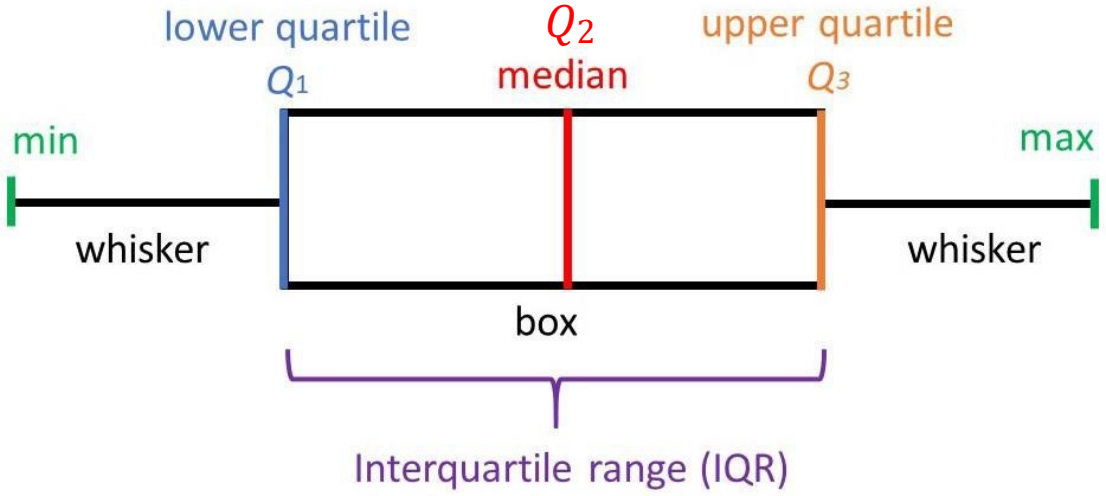
- **Quartiles** split the ordered data (in ascending order) into 4 segments with an equal number of data per segment.

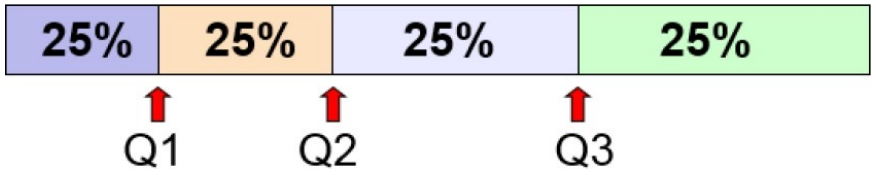


- The **median** (denoted by Q2), also known as the middle quartile, divides all data into two halves.
- **The lower and upper quartiles** (denoted by Q1 and Q3, respectively) are defined as follows:
  - 25% of the data lie below Q1, while
  - 25% of the data lie above Q3.

Data Presentation (Graphical) - Boxplot

- In practice, Q2 is used to be a measure of central tendency of the data, like the sample mean
- Q3-Q1 (known as the ***interquartile range (IQR)***) is used to be a measure of the data dispersion, like the sample variance or sample standard deviation.





Data Presentation (Graphical) - Boxplot

- Note that most often the number of data is NOT divisible by 4 so that we need a rule like the following one to determine the sample quartiles.

Let n be the total number of data, **p** = 0.25, 0.5 and 0.75 for the lower, middle and upper quartiles, respectively.

- ✓ **Case 1: If np+0.5 is an integer**, say m, then **the m_{th} smallest data point is the corresponding quartile**.

- ✓ **Case 2: If np+0.5 is NOT an integer**, then find the greatest integer just less than it, say, m and **the quartile is defined to be the average of the m_{th} and $(m+1)_{th}$ smallest data points**.

For example:

- ❖ n=10; 10 × 0.25 + 0.5 = 3 (case 1) ➔ m = 3; Q1 = the value of the 3rd datum
- ❖ n=9; 9 × 0.25 +0.5 = 2.75 (case 2) ➔ m = 2; Q1 = the average value of the 2nd datum and 3rd datum

Data Presentation (Graphical) - Boxplot

Boxplot is commonly used to detect the so-called ***outliers*** (unusual data points) which may be produced by measurement error and occurs with a small probability. Normally, outliers are far away from the majority of the data.

- The data point is labeled to be an outliers if its value is
  $\leq Q1 - 1.5IQR$ OR $\geq Q3 + 1.5IQR$

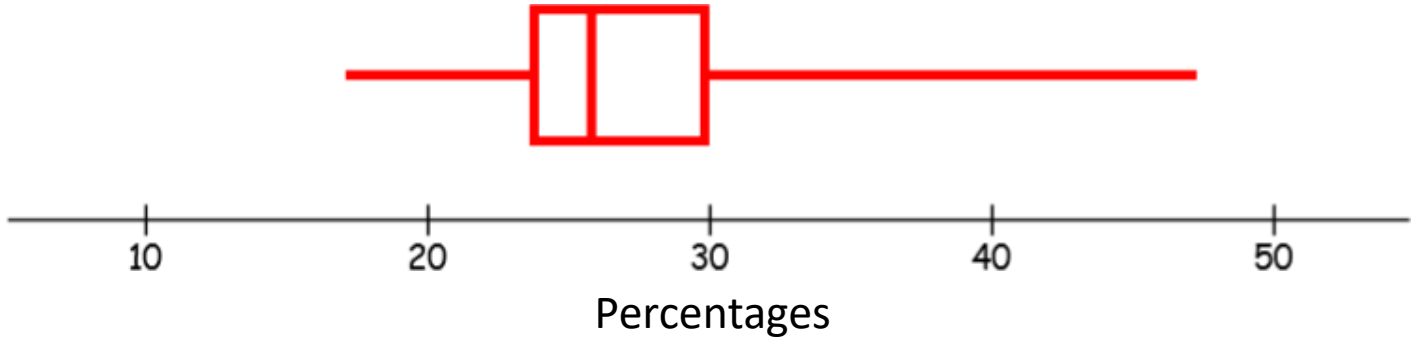
- According to this rule, if the data are from the normal distribution, we can show that there is a 0.0076 (0.76%) chance that the data points fall into the above regions.

- An outlier is extreme if it is more than 3(IQR) away from the nearest quartile
  $\leq Q1 - 3IQR$ OR $\geq Q3 + 3IQR$

Example : Boxplot

The percentage of waste reduction at the 51 monitoring sites

17	19	19	20	20	21	22	22	22	23
23	23	24	24	24	24	25	25	25	25
25	26	26	26	26	26	26	27	27	27
27	27	28	29	29	29	30	30	30	30
31	32	33	34	34	34	35	35	35	38
47									

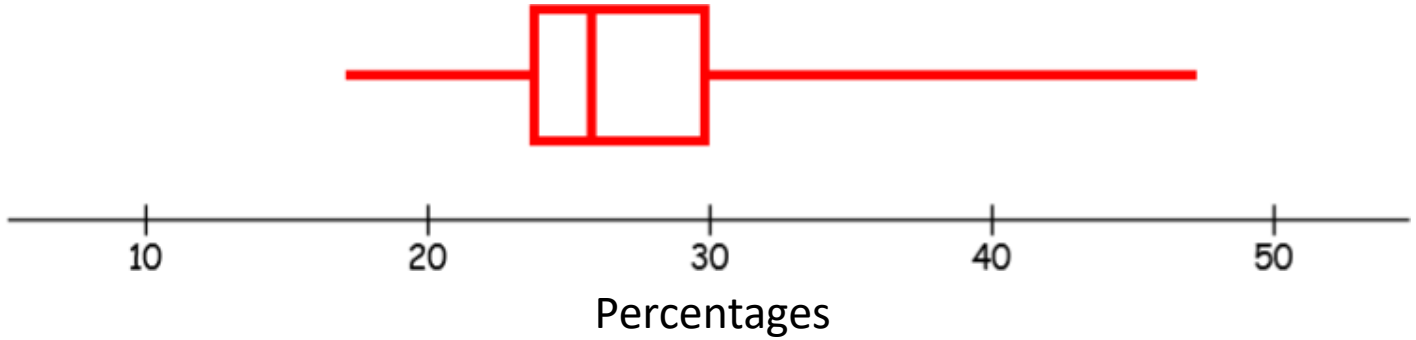


Example : Boxplot

The percentage of waste reduction at the 51 monitoring sites

17	19	19	20	20	21	22	22	22	23
23	23	24	24	24	24	25	25	25	25
25	26	26	26	26	26	26	27	27	27
27	27	28	29	29	29	30	30	30	30
31	32	33	34	34	34	35	35	35	38
47									

Q1=24
Q2=26
Q3=30
IQR=6



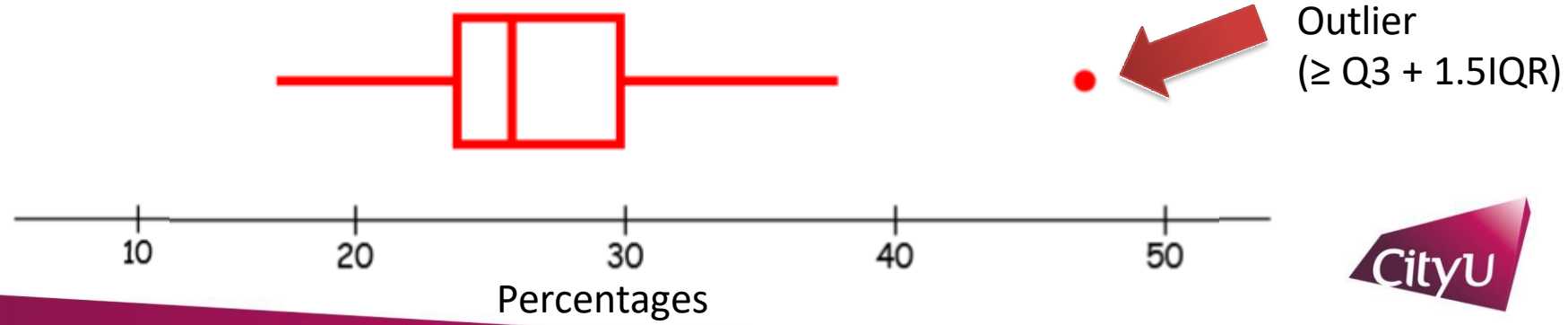
Outlier
(≥ Q3 + 1.5IQR=30 +1.5*6=39)

Data Presentation (Graphical) –Boxplot (Modified)

Modified Boxplot: whiskers extend to largest (or smallest) data observation that is not an outlier

The percentage of waste reduction at the 51 monitoring sites

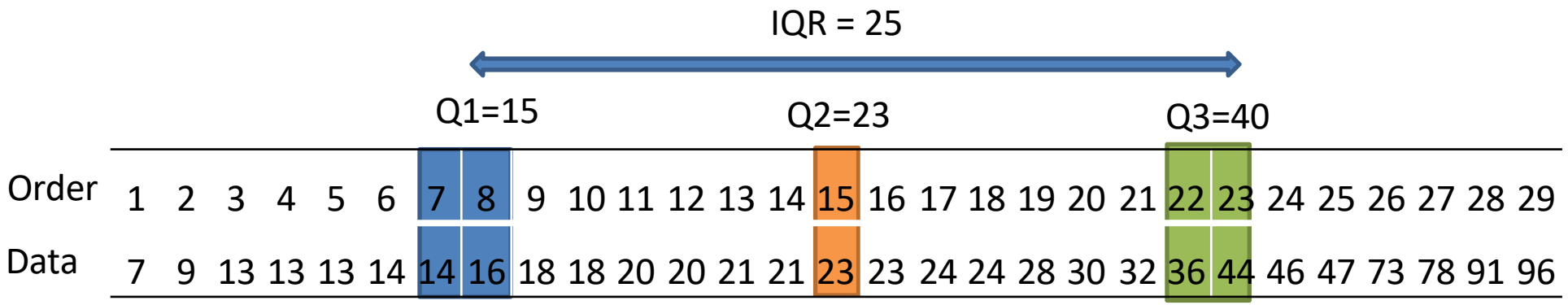
17	19	19	20	20	21	22	22	22	23
23	23	24	24	24	24	25	25	25	25
25	26	26	26	26	26	26	27	27	27
27	27	28	29	29	29	30	30	30	30
31	32	33	34	34	34	35	35	35	38
47									



Example : Modified Boxplot

Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Data	7	9	13	13	13	14	14	16	18	18	20	20	21	21	23	23	24	24	28	30	32	36	44	46	47	73	78	91	96

Example : Modified Boxplot



Example : Modified Boxplot

IQR = 25

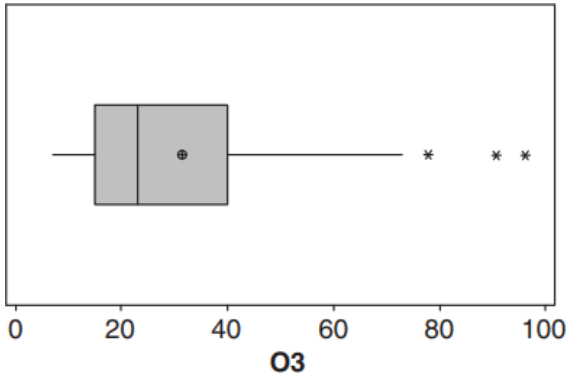
Q1=15 Q2=23 Q3=40

Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Data	7	9	13	13	13	14	14	16	18	18	20	20	21	21	23	23	24	24	28	30	32	36	44	46	47	73	78	91	96

Check outliers:

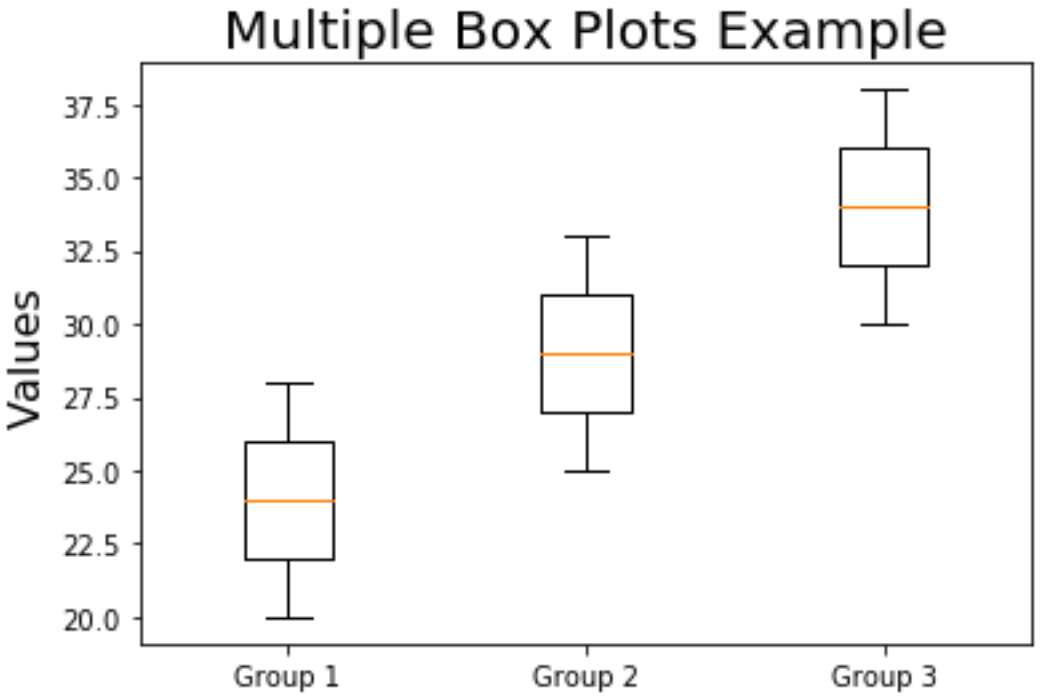
Outlier ≤ Q1 - 1.5IQR =15-1.5×25=-22.5.

Outlier ≥ Q3 + 1.5IQR=40+1.5×25=77.5

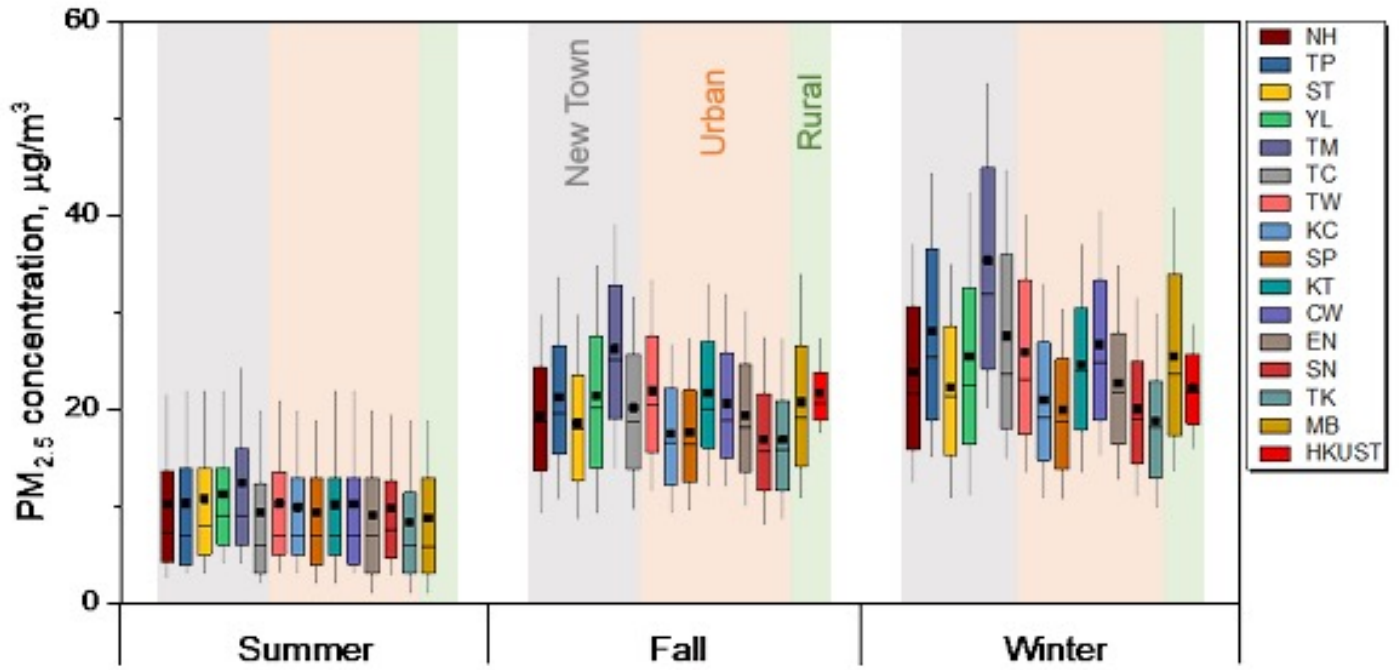


Example –Boxplot with python

```
# Multiple sets of example data
data1 = [20, 21, 22, 23, 24, 25, 26, 27, 28]
data2 = [25, 26, 27, 28, 29, 30, 31, 32, 33]
data3 = [30, 31, 32, 33, 34, 35, 36, 37, 38]

# Create side-by-side box plots
plt.boxplot([data1, data2, data3])

# Set the title and labels
plt.title('Multiple Box Plots Example', fontsize=20)
plt.ylabel('Values', fontsize=16)
plt.xticks([1, 2, 3], ['Group 1', 'Group 2', 'Group 3'])
plt.show()
```



Example –Boxplot



Box plot of the PM2.5 concentration among the 15 HKEPD general air quality monitoring stations and the HKUST supersite during 10 Jul.-31 Dec. 2020 (squares and solid lines correspond to mean and median values, respectively; boxes indicate the 25th and 75th percentile, and whiskers are the 10th and 90th percentile)

THANK YOU!