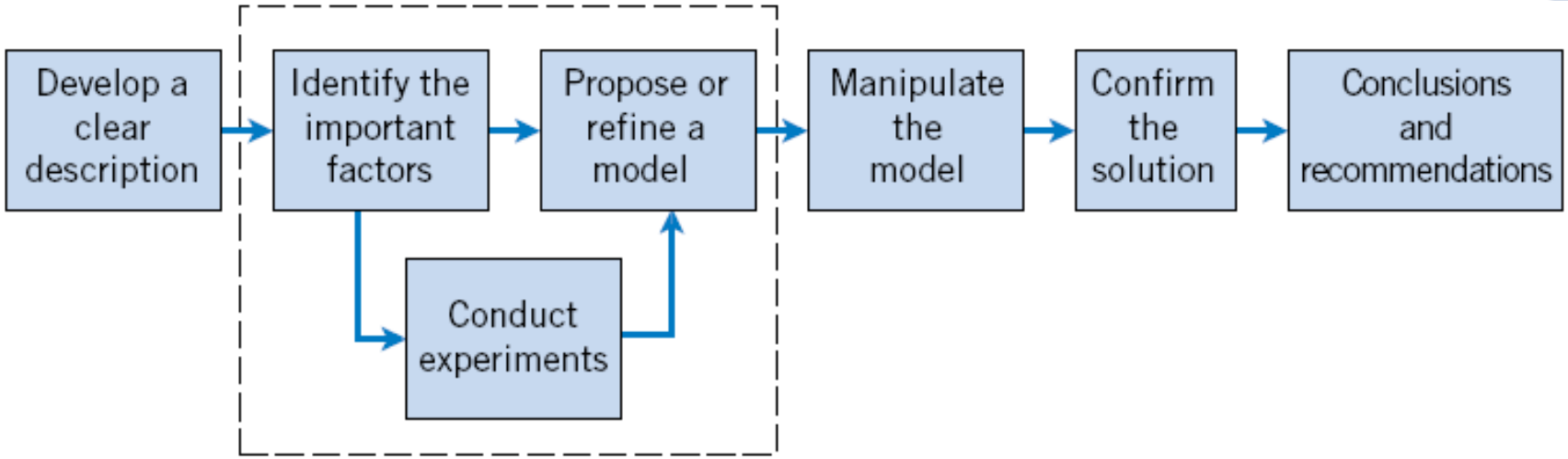
CIVL 7012/8012

Collection and Analysis of Information

Uncertainty in Engineering

"Statistics deals with the collection and analysis of data to solve real-world problems."

- Uncertainty is inherent in all real world problems.
- Two types of experiments:
  - Designed experiments
  - Observational studies
- Two types of data:
  - Continuous
  - Categorical
- Statistical methods are needed for describing information and forming the bases for design and decision making under uncertainty.


Develop a clear description
Identify the important factors
Propose or refine a model
Conduct experiments
Manipulate the model
Confirm the solution
Conclusions and recommendations

Variability

Variability results from three sources:

- Process
- Measurement
- Sampling

We must take care to select representative samples.

Well designed data collection and proper analysis techniques are essential for finding causes of variability and improving quality/productivity.

*Avoid over-controlling processes.

Variability

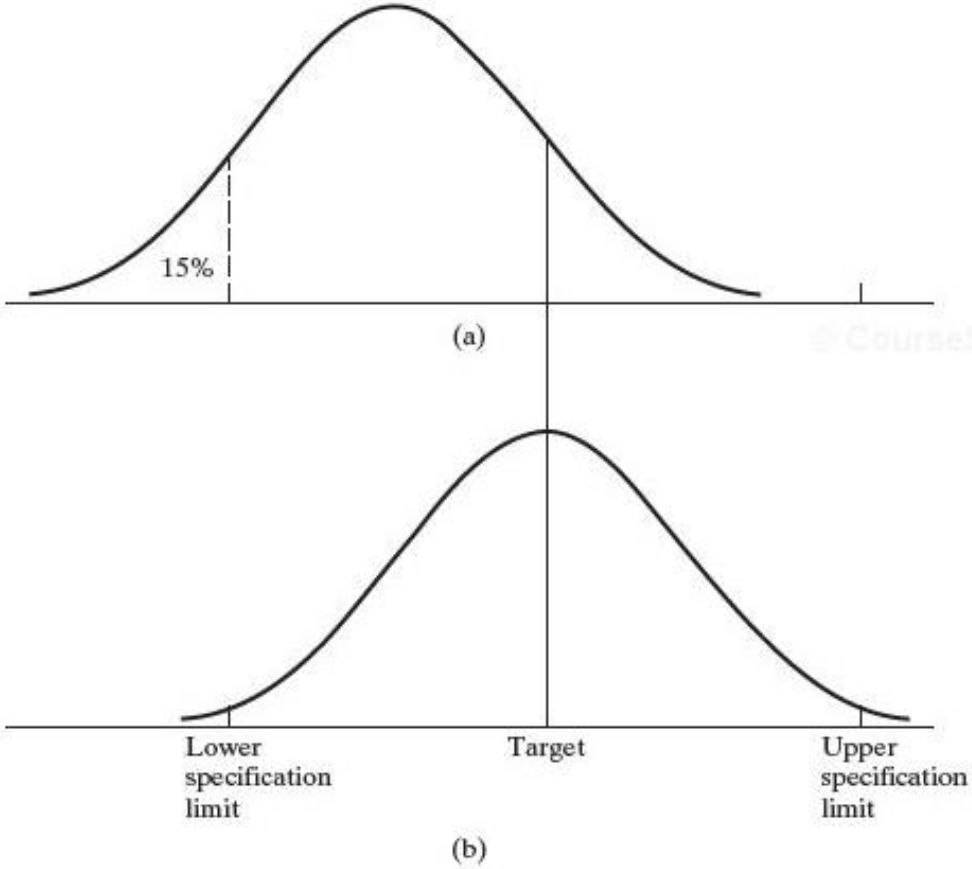


Figure 1.1-3 Two distributions compared with the target and lower/upper specification limits

Variability

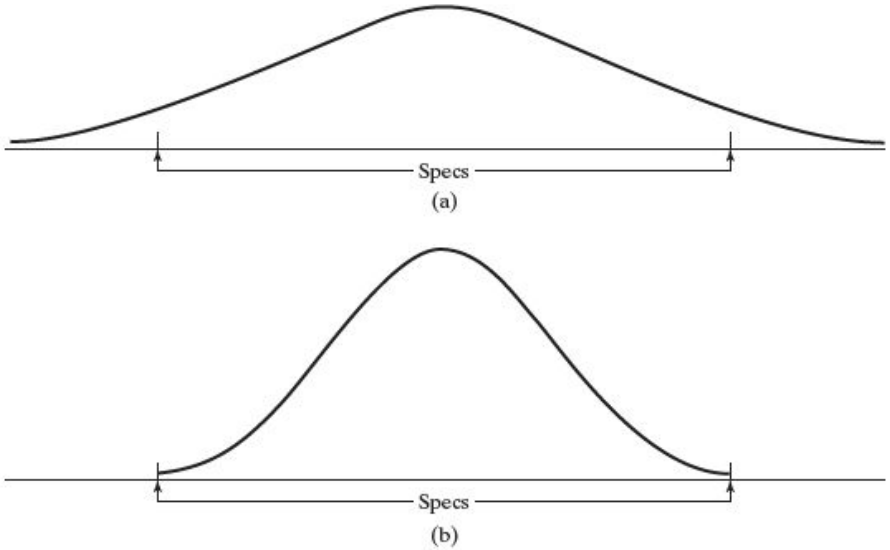


Figure 1.1-4 Two distributions compared with their specification limits

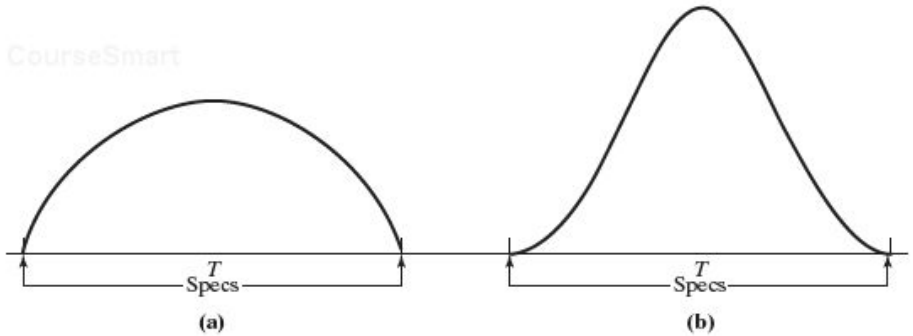


Figure 1.1-5 Two distributions resulting in zero defects

Measurements Collected Over Time

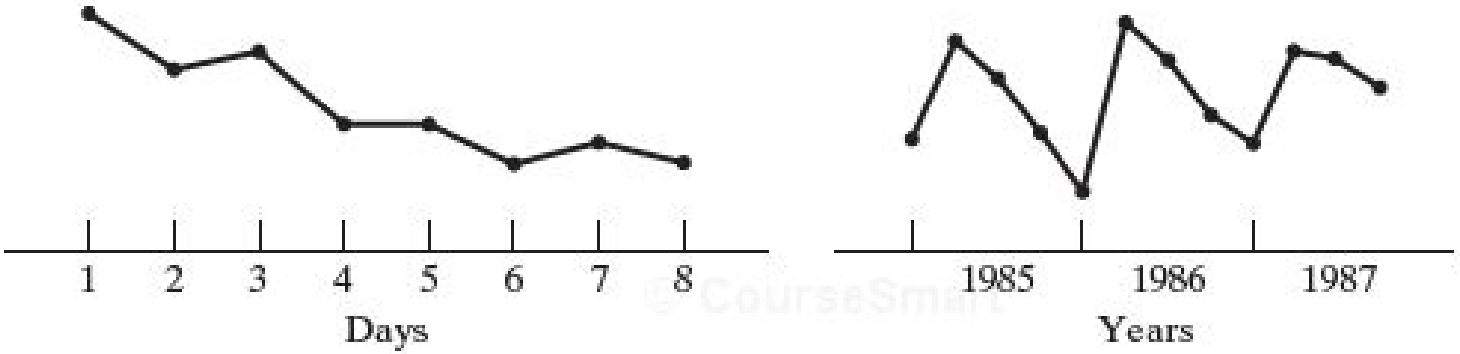


Figure 1.2-1 Two time sequences

Measurements Collected Over Time

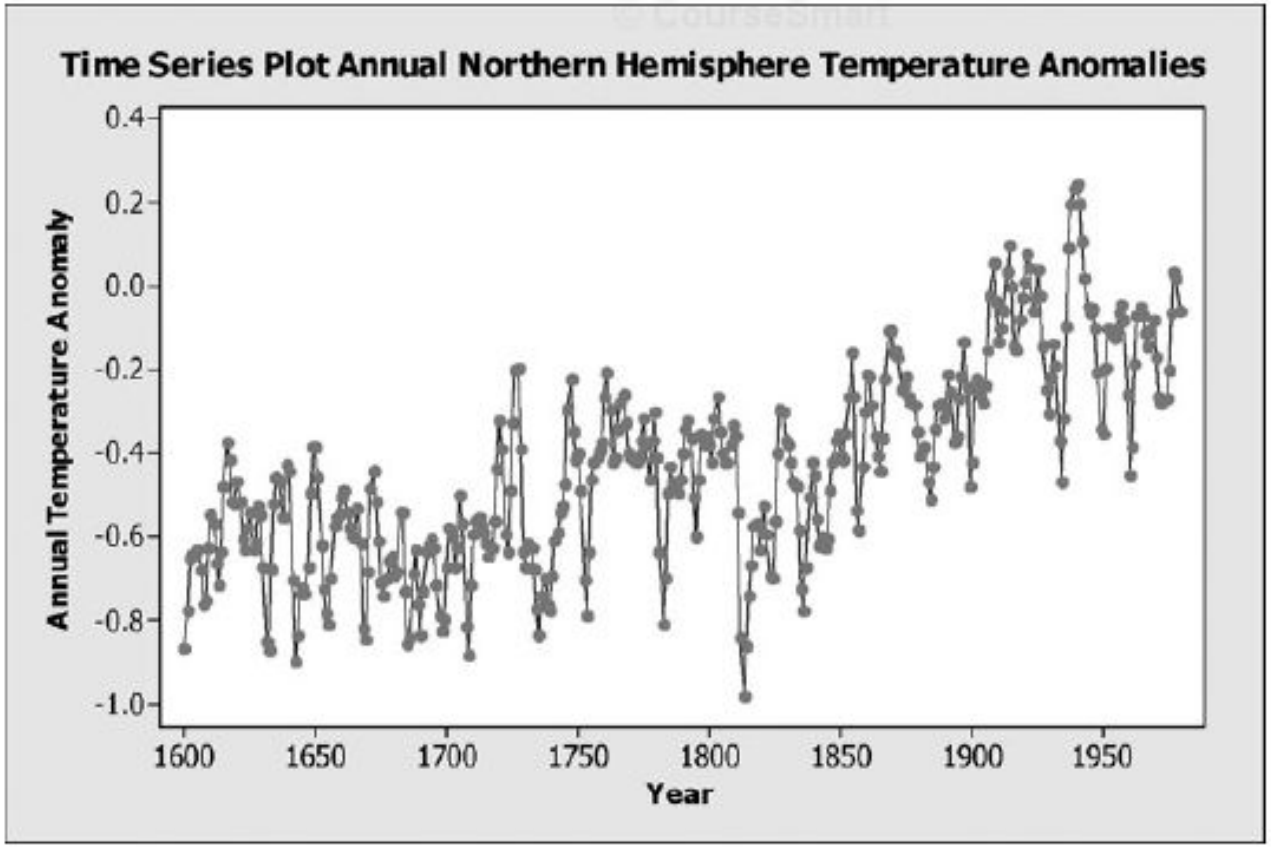


Figure 1.2-3 Time sequence plot of annual northern hemisphere temperature anomalies (1600 to 1997). This graph was created with Minitab

Measurements Collected Over Time

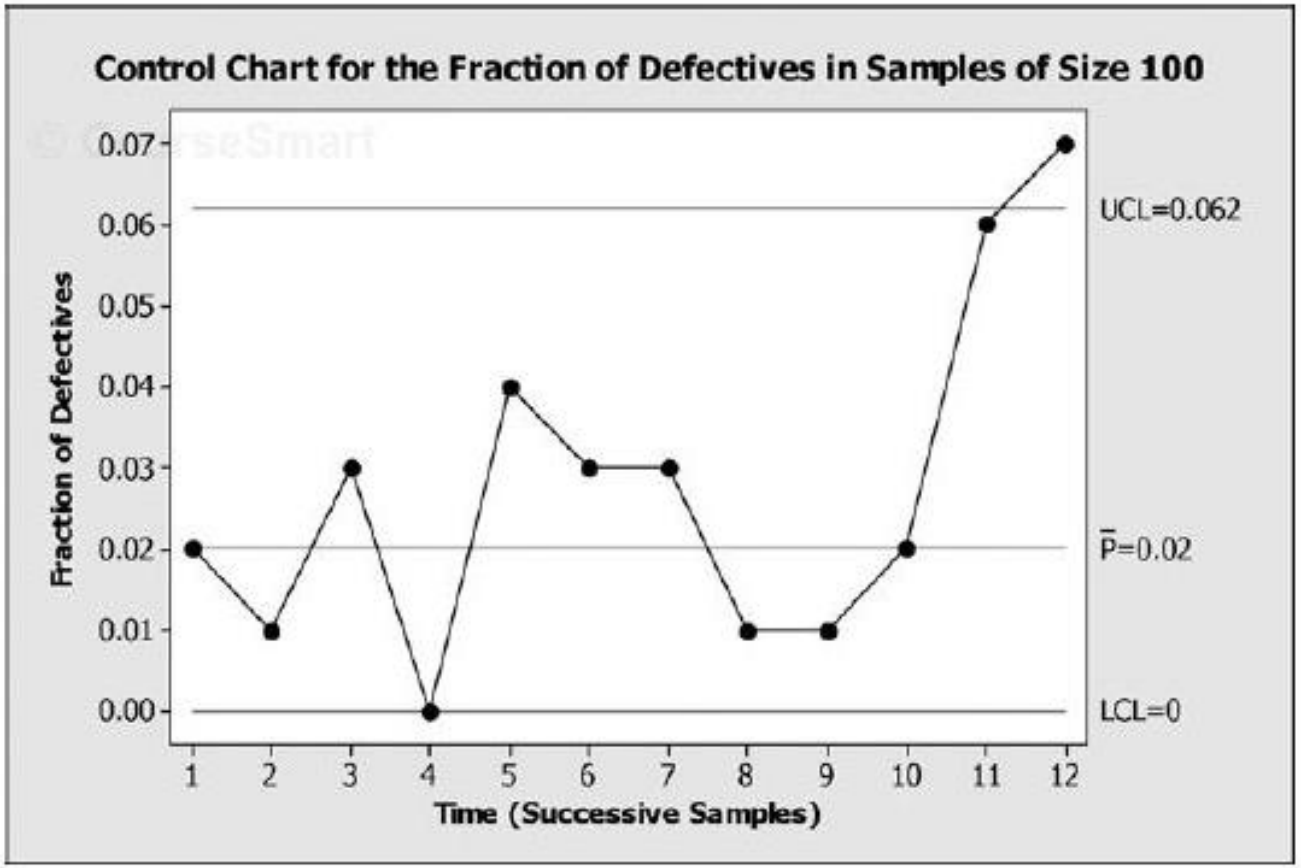


Figure 1.2-4 Control chart for the fraction of defectives from consecutive samples of size $n = 100$

Data Collection/Analysis Best Practices

- Proper design of experiments:
  - Plan BEFORE implementation
  - Avoid confounding factors
  - Blocking and randomization
- **Relevant** data
- Stratification
- Statistical significance
- Display data with clarity, efficiency

Types of data

- Cross-sectional data (a)
- Time series data (b)
- Pooled cross-sectional data (a+b)
- Panel data

Types of Data – Cross Sectional

- Cross-sectional data is a random sample

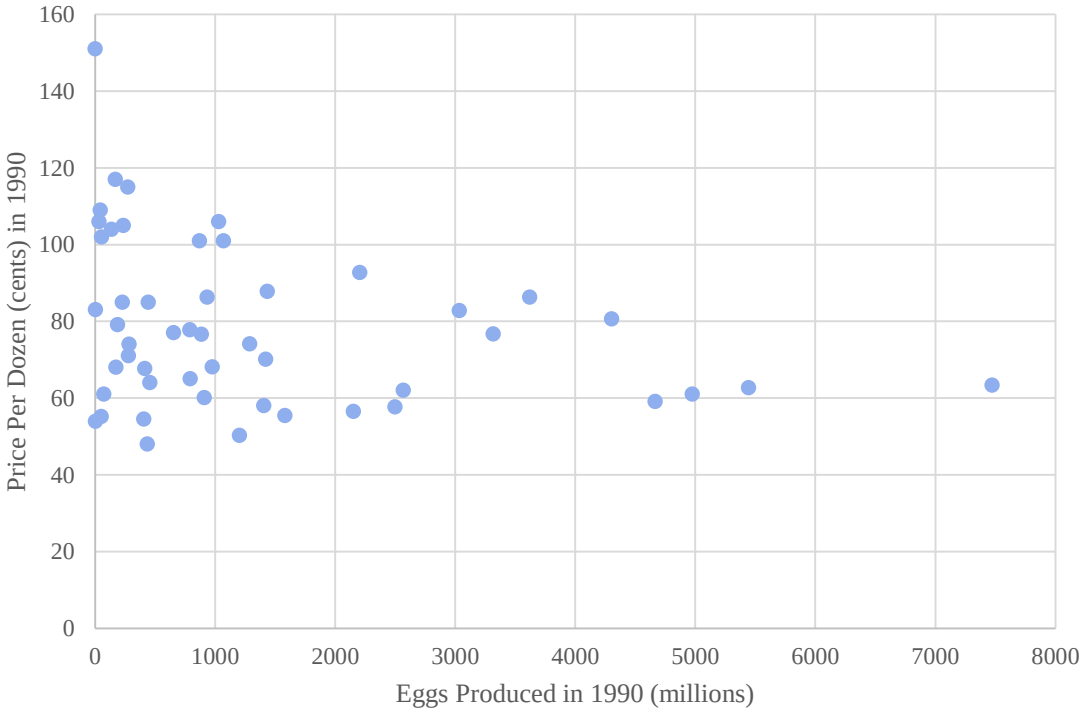
- Each observation is a new individual, firm, etc. with information at the same point in time (no temporal variation)
  - Example-1: census survey (every 10 years)
  - Example-2: National household travel survey (every five years)
  - Example-3: Commodity flow survey (every five years)

- If the data is not a random sample, we have a sample-selection problem

Example: cross-sectional data

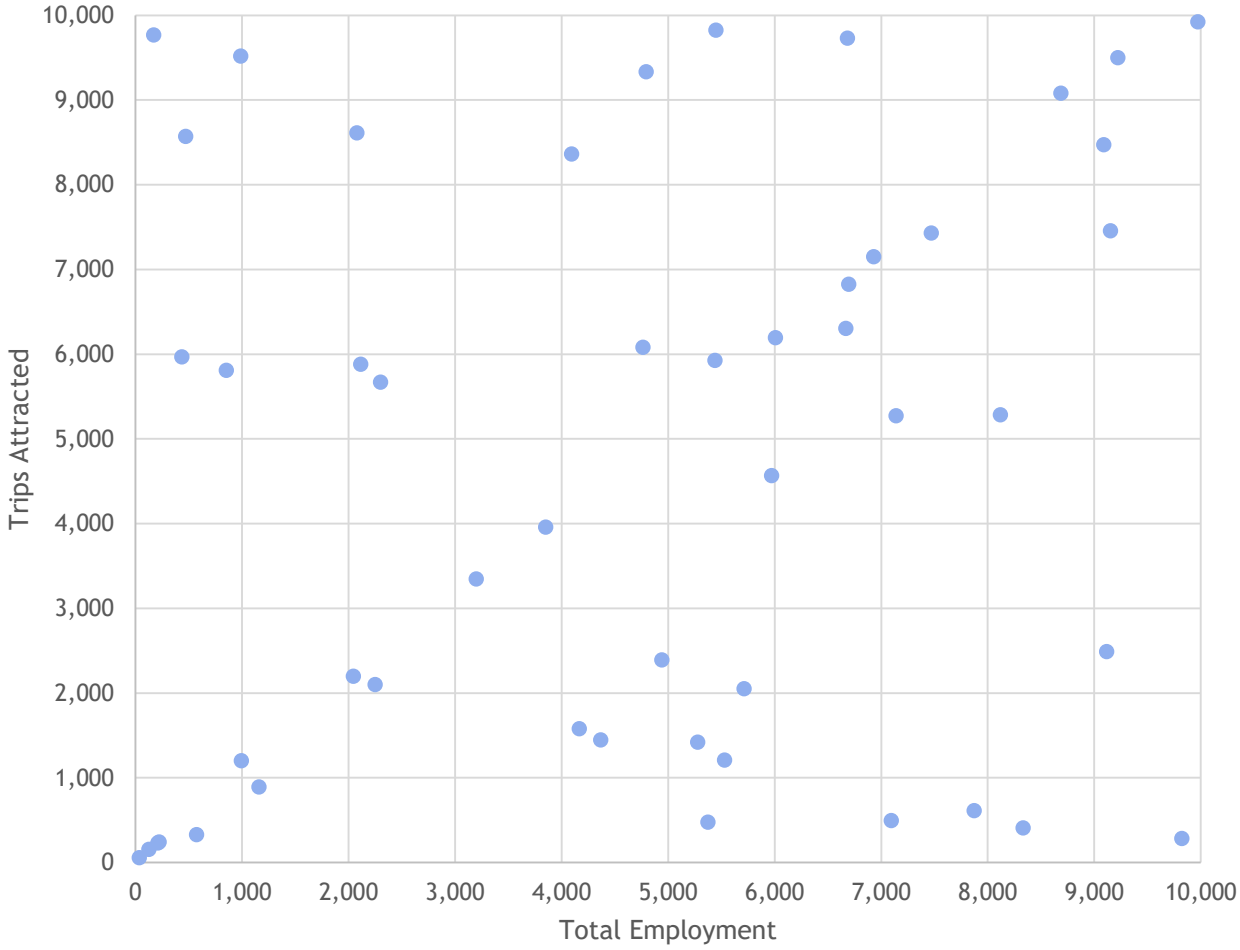
Zone	Peak Hr Trips Attracted(Y)	Total (X_1)	Manufacturing (X_2)	Retail & Services (X_3)	Other (X_4)
1	9,500	9,220	6,600	2,500	120
2	2,200	2,045	125	1,905	15
3	330	574	228	87	259
4	153	127	0	127	0
5	3,960	3,850	2,750	800	300
6	1,200	995	105	805	85
7	240	223	165	58	0
8	55	36	6	30	0
9	2,100	2,250	1,560	515	175
10	230	209	36	173	0

Cross-sectional data (cautions) -*1*



- Heterogeneity
- *Size* and *scale* effect must be taken into consideration otherwise model building will be erroneous
- Not to mix apples with oranges

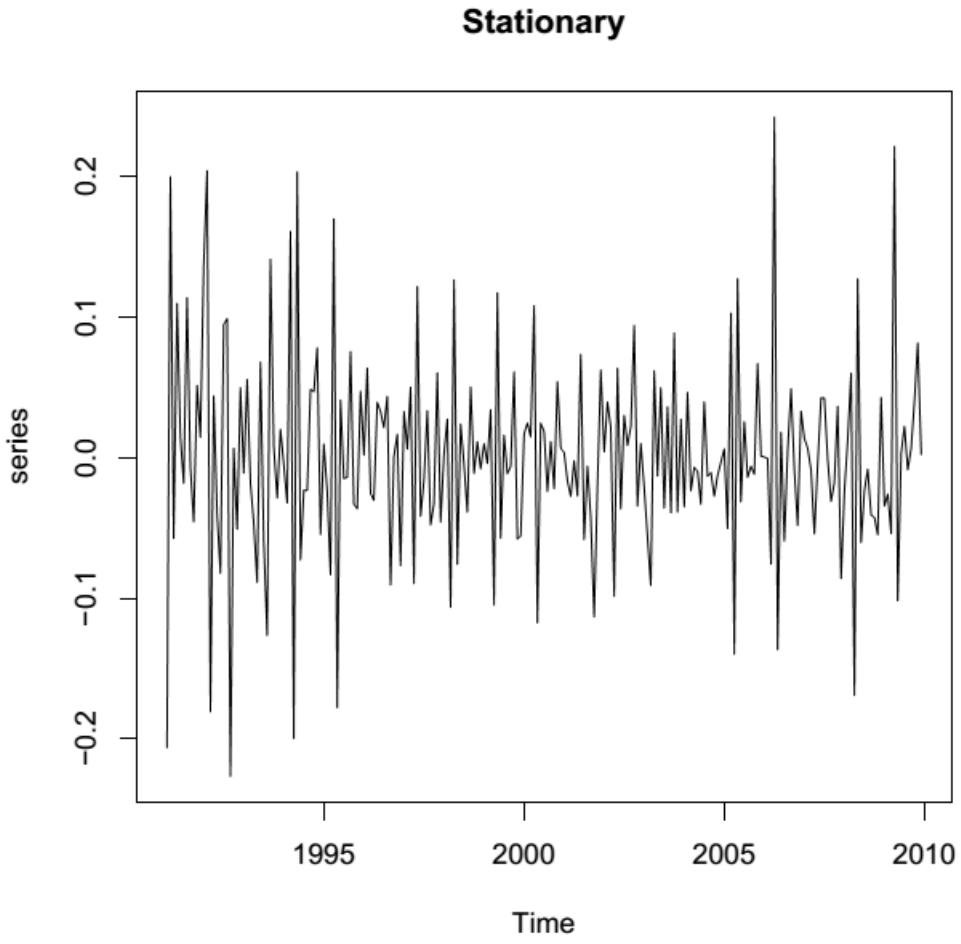
Cross-sectional data (cautions) -2



Types of Data – Time Series

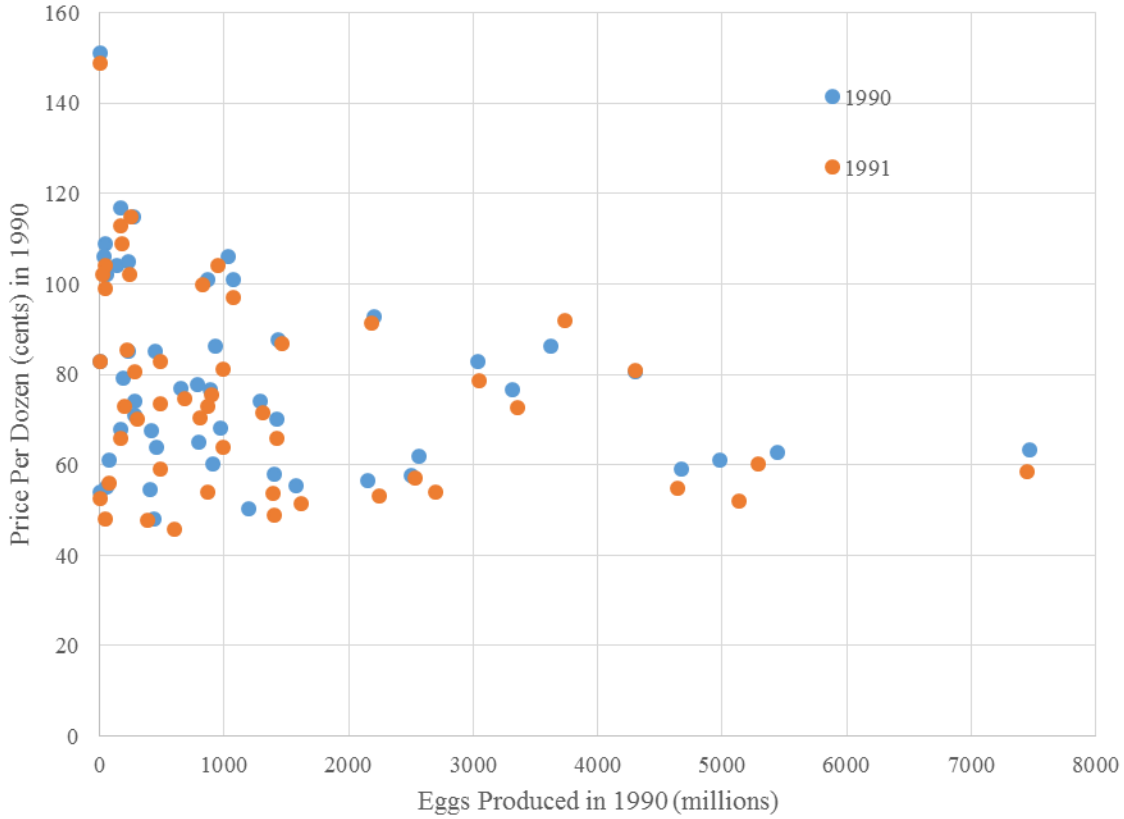
- Time series data has a separate observation for each time period – e.g. annual traffic volume on a corridor, census observations over multiple decades
- Time periods to consider
  - Daily, Weekly, Monthly, Quarterly, Annually, Quinquennially (every five years), Decennially (every 10 years)
- Since not a random sample, different problems to consider
- Trends and seasonality will be important
- Stationary issue
  - Loosely speaking a time series is stationary if its mean and standard deviation does not vary systematically over time

Stationary time series-example



- Stationarity
- Seasonality
- Trend effects
- We will explore more in time series lectures

Types of Data-pooled data



- Data elements consists of both cross-sectional and time series features
- Random samples over time are different (the same observations in sample may not remain same)

Types of Data - Panel

- A panel datasets consists of a time series for each cross-sectional member of the dataset.
- Can pool random cross sections and treat similar to a normal cross section. Will just need to account for time differences.
- Can follow the same random individual observations over time – known as panel data or longitudinal data
- Also known as longitudinal and micropanel data

Panel data - example

Zone	Peak Hr Trips Attracted(Y)	Total Employment (X_1)	Manufacturing (X_2)	Retail & Services (X_3)	Other (X_4)
1	9500	9220	6600	2500	120
1	5649	4996	3679	666	1288
2	2200	2045	125	1905	15
2	1839	4239	4822	1595	930
3	330	574	228	87	259
3	7057	2326	1238	487	628
4	153	127	0	127	0
4	8754	2128	1128	778	487
5	3960	3850	2750	800	300
5	1408	5656	4126	1929	2361
6	1200	995	105	805	85
6	2328	5209	1178	1062	1539
7	240	223	165	58	0
7	1095	4646	1858	848	1842
8	55	36	6	30	0
8	9315	2433	1826	1092	2733
9	2100	2250	1560	515	175
9	5642	3662	1671	1700	1146
10	6738	9493	3490	1701	2285
10	230	209	36	173	0

Measurement scale of variables

- Ratio scale
- Interval scale
- Ordinal scale
- Nominal scale

Ratio scale

- Measurements made with ratio scale can be added, subtracted, multiplied and divided.
- For a variable X taking two values, X1 and X2, the ratio X1/X2 and X1-X2 are meaningful quantities
- Also there is a natural ordering (ascending and descending) of the values along scale (X1 <= X2)
- Most of the variables belong to this category
- Example: Height, weight, speed, distance, etc.

Interval scale

- In interval scale the difference is meaningful but it does not satisfy the ratio property.
- Example: A 50 mph speed limit is higher than 40mph but lower than 6o mph. The difference is meaningful (equal intervals)
- Interval scales normally have a minimum and maximum point.

Ordinal scale

- A variable is ordinal if it only satisfies the natural ordering.
- Ordinal scales classify subjects and rank them in terms of how they possesses characteristics of interest
- Example
  - Grading system
    - A, B, C grades
  - Income class
    - Upper, middle, and lower
- Ordering exists but not the differences and ratios

Nominal scale

- Variables in this category does not have any of the features of ratio scale
  - Ratio
  - Difference
  - Natural ordering
- Variables such as
  - Gender (male, female);
  - Facility type (freeway, arterial)
- Also called as categorical variables

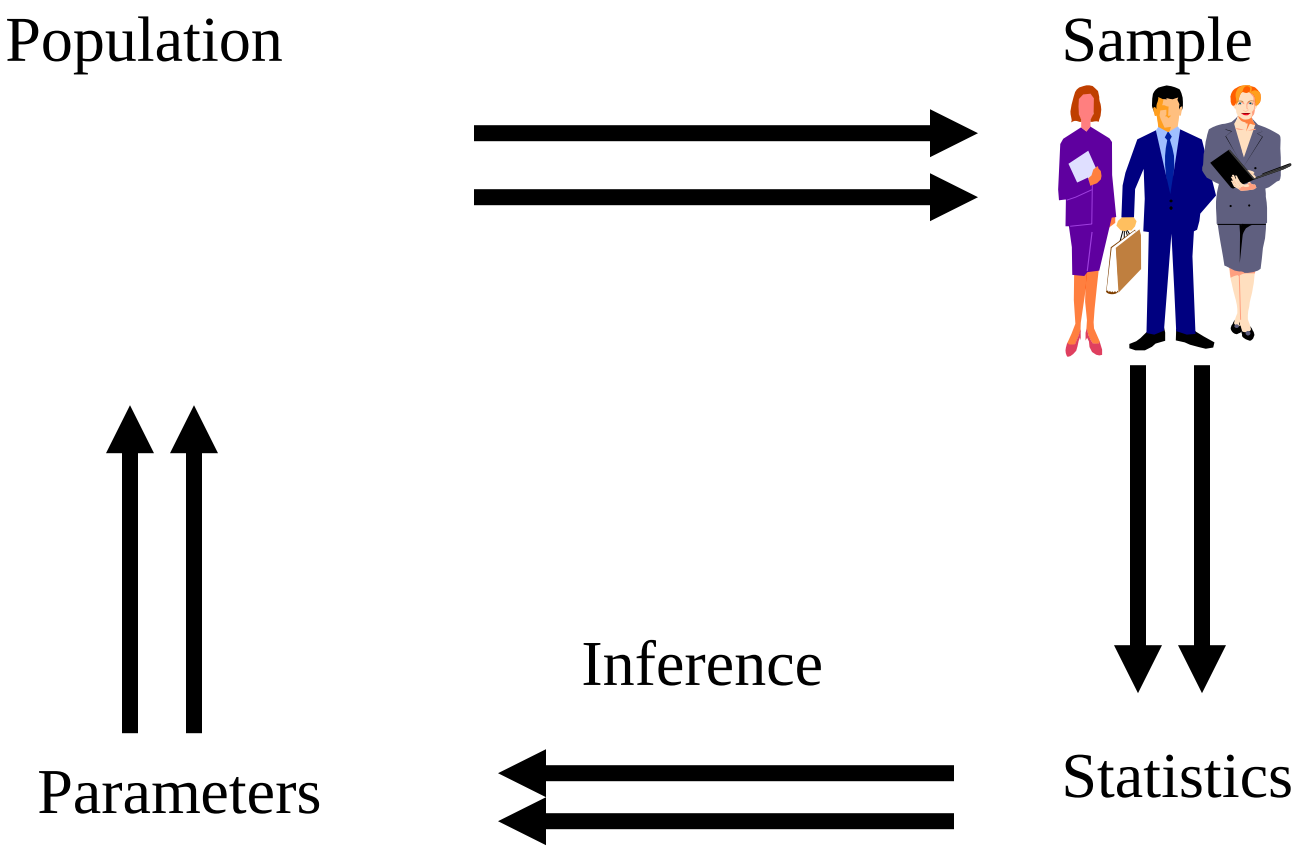
Descriptive Statistics

Descriptive vs. Inferential Statistics

DEFINITIONS

- **Population** – all members of a class or category of interest
- **Parameter** – a summary measure of the population (e.g. average)
- **Sample** – a portion or subset of the population collected as data
- **Observation** – an individual member of the sample (i.e., a data point)
- **Statistic** – a summary measure of the observations in a sample

Populations and Samples



Summary Statistics

- Measures of Central Tendency
  - Arithmetic mean
  - Median
  - Mode

- Measures of Dispersion
  - Variance
  - Standard deviation
  - Coefficient of variation (COV)

Measures of Central Tendency

8 6 4 10 3 8 4 8 5

The ***sample mean*** is given by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{8+6+4+10+3+8+4+8+5}{9} = 6.22$$

The ***sample median*** is given by:

3 4 4 5 (6) 8 8 8 10

Measures of Central Tendency

The ***mode*** of the sample is the value that occurs most frequently.

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

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

Bimodal

Measures of Dispersion

The most common measure of dispersion is the ***sample variance***:

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n-1}$$

The ***sample standard deviation*** is the square root of sample variance:

$$s = \sqrt{s^2}$$

Measures of Dispersion

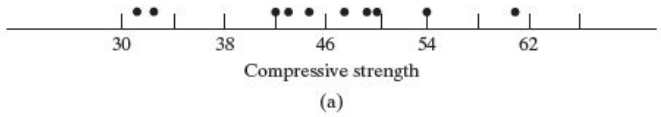
Coefficient of variation **(CV):**

$$CV = \frac{100s}{\bar{x}}$$

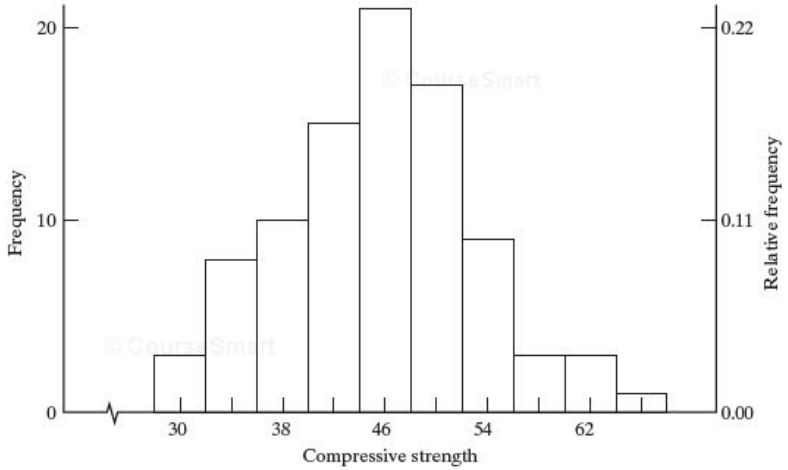
This is a good way to compare measures of dispersion between different samples whose values don't necessarily have the same magnitude (or, for that matter, the same units!).

Data Summary

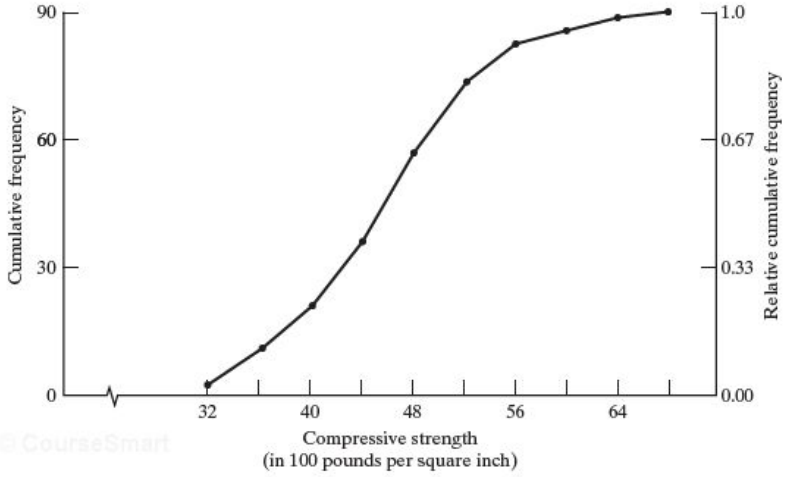
Table 1.3-1 Compressive Strength of Concrete Blocks (100 Pounds per Square Inch)									
49.2	53.9	50.0	44.5	42.2	42.3	32.3	31.3	60.9	47.5
43.5	37.9	41.1	57.6	40.2	45.3	51.7	52.3	45.7	53.4
51.0	45.7	45.9	50.0	32.5	67.2	55.1	59.6	48.6	50.3
45.1	46.8	47.4	38.3	41.5	44.0	62.2	62.9	56.3	35.8
38.3	33.5	48.5	47.4	49.6	41.3	55.2	52.1	34.3	31.6
38.2	46.0	47.0	41.2	39.8	48.4	49.2	32.8	47.9	43.3
49.3	54.5	54.1	44.5	46.2	44.4	45.1	41.5	43.4	39.1
39.1	41.6	43.1	43.7	48.8	37.2	33.6	28.7	33.8	37.4
43.5	44.2	53.0	45.1	51.9	50.6	48.5	39.0	47.3	48.8


30
38
46
54
62
Compressive strength

(a)


20
10
0
Frequency
0.22
0.11
0.00
Relative frequency
30
38
46
54
62
Compressive strength

(b)


90
60
30
0
Cumulative frequency
1.0
0.67
0.33
0.00
Relative cumulative frequency
32
40
48
56
64
Compressive strength
(in 100 pounds per square inch)

(c)

Frequency Distribution

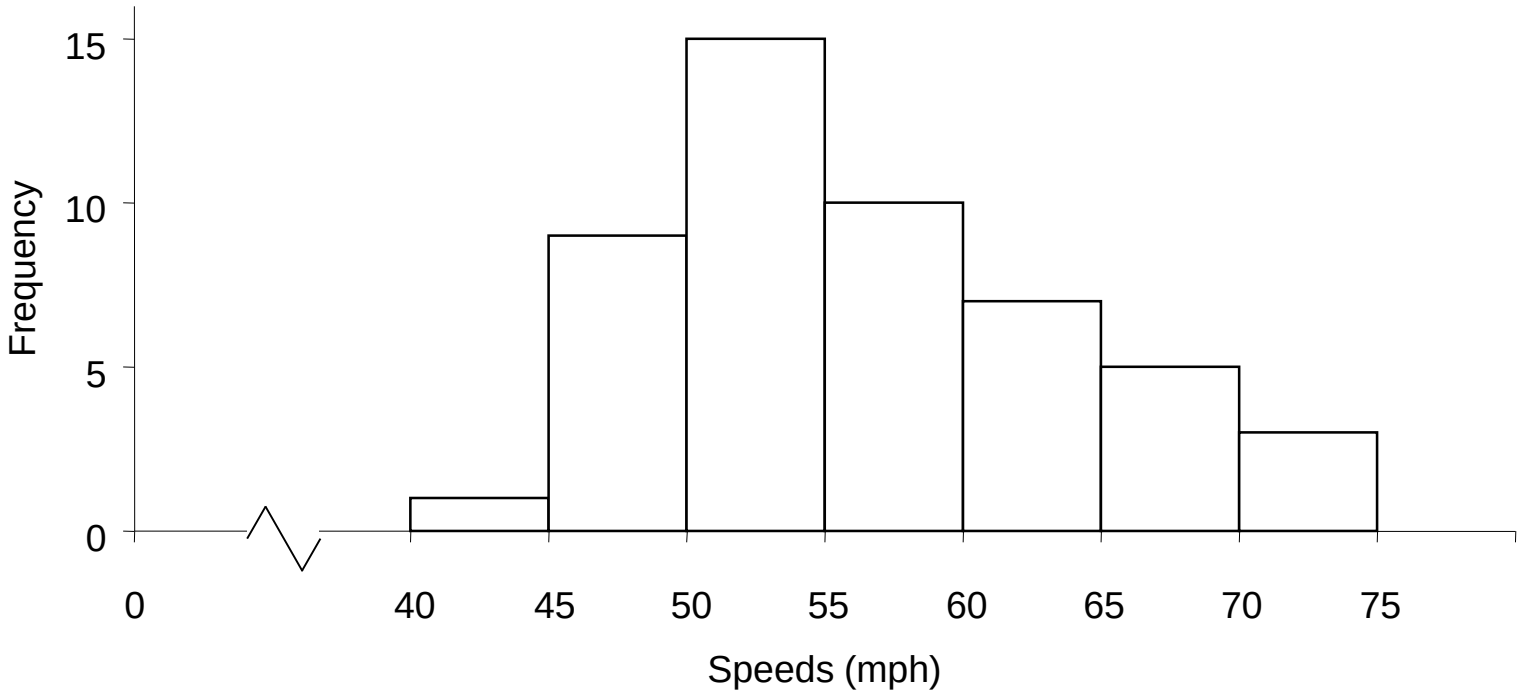
Vehicle Speeds on Central Avenue

Speeds (mph)	Vehicles Counted
40–45	1
45–50	9
50–55	15
55–60	10
60–65	7
65–70	5
70–75	3
TOTAL	50

Class Intervals

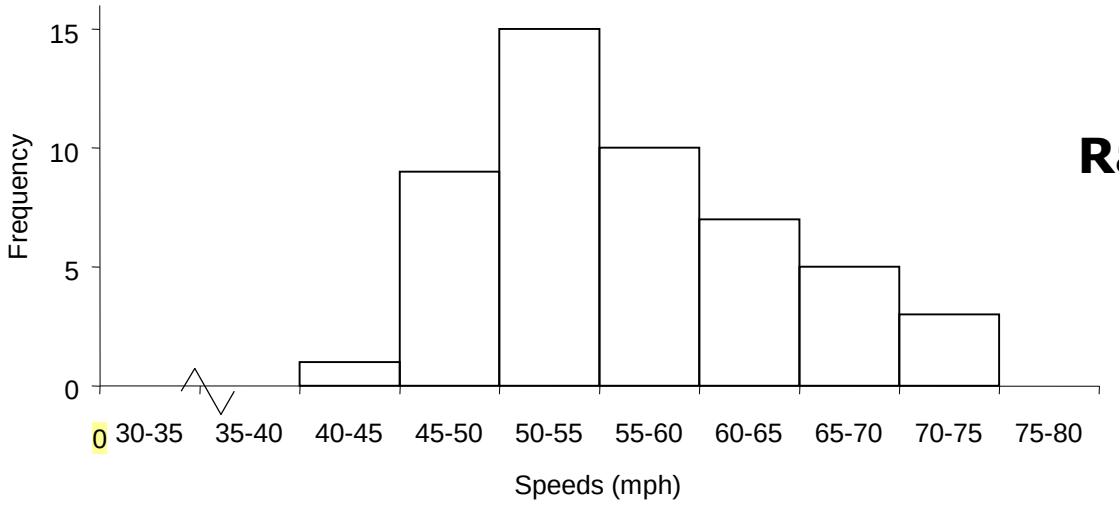
Class Frequencies

Histogram

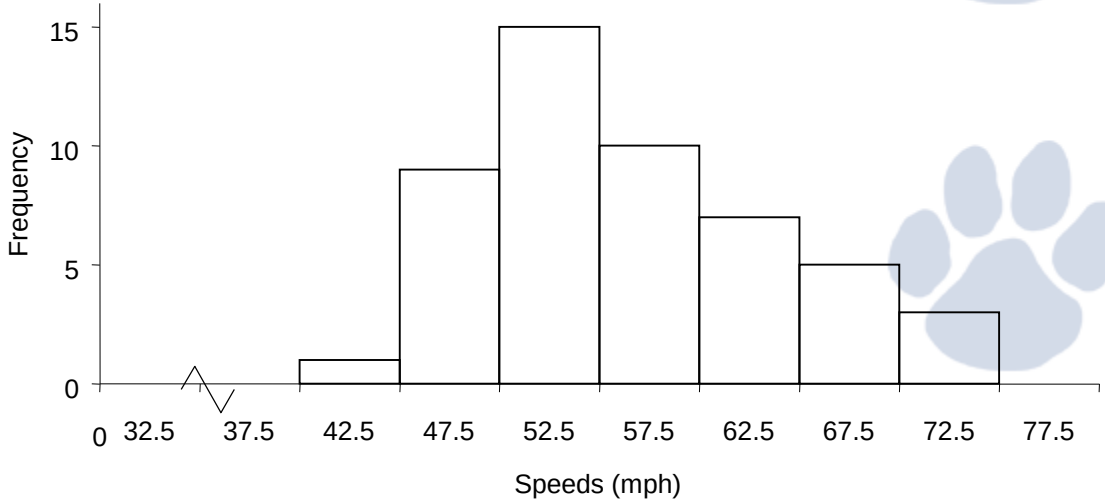


A histogram is a graphical representation of a frequency distribution. Each class includes those observations who's value is ***greater*** than the lower bound and ***less than or equal to*** the upper bound of the class.

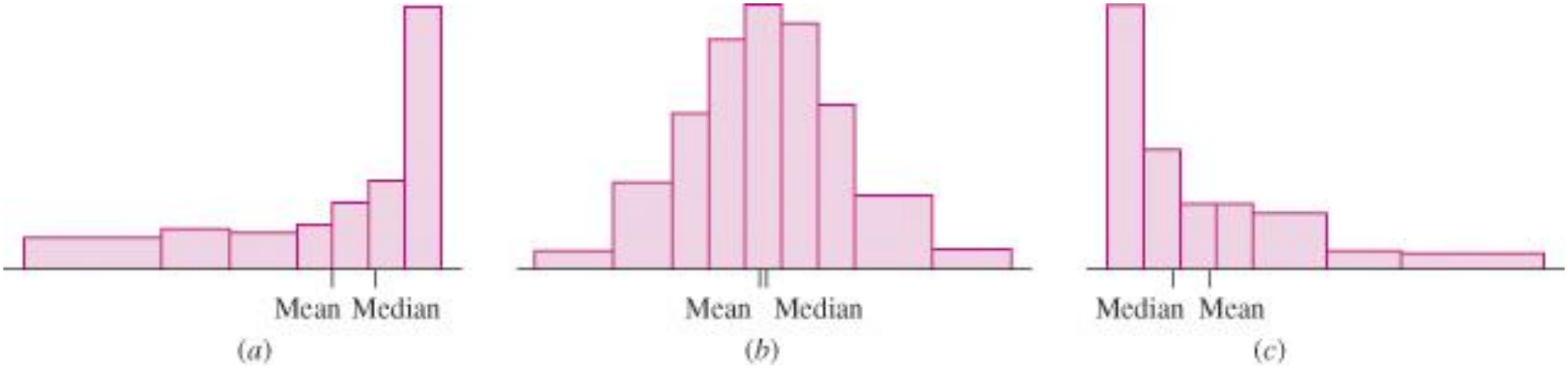
Histogram



Range



Symmetry and Skewness

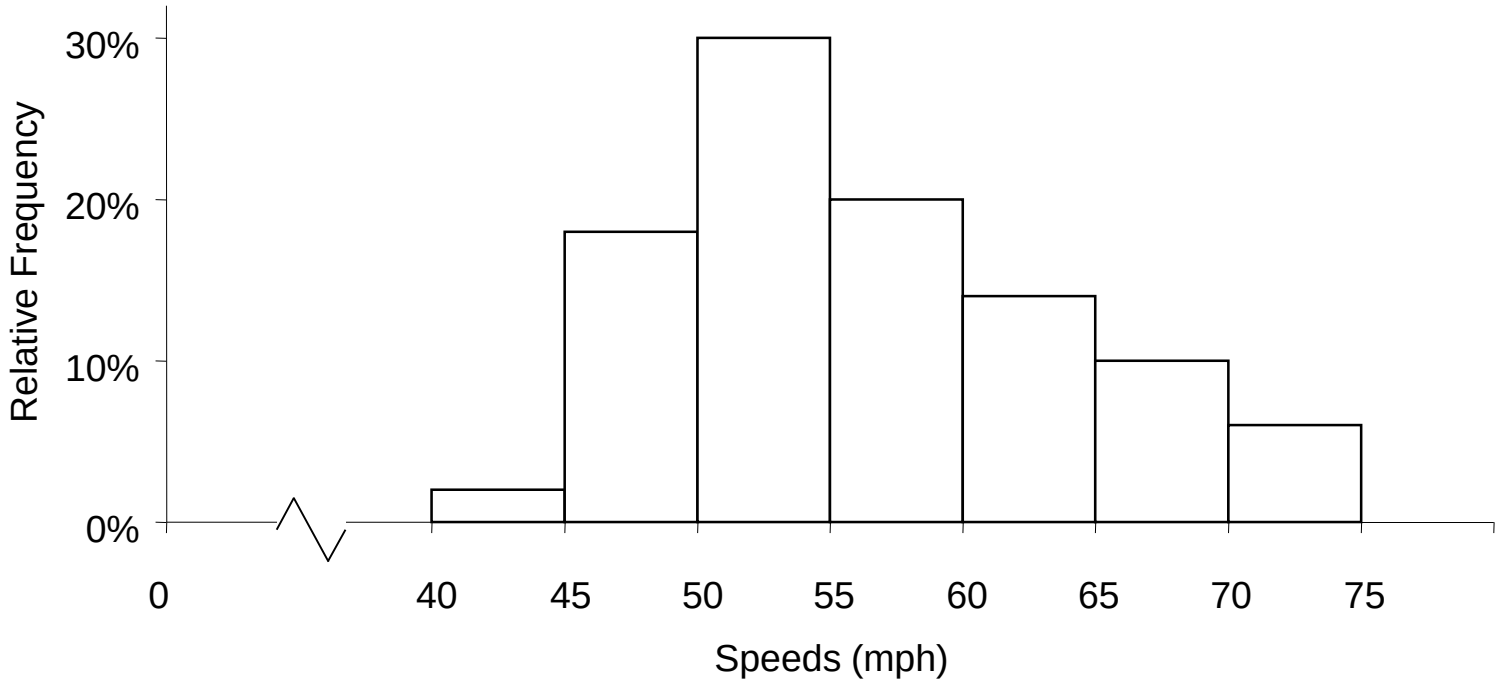


Relative Frequency Distribution

Vehicle Speeds on Poplar Avenue

Speeds (kph)	Vehicles Counted	Pecentage of Sample
40–45	1	2
45–50	9	18
50–55	15	30
55–60	10	20
60–65	7	14
65–70	5	10
70–75	3	6
TOTAL	50	100

Relative Frequency Histogram

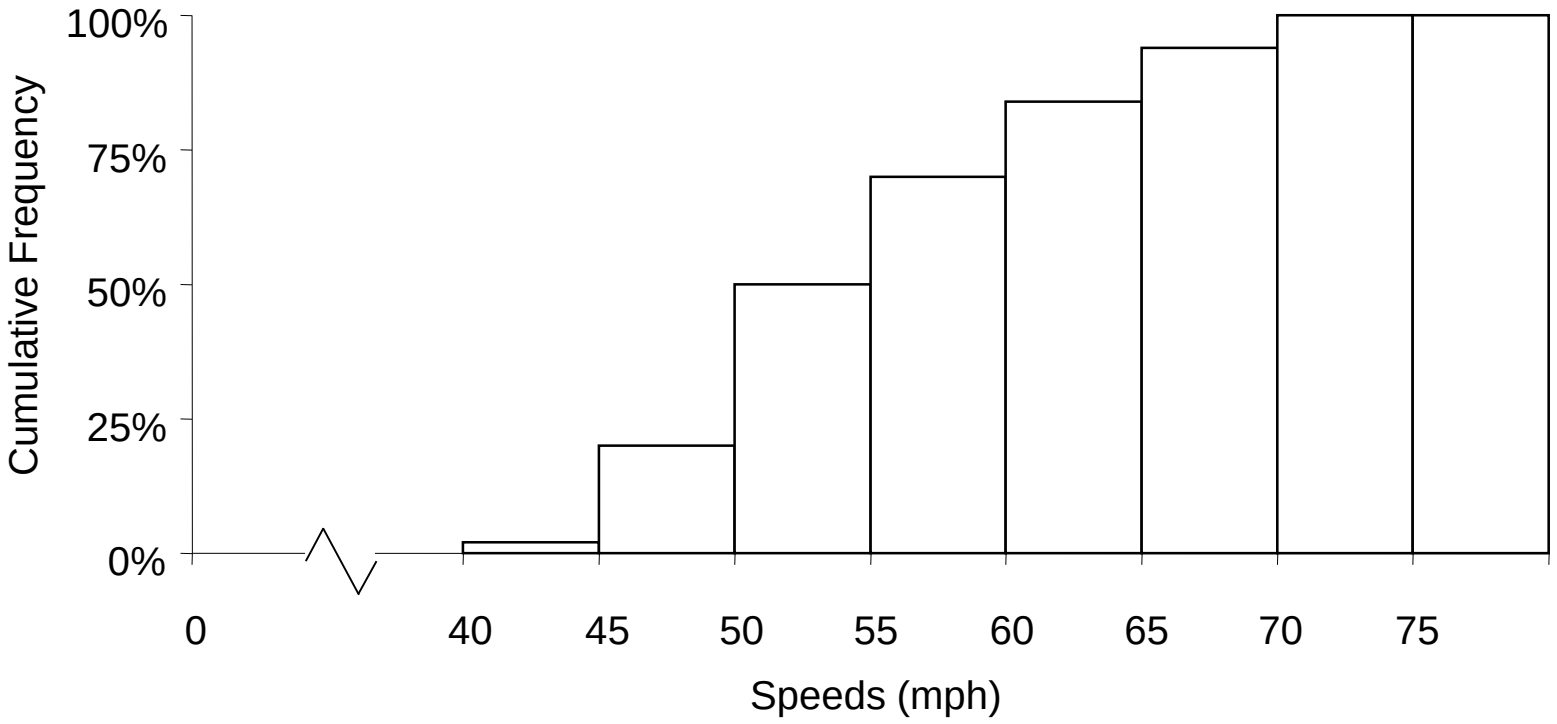


Cumulative Frequency Distributions

Vehicle Speeds on Poplar Avenue

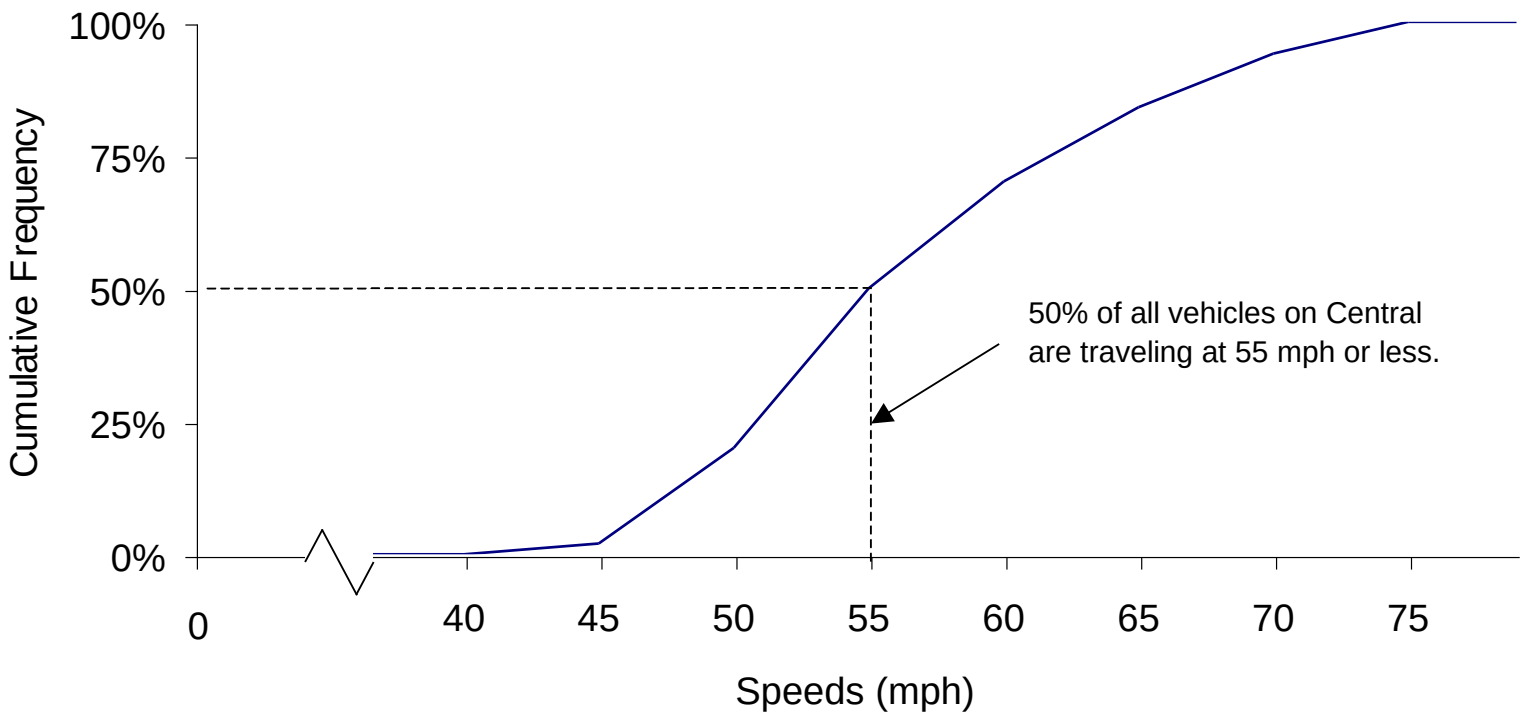
Speeds (mph)	Vehicles Counted	Percentage of Sample	Cumulative Percentage
40–45	1	2	2
45–50	9	18	20
50–55	15	30	50
55–60	10	20	70
60–65	7	14	84
65–70	5	10	94
70–75	3	6	100
TOTAL	50	100	

Cumulative Frequency Diagram



A good rule of thumb is that the number of classes should be approximately equal to the square root of the number of observations.

Cumulative Frequency Distributions



Boxplots

- A boxplot is a graphic that presents the median, the first and third quartiles, and any outliers present in the sample.

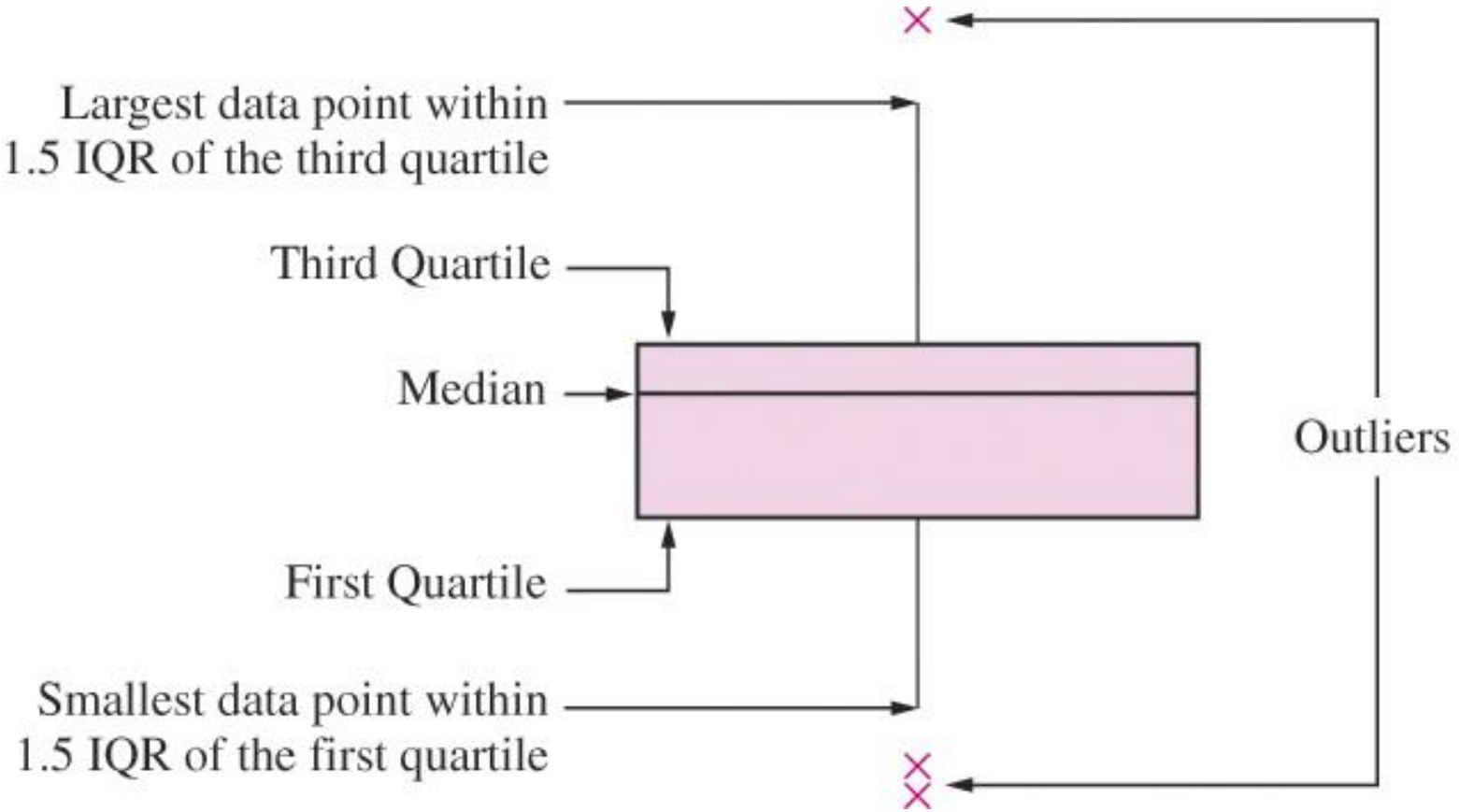
- The **interquartile range (IQR)** is the difference between the third and first quartile. This is the distance needed to span the middle half of the data.

Boxplots

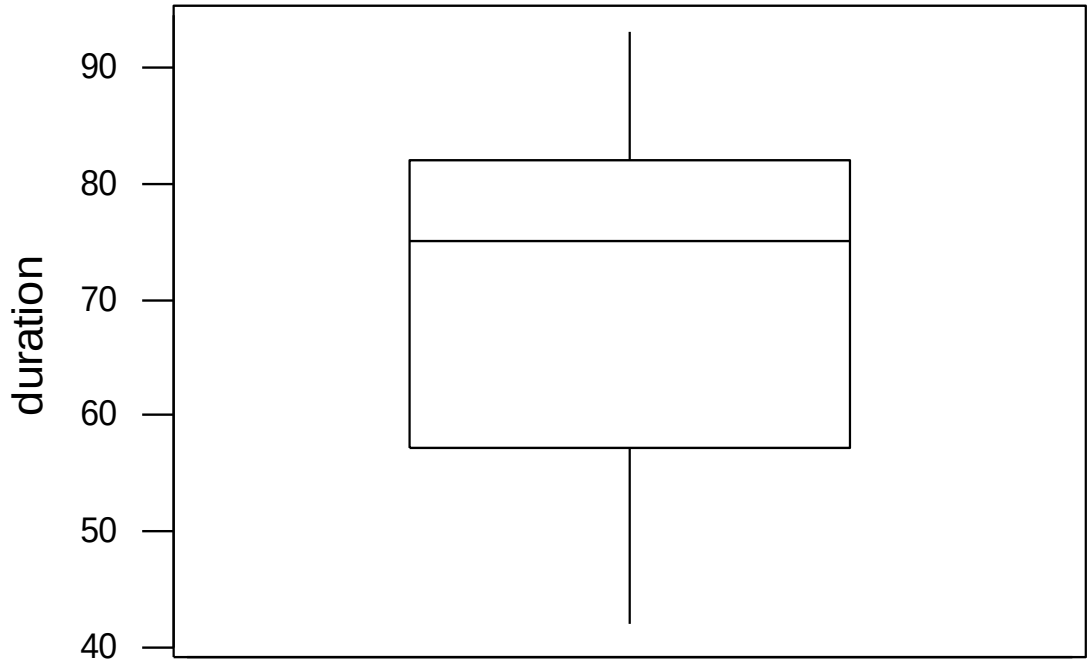
Steps in the Construction of a Boxplot:

- Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box.
- Find the largest sample value that is no more than 1.5 IQR above the third quartile, and the smallest sample value that is not more than 1.5 IQR below the first quartile. Extend vertical lines (whiskers) from the quartile lines to these points.
- Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile are designated as outliers. Plot each outlier individually.

Boxplots


Largest data point within
1.5 IQR of the third quartile
Third Quartile
Median
First Quartile
Smallest data point within
1.5 IQR of the first quartile
Outliers

Boxplots - Example



Boxplots - Example

- Notice there are no outliers in these data.
- Looking at the four pieces of the boxplot, we can tell that the sample values are comparatively densely packed between the median and the third quartile.
- The lower whisker is a bit longer than the upper one, indicating that the data has a slightly longer lower tail than an upper tail.
- The distance between the first quartile and the median is greater than the distance between the median and the third quartile.
- This boxplot suggests that the data are skewed to the left.

Stem and Leaf Plot

- Statistics: The branch of mathematics that deals with collecting, organizing, and analyzing or interpreting data.
- Data: Numerical facts or numerical information.
- Stem-and-Leaf Plots: A convenient method to display every piece of data by showing the digits of each number.

Stem and Leaf Plot

- In a stem-and leaf plot, the greatest common place value of the data is used to form *stems*.
- The numbers in the next greatest place-value position are then used to form the *leaves*.

Stem and Leaf Plot

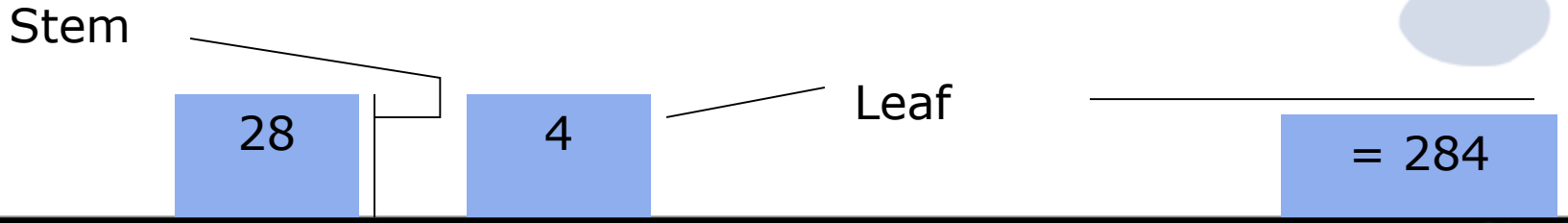
- Leaf: The last digit on the right of the number.
- Stem: The digit or digits that remain when the leaf is dropped.

Look at the number 284

The leaf is the last digit formed: the number 4.

The stem is the remaining digits when the leaf is dropped: the number 28.

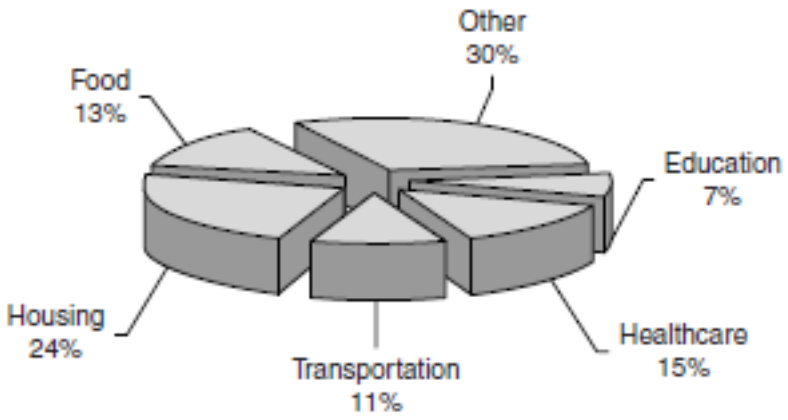
The stem with the leaf forms the number 284.



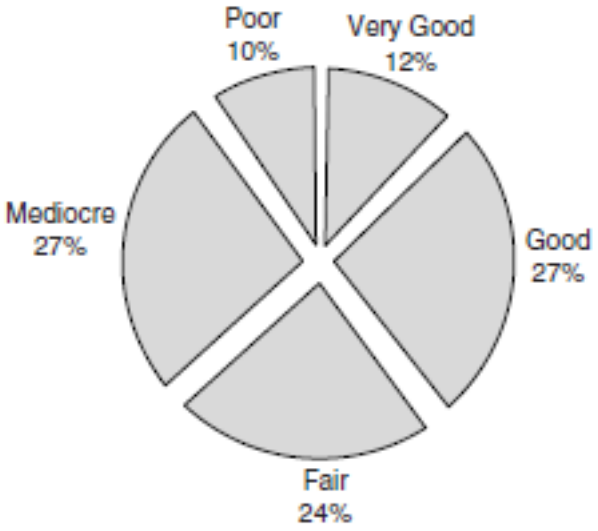
Stem and Leaf Plot

- Here are the scores from two periods of math class. Students took the same test.
- Period 1: 77 79 85 58 97 94 82 81 75 63 60 92 75 98 83 58 72 57 70 81
- Period 2: 57 60 88 85 79 70 65 98 97 59 58 65 62 77 77 75 73 69 82 81

Pi-chart

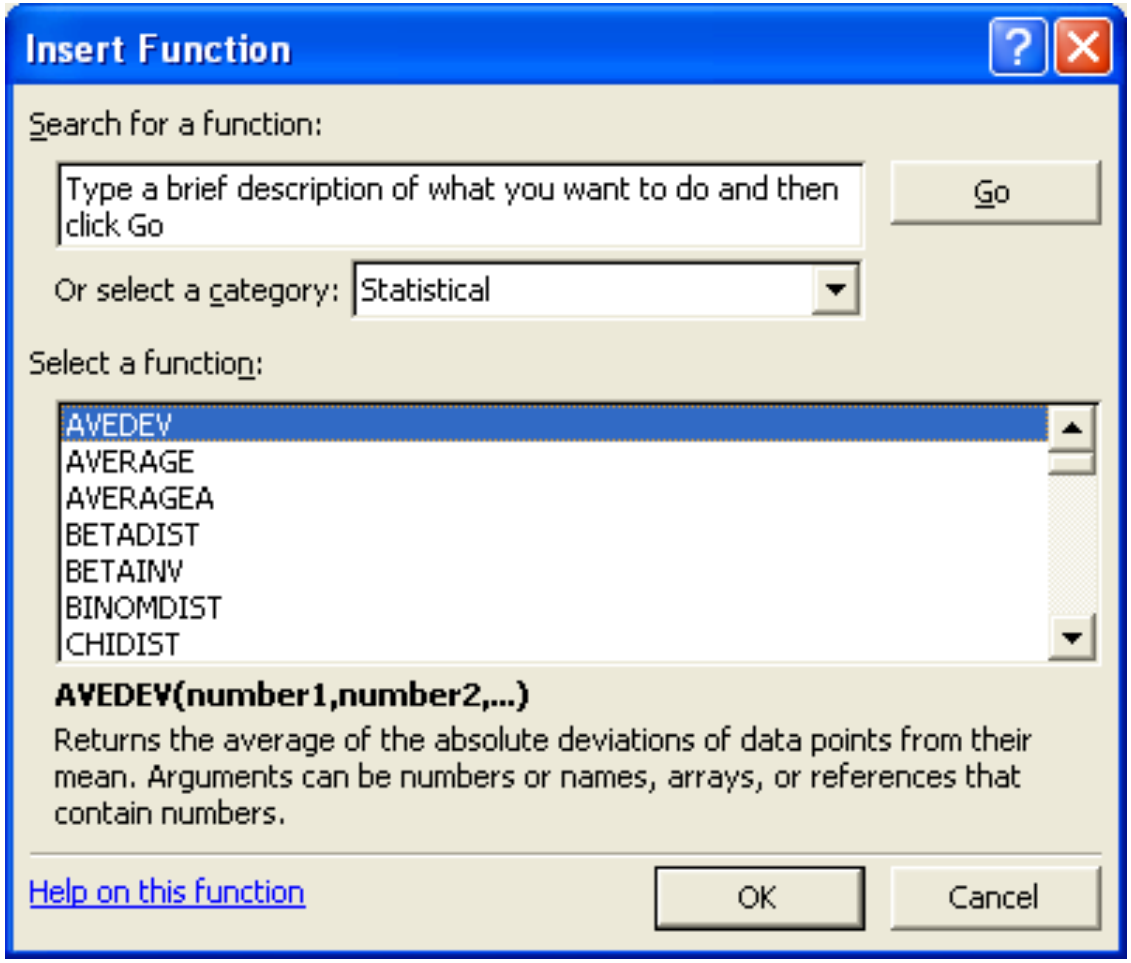


US GDP by Function

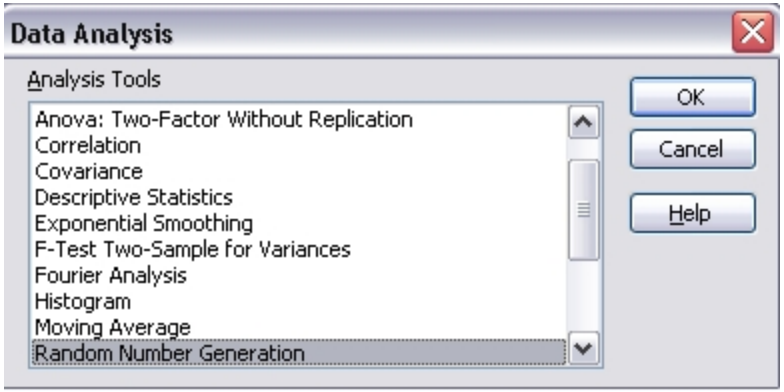


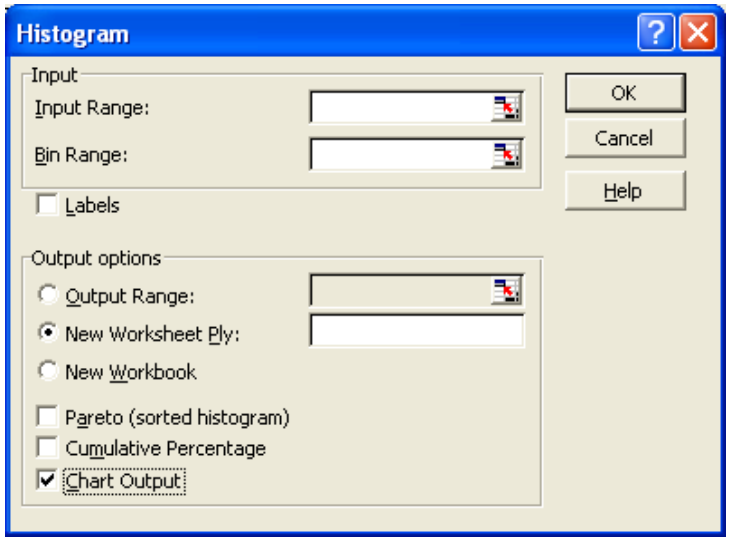
US Interstate Pavement roughness

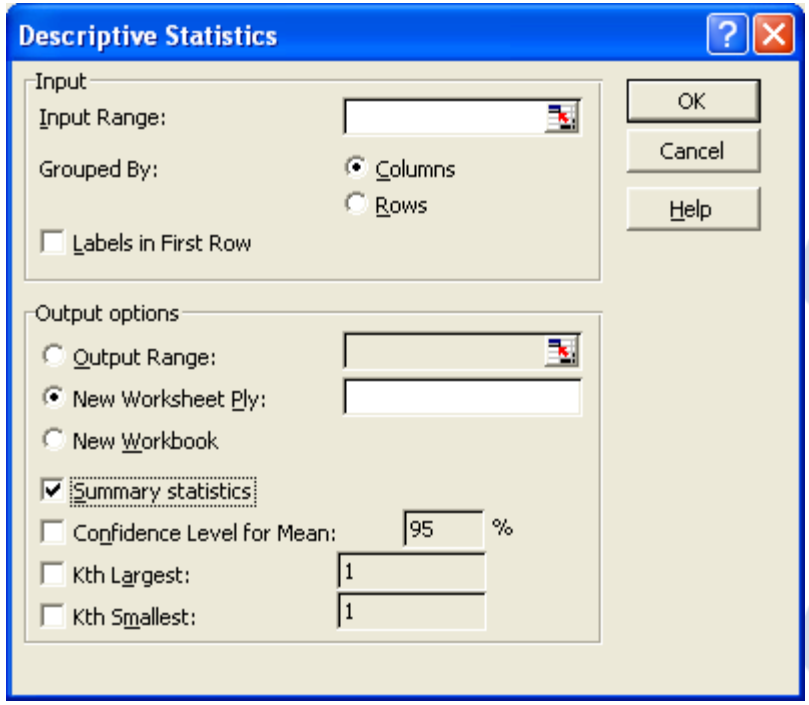
Excel Tools


Insert Function
Search for a function:
Type a brief description of what you want to do and then click Go
Go
Or select a category: Statistical
Select a function:
AVEDEV
AVERAGE
AVERAGEA
BETADIST
BETAINV
BINOMDIST
CHIDIST
AVEDEV(number1,number2,...)
Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.
Help on this function
OK
Cancel

Analysis Toolpack in Excel







Grouped Data-Mean

Grouped Data-Mean

Example: The following table gives the frequency distribution of the number of orders received each day during the past 50 days in a manufacturing company. Calculate the mean.

Number of order	*f*
10 – 12	**4**
13 – 15	**12**
16 – 18	**20**
19 – 21	**14**
	***n* = 50**

Solution:

X is the midpoint of the class. It is adding the class limits and divide by 2.

Number of order	*f*	*x*	*fx*
10 – 12	**4**	**11**	**44**
13 – 15	**12**	14	168
16 – 18	**20**	17	340
19 – 21	**14**	20	280
	***n* = 50**		**= 832**

$$\bar{x} = \frac{\sum fx}{n} = \frac{832}{50} = 16.64$$

Grouped Data-Median

Step 1: Construct the cumulative frequency distribution.
Step 2: Decide the class that contain the median.
Class Median is the first class with the value of cumulative frequency equal at least n/2.
Step 3: Find the median by using the following formula:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

Where:

n = the **total frequency**

F = the **cumulative frequency *before*** class median

f_m = the **frequency** of the class median

i = the class width

L_m = the **lower boundary** of the class median

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:

1st Step: Construct the cumulative frequency distribution

Time to travel to work	Frequency	Cumulative Frequency
1 – 10	8	8
11 – 20	14	22
21 – 30	12	34
31 – 40	9	43
41 – 50	7	50

$\frac{n}{2} = \frac{50}{2} = 25 \longrightarrow$ class median is the 3rd class

So, $F = 22$, $f_m = 12$, $L_m = 20.5$ and $i = 10$

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

$$= 21.5 + \left(\frac{25 - 22}{12} \right) 10$$

$$= 24$$

Thus, 25 persons take less than 24 minutes to travel to work and another 25 persons take more than 24 minutes to travel to work.

Quartiles

Using the same method of calculation as in the Median, we can get Q_1 and Q_3 equation as follows:

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) i \qquad Q_3 = L_{Q_3} + \left(\frac{\frac{3n}{4} - F}{f_{Q_3}} \right) i$$

Example: Based on the grouped data below, find the Interquartile Range

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:

1st Step: *Construct the cumulative frequency distribution*

Time to travel to work	***Frequency***	***Cumulative Frequency***
1 – 10	***8***	***8***
11 – 20	***14***	***22***
21 – 30	***12***	***34***
31 – 40	***9***	***43***
41 – 50	*7*	*50*

2nd Step: Determine the Q_1 and Q_3

$$\text{Class } Q_1 = \frac{n}{4} = \frac{50}{4} = 12.5$$

Class Q_1 is the 2nd class

Therefore,

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) i$$

$$= 10.5 + \left(\frac{12.5 - 8}{14} \right) 10$$

$$= 13.7143$$

$$\text{Class } Q_3 = \frac{3n}{4} = \frac{3(50)}{4} = 37.5$$

Class Q_3 is the 4th class
Therefore,

$$Q_3 = L_{Q_3} + \left(\frac{\frac{n}{4} - F}{f_{Q_3}} \right) i$$

$$= 30.5 + \left(\frac{37.5 - 34}{9} \right) 10$$

$$= 34.3889$$

Interquartile Range

$$IQR = Q_3 - Q_1$$

$IQR = Q_3 - Q_1$

calculate the IQ

$IQR = Q_3 - Q_1 = 34.3889 - 13.7143 = 20.6746$

Mode-Grouped Data

- Mode is the value that has the highest frequency in a data set.
- For grouped data, class mode (or, modal class) is the class with the highest frequency.
- To find mode for grouped data, use the following formula:

$$\text{Mode} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i$$

Where:

i is the class width

Δ_1 is the difference between the frequency of class mode and the frequency of the class **after** the class mode

Δ_2 is the difference between the frequency of class mode and the frequency of the class **before** the class mode

L_{mo} is the **lower boundary** of class mode

Calculation of Grouped Data - Mode

Example: Based on the grouped data below, find the mode

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:

Based on the table,

$L_{mo} = 10.5,\ \Delta_1 = (14 - 8) = 6,\ \Delta_2 = (14 - 12) = 2$ and $i = 10$

$$\text{Mode} = 10.5 + \left(\frac{6}{6 + 2}\right)10 = 17.5$$

Variance and SD - Grouped Data

Population Variance: $$\sigma^2 = \frac{\sum fx^2 - \frac{\left(\sum fx\right)^2}{N}}{N}$$

Variance for sample data: $$s^2 = \frac{\sum fx^2 - \frac{\left(\sum fx\right)^2}{n}}{n-1}$$

Standard Deviation:

Population: $$\sigma^2 = \sqrt{\sigma^2}$$

Sample: $$s^2 = \sqrt{s^2}$$

Variance and SD - Grouped Data

No. of order	*f*
10 – 12	4
13 – 15	12
16 – 18	20
19 – 21	14
Total	n = 50

Solution:

No. of order	*f*	*x*	*fx*	fx^2
10 – 12	4	11	44	484
13 – 15	12	14	168	2352
16 – 18	20	17	340	5780
19 – 21	14	20	280	5600
Total	n = 50		832	14216

Variance and SD - Grouped Data

$$\text{Variance,} \quad s^2 = \frac{\sum fx^2 - \frac{\left(\sum fx\right)^2}{n}}{n-1}$$

$$= \frac{14216 - \frac{(832)^2}{50}}{50-1}$$

$$= 7.5820$$

$$\text{Standard Deviation,} \quad s = \sqrt{s^2} = \sqrt{7.5820} = 2.75$$

Thus, the standard deviation of the number of orders received at the office of this mail-order company during the past 50 days is 2.75.