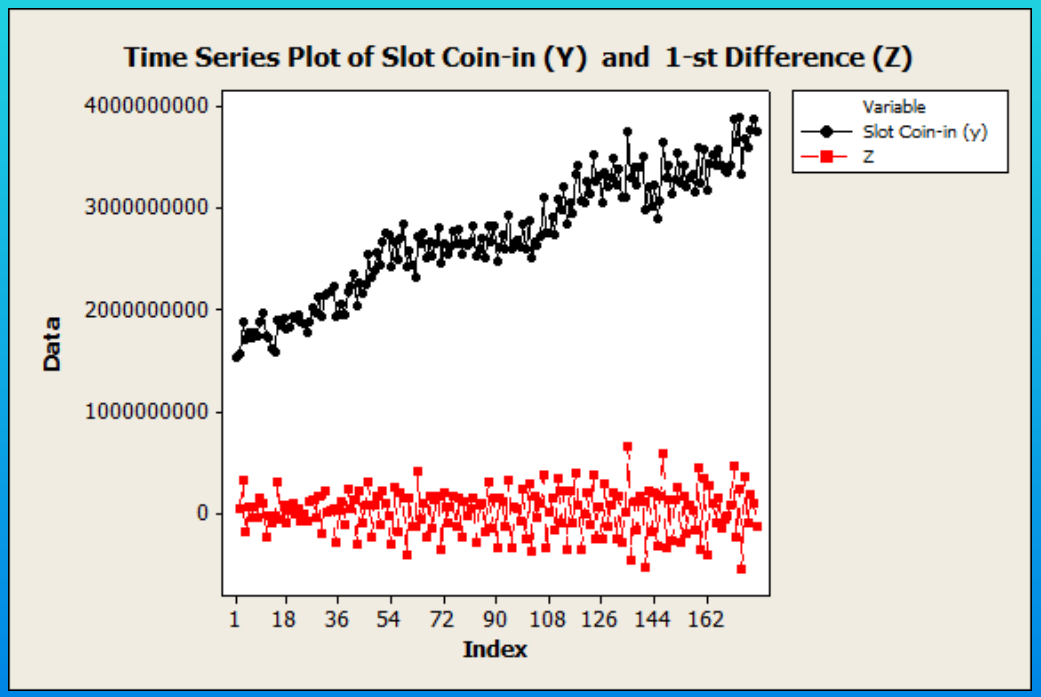
Time Series Modeling
Lecture 1a: Introduction



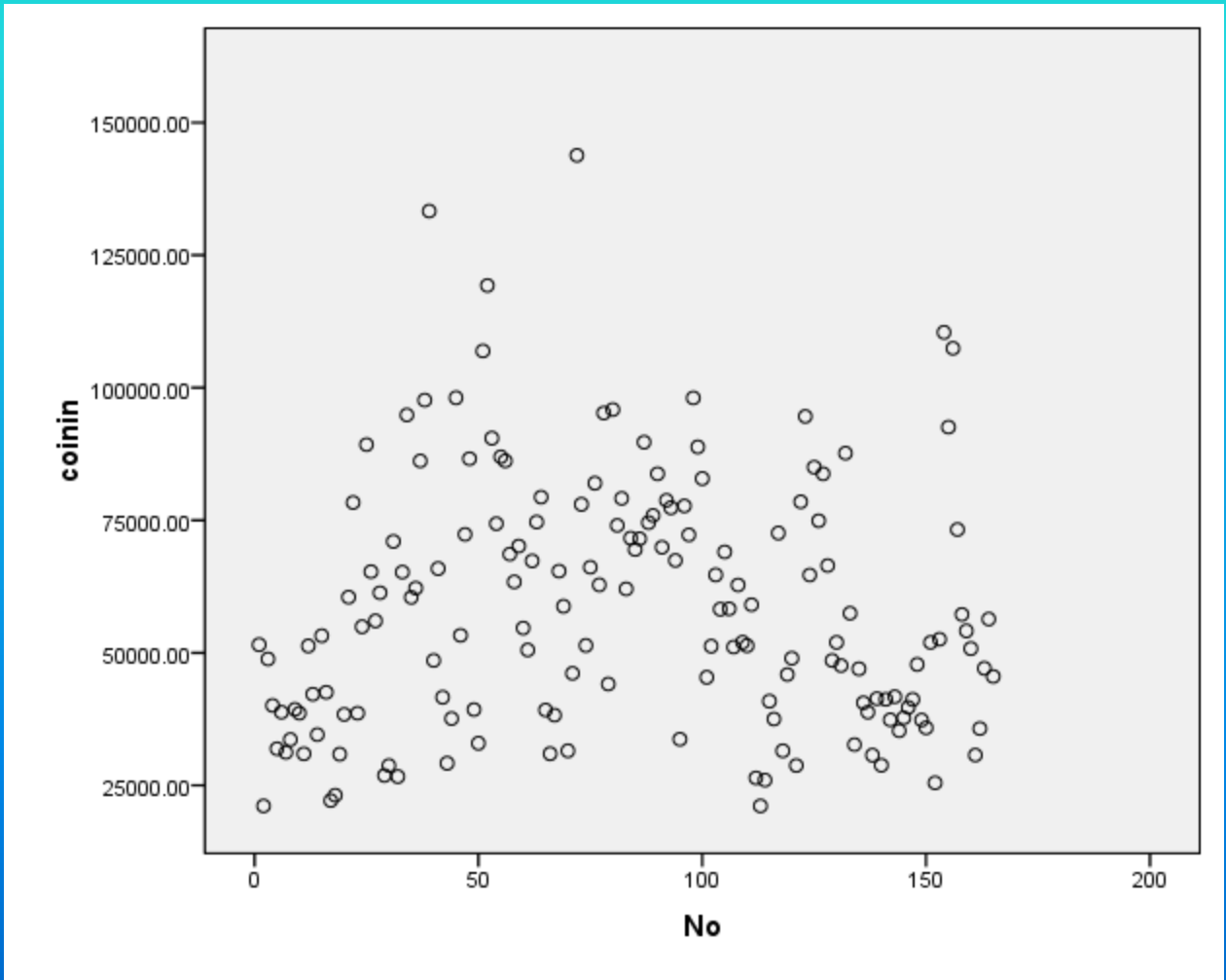
A common assumption in statistical data analysis is –

the observations in a random sample are INDEPENDENT.

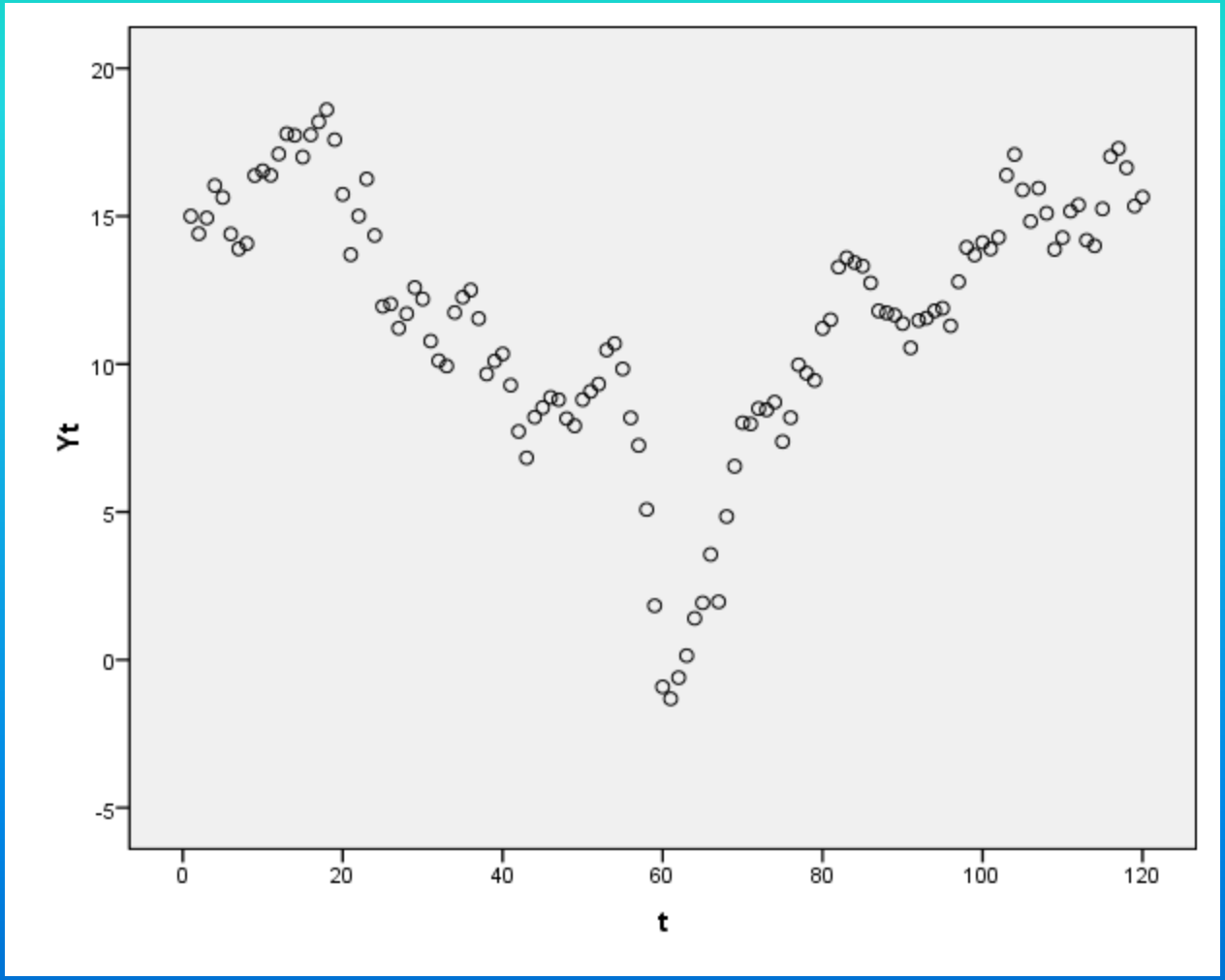
In a time series (observations of a variable of interest recorded over time), observations are typically CORRELATED.

Let us look at graphs of two different data sets.

Scatterplot of COININ vs SERIAL NUMBER for 165 slot machines



Scatterplot of PAPER TOWEL SALES vs TIME for 165 slot machines



Time Series plot of Paper Towel Sales Data
x-axis = time, y-axis = sales Yt

In this class you will learn how to use the correlation that exists in a time series data to obtain better forecasts of the time series.

Before we begin time series methods, we will review statistical concepts and methods.

Review of Basic Statistical Concepts

POPULATION in statistics refers to the entire collection of experimental units or subjects in which the experimenter is interested.

Typically the population is too large to collect measurements on (too expensive or may take too long), so the experimenter collects a RANDOM SAMPLE of experimental units from the population of interest.

Let us look at a few examples.

Example 1: CardWeb tracked all credit or debit card purchases in US in 2005. The amount of each purchase was recorded and classified according to the type of card used (AX, DISCOVER, MC, VISA).

What is the variables of interest?
Card Type

Does the data set collected represents a population or a sample?
Population

Example 2: Opinion polls are regularly conducted to determine the popularity of the current president. Suppose a poll is to be conducted next month in which 2000 residents of the country will be asked if the president is doing a good job or a bad job. The 2000 individuals will be selected by random-digit telephone dialing and asked the question over the phone.

What is the relevant population?
Entire population of the country.

What is the variable of interest?
Voter opinion

What is the sample?
2000 individuals selected

Is this sample representative of the population?
No, it only represents residents with land-line phones

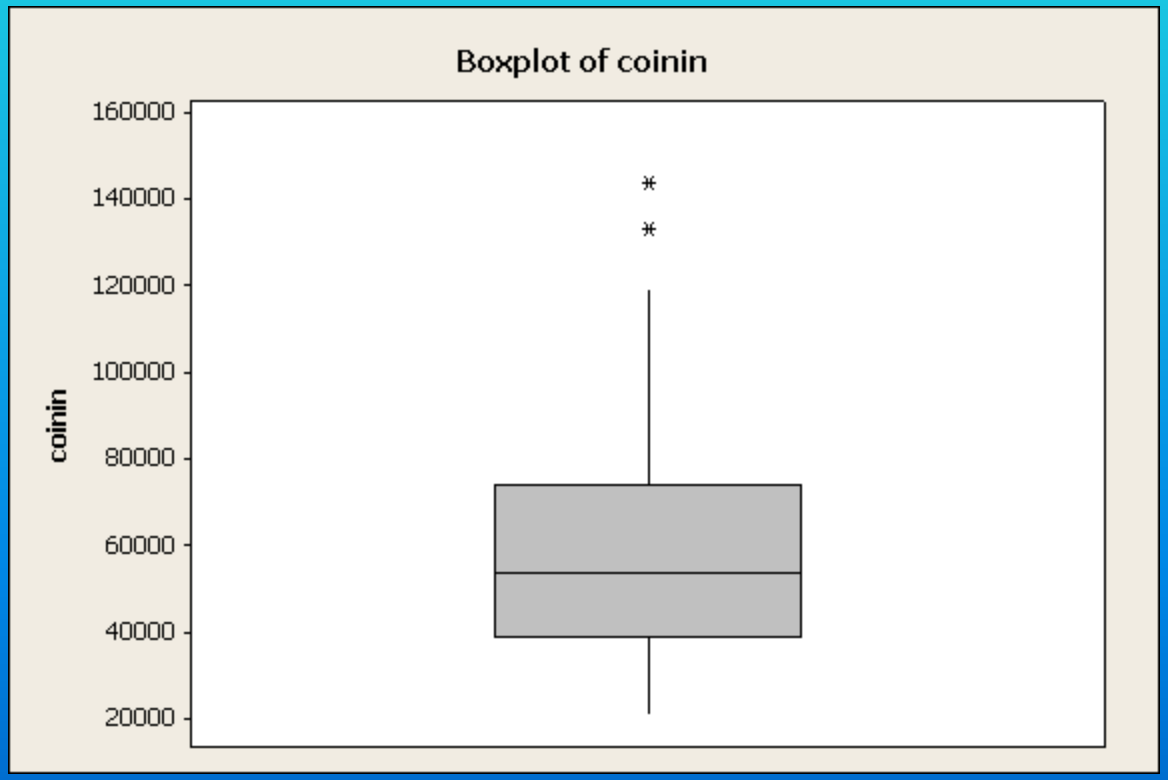
Example 3: Life Testing of Light Bulbs
Each light bulb we buy has its life (in hours) written on it. This number is an estimate of how long such a light bulb is expected to last. To obtain this number, one must put a sample of (say 50) light bulbs through a life test, and record the time at which the bulb fails.

What is the population in this example?
All light bulbs

Can we sample the entire population, or do we have to take a random sample?
Test is destructive, hence sampling is necessary.

Data File = Time Series\Data\Example 1 Coinin

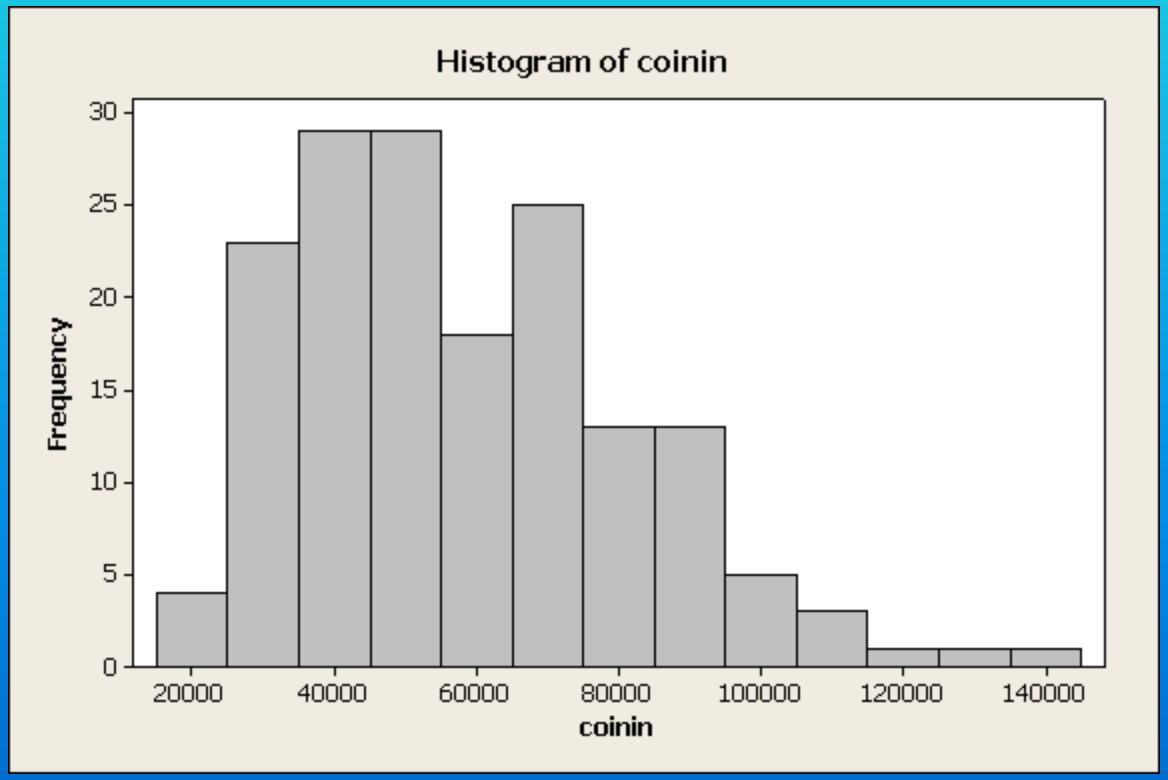
BOX PLOT is a 5 point summary of a data set, which shows the following measures of a data set:
Minimum, 1st Quartile Q_1, Median = 2nd Quartile Q_2 , 3rd Quartile Q_3, Maximum.



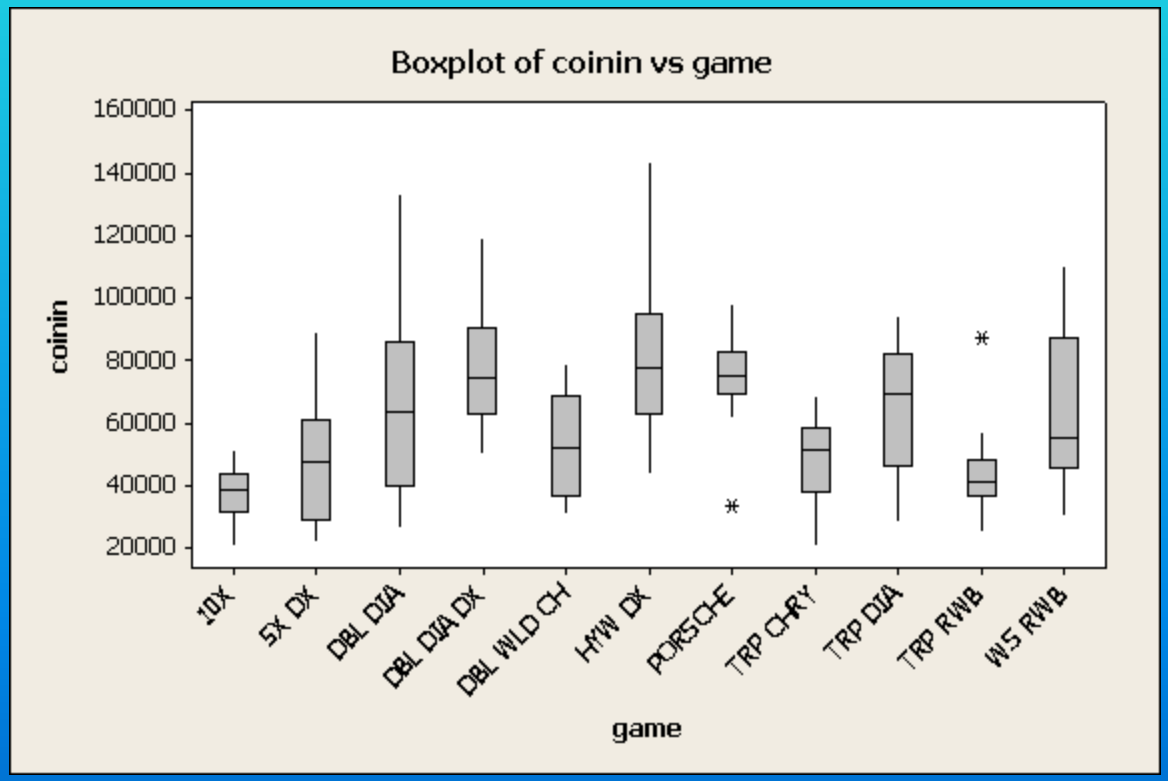
A FREQUENCY TABLE is created by dividing
RANGE = Sample Maximum – Sample Minimum by
k = # of class intervals (selected by user)
then
forming k class intervals, and counting # of observartions falling in each class interval (FREQUENCY of class interval). Frequency Table for Coinin is shown below:

Class Interval		
Low	High	Frequency
15000	25000	4
25000	35000	23
35000	45000	29
45000	55000	29
55000	65000	18
65000	75000	25
75000	85000	13
85000	95000	13
95000	105000	5
105000	115000	3
115000	125000	1
125000	135000	1
135000	145000	1

A sample histogram is a plot of frequencies of class intevals.



Note that the coinin data set has two columns, with the 1st column showing GAME-TYPE. The user will be interested in describing the data by the GAME-TYPE.



game	Count	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
10X	14	38149	8546	21122	31745	38705	43850	51520
5XDX	16	48140	20009	22134	29268	47895	61130	89280
DBLDIA	20	65148	28004	26662	39886	63707	86500	133310
DBLDIDX	11	79232	21220	50536	63381	74368	90486	119286
DBLWLDCH	10	53160	18136	30964	36549	52446	69173	79368
HYWDX	11	79320	26818	44113	62816	78000	95235	143830
PORSCHE	18	74749	13554	33711	69802	75235	83085	98068
TRPCHRY	16	48442	14509	21123	38353	51297	58872	69033
TRPDIA	12	64638	21369	28746	46679	69526	82451	94591
TRPRWB	25	43134	12177	25479	36583	41208	48184	87649
WSRWB	12	63438	26717	30704	45926	55249	87735	110438

We will next go over the descriptive statistics that are shown in the above table.

DESCRIPTIVE STATISTICS

MEASURES OF CENTRALITY

Sample mean $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

To calculate the sample median and the 1-st and 3-rd quartiles, the observations in the sample are first sorted in an increasing (or decreasing) order.

$$\text{Sample median} = \begin{cases} (m+1)\text{-st value if } n = 2m + 1 \text{ (odd integer)} \\ \text{average of the m-th and (m+1)-st value,} \\ n = 2m \text{ (even integer)} \end{cases}$$

MEASURES OF DISPERSION OR SPREAD

Sample range = sample max – sample min

Another measure of spread is -

1-st Quartile = Q_1 = value that splits data in range (min, median) into two equal halves.

3-rd Quartile = Q_3 = value that splits data in range (median.max) into two equal halves.

Interquartile Range (IQR) = $Q_3 - Q_1$

EXAMPLE:

The following are wingspan measurements (in cms) of two different samples of 5 birds each:

Sample 1 = {10, 9, 11, 10, 15}
Sample 2 = {5, 9, 11, 10, 20}

$$\overline{x}_1 = \frac{10+9+11+10+15}{5} = \frac{55}{5} = 11$$

$$\overline{x}_2 = \frac{5+9+11+10+20}{5} = \frac{55}{5} = 11$$

The two data sets have the same mean, but the two sets are different in terms of SPREAD.

Which one seems more variable?

THE SECOND ONE:

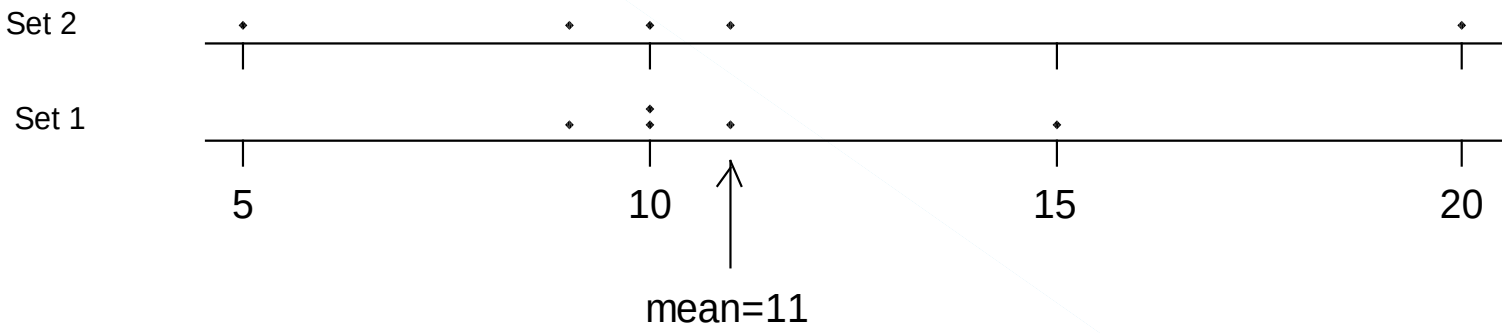
range of set 1 = 15-9=6,

range of set 2 = 20-5 = 15

IQR = 11 – 10 = 1, first set

= 11 – 9 = 2, second set

Dotplot for the two data sets



Now calculate the deviations from the mean: $d_i = x_i - \bar{x}$

Set 1:

W can try to get a measure of spread by calculating the average value of deviations from the mean

$$d_1 = 10 - 11 = -1, d_2 = 9 - 11 = -2, d_3 = 11 - 11 = 0, d_4 = 10 - 11 = -1, d_5 = 15 - 11 = 4$$

$$d_1 = 5 - 11 = -6, d_2 = 9 - 11 = -2, d_3 = 11 - 11 = 0, d_4 = 10 - 11 = -1, d_5 = 20 - 11 = 9$$

For the first set: average deviation = $\frac{-6-2+0-1+9}{5} = 0$

For the second set: average deviation = $\frac{-1-2+0-1+4}{5} = 0$

We can, in fact, easily show that the sum of deviations from the sample mean FOR ANY DATA SET equals 0:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

So what do we do now?

There are two options:

Use the average of SQUARRED DEVIATIONS:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} \qquad or$$

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

The second formula (divide by n-1) gives an UNBIASED ESTIMATOR of the POPULATION VARIANCE (we will discuss this further in Chapter 5), and is the formula used by most statistical software packages. We will now calculate s^2 (called the SAMPLE VARIANCE) using the second formula (divide by n-1) for the data set given above:

$$s^2 = \frac{(-1)^2 + (-2^{\ 2} +)(0^{\ 2} +)(-1)^2 + (4^{\ 2}\ )}{5-1} = 5.5 \qquad set-1$$

$$s^2 = \frac{(-6)^2 + (-2)^2 + (0)^2 + (-1)^2 + (9)^2}{5-1} = 30.5 \qquad set-2$$

We can also use the AVERAGE of ABSOLUTE DEVIATIONS:

$$mad = \frac{\sum_{i=1}^{n} | x_i - \overline{x} |}{n}$$

The use of AVERAGE ABSOLUTE DEVIATION is not very common.

The SAMPLE STANDARD DEVIATION (sd), denoted by *s*, is the positive square root of the sample variance.

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

Example

$$s_1 = \sqrt{5.5} = 2.35$$

$$s_2 = \sqrt{30.5} = 5.52$$

DESCRIPTIVE STATISTICS FROM GROUPED DATA or FREQUENCY DATA

Class Interval			
Low	High	Mid-point	Frequency
15000	25000	20000	4
25000	35000	30000	23
35000	45000	40000	29
45000	55000	50000	29
55000	65000	60000	18
65000	75000	70000	25
75000	85000	80000	13
85000	95000	90000	13
95000	105000	100000	5
105000	115000	110000	3
115000	125000	120000	1
125000	135000	130000	1
135000	145000	140000	1

$$\bar{x} = \frac{\sum_{i=1}^{K} M_i f_i}{\sum_{i=1}^{K} f_i} = \frac{20000 \times 4 + 30000 \times 23 + \ldots + 140000 \times 1}{4 + 23 + 29 + \ldots + 1 + 1 + 1}$$

$$= \quad 58424.242$$

$$s^2 = \frac{\sum_{i=1}^{K} (M_i - \bar{x})^2 f_i}{\sum_{i=1}^{K} f_i} = \frac{\sum_{i=1}^{K} (M_i)^2 f_i - n(\bar{x})^2}{n} = \quad 554818921$$

Descriptive Statistics of coinin from raw data

Variable	n	Mean	sd
coinin	165	58296	23480