

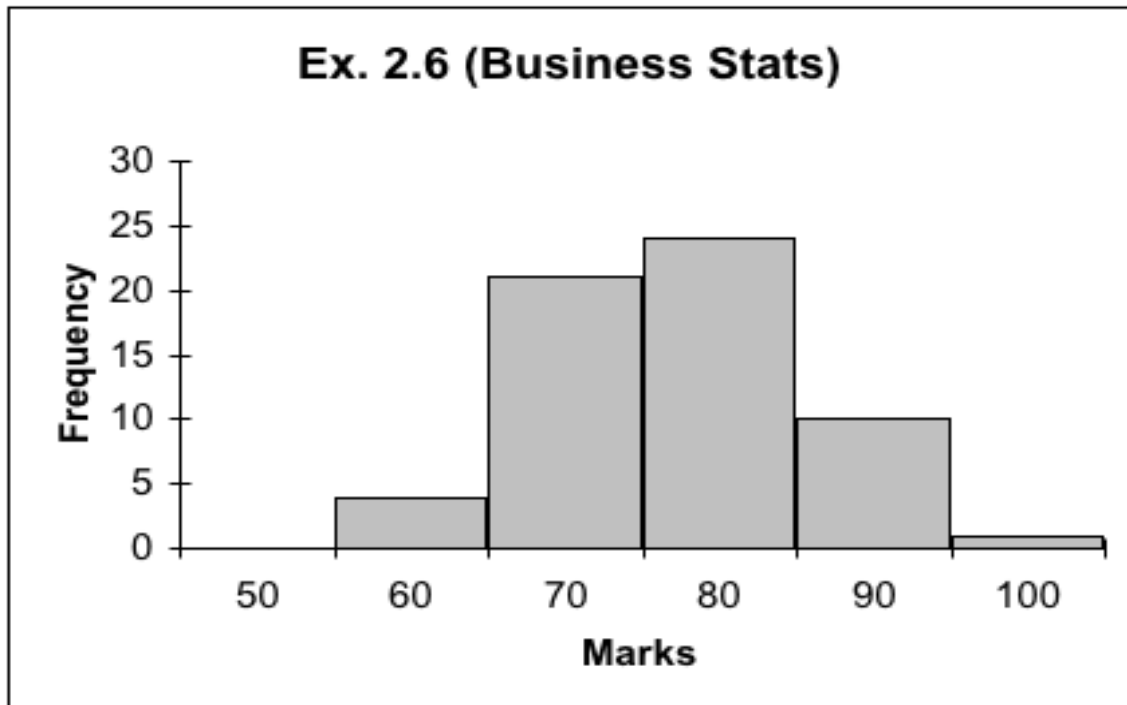
# Numerical Summarization of Data

OPRE 6301

## Motivation...

---

In the previous session, we used graphical techniques to describe data. For example:



While this histogram provides useful insight, other interesting questions are left unanswered. The most important ones are:

What is the class “average”?

What is the “spread” of the marks?

# Numerical Descriptive Techniques...

---

Measures of Central Location

— Mean, Median, Mode

Measures of Variability

— Range, Mean Absolute Deviation, Variance,  
Standard Deviation, Coefficient of Variation

Measures of Relative Standing

— Percentiles, Quartiles

Measures of Relationship

— Covariance, Correlation, Least-Squares Line

Details...

# Measures of Central Location...

---

## Mean...

The **arithmetic mean**, a.k.a. *average*, shortened to *mean*, is the most popular and useful measure of central location.

It is computed by simply adding up all the observations and dividing by the total number of observations:

$$\text{Mean} = \frac{\text{Sum of the Observations}}{\text{Number of Observations}}$$

To define things formally, we need a bit of notation...

When referring to the number of observations in a **population**, we use uppercase letter  $N$ .

When referring to the number of observations in a **sample**, we use lower case letter  $n$ .

The arithmetic mean for a **population** is denoted by the Greek letter “mu”:  $\mu$ .

The arithmetic mean for a **sample** is denoted by “ $x$ -bar”:  $\bar{x}$ .

Thus,

|      | <b>Population</b> | <b>Sample</b> |
|------|-------------------|---------------|
| Size | $N$               | $n$           |
| Mean | $\mu$             | $\bar{x}$     |

The formal definitions of  $\mu$  and  $\bar{x}$  are:

$$\mu \equiv \frac{1}{N} \sum_{i=1}^N x_i$$

and

$$\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i,$$

where  $x_i$  denotes the  $i$ th value in the population or in the sample.

## Other Definitions

The above definitions are for “raw” data. For grouped data or for “theoretical” population distributions, we have slightly different formulas. The idea, however, is the same.

For grouped data (i.e., a histogram), we use the following approximation:

$$\bar{x} \equiv \frac{1}{n} \sum_{k=1}^K m(k) f(k), \quad (1)$$

where  $K$  denotes the total number of bins (or classes),  $m(k)$  denotes the midpoint of the  $k$ th bin, and  $f(k)$  denotes the number of observations in, or the frequency for, the  $k$ th bin.

The idea here is to approximate all observations in a bin by the midpoint for that bin. We are forced into doing this because raw data are not available.

For a population that is described by a *probability density function*  $f(x)$ , we have:

$$\mu \equiv \int_{-\infty}^{\infty} x f(x) dx . \quad (2)$$

To motivate this, observe that (1) can be written as:

$$\bar{x} = \sum_{k=1}^K m(k) p(k) ,$$

where  $p(k) \equiv f(k)/n$  denotes the  $k$ th relative frequency. Now, if  $f(x)$  is a probability *density*, then “ $f(x) dx$ ” approximately equals the probability for a random sample from the population to fall in the interval  $(x, x + dx)$ . Replacing the sum by an integral then yields (2).

As a simple example, consider the **uniform** density, which has  $f(x) = 1$  for  $x$  in the interval  $(0, 1)$ . Then,

$$\mu = \int_0^1 x \cdot 1 dx = \frac{1}{2} x^2 \Big|_0^1 = \frac{1}{2} .$$

The Excel function RAND() generates (pseudo) samples, or realizations, from this theoretical density.

## Pooled Mean

We often pool two data sets together. Let the sizes of two samples be  $n_1$  and  $n_2$ ; and let the respective sample means be  $\bar{x}_1$  and  $\bar{x}_2$ . Suppose the two samples are pooled together to form one of size  $n_1 + n_2$ . Can we determine the mean of the combined group? The answer is:

$$\bar{x}_{pooled} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}.$$

Example: Suppose we have  $n_1 = 50$ ,  $\bar{x}_1 = \$15$ ,  $n_2 = 100$ , and  $\bar{x}_2 = \$17$ . These are samples of hourly wages. Then, the pooled average wage of the combined 150 individuals is:

$$\frac{50 \cdot 15 + 100 \cdot 17}{50 + 100} = \frac{750 + 1700}{50 + 100} = \$16.33.$$

The above readily extends to any number of data sets.

## Geometric Mean

The **geometric mean** is used when we are interested in the growth rate or rate of change of a variable. A good example is the “average” return of an investment.

Formally, let  $r_i$  be the rate of return for time period  $i$ . Then, the geometric mean  $r_g$  of the rates  $r_1, r_2, \dots, r_n$  is defined by:

$$(1 + r_g)^n = (1 + r_1)(1 + r_2) \cdots (1 + r_n).$$

Solving for  $r_g$  yields:

$$r_g = [(1 + r_1)(1 + r_2) \cdots (1 + r_n)]^{1/n} - 1.$$

Example: Suppose the rates of return are

| Year 1 | Year 2 | Year 3 | Year 4 | Year 5 |
|--------|--------|--------|--------|--------|
| 0.07   | 0.10   | 0.12   | 0.30   | 0.15   |

Then, the *geometric* mean of these rates is

$$\begin{aligned} r_g &= [(1 + 0.07)(1 + 0.1) \cdots (1 + 0.15)]^{1/5} - 1 \\ &= 0.145, \end{aligned}$$

which is different from the *arithmetic* mean 0.148.

## Median...

Finding the **median** is a two-step process. First, arrange all observations in ascending order; then, identify the observation that is located in the *middle*.

Example 1:  $n$  is odd.

$$\text{Data} = \{3, -1, 6, 10, 11\}$$

$$\text{Ordered Data} = \{-1, 3, 6, 10, 11\}$$

$$\text{Median} = 6$$

Thus, the median is the value at position  $(n + 1)/2$ .

Example 2:  $n$  is even.

$$\text{Data} = \{3, -1, 6, 10, 11, 7\}$$

$$\text{Ordered Data} = \{-1, 3, 6, 7, 10, 11\}$$

Median = any value between 6 and 7. The usual convention is to take the average of these two values, yielding 6.5.

Thus, the median is the average of the two values at positions  $n/2$  and  $(n/2) + 1$ .

## Comments

- For raw data, median can be found using the Excel function MEDIAN().
- The median is based on order. For example, if the value 11 in Example 2 is replaced by 12,000, the median would still remain at 6.5. The arithmetic mean, on the other hand, would be severely affected by the presence of extreme values, or “outliers.”
- The median cannot be manipulated algebraically.
- The median of pooled data sets cannot be determined by those of the original data sets. Hence, reordering is necessary.

## Grouped Data

The following table gives the frequency distribution of the thicknesses of steel plates produced by a machine.

| Lower<br>Limit | Upper<br>Limit | Midpoint,<br>$m(k)$ | Frequency,<br>$f(k)$ | Cum. Freq.,<br>$F(k)$ |
|----------------|----------------|---------------------|----------------------|-----------------------|
| 341.5          | 344.5          | 343                 | 1                    | 1                     |
| 344.5          | 347.5          | 346                 | 3                    | 4                     |
| 347.5          | 350.5          | 349                 | 8                    | 12                    |
| 350.5          | 353.5          | 352                 | 8                    | 20                    |
| 353.5          | 356.5          | 355                 | 20                   | 40                    |
| 356.5          | 359.5          | 358                 | 13                   | 53                    |
| 359.5          | 362.5          | 361                 | 5                    | 58                    |
| 362.5          | 365.5          | 364                 | 2                    | 60                    |

$$n = 60$$

Can we determine the median? Clearly, we need an approximation scheme.

From the  $F(k)$  column, we see that the 30th and the 31th observations ( $n$  is even in this case) lie in the interval  $(353.5, 356.5]$ . Since we have 20 observations in this interval, we now make the following *assumption*:

The data points in a bin are equi-spaced.

It follows that the 30th value is approximately at

$$353.5 + 10 \cdot (3/20) = 355.$$

Similarly, the 31th value  $\approx 355.15$  (the notation  $\approx$  denotes an approximation).

Hence, the median  $\approx (355 + 355.15)/2 = 355.075$ .

## Theoretical Density

Let  $f(x)$  be the probability density function that describes a population. Denote the median by  $\mu_{med}$ . Then,  $\mu_{med}$  satisfies the integral equation:

$$\int_{-\infty}^{\mu_{med}} f(x) dx = 0.5. \quad (3)$$

For a *standard*  $f(x)$ , this can be read from a table. Excel can also be used.

Equation (3) can be visualized as follows. First, compute

$$F(y) \equiv \int_{-\infty}^y f(x) dx ;$$

then, pick  $\mu_{med}$  so that  $F(\mu_{med}) = 0.5$ . This is similar to finding a percentile using an ogive (see figure on p. 24 of the previous set of notes).

## Mode . . .

The **mode** of a set of observations is the value that occurs most *frequently*.

A set of data may have one mode, or two, or more modes.

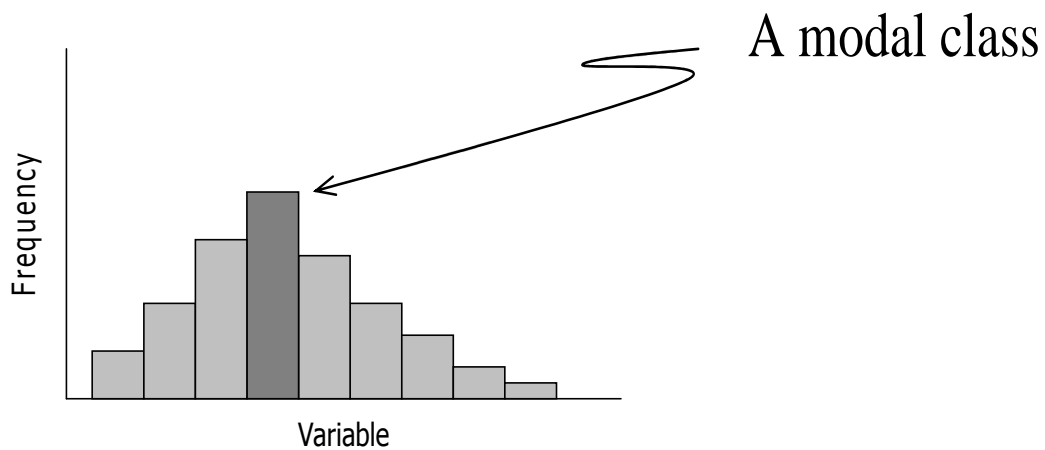
For raw data, there may not be any repeated values; therefore, mode may not be meaningful. The Excel function MODE() can be used to find the mode, but note that this function only finds the *smallest/first* mode; hence, you will need, e.g., a histogram to find modes.

Mode is useful for all data types, though mainly used for *nominal data*.

For grouped data, mode is taken as the midpoint of the bin with the greatest frequency. This bin itself is called the *modal class*.

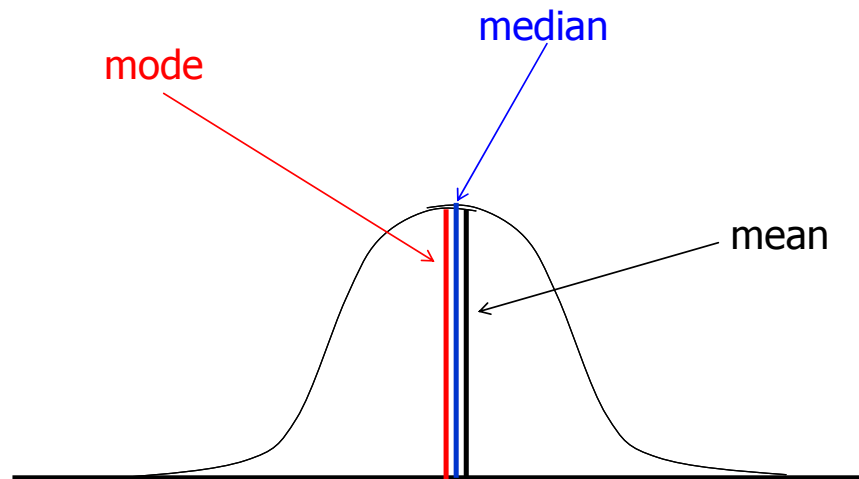
For a theoretical density, mode is the location of a peak of the density curve.

For large data sets the modal class is much more relevant than a single-value mode. Example: Data = {0, 7, 12, 5, 14, 8, 0, 9, 22, 33}. The mode is 0, but this is certainly not a good measure of *central* location. Identifying the modal class would be much better, as in:

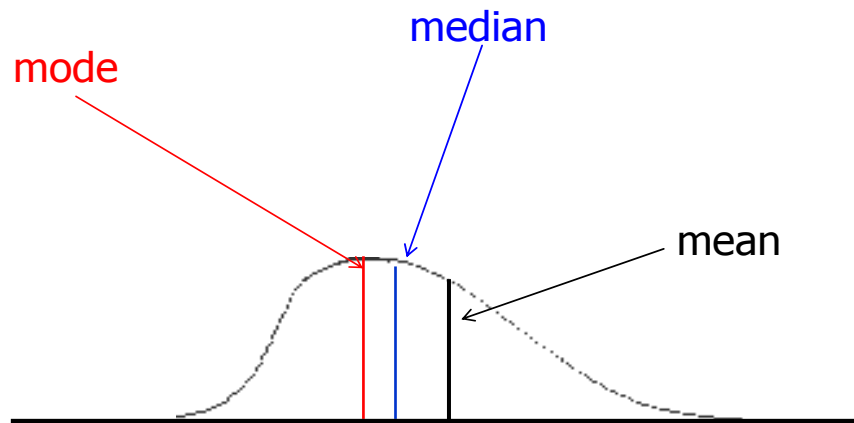


# Mean, Median, Mode — Relationship...

If a distribution is *symmetrical*, the mean, median and mode may coincide:



If a distribution is *asymmetrical*, say skewed to the left or to the right, the three measures may differ:



If data are skewed, reporting the median is recommended.

## Comments

- For ordinal and nominal data the calculation of the mean is not valid.
- Median is appropriate for ordinal data.
- For nominal data, a mode calculation is useful for determining the highest frequency but not “central location.” (Recall the beverage-preference example from last session.)

## Summary...

Use the mean to:

- Describe the central location of a single set of interval data

Use the median to:

- Describe the central location of a single set of interval or ordinal data

Use the mode to:

- Describe a single set of nominal data

Use the geometric mean to:

- Describe a single set of interval data based on growth rates

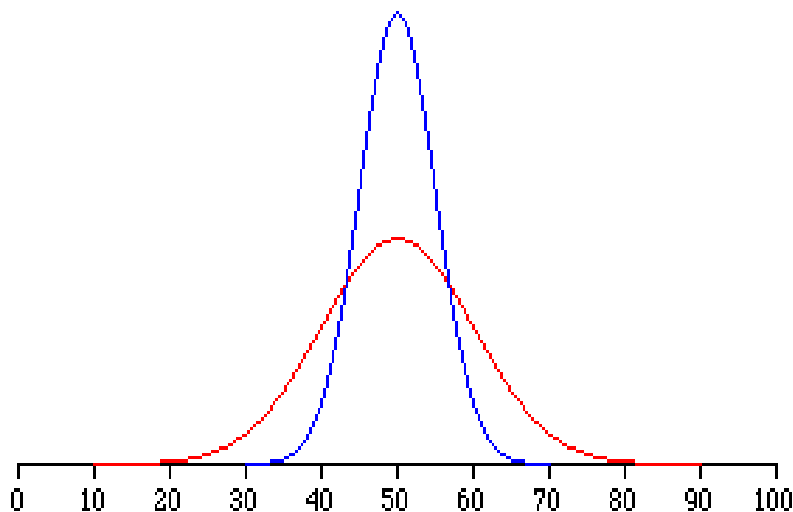
# Measures of Variability...

---

Measures of central location fail to tell the whole story about the distribution. Question: How much are the observations *spread out* around the mean value?

This question makes sense only for *interval* data.

Example: Shown below are the curves for two sets of grades. The means are the same at 50. However, the **red** class has greater variability than the **blue** class.



How do we quantify variability? This is a most important question in statistics.

## Range. . .

The **range** is the simplest measure of variability, calculated as:

$$\text{Range} = \text{Largest Observation} - \text{Smallest Observation}$$

Example 1:

$$\text{Data} = \{4, 4, 4, 4, 50\}$$

$$\text{Range} = 46$$

Example 2:

$$\text{Data} = \{4, 8, 15, 24, 39, 50\}$$

$$\text{Range} = 46$$

The range is the same in both cases, but the data sets have very different distributions. The problem is that the range fails to provide information on the dispersion of the observations between the two end points.

## Mean Absolute Deviation...

The **mean absolute deviation**, or MAD, is the average of all *absolute* deviations from the arithmetic mean. That is,

$$\text{MAD} \equiv \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| .$$

Note that if we did not take absolute values, the average deviation would equal to 0, which offers no information!

In the last two examples, the MADs are 14.72 and 14.33, respectively. (Exercise: Do this calculation in Excel.) This shows that Example 1 has greater dispersion.

The MAD incorporates information from all data points, and therefore is better than the range. It is, however, not easily manipulated algebraically.

Hence...

## Variance...

The **variance** is the average of all *squared* deviations from the arithmetic mean. The *population* variance is denoted by  $\sigma^2$  (“sigma” squared) and the *sample* variance by  $s^2$ . Their definitions differ slightly:

$$\sigma^2 \equiv \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2. \quad (4)$$

$$s^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (5)$$

For  $s^2$ , the division is by  $n-1$  because this makes *repeated* realizations of the sample variance *unbiased* (more on this later).

An equivalent formula for  $s^2$  is:

$$s^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right]. \quad (6)$$

This is somewhat easier to compute, as it relies directly on raw data (as opposed to via deviations from  $\bar{x}$ ).

Example: Job application counts of students.

Data = {17, 15, 23, 7, 9, 13}

Sample Size = 6

Sample Mean = 14 jobs

Sample Variance — Formula (5)

$$s^2 = \frac{1}{6 - 1} [(17 - 14)^2 + \dots + (13 - 14)^2] = 33.2$$

Sample Variance — Formula (6)

$$\begin{aligned} s^2 &= \frac{1}{6 - 1} \left[ (17^2 + \dots + 13^2) - \frac{(17 + \dots + 13)^2}{6} \right] \\ &= 33.2 \end{aligned}$$

Note that we did not use formula (4) in the above example, because what we have is a *sample* of six students.

The Excel function VAR() can be used to compute the variance.

## Grouped Data

Excel does not have a function that computes the standard deviation if the data is grouped. The computing formula to use in this case is (similar to (6)) given by:

$$s^2 = \frac{1}{n - 1} \left\{ \sum_{k=1}^K [m(k)]^2 f(k) - n\bar{x}^2 \right\}, \quad (7)$$

where  $\bar{x}$  is computed by formula (1). Like (1), this is only an approximation.

## Theoretical Density

The variance for a population governed by a theoretical density function  $f(x)$  is defined by:

$$\sigma^2 \equiv \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2. \quad (8)$$

where  $\mu$  is computed by (2).

This can be understood as the “limiting” version of (7).

## Standard Deviation...

The standard deviation is simply the *square root* of the variance.

Population standard deviation:

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

Sample standard deviation:

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

For raw data, the Excel function STDEV() can be used to compute the standard deviation (directly).

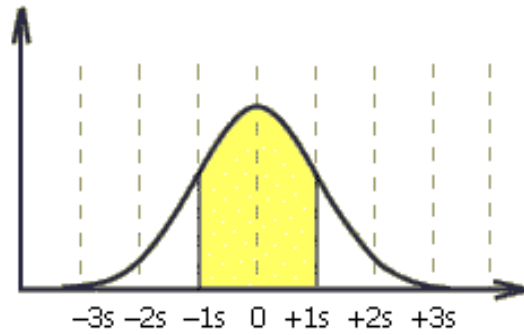
## Interpreting Standard Deviation

The standard deviation can be used to compare the variability of several distributions and make a statement about the general shape of a distribution. If the histogram is reasonably *bell shaped*, we can use the **Empirical Rule**, which states that:

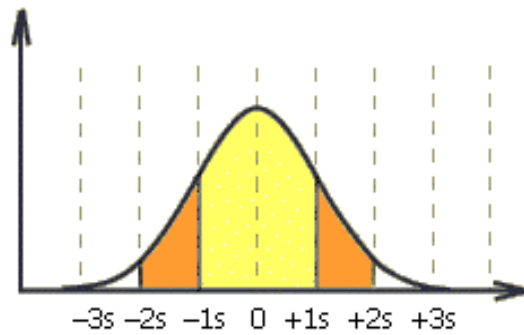
1. Approximately 68% of all observations fall within one standard deviation of the mean.
2. Approximately 95% of all observations fall within two standard deviations of the mean.
3. Approximately 99.7% of all observations fall within three standard deviations of the mean.

Visually...

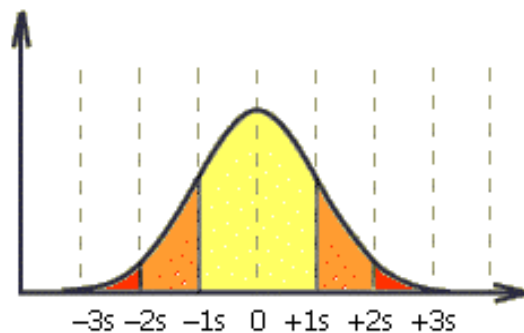
One standard deviation (68%):



Two standard deviations (95%):



Three standard deviations (99.7%):



A more general result based on the standard deviation is the **Chebysheff's Inequality** (or Chebyshev's). This inequality, which applies to *all* histograms (not just bell shaped) states that:

The *proportion* of observations in any sample that lie within **k** standard deviations of the mean is *at least*:

$$1 - \frac{1}{k^2} \quad \text{for } k > 1.$$

Example: For  $k = 2$ , the inequality states that at least  $3/4$  (or 75%) of all observations lie within 2 standard deviations of the mean. Note that this bound is *lower* compared to Empirical Rule's approximation (95%). Thus, although the Chebysheff's inequality is more general, it may not be very "tight."

## Coefficient of Variation...

The **coefficient of variation** of a set of observations is the standard deviation of the observations divided by their mean. For a *population*, this is denoted by CV; and for a *sample*, by cv. That is,

Population Coefficient of Variation:

$$CV \equiv \frac{\sigma}{\mu}.$$

Sample Coefficient of Variation:

$$cv \equiv \frac{s}{\bar{x}}.$$

This coefficient provides a **proportionate** measure of variation; that is, the extent of variation is now measured in units of the mean.

For example, a standard deviation of 10 may be perceived as large when the mean value is 100, but only moderately large when the mean value is 500, or even negligible when the mean value is 5,000.

## Measures of Relative Standing...

---

Measures of relative standing are designed to provide information about the **position** of particular values **relative** to the entire data set.

**Percentile:** The  $P$ th percentile is the value for which  $P\%$  of the data are less than that value and  $(100 - P)\%$  are greater than that value.

Example: Suppose you scored in the 60th percentile on the GMAT, that means 60% of the other scores were below yours, while 40% of scores were above yours.

**Quartiles** are the 25th, 50th, and 75th percentiles.

Notation:

The First Quartile,  $Q_1 = 25$ th percentile

The Second Quartile,  $Q_2 = 50$ th percentile

The Third Quartile,  $Q_3 = 75$ th percentile

## Interquartile Range

The quartiles can be used to create another measure of variability, the **interquartile range**, which is defined as follows:

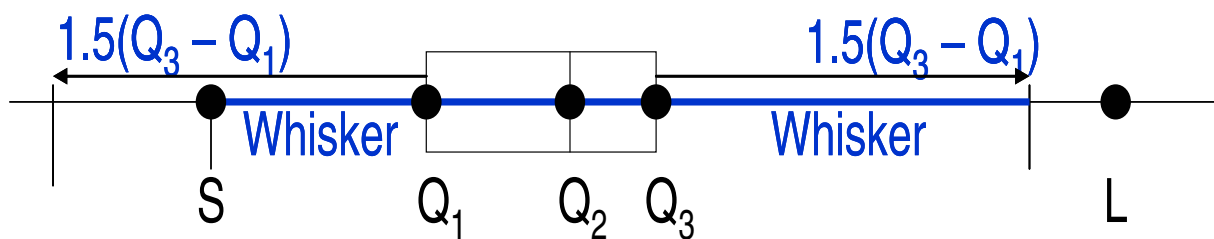
$$\text{Interquartile Range} = Q_3 - Q_1 .$$

The interquartile range measures the spread of the middle 50% of the observations. Large values of this statistic mean that the 1st and 3rd quartiles are far apart, indicating a high level of variability.

## Box Plots

A **box plot** is a pictorial representation of *five* descriptive statistics of a data set:

- $S$ , the smallest observation
- $Q_1$ , the lower quartile
- $Q_2$ , the median
- $Q_3$ , the upper quartile
- $L$ , the largest observation



The lines extending to the left and right are called whiskers. Any points that lie outside the whiskers are called *outliers*. The whiskers extend outward to the smaller of 1.5 times the interquartile range or to the most extreme point that is not an outlier.

## Measures of Relationship...

---

We now define two important numerical measures that provide information as to the *direction and strength* of the relationship between two variables. These are the **co-variance** and the **coefficient of correlation**. They are closely related to each other.

The covariance is designed to measure whether or not two variables move *together*.

The coefficient of correlation is designed to measure the *strength* of the relationship between two variables.

## Covariance...

Again, we have two definitions, one for the population and one for a sample. The former is denoted by  $\sigma_{XY}$ , where  $X$  and  $Y$  are two variables of interest (e.g.,  $Y$  = selling price of a house, and  $X$  = house size); and the latter is denoted by  $s_{XY}$ .

Population Covariance:

$$\sigma_{XY} \equiv \frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \quad (9)$$

Sample Covariance:

$$s_{XY} \equiv \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (10)$$

In the definitions above,

$\mu_X$  = population mean of  $X$

$\mu_Y$  = population mean of  $Y$

$\bar{x}$  = sample mean of  $X$

$\bar{y}$  = sample mean of  $Y$

Similar to sample variance, formula (10) also has an equivalent form that is easier to calculate:

$$s_{XY} = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n} \right] \quad (11)$$

Illustration:

|               | $x_i$ | $y_i$          | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---------------|-------|----------------|-----------------|-----------------|----------------------------------|
| Set 1         | 2     | 13             | -3              | -7              | 21                               |
|               | 6     | 20             | 1               | 0               | 0                                |
|               | 7     | 27             | 2               | 7               | 14                               |
| $\bar{x} = 5$ |       | $\bar{y} = 20$ | $s_{XY} = 17.5$ |                 |                                  |

|               | $x_i$ | $y_i$          | $x_i - \bar{x}$  | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---------------|-------|----------------|------------------|-----------------|----------------------------------|
| Set 2         | 2     | 27             | -3               | 7               | -21                              |
|               | 6     | 20             | 1                | 0               | 0                                |
|               | 7     | 13             | 2                | -7              | -14                              |
| $\bar{x} = 5$ |       | $\bar{y} = 20$ | $s_{XY} = -17.5$ |                 |                                  |

|               | $x_i$ | $y_i$          | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---------------|-------|----------------|-----------------|-----------------|----------------------------------|
| Set 3         | 2     | 20             | -3              | 0               | 0                                |
|               | 6     | 27             | 1               | 7               | 7                                |
|               | 7     | 13             | 2               | -7              | -14                              |
| $\bar{x} = 5$ |       | $\bar{y} = 20$ | $s_{XY} = -3.5$ |                 |                                  |

In these data sets, the sampled  $X$  and  $Y$  values are the same. The only differences are the *order* of the  $y_i$ s.

- In Set 1, as  $X$  increases, so does  $Y$ ; we see that  $s_{XY}$  is large and positive.
- In Set 2, as  $X$  increases,  $Y$  decreases; we see that  $s_{XY}$  is large and negative.
- In Set 3, as  $X$  increases,  $Y$ 's movement is unclear; we see that  $s_{XY}$  is “small”

Generally speaking. . .

When two variables move in the *same direction* (both increase or both decrease), the covariance will be a *large positive number*.

When two variables move in *opposite directions*, the covariance is a *large negative number*.

When there is *no particular pattern*, the covariance is a *small number*.

## Coefficient of Correlation...

The coefficient of correlation is defined as the covariance divided by the product of the standard deviations of the variables. That is,

Population Coefficient of Correlation:

$$\rho_{XY} \equiv \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (12)$$

Sample Coefficient of Correlation:

$$r_{XY} \equiv \frac{s_{XY}}{s_X s_Y} \quad (13)$$

The scaling in the denominators means that we are now measuring the size of covariance relative to the standard deviations of  $X$  and  $Y$ . This is similar in spirit to the coefficient of variation.

The result of this scaling is that the coefficient of correlation will always have a value between  $-1$  and  $1$ . When its value is close to  $1$ , there is a strong positive relationship; when its value is close to  $-1$ , a strong negative relationship; and when close to  $0$ , a weak relationship.

## Linear Relationship. . .

Of particular interest is whether or not there exists a *linear* relationship between two variables.

Recall that the equation of a line can be written as:

$$y = mx + b,$$

where  $m$  is the slope and  $b$  is the y-axis intercept.

If we believe a linear relationship is a good approximation of the relationship between two variables, what should be our choices for  $m$  and  $b$ ?

The answer to this question is based on the **least-squares method**. . .

## The Least-Squares Method

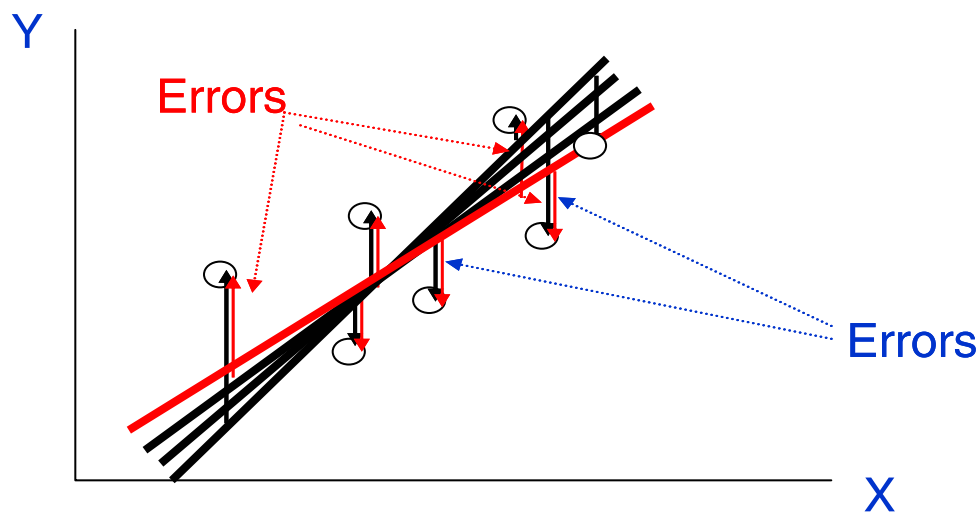
The idea is to construct a straight line through the points (i.e., the  $(x_i, y_i)$  pairs) so that the sum of squared deviations between the points and the line is minimized.

Suppose this line has the form:

$$\hat{y} = b_0 + b_1x ,$$

where  $b_0$  is the  $y$ -intercept,  $b_1$  is the slope, and  $\hat{y}$  can be thought of as the value of  $y$  “estimated” by the line. Clearly, the estimate  $\hat{y}$  would typically be different from the observed  $y_i$ , for a given  $x_i$ .

Visually...



It can be shown (by standard methods in calculus) that the best choices for  $b_0$  and  $b_1$  are given by:

$$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \frac{s_Y}{s_X}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}.$$

It is interesting to observe that  $b_1$  is proportional to  $r_{XY}$  and to the ratio  $s_Y/s_X$ .

We will return to this topic in later chapters.

# Summary of Notation...

---

|                            | Population    | Sample    |
|----------------------------|---------------|-----------|
| Size                       | $N$           | $n$       |
| Mean                       | $\mu$         | $\bar{x}$ |
| Variance                   | $\sigma^2$    | $s^2$     |
| Standard Deviation         | $\sigma$      | $s$       |
| Coefficient of Variation   | CV            | cv        |
| Covariance                 | $\sigma_{XY}$ | $s_{XY}$  |
| Coefficient of Correlation | $\rho_{XY}$   | $r_{XY}$  |