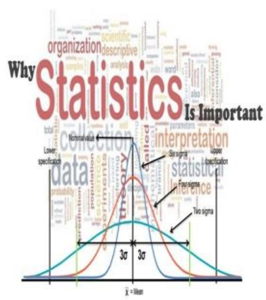
# ECON 2110
# Fall 2025

**Chapter 4: Numerical Descriptive Techniques**




### Sample Statistic or Population Parameter

A **parameter** is a descriptive measurement about a population, and a **statistic** is a descriptive measurement about a sample.

In this chapter, we introduce a dozen descriptive measurements. For each one, we describe how to calculate both the population parameter and the sample statistic.

In most realistic applications, the calculation of parameters are not practical, and they are provided here primarily to teach the concept and the notation.

2

### Measures of Location

**Univariate analysis is the simplest form of** quantitative (statistical) analysis. The analysis is carried out with the description of a single variable and its attributes of the applicable unit of analysis.

**Univariate analysis** contrasts with **bivariate** analysis - the analysis of two variables simultaneously - or **multivariate analysis** - the analysis of multiple variables simultaneously.

**Univariate** analysis is also used primarily for descriptive purposes, while **bivariate** and **multivariate** analyses are geared more towards explanatory purposes.

3

### Arithmetic 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 adding up all the observations and dividing by the total number of observations:

**Population mean**: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

**Sample mean**: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

In the notation presented here, N is the population size, and n the sample size.

4

# Chapter 4: Numerical Descriptive Techniques

## Weighted Mean

- The weighted mean of a set of numbers $X_1$, $X_2$, ..., $X_n$, with corresponding weights $w_1$, $w_2$, ...,$w_n$, is computed from the following formula:

WEIGHTED MEAN $$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n}{w_1 + w_2 + w_3 + \cdots + w_n}$$

5

## Weighted Mean EXAMPLE – Weighted Mean

The Carter Construction Company pays its hourly employees \$16.50, \$19.00, or \$25.00 per hour. There are 26 hourly employees, 14 of which are paid at the \$16.50 rate, 10 at the \$19.00 rate, and 2 at the \$25.00 rate. What is the mean hourly rate paid the 26 employees?

$$\bar{X}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$26.00)}{14 + 10 + 2}$$
$$= \frac{\$471.00}{26} = \$18.1154$$

6

### Question

Find the mean age of the 200 ACBL members

DATA: Xm03-01

**Solution:**

To calculate the mean, we add the observations and divide by the size of the sample.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{73 + 53 + 66 + \cdots + 17}{200} = \frac{10,753}{200}$$
$$\bar{x} = 53.765$$

### EXCEL Function

- If we want to compute the mean and no other statistics, we can use the AVERAGE function.
- **Instructions**:
- Type or import the data into one or more columns. (Open Xm03-01.) Type into any empty cell:
- = **AVERAGE**([Input Range])
- We would type into any empty cell:
- = **AVERAGE**(A2:A201)
- The active cell would store the mean as 53.765.



7

The **median** is calculated by placing all the observations in order (ascending or descending).

The observation that falls in the middle is the median.

The sample and population medians are computed in the same way.

When there is an even number of observations, the median is determined by averaging the two observations in the middle.

The **mode** is defined as the observation (or observations) that occurs with the greatest frequency.

Both the statistic and parameter are computed in the same way.



8

### Question

Find the median age of the 200 ACBL members

DATA: Xm03-01

**Solution:**

Because there is an even number of observations, the median is the average of the two middle ones.

When all the observations are placed in order, the 100th and 101st observations are 54 and 55, respectively.

$$\text{Median} = \frac{54 + 55}{2} = 54.5$$

### EXCEL Function

- If we want to compute the median and no other statistics, we can use the MEDIAN function.
- **Instructions**:
- Type or import the data into one or more columns. (Open Xm03-01.) Type into any empty cell:
- = **MEDIAN**([Input Range])
- In Example 4.4, we would type into any empty cell:
- = **MEDIAN**(A2:A201)
- The active cell would store the median as 54.5.




## Median

- For an odd number of observations:

| 26 | 18 | 27 | 12 | 14 | 27 | 19 |
|---|---|---|---|---|---|---|

7 observations

| 12 | 14 | 18 | 19 | 26 | 27 | 27 |
|---|---|---|---|---|---|---|

in ascending order

the median is the middle value. Median = 19

- For an even number of observations:

| 26 | 18 | 27 | 12 | 14 | 27 | 30 | 19 |
|---|---|---|---|---|---|---|---|

8 observations

| 12 | 14 | 18 | 19 | 26 | 27 | 27 | 30 |
|---|---|---|---|---|---|---|---|

in ascending order

the median is the average of the middle two values.
Median = (19 + 26)/2 = 22.5



### Question

Find the mode of ages of the 200 ACBL members.

DATA: Xm03-01

**Solution:**

The observation that occurs with the greatest frequency is 60, which occurs 8 times.

### EXCEL Function

- If we want to compute the mode and no other statistics, we can use the MODE.SNGL function.
- **Instructions**:
- Type or import the data into one or more columns. (Open Xm03-01
- Type into any empty cell:
- = **MODE.SNGL**([Input Range])
- We would type into any empty cell:
- = **MODE.SNGL**(A1:A201)
- The active cell would store the mode as 60.
- Note that if there is more than one mode, Excel prints only the smallest one, without indicating that there are other modes.
- For multiple modes, use the Excel function MODE.MULT.




With three measures from which to choose, which one should we use?

- The mean is generally our first selection.
- However, one advantage the median holds is that it not as sensitive to skewed distributions or to extreme values as is the mean.
- The mode is seldom the best measure of central location.

# Chapter 4: Numerical Descriptive Techniques

## Geometric Mean

- The geometric mean is calculated by finding the $n$th root of the product of $n$ values.
- It is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results).
- It should be applied anytime you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, . . .).
- Other common applications include: changes in populations of species, crop yields, pollution levels, and birth and death rates.

13

## Geometric Mean

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\dots(x_n)} = [(x_1)(x_2)\dots(x_n)]^{1/n}$$

- Excel's geometric mean function is:

=GEOMEAN(data cell range)

14

## Geometric Mean

- Example: Rate of Return

| Period | Return (%) |
|---|---|
| 1 | -6.0 |
| 2 | -8.0 |
| 3 | -4.0 |
| 4 | 2.0 |
| 5 | 5.4 |

| Growth Factor |
|---|
| 0.940 |
| 0.920 |
| 0.960 |
| 1.020 |
| 1.054 |

$$\bar{x}_g = \sqrt[5]{(.94).92)(.96)(1.02)(1.054)}$$

$$= [.89254]^{1/5} = .97752$$

Average growth rate per period is (.97752 - 1) (100) = -2.248%

15

## Measures of Central Location for Ordinal and Nominal Data

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 the "central location".

Summary of the measures discussed in this section and when to use them:

| Measure | Type of Data | Descriptive Measurement |
|---|---|---|
| Mean | Interval | Central location |
| Median | Ordinal or interval | Central location |
| Mode | Nominal, ordinal, or interval | |
| Geometric Mean | Interval, growth rates | |

16

## Measures of Variability

Besides the measures of central location, there are other characteristics of data that are of interest to us, such as the spread or variability of the data.

In this section, we introduce **four measures of variability** for interval data.
We begin with the simplest:

**Range** = Largest observation – Smallest observation

The range is a very simple measure, but because it only considers the two extreme values in the data, we need to consider other statistics that incorporate all the values.

Example:

| | | | | | | |
|---|---|---|---|---|---|---|
| Set 1: | 4 | 4 | 4 | 4 | 4 | 50 |
| Set 2: | 4 | 8 | 15 | 24 | 39 | 50 |

17

## Variance

The **variance** and its related measure, the **standard deviation**, are arguably the most important measures of variability. They are used to measure variability, but, as you will discover, they play a vital role in almost all statistical inference procedures.

**Population variance**: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$

**Sample variance**: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

The population variance is represented by $\sigma^2$ (Greek letter sigma squared).

*Why n-1?*

18

## Summer Jobs

**Sample Variance**

The following are the number of summer jobs a sample of six students applied for:

17 15 23 7 9 13

Find the mean and variance of these data.

The mean of the six observations is: $\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14$ jobs

The sample variance is: $s^2 = \frac{\sum_{i=1}^{6}(x_i-\bar{x})^2}{6-1} = \frac{(17-14)^2+(15-14)^2+(23-14)^2+(7-14)^2+(9-14)^2+(13-14)^2}{5}$

$$s^2 = \frac{9+1+81+49+25+1}{5} = \frac{166}{5} = 33.2\ \boldsymbol{jobs^2}$$

19

## Shortcut calculation of the sample variance

**Shortcut calculation of the sample variance:**

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$$ **Data** 17 15 23 7 9 13

Find the mean and variance of these data.

Where:

$\sum x_i^2 = 17^2 + 15^2 + 23^2 + 7^2 + 9^2 + 13^2 = 1{,}342$ $\sum x_i = 17+15+23+7+9+13=84$

$\left(\sum x_i\right)^2 = 84^2 = 7{,}056$

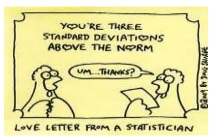
Thus:

$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] = \frac{1}{6-1}\left(1{,}342 - \frac{7.056}{6}\right) = 33.2\text{ jobs}^2$

20

Because to calculate the variance we squared the deviations from the mean, the unit of the variance is the square of the unit of the original observations.

We resolve this complication by introducing the **standard deviation**, defined as the square root of the variance.








## Standard Deviation

- The **standard deviation**, is defined as the square root of the variance.

  **Population standard deviation**: $\sigma = \sqrt{\sigma^2}$

  **Sample standard deviation**: $s = \sqrt{s^2}$

- The population standard deviation is represented by σ (Greek letter sigma).
- In Example 4.1, the sample standard deviation is:
- $s = \sqrt{s^2} = \sqrt{33.2} = 5.76$ jobs
- Notice how the unit associated with the standard deviation is the unit of the original data set.



A golfer was asked to hit 150 shots using a 7 iron, 75 of which were hit with the current club and 75 with the new innovative 7 iron. The distances were measured and recorded in file Xm04-08. Which 7 iron is more consistent?

**Solution:**
We use the standard deviations to gauge consistency. We can use the Excel formula STDEV.S or we can use the Descriptive Statistics tool in the Data Analysis, as shown to the right.

**Interpret:**
Based on this sample, the innovative club is more consistent (5.79 > 3.09.) Because the mean distances are also very similar (150.55 vs. 150.15), it appears that the new club is indeed superior.





## The Empirical Rule

- Knowing the mean and standard deviation allows the statistics practitioner to extract useful bits of information from the data.
- The information depends on the shape of the histogram. If the histogram is bell shaped, we can use the **Empirical Rule**:

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.

### Using the Empirical Rule to Interpret Standard Deviation

After an analysis of the returns on an investment, a statistics practitioner discovered that the histogram is bell shaped and that the mean and standard deviation are 10% and 8% , respectively.

What can you say about the way the returns are distributed?

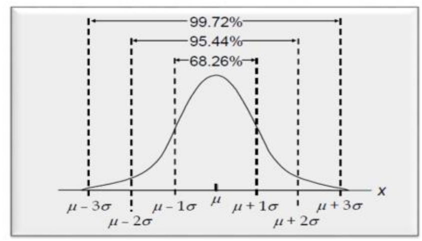
**Solution:**

Because the histogram is bell shaped, we can apply the Empirical Rule:

1. Approximately 68% of the returns lie between: (10 – 8)% = 2% and 10 + 8)% = 18%
2. Approximately 95% of the returns lie between: [10 – 2(8)]% = –6% and [10 + 2(8)]% = 26%
3. Approximately 99.7% of the returns lie between: (10 – 3(8)]% = –14% and [10 + 3(8)]% = 34%



### Empirical Rule






### Chebyshev's Theorem

At least $(1 - 1/z^2)$ of the items in any data set will be within $z$ standard deviations of the mean, where $z$ is any value greater than 1.

Chebyshev's theorem requires $z > 1$, but $z$ need not be an integer.

At least 75% of the data values must be within $z = 2$ standard deviations of the mean.

At least 89% of the data values must be within $z = 3$ standard deviations of the mean.

At least 94% of the data values must be within $z = 4$ standard deviations of the mean.



### Using Chebysheff's Theorem to Interpret Standard Deviation

The annual salaries of the employees of a chain of computer stores produced a positively skewed histogram. The mean and standard deviation are $28,000 and $3,000, respectively.

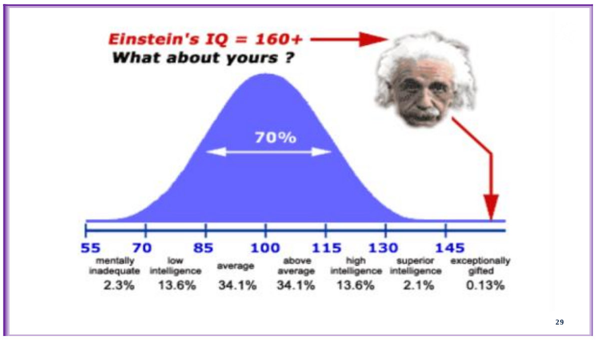
What can you say about the salaries at this chain?

**Solution:**

Because the histogram is not bell shaped, we must apply Chebysheff's Theorem instead.

The intervals created by adding and subtracting two and three standard deviations to and from the mean are as follows:

1. At least 75% of the salaries lie between ($28k – 2 · $3k) = $22k and ($28k + 2 · $3k) = $34k
2. At least 88.9% of the salaries lie between ($28k – 3 · $3k) = $19k and ($28k + 3 · $3k) = $37k



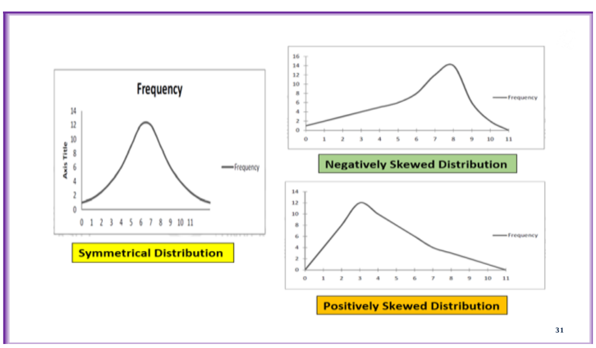

## Distribution Shape: Skewness

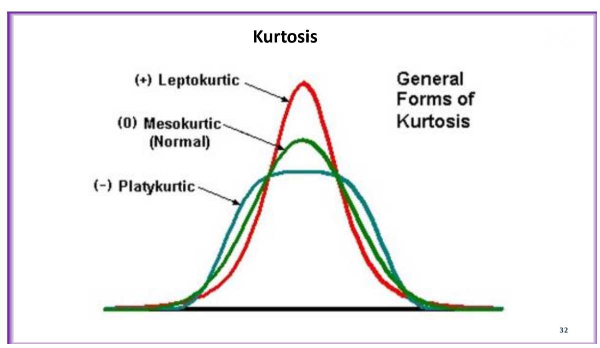
- An important measure of the shape of a distribution is called skewness.
- The formula for the skewness of sample data is

$$\textbf{Skewness} = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$$

- Skewness can be easily computed using statistical software.
- Excel's SKEW function can be used to compute the skewness of a data set.
- A skewness value of zero generally indicates a symmetric distribution.

30



## Coefficient of Variation

Because 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, it makes sense to introduce a measure of variability that measures the standard deviation as a proportion of the mean.

The **coefficient of variation** of a set of observations is the standard deviation of the observations divided by their mean:

**Population coefficient of variation**: $\mathrm{CV} = \frac{\sigma}{\mu}$

**Sample coefficient of variation**: $\mathrm{cv} = \frac{s}{\bar{x}}$

*A ratio of 0.5 indicate means the standard deviation is half as large as the mean.*

33

## Measures of Relative Standing

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

**Percentile**

The $P^{th}$ **percentile** is the value for which *P%* are less than that value and *(100-P)%* are greater than that value.

Suppose you scored in the 60$^{th}$ percentile on the SAT, that means 60% of the other scores were below yours, while 40% were above yours.

Of special importance are the 25$^{th}$, 50$^{th}$, and 75$^{th}$ percentiles, named first quartile ($Q_1$), second quartile ($Q_2$ or the median), and third quartile ($Q_3$), respectively. We can also convert percentiles into quintiles and deciles.

34

## Locating Percentiles

The following formula allows us to approximate the location of any percentile:

$$L_p = (n+1)\frac{p}{100}$$

Where $L_p$ is the location of the $P^{th}$ percentile.

Once the location is known, the percentile is calculated by linearly interpolating the two data points preceding and following the location.

35

## Percentiles of Time Spent on Internet

A sample of 10 adults was asked to report the number of hours they spent on the Internet the previous month. These were the results:

0 7 12 5 33 14 8 0 9 22

Calculate the *25$^{th}$*, *50$^{th}$*, and *75$^{th}$* percentiles ($Q_1$, $Q_2$, and $Q_3$).

Let us first sort the data in ascending order:

0 0 5 7 8 9 12 14 22 33

The location of the *25$^{th}$* percentile is:

$$L_{25} = (10+1)\frac{25}{100} = 11(.25) = 2.75$$

The *25$^{th}$* percentile is three-quarters of the distance between the 2$^{nd}$ and 3$^{rd}$ observations:

0 0 5 7 8 9 12 14 22 33

Thus, the *25$^{th}$* percentile is: 0 + 0.75(5 – 0) = 3.75

36

# Chapter 4: Numerical Descriptive Techniques

## Percentiles of Time Spent on Internet

**The location of the *50th* percentile is:**

$$L_{50} = (10+1)\frac{50}{100} = 11(.50) = 5.5$$

The *50th* percentile is half-way between the 5th and 6th observations:

0 0 5 7 8 9 12 14 22 33

Thus, the *50th* percentile is: 8 + 0.5(9 – 8) = 8.5

**The location of the *75th* percentile is:**

$$L_{75} = (10+1)\frac{75}{100} = 11(.75) = 8.25$$

The *75th* percentile is one-quarter of the distance between the 8th and 9th observations:

0 0 5 7 8 9 12 14 22 33

Thus, the *50th* percentile is: 14 + 0.25(22 – 14) = 16

37

## Quartiles and Percentiles of the Ages of ACBL Members

### Example 1

Determine the quartiles for Example 3.1 with Excel.

DATA: Xm03-01.

The simplest way to instruct Excel to produce quartiles is to use the QUARTILE function.

Type or import the data into one or more columns. (Open Xm03-01.) Type into any empty cell:

= QUARTILE ([Input Range], [Q])

where Q specifies the following: 0 = *min*, 1 = $Q_1$, 2 = $Q_2$ or median, 3 = $Q_3$, 4 = *max*.

Here are the outputs:

$Q_1$ = 36; $Q_2$ = 54.5; $Q_3$ = 69

### Example 2

- Determine the 5th, 10th, 90th, and 95th percentiles for Example 3.1 with Excel.
- DATA: Xm03-01.
- Let us use the PERCENTILE function.
- Type or import the data into one or more columns. (Open Xm03-01.) Type into any empty cell:
  - = PERCENTILE ([Input Range], [P])
- where P is the percentile, a value between 0 and 1.
- The outputs for the 5th, 10th, 90th, and 95th percentiles are, respectively:
  - 24, 27, 78.1, 83



38

## Interquartile Range

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

*Interquartile Range* = $Q_3 - Q_1$

The interquartile range measures the spread of the middle 50% of the observations.

**Interquartile Range**

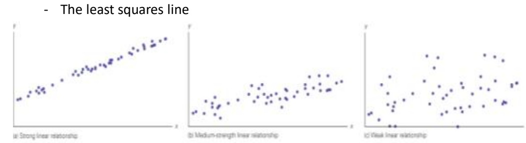
- 3rd Quartile (Q3) = 525
- 1st Quartile (Q1) = 445
- Interquartile Range = Q3 - Q1 = 525 - 445 = 80

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 425 | 430 | 430 | 435 | 435 | 435 | 435 | 435 | 440 | 440 |
| 440 | 440 | 440 | 445 | 445 | 445 | 445 | $\boxed{445}$ | 450 | 450 |
| 450 | 450 | 450 | 450 | 450 | 460 | 460 | 460 | 465 | 465 |
| 465 | 470 | 470 | 472 | 475 | 475 | 475 | 480 | 480 | 480 |
| 480 | 485 | 490 | 490 | 490 | 500 | 500 | 500 | 500 | 510 |
| 510 | 515 | $\boxed{525}$ | 525 | 525 | 535 | 549 | 550 | 570 | 570 |
| 575 | 575 | 580 | 590 | 600 | 600 | 600 | 600 | 615 | 615 |

39

## Measures of Linear Relationship

- Establish a relationship between two interval variables
- Direction and strength of the linear relationship
- Three numerical measures:
  1. Covariance
  2. Coefficient of correlation
  3. Coefficient of determination
- The least squares line



40

## Covariance

**Population covariance**: $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}$

**Sample covariance**: $s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

The denominator in the calculation of the sample covariance is *n–1*

There is also a shortcut for the calculation of the sample covariance:

$$s_{xy} = \frac{1}{n-1}\left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right]$$

41

## Calculating the Covariance

Compute the variance for the following datasets:

| Set 1 | $x_i$ | $y_i$ |
|---|---|---|
| | 2 | 13 |
| | 6 | 20 |
| | 7 | 27 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ |
| **Set 2** | $x_i$ | $y_i$ |
| | 2 | 27 |
| | 6 | 20 |
| | 7 | 13 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ |
| **Set 3** | $x_i$ | $y_i$ |
| | 2 | 20 |
| | 6 | 27 |
| | 7 | 13 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ |

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

The $x_i$ and $y_i$ values are the same across the three sets.

The only difference is the order of the $y_i$ values.

42

## Calculating the Covariance

| Set 1 | $x_i$ | $y_i$ | $(x_i-\bar{x})$ | $(y_i-\bar{y})$ | $(x_i-\bar{x})(y_i-\bar{y})$ |
|---|---|---|---|---|---|
| | 2 | 13 | –3 | –7 | 21 |
| | 6 | 20 | 1 | 0 | 0 |
| | 7 | 27 | 2 | 7 | 14 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ | | | $s_{xy} = 35/2 = 17.5$ |
| **Set 2** | $x_i$ | $y_i$ | $(x_i-\bar{x})$ | $(y_i-\bar{y})$ | $(x_i-\bar{x})(y_i-\bar{y})$ |
| | 2 | 27 | –3 | 7 | –21 |
| | 6 | 20 | 1 | 0 | 0 |
| | 7 | 13 | 2 | –7 | –14 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ | | | $s_{xy} = -35/2 = -17.5$ |
| **Set 3** | $x_i$ | $y_i$ | $(x_i-\bar{x})$ | $(y_i-\bar{y})$ | $(x_i-\bar{x})(y_i-\bar{y})$ |
| | 2 | 20 | –3 | 0 | 0 |
| | 6 | 27 | 1 | 7 | 7 |
| | 7 | 13 | 2 | –7 | –14 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ | | | $s_{xy} = -7/2 = -3.5$ |

Interpreting **sign** and **magnitude**:

- In Set 1, as *x* increases so does *y*. As a result, the products of the deviations are also positive, and $s_{xy}$ is large and positive.
- In Set 2, as *x* increases, *y* decreases. As a result, the products of the deviations are negative, and $s_{xy}$ is large and negative.
- In Set 3, as *x* increases, *y* does not exhibit any particular direction. As a result, the contributions to the products of the deviations largely cancel each other, and $s_{xy}$ is small.

*Limitation: magnitude is hard to interpret!*

43

## Correlation Coefficient

The coefficient can take on values between -1 and +1.

Values near -1 indicate a <u>strong negative linear relationship</u>.

Values near +1 indicate a <u>strong positive linear relationship</u>.

The closer the correlation is to zero, the weaker the relationship.

44

## Coefficient of Correlation

Because the magnitude of the covariance is difficult to evaluate, we can "normalize" the covariance dividing it by the product of the standard deviations.

**Population coefficient of correlation**: $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

**Sample coefficient of correlation**: $\mathrm{r} = \frac{s_{xy}}{s_x s_y}$

The coefficient of correlation is **always between –1 and +1**. The sign provides the direction of the association, and the magnitude measures the *strength of the linear relationship* between *x* and *y*.

*How do you interpret a coefficient of correlation of -0.4?*

45

## Calculating the Coefficient of Correlation

Calculate the coefficient of correlation for the three sets shown previously.

| $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - y)^2$ |
|---|---|---|---|---|---|
| 2 | 13 | –3 | –7 | 9 | 49 |
| 6 | 20 | 1 | 0 | 1 | 0 |
| 7 | 27 | 2 | 7 | 4 | 49 |
| $\bar{x} = 5$ | $\bar{y} = 20$ | | | $s_x = \sqrt{14/2} = 2.65$ | $s_y = \sqrt{98/2} = 7$ |

The coefficients of correlation are:

Set 1: $\mathrm{r} = \frac{s_{xy}}{s_x s_y} = \frac{17.5}{2.65 \cdot 7} = +0.943$

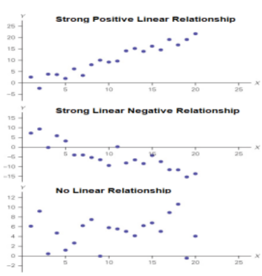
Set 2: $\mathrm{r} = \frac{s_{xy}}{s_x s_y} = \frac{-17.5}{2.65 \cdot 7} = -0.943$

Set 3: $\mathrm{r} = \frac{s_{xy}}{s_x s_y} = \frac{-3.5}{2.65 \cdot 7} = -0.189$

It is now easier to see the strength of the linear relationship between X and Y.

46

## Comparing Scatter Diagrams




- **Strong positive linear relationship:**
  - $\sigma_{xy} = 36.87$
  - $\rho = 0.964$
- **Strong negative linear relationship:**
  - $\sigma_{xy} = -34.18$
  - $\rho = -0.879$
- **No linear relationship:**
  - $\sigma_{xy} = 2.07$
  - $\rho = 0.121$

47

## Least Squares Method

The strength and direction of a linear relationship can be better judged by drawing a straight line through the data.

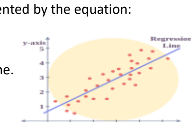
The objective method of producing a straight line that fits the data is called the **least squares method**.

The line produced by the least square method is represented by the equation:

$\hat{y} = b_0 + b_1 x$

Where:

- $\hat{y}$ ("*y hat*") is the value of *y* determined by the line.
- $b_0$ ("b naught") is the y-intercept
- $b_1$ is the slope.




48

## Derivation of Least Line Square Coefficients

The least line square coefficients $b_0$ and $b_1$ are derived so that they minimize the sum of the squared deviations:

$$\sum_{i=1}^{n}(y_i - \hat{y})^2$$

With the least line square coefficients are calculated as:

$$b_1 = \frac{s_{xy}}{s_x^2}$$

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

49

## Application in Economics

Fixed costs are costs that must be paid whether or not any units are produced.

These costs are "fixed" over a specified period of time or range of production.

Variable costs are costs that vary directly with the number of products produced.

There are some expenses that are mixed.

One method to break the mixed costs in its fixed and variable components is the least squares line, with the total cost expressed as:

$$y = b_0 + b_1 x$$

Where:

$y$ = total mixed cost

$b_0$ = fixed cost

$b_1$ = variable cost

$x$ = number of units.

50

## Estimating Fixed and Variable Costs

- A tool and die maker is considering increasing the size of his business and needs to know more about the cost of electricity to operate his machines.
- The daily electricity costs and the number of tools made on that day are shown in the first two columns below. DATA Xm04-15
- Determine the fixed and variable electricity costs.

| Day | X | Y | XY | X^2 | Y^2 |
|---|---|---|---|---|---|
| 1 | 7.00 | 23.80 | 166.60 | 49.00 | 566.44 |
| 2 | 3.00 | 11.89 | 35.67 | 9.00 | 141.37 |
| 3 | 2.00 | 15.98 | 31.96 | 4.00 | 255.36 |
| 4 | 5.00 | 26.11 | 130.55 | 25.00 | 681.73 |
| 5 | 8.00 | 31.79 | 254.35 | 64.00 | 1,010.60 |
| 6 | 11.00 | 39.93 | 439.23 | 121.00 | 1,594.40 |
| 7 | 5.00 | 12.27 | 61.35 | 25.00 | 150.55 |
| 8 | 15.00 | 40.06 | 600.90 | 225.00 | 1,604.80 |
| 9 | 3.00 | 21.38 | 64.14 | 9.00 | 457.10 |
| 10 | 6.00 | 18.65 | 111.90 | 36.00 | 347.82 |
| **Total** | 65.00 | 241.86 | 1,896.62 | 567.00 | 6,810.20 |

$$s_{xy} = \frac{1}{n-1}\left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right]$$

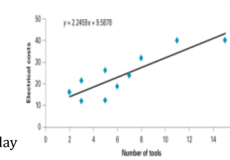
$$s_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$$

51

## Estimating Fixed and Variable Costs

We can use those sums to calculate:

$s_{xy}$ = 36.06; $s_x^2$ = 16.06; $s_y^2$ = 106.7; $\bar{x}$ = 6.5; $\bar{y}$ = 24.19




Thus, the least line square coefficients are:

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{36.06}{16.06} = \$2.25 \text{ per tool per day}$$

$$b_0 = \bar{y} - b_1\bar{x} = 24.19 - 2.25 \cdot 6.5 = \$9.57 \text{ per day}$$

**Interpretation:**

$b_1$: A one unit rise in the number of tools (x) is associated with a \$ 2.25 increase in cost (y) →**marginal rate of change** or **marginal increase**

$b_0$: When no tools are produced (i.e., x=0), the estimated fixed cost of electricity is \$9.57

52

## Measuring the Strength of the Linear Relationship

To calculate the coefficient of correlation, in addition to the covariance, we also need the standard deviations of both variables.

$$s_x = \sqrt{s_x^2} = \sqrt{16.06} = 4.01 \quad s_y = \sqrt{s_y^2} = \sqrt{106.7} = 10.33$$

The coefficient of correlation is:

$$\mathrm{r} = \frac{s_{xy}}{s_x s_y} = \frac{36.06}{4.01 \cdot 10.33} = 0.871$$

**Interpret**: The coefficient tells us the relationship is strong and positive.

53

## Coefficient of Determination

Except in the proximity to −1, 0, and +1, we cannot precisely interpret the meaning of the coefficient of correlation.

The **coefficient of determination**, which is calculated by squaring the coefficient of correlation, and denoted $r^2$, can be precisely interpreted as:
"The amount of variation in the dependent variable that is explained by the independent variable."

In Example 4.16, the coefficient of correlation was 0.871. Thus, the coefficient of determination is:

$$r^2 = (0.871)^2 = 0.759$$

About 76% of the variation in electrical costs is explained by the number of tools. The remaining 24% is unexplained.

54

## Interpreting Correlation

We learned that if two variables are linearly related, it does not mean that *X* causes *Y*. Remember:

**Correlation is not causation**

In the next section we show a few examples of **spurious correlation**, which happens when there is a large coefficient of correlation, but no reason to be a linear relationship between X and *Y* because another variable may cause both *X* and *Y*, or *Y* cause *X* instead.

55

## Examples of Spurious Correlation

**Example 1:**

*X*: Millions of pounds of uranium stored annually in the US (years 1996–2008)

*Y*: Number of Mathematics doctorates awarded annually in the United States

*r* = 0.952 Spurious Correlation 1.xlsx

**Example 2**

*X*: Per capita consumption of margarines annually in the US (years 2000–2008)

*Y*: Divorce rate in Maine (per 1,000)

*r* = 0.993 Spurious Correlation 2.xlsx

**Example 3:**

*X*: Total U.S. imports of crude oil annually (billions of barrels) (years 2000–2009)

*Y*: Per capita consumption of chicken (pounds)

*r* = 0.900 Spurious Correlation 3.xlsx

**Explanation:**

The coefficients of correlation in all three examples are large enough to infer that there is a linear relationship between each pair of variables.
However, common sense tells us that no such relationship may actually exist.

To avoid concluding that a relationship exists when there is spurious correlation, follow these steps:

1. Start with a theory of the relationship between the two variables, not with the coefficient of correlation.
2. If a relationship may justifiably exist, a large coefficient of correlation would confirm our belief

56

## Covariance and Correlation Coefficient

- Example: Golfing Study

A golfer is interested in investigating the relationship, if any, between driving distance and 18-hole score.

| Average Driving Distance (yds.) | Average 18-Hole Score |
|---|---|
| 277.6 | 69 |
| 259.5 | 71 |
| 269.1 | 70 |
| 267.0 | 70 |
| 255.6 | 71 |
| 272.9 | 69 |

57

## Covariance and Correlation Coefficient

- **Example: Golfing Study**

| | $x$ | $y$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|---|
| | 277.6 | 69 | 10.65 | -1.0 | -10.65 |
| | 259.5 | 71 | -7.45 | 1.0 | -7.45 |
| | 269.1 | 70 | 2.15 | 0 | 0 |
| | 267.0 | 70 | 0.05 | 0 | 0 |
| | 255.6 | 71 | -11.35 | 1.0 | -11.35 |
| | 272.9 | 69 | 5.95 | -1.0 | -5.95 |
| Average | 267.0 | 70.0 | | Total | -35.40 |
| Std. Dev. | 8.2192 | .8944 | | | |

58

## Covariance and Correlation Coefficient

- Example: Golfing Study
  - Sample Covariance

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-35.40}{6-1} = -7.08$$

  - Sample Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-7.08}{(8.2192)(.8944)} = -.9631$$

59

The first step in describing and summarizing data is to apply a graphical technique to learn the shape of the distribution.

The shape of the distribution helps answering the following questions.

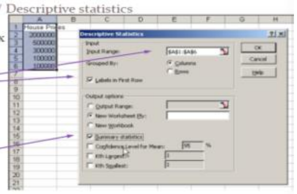
1. Where is the approximate center of the distribution?
2. Are the observations close to one another, or are they widely dispersed?
3. Is the distribution unimodal, bimodal, or multimodal? If there is more than one mode, where are the peaks, and where are the valleys?
4. Is the distribution symmetric or skewed? If symmetric, is it bell shaped?



60

## Using Microsoft Excel

- Simple Descriptive Statistics can be obtained from Microsoft® Excel
  - Use menu choice:
    Data Tab / Data analysis / Descriptive statistics
  - Enter details in dialog box



- Enter dialog box details
- Check box for summary statistics
- Click OK

61

## Stata Output and Instructions

Instructions for Descriptive Statistics:

1. Import or type the data into one column. (Click File/Import/Excel spreadsheet (*xls,*xlsx)/Xm03-01 Check **Import first row as variable names**.
2. Click **Statistics**, **Summary, tables, and tests**, **Summary and descriptive statistics**, and **Summary statistics**.
3. In the **Variables** (leave empty for all variables) select **Ages**. Under **Options** select **Display additional statistics**.
4. STATA prints a variety of statistics. Mean = 53.765; variance = 390.5224; standard deviation = 19.76164; first quartile (25th percentile) = 36, second quartile (50th percentile, median) = 54.5; third quartile (75th percentile) = 69.

Instructions for Least Square Line:

1. Import or type the data into one column. (Click File/Import/Excel spreadsheet (*xls,*xlsx)/Xm04-15 ) Check **Import first row as variable names**.
2. Click **Statistics**, **Linear models and related**, and **Linear regression**.
3. Select **Electrical costs** as the **Dependent variable**: and in the **Independent variables**: box select **Number of tools.**

The least squares line is: $\hat{y} = 9.587765 + 2.245882 \times \text{Number of Tools}$

**Covariance and Coefficient of Correlation**

Coefficient of correlation = 0.8711, covariance = 36.0589

62

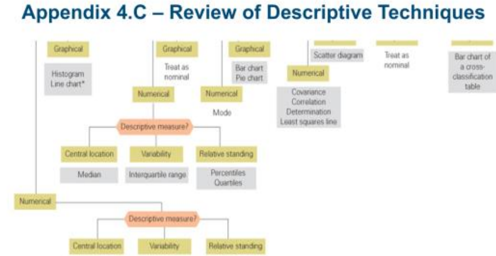
## Stata Output and Instructions

Instructions for Covariance and Coefficient of Correlation:

1. Import or type the data into one column. (Click File/Import/Excel spreadsheet (*xls,*xlsx)/Xm03-01 Check **Import first row as variable names**.
2. Click **Statistics**, **Summary, tables, and tests**, **Summary and descriptive statistics**, and **Correlations and Covariances.**
3. In the **Variables (leave empty for all):** box select **Number of tools**, **Electrical costs**. Click **OK**.
4. To output the covariance, repeat steps 1–3 and click **Options** and check **Display covariances**.

63

## Appendix 4.C – Review of Descriptive Techniques




CENGAGE