

# Numerical Descriptive Measures

Eco 2470: Economic Statistics

Fall, 2019. Chaoyi Chen

(Chapter 4)

# Numerical Descriptive Techniques...

Description of Distribution/Measures of Relative Standing

Percentiles, Quartiles, Median

Measures of Central Location

Mean, Median, Mode

Measures of Variability

Range, Interquartile Range, Standard Deviation, Variance,  
Coefficient of Variation

Measures of Linear Relationship

Covariance, Correlation, Coefficient of Determination

# Why numerical statistics

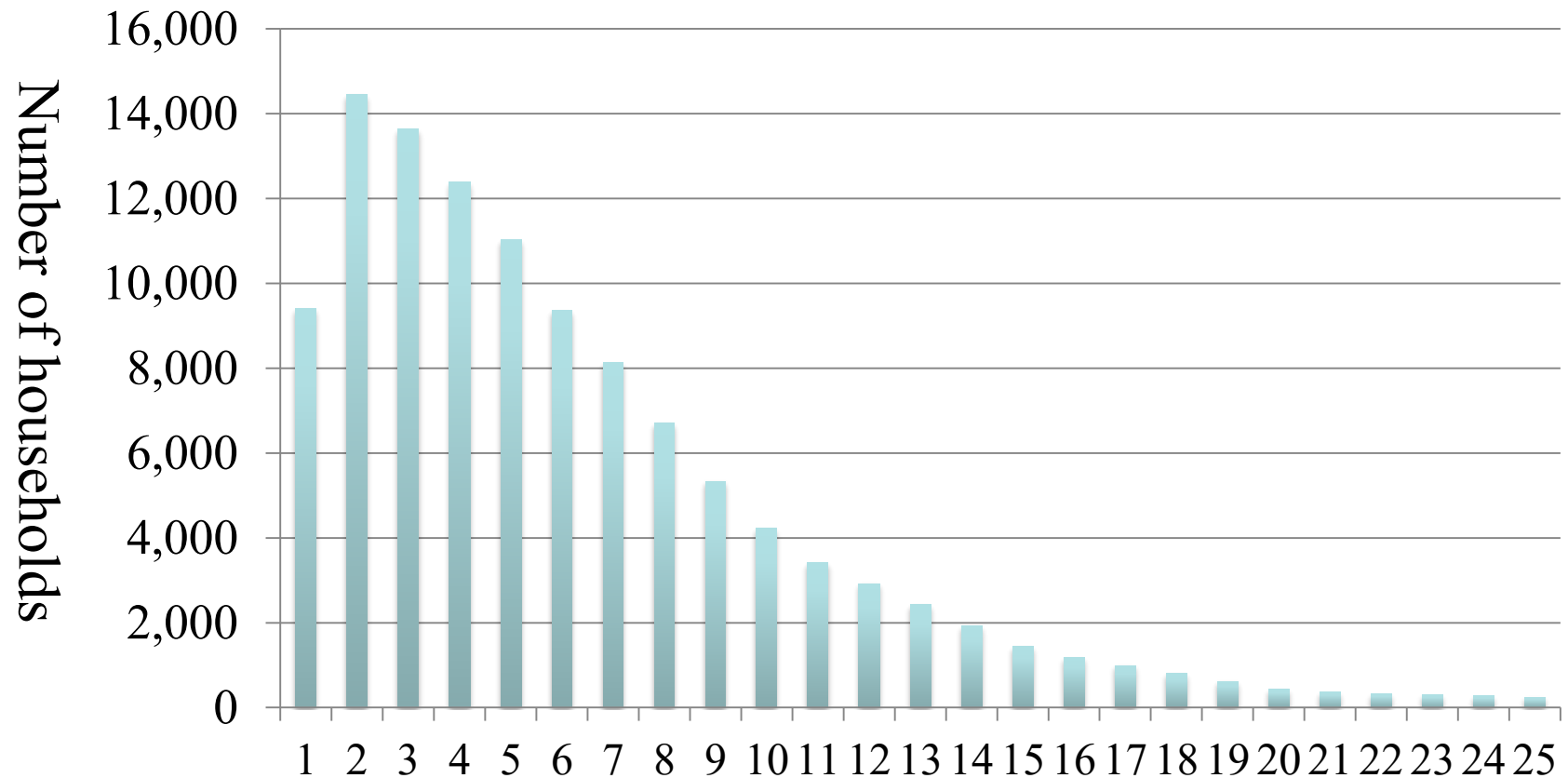
---

- Graphical Measures are great for visualizing the data. And we can learn a lot from them.
- But, they are also inherently imprecise.
- And they can be easily manipulated
- Therefore we complement them with numerical measures
- Usually a good analysis includes **both** graphical and numerical methods.

# Recall: histogram lets us view the distribution:

U.S household income (below 250K)

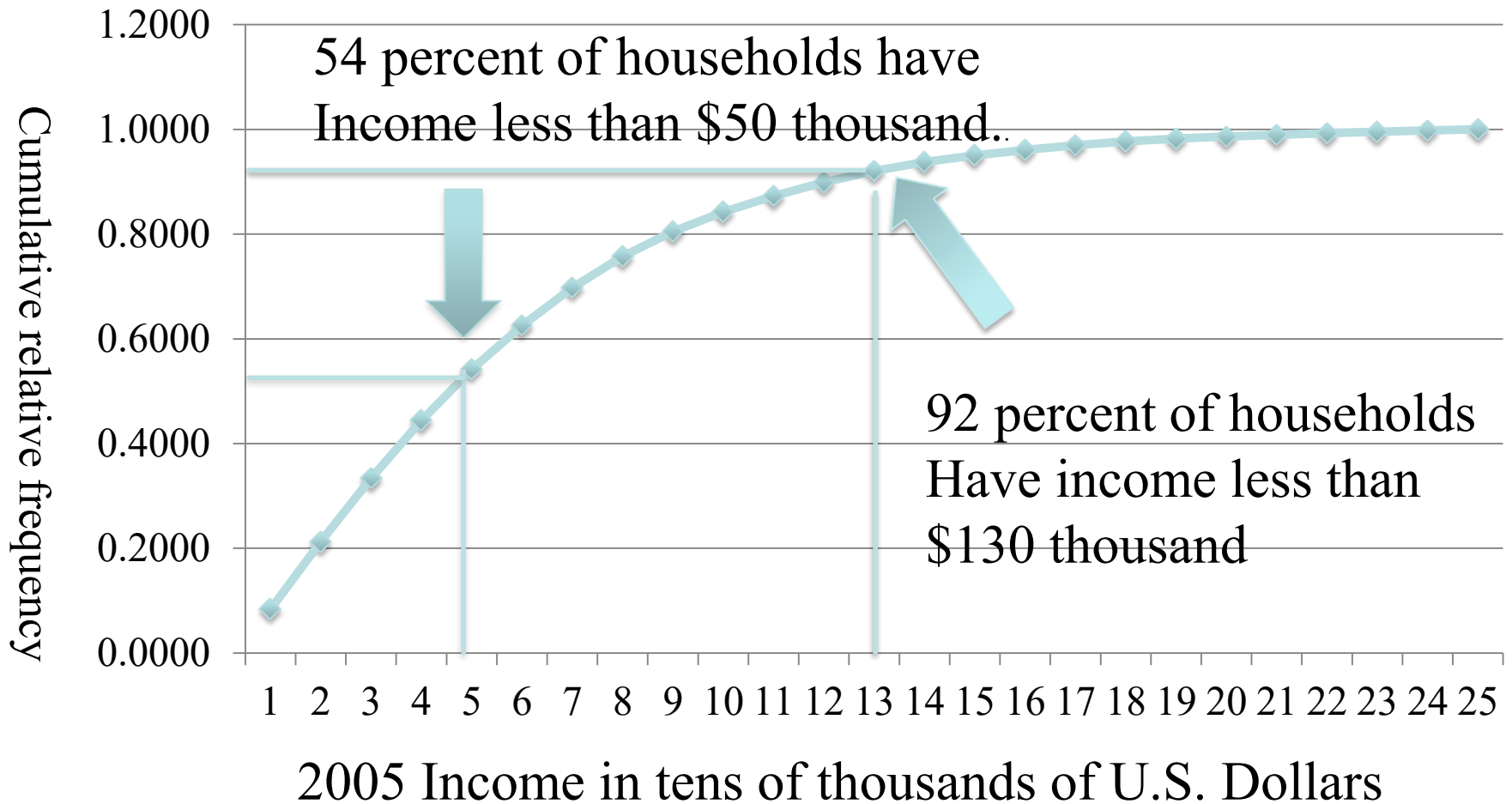
frequency (no. of households)



2005 Income in tens of thousands of U.S. Dollars

The Ogive gives us a second view from which we can read off the percentage of households under a given income level

Ogive: cumulative relative frequency

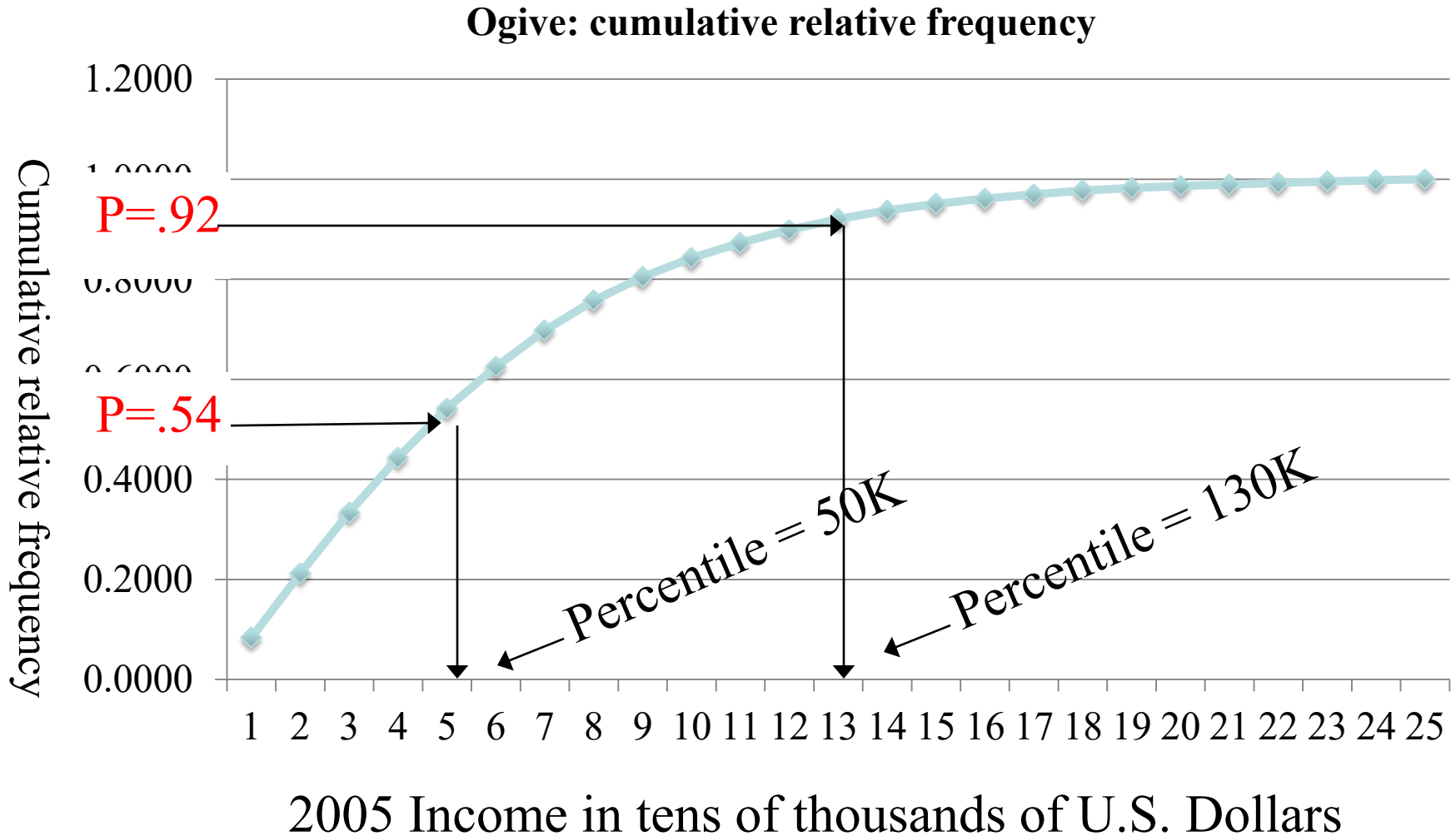


# Percentiles

---

- This is what we just read off of the Ogive (in reverse).
- Notation:
  - P – A percent (between 0% and 100%)
  - P<sup>th</sup> Percentile. The value for which P% of the observations are less than that value and (100-P)% are greater than that value.
  - L<sub>p</sub> – The location (or rank order) of the P<sup>th</sup> Percentile when ranked from low to high.

# Findings percentiles by reading Ogive in reverse



# Calculating percentiles

---

- P – Given to you (usually)
- n – Number of observations (count them)
- $L_p = (n+1) P/100$  (Formula for location of percentile)
- $P_{th}$  Percentile: Obtain as follows:
  1. Re-order data from smallest to largest value
  2. Put a check next to the first  $L_p$  values and Stop
  3. The value next to you last check mark is your  $P_{th}$  Percentile



# Calculation Example: Find 50<sup>th</sup> percentile

Y-Variable original	Y-Variable re-ordered low to high	Rank order (low to high) of Y-Variable
12	0	1
14	5	2
5	7	3
0	8	4
22	9	5
9	12	6
33	14	7
7	22	8
8	33	9

$$P = 50\% \text{ (given)}$$

$$n = 9 \text{ observations (count)}$$

$$L_p = (n+1) P/100$$

$$= (9+1)50/100 = 10/2 = 5$$

$$\left. \begin{array}{l} \\ \\ \\ \\ \end{array} \right\} L_p = 5$$

$$P_{th} \text{ Percentile} = 9$$

# Calculation 2: Find 25<sup>th</sup> percentile

Y-Variable original	Y-Variable re-ordered low to high	Rank order (low to high) of Y-Variable
12	0	1
14	5	2
5	7	3
0	8	4
22	9	5
9	12	6
33	14	7
7	22	8
8	33	9

$$P = 25\% \text{ (given)}$$

$$n = 9 \text{ observations (count)}$$

$$L_p = (n+1) P/100$$

$$= (9+1)25/100 = 10/4 = 2.5$$

}  $L_p$  halfway between 2 & 3

Percentile between 5 & 7

$$P_{th} \text{ Percentile} = 6$$

# Percentiles as 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.

If you're exam mark is in the 40<sup>th</sup> percentile, it means that you did better than 40% of the students, but worse than 60%.

Note: The 40<sup>th</sup> percentile does not mean that you obtained a 40% mark. If that were the case then over 40% of the class would have a failing mark!

# Graduate Management Admission Test (GMAT)

You are one of 9,999 students take the GMAT. You score a 550 (out of 800) placing in the 60<sup>th</sup> percentile. Find your P, L<sub>p</sub>, and P<sup>th</sup> Percentile and interpret them:

- P = 60 (the percentage): 60 Percent of students scored below you. 40 percent scored higher.
- P<sup>th</sup> Percentile = 550 (your score)
- $L_p = (n+1)*P/100 = (9,999+1)*60/100 = 10,000*0.6$   
= 6,000 (your rank order sorting low to high):  
5,999 students scored below you. 3,999 scored higher.

# Percentiles & Risk Management: Value at Risk (VaR)

- In 2008 Lehman Brothers declared bankruptcy and other financial giants, such as AIG, went on life support with government bailouts.
- Canada's financial institutions fared relatively better and the Canadian economy suffered less as a result.
- The 2008 crisis underscores the importance of proper risk management.
- One of the primary tools used by risk managers is based on Percentile – it is called Value at Risk (VaR)

# Value at Risk (continued)

---

- A bank's VaR is usually defined as the 5<sup>th</sup> Percentile of the return on its capital.
- The 5<sup>th</sup> percentile is designed to capture a “bad case” scenario, such as a financial crisis.
- But it is not a doomsday scenario
- Think of it the worst yearly return for the bank in 20 years.
- This is usually a big loss.
- Regulators require banks to ensure that their VaR is not so large that they will require a bailout.
- When calculating their VaR many U.S. banks apparently assumed that the housing market wouldn't fall ...

# Special Percentiles:

---

- Median      Median = 50<sup>th</sup> percentile.
  
- Quartiles: 1<sup>st</sup> Quartile = 25<sup>th</sup> Percentile  
2<sup>nd</sup> Quartile = 50<sup>th</sup> Percentile = Median  
3<sup>rd</sup> Quartile = 75<sup>th</sup> Percentile
- Interquartile Range = 3<sup>rd</sup> Quartile – 1<sup>st</sup> Quartile
  
- Deciles: 1<sup>st</sup> Decile = 10<sup>th</sup> Percentile  
2<sup>nd</sup> Decile = 20<sup>th</sup> Percentile  
....  
9<sup>th</sup> Decile = 90<sup>th</sup> Percentile

# Median as a Measure of Central Location

- The Median is the middle value: It is always the case that half of the values lie above the median and half lie below the median.
- Median exam score gives a gage of how well students did (or how hard the exam was). It tells the score of the middle student. Half of the class did better, half did less well.



# Another Measure of Central Location...

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

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

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

# Comparison of Mean and Median

- Advantages of Mean

- Often what we want to estimate:

- E.g. in financial models investors care about mean returns, not median returns

- A casino's profits depends on the mean winnings (total winnings/no. of gamblers), not the median winnings.

- Advantages of Median

- Less sensitive to extreme values

- Can be applied to ordinal data, whereas the mean cannot. Why? Because this calculation only involves the ordering of the data.

- Note** neither the mean nor median can be meaningfully applied to nominal data, since they have no numeric meaning

# Mean, Median, & Extreme values

Example: High earning CEO moves into your neighborhood.

Neighborhood Incomes  
(in 1,000 of \$)

Pre-CEO	Post-CEO
20	20
30	30
40	40
50	50
60	60
	10,000

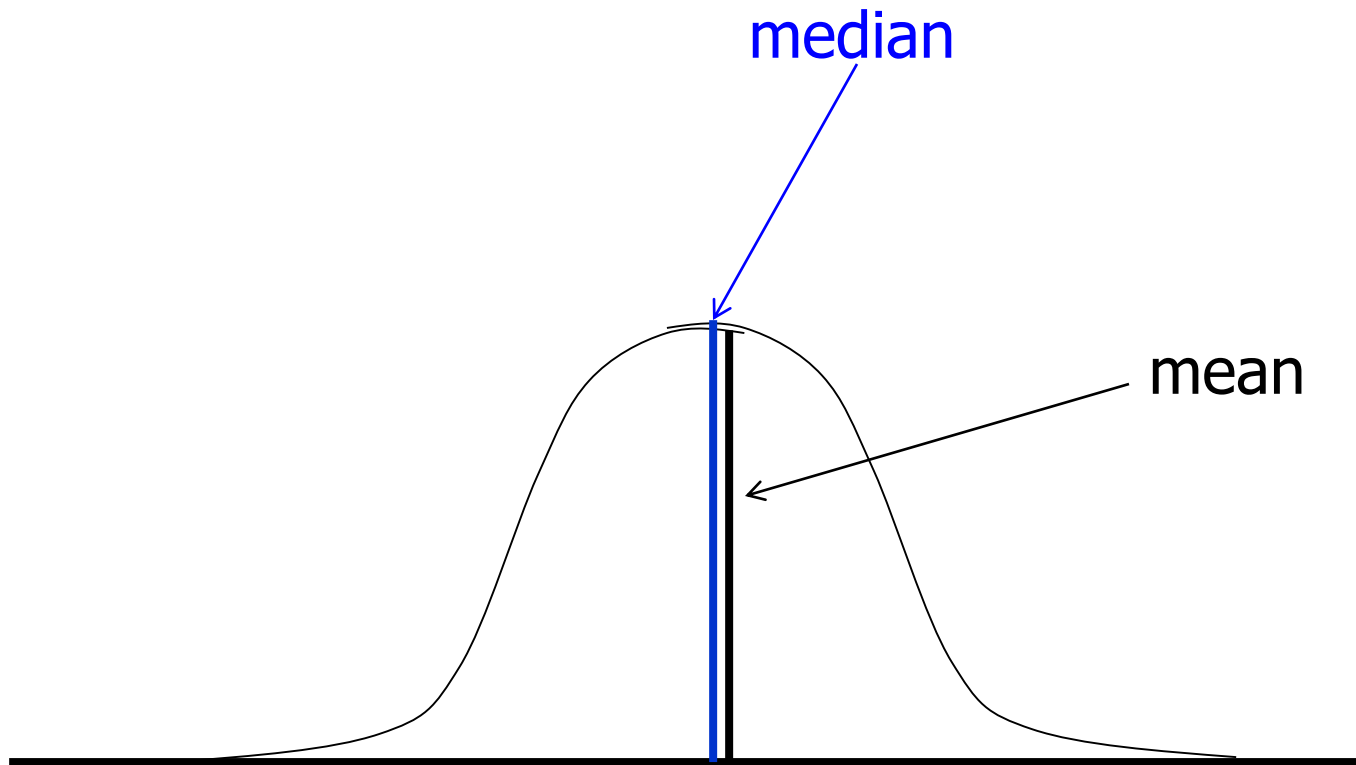
	Pre-CEO	Post-CEO
Sum Total	200	10,200
n	5	6
Mean	40	1,700
Median	40	45

**Conclusion:** Mean is sensitive to outliers (extreme value),  
Median is not

# Mean and Median

---

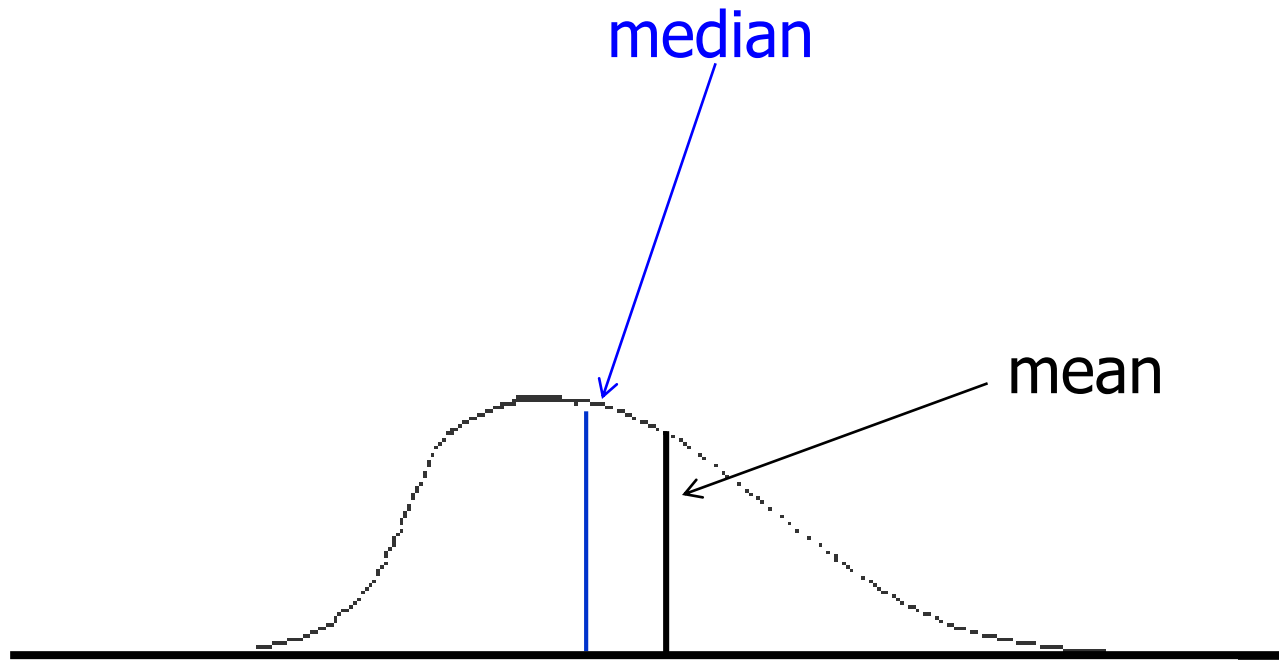
If a distribution is symmetrical the mean and median may coincide...



# Mean and Median

---

If a distribution is asymmetrical, say skewed to the left or to the right, the two measures may differ. E.g.:



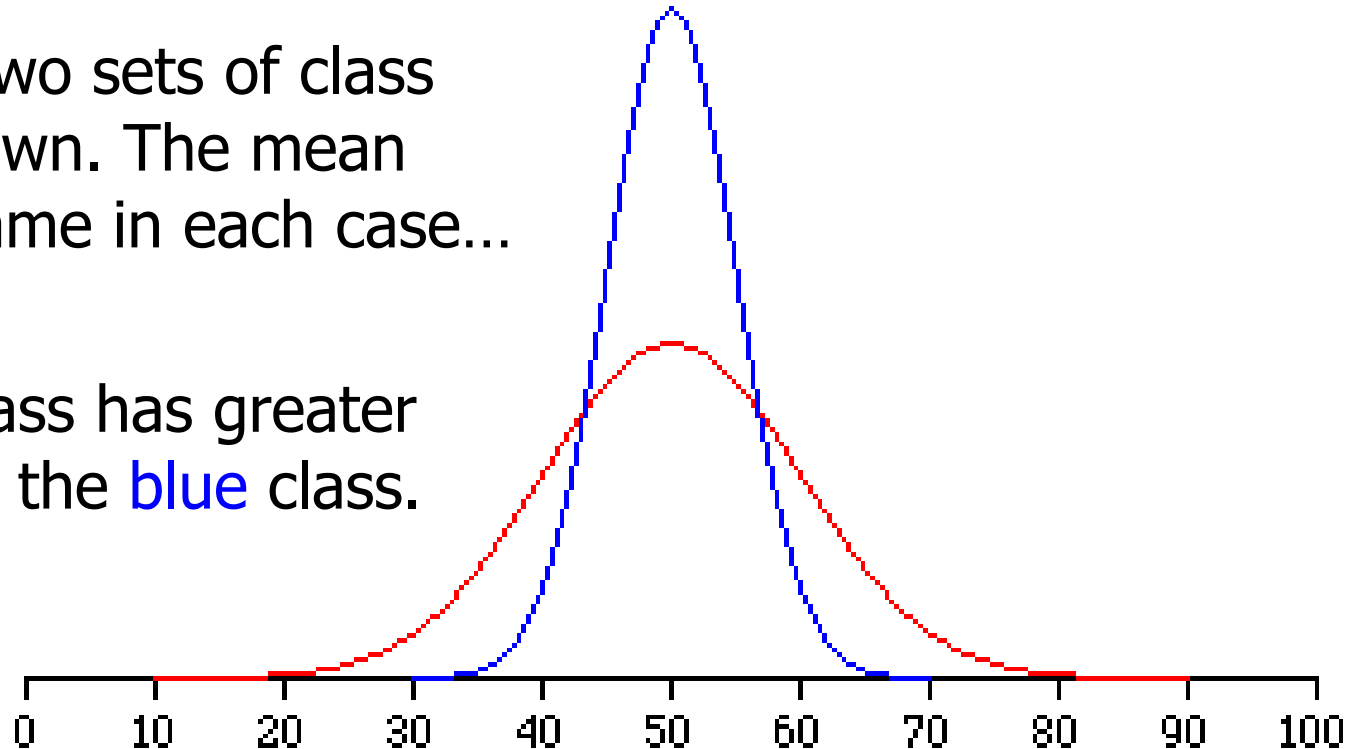
# Measures of Variability...

---

Measures of central location fail to tell the whole story about the distribution; that is, how much are the observations spread out around the mean value?

For example, two sets of class grades are shown. The mean (=50) is the same in each case...

But, the **red** class has greater variability than the **blue** class.



# Why we care about variability: Examples

- Investing: An investment with a highly variable investment return may be considered risky.
- Throwing Darts: If your accuracy is highly variable you may obtain a low score, even if the average (central) location of your darts is the bulls eye.
- Laser Eye Surgery: If the accuracy of the laser is variable, then you're in trouble.

# Range and Interquartile Range

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

Range = Largest observation – Smallest observation

Data: {4, 8, 15, 24, 39, 50}      Range = 50-4 = 46

Shortcoming: Only takes into account the two most extreme observations.

Interquartile Range = 3<sup>rd</sup> Quartile – 1<sup>st</sup> Quartile (Recall)

Still not a complete description of variability.



# Sample Variance: Intuition


---

- Data Set w/ Small variation: Most values “close” to the mean
- Data Set w/ Large variation: Many values “far” from mean
- How do we measure distance of an observation from the mean?
  - $x_i - \bar{x}$  is the distance if  $x_i > \bar{x}$
  - $\bar{x} - x_i$  is the distance if  $x_i < \bar{x}$
  - Which starts to get confusing ...
  - But,  $(x_i - \bar{x})^2$  is always the squared distance from the mean
- **Sample Variance = Average Squared Distance from the Mean** (with minor degrees of freedom adjustment.)

# Sample Variance Formula

---

- Sample Variance = Average Squared Distance from Mean w/ degrees of freedom adjustment.

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


Note: The denominator is sample size (n) minus one !  
This minus one is the degrees of freedom adjustment

# What is a Degree of Freedom & How is it Lost?

---

- To keep it simple, suppose our data set has just two observations:  $x_1$  and  $x_2$
- Together  $x_1$  and  $x_2$  have two degrees of freedom, because they are both “free” to take on whatever value they want.
- The sample mean is  $\bar{x} = \frac{1}{2}(x_1 + x_2)$   
so  $2\bar{x} = (x_1 + x_2)$
- Now consider  $x_1 - \bar{x}$   
and  $\frac{x_2 - \bar{x}}{1}$
- The add to:  $(x_1 + x_2) - 2\bar{x} = 2\bar{x} - 2\bar{x} = 0$
- So  $x_2 - \bar{x} = -(x_1 - \bar{x})$  is **not** “free” to take any value it wants. A degree of freedom has been lost.

# Sample Variance...

---

As you can see, you have to calculate the sample mean ( $\bar{x}$ ) in order to calculate the sample variance.

Alternatively, there is a short-cut formulation to calculate sample variance directly from the data without the intermediate step of calculating the mean. Its given by:

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

# Application...


---

Example 4.7. The following **sample** consists of the number of jobs six students applied for: 17, 15, 23, 7, 9, 13.

Finds its **mean** and **variance**.



$\bar{x}$



$s^2$

# Sample Mean & Variance...

---

## Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{17 + 15 + 23 + 7 + 9 + 13}{6} = \frac{84}{6} = 14 \text{ jobs}$$

## Sample Variance

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{1}{6 - 1} \left[ (17 - 14)^2 + (15 - 14)^2 + \dots + (13 - 14)^2 \right] = 33.2$$

## Sample Variance (shortcut method)

$$s^2 = \frac{1}{n - 1} \left[ \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right] = \frac{1}{6 - 1} \left[ (17^2 + 15^2 + \dots + 13^2) - \frac{(17 + 15 + \dots + 13)^2}{6} \right] = 33.2$$

# Sample Standard Deviation...

- To obtain sample variance we squared distance of each observation from the mean
- Now we “undue” the squaring by taking the square root.
- The standard deviation is simply the square root of the variance, thus:

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

# Application of Standard Deviation...

Consider Example 4.8 [[Xm04-08](#)] where a golf club manufacturer has designed a new club and wants to determine if it is hit more consistently (i.e. with less variability) than with an old club.

The following tables were produced for interpretation...

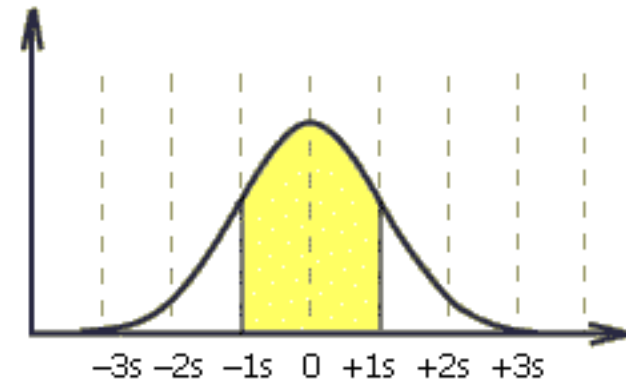
<i>Current 7-iron</i>		<i>New 7-iron</i>	
Mean	150.55	Mean	150.15
Standard Error	0.67	Standard Error	0.36
Median	151	Median	150
Mode	150	Mode	149
<b>Standard Deviation</b>	<b>5.79</b>	<b>Standard Deviation</b>	<b>3.09</b>
Sample Variance	33.55	Sample Variance	9.56
Kurtosis	0.13	Kurtosis	-0.89
Skewness	-0.43	Skewness	0.18
Range	28	Range	12
Minimum	134	Minimum	144
Maximum	162	Maximum	156
Sum	11291	Sum	11261
Count	75	Count	75

You get more consistent distance with the new club.

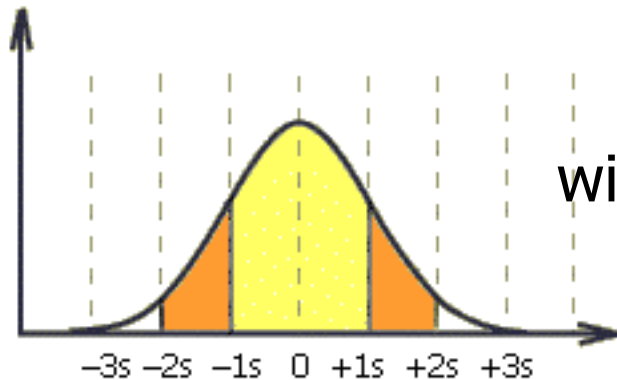


# The Empirical Rule for Bell Shaped Histograms

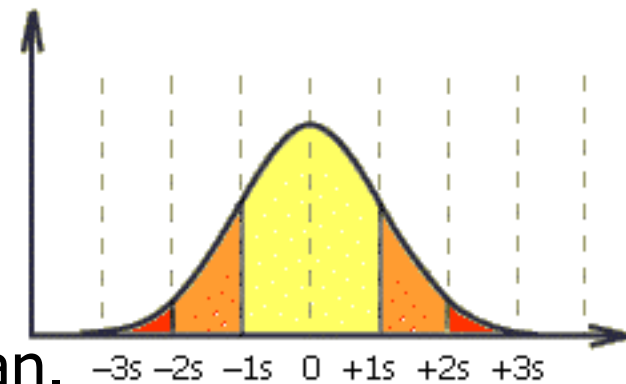
Approximately 68% of all observations fall within **one** standard deviation of the mean.



Approximately 95% of all observations fall within **two** standard deviations of the mean.



Approximately 99.7% of all observations fall within **three** standard deviations of the mean.



# Chebysheff' s Theorem

---

A more general interpretation of the standard deviation is derived from *Chebysheff' s Theorem*, which applies to all shapes of histograms (not just bell shaped).

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

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

For  $k=2$  (say), the theorem states that *at least* 3/4 of all observations lie within 2 standard deviations of the mean. This is a “lower bound” compared to Empirical Rule' s approximation (95%).

# Application: Interpreting Standard Deviation

Suppose that the mean and standard deviation of last year's midterm test marks were 70 and 5, respectively. How many marks fell between 60 and 80?

1. Calculate distance from mean ( $\bar{x} = 70$ ):

$$\bar{x} - 60 = 70 - 60 = 10 \quad (60 \text{ is ten marks below the mean})$$

$$80 - \bar{x} = 80 - 70 = 10 \quad (80 \text{ is ten marks above the mean})$$

2. Divide distance by the standard deviation ( $s = 5$ ):

$$k = (\bar{x} - 60) / s = 10 / 5 = 2 \quad (60 \text{ is 2 standard deviations below the mean})$$

$$k = (80 - \bar{x}) / s = 10 / 5 = 2 \quad (80 \text{ is 2 standard deviations above the mean})$$

3. Rephrase Question: How many marks lie within  $k = 2$  standard deviations of the mean.

# Application Continued

---

How many marks lie within  $k = 2$  standard deviations of the mean?

4. If the histogram is bell shaped: Use the Empirical Rule:  
Approximately 95 percent of marks lie within 2 standard deviations of the mean. **Answer: Approximately 95%**

5. If the histogram is **not** bell shaped: Use Chebyshev's bound: At least

$$1 - 1/k^2 = 1 - 1/2^2 = 1 - 1/4 = 3/4 = 75\%$$

of marks fall within 2 standard deviations of the mean.

**Answer: At least 75%**

# Coefficient of Variation...

---

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

$$\text{Sample coefficient of variation} = cv = \frac{s}{\bar{x}}$$

This coefficient provides a *proportionate & unit free* measure of variation, e.g.

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.

# (Linear) Relationships: Why are we interested?

- We may suspect that one variable help explains another
  - E.g. batting averages might help explain baseball salaries
  - E.g. Interest rates might help explain housing prices
  - E.g. Oil prices might help explain strength of Canadian dollar
- One variable might help predict another
  - E.g. Debate over whether earnings/price ratios help predict future stock returns.
  - E.g. Leading indicators (stock market, consumer expectations, building permits, and the money supply) may help predict business cycles.

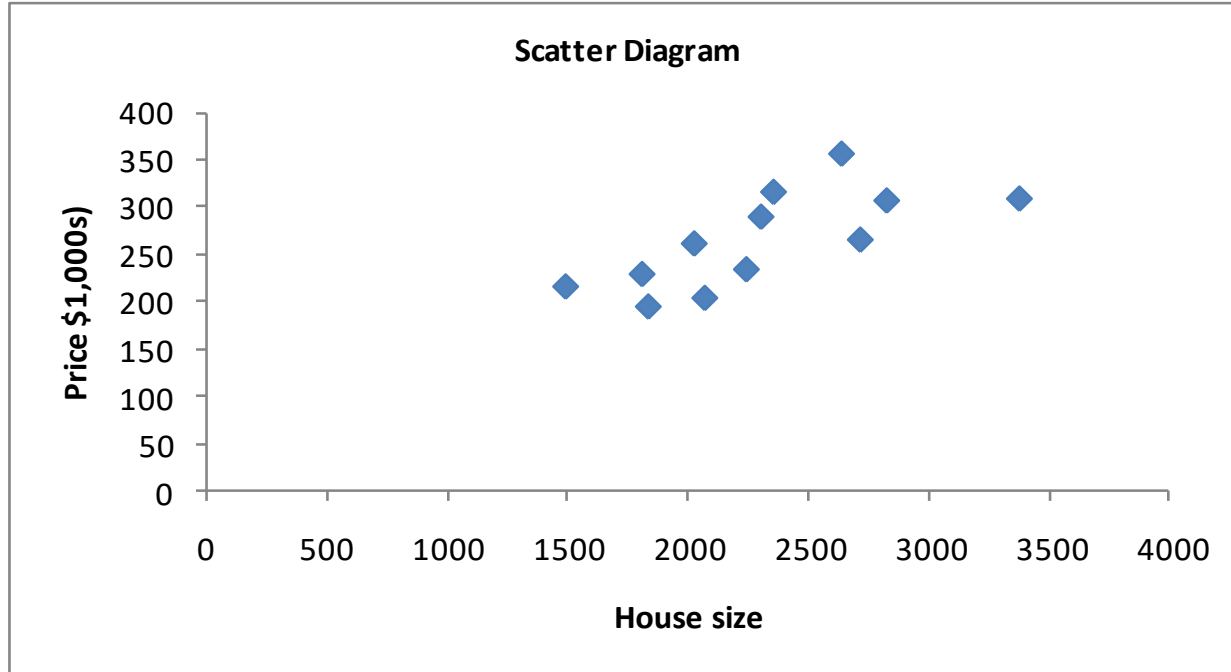
## **Caution:** Association does not imply cause and effect

---

- Often, our ultimate goal is to try explain  $Y$  with  $X$
- However, the descriptive statistics described below can only show us that  $Y$  and  $X$  are associated or related.
- They do not distinguish between direct causation, indirect causation, and reverse causation.
- In other words, if we see a strong association between  $X$  and  $Y$ : it could be that  $X$  causes  $Y$  or that  $Y$  causes  $X$  or that some  $Z$  causes both, etc.
- Logic, common sense, and economics are often needed to help you sort out which are the likeliest directions of causation.
- So, be careful not to jump to conclusions.

# Measure of Linear Relationship

- Recall Graphical Descriptive Statistics
  - Cross classification tables (nominal and ordinal data)
  - Scatter plots (interval data)



- Complement scatter plot with numerical descriptive measures



# Sample Covariance: Intuition and Formula

- What do we mean by a positive relationship between  $x$  and  $y$ ?
- Does  $y$  tend to be above its mean ( $y - \bar{y} > 0$ ) when  $x$  is above its mean ( $x - \bar{x} > 0$ )?
- When this happens the following product is positive:

$$(x - \bar{x})(y - \bar{y}) = (+)(+) > 0$$

- The **Sample Covariance** is the average of these products (adjusted for degrees of freedom)

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

sample mean of X and Y

Note: divisor is  $n-1$ , due to degrees of freedom adjustment.

# Sample Covariance: Shortcut Formula

In much the same way there was a “shortcut” for calculating sample variance without having to calculate the sample mean, there is also a shortcut for calculating sample covariance without having to first calculate the means:

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

# Covariance Illustrated...

Consider the following three sets of data (textbook § 4.5)...

	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	covariance
Set #1	2	13	-3	-7	21	$S_{xy} = 17.5$
	6	20	1	0	0	
	7	27	2	7	14	
Set #2	2	27	-3	7	-21	$S_{xy} = -17.5$
	6	20	1	0	0	
	7	13	2	-7	-14	
Set #3	2	20	-3	0	0	$S_{xy} = -3.5$
	6	27	1	7	7	
	7	13	2	-7	-14	

For each set:  $\bar{X} = 5$        $\bar{Y} = 20$

In each set, the values of X are the same, and the value for Y are the same; the only thing that's changed is the order of the Y's.

In set #1, as X increases so does Y;  $S_{xy}$  is large & positive

In set #2, as X increases, Y decreases;  $S_{xy}$  is large & negative

In set #3, as X increases, Y doesn't move in any particular way;  $S_{xy}$  is "small"

# Covariance... (Generally speaking)

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

NOTE: It is extremely rare that two variables always move in the same direction. Positive covariance is a just a tendency to move together.

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

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

However, it is often difficult to determine whether a particular covariance is large or small.

# Sample Coefficient of Correlation...

The coefficient of correlation is defined as the covariance divided by the standard deviations of the variables:

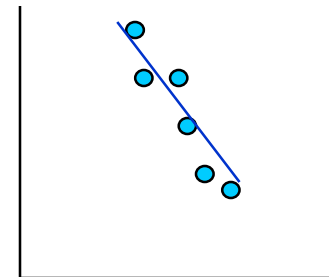
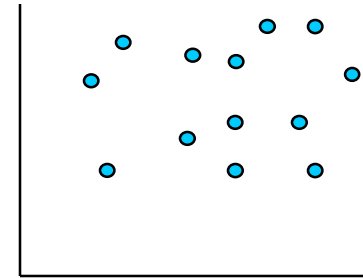
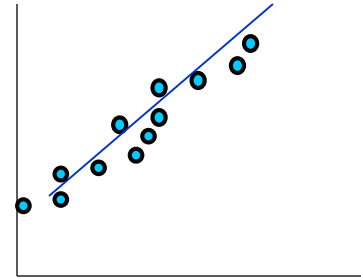
$$\text{Sample coefficient of correlation: } r = \frac{s_{xy}}{s_x s_y}$$

$$\text{Property of correlation: } -1 \leq r \leq 1$$

This coefficient answers the question:  
How **strong** is the association between X and Y?

# Sample Coefficient of Correlation...

$\rho$  or  $r =$   $\left\{ \begin{array}{l} +1 \text{ Strong positive linear relationship} \\ 0 \text{ No linear relationship} \\ -1 \text{ Strong negative linear relationship} \end{array} \right.$



# Coefficient of Determination

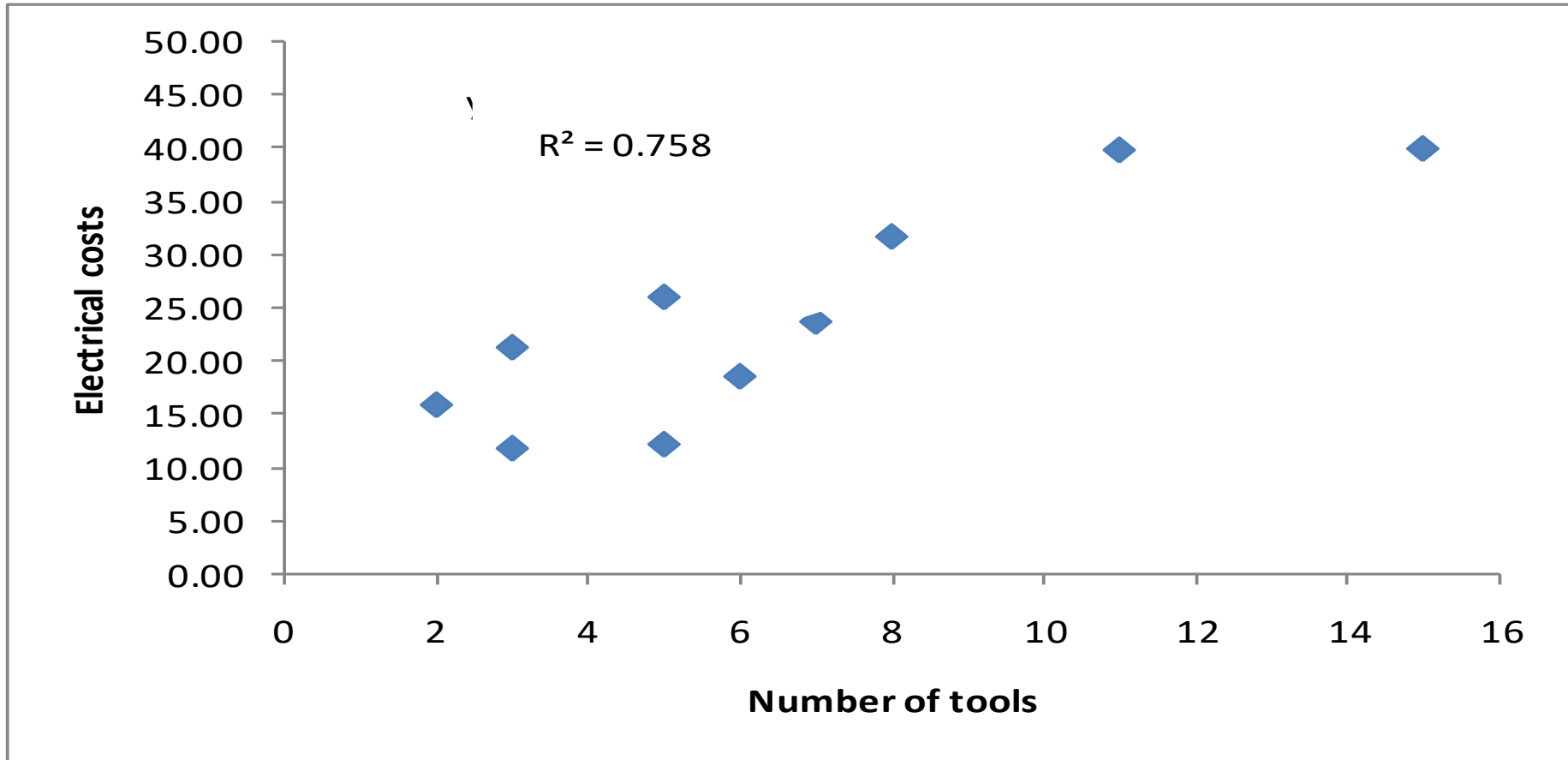
---

- Except for  $r = -1, 0,$  and  $+1$  we cannot precisely interpret the coefficient of correlation. We can judge it in relation to its proximity to  $-1, 0,$  and  $+1$  only.
- Fortunately, we have another measure that can be precisely interpreted. It is the *coefficient of determination*,
- It is calculated by squaring the coefficient of correlation. For this reason we denote it  $R^2$  . ( $R^2 = r^2$ )
- The coefficient of determination measures the amount of variation in the dependent variable that is explained by (or associated with) the variation in the independent variable.

# Example 4.18

---

Calculate the coefficient of determination for Example 4.17





# Example 4.18

---

The coefficient of determination is

$$R^2 = .758$$

This tells us that 75.8% of the variation in electrical costs is explained by (or associated) the number of tools. The remaining 24.2% is unexplained.

**Note on Causation and  $R^2$ :** Traditionally we use the terminology “explained by” for the coefficient of determination. This is appropriate when  $X$  causes/explains  $Y$ . I add (or “associated with”) as a reminder that the association between  $X$  and  $Y$  measured by  $R^2$  is not always due to the fact that  $X$  causes/explains  $Y$ .

# Chapter 4 Topics Skipped & Deferred

Major Topics Skipped (You are not responsible for them):

1. Box Plots

Major Topics Deferred Until Later (you are not responsible for them yet. You will become responsible for them later after we cover them)

1. Population versions of mean, variance, covariance, etc.
2. Regression and line of best fit.