MGMT 524

Management Science

Statistics & Spreadsheet Review

Dan Brandon, Ph.D., PMP

Copyright Dan Brandon, PhD, PMP

What is statistics?

- You may think of ‘statistics’ in the narrow sense as ‘numerical facts’:
  - The population of the US
  - The average household income
  - The average price of a stock
- But in a broader sense, statistics is the art and science of collecting, analyzing, presenting, and interpreting data

Descriptive and Predictive Statistics

- **Descriptive statistics** quantitatively describe the main features of a collection of data
- Descriptive statistics are distinguished from inferential statistics (or **predictive or inductive** statistics), in that descriptive statistics aim to summarize a data set, rather than use the data to learn about the population that the data are thought to represent
- This generally means that descriptive statistics, unlike inferential statistics, are not developed on the basis of probability theory

Copyright Dan Brandon, PhD, PMP

Data, Data Sets, Elements, Variables, and Observations

Column names identify the variables that were captured about each element.

Each company name identifies an element.

Company	Stock Exchange	Annual Sales($M)	Earn/ Share($)
Dataram	AMEX	73.10	0.86
EnergySouth	OTC	74.00	1.67
Keystone	NYSE	365.70	0.86
LandCare	NYSE	111.40	0.33
Psychemedics	AMEX	17.60	0.13

Each row records an observation.

The whole thing is called the data set.

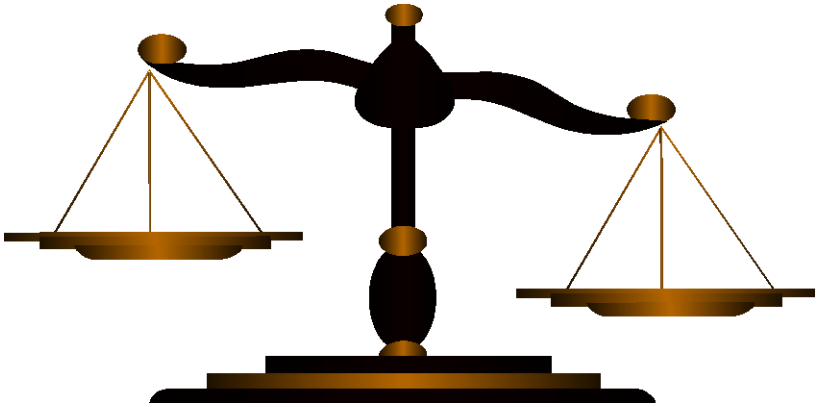
Scales of Measurement

- Each variable is measured using one of these scales of measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio

- The scale *determines the amount of information contained in the data* - and determines the statistical analysis techniques that can be used with this type of data

Types of Scales

- **Nominal**
- **Ordinal**
- **Interval**
- **Ratio**



There are four types of measurement scales as shown here. One needs to know the difference in order to choose the appropriate type and then use the appropriate analysis for that type. Each scale type has its own characteristics.

Levels of Measurement

Nominal

Classification

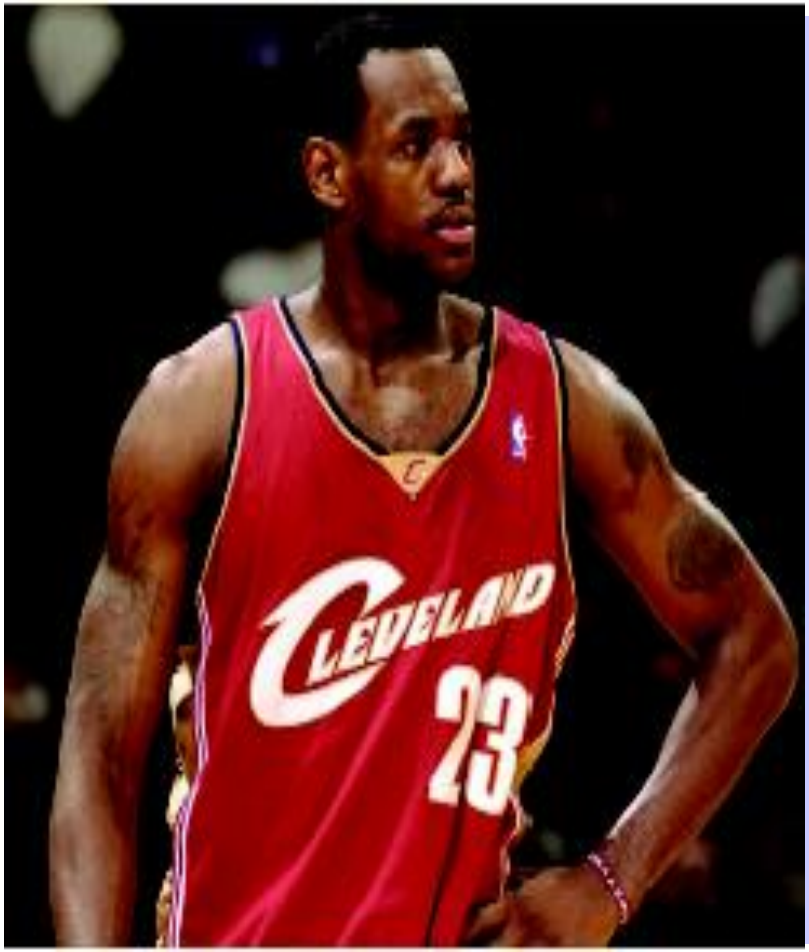
Ordinal

Interval

Ratio

Nominal (or classification) scales provide only classification of the person, event, or property. Classification means that numbers are used to group or sort responses.

Nominal Scales



- Mutually exclusive and collectively exhaustive categories – each entity placed into one and only one of the categories
- Examples:
  - Gender
  - Political party
  - Religion
  - Ethnicity

Nominal Scales (con’)

- For example, basketball fans could be classified by whether or not they had ever attended a professional game before
- The counting of members in each group is the only possible arithmetic operation when a nominal scale is employed
- If we use numerical symbols within our mapping rule to identify categories, these numbers are recognized as labels only and have no quantitative value
- Nominal scales are the least powerful of the four data types
- They suggest no order or distance relationship and have no arithmetic origin
- A researcher is restricted to use of the mode as a measure of central tendency.
  - The mode is the most frequently occurring value. There is no generally used measure of dispersion for nominal scales. Dispersion describes how scores cluster or scatter in a distribution.

Levels of Measurement

Nominal	Classification
Ordinal	Classification Order
Interval	
Ratio	

Order means that the numbers are ordered. One number is greater than, less than, or equal to another number.

Ordinal Scales

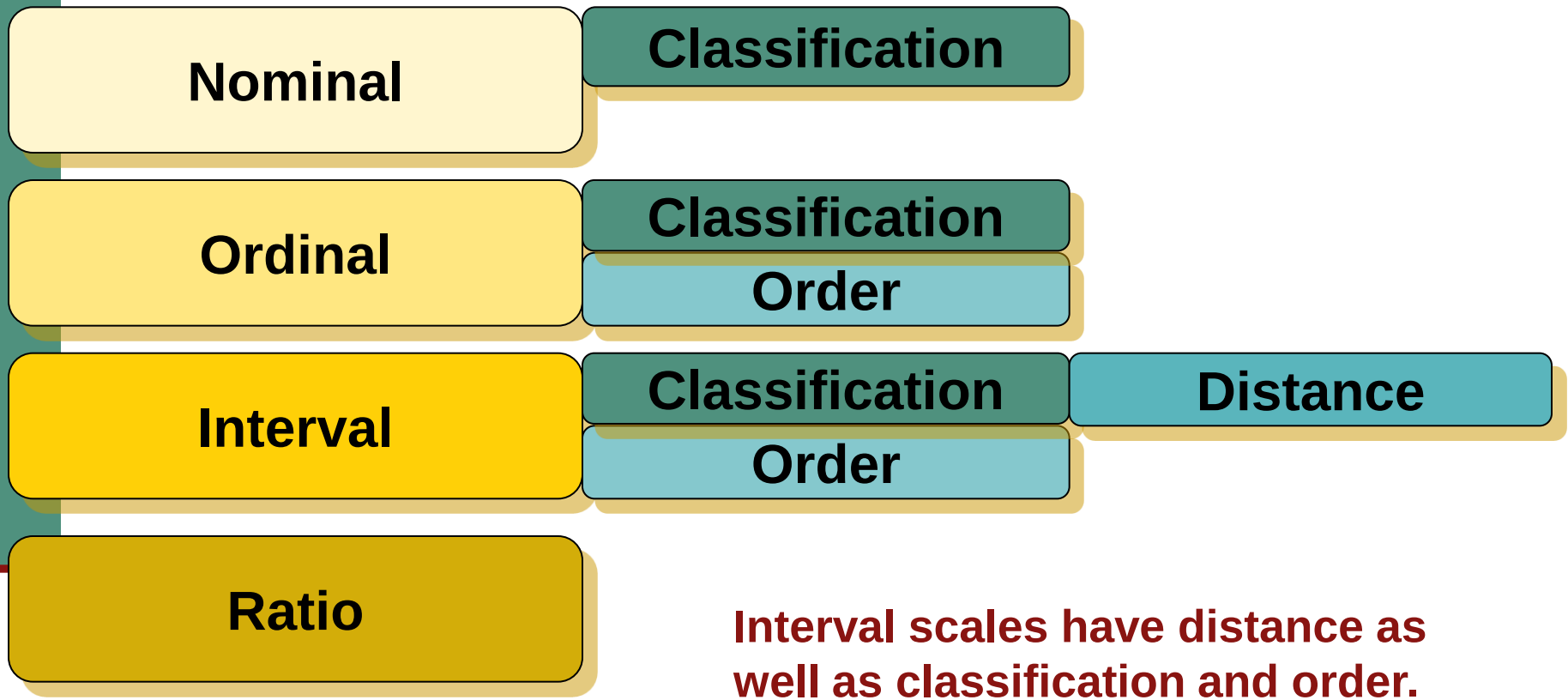
- Characteristics of nominal scale plus an indication of order
- Implies statements of greater than and less than
- Example:
  - Socioeconomic class
    - Low, middle, upper
  - Attitudes (i.e. Likert Scale):
    - Strongly favorable, favorable, neutral, unfavorable, strongly unfavorable

All material, both content and format, contained herein including the slide notes but excepting the direct textbook slides are the property of and copyright by Dan Brandon Jr. Memphis TN USA

Ordinal Scales (con't)

- Rankings are examples of ordinal scales
- Attitude and preference scales are also ordinal
- The appropriate measure of central tendency is the median - the median is the midpoint of a distribution
- A percentile or quartile reveals the dispersion
- "Nonparametric" tests should be used with nominal and ordinal data
  - This is due to their simplicity, statistical power, and lack of requirements to accept the assumptions of parametric testing

Levels of Measurement



Interval scales have distance as well as classification and order. Distance means that differences between numbers can be measured.

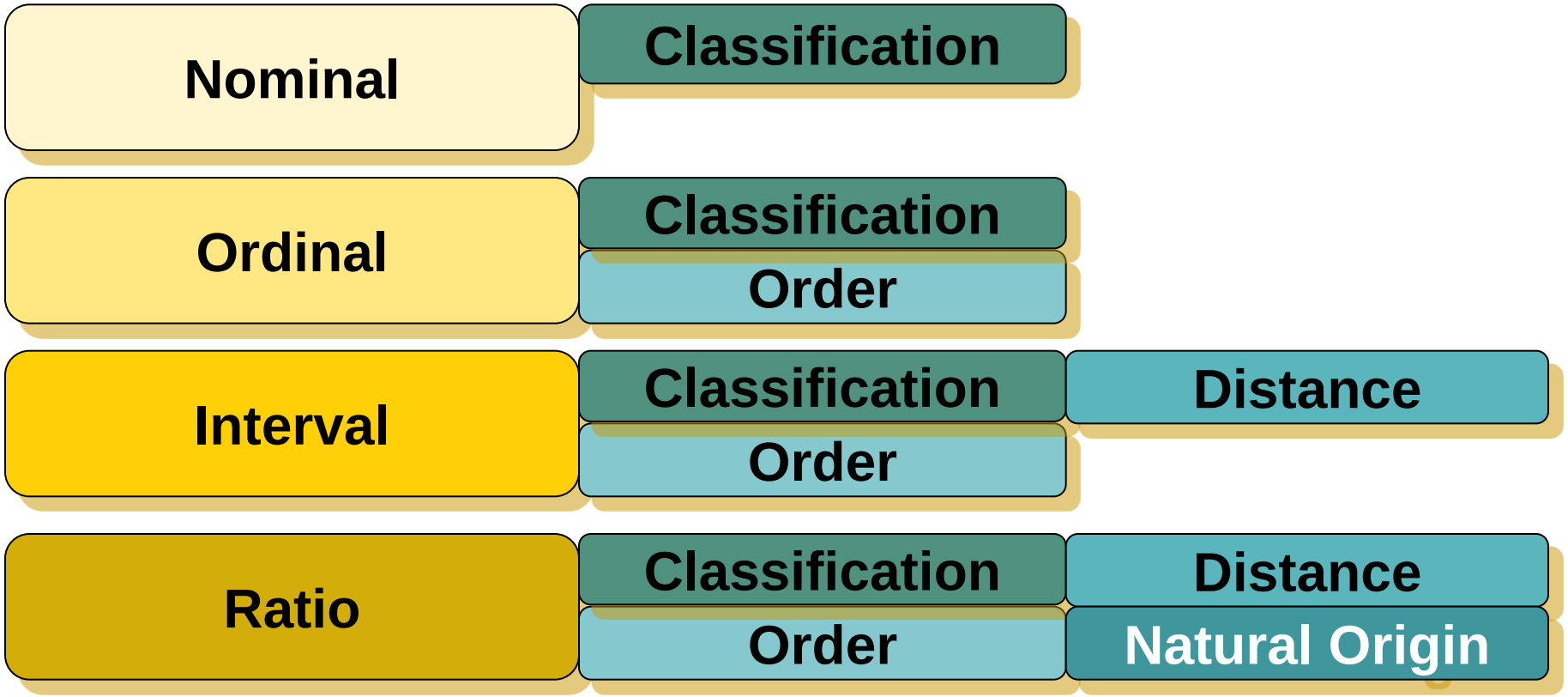
Interval Scales

- Characteristics of nominal and ordinal scales plus the concept of equality of interval
- Equal distance exists between numbers, but zero point is relative (not absolute)
- Examples:
  - Attitudes (Thurston Scale) 10-20, 21-30, …
  - Temperature (F or C)

Interval Scales (con't)

- There is no absolute origin, so for example one cannot calculate ratios; for example, say that 60 degrees is twice as hot as 30 degrees
- Thus while no mathematical operation can be performed on the readings themselves, operations can be performed on the difference in temperature readings
- When a scale is interval and the data are relatively symmetric with one mode, one can use the arithmetic mean as the measure of central tendency
- The standard deviation is the measure of dispersion
  - The product-moment correlation, t-tests, F-tests, and other parametric tests are the statistical procedures of choice for interval data

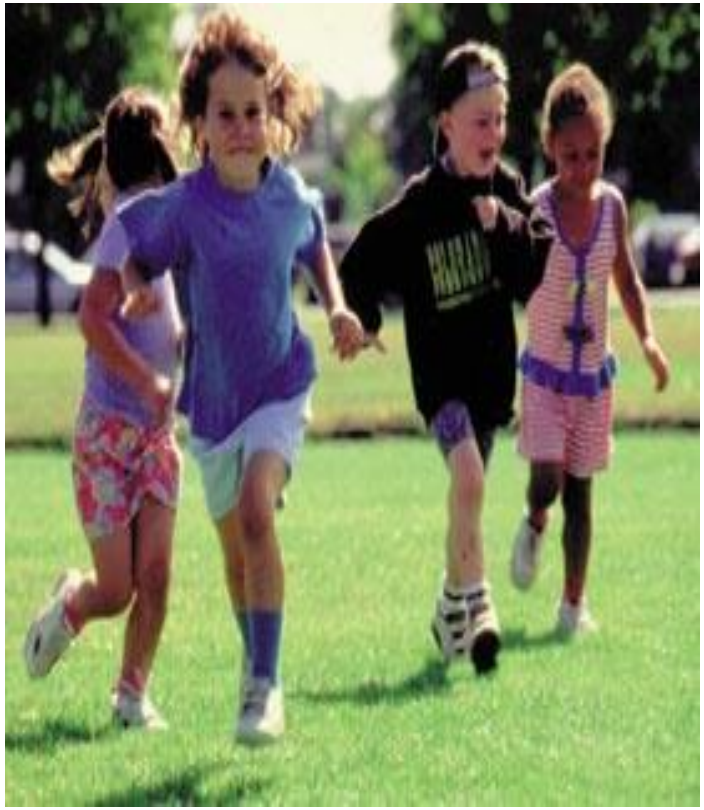
Levels of Measurement



Ratio scales also have origin. Origin means that the number series has a unique origin indicated by the number zero.

Ratio Scales

- Characteristics of previous scales plus an absolute zero point allowing ratios
- Examples:
  - Weight
  - Height
  - Age
  - Number of children



All material, both content and format,
contained herein including the slide notes but

Ratio Scales (con't)

- In research, there are many examples such as
  - Monetary values
  - Population counts
  - Distances
  - Rates
  - Amounts of time

- All statistical techniques mentioned up to this point are usable with ratio scales; geometric and harmonic means which are other measures of central tendency and coefficients of variation may also be calculated
- Higher levels of measurement generally yield more information and are appropriate for more powerful statistical procedures

Summary of Data Types

- Nominal - categories only (male/female)
- Ordinal - categories with some rank order (letter grades)
- Interval - differences but no natural starting point (years)
- Ratio - differences and a natural starting point (dollar amounts)

Quantitative vs Qualitative

- Data can be also be classified as being qualitative or quantitative
- The statistical analysis that is appropriate depends on whether the data for the variable are qualitative or quantitative
- In general, there are more alternatives for statistical analysis when the data are quantitative

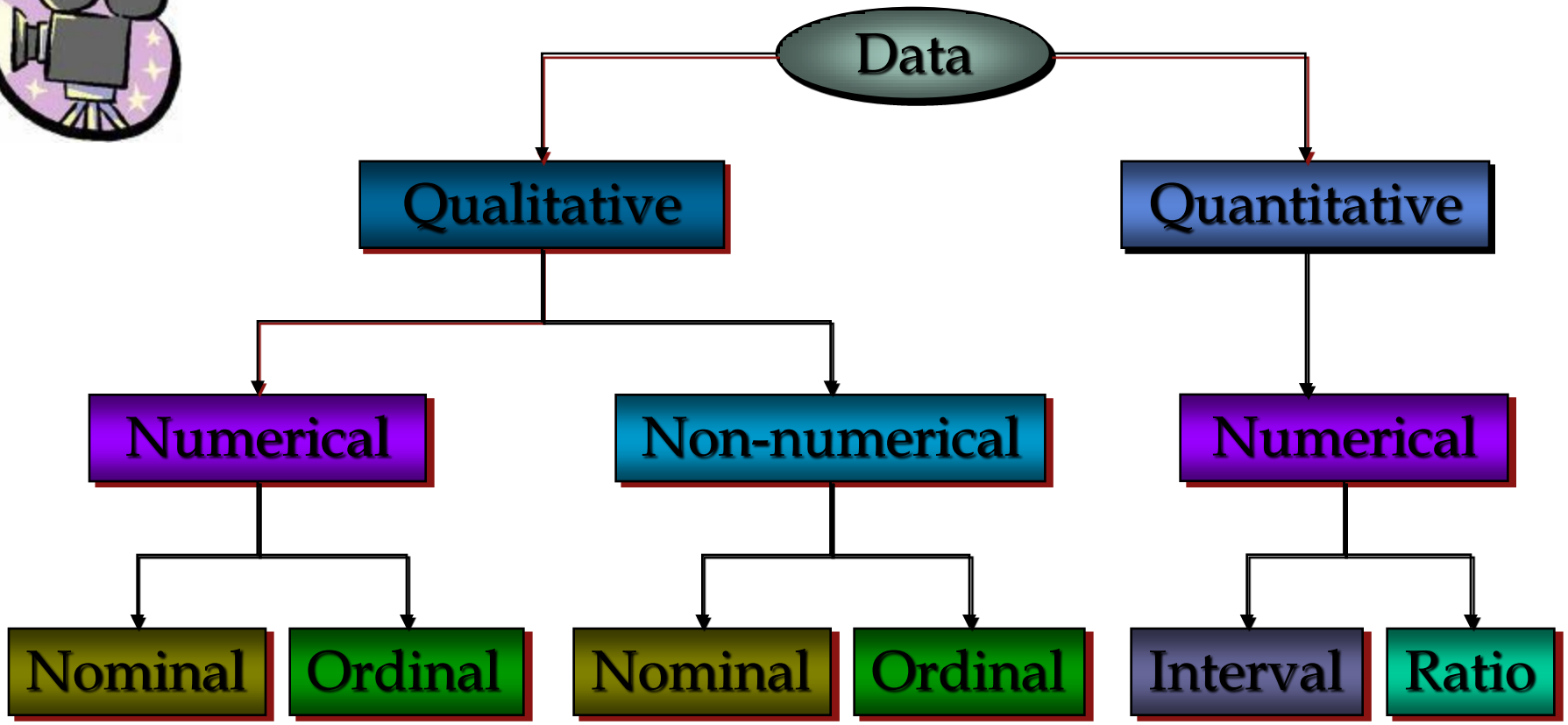
Qualitative Data

- Qualitative data are labels, names, or numeric codes use to identify an attribute of each element
- Qualitative data use either the nominal or ordinal scale of measurement
- Qualitative data can be either numeric or nonnumeric
- The statistical analysis for qualitative data are more limited

Quantitative Data

- Quantitative data indicate either how many or how much
- Quantitative data use either the interval or ratio scale of measurement
- Quantitative data are always numeric
- Ordinary arithmetic operations are meaningful only with quantitative data

Scales of Measurement



Central Tendency & Dispersion

Scale	Central Tendency	Dispersion
Nominal	Mode	None
Ordinal	Median	Percentile or quartile
Interval	Arithmetic mean (symmetric with one mode)	Standard deviation
Ratio	Arithmetic, geometric, and harmonic means	Standard deviation

Exercise

- Answer the following questions for the data set on the next slide

1. How many elements are there?
2. How many variables are there?
3. How many observations are there?
4. Which variables are qualitative?
5. Which variables are quantitative?
6. What measurement scale is being used for each variable?

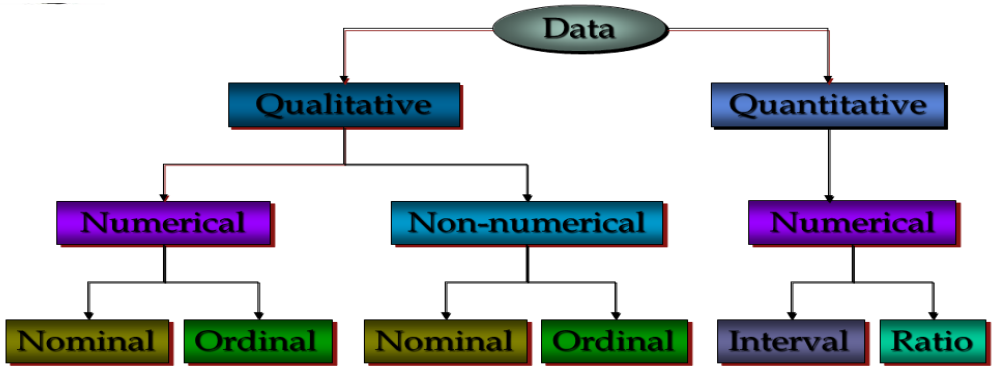
University Professors

Name	Overall Rating (out of 5.0)	Number of Years Teaching	Department	Age
Byte, Ima	4.9	40	Accounting	75
Lawrence, Joe	4.8	28	Accounting	58
Wade, Larry	4.6	12	ITM	62
Kelly, Anne	4.4	10	Accounting	42
McKinney, Alice	4.3	6	Marketing	40
Cooper, Karen	4.1	15	ITM	52
Parker, Jeff	4.1	2	ITM	46
Dripper, Ivy	3.9	20	Finance	57
Crennard, Rob	3.5	5	Marketing	36
Riviera, Jose	3.0	30	Management	68

Do not look ahead !

Copyright Dan Brandon, PhD, PMP

Answers:



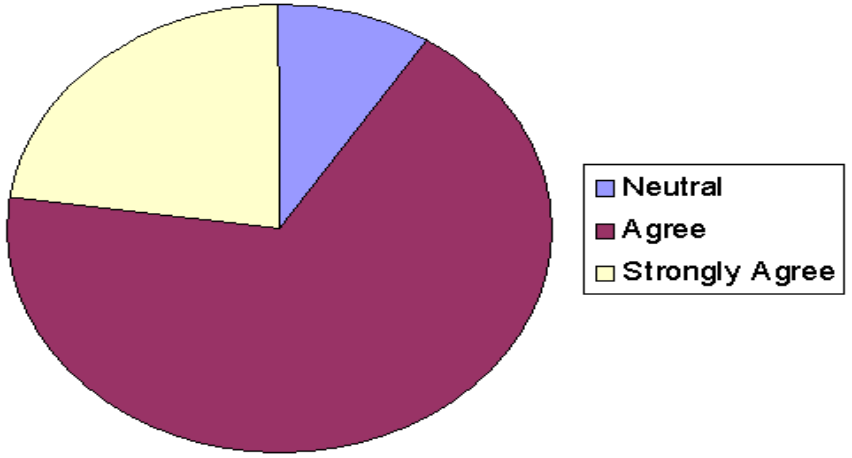
1. 10 elements
2. 4 variables
3. 10 observations
4. ‘number of years teaching’ and ‘age’ are quantitative
5. ‘rating’ and ‘department’ are qualitative
6. ‘rating’ is ordinal (numeric), ‘department’ is nominal (non-numeric), ‘number of years teaching’, and ‘age’ are ratio (numeric)

Summarization Techniques

- After the data has been collected, it must first be summarized and then analyzed
- Recall that data can be quantitative or qualitative
  - Qualitative data are described using labels or names
  - Quantitative data are described using numeric quantities
- Some summarization techniques can be applied to both types of data and others can be applied to just one type of data

Tabular And Graphical Methods Of Summarizing Qualitative Data

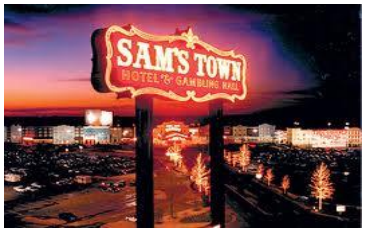
- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Bar Graph
- Pie Chart



Frequency Distribution

- A frequency distribution is a tabular summary of data showing the total number of occurrences of each unique value in a data set
- The purpose of creating a frequency distribution is to provide insights about the data that cannot be quickly obtained by looking only at the original data set

Frequency Distribution (con't)



Guests staying at Sam's Town were asked to rate the quality of their accommodations as being ***excellent***, ***above average***, ***average***, ***below average***, or ***poor***. The ratings provided by a sample of 20 guests are shown below.

Below Average	Average	Above Average
Above Average	Above Average	Above Average
Above Average	Below Average	Below Average
Average	Poor	Poor
Above Average	Excellent	Above Average
Average	Above Average	Average
Above Average	Average	

Data Summarized By A Frequency Distribution

Guest Quality Ratings

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

Frequency Distributions

- The relative frequency distribution is a tabular summary of data showing the *proportion* of times each unique value occurred out of the total number of values in a data set
- A percent frequency distribution is a tabular summary of data showing the *percentage* of times each unique value occurred out of the total number of values in a data set

Expressing Frequencies As Relative Or Percentage Frequencies

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

To express this value as a relative frequency, divide 2 by the total (20) to get .10

The percentage frequency would be 10%

Example of Relative Frequency and Percent Frequency Distributions

Guest Quality Ratings

Rating	Relative Frequency	Percent Frequency
Poor	.10	10%
Below Average	.15	15%
Average	.25	25%
Above Average	.45	45%
Excellent	.05	5%
Total	1.00	100%

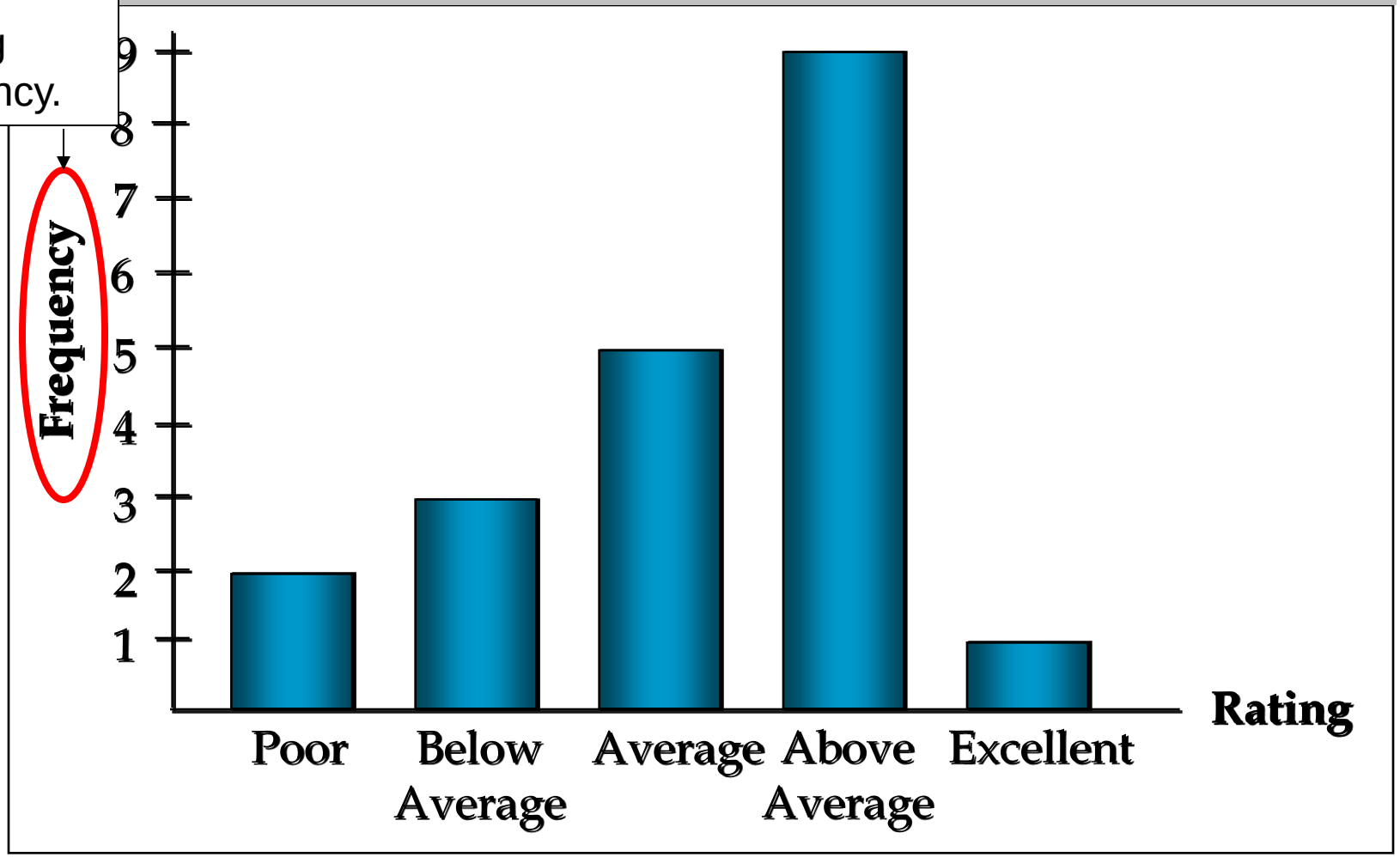
Relative frequencies are expressed as proportions; percent frequencies are expressed as percentages.

Bar Graph

- A bar graph is a graphical device for depicting qualitative data that have been summarized in a frequency, relative frequency, or percent frequency distribution
- On the horizontal axis (X) we specify the labels that are used for each of the classes
- A frequency, relative frequency, or percent frequency scale can be used for the vertical (Y) axis
- Using a bar of fixed width drawn above each class label, we extend the height appropriately
- The bars are not touching each other to emphasize the fact that each class is a separate category

Bar Graph Presenting Qualitative Data As A Frequency Distribution

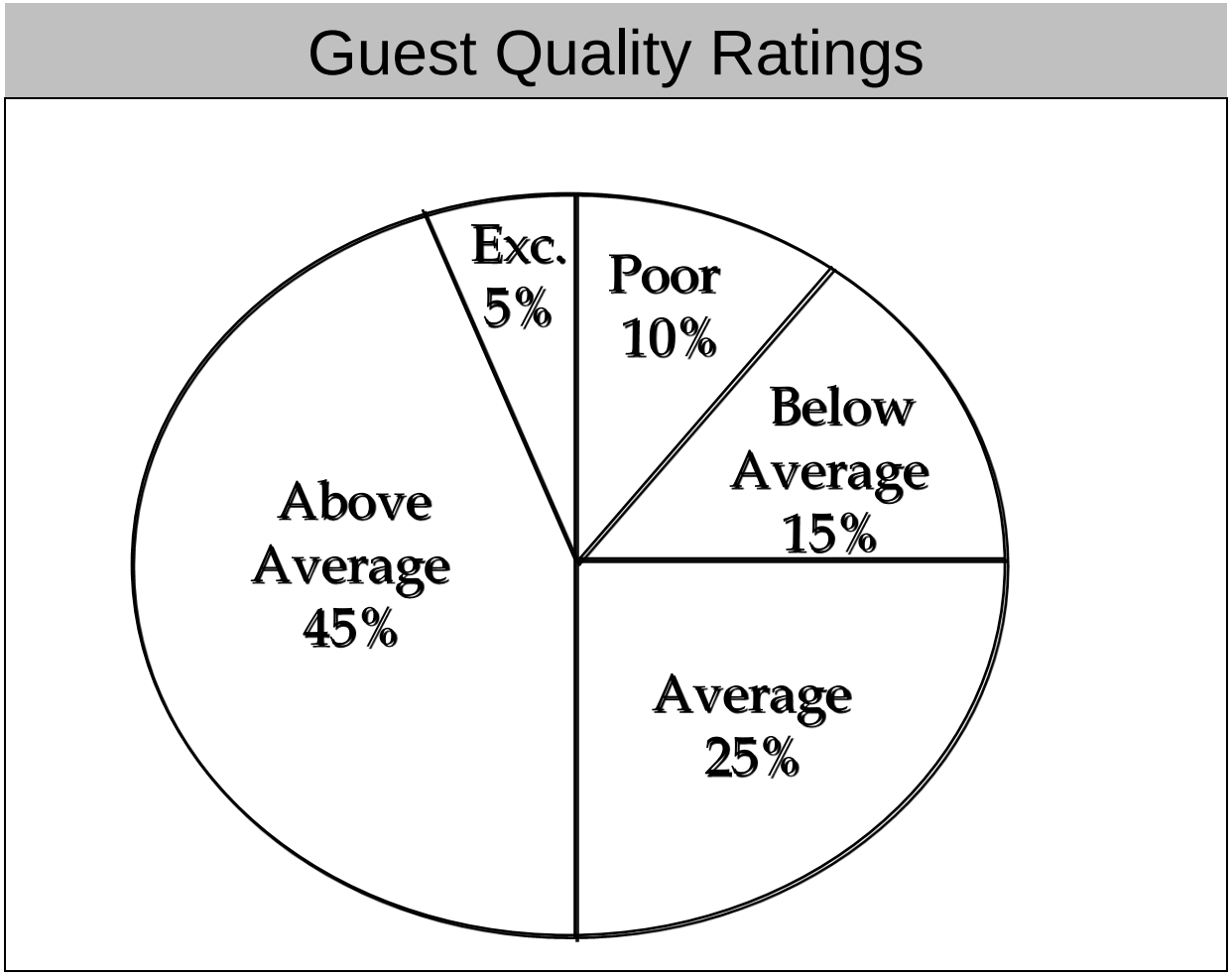
In this case, we are plotting frequency.



Pie Chart

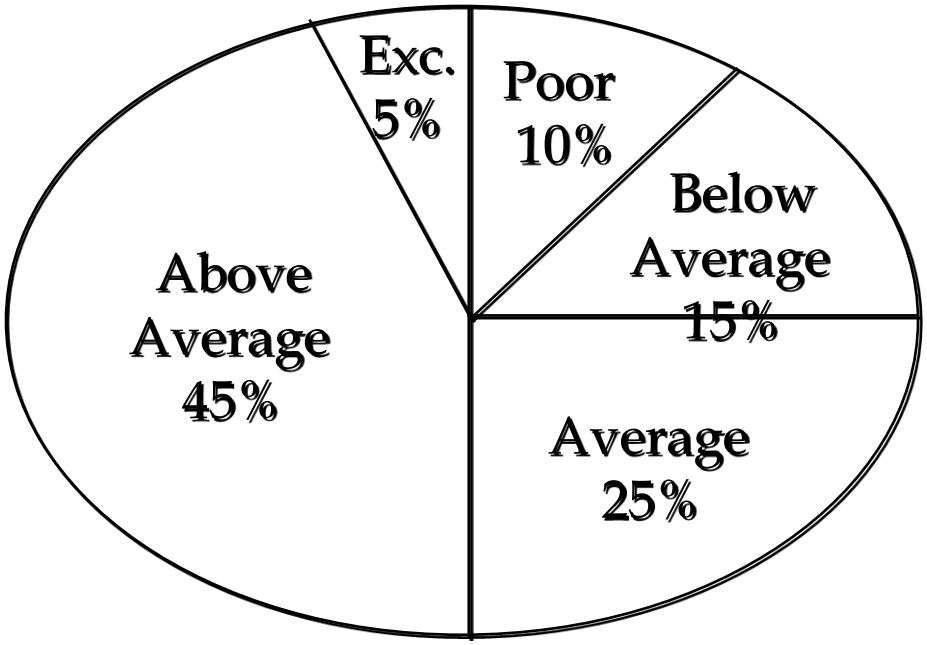
- The pie chart is a commonly used graphical device for presenting relative frequency distributions for qualitative data
- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class
- Since there are 360 degrees in a circle, a class with a relative frequency of .25 would consume .25(360) = 90 degrees of the circle

Pie Chart Presenting Qualitative Data



Some Insights Gained From Examining Descriptive Statistics

- One-half of the customers surveyed gave the Inn a quality rating of “above average” or “excellent”
- For each customer who gave an “excellent” rating, there were two customers who gave a “poor” rating



Reported TV Show Watching (sample of 50 TV viewers)

Millionaire	Millionaire	Millionaire	Fraiser	Charmed
Fraiser	Fraiser	Millionaire	Millionaire	Fraiser
Fraiser	Millionaire	Millionaire	Hope	Millionaire
Charmed	Millionaire	Fraiser	Hope	Millionaire
Hope	Charmed	Fraiser	Fraiser	Millionaire
Millionaire	Fraiser	Millionaire	Millionaire	Hope
Fraiser	Millionaire	Millionaire	Charmed	Hope
Hope	Millionaire	Millionaire	Millionaire	Millionaire
Fraiser	Fraiser	Millionaire	Fraiser	Fraiser
Millionaire	Millionaire	Hope	Millionaire	Fraiser

Exercise

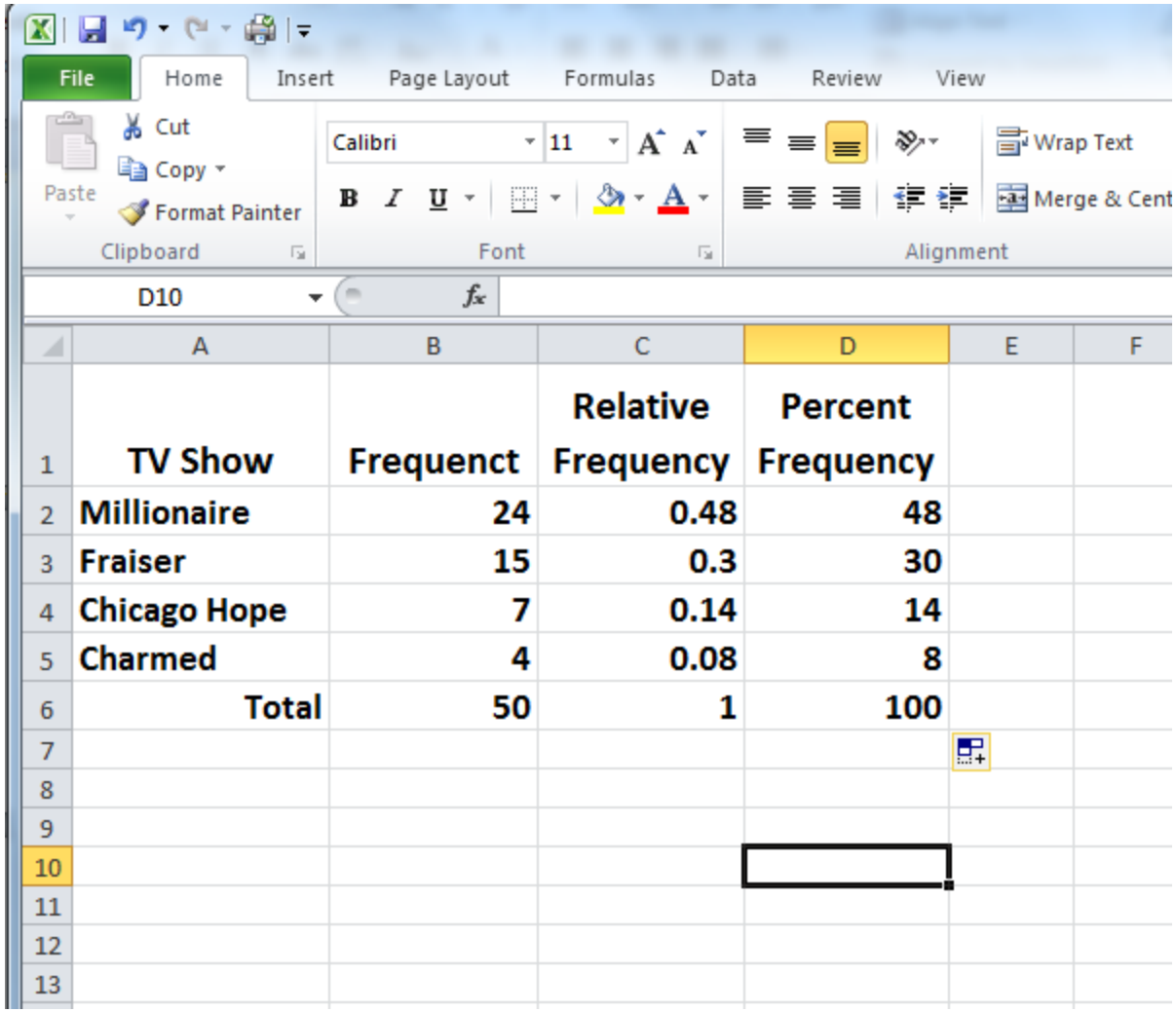
- **Using the data on the previous slide**
- **Build an Excel spreadsheet with columns for frequency, relative frequency, and percent frequency**
- **Using frequency, construct a pie chart and bar chart in excel**

Copyright Dan Brandon, PhD, PMP

Do not look ahead !

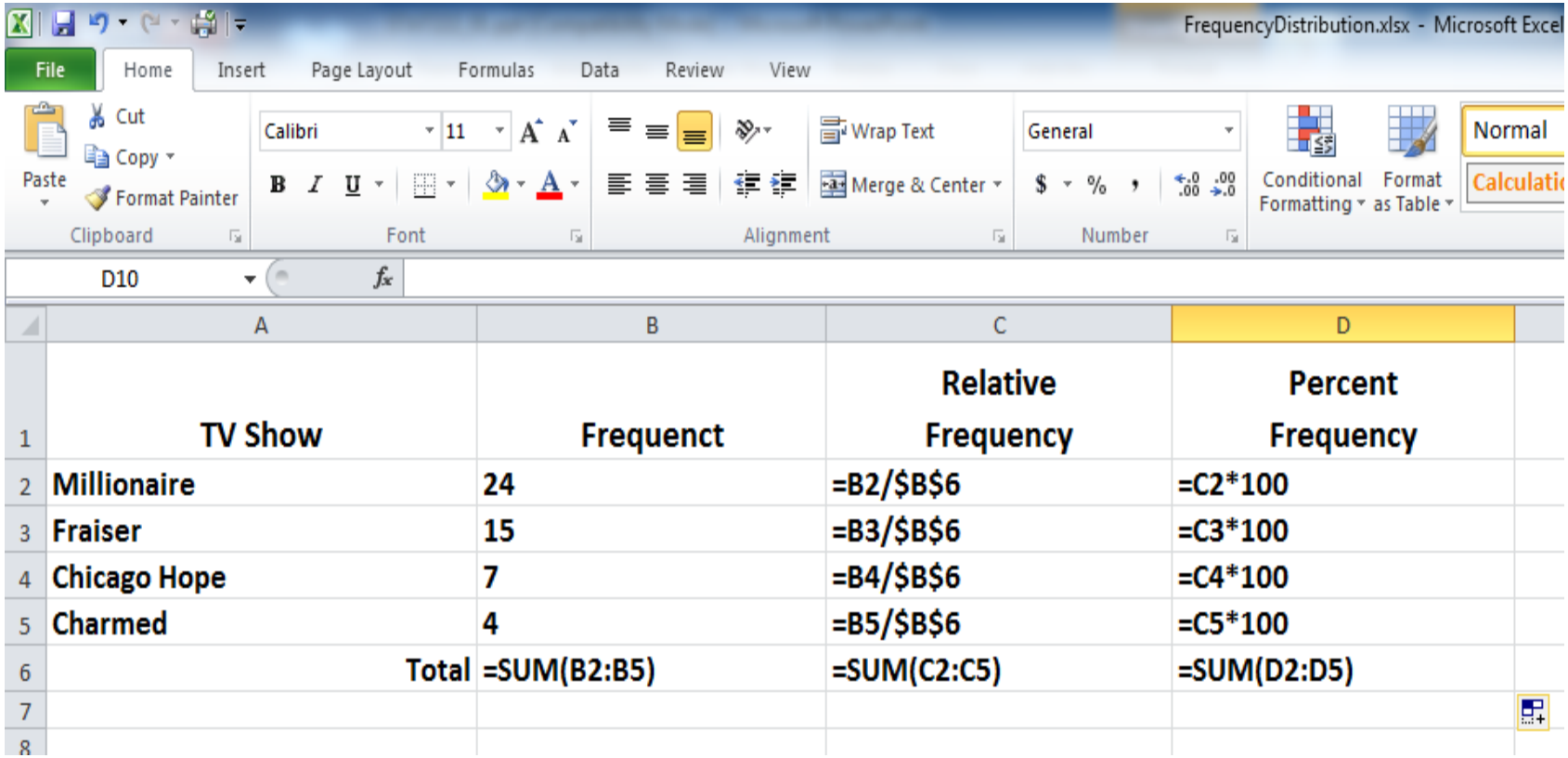
Copyright Dan Brandon, PhD, PMP

Frequency Distributions in Excel



	A	B	C	D
1	TV Show	Frequenct	Relative Frequency	Percent Frequency
2	Millionaire	24	0.48	48
3	Fraiser	15	0.3	30
4	Chicago Hope	7	0.14	14
5	Charmed	4	0.08	8
6	Total	50	1	100
7				
8				
9				
10				
11				
12				
13				

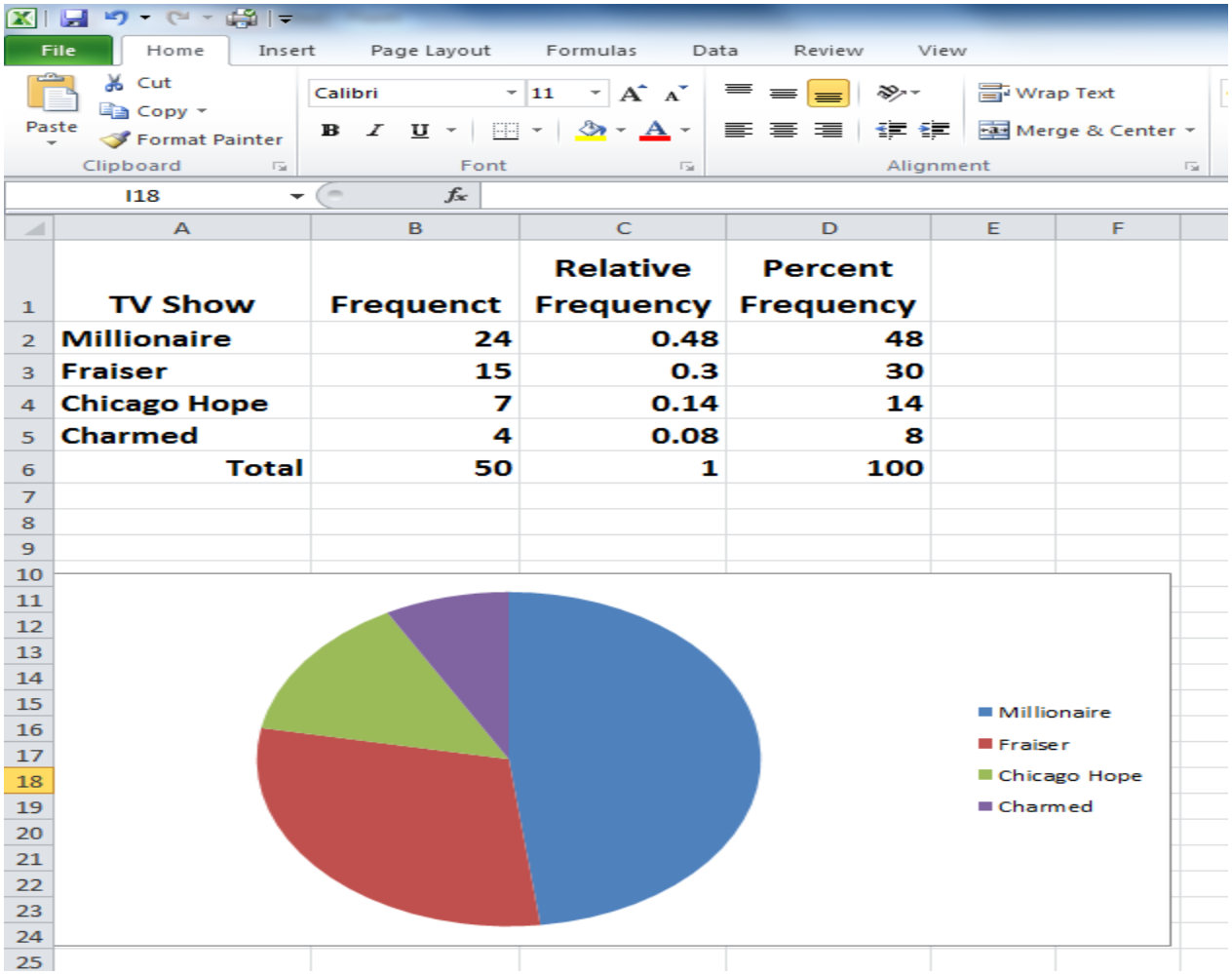
Frequency Distributions in Excel (Formulas)



	A	B	C	D
1	TV Show	Frequenct	Relative Frequency	Percent Frequency
2	Millionaire	24	=B2/$B$6	=C2*100
3	Fraiser	15	=B3/$B$6	=C3*100
4	Chicago Hope	7	=B4/$B$6	=C4*100
5	Charmed	4	=B5/$B$6	=C5*100
6	Total	=SUM(B2:B5)	=SUM(C2:C5)	=SUM(D2:D5)
7				
8				

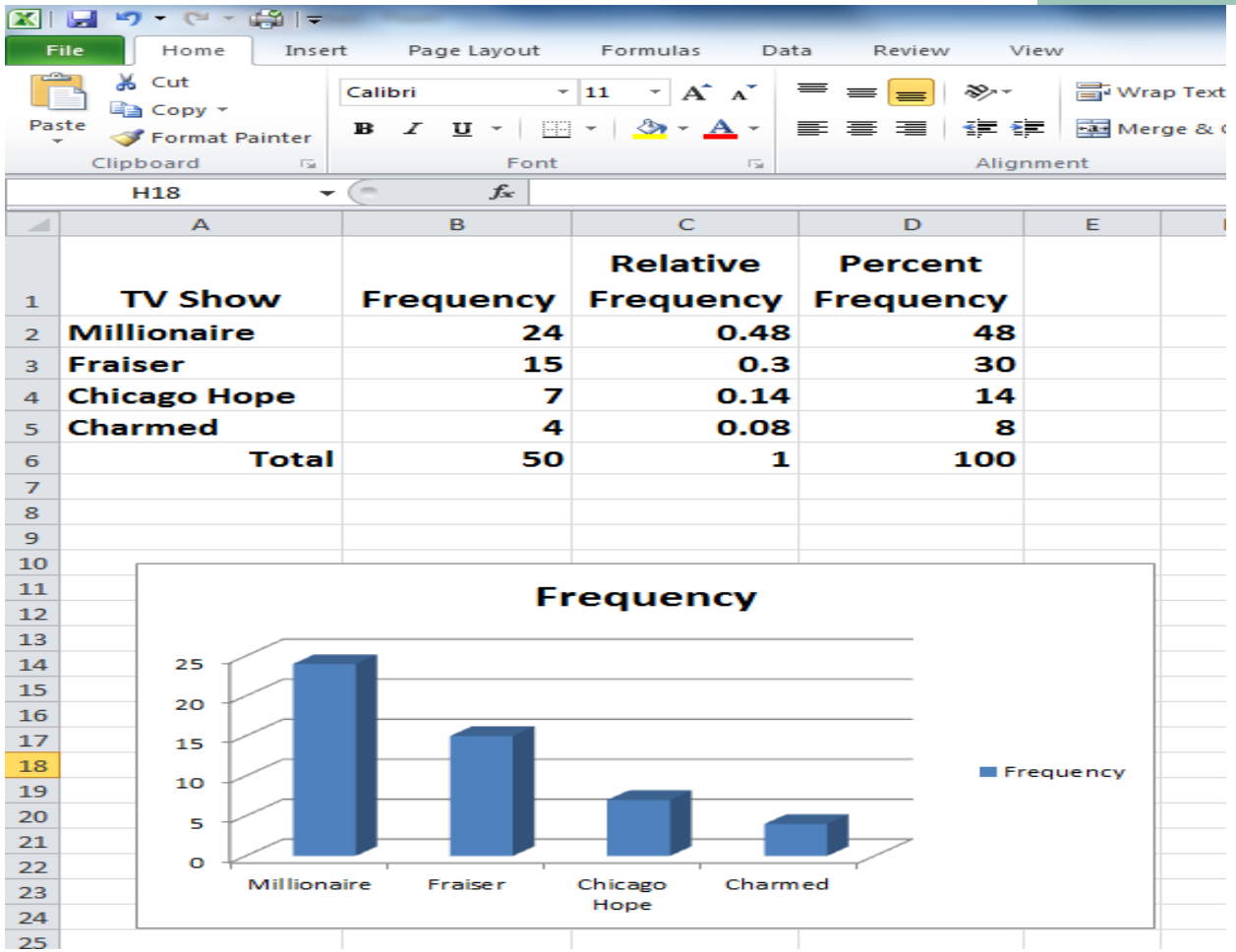
Copyright Dan Brandon, PhD, PMP

Excel Pie Chart



Copyright Dan Brandon, PhD, PMP

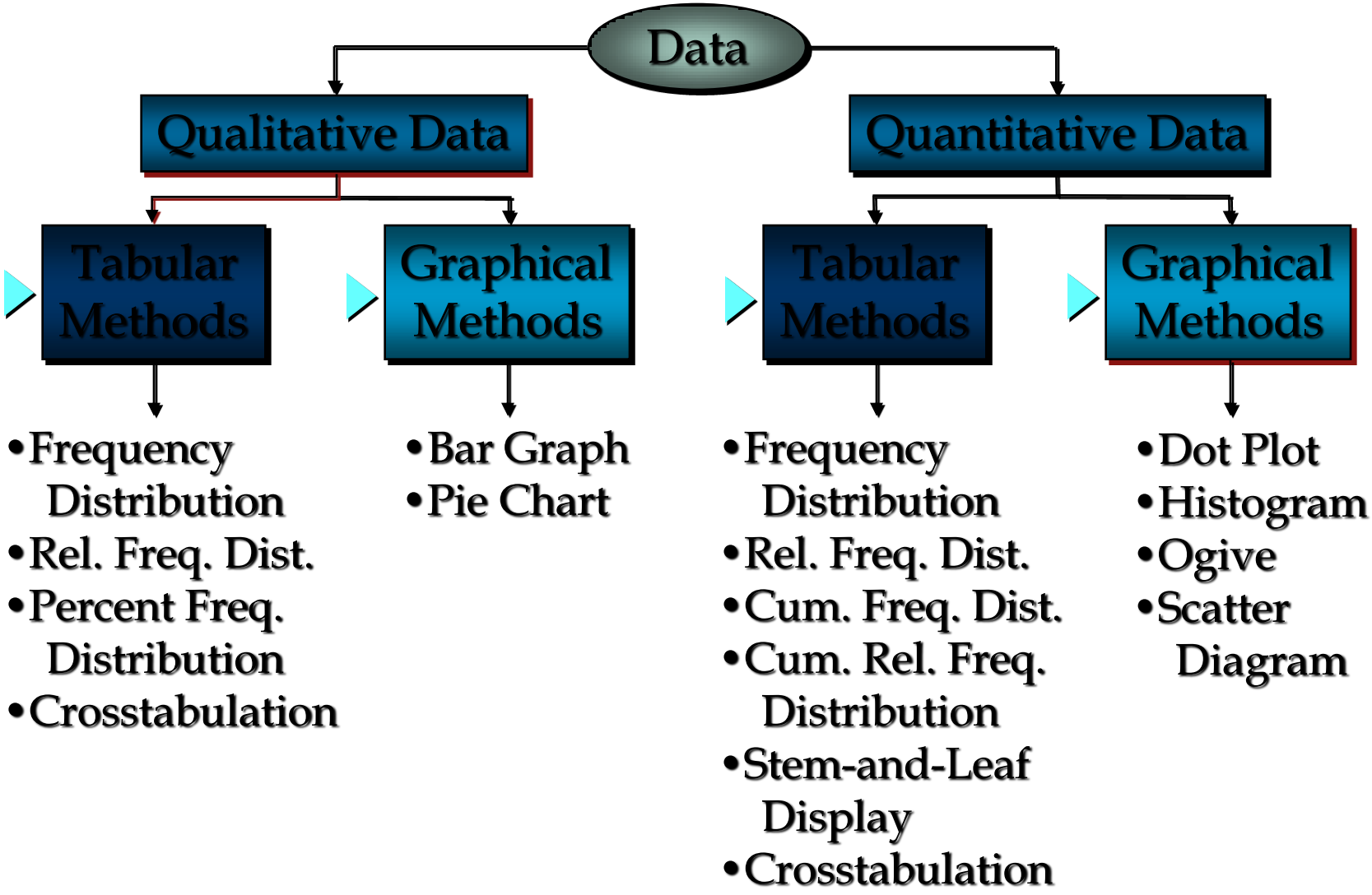
Excel Bar Chart



TV Show	Frequency	Relative Frequency	Percent Frequency
Millionaire	24	0.48	48
Fraiser	15	0.3	30
Chicago Hope	7	0.14	14
Charmed	4	0.08	8
Total	50	1	100

Copyright Dan Brandon, PhD, PMP

Tabular and Graphical Procedures



Techniques For Summarizing Numeric Data

- Measures of Central Tendency [location] (mean, median, mode, percentiles, quartiles)
- Measures of Variability (range, inter-quartile range, variance, standard deviation, coefficient of variation)
- Measures of Relative Location (z-scores) and Detecting Outliers
- Exploratory Data Analysis (the 5-number summary and box plot)
- Measures of Association Between Two Variables (covariance and correlation coefficient)

Measures of Location (aka measures of central tendency)

- Mean
  - Arithmetic
  - Weighted
  - Geometric
  - Harmonic
- Median
- Mode
- Percentiles
- Quartiles

The Mean

[arithmetic average – see appendix for other means]

- The mean of a data set is the arithmetic average of all the data values
- If the data are from a sample, the mean is denoted by $\bar{x}$ (x-bar)

$$\bar{x} = \frac{\sum x_i}{n}$$

n = sample size

- If the data are from a population, the mean is denoted by μ (mu).

$$\mu = \frac{\sum x_i}{N}$$

N = population size

Copyright Dan Brandon, PhD, PMP

Example: Apartment Rents

Given below is a sample of monthly rent values ($) for one-bedroom apartments. The data is a sample of 70 apartments in a particular city. The data are presented in ascending order.

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	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	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Copyright Dan Brandon, PhD, PMP

Calculating The Mean Rent

- Add up all the rents and divide by the number of rents
- The mean is denoted by $\bar{x}$ (x-bar)

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{34{,}356}{70}$$

$$= 490.8$$

Copyright Dan Brandon, PhD, PMP

The Median

- The median of a data set is the value in the middle when the data items are arranged in ascending order
- For an odd number of observations (n), the median is the middle value
- For an even number of observations (n), the median is the average of the two middle values
- The median may useful instead of the mean when the data set includes a few extreme values (outliers), until these outliers are removed

The Median Rent

- ***i*** refers to index - Index is the position number of a value in a data set that has been arranged into ascending order.
- ***i*** = 50% * 70 = 35
- Since 70 is even, we average the values in the 35th and 36th positions: Median = (475 + 475)/2 = **475**

The Median Rent (con't)

- What would be the median rent if n= 25?

25 /2 = 12.5, round up to 13. The 13th value is 440.

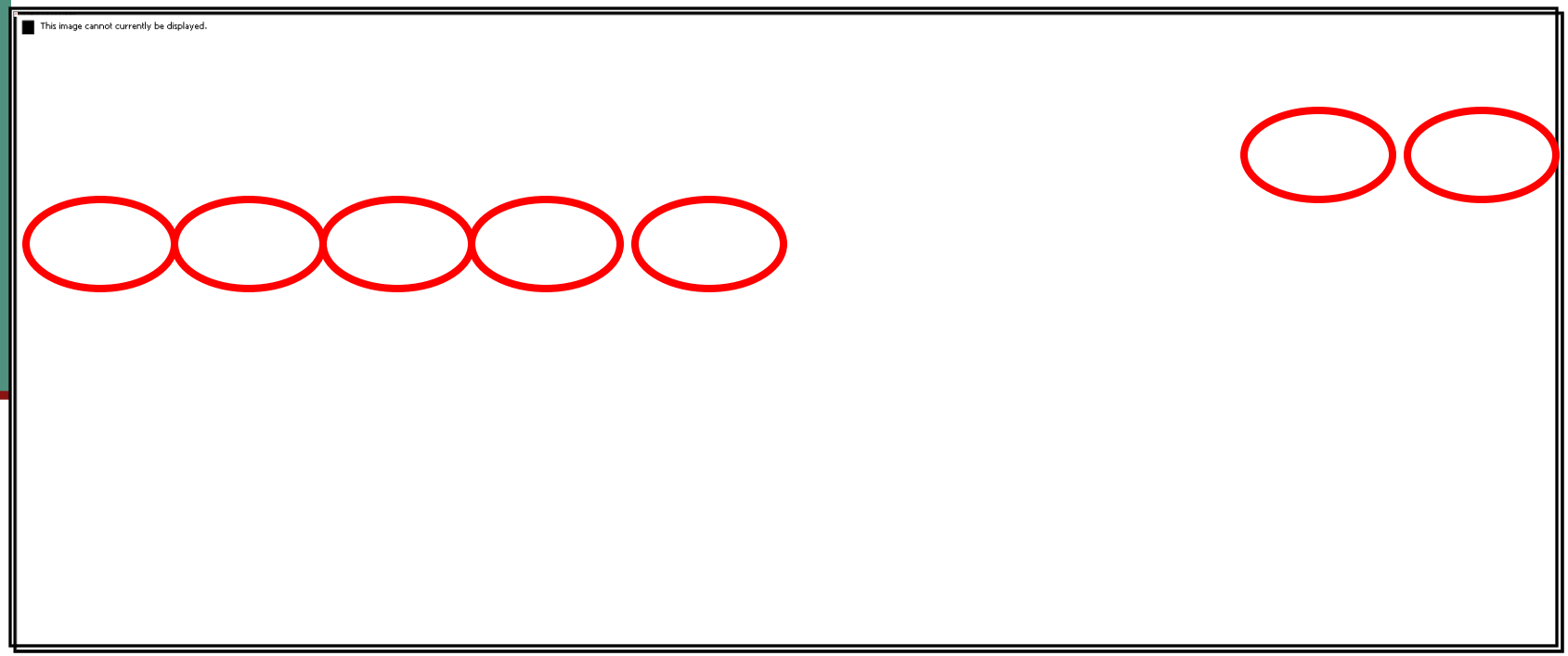
(The middle value)

The Mode

- The mode of a data set is the value that occurs with greatest frequency
- The greatest frequency can occur at two or more different values
- If the data have exactly two modes, the data are bimodal
- If the data have more than two modes, the data are multimodal

The Mode Rent

- 450 occurred most frequently (7 times) so the Mode = 450



Exercise: Mean, Median, Mode

- Consider the set of numbers 80, 90, 90, 100, 85, 90. They could be grades, for example.
- Calculate, by hand, the mean, median, and mode ?

Exercise: Mean, Median, Mode (con't)

- Do not look ahead !

Copyright – D. Brandon, Ph.D., PMP

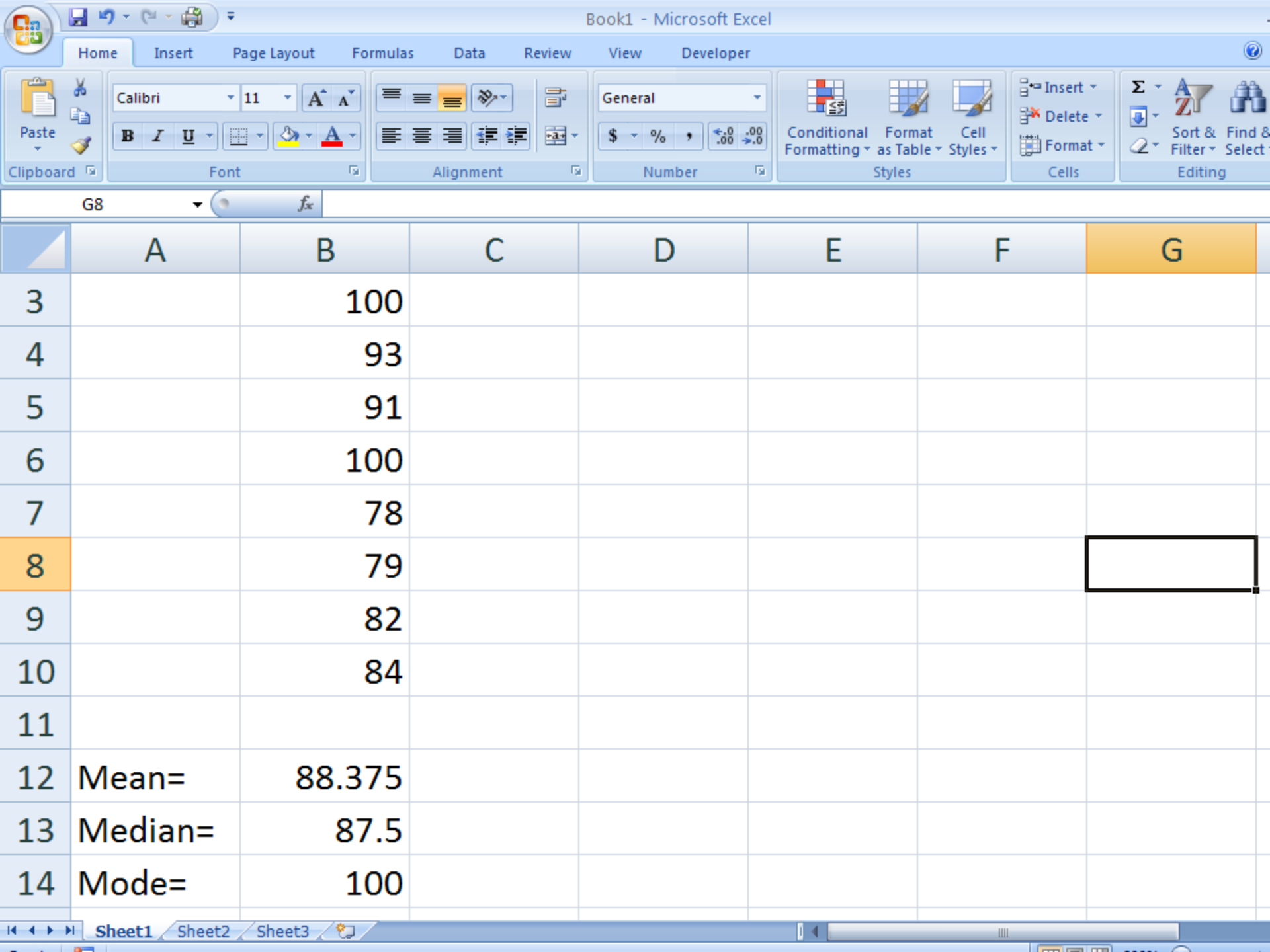
Exercise: Mean, Median, Mode (con't)

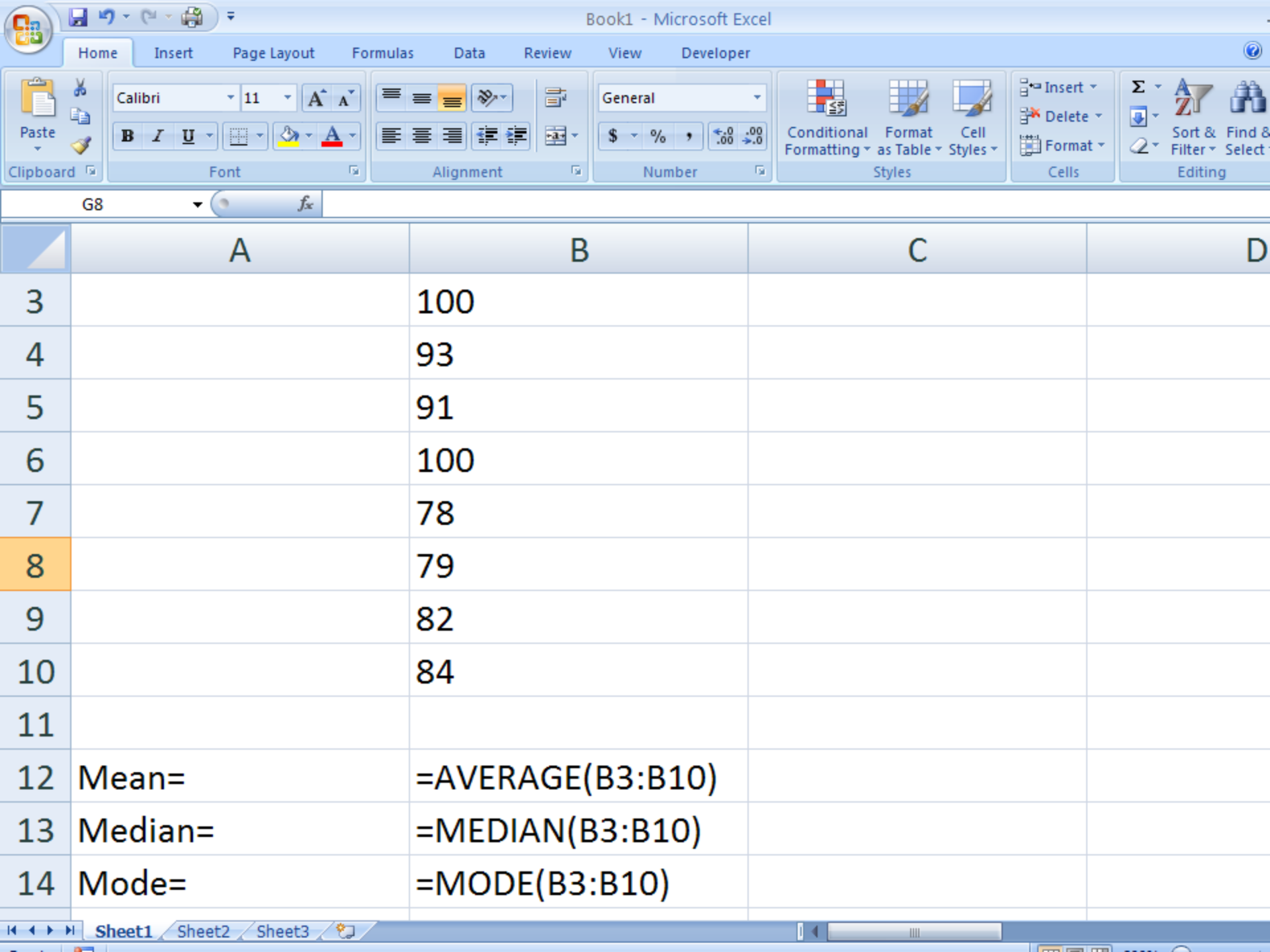
- The MEAN is the arithmetic average, the average you are probably used to finding for a set of numbers - add up the numbers and divide by how many there are: (80 + 90 + 90 + 100 + 85 + 90) / 6 = 89 1/6
- The MEDIAN is the number in the middle
- In order to find the median, you have to put the values in order from lowest to highest, then find the number that is exactly in the middle: 80 85 90 90 90 100
- Since there is an even number of values, the MEDIAN is between these two, or it is 90
- The MODE is the value that occurs most often
- Since there are 3 90's, the mode is 90 - a set of data can have more than one mode

Copyright Dan Brandon, PhD, PMP

Excel's Built-in Functions

- Excel has built-in formulas to calculate mean, median, and mode:
- =average( )
- =median( )
- =mode( )
  - Note that Excel's mode function only finds the first mode


Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Paste
Clipboard
Calibri
11
Font
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
Sort & Filter
Find & Select
Editing
G8
A
B
C
D
E
F
G
3 100
4 93
5 91
6 100
7 78
8 79
9 82
10 84
11
12 Mean= 88.375
13 Median= 87.5
14 Mode= 100
Sheet1
Sheet2
Sheet3


Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Paste
Clipboard
Calibri
11
Font
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
Sort & Filter
Find & Select
Editing
G8
A
B
C
D
3 100
4 93
5 91
6 100
7 78
8 79
9 82
10 84
11
12 Mean= =AVERAGE(B3:B10)
13 Median= =MEDIAN(B3:B10)
14 Mode= =MODE(B3:B10)
Sheet1
Sheet2
Sheet3

Which measure of central tendency?

- If there are a few extreme values in your data set, extreme values may distort the mean but not the median or the mode
- Let's say you are a fund-raiser. Your last 10 donations were:
- $5, $5, $15, $5, $10, $5, $10, $15, $10 and $1,000.
- What do you want to tell the next person you solicit for a donation?
- 1. That the average donation is over $100 (actually its $103.50)
- 2. The median donation is $10
- 3. The mode donation is $5

Which measure (con’t)

- Fashion is generally an answer to the question, "What is the mode?" e.g. "What's the most popular sneaker?" or "What shirt style would most people want to wear?“
- Of the two outcomes: "strike out" or "base hit or better," assigning 0 to "strike out" and 1 to "base hit or better" and then averaging may be more useful to a manager than determining the mode
  - Which in pro baseball is for everybody to "get out" (we do not often see a batting average over 500)
- Knowing the median test score is important to people who want to think they are in the "better half of the population"

Mean, Median, Mode

- The mean of the numbers can be misleading
- If I tell you the mean income in my neighborhood is 1 million dollars a year you might think that I am wealthy -- but maybe the neighborhood is a poor one with one very rich person making many millions of dollars a year
- The average will be large because one number is very large, the median and the mean together give a better idea of the spread of the numbers; if there were one really wealthy person then the median income would be quite low

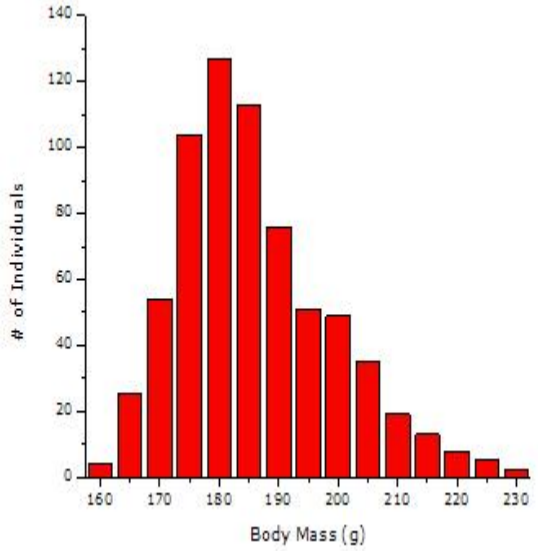
Mean, Median, Mode (con't)

- For this example: 100 100 93 91 84 82 79 78, the median is quite near to the mean, which suggests that the spread is probably evenly balanced
- The mode of a set of numbers is the number that occurs most often, in this example: mode = 100
- The mode may help to correct false impressions if you know the mean and the median but don't actually see the data
- For example, if I tell you the mean of four numbers is 50.5 and the median is 50 you may think that the four numbers are close to one another, like the numbers in the previous example
- In fact, however, if the numbers are: 101 99 1 1 The mean is (101 + 99 + 1 + 1)/4 = 202/4 = 50.5 and the median = (99+1)/2 = 50
- But the mode here is 1; if you know that, you will know that 1 is repeated more than once in the data and, since the mean is near 50 with only four numbers, that might give you the idea that the data include several large numbers and several small numbers rather than four numbers close to one another

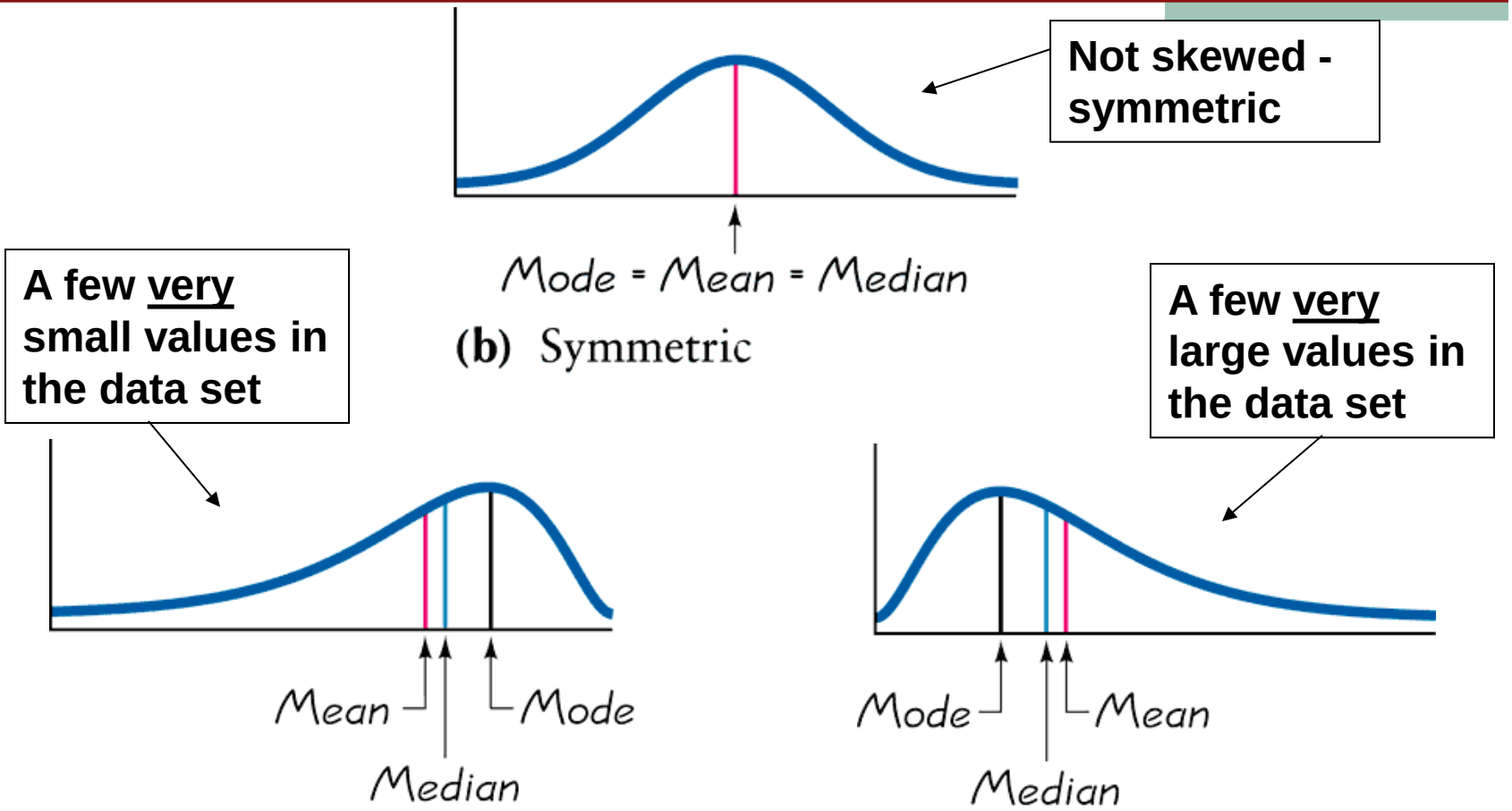
Copyright Dan Brandon, PhD, PMP

Skewed Data

- When the mean, median, and mode are not “aligned, the data is said to be “skewed”
- Data is skewed if it is not symmetric, and if it extends more to one side than the other



Skewness


Not skewed - symmetric
Mode = Mean = Median
(b) Symmetric
A few very small values in the data set
Mean
Mode
Median
(a) Skewed to the Left (Negatively)
A few very large values in the data set
Mode
Mean
Median
(c) Skewed to the Right (Positively)

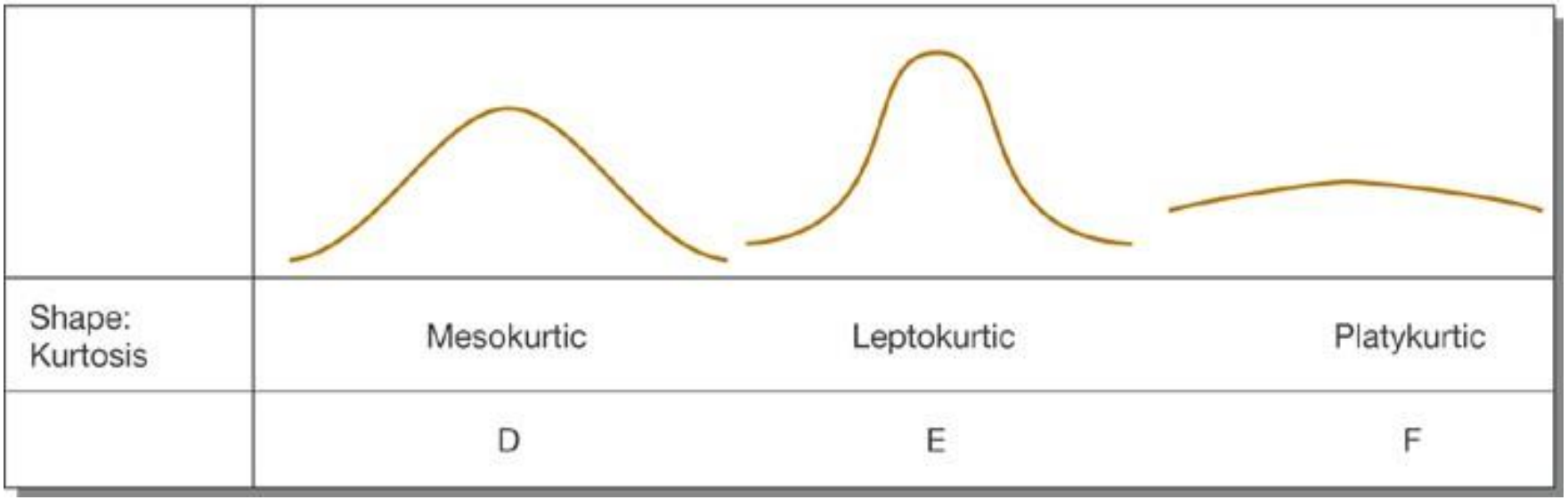
Kurtosis

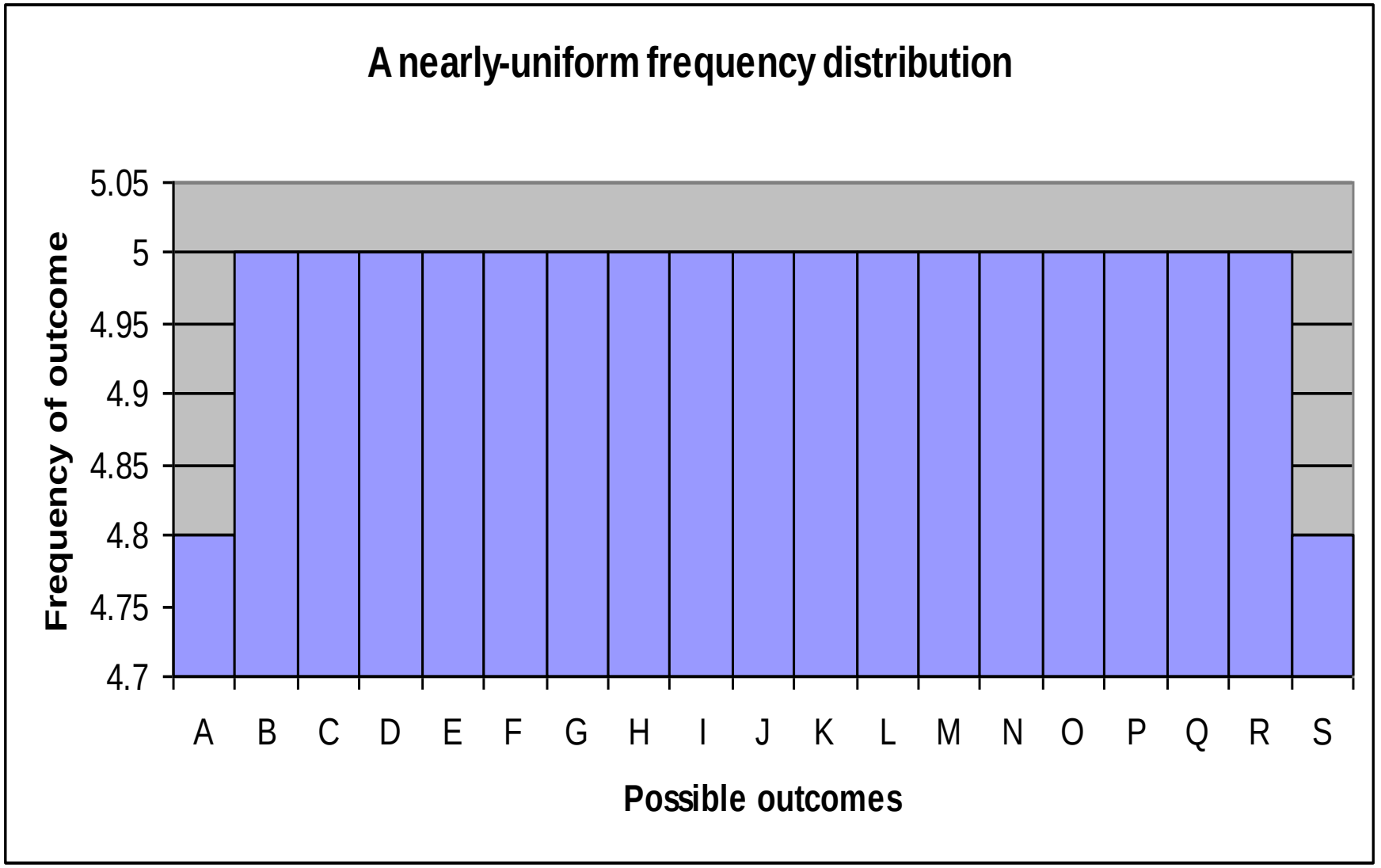
- **Kurtosis** is a measure of a distribution's "peakedness" - a measure of the extent to which observations cluster around a central point
- The symbol for kurtosis is ku
- For a normal distribution, the value of the kurtosis statistic is 0
- Positive kurtosis indicates that the observations cluster more and have longer tails than those in the normal distribution
- Negative kurtosis indicates the observations cluster less and have shorter tails

Kurtosis (con't)

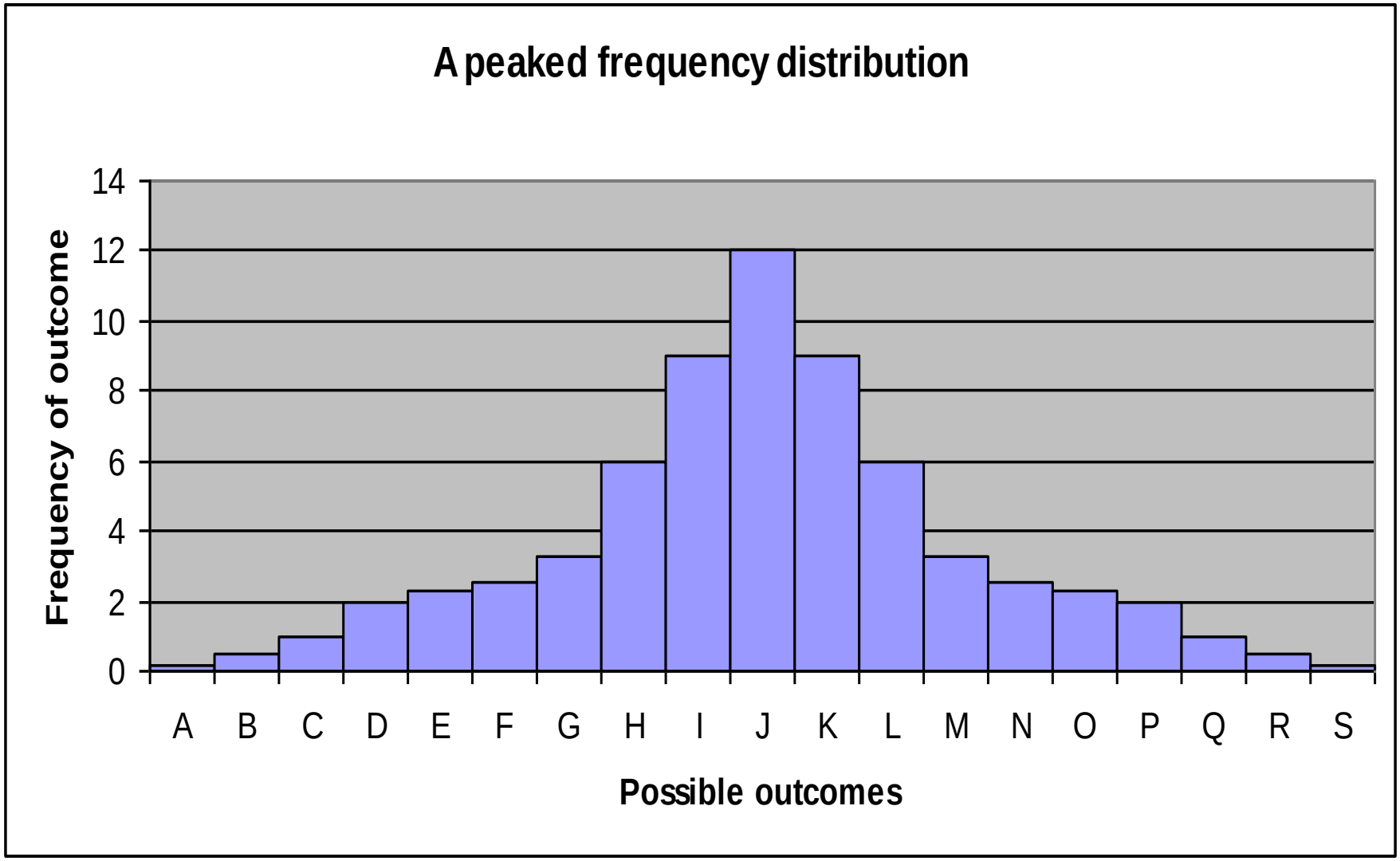
- Intermediate or **mesokurtic** distributions approach normal - the value of **ku** for a normal distribution is close to 0
- A **leptokurtic** distribution will have a positive value - distributions that have scores which cluster heavily or pile up in the center are peaked or leptokurtic
- Flat distributions are called **platykurtic -** the platykurtic distribution will be negative

Summarizing Distributions with Shape





Kurtosis would be negative because of the flat top.



Kurtosis would be positive because of the peaked top.

Measures of Variability

- Range
- Variance
- Standard Deviation
- Coefficient of Variation
- Mean Absolute Deviation
- Inter-quartile Range
- Etc…

The Range

- The range of a data set is the difference between the largest and smallest data values
- It is the simplest measure of variability
- It is very sensitive to the smallest and largest data values

The Range of Apartment Rents

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	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	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

- The range is 615 – 425 or 190

Copyright Dan Brandon, PhD, PMP

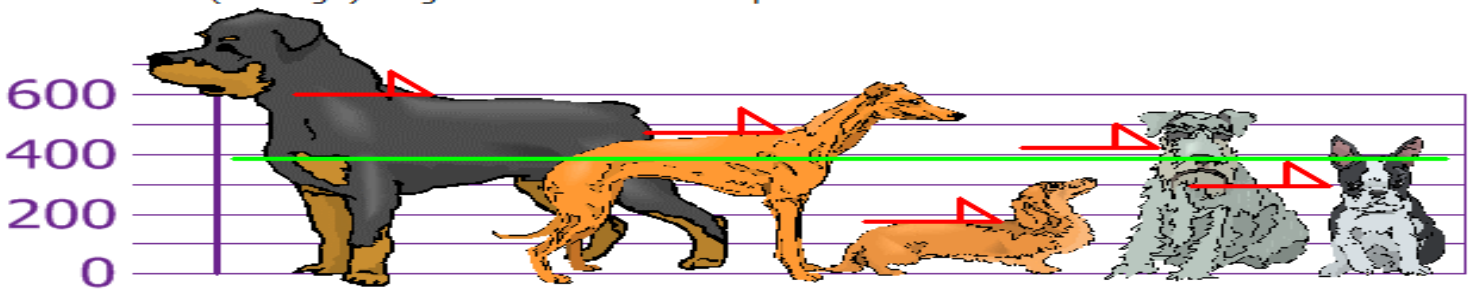
Variance

- The variance is a measure of variability that utilizes all the data
- It is based on the difference between the value of each observation (x_i) and the mean

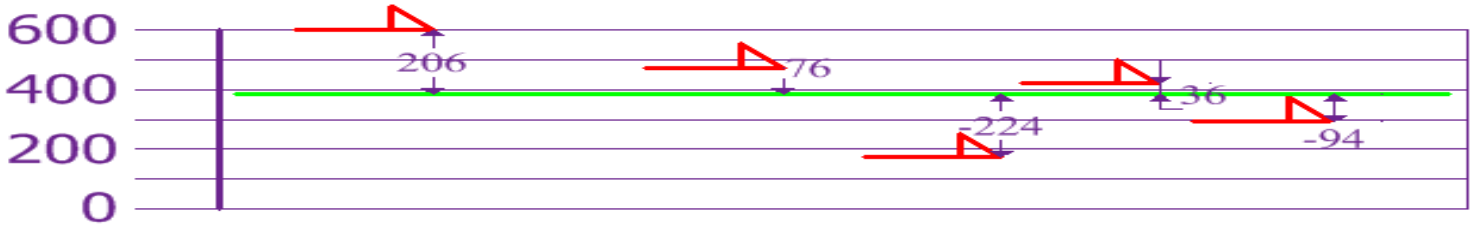
Variance (con't)

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



Now, we calculate each dogs difference from the Mean:



To calculate the Variance, take each difference, square it, and then average the result:

$$\text{Variance: } \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$
$$= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5}$$
$$= \frac{108,520}{5} = 21,704$$

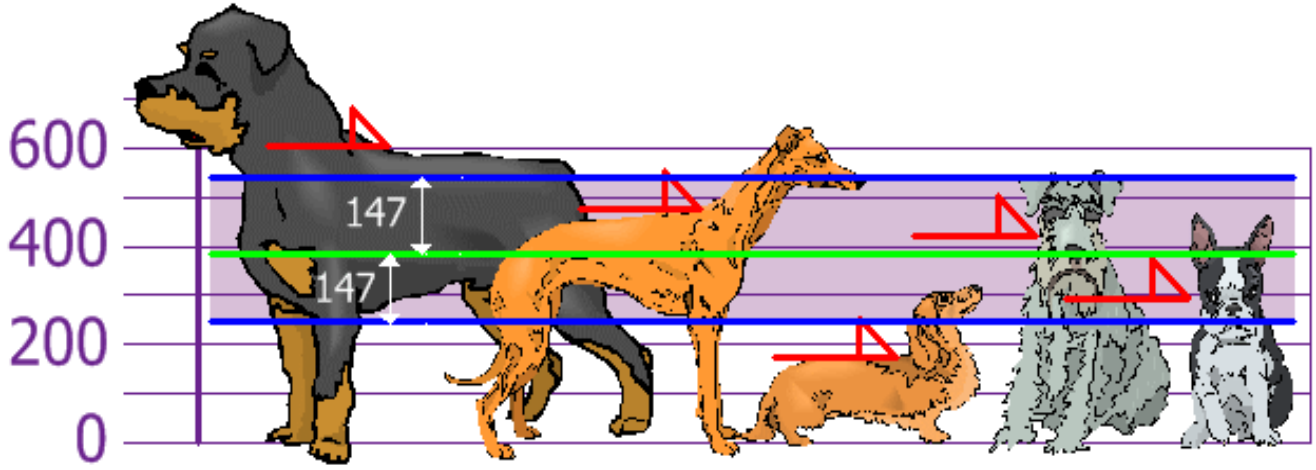
Copyright Dan Brandon, PhD, PMP

Standard Deviation

And the Standard Deviation is just the square root of Variance, so:

Standard Deviation: $\sigma = \sqrt{21{,}704} = 147.32... = 147$ (to the nearest mm)

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Copyright Dan Brandon, PhD, PMP

Standard Deviation (con't)

Our example was for a **Population** (the 5 dogs were the only dogs we were interested in).

But if the data is a **Sample** (a selection taken from a bigger Population), then the calculation changes!

When you have "N" data values that are:

- **The Population**: divide by **N** when calculating Variance (like we did)
- **A Sample**: divide by **N-1** when calculating Variance

All other calculations stay the same, including how we calculated the mean.

Example: if our 5 dogs were just a **sample** of a bigger population of dogs, we would divide by **4 instead of 5** like this:

→ Sample Variance = 108,520 / 4 = 27,130

→ Sample Standard Deviation = √27,130 = 164 (to the nearest mm)

Think of it as a "correction" when your data is only a sample.

Copyright Dan Brandon, PhD, PMP

Standard Deviation (con't)

The "**Population** Standard Deviation":

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The "**Sample** Standard Deviation":

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

Copyright Dan Brandon, PhD, PMP

Coefficient of Variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean
- Which is the inverse of one definition of the signal-to-noise ratio
- It shows the extent of variability in relation to mean of the population
- It is calculated as:
  - (standard deviation)/(mean) * 100

File Edit View Insert Format Tools Data Window Help
Type a question for help
100%
12
A1
Cereal

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03		
Corn Flakes	0.07		
Rice Krispies	0.09		
Wheaties	0.13		
Special K	0.13		
Bran Flakes	0.17		
Harmony	0.24		
Smart Start	0.30		
Frosted Flakes	0.39		
Lucky Charms	0.43		
Trix	0.43		
Cap'n Crunch	0.44		
Corn Pops	0.45		
Cocoa Puffs	0.47		
Fruit Loops	0.47		
Apple Jacks	0.48		
Mean ($\bar{x}$):		**$\Sigma(x_i - \bar{x})^2$:**	
Count (n):			
Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1			
Standard Deviation (s): ($\sqrt{s^2}$)			

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3
AutoShapes

File Edit View Insert Format Tools Data Window Help

Type a question for help

100%

12

B18 =AVERAGE(B2:B17)

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03		
Corn Flakes	0.07		
Rice Krispies	0.09		
Wheaties	0.13		
Special K	0.13		
Bran Flakes	0.17		
Harmony	0.24		
Smart Start	0.30		
Frosted Flakes	0.39		
Lucky Charms	0.43		
Trix	0.43		
Cap'n Crunch	0.44		
Corn Pops	0.45		
Cocoa Puffs	0.47		
Fruit Loops	0.47		
Apple Jacks	0.48		
Mean ($\bar{x}$):	0.30	**$\Sigma(x_i - \bar{x})^2$:**	
Count (n):			
Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1			
Standard Deviation (s): ($\sqrt{s^2}$)			

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

File Edit View Insert Format Tools Data Window Help

Type a question for help

100%

B19 =COUNT(B2:B17)

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03		
Corn Flakes	0.07		
Rice Krispies	0.09		
Wheaties	0.13		
Special K	0.13		
Bran Flakes	0.17		
Harmony	0.24		
Smart Start	0.30		
Frosted Flakes	0.39		
Lucky Charms	0.43		
Trix	0.43		
Cap'n Crunch	0.44		
Corn Pops	0.45		
Cocoa Puffs	0.47		
Fruit Loops	0.47		
Apple Jacks	0.48		
Mean ($\bar{x}$):	0.30	**$\Sigma(x_i - \bar{x})^2$:**	
Count (n):	16		
Variance (s^2): $\Sigma(x_i - \bar{x})^2 / n-1$			
Standard Deviation (s): ($\sqrt{s^2}$)			

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

C2 =B2-$B$18

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03	-0.27	
Corn Flakes	0.07		
Rice Krispies	0.09		
Wheaties	0.13		
Special K	0.13		
Bran Flakes	0.17		
Harmony	0.24		
Smart Start	0.30		
Frosted Flakes	0.39		
Lucky Charms	0.43		
Trix	0.43		
Cap'n Crunch	0.44		
Corn Pops	0.45		
Cocoa Puffs	0.47		
Fruit Loops	0.47		
Apple Jacks	0.48		
Mean ($\bar{x}$):	**0.30**	**$\Sigma(x_i - \bar{x})^2$:**	
Count (n):	**16**		

Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1

Standard Deviation (s): ($\sqrt{s^2}$)

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

Stat221Ne... | Earn your d... | 221CHap3nb | 221NewBo... | 221Chapte... | DataSetsF... | 5:02

C2 =B2-$B$18

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03	-0.27	
Corn Flakes	0.07	-0.23	
Rice Krispies	0.09	-0.21	
Wheaties	0.13	-0.17	
Special K	0.13	-0.17	
Bran Flakes	0.17	-0.13	
Harmony	0.24	-0.05	
Smart Start	0.30	0.01	
Frosted Flakes	0.39	0.10	
Lucky Charms	0.43	0.14	
Trix	0.43	0.14	
Cap'n Crunch	0.44	0.15	
Corn Pops	0.45	0.16	
Cocoa Puffs	0.47	0.18	
Fruit Loops	0.47	0.18	
Apple Jacks	0.48	0.19	
Mean ($\bar{x}$):	0.30	**$\Sigma(x_i - \bar{x})^2$:**	
Count (n):	16		

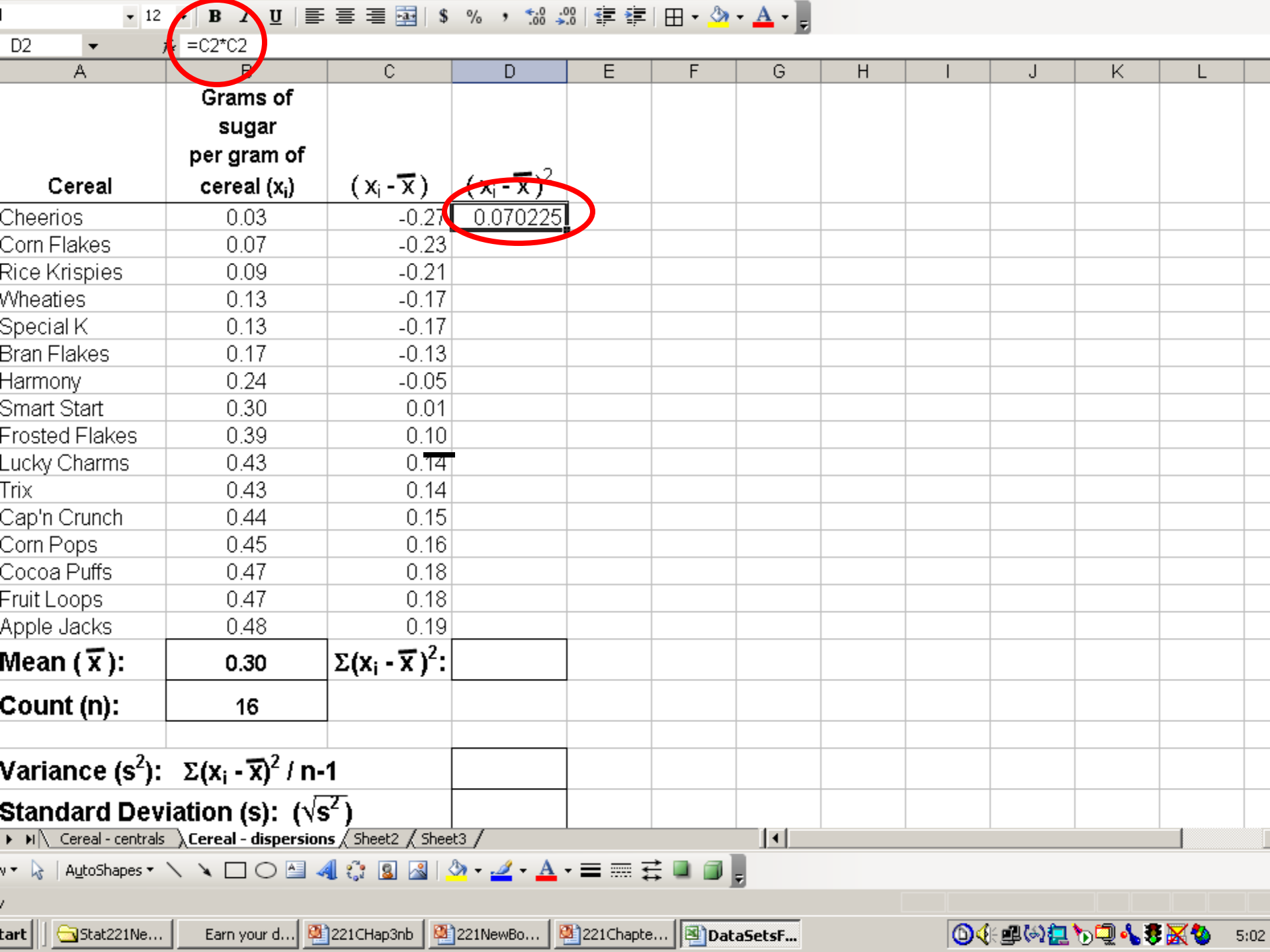
Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1

Standard Deviation (s): ($\sqrt{s^2}$)

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

Sum=0.00

Stat221Ne... | Earn your d... | 221CHap3nb | 221NewBo... | 221Chapte... | DataSetsF... | 5:02



D2 =C2*C2

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03	-0.27	0.070225
Corn Flakes	0.07	-0.23	
Rice Krispies	0.09	-0.21	
Wheaties	0.13	-0.17	
Special K	0.13	-0.17	
Bran Flakes	0.17	-0.13	
Harmony	0.24	-0.05	
Smart Start	0.30	0.01	
Frosted Flakes	0.39	0.10	
Lucky Charms	0.43	0.14	
Trix	0.43	0.14	
Cap'n Crunch	0.44	0.15	
Corn Pops	0.45	0.16	
Cocoa Puffs	0.47	0.18	
Fruit Loops	0.47	0.18	
Apple Jacks	0.48	0.19	
Mean ($\bar{x}$):	**0.30**	**$\Sigma(x_i - \bar{x})^2$:**	
Count (n):	**16**		

Variance (s^2): $\Sigma(x_i - \bar{x})^2 / n-1$

Standard Deviation (s): $(\sqrt{s^2})$

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

File Edit View Insert Format Tools Data Window Help

Type a question for help

D2 =C2*C2

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03	-0.27	0.070225
Corn Flakes	0.07	-0.23	0.050625
Rice Krispies	0.09	-0.21	0.042025
Wheaties	0.13	-0.17	0.027225
Special K	0.13	-0.17	0.027225
Bran Flakes	0.17	-0.13	0.015625
Harmony	0.24	-0.05	0.003025
Smart Start	0.30	0.01	2.5E-05
Frosted Flakes	0.39	0.10	0.009025
Lucky Charms	0.43	0.14	0.018225
Trix	0.43	0.14	0.018225
Cap'n Crunch	0.44	0.15	0.021025
Corn Pops	0.45	0.16	0.024025
Cocoa Puffs	0.47	0.18	0.030625
Fruit Loops	0.47	0.18	0.030625
Apple Jacks	0.48	0.19	0.034225
Mean ($\bar{x}$):	**0.30**	**$\Sigma(x_i - \bar{x})^2$:**	
Count (n):	**16**		

Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1

Standard Deviation (s): ($\sqrt{s^2}$)

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

Sum=0.422

D18 =SUM(D2:D17)

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03	-0.27	0.0702
Corn Flakes	0.07	-0.23	0.0506
Rice Krispies	0.09	-0.21	0.0420
Wheaties	0.13	-0.17	0.0272
Special K	0.13	-0.17	0.0272
Bran Flakes	0.17	-0.13	0.0156
Harmony	0.24	-0.05	0.0030
Smart Start	0.30	0.01	0.0000
Frosted Flakes	0.39	0.10	0.0090
Lucky Charms	0.43	0.14	0.0182
Trix	0.43	0.14	0.0182
Cap'n Crunch	0.44	0.15	0.0210
Corn Pops	0.45	0.16	0.0240
Cocoa Puffs	0.47	0.18	0.0306
Fruit Loops	0.47	0.18	0.0306
Apple Jacks	0.48	0.19	0.0342
Mean ($\bar{x}$):	0.30	**$\Sigma(x_i - \bar{x})^2$:**	0.4220
Count (n):	16		
Variance (s^2): $\Sigma(x_i - \bar{x})^2 / n-1$			
Standard Deviation (s): ($\sqrt{s^2}$)			

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

Start | Stat221Ne... | Earn your d... | 221CHap3nb | 221NewBo... | 221Chapte... | DataSetsF... | 5:0

File Edit View Insert Format Tools Data Window Help

Type a question for help

D21 =D18/(B19-1)

	A	B	C	D	E	F	G	H	I	J	K	L
	Corn Flakes	0.07	-0.23	0.0506								
	Rice Krispies	0.09	-0.21	0.0420								
	Wheaties	0.13	-0.17	0.0272								
	Special K	0.13	-0.17	0.0272								
	Bran Flakes	0.17	-0.13	0.0156								
	Harmony	0.24	-0.05	0.0030								
	Smart Start	0.30	0.01	0.0000								
	Frosted Flakes	0.39	0.10	0.0090								
	Lucky Charms	0.43	0.14	0.0182								
	Trix	0.43	0.14	0.0182								
	Cap'n Crunch	0.44	0.15	0.0210								
	Corn Pops	0.45	0.16	0.0240								
	Cocoa Puffs	0.47	0.18	0.0306								
	Fruit Loops	0.47	0.18	0.0306								
	Apple Jacks	0.48	0.19	0.0342								
17	**Mean ($\bar{x}$):**	0.30	**$\Sigma(x_i - \bar{x})^2$:**	0.4220								
18	**Count (n):**	16										
19												
20	**Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1**			0.0281								
21	**Standard Deviation (s): ($\sqrt{s^2}$)**											
22	**Coefficient of Variation: (s / $\bar{x}$) * 100**											

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

File Edit View Insert Format Tools Data Window Help

Type a question for help

D22 =SQRT(D21)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Corn Flakes	0.07	-0.23	0.0506								
2	Rice Krispies	0.09	-0.21	0.0420								
3	Wheaties	0.13	-0.17	0.0272								
4	Special K	0.13	-0.17	0.0272								
5	Bran Flakes	0.17	-0.13	0.0156								
6	Harmony	0.24	-0.05	0.0030								
7	Smart Start	0.30	0.01	0.0000								
8	Frosted Flakes	0.39	0.10	0.0090								
9	Lucky Charms	0.43	0.14	0.0182								
10	Trix	0.43	0.14	0.0182								
11	Cap'n Crunch	0.44	0.15	0.0210								
12	Corn Pops	0.45	0.16	0.0240								
13	Cocoa Puffs	0.47	0.18	0.0306								
14	Fruit Loops	0.47	0.18	0.0306								
15	Apple Jacks	0.48	0.19	0.0342								
16	**Mean ($\bar{x}$):**	0.30	**$\Sigma(x_i - \bar{x})^2$:**	0.4220								
17	**Count (n):**	16										
18												
19	**Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1**			0.0281								
20	**Standard Deviation (s): ($\sqrt{s^2}$)**			0.1677								
21	**Coefficient of Variation: (s / $\bar{x}$) * 100**											

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

D23 =(D22/B18)

	A	B	C	D	E	F	G	H	I	J	K	L
	Corn Flakes	0.07	-0.23	0.0506								
	Rice Krispies	0.09	-0.21	0.0420								
	Wheaties	0.13	-0.17	0.0272								
	Special K	0.13	-0.17	0.0272								
	Bran Flakes	0.17	-0.13	0.0156								
	Harmony	0.24	-0.05	0.0030								
	Smart Start	0.30	0.01	0.0000								
	Frosted Flakes	0.39	0.10	0.0090								
	Lucky Charms	0.43	0.14	0.0182								
	Trix	0.43	0.14	0.0182								
	Cap'n Crunch	0.44	0.15	0.0210								
	Corn Pops	0.45	0.16	0.0240								
	Cocoa Puffs	0.47	0.18	0.0306								
	Fruit Loops	0.47	0.18	0.0306								
	Apple Jacks	0.48	0.19	0.0342								
	Mean ($\bar{x}$):	0.30	**$\Sigma(x_i - \bar{x})^2$:**	0.4220								
	Count (n):	16										
	Variance (s^2): $\Sigma(x_i - \bar{x})^2$ / n-1			0.0281								
	Standard Deviation (s): ($\sqrt{s^2}$)			0.1677								
	Coefficient of Variation: (s / $\bar{x}$)			57%								

Cereal - centrals | Cereal - dispersions | Sheet2 | Sheet3

AutoShapes

Start | Stat221Ne... | Earn your d... | 221CHap3nb | 221NewBo... | 221Chapte... | DataSetsF... | 5:0

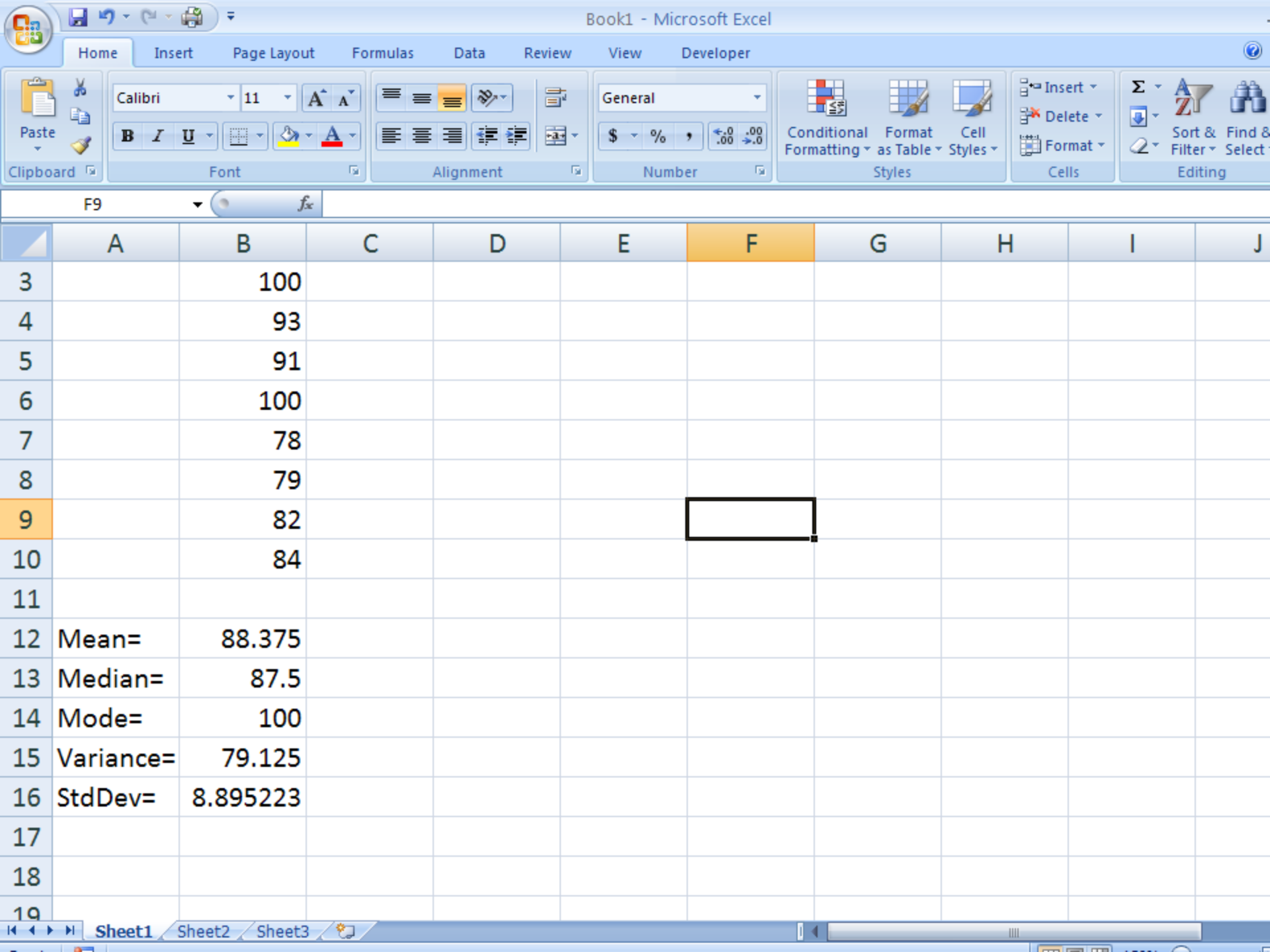
Excel's Built-in Formulas

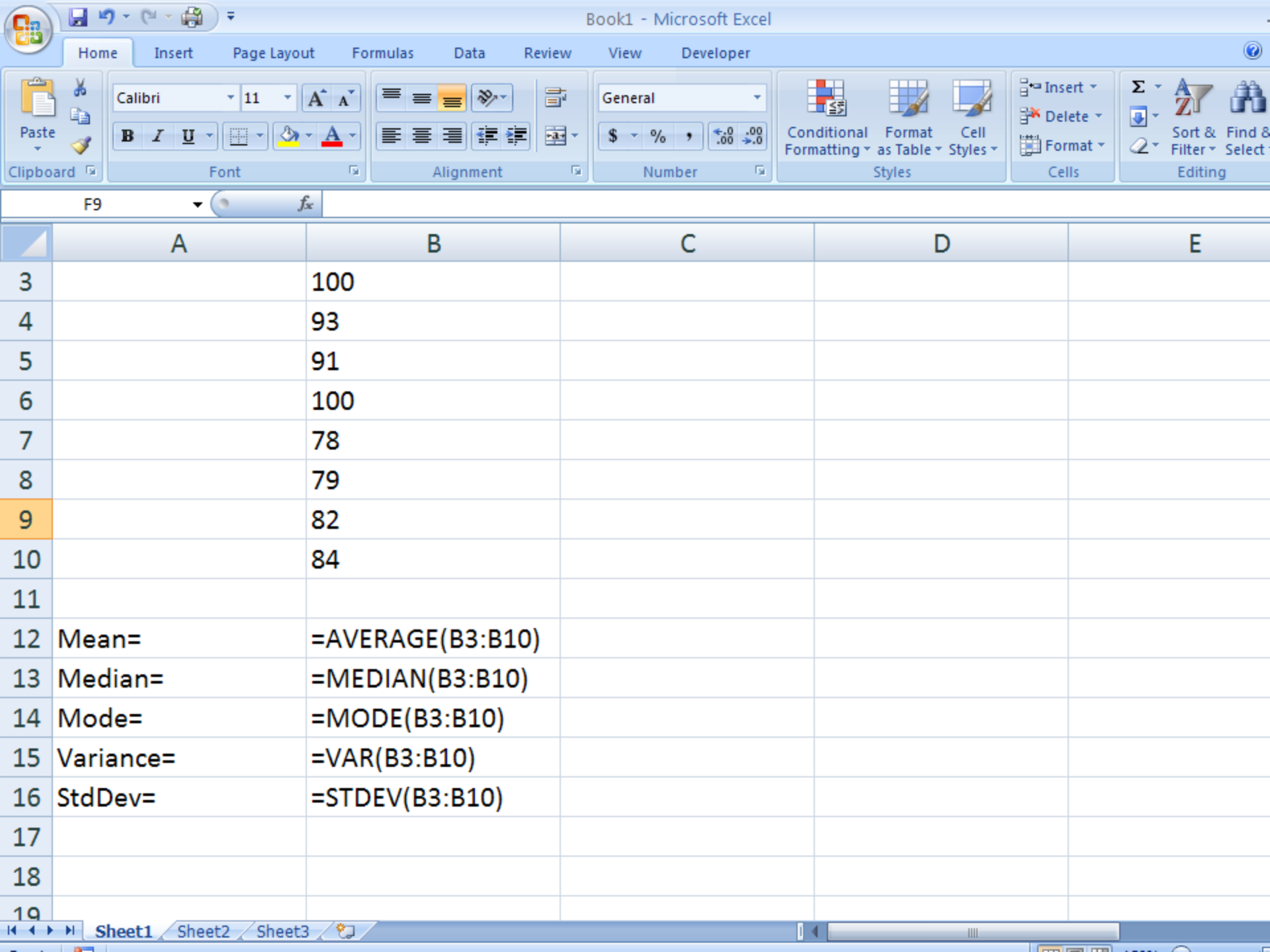
- Standard deviation of a sample:

  =stdev( )

- Variance of a sample:

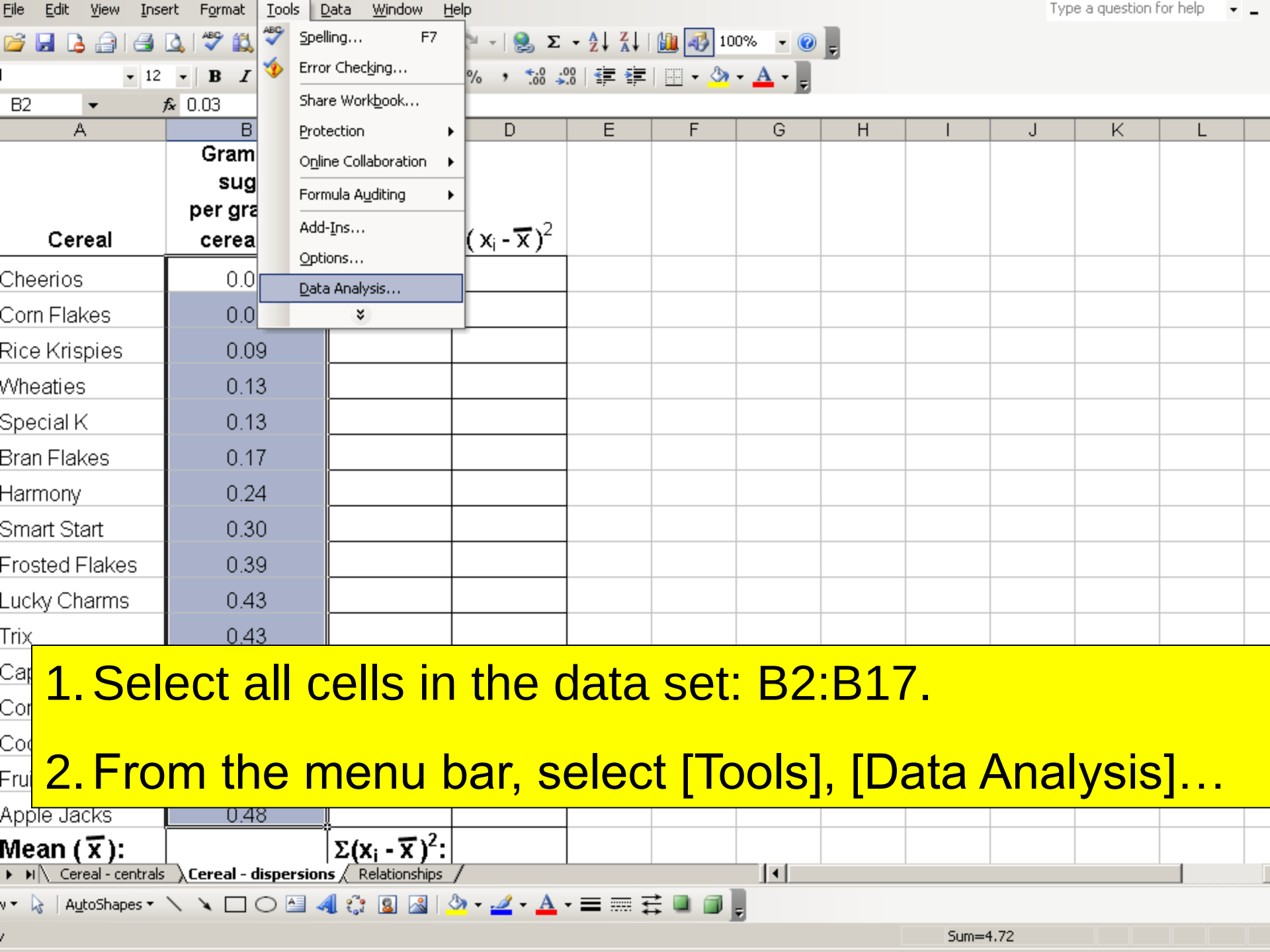
  =var( )


Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Paste
Clipboard
Calibri
11
Font
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
Sort & Filter
Find & Select
Editing
F9
fx
A
B
C
D
E
F
G
H
I
J
3 100
4 93
5 91
6 100
7 78
8 79
9 82
10 84
11
12 Mean= 88.375
13 Median= 87.5
14 Mode= 100
15 Variance= 79.125
16 StdDev= 8.895223
17
18
19
Sheet1
Sheet2
Sheet3


Book1 - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Paste
Clipboard
Calibri
11
Font
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
Sort & Filter
Find & Select
Editing
F9
fx
A
B
C
D
E
3 100
4 93
5 91
6 100
7 78
8 79
9 82
10 84
11
12 Mean= =AVERAGE(B3:B10)
13 Median= =MEDIAN(B3:B10)
14 Mode= =MODE(B3:B10)
15 Variance= =VAR(B3:B10)
16 StdDev= =STDEV(B3:B10)
17
18
19
Sheet1
Sheet2
Sheet3

Excel's Descriptive Statistics

- We can also use Excel's data analysis tool to generate a table of all the descriptive statistics


File Edit View Insert Format Tools Data Window Help
Type a question for help
Spelling... F7
Error Checking...
Share Workbook...
Protection
Online Collaboration
Formula Auditing
Add-Ins...
Options...
Data Analysis...
100%
B2
0.03
Gram sug per gra cerea
Cereal
$(x_i - \overline{x})^2$
Cheerios 0.0
Corn Flakes 0.0
Rice Krispies 0.09
Wheaties 0.13
Special K 0.13
Bran Flakes 0.17
Harmony 0.24
Smart Start 0.30
Frosted Flakes 0.39
Lucky Charms 0.43
Trix 0.43
Apple Jacks 0.48
Mean ($\overline{x}$):
$\Sigma(x_i - \overline{x})^2$:
1. Select all cells in the data set: B2:B17.
2. From the menu bar, select [Tools], [Data Analysis]…
Cereal - centrals
Cereal - dispersions
Relationships
AutoShapes
Sum=4.72

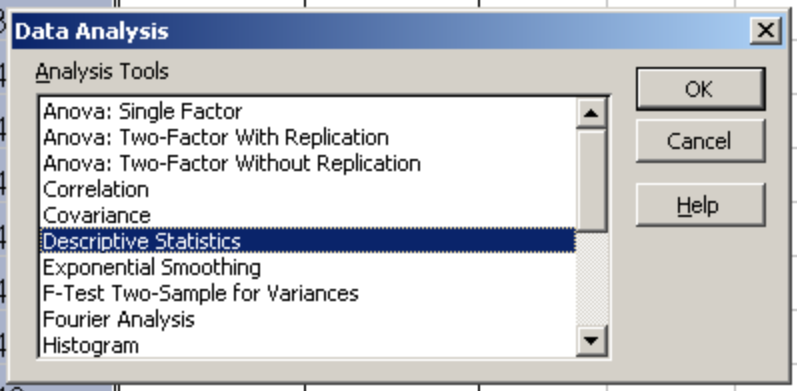
File Edit View Insert Format Tools Data Window Help

Type a question for help

B2 0.03

Cereal	Grams of sugar per gram of cereal (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Cheerios	0.03		
Corn Flakes	0.07		
Harmony	0.24		
Smart Start	0.30		
Frosted Flakes	0.3		
Lucky Charms	0.4		
Trix	0.4		
Cap'n Crunch	0.4		
Corn Pops	0.4		
Cocoa Puffs	0.4		
Fruit Loops	0.4		
Apple Jacks	0.48		
Mean ($\bar{x}$):		**$\Sigma(x_i - \bar{x})^2$:**	

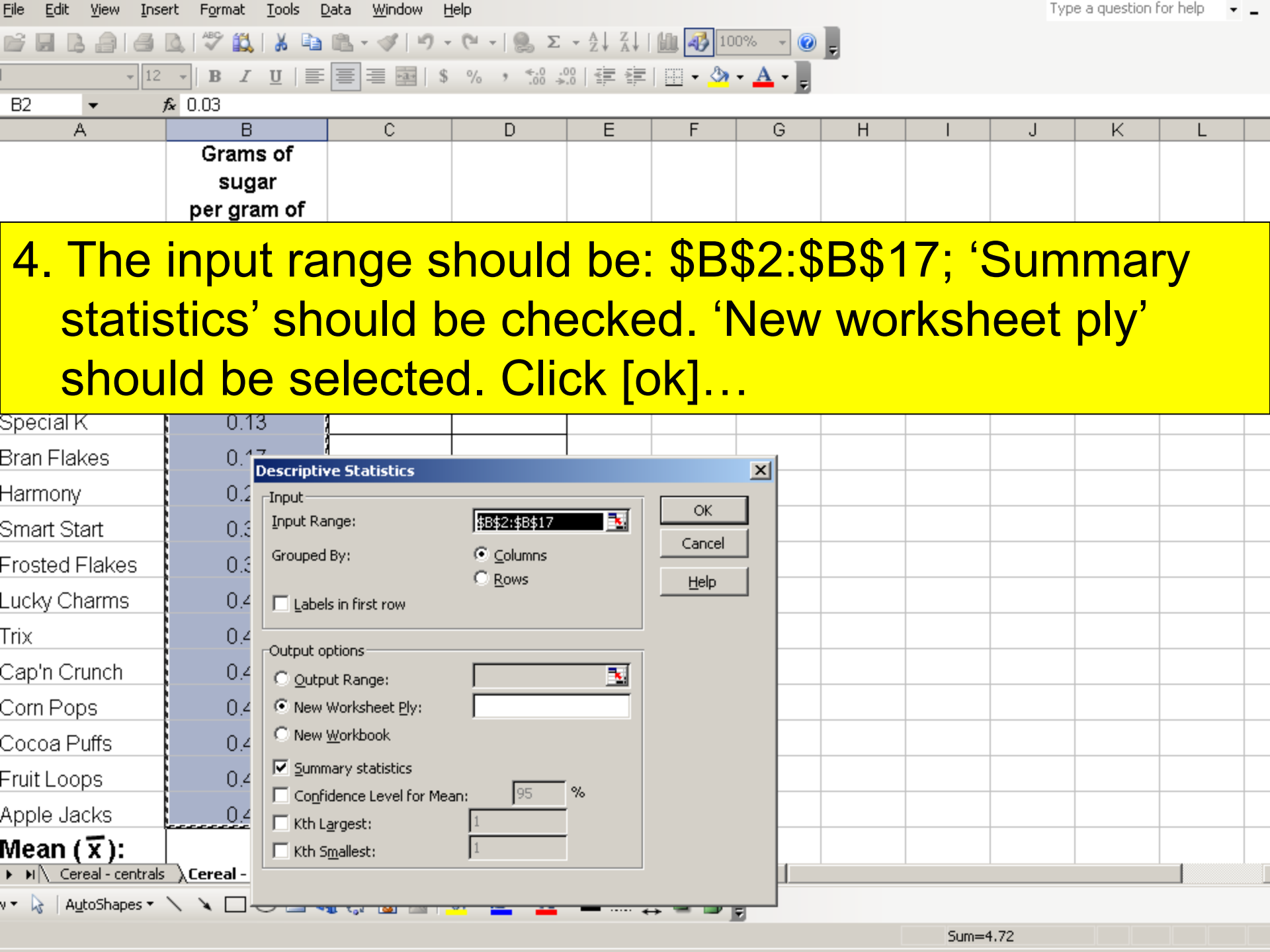
3. In the ‘data analysis’ window, select [Descriptive Statistics] and click [ok]…

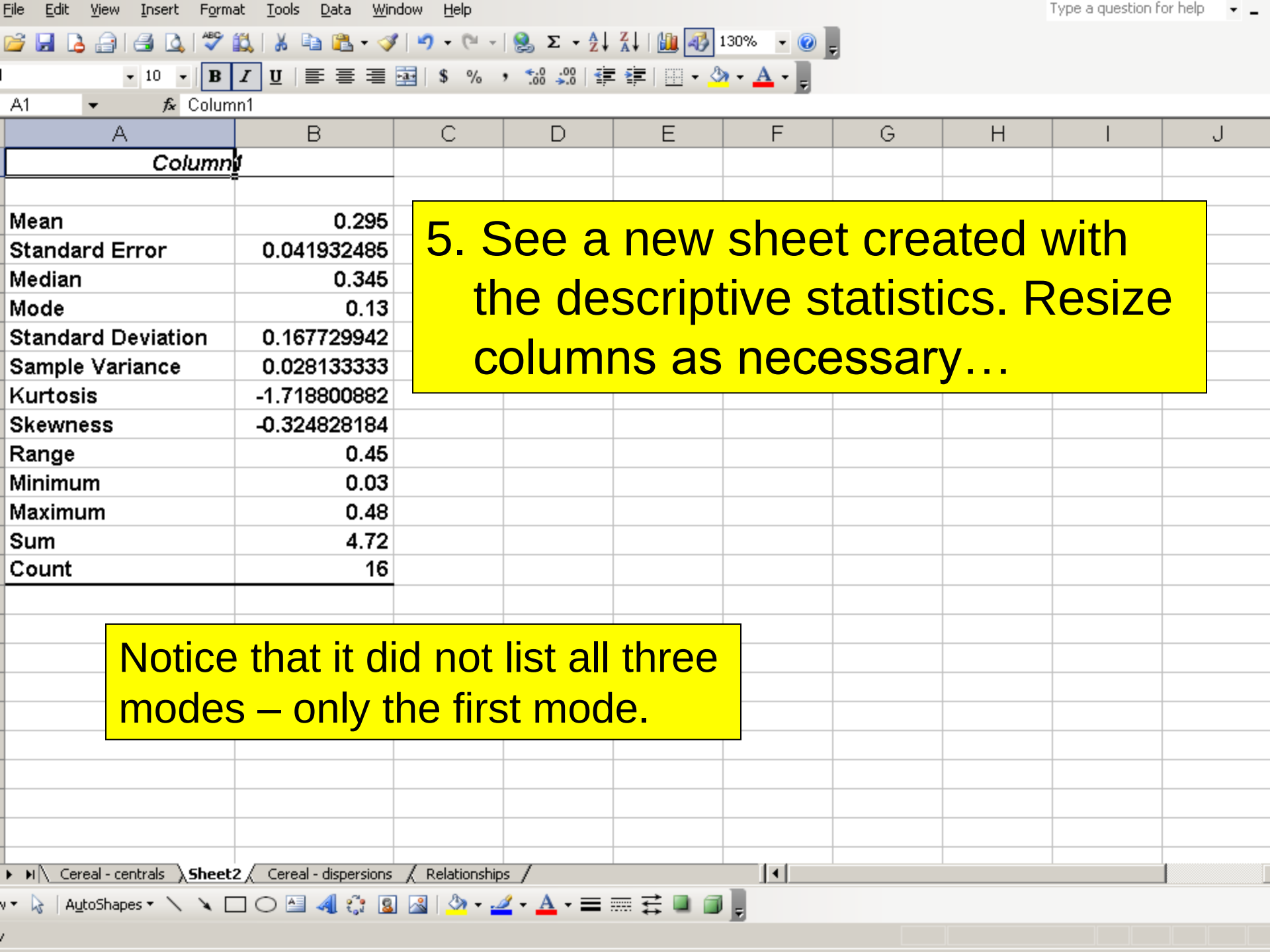


Cereal - centrals | Cereal - dispersions | Relationships

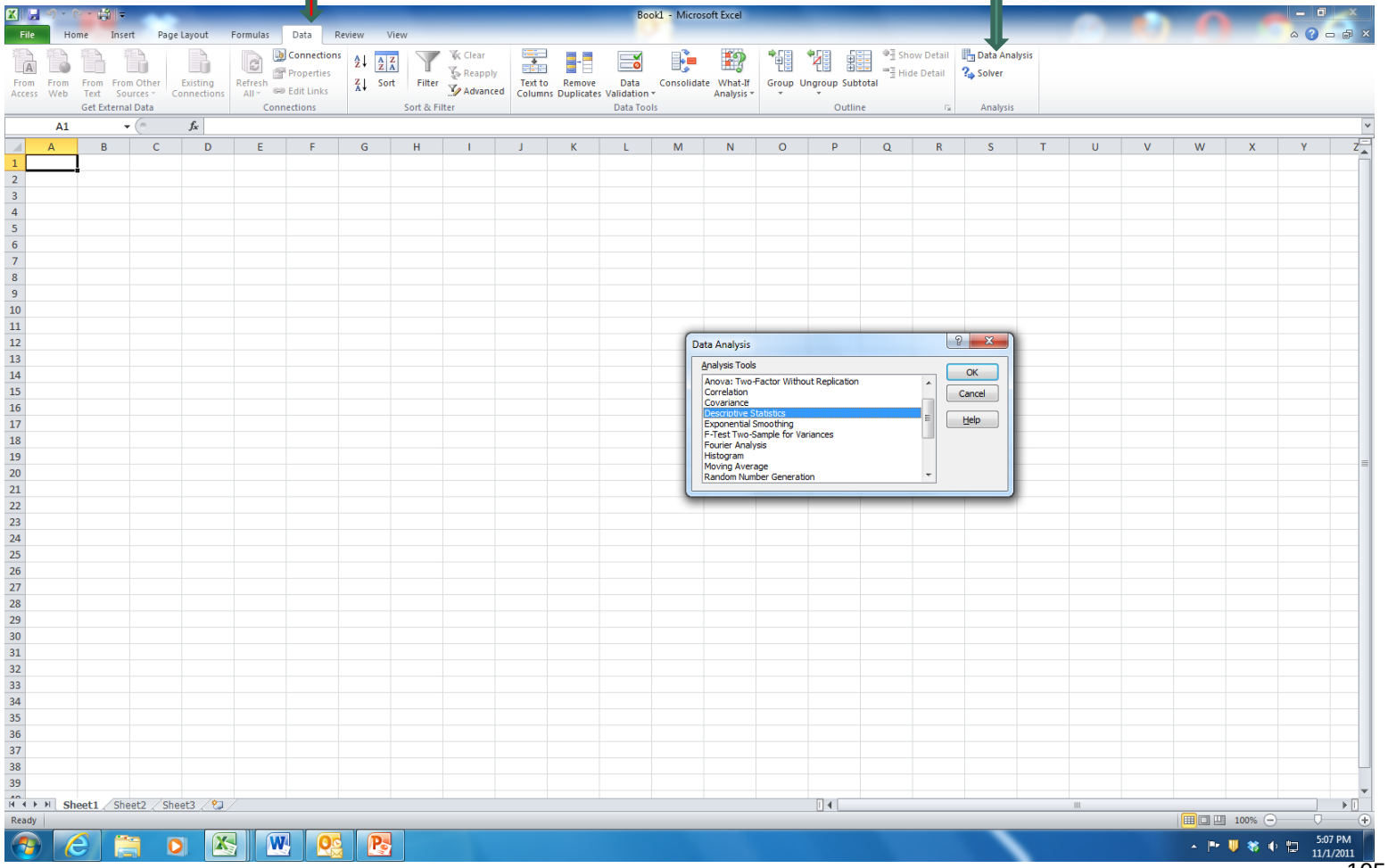
AutoShapes

Sum=4.72


File Edit View Insert Format Tools Data Window Help
Type a question for help
B2
0.03
Grams of sugar per gram of
4. The input range should be: $B$2:$B$17; ‘Summary statistics’ should be checked. ‘New worksheet ply’ should be selected. Click [ok]…
Special K 0.13
Bran Flakes
Harmony
Smart Start
Frosted Flakes
Lucky Charms
Trix
Cap'n Crunch
Corn Pops
Cocoa Puffs
Fruit Loops
Apple Jacks
Mean (x̄):
Descriptive Statistics
Input
Input Range: $B$2:$B$17
Grouped By: Columns Rows
Labels in first row
OK
Cancel
Help
Output options
Output Range:
New Worksheet Ply:
New Workbook
Summary statistics
Confidence Level for Mean: 95 %
Kth Largest: 1
Kth Smallest: 1
Cereal - centrals
Cereal -
AutoShapes
Sum=4.72


File Edit View Insert Format Tools Data Window Help
Type a question for help
130%
A1
Column1
Column1
Mean 0.295
Standard Error 0.041932485
Median 0.345
Mode 0.13
Standard Deviation 0.167729942
Sample Variance 0.028133333
Kurtosis -1.718800882
Skewness -0.324828184
Range 0.45
Minimum 0.03
Maximum 0.48
Sum 4.72
Count 16
5. See a new sheet created with the descriptive statistics. Resize columns as necessary…
Notice that it did not list all three modes – only the first mode.
Cereal - centrals
Sheet2
Cereal - dispersions
Relationships
AutoShapes

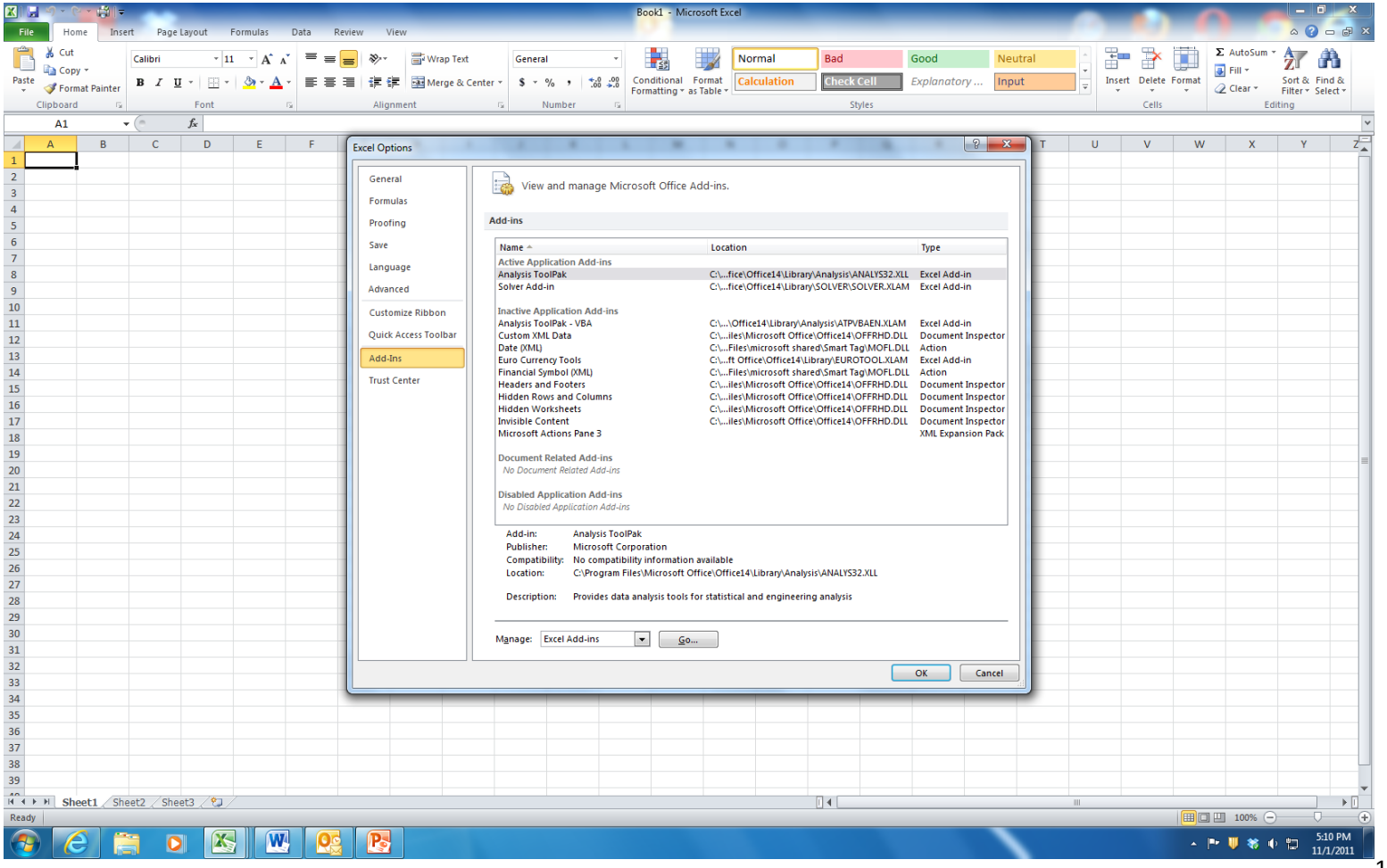
Data Analysis Tools in Excel 2007/10



Copyright Dan Brandon, PhD, PMP

Adding ToolPak to Tools Group

File → Options → Add Ins → Analysis ToolPak



Copyright Dan Brandon, PhD, PMP

Statistics Symbols

Symbol	Text Equivalent	Meaning	Formula	Link to Glossary (if appropriate)
α	Alpha	Type I error or Level of Significance.	α = P [Rejecting the null hypothesis | Null hypothesis is true].	Hypothesis Testing
β	Beta	Type II error or Power of the test.	β = P [Accepting the null hypothesis | Null hypothesis is False].	Hypothesis Testing
ϵ	Epsilon	"Error Term" in regression/statistics; more generally used to denote an arbitrarily small positive number	$y = \beta_0 + \beta_1 * x + \epsilon$	Regression
χ^2	Chi-square	Chi-square distribution	χ^2 = Sum of n independent Standard normal variables	Chi-square distribution.
χ^2	Chi-square	Chi-square distribution	$\chi^2 = \sum \frac{(O-E)^2}{E}$ where O is the observed frequency and E is the expected frequency. Or $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ (?)	Goodness of fit test
$\Gamma(n)$	Gamma-n	Gamma function	$\Gamma(n)$ = (n-1) !	
λ	Lambda	Parameter used for Poisson distribution	λ = Mean of Poisson distribution	Poisson distribution

Statistics Symbols (con't)

Symbol	Text Equivalent	Meaning	Formula	Link to Glossary (if appropriate)
μ	Mu	Arithmetic mean or Average of the population.	$\mu = \frac{\sum x}{N}$ $\mu = E(x) = \sum xP(x)$	
μ_r	Mu-r	r^{th} central moment	$\mu_r = E[(X - \mu)^r]$	Measures of central tendency.
μ_r'	Mu-r-dash	r^{th} Raw moment	$\mu_r' = E(X^r)$	Measures of central tendency.
ρ	Rho	Population correlation coefficient	$\rho = \frac{Co\,\mathrm{var}\,iamvce(X,Y)}{SD(X) * SD(Y)}$	
$\sum$	Sigma	Summation	$\sum x$ = Sum of x scores.	
σ	Sigma	Population Standard Deviation	$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ $\sigma = \sqrt{E[(x-\mu)^2]} = \sqrt{\sum (x-\mu)^2 P(x)}$	Measures of dispersion
σ^2	Sigma square	Population variance	$\sigma^2 = \frac{\Sigma (x-\mu)^2}{N}$	Measures of dispersion

Statistics Symbols (con't)

Symbol	Text Equivalent	Meaning	Formula	Link to Glossary (if appropriate)
a		Y- intercept of least square regression line	$a = \bar{y} - b\bar{x}$, for line y = a + bx	Regression: y on x
b		Slope of least squares regression line	$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ for line y = a + bx	Regression: y on x
B (n, p)	Binomial distribution with parameters n and p	Discrete probability distribution for the probability of number of successes in n independent random trials under the identical conditions.	If X follows B (n, p) then, $P(X = r) = {}^{n}C_{r} p^{r}(1-p)^{n-r}$, Where, $0 < p < 1$, $r = 0,1,2, \ldots n$	Binomial Distribution
c		Confidence level	$c = P(-z_c < Normal(0,1) < z_c)$	Confidence interval
${}^{n}C_{r}$	n-c-r	Combinations (number of combinations of n objects taken r at a time)	${}^{n}C_{r} = \frac{n!}{r!(n-r)!}$, where n ≥ r	
$C_{n,r}$	n-c-r	Combinations (number of combinations of n objects taken r at a time)	$C_{n,r} = \frac{n!}{r!(n-r)!}$, where n ≥ r	
Cov (X, Y)	Covariance between X and Y	Covariance between X & Y	Cov (X) =E [(X-E (X))(Y- E (Y)]	
CV		Coefficient of variation	$CV = \frac{Standard\ Deviation}{Arithmatic\ mean}$.	
df		Degree(s) of freedom		

Statistics Symbols (con't)

Symbol	Text Equivalent	Meaning	Formula	Link to Glossary (if appropriate)
E		Maximal error tolerance	$E = z_c \frac{\sigma}{\sqrt{n}}$ for large samples.	
E (f (x))	Expected value of f (x)		E (f (x)) = $\sum f(x)P(x)$	
f		Frequency	f = number of times score.	
F		F-distribution variable	F= $\frac{\chi_1^2 / n_1}{\chi_2^2 / n_2}$ where n_1 and n_2 are the corresponding degrees of freedom.	F-distribution, Hypothesis testing for equality of 2 variances.
F (x) or F_x		Distribution function	$F_x = \int_{-\infty}^{x} f_x dx$	
f (x) or f_x		Probability mass function	Depends on the distribution. $f_x \geq 0$ & $\int_x f_x dx = 1$.	
H_0	H-naught	Null hypothesis.	The null hypothesis is the hypothesis about the population parameter.	Testing of hypothesis
H_1	H-one	Alternate hypothesis.	An alternate hypothesis is constructed in such a way that it is the one to be accepted when the null hypothesis must be rejected.	Testing of hypothesis
IQR		Interquartile range	IQR = Q_3 - Q_1	Measures of central tendency.
MS	M-S	Mean square	MS= $\frac{SS}{df}$	Analysis of variance (ANOVA)
n		Sample size.	n = number of units in a sample.	
N		Population size	N = Number of units in the population.	

Statistics Symbols (con't)

Symbol	Text Equivalent	Meaning	Formula	Link to Glossary (if appropriate)
$P_{n,r}$	n-p-r	Permutation (number of ways to arrange **in order** n distinct objects taking them r at a time)	$P_{n,r} = \frac{n!}{(n-r)!}$, where n ≥ r	
${}_nP_r$	n-p-r	Permutation (number of ways to arrange **in order** n distinct objects taking them r at a time)	${}_nP_r = \frac{n!}{(n-r)!}$, where n ≥ r	
$\hat{p}$	p-hat	Sample proportion	$\hat{p} = \frac{number\ of\ success}{number\ of\ trials}$.	Binomial distribution
P (A | B)	Probability of A given B	Conditional probability	$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$	
P (x)	Probability of x	Probability of x	$P(x) = \frac{No.of\ favorable\ outcomes}{Total\ no.of\ outcomes}$	
p-value		The attained level of significance.	P value is the smallest level of significance for which the observed sample statistic tells us to reject the null hypothesis.	
Q		Probability of not happening of the event	$q = 1 - p$	
Q_1	Q-one	First quartile	Q_1 = Median of the lower half of the data that is data below median.	Measures of central tendency
Q_2	Q-two	Second quartile Or Median	Q_2 = Central value of an ordered data.	Measures of central tendency
Q_3	Q-three	Third quartile	Q_3 = Median of the upper half of the data that is data above the median.	Measures of central tendency

Statistics Symbols (con't)

Symbol	Text Equivalent	Meaning	Formula	Link to Glossary (if appropriate)
R		Sample Correlation coefficient	$r = \frac{Covariance(X,Y)}{[SD(X)]*[SD(Y)]}$	
r^2	r-square	Coefficient of determination	r^2 - *(Correlation coefficient)*^{2}	
R^2	r-square	Multiple correlation coefficient	$R^2 = 1 - \frac{mean\ square\ error}{S_y^2}$	
S		Sample standard deviation	$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ for ungrouped data. $s = \sqrt{\frac{\sum f(x-\bar{x})^2}{(\sum f)-1}}$ for grouped data.	Measures of dispersion
S^2	S-square	Sample variance	$S^2 = \frac{\sum (x-\bar{x})^2}{n-1}$ for ungrouped data. $S^2 = \frac{\sum f(x-\bar{x})^2}{(\sum f)-1}$ for grouped data	Measures of dispersion
S_e^2	s-e- square	Error variance	$S_e^2 = \frac{sum\ of\ squares\ of\ residuals}{n}$	
SD		Sample standard deviation	$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ for ungrouped data. $s = \sqrt{\frac{\sum f(x-\bar{x})^2}{(\sum f)-1}}$ for grouped data.	
sk_b		Bowley's coefficient of skewness	$sk_b = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$	Measures of skew ness
sk_p		Pearson's coefficient of skewness	$sk_p = \frac{Mean - Mode}{S\tan dard\ Deviation}$	Measures of skew ness

Statistics Symbols (con't)

Symbol	Text Equivalent	Meaning	Formula	Link to Glossary (if appropriate)
SS_x		Sum of Squares	$SS_x = \sum (x - \bar{x})^2$ for ungrouped data. $SS_x = \sum f(x - \bar{x})^2$ for grouped data.	
t		Student's t variable.	$t = \frac{Normal(0,1)}{\sqrt{\chi^2_n / n}}$	t-distribution
t_c	t critical	The critical value for a confidence level c.	t_c =Number such that the area under the t distribution for a given number of degrees of freedom falling between $-t_c$ and t_c is equal to c.	Testing of hypothesis
Var (X)	Variance of X	Variance of X	Var (X) = E (X- μ) ^{2}	
x		Independent variable or explanatory variable in regression analysis	Eg. In the study of, yield obtained & the irrigation level, independent variable is, X= Irrigation level.	
$\bar{x}$	x-bar	Arithmetic mean or Average of X scores.	$\bar{x} = \frac{\sum x}{n}$ for ungrouped data. $\bar{x} = \frac{\sum fx}{\sum f}$ for grouped data.	Measures of central tendency
y		Dependent variable or response variable in regression analysis	Eg. In the study of, yield obtained & the irrigation level, dependent variable is, Y= Yield obtained.	
Z	Z-score	Standard normal variable (Normal variable with mean = 0 & SD = 1)	$z = \frac{x - \mu}{\sigma}$, where X follows Normal (μ, σ).	Standard normal distribution
z_c	z critical	The critical value for a confidence level c.	z_c = Number such that the area under the standard normal curve falling between $-z_c$ and z_c is equal to c.	Testing of hypothesis Confidence interval

Key Statistics Symbols

Variable	Population	Sample
Mean	μ	$\bar{X}$
Proportion	Π	p
Variance	σ^2	s^2
Standard deviation	σ	s
Size	N	n
Standard error of the mean	$\sigma_{\bar{x}}$	$S_{\bar{x}}$
Standard error of the proportion	σ_p	S_p

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dan Brandon. Memphis TN USA

References

- The Complete Idiot's Guide to Statistics, 2nd Edition by Jr., Ph.D., Robert A. Donnelly
- Statistics, 4th Edition by David Freedman, Robert Pisani, and Roger Purves
- Statistics for Dummies by Deborah J. Rumsey
- Statistics in Plain English, 2nd Edition by Timothy C. Urdan
- Head First Statistics by Dawn Griffiths

Copyright Dan Brandon, PhD, PMP

OTHER MEANS

Copyright Dan Brandon, PhD, PMP

Trimmed Mean

- Another measure, often used when extreme values are present, is the trimmed mean
- It is obtained by deleting a percentage of the smallest and largest values from a data set and then computing the mean of the remaining values
- For example, the 5% trimmed mean is obtained by removing the smallest 5% and the largest 5% of the data values and then computing the mean of the remaining values

Copyright Dan Brandon, PhD, PMP

Geometric Mean

- The profit of Company A has grown over the last three years by 10 million, 12 million, and 14 million dollars
- It is appropriate to say that it has grown by an average of 12 million dollars yearly, for which we use the arithmetic mean
- The profit of Company B has grown the over last three years by 2.5%, 3%, and 3.5%
- Here we cannot use the arithmetic mean and say that the average growth was 3%
- Why not?

Geometric Mean (con't)

- Suppose that Company B started with a 100-million-dollar profit. Three years later it will have become: $100,000,000 * 1.025 * 1.03 * 1.035 = $109,270,125
- This is less than a yearly increase of 3% would yield, since: $100,000,000 * 1.03 * 1.03 * 1.03 = $109,272,700
- Here we see that we should use the geometric mean of the growth factors 1.025, 1.03, and 1.035 (the cube root of their product – 1.02999) to find the average percentage. That is always less than the arithmetic mean would yield.
- How do you find the nth root of a number in Excel ?

Copyright Dan Brandon, PhD, PMP

Do not look ahead !

Copyright – D. Brandon, Ph.D., PMP

- Raise the number to the 1/n power:
  - X^(1/n)

Copyright Dan Brandon, PhD, PMP

Harmonic Mean

- In certain situations, the harmonic mean provides the correct notion of "average“
- For instance, if for half the *distance* of a trip you travel at 40 miles per hour and for the other half of the *distance* you travel at 60 miles per hour, then your average speed for the trip is given by the harmonic mean of 40 and 60, which is 48
- H = (2a1*a2)/(a1+a2)
- That is, the total amount of time for the trip is the same as if you traveled the entire trip at 48 miles per hour
- Note however that if you had traveled for half the *time* at one speed and the other half at another, the arithmetic mean, 50 miles per hour, would provide the correct notion of "average“

Weighted Average

- Your GPA (grade point average) is a “weighted” average, and it is weighted by the credits awarded for each class
- A weighted average is one in which different data in the data set are given different "weights“
- “Slugging Percentage” in baseball:
  - A batter's slugging percentage in baseball is computed by: SLG = (1*SI + 2*DO + 3*TR + 4*HR) / H where:
    - SLG = slugging percentage
    - SI = number of singles
    - DO = number of doubles
    - TR = number of triples
    - HR = number of home runs
    - H = total number of hits (H = SI + DO + TR + HR)
  - Here, each single has a "weight" of 1, each double has a "weight" of 2, etc. The average counts home runs four times as important as singles, and so on.

Moving Averages

- A batting average is the total number of hits divided by the total number of official at-bats
- There are two ways you can compute the batting average when the data change
- One is to do the total division over again
- The other is to use the previously computed average in the following way:
  - Let H be the total number of hits up through yesterday, and h the number of hits today, A the total number of at-bats up through yesterday, and a the number of at-bats today
  - The old batting average was H/A. The new batting average is (H+h)/(A+a) = (H/A)*(A/(A+a)) + (h/a)*(a/(A+a)). That means that you multiply the old batting average by A/(A+a), and multiply today's batting average (which is h/a) by a/(A+a), and add them together
  - Example: Suppose Mark McGwire had 23 hits in 70 at-bats through yesterday, and today he was 2-for-5. Then H = 23, A = 70, h = 2, a = 5, and H/A = 0.32857, h/a = .40000. Then A/(A+a) = 70/75 = 14/15, and a/(A+a) = 5/75 = 1/15, so the new batting average is 0.32857*14/15 + 0.40000*1/15 = 0.30667 + 0.02666 = 0.33333, and McGwire has raised his batting average by 0.00476. Or you could divide H+h = 25 by A+a = 75, and get 0.33333 for his new batting average.