Chapter 2: Descriptive Statistics and Data Analysis

Statistics, Data Analysis, and Decision Modeling, Fifth Edition

James R. Evans

Descriptive Statistics

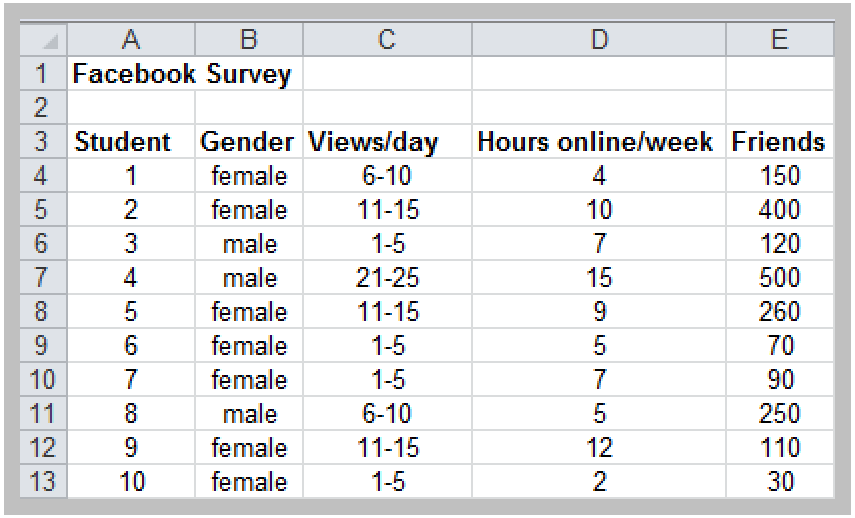
- Quantitative measures and ways of describing data.
  - *measures of central tendency* (mean, median, mode, proportion),
  - *measures of dispersion* (range, variance, standard deviation), and
  - *frequency distributions and histograms* .

Statistical Support in Excel

- Using statistical functions that are entered in worksheet cells directly or embedded in formulas.
- Using the Excel *Analysis Toolpak* add-in to perform more complex statistical computations.
- Using the *Prentice-Hall* statistics add-in, *PHStat,* to perform analyses not designed into Excel.
- See Table 2.1.

Frequency Distribution

- Tabular summary showing the frequency of observations in each of several non-overlapping classes, or cells



	A	B	C	D	E
1	Facebook Survey				
2					
3	Student	Gender	Views/day	Hours online/week	Friends
4	1	female	6-10	4	150
5	2	female	11-15	10	400
6	3	male	1-5	7	120
7	4	male	21-25	15	500
8	5	female	11-15	9	260
9	6	female	1-5	5	70
10	7	female	1-5	7	90
11	8	male	6-10	5	250
12	9	female	11-15	12	110
13	10	female	1-5	2	30

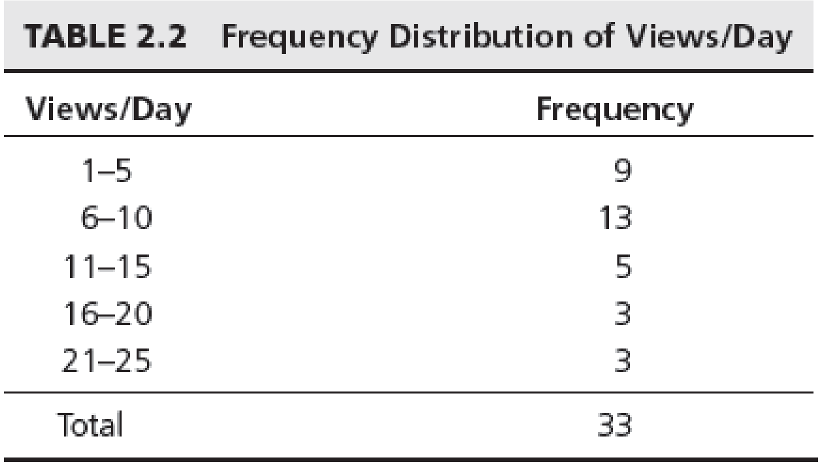


TABLE 2.2 **Frequency Distribution of Views/Day**

Views/Day	Frequency
1–5	9
6–10	13
11–15	5
16–20	3
21–25	3
Total	33

Relative Frequency Distribution

- Relative frequency – fraction or proportion of observations that fall within a cell

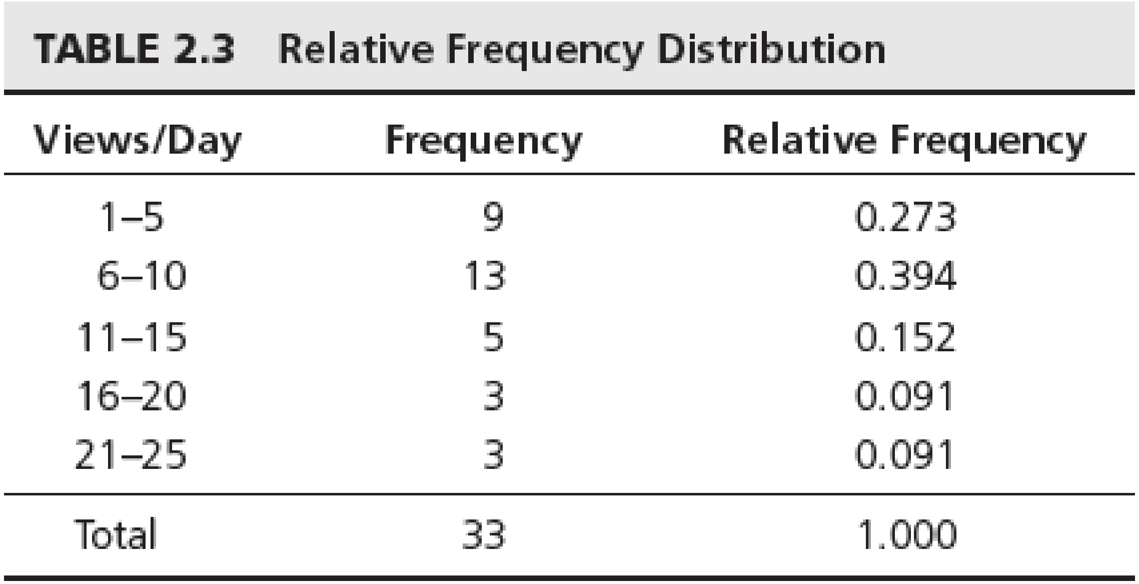
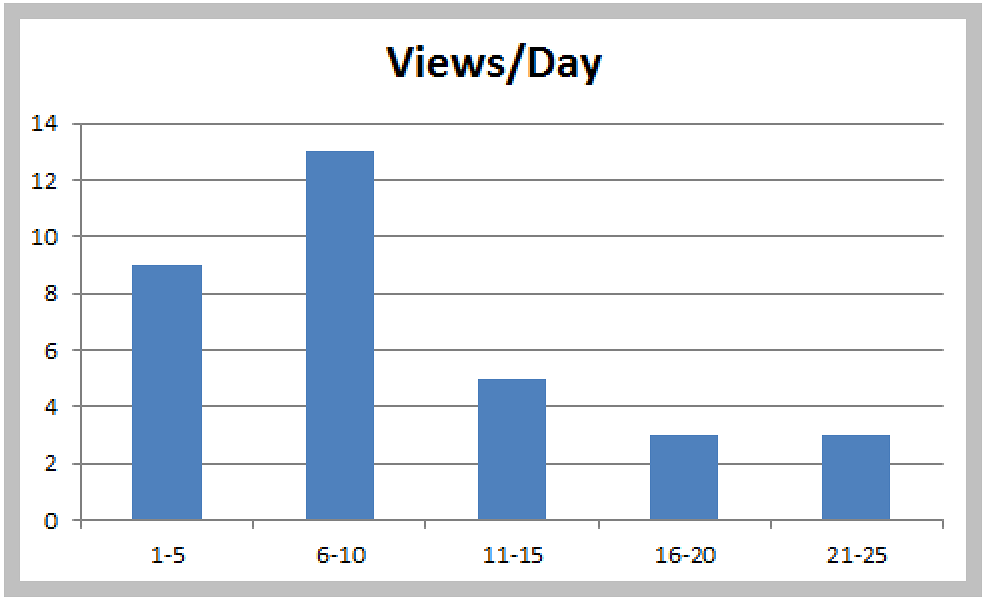


TABLE 2.3 Relative Frequency Distribution

Views/Day	Frequency	Relative Frequency
1–5	9	0.273
6–10	13	0.394
11–15	5	0.152
16–20	3	0.091
21–25	3	0.091
Total	33	1.000

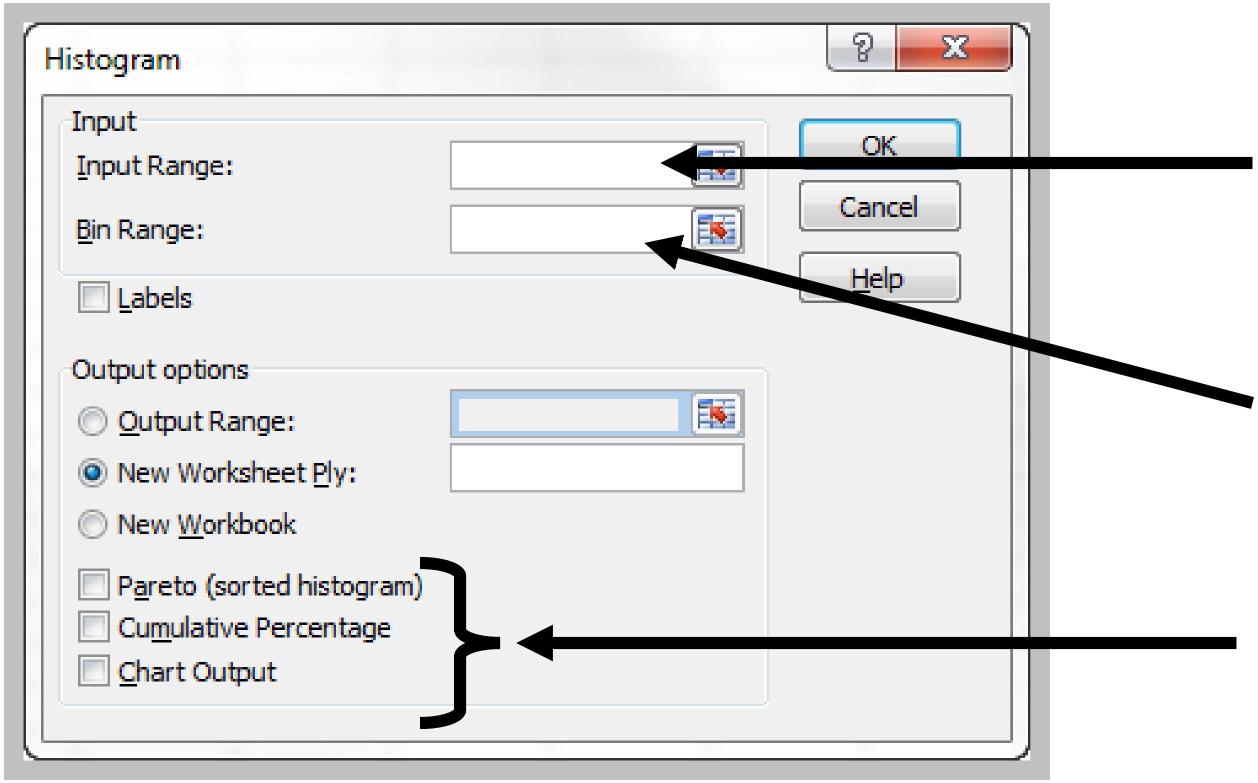
Histogram

- A graphical representation of a frequency distribution



Excel Tool: Histogram

- Excel Menu > *Tools* > *Data Analysis* > *Histogram*



Histogram

Input

Input Range:

Bin Range:

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

□ Chart Output

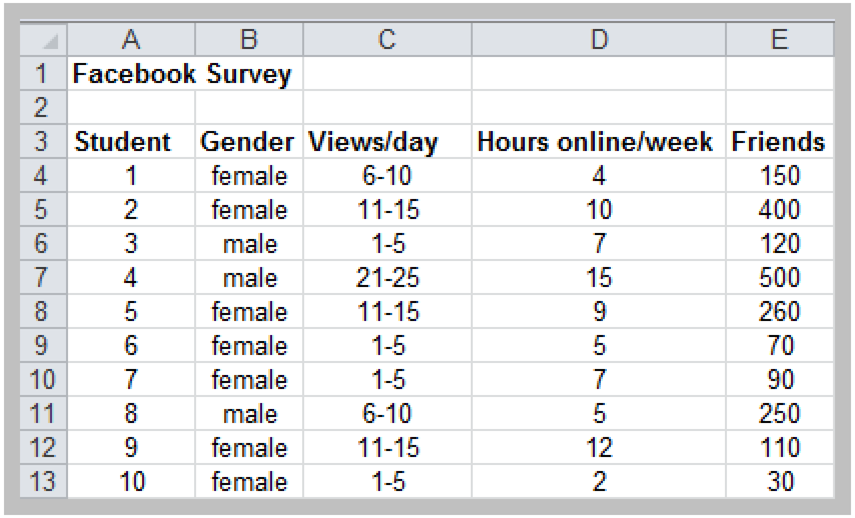
Specify range of data

Define and specify bin range (recommended)

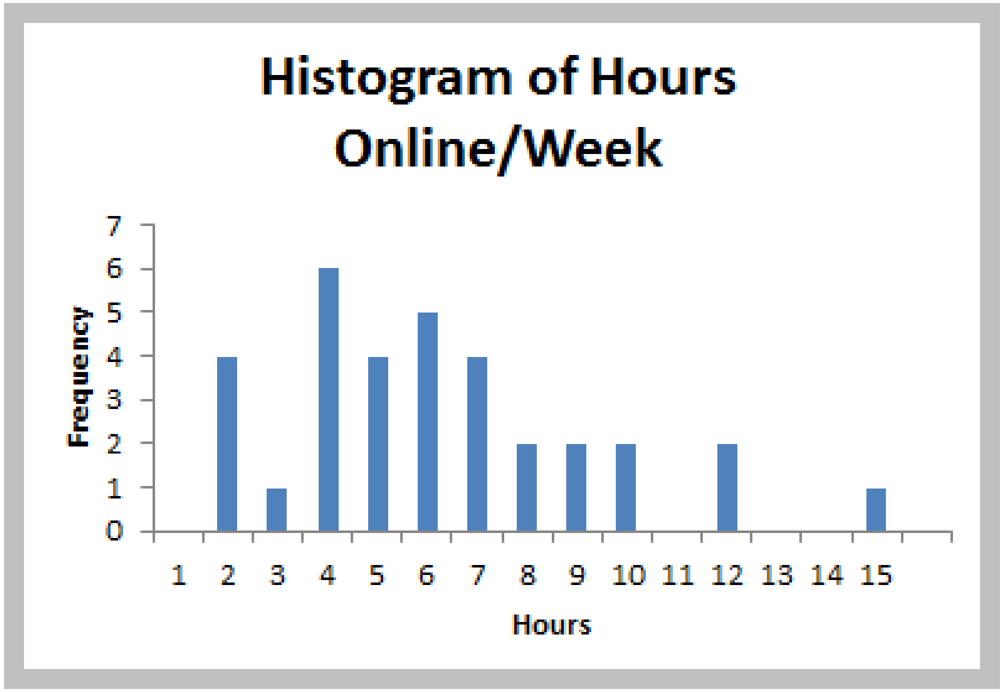
Select output options (always check Chart Output

Histograms for Numerical Data – Few Discrete Values

- Leave *Bin Range* blank in Excel dialog.

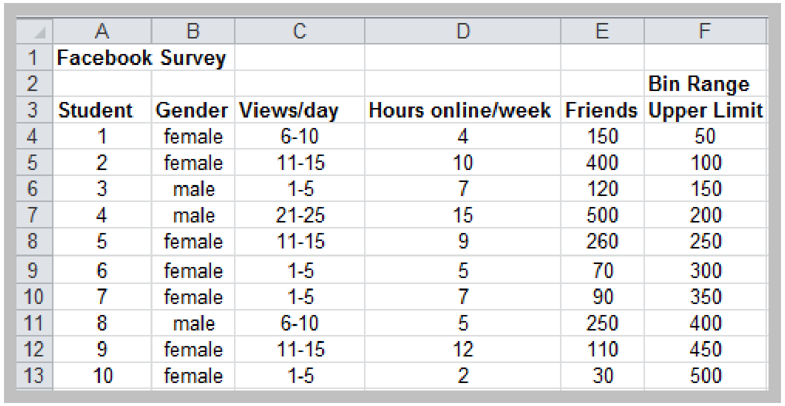


	A	B	C	D	E
1	Facebook Survey				
2					
3	Student	Gender	Views/day	Hours online/week	Friends
4	1	female	6-10	4	150
5	2	female	11-15	10	400
6	3	male	1-5	7	120
7	4	male	21-25	15	500
8	5	female	11-15	9	260
9	6	female	1-5	5	70
10	7	female	1-5	7	90
11	8	male	6-10	5	250
12	9	female	11-15	12	110
13	10	female	1-5	2	30

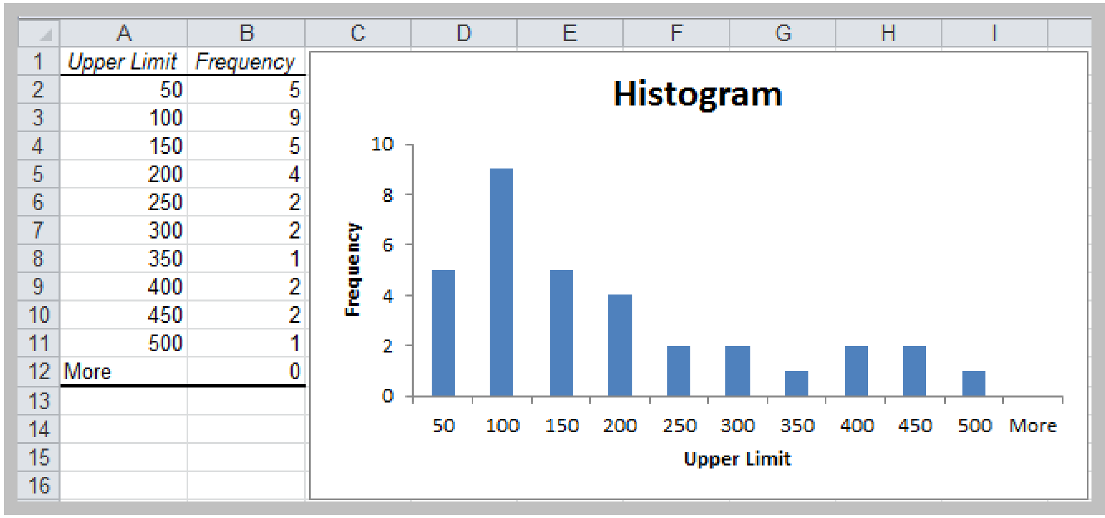


Histograms for Numerical Data – Many Discrete or Continuous Values

- Define a *Bin Range* in your spreadsheet



	A	B	C	D	E	F
1	Facebook Survey					
2						Bin Range
3	Student	Gender	Views/day	Hours online/week	Friends	Upper Limit
4	1	female	6-10	4	150	50
5	2	female	11-15	10	400	100
6	3	male	1-5	7	120	150
7	4	male	21-25	15	500	200
8	5	female	11-15	9	260	250
9	6	female	1-5	5	70	300
10	7	female	1-5	7	90	350
11	8	male	6-10	5	250	400
12	9	female	11-15	12	110	450
13	10	female	1-5	2	30	500



	A	B
1	*Upper Limit*	*Frequency*
2	50	5
3	100	9
4	150	5
5	200	4
6	250	2
7	300	2
8	350	1
9	400	2
10	450	2
11	500	1
12	More	0
13		
14		
15		
16		

Good Practice Guidelines

- Cell intervals should be of equal width.
- Choose the width using the formula

  (largest value – smallest value)/number of cells

  but round to reasonable values

  (e.g., 97   to 100)
- Choose somewhere between 5 to 15 cells to provide a useful picture of the data

Cumulative Relative Frequency

- Cumulative relative frequency – proportion or percentage of observations that fall below the upper limit of a cell

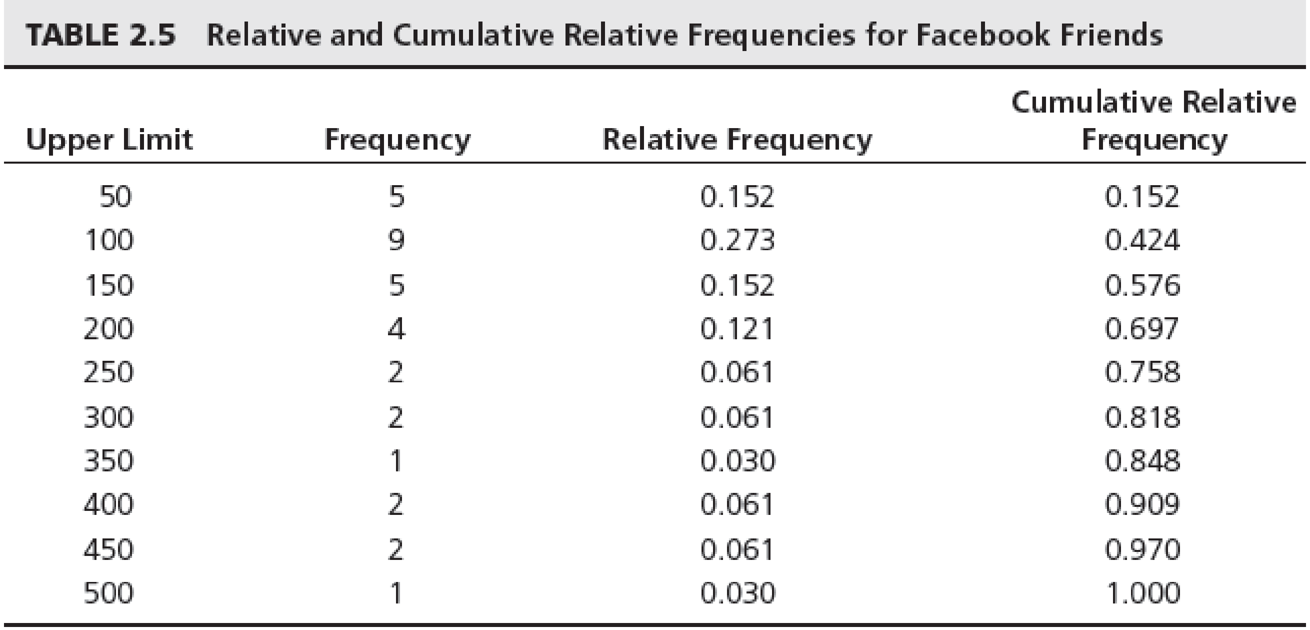
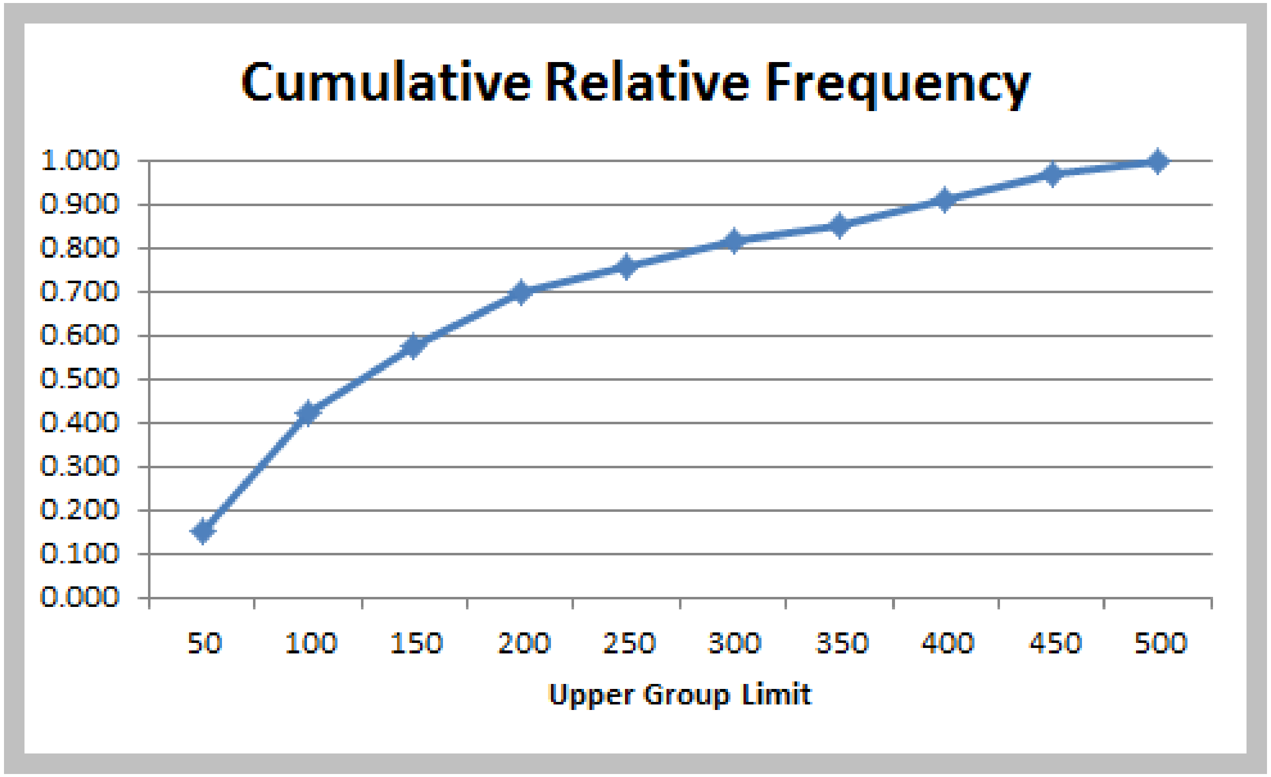


TABLE 2.5 Relative and Cumulative Relative Frequencies for Facebook Friends

Upper Limit	Frequency	Relative Frequency	Cumulative Relative Frequency
50	5	0.152	0.152
100	9	0.273	0.424
150	5	0.152	0.576
200	4	0.121	0.697
250	2	0.061	0.758
300	2	0.061	0.818
350	1	0.030	0.848
400	2	0.061	0.909
450	2	0.061	0.970
500	1	0.030	1.000

Chart of Cumulative Relative Frequency



Using Excel's Frequency Function

- Define bins
- Select a range of cells adjacent to the bin range (if continuous data, add one empty cell below this range as an overflow cell)
- Enter the formula =FREQUENCY(*range of data, range of bins*) and press *Ctrl-Shift-Enter* simultaneously*.*
- Construct a histogram using the *Chart Wizard* for a column chart.

Data Profiles (Fractiles)

- Describe the location and spread of data over its range
  - Quartiles – a division of a data set into four equal parts; shows the points below which 25%, 50%, 75% and 100% of the observations lie (25% is the first quartile, 75% is the third quartile, etc.)
  - Deciles – a division of a data set into 10 equal parts; shows the points below which 10%, 20%, etc. of the observations lie
  - Percentiles – a division of a data set into 100 equal parts; shows the points below which “k” percent of the observations lie

Descriptive Statistics for Numerical Data

- Measures of location
- Measures of dispersion
- Measures of shape
- Measures of association

Arithmetic Mean

- Population

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

- Sample

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

- Excel function AVERAGE(*data range*)

Properties of the Mean

- Meaningful for interval and ratio data
- All data used in the calculation
- Unique for every set of data
- Affected by unusually large or small observations (outliers)
- The only measure of central tendency where the sum of the deviations of each value from the measure is zero; i.e.,

$$\sum(x_i - \bar{x}) = 0$$

Median

- Middle value when data are ordered from smallest to largest. This results in an equal number of observations above the median as below it.
  - Unique for each set of data
  - Not affected by extremes
  - Meaningful for ratio, interval, and ordinal data
- Excel function MEDIAN(*data range*)

Mode

- Observation that occurs most frequently; for grouped data, the midpoint of the cell with the largest frequency (approximate value)
  - Useful when data consist of a small number of unique values
- Excel functions MODE.SNGL(*data range*) and MODE.MULT(*data range*)

Midrange

- Average of the largest and smallest observations
  - Useful for very small samples, but extreme values can distort the result

Measures of Dispersion

- Dispersion – the degree of variation in the data.
  - Example:

    {48, 49, 50, 51, 52} versus
    {10, 30, 50, 70, 90}
  - Both means are 50, but the second data set has larger dispersion

Range Measures

- Range – difference between the maximum and minimum observations
  - Useful for very small samples, but extreme values can distort the result
- Interquartile range: $Q_3 - Q_1$
  - Avoids problems with outliers

Variance

- Population

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

- Sample

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

- Excel functions VAR.P(*data range*), VAR.S(*data range*)

Standard Deviation

- Population

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

- Sample

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

- The standard deviation has the same units of measurement as the original data, unlike the variance
- Excel functions STDEV.P(*data range*), STDEV.S(*data range*)

Chebyshev’ s Theorem

- For any set of data, the proportion of values that lie within k standard deviations of the mean is at least 1 – 1/k^2, for any k > 1
  - For k = 2, at least ¾ of the data lie within 2 standard deviations of the mean
  - For k = 3, at least 8/9, or 89% lie within 3 standard deviations of the mean
  - For k = 10, at least 99/100, or 99% of the data lie within 10 standard deviations of the mean

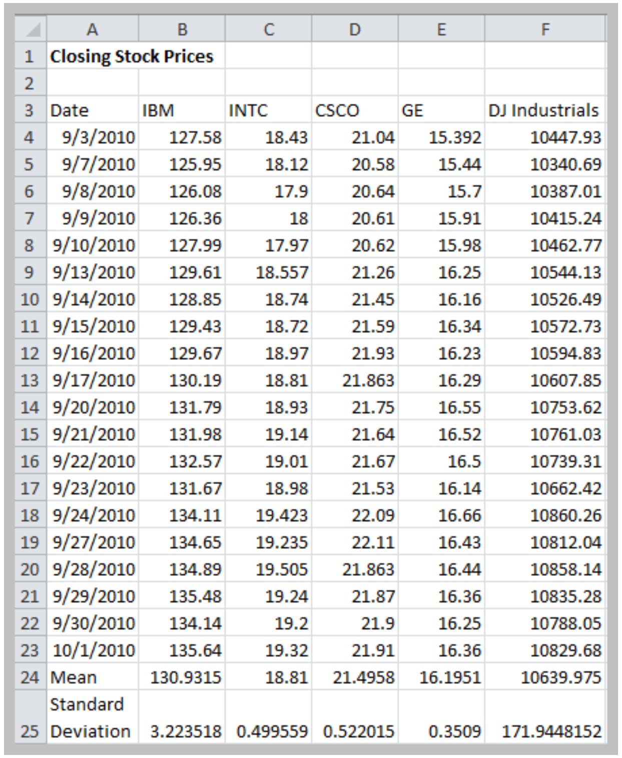
Empirical Rules

- Approximately 68% of the observations will fall within one standard deviation of the mean.
- Approximately 95% of the observations will fall within two standard deviations of the mean.
- Approximately 99.7% of the observations will fall within three standard deviations of the mean.

Coefficient of Variation

CV = Standard Deviation / Mean

- CV is dimensionless, and therefore is useful when comparing data sets that are scaled differently.



	A	B	C	D	E	F
1	**Closing Stock Prices**					
2						
3	Date	IBM	INTC	CSCO	GE	DJ Industrials
4	9/3/2010	127.58	18.43	21.04	15.392	10447.93
5	9/7/2010	125.95	18.12	20.58	15.44	10340.69
6	9/8/2010	126.08	17.9	20.64	15.7	10387.01
7	9/9/2010	126.36	18	20.61	15.91	10415.24
8	9/10/2010	127.99	17.97	20.62	15.98	10462.77
9	9/13/2010	129.61	18.557	21.26	16.25	10544.13
10	9/14/2010	128.85	18.74	21.45	16.16	10526.49
11	9/15/2010	129.43	18.72	21.59	16.34	10572.73
12	9/16/2010	129.67	18.97	21.93	16.23	10594.83
13	9/17/2010	130.19	18.81	21.863	16.29	10607.85
14	9/20/2010	131.79	18.93	21.75	16.55	10753.62
15	9/21/2010	131.98	19.14	21.64	16.52	10761.03
16	9/22/2010	132.57	19.01	21.67	16.5	10739.31
17	9/23/2010	131.67	18.98	21.53	16.14	10662.42
18	9/24/2010	134.11	19.423	22.09	16.66	10860.26
19	9/27/2010	134.65	19.235	22.11	16.43	10812.04
20	9/28/2010	134.89	19.505	21.863	16.44	10858.14
21	9/29/2010	135.48	19.24	21.87	16.36	10835.28
22	9/30/2010	134.14	19.2	21.9	16.25	10788.05
23	10/1/2010	135.64	19.32	21.91	16.36	10829.68
24	Mean	130.9315	18.81	21.4958	16.1951	10639.975
25	Standard Deviation	3.223518	0.499559	0.522015	0.3509	171.9448152

CV(IBM) = 0.025

CV(INTC) = 0.027

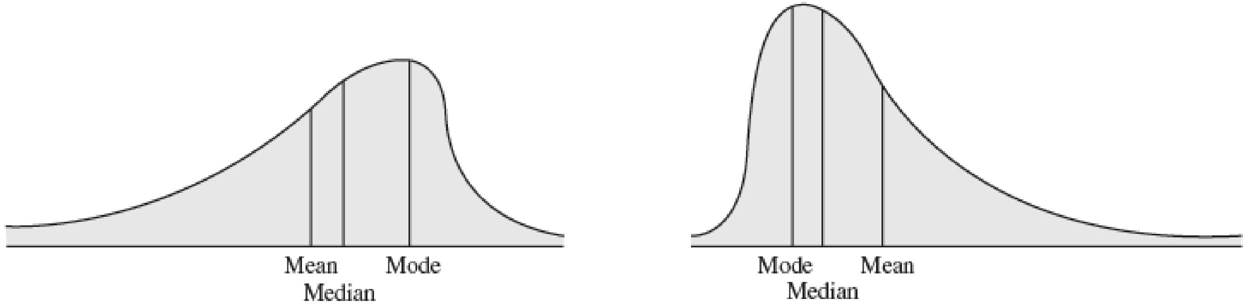
CV(CSCO) = 0.024

CV(GE) = 0.022

CV(DJI) = 0.016

Skewness

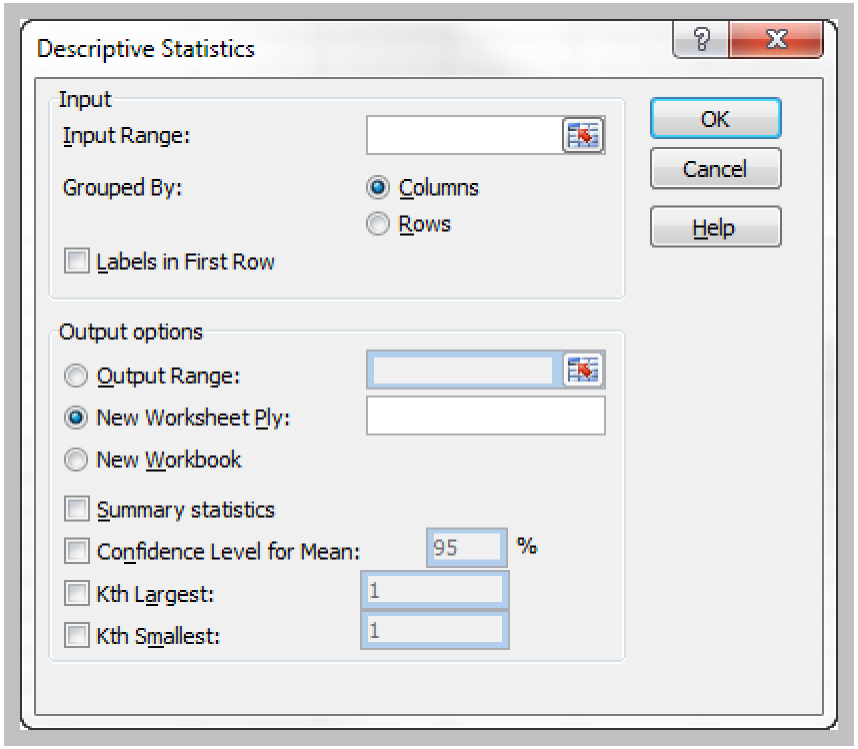
- Coefficient of skewness (CS)
  - -0.5 < CS < 0.5 indicates relative symmetry
  - CS > 1 or CS < -1 indicates a high degree of skewness
- Excel function SKEW(*data range*)



Kurtosis

- Refers to the peakedness or flatness of a distribution.
- Coefficient of kurtosis (CK)
  - CK < 3: more flat with wide degree of dispersion
  - CK >3 more peaked with less dispersion
- The higher the kurtosis, the more area in the tails of the distribution
- Excel function KURT(*data range*)

Excel *Descriptive Statistics* Tool



Descriptive Statistics

Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

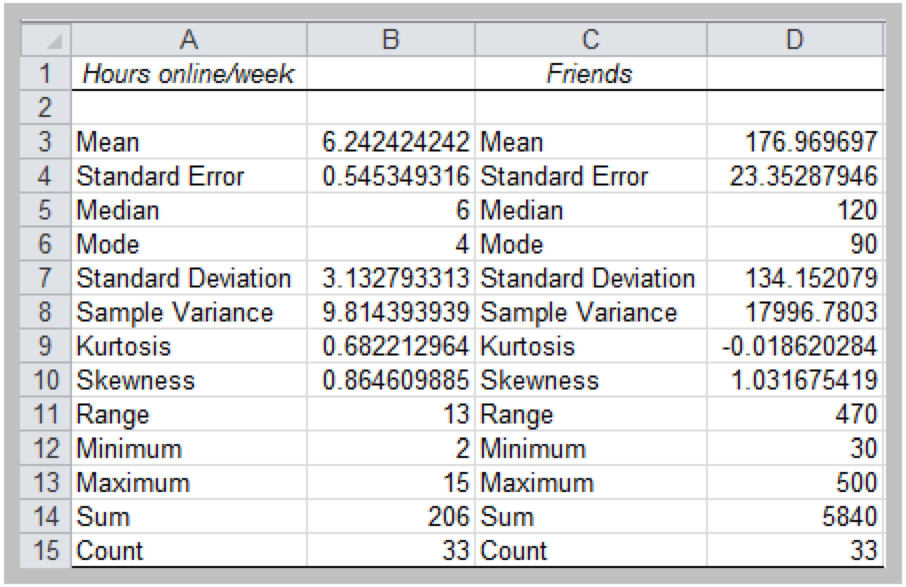
○ New Workbook

□ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

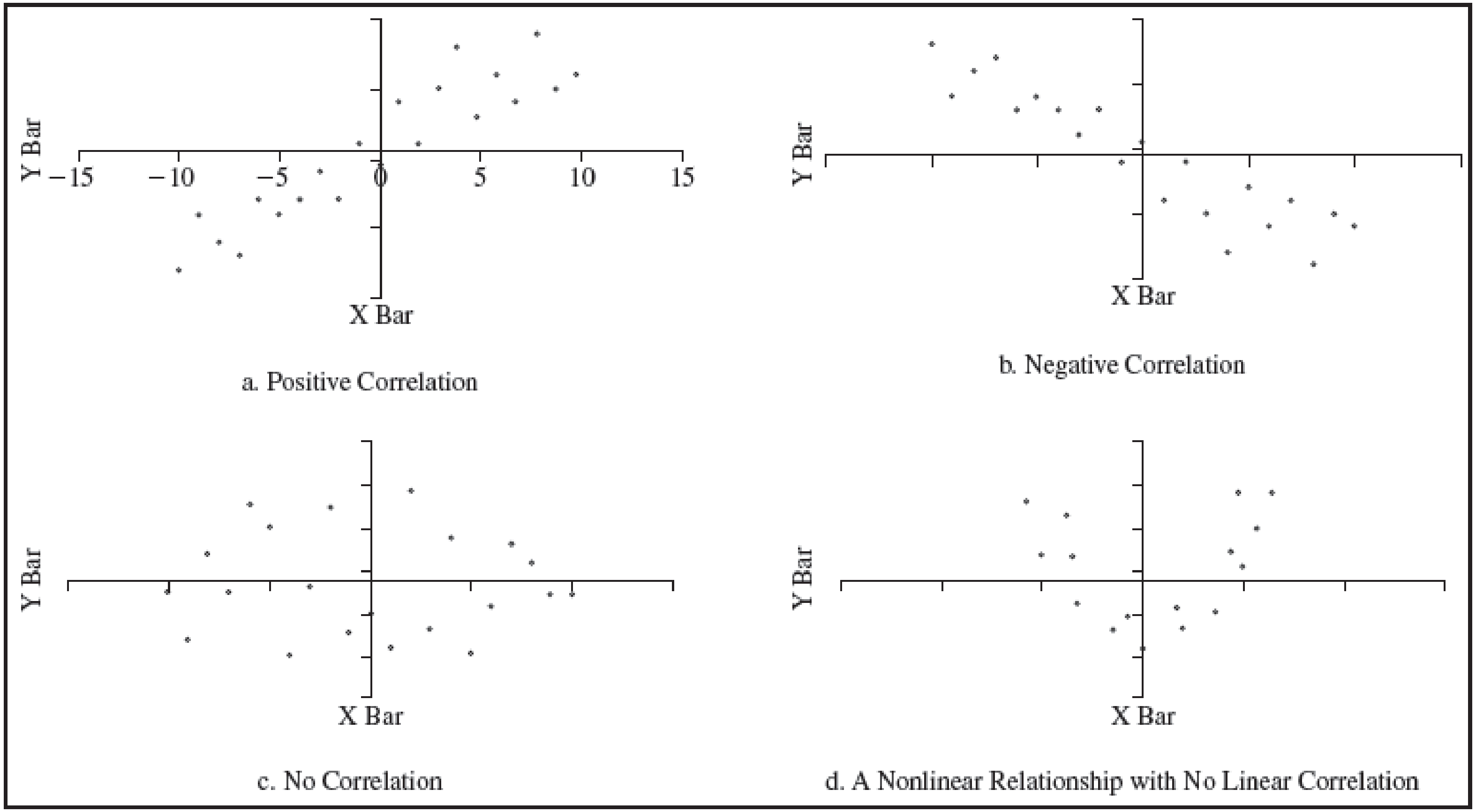


	A	B	C	D
1	*Hours online/week*		*Friends*	
2				
3	Mean	6.242424242	Mean	176.969697
4	Standard Error	0.545349316	Standard Error	23.35287946
5	Median	6	Median	120
6	Mode	4	Mode	90
7	Standard Deviation	3.132793313	Standard Deviation	134.152079
8	Sample Variance	9.814393939	Sample Variance	17996.7803
9	Kurtosis	0.682212964	Kurtosis	-0.018620284
10	Skewness	0.864609885	Skewness	1.031675419
11	Range	13	Range	470
12	Minimum	2	Minimum	30
13	Maximum	15	Maximum	500
14	Sum	206	Sum	5840
15	Count	33	Count	33

Measures of Association

- Correlation – a measure of strength of linear relationship between two variables
- Correlation coefficient – a number between -1 and 1.
  - A correlation of 0 indicates that the two variables have no linear relationship to each other.
  - A positive correlation coefficient indicates a linear relationship for which one variable increases as the other also increases.
  - A negative correlation coefficient indicates a linear relationship for one variable that increases while the other decreases.
- Excel function CORREL or *Data Analysis Correlation* tool

Examples of Correlation



a. Positive Correlation

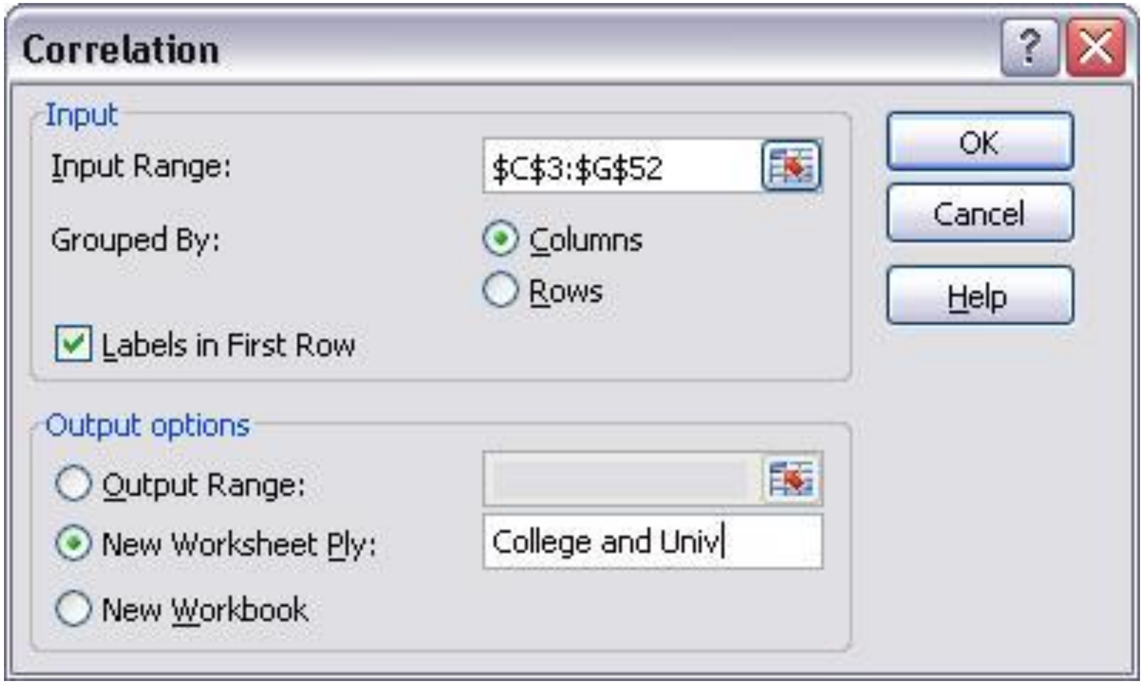
b. Negative Correlation

c. No Correlation

d. A Nonlinear Relationship with No Linear Correlation

Excel Tool: Correlation

- Excel menu > *Tools* > *Data Analysis* > *Correlation*



Input Range: $C$3:$G$52

Grouped By: ● Columns ○ Rows

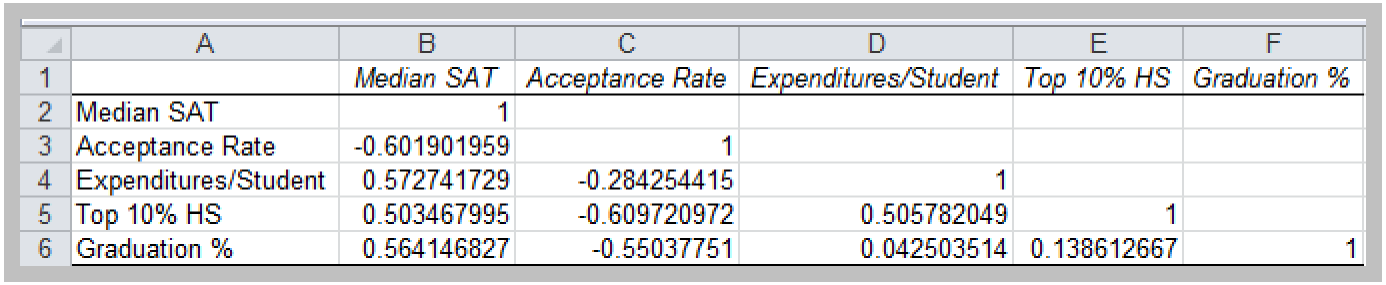
☑ Labels in First Row

○ Output Range:

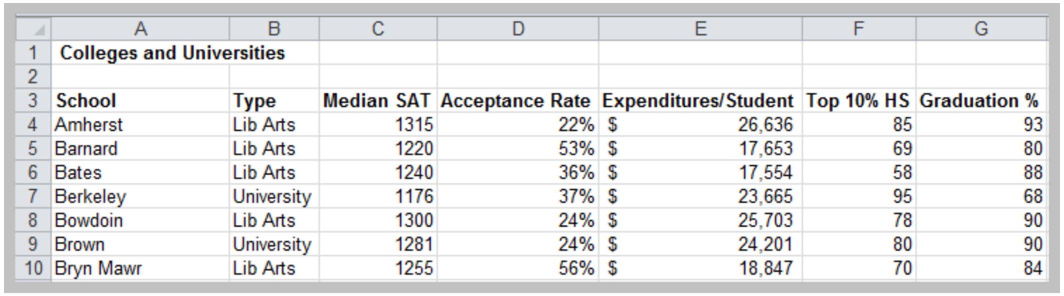
● New Worksheet Ply: College and Univ

○ New Workbook

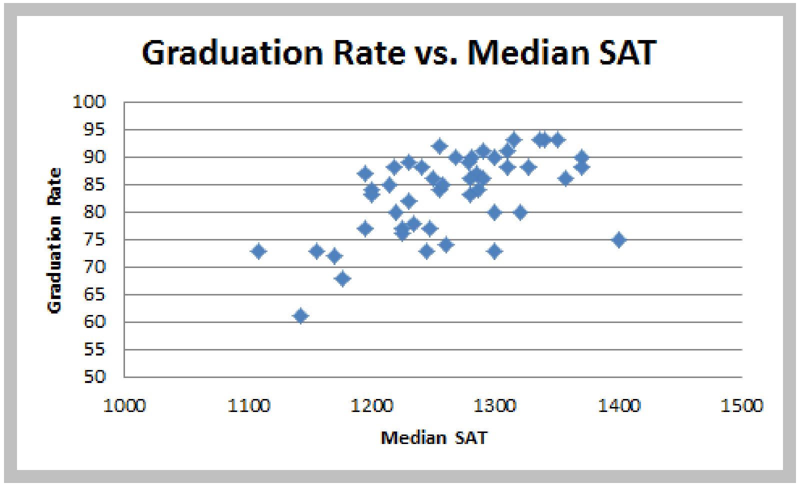
Colleges and Universities Data



	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1



	A	B	C	D	E	F	G
1	**Colleges and Universities**						
2							
3	**School**	**Type**	**Median SAT**	**Acceptance Rate**	**Expenditures/Student**	**Top 10% HS**	**Graduation %**
4	Amherst	Lib Arts	1315	22%	$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	$ 17,554	58	88
7	Berkeley	University	1176	37%	$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	$ 25,703	78	90
9	Brown	University	1281	24%	$ 24,201	80	90
10	Bryn Mawr	Lib Arts	1255	56%	$ 18,847	70	84



Descriptive Statistics for Categorical Data

- Sample proportion, *p* - fraction of data that has a certain characteristic
- Use the Excel function COUNTIF(*data range, criteria*) to count observations meeting a criterion to compute proportions.

Cross-Tabulation (Contingency Table)

- A tabular method that displays the number of observations in a data set for different subcategories of two categorical variables.
- The subcategories of the variables must be mutually exclusive and exhaustive, meaning that each observation can be classified into only one subcategory and, taken together over all subcategories, they must constitute the complete data set.

Example: *Facebook Survey*

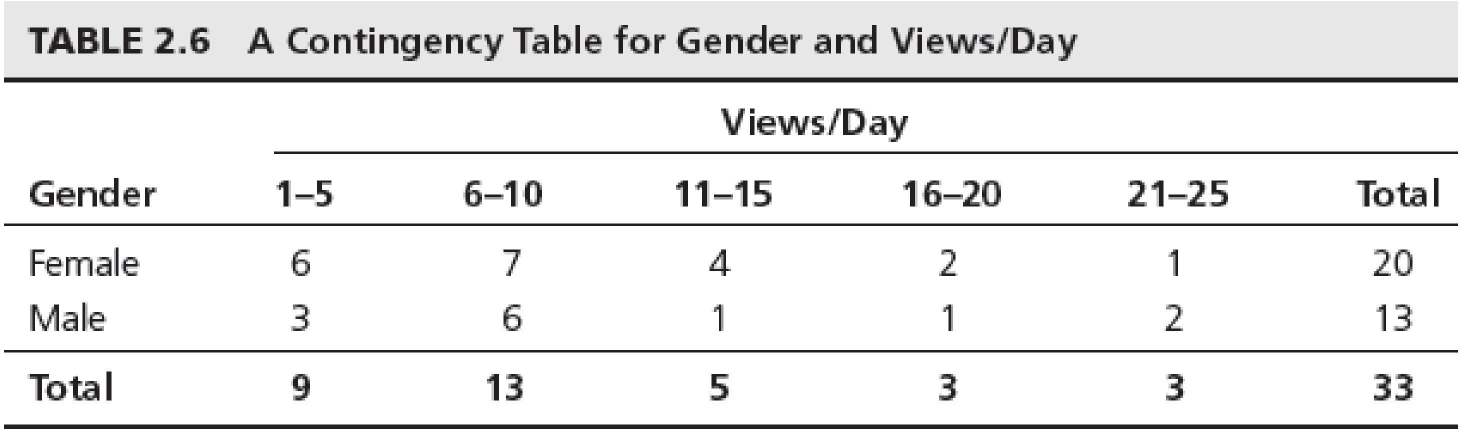


TABLE 2.6 A Contingency Table for Gender and Views/Day

	Views/Day					
Gender	1–5	6–10	11–15	16–20	21–25	Total
Female	6	7	4	2	1	20
Male	3	6	1	1	2	13
Total	**9**	**13**	**5**	**3**	**3**	**33**

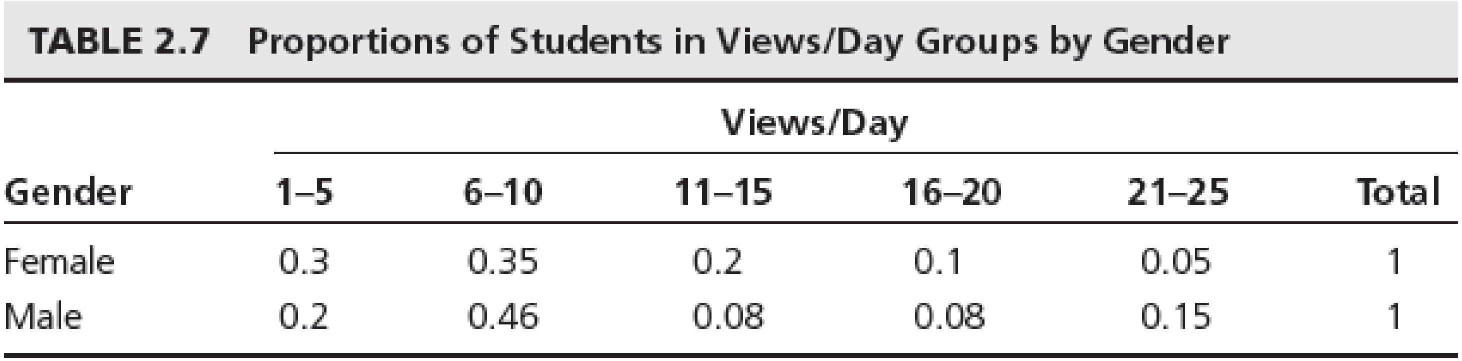
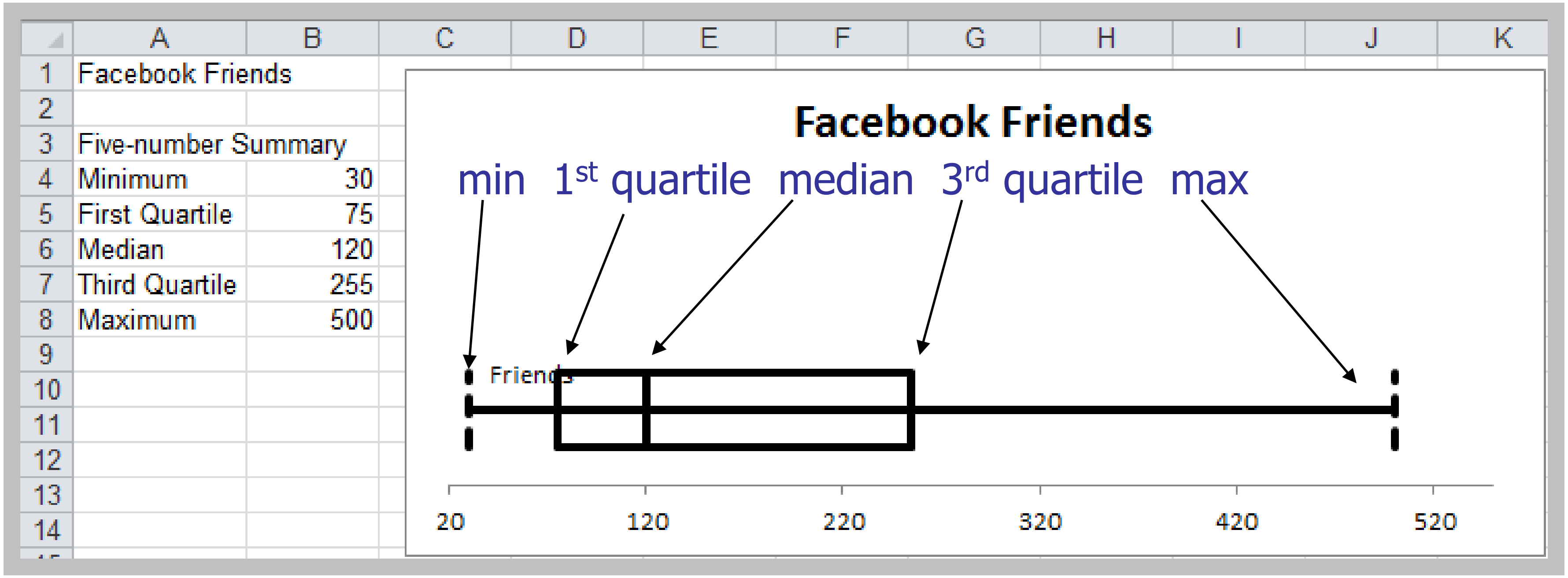


TABLE 2.7 Proportions of Students in Views/Day Groups by Gender

	Views/Day					
Gender	1–5	6–10	11–15	16–20	21–25	Total
Female	0.3	0.35	0.2	0.1	0.05	1
Male	0.2	0.46	0.08	0.08	0.15	1

Box Plots

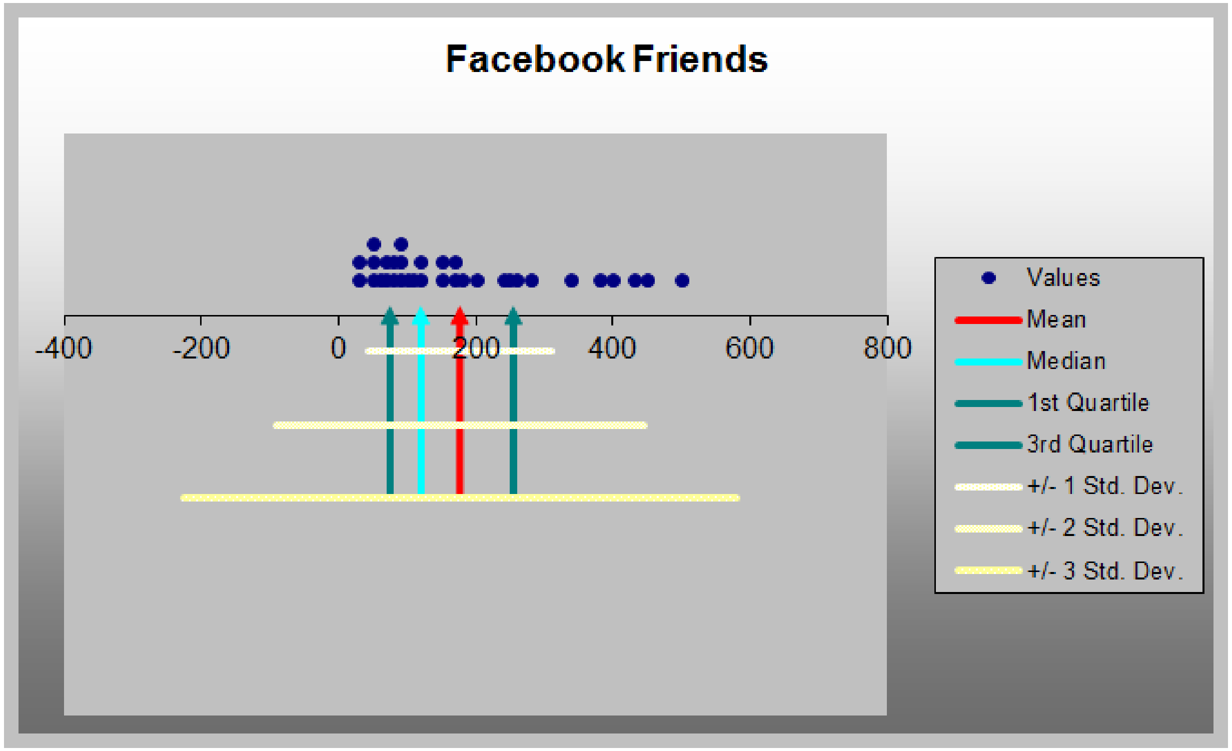
- Display minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum values graphically



	A	B
1	Facebook Friends	
2		
3	Five-number Summary	
4	Minimum	30
5	First Quartile	75
6	Median	120
7	Third Quartile	255
8	Maximum	500

Dot Scale Diagram

- *PHStat* menu > *Descriptive Statistics* > *Dot Scale Diagram*

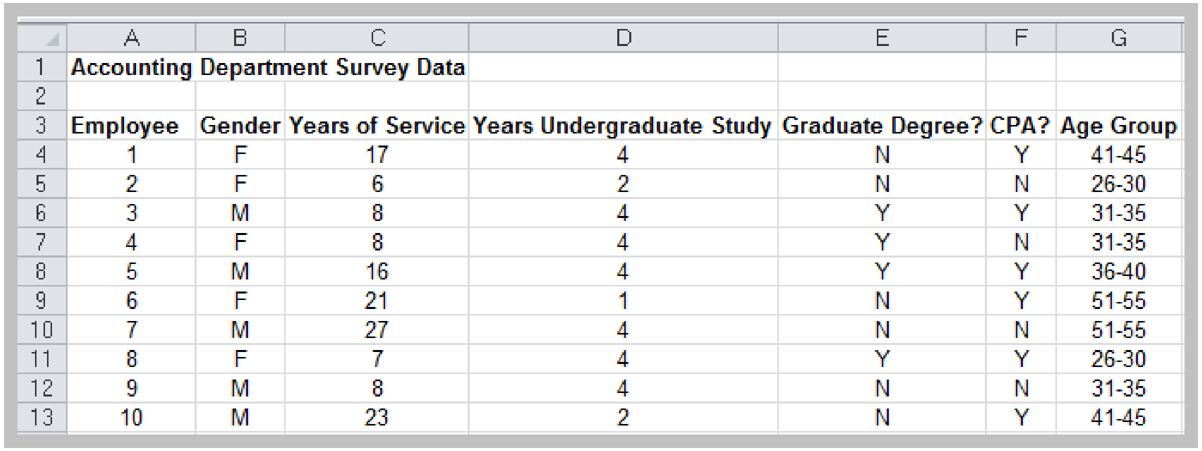


Outliers

- Outliers can make a significant difference in the results we obtain from statistical analyses.
- Box plots and dot-scale diagrams can help identify possible outliers visually.
- Other approaches:
  - Use the empirical rule to identify an outlier as one that is more than three standard deviations from the mean.
  - Use the IQR. “Mild” outliers are often defined as being between 1.5*IQR and 3*IQR to the left of Q 1 or to the right of Q 3 , and “extreme” outliers as more than 3*IQR away from these quartiles.

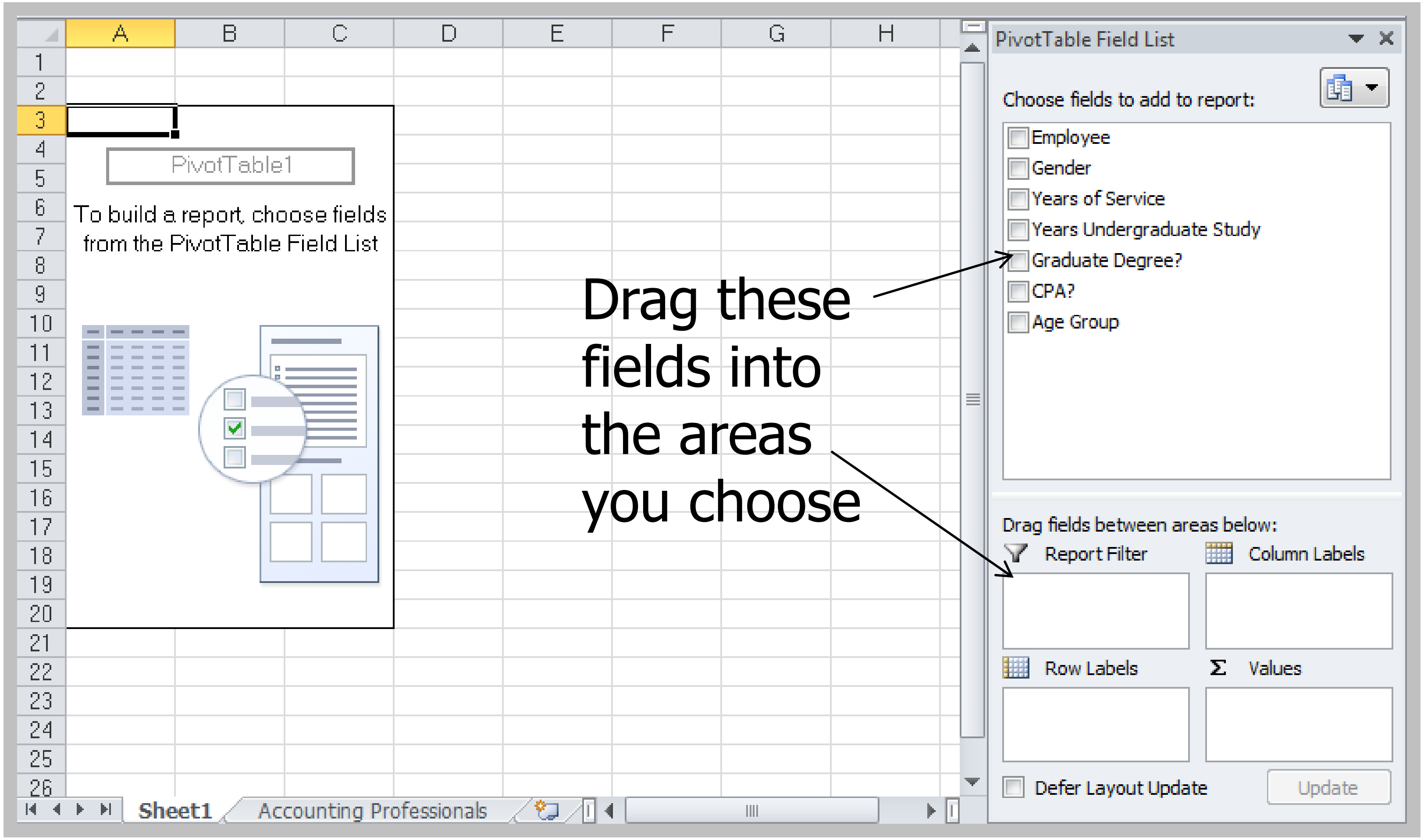
PivotTables

- Create custom summaries and charts from data
- Need a data set with column labels. Select any cell and choose *PivotTable Report* from *Data* menu. Follow the wizard steps.



	A	B	C	D	E	F	G
1	Accounting Department Survey Data						
2							
3	Employee	Gender	Years of Service	Years Undergraduate Study	Graduate Degree?	CPA?	Age Group
4	1	F	17	4	N	Y	41-45
5	2	F	6	2	N	N	26-30
6	3	M	8	4	Y	Y	31-35
7	4	F	8	4	Y	N	31-35
8	5	M	16	4	Y	Y	36-40
9	6	F	21	1	N	Y	51-55
10	7	M	27	4	N	N	51-55
11	8	F	7	4	Y	Y	26-30
12	9	M	8	4	N	N	31-35
13	10	M	23	2	N	Y	41-45

Blank PivotTable



PivotTable1

To build a report, choose fields from the PivotTable Field List

PivotTable Field List

Choose fields to add to report:

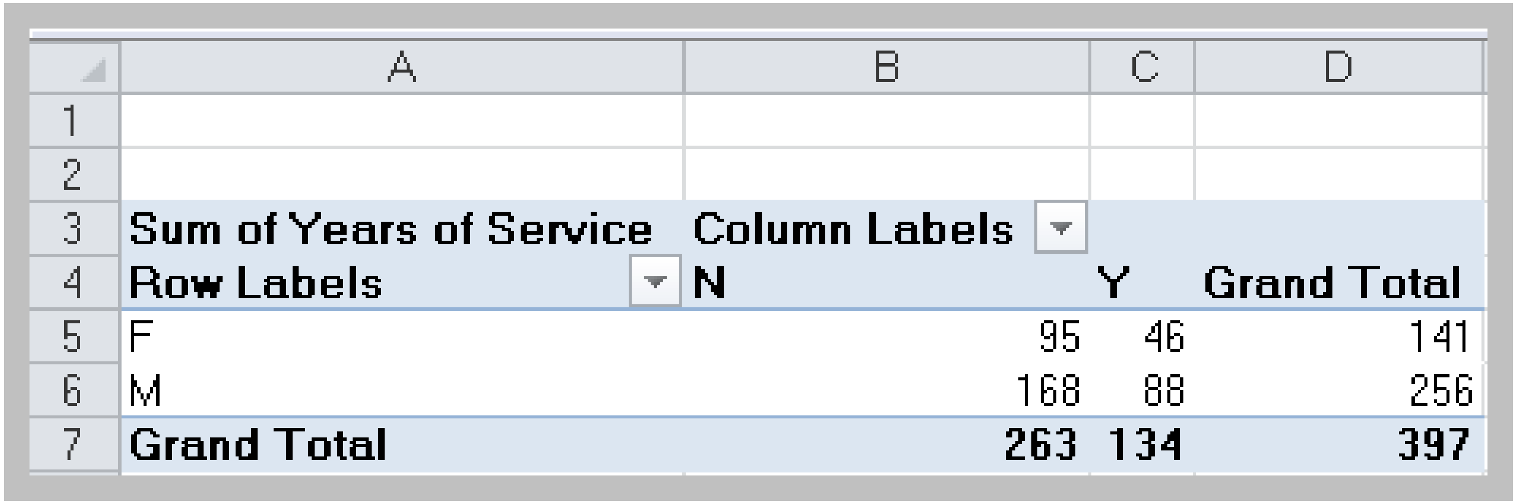
- □ Employee
- □ Gender
- □ Years of Service
- □ Years Undergraduate Study
- □ Graduate Degree?
- □ CPA?
- □ Age Group

Drag fields between areas below:

Drag these fields into the areas you choose

Example

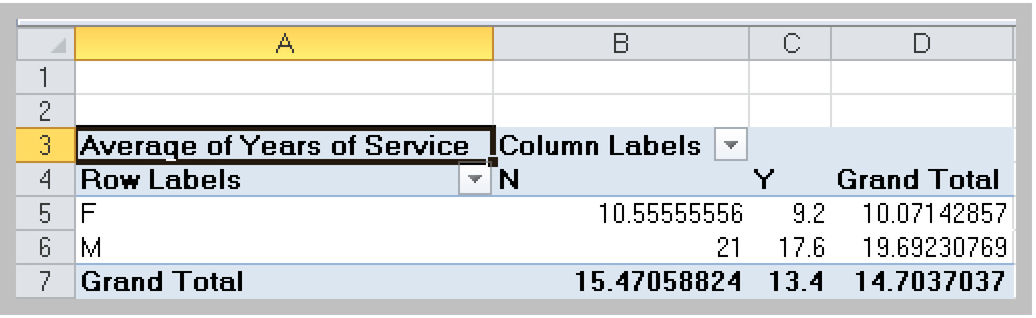
- Drag *Gender* from the *PivotTable Field List* to the *Row Labels* area, *Graduate Degree?* into the *Column Labels* area, and *Years of Service* into the *Values* area:



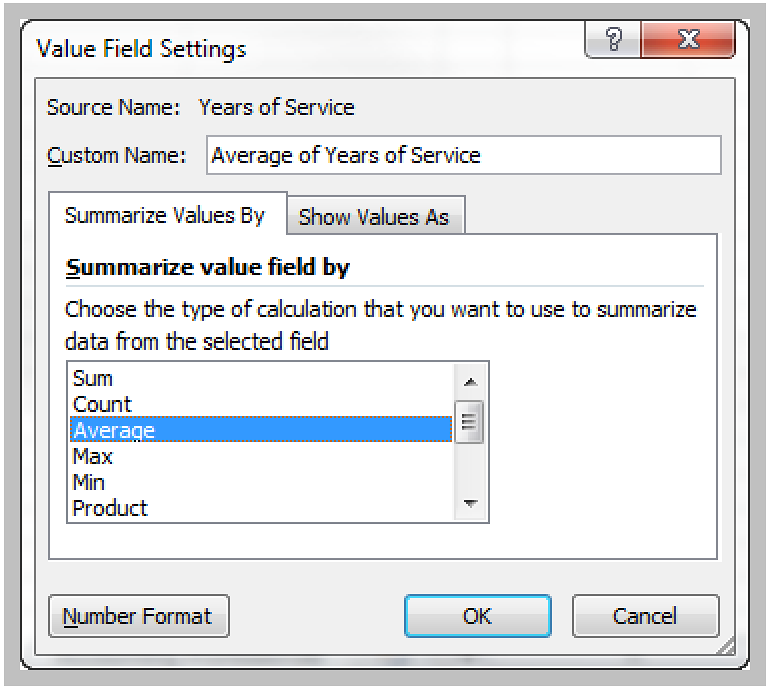
	A	B	C	D
1				
2				
3	**Sum of Years of Service**	**Column Labels** ▾		
4	**Row Labels** ▾	**N**	**Y**	**Grand Total**
5	F	95	46	141
6	M	168	88	256
7	**Grand Total**	**263**	**134**	**397**

Value Field Settings

In the *Options* tab under *PivotTable Tools* in the menu bar, click on the *Active Field* group and choose *Value Field Settings* to change type of summary



	A	B	C	D
1				
2				
3	**Average of Years of Service**	Column Labels		
4	**Row Labels**	**N**	**Y**	**Grand Total**
5	F	10.55555556	9.2	10.07142857
6	M	21	17.6	19.69230769
7	**Grand Total**	**15.47058824**	**13.4**	**14.7037037**



Value Field Settings

Source Name: Years of Service

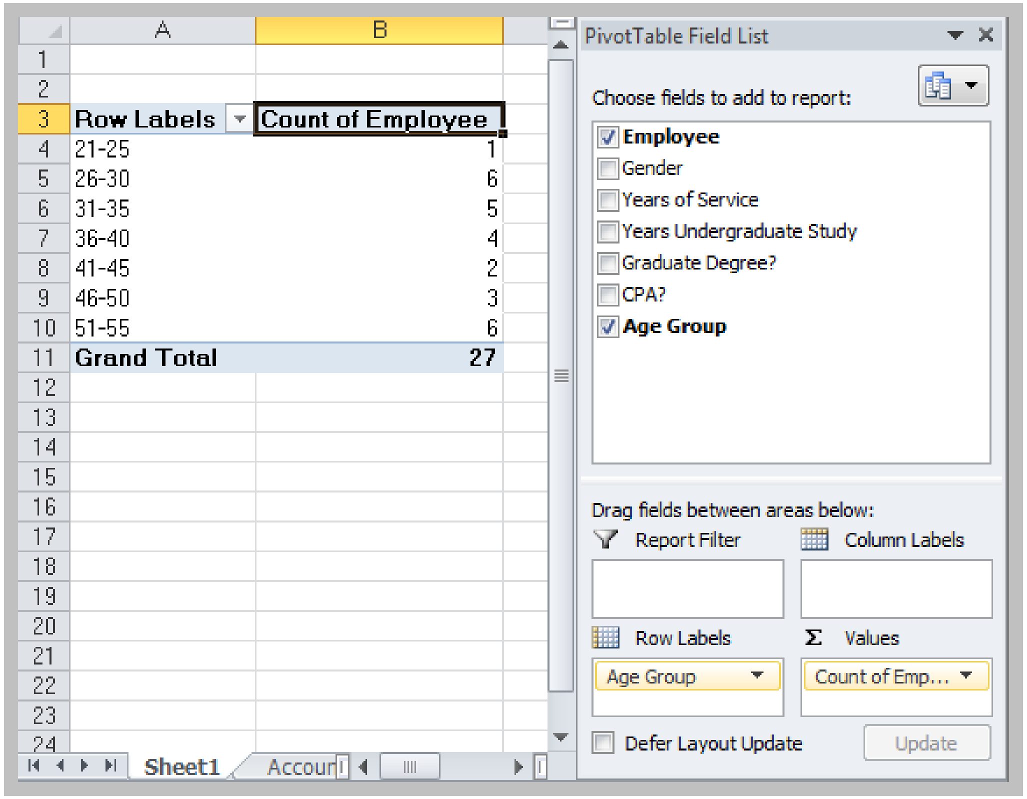
Custom Name: Average of Years of Service

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

Changing PivotTable Views

Uncheck the boxes in the *PivotTable Field List or drag the variable names to different* field areas.



	A	B
1		
2		
3	Row Labels	Count of Employee
4	21-25	1
5	26-30	6
6	31-35	5
7	36-40	4
8	41-45	2
9	46-50	3
10	51-55	6
11	**Grand Total**	**27**

PivotTable Field List

Choose fields to add to report:

- ☑ **Employee**
- □ Gender
- □ Years of Service
- □ Years Undergraduate Study
- □ Graduate Degree?
- □ CPA?
- ☑ **Age Group**

Drag fields between areas below:

Report Filter:

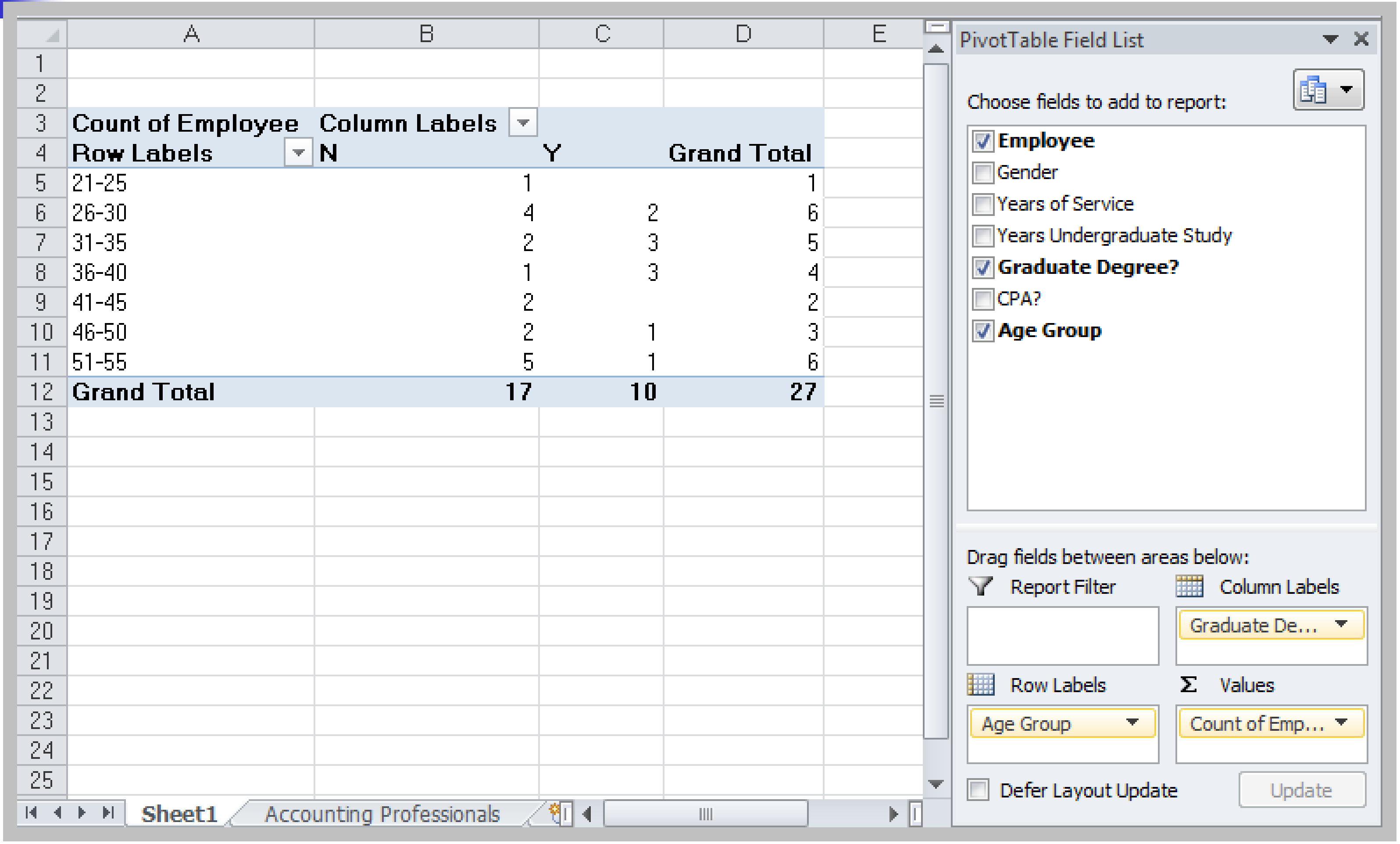
Column Labels:

Row Labels: Age Group

Values: Count of Emp...

□ Defer Layout Update

PivotTables for Cross Tabulation



Count of Employee	Column Labels		
Row Labels	N	Y	Grand Total
21-25	1		1
26-30	4	2	6
31-35	2	3	5
36-40	1	3	4
41-45	2		2
46-50	2	1	3
51-55	5	1	6
Grand Total	**17**	**10**	**27**

PivotTable Field List

Choose fields to add to report:

- ☑ **Employee**
- □ Gender
- □ Years of Service
- □ Years Undergraduate Study
- ☑ **Graduate Degree?**
- □ CPA?
- ☑ **Age Group**

Drag fields between areas below:

Report Filter:

Column Labels: Graduate De...

Row Labels: Age Group

Values: Count of Emp...

□ Defer Layout Update

Grouped Data: Calculation of Mean

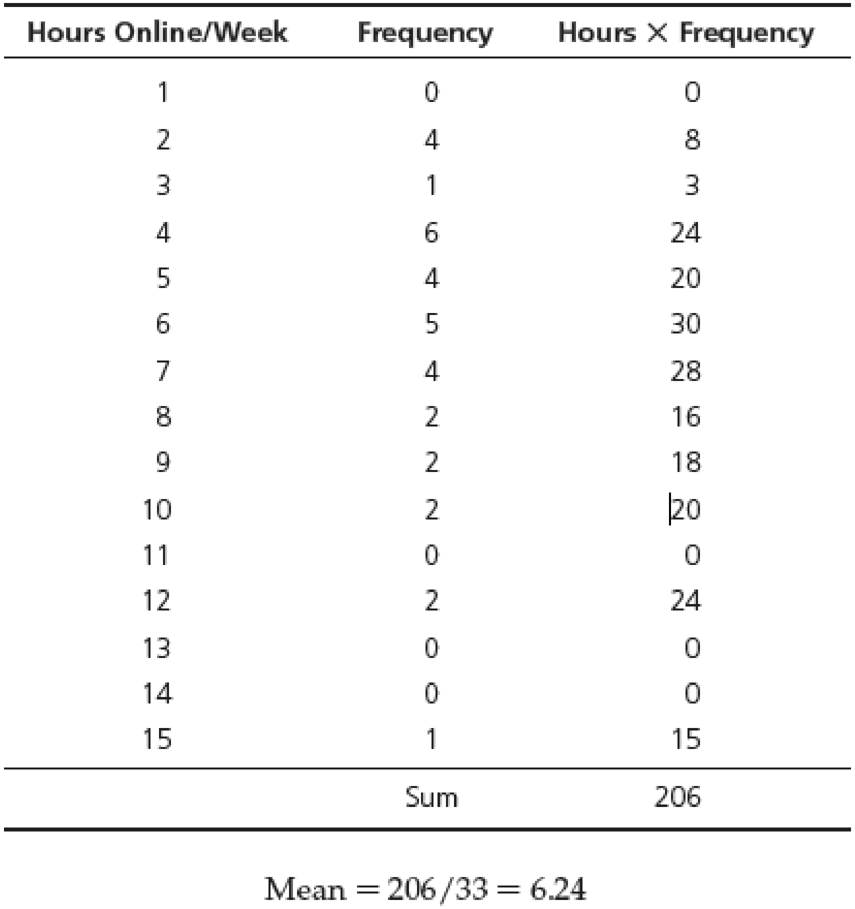
- Sample

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

- Population

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

Example

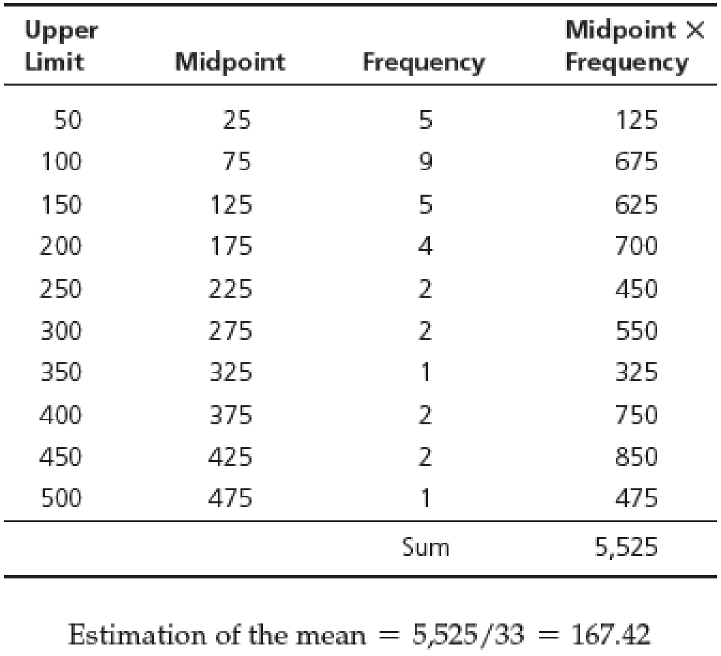


Hours Online/Week	Frequency	Hours × Frequency
1	0	0
2	4	8
3	1	3
4	6	24
5	4	20
6	5	30
7	4	28
8	2	16
9	2	18
10	2	20
11	0	0
12	2	24
13	0	0
14	0	0
15	1	15
	Sum	206

$$\text{Mean} = 206/33 = 6.24$$

Grouped Frequency Distribution

- We may estimate the mean by replacing x_i with a representative value (such as the midpoint) for all the observations in each cell.



Upper Limit	Midpoint	Frequency	Midpoint × Frequency
50	25	5	125
100	75	9	675
150	125	5	625
200	175	4	700
250	225	2	450
300	275	2	550
350	325	1	325
400	375	2	750
450	425	2	850
500	475	1	475
		Sum	5,525

Estimation of the mean $= 5{,}525/33 = 167.42$

Grouped Data: Calculation of Variance

- Sample

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

- Population

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