

## Chapter 2

### Describing Your Data

This chapter introduces *descriptive* statistics, center, spread, and distribution shape, which are almost always included with any statistical analysis to characterize a dataset. The particular descriptive statistics used depend on the *scale* that has been used to assign numbers to represent the characteristics of entities being studied. When the distribution of continuous data is bell shaped, we have convenient properties that make description easier. Chapter 2 looks at dataset types and their description.

#### 2.1 Describe Data with Summary Statistics and Histograms

We use numbers to measure aspects of businesses, customers and competitors. These measured aspects are *data*. Data become meaningful when we use statistics to describe patterns within particular *samples* or collections of businesses, customers, competitors, or other entities.

*Example 2.1 Yankees' Salaries: Is It a Winning Offer?* Suppose that the Yankees want to sign a promising rookie. They expect to offer \$1M, and they want to be sure they are neither paying too much nor too little. What would the General Manager need to know to decide whether or not this is the right offer?

He might first look at how much the other Yankees earn. Their 2005 salaries are in [Table 2.1](#):

Table 2.1 Yankees' salaries (in \$MM) in alphabetical order

Crosby	\$.3	Johnson	\$16.0	Posada	\$11.0	Sierra	\$1.5
Flaherty	.8	Martinez	2.8	Rivera	10.5	Sturtze	.9
Giambi	1.34	Matsui	8.0	Rodriguez	21.7	Williams	12.4
Gordon	3.8	Mussina	19.0	Rodriguez F	3.2	Womack	2.0
Jeter	19.6	Phillips	.3	Sheffield	13.0		

What should he do with this data?

Data are more useful if they are ordered by the aspect of interest. In this case, the Manager would re-sort the data by salary ([Table 2.2](#)):

Table 2.2 Yankees sorted by salary (in \$MM)

Rodriguez	\$21.7	Williams	\$12.4	Rodriguez F	\$3.2	Sturtze	\$.9
Jeter	19.6	Posada	11.0	Martinez	2.8	Flaherty	.8
Mussina	19.0	Rivera	10.5	Womack	2.0	Crosby	.3
Johnson	16.0	Matsui	8.0	Sierra	1.5	Phillips	.3
Sheffield	13.0	Gordon	3.8	Giambi	1.3		

Now he can see that the lowest Yankee salary, the *minimum*, is \$300,000, and the highest salary, the *maximum*, is \$21.7M. The difference between the maximum and the minimum is the *range* in salaries, which is \$21.4M, in this example. From these statistics, we know that the salary offer of \$1M falls in the lower portion of this range. Additionally, however, he needs to know just how unusual the extreme salaries are to better assess the offer.

He'd like to know whether or not the rookie would be in the better paid half of the Team. This could affect morale of other players with lower salaries. The *median*, or middle, salary is \$3.8M. The lower paid half of the team earns between \$300,000 and \$3.8M, and the higher paid half of the team earns between \$3.8M and \$21.7M. Thus, the rookie would be in the bottom half. The Manager needs to know more to fully assess the offer.

Often, a *histogram* and a *cumulative distribution plot* are used to visually assess data, as shown in [Figures 2.1](#) and [2.2](#). A histogram illustrates central tendency, dispersion, and symmetry.

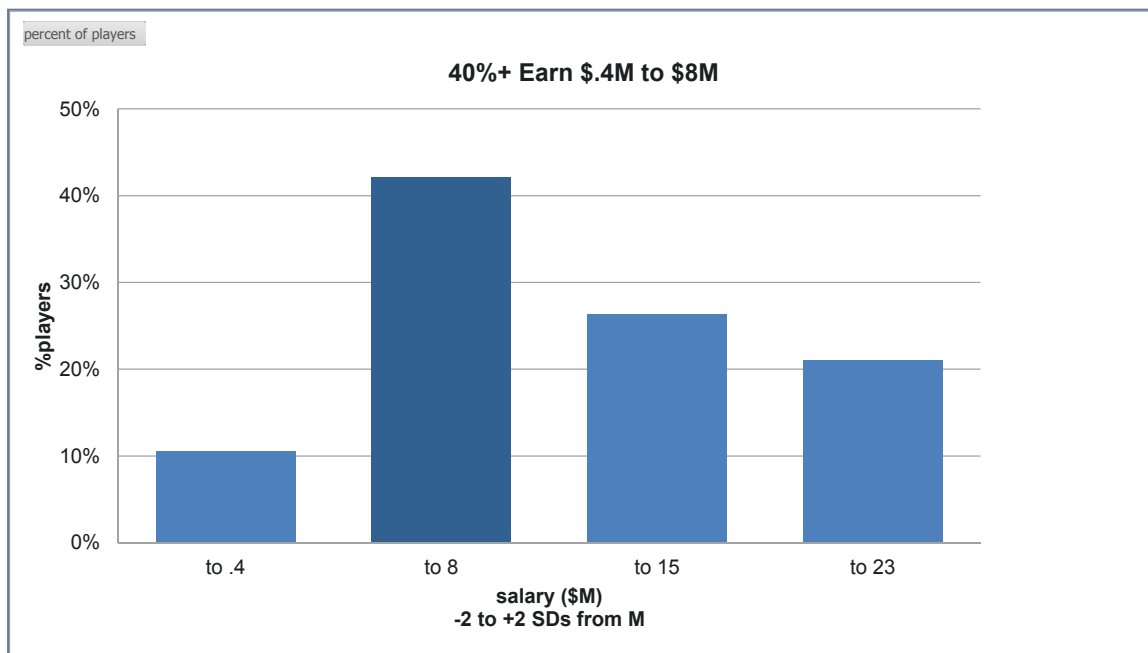


Figure 2.1 Histogram of Yankee salaries

The histogram of team salaries shows us that a large proportion, more than 40%, earn more than \$400,000, but less than the average, or *mean*, salary of \$8M.

The cumulative distribution makes it easy to see the median, or 50th percentile, which is one measure of central tendency. It is also easy to find the *interquartile range*, the range of values that the middle 50% of the datapoints occupy, providing a measure of the data dispersion.

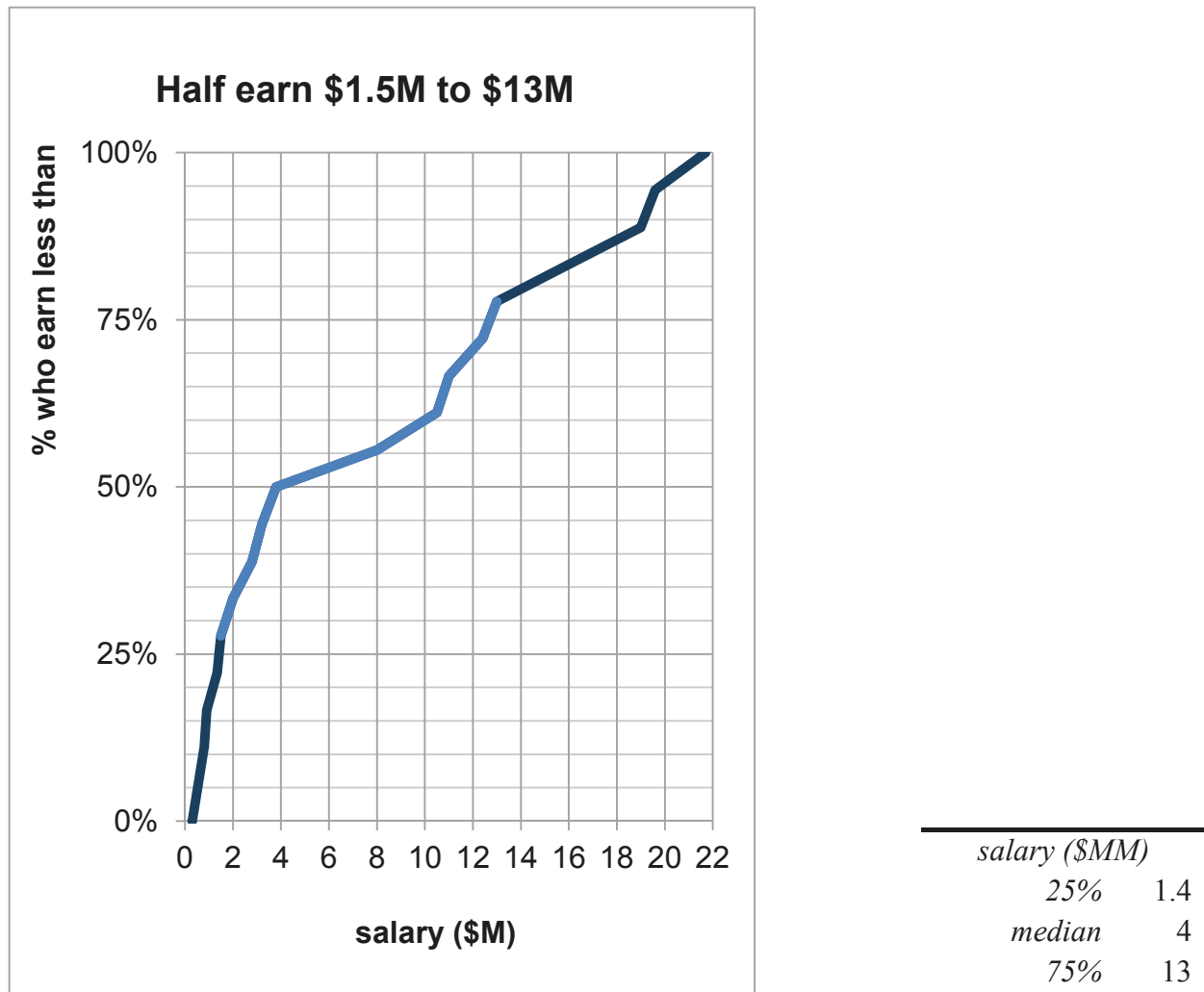
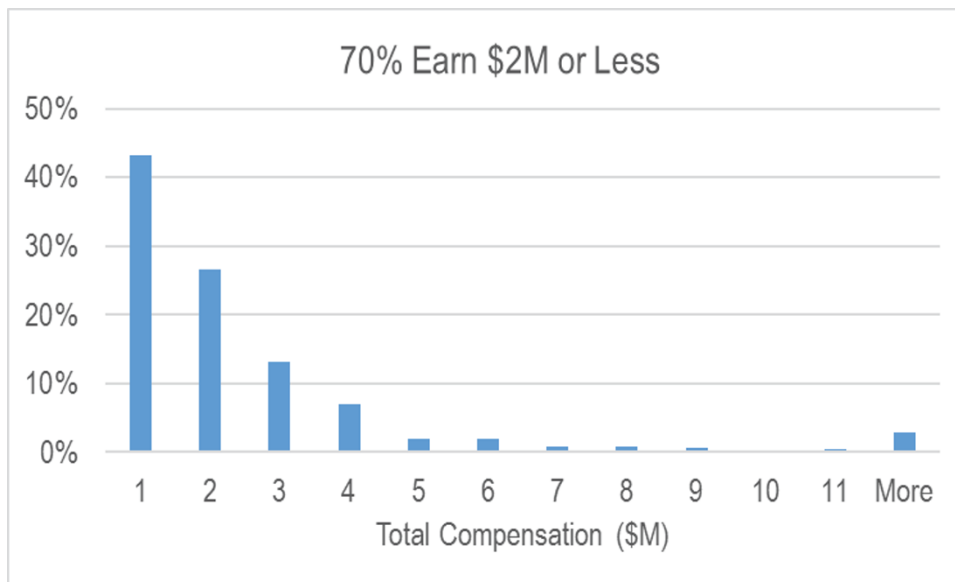


Figure 2.2 Cumulative distribution of salaries

The cumulative distribution reveals that the *Interquartile Range*, between the 25th percentile and the 75th percentile, is more than \$10M. A quarter earns less than \$1.4M, the 25th percentile, about half earn between \$1.5 and \$13M, and a quarter earns more than \$13M, the 75th percentile. Half of the players have salaries below the *median* of \$4M and half have salaries above \$4M.

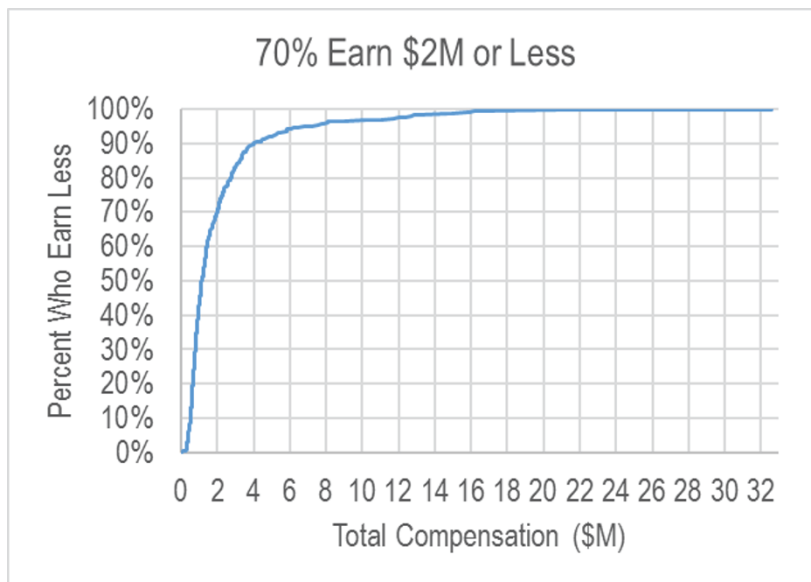
**Example 2.2 Executive Compensation: Is the Board's Offer on Target?** The Board of a large corporation is pondering the total compensation package of the CEO, which includes salary, stock ownership, and fringe benefits. Last year, the CEO earned \$2,000,000. For comparison, The Board consulted Forbes' summary of the total compensation of the 500 largest corporations. The histogram, cumulative frequency distribution and descriptive statistics are shown in [Figures 2.3](#) and [2.4](#).



<i>Total Compensation (\$M)</i>	<i>% Executives</i>
< 1.0	43
1.1 to 2.0	27
2.1 to 3.0	13
3.1 to 4.0	7
4.1 to 5.0	2
5.1 to 6.0	2
6.1 to 7.0	1
7.1 to 8.0	1
8.1 to 9.0	1
9.1 to 10.0	0
10.1 to 11.0	0
>11.0	3

Figure 2.3 Histogram of executive compensation





Total Compensation (\$M)	
<i>M</i>	2.1
<i>SD</i>	3.0
75th percentile	2.3
median	1.1
25th percentile	.7

Figure 2.4 Cumulative distribution of total compensation

The average executive compensation in this sample of large corporations is \$2.1M. Half the sample of 447 executives earns \$1.1M (the median) or less. One quarter earns less than \$.7M, the middle half, or *interquartile range*, earns between \$.7M and \$2.3M, and one quarter earns more than \$2.3M.

## 2.2 Round Descriptive Statistics

In the examples above, statistics in the output from statistical packages are presented with many decimal points of accuracy. The Yankee manager in Example 2.1 and The Board considering executive compensation in Example 2.2 will most likely be negotiating in hundred thousands. It would be distracting and unnecessary to report descriptive statistics with significant digits more than two or three. In the **Yankees** example, the average salary is \$8,000,000 (*not* \$7,797,000). In the **Executive Compensation** example, average total compensation is \$2,200,000 (*not* \$2,215,262.66). It is deceptive to present results with many significant digits, creating an illusion of precision. In addition to being honest, statistics in two or three significant digits are much easier for decision makers to process and remember. If more significant digits don't affect a decision, round to fewer and make your statistics easier to process and remember.

## 2.3 Share the Story That Your Graphics Illustrate

Use your graphics to support the conclusion you have reached from your analysis. Choose a "bottom line" title that shares with your audience what it is that they should be able to see. Often this title should relate specifically to your reasons for analyzing data. In the executive compensation example, The Board is considering a \$2M offer. The chart titles capture Board interest by highlighting this critical value. The "bottom line," that a \$2M offer is relatively high, when compared with similar firms, makes the illustrations relevant.

Many have the unfortunate and unimaginative habit of choosing chart titles which name the type of chart. “Histogram of executive salaries” tells the audience little, beyond the obvious realization that they must form their own, independent conclusions from the analysis. Choose a “bottom line” title so that decision makers can take away your conclusion from the analysis. Develop the good habit of titling your graphics to enhance their relevance and interest.

## 2.4 Data Is Measured with Quantitative or Categorical Scales

If the numbers in a dataset represent amount, or magnitude of an aspect, **and** if differences between adjacent numbers are equivalent, the data are *quantitative* or *continuous*. Data measured in dollars (i.e., revenues, costs, prices and profits) or percents (i.e., market share, rate of return, and exam scores) are continuous. Quantitative numbers can be added, subtracted, divided or multiplied to produce meaningful results.

With quantitative data, report central tendency with the *mean*,  $M$ :

$$\mu = \frac{\sum x_i}{N} \text{ for describing a } \textit{population} \text{ and}$$

$$\bar{X} = \frac{\sum x_i}{N} \text{ for describing a } \textit{sample} \text{ from a population,}$$

where  $x_i$  are data point values, and

$N$  is the number of data points that we are describing.

The *median* can also be used to assess central tendency, and the *range*, *variance*, and *standard deviation* can be used to assess dispersion.

The *variance* is the average squared difference between each of the data points and the mean:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \text{ for a population and}$$

$$s^2 = \frac{\sum (x_i - \bar{X})^2}{(N - 1)} \text{ for a sample from a population.}$$

The *standard deviation*  $SD$ ,  $\sigma$  for a population and  $s$  for a sample, is the square root of the variance, which gives us a measure of dispersion in the more easily interpreted, original units, rather than squared units.

To assess distribution symmetry, assess its skewness:

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

Skewness of zero indicates a symmetric distribution, and skewness between  $-1$  and  $+1$  is evidence of an approximately symmetric distribution.

If numbers in a dataset are arbitrary and used to distinguish categories, the data are *nominal*, or *categorical*. Football jersey numbers and your student ID are nominal. A larger number doesn't mean that a player is better or a student is older or smarter. Categorical numbers can be tabulated to identify the most popular number, occurring most frequently, the *mode*, to report central tendency. Categorical numbers cannot be added, subtracted, divided or multiplied.

Quantitative measures convey the more information, including direction and magnitude, while categorical measures convey the less, sometimes direction, and sometimes, merely category membership. One, more informative type of categorical data are *ordinal* scales that used to rank order data, or to convey direction, but not magnitude. With ordinal data, an element (which could be a business, a person, a country) with the most or best is coded as '1', second place as '2', etc. With ordinal numbers, or rankings, data can sorted, but not added, subtracted, divided or multiplied. As with other categorical data, the mode represents the central tendency of ordinal data.

When focus is on membership in a particular category, the *proportion* of sample elements in the category is a continuous measure of central tendency. Proportions are quantitative and can be added, subtracted, divided or multiplied, though they are bounded by zero, below, and by one, above.

## 2.5 Continuous Data Are Sometimes Normal

Continuous variables are often *Normally distributed*, and their histograms resemble symmetric, bell shaped curves, with the majority of data points clustered around the mean. Most elements are "average" with values near the mean; fewer elements are unusual and far from the mean.

Skewness reflects lack of symmetry. Normally distributed data have skewness of zero, and approximately Normal data have skewness between  $-1$  and  $+1$ .

If continuous data are Normally distributed, we need only the mean and standard deviation to describe this data and description is simplified.

*Example 2.3 Normal SAT Scores.* Standardized tests, such as SAT, capitalize on Normality. Math and verbal SATs are both specifically constructed to produce Normally distributed scores with *mean*  $M = 500$  and *standard deviation*  $SD = 100$  over the population of students (Figure 2.5):

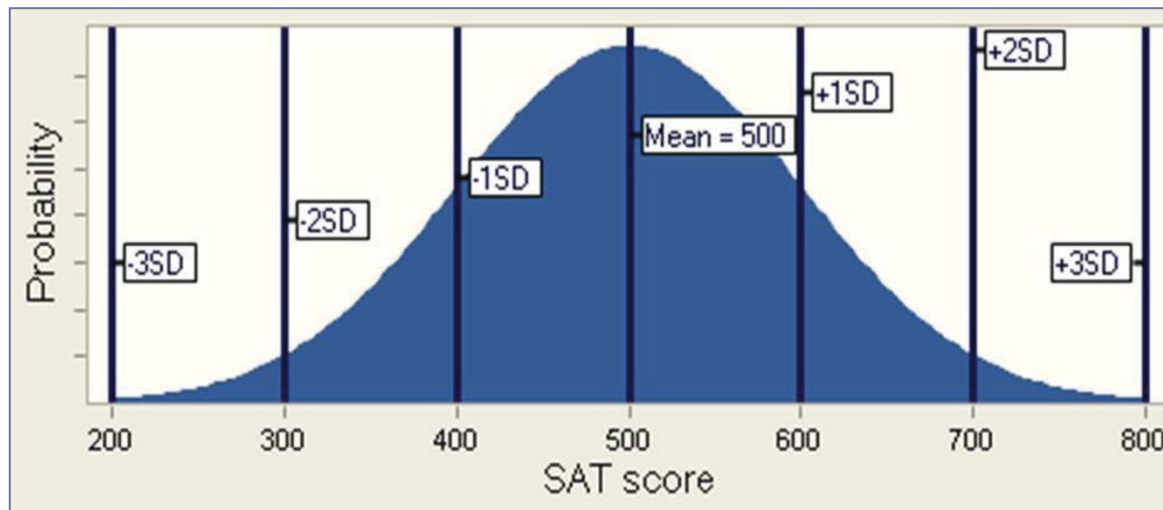


Figure 2.5 Normally distributed SAT scores

## 2.6 The Empirical Rule Simplifies Description

Normally distributed data have a very useful property described by the *Empirical Rule*:

- 2/3 of the data lie within one standard deviation of the mean
- 95% of the data lie within two standard deviations of the mean

This is a powerful rule! *If data are Normally distributed, data can be described with just two statistics: the mean and the standard deviation.*

Returning to SAT scores, if we know that the average score is 500 and the standard deviation is 100, we also know that

- 2/3 of SAT scores will fall within 100 points of the mean of 500, or between 400 and 600,
- 95% of SAT scores will fall within 200 points of the mean of 500, or between 300 and 700.

*Example 2.4 Class of Business Students' SATs: Normal & Exceptional.* Descriptive statistics and a histogram of Math SATs of a third year class of business students reveal an interquartile range from 640 to 730, with mean of 690 and standard deviation of 70, as shown in [Figure 2.6](#). Skewness is  $-0.5$ , indicating approximate symmetry, an approximately Normal distribution.

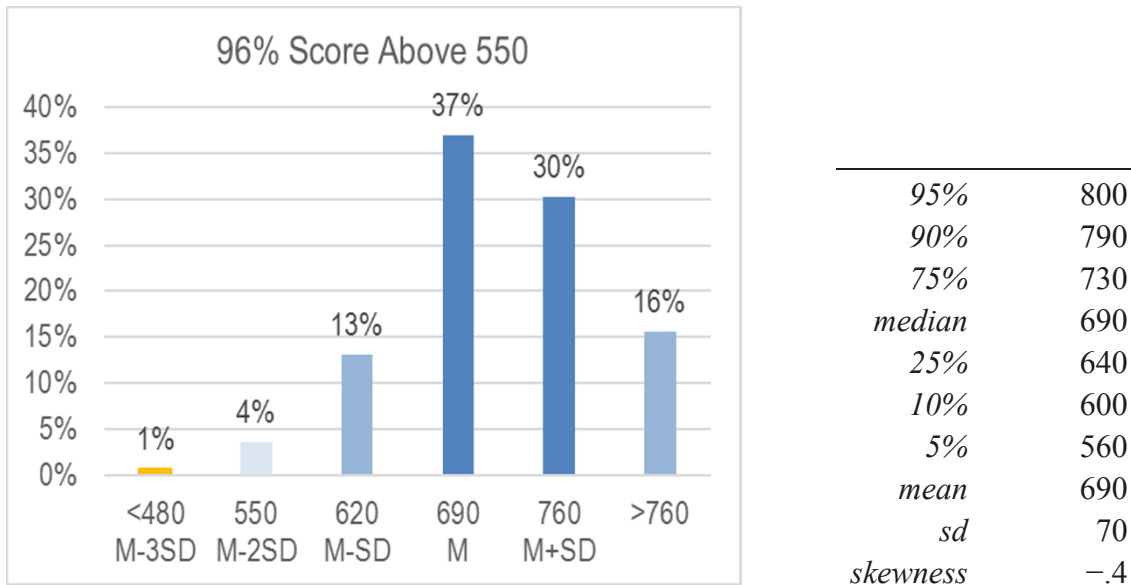


Figure 2.6 Histograms and descriptive statistics of class of business students' math SATs

These scores are bell shaped. However, there are “too many” perfect scores of 800.

The Empirical Rule would predict that 2/3 of the class would have scores within one standard deviation, 70 points, of the mean of 690, or within the interval 620 to 760. There actually 67% (=37%+30%).

The Empirical Rule would also predict that only 2-1/2% of the class would have scores more than two standard deviations below or above the mean of 690: scores below 550 and above 830. We find that 5% actually do have scores below 550, though none score above 830 (since a perfect SAT score is 800). This class of business students has Math SATs that are nearly Normal, but not exactly Normal.

To summarize students' SAT scores, report:

- Business students' Math SAT scores are approximately Normally distributed with *mean* of 690 and *standard deviation* of 70.
- Relative to the larger population of all SAT takers, the smaller *standard deviation* in business students' Math SAT scores, 70 versus 100, indicates that this class of business students is a more homogeneous group than the more varied population.

## 2.7 Outliers Can Distort the Picture

Outliers are extreme elements, considered unusual when compared with other sample elements. Because they are extraordinary, they can distort descriptive statistics.

Revisiting the **Executive Compensation** example, why is the *mean*, \$2.2M, so much larger than the *median*, \$1.1M? There is a group of *outliers*, shown as *MORE* than three standard deviations above the mean in Figure 2.7, who are compensated extraordinarily well. Each collects a compensation package of more than \$11.1M, a compensation level that is more than three standard deviations greater than the mean.

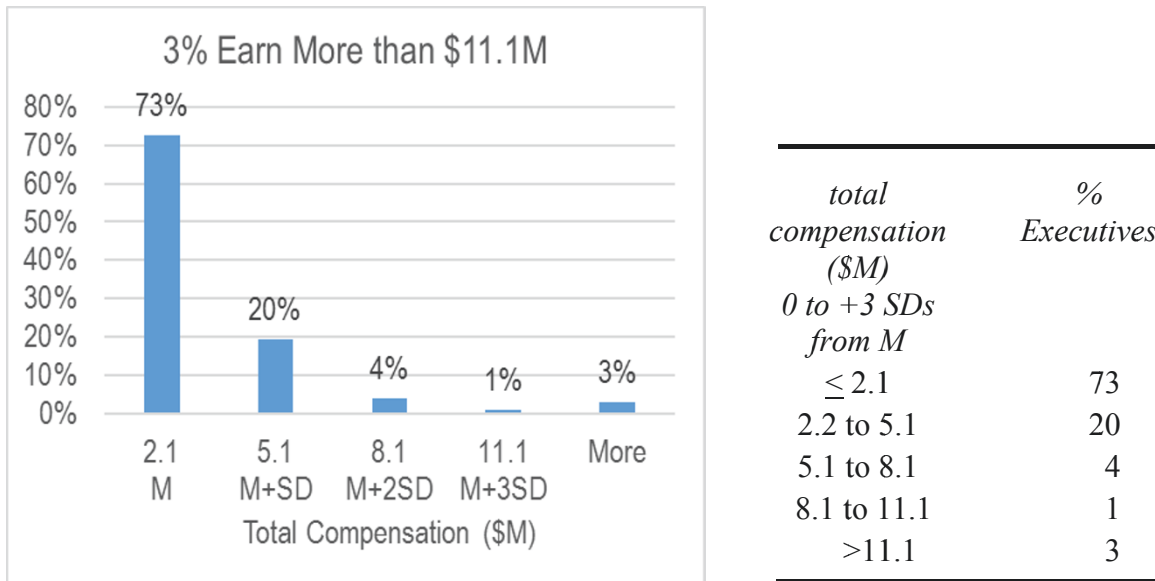


Figure 2.7 Histogram and descriptive statistics by SDs from M

Because extraordinary executives exist, the distribution of compensation is *skewed*, with relatively few exceptional executives being exceptionally well compensated.

## 2.8 Central Tendency, Dispersion and Skewness Describe Data

The baseball salaries and executive compensation examples focused on two measures of *central tendency*: the *mean*, or average, and the *median*, or middle. Both examples also refer to a measure of *dispersion* or variability: the *range* separating the minimum and maximum. *Skewness* reflects distribution symmetry. SATs are approximately symmetric and Normal; Executive compensation values are skewed. To describe data, we need statistics to assess central tendency, dispersion, and skewness. The statistics we choose depends on the *scale* which has been used to code the data we are analyzing.

## 2.9 Describe Categorical Variables Graphically

Numbers representing category membership in nominal, or categorical, data are described by tabulating their frequencies. The most popular category is the *mode*. Visually, we show our tabulations with a *Pareto* chart, which orders categories by their popularity.

**Example 2.5 Who Is Honest & Ethical?** Figure 2.8 shows a column chart of results of a survey of 1014 adults by Gallup:

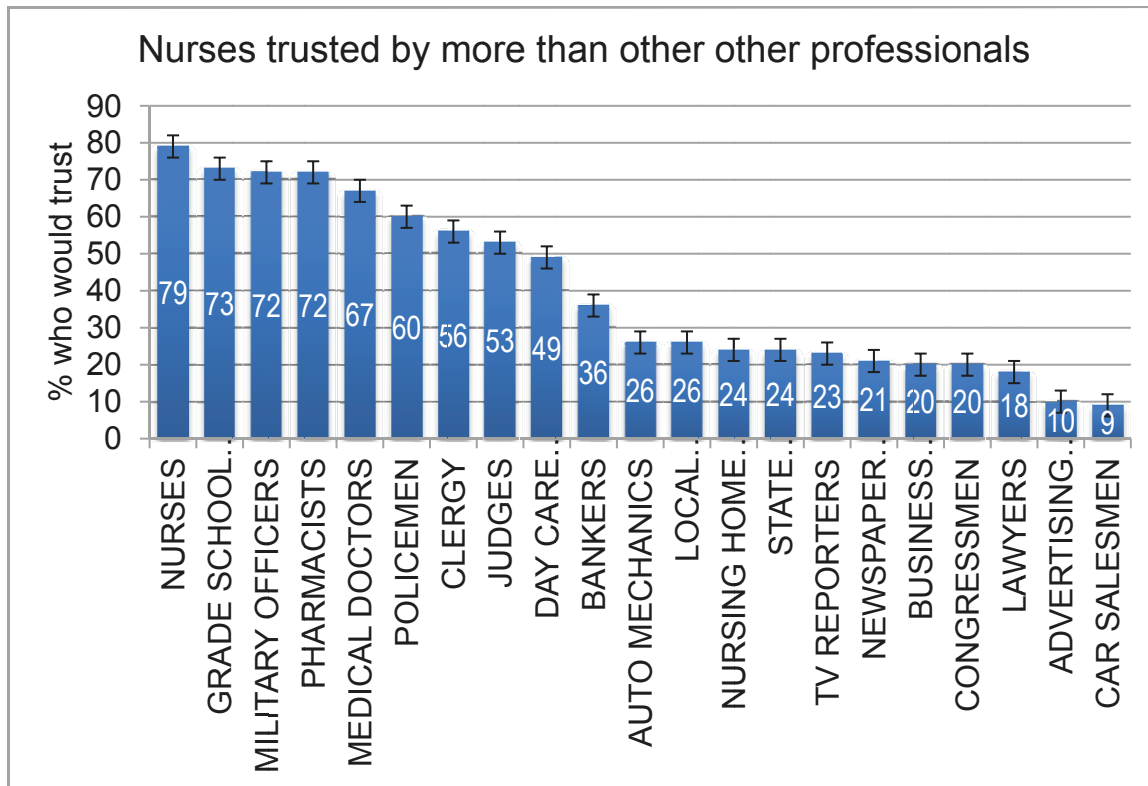


Figure 2.8 Pareto charts of the percents who judge professions honest

More Americans trust and respect nurses (79%, the *modal* response) than people in other professions, including doctors, clergy and teachers. Though a small minority judge business executives (20%) and advertising professionals (10%) as honest and ethical, most do not judge people in those fields to be honest (which highlights the importance of ethical business behavior in the future).

## 2.10 Descriptive Statistics Depend On The Data and Rely on Your Packaging

Descriptive statistics, graphics, central tendency and dispersion, depend upon the type of scale used to measure data characteristics (i.e., quantitative or categorical).

Table 2.3 summarizes the descriptive statistics (graph, central tendency, dispersion, shape) used for both types of data:

Table 2.3 Descriptive statistics (central tendency, dispersion, graphics) for two types of data

	<b>Quantitative</b>	<b>Categorical</b>
<b>Central tendency</b>	<i>mean</i> <i>median</i>	<i>mode</i> <i>proportion</i>
<b>Dispersion</b>	<i>range</i> <i>standard deviation</i>	
<b>Symmetry</b>	<i>skewness</i>	
<b>Graphics</b>	<i>histogram</i> <i>cumulative distribution</i>	<i>Pareto chart</i> <i>pie chart</i> <i>column chart</i>

If continuous data are Normally distributed, a dataset can be completely described with just the mean and standard deviation. We know from the *Empirical Rule* that 2/3 of the data will lie within one standard deviation of the mean and that 95% of the data will lie within two standard deviations of the mean.

Effective results are those which are remembered and used to improve decision making. Your presentation of results will influence whether or not decision makers remember and use your results. Round statistics to two or three significant digits to make them honest, digestible, and memorable. Title your graphics with the “bottom line,” to guide and facilitate decision makers’ conclusions.



## Excel 2.1 Produce Descriptive Statistics

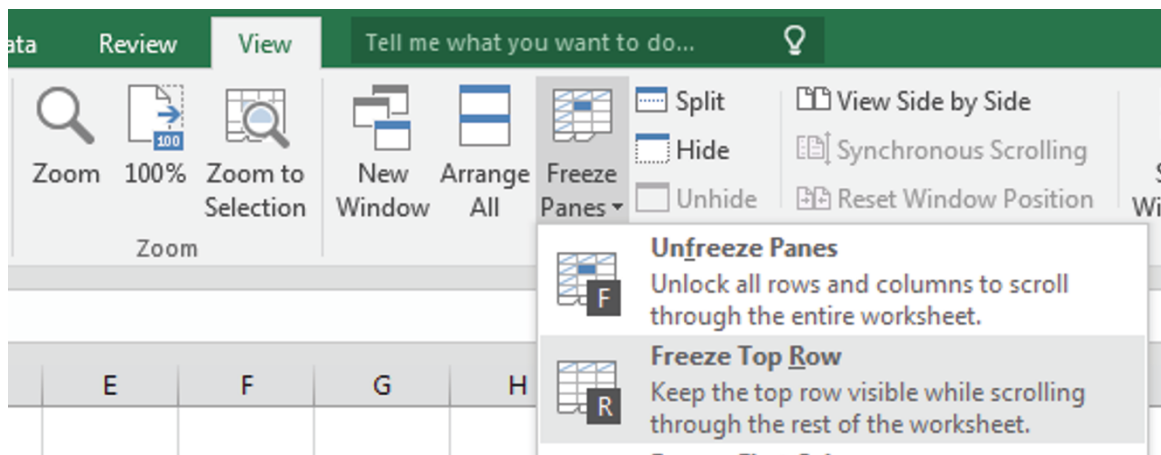
*Executive Compensation.* We will describe executive compensation packages by producing descriptive statistics, a histogram and cumulative distribution.

First, freeze the top row of **Excel 2.1 Executive Compensation** so that column labels are visible when you are at the bottom of the dataset.

From the first cell, **A1**,

**Alt WFR**

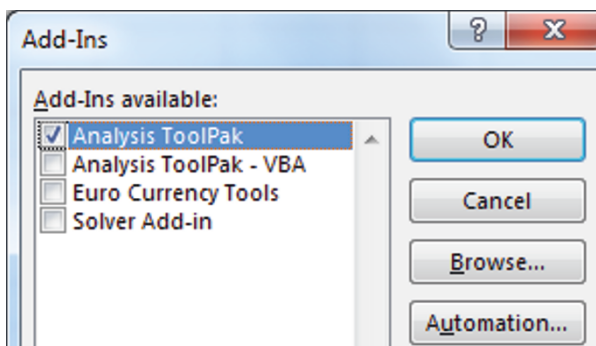
(The shortcuts, activated with **Alt** select the **vieW** menu , the **Freeze panes** menu, and then freeze **Rows**.)



## Descriptive Statistics.

Turn on the Excel statistics add-in, Analysis ToolPak.

**Alt TI**

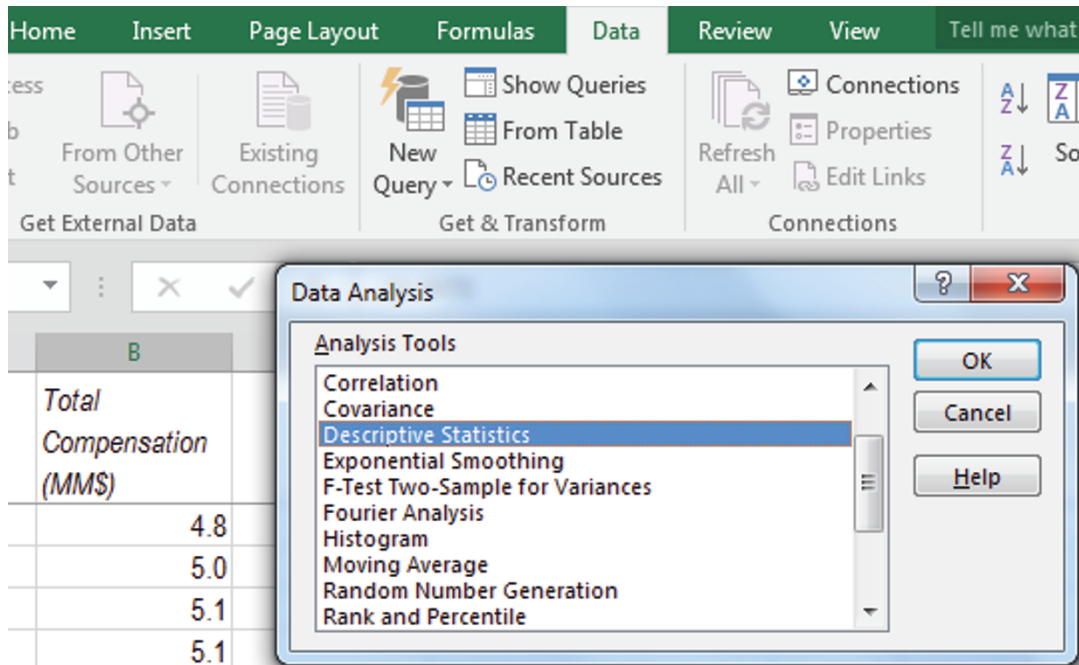
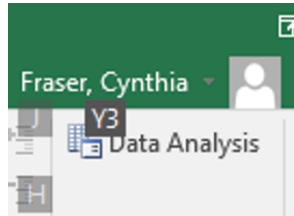


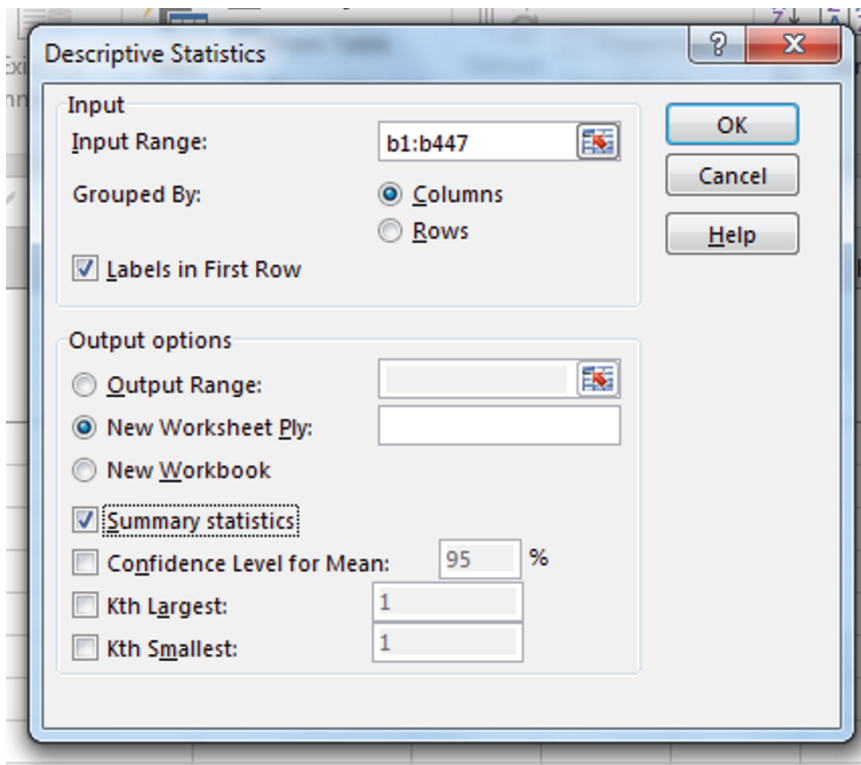
Request Descriptive Statistics.

**Alt AYn D**

The number *n* varies. Enter what you see in the menu, which is 3 on this computer.

**b1:b447 tab LS**





	A	B
1	Total Compensation (M\$)	
2		
3	Mean	2.101147
4	Standard Error	0.141497
5	Median	1.12647
6	Mode	#N/A
7	Standard Deviation	2.988242
8	Sample Variance	8.929591
9	Kurtosis	32.2745
10	Skewness	4.756857
11	Range	32.55348
12	Minimum	0.028816
13	Maximum	32.5823
14	Sum	937.1115
15	Count	446

**Set up Histogram Bins.** To make a histogram of compensation, Excel needs to know what ranges of values to combine. To take advantage of the *Empirical Rule*, create *bins*, or categories, using differences from the approximate sample mean that are in widths of approximate standard deviations. In this case, the mean is about 2 and the standard deviation is about 3.

Move back to the data page, and then move to the bottom of the data.

**Cntl+Page Down**

In column B

**Cntl+down arrow**

Excel uses bin values to set the upper limit for each category. Start with a bin with upper limit equal to 2, which will include compensation values that are at less than or equal to 2.

	A	B
		Total Compensation (M\$)
1		
445		16.2
446		20.7
447		32.6
448		
449		Total Compensation
450	M	2
451	M+SD	5
452	M+2SD	8
453	M+3SD	11

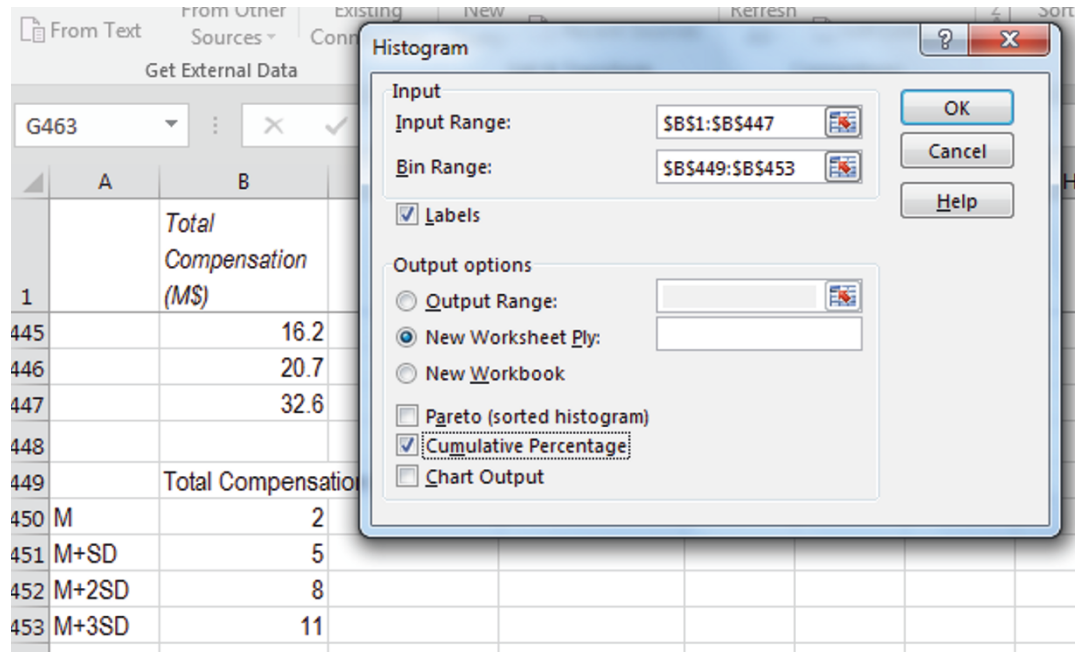
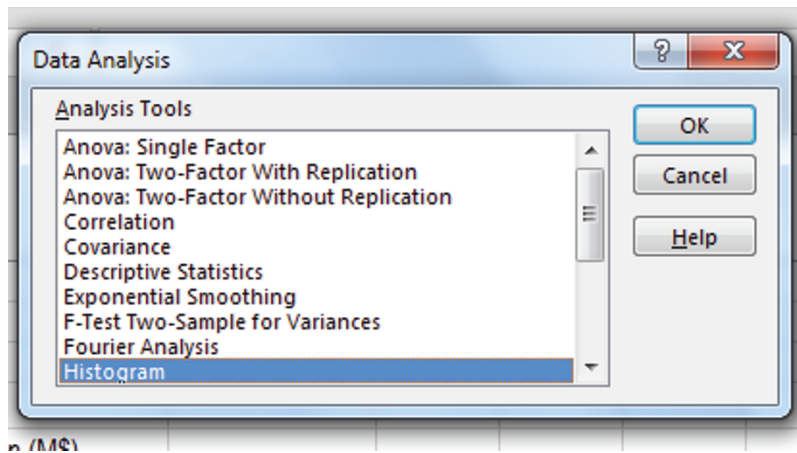
This will be the first bin, since subtracting one standard deviation from the mean produces a negative number, and none of the executives earns negative salary dollars

In each of the three cells below this first bin, add one SD to the cell above, 3, creating bins with upper limits of  $M + 1SD$ ,  $M + 2SD$  and  $M + 3SD$ .

Request a tabulation.

**Alt AYn H**

**b1:b447 tab b449:b453 tab LM**



	A	B	C
1	Compensation	Frequency	Cumulative
2	2.0	312	69.96%
3	5.0	99	92.15%
4	8.0	17	95.96%
5	11.0	5	97.09%
6	More	13	100.00%

To produce a histogram showing percents of the sample in each compensation category, add a column D of the change in cumulative%.

**d2=c2**

**d3=c3-c2**

Select d3

**Shift+down arrow to d6**

**Cntl+D**

D3					<b>=C3-C2</b>
	A	B	C	D	
1	Compensation	Frequency	Cumulative	%	
2	2.0	312	69.96%	69.96%	
3	5.0	99	92.15%	22.20%	
4	8.0	17	95.96%	3.81%	
5	11.0	5	97.09%	1.12%	
6	More	13	100.00%	2.91%	

	A	
	Total Compensation (M\$)	411
1		Frequency
2	<2 M	
3	3 to 5 M+SD	
4	6 to 8 M+2SD	
5	9 to 11 M+3SD	
6	More	
7		

To increase readability of the histogram, change category labels in column A to indicate ranges and add M, M+SD, M+2SD, M+3SD:

The histogram will plot percents in column D by categories in column A. Move column D to column B.

From column D  
**Cntl+spacebar**,  
**Cntl+X**  
 left arrow key to column B **Cntl+spacebar**  
**Alt HIE**

B	C	D
		<i>Cumulative</i>
<i>%</i>	<i>Frequency</i>	<i>%</i>
69.96%	312	69.96%
22.20%	99	92.15%
3.81%	17	95.96%
1.12%	5	97.09%
2.91%	13	100.00%

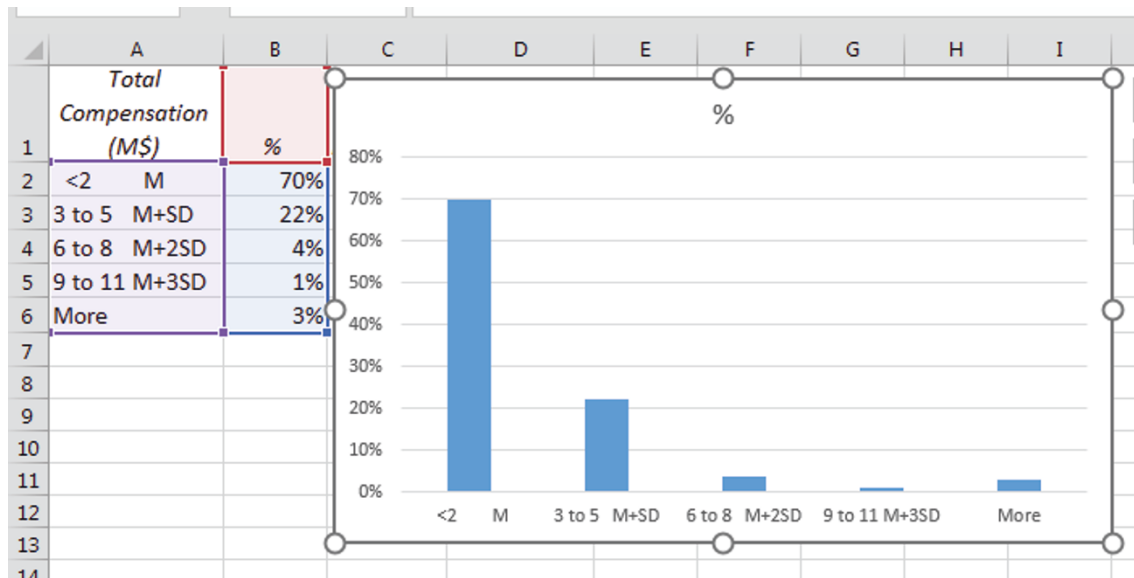
Excel often shows more decimals than are desired. Select the percents in column B and reduce decimals.

From B2  
**Cntl+shift+down**  
**Alt H9**

B
<i>%</i>
<i>Fre</i>
70%
22%
4%
1%
3%

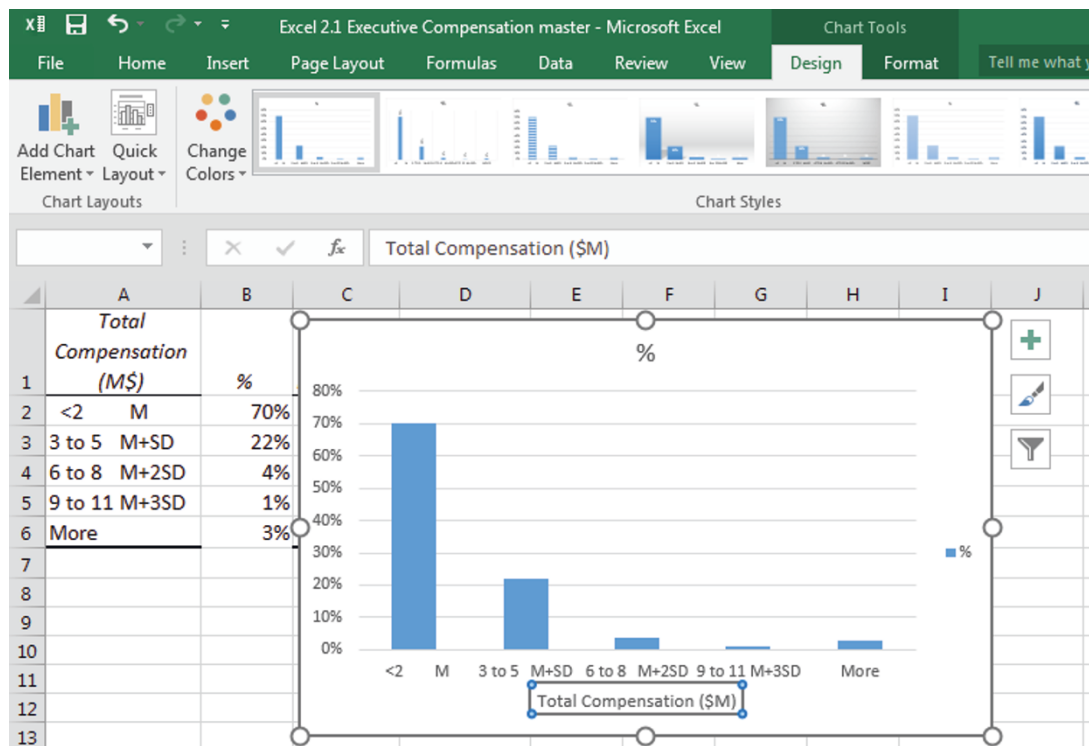
Produce the histogram by selecting data in columns A and B, then request a column chart.

From A1  
**Cntl+shift+down**  
**Shift+right**  
**Alt NC**



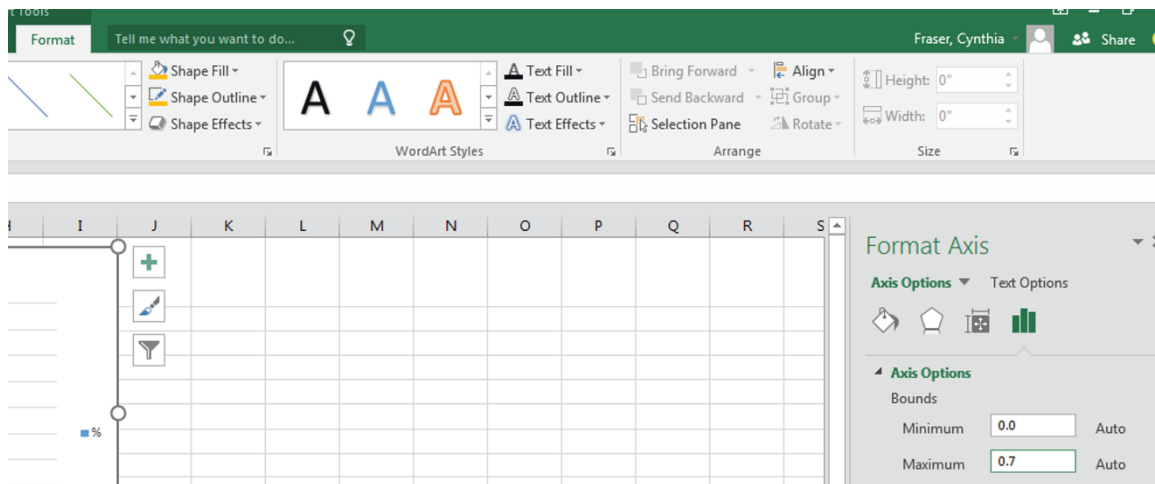
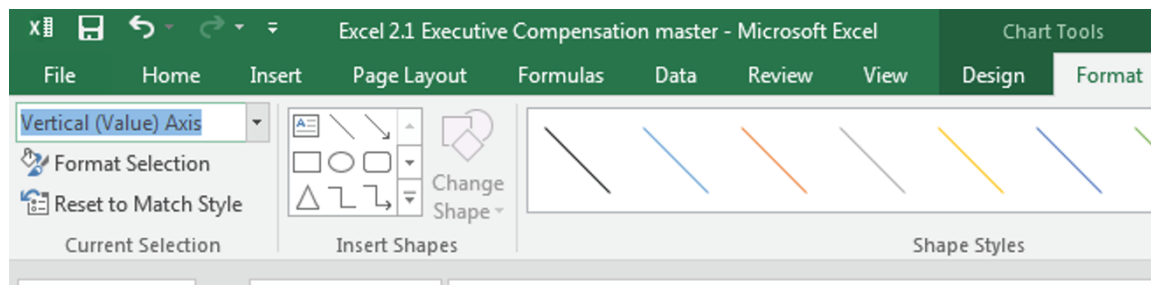
Add a horizontal axis title.

| Alt JCAAH



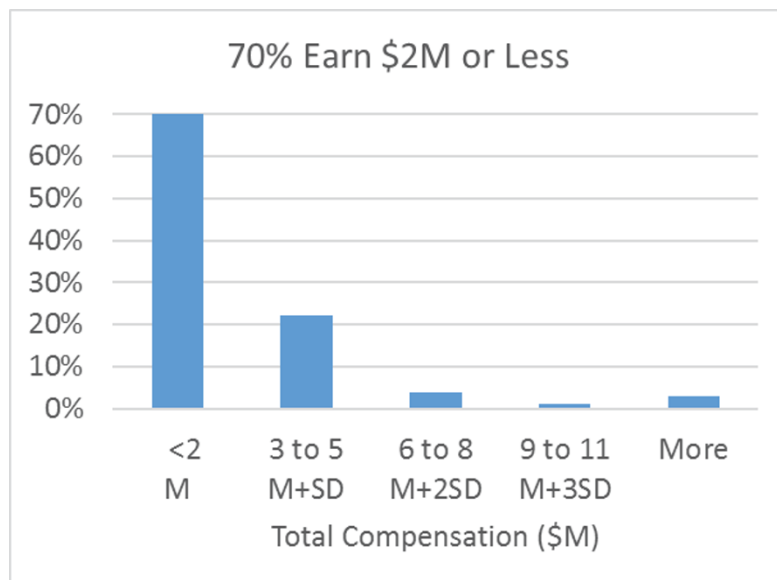
To make better use of the chart space, reformat the vertical axis, setting the maximum to 70%.

| Alt JAE down to Vertical (Value) Axis Alt JAM



Increase the font size.

Click the outside edge of the chart, then  
**Alt HFS 12**



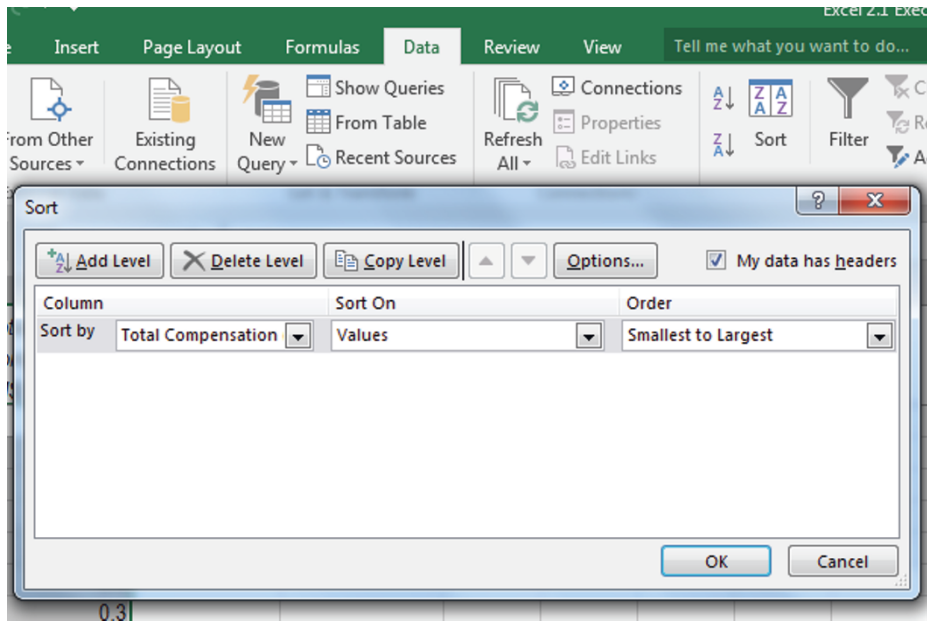
Delete the legend and replace the chart title with a stand alone title:



## Excel 2.2 Sort to Produce Descriptives Without Outliers

To easily identify and remove outliers, sort the rows from lowest to highest *total compensation (\$M)*. Move back to the data page, select *total compensation* data in column **B** (but not the histogram bins below the data), then use shortcuts to sort:

**Cntl+Page Down**  
From B1,  
**Cntl+Shift+Down**  
**Alt ASS**



Move to the end of B, and then scroll up to identify the rows with compensation within 3SDs of the mean, 11.0 or less.

**Cntl+Down**  
**Up arrow**

	A	B	C
		<i>Total Compensation (M\$)</i>	
1			
433		10.1	
434		11.0	
435		11.7	

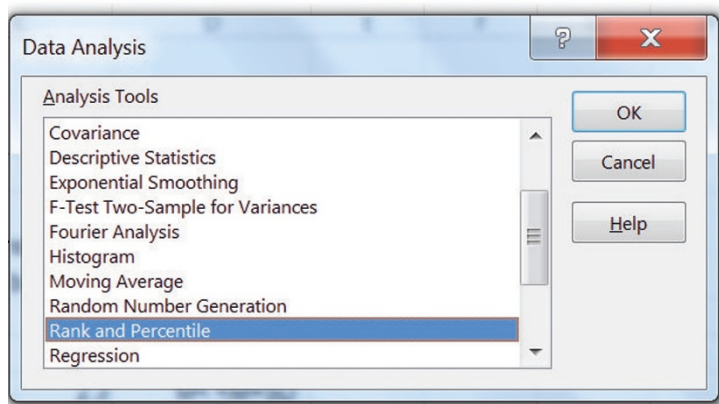
Run descriptives, again, changing the input range to b1:b434.

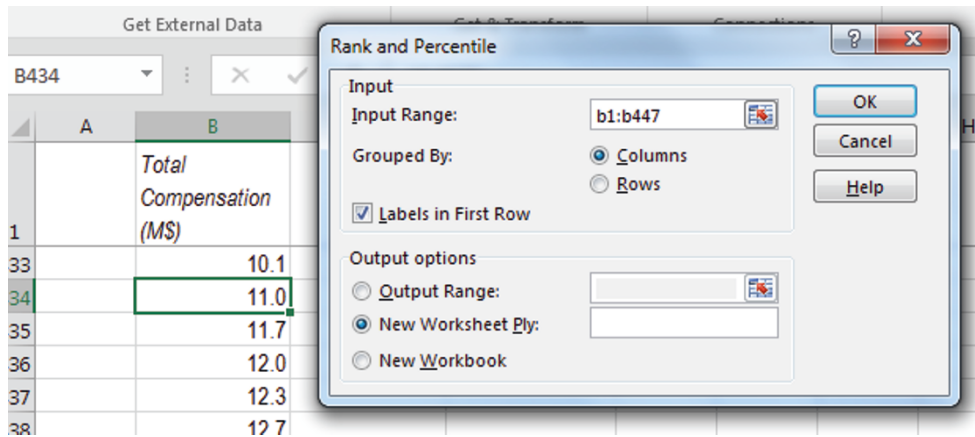
	A	B
1	Total Compensation (M\$)	
2		
3	Mean	1.69028
4	Standard Error	0.076304
5	Median	1.11245
6	Mode	#N/A
7	Standard Deviation	1.58779
8	Sample Variance	2.521079
9	Kurtosis	7.517214
10	Skewness	2.441577
11	Range	10.93838
12	Minimum	0.028816
13	Maximum	10.9672
14	Sum	731.8914
15	Count	433
16		

## Excel 2.3 Plot a Cumulative Distribution

Return to the data page and request the cumulative distribution of total compensation.

**Ctrl+Page Dn**  
**Alt AYn, R down**  
**B1:b447 tab L**





	A	B	C	D
1	Point	Compensation	Rank	Percent
2	446	32.6	1	100.00%
3	445	20.7	2	99.70%
4	444	16.2	3	99.50%
5	443	15.9	4	99.30%
6	442	15.7	5	99.10%
7	441	14.9	6	98.80%
8	440	14.7	7	98.60%

Excel will plot cumulative percents in column D by compensation in column B. For convenience, delete column C. Select the cumulative percent data now in column C and reduce decimals, and then select columns B and C and insert a scatterplot showing the cumulative distribution.

From column C,

**Alt HDC.**

(**H** selects the **H**ome menu, **D** selects the **D**elete menu, and **C** deletes the **C**olumn.)

From C2

**Cntl+shift+down**

**Alt H9**

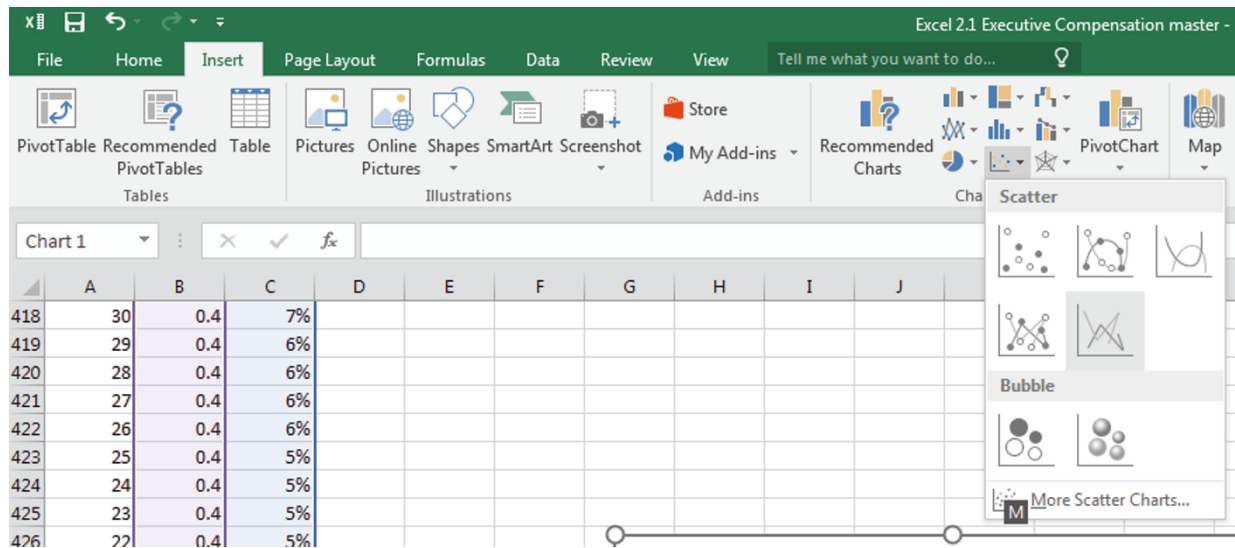
From B1

**Cntl+shift+down**

**Shift+right**

**Alt ND**

Choose the last scatter option.



Use shortcuts to choose design layout 1 to add axes labels and title. Delete the legend and adjust font size to 12.

**Alt JCL**  
**Alt HFS 12**

Use shortcuts to select and format axes. Set the vertical axis maximum at 100%. Set the horizontal axis maximum at 11, with 0 decimals.

**Alt JAE**  
**Alt JAM**  
**Alt JCAGV**

Use shortcuts to add vertical gridlines:

**Format Axis**

**Axis Options** ▾ Text Options

Maximum:  Reset

Units

Major:  Auto

Minor:  Auto

Vertical axis crosses

☒ Automatic

☐ Axis value

☐ Maximum axis value

Display units:

☐ Show display units label on chart

☐ Logarithmic scale Base:

☐ Values in reverse order

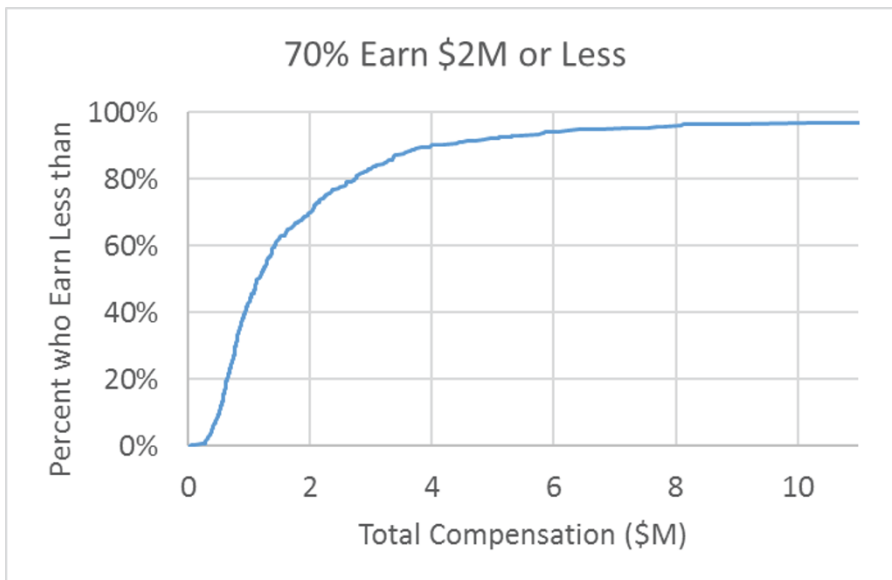
▸ Tick Marks

▸ Labels

▾ Number

Category:

Decimal places:



## Excel 2.4 Use a PivotTable to Sort by Industry

Use a PivotTable to sort the 200 best paid CEOs by industry. Open **Lab 2 Highest Paid CEOs 2014**.

Select compensation and gender data in columns C and D. Insert a PivotTable to compare compensation by gender.

From C1,  
**Cntl+shift+down**  
**Shift+right**  
**Alt NV**  
 Drag Gender to the Rows.  
 Drag Compensation to the Sum Values

File Home Insert Page Layout Formulas Data Review				
PivotTable Recommended Table Pictures Online Pictures Screenshot Illustrations Store My Add-ins				
C1 : x ✓ fx CEO Compensation (\$M)				
	A	B	C	D
	firm	2014 Revenue (\$B)	CEO Compensation (\$M)	Gender
1	21st Centu	11.1	23.9	Male
2	3M	56.2	14.3	Female
3	Abbott Lal	9.1	16.2	Male

**PivotTable Fields**

Choose fields to add to report:

Search

☒ CEO Compensation (\$M)  
☒ Gender  
 MORE TABLES...

Drag fields between areas below:

**FILTERS**

**ROWS**  
 Gender

**COLUMNS**

**VALUES**  
 Sum of CEO Compensation (\$M)

Row Labels	Sum of CEO Compensation (\$M)
Female	225.7
Male	4409.2
<b>Grand Total</b>	<b>4634.9</b>

Excel shows the sums in PivotTables.

Convert the sums to averages.  
 Alt JTG tab tab down to Average

**PivotTable**

Active Field

B5

23.4531914893617

Row Labels	Average of CEO Compensation (\$M)
Female	18.80833333
Male	23.45319149
<b>Grand Total</b>	<b>23.1745</b>

**Value Field Settings**

Source Name: CEO Compensation (\$M)

Custom Name: Average of CEO Compensation (\$M)

Summarize Values By: Show Values As

**Summarize value field by**

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

Sum  
 Count  
**Average**  
 Max  
 Min  
 Product

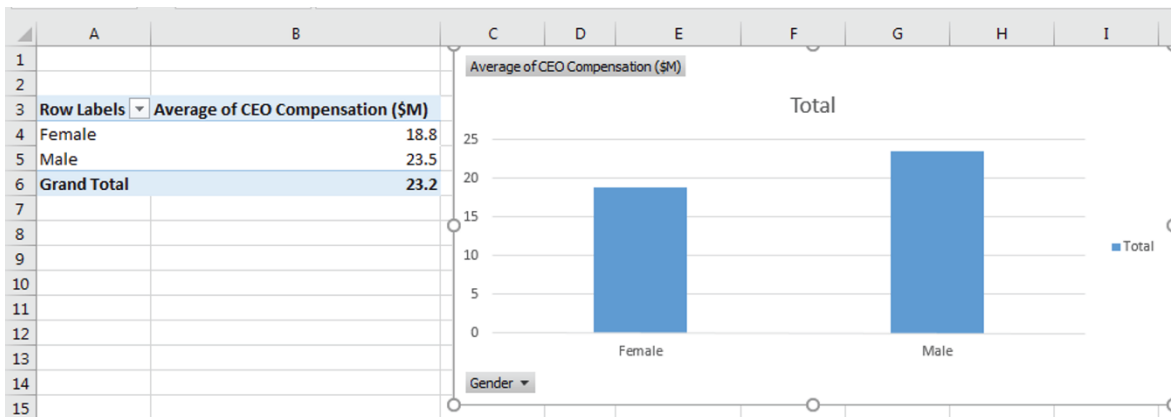
Number Format OK Cancel

Reduce decimals.

From B4,  
**Shift+down**  
**Alt H9**

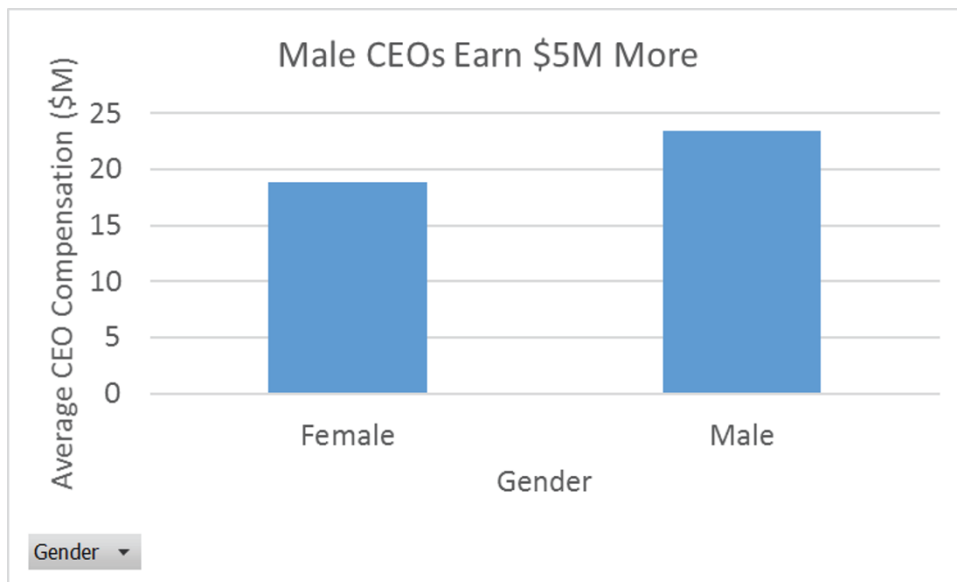
Show compensation by gender in a PivotChart.

**Alt JTC**



Choose the ninth design layout, add axis titles and a stand alone chart title, delete the legend, and adjust fontsize to 12.

**Alt JCL**  
**Alt HFS 12**

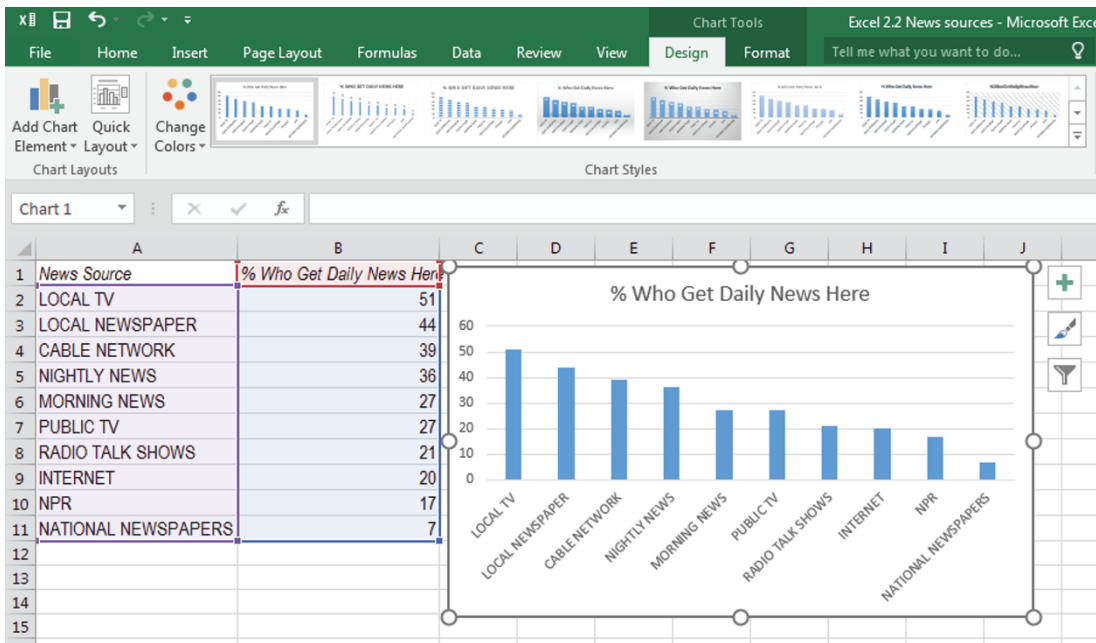


## Excel 2.5 Produce a Column Chart of a Nominal Variable

To show percents who choose alternate media for daily news, produce a column chart from Gallup Poll of 992 Americans.

Open **Excel 2.2 News Sources**, select the **News Source** and **% Who Get Daily News Here** data, and insert a column chart.

**Alt NC**

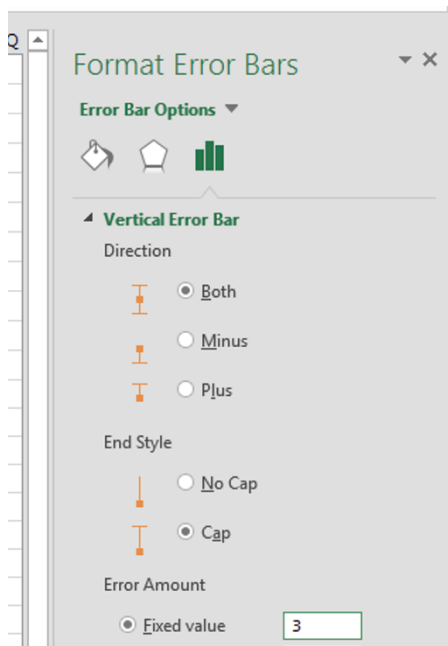


Choose **Design Chart Layout 9** and type in a stand alone title and axes titles.

| Alt JCL

Add vertical margin of error bars fixed at the approximate margin of error, 3.

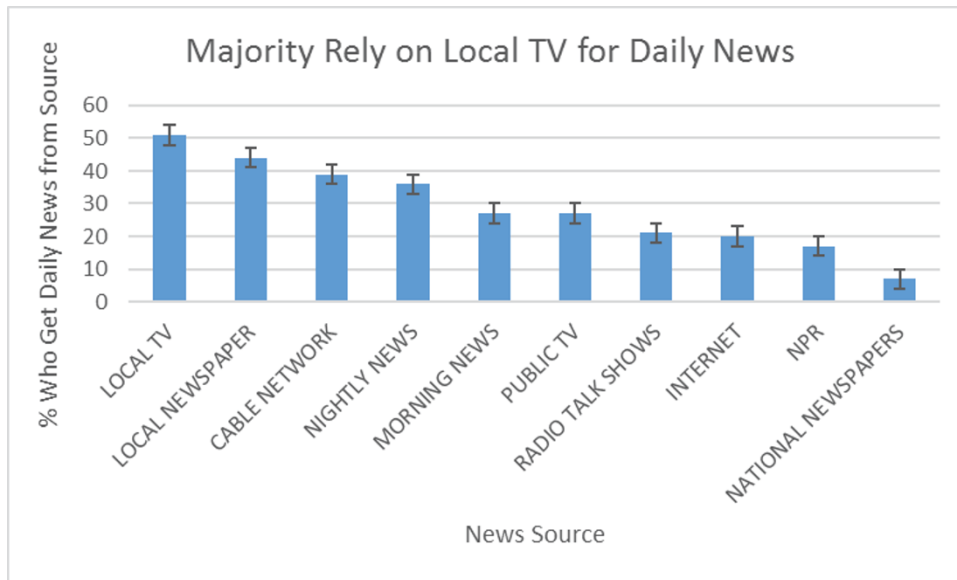
| Alt JCAEM



Set fontsize to 12.

| Alt HFS 12





## Excel Shortcuts Used in Chapter 2

### *Home menu shortcuts*

Insert cElls that were copied or cut  
set FontSize  
reduce decimals  
Delete a Column

Alt HIE  
Alt HFS  
Alt H9  
Alt HDC

### *Insert menu shortcuts*

iNsert a Column chart  
iNsert a scatterplot  
iNsert a PiVotTable

Alt NC  
Alt ND  
Alt NV

### *Data menu shortcuts*

analyze dAta  
Sort selected dAta

Alt AYn  
Alt ASS

### *View menu shortcuts*

Freeze top Row

Alt WFR

*Cntl+ to move, select an array, extend formula down an array, or cut selected array*

move to the page right

Cntl+Page  
Down

move to bottom of data array

Cntl+down  
arrow

select data below

cntl+shift+down

select a column

cntl+spacebar

fill down

cntl+D

cut selected cells

cntl+X

*Shift+ to select adjacent cells*

select adjacent cells

shift+down  
arrow  
shift+right

*Chart or scatterplot design*

add a horizontal axis title  
select a design layout  
add vertical gridlines  
add vertical margin of error bars

**Alt JCAAH**  
**Alt JCL**  
**Alt JCAGV**  
**Alt JCAEM**

*Chart or scatterplot element selection, formatting*

select an axis  
format selected chart element

**Alt JAE**  
**Alt JAM**

*Reformat or graph a PivotTable*

show averages instead of sums  
produce a PivotChart

**Alt JTG**  
**Alt JTC**

*Other*

turn on Add-In

**Alt TI**

**Alt** activates shortcuts menus, linking keyboard letters to Excel menus. Press and press letters linked to the menus you want.

**Alt H**ome:



Home menu leys, from left to right, include:

V	Paste	FF	Choose a font	FS	Choose a fontsize	W	Wrap text	9	Reduce decimals	I	Insert
X	Cut	1	Bold	FC	Choose font color					D	Delete
C	Copy	2	Italicize								
		3	Underline								

Other useful menus activated with **Alt** include:

A	Data	N	Insert	W	View
---	------	---	--------	---	------

From a chart or plot, **Alt** provides access to chart menus:

JC	Chart design	JA	Chart format	JT	Reformat PivotTable data or chart
----	--------------	----	--------------	----	-----------------------------------

## Significant Digits Guidelines

The number of significant digits in a number are those which convey information. Significant digits include:

1. All nonzero numbers
2. Zeros between nonzero numbers, and
3. Trailing zeros.

Zeros acting as placeholders aren't counted.

The number 2061 has four significant digits, while the number 2610 has three, since the zero is merely a placeholder. The number 0.0920 has three significant digits, "9," "2," and the final, trailing "0." The first two zeros are placeholders that aren't counted.

In rare cases, it is not clear whether zero is a placeholder or a significant digit. The number 40,000 could represent the range 39,500 to 40,499. In that case, the number of significant digits is one, and the zeros are placeholders. Alternatively, 40,000 could represent the range 39,995 to 40,004. In this latter case, the number of significant digits is four, since the zeros convey meaning. When in doubt, a number could be written in scientific notation, which is unambiguous. For one significant digit, 40,000 becomes  $4 \times E^4$ . For four significant digits, 40,000 becomes  $4.000 \times E^4$ .



## Lab 2 Description

### Compensation of 200 Best Paid CEOs

The New York Times recently published the compensation packages of the 200 best compensated CEOs of publicly traded firms in the U.S. These data are in **Lab 2 Compensation of Best Paid CEOs**.

#### I. Describe the compensation of the best paid CEOs.

1. Find the average compensation (M) among the best compensated CEOs: \_\_\_\_\_
2. Find the standard deviation (SD) of compensation: \_\_\_\_\_
3. Is the distribution of compensation among the best paid CEOs approximately Normal?

Y or N Evidence: \_\_\_\_\_

#### II. Identify outlier(s) who earn(s) more than 3 SDs above the M and describe compensation of best paid CEOs excluding outliers.

1. Find average compensation, M, excluding outlier(s): \_\_\_\_\_
2. Find the standard deviation of compensation, SD, excluding outlier(s): \_\_\_\_\_

#### III. Make a histogram and cdf to illustrate distribution of CEO compensation

1. Make the histogram of compensation for top paid CEOs.
2. Plot the cumulative distribution of compensation.
3. What is median compensation among the best paid CEOs? \_\_\_\_\_
4. What is the Interquartile Range of compensation among the 25 best paid CEOs?

\_\_\_\_\_

#### IV. Compare the Distribution of CEO Compensation to Normal

1. By the Empirical Rule, Normal is distributed as in second and third columns, below. Fill in the third through fifth columns to compare the distribution of CEO compensation to Normal:

	Normal		CEO compensation		
Range	%	Cum %	Range (\$M)	%	Cumulative %
Below M-3SD	.1	.1			
M-3SD to M-2SD	2.1	2.2			
M-2SD to M-1SD	13.6	15.8			
M-1SD to M	34.1	50			
M to M+1SD	34.1	84.1			
M+1SD to M+2SD	13.6	97.7			
M+2SD to M+3SD	2.1	99.9			
Above M+3SD	.1	100			

2. By the Empirical Rule, Normal is distributed as in the second column, below. Fill in the third and fourth columns to compare the distribution of CEO compensation to Normal:

	Normal	CEO compensation	
Range	%	Range (\$M)	%
Within 1SD of M			
Within 2SD of M			
Within 3 SD of M			

#### V. Identify Industries where CEOs are Best Compensated

Use a PivotTable to determine the best paid industry.

1. What is the best paid industry? \_\_\_\_\_
2. Average CEO compensation within industries range: \_\_\_\_ to \_\_\_\_



## Assignment 2.1 Procter & Gamble's Global Advertising

Procter & Gamble spent \$5,960,000 on advertising in 51 global markets. This data, from *Advertising Age*, Global Marketing is in **Assignment 2.1 P&G Global Advertising**.

P&G Corporate is reviewing the firm's global advertising strategy, which is the result of decisions made by many brand management teams. Corporate wants to be sure that these many brand level decisions produce an effective allocation when viewed together.

Describe *Procter & Gamble's* advertising spending across the 51 *countries* that make up the global markets.

*Note: Be specific: label responses with appropriate units! Important! Also round responses to two or three significant digits. Points deducted for missing/incorrect units or too few/too many significant digits.*

### I. Describe P&G's global advertising.

1. Find the average advertising (M) in countries Worldwide: \_\_\_\_\_
2. Find the standard deviation (SD) of advertising in countries Worldwide: \_\_\_\_\_
3. Is the distribution of advertising across countries Worldwide approximately Normal?

Y or N Evidence: \_\_\_\_\_

### II. Identify *countries* which are outliers and list them here:

### III. Illustrate the distribution of advertising levels in countries with a histogram and a cdf. Reduce decimals, add axis labels, and add a "stand alone" chart title. *Points deducted for titles that aren't "stand alone."*

1. What is median advertising across countries Worldwide? \_\_\_\_\_
2. What is the Interquartile Range of across countries Worldwide? \_\_\_\_\_

#### IV. Compare the Distribution of P&G's Advertising to Normal

1. By the Empirical Rule, Normal is distributed as in second and third columns, below. Fill in the third through fifth columns to compare the distribution of CEO compensation to Normal:

	Normal		P&G's Advertising		
Range	%	Cum %	Range (\$M)	%	Cumulative %
Below M-3SD	.1	.1			
M-3SD to M-2SD	2.1	2.2			
M-2SD to M-1SD	13.6	15.8			
M-1SD to M	34.1	50			
M to M+1SD	34.1	84.1			
M+1SD to M+2SD	13.6	97.7			
M+2SD to M+3SD	2.1	99.9			
Above M+3SD	.1	100			

2. By the Empirical Rule, Normal is distributed as in the second column, below. Fill in the third and fourth columns to compare the distribution of P&G advertising to Normal:

	Normal	CEO compensation	
Range	%	Range (\$M)	%
Within 1SD of M	68		
Within 2SD of M	95		
Within 3 SD of M	99.7		

#### V. Conclusions

1. Make a Pivot table of advertising by level of development, and then make a Pivot chart and paste here, after reducing decimals, adding axis labels, and a SAS title:
2. Which advertising strategy describes the P&G strategy best: (i) advertise at a moderate level in many countries, (ii) advertise heavily in a small number of key countries and spend much less in many other markets.

### Assignment 2.2 Best Practices Survey

Firm managers use statistics to advantage. Sometimes when results are lackluster, more significant digits are used, since readers will spend less time digesting results, and results with more significant digits are less likely to be remembered. Sometimes when results are impressive, fewer significant digits are used to motivate readers to digest and remember.

Choose an Annual Report and cite the firm and the year.

1. In the body of the report, what range of significant digits are used to report numerical results? Cite two examples, one with the smallest number of significant digits, one with the largest number of significant digits.
2. In the Financial Exhibits at the end, what range of significant digits are used? Cite two examples, one with the smallest number of significant digits, and one with the largest number of significant digits.
3. Survey the graphics. Cite an example where stand alone title is used to help readers interpret. Cite an example where the title could be more effective, and provide a suggestion for a better title.

## Assignment 2.3 Shortcut Challenge

Complete the steps in the first Excel page of Lab 2 (find descriptive statistics, sort to identify and remove outliers, find descriptive statistics without outliers, make a histogram, plot the cdf, make a PivotTable, make a PivotChart), and record your time. If your time is more than 5 minutes, repeat twice, and then record your best time.

## Case 2.1 VW Backgrounds

Volkswagon management comissioned background music for New Beetle commercials. The advertising message is that the New Beetle is unique... “round in a world of squares.” To be effective, the background music must support this message.

Thirty customers were asked to write down the first word that came to mind when they listened to the music. The clip is in **Case 2.1 VW background.MP3** and words evoked are contained in **Case 2-1 VW background**. Listen to the clip, then describe market response.

Create a PivotTable of the percent who associate each image with the music and sort rows so that the modal image is first.

1. Create a PivotChart to illustrate the images associated with the background music. (Add a stand alone title and round percentages to two significant digits.)
2. What is the modal image created by the VW commercial’s background music?
3. Is this music is a good choice for the VW commercial? Explain.

## Case 2.2 Global Smelter Costs at Alcoa

Faced with recent expansion in Chinese aluminum production, Alcoa seeks to identify smelters that are less profitable. Data in **Alcoa Smelter Costs** contain costs for nine smelters... four which have been closed or curtailed, four which are candidates for closure or curtailment, and one benchmark smelter in which management plans to continue operations. Profit drivers are largely cost based and include total unit costs, labor and power costs per unit. Describe smelter costs at the nine Alcoa smelters. Data are in **AlcoaSmelterCosts**.

*Note: label responses with appropriate units! Important! Also round responses to two or three significant digits.*

**I. Describe smelter unit costs.**

1. Find the average (M) total, labor and power costs per unit: \_\_\_\_\_
2. Find the standard deviation (SD) of total, labor and power costs per unit: \_\_\_\_\_
3. Are the distributions of total, labor and power costs per unit approximately Normal?

Total: Y or N Evidence: \_\_\_\_\_ Labor: Y or N Evidence: \_\_\_\_\_

Power: Y or N Evidence: \_\_\_\_\_

**II. Identify *smelters* which are outliers in terms of total, labor or power costs per unit and list them here:**

**III. Illustrate the distributions of total, labor and power costs per unit with histograms.**

1. Round the unit cost bins to two significant digits. Reduce decimals, add axis labels, and add a “stand alone” chart titles. Paste your three graphs here:
2. By the Empirical Rule, Normal is distributed as in second and third columns, below. Fill in the third through fifth columns to compare the distribution of units costs to Normal:

	Normal		Total Unit Cost		
Range	%	Cum %	Range (\$/t)	%	Cumulative %
Below M-3SD	.1	.1			
M-3SD to M-2SD	2.1	2.2			
M-2SD to M-1SD	13.6	15.8			
M-1SD to M	34.1	50			
M to M+1SD	34.1	84.1			
M+1SD to M+2SD	13.6	97.7			
M+2SD to M+3SD	2.1	99.9			
Above M+3SD	.1	100			

	Normal		Unit Labor Cost		
Range	%	Cum %	Range (\$/t)	%	Cumulative %
Below M-3SD	.1	.1			
M-3SD to M-2SD	2.1	2.2			
M-2SD to M-1SD	13.6	15.8			
M-1SD to M	34.1	50			
M to M+1SD	34.1	84.1			
M+1SD to M+2SD	13.6	97.7			
M+2SD to M+3SD	2.1	99.9			
Above M+3SD	.1	100			

	Normal		Unit Power Cost		
Range	%	Cum %	Range (\$/t)	%	Cumulative %
Below M-3SD	.1	.1			
M-3SD to M-2SD	2.1	2.2			
M-2SD to M-1SD	13.6	15.8			
M-1SD to M	34.1	50			
M to M+1SD	34.1	84.1			
M+1SD to M+2SD	13.6	97.7			
M+2SD to M+3SD	2.1	99.9			
Above M+3SD	.1	100			

3. By the Empirical Rule, Normal is distributed as in the second column, below. Fill in the third and fourth columns to compare the distribution of units costs to Normal:

	Normal	Total Unit Cost	
Range	%	Range (\$/t)	%
Within 1SD of M	68		
Within 2SD of M	95		
H1	99.7		

	Normal	Unit Labor Cost	
Range	%	Range (\$/t)	%
Within 1SD of M	68		
Within 2SD of M	95		
Within 3 SD of M	99.7		

	Normal	Unit Power Cost	
Range	%	Range (\$/t)	%
Within 1SD of M	68		
Within 2SD of M	95		
Within 3 SD of M	99.7		

#### IV. Illustrate the distributions of total, labor and power costs per unit with cdfs.

- Paste your cdfs plots here:
- What are median unit costs across smelters? Total: \_\_\_\_\_ Labor: \_\_\_\_\_  
Power: \_\_\_\_\_
- What is are the Interquartile Ranges of unit costs across smelters?  
Total: \_\_\_\_\_ to \_\_\_\_\_ Labor: \_\_\_\_\_ to \_\_\_\_\_ Power: \_\_\_\_\_ to \_\_\_\_\_

## **V. Conclusions**

1. Make Pivot tables of total, labor and power costs per unit by operating status, location, and process ( $3 \times 3 = 9$  total), and then make Pivot charts (9 total) and paste here, after reducing decimals, adding axis labels, and stand alone titles:
2. Describe an ideal, smelter with differential advantages which may lead to lower unit costs:

<http://www.springer.com/978-3-319-32184-4>

Business Statistics for Competitive Advantage with  
Excel 2016

Basics, Model Building, Simulation and Cases

Fraser, C.

2016, XIV, 475 p. 375 illus., 370 illus. in color.,

Hardcover

ISBN: 978-3-319-32184-4