

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 Figs. 2.1 and 2.2. A histogram illustrates central tendency, dispersion, and symmetry.

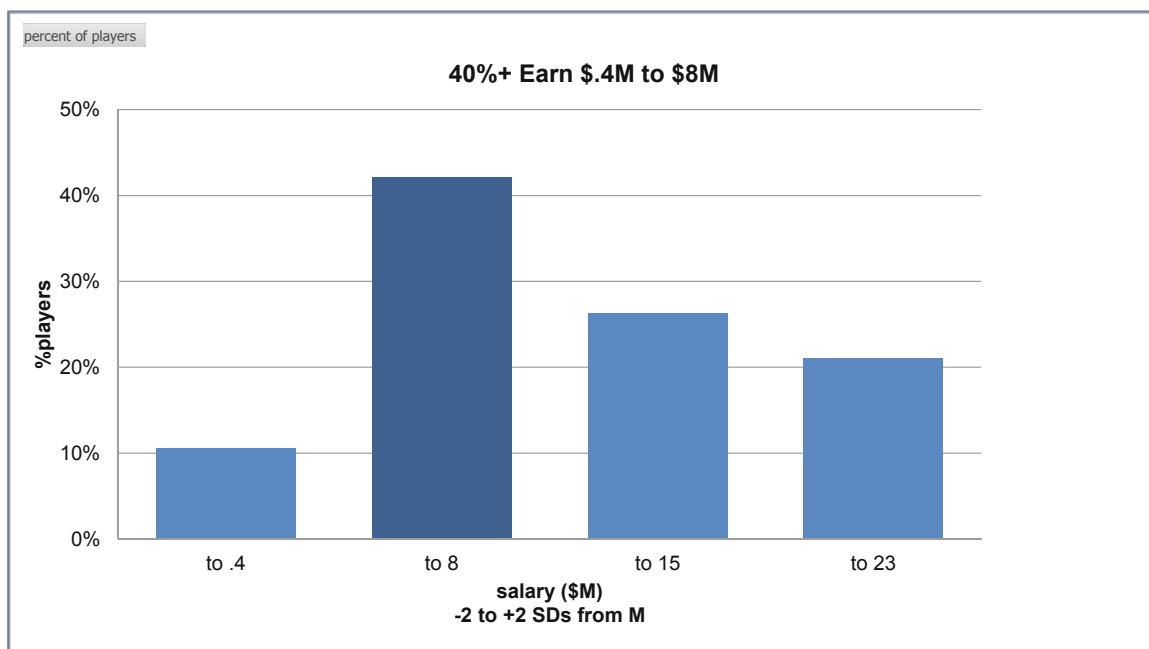


Fig. 2.1 Histogram of Yankee salaries

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.

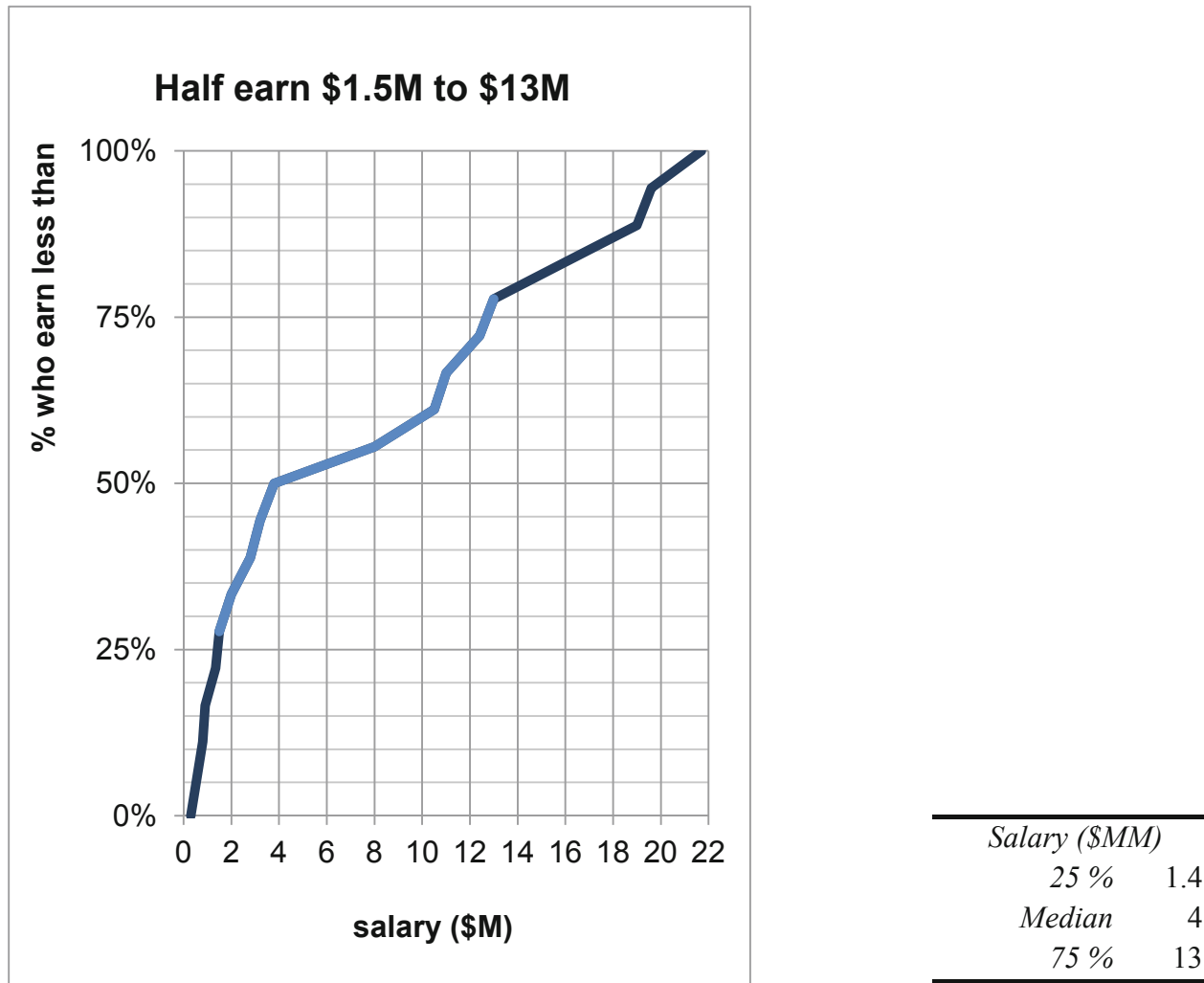
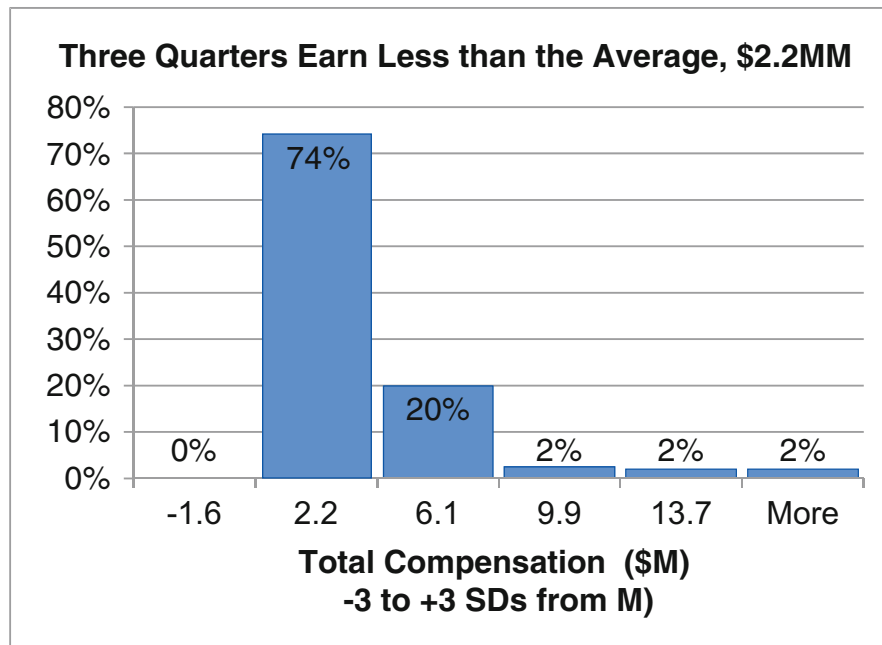


Fig. 2.2 Cumulative distribution of salaries

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 \$1.5M 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 [Figs. 2.3](#) and [2.4](#).



<i>Total compensation (-1 to +3 sds from M)</i>	<i>% Executives</i>
2	74 %
6	20 %
10	2 %
14	2 %
More	2 %

Fig. 2.3 Histogram of executive compensation

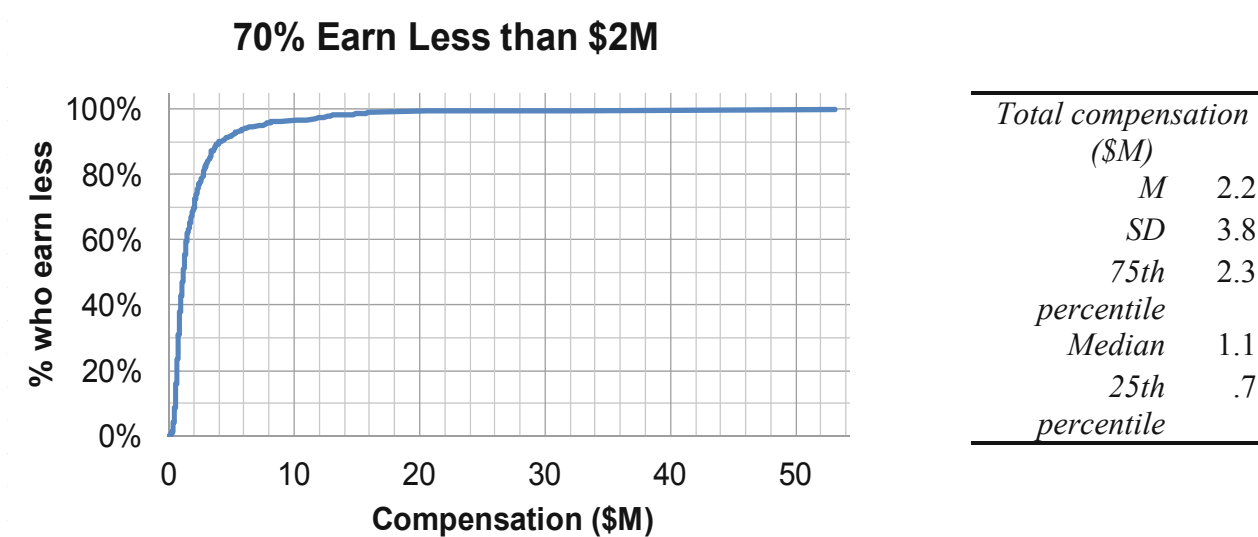


Fig. 2.4 Cumulative distribution of total compensation

The average executive compensation in this sample of large corporations is \$2.2M. 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 :

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

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

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 , σ 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 (Fig. 2.5):

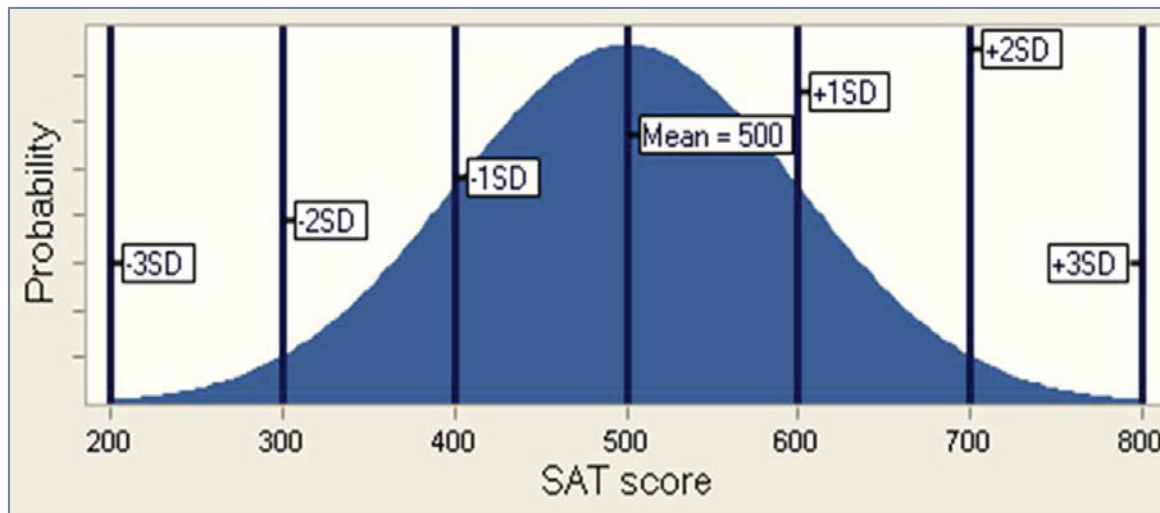


Fig. 2.5 Normally distributed SAT scores

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 '10 SATs: This Class Is 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 Fig. 2.6. Skewness is $-.5$, indicating approximate symmetry, an approximately Normal distribution.

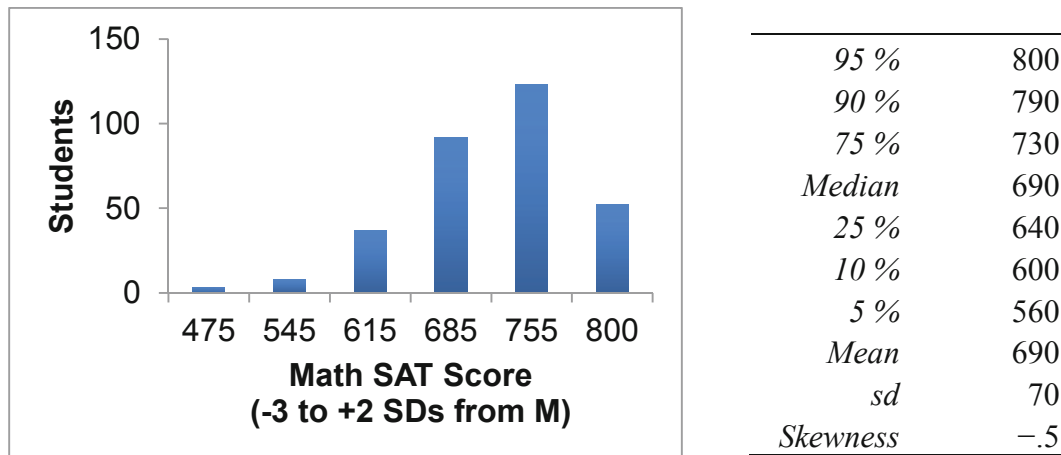


Fig. 2.6 Histograms and descriptive statistics of class '06 math SATs

Class '10 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 of 70 points of the mean of 690, or within the interval 620–760. 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 4 % actually do have scores below 530, 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 Class '10 students' SAT scores, report:

- Class '10 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 Class '10 students' Math SAT scores, 70 versus 100, indicates that Class '06 students are 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 eight *outliers*, shown as *MORE* than three standard deviations above the mean in Fig. 2.3, who are compensated extraordinarily well. Each collects a compensation package of more than \$14M, a compensation level that is more than three standard deviations greater than the mean.

When we exclude these eight outliers, eleven additional outliers emerge. This cycle repeats, since the distribution is highly skewed. When we remove outliers, the new mean is adjusted, making other executives appear to be more extreme. After removing about ten percent, or the 44 best compensated executives, we see a clearer picture of what “typical” compensation is, shown in Fig. 2.7:

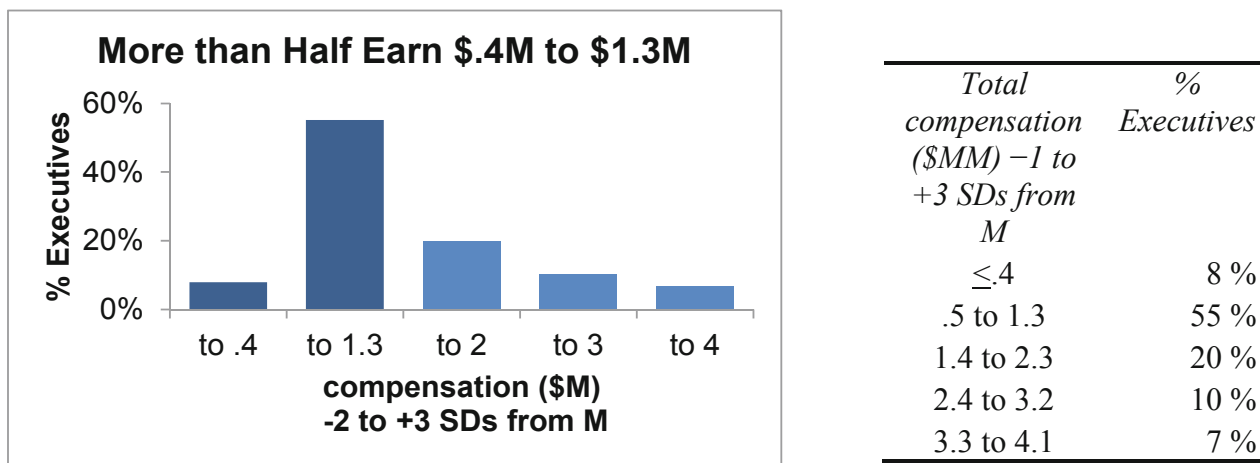


Fig. 2.7 Histogram and descriptive statistics with 44 outliers excluded

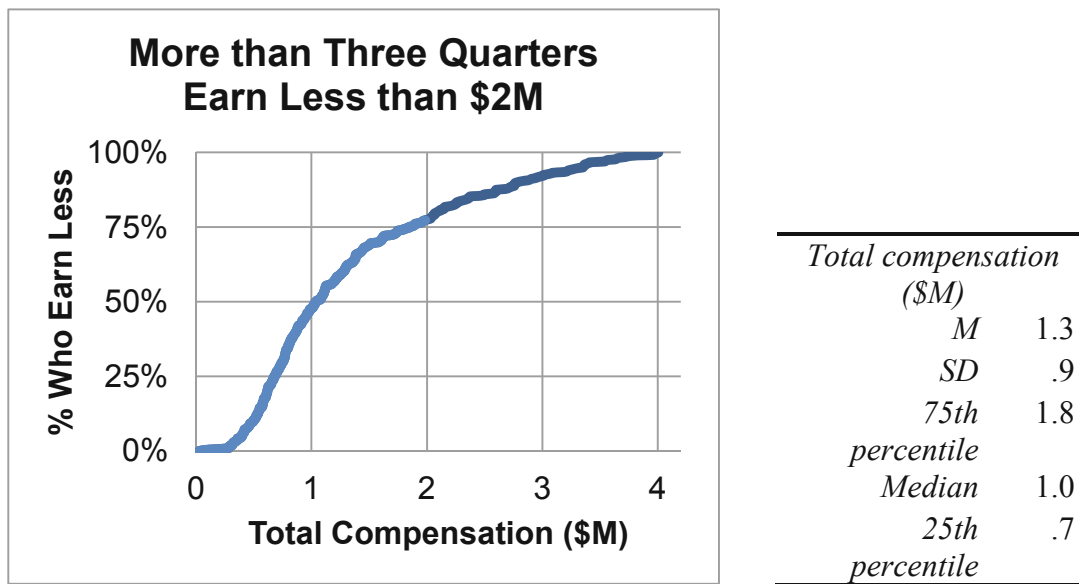


Fig. 2.8 Cumulative distribution of total compensation

Ignoring the 44 outliers, the average compensation is about \$1.3M, and the *median* compensation is about \$1M, shown in Fig. 2.8. The *mean* and *median* are closer. With this more representative description of executive compensation in large corporations, The Board has an indication that the \$2M package is well above average. More than three quarters of executives earn less. Because extraordinary executives exist, the original 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 were approximately symmetric and Normal; Executive compensation values were skewed, until outliers were removed. 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: Column and PivotCharts

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 2.9 shows a column chart of results of a survey of 1,014 adults by Gallup:

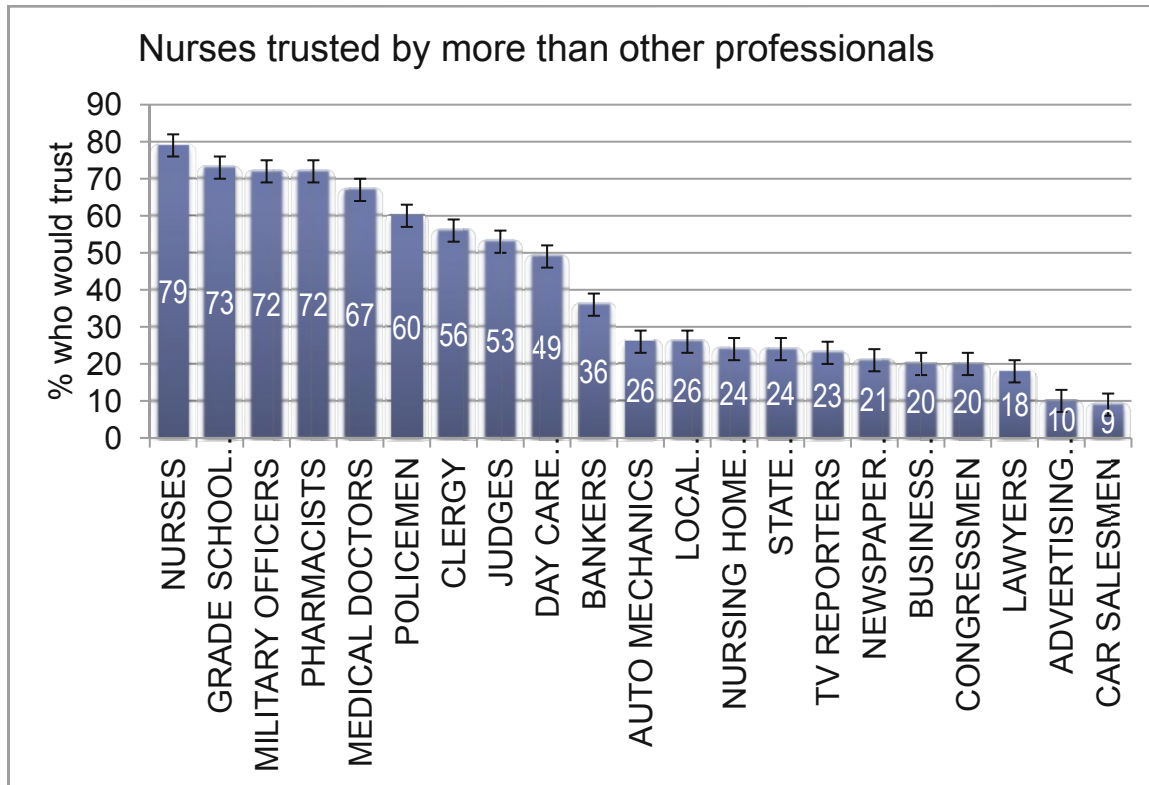


Fig. 2.9 Pareto charts of the percents who judge professions honest

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

	Quantitative	Categorical
Central tendency	<i>Mean</i>	<i>Mode</i>
	<i>Median</i>	<i>Proportion</i>
Dispersion	<i>Range</i>	
	<i>Standard deviation</i>	
Symmetry	<i>Skewness</i>	
Graphics	<i>Histogram</i>	<i>Pareto chart</i>
	<i>Cumulative distribution</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.

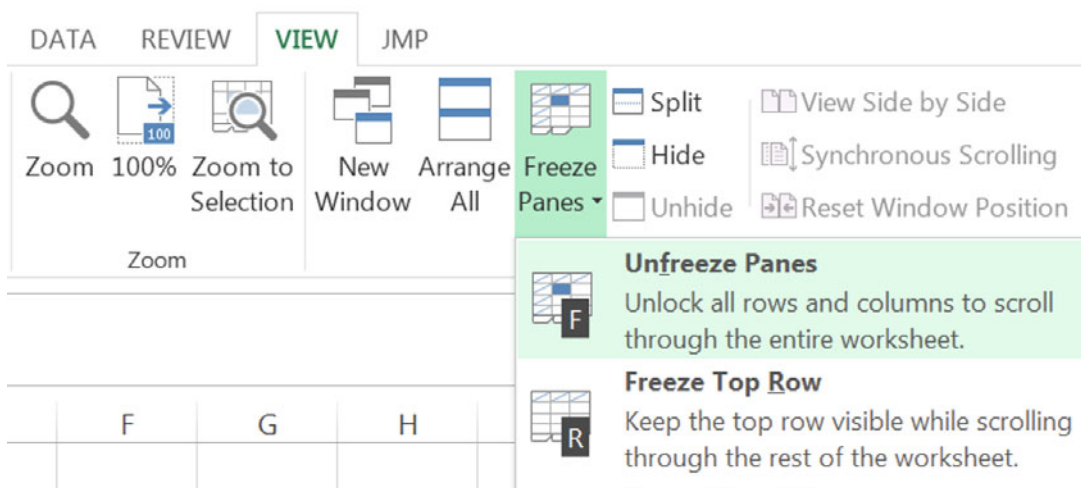
Effective results are those which are remembered and used to improve decision making. Your presentation of results will influence whether or not decision makers remember and use your results. Round statistics to two or three significant digits to make them honest, digestible, and memorable. Title your 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.xls** so that column labels are visible when you are at the bottom of the dataset. Select the first cell, **A1**, and then use Excel shortcuts **Alt WFR**. (The shortcuts, activated with **Alt** select the vieW menu, the Freeze panes menu, and then freeze Rows.)



Use shortcuts to move to the end of the file where we will add descriptive statistics. **Cntl+down arrow** scrolls through all cells containing data in the same column and stops at the last filled cell.

Descriptive statistics. In the first empty cell in the column, below the data, use shortcuts to find the sample mean: **Alt MUA**.

Use the:

STDEV.S(array) function to find the standard deviation,

PERCENTILE.INC(array, .75) and **PERCENTILE.INC(array, .25)** to find the 75th and 25th percentile values,

MEDIAN(array) function to find the median, and

SKEW(array) function to find skewness:

B454 : =SKEW(B2:B448)				
	A	B	C	D
1		Total Compensation (MM\$)		
448		53.1		
449	M	2.2		
450	SD	3.8		
451	75%	2.3		
452	median	1.1		
453	25%	0.72		
454	skew	7.4		

Set up Histogram Bins. To make a histogram of salaries, 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 sample mean that are in widths of standard deviations.

Excel uses bin values to set the upper limit for each category. Start with a bin with upper limit equal to the mean, which will include compensation values that are at less than or equal to the mean. Specify the cell containing the mean rather than typing in the mean value:

= B449

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.

C449 : =B449				
	A	B	C	D
1		Total Compensation (MM\$)		
448		53.1	total compensation (MM\$) bin	
449	M	2.2	2.2	≤ M

In each of the three cells below this first bin, add one SD to the cell above, creating bins with upper limits of

$M + 1SD$,
 $M + 2SD$ and
 $M + 3SD$.

(Lock the cell reference to the SD in B450 by pressing fn4.)

T.IN...		✕ ✓ fx	=C451+\$B\$450	
	A	B	C	D
		Total Compensation (MM\$)		
1				
			total compensation	
448		53.1	(MM\$) bin	
449	M	2.2	2.2	$\leq M$
450	SD	3.8	6.1	$M < \leq M + SD$
451	75%	2.3	9.9	$M + SD < \leq M + 2SD$
452	median	1.1	=C451+\$B\$450	$M + 2SD < \leq M + 3SD$

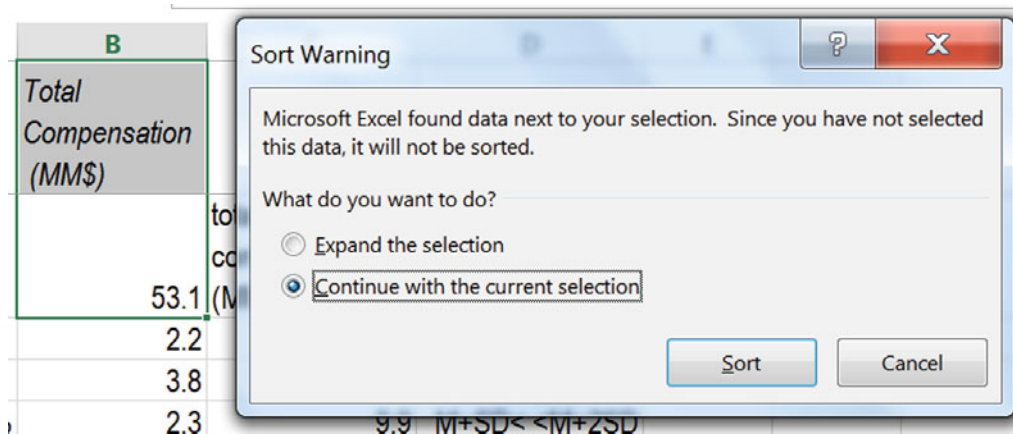
Excel 2.2 Sort to Produce Descriptives Without Outliers

Outliers are executives whose total compensation is more than three standard deviations greater than the mean.

To easily identify and remove outliers, sort the rows from lowest to highest *total compensation* (\$M):

Select *total compensation* data in column **B** (but not statistics below the data), then use shortcuts to sort:





Alt ASA, **C**ontinue with the current selection, **S**ort. (**A** selects the data, **A** menu, **S** selects the **S**ort menu, and **A** specifies **A**scending.)



Scroll up from the end of **B** to identify the rows with outlier values more than 13.7:

	A	B
		<i>Total Compensation (MM\$)</i>
1		
440		13.1
441		14.7
442		14.9
443		15.7
444		15.9
445		16.2
446		20.7
447		32.6
448		53.1

Recalculate the mean and standard deviation, including only rows with *total compensation* less than 14 million, by changing the end of the array in both Excel functions, which will update your bin upper limits:

B450		:	  	=STDEV.S(B2:B440)
	A	B	C	D
		<i>Total Compensation (MM\$)</i>		
1				
448		53.1	total compensation (MM\$) bin	
449	M	1.8	1.8	$\leq M$
450	SD 	2.0	3.8	$M < \leq M+SD$
451	75%	2.3	5.9	$M+SD < \leq M+2SD$
452	median	1.1	7.9	$M+2SD < \leq M+3SD$

Repeat this process to continue excluding outliers until there are no outliers. Since the distribution of total compensation is highly skewed, *outliers* will continue to appear.

Update the 75 % percentile, median, 25th percentile, and skewness.

B454		=SKEW(B2:B404)		
	A	B	C	D
1		Total Compensation (MM\$)		
447		32.6		
448		53.1	total compensation (MM\$) bin	
449	M	1.3	1.3	$\leq M$
450	SD	0.9	2.3	$M < \leq M+SD$
451	75%	1.8	3.2	$M+SD < \leq M+2SD$
452	median	1.0	4.1	$M+2SD < \leq M+3SD$
453	25%	0.68		
454	ske	1.1		

Including only executives whose total compensation is less than \$4.1 million, the descriptive statistics are more representative.

Excel 2.3 Make a Histogram

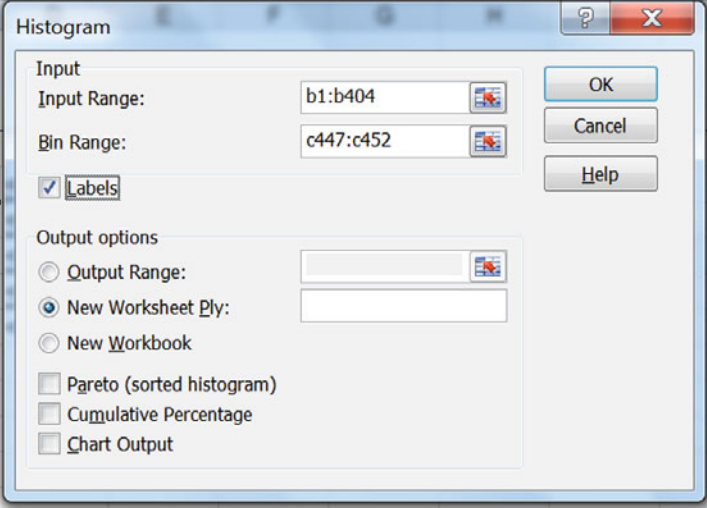
Excluding outliers, there are now executives whose compensation is more than one standard deviation beneath the mean. Add a histogram bin $M-SD$:

T.IN...		=B449-B450		
	A	B	C	D
1		Total Compensation (MM\$)		
447		32.6		
448		53.1	total compensation (MM\$) bin	
449	M	1.3	=B449-B450	<=M-SD
450	SD	0.9	1.3	M-SD<=M

To see the distribution of *Total Compensation*, use shortcuts to request a histogram:

Alt AY3 H Enter (Alt AY3 selects the data menu and the data Analysis menu.)

	A	B	C
1		Total Compensation (MM\$)	
447		32.6	bins
448		53.1	0.45
449	M	1.35	1.35
450	SD	0.90	2.25
451	75%	1.85	3.16
452	median	1.04	4.06
453	25%	0.68	
454	skewness	1.11	
455			
456			
457			

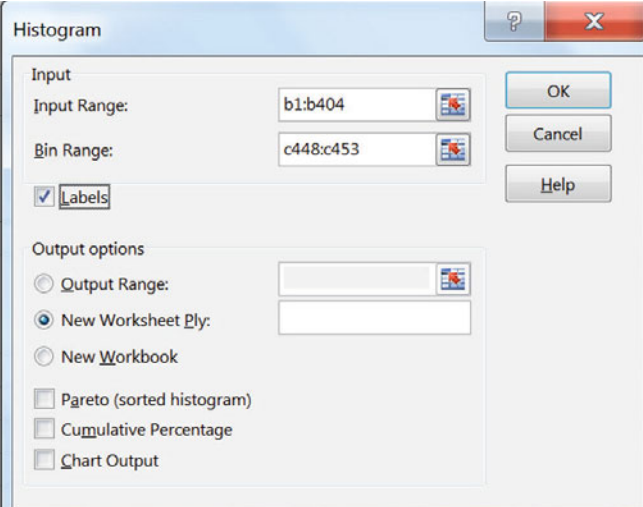


The Histogram dialog box is shown with the following settings:

- Input**
 - Input Range: b1:b404
 - Bin Range: c447:c452
 - ☒ Labels
- Output options**
 - ☐ Output Range:
 - ☒ New Worksheet Ply:
 - ☐ New Workbook
 - ☐ Pareto (sorted histogram)
 - ☐ Cumulative Percentage
 - ☐ Chart Output

For **Input Range**, enter the *total compensation* cells, including the label, excluding outliers; for **Bin Range**, enter the *total compensation bin* cells, including the label, and choose **L**abels, **O**K.

	B	C	D
	Total Compensation (MM\$)		
	32.6		
	total compensation		
	53.1 (MM\$) bin		
	1.3	0.4	$\leq M - SD$
	0.9	1.3	$M - SD < \leq M$
	1.8	2.3	$M < \leq M + SD$
	1.0	3.2	$M + SD < \leq M + 2SD$
	0.68	4.1	$M + 2SD < \leq M + 3SD$



The Histogram dialog box is shown with the following settings:

- Input**
 - Input Range: b1:b404
 - Bin Range: c448:c453
 - ☒ Labels
- Output options**
 - ☐ Output Range:
 - ☒ New Worksheet Ply:
 - ☐ New Workbook
 - ☐ Pareto (sorted histogram)
 - ☐ Cumulative Percentage
 - ☐ Chart Output

Select the bins cells with more than two significant digits and use shortcuts **Alt H 9** to reduce the unnecessary decimals. (**H** selects the **H**ome menu and **9** selects the reduce decimals function of the Number menu.)

FILE

HOME

INSERT

PAGE LAYOUT

FORMULAS

DATA

REVIEW

VIEW

JMP

Clipboard

Paste

Font

Arial Narrow

12

A

A

Alignment

Wrap Text

Number

Clipboard

Font

Alignment

Number

A2

✕

✓

f_x

0.447091619274218

	A	B	C	D	E	F	G	H	I
	total compensation (MMS) bin	Frequency							
1									
2	0.4	32							
3	1.3	222							
4	2.3	80							
5	3.2	42							
6	4.1	27							
7	More	0							

Excel 2.4 Use a PivotTable to Plot the Distribution

Select the histogram table, excluding the More row, and use shortcuts to produce a PivotTable:

Alt N V Enter (Shortcuts for iNsert piVot.)

Book1 - Microsoft Excel Preview

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW JMP

PivotTable Fields

Choose fields to add to report:

- ☐ total compensation (MMS) bin
- ☐ Frequency

Drag fields between areas below:

FILTERS COLUMNS

ROWS Σ VALUES

PivotTable1

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

Drag *total compensation bins* to ROWS and drag *frequency* to the Σ VALUES:

A		B
1		
2		
3	Row Labels	Sum of Frequency
4	0.4	32
5	1.3	222
6	2.3	80
7	3.2	42
8	4.1	27
9	Grand Total	403
10		
11		
12		
13		
14		

PivotTable Fields

Choose fields to add to report:

☒ total compensation (MM\$) bin
☒ Frequency

MORE TABLES

Drag fields between areas below:

FILTERS

COLUMNS

ROWS

total compensation (MM\$) bin

VALUES

Sum of Frequency

From a cell in the *Sum of Frequency* column, use shortcuts to change *Frequency* to percents:

Alt JT G Tab > Tab dn to %Grand Total Enter

B3

✕ ✓ fx

Sum of Freq

A	B
1	
2	
3	Row Labels Sum of Frequency
4	0.4 32
5	1.3 222
6	2.3 80
7	3.2 42
8	4.1 27
9	Grand Total 403
10	
11	
12	
13	
14	

Value Field Settings

Source Name: Frequency

Custom Name: Sum of Frequency

Summarize Values By Show Values As

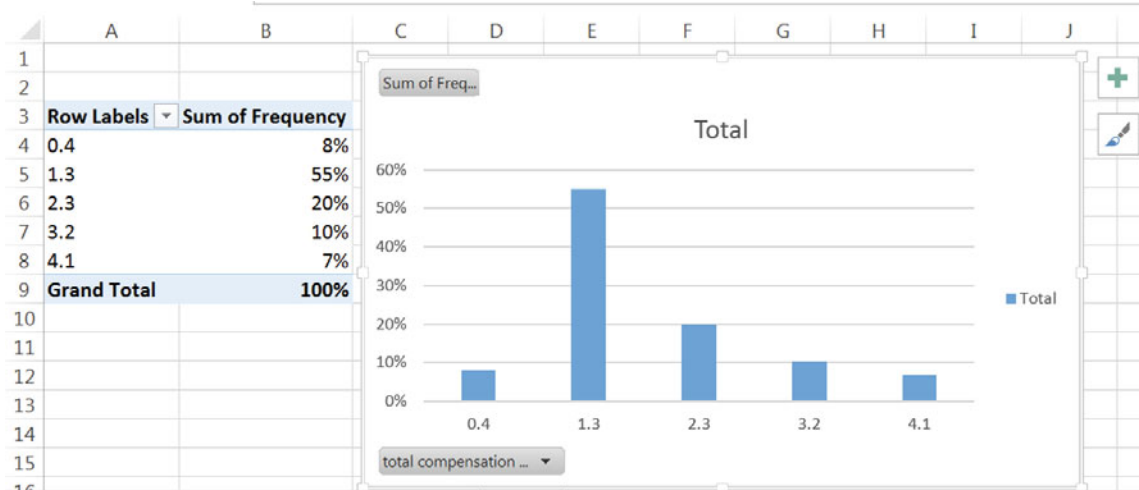
Show values as

- No Calculation
- No Calculation
- % of Grand Total**
- % of Column Total
- % of Row Total
- % Of
- % of Parent Row Total

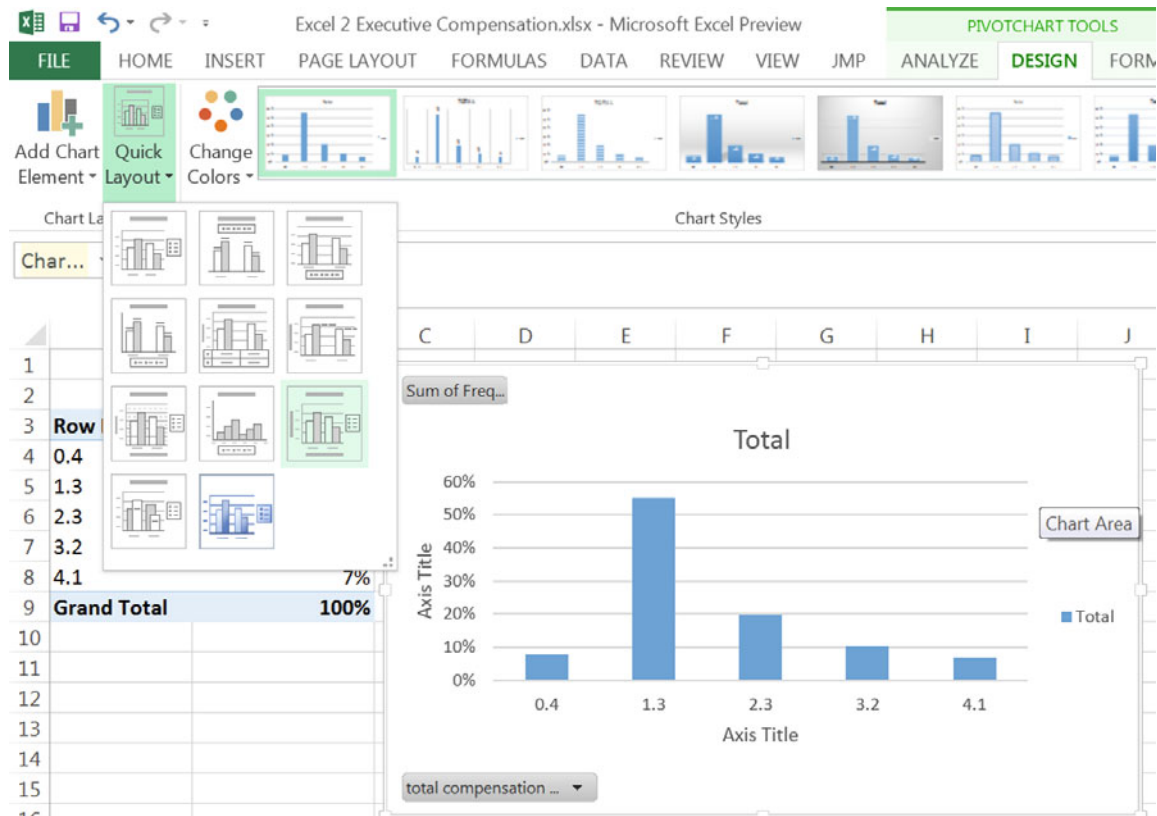
Number Format OK Cancel

Reduce decimals.

Use shortcuts to make a PivotChart of the distribution:

Alt JT C Enter

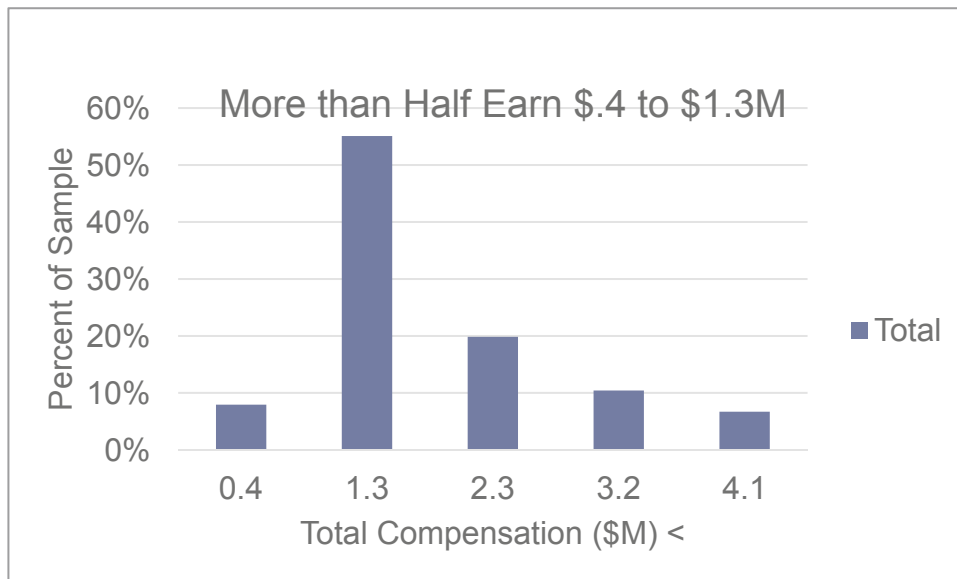
Use shortcuts to choose the ninth layout, which will add axes labels and a title:

Alt JC L Enter

Type in axes labels and title, delete the legend, and use shortcuts to set font size at 12:

Alt H F S 12 Enter

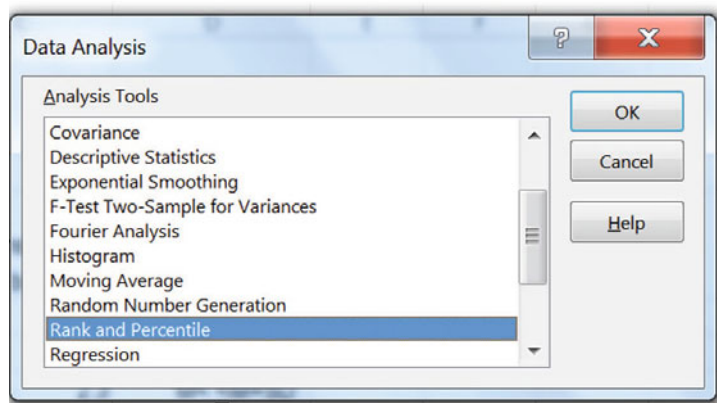
(where shortcuts stand for **H**ome **F**ont **S**ize **12**)

**Excel 2.5 Plot a Cumulative Distribution**

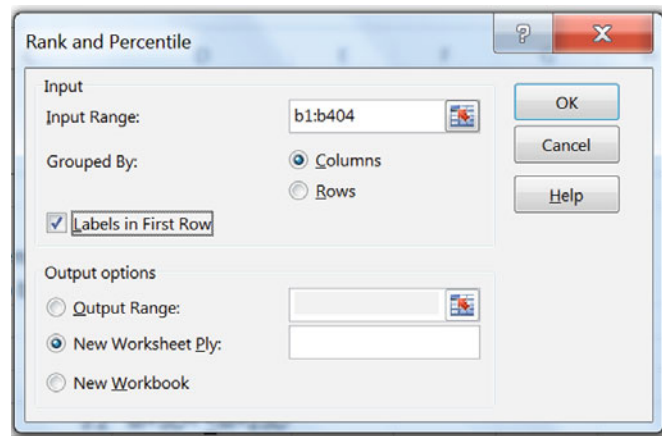
Return to the data sheet:

CNTL+Page Dn Page Dn

To see the cumulative distribution of *total compensation*, choose **Rank and Percentile**, **Alt AY3, R dn Enter**



Enter *total compensation* cells, including the label, excluding outliers, **L**abel, **E**nter:

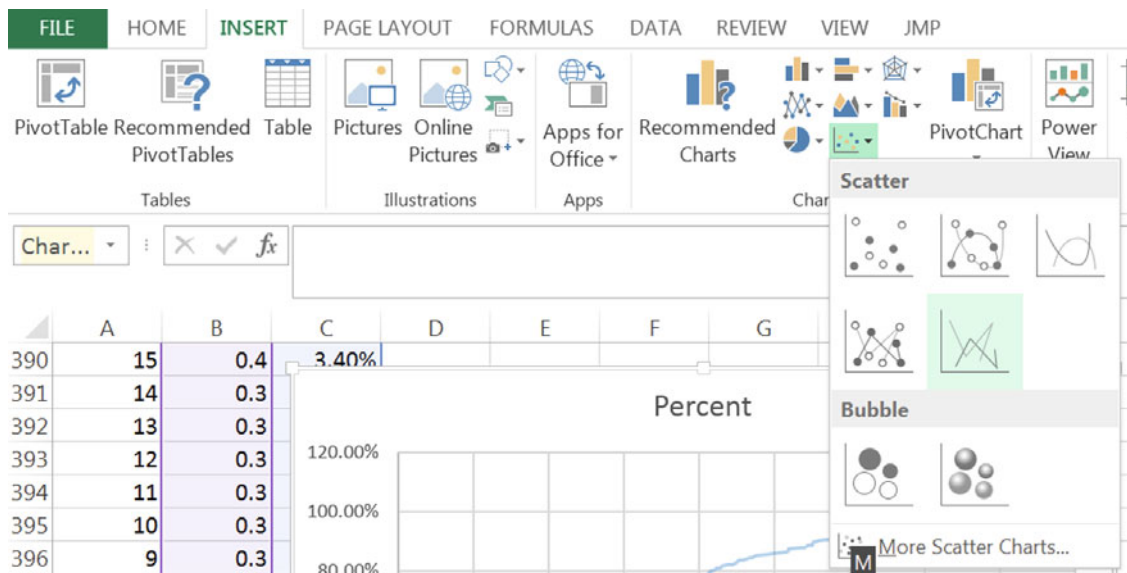


For convenience, select and delete column **C**, **Alt HDC**. (**H** selects the **H**ome menu, **D** selects the **D**elete menu, and **C** deletes the **C**olumn.)

	A	B	C
1	Point	mpensation	Percent
2	403	4.0	100.00%
3	402	4.0	99.70%
4	401	4.0	99.50%
5	400	4.0	99.20%

Reduce decimals in the *Percent* column **C**.

Select *Total Compensation* in **B** and *Percent* in **C** and then use shortcuts to create a scatterplot of the cumulative distribution, **Alt N D**



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

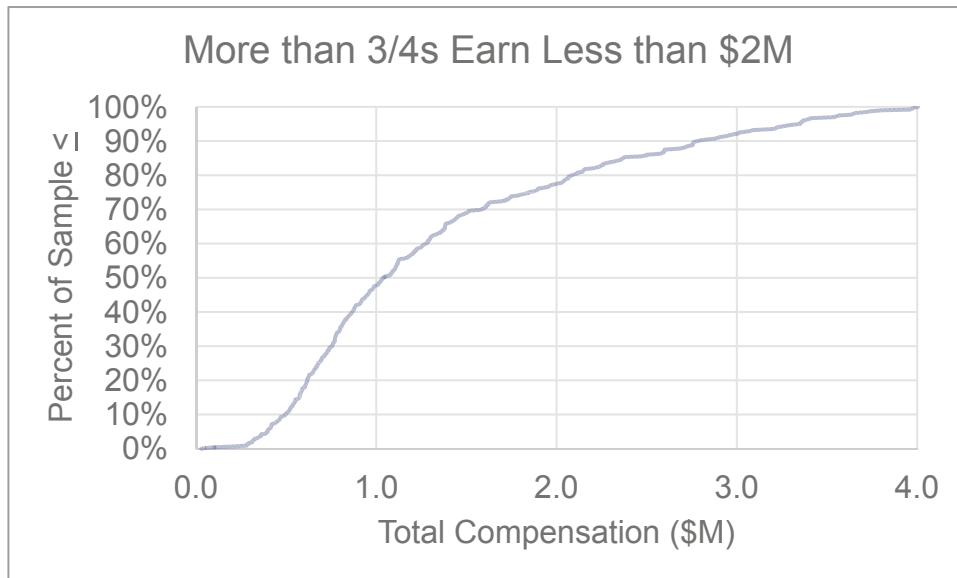
Use shortcuts to select and format axes:

Alt JA E

Alt JA M

Use shortcuts to add vertical gridlines:

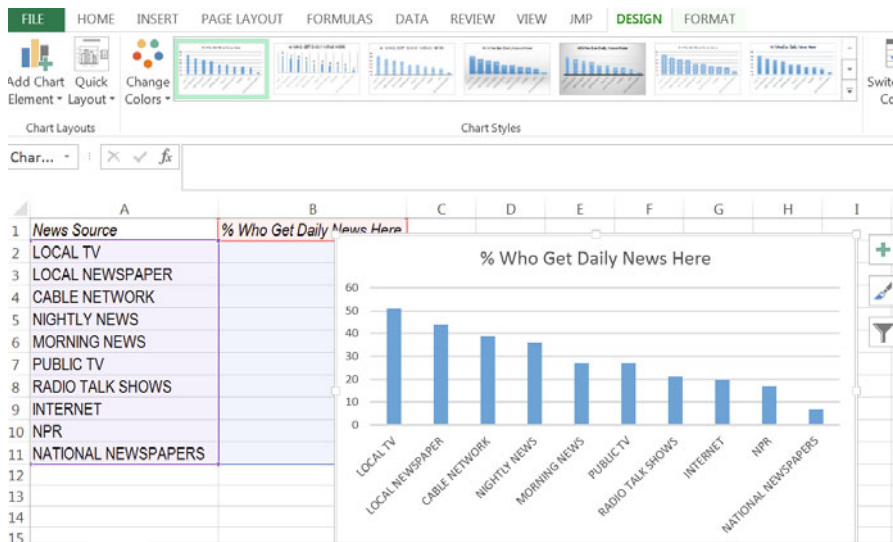
Alt JC A V G V



Excel 2.6 Produce a Column Chart of a Nominal Variable

A firm is targeting customers who consult a news source daily. Management wants to compare the popularity of news sources. To facilitate comparisons, we will make a PivotChart from a Gallup Poll of 992 Americans. Data are in **Excel 2.2 News Sources.xls**.

Open **Excel 2.2 News Sources.xls**, select the *News Source* and *% Who Get Daily News Here* data, and insert a column chart **Alt NC**. (where shortcuts activate insert a Column chart.)



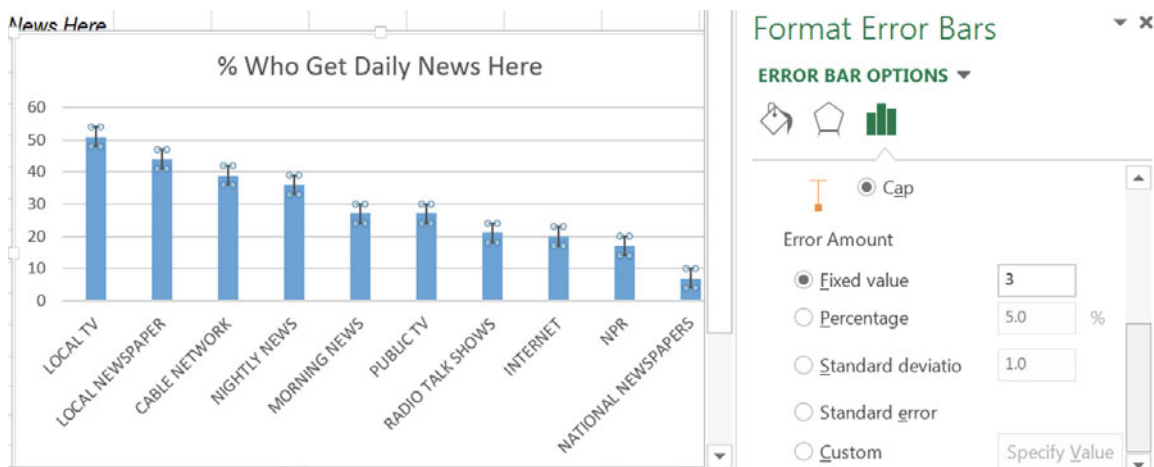
To add vertical margin of error bars, using click inside a column, then use shortcuts

Alt JC A E M

Then enter

Fixed: 3

(where 3 is the approximate margin of error).

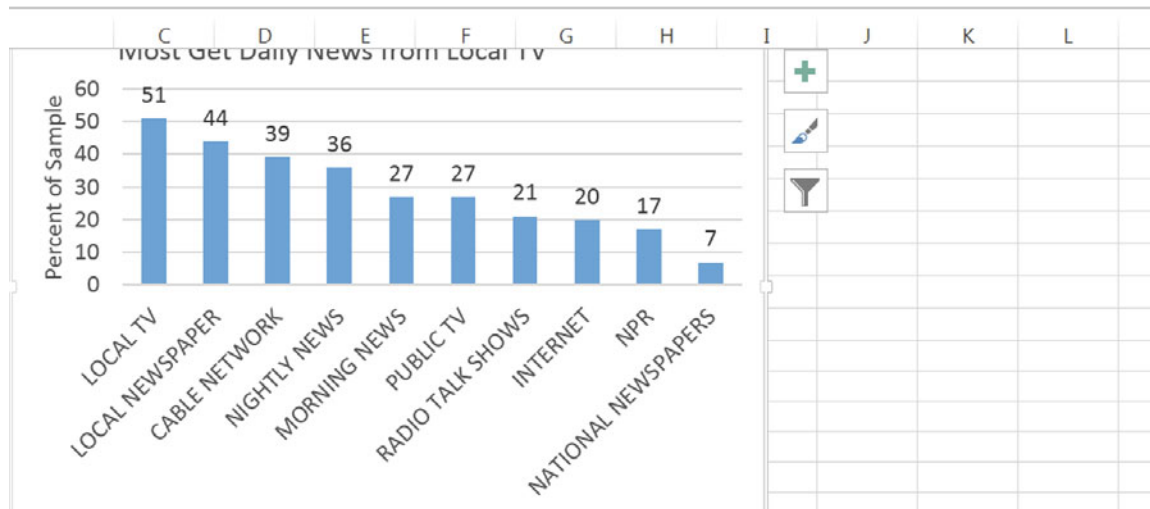


Choose **Design Chart Layout 9**,

Type in a stand alone title and axes titles and add data labels:

Alt JC A D O

Adjust chart height so that all of the *News Source* labels show:

Alt JA H up arrow

Excel Shortcuts at Your Fingertips

To navigate

to the end of the data

Cntl+Dn

To select

all of the data

A1

Cntl+Shift+right >

Cntl+Shift+Dn

a column

Cntl+space bar

a row

Shift+space bar

To do something with the data

Sort

Alt A S S

Tab to Sort by

Dn to Salary (M\$)

Find

Sample mean from the end of an array

Alt M U A

Standard deviation

=STDEV.S(array)

25 %

=PERCENTILE.INC(array,.25)

75 %

=PERCENTILE.INC(array,.75)

Median

=MEDIAN(array)

Skewness

=SKEW(array)

Make a histogram

Alt AY3

H Enter

array Tab

array Tab

L Enter

Reduce decimals

Select data in column A

Alt H 9

Make a PivotTable

Alt N V

Change from counts to percents

Alt JT G Tab > Tab to % Grand total

Make a PivotChart

Alt JT C Enter

Find the cumulative distribution

Alt AY3 R down Enter

array Tab

L Enter

Delete a column

Alt H D C

Plot the cumulative distribution

Select values and cumulative percents

Alt N D

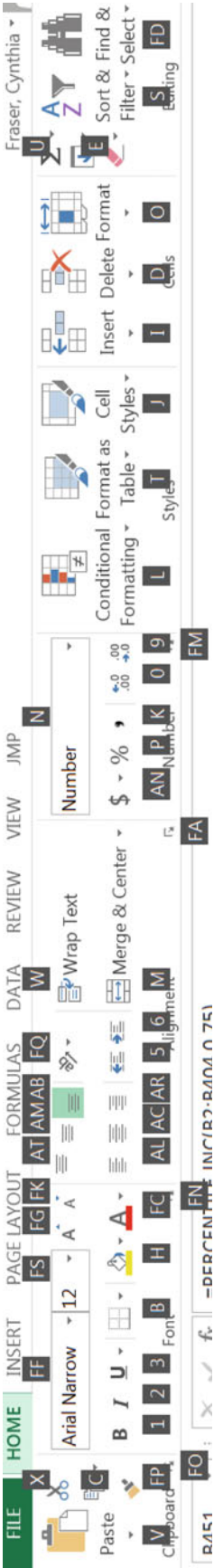
Make a column chart

Select categories and values or
percents

Alt N C

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

Alt H Home:



Home menu leys, from left to right, include:

V	paste	FF	Choose a font	FS	Choose a font size	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
----	--------------	----	--------------

From a PivotTable, **Alt** provides access to PivotTable tool menus, including **JT** for PivotTable ANALYZE (to change how cell values are shown).

From a PivotChart, **Alt** provides access to PivotChart tool menus, including:

JC	PivotChart Design	JA	PivotChart format
----	-------------------	----	-------------------

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 2,061 has four significant digits, while the number 2,610 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 39,500–40,499. In that case, the number of significant digits is one, and the zeros are placeholders. Alternatively, 40,000 could represent the range 39,995–40,004. 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 Descriptive Statistics

Compensation of 25 Best Paid CEOs

Forbes recently published the compensation packages of the 25 best compensated CEOs in the U.S. These data are in **Lab 2 Compensation of 25 Best Paid CEOs.xlsx**.

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. Identify outlier(s) who earn(s) more than 3 SDs above the M: _____
4. Find average compensation, M, excluding outlier(s): _____
5. Find the standard deviation of compensation, SD, excluding outlier(s): _____
6. Is the distribution of compensation among the best paid CEOs (excluding outlier(s)) approximately Normal? Y or N Evidence: _____
7. Make the histogram of compensation for top paid CEOs (excluding outlier(s)).
8. Plot the cumulative distribution of compensation (excluding outlier(s)).
9. What is median compensation among the 25 best paid CEOs? _____
10. What is the Interquartile Range of compensation among the 25 best paid CEOs?

II. Identify Industries where CEOs are Best Compensated

1. Use a PivotTable to determine the best paid industry: _____
2. What is the best paid industry, excluding outlier(s)? _____
3. In which industries do CEOs earn more than average (excluding outlier(s))?

Candidate Campaign Contributions

2012 Presidential Candidates' fundraising to date was published in the New York Times in October 2012. These data are in **Lab 2 Candidate Funds.xlsx**.

1. Plot the Candidates' donations by size.
2. What was the modal donation to President Obama? _____
3. What was the modal donation to Romney? _____
4. Which Candidate collected more donations under \$200? _____
5. Which Candidate collected more donations of \$2,500? _____

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.xls**.

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.

1. Identify *countries* which are **outliers**:
2. Illustrate advertising levels in countries that are not outliers. Add a "bottom line" chart title.
3. Summarize your analysis by describing *P&G's* advertising in *countries* around the world, excluding outliers.

Include:

- one or more measures of central tendency, such as the mean and median,
- one or more measures of dispersion, such as the standard deviation and range,
- the similarity of the distribution to a *Normal* distribution

Be sure to round your answers to two or three significant digits.

4. Considering the entire sample, which advertising strategy describes the P&G strategy better: (i) advertise at a moderate level in many global markets, (ii) advertise heavily to a small number of key markets and spend a little 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, make a PivotTable, make a PivotChart, plot the cumulative distribution) and record your time. If your time is more than 5 min, repeat twice, and then record your best time.

CASE 2-1 VW Backgrounds

Volkswagon management commissioned 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.xls**. 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 “bottom line” 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.

Business Statistics for Competitive Advantage with
Excel 2013

Basics, Model Building, Simulation and Cases

Fraser, C.

2013, XIV, 449 p. 406 illus. in color., Softcover

ISBN: 978-1-4614-7380-0