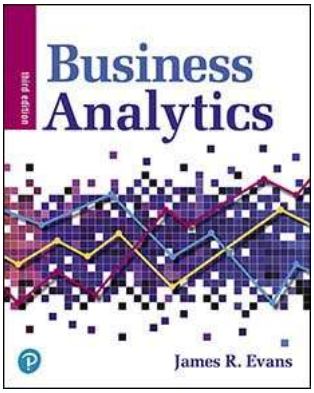
Business Analytics: Methods, Models, and Decisions

Third Edition



Chapter 4

Descriptive Statistics

Statistics

- **Statistics**, as defined by David Hand, past president of the Royal Statistical Society in the UK, is both the science of uncertainty and the technology of extracting information from data.
  - Statistics involves collecting, organizing, analyzing, interpreting, and presenting data.
  - A **statistic** is a summary measure of data.
- **Descriptive statistics** refers to methods of describing and summarizing data using tabular, visual, and quantitative techniques.

Metrics and Data Classification

- **Metric** - a unit of measurement that provides a way to objectively quantify performance.
- **Measurement** - the act of obtaining data associated with a metric.
- **Measures** - numerical values associated with a metric.

Types of Metrics

- **Discrete metric** - one that is derived from counting something.
  - For example, a delivery is either on time or not; an order is complete or incomplete; or an invoice can have one, two, three, or any number of errors. Some discrete metrics would be the proportion of on-time deliveries; the number of incomplete orders each day, and the number of errors per invoice.
- **Continuous metrics** are based on a continuous scale of measurement.
  - Any metrics involving dollars, length, time, volume, or weight, for example, are continuous.

Measurement Scales

- **Categorical (nominal) data** - sorted into categories according to specified characteristics.
- **Ordinal data** - can be ordered or ranked according to some relationship to one another.
- **Interval data** - ordinal but have constant differences between observations and have arbitrary zero points.
- **Ratio data** - continuous and have a natural zero.

 Slide - 5

Example 4.1: Classifying Data Elements

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
11	Durrable Products	Aug11008	7258	Pressure Gauge	$ 90.00	100	$ 9,000.00	45	08/25/11	08/28/11
12	Fast-Tie Aerospace	Aug11009	6321	O-Ring	$ 2.45	1,300	$ 3,185.00	30	08/25/11	09/04/11
13	Fast-Tie Aerospace	Aug11010	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/25/11	09/02/11
14	Steelpin Inc.	Aug11011	5319	Shielded Cable/ft.	$ 1.10	18,100	$ 19,910.00	30	08/25/11	09/05/11
15	Hulkey Fasteners	Aug11012	3166	Electrical Connector	$ 1.25	5,600	$ 7,000.00	30	08/25/11	08/29/11

Categorical, Ordinal, Categorical, Categorical, Ratio, Ratio, Ratio, Ratio, Interval, Interval

 Slide - 6

Frequency Distributions and Histograms

- A **frequency distribution** is a table that shows the number of observations in each of several nonoverlapping groups.
- A graphical depiction of a frequency distribution in the form of a column chart is called a **histogram**.

Frequency Distributions for Categorical Data

- Categorical variables naturally define the groups in a frequency distribution.
- To construct a frequency distribution, we need only count the number of observations that appear in each category.
  - This can be done using the Excel COUNTIF function.

Example 4.2: Constructing a Frequency Distribution for Items in the Purchase Orders Database

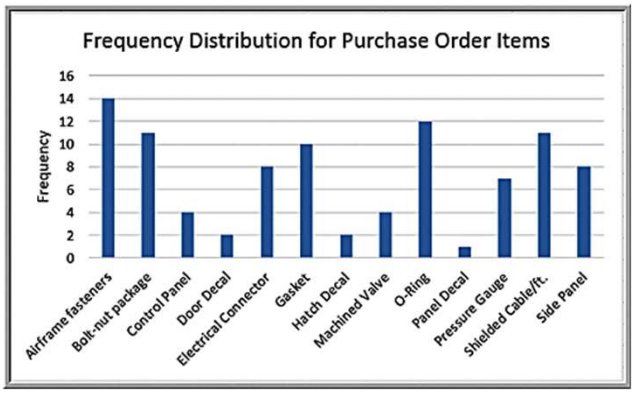
- List the item names in a column on the spreadsheet.
- Use the function =COUNTIF($D$4:$D$97,*cell_reference*), where *cell_reference* is the cell containing the item name.

	A	B
100	Item Description	Frequency
101	Airframe fasteners	=COUNTIF($D$4:$D$97,A101)
102	Bolt-nut package	=COUNTIF($D$4:$D$97,A102)
103	Control Panel	=COUNTIF($D$4:$D$97,A103)
104	Door Decal	=COUNTIF($D$4:$D$97,A104)
105	Electrical Connector	=COUNTIF($D$4:$D$97,A105)
106	Gasket	=COUNTIF($D$4:$D$97,A106)
107	Hatch Decal	=COUNTIF($D$4:$D$97,A107)
108	Machined Valve	=COUNTIF($D$4:$D$97,A108)
109	O-Ring	=COUNTIF($D$4:$D$97,A109)
110	Panel Decal	=COUNTIF($D$4:$D$97,A110)
111	Pressure Gauge	=COUNTIF($D$4:$D$97,A111)
112	Shielded Cable/ft.	=COUNTIF($D$4:$D$97,A112)
113	Side Panel	=COUNTIF($D$4:$D$97,A113)

	A	B
100	Item Description	Frequency
101	Airframe fasteners	14
102	Bolt-nut package	11
103	Control Panel	4
104	Door Decal	2
105	Electrical Connector	8
106	Gasket	10
107	Hatch Decal	2
108	Machined Valve	4
109	O-Ring	12
110	Panel Decal	1
111	Pressure Gauge	7
112	Shielded Cable/ft.	11
113	Side Panel	8

Example 4.2 Continued

- Construct a column chart to visualize the frequencies.



Relative Frequency Distributions

- **Relative frequency** is the fraction, or proportion, of the total.
- If a data set has *n* observations, the relative frequency of category *i* is:

$$\text{Relative Frequency of Category } i = \frac{\text{Frequency of Category } i}{n} \qquad (4.1)$$

- We often multiply the relative frequencies by 100 to express them as percentages.
- A **relative frequency distribution** is a tabular summary of the relative frequencies of all categories.

Example 4.3: Constructing a Relative Frequency Distribution for Items in the Purchase Orders Database

- First, sum the frequencies to find the total number (note that the sum of the frequencies must be the same as the total number of observations, *n*).
- Then divide the frequency of each category by this value.

	A	B	C
100	Item Description	Frequency	Relative Frequency
101	Airframe fasteners	14	0.1489
102	Bolt-nut package	11	0.1170
103	Control Panel	4	0.0426
104	Door Decal	2	0.0213
105	Electrical Connector	8	0.0851
106	Gasket	10	0.1064
107	Hatch Decal	2	0.0213
108	Machined Valve	4	0.0426
109	O-Ring	12	0.1277
110	Panel Decal	1	0.0106
111	Pressure Gauge	7	0.0745
112	Shielded Cable/ft.	11	0.1170
113	Side Panel	8	0.0851
114	Total	94	1.0000

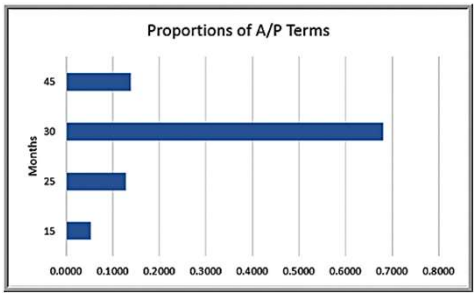
Frequency Distributions for Numerical Data

- For numerical data that consist of a small number of discrete values, we may construct a frequency distribution similar to the way we did for categorical data; that is, we simply use COUNTIF to count the frequencies of each discrete value.

Example 4.4: Frequency and Relative Frequency Distribution for A/P Terms

- In the Purchase Orders data, the A/P terms are all whole numbers 15, 25, 30, and 45.

	A	B	C
117	A/P Terms	Frequency	Relative Frequency
118	15	5	0.0532
119	25	12	0.1277
120	30	64	0.6809
121	45	13	0.1383
122	Total	94	1.0000

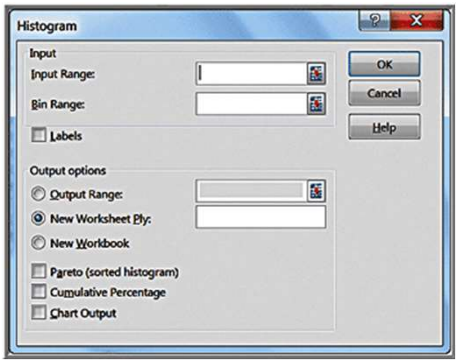


Excel Histogram Tool

- Frequency distributions and histograms can be created using the Analysis Toolpak in Excel.
  - Click the *Data Analysis* tools button in the *Analysis* group under the *Data* tab in the Excel menu bar and select *Histogram* from the list.

Histogram Dialog

- Specify the *Input Range* corresponding to the data. If you include the column header, then also check the *Labels* box so Excel knows that the range contains a label. The *Bin Range* defines the groups (Excel calls these "bins") used for the frequency distribution.



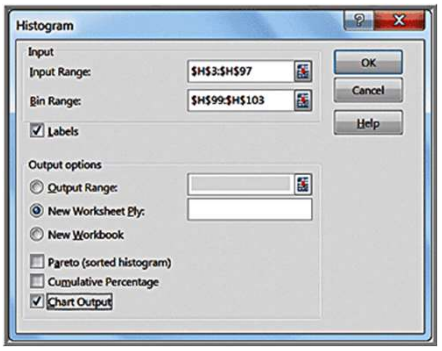
Using Bin Ranges

- If you do not specify a bin range, Excel will automatically determine bin values for the frequency distribution and histogram, which often results in a rather poor choice.
- If you have discrete values, set up a column of these values in your spreadsheet for the bin range and specify this range in the *Bin Range* field.

Example 4.5: Using the Histogram Tool

- We will create a frequency distribution and histogram for the A/P Terms variable in the Purchase Orders database.
- We defined the bin range below the data in cells H99:H103 as follows:

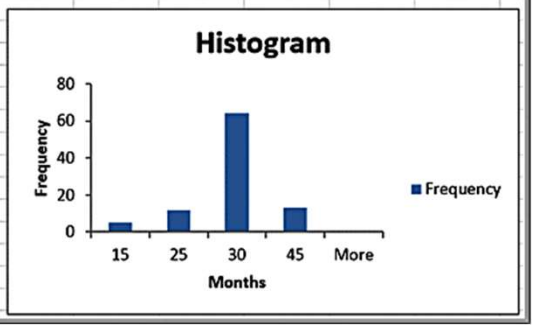
Month
15
25
30
45



Example 4.5 Continued

- Histogram tool results:

	A	B
1	*Months*	*Frequency*
2	15	5
3	25	12
4	30	64
5	45	13
6	More	0



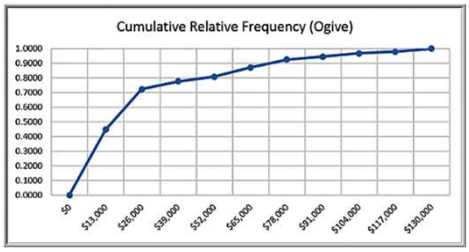
Cumulative Relative Frequency Distributions

- The **cumulative relative frequency** represents the proportion of the total number of observations that fall at or below the upper limit of each group.
- A tabular summary of cumulative relative frequencies is called a **cumulative relative frequency distribution**.

Example 4.7: Computing Cumulative Relative Frequencies

- Set the cumulative relative frequency of the first group equal to its relative frequency. Then add the relative frequency of the next group to the cumulative relative frequency.
  - For example, the cumulative relative frequency in cell D3 is computed as =D2+C3 = 0.000 + 0.4468 = 0.4468.

	A	B	C	D
1	*Upper Group Limit*	*Frequency*	*Relative Frequency*	*Cumulative Relative Frequency*
2	0	0	0.0000	0.0000
3	13000	42	0.4468	0.4468
4	26000	26	0.2766	0.7234
5	39000	5	0.0532	0.7766
6	52000	3	0.0319	0.8085
7	65000	6	0.0638	0.8723
8	78000	5	0.0532	0.9255
9	91000	2	0.0213	0.9468
10	104000	2	0.0213	0.9681
11	117000	1	0.0106	0.9787
12	130000	2	0.0213	1.0000
13	More	0	0.0000	1.0000
14	Total	94		

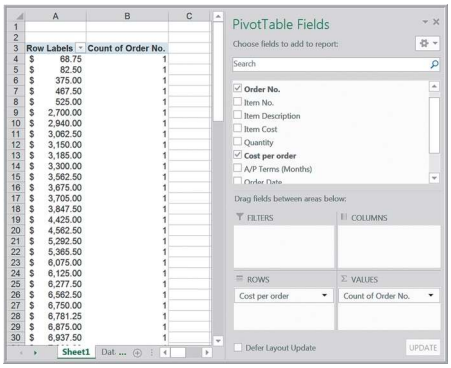


Constructing Frequency Distributions Using PivotTables

- In the *Purchase Orders* data, we can simply build a PivotTable to find a count of the number of orders for each item.
- For continuous numerical data, we can also use PivotTables to construct a grouped frequency distribution.

Example 4.8: Constructing a Grouped Frequency Distribution Using PivotTables

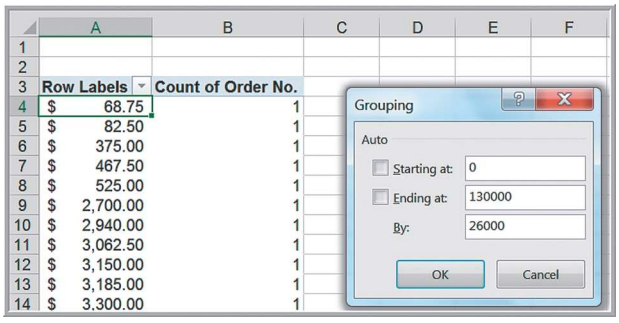
1. Using the Purchase Orders database, create a PivotTable as shown:



 Slide - 23

Example 4.8 Continued

2. Click on any value in the *Row Labels* column, and from the *Analyze* tab for *PivotTable Tools*, select *Group Field*. Edit the dialog to start at 0 and end at 130000, and use 26000 as the group range.



 Slide - 24

Example 4.8 Continued

- Grouped frequency distribution results:

	A	B
1		
2		
3	**Row Labels**	**Count of Order No.**
4	0-26000	68
5	26000-52000	8
6	52000-78000	11
7	78000-104000	4
8	104000-130000	3
9	**Grand Total**	**94**

Percentiles

- The ***k*th percentile** is a value at or below which at least *k* percent of the observations lie. The most common way to compute the *k*th percentile is to order the data values from smallest to largest and calculate the rank of the *k*th percentile using the formula:

$$\frac{nk}{100} + 0.5 \qquad (4.3)$$

- Statistical software use different methods that often involve interpolating between ranks instead of rounding, thus producing different results.
  - The Excel function PERCENTILE.INC(*array,k*) computes the *k*th percentile of data in the range specified in the array field, where *k* is in the range 0 to 1, inclusive (i.e., including 0 and 1).

Examples 4.9 and 4.10: Computing Percentiles

- Compute the 90th percentile for Cost per order in the Purchase Orders data.
- Rank of k^{th} percentile $= \dfrac{nk}{100 + 0.5}$
- $n = 94$; $k = 90$
- For the 90th percentile, the rank is

$$= \frac{94(90)}{100} + 0.5 = 85.1 \ (\text{round to } 85)$$

  - Value of the 85th observation = \$74,375
- Using the Excel function PERCENTILE.INC(G4:G97,0.9), the 90th percentile is \$73,737.50, which is different from using formula (3.3).

Example 4.11: Excel Rank and Percentile Tool

Data > Data Analysis > Rank and Percentile

90.3rd percentile

= \$74,375

(same result as manually computing the 90th percentile)

	A	B	C	D
1	*Point*	*Cost per order*	*Rank*	*Percent*
2	74	\$127,500.00	1	100.00%
3	62	\$121,000.00	2	98.90%
4	71	\$110,000.00	3	97.80%
5	16	\$103,530.00	4	96.70%
6	73	\$ 96,750.00	5	95.60%
7	1	\$ 82,875.00	6	94.60%
8	67	\$ 81,937.50	7	93.50%
9	82	\$ 77,400.00	8	92.40%
10	54	\$ 76,500.00	9	91.30%
11	80	\$ 74,375.00	10	90.30%
12	68	\$ 72,250.00	11	89.20%
13	20	\$ 65,875.00	12	88.10%
14	65	\$ 64,500.00	13	87.00%
15	28	\$ 63,750.00	14	86.00%

The Excel value of the 90th percentile that was computed in Example 4.9 as \$74,375 is the 90.3rd percentile value.

Quartiles

- **Quartiles** break the data into four parts.
  - The 25th percentile is called the first quartile,Q1;
  - the 50th percentile is called the second quartile, Q2;
  - the 75th percentile is called the third quartile, Q3; and
  - the 100th percentile is the fourth quartile, Q4.
- One-fourth of the data fall below the first quartile, one-half are below the second quartile, and three-fourths are below the third quartile.
- Excel function QUARTILE.INC(*array, quart*), where array specifies the range of the data and quart is a whole number between 1 and 4, designating the desired quartile

Example 4.12: Computing Quartiles in Excel

- Compute the Quartiles of the Cost per Order data
  - First quartile: =QUARTILE.INC(G4:G97,1) = $6,757.81
  - Second quartile: =QUARTILE.INC(G4:G97,2) = $15,656.25
  - Third quartile: =QUARTILE.INC(G4:G97,3) = $27,593.75
  - Fourth quartile: =QUARTILE.INC(G4:G97,4) = $127,500.00

