MgtOp 215 **Chapter 2** **Dr. Ahn**

In this chapter we will study how to organize and summarize (raw) data into a meaningful way.

Categorical Data

A **Summary Table** presents tallied responses as frequencies or percentages for each category.

Example 1: Refer to Table 2.3 of Example 2.1 on p.38 of the textbook which used the data set "Retirement Funds."

A **Contingency Table** summarizes data with two or more categories by cross-tabulating or tallying jointly.

Example 2: Refer to Tables 2.3 & 2.4 on p. 39 of the textbook which used the data set "Retirement Funds."

A **Bar chart** and a **Pareto diagram** are used for graphical representation for categorical data

A **Pie chart** is a circle divided into portions that represent the relative frequencies of categorical data.

Example 3: Refer to Figure 2.1 on p. 51 of the textbook.

These can be easily constructed using PHStat: Refer to Excel Guide on p. 85.

MSL Homework: 2.1, 2.5, 2.9 (on pp. 40-41), 2.27 (on p. 56)
HC Homework: 2.25 (on. p.56)

Example 4. The following data are 227 observations of the 1-year return % of the Growth Funds in the data file "Retirement Funds." (Refer to p. 34 of the textbook for further detail and Appendix C on p. 729 for how to access the data.).

28.99	33.40	33.98	33.78	21.62	22.87	17.41	16.54
15.72	14.50	17.14	15.23	18.62	22.45	15.66	13.75
14.62	26.06	15.72	11.86	18.64	15.71	17.77	15.36
17.33	21.08	18.12	9.16	22.02	16.63	13.08	12.49
17.72	13.42	14.18	16.99	13.12	14.74	17.93	12.27
11.29	10.75	12.57	17.67	13.60	12.18	18.31	17.07
14.88	15.13	15.02	11.17	20.67	16.79	19.91	13.23
.							
.							
.							
10.90	9.82	12.26	13.78	15.46	11.97	14.93	14.45
10.88	12.18	12.20	9.45	7.74	13.80	15.31	21.46
13.69	13.93	12.91	2.43	12.86	12.94	15.46	11.39
5.70	9.97	11.78	8.13	4.23	5.97	12.60	7.85
6.77	11.47	3.80	22.44	-11.28	2.03	9.11	24.30
16.46	19.19	11.59	17.84	10.35	17.08	14.45	12.67
15.92	18.34	12.06	10.20	11.99	12.45	11.00	12.48
17.98	12.93	13.58	10.42	13.95	8.88	14.50	14.59
13.04	16.52	14.61	12.43	13.39	18.35	11.85	14.31
14.15	8.24	14.30	22.10	4.14	11.52	10.38	8.33
15.25	8.07	12.08	10.75	18.15	4.48	10.61	8.66
7.63	5.30	10.97					

As we cannot make sense out of the (raw) data as they are, we will "transform" the data into meaningful forms.

Ordered data: data arranged in ascending order or in descending order; easy to pick out extremes, typical values, and concentration of values.
To order data in MS Excel, **Data > Sort…**

-11.28	2.03	2.43	3.80	4.14	4.23	4.48	5.30
5.70	5.97	6.77	7.63	7.74	7.84	7.85	8.07
8.13	8.24	8.33	8.66	8.85	8.88	9.11	9.16
9.23	9.45	9.77	9.82	9.97	9.99	10.20	10.32
10.34	10.35	10.38	10.42	10.46	10.61	10.75	10.75
10.88	10.90	10.97	11.00	11.17	11.21	11.28	11.29
11.39	11.40	11.47	11.52	11.56	11.59	11.72	11.75
11.78	11.80	11.80	11.85	11.86	11.97	11.99	12.06
12.08	12.16	12.18	12.18	12.20	12.25	12.26	12.27
12.40	12.43	12.45	12.48	12.49	12.57	12.60	12.67
12.70	12.72	12.86	12.91	12.93	12.94	12.97	12.99
12.99	13.02	13.04	13.08	13.12	13.23	13.39	13.39
13.42	13.42	13.44	13.54	13.58	13.60	13.69	13.70
13.75	13.78	13.80	13.93	13.95	13.96	14.05	14.09
14.15	14.18	14.26	14.29	14.30	14.31	14.40	14.45
14.45	14.50	14.50	14.59	14.60	14.61	14.62	14.65
14.71	14.74	14.75	14.82	14.88	14.93	15.02	15.08
15.13	15.23	15.25	15.27	15.31	15.36	15.46	15.46
15.61	15.61	15.66	15.67	15.69	15.69	15.71	15.72
15.72	15.77	15.77	15.83	15.92	15.96	16.04	16.04
16.13	16.16	16.27	16.40	16.43	16.46	16.52	16.54
16.57	16.63	16.64	16.65	16.79	16.94	16.95	16.95
16.95	16.99	17.07	17.08	17.14	17.23	17.33	17.41
17.61	17.67	17.72	17.74	17.77	17.80	17.80	17.84
17.93	17.98	18.12	18.15	18.28	18.31	18.34	18.35
18.36	18.57	18.62	18.64	19.19	19.40	19.73	19.91
19.95	20.67	20.85	21.08	21.46	21.62	22.02	22.10
22.44	22.45	22.87	23.24	24.30	26.06	26.38	28.99
33.40	33.78	33.98					

A **frequency distribution** (table) is a summary table in which data are arranged into conveniently established, numerically ordered class groupings or categories. The number of observations falling in each **class** is called the **class frequency**. It is a way of organizing raw data (that are considered continuous) into a meaningful way.

Example 5. Refer to the data in Example 4. Construct a frequency table.

Construction of the frequency table:

1 Determine the range of the raw data
Range = largest observation - smallest observation, for example, range=$33.98 - (-11.28) = 45.26$

2. Determine, approximately, the number of classes *K* such that *K* is the smallest whole number that makes 2^K greater than the number of observations, 227 in this case. Here $2^7 = 128$, and $2^8 = 256$ thus *K*=8.

3. Divide the range by this *K* to determine, approximately the class width. 45.26/8=5.65≈5.

(Note that for convenience we use 5 as a class width. But you may try 6 or even 10 as class width.)

4. Determine class limits by beginning with a whole number. Make sure to include the smallest and largest observation.

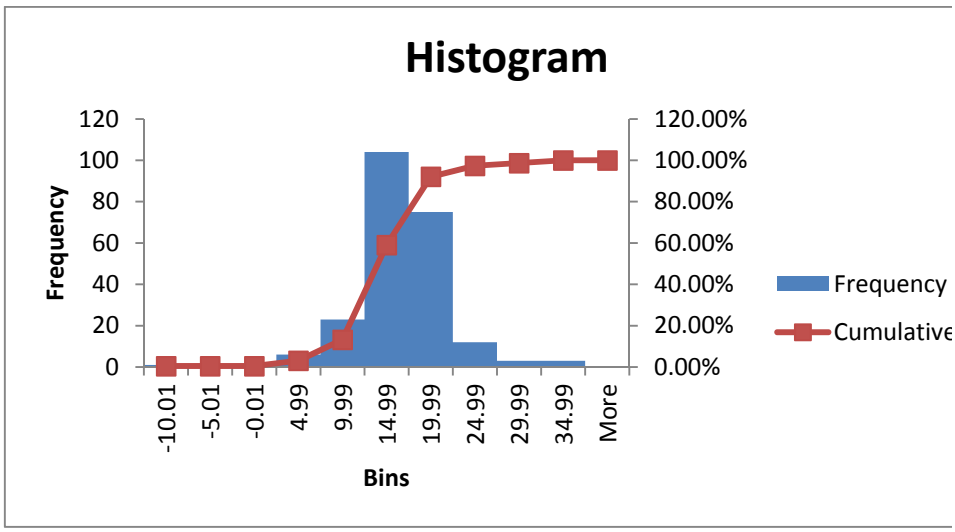
Class			Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
-15	to under	-10	1	0.0044	1	0.0044
-10	to under	-5	0	0	1	0.0044
-5	to	0	0	0	1	0.0044
0		5	6	0.0264	7	0.0308
5		10	23	0.1013	30	0.1322
10		15	104	0.4582	134	0.5903
15		20	75	0.3304	209	0.9207
20		25	12	0.0529	221	0.9736
25		30	3	0.0132	224	0.9868
30		35	3	0.0132	227	1

5. If it is necessary, divide the data into more (fewer) groups, decrease (increase) the class width.

Histogram: a graph made of rectangles whose areas are proportional to the relative frequencies of respective classes.

To make histograms and ogives in MS Excel, **Data > Data Analysis… > Histogram.** Then, enter the data range and the bin range. The bin range consists of the upper limits of the classes which in this example are -10.0001, -5.0001, -0.0001, 4.9999, 9.999, …, 29.9999, 34.9999. (What happens if you use -5, 0, 5, 10, …, instead?)

Bins	*Frequency*	*Cumulative %*
-10.01	1	0.44%
-5.01	0	0.44%
-0.01	0	0.44%
4.99	6	3.08%
9.99	23	13.22%
14.99	104	59.03%
19.99	75	92.07%
24.99	12	97.36%
29.99	3	98.68%
34.99	3	100.00%
More	0	100.00%



When you make histograms, you assume that the data are continuous data. Thus, you do **not** want to have a gap between two rectangles. To remove gaps, click on the histogram. Then you will see the panel "Format Data Series." Next click "Options" and set the Gap Width to zero.

You can get these using PHStat, which yields the correct cumulative ogive.

For details of how to make frequency distributions, histograms, and cumulative distribution ogives using MS Excel, refer to Chapter 2 Excel Guide (EG) on p. 85.

HC Homework: In Example 5, construct a relative frequency distribution and sketch a corresponding histogram and the relative cumulative frequency using a class width of 10. Then compare this histogram with the histogram constructed in Example 4 and discuss which summarize the data best.

MSL Homework: 2.11, 2.14, 2.16, 2.20

Open ended class: a class without either a lower limit or an upper limit.

Polygon: another way of displaying a (relative) frequency distribution

Ogive: a graph of a (relative) cumulative frequency distribution

Example 6: The following is a frequency distribution of the number of cups of coffee sold by an Espresso stand on the corner of Grand and Main during a period of 50 days. Sketch a relative frequency histogram and a cumulative relative frequency ogive.

Number of Cups			Frequency
0	to under	100	5
100	to under	200	10
200	to under	300	30
300	to under	500	5

The height (y-axis) of a relative frequency histogram is called the **density**. The relative frequency of a class is the density (height) times the width of the class.

HC Homework:

Sketch a relative frequency histogram and a cumulative relative frequency ogive for the following frequency distribution.

Number of Cups			Frequency
0	to under	50	15
50	to under	100	30
100	to under	200	60
200	to under	300	15

Dot plot:

A **stem and leaf plot** is partly tabular and partly graphical way of summarizing data, and suitable for smaller data sets. In this plot the first or leading digits serve as the stem and the trailing digits as the leaf

Example 7. The following data represent the bounced check fee in dollars for a sample of 23 banks for direct-deposit customers who maintain a $100 balance. Construct a dot plot and stem-and-leaf plot for the data.

```
26  28  20  20  21  22  25  25  18
25  15  20  18  20  25  25  22  30
30  30  15  20  29
```

We can obtain the following stem-and-leaf plot of the data in Example 4 using PHStat:

			Stem-and-Leaf Display	
			Stem unit:	**10**
Statistics			**1**	**5 5 8 8**
Sample Size	**23**		**2**	**0 0 0 0 0 1 2 2 5 5 5 5 5 6 8 9**
Mean	**23**		**3**	**0 0 0**
Median	**22**			
Std. Deviation	**4.602371**			
Minimum	**15**			
Maximum	**30**			

MSL Homework: 2.33, 2.37, 2.40, 2.44

Common shapes of distribution:
- symmetric
- skewed to the right (positively skewed)
- skewed to the left (negatively skewed)

Mode: the location at which a relative frequency distribution peaks

Modal class: the class that contains a mode.

Unimodal

Bimodal

HC Homework: Read Section 2.7 on misuses and common errors in visualizing data. Then find an example of graphics that is relevant to this section from newspapers, magazines, or the Internet, and make relevant comments on the example that you find. Good sources include Wall Street Journal, Business Week, and USA Today. This particular problem is due week from Friday of the week this Chapter 2 is finished. You must include the source of your example and a copy of graph(s). You comments need to be one or two pages long and typed in double space.