Section 1.5 – Organizing Qualitative Data

When data is collected from a survey or designed experiment, they must be organized into a manageable form. Data that is not organized is referred to as ***raw data.***

Ways to Organize Data

- Tables
- Graphs
- Numerical Summaries

Definition

A ***frequency distribution*** (or ***frequency table***) shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number of data values in each of the categories.

Example

Consider pulse rate measurements (in beats per minute) obtained from a simple random sample of 40 males and another simple random sample of 40 females, with the results listed in the table below.

Pulse Rates (*beats per minute*) of Females and Males

Females																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	93	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
Males																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

The frequency distribution summarizing the pulse rate of females listed in table below.

Pulse Rates of Females

Pulse Rate	*Frequency*
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

The frequency for a particular class is the number of original values that fall into that class. That is the frequency of 12, indicating that 12 of the original pulse rates are between 60 and 69 beats per minute.

Relative Frequency Distribution

The ***relative frequency*** is the proportion (or percent) of observations within a category and is found using the formula:

$$relative\ frequency = \frac{class\ frequency}{sum\ of\ all\ frequency} \qquad percentage\ frequency = \frac{class\ frequency}{sum\ of\ all\ frequency} \times 100\%$$

A variation of the basic frequency distribution is a ***relative frequency distribution***.

Pulse Rate	*Frequency*	*Relative Frequency*	*Relative Frequency %*
60 – 69	12	$\frac{12}{40} = 0.3$	$\frac{12}{40} \times 100\% = 30\%$
70 – 79	14	$\frac{14}{40} = 0.35$	$\frac{14}{40} \times 100\% = 35\%$
80 – 89	11	$\frac{11}{40} = 0.275$	$\frac{11}{40} \times 100\% = 27.5\%$
90 – 99	1	$\frac{1}{40} = 0.25$	$\frac{1}{40} \times 100\% = 2.5\%$
100 – 109	1	$\frac{1}{40} = 0.25$	$\frac{1}{40} \times 100\% = 2.5\%$
110 – 119	0	0	0
120 – 129	1	$\frac{1}{40} = 0.25$	$\frac{1}{40} \times 100\% = 2.5\%$
	40		

Example

The data below represent the color of M&Ms in a bag of plain M&Ms.

brown, brown, yellow, red, red, red, brown, orange, blue, green, blue, brown, yellow, yellow, brown, red, red, brown, brown, brown, green, blue, green, orange, orange, yellow, yellow, yellow, red, brown, red, brown, orange, green, red, brown, yellow, orange, red, green, yellow, yellow, brown, yellow, orange

Construct a frequency distribution and a relative frequency distribution of the color of plain M&Ms.

Solution

Color	*Tally*	*Frequency*	*Relative Frequency*
Brown	||||| ||||| ||	12	$\frac{12}{45} \approx 0.2667$
Yellow	||||| |||||	10	$\frac{10}{45} \approx 0.2222$
Red	||||| ||||	9	$\frac{9}{45} = 0.2$
Orange	||||| |	6	$\frac{6}{45} \approx 0.1333$
Blue	|||	3	$\frac{3}{45} \approx 0.0667$
Green	|||||	5	$\frac{5}{45} \approx 0.1111$
		45	

Construct Bar Graphs

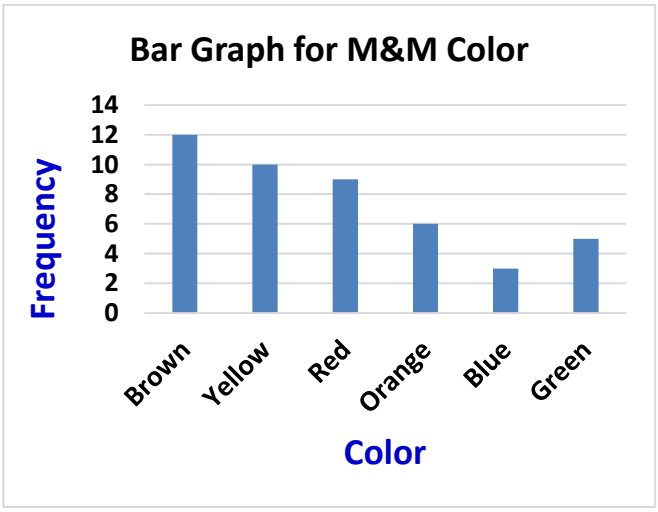
A ***bar graph*** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represents the category's frequency or relative frequency.

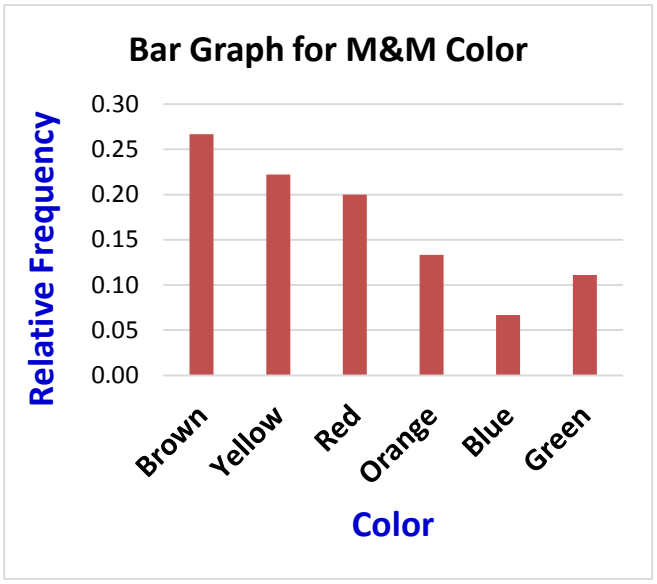
Example

Use the M&M data to construct

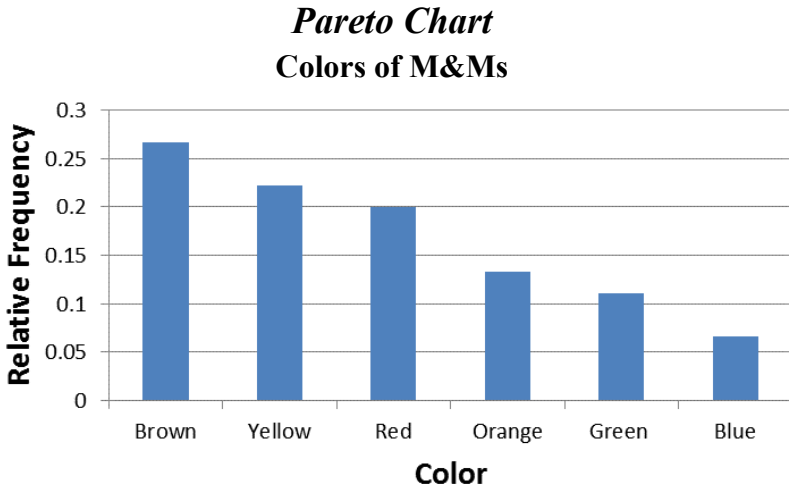
a) a frequency bar graph and
b) a relative frequency bar graph.

Solution





A ***Pareto chart*** is a bar graph where the bars are drawn in decreasing order of frequency or relative frequency.



Construct Pie Charts

A ***pie chart*** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

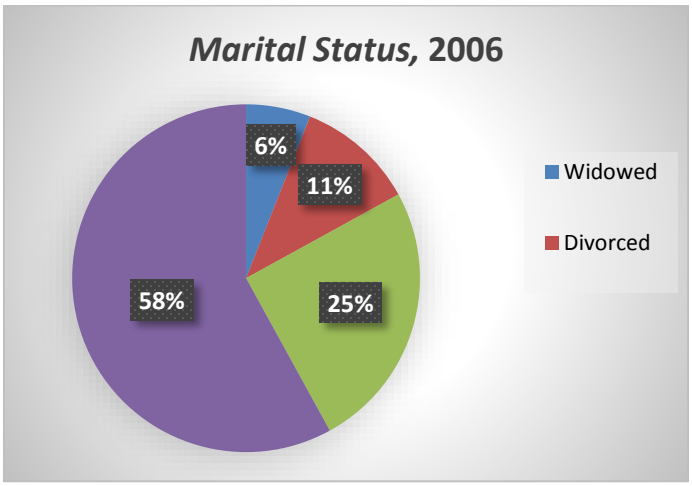
Example

The following data represent the marital status (in millions) of U.S. residents 18 years of age or older in 2006. Draw a pie chart of the data.

Marital Status	*Frequency*
Never married	55.3
Married	127.7
Widowed	13.9
Divorced	22.8

Solution

Marital Status	*Frequency*	
Never married	55.3	$\frac{55.3}{219.7}\times 100 \approx 25\%$
Married	127.7	$\frac{127.7}{219.7}\times 100 \approx 58\%$
Widowed	13.9	$\frac{13.9}{219.7}\times 100 \approx 6\%$
Divorced	22.8	$\frac{22.8}{219.7}\times 100 \approx 11\%$
	219.7	



Normal Distribution

- The ***frequencies*** start low, then increase to one or two high frequencies, then decrease to a low frequency.
- The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

Example

IQ scores from 1000 adults were randomly selected. The results are summarized in the frequency distribution table

IQ Score	*Frequency*	*Normal Distribution*
50 – 69	24	← Frequencies start low
70 – 89	228	
90 – 109	490	← Increase to a maximum, …
110 – 129	232	
130 – 149	26	← Decrease to become low again

The frequencies start low, then increase to a maximum frequency of 490, then decrease to low frequencies. Also, the frequencies are roughly symmetric about the maximum frequency of 490. It appears that the distribution is approximately a normal distribution.

Exercises *Section* 1.5 – Organizing Qualitative Data

1. In a study, researchers treated 570 people who smoke with either nicotine gum or a nicotine patch. Among those treated with nicotine gum, 191 continued to smoke and the other 59 stopped smoking. Among those treated with nicotine patch, 263 continued to smoke and the other 57 stopped smoking (based on data from the Center for Disease Control and Prevention). Construct the relative frequency distribution.

2. Heights of statistics students were obtained by the author as part of a study conducted for class. The last digits of those heights are listed below. Construct a frequency distribution with 10 classes. Based on the distribution, do the heights appear to be reported or actually measured? What do you know about the accuracy of the results?

0 0 0 0 0 0 0 0 0 1 1 2 3 3 3 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 8 8 8 9

3. Listed below are amounts of strontium-90 (in millibecquerels) in a simple random sample of baby teeth obtained from Pennsylvania residents born after 1979, Construct a frequency distribution with eight classes. Begin with a lower class limit of 110, and use a class width of 0. Cite a reason why such data are important.

155	142	149	130	151	163	151	142	156	133
138	161	128	144	172	137	151	166	147	163
145	116	136	158	114	165	169	145	150	150
150	158	151	145	152	140	170	129	188	156

4. Refer to the data below and use the 40 voltage measurements from the generator. Construct a frequency distribution with seven classes. Begin with a lower class limit of 123.9 volts, and use a class width of 0.20 volt. Using a very loose interpretation of the relevant criteria, does the result appear to have a normal distribution?

Day	Home	Generator	UPS	Day	Home	Generator	UPS
1	123.8	124.8	123.1	21	124.0	125.0	123.8
2	123.9	124.3	123.1	22	123.9	124.7	123.8
3	123.9	125.2	123.6	23	123.6	124.9	123.7
4	123.3	124.5	123.6	24	123.5	124.9	123.8
5	123.4	125.1	123.6	25	123.4	124.7	123.7
6	123.3	124.8	123.7	26	123.4	124.2	123.8
7	123.3	125.1	123.7	27	123.4	124.7	123.8
8	123.6	125.0	123.6	28	123.4	124.8	123.8
9	123.5	124.8	123.6	29	123.3	124.4	123.9
10	123.5	124.7	123.8	30	123.3	124.6	123.8
11	123.5	124.5	123.7	31	123.5	124.4	123.9
12	123.7	125.2	123.8	32	123.6	124.0	123.9
13	123.6	124.4	123.5	33	123.8	124.7	123.9
14	123.7	124.7	123.7	34	123.9	124.4	123.9
15	123.9	124.9	123.0	35	123.9	124.6	123.6
16	124.0	124.5	123.8	36	123.8	124.6	123.2
17	124.2	124.8	123.8	37	123.9	124.6	123.1
18	123.9	124.8	123.1	38	123.7	124.8	123.0

19	123.8	124.5	123.7	39	123.8	124.3	122.9
20	123.8	124.6	123.7	40	123.8	124.0	123.0

5. As part of the Garbage Project at the University of Arizona, the discarded garbage for 62 households was analyzed. Refers to the 62 weights from table below and construct a frequency distribution. Begin with a lower class of 1.00 lb., and use a class width of 4.00 lb. Do the weights of discarded paper appear to have a normal distribution?

2.41	11.08	9.45	5.88
7.57	12.43	12.32	8.26
9.55	6.05	20.12	12.45
8.82	13.61	7.72	10.58
8.72	6.98	6.16	5.87
6.96	14.33	7.98	8.78
6.83	13.31	9.64	11.03
11.42	3.27	8.08	12.29
16.08	6.67	10.99	20.58
6.38	17.65	13.11	12.56
13.05	12.73	3.26	9.92
11.36	9.83	1.65	3.45
15.09	16.39	10	9.09
2.8	6.33	8.96	3.69
6.44	9.19	9.46	2.61
5.86	9.41		

6. *a*) Refer to the data below for the FICO credit rating scores. Construct a frequency distribution beginning with a lower class limit of 400, and use a class width of 50. Does the result appear to have a normal distribution? Why or why not?

708	713	781	809	797	793	711	681	768	611	698	729	829
836	768	532	657	559	741	792	701	753	745	681	594	744
598	693	743	444	502	739	755	835	714	517	787	706	752
714	497	636	637	797	568	714	618	830	579	818	722	783
751	731	850	591	802	756	689	789	654	617	849	604	630
628	692	779	756	782	760	503	784	798	611	709	661	579
591	834	694	795	660	651	696	638	697	732	796	753	782
635	795	519	682	824	603	709	777	664				

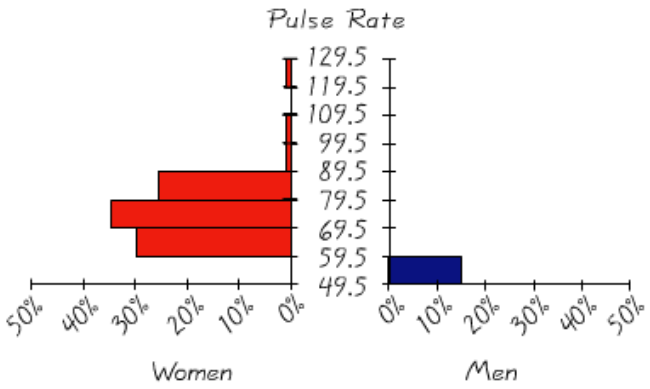
b) Use the table to construct a histogram. Does the result appear to be normal distribution? Why or why not?

7. *a*) Refer to the data in the table below. Construct a frequency distribution. Begin with lower class limit of 6.0000 g, and use a class width of 0.0500 g.

6.2771	6.2371	6.1501	6.0002	6.1275	6.2151	6.1947	6.1940
6.2866	6.0760	6.1426	6.3415	6.1309	6.2412	6.2130	6.0257
6.1442	6.1073	6.1181	6.1352	6.2821	6.2647	6.1787	6.1719
6.2908	6.1661	6.2674	6.2718	6.1949	6.2465	6.1095	6.3278
6.3172	6.1487	6.0829	6.1423	6.1970	6.2441	6.0775	6.3669

b) Use the table to construct a histogram.

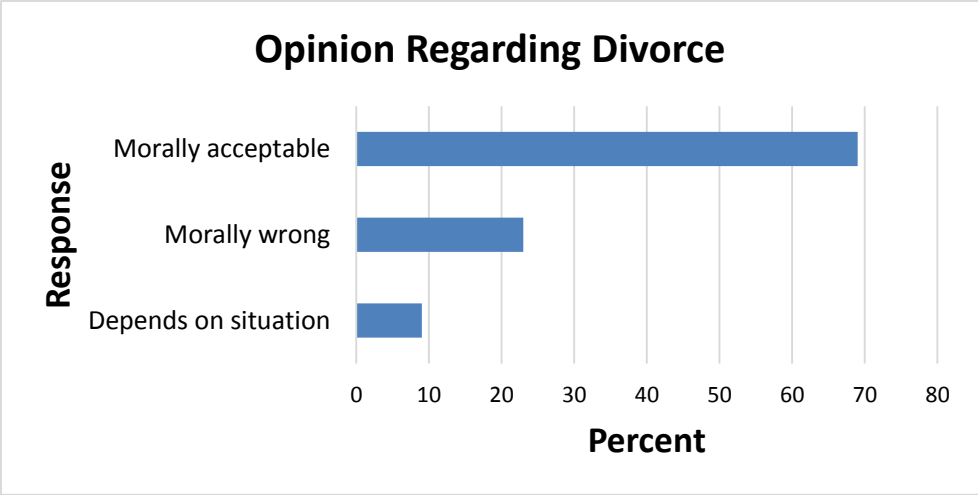
8. When using histograms to compare two data sets, it is sometimes difficult to make comparisons by looking back and forth between the two histograms. A back-to-back relative frequencies histogram uses a format that makes the comparison much easier. Instead of frequencies, we should use relative frequencies (percentages or proportions) so that the comparisons are not distorted by different sample sizes. Complete the back-to-back relative frequency histograms shown below by using the data below. Then use the result to compare the two data sets.



Pulse Rates (*beats per minute*) of Females and Males

Females																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	93	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
Males																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

9. The following graph represents the results of a survey, by Gallup in May 2010, in which a random sample of adult Americans was asked, "Please tell me whether you personally believe that in general divorce is morally accepted or morally wrong."



a) What percent of the respondents believe divorce is morally acceptable?
b) If there were 240 million adult Americans, how many believe that divorce is morally wrong?
c) If Gallup claimed that the results of the survey indicate that 8% of adult Americans believe that divorce is acceptable in certain situations, would you say this statement is descriptive or inferential? Why?

10. In a national survey conducted by the Centers for Disease Control to determine health-risk behaviors among college students, college students were asked, "How often do you wear a seat belt when driving a car?" The frequencies were as follow:

Response	*Frequency*
I do not drive a car	249
Never	118
Rarely	249
Sometimes	345
Most of the time	716
Always	3093

a) Construct a relative frequency distribution
b) What percentage of respondents answered "Always"?
c) What percentage of respondents answered "Never" or "Rarely"?
d) Construct a frequency bar graph.
e) Construct a relative frequency bar graph.
f) Construct a pie chart
g) Suppose that a representative from the Centers for Disease Control says, "2.5% of the college students in this survey responded that they never wear a seat belt." Is this a descriptive or inferential statement?

11. A phlebotomist draws the blood of a random sample of 50 patients and determines their blood types as shown.

O	*B*	*AB*	*O*	*AB*	*O*	*O*	*O*	*O*	*O*
O	*O*	*B*	*O*	*O*	*A*	*A*	*B*	*O*	*A*
A	*B*	*A*	*A*	*A*	*A*	*O*	*A*	*O*	*O*
A	*A*	*B*	*A*	*B*	*O*	*AB*	*A*	*A*	*A*
O	*O*	*AB*	*O*	*A*	*A*	*A*	*O*	*O*	*O*

a) Construct a frequency distribution
b) Construct a relative frequency distribution
c) According to the data, which blood type is most common?
d) According to the data, which blood type is least common?
e) Use the results of the sample to conjecture the percentage of the population that has type *O* blood. Is this an example of descriptive or inferential statistics?
f) Contact a local hospital and ask them the percentage of the population that us blood type *O*. Why might the results differ?
g) Draw a frequency bar graph
h) Draw a relative frequency bar graph
i) Draw a pie chart