Organizing Data Using Tables and Graphs

Organizing Data

Raw, unorganized leadership scores of 35 managers:

54	43	48	50	52	44	49
44	46	51	42	50	46	50
51	55	57	48	53	51	48
46	52	50	45	55	49	53
48	50	49	51	48	43	52

Frequency Distributions – provide organization to a set of scores.

Simple frequency distributions – list the frequencies with which each raw score occurs.

Simple Frequency Distribution for Previous Leadership Scores

X	Tally	*f*
57	|	1
56		0
55	||	2
54	|	1
53	||	2
52	|||	3
51	||||	4
50	𝍸	5
49	|||	3
48	𝍸	5
47		0
46	|||	3
45	|	1
44	||	2
43	||	2
42	|	1
		N = 35

To create a simple frequency distribution:

1. Create labels for three columns, as follows: *X*, Tally, and *f*.
2. Locate the highest and lowest scores in the unorganized list of scores.
3. Beginning with the highest score at the top, list the score values in descending order in the "*X*" column of your frequency distribution. Do not skip any values even if there were no occurrences of some of the values in your list of scores. Stop at the lowest obtained score.
4. Underline the first score in your unorganized list and place a tally mark for that score in the "Tally" column of your frequency distribution. Underlining the scores helps you to keep track of your place on the list. Continue this process until all of the scores in your list have been underlined.
5. Count the number of tally marks for each score and record this number in the "*f*" column.
6. Count the number of scores in the "*f*" column. The sum should be equal to the total number of scores (*N*).

For Example,

Raw scores on a 10-point statistics quiz:

8	3	5	6	10
6	9	8	9	6
5	7	10	5	8
7	8	3	8	10
6	7	8	7	6
10				

Simple frequency distribution of scores for the statistics quiz:

X	Tally	*f*
10	||||	4
9	||	2
8	𝍸 |	6
7	||||	4
6	𝍸	5
5	|||	3
4		0
3	||	2
		N =26

Notice that even though there were no occurrences of 4, it is included in the list of scores.

Relative Frequency Distribution

Simple frequency distribution

- tells *how many times* a particular scores occurs.

Relative frequency distribution

- tells the *proportion of time* that the score occurs.

$$Rel.f = \frac{f}{N}$$

Simply add a column to the simple frequency distribution table and divide each *f* by *N*.

For Example,

Relative Frequency Distribution for scores on the statistics quiz

X	*f*	Rel. *f*
10	4	.15
9	2	.08
8	6	.23
7	4	.15
6	5	.19
5	3	.12
4	0	.00
3	2	.08
	N = 26	1.00

In the above distribution, the relative frequency of a score of 6 was .19. In other words, the score of 6 occurred 19% of the time.

Cumulative Frequency Distribution

Cumulative frequency distribution

- tells the *frequency* of scores that fall at or below a particular score value.
- useful if you want to know how many people scored below someone on a test.

To create a cumulative frequency distribution:

1. Begin with a simple frequency distribution table.
2. Add on a cumulative frequency (*cf*) column.
3. Work from the bottom up in the *f* and *cf* columns. Take the bottom score in the *f* column and enter that number in the corresponding space in the *cf* column. Then take that number (in *cf* column) and add it to the next number up in the *f* column and record the total in the next space in the *cf* column. Take the last number entered in the *cf* column and add it to the next number up in the *f* column. Repeat this process until the *cf* column is complete. The number at the top of the *cf* column should be equal to *N*.

For Example,

Cumulative Frequency Distribution for scores on the statistics quiz:

X	*f*	*cf*	
10	4	26	(the *cf* at the top is equal to *N*)
9	2	22	
8	6	20	
7	4	14	
6	5	10	
5	3	5	(*cf* of 2 below + *f* of 3 = 5)
4	0	2	(since no students scored 4, the *cf* is still 2)
3	2	2	

N = 26

Percentile Rank

Percentile rank (P.R.)

- tells the *percentage* of scores that fall at or below a given score.

$$P.R. = \frac{cf}{N} \times 100$$

For Example,

Percentile Rank for Scores on Statistics Quiz:

X	*f*	*cf*	*P.R.*
10	4	26	100.00
9	2	22	84.62
8	6	20	76.92
7	4	14	53.85
6	5	10	38.46
5	3	5	19.23
4	0	2	7.69
3	2	2	7.69
	N = 26		

For example, the percentile rank of a student who scored 8 is 76.92. In other words, approximately 77% of the students scored at or below that score and only 23% of the students achieved scores above 8.

For Example,

Percentile Rank for Scores on Statistics Quiz:

$$P.R. = \frac{cf}{N} \times 100$$

X	*f*	*cf*	*P.R.*
10	4	26	100.00
9	2	22	84.62
8	6	20	76.92
7	4	14	53.85
6	5	10	38.46
5	3	5	19.23
4	0	2	7.69
3	2	2	7.69
	N = 26		

For example, the percentile rank of a student who scored 8 is 76.92. In other words, approximately 77% of the students scored at or below that score and only 23% of the students achieved scores above 8.

Combining the Tables!

It is more efficient to create one table with several columns.

For Example,
Frequency Distribution for Scores on Statistics Quiz

X	*f*	*Rel. f*	*cf*	*P.R.*
10	4	.15	26	100.00
9	2	.08	22	84.62
8	6	.23	20	76.92
7	4	.15	14	53.85
6	5	.19	10	38.46
5	3	.12	5	19.23
4	0	.00	2	7.69
3	2	.08	2	7.69

Grouped Frequency Distributions

With a lot of scores, ungrouped frequency distributions are:

- cumbersome.
- data are unremarkable.

Grouped frequency distributions:

- combine scores into groups (class intervals).
- makes trends more apparent.

Ungrouped Simple Frequency Distribution

X	*f*	*X*	*f*	*X*	*f*
81	1	69	3	57	3
80	2	68	4	56	3
79	0	67	2	55	1
78	1	66	3	54	0
77	2	65	9	53	2
76	1	64	8	52	1
75	3	63	4	51	0
74	5	62	5	50	0
73	4	61	6	49	3
72	2	60	4	48	2
71	0	59	6	47	1
70	4	58	4	46	1

$N = 100$

Difficult to see any meaningful patterns.

Grouped frequency distribution for previous simple frequency distribution:

Class Interval	*f*
81 – 83	1
78 – 80	3
75 – 77	6
72 – 74	11
69 – 71	7
66 – 68	9
63 – 65	21
60 – 62	15
57 – 59	13
54 – 56	4
51 – 53	3
48 – 50	5
45 – 47	2

We can now more easily pick up patterns of the scores. Most are in the middle range, while just a few were very high or very low.

Before the looking at the actual steps for creating a grouped frequency distribution, we will consider some new terms and ground rules.

New terms:

- Class intervals refer to groups of scores, such as 6–7 or 21–23.
- Interval size (*i*) - number of scores in the class interval. For example, *i* = 3 for the class interval of 21–23 since there are three scores in the interval.
- Range (R) - amount of spread in a distribution of scores; determined by subtracting the lower limit (LL) of the lowest score from the upper limit (UL) of the highest score.

$$R = X_{UL\text{-}High} - X_{LL\text{-}Low}$$

For example, if the lowest score in a distribution is 53 and the highest score is 81, then the range for that distribution would be 29 (R = 81.5 – 52.5).

Ground rules:

- Keep the number of class intervals both meaningful and manageable (i.e., between 10 and 20).
- Use interval sizes (*i*) of 2, 3, or 5.
- Begin the class interval column at the bottom with a multiple of *i*. For example, if *i* = 3 in a distribution where the lowest raw score is 31, the first class-interval should be 30-32 because 30 is a multiple of 3.
- The largest scores go at the top of the distribution.

Finally, the steps:

1. Locate the highest and lowest scores in the distribution and find the range of scores.
2. The size of the class interval (*i*) to use will be determined by trial-and-error, keeping the above ground rules in mind. Divide the range by a potential *i* value. This will tell you approximately how many class intervals would result (remember, we will keep this number between 10 and 20). We will experiment with the standard sizes of 2, 3, and 5.

    - Suppose the highest score in our distribution is 64 and the lowest is 23. The range, then, is 64.5 – 22.5 = 42.
    - Experiment by dividing the range by potential *i* values:

        42 ÷ 2 = 21 (too many)
        42 ÷ 3 = 14 (just right)
        42 ÷ 5 = 8.4 (too few)

    - In this case, we would use 3 as our interval size with a result of approximately 14 class intervals.

3. Create labels for three columns as follows: Class Interval, Tally, and *f*.
4. Using lined paper, count down the number of spaces needed to accommodate all of your class intervals. It's a good idea to add a couple of extra spaces since the division in step 2 only approximates the number of class intervals.
5. Start at the bottom space of the class interval column and being creating your class intervals. Remember that your first entry should begin with a multiple of *i* and should contain the lowest score in the data set. You may need to begin with a value that is lower than the lowest score to accomplish this.

For example,

Bowling scores of a high school gym class:

82	128	110	140	127
109	92	119	142	111
126	85	124	138	92
83	114	146	112	86
98	132	128	95	120
122	115	92	116	119
79	113	81	112	115

Step 1: The high score is 146 and the low score is 79.
Thus, R = 146.5 – 78.5 = 68.

Step 2: 68 ÷ 2 = 34 (too many)
68 ÷ 3 = 22.67 (too many)
68 ÷ 5 = 13.60 (just right, $i = 5$)

Step 3: ⟶

Class Interval	Tally	*f*
145 - 149	|	1
140 – 144	||	2
135 – 139	|	1
130 – 134	|	1
125 – 129	||||	4
120 – 124	|||	3
115 – 119	~~||||~~	5
110 – 114	~~||||~~ |	6
105 – 109	|	1
100 – 104		0
95 – 99	||	2
90 – 94	|||	3
85 – 89	||	2
80 – 84	|||	3
75 – 79	|	1

Steps 4 & 5: ⟶

N = 35

Graphic Representation

Graphs are usually displayed on two axes, one horizontal and one vertical.

- The horizontal axis is the x-axis, also called the baseline or abscissa, and it usually represents the values or categories of the variable being measured, such as scores on a test or military rank.

- The vertical axis is the y-axis, also called the ordinate, and it usually represents the frequencies of those values or categories.

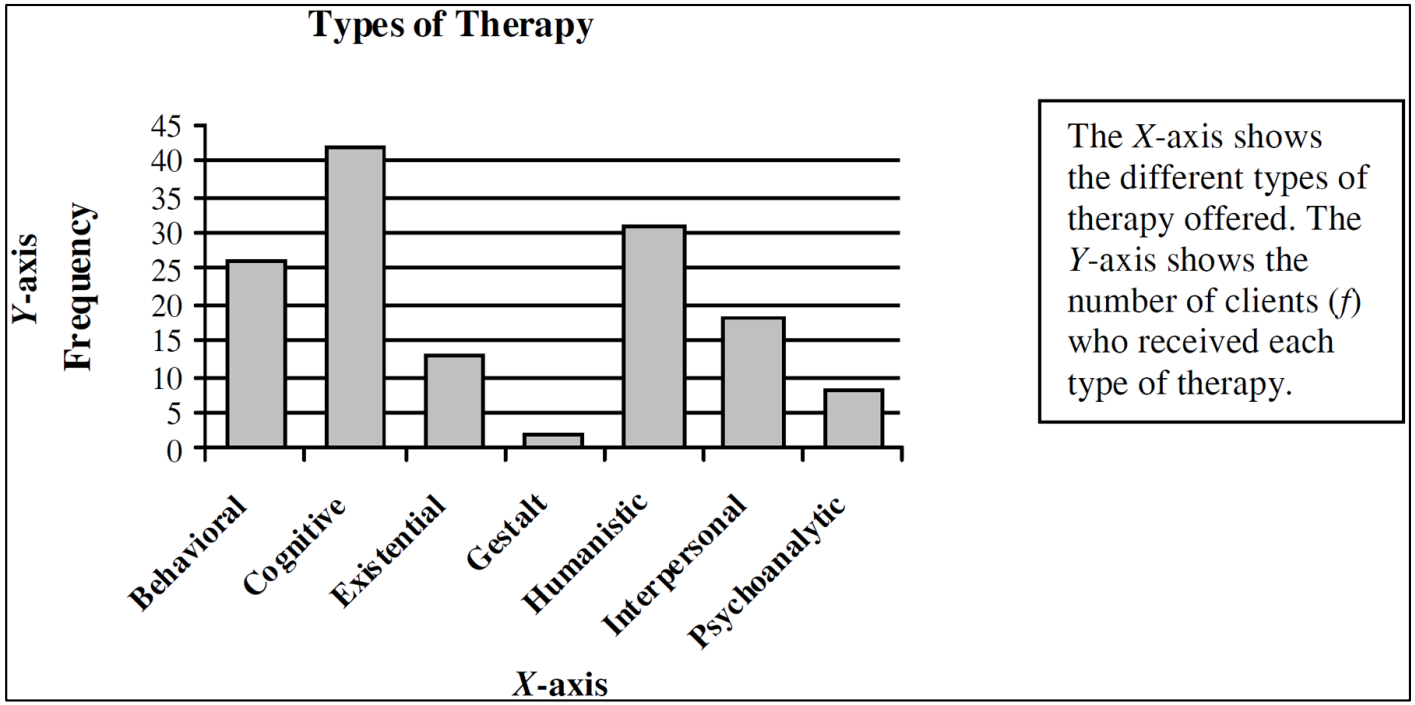
- The most common graphs are bar graphs, histograms, and frequency polygons.

Bar graphs

- used for qualitative variables that differ in kind (nominal and ordinal scales).
- bars are spatially separated.
- heights of the bars reflect the frequencies of the event.

For Example,

Frequencies of seven different types of therapy received by 140 clients at a mental health clinic in one month:

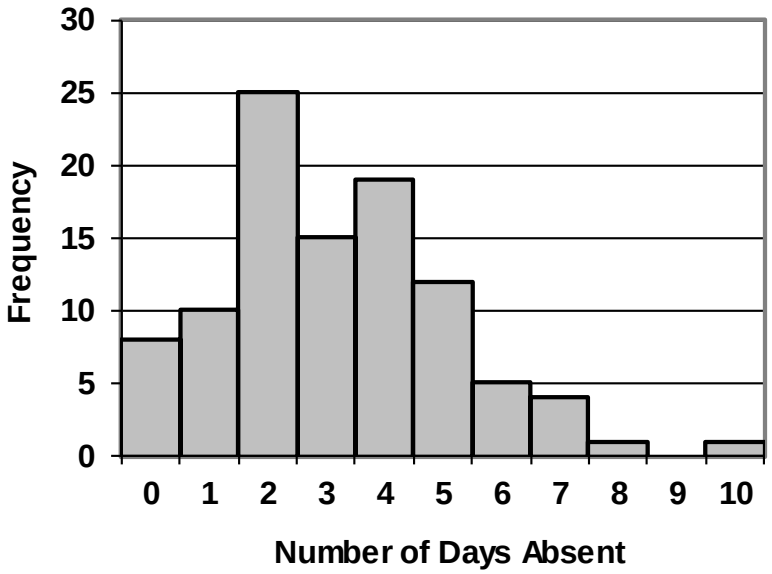


Histograms

- used for quantitative variables that differ in amount (interval and ratio).
- bars touch each other.

For Example,

Histogram for number of days absent of 100 employees over a 12 month period:

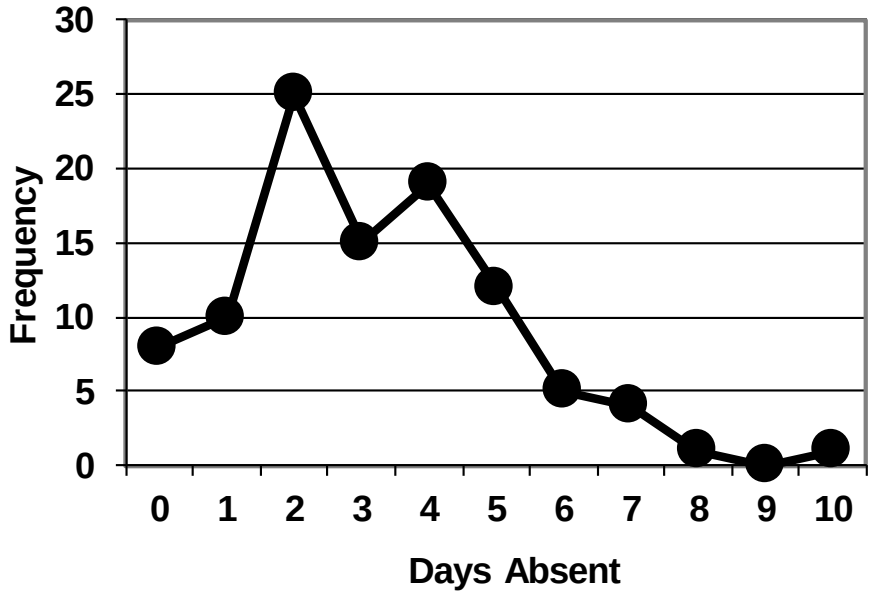


Frequency polygons

- also used for quantitative data.
- except that dots are used instead of bars.
- dots are then connected by a straight line.
- useful for making comparisons between different distributions.

For Example,

Frequency Polygon for number of days absent of 100 employees over a 12 month period:



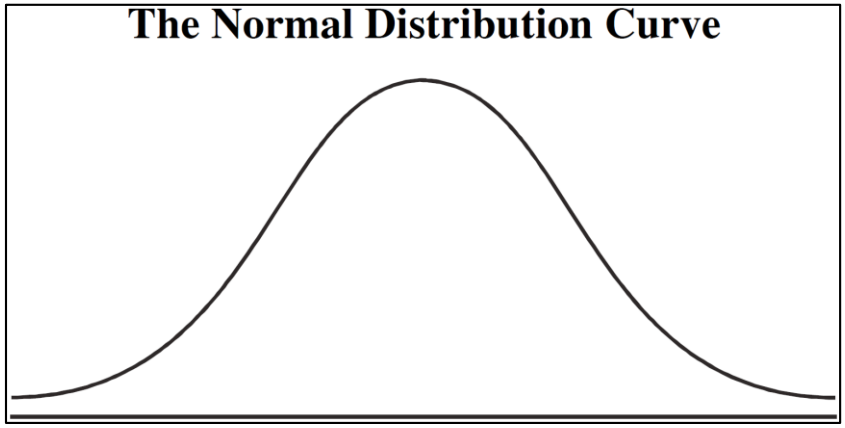
Types and Shapes of Frequency Polygons

Empirical distributions

- based on frequencies of actual scores represented by dots.

Theoretical distributions

- based on the mathematical probability of the frequencies of scores in a population.
- drawn with smooth lines without dots since actual scores are not represented.



Important characteristics of the bell curve are as follows:

- It is symmetrical.
- Tails are asymptotic (never touch the baseline).
- The most frequently occurring scores are in the middle.
- The least frequently occurring scores are furthest away from the middle.

Skewed distributions

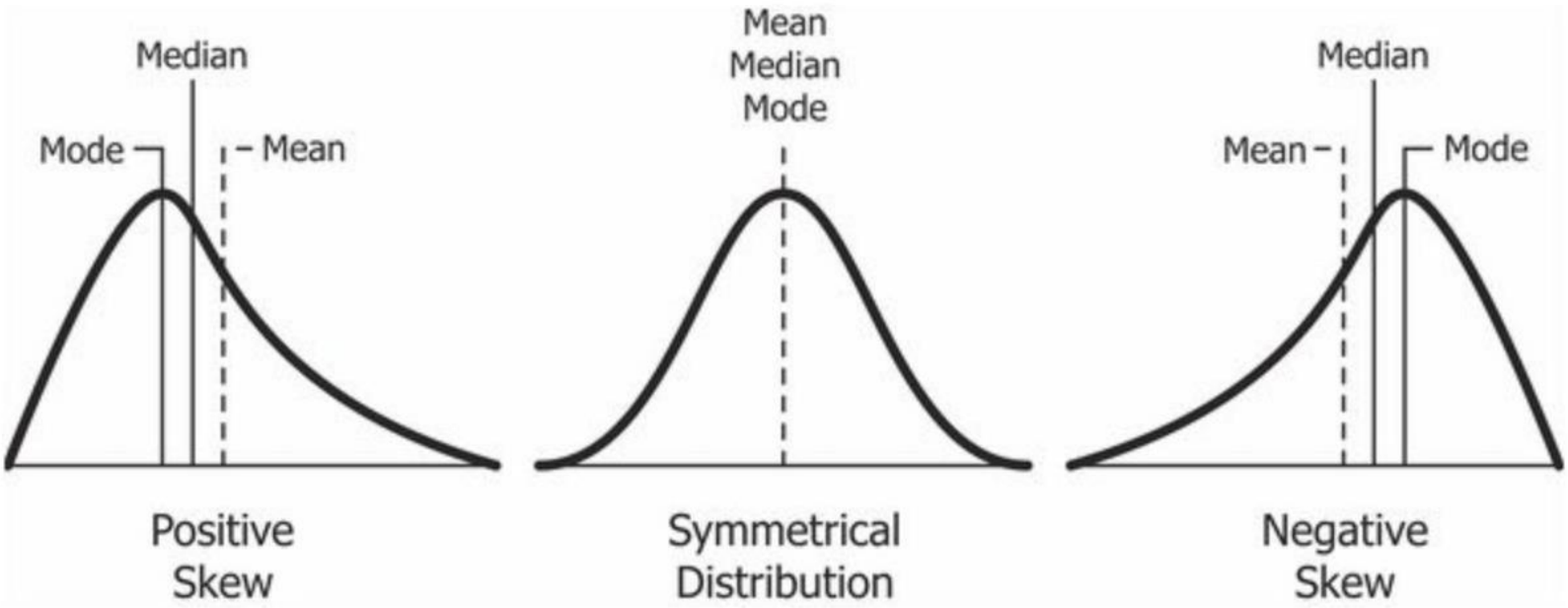
have scores that tend to stack up at either the high or low end of the distribution and trail off at the other end.

Positively skewed distribution

- more low scores than high scores.

Negatively skewed distribution

- more high scores than low scores.


Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

A. Qualitative Variables
B. Quantitative Variables
1. Stem and Leaf Displays
2. Histograms
3. Frequency Polygons
4. Box Plots
5. Bar Charts
6. Line Graphs
7. Dot Plots

Graphing Qualitative Variables

Learning Objectives

1. Create a frequency table
2. Determine when pie charts are valuable and when they are not
3. Create and interpret bar charts
4. Identify common graphical mistakes

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" of 0.17 = 85/500.

Table 1. Frequency Table for the iMac Data.

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1

Pie Charts

The pie chart in Figure 1 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

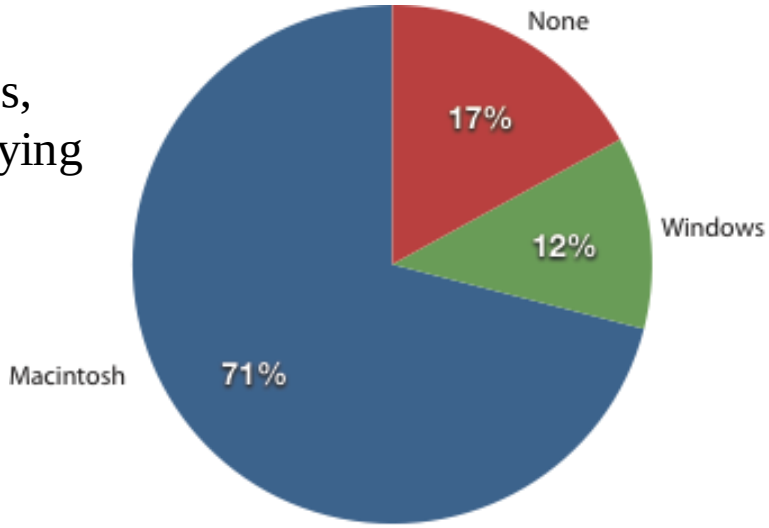


Figure 1. Pie chart of iMac purchases illustrating frequencies of previous computer ownership.

Bar charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure 2. Frequencies are shown on the Y-axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations in each category as is typical in pie charts.

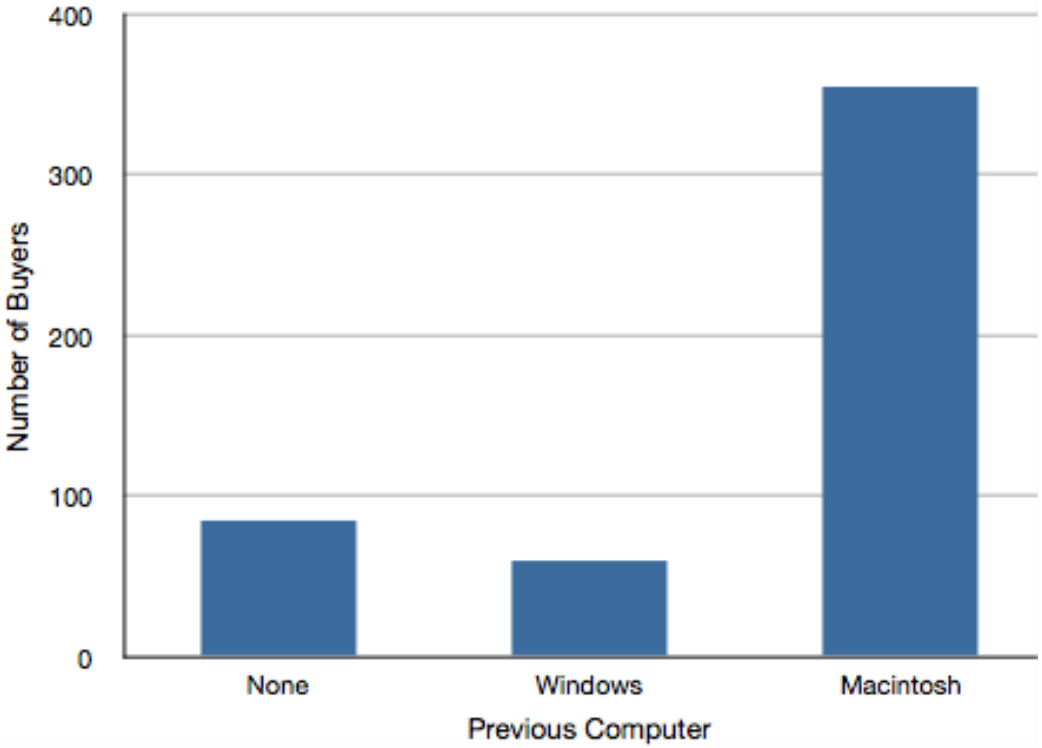


Figure 2. Bar chart of iMac purchases as a function of previous computer ownership.

Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 3 shows the number of people playing card games at the Yahoo web site on a Sunday and on a Wednesday in the spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

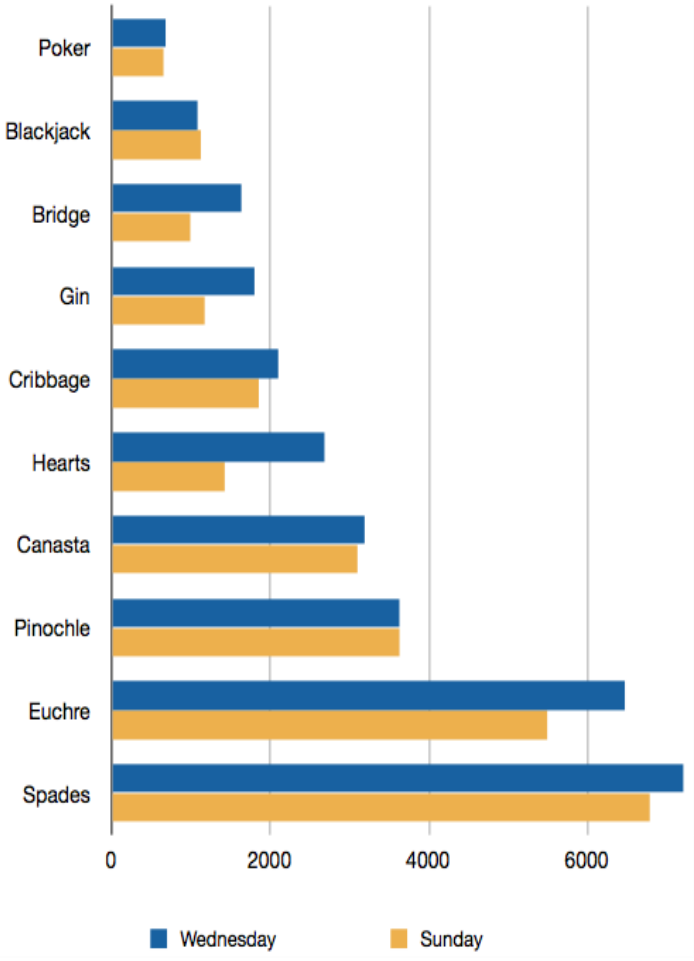
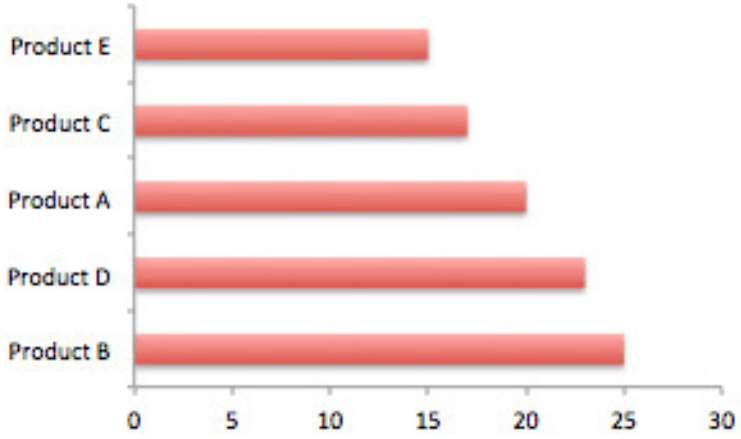


Figure 3. A bar chart of the number of people playing different card games on Sunday and Wednesday

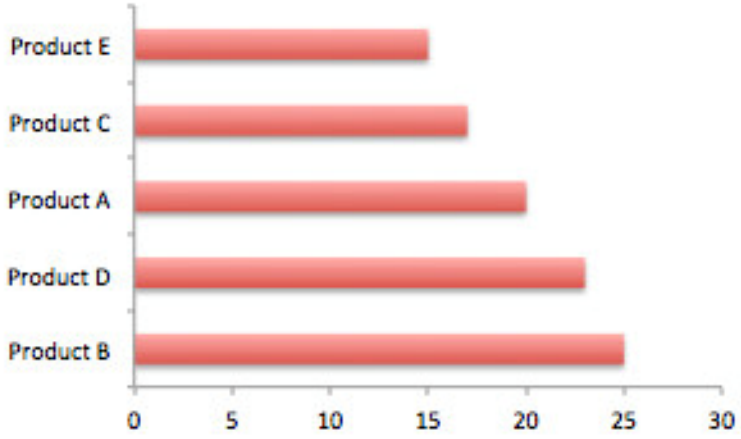
Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.


Product A
Product B
Product C
Product D
Product E


Product E
Product C
Product A
Product D
Product B
0
5
10
15
20
25
30


Product A
Product B
Product C
Product D
Product E


Product E
Product C
Product A
Product D
Product B
0
5
10
15
20
25
30