Exploring Data with Graphs

Glossary

1. Frequency Distribution: A table which shows how data is partitioned among several categories by specifying the frequency in that category.
2. Lower class limits: are the smallest numbers that belong to each class
3. Upper class limits: are the largest numbers that belong to each class
4. Class boundaries: are the numbers used to separate the classes, but without the gaps created by class limits.
5. Class midpoints: values in the middle of the classes
6. Class width: Upper class - lower class also (max - min)/no of classes
7. We construct frequency distributions to

   (1) summarize large data sets,

   (2) see the distribution and identify outliers, and

   (3) have a basis for constructing graphs (such as histograms)
8. Manual Steps to construct a frequency distribution
   a. find no of classes { 1 + (log n)/(log 2) }

b. find class width { Upper class - lower class also (max - min)/no of classes } ... **Round UP**

c. Choose the value for the first lower class limit by using either the minimum value or a convenient value below the minimum.

d. Using the first lower class limit and the class width, list the other lower class limits.

e. List the lower class limits in a vertical column and then determine and enter the upper class limits.

f. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.

9. T**ool for Frequency Distribution**
{https://www.socscistatistics.com/descriptive/frequencydistribution/default.aspx }

Time (seconds)	Frequency
75–124	11
125–174	24
175–224	10
225–274	3
275–324	2

10. **Relative Frequency Distribution**

a. Relative : frequency of a class / sum of all frequencies

b. Percentage : frequency of a class / sum of all frequencies * 100

Time (seconds)	Relative Frequency
75–124	22%
125–174	48%
175–224	20%
225–274	6%
275–324	4%

11. **Cumulative Frequency Distribution**
    Another variation of a frequency distribution is a cumulative frequency distribution in which the frequency for each class is the sum of the frequencies for that class and all previous classes.

Time (seconds)	Cumulative Frequency
Less than 125	11
Less than 175	35
Less than 225	45
Less than 275	48
Less than 325	50

12. N**OTE:**
    a. **Frequencies of last digits sometimes reveal how the data was collected or measured. The presence of gaps implies that the data was collected from two or more different populations.**
    b. **Combining two or more relative frequency distributions in one table makes comparisons of data much easier.**

Time (seconds)	McDonald's	Dunkin' Donuts
25–74		22%
75–124	22%	44%
125–174	48%	28%
175–224	20%	6%
225–274	6%	
275–324	4%	

13. **HISTOGRAM**

a. A histogram is a graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data). The horizontal scale represents classes of quantitative data values, and the vertical scale represents frequencies. The heights of the bars correspond to frequency values.

b. **Important Uses of a Histogram**

- Visually displays the shape of the distribution of the data
- Shows the location of the center of the data
- Shows the spread of the data
- Identifies outliers

c. A **relative frequency histogram** has the same shape and horizontal scale as a histogram, but the vertical scale uses relative frequencies (as percentages or proportions)

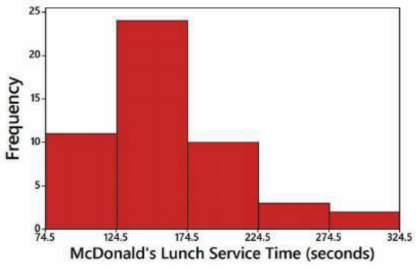


FIGURE 2-2 Histogram of McDonald's Drive-Through Lunch Service Times (seconds)

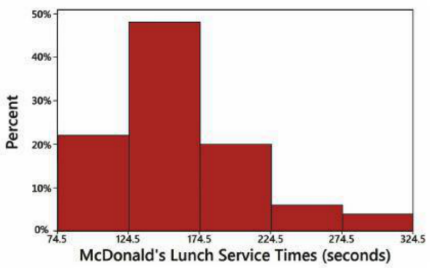


FIGURE 2-3 Relative Frequency Histogram of McDonald's Drive-Through Lunch Service Times (seconds)

14. "CVDOT": the

a. center of the data,

b. the variation (which will be discussed at length in Section 3-2),

c. the shape of the distribution,

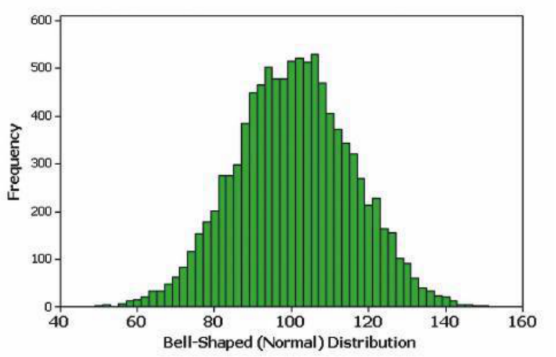
d. whether there are any outliers (values far away from the other values), and

e. time (whether there is any change in the characteristics of the data

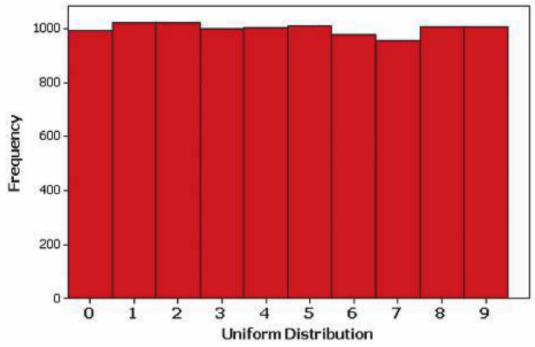
15.

Common Distribution Shapes

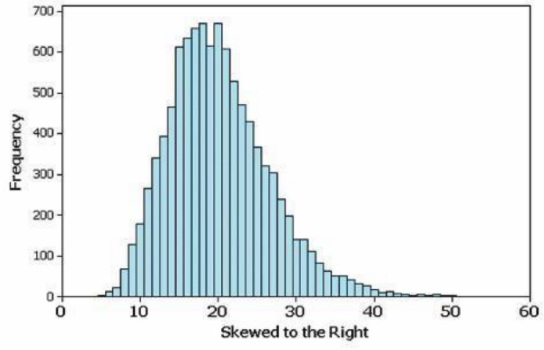
The histograms shown in Figure 2-4 depict four common distribution shapes.

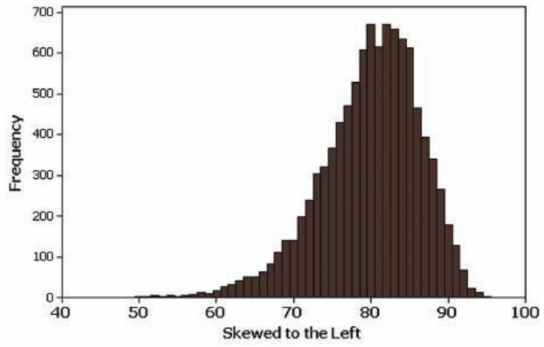


(a)



(b)





16. **Skewness**

a. A distribution of data is skewed if it is not symmetric and extends more to one side than to the other.

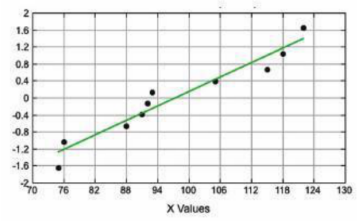
b. Data skewed to the right (also called positively skewed) have a longer right tail,

c. Data skewed to the left (also called negatively skewed) have a longer left tail,

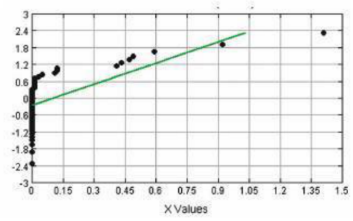
17. A**ssessing Normality with Normal Quantile Plots**

a. Normal Distribution: The population distribution is normal
if the pattern of the points in the normal quantile plot is reasonably close to a straight line,
the points do not show some systematic pattern that is not a straight-line pattern.

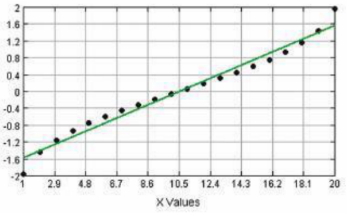
b. Not a Normal Distribution: The population distribution is not normal if the normal quantile plot has either or both of these two conditions:
• The points do not lie reasonably close to a straight-line pattern.
• The points show some systematic pattern that is not a straight-line pattern.



Normal Distribution: The points are reasonably close to a straight-line pattern, and there is no other systematic pattern that is not a straight-line pattern.



Not a Normal Distribution: The points do not lie reasonably close to a straight line.

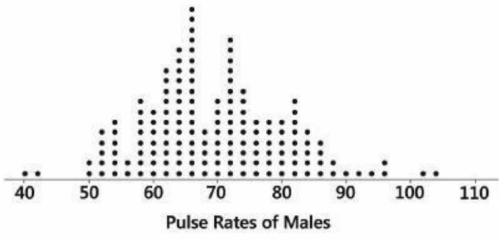


Not a Normal Distribution: The points show a systematic pattern that is not a straight-line pattern.

18. **GRAPHS THAT ENLIGHTEN**

a. **Dotplots: A dotplot consists of a graph of quantitative data in which each data value is plotted as a point (or dot) above a horizontal scale of values. Dots representing equal values are stacked.**
1. Displays the shape of the distribution of data.
2. It is usually possible to recreate the original list of data

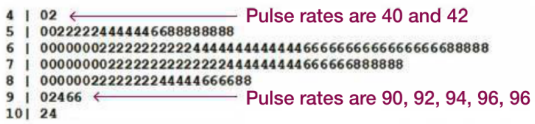
Dotplot



values.

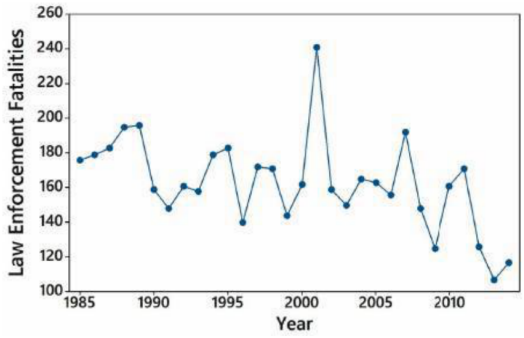
b. Stemplots: A stemplot (or stem-and-leaf plot) represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit).
   1. Shows the shape of the distribution of the data.
   2. Retains the original data values.
   3. The sample data are sorted (arranged in order).

Stemplot



c. Time-Series graph: A time-series graph is a graph of time-series data, which are quantitative data that have been collected at different points in time, such as monthly or yearly.
   1. Reveals information about trends over time

Time-series graph

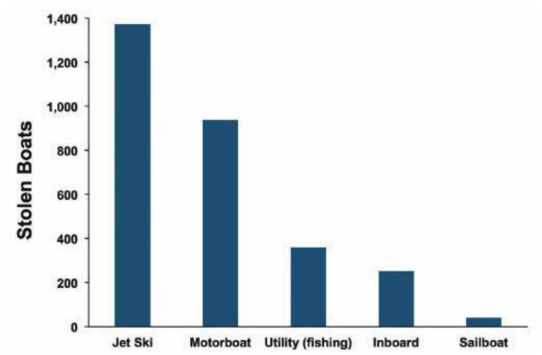


d. Bar Graph: A bar graph uses bars of equal width to show frequencies of categories of categorical (or qualitative) data. The bars may or may not be separated by small gaps.
   1. Shows the relative distribution of categorical data so that it is

Pareto Chart

easier to compare the different categories

e. Pareto Chart: A Pareto chart is a bar graph for categorical data, with the added stipulation that the bars are arranged in descending order according to frequencies, so the bars decrease in height from left to right.
   1. Shows the relative distribution of categorical data so that it is easier to compare the different categories
   2. Draws attention to the more important categories

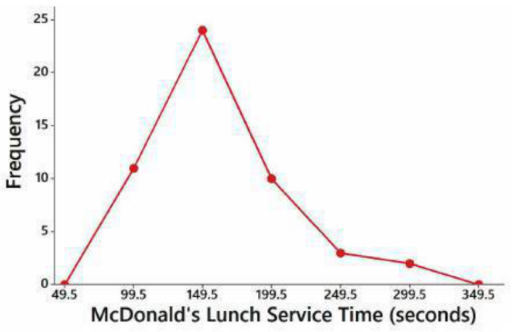


f. Pie Chart: A pie chart is a very common graph that depicts categorical data as slices of a circle, in which the size of each slice is proportional to the frequency count for the category.

   Never use pie charts because they waste ink on components that are not data, and they lack an appropriate scale.

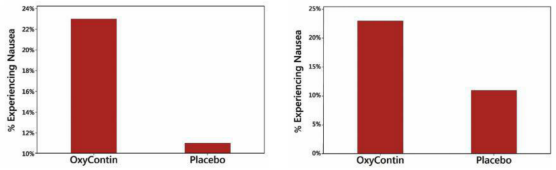
g. Frequency Polygon: A frequency polygon uses line segments connected to points located directly above class midpoint values. A frequency polygon is very similar to a histogram, but a frequency polygon uses line segments instead of bars.

Frequency Polygon

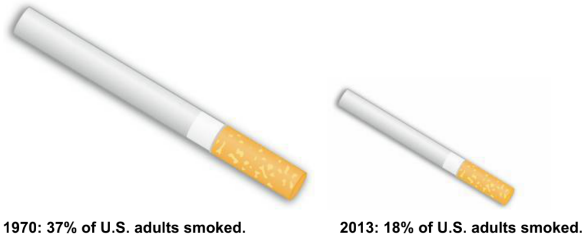


19. GRAPHS THAT DECEIVE

a. NONZERO AXIS: Always examine a graph carefully to see whether a vertical axis begins at some point other than zero so that differences are exaggerated.



b. Drawings of objects, called pictographs, are often misleading. Data that are one dimensional in nature (such as budget amounts) are often depicted with two dimensional objects (such as dollar bills) or three-dimensional objects (such as stacks of coins, homes, or barrels).



1970: 37% of U.S. adults smoked. 2013: 18% of U.S. adults smoked.

20. In The Visual Display of Quantitative Information, Edward Tufte offers these principles:

■ For small data sets of 20 values or fewer, use a table instead of a graph.

■ A graph of data should make us focus on the true nature of the data, not on other elements, such

as eye-catching but distracting design features.

■ Do not distort data; construct a graph to reveal the true nature of the data.

■ Almost all of the ink in a graph should be used for the data, not for other design

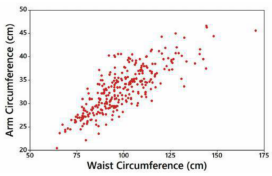


FIGURE 2-14 **Waist and Arm Circumferences**
Correlation: The distinct pattern of the plotted points suggests that there is a correlation between waist circumferences and arm circumferences.

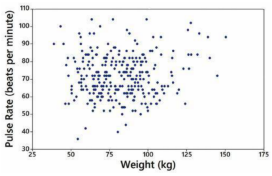


FIGURE 2-15 **Weights and Pulse Rates**
No Correlation: The plotted points do not show a distinct pattern, so it appears that there is no correlation between weights and pulse rates.

21. Scatterplot and Correlation

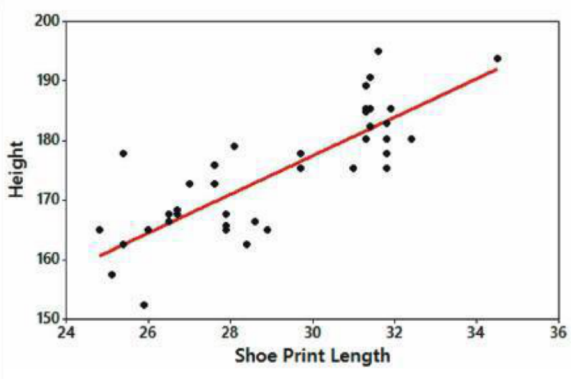
    a. A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

    b. A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

    c. A **scatterplot** (or **scatter diagram**) is a plot of paired (*x*, *y*) quantitative data with a horizontal *x*-axis and a vertical *y*-axis. The horizontal axis is used for the first variable (*x*), and the vertical axis is used for the second variable (*y*).

22. Linear Correlation Coefficient r

a. The **linear correlation coefficient** is denoted by *r*, and it measures the strength of the linear association between two variables.



Correlation If the computed linear correlation coefficient *r* lies in the left or right tail region beyond the table value for that tail, conclude that there is sufficient evidence to support the claim of a linear correlation.

No Correlation If the computed linear correlation coefficient *r* lies between the two critical values, conclude that there is not sufficient evidence to support the claim of a linear correlation.

TABLE 2-11 Critical Values of the Linear Correlation Coefficient *r*

Number of Pairs of Data *n*	Critical Value of *r*
4	0.950
5	**0.878**
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
11	0.602
12	0.576

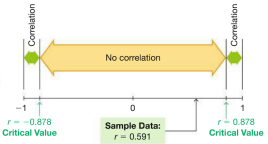


FIGURE 2-17 **Critical Values from Table 2-11 and the Computed Value of *r***

23. P- values for determining correlation

    a. Interpretation

Only a *small* *P*-value, such as 0.05 or less (or a 5% chance or less), suggests that the sample results are *not* likely to occur by chance when there is no linear correlation, so a small *P*-value supports a conclusion that there is a linear correlation between the two variables.

24. Regression

    a. Given a collection of paired sample data, the regression line (or line of best fit, or least-squares line) is the straight line that “best” fits the scatterplot of the data.