GREATER JAKARTA • BANDUNG • MALANG

STAT6171001 Basic Statistics

Graphing Distributions
Session 2

Raymond Bahana
rbahana@binus.ac.id

People
Innovation
Excellence

Session Learning Outcomes

Upon completion of this session, students are expected to be able to

- LO 2. Analyze a problem by using the basic concept of descriptive and inferential statistics
- LO 3. Design a descriptive and inferential statistics solution to meet a given set of computing requirements in the context of computer science
- LO4. Produce descriptive and inferential statistics solutions

Topic

- Graphing distributions

Graphing Distributions

Introduction

- A frequency distribution helps us to detect any pattern in the data (assuming a pattern exists) by superimposing some order on the inevitable variability among observations.

- Graphs of frequency distributions further aid our effort to detect data patterns and make sense out of the data.

Graphing Data

- Graphing data is the first and often most important step in data analysis.

- In this day of computers, researchers all too often see only the results of complex computer analyses without ever taking a close look at the data themselves.

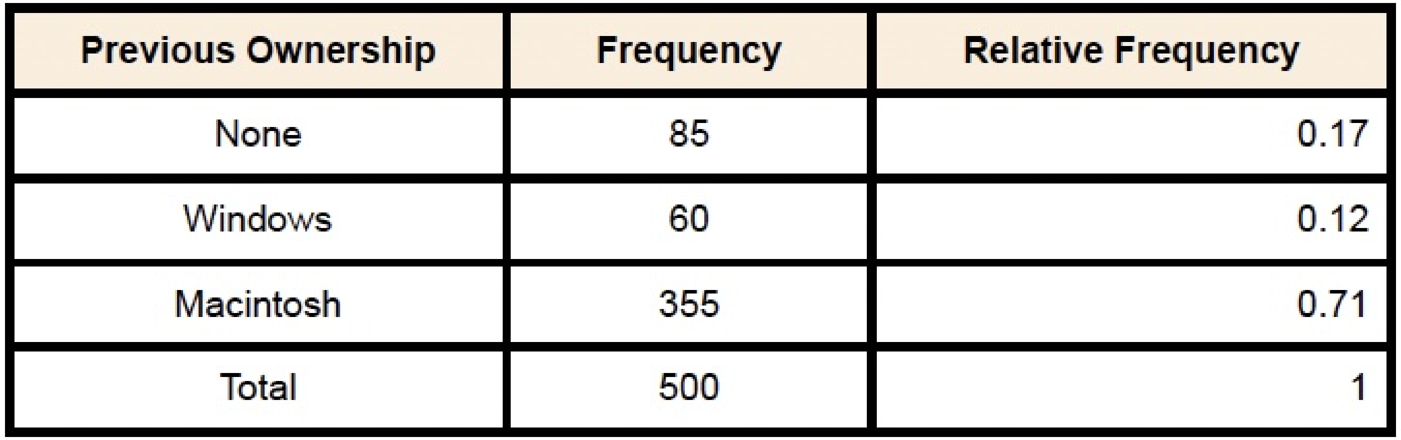
- This is all the more unfortunate because computers can create many types of graphs quickly and easily.

Graphing Qualitative Variables

Graphing Qualitative Variables

- Frequency Tables

  Example: Frequency Table for the iMac Data

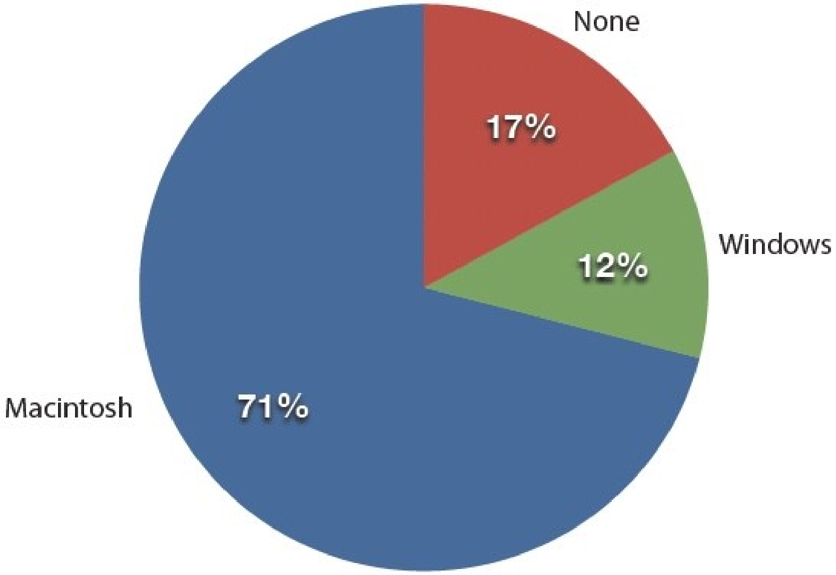


Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1

0.17 = 85/500

Pie Charts

Example: Pie chart of iMac purchases illustrating frequencies of previous computer ownership



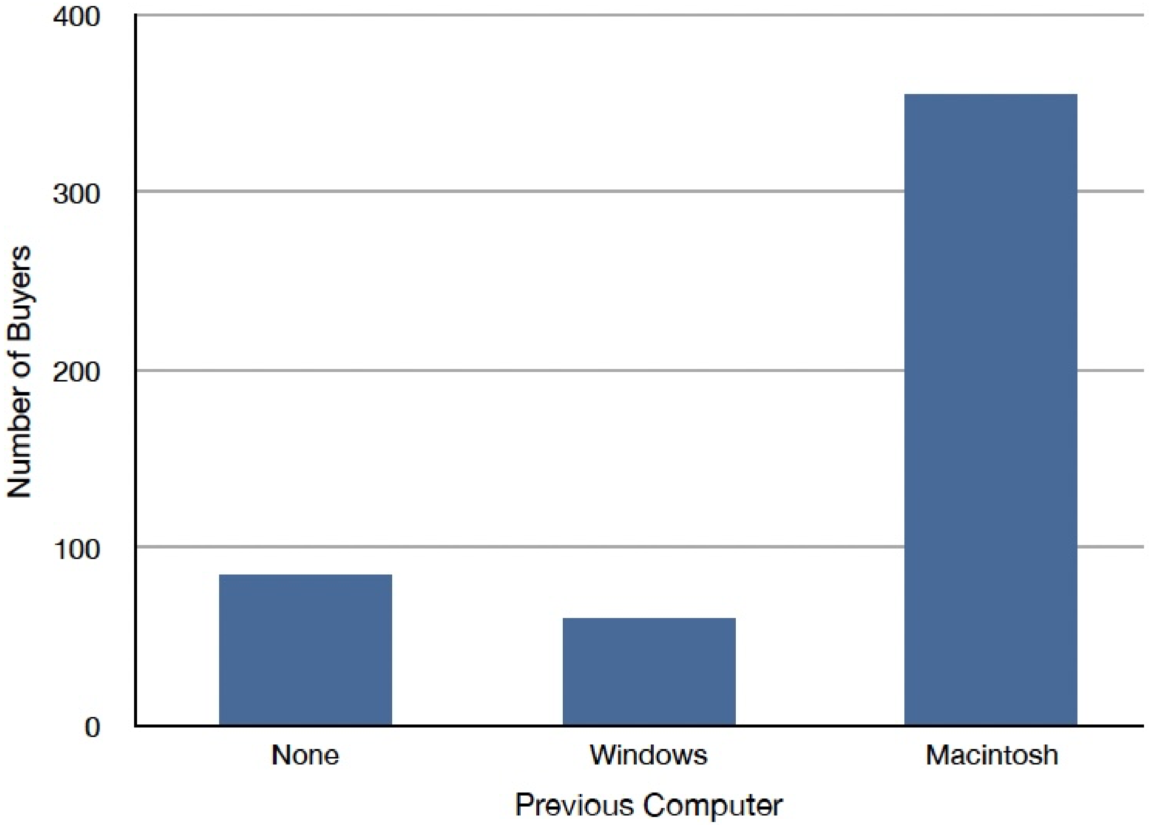
Pie Charts

- Pie charts are effective for displaying the relative frequencies of a small number of categories.
- They are not recommended, however, when you have a large number of categories.
- Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments.
- If the data is small, it’s better to alert the user of the pie chart to the actual numbers involved, labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar Charts

- Bar charts can also be used to represent frequencies of different categories.

Example: Bar chart of iMac purchases as a function of previous computer ownership.



Bar Charts

- 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.

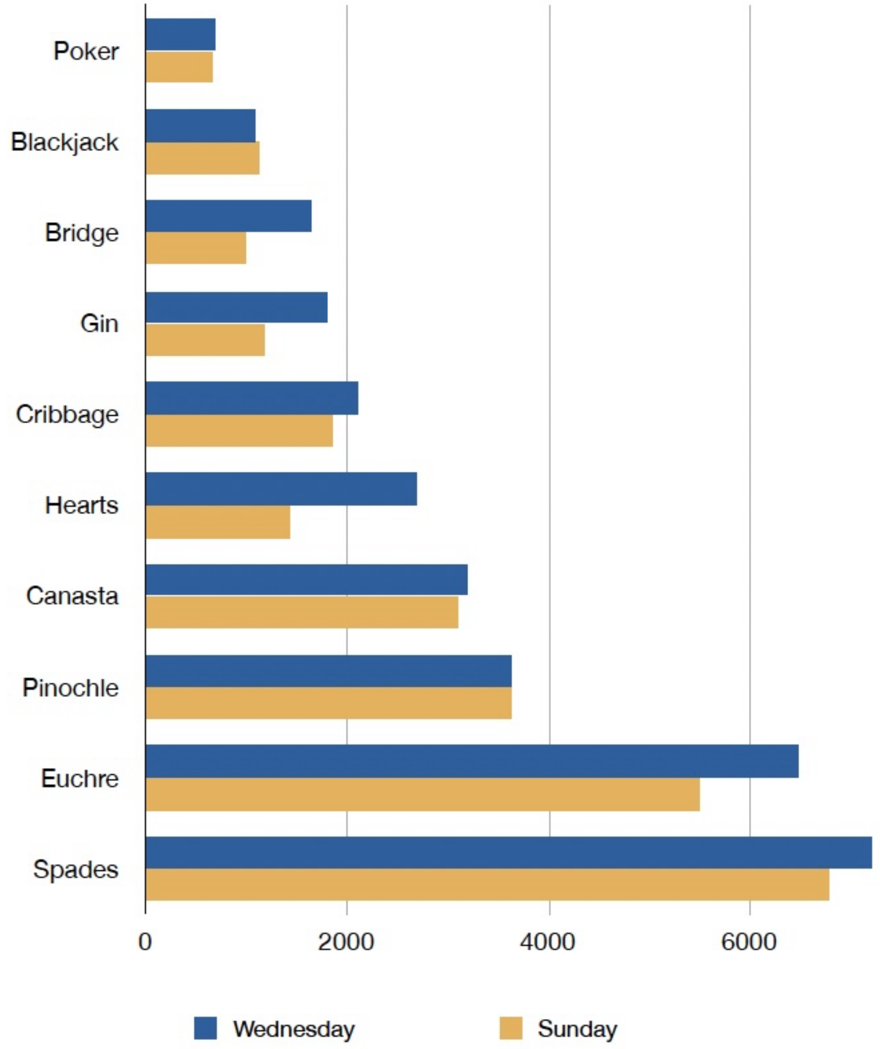
Bar Charts - Comparing Distributions

- Often, we need to compare the results of different surveys, or of different conditions within the same overall survey.
- Bar charts are often excellent for illustrating differences between two distributions
- Next Figure shows the number of people playing card games on a Sunday and on a Wednesday. Overall, more players on Wednesday compared to Sunday.
- The number of people playing Pinochle was nonetheless the same on these two days. Facts like these emerge clearly from a well-designed bar chart.

Bar Charts - Comparing Distributions

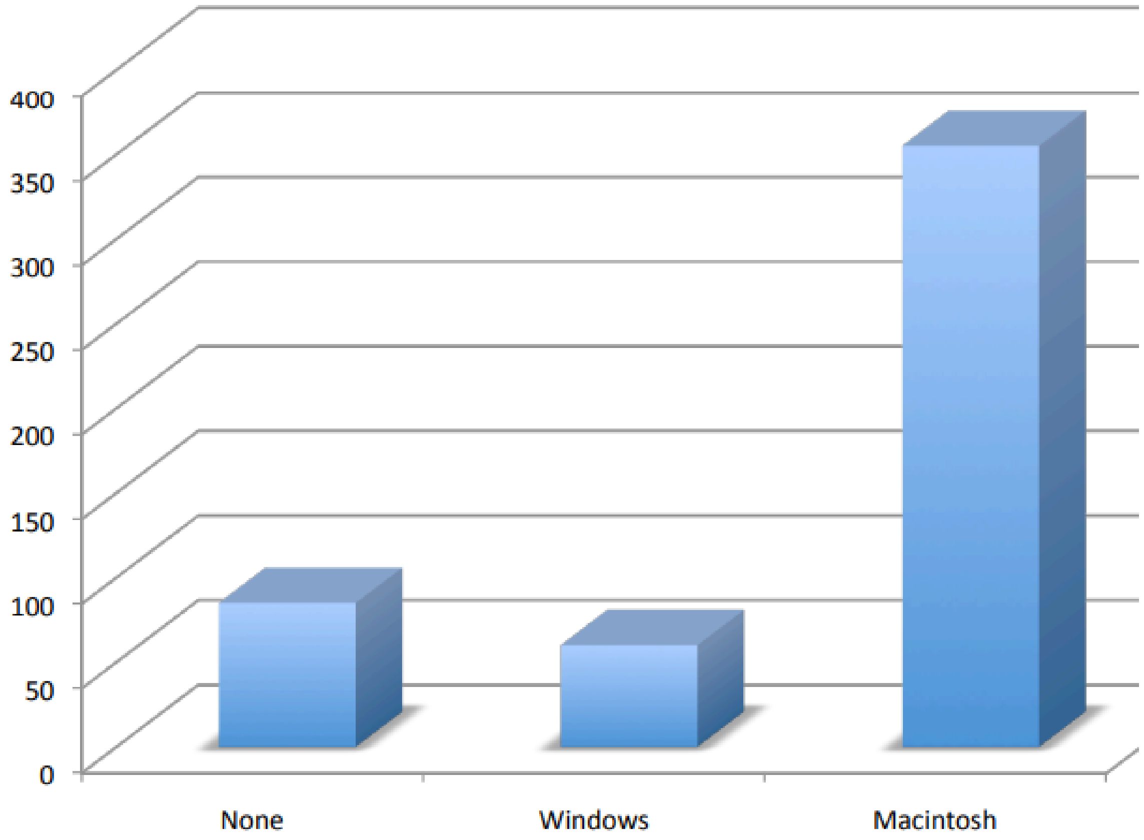
- Example: A bar chart of the number of people playing different card games on Sunday and Wednesday.

- The bars are oriented horizontally rather than vertically.
- The horizontal format is useful when you have many categories because there is more room for the category labels.



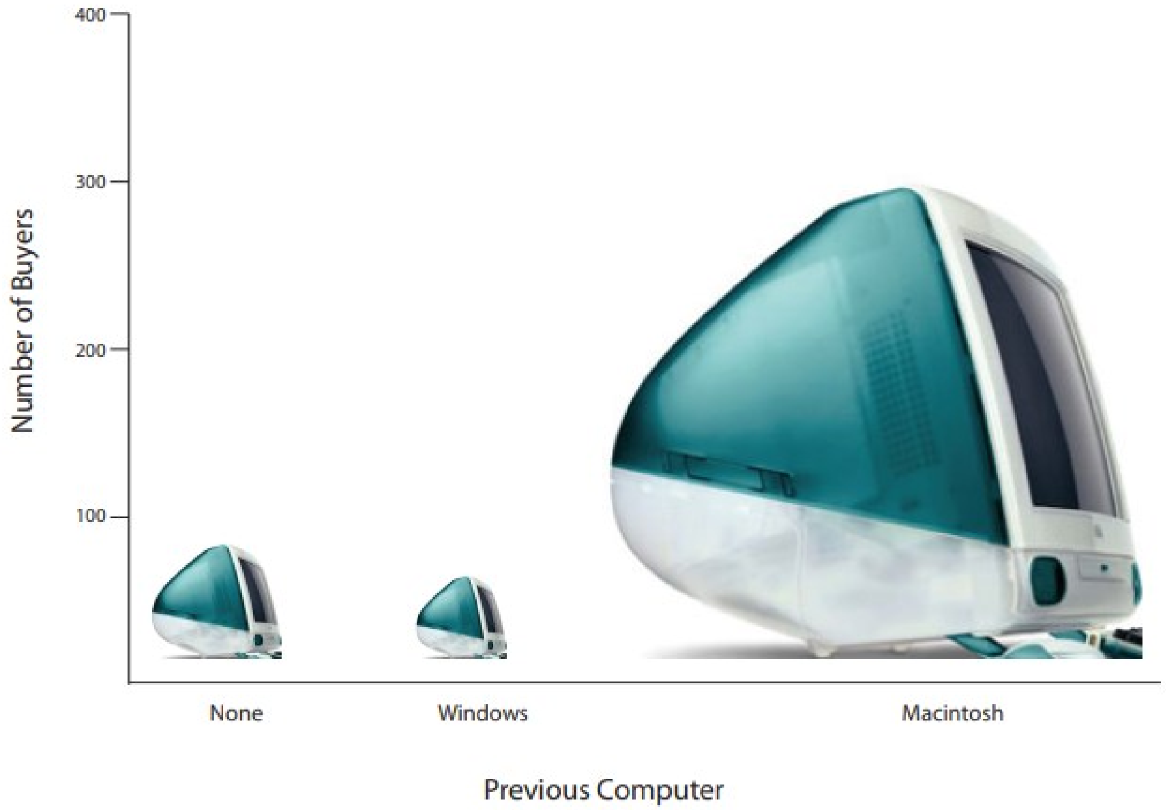
Some Graphical Mistakes to Avoid

- Don’t get fancy!
- People sometimes add features to graphs that don’t help to convey their information.
- For example, 3-dimensional bar charts are usually not as effective as their two-dimensional counterparts.



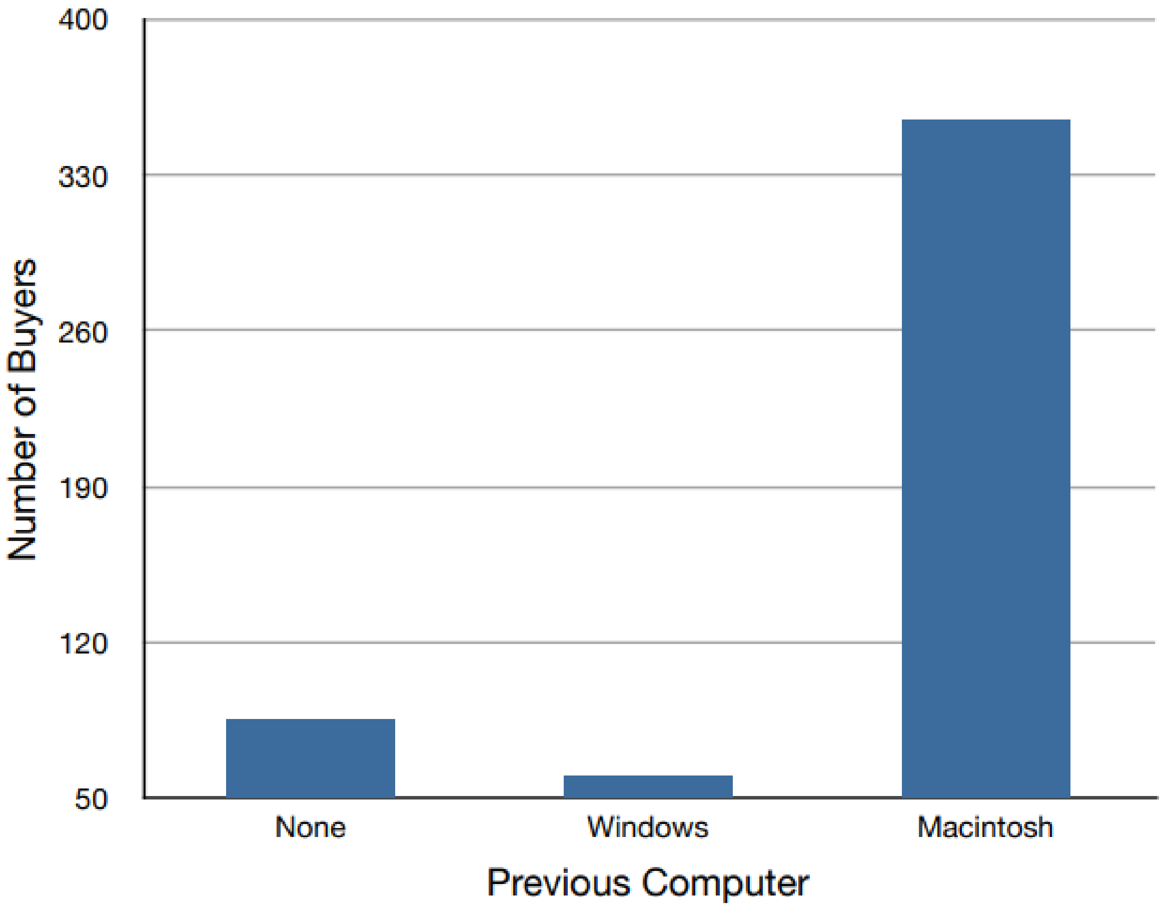
Some Graphical Mistakes to Avoid

- This figure presents the data using pictures of computers.
- The heights of the pictures accurately represent the number of buyers, yet it’s misleading because the viewer's attention will be captured by areas.
- The areas can exaggerate the size differences between the groups. In terms of %, the ratio of previous Mac owners to previous Windows owners is about 6 to 1.
- But the ratio of the two areas in this figure is about 35 to 1.



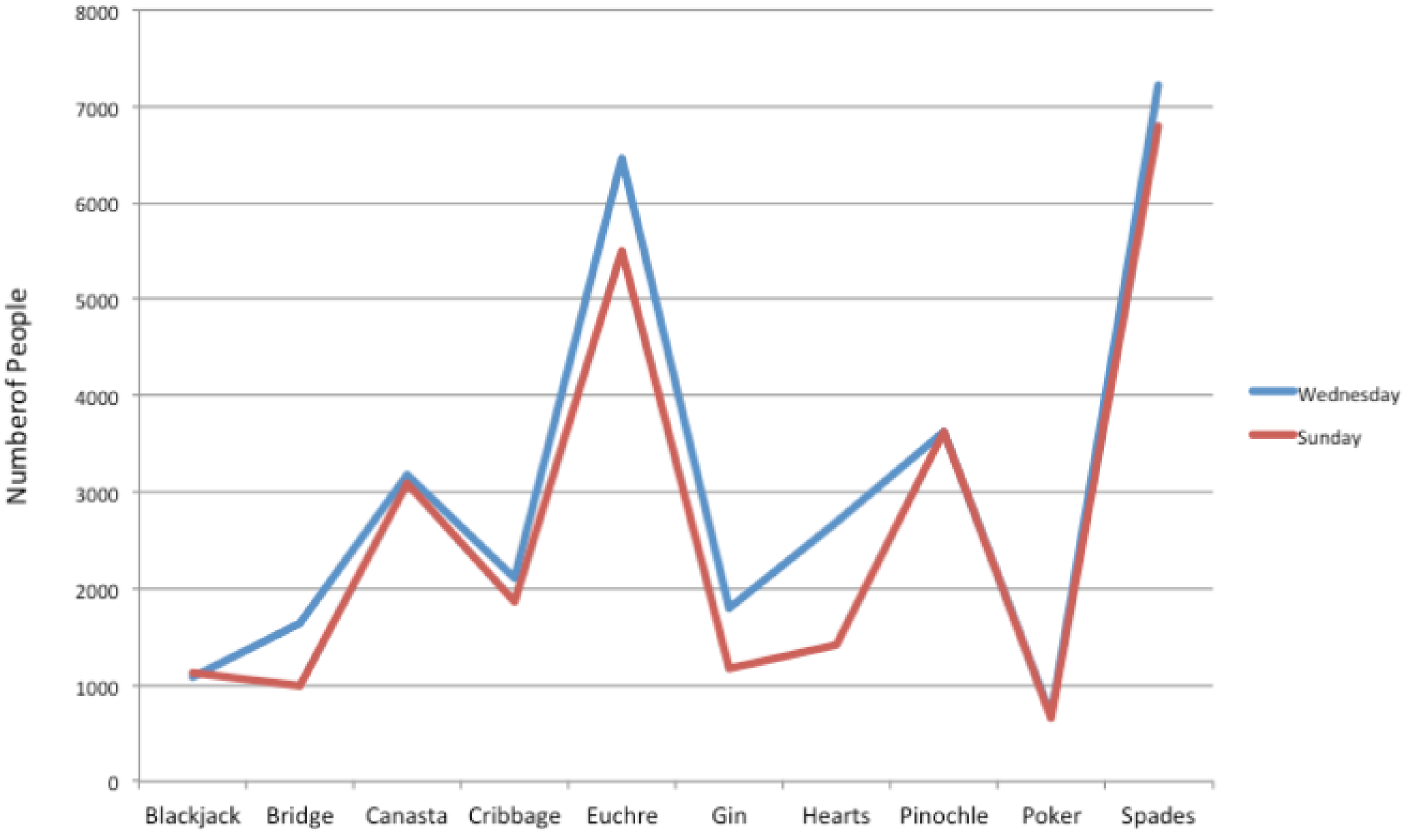
Some Graphical Mistakes to Avoid

- Another distortion in bar charts results from setting the baseline to a value other than zero.
- The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category.



Some Graphical Mistakes to Avoid

- A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines.
- Figure inappropriately shows a line graph of the card game data.
- The drawback is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically



Graphing Quantitative Variables

Graphing Quantitative Variables

- Stem and Leaf Displays
- Histograms
- Frequency Polygons
- Box Plots
- Bar Charts
- Line Graphs
- Dot Plots
- Statistical Literacy

Graphing Quantitative Variables

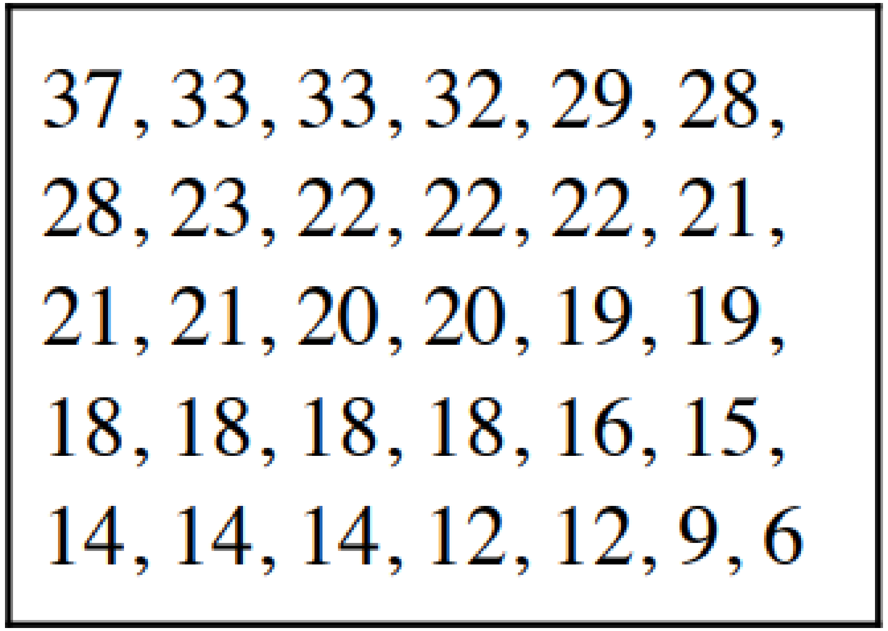
- Quantitative variables are variables measured on a numeric scale.
- Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables.
- Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, favorite sport in which there is no ordering or measuring involved.

Graphing Quantitative Variables

- Stem and leaf displays are best-suited for small to moderate amounts of data, whereas others such as histograms are best suited for large amounts of data.
- Graph types such as box plots are good at depicting differences between distributions.
- Scatter plots are used to show the relationship between two variables.

Stem and Leaf Displays

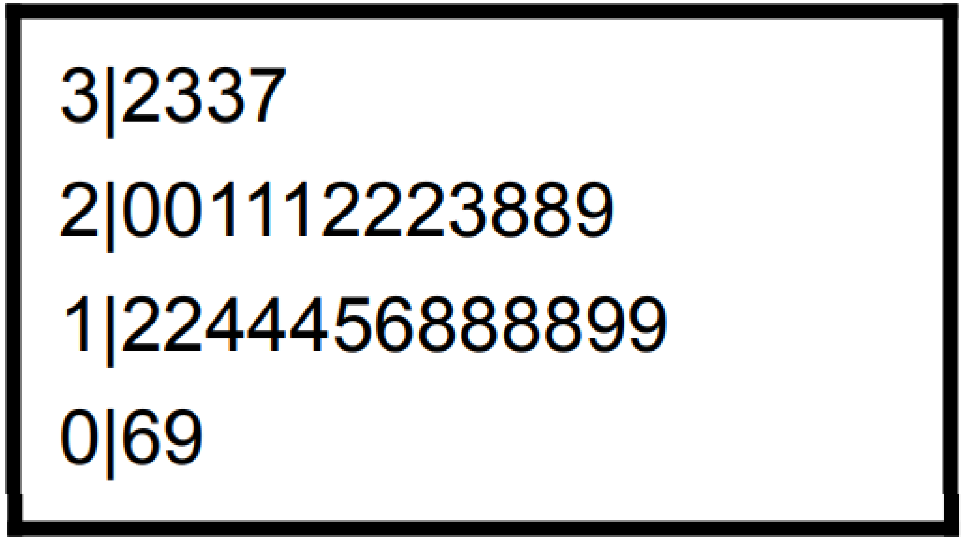
- A stem and leaf display is a graphical method of displaying data.
- It is particularly useful when your data are not too numerous.
- Consider the Table that shows the number of touchdown passes thrown by each of the 31 teams in the NFL in the 2000 season



37, 33, 33, 32, 29, 28,
28, 23, 22, 22, 22, 21,
21, 21, 20, 20, 19, 19,
18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6

Stem and Leaf Displays

- A stem and leaf display of the data is shown in next figure .
- The left portion of figure contains the stems.
- They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10’s digits
- The numbers to the right of the bar are leaves, and they represent the 1’s digits.
- Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.



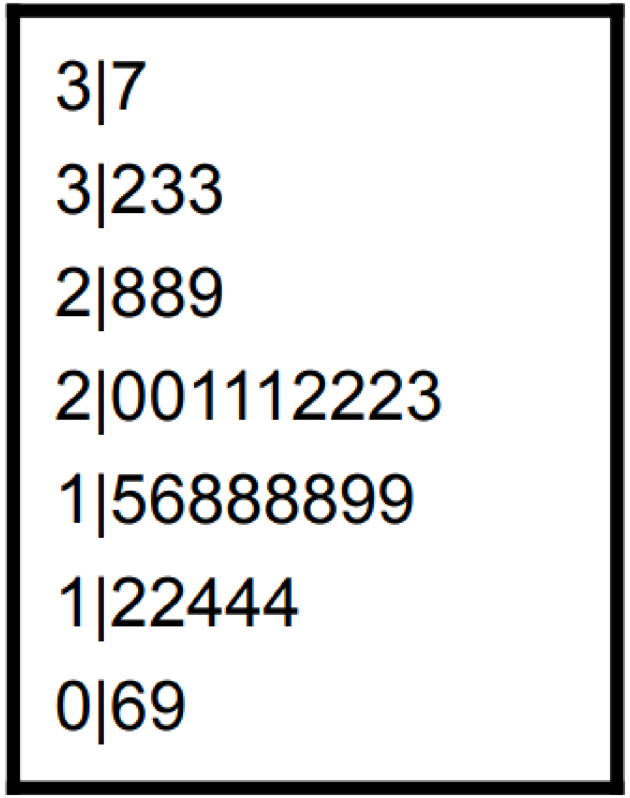
```
3|2337
2|001112223889
1|2244456888899
0|69
```

Stem and Leaf Displays

- One purpose of a stem and leaf display is to clarify the shape of the distribution.
- You can see many facts about TD passes more easily in Figure than in Table.
- For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TD's, with a few having more and a few having less.
- The precise numbers of TD passes can be determined by examining the leaves

Stem and Leaf Displays

- We can make our figure even more revealing by splitting each stem into two parts
- The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table.
- The second row is reserved for the numbers from 30 to 34 and holds the 32, 33, and 33 TD passes made by the next three teams in the table.



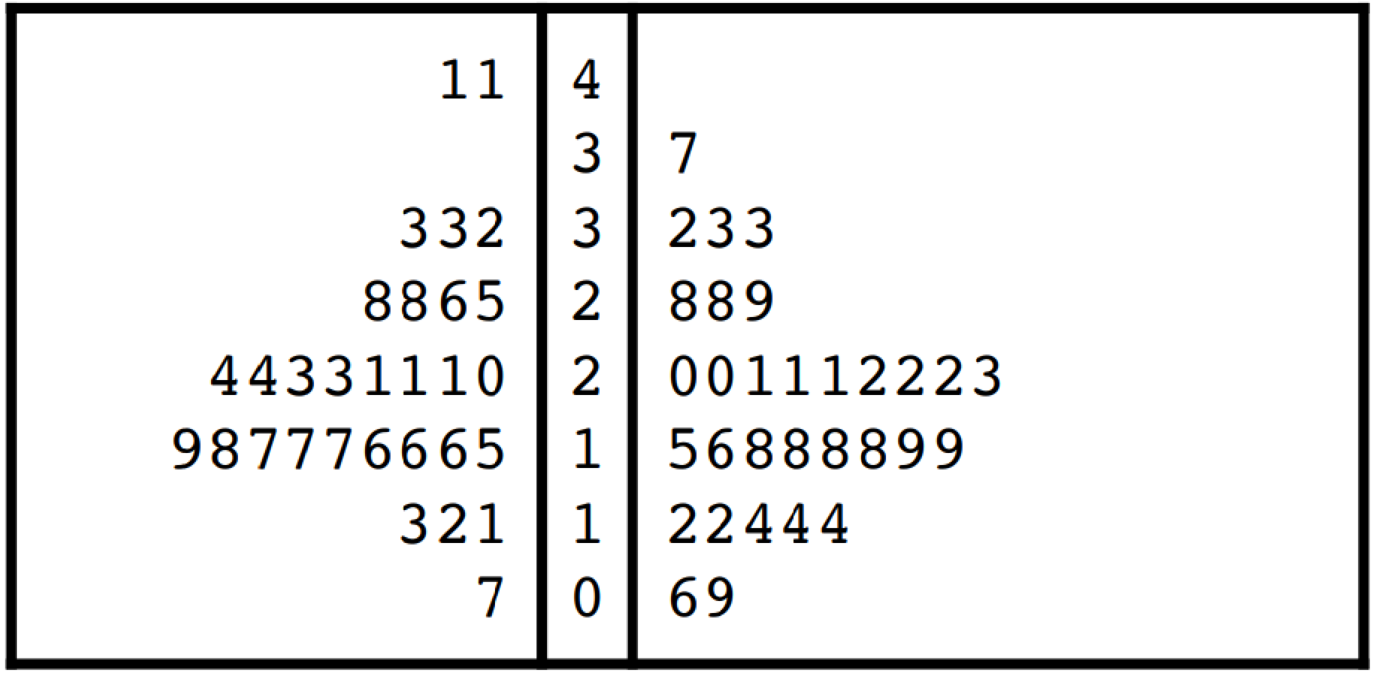
```
3|7
3|233
2|889
2|001112223
1|56888899
1|22444
0|69
```

Stem and Leaf Displays

- This figure is more revealing than previous figure because the latter figure lumps too many values into a single row.
- Whether you should split stems in a display depends on the exact form of your data.
- If rows get too long with single stems, you might try splitting them into two or more parts

Stem and Leaf Displays

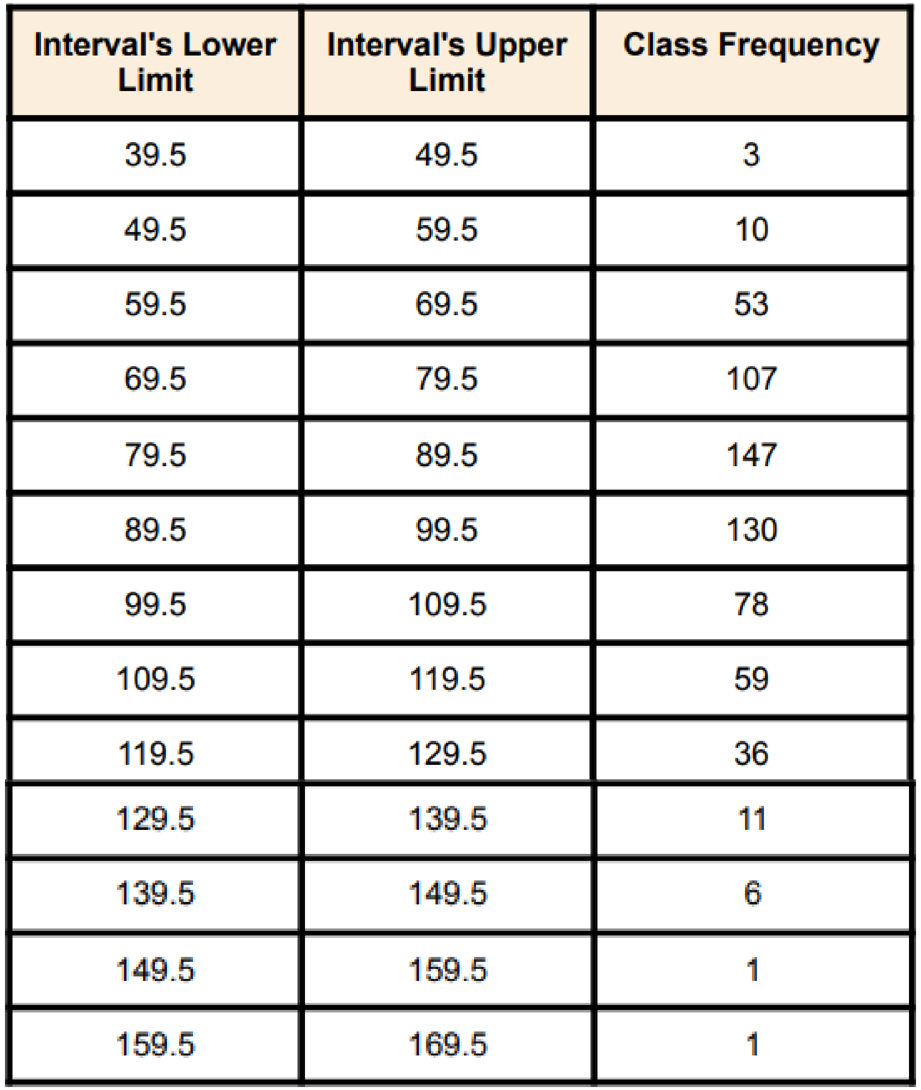
- This figure compares the numbers of TD passes in the 1998 and 2000 seasons.
- The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data.



11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69

Histograms

- A histogram is a graphical method for displaying the shape of a distribution.
- It is particularly useful when there are a large number of observations.
- An example consisting of the scores of 642 students on a psychology test.
- The test consists of 197 items each graded as “correct” or “incorrect.”
- The students' scores ranged from 46 to 167



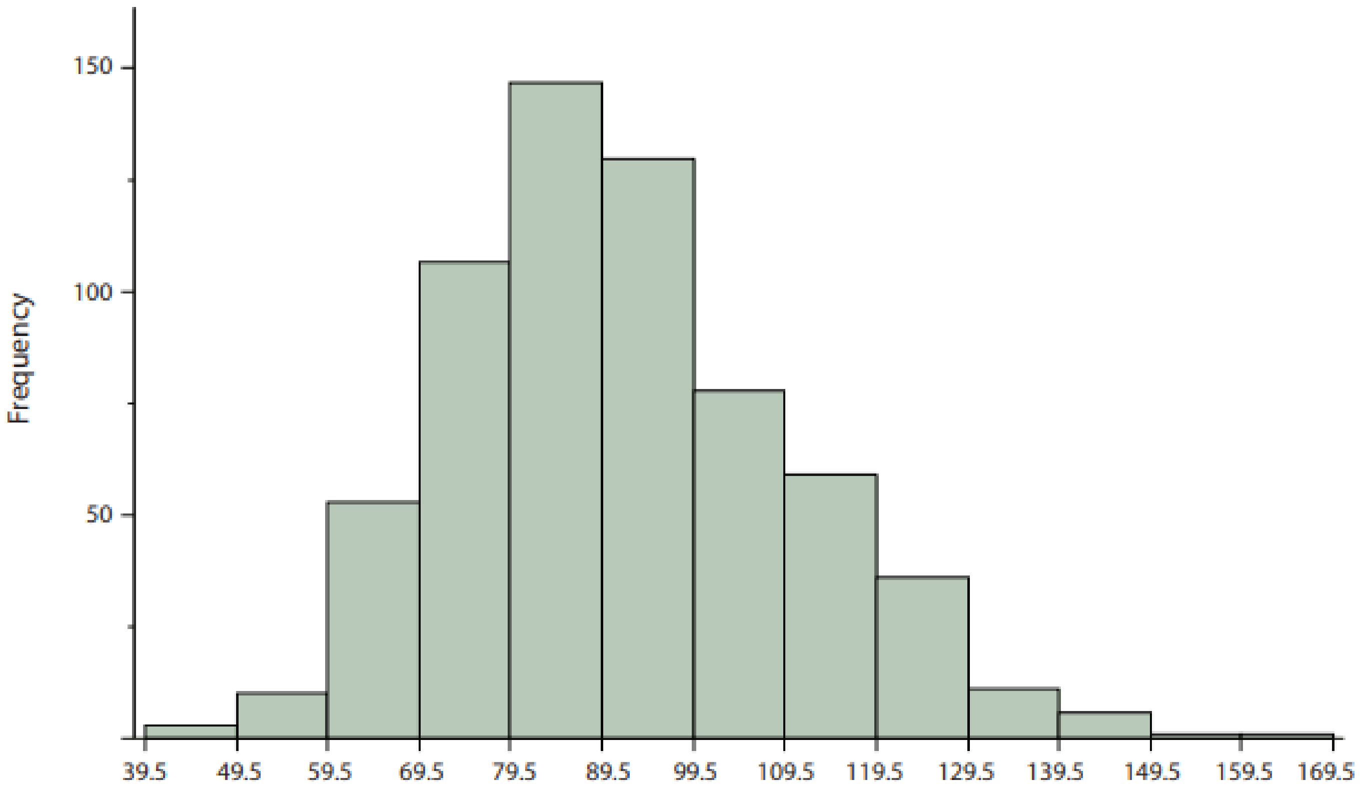
Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

Histograms

- The first step is to create a frequency table.
- Unfortunately, a simple frequency table would be too big, containing over 100 rows.
- To create this table, the range of scores was broken into intervals, called class intervals.
- The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc.
- Next, the number of scores falling into each interval was counted to obtain the class frequencies.
- There are three scores in the first interval, 10 in the second, etc.

Histograms

- In a histogram, the class frequencies are represented by bars.
- The height of each bar corresponds to its class frequency.

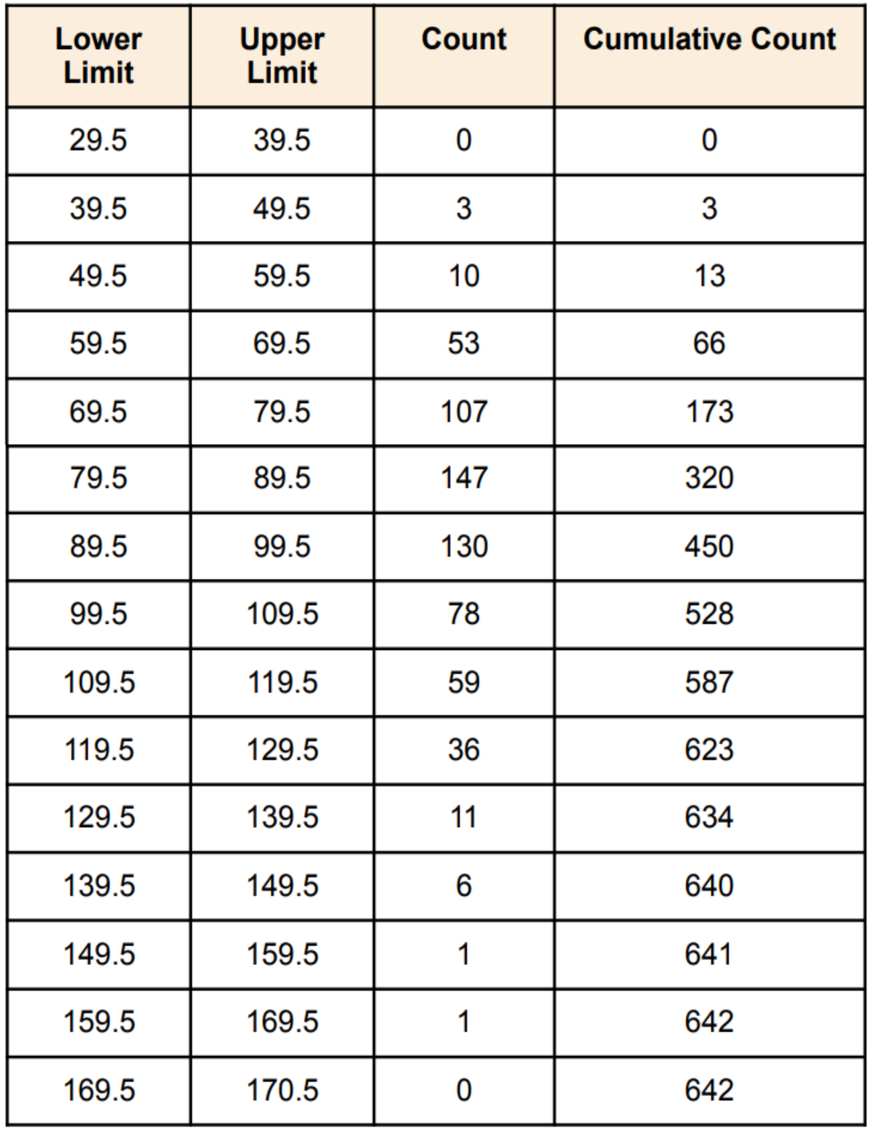


Histograms

- The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes.
- You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left.
- The distribution is therefore said to be skewed.

Frequency Polygons

- Frequency polygons are a graphical device for understanding the shapes of distributions.
- They serve the same purpose as histograms but are especially helpful for comparing sets of data.
- Frequency polygons are also a good choice for displaying cumulative frequency distributions.



Lower Limit	Upper Limit	Count	Cumulative Count
29.5	39.5	0	0
39.5	49.5	3	3
49.5	59.5	10	13
59.5	69.5	53	66
69.5	79.5	107	173
79.5	89.5	147	320
89.5	99.5	130	450
99.5	109.5	78	528
109.5	119.5	59	587
119.5	129.5	36	623
129.5	139.5	11	634
139.5	149.5	6	640
149.5	159.5	1	641
159.5	169.5	1	642
169.5	170.5	0	642

Frequency Polygons

- To create a frequency polygon, start by choosing a class interval.
- Then draw an X-axis representing the values of the scores.
- Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class.
- Draw the Y-axis to indicate the frequency of each class.
- Place a point in the middle of each class interval at the height- corresponding to its frequency.
- Finally, connect the points

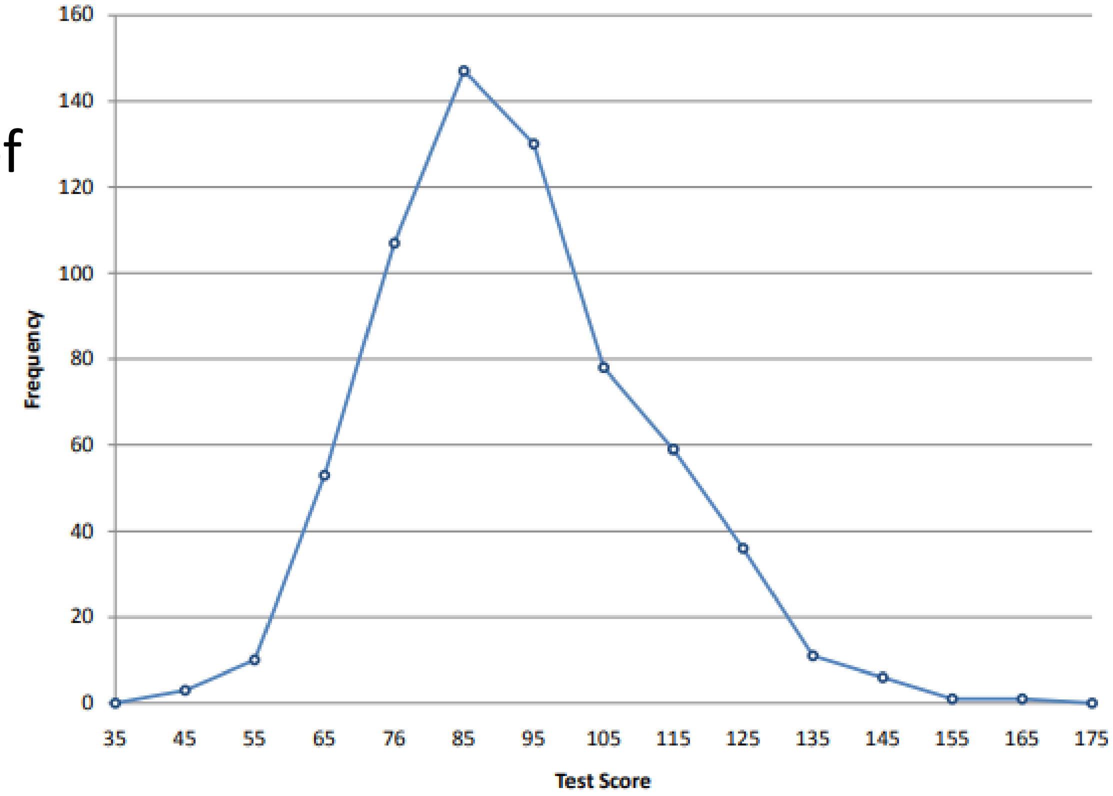


Figure 1. Frequency polygon for the psychology test scores.

Frequency Polygons

- A cumulative frequency polygon for the same test scores is shown in Figure 2.
- The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals.

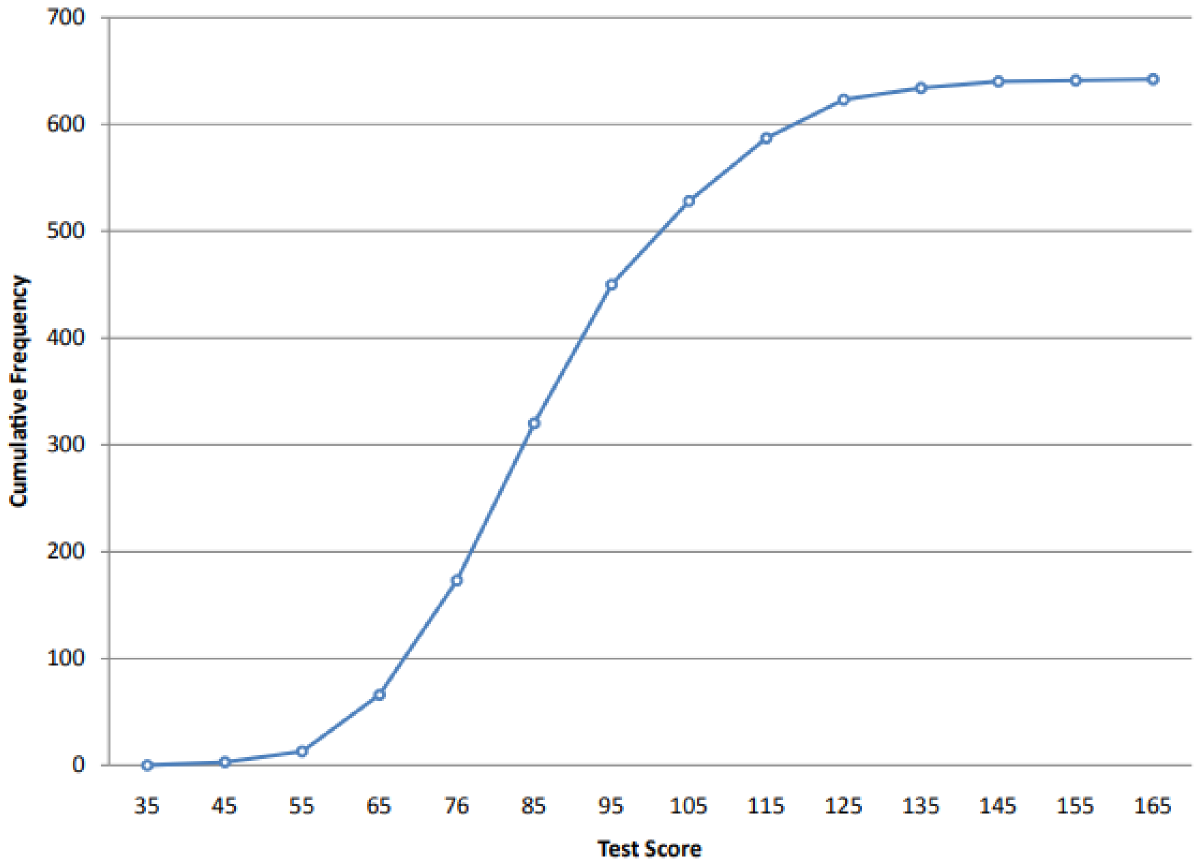
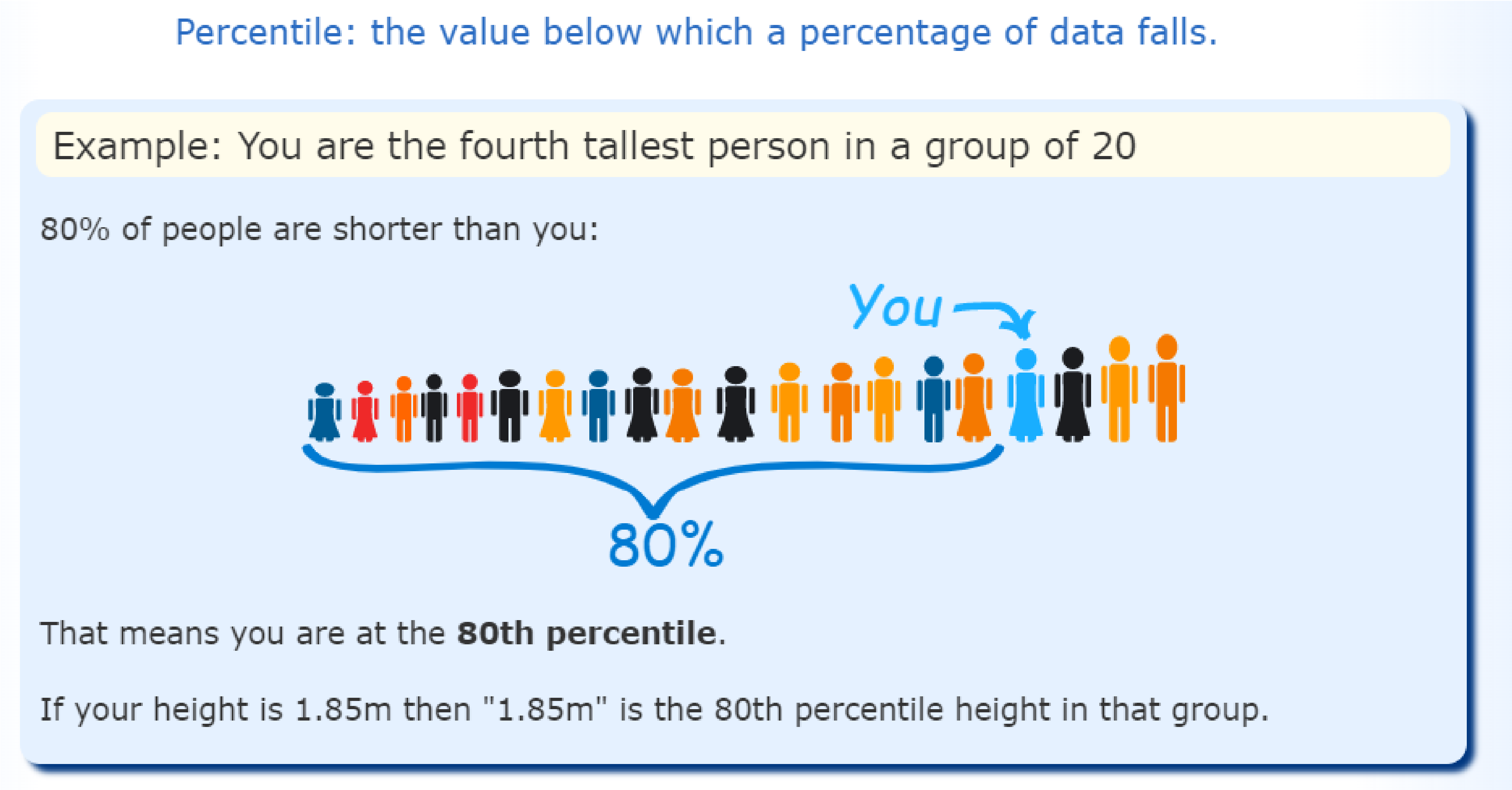


Figure 2. Cumulative frequency polygon for the psychology test scores.

Box Plots

- Box plots are useful for identifying outliers and for comparing distributions.
- Example: students in Introductory Statistics were presented with a page containing 30 colored rectangles.
- Their task was to name the colors as quickly as possible.
- Their times (in seconds) were recorded.
- We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender.
- Such a display is said to involve parallel box plots.

Percentile



Percentile: the value below which a percentage of data falls.

Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:

That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Box Plots

- There are several steps in constructing a box plot.
- The first relies on the **25th**, **50th**, and **75th percentiles** in the distribution of scores.
- Figure 1 shows how these three statistics are used.
- For each gender we draw a box extending from the 25th percentile to the 75th percentile.
- The 50th percentile is drawn inside the box.
- Therefore, the bottom of each box is the 25th percentile, the top is the 75th percentile, and the line in the middle is the 50th percentile.

Box Plots

- For these data, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20.
- For the men, the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5

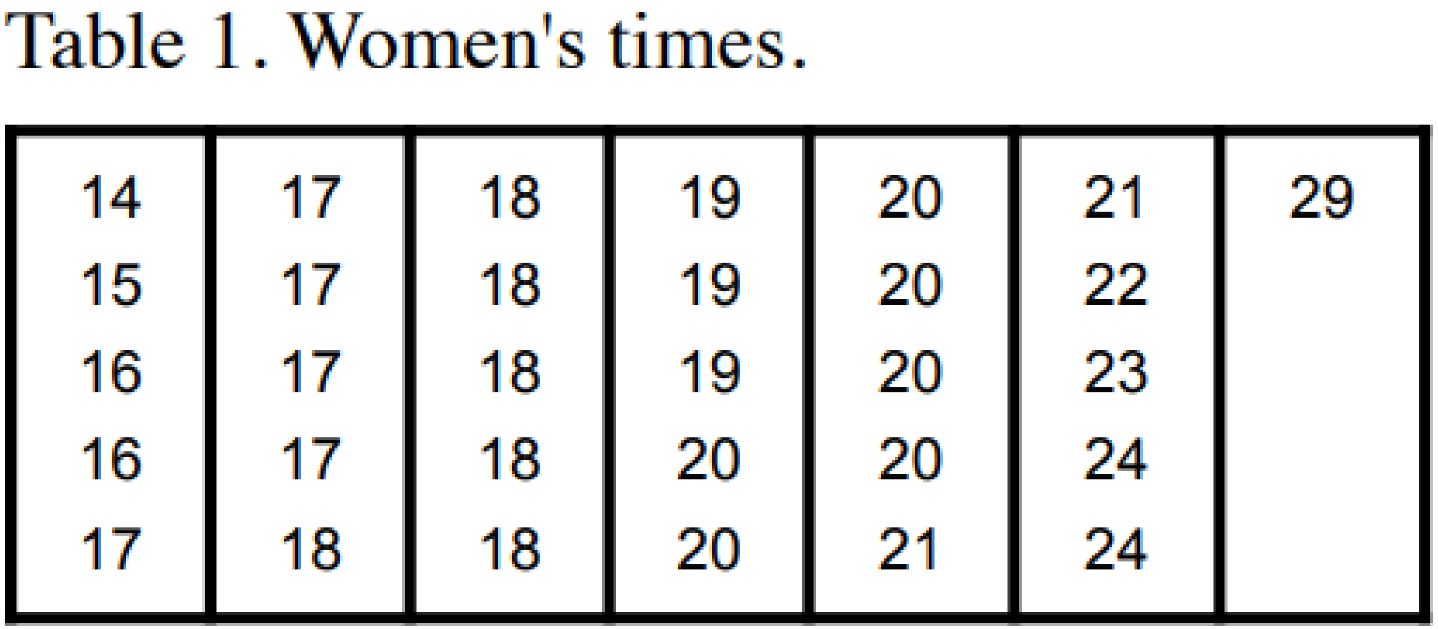


Table 1. Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

Box Plots

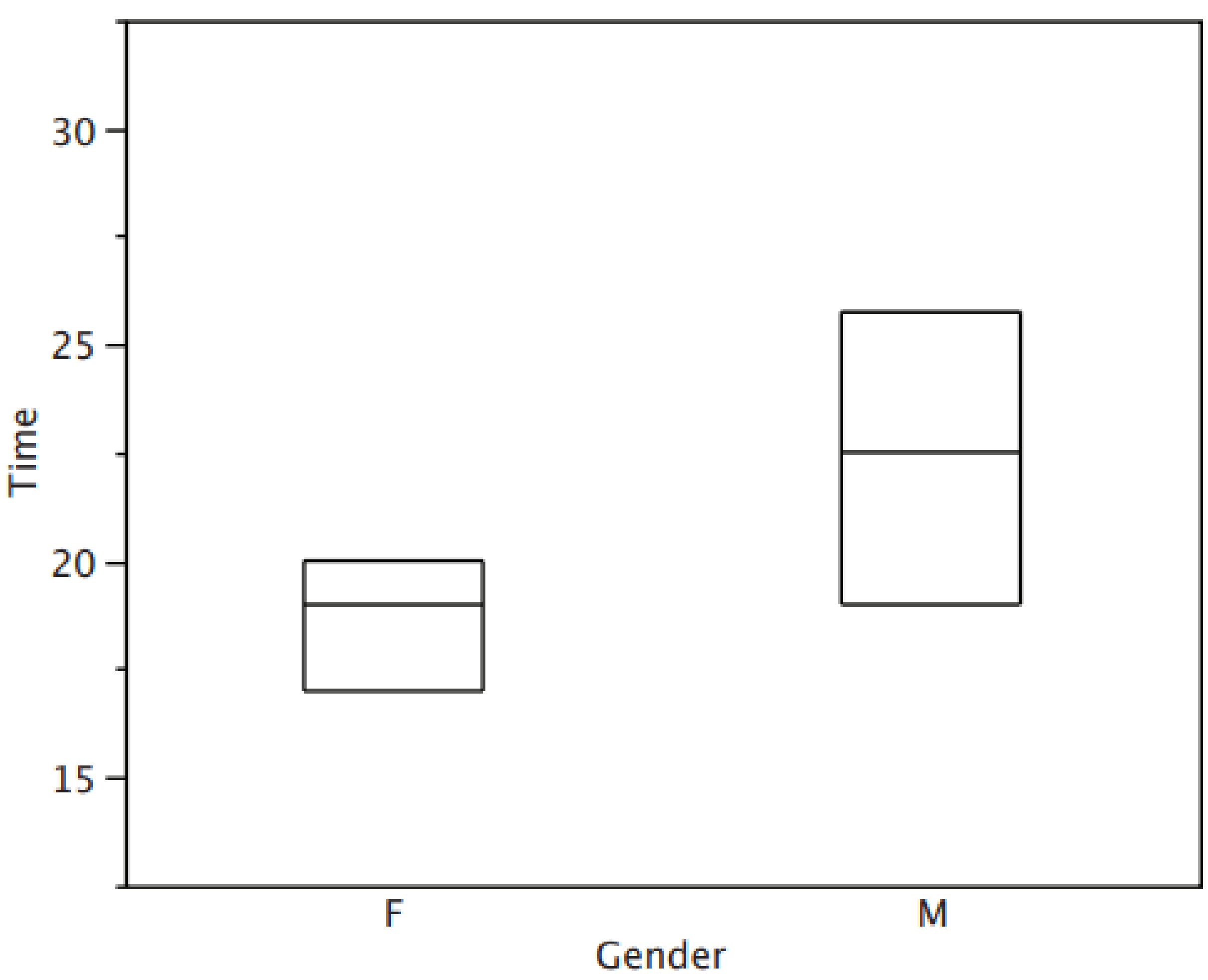
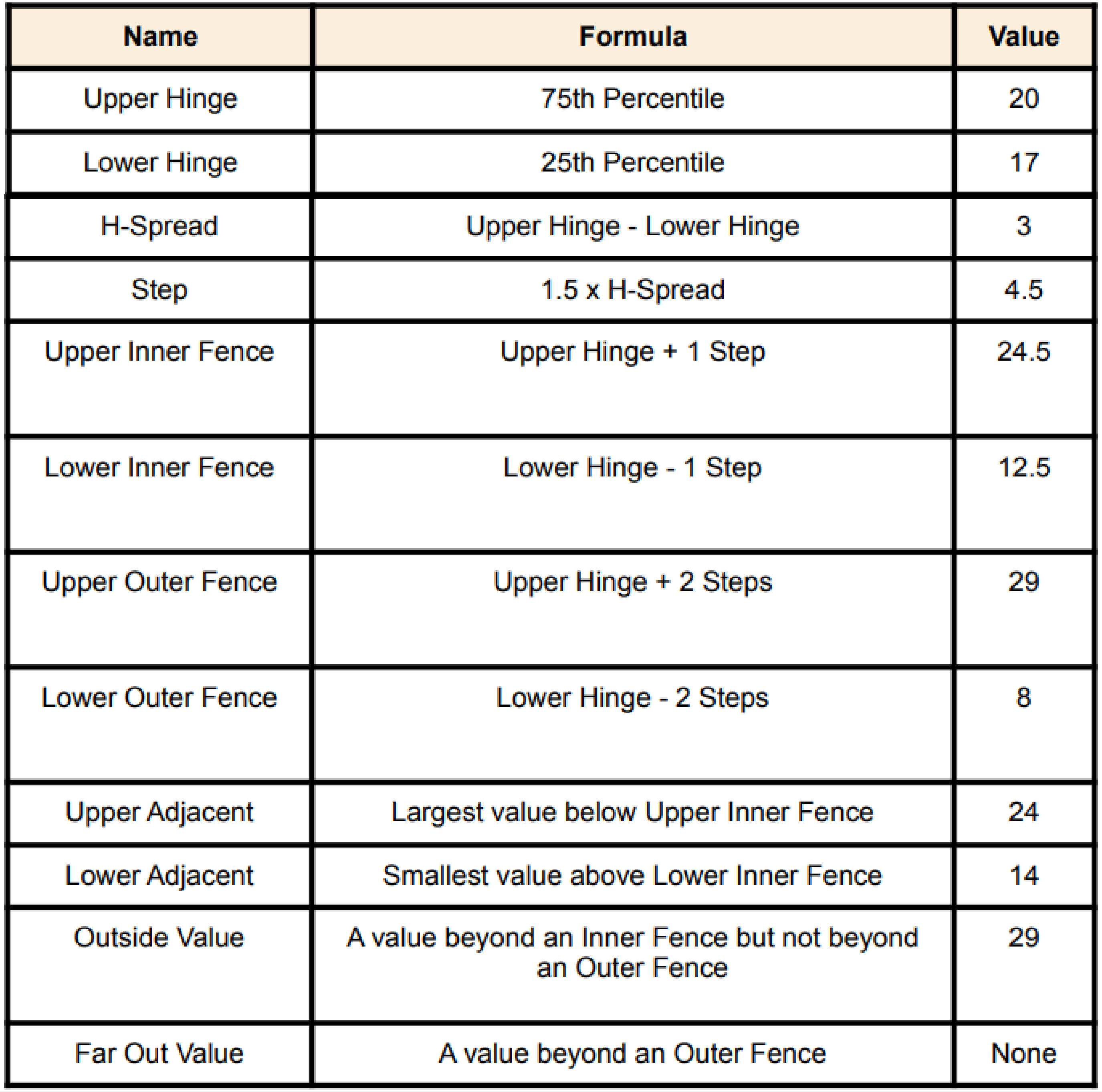


Figure 1. The first step in creating box plots.



Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17
H-Spread	Upper Hinge - Lower Hinge	3
Step	1.5 x H-Spread	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Box Plots

- The H-spread range is a measure of variability, spread or dispersion. It is the difference between the 75th percentile (often called Q3) and the 25th percentile (Q1) → Q3-Q1.
- Inner fences are values of separating data that are a predictable part of the distribution from data that are outside the distribution. Inner fences are located beyond each hinge at 1½ times the H-spread, a distance called a step.
- Outer Fences, data beyond these values are far outside the distribution. They are step beyond the inner fences.

Box Plots

- Put "whiskers" above and below each box to give additional information about the spread of data.
- Whiskers are vertical lines that end in a horizontal stroke.
- Whiskers are drawn from the upper and lower hinges (20 and 17) to the upper and lower adjacent values (24 and 14)

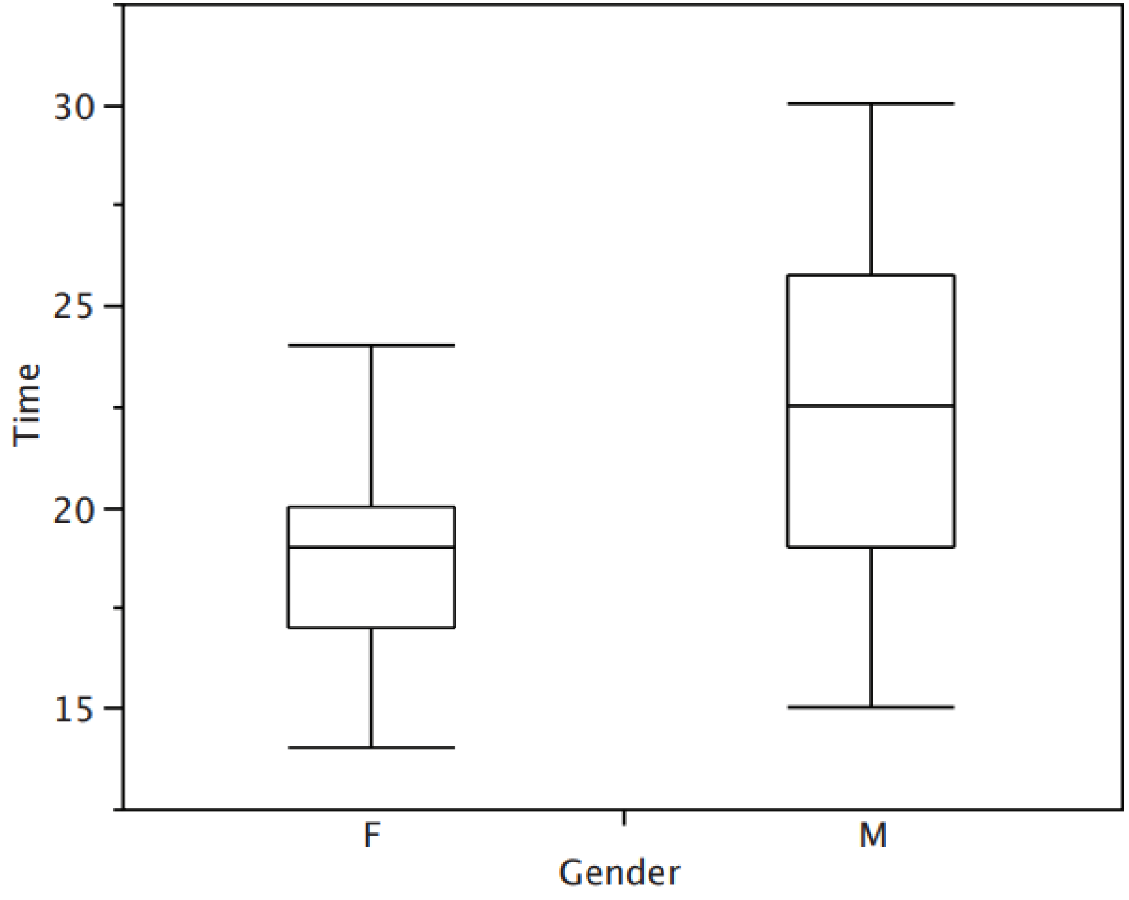


Figure 2. The box plots with the whiskers drawn.

Box Plots

- Outside values are indicated by small “o's” and far out values are indicated by asterisks (*).
- In data, there are no far out values and just one outside value.
- This outside value of 29 is for the women.

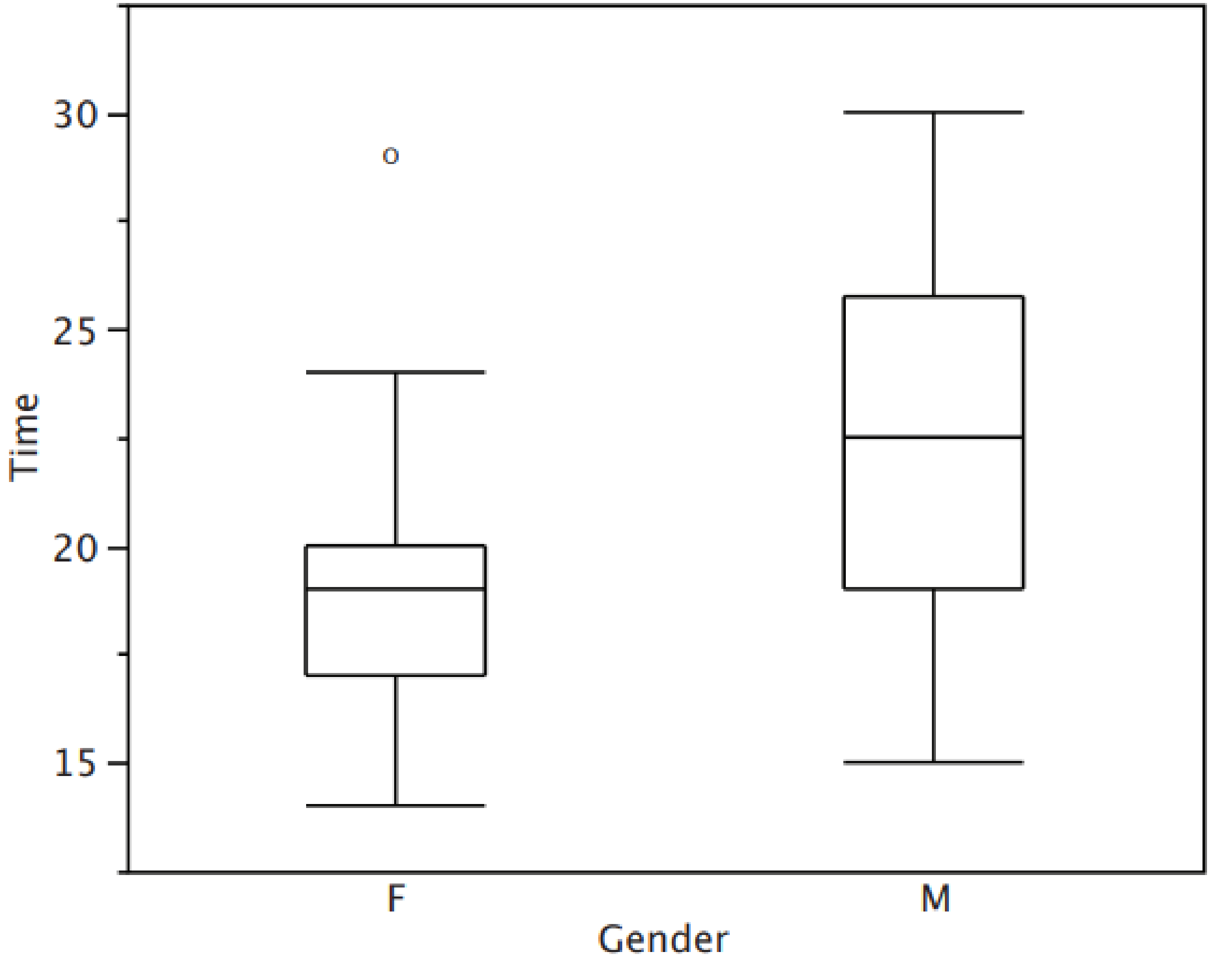


Figure 3. The box plots with the outside value shown.

Box Plots

- There is one more mark to include in box plots (although sometimes it is omitted).
- We indicate the mean score for a group by inserting a plus sign.
- Figure 4 shows the result of adding means to our box plots.
- Mean refers to the average of a set of values.

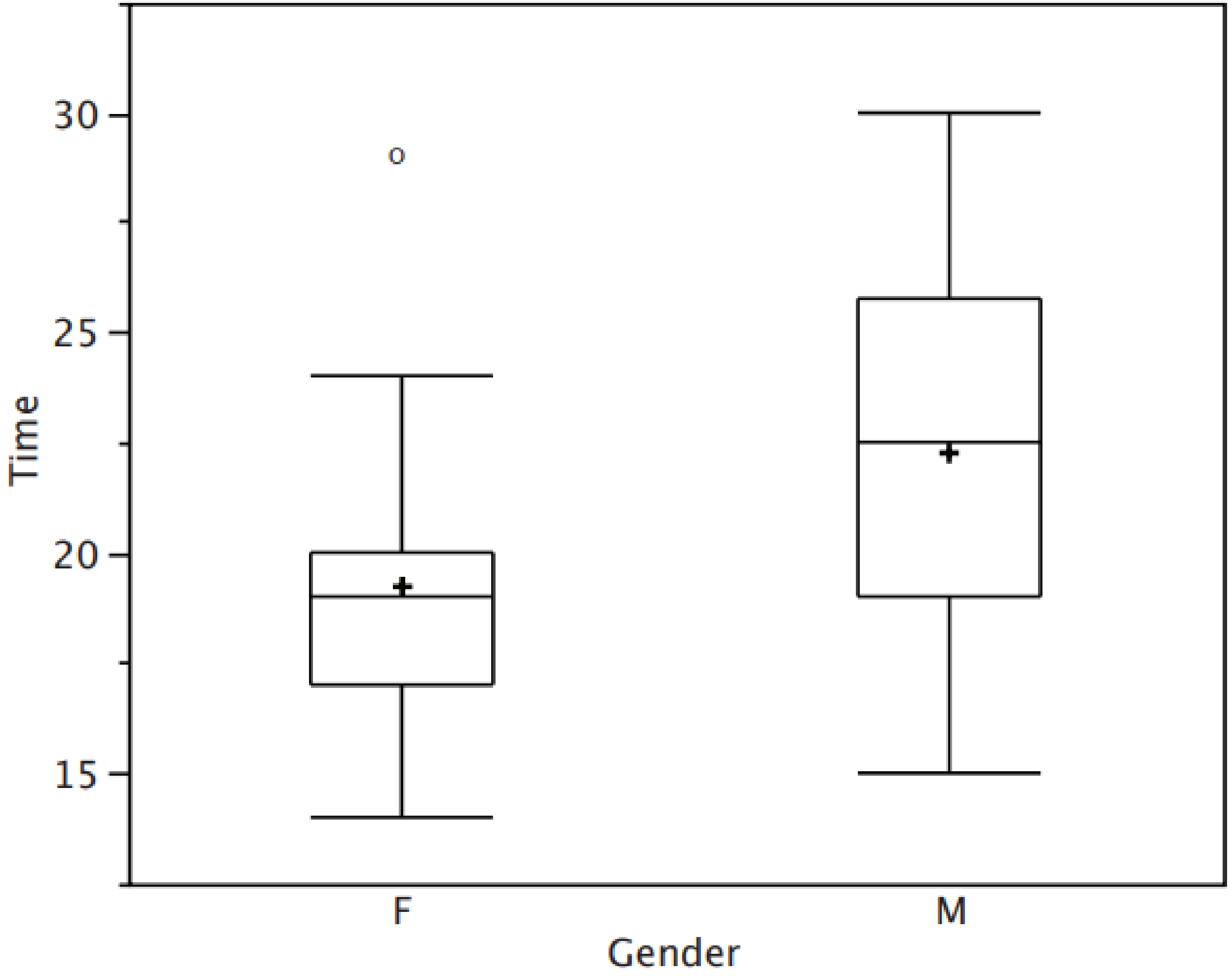


Figure 4. The completed box plots.

Box Plots

- Figure 4 provides a revealing summary of the data.
- Since half the scores in a distribution are between the hinges, we see that half the women's times are between 17 and 20 seconds whereas half the men's times are between 19 and 25.5 seconds.
- We also see that women generally named the colors faster than the men did, although one woman was slower than almost all the men.

Box Plots

- Figure 5 shows the box plot for the women’s data with detailed labels

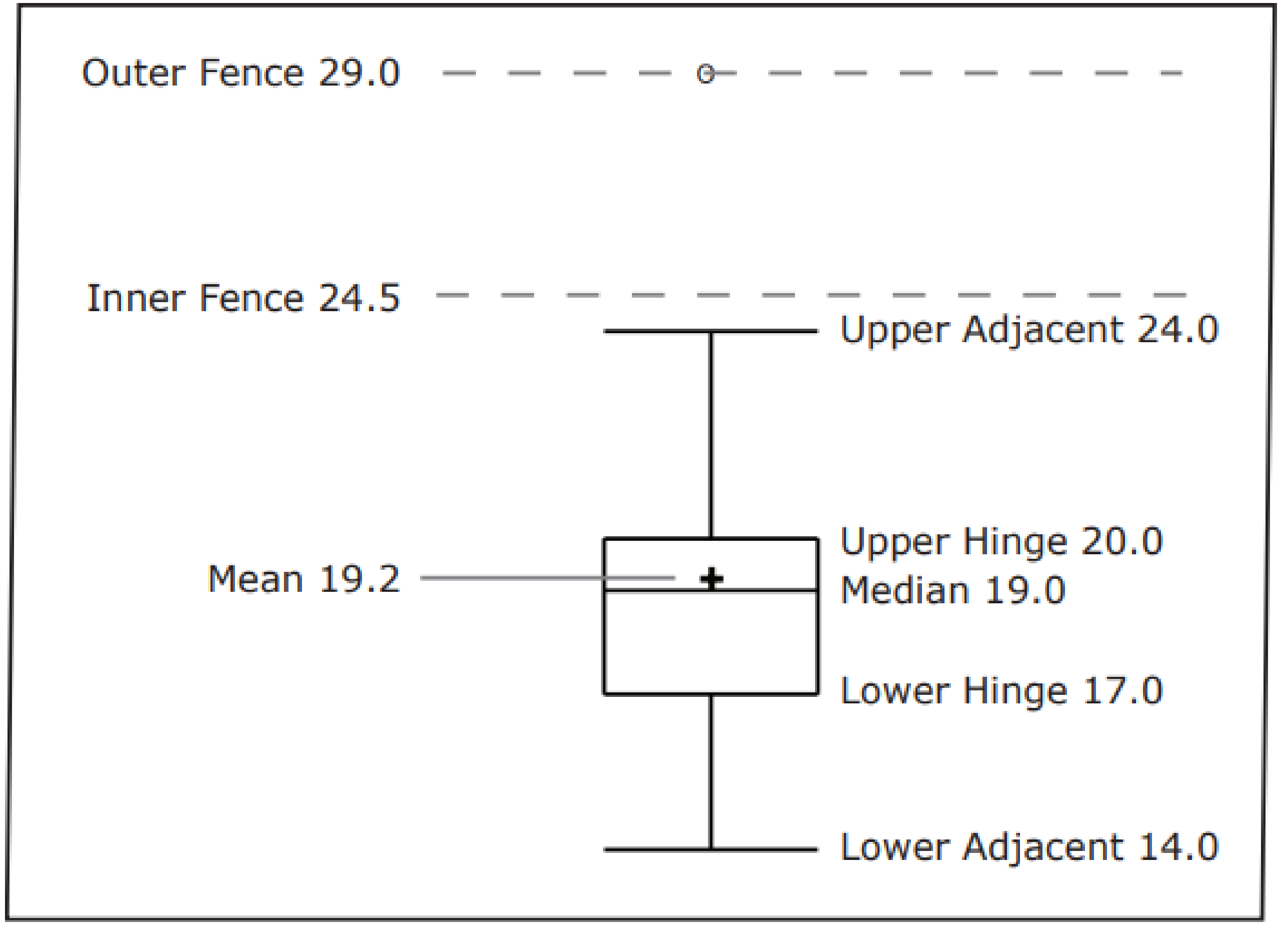


Figure 5. The box plots for the women's data with detailed labels.

Box Plots

- Box plots provide basic information about a distribution.
- For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction.
- A larger mean than median would also indicate a positive skew.
- Box plots are good at portraying extreme values and are especially good at showing differences between distributions.
- However, many of the details of a distribution are not revealed in a box plot and to examine these details one should use create a histogram and/or a stem and leaf display.

Bar Charts

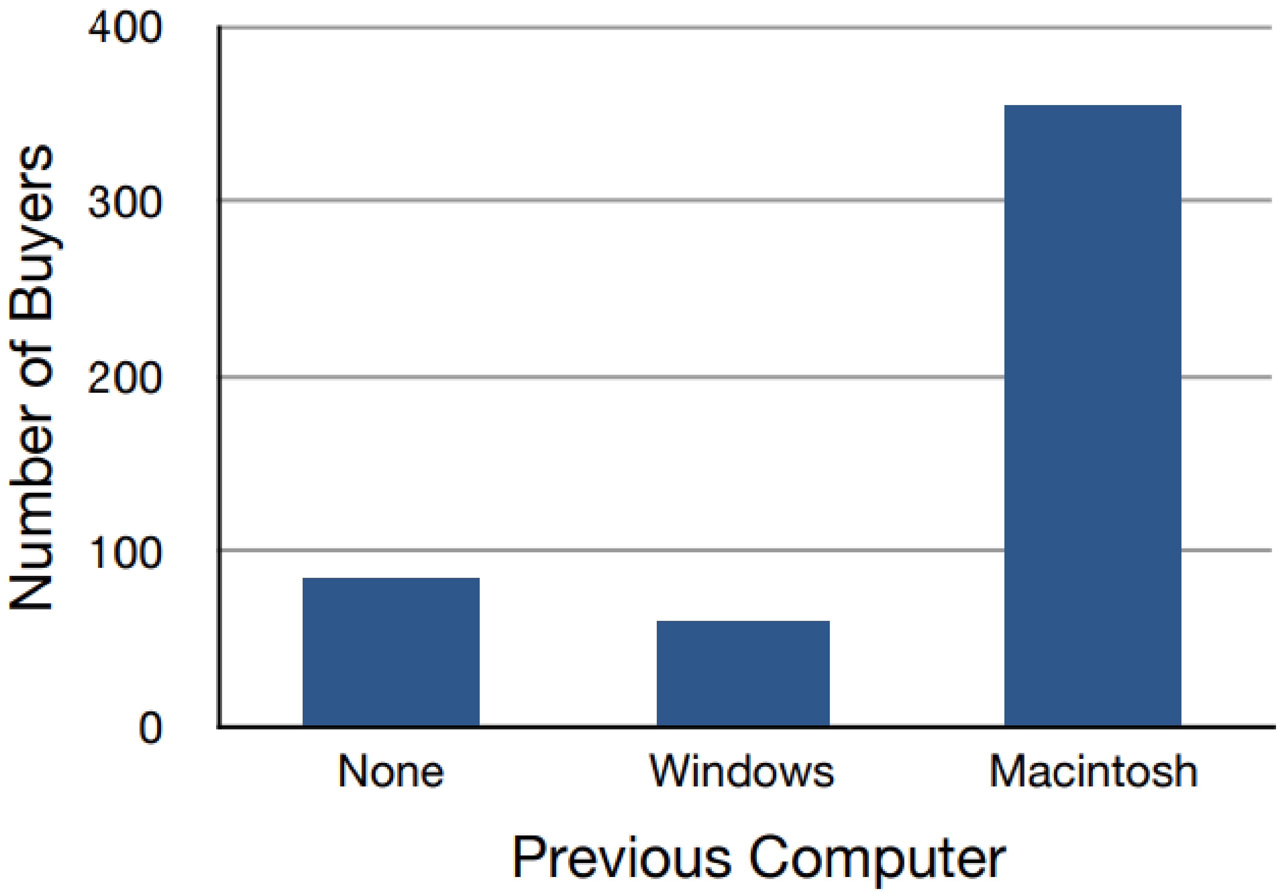
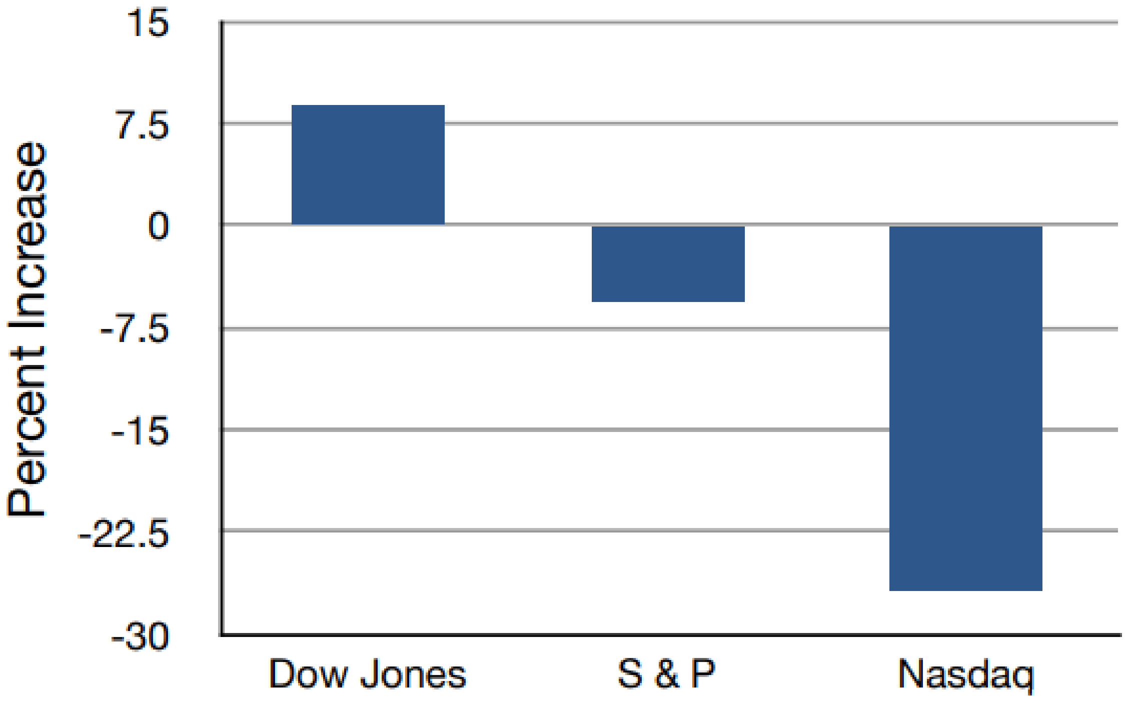


Figure 1. iMac buyers as a function of previous computer ownership.

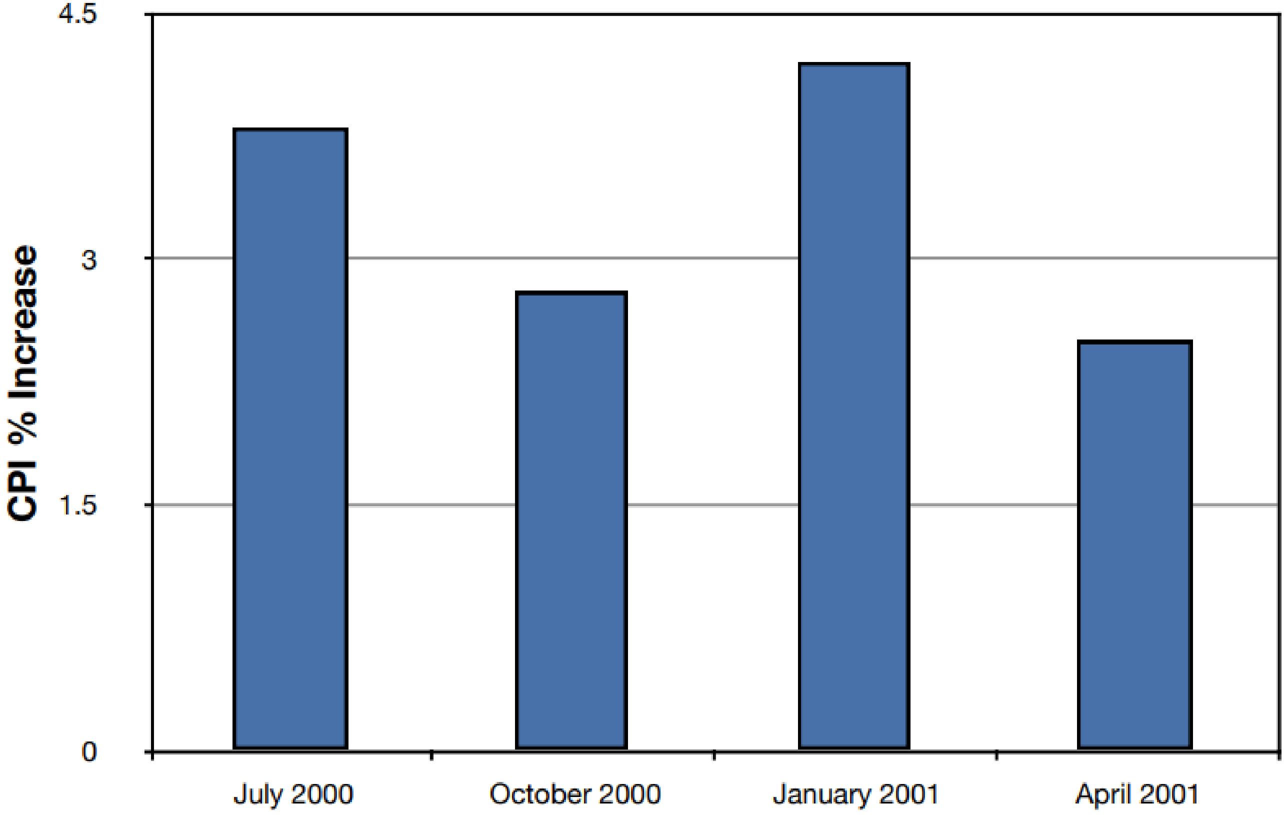
Bar Charts

- Figure shows the percent increases in the Dow Jones, S & P, and Nasdaq stock indexes from 24 May 2000 to 24 May 2001.
- Notice that both the S & P and the Nasdaq had “negative increases” which means that they decreased in value.
- In this bar chart, the Y-axis is not frequency but rather the signed quantity percentage increase.



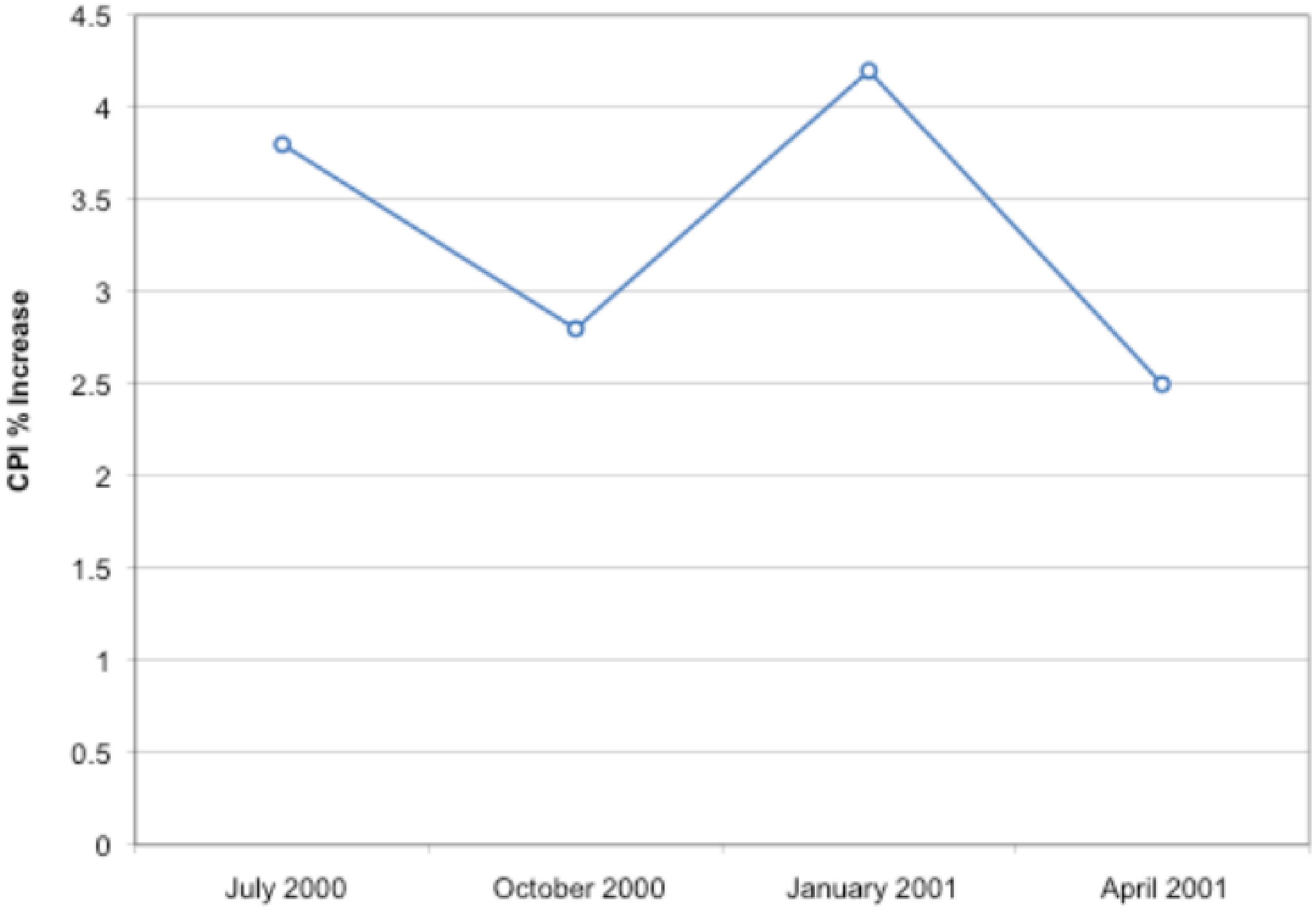
Bar Charts

- Bar charts are particularly effective for showing change over time.
- Figure shows the percent increase in the Consumer Price Index (CPI) over four three-month periods.
- The fluctuation in inflation is apparent in the graph.



Line Graphs

- A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed).
- The line graph emphasizes the change from period to period



Line Graphs

- Line graphs are appropriate only when both the X- and Y-axes display ordered (rather than qualitative) variables.
- Although bar charts can also be used in this situation, line graphs are generally better at comparing changes over time.
- CPI (Consumer Price Index)

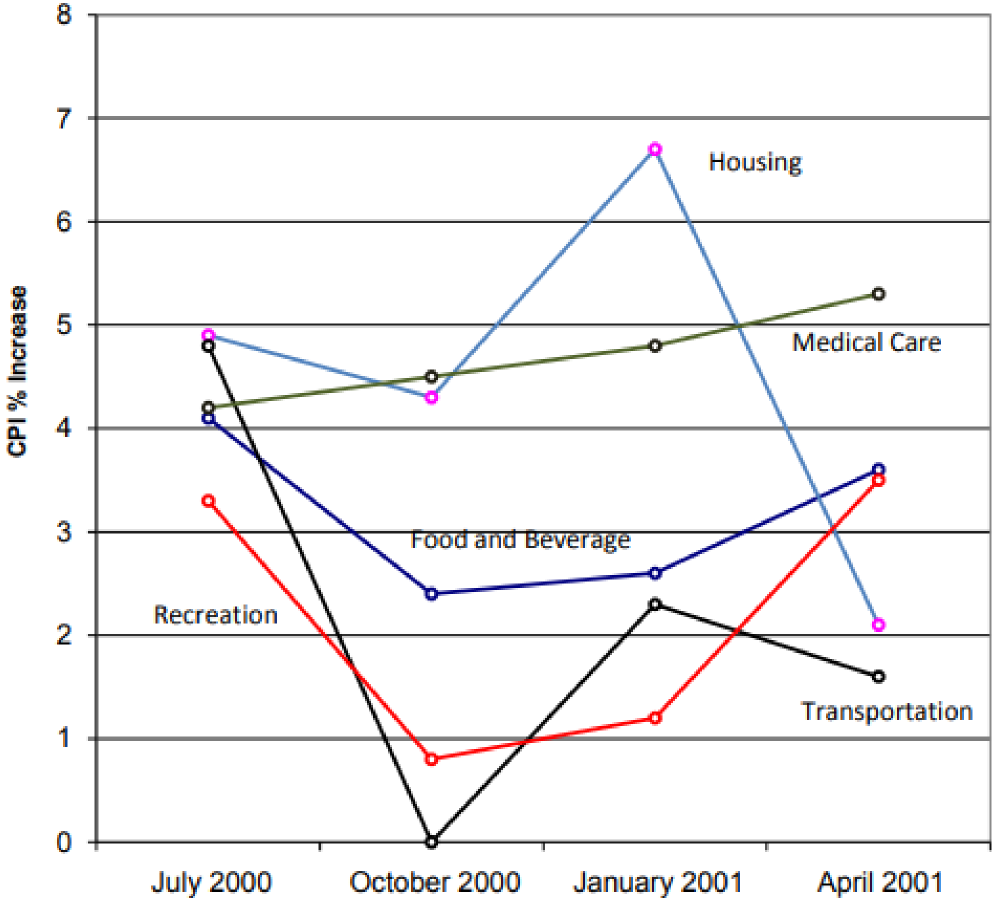


Figure 3. A line graph of the percent change in five components of the CPI over time.

Line Graphs

- Let us stress that it is misleading to use a line graph when the X-axis contains merely qualitative variables.
- The defect in this figure is that it gives the false impression that the games are naturally ordered in a numerical way.

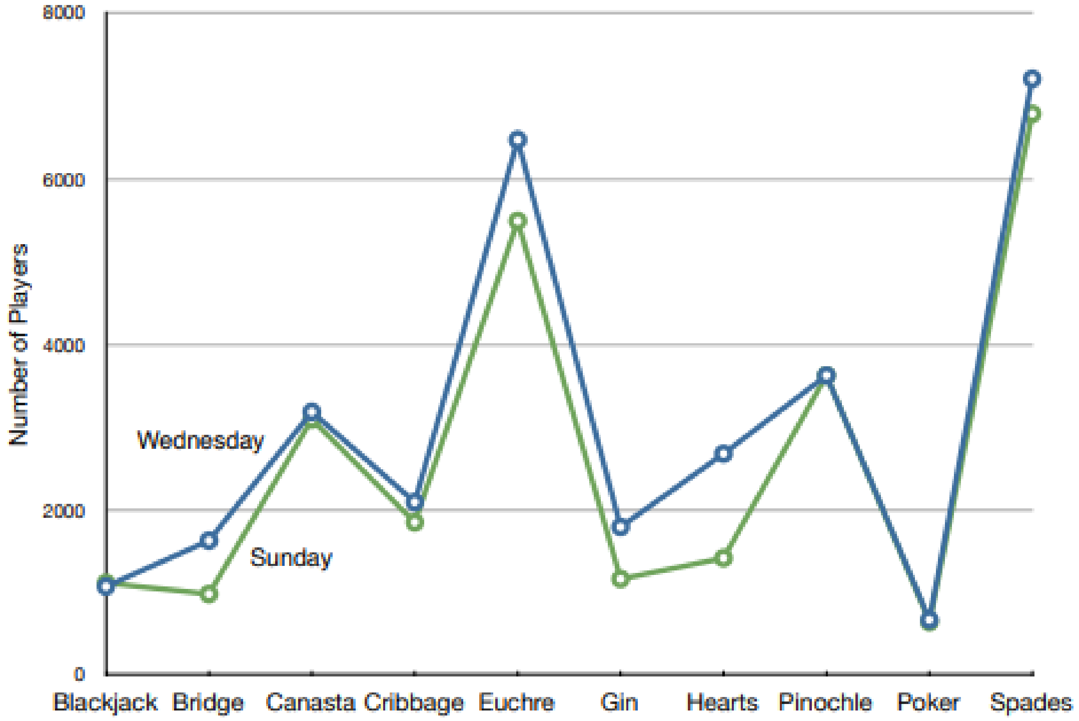
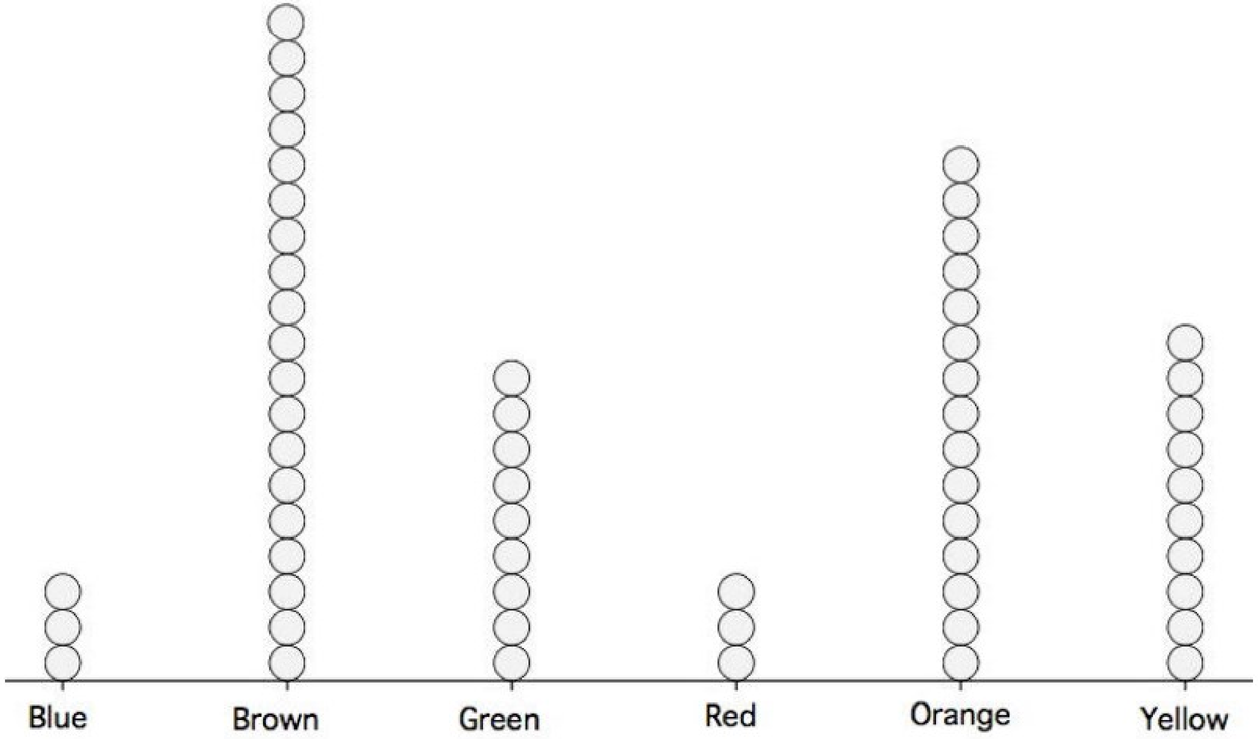


Figure 4. A line graph, inappropriately used, depicting the number of people playing different card games on Wednesday and Sunday.

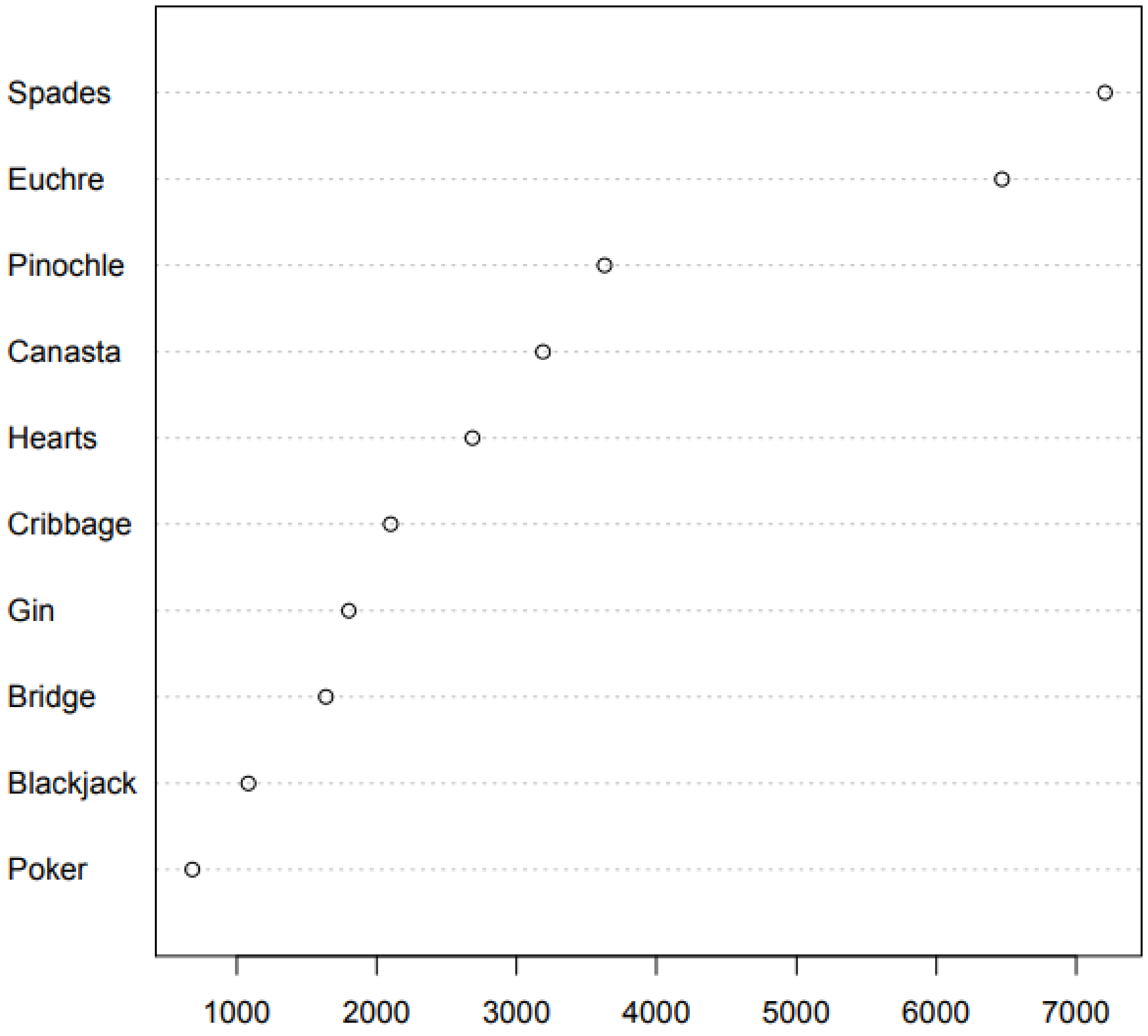
Dot Plots

- Dot plots can be used to display various types of information.
- Figure uses a dot plot to display the number of M & M's of each color found in a bag of M & M's.
- Each dot represents a single M & M.
- From the figure, you can see that there were 3 blue M & M's, 19 brown M & M's, etc.



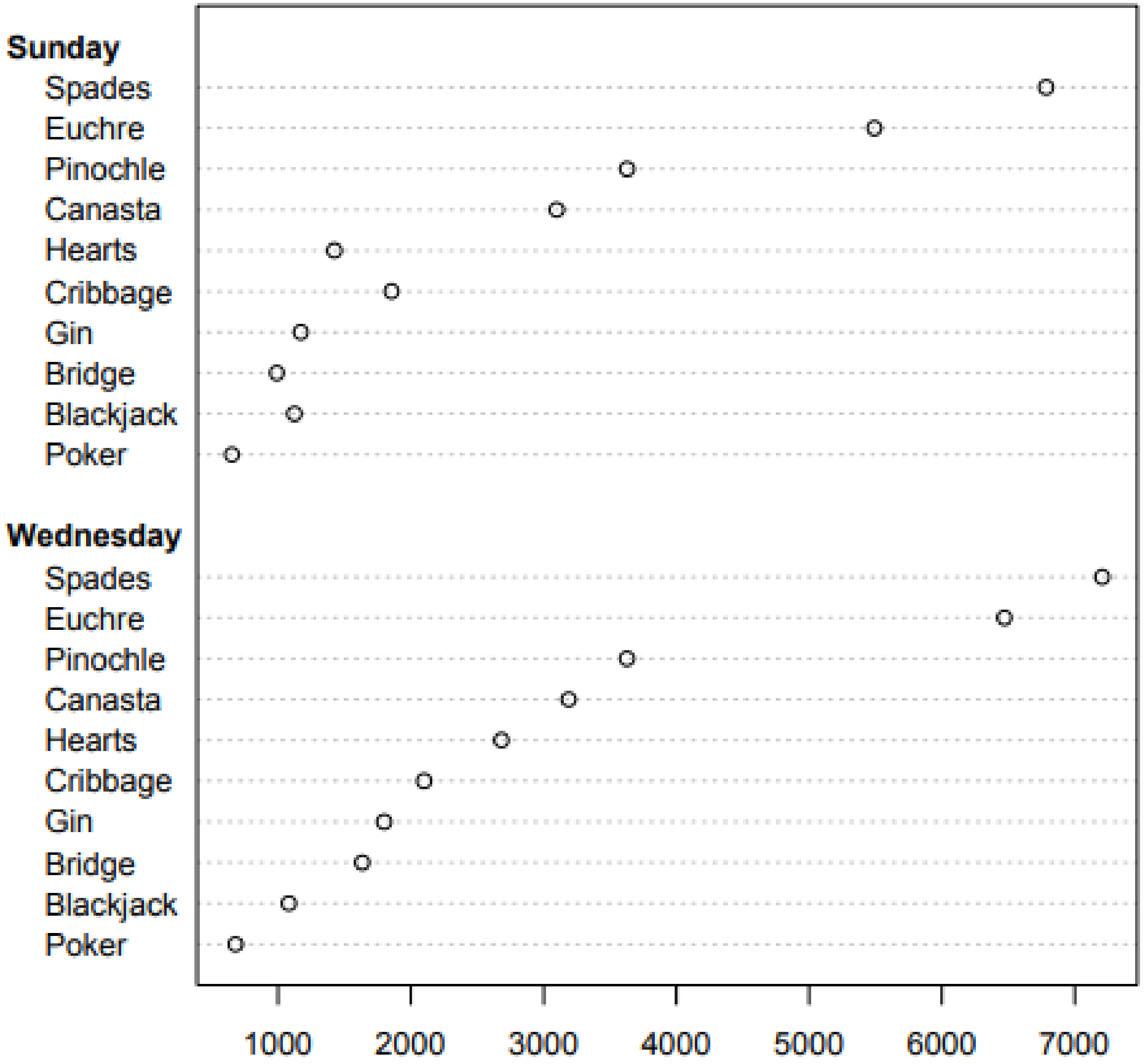
Dot Plots

- The dot plot in Figure shows the number of people playing various card games on the Yahoo website on a Wednesday.
- Unlike previous figure, the location rather than the number of dots represents the frequency.



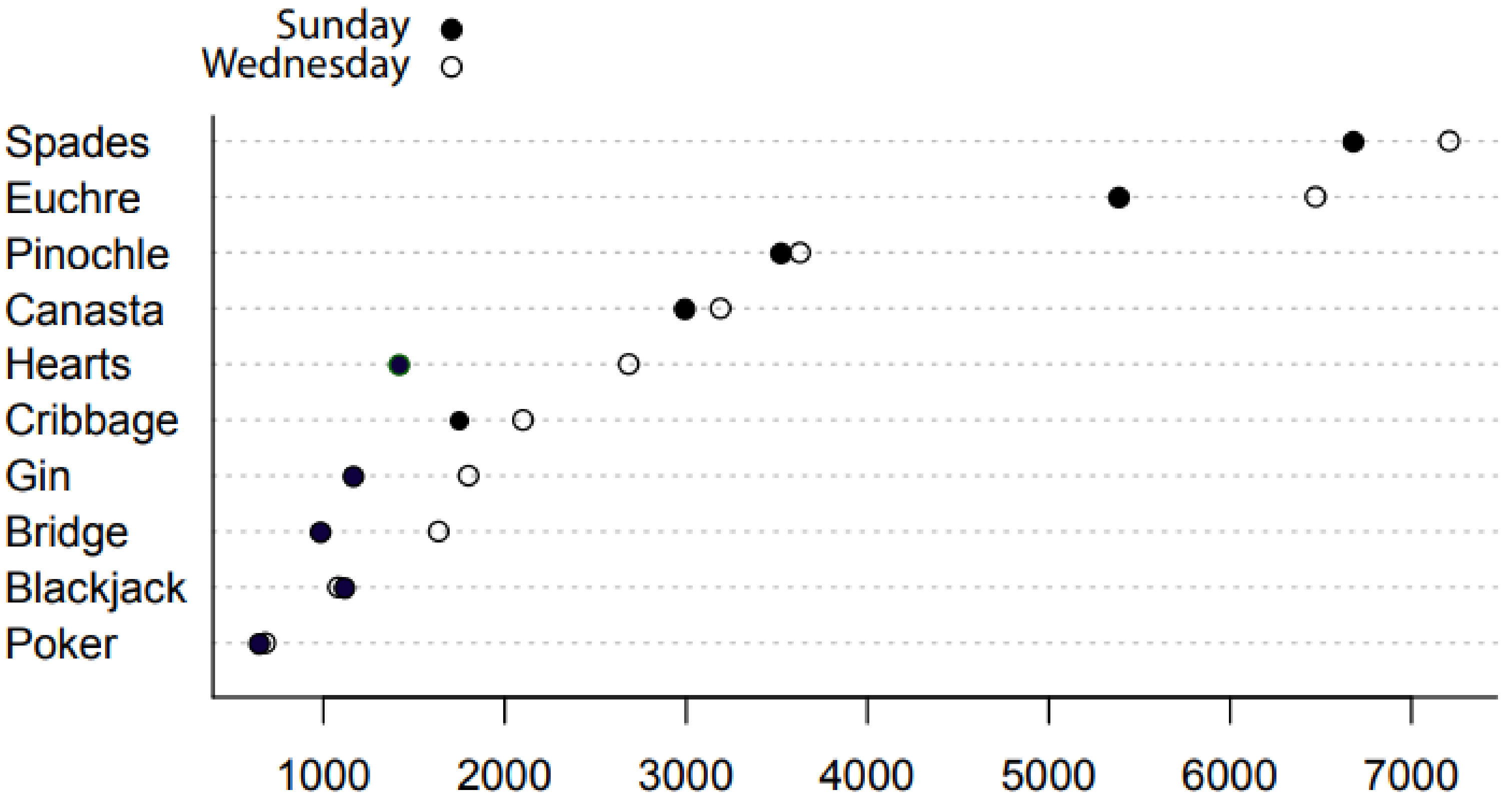
Dot Plots

- The dot plot in this figure shows the number of people playing on a Sunday and on a Wednesday.
- This graph makes it easy to compare the popularity of the games separately for the two days but does not make it easy to compare the popularity of a given game on the two days.



Dot Plots

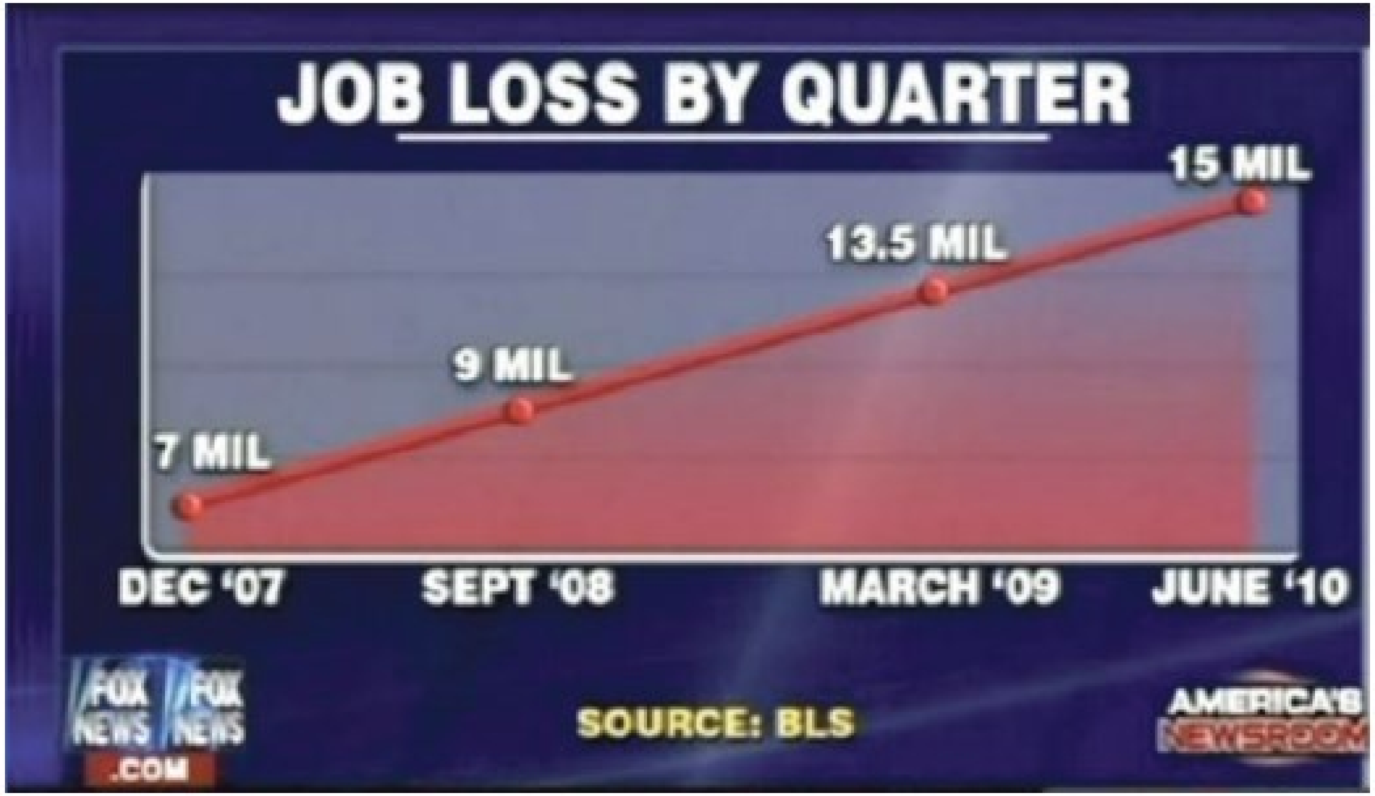
- The dot plot in this figure makes it easy to compare the days of the week for specific games while still portraying differences among games.



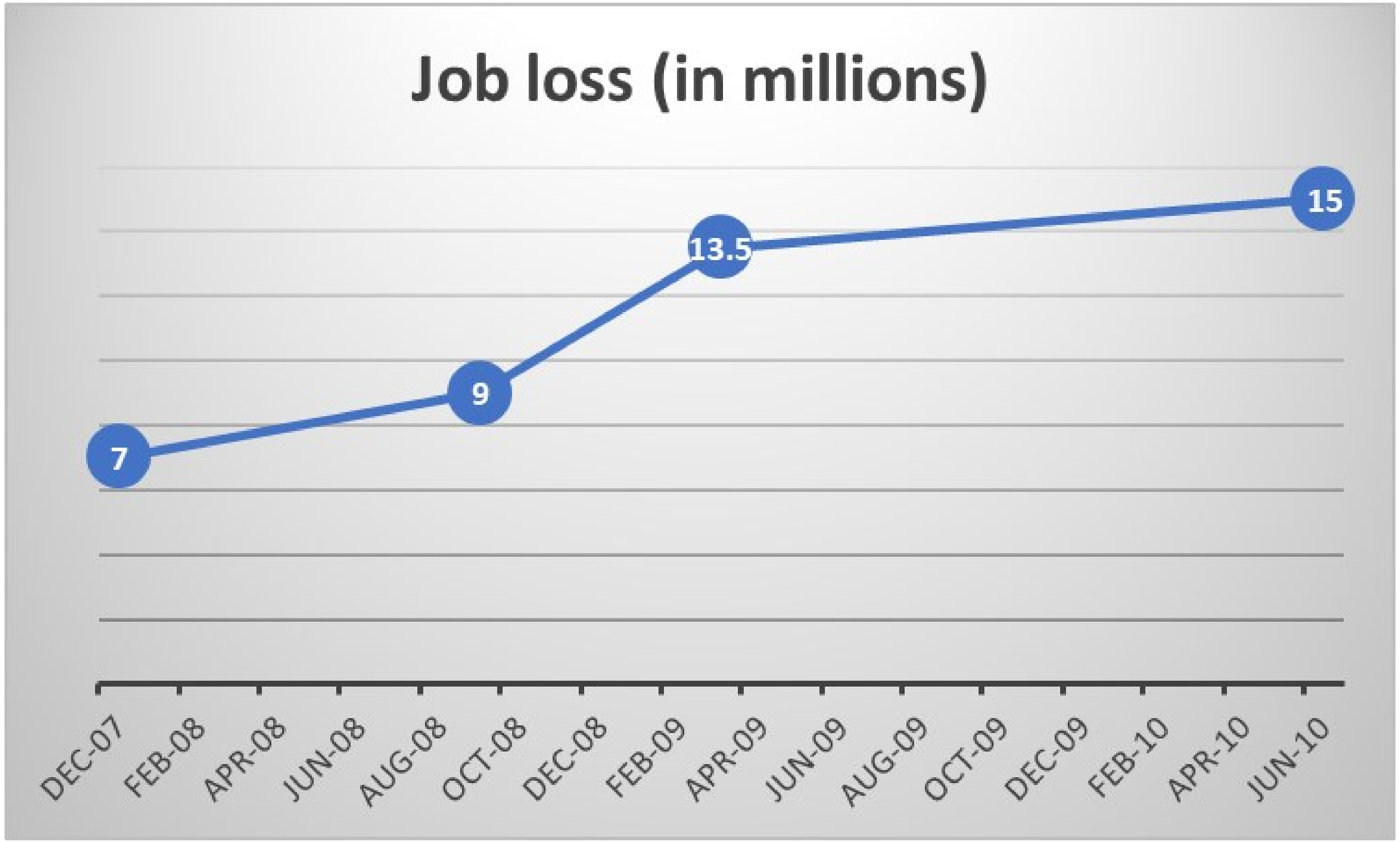
Statistical Literacy

Statistical Literacy

- Fox News aired the line graph below showing the number unemployed during four quarters between 2007 and 2010.
- Does Fox News' line graph provide misleading information?
- Why or Why not?



Statistical Literacy



References

- Witte, R.S.&Witte, J.S. (2017). Statistics (11th ed.). Wiley. ISBN: 978-1119386056.
- Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D.& Zimmer, H. (2003). Introduction to Statistics. Online edition at https://open.umn.edu/opentextbooks/textbooks/459
- Levine, D.M., Stephan, D.F. & Szabat, K.A. (2017). Statistics for Managers Using Microsoft Excel (8th ed.). Pearson. ISBN: 978-0134566672

Thank you