Chapter 3 - Displaying and Summarizing Quantitative Data

August 25, 2014

3.1 Graphs for Quantitative Data (LABEL GRAPHS)

Histogram (p. 44) - Graph that uses bars to represent different frequencies or relative frequencies in specific intervals called classes or bins. The intervals are the same width and can be chosen for convenience. Look similar to Bar Charts, but there are no spaces between bars unless there is no data in that interval.

Stem and Leaf Displays (p. 46) - (Also called Stem and Leaf Plots or Stem Plots) Graph that is created by making a stem out of the left most digits and writing them in order in a vertical column. Then creating a leaf out of the right most digit and putting it in the row that corresponds to the appropriate stem.   Leaves must be arranged in numerical order. This graph gives a quick picture of the distribution while including the numerical values.

If a row does not have any leaves to put in the stems must still be included to show where gaps might be.

If decimal values are used, the decimal values can be rounded.

Example: Length of ownership (in months) of cars before they are traded in.

72   15   70   40   42   74   64   68   50   53   62   64   45

72   15   70   40   42   74   64   68   50   53   62   64   45

Split Stem and Leaf Plot (p. 46) - Plot that is created by splitting the stems into 2 or more rows. If 2 rows per stem than one will be used for digits 0-4 and the other for digits 5-9. This is useful for a larger sets or sets with large amounts of data in each row.

We will see **Back to Back Stem and Leaf Plots** in Chapter 4.

Dotplots or Line Plot (p. 47) – Graph constructed by placing a dot along the axis for each instance of a data value. Axis can be vertical or horizontal.

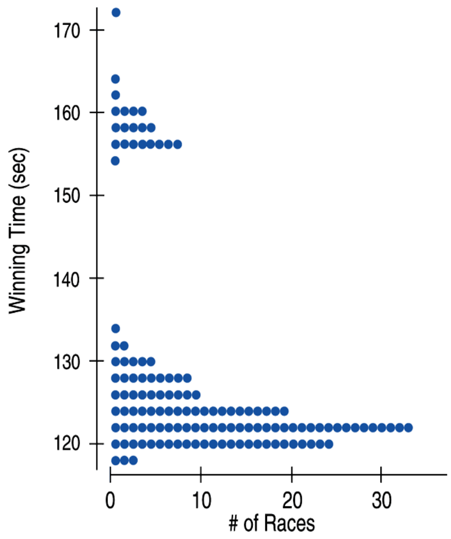
Distribution for Graphs of Quantitative Variables

Whenever a distribution of a graph for a quantitative variable is described it should include its shape, center, and spread.

3.2 Shape of a Distribution

- Is the graph symmetric, one side is close to the mirror image of the other or if folded along a vertical line through the middle the graph would match pretty closely.
- Is the graph skewed, which means one tail stretches out farther than the other. The graph is said to be skewed to the side of the longer tail
- Are there peaks or modes on the graph? This is where a large part of the data is occurring. Unimodal - one peak. Bimodal - two peaks. Multimodal - 3 or more peaks. Uniform if no peaks and fairly straight across graph.
- Are there any outliers, extreme values that are away from the rest of the data. Sometimes outliers will be ignored.
- Are there any gaps or spaces in the graph?

Example:
Dotplot of Kentucky Derby winning times over the years.



3.3 The Center of the Distribution: The Midrange and The Median

For a unimodal, symmetric distribution the **center** is the point on the graph where graph folds to give mirror image of each side.

What if the distribution is skewed or multimodal?

Midrange (or Midpoint) - $\left(\frac{\text{Maximum Value} + \text{Minimum Value}}{2}\right)$. Easy to calculate, but it is very sensitive to extreme outlying values. It is not usually used to summarize a distribution.

Median (p. 51) - Score in the middle when values are arranged in numerical order

If n is odd then median is in the $\left(\frac{n+1}{2}\right)^{\text{st}}$ position

If n is even then median is the average of the numbers in the $\left(\frac{n}{2}\right)^{\text{st}}$ position and the $\left(\frac{n}{2}+1\right)^{\text{st}}$ position.

Example: 10 25 30 35 40 60 60 65 100

10 25 30 35 40 60 60 65 80 100

The median is one of the many ways to find the center of the data. Another important way will be mentioned later.

3.4 The Spread of the Distribution: The Range and The Interquartile Range

The **spread** measures how much the data varies from the center

Range (p. 52) - Maximum – Minimum. Sensitive to outlying values so does not always represent data properly.

Instead of looking at the ends of the data the range of the middle of the data could be measured.

Interquartile Range (p. 52) - IQR = $Q_3 - Q_1$

Quartiles split the sorted data into Quarters.
The median is the middle quartile or Q_2.
The Third Quartile, Q_3 or the Upper Quartile, is the median of the upper half of the numbers.
The First Quartile, Q_1 or the Lower Quartile, is the median of the lower half of the numbers
The lower and upper quartiles are also known as the 25th and 75th percentiles, respectively.

(If n is odd the book includes the median in both halves when figuring up the quartiles. I will not, since that is the way the calculator leaves it out if n is odd.)

Example: 10 25 30 35 40 60 60 65 100

10 25 30 35 40 60 60 65 80 100

3.5 Boxplots and 5-Number Summaries

The **5-number summary** (p. 54) is the following

Minimum Score, Q_1, Median, Q_3, Maximum Score

The **Boxplot or Box and Whisker Plot** (p. 54) is a graphical representation of the 5-number summary.

Boxplots are useful when comparing different groups of data.

Creating a Boxplot:

1. Draw an axis and label appropriately.

2. Draw a box with the ends being the Q_1 and Q_3 and a line in the middle of the box for the median. The box shows where the data in the 25th percentile to the 75th percentile are located or the middle 50%.

3. Calculate **Upper and Lower "Fences" or Limits** to determine which points are outliers. Outliers will be any values outside interval formed by fences.

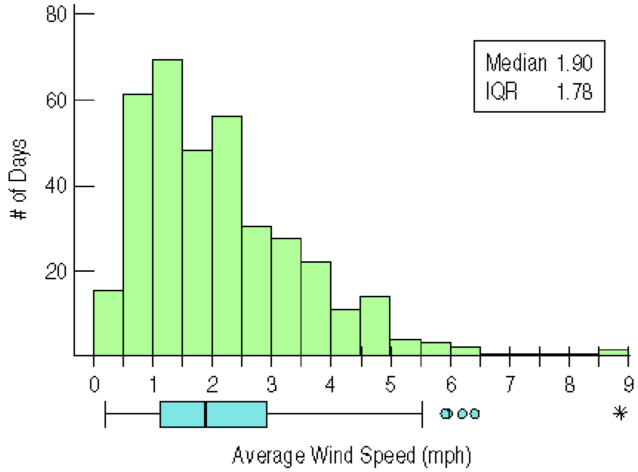
$$\text{Upper Fence} = Q_3 + 1.5 \times \text{I.Q.R.}$$
$$\text{Lower Fence} = Q_1 - 1.5 \times \text{I.Q.R.}$$

4. Mark outliers with a special symbol (*). Draw whiskers out to smallest and largest values of data that are not outliers. (Sometimes a different symbol is used for "Far Outliers" – data values farther than 3 IQRs from the quartiles.)

Example: The following data represent the weight (in pounds) of 15 five year old girls.

30 32 34 35 36 36 37 38 41 42 44 45 47 62 65

Histogram and Boxplot for daily wind speeds (see also Fig 3.13, p. 55).



How does each represent distribution?

3.6 The Center of Symmetric Distributions: The Mean

If the distribution is symmetric then calculations for the center and spread can be used that include all data values.

Mean(p. 57) – The value found by summing up the numbers and dividing by n.

$$\overline{x} = \frac{\Sigma x}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Typically will see mean rounded to one more digit than data has.

Example: 10 25 30 35 40 60 60 65 100

Median is point where there are half of the scores above and half of the scores below. Median is not sensitive to outlying values.

Mean is point where histogram would "balance". Mean is more sensitive to outlying values.

If distribution is symmetric then mean = median.

If distribution is skewed the mean is pulled closer to the tail than the median.

If distribution is symmetric and there are no outliers mean is preferred measure of center. If distribution is skewed or has outliers than median is preferred since the median is resistant to extreme large or small values.

3.7 The Spread of Symmetric Distributions: The Standard Deviation

Variance and Standard Deviation (p. 59)
IQR only uses two quartiles (or 50%) of the data, so it ignores how individual values vary. The Standard Deviation takes into account how far each value is from the mean. These differences are called deviations.

Variance: $s^2 = \frac{\Sigma\left(x-\bar{x}\right)^2}{n-1}$ **Standard Deviation:** $s = \sqrt{\frac{\Sigma\left(x-\bar{x}\right)^2}{n-1}}$

Why $(\ )^2$ each term?

Why Standard Deviation instead of Variance?

Like the mean the standard deviation is only appropriate for symmetric data.

Also as with the mean typically round standard deviation to one more digit than original data contains.

Example: Find the Standard Deviation of 1, 3, 4, 8, 9.

What to "Tell" about a Quantitative Variable (p. 61)

What can go wrong? (p. 64-76)