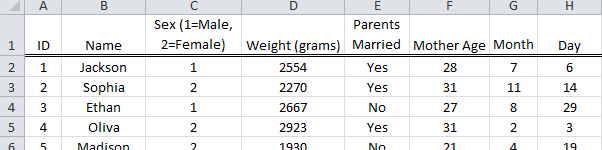
Chapter 5: The Analysis of a Single Numerical Variable

In this chapter we will consider the analysis of a single set of numerical measurements. The approach and methods used in this chapter are somewhat different as summaries beyond counts and percentages will be necessary.

|  |
| --- |
| Definition |
| Numerical Variable: A variable, i.e. set of measurements, that are on a *naturally* numeric scale. |

Consider the following list of data regarind the newborn births.



Questions:

1. Is weight measured on a naturally numeric scale?
2. Why is Sex of child not a naturally numeric variable?
3. Would Month be consider naturally numeric? How about Day?
4. Suppose your friend decides to compute the average for Sex. What would this average tell us? Discuss.

For categorical variables, appropriate summaries include counts, percentages, bar graphs, and mosaic plots. There are a wider variety of summaries that are appropriate for numerical variables. In particular, numerical variables have several characteristics that are of interest. In the next section, we will consider the following general characteristics for a single set of numerical variables.

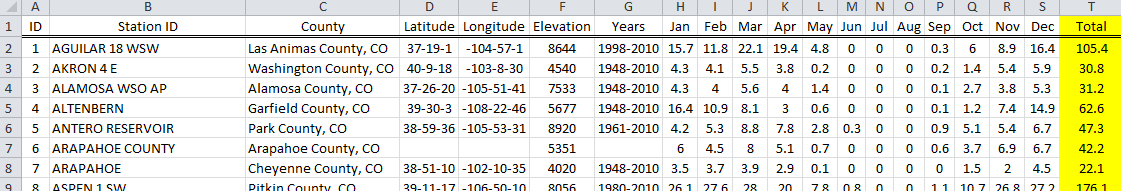
Characteristsics of Interest for Numerical Variables

* Location / Position
* Spread
* Shape
* Outliers / Unusal Observations

5.1: Descriptive Summaries for a Numerical Variable

In this section, we will consider methods used to describe various characteristics of a numerical set of mesurements.

Example 5.1.1 The following data includes snowfall amounts for various weather stations across the state of Coloarado. These measurements were obtained over several years. The data also includes monthly breakdown of the snowfall amounts.

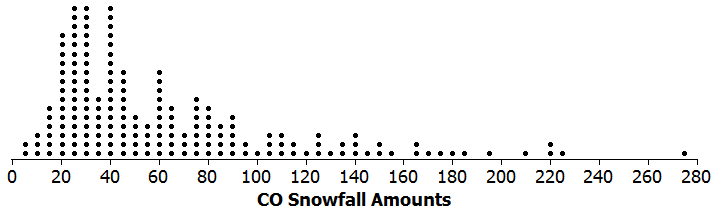


Sources:

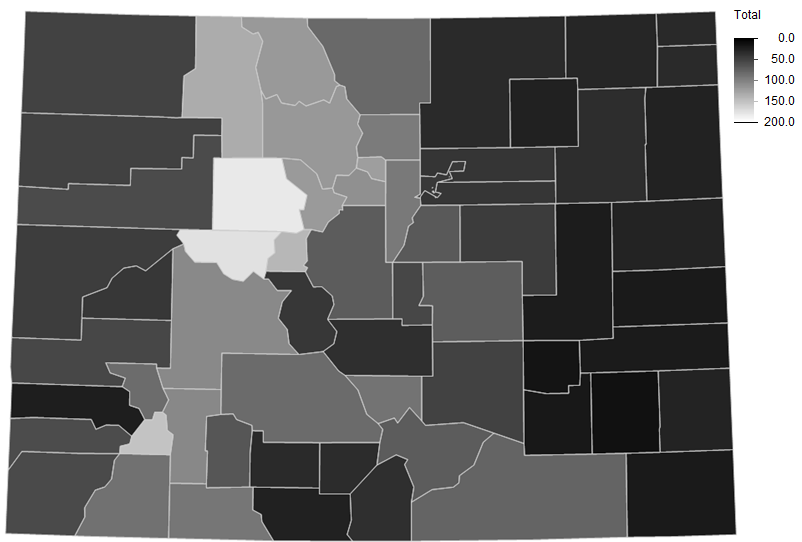
* <http://www.wrcc.dri.edu/htmlfiles/co/co.sno.html>
* <http://www.wrcc.dri.edu/summary/Climsmco.html>

|  |  |
| --- | --- |
| Weather Stations in CO | County Names for CO |

A dotplot of the snowfall amounts from the 180 weather stations across for Colorado.

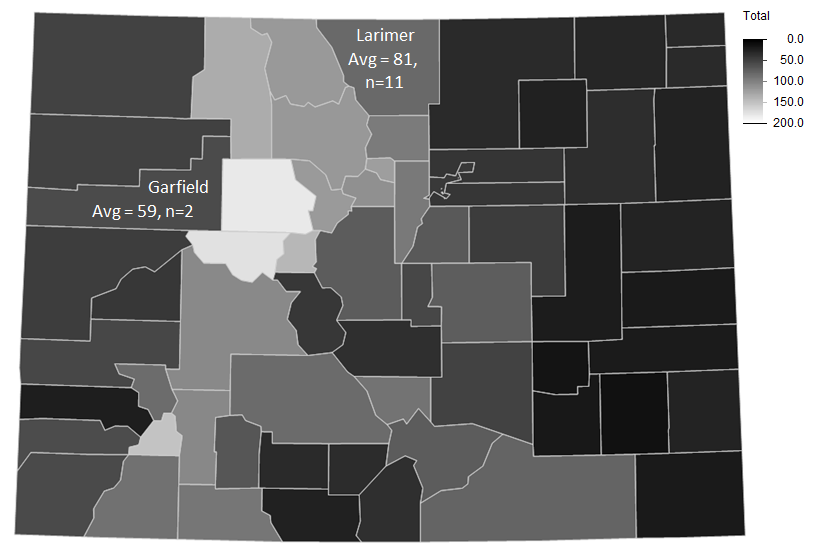


Maps such as the following can be used to visualize snowfall amounts geographically.



Questions:

1. Recall, data was collected on 180 weather stations across Colorado. This map shows the outcomes for each of the 64 counties. What do you think the map did in order to reduce the 180 weater station outcomes down to the 64, ie.. one measurement for each county? Discuss.
2. Consider the fact that Larimer County has 11 weather stations and Garfield County only had 2.



1. Is the average for Larimer higher because it has more weather stations? Discuss.
2. What impact might the number of stations have on the margin-of-error for the average for each of these counties? Discuss.

Measures of Location / Position

The most common measures of location are for the middle of dataset. The middle is typically described using one or more of the following: Mean, Median, and ~~Mode~~.

* Mean (or Average):
* Median: The middle value of a dataset (after the numerical values have been put in order). If the dataset contains an even number of observations, then the median is the average fo middle two observations.
* ~~Mode: The measurement that occurs most often.~~

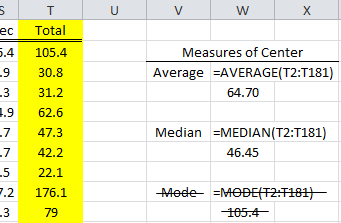
Comment: I have used strikethrough font for Mode as it is not necessarily a good measure of center. In fact, for asymmetric data, the mode is nowhere near the center.

Gettiing these summaries in Excel

In Excel, the mean, median, and mode can be found using the following functions.

|  |  |
| --- | --- |
| Quantity | Excel Function |
| Mean | =AVERAGE() |
| Median | =MEDIAN() |
| ~~Mode~~ | =~~MODE()~~ |

Getting these quantities for the CO snowfall data.

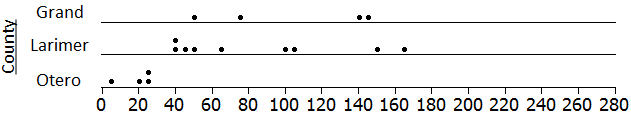


When an average is computed for a set of numbers, the information contained in the measurements is reduced or condensed into a single number.

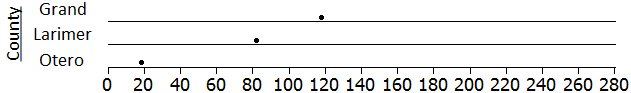
|  |  |
| --- | --- |
| The average is the *balance point* in the distribution. |  |
| Measurements get *condensed* to an average |

There are advantages (and disadvantages) to reducing this information down to a single number. For example, consider the snowfall amounts for the following counties.

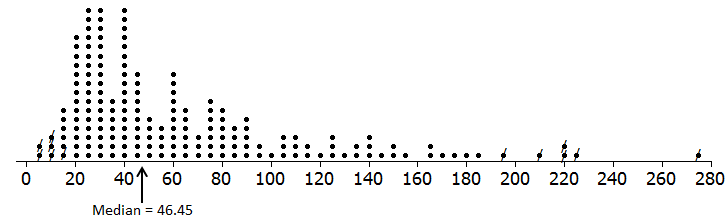
Question: How different are the snowfall amounts in these counties?



If the average of each county is computed, then measuring the amount of difference between the three counties is straight forward.



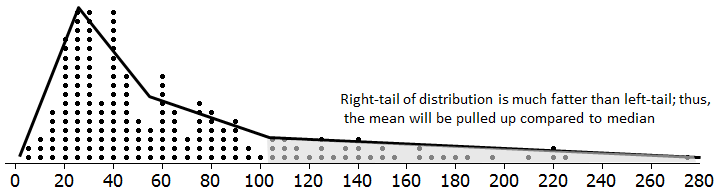
The median is guaranteed to be the middle value in a set of measurements. There are 180 measurements in the following dotplot, thus the middle value would be halfway between the 90th and 91st measurement, which is 46.45 inches for the CO snowfall data.



For this particular example, the mean and median are somewhat different. In fact, the difference between these two measures of center is just about 20 inches (which is a lot considering these relative size of our measurements).

|  |  |
| --- | --- |
| Mean = | 64.70 inches |
| Median = | 46.45 inches |

The average snowfall amount is somewhat larger than median because the snowfall measurements tend to be more spread out in the right-tail of the distribution compared to the left-tail. Statisticians affectionately say this distribution would have a fat right-tail or upper-tail.

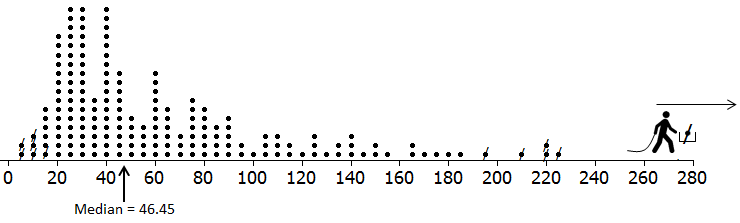


Comment: There is a natural lower boundary for snowfall amounts (snowfall cannot be less than zero); thus, it might be expected that measurements of this nature have a more substantial right-tail than left.

A single observation may adversely affect the mean. For example, the most extreme snowfall has a measurement of 274.5. If this measurement is increased substantially, then the average will be adversely effected. In some sense, each observation is tethered to the mean (the actual measurement is used in the calculation of the mean). So when a single observation is increased, it will have a direct and adverse impact on the mean. In short, the mean is adversely affected by outliers.

|  |  |
| --- | --- |
|  | Outliers have a direct impact on the formula for the mean. |

On the other hand, the median is \*not\* affected by single observations. For example, the largest observation can be increased substantially, but it will have \*no\* impact on the median. In this sense, the median is less affected by outliers, uneven tails, etc.



|  |
| --- |
| Excel Tips |
| Naming a range of cells.  A set of cells can be assigned a name. This name can then be used in functions instead of cell ranges.    This name can then be used in any formula, etc.    Filtering  To Filter in Excel, select Data > Filter. Each column in your dataset should now include a drop-down arrow. To select rows for stations in Adams County, CO, select only Adams County from the County drop-down menu as is shown here.    The resulting rows are shown. Only stations in rows 17, 22, and 126 (row labels shown in blue) are in Adams County, CO.    Note: Formuals do \*not\* ignore the hidden rows when filtering as been applied. |

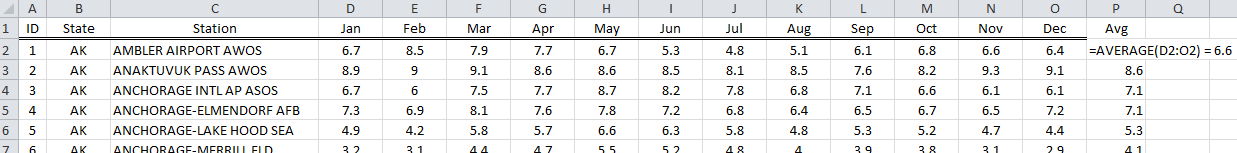
Example 5.1.2 The following data includes wind speeds from weather stations in the Western United States. These measurements were obtained over a 10 year. Once again, this data has been condensed from its orginial form to monthly averages for each weather station.

Source: <http://www.wrcc.dri.edu/wind-information/>

|  |  |
| --- | --- |
| Wind speeds for 445 weather stations in Western States of US    Question: Where to build?  http://ts4.mm.bing.net/th?id=H.4578101342243463&pid=1.7 http://windturbinezone.com/wp-content/uploads/2010/02/wind-farms1.jpg |  |

The map above condenses the wind speeds substainitally by computing an average for each state. If the goal is to find an optimal place to build a wind farm, we should not condense the wind speed data too much. In particular, we should probably consider the average wind speed at each weather station.

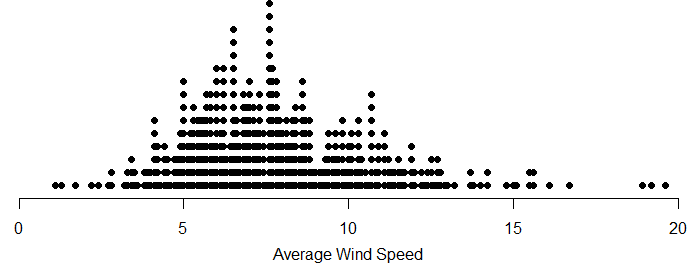
Getting the average wind speed for each weather station in Excel can be easily using either the =AVERAGE() function or the PivotTable feature.



Getting the averages using PivotTables with the *stacked* version of the dataset.

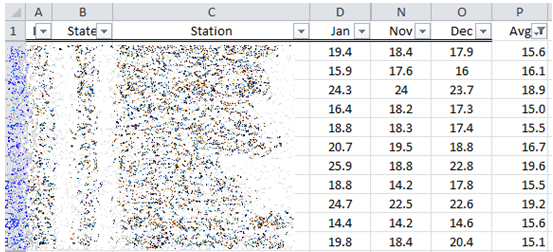
|  |  |
| --- | --- |
|  | The setup for getting the averages |

A dotplot of the average wind speed for all weather stations is provided here.



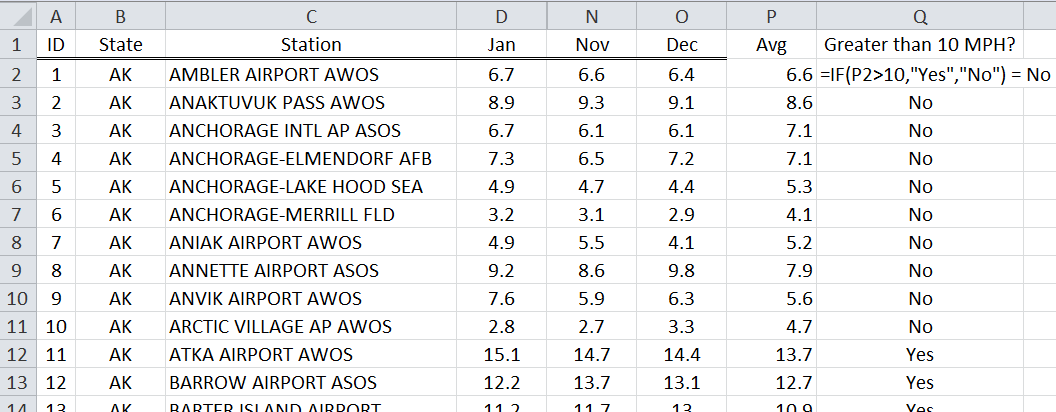
Questions

1. What weather stations have an average wind speed greater than, say 15? Use a Filter in Excel to obtain a list similar to the one below.

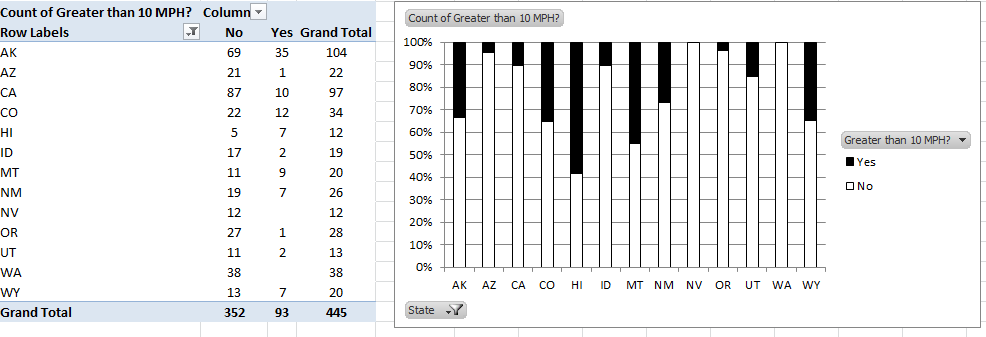


1. From the map created above, HI appears to have the largest average wind speed. How many weather stations in the list above were from HI?

Consider the following use of an =IF() statement in Excel. Here, I have created a new variable that identifies weather or not the average wind speed is greater than 10. This new variable is simple called Greater than 10 MPH?



This new variable was then used to construct the following table and graph. These summaries include the number and percentage of weather stations in each state whose average wind speed is greater than 10 mph.



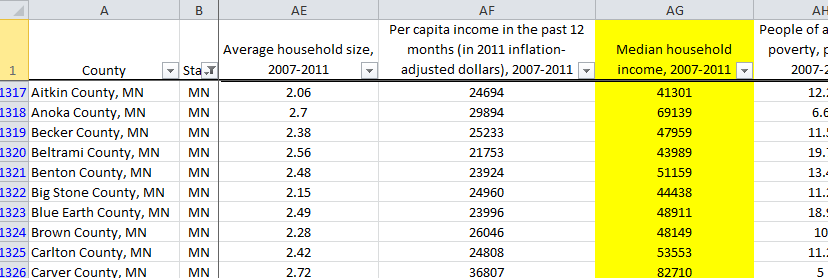
1. Which state tends to have the most weather stations whose average wind speed is greather than 10 mph?
2. Does your answer above agree with the initial map that was drawn?

General Measures of Location (not necessarily center)

Example 5.1.3 The Census Bureau provies a variety of information at the county level (for all counties across the United States) in it’s State and County QuickFacts datasets. In this example, we will consider the County level quickfacts data for the state of MN.

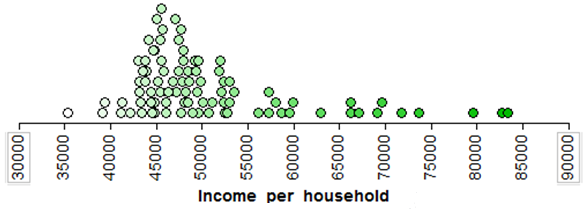
Source: <http://quickfacts.census.gov/qfd/download_data.html>

In Excel, you can use the Filter option to In the following, I have filtered down to just counties from MN. An estimate of Household Income is provided in column AG as is shown here.



**Percentile**: The pth percentile of a set of measurements is defined to be the *point* in the data set where *p%* of the measurements fall at or below.

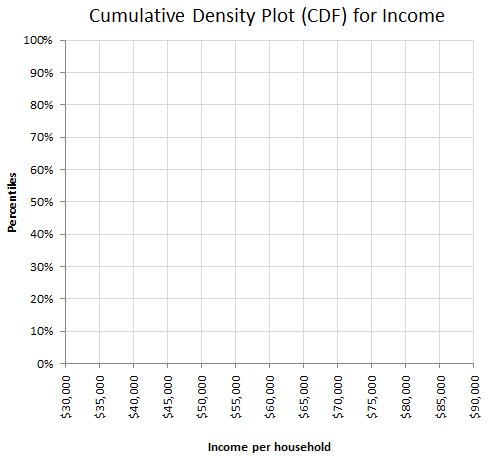
The income per household for all 87 counties in MN is provided here.



One way to understand percentiles is to determine the percentage of observations less than a particular point. For example, about 3% of the counties have income levels below $40,000.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Income per household | Percentiles | | $35,000 | 0% | | $40,000 | 3% | | $45,000 | 26% | | $50,000 | 64% | | $55,000 | 79% | | $60,000 | 87% | | $65,000 | 89% | | $70,000 | 94% | | $75,000 | 97% | | $80,000 | 98% | | $85,000 | 100% | |
|  |
|  |

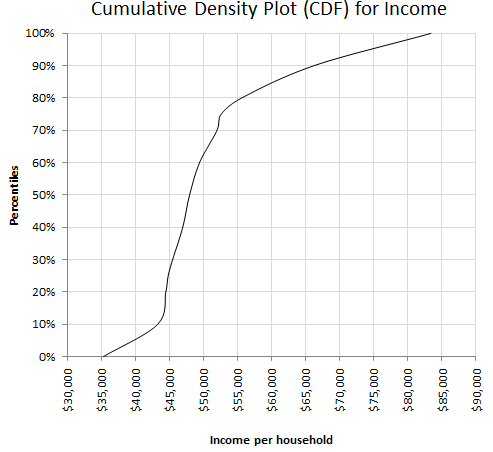
A cumulative density function (CDF) plot is used to display the spectrum of percentiles from a set of data. Plot the income levels (x-axis) and their respective percentiles (y-axis) below.



An equivalent approach would be to pre-determine certain percentages and then determine the income level for that percentile. For example, the bottom 10% of the incomes would fall below $43,285.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | |  | Income per household | Percentiles | |  | $35,307 | 0% | |  | $43,285 | 10% | |  | $44,472 | 20% | | -> | $44,820 | 25% | |  | $45,475 | 30% | |  | $46,960 | 40% | | Median -> | $47,959 | 50% | |  | $49,420 | 60% | |  | $51,987 | 70% | | -> | $52,598 | 75% | |  | $55,590 | 80% | |  | $66,208 | 90% | |  | $83,415 | 100% | |
|  |
|  |

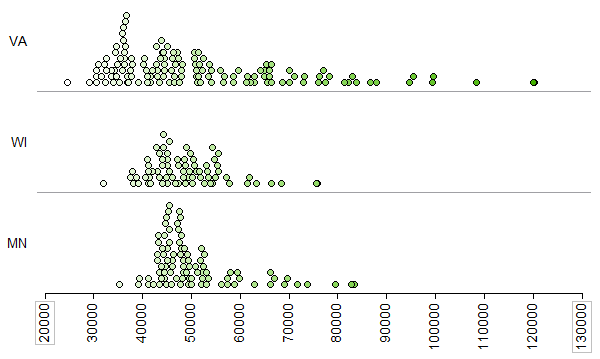
The CDF plot is equivalent to the one sketched above.



Questions: Use the above table of percentiles and CDF plot to answer the following questions.

1. What is the median income level per household for MN?
2. How could you determine the median from the CDF plot? Discuss.
3. What is the minimum income level per household in MN? How about maximum?
4. The CDF plot has a longer tail on the upper-end then on the lower-end. What does this imply about income levels per household in MN? Discuss.
5. The CDF plot is fairly steep $45,000 and $55,000. What does this imply about income per household in MN? Discuss.

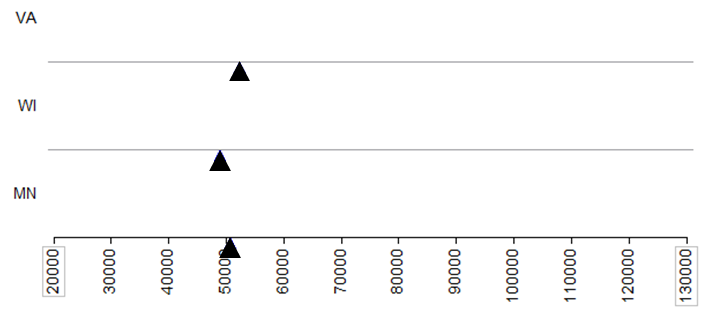
The following stacked dotplots show the income per household for Minnesota, Wisconsin, and Virginia.



Questions:

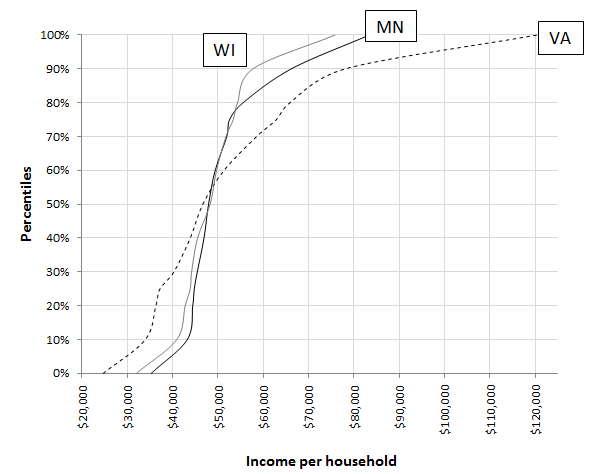
1. What difference exist in the income per household across these three states? Discuss.
2. VA dots are more spread out than MN and WI. Practically speaking, what does this imply about the income levels across the counties in VA? Disucss.

As previously stated, the average is used to condense information down so that comparisons can be made. In some situations, the average may not be sufficient because averages remove the sense of disparity in the income levels. In a sense, computing the averages has condensed the information too much.



In comparison, a CDF plot clearly shows the entire spectrum of income levels. Putting all three states onto a single CDF plot allows for comparisons to be made across *all* incomes levels. For example, we can compare the poor people from each state against other states, and compare the rich people from each state against other states, etc.

CDF to Compare Incomes



Questions:

1. Consider the poorest people in each state. In which state is the incomes levels the lowest for the population of people? Likewise, for which state is the income levels the highest for the richest people in the population?
2. How do the income levels of MN and WI compare?
3. What is an advantage to using the CDF plot to make comparisons?

5.2: Measures of Spread

As previously mentioned, there are additional characteristics of a set of measurements that are of particular interest other than location. In this section, we will consider the variation, i.e. spread, in a set of numerical measurments.

Consider Situation 1) which is the dotplot again for Income per Household for the counties in Minnesota. Give a description of the spread in incomes across the counties in Minnesota. Compare and constrast this description against Situation 2) whose average Income per Household is the same as Situation 1).

|  |  |  |
| --- | --- | --- |
| Situation | Dotplot of Income per HouseholdGraph | Give a description of variation, i.e. spread, for each situation |
| 1) |  |  |
| 2) |  |  |

|  |  |
| --- | --- |
| A variety of terms are used in place of variation for a set of measurements. | |
| *Disparity*: If a state has high income *disparity*, then incomes are dissimilar. | *Volatility*: Stocks with high *volatility* are considered risky as the rate of return is uncertain.  http://www.bigfatpurse.com/wp-content/uploads/Volatility.png |

**RANGE**

The range is easy to compute and understand; however, the range is very limiting in the information it provides about a set of measurements.

Range = Maximum Value – Minimum Value

|  |  |
| --- | --- |
| Range encompasses all the measurements |  |
|  | |  |  |  | | --- | --- | --- | |  | Income per household | Percentiles | |  | $35,307 | 0% | |  | $43,285 | 10% | |  | $44,472 | 20% | | -> | $44,820 | 25% | |  | $45,475 | 30% | |  | $46,960 | 40% | | Median -> | $47,959 | 50% | |  | $49,420 | 60% | |  | $51,987 | 70% | | -> | $52,598 | 75% | |  | $55,590 | 80% | |  | $66,208 | 90% | |  | $83,415 | 100% | |

The range is \*not\* a commonly used measure of variation or spread by statisticians. There are two main reasons for this.

* Range only uses two observations in the entire dataset

Note: You would never use only two obervations to compute a mean, so don’t do this to compute a measure of variation.

* Range, by definition, is directly and adversely affected by outliers.

|  |  |
| --- | --- |
| Range only used two observations – all other mesaurements are ignored | The range would not adequately measure income disparity across all counties |

Questions:

1. How many observations from the data set are used in the computation of the range?
2. What is the smallest possible value for the range? What does it mean if the range is at this value?

**INTER-QUARTILE RANGE**

John Tukey, the inventor of the box-and-whisker plot, used a quanitity called the inter-quartile range which similar to the range.

|  |  |
| --- | --- |
| The Box-and-Whisker Plot  http://tigger.uic.edu/~jlarson/p343/xtra/p343BoxPlot/Box-Whisker.gif | John Tukey  https://fbcdn-profile-a.akamaihd.net/hprofile-ak-prn1/c10.10.160.160/378504_116574675127007_14821520_a.jpg |

The inter-quartile range, dentoed IQR, is computed as follows.

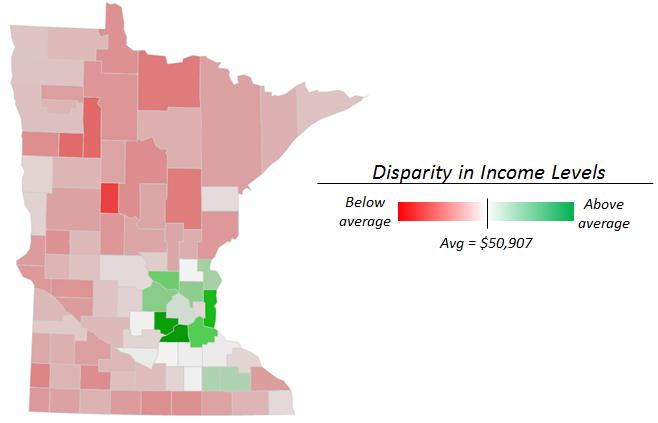
|  |  |
| --- | --- |
| IQR encompasses the middle 50% of the measurements | |
|  | |  |  |  | | --- | --- | --- | |  | Income per household | Percentiles | |  | $35,307 | 0% | |  | $43,285 | 10% | |  | $44,472 | 20% | | -> | $44,820 | 25% | |  | $45,475 | 30% | |  | $46,960 | 40% | | Median -> | $47,959 | 50% | |  | $49,420 | 60% | |  | $51,987 | 70% | | -> | $52,598 | 75% | |  | $55,590 | 80% | |  | $66,208 | 90% | |  | $83,415 | 100% | |

Note: Suppose WI has a smaller IQR value than Minnesota. The interpretation for WI is that their middle 50% of the data is more similar than Minnesota’s middle 50%.

**“DISTANCE-TO-MIDDLE” CONCEPT**

The standard approach to measuring spread or variation in data relies on the concept of distance-to-middle. In particular, a set of measurements that have a smaller total or average distance-to-middle is said to have less spread.

Consider again the following map which shows the Income per Household for counties across Minnesota. Counties whose household incomes are near the average are white. A more instense color indicates a further distance from the average – red indicating income below average and green indicating above.



For simplicity, consider the five counties from Rhode Island and the following hypothetical incomes levels for each county.

|  |  |  |
| --- | --- | --- |
| Data Set A |  |  |
| Data Set B |  |  |
| Data Set C |  |  |

A **residual** is defined to be the distance between a data point and it’s average and is used to measure distance-to-middle.

Residual = Data Point - Average

The residual values for each data point are shown below. Unfortantely, when residuals values from an entire dataset are summed to compute the total distance-to-middle is 0. This happens because the negative residuals cancel out the positive residuals.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data Set A |  | |  | | --- | | Residual | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | |
| Data Set B |  | |  | | --- | | Residual | | -20,000 | | 0 | | 0 | | 0 | | 20,000 | | 0 | |
| Data Set C |  | |  | | --- | | Residual | | -20,000 | | -10,000 | | 0 | | 10,000 | | 20,000 | | 0 | |

Akin to what was done for the Chi-Square test statistic, we can prevent the cancelling out effect when summing the residuals by considering the absolute value of the residuals.

|  |  |  |
| --- | --- | --- |
| Data Set A | Data Set B | Data Set C |
| |  |  | | --- | --- | | Residual | |Residual| | | $0 | $0 | | $0 | $0 | | $0 | $0 | | $0 | $0 | | $0 | $0 | | $0 | $0 | | |  |  | | --- | --- | | Residual | |Residual| | | -$20,000 | $20,000 | | $0 | $0 | | $0 | $0 | | $0 | $0 | | $20,000 | $20,000 | | $0 | $40,000 | | |  |  | | --- | --- | | Residual | |Residual| | | -$20,000 | $20,000 | | -$10,000 | $10,000 | | $0 | $0 | | $10,000 | $10,000 | | $20,000 | $20,000 | | $0 | $60,000 | |

The mean absolute residuals (or deviance), denoted MAD, is computed as follows and is one way to measure distance-to-middle.

In addition to using the absolute value of the residuals, the residuals could have been squared to prevent the cancelling out effect when summing the residuals.

|  |  |  |
| --- | --- | --- |
| Data Set A | Data Set B | Data Set C |
| |  |  | | --- | --- | | Residual | Residual2 | | $0 | 0 | | $0 | 0 | | $0 | 0 | | $0 | 0 | | $0 | 0 | | $0 | 0 | | |  |  | | --- | --- | | Residual | Residual2 | | -$20,000 | 400,000,000 | | $0 | 0 | | $0 | 0 | | $0 | 0 | | $20,000 | 400,000,000 | | $0 | 800,000,000 | | |  |  | | --- | --- | | Residual | Residual2 | | -$20,000 | 400,000,000 | | -$10,000 | 100,000,000 | | $0 | 0 | | $10,000 | 100,000,000 | | $20,000 | 400,000,000 | | $0 | 1,000,000,000 | |

The population variance is defined to be the average distance-to-middle when residual values are squared to alleviate the cancelling out effect described above.

The (sample) variance is used more often than the population variance. The denominator for the sample variance is adjusted slightly due. This adjustmen is necessary as only of the residuals are free-to-vary due to the fact that their sum must be zero.

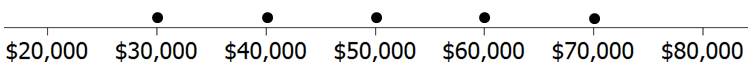
One issue with the variance is the scale as it is \*not\* on the same scale as the original data. To overcome this problem, we can simply take the square root of the variance. The result is called the standard deviation.

The standard deviation measurement has been computed for each data set provided above.

|  |  |  |
| --- | --- | --- |
| Data Set A | Data Set B | Data Set C |
|  |  |  |

Consider the following summaries regarding our desire to measure spread as distance-to-middle.

* The range should \*not\* be used to measure spread in data.
* The mean absolute deviation or standard deviation are acceptable ways to measure spread. The standard deviation is most commonly used by statisticians because of it’s optimal theoretical properties.
* The mean absolute deviation and standard deviation measurements cannot go below zero. A value near zero implies less spread and a large value implies more spread.
* Meauring distance-to-middle is different than a uniform pattern in the data. For example, the following data is uniform, but has a great deal of spread as measured by distance-to-middle.



* Provide fictious data points for each of the following situations.

|  |  |
| --- | --- |
| Original Data |  |
| Five new data points with less spread than orginal data |  |
| Five new data points with a larger average, but has the same amount of spread as the original data |  |
| Five new data points with a larger average and less spread than orginal data. |  |

Getting Mean Absolute Deviation and Standard Deviation Values in Excel

Excel has formulas that will automatically compute the mean absolute deviation and the standard deviation for a set of numerical measurements. These are shown here.

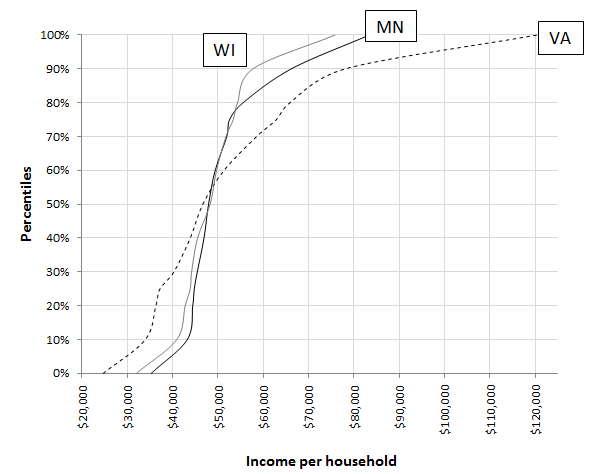
|  |  |
| --- | --- |
| Mean Absolute Deviation | Standard Deviation |

As previously mentioned, the standard deviation is more commonly used than mean absolute deviation. A standar deviation can be computed by PivotTables in Excel.

|  |  |
| --- | --- |
| Getting the standard deviation using PivotTables | Dotplots for each of these states |

Questions

1. Does the fact that these states have a different number of counites adversely affect our ability to make fair comparisons between these states? Explain.
2. How does the average income compare across these three states? Discuss.
3. How does the spread in income, i.e. income disparity, vary across these three states? Which state has the largest amount of income disparity? Discuss.
4. Consider once again the CDF plots for each of these states. Does the appearance of this plot agree with your answers for question 2 and 3 above? Dicsuss.

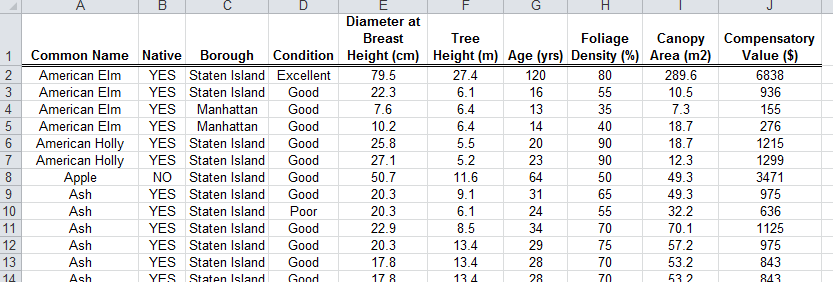


5.3: Shape in Data

Determining the shape of a data distribution is a very important step in many statistical procedures. For example, some procedures require that the data distribution be bell-shaped. Most often, graphical techniques are used to determine the shape of a distribution; however, a few numerical measures exist and will be discussed later.

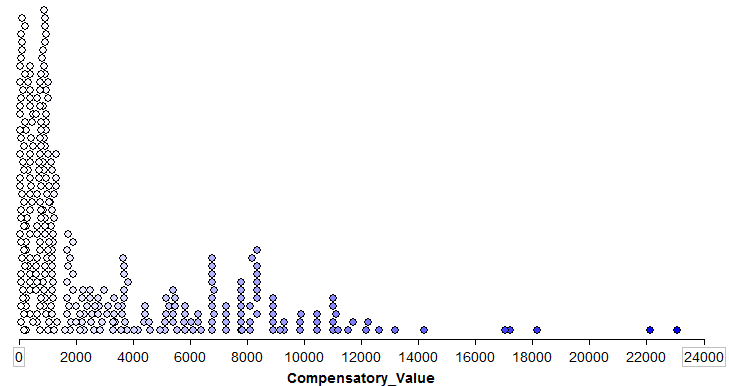
Graphical Summaries for Shape

**Example 5.3:** Consider the NYC\_Trees data. Recall, these data are a sample of tree characteristics from three different boroughs in New York City. A portion of the data set is shown below.



We will use this example to introduce several graphical techniques for numerical data.

* **Dot Plot**: This contains one axis for the numerical variable, and the plot uses dots to represent each measurement in the data set.



Questions:

* 1. Where are most trees located in terms of Compensatory Value?
  2. Notice that a few trees have a somewhat high Compensatory Value. Would you consider any of the trees on the upper end as extreme in terms of their Compensatory Value? That is, would you consider any of the trees on the upper-end of the distribution as potential outliers? Explain.
  3. Can you determine a cut-off for Compensatory Value so that trees above your cutoff would be considered potential outliers?
* **Boxplot**: This displays the quartiles of a data set.

The procedure for constructing a boxplot is as follows:

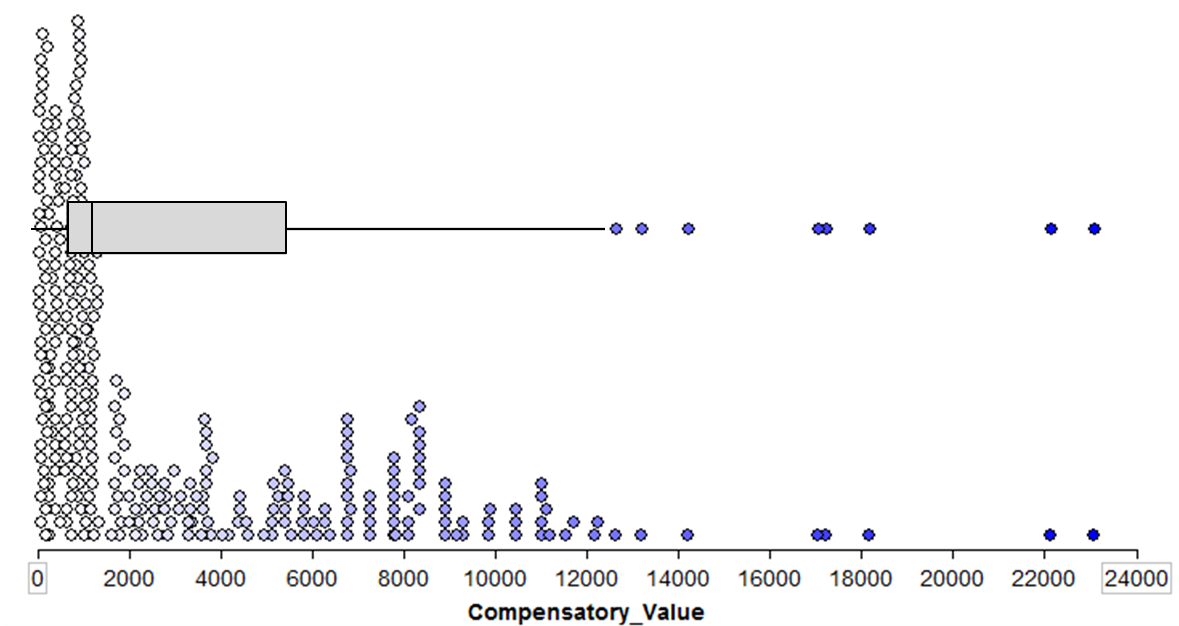
* Step 1: Draw vertical lines at Q1, Q2, and Q3. Enclose these lines in a box.
* Step 2: Find the lower and upper endpoints (or whiskers as they were originally called):

- The lower endpoint or whisker is the larger of the minimum and   
 (Q1 – 1.5\*IQR).

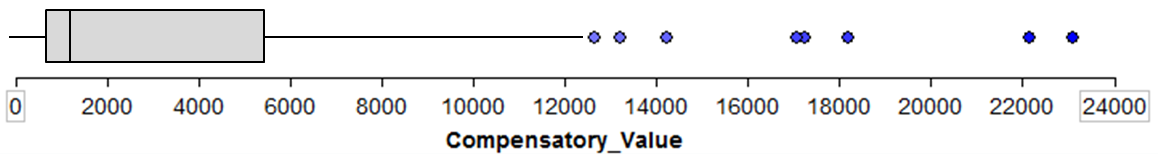
- The upper endpoint or whisker is the smaller of the maximum and   
 (Q3 + 1.5\*IQR).

Comment: Any measurement beyond the endpoint of either whisker is classified as a potential outlier (an extreme observation).

The boxplot shown with the dotplot.



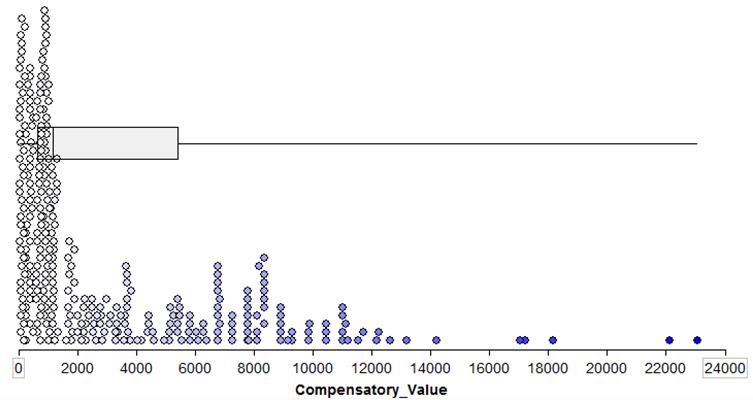
A boxplot is typically drawn without the dotplot as is shown here.



Questions:

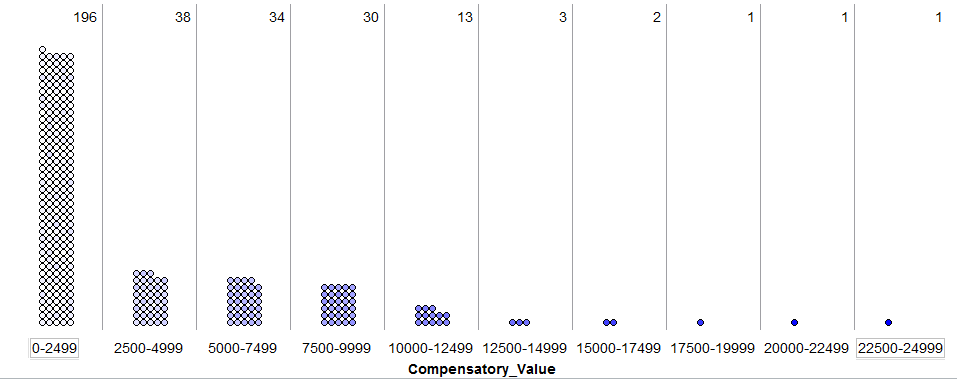
1. The median in this boxplot is shifted to the left. What does this imply about the Compensatory Value of trees in New York City?
2. Notice that Q1 and the endpoint of the lower whisker are very close. What does this imply about the Compensatory Value of trees in New York City?
3. What do you think the dots on this boxplot represent?
4. What value does the boxplot use to identify potential outliers? Does this agree with your outlier rule based on the dotplot?

Comment: Some textbooks and software packages, e.g. the default setting in Tinkerplots, use the minimum and maximum as the lower and upper endpoints. Statisticians do not do this. The reason is that boxplots are often used to identify potential outliers in a set of data and when the endpoints extend to the minimum and maximum, outliers may be masked.

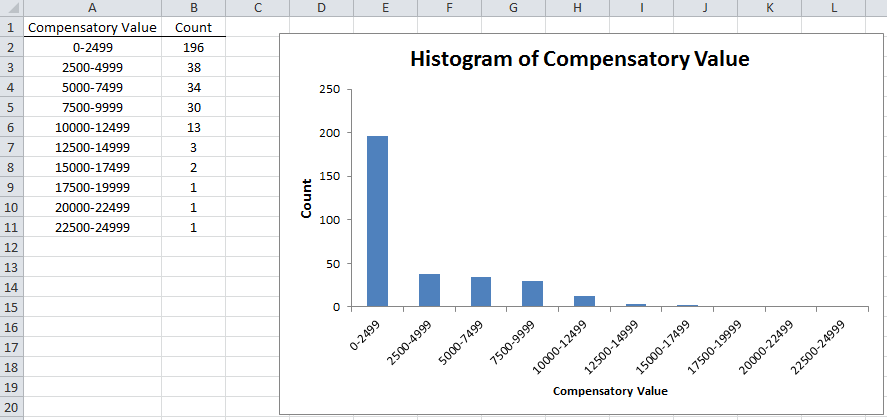


* **Histogram**: This divides the range of the variable into class intervals for which the frequency of occurrence is represented by a rectangular column. The height of the column is proportional to the frequency of observations within the interval.

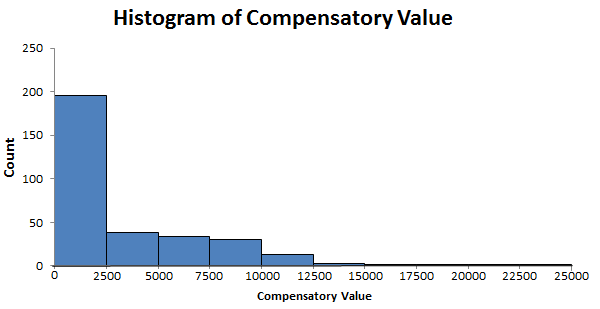
You can think of a histogram as a bar chart for numerical data. First, classes (or categories) are created. The histogram then displays the counts of the number of observations that fall in each class (the frequency). Just like with qualitative data, we can also create a relative frequency histogram by plotting the percentage of the observations that fall in each class. For example, we could start with our dotplot and divide our data into the following classes. Then, we can count the number of data points in each class.



The histogram would then be displayed as follows:



The gaps are often removed between the bars to emphasize the fact that the data is numeric and on a continuous scale.



Questions:

1. Does the information given by the histogram tend to agree with what you saw in the dotplot and boxplot? Explain.
2. Is it easier to identify outliers from the dotplot, boxplot , or histogram? Explain.

Comments:

* Excel can create histograms, but it takes some work. First, the easiest is to use the Data Analysis Toolpak (which is not available anymore for Mac’s).
* When making a histogram in Excel, you should specify the bins. If this is not done you end up with inappropriate divisions as is shown here.

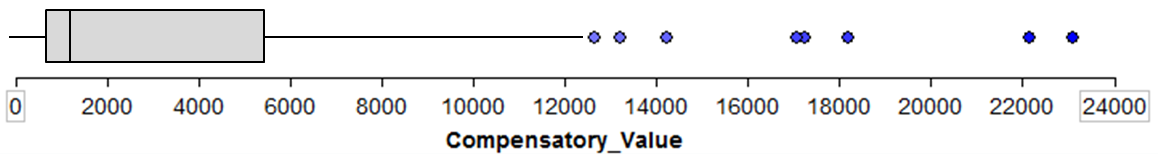


5.4: Outlier Detection

There are two commonly accepted methods of determining outliers in a set of numerical data.

* Boxplots
* Z Scores

**Boxplot**: Boxplot of the Compensatory Value of trees from the NYC Tree dataset.



The lower and upper whiskers are computed using

* The endpoint of the lower whisker is the larger of the minimum and (Q1 – 1.5\*IQR).
* The endpoint of the upper whisker is the smaller of the maximum and (Q3 + 1.5\*IQR).

**Z Scores**:

The standardized value, commonly called a Z-Score or Z-Value, is the most common method of determining outliers. The Z-Score has many uses in statistics: i) used as a measure of position, ii) used to determine if a data value is to be considered an outlier, and iii) used in the computation of a test statistic when completing a hypothesis test.

The standardized value for a particular data value is obtained using the following formula.



Comments:

• A Z-Score measures how many standard deviation a data point is from the mean.

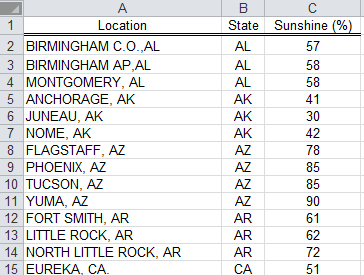
• Each value in a data set has a corresponding Z-Score.

• The smallest value in a data set will have the smallest Z-Score. The largest value will have the largest Z-Score.

• If a Z-Score is negative, then the data value is below the mean.

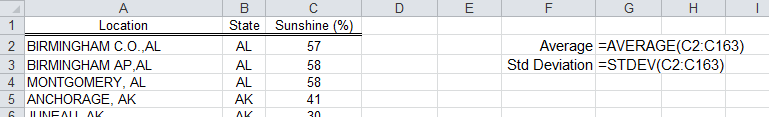
• If the Z-Score is positive, then the data value is above the mean.

• If the Z-Score is 0, then the data value is the same as the mean.

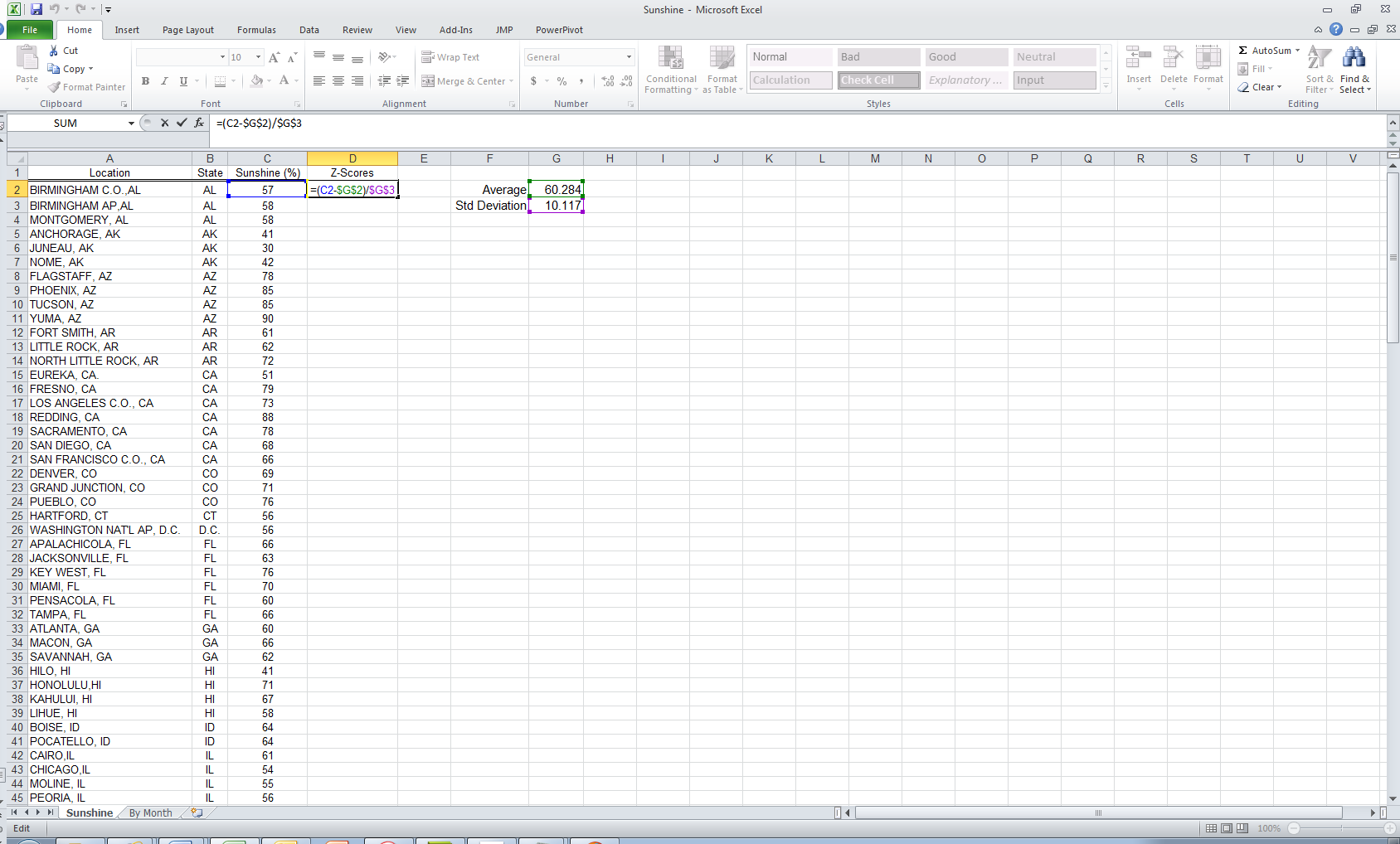


Getting Z-Scores in Excel…

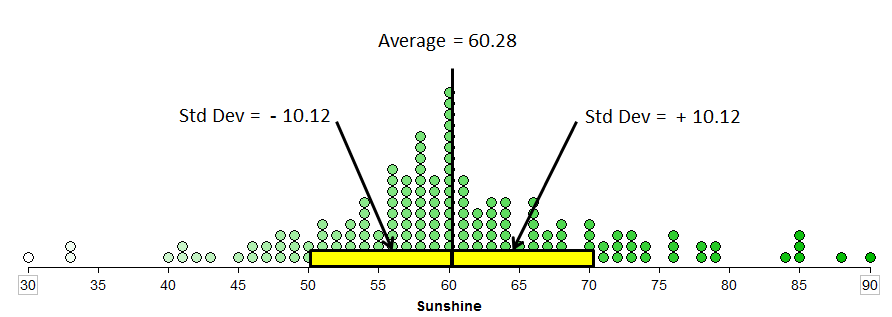
First, we need to get the average and standard deviation of your data.



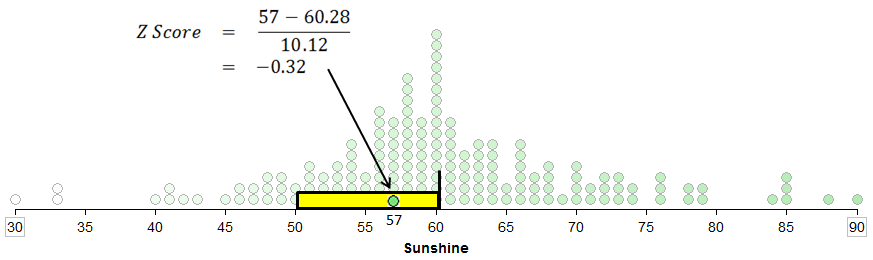
Next, type in the Z-Score formula in a column next to the data values. You should use absolute cell referencing for the Average and Std Deviation.



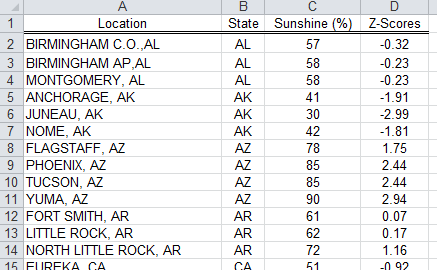
The following shows the data values that are within 1 standard deviation of the average.



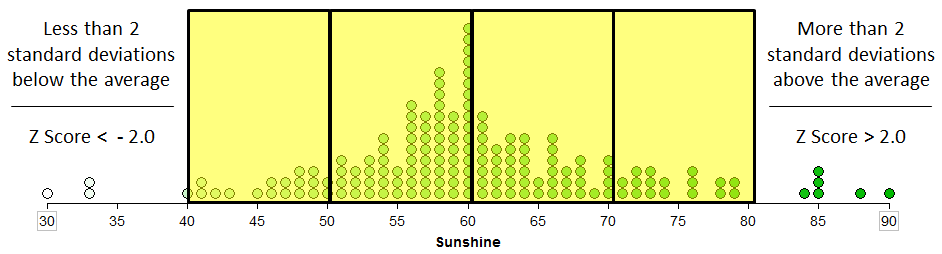
The first observations, Birmingham, C.O. AL is only slightly below the average. Thus, it’s Z-Score is negative and fairly close to zero.



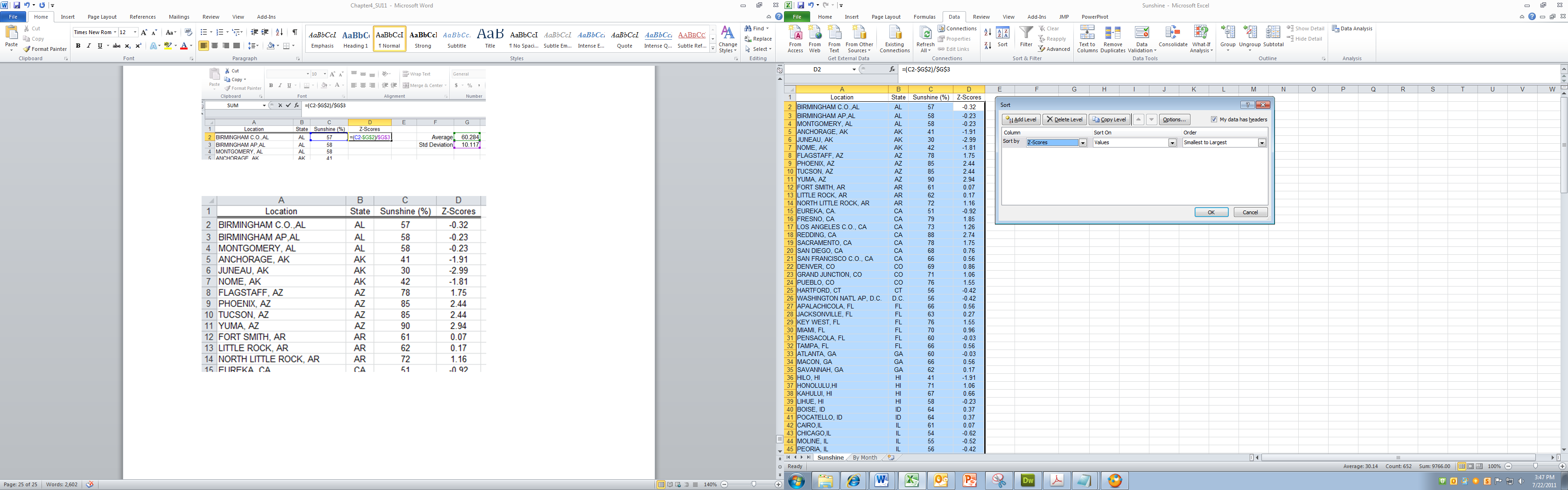
Z-Scores for the first few observations in this dataset.



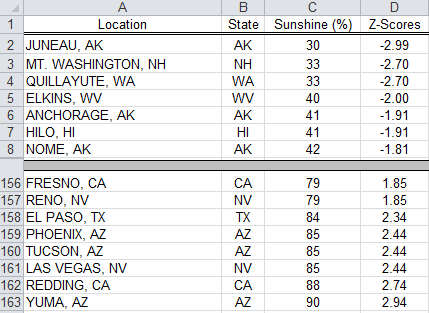
**Outlier Rule**: Any data values beyond 2 standard deviations of the average. A data value can be an outlier on the upper-side or lower-side. There appears to be three or four observations that would be classified as outliers on the lower-side and six on the upper-side.



In the following, the dataset is sorted with respect to the Z-Score values.



The smallest Z-Scores are listed first and the largest last here.



The boxplot and Z-Score approach across to identifying outliers agree for the most part.

