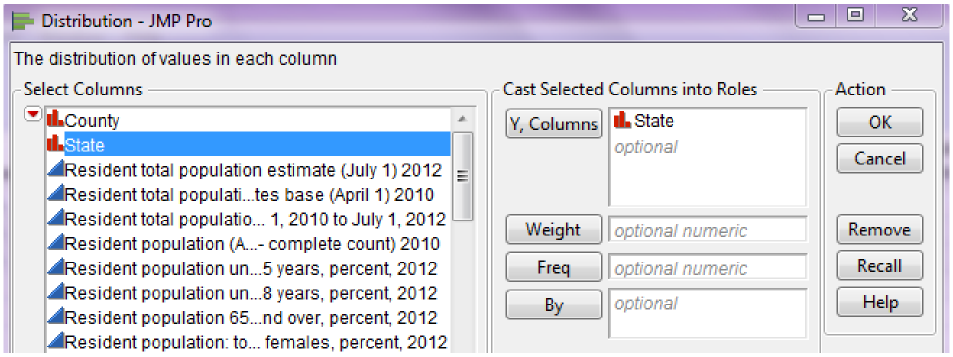
In this chapter, we will consider descriptive methods appropriate for summarizing numerical variables.

Example 4.1: Summarizing Household Income Levels in Minnesota Counties

The data in the file **USQuickFacts.jmp** was recorded by the U.S. Census Bureau (*Source: http://quickfacts.census.gov*). This file includes information on several demographic variables for all counties in the U.S. In this example, we will consider only the counties of Minnesota.

To subset the Minnesota counties in JMP, we first must select Analyze > Distribution and enter the following:

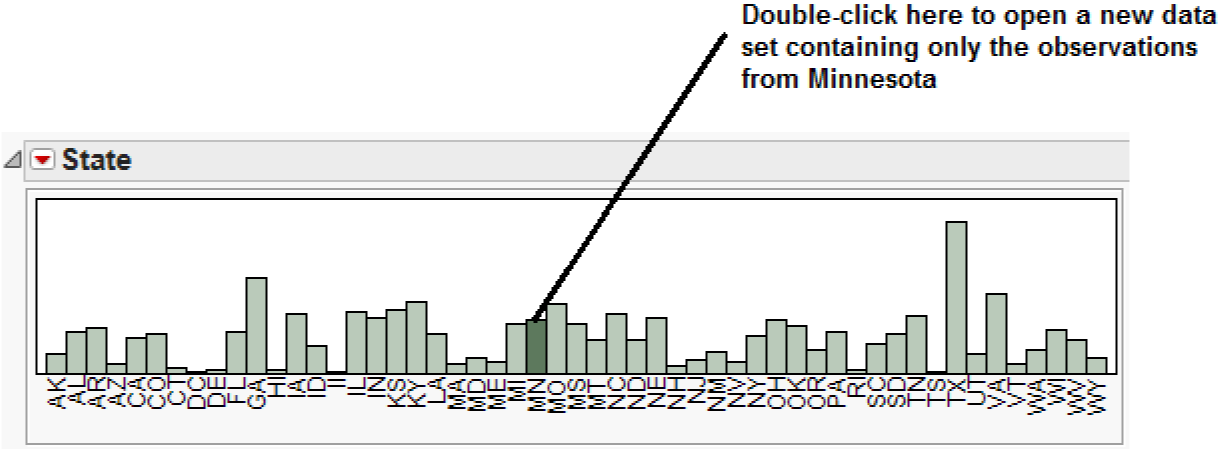

Distribution - JMP Pro

The distribution of values in each column

Select Columns
- County
- State
- Resident total population estimate (July 1) 2012
- Resident total populati...tes base (April 1) 2010
- Resident total populatio... 1, 2010 to July 1, 2012
- Resident population (A...- complete count) 2010
- Resident population un...5 years, percent, 2012
- Resident population un...8 years, percent, 2012
- Resident population 65...nd over, percent, 2012
- Resident population: to... females, percent, 2012

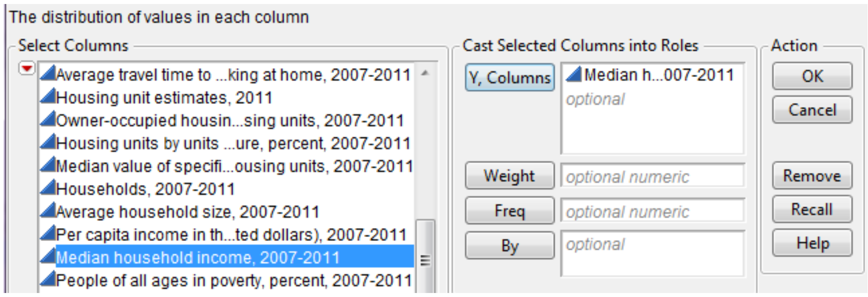
Cast Selected Columns into Roles
- Y, Columns: State
- Weight: optional numeric
- Freq: optional numeric
- By: optional

Next, on the resulting bar chart, double-click on the bar for Minnesota.


Double-click here to open a new data set containing only the observations from Minnesota

A new data table should open in JMP which contains only the 87 counties in Minnesota. This is the data file we will consider in this example. We will begin by examining the distribution of *Median Household Income, 2007-2011.*

To do this, once again select Analyze > Distribution. Place the variable of interest, *Median Household Income, 2007-2011,* in the Y, Columns box as shown below.



The distribution of values in each column

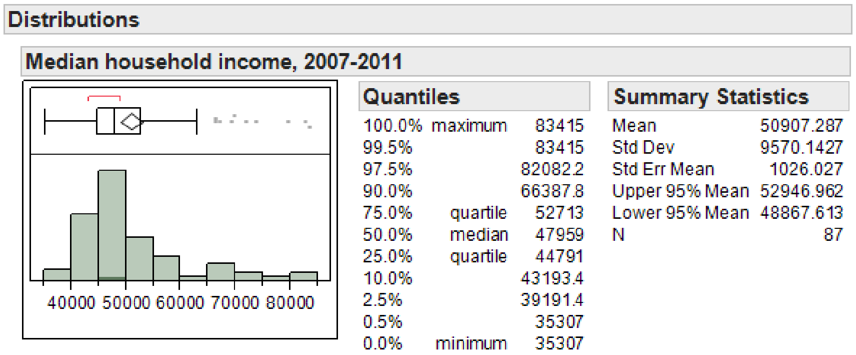
Select Columns

- Average travel time to ...king at home, 2007-2011
- Housing unit estimates, 2011
- Owner-occupied housin...sing units, 2007-2011
- Housing units by units ...ure, percent, 2007-2011
- Median value of specifi...ousing units, 2007-2011
- Households, 2007-2011
- Average household size, 2007-2011
- Per capita income in th...ted dollars), 2007-2011
- Median household income, 2007-2011
- People of all ages in poverty, percent, 2007-2011

Cast Selected Columns into Roles

Y, Columns: Median h...007-2011 / optional
Weight: optional numeric
Freq: optional numeric
By: optional

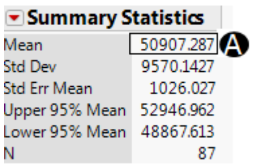
JMP returns the following output by default:



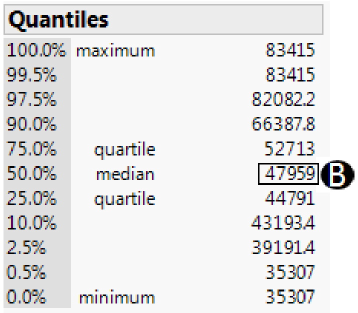
```
Distributions
Median household income, 2007-2011

Quantiles                         Summary Statistics
100.0%  maximum    83415          Mean             50907.287
99.5%              83415          Std Dev          9570.1427
97.5%            82082.2          Std Err Mean      1026.027
90.0%            66387.8          Upper 95% Mean   52946.962
75.0%   quartile   52713          Lower 95% Mean   48867.613
50.0%   median     47959          N                       87
25.0%   quartile   44791
10.0%            43193.4
2.5%             39191.4
0.5%               35307
0.0%    minimum    35307
```

Next, we'll discuss some of the summary statistics provided in each piece of the output.

Summary Statistics Representing a "Typical Value" in a Data Set



```
Summary Statistics
Mean            50907.287 (A)
Std Dev         9570.1427
Std Err Mean     1026.027
Upper 95% Mean  52946.962
Lower 95% Mean  48867.613
N                      87
```



```
Quantiles
100.0%  maximum    83415
99.5%              83415
97.5%            82082.2
90.0%            66387.8
75.0%   quartile   52713
50.0%   median     47959 (B)
25.0%   quartile   44791
10.0%            43193.4
2.5%             39191.4
0.5%               35307
0.0%    minimum    35307
```

(A) **Mean**: The arithmetic average of all of the values in a data set.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(B) **Median**: The middle term of a data set (after the numerical values have been ordered). If the data set contains an even number of observations, then the median is the average of the middle two observations.

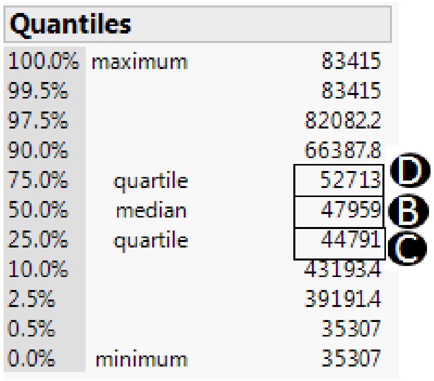
Questions:

1. Is it necessary that the median be an actual measurement from the data set? Explain.

2. Is it necessary that the mean be an actual measurement from the data set? Explain.

3. Suppose the maximum value in this data set was changed from $83,415 to $100,000. What impact would this have on the mean? What impact would this have on the median?

Other Summary Statistics for Describing the "Location" of a Data Set



Quantiles

100.0%	maximum	83415	
99.5%		83415	
97.5%		82082.2	
90.0%		66387.8	
75.0%	quartile	52713	D
50.0%	median	47959	B
25.0%	quartile	44791	C
10.0%		43193.4	
2.5%		39191.4	
0.5%		35307	
0.0%	minimum	35307	

Percentiles (JMP calls these Quantiles) give us insight into the entire spectrum of our data set. 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.

Quartiles are special cases of percentiles.

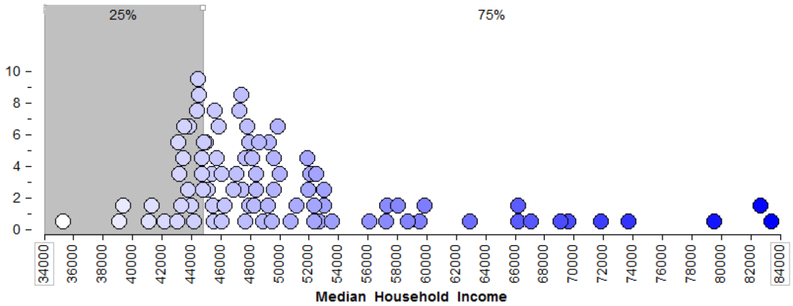
C Q_1 – The median of the lower half of the data (i.e., the 25th percentile)

B Q_2 – The median (i.e., the 50th percentile)

D Q_3 – The median of the upper half of the data (i.e., the 75th percentile)

Questions:

1. What percent of counties in Minnesota have a median household income of $44,791 or less?



2. What percent of counties in Minnesota have a median household income of $52,713 or less?

3. What percent of counties in Minnesota have a median household income above $52,713?

4. The 2.5th percentile is about ______________ and the 97.5th percentile is _______________. What proportion of Minnesota counties has a median household income between these two values?

MEASURES OF LOCATION

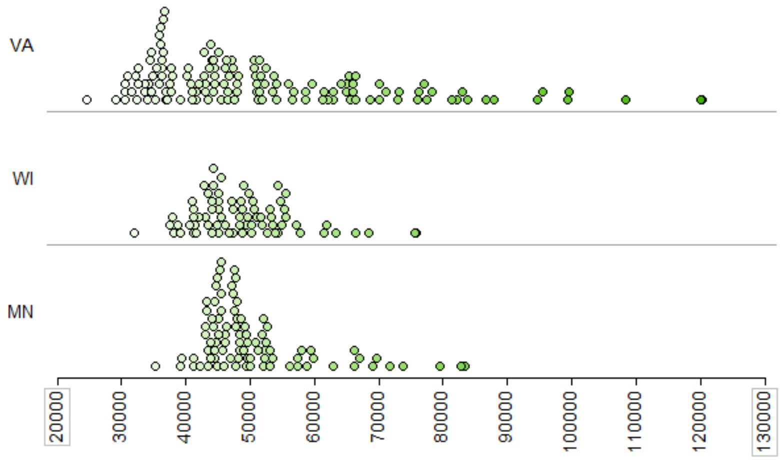
The summary statistics discussed above are often all referred to as measures of location. These summaries give us an idea of where a data distribution lies. In particular, the mean and median give us an idea of the *center* (or middle) of the distribution (this summarizes the data set with a single value representing a "typical" value in the data set).

The percentiles (called quantiles in JMP) give us an idea of what percent of the data distribution lies at or below a particular value.

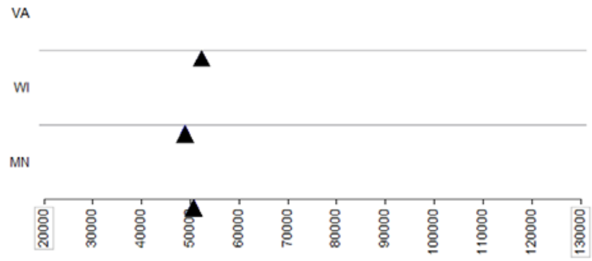
What summary (or summaries) we choose to describe the entire data set depends on our objective. If the goal is to describe a typical value in the data set, then the mean or median may be an appropriate summary statistic. However, if interest lies in what value is exceeded by only 5% of the data distribution, for example, then we would use the 95th percentile.

MEASURES OF VARIABILITY

Sometimes a measure of "center" does not adequately tell a data set's story. For example, consider the median household income across counties for three different states (Minnesota, Wisconsin, and Virginia):



The following picture shows the average for each state.

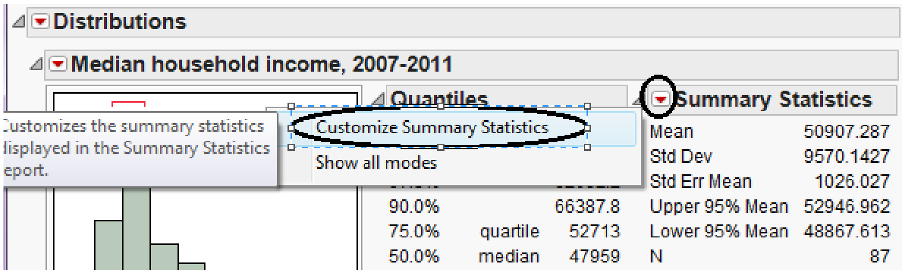


Questions:

1. What differences exist in the *Typical Household Income* values across these three states? Discuss.

2. Suppose that your friend tries to summarize the differences across these three states using only the mean (i.e., average) from each state. Do you think that this single summary (the mean) tells the whole story well? Why or why not?

To adequately describe a data set, we must also describe the amount of variability present in that data set. Some of these measures are described below. Also, some of these summary statistics don't appear by default in JMP. To see them, click on the red drop-down arrow next to "Summary Statistics" and choose "Customize Summary Statistics" as shown below.

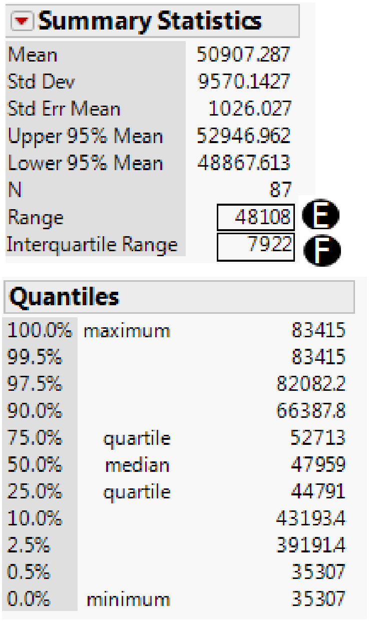

Distributions
Median household income, 2007-2011
Quantiles
Summary Statistics
Customize Summary Statistics
Show all modes
ustomizes the summary statistics
isplayed in the Summary Statistics
eport.

90.0%		66387.8
75.0%	quartile	52713
50.0%	median	47959

Mean	50907.287
Std Dev	9570.1427
Std Err Mean	1026.027
Upper 95% Mean	52946.962
Lower 95% Mean	48867.613
N	87

To get the following output, I requested that JMP display both the **Range** and the **Interquartile Range**.

Summary Statistics for Describing the "Variability" of a Data Set



Summary Statistics

Mean	50907.287
Std Dev	9570.1427
Std Err Mean	1026.027
Upper 95% Mean	52946.962
Lower 95% Mean	48867.613
N	87
Range	48108 E
Interquartile Range	7922 F

Quantiles

100.0%	maximum	83415
99.5%		83415
97.5%		82082.2
90.0%		66387.8
75.0%	quartile	52713
50.0%	median	47959
25.0%	quartile	44791
10.0%		43193.4
2.5%		39191.4
0.5%		35307
0.0%	minimum	35307

E **Range**: The difference between the largest value and the smallest value in a data set.

Range = Maximum – Minimum

F **Interquartile Range (IQR)**: The IQR is computed as the difference between the first and third quartiles.

IQR = $Q_3 - Q_1$

Questions:

1. How many observations from the data set are used in the computation of the range?

2. Outliers (which we will discuss later) are extreme observations which need to be handled with care in an analysis. How will outliers affect the range?

3. What is the smallest possible value for the range? What does it mean if the range is at this value?

4. What percent of the data lies between Q_1 and Q_3?

5. Which is more affected by outliers: the range or the IQR? Explain.

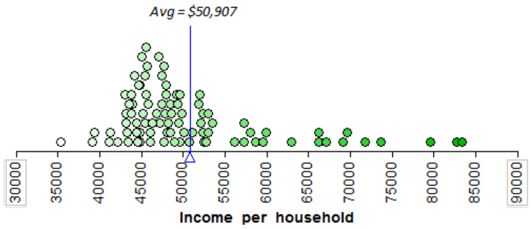
Some statisticians advocate that one of the most intuitive ways to measure variability is to consider what's called the **Mean Absolute Deviation**.

Definition

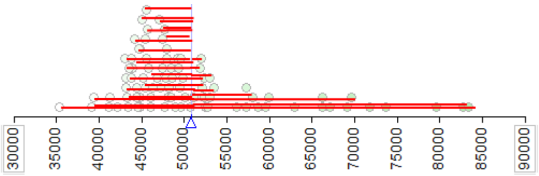
Mean Absolute Deviation: For each measurement, calculate how far away that measurement is from the mean of the data set. The mean absolute deviation is the average of these absolute distances.

$$\text{MAD} = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n}$$

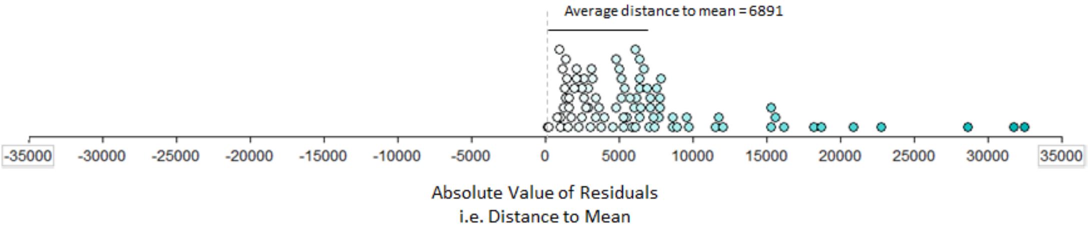
To see how this is calculated, first consider the mean for Minnesota:



Then, consider the distance between each measurement and the mean:



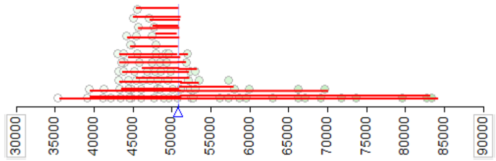
Next, we consider the length of each of these distances (i.e., the absolute value of each of the distances) on the following plot. The average of these lengths is the **mean absolute deviation**.



Question: Why is it important that we take the absolute value of the distances before finding their average?

JMP doesn't compute this measure of variability, which tells you how often it's used in practice. This concept is worth considering, however, because it helps you understand how both the **variance** and the **standard deviation** are computed (these measures are commonly used to describe the amount of variability in a data set).

For example, like the mean absolute deviation, the variance also measures how far a "typical" observation is from the mean.



Instead of working with the *absolute value* of these distances, however, the variance works with the *squared* distances (note that this still prevents positive distances from cancelling out negative distances).

Definition

Variance: For each data point, calculate how far away that value is from the mean of the data set. Then, square each of these distances and add them up. Finally, divide by n-1.

$$\text{Variance: } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Note: We divide by n –1 because dividing by n tends to underestimate the true population variance. So, our estimate of the variance is better when we divide by n-1.

Finally, because the original distances were squared, the variance is in terms of squared units. To get back into the original scale of our data set, we take the square root of the variance.

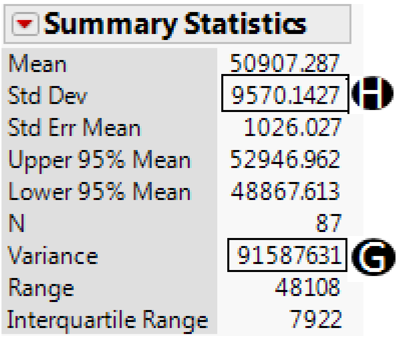
Definition

Standard Deviation: The square root of the variance.

$$\text{Standard deviation: } s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

The following output shows both the variance and the standard deviation.

Summary Statistics for Describing the "Variability" of a Data Set



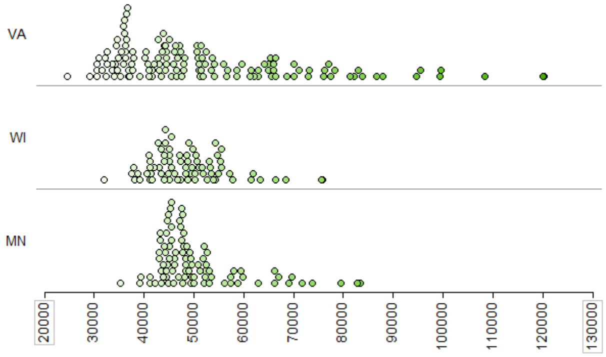
Summary Statistics

Mean	50907.287	
Std Dev	$\boxed{9570.1427}$	Ⓗ
Std Err Mean	1026.027	
Upper 95% Mean	52946.962	
Lower 95% Mean	48867.613	
N	87	
Variance	$\boxed{91587631}$	Ⓖ
Range	48108	
Interquartile Range	7922	

Ⓖ **Variance**: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

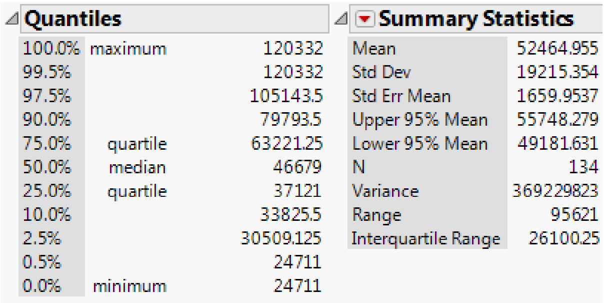
Ⓗ **Standard deviation**: $s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Next, reconsider the following three states.



JMP was used to obtained summary statistics for all three:

Virginia

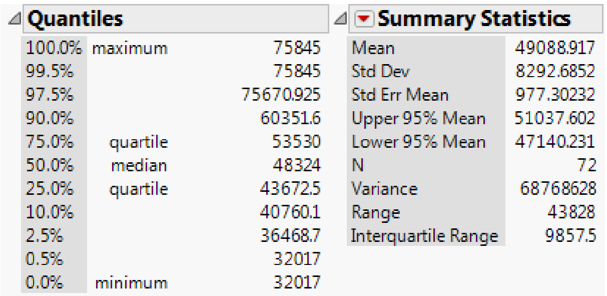
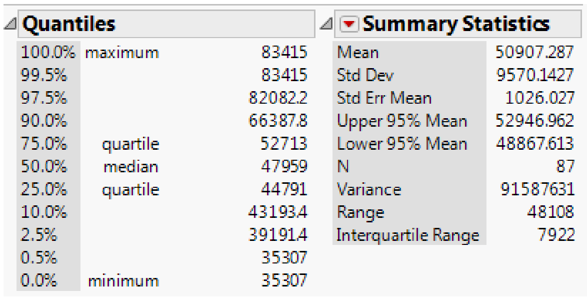


Quantiles

100.0%	maximum	120332
99.5%		120332
97.5%		105143.5
90.0%		79793.5
75.0%	quartile	63221.25
50.0%	median	46679
25.0%	quartile	37121
10.0%		33825.5
2.5%		30509.125
0.5%		24711
0.0%	minimum	24711

Summary Statistics

Mean	52464.955
Std Dev	19215.354
Std Err Mean	1659.9537
Upper 95% Mean	55748.279
Lower 95% Mean	49181.631
N	134
Variance	369229823
Range	95621
Interquartile Range	26100.25

<table>
<tr>
<td><b>Wisconsin</b></td>
<td>

<table>
<tr><th colspan="3">Quantiles</th></tr>
<tr><td>100.0%</td><td>maximum</td><td>75845</td></tr>
<tr><td>99.5%</td><td></td><td>75845</td></tr>
<tr><td>97.5%</td><td></td><td>75670.925</td></tr>
<tr><td>90.0%</td><td></td><td>60351.6</td></tr>
<tr><td>75.0%</td><td>quartile</td><td>53530</td></tr>
<tr><td>50.0%</td><td>median</td><td>48324</td></tr>
<tr><td>25.0%</td><td>quartile</td><td>43672.5</td></tr>
<tr><td>10.0%</td><td></td><td>40760.1</td></tr>
<tr><td>2.5%</td><td></td><td>36468.7</td></tr>
<tr><td>0.5%</td><td></td><td>32017</td></tr>
<tr><td>0.0%</td><td>minimum</td><td>32017</td></tr>
</table>
<table>
<tr><th colspan="2">Summary Statistics</th></tr>
<tr><td>Mean</td><td>49088.917</td></tr>
<tr><td>Std Dev</td><td>8292.6852</td></tr>
<tr><td>Std Err Mean</td><td>977.30232</td></tr>
<tr><td>Upper 95% Mean</td><td>51037.602</td></tr>
<tr><td>Lower 95% Mean</td><td>47140.231</td></tr>
<tr><td>N</td><td>72</td></tr>
<tr><td>Variance</td><td>68768628</td></tr>
<tr><td>Range</td><td>43828</td></tr>
<tr><td>Interquartile Range</td><td>9857.5</td></tr>
</table>
</td>
</tr>
<tr>
<td><b>Minnesota</b></td>
<td>

<table>
<tr><th colspan="3">Quantiles</th></tr>
<tr><td>100.0%</td><td>maximum</td><td>83415</td></tr>
<tr><td>99.5%</td><td></td><td>83415</td></tr>
<tr><td>97.5%</td><td></td><td>82082.2</td></tr>
<tr><td>90.0%</td><td></td><td>66387.8</td></tr>
<tr><td>75.0%</td><td>quartile</td><td>52713</td></tr>
<tr><td>50.0%</td><td>median</td><td>47959</td></tr>
<tr><td>25.0%</td><td>quartile</td><td>44791</td></tr>
<tr><td>10.0%</td><td></td><td>43193.4</td></tr>
<tr><td>2.5%</td><td></td><td>39191.4</td></tr>
<tr><td>0.5%</td><td></td><td>35307</td></tr>
<tr><td>0.0%</td><td>minimum</td><td>35307</td></tr>
</table>
<table>
<tr><th colspan="2">Summary Statistics</th></tr>
<tr><td>Mean</td><td>50907.287</td></tr>
<tr><td>Std Dev</td><td>9570.1427</td></tr>
<tr><td>Std Err Mean</td><td>1026.027</td></tr>
<tr><td>Upper 95% Mean</td><td>52946.962</td></tr>
<tr><td>Lower 95% Mean</td><td>48867.613</td></tr>
<tr><td>N</td><td>87</td></tr>
<tr><td>Variance</td><td>91587631</td></tr>
<tr><td>Range</td><td>48108</td></tr>
<tr><td>Interquartile Range</td><td>7922</td></tr>
</table>
</td>
</tr>
</table>

Questions:

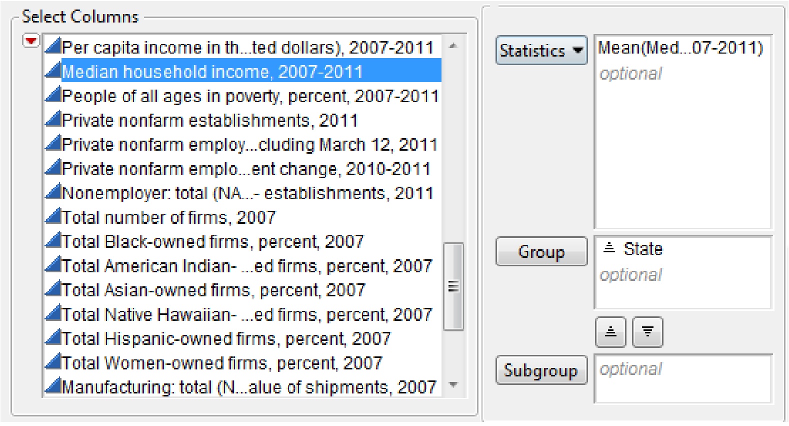
1. Compare each of the aforementioned statistics appropriate for summarizing variability across the three states.

2. Which state has the *most* variability in its county-level median household incomes? The *least*? How did you decide this?

Example 4.2: Comparing Household Income Levels Across States

Next, suppose we want to determine which states tend to have the highest household income levels. To do this, we will go back to our original data set, **QuickFacts.jmp**. Recall that the data were collected at the county level, so we will begin by summarizing the household incomes for each state.

For example, we could use JMP to find the mean of *Median household income* for each state as follows. Select Tables > Summary and enter the following:



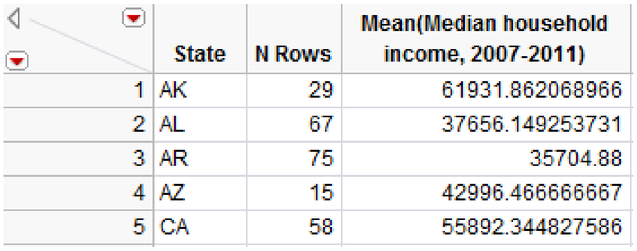
Select Columns
Per capita income in th...ted dollars), 2007-2011
Median household income, 2007-2011
People of all ages in poverty, percent, 2007-2011
Private nonfarm establishments, 2011
Private nonfarm employ...cluding March 12, 2011
Private nonfarm emplo...ent change, 2010-2011
Nonemployer: total (NA...- establishments, 2011
Total number of firms, 2007
Total Black-owned firms, percent, 2007
Total American Indian- ...ed firms, percent, 2007
Total Asian-owned firms, percent, 2007
Total Native Hawaiian- ...ed firms, percent, 2007
Total Hispanic-owned firms, percent, 2007
Total Women-owned firms, percent, 2007
Manufacturing: total (N...alue of shipments, 2007

Statistics: Mean(Med...07-2011)
optional

Group: State
optional

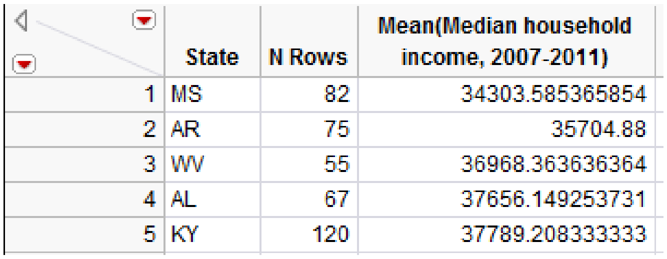
Subgroup: optional

JMP will then open a new data table of which a portion is shown below:



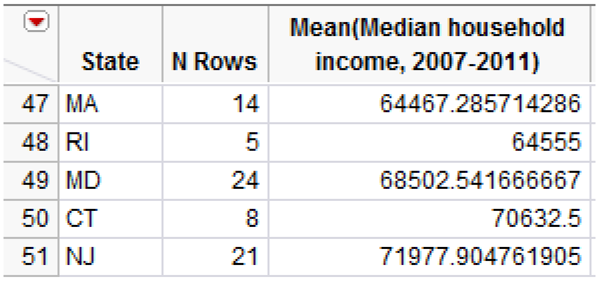
	State	N Rows	Mean(Median household income, 2007-2011)
1	AK	29	61931.862068966
2	AL	67	37656.149253731
3	AR	75	35704.88
4	AZ	15	42996.466666667
5	CA	58	55892.344827586

This will allow us to make comparisons across states. Note that you can select Tables > Sort to sort the results in numerical order. After sorting, we see that the five states with the lowest household incomes as measured by the mean are shown below:



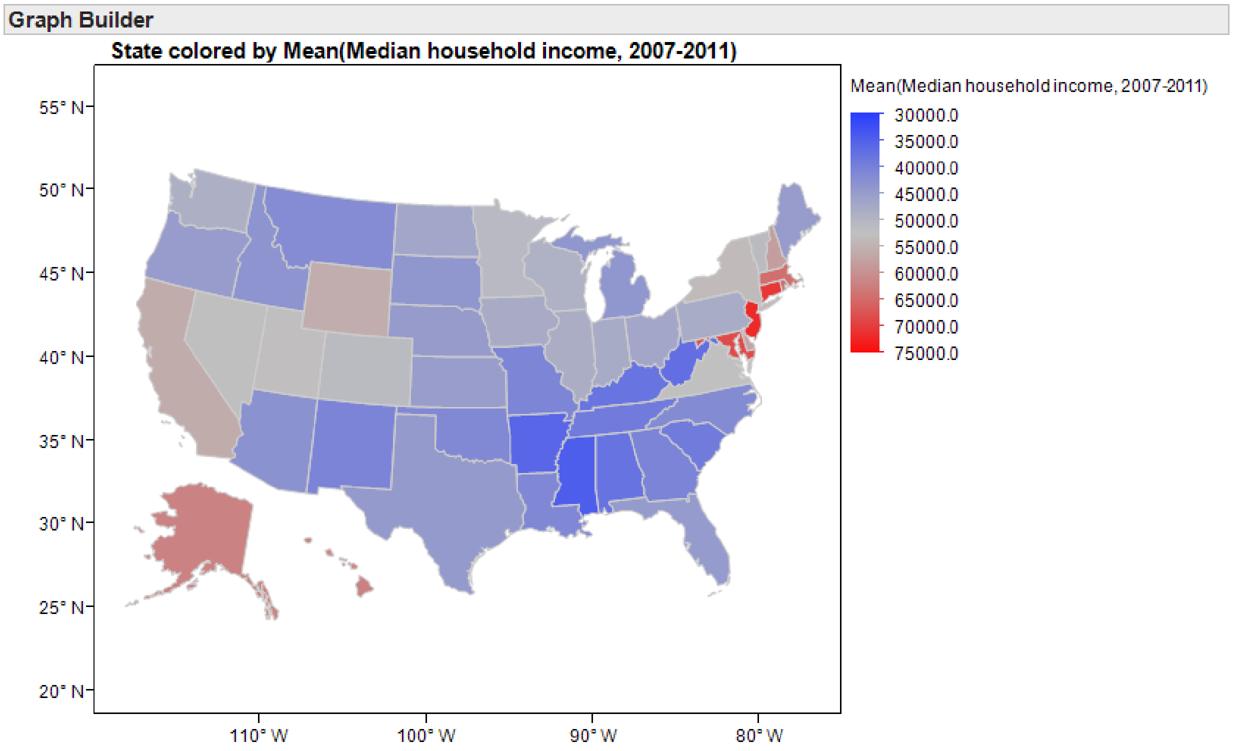
	State	N Rows	Mean(Median household income, 2007-2011)
1	MS	82	34303.585365854
2	AR	75	35704.88
3	WV	55	36968.363636364
4	AL	67	37656.149253731
5	KY	120	37789.208333333

The top five states are shown here:



	State	N Rows	Mean(Median household income, 2007-2011)
47	MA	14	64467.285714286
48	RI	5	64555
49	MD	24	68502.541666667
50	CT	8	70632.5
51	NJ	21	71977.904761905

Instead of sorting the data, you could also use Graph Builder to create graphs such as the following. This makes a visual comparison of household incomes across states much easier.



Note that you could create similar maps with other summary statistics such as the median or percentiles of the household income data, as well.

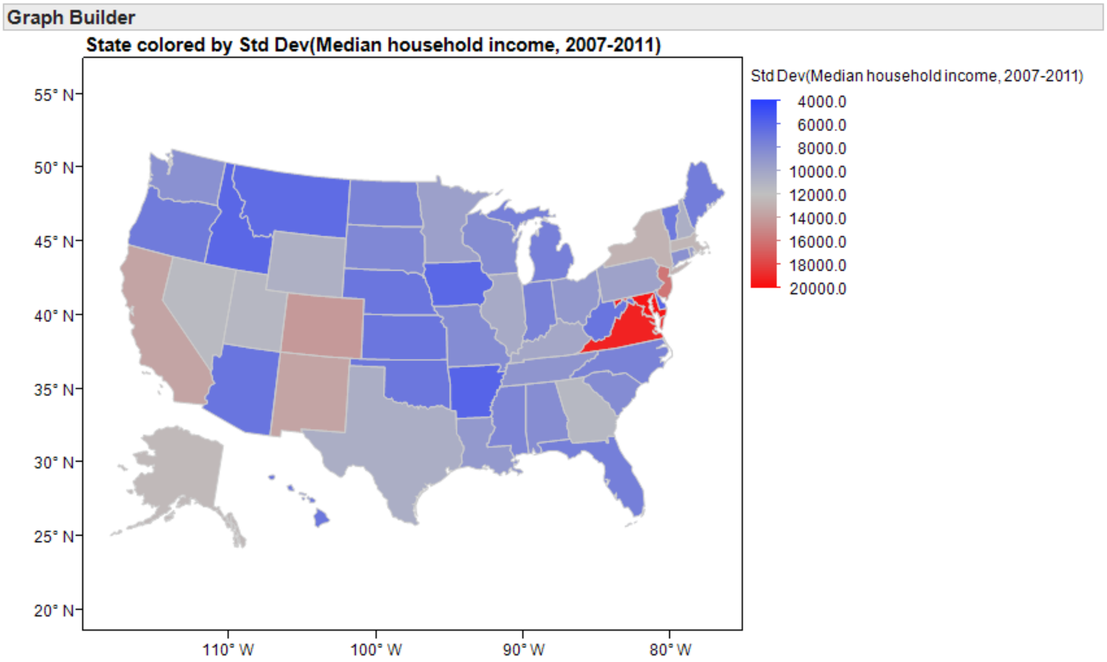
Next, suppose that instead of identifying the highest-earning states, you were more interested in which states have the most problems with *income inequality*.

Questions:

1. The map from the previous page will *not* help you investigate this question. Why not?

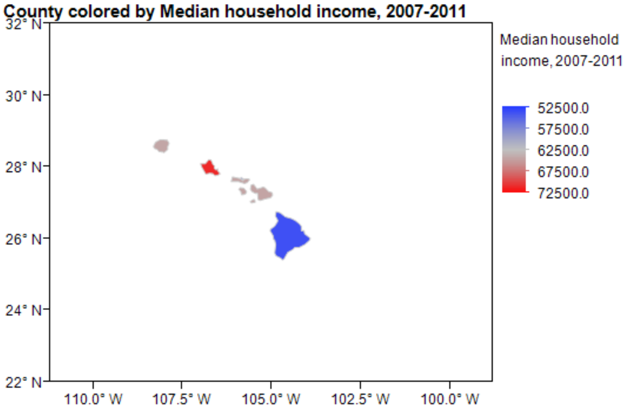
2. What summary statistics *will* help you investigate income inequality within each state? Explain your reasoning.

3. Use the following map to identify a few states that appear to have the most problems with income inequality and a few states that appear to have the least problems with income inequality.

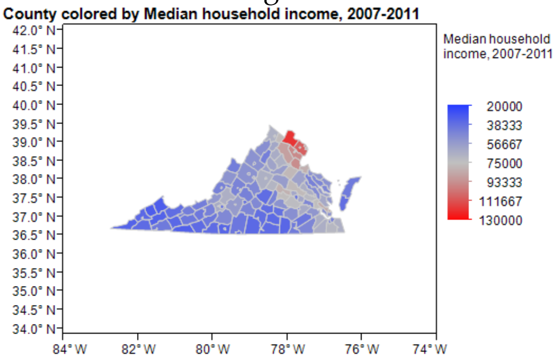


Another look at the household income levels by county in a few individual states:

Hawaii



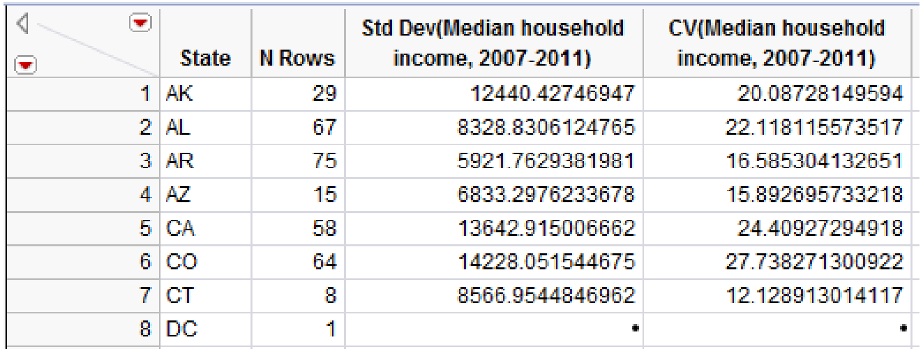
Virginia



Questions:

4. Suppose a researcher claims there will obviously be less variability in incomes in Hawaii than in Virginia simply because Hawaii has only 5 counties versus Virginia's 134 counties. In other words, he or she is arguing that the measures of variability are smaller for Hawaii simply because we have a smaller sample size in that state. Why is this reasoning *incorrect*? What is the real reason there is less variability in Hawaii?

5. Note that when I used JMP to compute the standard deviation for each state to create the map on page 124, JMP did not report this summary statistic for the District of Columbia.



	State	N Rows	Std Dev(Median household income, 2007-2011)	CV(Median household income, 2007-2011)
1	AK	29	12440.42746947	20.08728149594
2	AL	67	8328.8306124765	22.118115573517
3	AR	75	5921.7629381981	16.585304132651
4	AZ	15	6833.2976233678	15.892695733218
5	CA	58	13642.915006662	24.40927294918
6	CO	64	14228.051544675	27.738271300922
7	CT	8	8566.9544846962	12.128913014117
8	DC	1	•	•

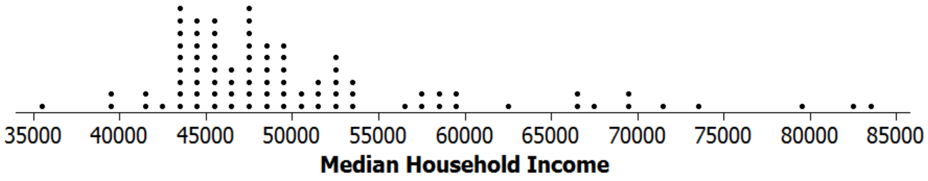
Why did this happen?

GRAPHICAL SUMMARIES OF NUMERICAL DATA

In this section, we will discuss common methods for graphing numerical data. Graphs conveniently allow us to examine both the location and the variability in a data set. Moreover, we gain insight into the *shape* of a data distribution.

Example 4.3: Graphing Household Income Levels in Minnesota Counties

In this example, we will once again consider summarizing the median household incomes of all 87 counties in Minnesota. Recall that we've already seen a dotplot of these data:

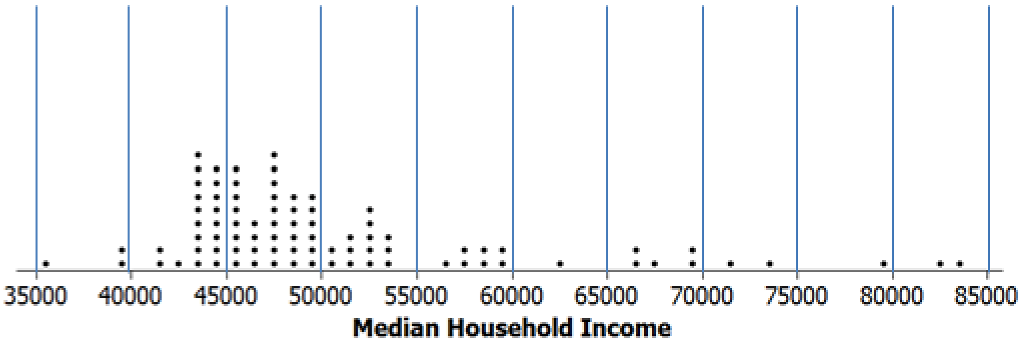


JMP does not create dotplots, but JMP can be used to create the following graphics.

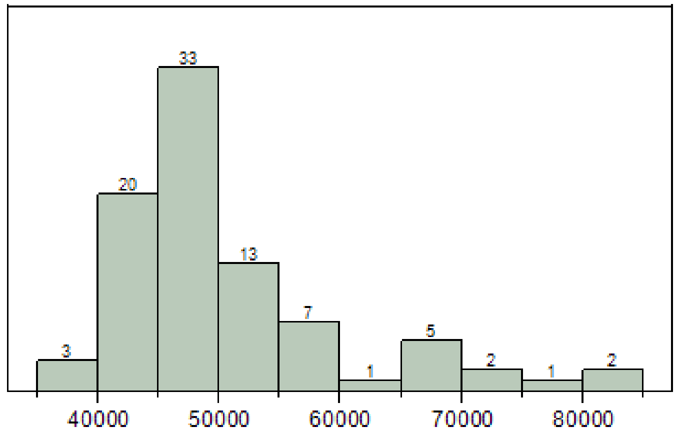
Histogram

A histogram is created by dividing the range of the data distribution into class intervals and then counting the number of observations that fall in each interval. A rectangular column is plotted in each interval, and the height of the column is proportional to the frequency of observations within the interval. The y-axis can be labeled with either the count or the percentage of the observations that fall in each interval.

To see this, note that we could start with our dotplot and divide the data distribution into the following classes. Then, we can count the number of data points in each interval.

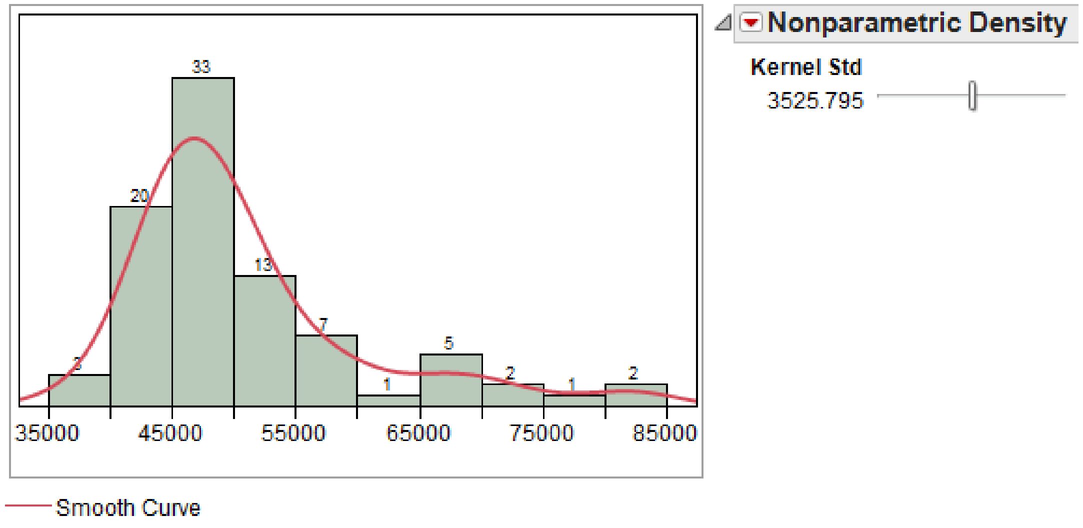


When you use the Analyze > Distribution platform, JMP provides the histogram by default. You can select “Histogram Options > Show counts” to display the counts for each interval.



Density Smoother

You can use JMP to overlay a density smoother on the histogram. To do this, select “Continuous Fit > Smooth Curve” from the red drop-down arrow next to the variable name. You can control the amount of smoothing with the slider bar.



Nonparametric Density

Kernel Std
3525.795

This curve ignores some of the minor irregularities that may appear in the histogram and provides us with a smooth estimate of the real trends in the data.

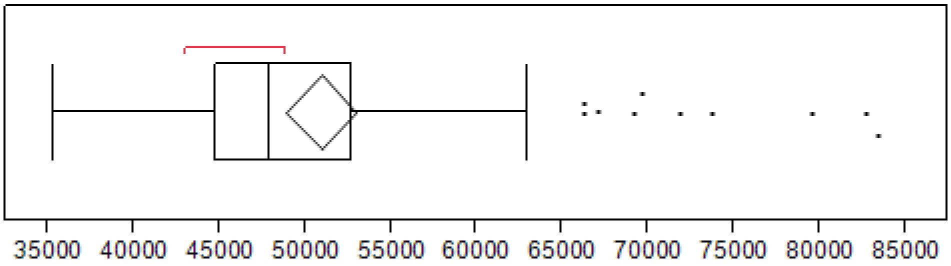
Boxplot

The procedure for constructing a boxplot is as follows:

1. Draw horizontal lines at Q_1, Q_2, and Q_3. Enclose these horizontal lines in a box.
2. Find the lower and upper whiskers:
   - The endpoint of the lower whisker is the larger of the minimum and (Q_1 – 1.5*IQR).
   - The endpoint of the upper whisker is the smaller of the maximum and (Q_3 + 1.5*IQR).

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

When you use the Analyze > Distribution platform in JMP, the boxplot appears by default. You can open and close the boxplot by clicking on the red drop-down menu next to the variable name and selecting "Outlier Box Plot."



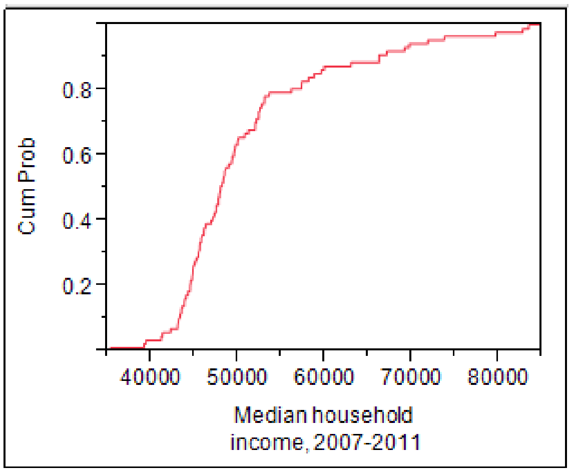
Questions:

1. Are there any counties in Minnesota in which the median household income is unusually low? If so, which ones?

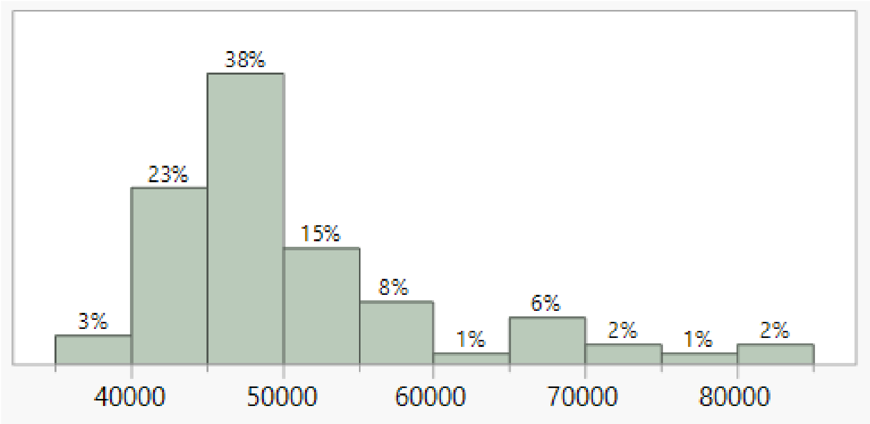
2. Are there any counties in Minnesota in which the median household income is unusually high? If so, which ones?

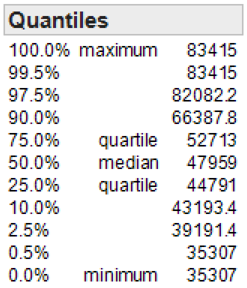
CDF Plot

To create a plot of the cumulative distribution function (i.e., a CDF plot) in JMP, select this option from the red drop-down arrow next to the variable name.



This plot allows you to easily determine the percentage of the data that falls at or below a given value on the x-axis. To see this, consider both a histogram of the data and the percentiles which were discussed earlier.





Quantiles

100.0%	maximum	83415
99.5%		83415
97.5%		82082.2
90.0%		66387.8
75.0%	quartile	52713
50.0%	median	47959
25.0%	quartile	44791
10.0%		43193.4
2.5%		39191.4
0.5%		35307
0.0%	minimum	35307

Questions:

1. What percentage of counties in Minnesota have a median household income level below $50,000?

2. What percentage of counties in Minnesota have a median household income level above $60,000?

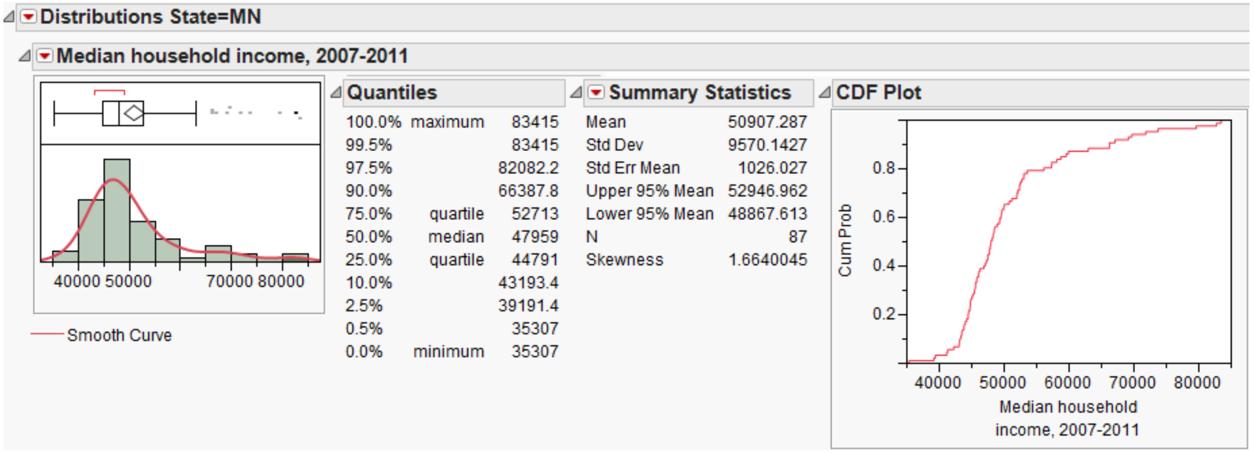
A DISCUSSION OF SKEWNESS

A data distribution is said to be symmetric if it has the same shape on both sides of the center. Skewness measures the amount of asymmetry in a data distribution.

The distribution is said to be *positively skewed* or *skewed to the right* if the measurements tend to trail off to the right. Similarly, a distribution is *negatively skewed* or *skewed to the left* if the measurements trail off to the left.

JMP provides a numerical measure of skewness, as well. This measure takes on the value zero when the data distribution is perfectly symmetric; skewness measures greater than zero indicate the data are positively skewed, and skewness measures less than zero indicate the data are negatively skewed. This skewness measure can be displayed by customizing the Summary Statistics section of the output.

We have already summarized the median household income for counties in Minnesota. Next, let's also consider summaries from two other states, Nevada and Maryland.



Distributions State=MN

Median household income, 2007-2011

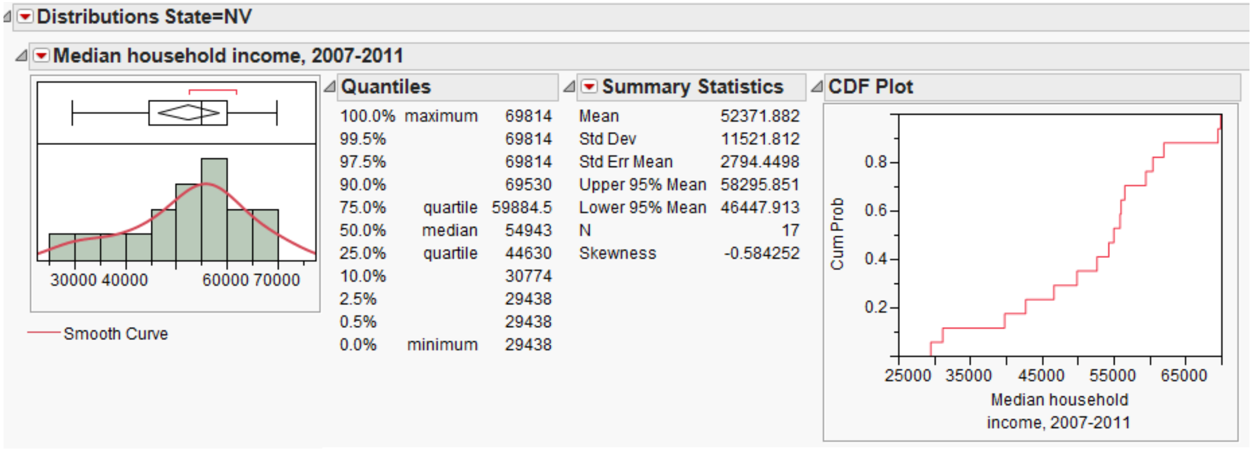
Quantiles

100.0%	maximum	83415
99.5%		83415
97.5%		82082.2
90.0%		66387.8
75.0%	quartile	52713
50.0%	median	47959
25.0%	quartile	44791
10.0%		43193.4
2.5%		39191.4
0.5%		35307
0.0%	minimum	35307

Summary Statistics

Mean	50907.287
Std Dev	9570.1427
Std Err Mean	1026.027
Upper 95% Mean	52946.962
Lower 95% Mean	48867.613
N	87
Skewness	1.6640045

CDF Plot



Distributions State=NV

Median household income, 2007-2011

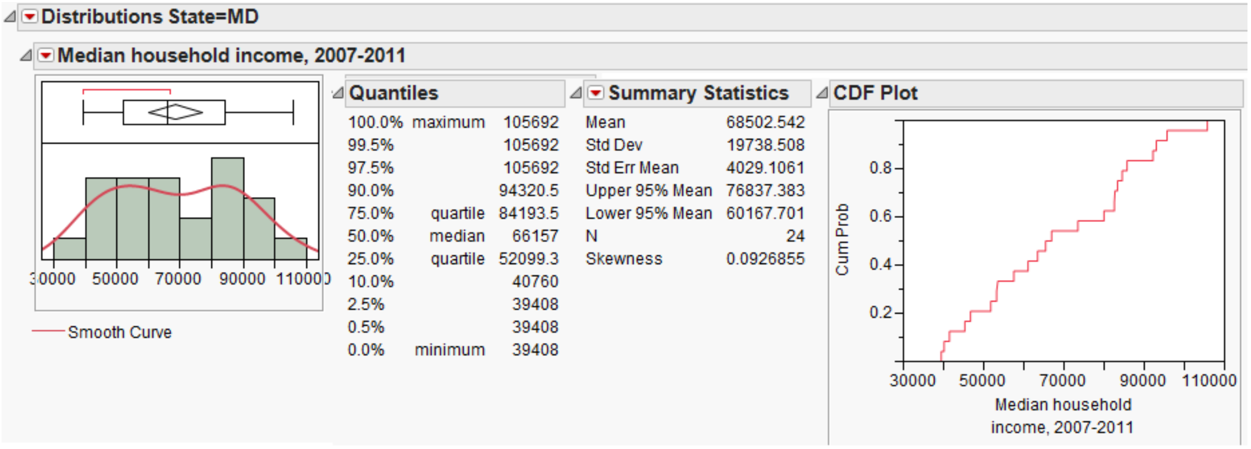
Quantiles

100.0%	maximum	69814
99.5%		69814
97.5%		69814
90.0%		69530
75.0%	quartile	59884.5
50.0%	median	54943
25.0%	quartile	44630
10.0%		30774
2.5%		29438
0.5%		29438
0.0%	minimum	29438

Summary Statistics

Mean	52371.882
Std Dev	11521.812
Std Err Mean	2794.4498
Upper 95% Mean	58295.851
Lower 95% Mean	46447.913
N	17
Skewness	-0.584252

CDF Plot



Distributions State=MD

Median household income, 2007-2011

Quantiles

100.0%	maximum	105692
99.5%		105692
97.5%		105692
90.0%		94320.5
75.0%	quartile	84193.5
50.0%	median	66157
25.0%	quartile	52099.3
10.0%		40760
2.5%		39408
0.5%		39408
0.0%	minimum	39408

Summary Statistics

Mean	68502.542
Std Dev	19738.508
Std Err Mean	4029.1061
Upper 95% Mean	76837.383
Lower 95% Mean	60167.701
N	24
Skewness	0.0926855

Questions:

1. How would you describe the shape of the distribution of median household incomes in Nevada counties?

2. How would you describe the shape of the distribution of median household incomes in Maryland counties?

Z-SCORES

A Z-score, often called a standardized value, measures the number of standard deviations an observation is away from the mean of the data distribution. The z-score can be used to transform observations to a dimensionless scale; in addition, it can be used to measure the position of an observation. Z-scores are calculated as shown below:

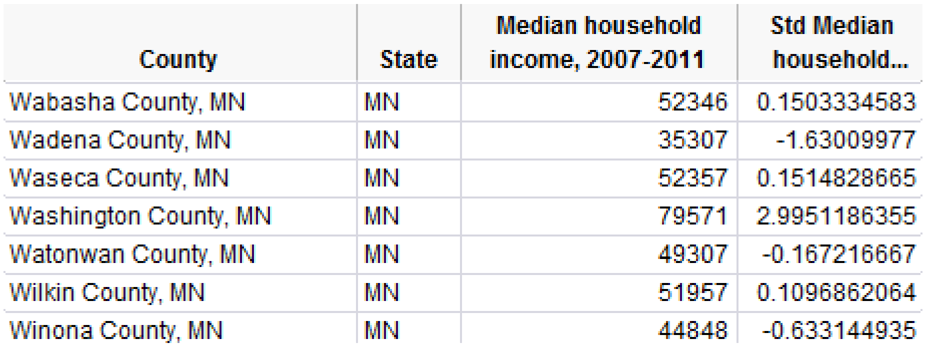
$$Z - \text{score} = \frac{\text{Observation - Mean}}{\text{Standard Deviation}}$$

Interpretation of Z-Scores:

- As mentioned, the standardized values transform the data so that the data is placed on a standard, dimensionless scale that has a mean of 0 and a standard deviation of 1.
- If a Z-Score is negative, then the observation is that many standard deviations below the mean.
- If the Z-Score is positive, then the observation is that many standard deviations above the mean.
- If the Z-Score is 0, then the data value is the same as the mean.
- If the Standard Deviation is 0, then the Z-Score is not defined and thus cannot be computed.

To obtain Z-scores for the median household incomes of Minnesota counties in JMP, select **Save > Standardized** from the red drop-down arrow next to the variable name once you're in the Analyze > Distribution platform.

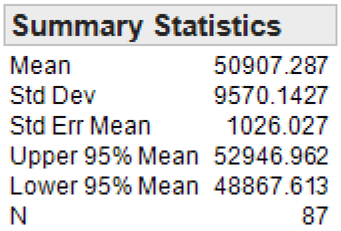
The Z-scores will be displayed in the last column of the original data set.



County	State	Median household income, 2007-2011	Std Median household...
Wabasha County, MN	MN	52346	0.1503334583
Wadena County, MN	MN	35307	-1.63009977
Waseca County, MN	MN	52357	0.1514828665
Washington County, MN	MN	79571	2.9951186355
Watonwan County, MN	MN	49307	-0.167216667
Wilkin County, MN	MN	51957	0.1096862064
Winona County, MN	MN	44848	-0.633144935

Questions:

1. Show how the Z-score for Winona County was calculated:



Summary Statistics	
Mean	50907.287
Std Dev	9570.1427
Std Err Mean	1026.027
Upper 95% Mean	52946.962
Lower 95% Mean	48867.613
N	87

2. What does this tell you about the relative position of Winona County in the data set?

THE IDENTIFICATION OF OUTLIERS

We have already discussed using boxplots to identify outliers. In addition, we can use Z-scores.

- Any data value whose Z-Score is **below −2 or above 2** is considered a potential outlier.
- Any data value whose Z-Score is **below -3 or above 3** is considered an outlier and warrants further investigation.

These guidelines come from the following theories.

1. **Empirical Rule**: If the probability distribution is bell-shaped and symmetric, then the Empirical Rule applies. This rule says that APPROXIMATELY…

   - 68% of the values fall within one standard deviation of the mean.
   - 95% of the values fall within two standard deviations of the mean.
   - 99.7% of the values fall within three standard deviations of the mean.

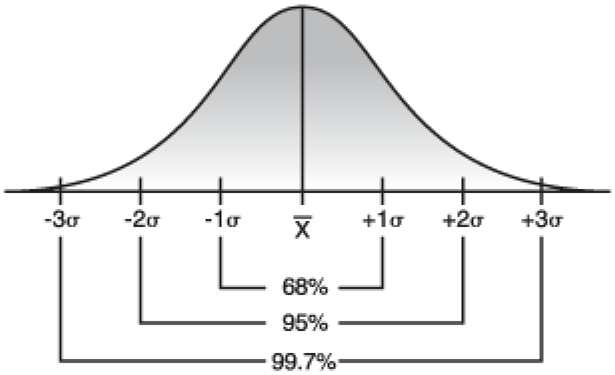


Image Source: http://threes.com

2. **Chebyshev's Rule**: For ANY probability distribution, Chebyshev's Rule tells us that AT LEAST…

   - 75% of the values fall within two standard deviations of the mean.
   - 89% of the values fall within three standard deviations of the mean.
   - 1 - 1/k^2 of the values fall within k standard deviation of the mean.