

Data presentation and summary

Consider the following table:

| State | Income | State | Income | State | Income |
|------------------|--------|----------------|--------|----------------|--------|
| Alabama | 18,010 | Kentucky | 17,807 | North Dakota | 18,546 |
| Alaska | 23,788 | Louisiana | 17,651 | Ohio | 20,928 |
| Arizona | 19,001 | Maine | 19,663 | Oklahoma | 17,744 |
| Arkansas | 16,898 | Maryland | 24,933 | Oregon | 20,419 |
| California | 22,493 | Massachusetts | 25,616 | Pennsylvania | 22,324 |
| Colorado | 22,333 | Michigan | 22,333 | Rhode Island | 22,251 |
| Connecticut | 29,402 | Minnesota | 22,453 | South Carolina | 17,695 |
| Delaware | 22,828 | Mississippi | 15,838 | South Dakota | 19,577 |
| Washington, D.C. | 31,136 | Missouri | 20,717 | Tennessee | 19,482 |
| Florida | 21,677 | Montana | 17,865 | Texas | 19,857 |
| Georgia | 20,251 | Nebraska | 20,488 | Utah | 17,043 |
| Hawaii | 24,057 | Nevada | 24,023 | Vermont | 20,224 |
| Idaho | 18,231 | New Hampshire | 23,434 | Virginia | 22,594 |
| Illinois | 23,784 | New Jersey | 28,038 | Washington | 22,610 |
| Indiana | 20,378 | New Mexico | 17,106 | West Virginia | 17,208 |
| Iowa | 20,265 | New York | 25,999 | Wisconsin | 21,019 |
| Kansas | 20,896 | North Carolina | 19,669 | Wyoming | 20,436 |

These values are 1994 per capita personal income dollar values for the 50 U.S. states and District of Columbia (as provided by the Bureau of Economic Analysis of the U.S. Department of Commerce, and given in the file *pcincome.mpj* in the `js` directory). These are the numbers — so what do we now know? The answer is, not too much. While this table is a perfectly valid presentation of the data, it’s not a very efficient one. We need to summarize the values somehow, and present them using effective tabular and graphical methods, if we are to have any hope of seeing the patterns that are there.

The form of such presentations is intimately connected to the type of data being examined. Data are often classified by whether they are **qualitative** or **quantitative**:

- (1) **Qualitative data:** these are data where the possible values fall into well-defined categories or groups. Examples of qualitative variables include gender (male or female), religion (Protestant, Catholic, Jewish, etc.) and type of school attended (public, non-sectarian private, sectarian private, etc.). These variables are all *nominal* variables, in that there is no ordering to the categories (a variable where reordering the categories does not change your impressions of the data is an example of a nominal variable). *Ordinal* qualitative variables are ones where there is an apparent ordering to the categories (and reordering the categories would result in loss of information). A response scale of the form “Strongly agree – Agree – Neutral – Disagree – Strongly disagree,” for example, leads to an ordinal qualitative variable (such scales are called Likert scales).
- (2) **Quantitative data:** these are data that come in numerical form. Such variables can be classified into ones that are *discrete* and ones that are *continuous*. A discrete variable

is one where the possible outcomes can be counted; for example, the number of children in a family, or the number of airline accidents this year. Qualitative variables where a number has been assigned to each category are sometimes thought of as discrete quantitative variables also. A continuous variable is one where the possible values are so numerous that they cannot be counted (or, there are so many that the count is effectively infinite). Examples of such a variable are the temperature in a room when you enter it, the gross national product of a country, and the net profits of a corporation. Quantitative variables are often viewed as being on an *interval* scale or on a *ratio* scale. Interval-scaled data are ones where the difference between values is meaningful; for example, the 20 degree difference between 40°F and 60°F means the same thing as the difference between 80°F and 100°F. Ratio-scaled data are data where there is a true zero point, so that ratios makes sense; for example, someone who is 70 inches tall is twice as tall as someone who is 35 inches tall (so height is ratio-scaled), but 80°F is not “twice as hot” as 40°F (so temperature Fahrenheit is not ratio-scaled).

Data presentation for qualitative data is pretty straightforward. The natural way of presenting this type of data is by using a **frequency distribution** — that is, a tabulation (or *tally* of the number of observations in the data set that fall into each group.

For example, in Fall, 1994, I asked the members of the Data Analysis and Modeling for Managers course the following three questions:

- (1) In your opinion, does smoking cause lung cancer in humans? (SMOKCANC)
- (2) In your opinion, does environmental tobacco (second-hand) smoke cause lung cancer in humans? (ETSCANC)
- (3) Please classify your usage of tobacco products into one of three groups: Never used, Previously used but do not currently use, Currently use. (USETOBAC)

These three variables are all qualitative ones. The first two are definitely nominal, while the third could possibly be considered ordinal. The following frequency distributions summarize the responses of the students to the questions:

Summary Statistics for Discrete Variables

| SMOKCANC | Count | Percent | ETSCANC | Count | Percent | USETOBAC | Count | Percent |
|----------|-------|---------|---------|-------|---------|----------|-------|---------|
| Yes | 50 | 84.75 | Yes | 42 | 71.19 | Never | 40 | 66.67 |
| No | 9 | 15.25 | No | 17 | 28.81 | Previous | 11 | 18.33 |
| N= | 59 | | N= | 59 | | Currentl | 9 | 15.00 |
| *= | 1 | | *= | 1 | | N= | 60 | |

These tables summarize what’s going on here. Clearly most students felt that smoking causes lung cancer, but opinions on second-hand smoke were a little less strong. Only 15% of the students were currently using tobacco products. One question that would be natural to ask is how these answers related to the national origin of the respondent, since smoking rates are considerably higher in Europe and Asia than they are in the United States. That

is, sensible data analysis will often focus on issues of the *association* between variables, in addition to properties of the variables separately. The term *correlation* is often used as well, particularly for quantitative variables. When we examine such associations, we must always remember that just because two events occur together, that doesn't mean that one causes the other; that is, *correlation does not imply causation*.

In Spring, 1994, a survey was administered to 61 Stern students regarding their opinions of the New York City subway system (the data are given in the file *subway.mpj* in the *chs* directory). Among other questions, they were asked to rate the cleanliness of the stations and the safety of the stations on the scale “Very unsatisfactory – Unsatisfactory – Neutral – Satisfactory – Very satisfactory.” These variables are ordinal qualitative variables. Once again a frequency distribution is very effective at summarizing the results of the survey, but now the ordering of the entries should be taken into account:

Summary Statistics for Discrete Variables

| ClnStat | Count | CumCnt | Percent | CumPct |
|---------|-------|--------|---------|--------|
| 1 | 23 | 23 | 37.70 | 37.70 |
| 2 | 27 | 50 | 44.26 | 81.97 |
| 3 | 3 | 53 | 4.92 | 86.89 |
| 4 | 7 | 60 | 11.48 | 98.36 |
| 5 | 1 | 61 | 1.64 | 100.00 |
| N= | 61 | | | |
| *= | 1 | | | |

Summary Statistics for Discrete Variables

| SafStat | Count | CumCnt | Percent | CumPct |
|---------|-------|--------|---------|--------|
| 1 | 17 | 17 | 27.87 | 27.87 |
| 2 | 15 | 32 | 24.59 | 52.46 |
| 3 | 17 | 49 | 27.87 | 80.33 |
| 4 | 10 | 59 | 16.39 | 96.72 |
| 5 | 2 | 61 | 3.28 | 100.00 |
| N= | 61 | | | |
| *= | 1 | | | |

It is obvious that the students were very dissatisfied with the cleanliness of the stations, as more than 80% of the respondents rated it “Very unsatisfactory” or “Unsatisfactory.”

Cleanliness can be viewed as a “quality of life” issue, and on that score the Metropolitan Transit Authority is apparently not doing the job. The picture is a little better as regards safety, but is still not good enough, as more than half the respondents rate safety in the “Unsatisfactory” categories. What is interesting about this is that crime rates in the subways are *lower* than they are above ground, but perceptions don’t necessarily follow the facts.

Let’s go back now to the per capita income data on page 1. In theory we could form a frequency distribution for this variable (although Minitab won’t allow this using the *tally* command), but there’s really no reason to try, since it would just be a list of 51 numbers. For this continuous variable, a raw frequency distribution isn’t very helpful, since it pretty much duplicates the original table. We need to form a set of categories for this variable, and then look at the resultant frequency distribution. Here’s an example:

Summary Statistics for Discrete Variables

| PCIncome | Count | CumCnt | Percent | CumPct |
|-------------|-------|--------|---------|--------|
| 15800–19800 | 18 | 18 | 35.29 | 35.29 |
| 19800–23800 | 25 | 43 | 49.02 | 84.31 |
| 23800–27800 | 5 | 48 | 9.80 | 94.12 |
| 27800–31800 | 3 | 51 | 5.88 | 100.00 |
| N= | 51 | | | |

This isn’t too great either. We’ve chosen categories that are too wide, and the summary frequency distribution is too coarse to be very useful. How do we know how many categories to use? A general rule of thumb is to use in the range of 5 to 15 categories, with usually no more than 7 or 8 for sample sizes less than 50. It’s also often helpful to define categories using round numbers, to make it easier to interpret the results. The best rule, however, is to just look at the frequency distribution and see what makes sense.

Here’s another frequency distribution for the per capita income data:

Summary Statistics for Discrete Variables

| PCIncome | Count | CumCnt | Percent | CumPct |
|-------------|-------|--------|---------|--------|
| 15800-16800 | 1 | 1 | 1.96 | 1.96 |
| 16800-17800 | 7 | 8 | 13.73 | 15.69 |
| 17800-18800 | 5 | 13 | 9.80 | 25.49 |
| 18800-19800 | 5 | 18 | 9.80 | 35.29 |
| 19800-20800 | 9 | 27 | 17.65 | 52.94 |
| 20800-21800 | 4 | 31 | 7.84 | 60.78 |
| 21800-22800 | 8 | 39 | 15.69 | 76.47 |
| 22800-23800 | 4 | 43 | 7.84 | 84.31 |
| 23800-24800 | 2 | 45 | 3.92 | 88.24 |
| 24800-25800 | 2 | 47 | 3.92 | 92.16 |
| 25800-26800 | 1 | 48 | 1.96 | 94.12 |
| 26800-27800 | 0 | 48 | 0.00 | 94.12 |
| 27800-28800 | 1 | 49 | 1.96 | 96.08 |
| 28800-29800 | 1 | 50 | 1.96 | 98.04 |
| 29800-30800 | 0 | 50 | 0.00 | 98.04 |
| 30800-31800 | 1 | 51 | 1.96 | 100.00 |

N= 51

This is a lot more informative than before. First, we see that income values range between roughly \$15,800 and \$31,800. A “typical” value seems to be a bit more than \$20,000. Interestingly, the distribution of income values drops much more quickly below \$20,000 than above it; income values stretch all the way out to over \$30,000, but there are no values below \$15,000.

This pattern can be seen more clearly in a graphical representation of the frequency distribution called a **histogram**. Here’s the histogram that corresponds to the frequency distribution above:



In a histogram, the number of observations that fall into each category is represented by the area of the bar for that category (since the category bins have equal width, this is the same as saying that the count is proportional to the height of the bin). This histogram confirms the impressions mentioned before: a “typical” value a bit above \$20,000, and a long right tail (compared to the left tail). This long-tailedness is not at all unusual for salary / income data. In fact, salaries and incomes are often modeled with a distribution called the *lognormal* distribution, which is characterized by this pattern. It’s easy to understand why this would be — there is a natural lower limit on incomes (zero, of course, but more realistically, a minimal societally accepted income), but not an upper limit. A distribution like this is said to be *asymmetric*; a *symmetric* distribution is one where the upper and lower halves of a histogram look roughly like mirror images of each other.

Another version of the histogram is the **stem-and-leaf display**. Here is an example:

Character Stem-and-Leaf Display

Stem-and-leaf of PCINCOME N = 51
Leaf Unit = 100

```

 1   15 8
 2   16 8
10   17 01266788
13   18 025
19   19 045668
(10) 20 2223444789
22   21 06
20   22 233344568
11   23 477
 8   24 009
 5   25 69
 3   26
 3   27
 3   28 0
 2   29 4
 1   30
 1   31 1
```

The stem-and-leaf display has a little more information than the histogram, in that rounded-off versions of the data values can be read from the display. For example, the “15” in the stem followed by the “8” on the leaves represents 15,800; the “16” “8” is 16,800; and so on. The first column is a running count of the total number of observations from either the low end or high end (there is 1 value up through the 15,000’s; there are 2 up through the 16,000’s; there are 10 up through the 17,000’s; there is one greater than or equal to the 31,000’s; there are 2 greater than or equal to the 29,000’s; and so on). The bin where the “middle” value falls has the number of observations in that bin given in parentheses.

The stem-and-leaf display is particularly useful for paper-and-pen analysis, since it is pretty easy to construct one quickly (try it!). It also provides a quick way to sort a set of numbers.

Still another variation on the histogram is the **frequency polygon**. In this display, instead of the bars of the histogram representing the number of observations in each bin, the midpoints of the tops of each bar are connected by straight lines. Even though the same information is given in the histogram and frequency polygon, the frequency polygon is smoother, and can be a more accurate representation of the true pattern in the data because of this.



These graphical displays are very helpful in getting a general feel for the data. Having done that, we would probably like to summarize what is going on with a few well-chosen numbers. Such numbers are called *summary statistics*, and they fall into two broad categories: measures of location, and measures of scale.

Measures of location

A measure of location, also known as a measure of central tendency, is a number designed to be a “typical value” for the set of data. We’ll represent our data as $\{x_1, \dots, x_n\}$, a set of n values. The most commonly used measure of location is the **sample mean** (often called the average), which equals

$$\bar{X} = \frac{x_1 + \dots + x_n}{n} \equiv \frac{1}{n} \sum_{i=1}^n x_i$$

(the latter notation is called “sigma notation,” and is the common mathematical shorthand for summation). The mean can always be calculated, is unique, and is reliable (in the sense that if we repeatedly take samples from a given population, and calculate \bar{X} , the values don’t move around very much). The sample mean for the income data is $\overline{\text{PCINCOME}} = \$21,157$, and we could report this as a “typical” value of 1994 state per capita personal income. In general, the sample mean can be useful for quantitative data, and is even sometimes used for ordinal qualitative data, if a number is assigned to each category of the variable.

The mean isn't perfect, however. If you go back to the histogram or stem-and-leaf display and see where the mean falls, you can see that it seems to be too high to be "typical," falling too far in the upper tail. The problem is that the mean is not *robust*. A robust statistic is one that is not severely affected by unusual observations, and the mean doesn't have this property — it is very severely affected by such observations. If a value is unusually high or low, the mean will "follow" it, and will correspondingly be too high or too low, respectively (such an unusual value is called an **outlier**, since it lies outside the region where most of the data are).

For the income data, there are three values that are noticeably larger than the other values: 28,038 (New Jersey), 29,402 (Connecticut) and 31,136 (Washington, D.C.). The first two of these states benefit from having many residents who earn high salaries in New York City, without having their average income drawn down by the poorer parts of the city (since the suburbs are in different states from the city). Washington, D.C. is a special case — a city where most of the population is relatively poor, but where a small portion of the population (connected with the Federal government) are very wealthy. It's not clear whether we should consider these values as outliers, or as just part of the generally long right tail of the distribution of the data. If we decided to consider them as outliers, we would examine them to see what makes them unusual (as I just did), and we might omit them from our analysis to see how things look without them.

What would be desirable is a measure of location that is robust. One such measure is the **sample median**. The median is defined as the "middle" value in the sample, if it has been ordered. More precisely, the median has two definitions, one for even n and one for odd n :

$$\begin{array}{c} \underline{n \text{ odd}} \\ x_{(\frac{n+1}{2})} \end{array} \qquad \begin{array}{c} \underline{n \text{ even}} \\ [x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}] / 2 \end{array}$$

The notation $x_{(i)}$ means the i^{th} smallest value in the sample (so $x_{(1)}$ is the minimum, and $x_{(n)}$ is the maximum).

Thus, for the income data, the median is the 26th largest (or smallest) value, or \$20,488. Note that it is almost \$700 smaller than the sample mean, since it is less affected by the long right tail. Since money data are often long right-tailed, it is generally good practice to report things like *median* income or *median* revenue, and *mean* rainfall or *mean* height (since these latter variables tend to have symmetric distributions). The difference between the mean and median here is not that large, since the income variable is not that long-tailed, but the difference between the two can sometimes be very large. That sometimes leads people to report one value or the other, depending on their own agenda. A noteworthy example of this was during the 1994 baseball season before and during the player strike, when owners would quote the mean player salary of \$1.2 million, while the players would quote the median player salary of \$500,000.

Not only were both of these figures correct, but they even both reflected the concerns of each group most effectively. The "units" of measurement for the mean are dollars, since it is based on a sum of dollars, while the units of measurement for the median are people, since it is based on ordering the observations. Since an owner cares about typical total payroll (25 times the mean player salary, since there are 25 players on a team), he or she

is most interested in the mean; since a player cares about what he can earn, he is most interested in the median.

The median is also a useful location measure for ordinal qualitative data, since the idea of a “middle” value is well-defined. The median values for both the station cleanliness and station safety variables in the subway survey are “Unsatisfactory,” which is certainly a reasonable representation of what that data says.

A different robust location measure is the **trimmed mean**. This estimator orders the observations, and then throws out a certain percentage of the observations from the bottom and from the top. *Minitab* gives a 5% trimmed mean. The 5% trimmed mean is clearly not as robust as the median, since if more than 5% of the observations are outliers in the low (or high) direction, they are not trimmed out. The estimator becomes more robust (but less accurate for clean data) as the trimming percentage increases, culminating in the median, which is the (almost) 50% trimmed mean.

A fourth location measure is the **sample mode**, which is simply the value that occurs most often in the data. This isn’t very useful for continuous data, but it is for discrete and especially qualitative data. Indeed, the mode is the *only* sensible location measure for nominal qualitative data. The modes for the smoking survey were “Smoking causes lung cancer,” “Second-hand smoke causes lung cancer,” and “I have never used tobacco products,” all eminently reasonable reflections of what would be considered “typical” responses. The mode for station cleanliness is “Unsatisfactory”, but there are two modes for station safety (“Very unsatisfactory” and “Neutral”), reflecting less unanimity of opinion for the latter question.

The concept of the mode can be extended to continuous data by allowing some “fuzziness” into the definition. For continuous data, a mode is a region that has a more concentrated distribution of values than contiguous regions nearby. The histogram (or stem-and-leaf display) can be used to identify these regions. For the income data, we might identify three modes — around \$17,500, \$20,500 and \$22,500, respectively. This has the appealing characterization of low, medium and high income states, respectively. Members of each group include Arkansas, Louisiana and West Virginia (low), Iowa, Kansas and Nebraska (medium), and Delaware, Pennsylvania and Washington (high), which ties in well with the known association between higher income and the presence of large corporations.

Measures of scale

Location measures only give part of the story, as it is also important to describe the variability in the sample. For example, a fixed income security that provides a constant return of 10% over five years and a real estate investment that averages a 10% return per year over five years both have the same location measure, but are clearly not equivalent investments. More specifically, measures of scale (also known as measures of variability or measures of dispersion) are the primary focus in two important business applications:

- (1) quality control: a manufacturing process that is stable is said to be “in control”; a process that is producing widgets, say, with the correct average diameter, but with too large variability around that average, is said to be “out of control”
- (2) the stock market: stocks that are more variable than the market (often measured using “beta”) are riskier, since they go up or down more than the market as a whole does

Many different measures of variability have been suggested. A simple one is the **range** of the data, or the difference between the largest and smallest values ($x_{(n)} - x_{(1)}$). This is easy to calculate, but not very useful, since it is **extremely** non-robust (an unusual value becomes the minimum or maximum, and causes the range to become much larger). [An aside: the range can also be used to calculate a location estimate. The **midrange** is the average of the minimum and maximum values in the sample, $[x_{(1)} + x_{(n)}]/2$, or the midpoint of the interval that contains all of the values.]

A better range-based scale estimate trims off the most extreme values, and uses the range of the “middle” values that are left. To define such an estimate, we first have to define the quartiles of the sample. Roughly speaking, the quartiles correspond to the 25% points of the sample; so, the first quartile is the value corresponding to one-fourth of the sample being below and three-fourths above, the second quartile corresponds to half above and below (that is, the median), and the third quartile corresponds to three-fourths of the sample below and one-fourth above. Unfortunately, for most sample sizes, “one-fourth” and “three-fourths” are not well-defined, so we need to make the definition more precise. Different statistical packages use slightly different definitions, but here’s a common one: the first quartile is the value corresponding to a rank of

$$\frac{1 + \lfloor \frac{n+1}{2} \rfloor}{2},$$

where $\lfloor \cdot \rfloor$ is the largest integer less than or equal to the value being evaluated. So, for example, say $n = 32$; then, the first quartile corresponds to a rank of

$$\frac{1 + \lfloor \frac{33}{2} \rfloor}{2} = \frac{1 + 16}{2} = 8\frac{1}{2}.$$

Thus, the first quartile is the average of the eighth and ninth smallest values in the sample. The third quartile is similarly defined, as the value corresponding to a rank of

$$n - \frac{\lfloor \frac{n+1}{2} \rfloor - 1}{2}.$$

Thus, for $n = 32$, the third quartile corresponds to a rank of $32 - (16 - 1)/2 = 32 - 7.5 = 24.5$, or the average of the 24th and 25th smallest values. If this seems a bit mysterious, perhaps it will be clearer if you notice that this is the average of the eighth and ninth *largest* values in the sample, just as the first quartile is the average of the eight and ninth *smallest* values.

The first and third quartiles are values that are relatively unaffected by outliers, since they do not change even if the top and bottom 25% of the sample values are sent off to $\pm\infty$. The difference between them, the **interquartile range** (IQR) is therefore a much more robust measure of scale, as it represents the width of the interval that covers the “middle half” of the data. [An aside: the IQR can also be used to calculate a more robust location estimate. The **midhinge** is the average of the first and third quartiles, or the midpoint of the interval that contains the middle half of the data. Another term for the quartiles is the hinges, which accounts for the name of this location estimate.]

The quartiles and interquartile range provide an easy-to-calculate set of values that can summarize the sample. The **five number summary** is simply the set of five numbers {minimum, first quartile, median, third quartile, maximum}, which can be useful to summarize the general position of the data values. Here is output for the income data that gives the five number summary:

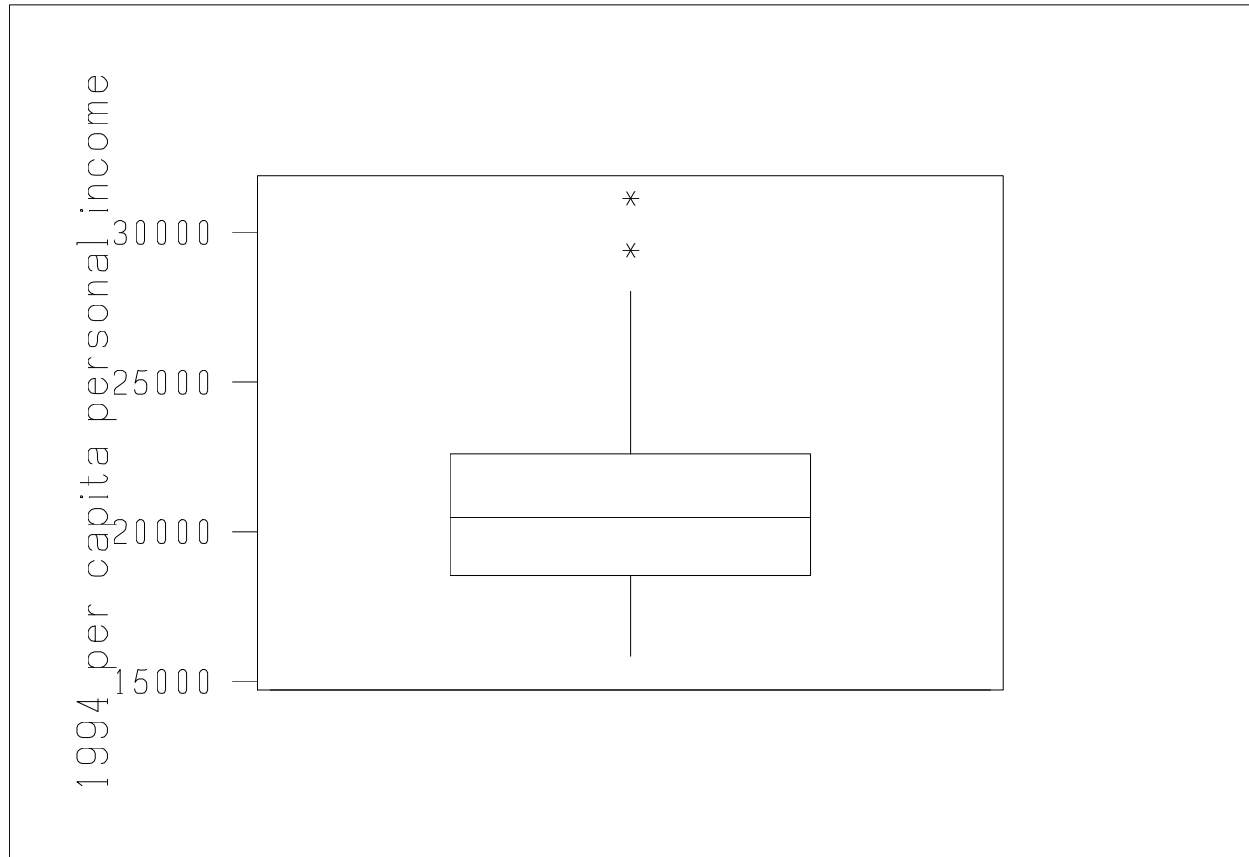
Descriptive Statistics

| Variable | N | Mean | Median | Tr Mean | StDev | SE Mean |
|----------|----|-------|--------|---------|-------|---------|
| PCINCOME | 51 | 21157 | 20488 | 20904 | 3237 | 453 |

| Variable | Min | Max | Q1 | Q3 |
|----------|-------|-------|-------|-------|
| PCINCOME | 15838 | 31136 | 18546 | 22610 |

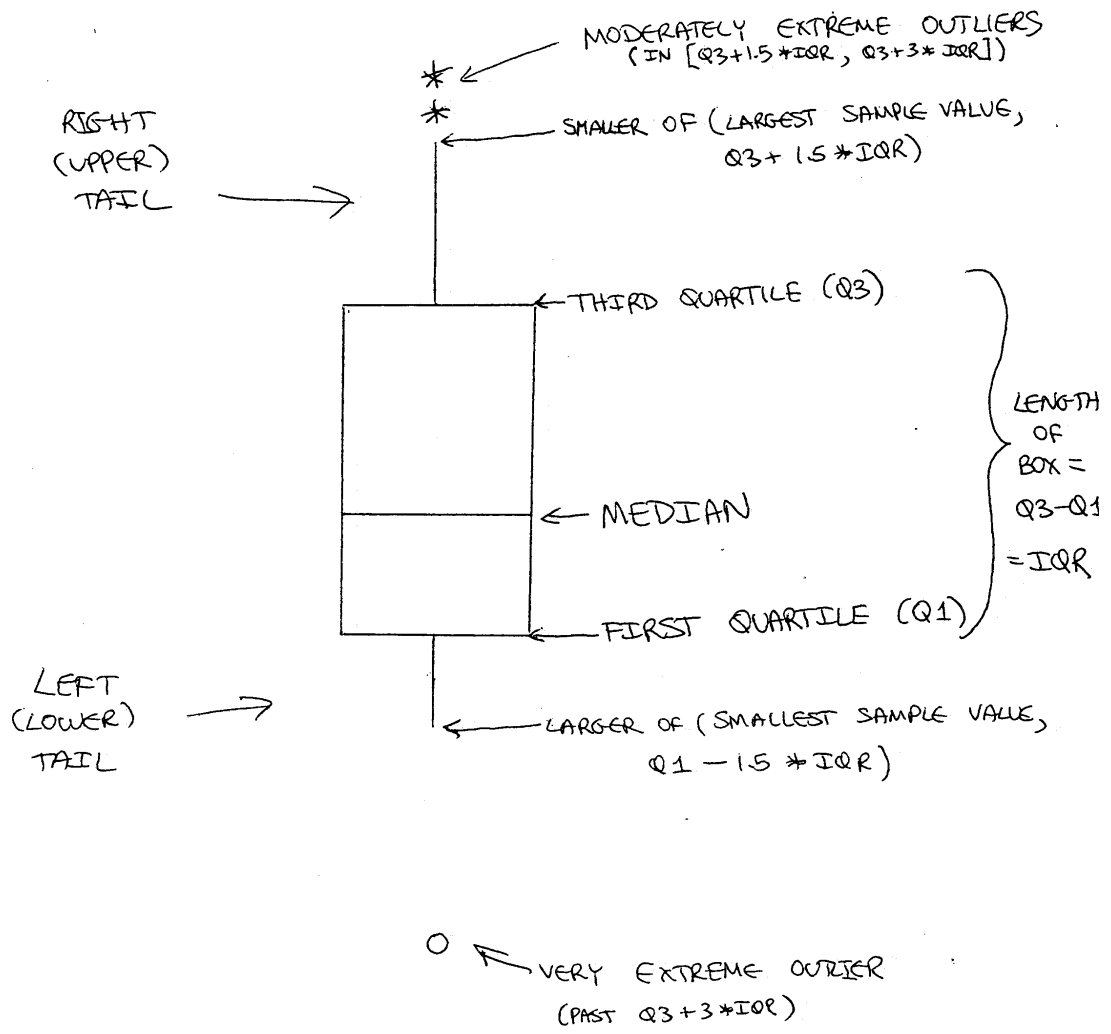
The five number summary is (15838, 18546, 20488, 22610, 31136). The summary shows that half of the values are in the interval (18546, 22610), with the IQR equaling \$4064. Since the median \$20488 is roughly in the middle of this interval (that is, it is not far from the midhinge $(18546 + 22610)/2 = \$20578$), this suggests that the middle of the income distribution is roughly symmetric (a conclusion that the histogram supports). The upper fourth of the data cover a much wider range than the lower fourth, however (\$8526 versus \$2708), suggesting a long right (upper) tail (a conclusion also supported by the histogram).

Using the IQR to measure variability leads to the construction of a useful graphical summary of a set of data called the **boxplot** (sometimes called it the box-and-whisker plot). Here is the boxplot for the income data:



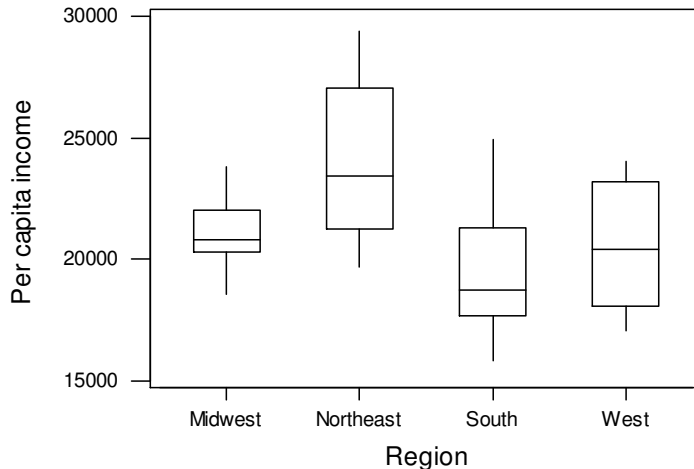
The boxplot consists of two parts. The box is a graphical representation of the middle half of the data. The top line of the box is the third quartile, the bottom line is the first quartile, and the middle line is the median. The lines that extend out from the box are called the whiskers, and they represent the variability in the data. The upper whisker extends out to the smaller of either the largest sample value or the third quartile plus 1.5 times the interquartile range, while the lower whisker extends out to the larger of either the smallest sample value or the first quartile minus 1.5 times the interquartile range. In this way, the whiskers represent how far the data values extend above and below the middle half, except for unusual values. These possible outliers are then identified using either stars (for moderately extreme values that fall between 1.5 and 3 interquartile ranges above and below the third and first quartiles, respectively) or circles (for extreme values more than three interquartile ranges above and below the third and first quartiles, respectively).

The following display summarizes the boxplot construction:



The boxplot for the income data indicates asymmetry, as the upper whisker is longer than the lower whisker. There are also two moderate upper outliers flagged, which correspond to Connecticut and Washington, D.C.

The boxplot is not overwhelmingly helpful in this case, since the histogram shows what's going on in more detail. Boxplots are much more useful for comparing different groups of data, as the following plot shows:



It is evident that there are differences in per capita income based on the region of the country. The Northeast has highest state average per capita income values, but also exhibits higher variability. The Midwest and West regions have similar median state average incomes, but the Midwest is noticeably more consistent (less variable). The South has lowest income values, being the poorest part of the country.

The IQR is not the most commonly used scale estimate, although it does have the advantage of being well-defined for ordinal qualitative data, as well as quantitative data. A data set is not highly variable if the data values are not highly dispersed around the mean, which suggests measuring variability using the deviations from the mean $x_i - \bar{X}$. We can't use $\sum(x_i - \bar{X})$, since the positive and negative values will cancel out, yielding zero. We could use absolute values, but these are hard to handle mathematically. The **sample variance** is a standardized sum of squared deviations, which must be positive:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2.$$

This is an average of the squared deviations, except that the sum is divided by $n - 1$ instead of n , for purely technical reasons (except for very small samples, there is virtually no difference between using n and $n - 1$ anyway). Unfortunately, the variance is in squared units, so it has no physical meaning (and is thus not a scale estimate), but this can be

fixed by taking its square root, resulting in the **sample standard deviation**:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2} = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n(\bar{X})^2 \right]}.$$

The usefulness of the standard deviation comes from the following rule of thumb: for roughly symmetric data sets without outliers (what we might call “well-behaved” data), we can expect that about two-thirds of the observations will fall within one standard deviation of the mean; about 95% will fall within two standard deviations of the mean; and roughly 99.7% will fall within three standard deviations of the mean. Besides summarizing what’s going on in the sample, the sample mean and standard deviation can also be used to construct a rough **prediction interval**; that is, an interval where we expect future values from the same population to fall.

Say we were told that annual rainfall in New York City is 35 inches, with a standard deviation of 3 inches, based on historical (sample) data. Since rainfall data are well-behaved, this tells us that roughly two-thirds of years have rainfall in the range $35 \pm 3 = (32, 38)$, while roughly 95% of years have rainfall in the range $35 \pm 6 = (29, 41)$. If a particular year had 29 inches of rainfall, we wouldn’t be very surprised, but if a year had only 23 inches, we would, since that is four standard deviations away from the mean, a highly unusual result.

The income data have a standard deviation of \$3236, so the rule of thumb would say, for example, that roughly two-thirds of all states have per capita incomes in the range 21157 ± 3236 , or $(17921, 24393)$. In fact, 35 of the 51 values, or 68.6% of the values, are in this interval, quite close to two-thirds. Roughly 95% of the values should be in the interval 21157 ± 6472 , or $(14685, 27629)$. In fact, 48 of the 51 values, or 94.1% are in the interval, again quite close to the hypothesized value. It should be noted, however, that some luck is involved here, since the income data are a bit too long-tailed to be considered “well-behaved.”

A good way to recognize the usefulness of the standard deviation is in description of the stock market crash of October 19, 1987 (“Black Monday”). From August 1 through October 9, the standard deviation of the daily change in the Dow Jones Industrial Average was 1.17%. On Black Monday the Dow lost 22.61% of its value, or **19.26 standard deviations!** This is an **unbelievably** large drop, and well justifies the description of a crash. What apparently happened is that the inherent volatility of the market suddenly changed, with the standard deviation of daily changes in the Dow going from 1.17% to 8.36% for the two-week period of October 12 through October 26, and then dropping back to 2.09% for October 27 through December 31. That is, the market went from stable to incredibly volatile to stable, with higher volatility in the second stable period than in the first period.

The second stable period wasn’t actually quite stable either. By the middle of 1988 (eight months after the crash) the standard deviation of daily changes in the Dow was down to a little more than 1% again. That is, volatility eventually settled down to historical levels. This is not unusual. Historical data show that since 1926, about 2% of the time stock volatility suddenly changes from stable to volatile. Once this volatile period is entered,

the market stays volatile about 20% of the time, often changing again back to the level of volatility that it started at. If these sudden “shocks” could be predicted, you could make a lot of money, but that doesn’t seem to be possible.

The standard deviation, being based on squared deviations around the mean, is not robust. We can construct an analogous, but more robust, version of the estimate using the same ideas as those that led to the standard deviation, but replacing nonrobust terms with more robust versions. So, instead of measuring variability using deviations around the (nonrobust) \bar{X} , we measure them using deviations around the (robust) median (call it M); instead of making positive and negative deviations positive by squaring them (a nonrobust operation), we make them all positive by taking absolute values (a robust operation); instead of averaging those squared deviations (a nonrobust operation), we take the median of the absolute deviations (a robust operation). This leads to the **median absolute deviation**, the median of the absolute deviations from the sample median; that is,

$$\text{MAD} = \text{median}|x_i - M|$$

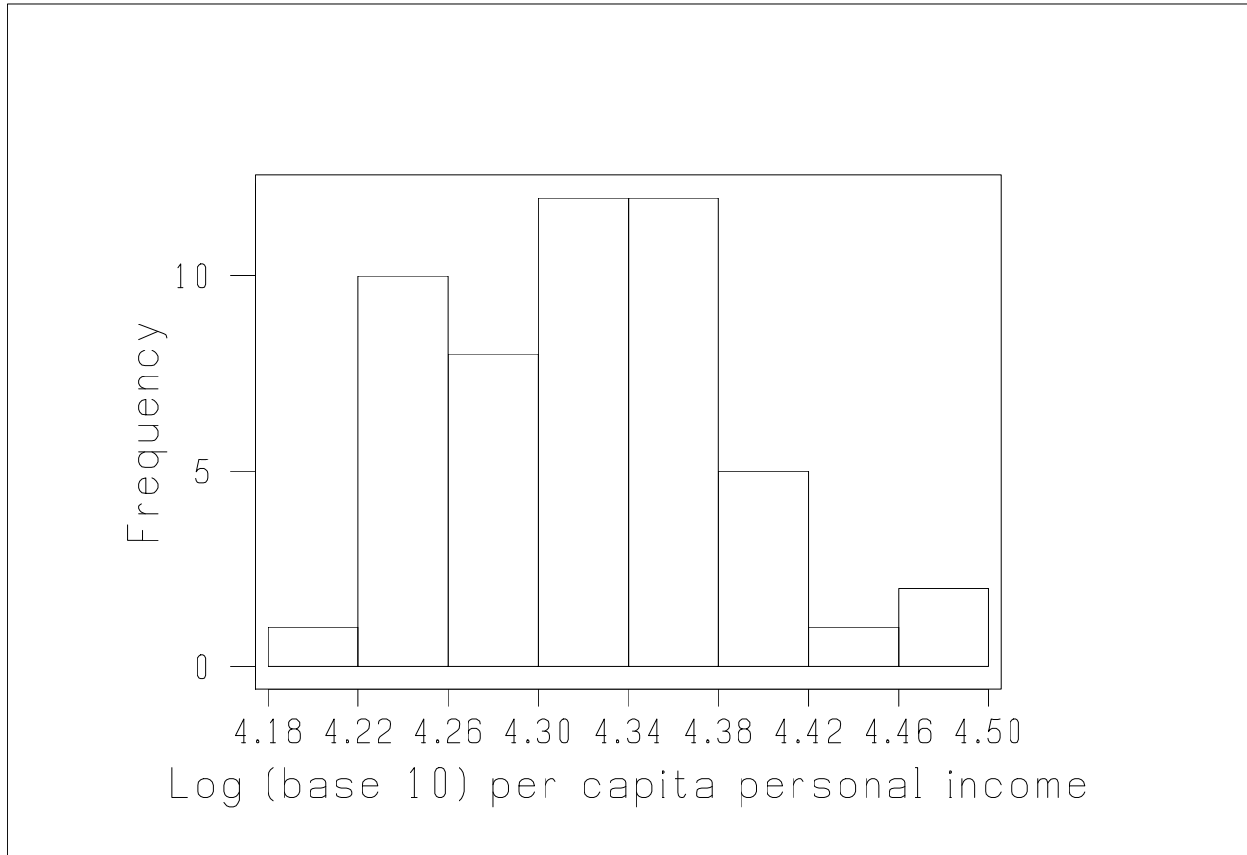
(the MAD is \$2106 for the income data). For well-behaved data, the sample standard deviation satisfies the relationships $s \approx \text{IQR}/1.35 \approx 1.5 \times \text{MAD}$, but for asymmetric data, or data with outliers, the MAD and IQR are much less affected by unusual values, while s is inflated by them.

One other important property of location and scale measures should be mentioned here. All location estimates are *linear* operators, in the following sense: if a constant is added to all of the data values, the location estimate is added to by that constant as well; if all of the data values are multiplied by constant, the location estimate is multiplied by that constant as well.

Scale estimates, on the other hand, are *multiplicative* operators, in the following sense: if a constant is added to all of the data values, the scale estimate is unaffected; if all of the data values are multiplied by a constant, the scale estimate is multiplied by the absolute value of that constant.

It is sometimes the case that data values are easier to analyze if they are examined using a scale that emphasizes the structure better. One such **transformation** that is often appropriate for long right-tailed data is the logarithm. I noted earlier that salary and income data are often modeled as following a lognormal distribution. Data values that follow this distribution have the important property that if they are analyzed in the log scale, they appear “well-behaved” (that is, symmetric, with no outliers; more technically, they look roughly *normally* distributed). There is a good reason why this pattern occurs for income data, which is related to the mathematical properties of the logarithm function. The logarithm has the property that it converts multiplications to additions; that is, $\log(a \times b) = \log(a) + \log(b)$ (the other important property of logarithms is that they convert exponentiation to multiplication; that is, $\log(a^b) = b \times \log(a)$). People typically receive wage increases as a percentage of current wage, rather than fixed dollar amounts, which implies that incomes are multiplicative, rather than additive. If incomes are examined in a log scale, these multiplications become additions. That is, log-incomes can be viewed as sums of (log) values. It turns out that sums of values are usually symmetric, without outliers (that is, well-behaved).

What this means is that if the income data are transformed, by taking logs, the resultant log-incomes will look more symmetric. Logarithms can be taken to any base; common choices are base 10, the *common logarithm*, and base e , the *natural logarithm*, where e is the transcendental number $e \approx 2.71828\dots$). It is easy to convert logarithms from one base to another. For example, $\log_{10}(x) = \log_e(x) \times \log_{10}(e) = .4343 \times \log_e(x)$. Here is a histogram of the log-incomes (base 10):



As expected, the long right tail has pretty much disappeared. This shows up in the summary statistics as well:

Descriptive Statistics

| Variable | N | Mean | Median | Tr Mean | StDev | SE Mean |
|-----------|----|--------|--------|---------|--------|---------|
| LogIncome | 51 | 4.3208 | 4.3115 | 4.3176 | 0.0638 | 0.0089 |

| Variable | Min | Max | Q1 | Q3 |
|-----------|--------|--------|--------|--------|
| LogIncome | 4.1997 | 4.4933 | 4.2683 | 4.3543 |

The mean 4.321 and median 4.312 are similar, as is typical for roughly symmetric data. The log transformation is a *monotone* transformation, which means that the median of

the log–incomes equals the log of the median of the incomes; that is, $4.3115 = \log(20488)$. This is not true for the mean, which leads to yet another location estimate. The **sample geometric mean** is calculated as follows:

- (1) Log the data (base 10, to keep things simple).
- (2) Determine the sample mean of the logged data.
- (3) Raise 10 to a power equaling the sample mean of the logged data.

This algorithm can be written as follows: the geometric mean G equals

$$G = 10^{(\sum_{i=1}^n \log_{10}(x_i)/n)}$$

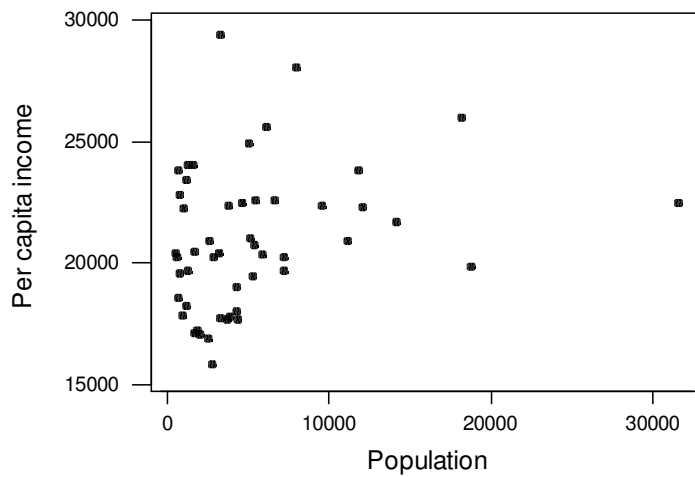
(to distinguish this from the average, the average is sometimes called the sample *arithmetic* mean).

Since the geometric mean is based on an arithmetic mean of logged data, it is less affected by long right tails than the arithmetic mean, and is likely to move in the direction of the median. The geometric mean of the income data is $10^{4.3208} = \$20931$, which is indeed in the direction of the median \$20488 from the (arithmetic) mean \$21157.

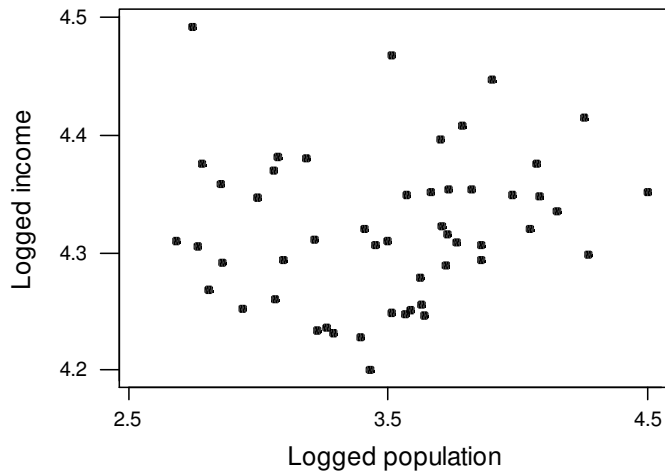
These results don't really show the power of transformation in data analysis. We will talk a good deal about transformations in class. In particular, we will use log transformations often during the semester in various contexts, where their benefits will become clearer. One place where this is true is when looking at the relationships between variables.

Scatter plots and correlation

All of the analyses of the income data so far have been *univariate*; that is, we've only looked at the one variable of average per capita income. It is usually much more interesting to look at how variables relate to each other; that is, look at *bivariate* (two variable) or *multivariate* (many variable) relationships. For example, we might wonder what factors might be related to the average per capita income of a state. Are larger states more likely to have higher per capita income, for example? A graphical way to investigate this question is to construct a **scatter plot**. In this plot each observation is represented by a symbol on the graph. Here is a scatter plot of state average per capita income versus state population (in thousands):



There appears to be a weak direct relationship between the two variables (that is, states with higher populations also have higher average per capita income; an inverse relationship exists if higher values of one variable are associated with lower values of the other). As we will discuss in (much) more detail later in the semester, straight line relationships are particularly easy to understand, but it's not clear that a straight line is appropriate for this relationship. Part of the problem is that most of the observations are crowded into the lower left corner of the plot. The reason for this is that both variables have long right tails. Here is a scatter plot of the same variables, only now both have been logged (base 10):

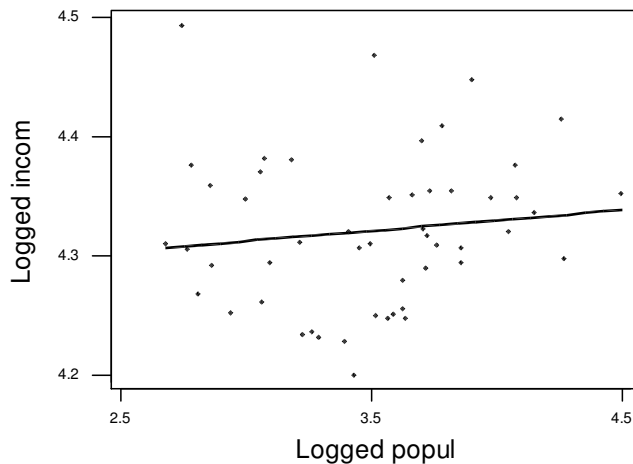


The relationship is still weak, but now the points are spread more evenly over the plot, and a straight line might be reasonable. One way to choose the “best” straight line that goes through the points is using the method of **least squares regression**, which we will spend a good deal of time talking about. The following plot shows the least squares line for these data:

Regression Plot

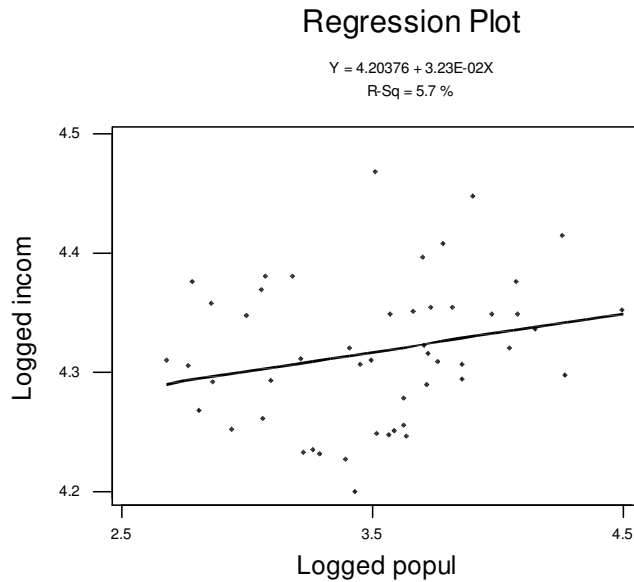
$$Y = 4.26021 + 1.73E-02X$$

R-Sq = 1.5 %



The line through the points is fairly flat, which is consistent with the weak relationship here. Something else we note from this graph is that there is one point that is very far from the regression line — the point in the upper right corner of the plot. This corresponds to a state with low population and high per capita income, and is an outlier. In fact, it’s not a state at all — it’s Washington, DC. As we noted earlier, DC is unusual because it is the

center of the federal government, but at an even more basic level, it doesn't belong on this list, since it's a city, not a state. It is thus quite natural to omit Washington, DC from the data set. Does this change things very much? As you can see, the relationship between the logged average per capita and logged population gets a bit stronger, since the outlier was pulling the line up towards it (and thus flattening it). It's still fairly weak, however.



One way of thinking about the identification of DC as an outlier is to say that its observed (logged) per capita income is far from where we would have expected it to be. The fitted least squares regression line is

$$\text{Logged income} = 4.26 + .0173 \times \text{Logged population};$$

substituting the logged population for DC (which is 2.74351) into the equation yields a *fitted* logged income of 4.307. This can be contrasted with the observed logged income for DC, which is 4.493. The difference between the observed target variable (here logged per capita income) value and the fitted value is called the **residual**, and this residual for DC ($4.493 - 4.307 = .186$) is larger than that for any other observation in the data set.

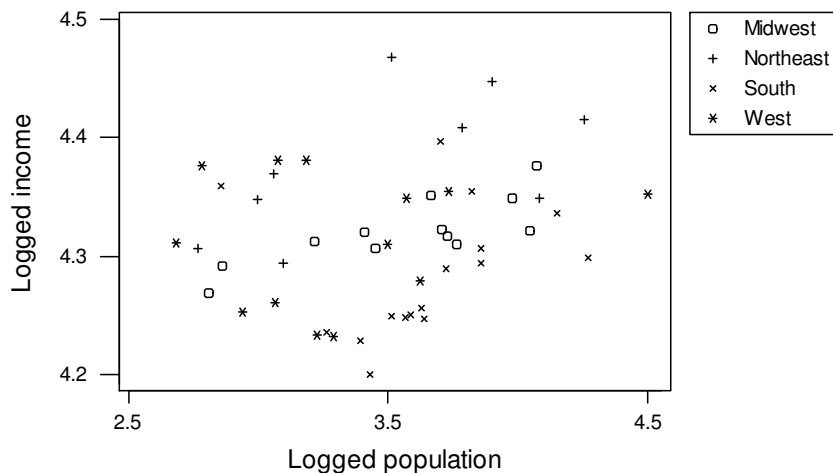
The strength of the straight-line relationship between two variables can be summarized using a statistic, the **correlation coefficient**. This statistic, usually called r , takes on values between -1 and 1 , with $r = 1$ representing a perfect direct linear relationship, $r = -1$ representing a perfect inverse linear relationship, and $r = 0$ representing no linear relationship. The correlation coefficient equals

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{(n - 1)s_X s_Y},$$

where x_i is the i th value of one of the variables, \bar{X} is the sample mean of that variable, y_i and \bar{Y} are the corresponding values for the other variable, and s_X and s_Y are the sample standard deviations of the X and Y variables, respectively. The weak relationship seen

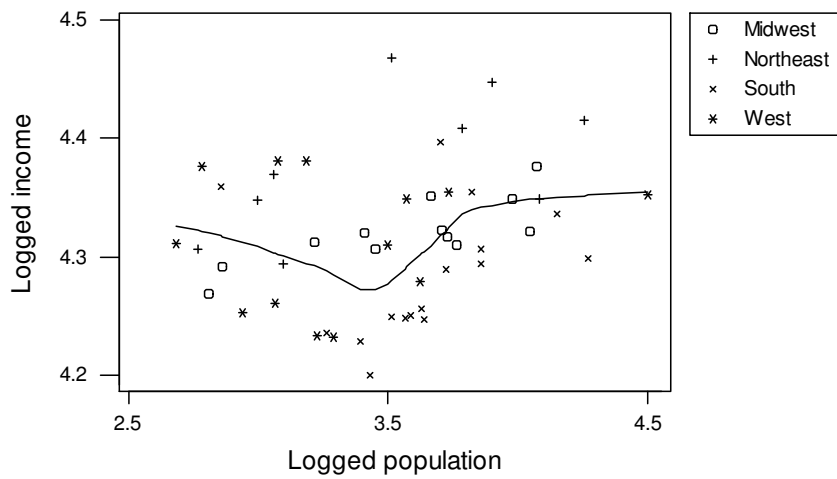
in the plots is also reflected in the correlation coefficient, since the correlation between logged average per capita income and logged population is only .239. The outlier can have a strong effect on r ; it equals only .122 when DC is included in the data.

There are other ways of constructing scatter plots that can be used to highlight interesting patterns in the data. For example, as we saw earlier, each state falls into a specific region. Constructing a scatter plot so that the symbol plotted for each observation differs based on group membership can help highlight patterns related to group membership:

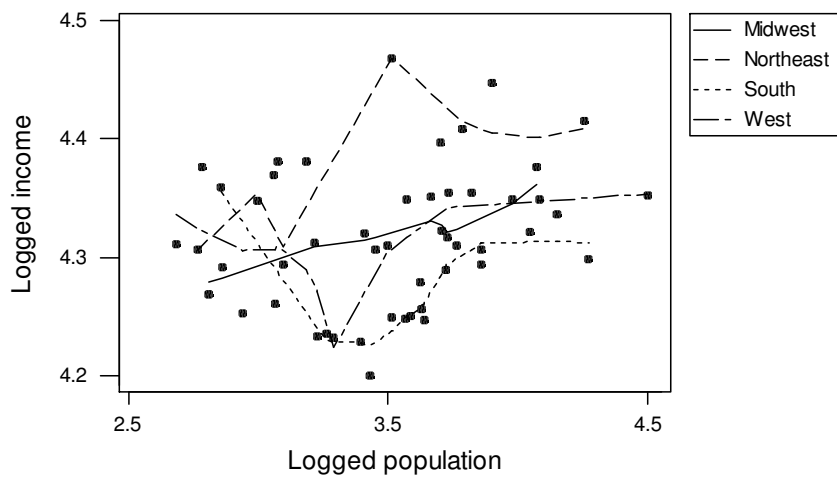


An immediate impression comes from looking at the circles, which correspond to the Midwest region. For these states larger population is associated with higher average per capita income, but what is even more striking is that there is very little spread off the straight line connecting population and income for these states. That is, midwestern states are more internally consistent with respect to this relationship than are states from other regions. Northeastern states also show a direct relationship between the two variables, although there is more variability off the line. We can also see that states from the South generally fall along the bottom of the plot (corresponding to lower average per capita income), with the exception of one observation (this is Delaware, which could, of course, have been treated as falling in the Northeast group).

As was noted earlier, straight line relationships are easy to understand, and are often good representations of relationships, but this is not always the case. A general area of statistical methods called *nonparametric regression* methods attempt to address this possibility by letting the data tell us what the relationship looks like, by putting a smooth curve through the points. **Minitab** includes one such method called **lowess**, and here is the scatter plot with the lowess curve superimposed:



This is remarkable picture. The lowess curve implies that there is an *inverse* relationship between population and income for less populous states (less than 3.5 million people or so), and a *direct* relationship between population and income for more populous states. I don't have a good argument to explain this, but it suggests that there might be other factors involved that we're missing. One of these might be region, as apparently different regions exhibit different relationships:



Some guidelines on thinking about data

The most effective way (by far) to get a feeling for how to think about data is to analyze it; experience is the best teacher. Still, there are some guidelines that can be helpful to consider.

- (1) Notice the title at the top of the page. Before you do **anything**, you should *think* about your data. What are the questions you would like to answer? What are the correct “units” of study? For example, if you’re interested in the purchasing behavior of individual consumers, data on statewide consumption patterns are unlikely to be as useful as data at the individual level, if you can get it.

This issue should come up before you ever gather a single data point, since it determines what kinds of data you’ll look for. It also comes up after you have the data, in that the question you want to answer will often determine how you should look at the data. For example, say you are interested in the process by which people decide whether to respond to a direct mail advertisement. Knowing the number of responses from a particular demographic group (or building a statistical model to predict that number) doesn’t answer the question you’re interested in. Rather, you should convert your data to *rates*, where you are examining the proportion of mailings sent out that result in a response.

- (2) Look at your individual variables. You are probably going to want to learn something about the general level and scale in your variables; construct descriptive statistics to do that. Graphical displays that show the distribution of values include the histogram and stem-and-leaf display. Boxplots summarize the positions of key values (quartiles, limits of the data), but are generally less useful when looking at variables one at a time.
- (3) Think about (possible) relationships between variables in your data. If you think two variables might be related to each other, construct a scatter plot of them (by convention, if one variable is naturally thought of as the “target,” while the other is the “predictor,” the former goes on the vertical axis, while the latter goes on the horizontal axis). A numerical summary of the association between two variables is the correlation coefficient. If the variables you’re looking at are categorical, or discrete with a small number of possible values, a cross-tabulation of the variables (including row or column percentages, as appropriate) allows you to see any relationships.
- (4) The best way to think about problems is often from a predictive point of view; that is, can I use information that I have to model, or predict, future values. In this context, a fitted regression line superimposed on a scatter plot provides a graphical representation of what a straight line prediction would look like. If you’re interested to see whether the relationship between two variables might be nonlinear, a scatterplot smoother like lowess can be superimposed on a scatter plot.
- (5) There are often subgroups in data, and it’s of interest to see if variables differ based on group membership. Side-by-side boxplots are an excellent graphical tool to use to assess this. Descriptive statistics separated by group membership give numerical summaries that can be used with the graphical summary provided by the boxplots. Scatter plots where different symbols are used to identify members of different groups also might be helpful.

- (6) Think about whether a transformation of the data makes sense. Long right-tailed data often benefit from taking logarithms, as do data that operate multiplicatively rather than additively (money data, for example).
- (7) Don't be afraid to ignore items (1)–(6)! Look at your data in any way that makes sense to you. A good way to think about analyzing data is that it involves two parts: exploratory analysis, which is detective work, where you try to figure out what you've got and what you're looking for; and inference, which is based on the scientific method, where you carefully sift through the available evidence to decide if generally held beliefs are true or not. The nature of exploratory analysis means that you'll spend a lot of time following false clues that lead to dead ends, but if you're patient and thorough, like Sherlock Holmes or Hercule Poirot you'll eventually get your man!

Minitab commands

To enter data by hand, click on the **Data** window, and enter the values in as you would in any spreadsheet. To then save the data as Minitab file, click on the **Session** window, and then click on the “diskette” icon and enter a file name in the appropriate box. This creates a Minitab “project” file in .mpj format. These files also can include statistical output, graphs, and multiple data files. To read in a previously saved data file, click on the “open folder” icon and enter the file name in the appropriate box. Data files also can be inputted to Minitab as worksheets. Click on **File** → **Open Worksheet** and choose the appropriate file type (possibilities include Excel, dBase, Quattro Pro, and text) and file name.

Frequency distributions for qualitative variables are obtained by clicking on **Stat** → **Tables** → **Tally Individual Variables**. Enter the variable names in the box under **Variables:** (note that you can double click on variable names to “pick them up”). You’ll probably want to click on the box next to **Percents** to get the percent of the sample with each value. If there is a natural ordering to the categories, clicking on the boxes next to **Cumulative counts** and **Cumulative percents** is also appropriate.

An inherently continuous variable can be converted to a discrete one (which can then be summarized using a frequency distribution) by clicking on **Data** → **Code**. If the code is to a set of numbers, click **Numeric** to **Numeric**; if it is to a set of labels, click **Numeric** to **Text**. Enter the original variable under **Code data from columns:**, and a new variable name or column number under **Store coded data in columns:** (you can put the same variable name in the latter box, but then the original values will be overwritten). Enter a set of original ranges in the boxes under **Original values:**, and the corresponding new values in the boxes under **New:**. If more than eight recodings are needed, the dialog box can be called repeatedly until all values are recoded.

To construct a histogram, click on **Graph** → **Histogram**. Clicking on **Simple** will give a histogram, while clicking on **With Fit** will superimpose the best-fitting normal curve. A technically more correct version of the plot would be obtained by first clicking on **Tools** → **Options** → **Individual graphs** → **Histograms** and clicking the radio button next to **CutPoint**. Right-clicking on the graph allows you to change the appearance of the histogram (this is true for most plots). You can add a frequency polygon to the histogram by right-clicking on it, clicking on **Add** → **Smoother**, and changing the **Degree of smoothing** value to 0. You can then omit the bars (leaving only the frequency polygon) by right-clicking on the graph, clicking **Select item** → **Bars**, and then pressing the delete key (you can select other parts of this, and most other, graphical displays this way, and delete them if you wish the same way).

To construct a stem-and-leaf display, click on **Stat** → **EDA** → **Stem-and-Leaf**. Enter the variable name under **Graph variables**. Note that the stem-and-leaf display appears in the Session Window, not in a Graphics Window. To get stem-and-leaf displays separated by groups, click the box next to **By variable:**, and enter the variable that defines the groups (the variable that defines the groups must use different integer values to define the groups).

To construct a boxplot, click on **Graph** → **Boxplot**. To construct separate boxplots of different variables, click on **Simple** and enter the variable name(s) under **Graph variables**.

To construct side-by-side boxplots, click on **Graph** → **Boxplot** → **With Groups**. Enter the variable name(s) under **Graph variables** and the variable(s) that determine the groups under **Categorical variables for grouping**. Note that these grouping variables can be either numerical or categorical.

To obtain descriptive statistics, click on **Stat** → **Basic Statistics** → **Display Descriptive Statistics**. Enter the variable name(s) under **Variables**. To get descriptive statistics separated by groups, click the box next to **By variables:**, and enter the variable that defines the groups (either a numerical or categorical grouping variable can be used here).

To create new variables or transform old ones, click on **Calc** → **Calculator**. Enter the new variable name under **Store result in variable** and the transformation desired under **Expression**. If you put a current variable name under **Store result in variable** you will overwrite the old values of the variable. So, for example, to take logs (base 10), use $LOGT()$ under **Expression** (Log base 10 under **Functions:**).

To construct a scatter plot, click on **Graph** → **Scatterplot** → **Simple**. Enter the variable name for the vertical axis under **Y variables** and the variable name for the horizontal axis under **X variables**. You can construct multiple scatter plots at the same time by adding variable names. To construct a plot with different symbols for observations from different groups, click on **Graph** → **Scatterplot** → **With Groups**. Enter the variable name for the vertical axis under **Y variables**, the variable name for the horizontal axis under **X variables**, and the name of the variable that defines the groups under **Categorical variables for grouping**. To superimpose a lowess curve on a plot, right-click on the plot, and click **Add** → **Smoother**, and click **OK** when the dialog box pops up. To superimpose separate lowess lines for different groups on the same plot, you would click in the box next to **Apply same groups of current displays to lowess smoother** (this only appears in the multiple group scatter plot situation).

To construct a scatter plot with a least squares regression line superimposed, click on **Stat** → **Regression** → **Fitted Line Plot**. Enter the variable name for the vertical axis under **Response (Y):**, and the variable name for the horizontal axis under **Predictor (X):**. Much more extensive regression output, including diagnostic statistics and graphics, is obtained by clicking **Stat** → **Regression** → **Regression**. Enter the target variable (Y) under **Response:** and the predictor (X) under **Predictors:**.

To obtain the correlation coefficient between two variables, or among a set of variables, click on **Stat** → **Basic Statistics** → **Correlation**. Enter the variable names under **Variables**.

“If you torture the data long enough, it will confess.”
— Ronald Coase

“‘Data, Data, Data!’ Holmes cried impatiently, ‘I can’t make bricks without clay!’”
— A. Conan Doyle, *The Adventure of the Copper Beeches*

“‘I can only suggest that, as we are practically without data, we should endeavour to obtain some.’”
— R. Austin Freeman, *A Certain Dr. Thorndike*

“Data, data, everywhere, and not a thought to think.”
— John Allen Paulos

“Everyone spoke of an information overload, but what there was in fact was a non-information overload.”
— Richard Saul Wurman, *What-If, Could-Be*

“In God we trust. All others bring data.”
— W. Edwards Deming

“It’s not what you don’t know that hurts, it’s what you know that ain’t so.”
— Anonymous