MODULE 6

UNIVARIATE EDA - QUANTITATIVE

Contents

6.1 Interpreting Shape **48**
6.2 Interpreting Outliers **49**
6.3 Comparing the Median and Mean **51**
6.4 Synthetic Interpretations **53**

A univariate EDA for a quantitative variable is concerned with describing the distribution of values for that variable; i.e., describing what values occurred and how often those values occurred. Specifically, the distribution is described by four specific attributes:

1. **shape** of the distribution,
2. presence of **outliers**,
3. **center** of the distribution, and
4. **dispersion** or spread of the distribution.

Graphs are used to identify shape and the presence of outliers and to get a general feel for center and dispersion. Numerical summaries, however, are used to specifically describe center and dispersion of the variable. Computing and constructing the required numerical and graphical summaries was described in Module 5. Those summaries are interpreted here to provide an overall description of the distribution of the quantitative variable.

The same three data sets used in Module 5 are used here.

- Number of open pit mines in countries with open pit mines (Table 5.1).
- Richter scale recordings for 15 major earthquakes (Table 5.2).
- The number of days of ice cover at ice gauge station 9004 in Lake Superior.

6.1 Interpreting Shape

A distribution has two tails – a left-tail of smaller or more negative values and a right-tail of larger or more positive values (Figure 6.1). The relative appearance of these two tails is used to identify three different shapes of distributions – symmetric, left-skewed, and right-skewed. If the left- and right-tail of a distribution are approximately equal in shape (length and height), then the distribution is said to be **symmetric** (or more specifically **approximately symmetric**). If the left-tail is stretched out or is longer and flatter than the right-tail, then the distribution is negatively- or **left-skewed**. If the right-tail is stretched out or is longer and flatter than the left-tail, then the distribution is positively- or **right-skewed**. The type of skew is defined by the longer tail; a longer right-tail means the distribution is right-skewed and a longer left-tail means it is left-skewed.

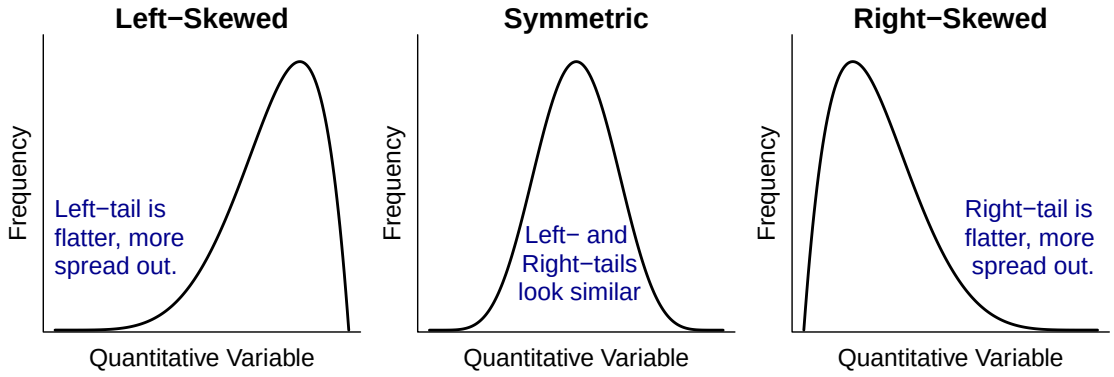

Figure 6.1. Examples of left-skewed (center), symmetric (left), and right-skewed (right) distributions.

◇ **The longer tail defines the type of skew.**

In practice, these labels form a continuum. For example, it may be difficult to discern whether the shape approximately symmetric or one of the skewed distributions. To partially address this issue, "slightly" or "strongly" may be used with "skewed" to distinguish whether the distribution is obviously skewed (i.e., "strongly skewed") or nearly symmetric (i.e., "slightly skewed").

◇ **Symmetric, left-skewed, and right-skewed descriptors are guides; many "real" distributions will not fall neatly into these categories.**

The shape of a distribution is most easily identified from a histogram. Histograms that are examples of each shape are in Figure 6.2. For the sets of skewed distributions, the distributions are less strongly skewed from left-to-right.

The shape of a distribution can also be determined from a boxplot. The relative length from the median to Q1 and the median to Q3 (i.e., the relative position of the median line in the box) indicates the shape of the distribution. If the distribution is left-skewed (i.e., lesser-valued individuals are "spread out"; Figure 6.3-Right), then median-Q1 will be greater than Q3-median. In contrast, if the distribution is right-skewed (i.e., larger-valued individuals are spread out; Figure 6.3-Middle), then Q3-median will be greater than median-Q1. Thus, the median is nearer the top of the box for a left-skewed distribution, nearer the bottom of the box for a right-skewed distribution, and nearer the center of the box for a symmetric distribution (Figure 6.3).

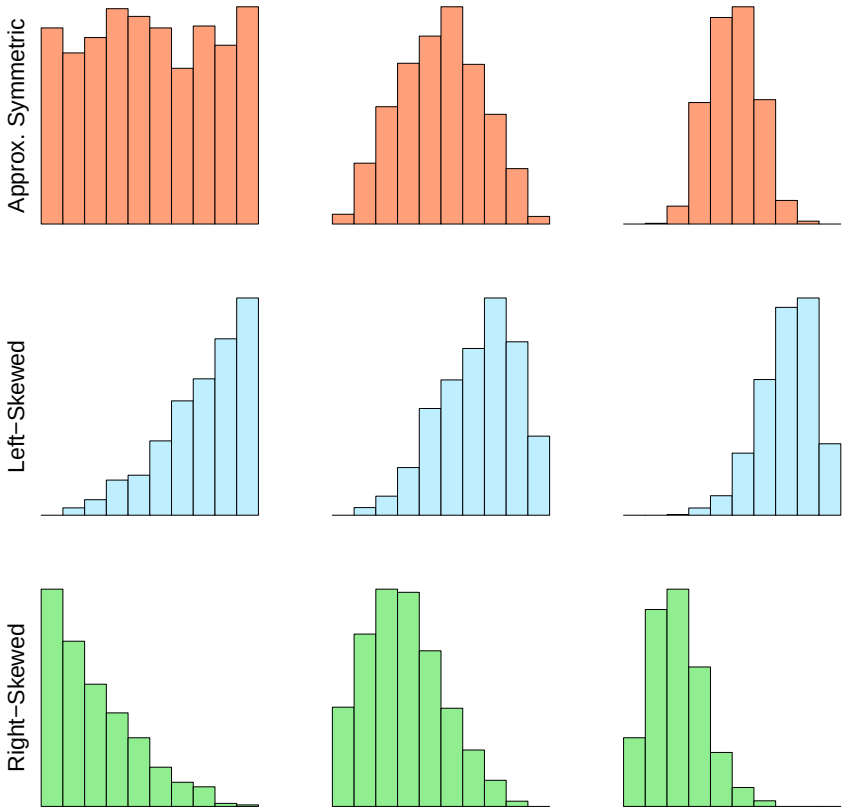

Figure 6.2. Examples of approximately symmetric (top, red), left-skewed (middle, blue), and right-skewed (bottom, green) histograms. Note that the axes labels were removed to focus on the shape of the histograms.

◇ **Even though shape can be described from a boxplot, it is always easier to describe shape from a histogram.**

6.2 Interpreting Outliers

An outlier is an individual whose value is widely separated from the main cluster of values in the sample. On histograms, outliers appear as bars that are separated from the main cluster of bars by "white space" or areas with no bars (Figure 6.4). In general, outliers must be **on the margins of the histogram, should be separated by one or two missing bars, and should only be one or two individuals.**
An outlier may be a result of human error in the sampling process. If this is the case, then the value should be corrected or removed. Other times an outlier may be an individual that was not part of the population of interest – e.g., an adult animal that was sampled when only immature animals were being considered. In this case, the individual should be removed from the sample. Still other times, an outlier is part of the population and should generally not be removed from the sample. In fact you may wish to highlight an outlier as an interesting observation! Regardless, it is important that you construct a histogram to determine if outliers are present or not.

Don't let outliers completely influence how you define the shape of a distribution. For example, if the main

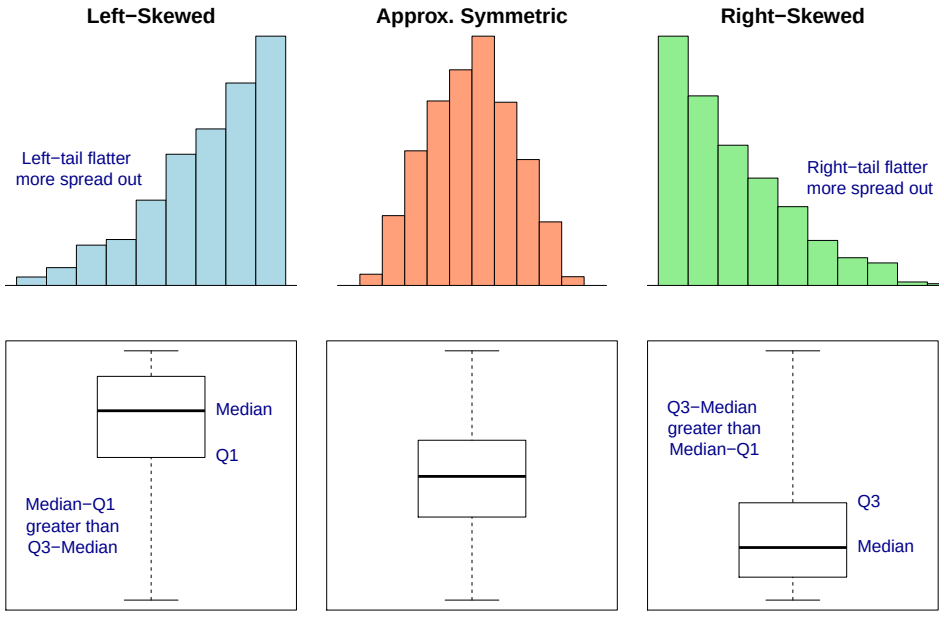

Figure 6.3. Histograms and boxplots for several different shapes of distributions.

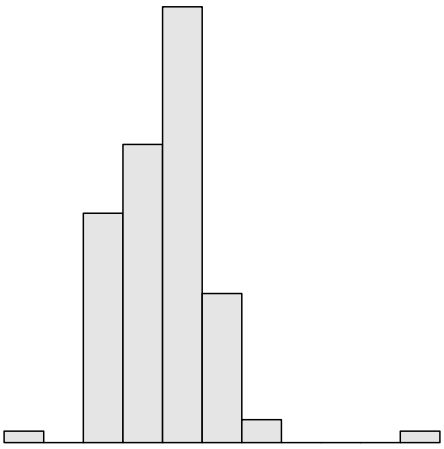

Figure 6.4. Example histogram with an outlier to the right.

cluster of values is approximately symmetric and there is one outlier to the right of the main cluster (as illustrated in Figure 6.4), **DON'T** call the distribution right-skewed. You should describe this distribution as approximately symmetric with an outlier to the right.

◇ **Not all outliers warrant removal from your sample.**

◇ **Don't let outliers completely influence how you define the shape of a distribution.**

6.3 Comparing the Median and Mean

As mentioned previously, numerical measures will be used to describe the center and dispersion of a distribution. However, which values should be used? Should one use the mean or the median as a measure of center? Should one use the IQR or the standard deviation as a measure of dispersion? Which measures are used depends on how the measures respond to skew and the presence of outliers. Thus, before stating a rule for which measures should be used, a fundamental difference among the measures discussed in Module 5 is explored here.

The following discussion is focused on comparing the mean and the median. However, note that the IQR is fundamentally linked to the median (i.e., to find the IQR, the median must first be found) and the standard deviation is fundamentally linked to the mean (i.e., to find the standard deviation, the mean must first be found). Thus, **the median and IQR will always be used together to measure center and dispersion, as will the mean and standard deviation.**

The mean and median measure center in different ways. The median balances the number of individuals smaller and larger than it. The mean, on the other hand, balances the sum of the distances from it to all points smaller than it and the sum of the distances from it to all points greater than it. Thus, the median is primarily concerned with the **position** of the value rather than the value itself, whereas the mean is very much concerned about the **values** for each individual (i.e., the values are used to find the "distance" from the mean).

◇ **The actual values of the data (beyond ordering the data) are not considered when calculating the median; whereas the actual values are very much considered when calculating the mean.**

A plot of the Richter scale data against the corresponding ordered individual number is shown in Figure 6.5-Left.[1] The median (blue line) is found by locating the middle position on the individual number axis and then finding the corresponding Richter scale value (move right until the point is intercepted and then move down to the x-axis). The vertical blue line represents the median; i.e., it has the same **number** of individuals (i.e., points) above and below it. In contrast, the mean finds the Richter scale value that has the same total distance to values below it as total distance to values above it. In other words, the mean is the vertical red line placed such that the total **length** of the horizontal dashed red lines is the same to the left as it is to the right. Thus, the median balances the number of individuals above and below the median, whereas the mean balances the total difference in values above and below the mean.

◇ **The mean balances the distance to individuals above and below the mean. The median balances the number of individuals above and below the median.**

◇ **The sum of all differences between individual values and the mean (as properly calculated) equals zero.**

The mean and median differ in their sensitivity to outliers (Figure 6.5-Right). For example, suppose that an incredible earthquake with a Richter Scale value of 19.0 was added to the earthquake data set. With this additional individual, the median increases from 7.1 to 7.2, but the mean increases from 7.1 to 7.8. The outlier impacts the value of the mean more than the value of the median because of the way that each statistic measures center. The mean will be pulled towards an outlier because it must "put" many values on the "side" of the mean away from the outlier so that the sum of the differences to the larger values and

[1] This is a rather non-standard graph but it is useful for comparing how the mean and median measure the center of the data.

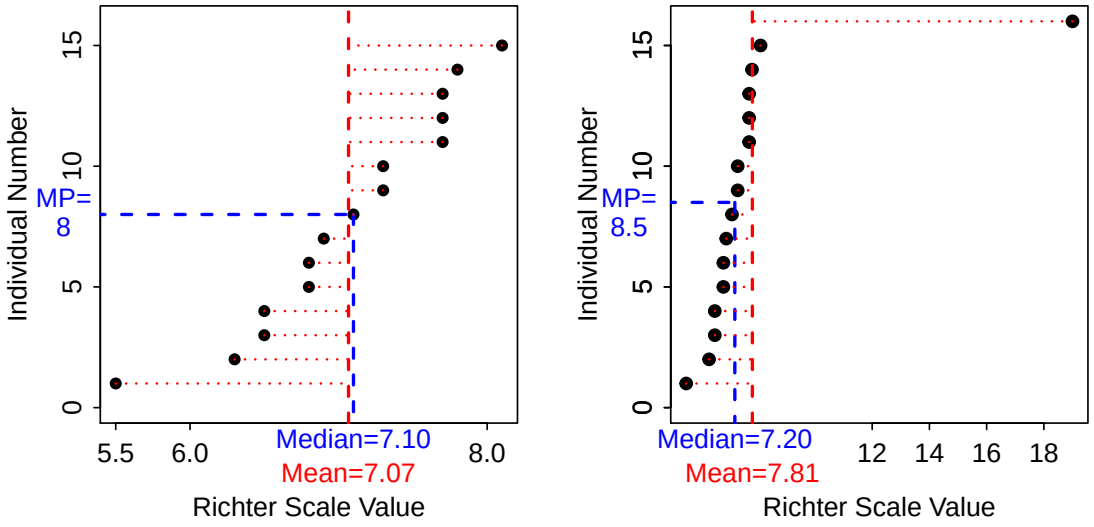

Figure 6.5. Plot of the individual number versus Richter scale values for the original earthquake data (**Left**) and the earthquake data with an extreme outlier (**Right**). The median value is shown as a blue vertical line and the mean value is shown as a red vertical line. Differences between each individual value and the mean value are shown with horizontal red lines.

the sum of the differences to the smaller values will be equal. In this example, the outlier creates a large difference to the right of the mean such that the mean has to "move" to the right to make this difference smaller, move more individuals to the left side of the mean, and increase the differences of individuals to the left of the mean to balance this one large individual. The median on the other hand will simply "put" one more individual on the side opposite of the outlier because it balances the number of individuals on each side of it. Thus, the median has to move very little to the right to accomplish this balance.

◇ **The mean is more sensitive (i.e., changes more) to outliers than the median; it will be "pulled" towards the outlier more than the median.**

The shape of the distribution, even if outliers are not present, also has an impact on the mean and median (Figure 6.6). If a distribution is approximately symmetric, then the median and mean (along with the mode) will be nearly identical. If the distribution is left-skewed, then the mean will be less than the median. Finally, if the distribution is right-skewed, then the mean will be greater than the median.

◇ **The mean is pulled towards the long tail of a skewed distribution. Thus, the mean is greater than the median for right-skewed distributions and the mean is less than the median for left-skewed distributions.**

As shown above, the mean and median measure center in different ways. The question now becomes "which measure of center is better?" The median is a "better" measure of center when outliers are present. In addition, the median gives a better measure of a typical individual when the data are skewed. Thus, in this course, the median is used when outliers are present or the distribution of the data is skewed. If the distribution is symmetric, then the purpose of the analysis will dictate which measure of center is "better." However, in this course, use the mean when the data are symmetric or, at least, not strongly skewed.

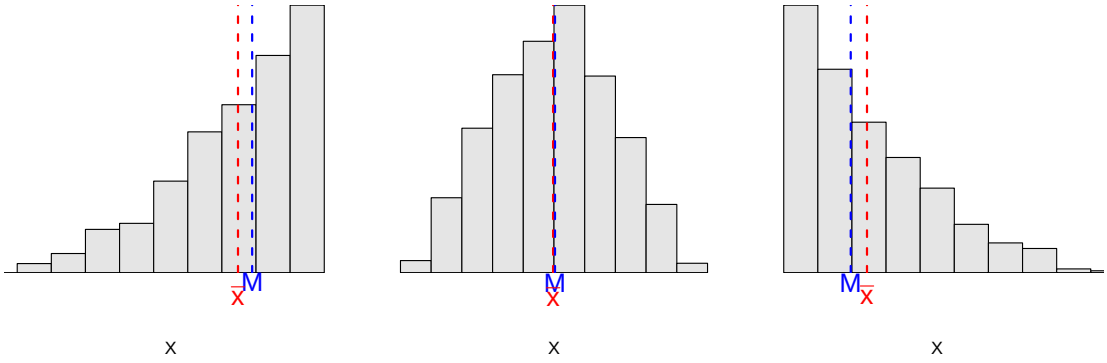

Figure 6.6. Three differently shaped histograms with vertical lines superimposed at the median (M; blue lines) and the mean ($\bar{x}$; red lines).

As note above, the IQR and standard deviation behave similarly to the median and mean, respectively, in the face of outliers and skews. Specifically, the IQR is less sensitive to outliers than the standard deviation.

6.4 Synthetic Interpretations

The graphical and numerical summaries from Module 5 and the rationale described above can be used to construct a synthetic description of the shape, outliers, center, and dispersion of the distribution of a quantitative variable. In the examples below specifically note the 1) reference to figures and tables, 2) labeling of the figures and tables, 3) that only the mean and standard deviation or the median and IQR are discussed, 4) the range was not used alone as a measure of dispersion, 5) the explanation for why either the median and IQR or the mean and standard deviation were used, and 6) an appendix of R code used was provided.

Number of Open Pit Mines

> *Construct a proper EDA for the following situation and data – "The number of open pit mines in countries that have open pit mines (Table 5.1)."*

The number of open pit mines in countries with open pit mines is strongly right-skewed with no outliers present (Figure 6.7). [*I did not call the group of four countries with 10 or more open pit mines outliers because there were more than one or two countries there.*] The center of the distribution is best measured by the median, which is 2 (Table 6.1). The range of open pit mines in the sample is from 1 to 15 while the dispersion as measured by the inter-quartile range (IQR) is from a Q1 of 1.0 to a Q3 of 4.0 (Table 6.1). I chose to use the median and IQR because the distribution was strongly skewed.

Table 6.1. Descriptive statistics of number of open pit mines in countries with open pit mines.

n	mean	sd	min	Q1	median	Q3	max
26.0	3.6	4.0	1.0	1.0	2.0	4.0	15.0

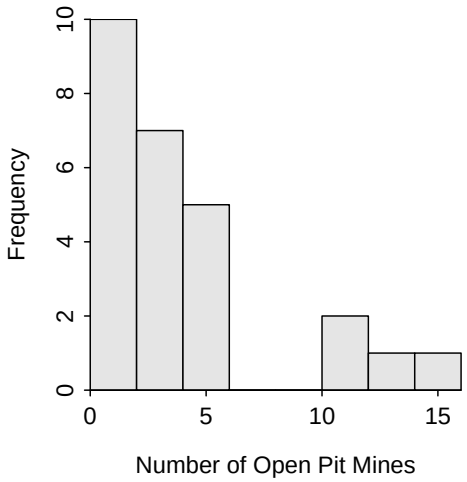

Figure 6.7. Histogram of number of open pit mines in countries with open pit mines.

R Code Appendix:

```
setwd("c:/data/")
mc <- read.csv("MineData.csv")
str(mc)
Summarize(~mines,data=mc,digits=1)
hist(~mines,data=mc,w=2,xlab="Number of open pit mines")
```

Lake Superior Ice Cover

Thoroughly describe the distribution of number of days of ice cover at ice gauge station 9004 in Lake Superior (data are in LakeSuperiorIce.csv).

The shape of number of days of ice cover at gauge 9004 in Lake Superior is approximately symmetric with no obvious outliers (Figure 6.8). The center is at a mean of 107.8 days and the dispersion is a standard deviation of 21.6 days (Table 6.2). The mean and standard deviation were used because the distribution was not strongly skewed and no outlier was present.

Table 6.2. Descriptive statistics of number of days of ice cover at ice gauge 9004 in Lake Superior..

n	nvalid	mean	sd	min	Q1	median	Q3	max
42.0	39.0	107.8	21.6	48.0	97.0	114.0	118.0	146.0

R Appendix:

```
setwd("c:/data/")
LSI <- read.csv("LakeSuperiorIce.csv")
str(LSI)
hist(~days,data=LSI,xlab="Day of Ice Cover",ylab="Frequency of Years",w=20)
Summarize(~days,data=LSI,digits=1)
```

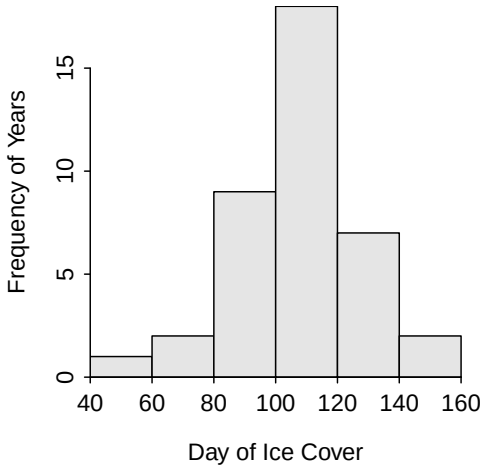

Figure 6.8. Histogram of number of days of ice cover at ice gauge 9004 in Lake Superior.

Crayfish Temperature Selection

Peck (1985) examined the temperature selection of dominant and subdominant crayfish (Orconectes virilis) together in an artificial stream. The temperature (^oC) selection by the dominant crayfish in the presence of subdominant crayfish in these experiments was recorded below. Thoroughly describe all aspects of the distribution of selected temperatures.

30	26	26	26	25	25	25	25	25	24	24	24	24	24	24	23
23	23	23	22	22	22	22	21	21	21	20	20	19	19	18	16

The shape of temperatures selected by the dominant crayfish is slightly left-skewed (Figure 6.9) with a possible weak outlier at the maximum value of 30^oC (Table 6.3). The center is best measured by the median, which is 23^oC (Table 6.3) and the dispersion is best measured by the IQR, which is from 21 to 25^oC (Table 6.3). I used the median and IQR because of the (combined) skewed shape and outlier present.

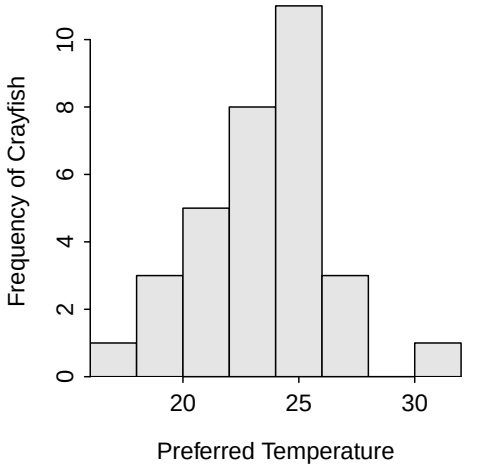

Figure 6.9. Histogram of crayfish temperature preferences.

Table 6.3. Descriptive statistics of crayfish temperature preferences.

n	mean	sd	min	Q1	median	Q3	max
32.00	22.88	2.79	16.00	21.00	23.00	25.00	30.00

R Appendix:

```
setwd("c:/data/")
cray <- read.csv("Crayfish.csv")
str(cray)
hist(~temp,data=cray,xlab="Preferred Temperature",ylab="Frequency of Crayfish",w=2)
Summarize(~temp,data=cray,digits=2)
```