MODULE 5

SUMMARIES FOR ONE QUANTITATIVE VARIABLE)

Contents

SUMMARIZING LARGE QUANTITIES OF DATA WITH few graphical or numerical summaries makes it is easier to identify meaning from data (discussed in Module 1). Numeric and graphical summaries specific to a single quantitative variable are described in this module. Interpretations from these numeric and graphical summaries are described in the next module.

Two data sets will be considered in this module when making calculations "by hand" (i.e., without using R). The first data set consists of the number of open pit mines in countries that have open pit mines (Table 5.1).[1] The second data set is Richter scale recordings for 15 major earthquakes (Table 5.2). A third data set – number of days of ice cover at ice gauge station 9004 in Lake Superior – will be used to demonstrate calculations with R. These data are in LakeSuperiorIce.csv and are loaded into `LSI` below.[2]

```
> LSI <- read.csv("data/LakeSuperiorIce.csv")
```

Table 5.1. Number of open pit mines in countries that have open pit mines.

2.0	11.0	4.0	1.0	15.0	12.0	1.0	1.0	3.0	2.0	2.0	1.0	1.0
1.0	1.0	2.0	4.0	1.0	4.0	2.0	4.0	2.0	1.0	4.0	11.0	1.0

Table 5.2. Richter scale recordings for 15 major earthquakes.

5.5	6.3	6.5	6.5	6.8	6.8	6.9	7.1	7.3	7.3	7.7	7.7	7.7	7.8	8.1

[1] These data were collected from this page. See Section 4.3.2 for how to enter these data into R.
[2] See Section 4.3.2 for how to access these data. These data are originally from the National Snow and Ice Data Center.

5.1 Numerical Summaries

A "typical" value and the "variability" of a quantitative variable are often described from numerical summaries. Calculation of these summaries is described in this module, whereas their interpretation is described in Module 5. As you will see in Module 5, "typical" values are measures of **center** and "variability" is often described as **dispersion** (or spread). Three measures of center are the median, mean, and mode. Three measures of dispersion are the inter-quartile range, standard deviation, and range.

All measures computed in this module are summary statistics – i.e., they are computed from individuals in a sample. Thus, the name of each measure should be preceded by "sample" – e.g., sample median, sample mean, and sample standard deviation. These measures could be computed from every individual, if the population was known. These values would then be parameters and would be preceded by "population" – e.g., population median, population mean, and population standard deviation.[3]

5.1.1 Median

The median is the value of the individual in the position that splits the **ordered** list of individuals into two equal-**sized** halves. In other words, if the data are ordered, half the values will be smaller than the median and half will be larger.

The process for finding the median consists of three steps,[4]

1. Order the data from smallest to largest.
2. Find the "middle **position**" (mp) with $mp = \frac{n+1}{2}$.
3. If mp is an integer (i.e., no decimal), then the median is the value of the individual in that position. If mp is not an integer, then the median is the average of the value immediately below and the value immediately above the mp.

As an example, the open pit data from Table 5.1 are,

1	1	1	1	1	1	1	1	1	1	2	2	2
2	2	2	3	4	4	4	4	4	11	11	12	15

Because $n = 26$, the $mp = \frac{26+1}{2} = 13.5$. The mp is not an integer so the median is the average of the values in the 13th and 14th ordered positions (i.e., the two positions closest to mp). Thus, the median number of open pit mines in this sample of countries is $\frac{2+2}{2} = 2$.

Consider finding the median of the Richter Scale magnitude recorded for fifteen major earthquakes as another example (ordered data are in Table 5.2). Because $n = 15$, the $mp = \frac{15+1}{2} = 8$. The mp is an integer so the median is the value of the individual in the 8th ordered position, which is 7.1.

◇ Don't forget to order the data when computing the median.

5.1.2 Inter-Quartile Range

Quartiles are the values for the three individuals that divide ordered data into four (approximately) equal parts. Finding the three quartiles consists of finding the median, splitting the data into two equal parts at

[3] See Module 2.1 for clarification on the differences between populations and samples and parameters and statistics.
[4] Most computer programs use a more sophisticated algorithm for computing the median and, thus, will produce different results than what will result from applying these steps.

the median, and then finding the medians of the two halves.[5] A concern in this process is that the median is NOT part of either half if there is an odd number of individuals. These steps are summarized as,

1. Order the data from smallest to largest.
2. Find the median – this is the second quartile (Q2).
3. Split the data into two halves at the median. If n is odd (so that the median is one of the observed values), then the median is not part of either half.[6]
4. Find the median of the lower half of data – this is the 1st quartile (Q1).
5. Find the median of the upper half of data – this is the third quartile (Q3).

These calculations are illustrated with the open pit mine data (the median was computed in Section 5.1.1). Because $n = 26$ is even, the halves of the data split naturally into two halves each with 13 individuals. Therefore, the $mp = \frac{13+1}{2} = 7$ and the median of each half is the value of the individual in the seventh position. Thus, $Q1 = 1$ and $Q3 = 4$.

1	1	1	1	1	1	1
1	1	1	2	2	2	
2	2	2	3	4	4	4
4	4	11	11	12	15	

In summary, the first, second, and third quartiles for the open pit mine data are 1, 2, and 4, respectively. These three values separate the ordered individuals into approximately four equally-sized groups – those with values less than (or equal to) 1, with values between (inclusive) 1 and 2, with values between (inclusive) 2 and 4, and with values greater (or equal to) than 4.

As another example, consider finding the quartiles for the earthquake data (Table 5.2). Recall from above (Section 5.1.1) that the median (=7.1) is in the eighth position of the ordered data. The value in the eighth position will not be included in either half. Thus, the two halves of the data are 5.5, 6.3, 6.5, 6.5, 6.8, 6.8, 6.9 and 7.3, 7.3, 7.7, 7.7, 7.7, 7.8, 8.1. The middle position for each half is then $mp = \frac{7+1}{2} = 4$. Thus, the median for each half is the individual in the fourth position. Therefore, the median of the first half is $Q1 = 6.5$ and the median of the second half is $Q3 = 7.7$.

The interquartile range (IQR) is the difference between $Q3$ and $Q1$, namely $Q3 - Q1$. However, the IQR (as strictly defined) suffers from a lack of information. For example, what does an IQR of 9 mean? It can have a completely different interpretation if the IQR is from values of 1 to 10 or if it is from values of 1000 to 1009. Thus, the IQR is more useful if presented as both $Q3$ and $Q1$, rather than as the difference. Thus, for example, the IQR for the open pit mine data is from a $Q3$ of 4 to a $Q1$ of 1 and the IQR for the earthquake data is from a $Q3$ of 7 to a $Q1$ of 6.5.

◇ **The IQR can be thought of as the "range of the middle half of the data."**

◇ **When reporting the IQR, explicitly state both $Q3$ and $Q1$ (i.e., do not subtract them).**

5.1.3 Mean

The mean is the arithmetic average of the data. The sample mean is denoted by $\bar{x}$ and the population mean by μ. The mean is simply computed by adding up all of the values and dividing by the number of individuals.

[5] You should review how a median is computed before proceeding with this section.
[6] Some authors put the median into both halves when n is odd. The difference between the two methods is minimal for large n.

If the measurement of the generic variable x on the ith individual is denoted as x_i, then the sample mean is computed with these two steps,

1. Sum (i.e., add together) all of the values – $\sum_{i=1}^{n} x_i$.
2. Divide by the number of individuals in the sample – n.

or more succinctly summarized with this equation,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.1.1}$$

For example, the sample mean of the open pit mine data is computed as follows:

$$\bar{x} = \frac{2 + 11 + 4 + 1 + 15 + ... + 2 + 1 + 4 + 11 + 1}{26} = \frac{94}{26} = 3.6$$

Note in this example with a discrete variable that it is possible (and reasonable) to present the mean with a decimal. For example, it is not possible for a country to have 3.6 open pit mines, but it IS possible for the mean of a sample of countries to be 3.6 open pit mines.

◇ **As a general rule-of-thumb, present the mean with one more decimal than the number of decimals it was recorded in.**

5.1.4 Standard Deviation

The sample standard deviation, denoted by s, is computed with these six steps:

1. Compute the sample mean (i.e., $\bar{x}$).
2. For each value (x_i), find the difference between the value and the mean (i.e., $x_i - \bar{x}$).
3. Square each difference (i.e., $(x_i - \bar{x})^2$).
4. Add together all the squared differences.
5. Divide this sum by $n - 1$. [*Stopping here gives the sample variance,* s^2.]
6. Square root the result from the previous step to get s.

These steps are neatly summarized with

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \tag{5.1.2}$$

The calculation of the standard deviation of the earthquake data (Table 5.2) is facilitated with the calculations shown in Table 5.3. In Table 5.3, note that

- $\bar{x}$ is the sum of the "Value" column divided by $n = 15$ (i.e., $\bar{x} = 7.07$).
- The "Diff" column is each observed value minus $\bar{x}$ (i.e., Step 2).
- The "Diff^2" column is the square of the differences (i.e., Step 3).
- The sum of the "Diff^2" column is Step 4.
- The sample variance (i.e., Step 5) is equal to this sum divided by $n - 1 = 14$ or $\frac{6.773}{14} = 0.484$.
- The sample standard deviation is the square root of the sample variance or $s = \sqrt{0.484} = 0.696$.

Table 5.3. Table showing an efficient calculation of the standard deviation of the earthquake data.

Indiv i	Value x_i	Diff $x_i - \bar{x}$	Diff^2 $(x_i - \bar{x})^2$
1	5.5	-1.57	2.454
2	6.3	-0.77	0.588
3	6.5	-0.57	0.321
4	6.5	-0.57	0.321
5	6.8	-0.27	0.071
6	6.8	-0.27	0.071
7	6.9	-0.17	0.028
8	7.1	0.03	0.001
9	7.3	0.23	0.054
10	7.3	0.23	0.054
11	7.7	0.63	0.401
12	7.7	0.63	0.401
13	7.7	0.63	0.401
14	7,8	0.73	0.538
15	8.1	1.03	1.068
Sum	106	0	6.773

From this, on average, each earthquake is approximately 0.7 Richter Scale units different than the average earthquake in these data.

◇ In the standard deviation calculations don't forget to take the square root of the variance.

◇ The standard deviation is greater than or equal to zero.

The standard deviation can be thought of as "the average difference between the values and the mean." This is, however, not a strict definition because the formula for the standard deviation does not simply add the differences and divide by n as this definition would imply. Notice in Table 5.3 that the sum of the differences from the mean is 0. This will be the case for all standard deviation calculations using the correct mean, because the mean balances the distance to individuals below the mean with the distance of individuals above the mean (see Section 6.3 in the next module). Thus, the mean difference will always be zero. This "problem" is corrected by squaring the differences before summing them. To get back to the original units, the squaring is later "reversed" by the square root. So, more accurately, the standard deviation is the square root of the average squared differences between the values and the mean. Therefore, "the average difference between the values and the mean" works as a practical definition of the meaning of the standard deviation, but it is not strictly correct.

◇ **Use the fact that the sum of all differences from the mean equals zero as a check of your standard deviation calculation.**

Further note that the mean is the value that minimizes the value of the standard deviation calculation – i.e., putting any other value besides the mean into the standard deviation equation will result in a larger value.

Finally, you may be wondering why the sum of the squared differences in the standard deviation calculation is divided by $n-1$, rather than n. Recall (from Section 2.1) that statistics are meant to estimate parameters. The sample standard deviation is supposed to estimate the population standard deviation (σ). Theorists have shown that if we divide by n, s will consistently underestimate σ. Thus, s calculated in this way would be a biased estimator of σ. Theorists have found, though, that dividing by $n-1$ will cause s to be an unbiased estimator of σ. Being unbiased is generally good – it means that on average our statistic estimates our parameter (this concept is discussed in more detail in Module **??**).

5.1.5 Mode

The mode is the value that occurs most often in a data set. For example, one open pit mine is the mode in the open pit mine data (Table 5.4).

Table 5.4. Frequency of countries by each number of open pit mines.

Number of Mines	1	2	3	4	11	12	15
Freq of Countries	10	6	1	5	2	1	1

The mode for a continuous variable is the class or bin with the highest frequency of individuals. For example, if 0.5-unit class widths are used in the Richter scale data, then the modal class is 6.5-6.9 (Table 5.5).

Table 5.5. Frequency of earthquakes by Richter Scale class.

Richter Scale Class	5.5-5.9	6-6.4	6.5-6.9	7-7.4	7.5-7.9	8-8.4
Freq of Earthquakes	1	1	5	3	4	1

Some data sets may have two values or classes with the maximum frequency. In these situations the variable is said to be **bimodal**.

5.1.6 Range

The range is the difference between the maximum and minimum values in the data and measures the ultimate dispersion or spread of the data. The range in the open pit mine data is 15-1 = 14.

The range should **never be used by itself** as a measure of dispersion. The range is extremely sensitive to outliers and is best used only to show all possible values present in the data. The range (as strictly defined) also suffers from a lack of information. For example, what does a range of 9 mean? It can have a completely different interpretation if it came from values of 1 to 10 or if it came from values of 1000 to 1009. Thus, the range is more instructive if presented as both the maximum and minimum value rather than the difference.

5.1.7 Computation of Summaries in R

All summary statistics described above, with the exception of the mode, is calculated in R with `Summarize()`. To summarize a single variable a one-sided formula of the form `~quant` is used, where `quant` generically

represents the quantitative variable, along with the `data=` argument. The number of digits after the decimal place is controlled with `digits=`.

```
> Summarize(~days,data=LSI,digits=2)
     n nvalid   mean     sd    min     Q1 median     Q3    max
 42.00  39.00 107.85  21.59  48.00  97.00 114.00 118.00 146.00
```

From this it is seen that the sample median is 114 days, sample mean is 107.8 days, sample IQR is from 97 to 118 days, the sample standard deviation is 21.59 days, and the range is from 48 to 146.

5.2 Graphical Summaries

5.2.1 Histogram

A histogram plots the frequency of individuals (y-axis) in classes of values of the quantitative variable (x-axis). Construction of a histogram begins by creating classes of values for the variable of interest. The easiest way to create a list of classes is to divide the range (i.e., maximum minus minimum value) by a "nice" number near eight to ten, and then round up to make classes that are easy to work with. The "nice" number between eight and ten is chosen to make the division easy and will be the number of classes. For example, the range of values in the open pit mine example is 15-1 = 14. A "nice" value near eight and ten to divide this range by is seven. Thus, the classes should be two units wide (=14/7) and, for ease, will begin at 0 (Table 5.6).

Table 5.6. Frequency table of number of countries in two-mine-wide classes.

Class	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15
Frequency	10	7	5	0	0	2	1	1

The frequency of individuals in each class is then counted (shown in the second row of Table 5.6). The plot is prepared with values of the classes forming the x-axis and frequencies forming the y-axis (Figure 5.1A). The first bar added to this skeleton plot has the bottom-left corner at 0 and the bottom-right corner at 2 on the x-axis, and a height equal to the frequency of individuals in the 0 and 1 class (Figure 5.1B). A second bar is then added with the bottom-left corner at 2 and the bottom-right corner at 4 on the x-axis, and a height equal to the frequency of individuals in the 2 and 3 class (Figure 5.1C). This process is continued with the remaining classes until the full histogram is constructed (Figure 5.1D).

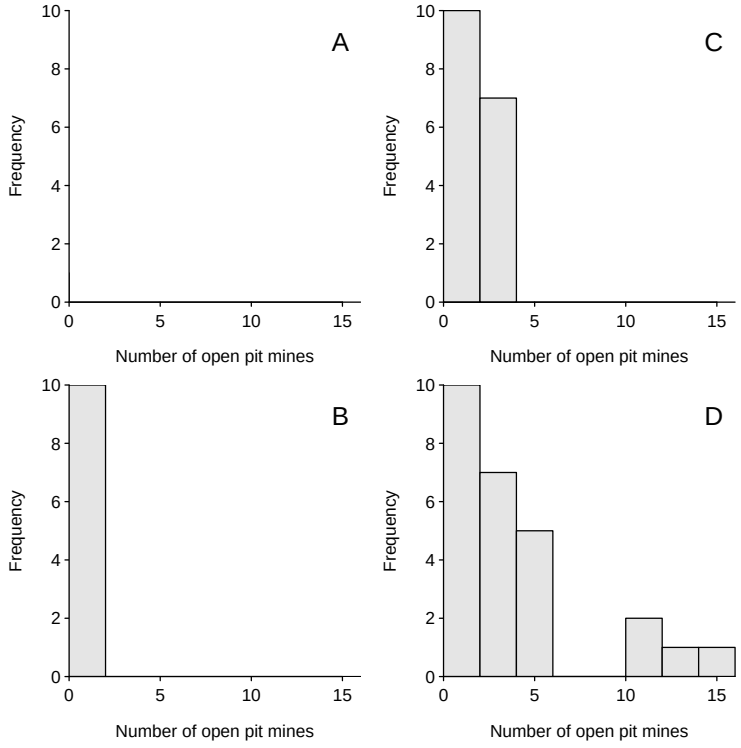

Figure 5.1. Steps (described in text) illustrating the construction of a histogram.

Ideally eight to ten classes are used in a histogram. Too many or too few bars make it difficult to identify the shape and may lead to different interpretations. A dramatic example of the effect of changing the number of classes is seen in histograms of the length of eruptions for the Old Faithful geyser (Figure 5.2).

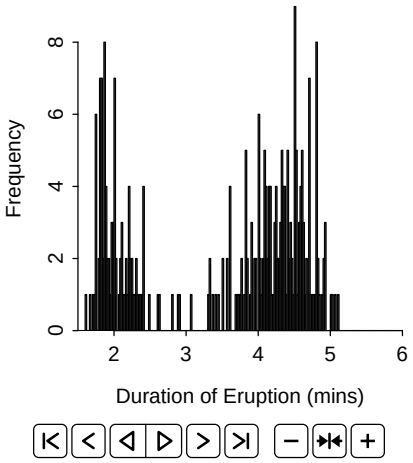

Figure 5.2. Histogram of length of eruptions for Old Faitfhul geyser with varying number of classes.

5.2.2 Boxplot

The **five-number summary** consists of the minimum, Q1, median, Q3, and maximum values (effectively contains the range, IQR, and median). For example, the five-number summary for the open pit mine data is 1, 1, 2, 4, and 15 (all values computed in the previous section). The five-number summary may be displayed as a **boxplot**. A traditional boxplot (Figure 5.3-Left) consists of a horizontal line at the median, horizontal lines at Q1 and Q3 that are connected with vertical lines to form a box, and vertical lines from Q1 to the minimum and from Q3 to the maximum. In modern boxplots (Figure 5.3-Right) the upper line extends from Q3 to the last observed value that is within 1.5 IQRs of Q3 and the lower line extends from Q1 to the last observed value that is within 1.5 IQRs of Q1. Observed values outside of the whiskers are termed "outliers" by this algorithm and are typically plotted with circles or asterisks. If no individuals are deemed "outliers" by this algorithm, then the two traditional and modern boxplots will be the same.

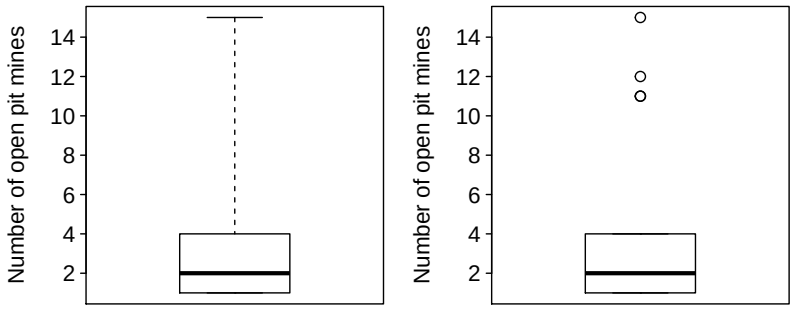

Figure 5.3. Traditional (Left) and modern (Right) boxplots of the open pit mine data.

5.2.3 Construction of Graphs in R

A simple (by default) histogram is constructed in R with `hist()` using a one-sided formula of the form `~quant`, where `quant` generically represents the quantitative variable, and the corresponding data frame in `data=`.[7] The x-axis label may be improved from the default value by including a label in `xlab=`. The width of the classes may be controlled with a positive integer in `w=`.[8]

```
> hist(~days,data=LSI,xlab="Days of Ice Cover")      # Fig 5.4-Left
> hist(~days,data=LSI,xlab="Days of Ice Cover",w=20) # Fig 5.4-Right
```

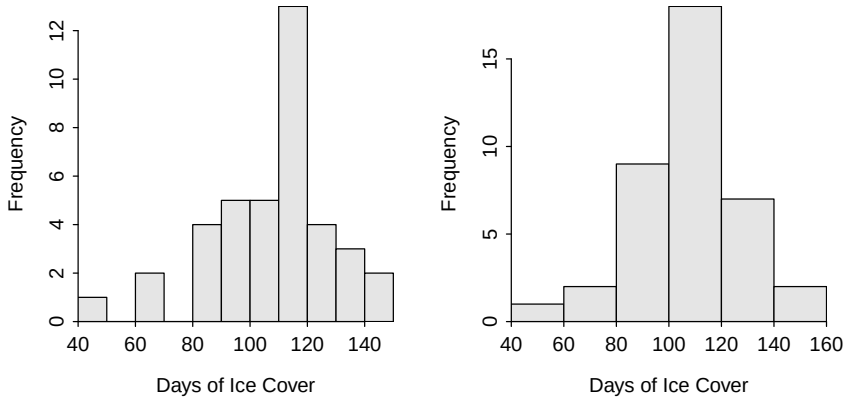

Figure 5.4. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior using the default class widths (Left) and widths of 20 days (Right).

A modern boxplot of a single variable is constructed in R with `boxplot()`, where the first argument is usually a specific variable in a data.frame. Additionally, the y-axis may be properly labeled with `ylab=`.

```
> boxplot(LSI$days,ylab="Days of Ice Cover")
```

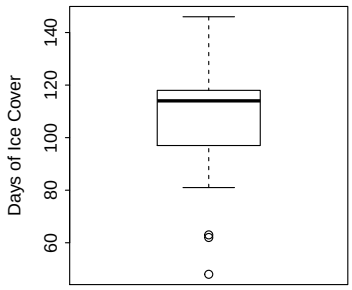

Figure 5.5. Boxplot of the duration of ice cover at ice gauge 9004 in Lake Superior.

[7] Note that this is the same formula used in `Summarize()`.
[8] The endpoints for the classes may also be set by giving a vector of endpoints to `breaks=`.

◇ **The default histogram and boxplot should be modified by properly labeling the axes.**

5.3 Multiple Groups

It is common to need to compute numerical or construct graphical summaries of a quantitative variable separately for groups of individuals. In these cases it is beneficial to have a function that will efficiently construct a histogram and compute summary statistics for the quantitative variable separated by the levels of a factor variable. Separate histograms are constructed with `hist()`, if the first argument is a "formula" of the type `quant~group` where `quant` represents the quantitative response variable of interest and `group` represents the factor variable that indicates to which group the individual belongs. The data.frame that contains `quant` and `group` is given to `data=`. Summary statistics are separated by group by supplying the same formula and `data=` arguments to `Summarize()`.

As an example, the LSI data.frame contains a `period` variable that indicates whether the ice season was "pre-1975" or "post-1975" (which included 1975). Thus, one may be interested in examining the distribution of annual days of ice for each of these periods period. Histograms (Figure 5.6) and summary statistics separated by period are constructed below.

```
> hist(days~period,data=LSI,ylab="Days of Ice Cover",w=20)
> Summarize(days~period,data=LSI,digits=2)
     period  n nvalid   mean    sd min Q1 median  Q3 max
1 post-1975 22     21 106.76 26.01  48 99  116.0 123 146
2  pre-1975 20     18 109.11 15.59  82 97  110.5 118 137
```

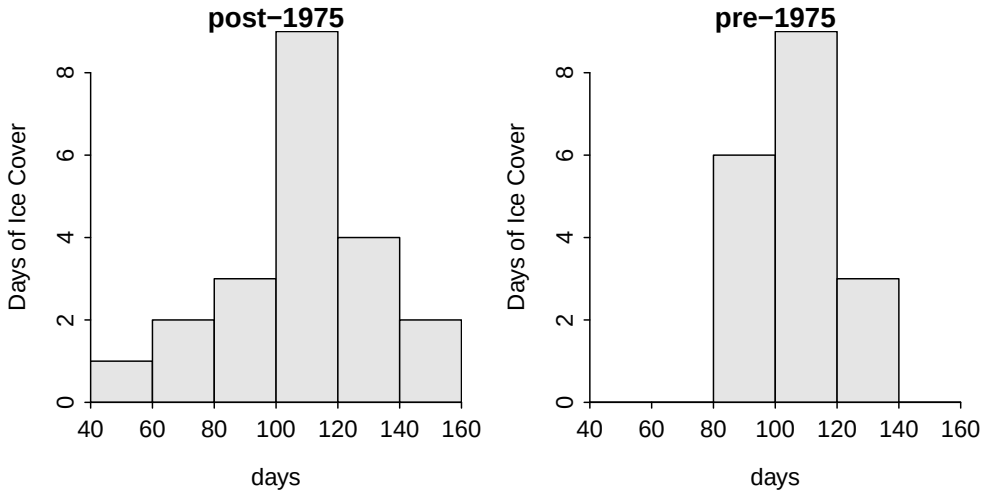

Figure 5.6. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior by period.

Side-by-side boxplots (Figure 5.7) are an alternative to separated histograms and are constructed by including the same formula and `data=` arguments to `boxplot()`.

```
> boxplot(days~period,data=LSI,ylab="Days of Ice Cover",xlab="Period")
```

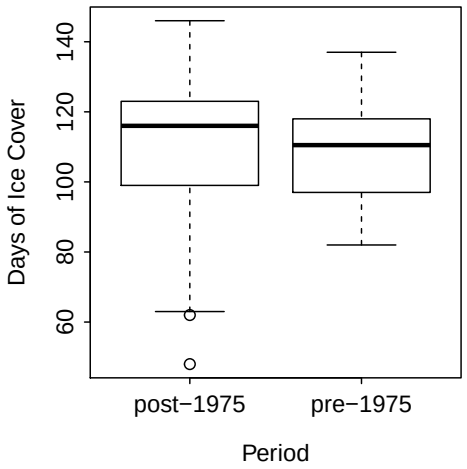

Figure 5.7. Boxplot of the duration of ice cover at ice gauge 9004 in Lake Superior by period.

Note that the formulae above required the grouping variable to be a factor. In some instances, a grouping variable may appear as an integer variable to R. For example, one may want to explore days of ice by decade, but the decade variable is not a factor variable.

```
> str(LSI)
'data.frame': 42 obs. of  5 variables:
 $ season: int  1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 ...
 $ decade: int  1950 1950 1950 1950 1950 1960 1960 1960 1960 1960 ...
 $ period: Factor w/ 2 levels "post-1975","pre-1975": 2 2 2 2 2 2 2 2 2 2 ...
 $ temp  : num  22.9 23 25.7 20 24.8 ...
 $ days  : int  87 137 106 97 105 118 118 136 91 NA ...
```

In these cases, the variale needs to be explicitly converted to a factor variable using `factor()`, as shown below. The use of `factor()` is not needed if R already recognizes the variable as a factor variable.

```
> LSI$decade <- factor(LSI$decade)
> str(LSI)
'data.frame': 42 obs. of  5 variables:
 $ season: int  1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 ...
 $ decade: Factor w/ 5 levels "1950","1960",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ period: Factor w/ 2 levels "post-1975","pre-1975": 2 2 2 2 2 2 2 2 2 2 ...
 $ temp  : num  22.9 23 25.7 20 24.8 ...
 $ days  : int  87 137 106 97 105 118 118 136 91 NA ...
```