Chapter 3

Populations and Statistics

This chapter covers two topics that are fundamental in statistics. The first is the concept of a statistical population, which is the basic unit on which statistics are conducted and inferences made. We then examine descriptive statistics and frequency distributions, which are used quantify the properties of samples from a statistical population.

3.1 Statistical populations

Suppose we want to estimate the body length of an insect species in a particular location, say a forest stand. We sample the insects in some way (traps, sweep nets, locate them visually, etc.), and average their lengths to obtain an estimate of insect length. We can therefore make some inference about insect lengths in this particular forest stand, which we can call a **statistical population**. A statistical population is defined by both the question of interest (insect length) as well as the sampling method. If we sample insects in only a single forest stand, then the statistical population is length in that stand, not other stands. This is commonly called the **scope of inference** of the study. If we sampled within multiple stands in a forest, then we could potentially examine length for the forest as whole, which would be a different statistical population and the scope of inference would be broader. The sampling technique itself can also affect the statistical population. For example, only a subset of insects might be caught with sweep nets (maybe slower, smaller ones) and this would be a different set than those found visually. The two sampling techniques might therefore define different statistical populations.

Biologists are continually searching for better methods of sampling organisms, ones that better represent their true properties. In many cases the idea is to approximate what is known as **random sample** of the statistical population (see Chapter 8).

In the insect length example above, the statistical population coincides with individual insects in a location. However, the quantities comprising a statistical population can be other quantities. For example, suppose we want to estimate the abundance of these insects using traps. We could deploy several traps in the stand, and then average the number of insects caught to estimate their abundance. The statistical population in this case would consist of number of insects caught in traps deployed at that location, rather than individual insects. Or one might be interested in soil nitrogen levels in the stand, estimated using core samples. In this case, the statistical population would be the nitrogen levels in core samples at this location.

Another type of statistical population involves experiments. Suppose we are interested in trapping the same insects in the forest stand, but now have traps baited with different attractants, say A, B, and C. Several traps are baited with each attractant, and the number of insects caught observed for each trap. We are interested in whether the number of insects caught varies with the attractant used. In this case, the statistical population would be trap catches for the different attractants. Similarly, suppose we were interested in the effect of different commercial diets on the growth rate of fish. Different fish would be fed the various diets and their growth rate observed. Here the statistical population would be the growth rate of individual fish for the different diets. Experiments also have a scope of inference. If we use four particular diets to grow fish, our conclusions are restricted to these four diets and not other diets. If the experiment used a particular strain of fish, our inferences would also be restricted to this strain.

3.2 Descriptive statistics and frequency

Given a sample from a statistical population, the first step in understanding its properties is to calculate a number of descriptive statistics. Some statistics give you an idea of the overall magnitude or location of the data, and are traditionally called **statistics of location**. We will examine two such statistics, the sample mean and the median. Other statistics give an indication of the scatter or spread of the data, and are called **statistics of**

dispersion. These include the sample variance, standard deviation, the coefficient of variation, and range of the data. Another important tool is the **frequency distribution** of the sample, often plotted as a histogram indicating the frequency of different values in the sample. Three other statistics, the mode, skewness, and kurtosis, provide information on the shape of this frequency distribution.

To illustrate how the various descriptive statistics are calculated, we will use a small subset of a larger data set on the elytra length for a predatory beetle, *Thanasimus dubius* (Coleoptera: Cleridae). This predator attacks insects known as bark beetles, some species of which are serious pests of coniferous forests (Berryman 1988). Beetles have two pairs of wings. The first pair, the elytra, act as covers for a membraneous second pair that are used in flight. The data are drawn from a rearing study of *T. dubius*, in which elytra length (mm) was used as a overall index of body size (Reeve et al. 2003). The subset data are for eight female *T. dubius* and are listed below:

```
5.2 4.2 5.7 5.4 4.0 4.5 5.2 4.2
```

We will later examine the full data set consisting of 130 individuals using SAS programs.

3.2.1 Sample mean

The sample mean is the average of the values in the sample, and is symbolized as $\bar{Y}$. It is commonly used as a measure of the location or center of the observations. If $Y_1, Y_2, \ldots, Y_n$ represent the observations in a sample from a statistical population, where n is the sample size, the sample mean is calculated using the formula

$$\bar{Y} = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{3.1}$$

The symbol $\sum_{i=1}^{n}$ stands for summing the observations, beginning with $i = 1$ and ending with $i = n$. The units of $\bar{Y}$ are the same as those for the Y_i values.

For our sample data set involving $n = 8$ elytra from female *T. dubius* beetles, we have

$$\bar{Y} = \frac{5.2 + 4.2 + 5.7 + 5.4 + 4.0 + 4.5 + 5.2 + 4.2}{8} = \frac{38.4}{8} = 4.8 \text{ mm}. \tag{3.2}$$

3.2.2 Median

The median is defined as the middle value of the sample, after ordering the sample from the smallest to the largest value. Suppose that $Y_{[j]}$ is the jth value in the ordered data set, with $Y_{[1]}$ the smallest value and $Y_{[n]}$ the largest. If n is odd, the median is equal to the middle value in the ordered data set, or $Y_{[n/2+1/2]}$. If n is even then the median is the average of the two middle values, or $(Y_{[n/2]} + Y_{[n/2+1]})/2$.

To find the median for the elytra data set, we first order the observations from smallest to largest. We have

j (order):	1	2	3	4	5	6	7	8
$Y_{[j]}$:	4.0	4.2	4.2	4.5	5.2	5.2	5.4	5.7

Because $n = 8$ is even, the median is the average of the middle two observations, or $(Y_{[n/2]} + Y_{[n/2+1]})/2 = (Y_{[8/2]} + Y_{[8/2+1]})/2 = (Y_{[4]} + Y_{[5]})/2 = (4.5 + 5.2)/2 = 4.85$.

Suppose now we had only $n = 7$ observations, with the ordered data set equal to

j (order):	1	2	3	4	5	6	7
$Y_{[j]}$:	4.0	4.2	4.2	4.5	5.2	5.2	5.4

Because $n = 7$ is odd, the median is the middle observation, or $Y_{[n/2+1/2]} = Y_{[7/2+1/2]} = Y_{[4]} = 4.5$ mm.

The median is also a measure of the location of the data, like the sample mean $\bar{Y}$, but is less sensitive to very large or small values in the sample. For example, suppose that the largest observation in the elytra data set was 100.0. The median would be unchanged because the ordering of the observations is unchanged, but now $\bar{Y} = 16.8$ mm, much larger than before.

The median represents a value that essentially divides the data in half, with 50% of the observations lying above or below it. This is an example of a statistic generically called **quantiles** or **percentiles**, with the median a 50% quantile. Other commonly used quantiles are the 25% and 75% quantiles. They and the median are sometime called **quartiles** because they divide the data into four quarters.

3.2.3 Sample variance

The sample variance, written as s^2, is a measure of the dispersion or scatter in the data around the sample mean. It is calculated using the formula

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \tag{3.3}$$

The sample variance s^2 will be small if the observations cluster tightly around $\bar{Y}$, because this makes $(Y_i - \bar{Y})^2$ small. Conversely, if the observations are widely scattered these terms will be large, making s^2 large. The units of s^2 are those of Y_i, but squared.

To find s^2 for the elytra data set, we first need to calculate the sample mean. We previously found that $\bar{Y} = 4.8$. We then calculate s^2 using the above formula. We have

$$s^2 = \frac{(5.2-4.8)^2 + (4.2-4.8)^2 + \ldots + (4.2-4.8)^2}{8-1} \tag{3.4}$$

$$= \frac{0.16 + 0.36 + 0.81 + 0.36 + 0.64 + 0.09 + 0.16 + 0.36}{7} \tag{3.5}$$

$$= \frac{2.94}{7} = 0.42 \text{ mm}^2. \tag{3.6}$$

3.2.4 Standard deviation

The sample standard deviation, written as s, is simply the square root of s^2. We have

$$s = \sqrt{s^2} \tag{3.7}$$

For the elytra example, we have $s = \sqrt{s^2} = \sqrt{0.42} = 0.645$ mm. The units of s are the same as those of Y_i, which makes it more comparable to statistics of location like $\bar{Y}$.

3.2.5 Coefficient of variation

The coefficient of variation, or CV, provides a measure of the variability of the observations expressed as a percentage of the sample mean. It is calculated using the formula

$$CV = 100\% \times \frac{s}{\bar{Y}}. \tag{3.8}$$

The CV allows one to compare the variability of observations on variables that have different means. For example, suppose that we want to compare variability in *T. dubius* elytra length with variability in another predator that has a longer overall length. For biological variables like length, the standard deviation s often seems proportional to the sample mean $\bar{Y}$. If we divide s by $\bar{Y}$, as in the CV, we can control to some extent the influence of $\bar{Y}$ on variability. This allows us to compare variability in length across the two predators on a more even basis.

3.2.6 Range

The range is defined as the difference between the largest and smallest observations, i.e.,

$$\text{range} = Y_{\max} - Y_{\min}, \tag{3.9}$$

where $Y_{\max}$ is the largest observation and $Y_{\min}$ is the smallest. For the elytra data, we have $Y_{\max} = 5.7$ and $Y_{\min} = 4.0$, so

$$\text{range} = 5.7 - 4.0 = 1.7. \tag{3.10}$$

The range is another statistic of dispersion, but has some problems. The range tends to increase in size as the sample size n increases, because larger samples are more likely to yield very small or large observations. This is not the case for s^2 or s.

3.2.7 Frequency distributions - SAS demo

Frequency distributions are another way of summarizing and describing a sample from a statistical population. They typically take the form of a histogram showing the frequency of different observations in the sample. We will use SAS to construct frequency distributions as well as calculate descriptive statistics like $\bar{Y}$, s^2, and so forth. We will use the full elytra data set for *T. dubius* (Reeve et al. 2003) to illustrate these calculations. This data set contains both male and female beetles, and we will conduct separate analyses for each sex. See also Chapter 21.

The program first uses a `data` step to read in the observations and make a data file (SAS Institute Inc. 2014a). The line

```
data elytra;
```

tells SAS to set up a data file named `elytra`. If you omit a name from this statement, SAS will automatically generate one for you. The line

```
input sex $ length;
```

tells SAS to read in two variables and give them the names `sex` and `length`. It also tells SAS to expect the data in the form of two columns. The `$` symbol after `sex` tells SAS that it is a character variable, consisting of a word or letters rather than a number. The default is for a numeric variable. The line

```
datalines;
```

tells SAS that following lines in the program are the actual data. The program then lists the data, followed by another semicolon and then a `run` statement (see below). The full data set is not listed here because it is extensive (see Chapter 21, Section 21.1). The `run` statement tells SAS the `data` step is over, and also that it should process the data and generate a SAS data file.

```
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
```

We are now ready to do something with our newly minted SAS data file, named `elytra`. It is usually a good idea just to print the data file to make sure SAS correctly read the data. This is accomplished using the `proc print` code listed below.

```
* Print data set;
proc print data=elytra;
run;
```

The final lines of the SAS program invoke `proc univariate` to generate the histogram and calculate a number of descriptive statistics (SAS Institute Inc. 2014b). The first and third lines are comments. The second line tells SAS to call `proc univariate` and requests that certain plots be made using the `plots` option. The `class` statement tells the procedure to conduct a separate analysis for each sex in the data set, while the `var` statements tells it which variable to analyze, in this case the variable `length`. The `histogram` statement asks for a histogram of `length`, with the statements after the forward slash (`/`) being options for the graph. The option `vscale=count` tells SAS to make the vertical axis using counts of the observations (the default uses percentages). The remaining options control the width of the lines in the graph as well as text height. The program would work without these options but would generate a different-looking histogram.

```
* Descriptive statistics and histograms;
proc univariate plots data=elytra;
    * Separate analyses for each sex;
    class sex;
    var length;
    histogram length / vscale=count wbarline=3 waxis=3 height=4;
run;
quit;
```

After running the program, we obtain output with various statistics of location and dispersion, including the sample mean, median range, variance, and standard deviation, as well as a graph showing the frequency distribution. A separate analysis is generated for each sex (`M` or `F`) of the beetles. We see that females have somewhat longer elytra than males ($\bar{Y} = 4.940$ mm vs. 4.713 mm), and there are small differences in other statistics. See a complete program listing below, and SAS output with some editing to reduce its length.

SAS Program

```
* descriptive.sas;
options pageno=1 linesize=80;
title 'Descriptive statistics for the elytra data';
data elytra;
    input sex $ length;
    datalines;
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
* Print data set;
proc print data=elytra;
run;
* Descriptive statistics and histograms;
proc univariate plots data=elytra;
    * Separate analyses for each sex;
    class sex;
    var length;
    histogram length / vscale=count wbarline=3 waxis=3 height=4;
run;
quit;
```

SAS Output

Descriptive statistics for the elytra data 1
09:32 Tuesday, May 18, 2010

Obs	sex	length
1	M	4.9
2	F	5.2
3	M	4.9
4	F	4.2
5	F	5.7

etc.

Descriptive statistics for the elytra data 4
09:32 Tuesday, May 18, 2010

The UNIVARIATE Procedure
Variable: length
sex = F

Moments

N	60	Sum Weights	60
Mean	4.94	Sum Observations	296.4
Std Deviation	0.48544929	Variance	0.23566102
Skewness	-0.521146	Kurtosis	0.16125847
Uncorrected SS	1478.12	Corrected SS	13.904
Coeff Variation	9.82690878	Std Error Mean	0.06267123

Basic Statistical Measures

Location		Variability	
Mean	4.940000	Std Deviation	0.48545
Median	5.000000	Variance	0.23566
Mode	5.200000	Range	2.20000
		Interquartile Range	0.70000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	78.82404	Pr > |t|	<.0001
Sign	M	30	Pr >= |M|	<.0001
Signed Rank	S	915	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	5.9
99%	5.9
95%	5.7
90%	5.5
75% Q3	5.3
50% Median	5.0
25% Q1	4.6
10%	4.3
5%	4.0
1%	3.7
0% Min	3.7

Descriptive statistics for the elytra data 7
09:32 Tuesday, May 18, 2010

The UNIVARIATE Procedure
Variable: length
sex = M

Moments

N	70	Sum Weights	70
Mean	4.71285714	Sum Observations	329.9
Std Deviation	0.44977335	Variance	0.20229607
Skewness	-0.896502	Kurtosis	1.00307174
Uncorrected SS	1568.73	Corrected SS	13.9584286
Coeff Variation	9.5435388	Std Error Mean	0.0537582

Basic Statistical Measures

Location Variability

Mean	4.712857	Std Deviation	0.44977
Median	4.800000	Variance	0.20230
Mode	5.000000	Range	2.40000
		Interquartile Range	0.50000

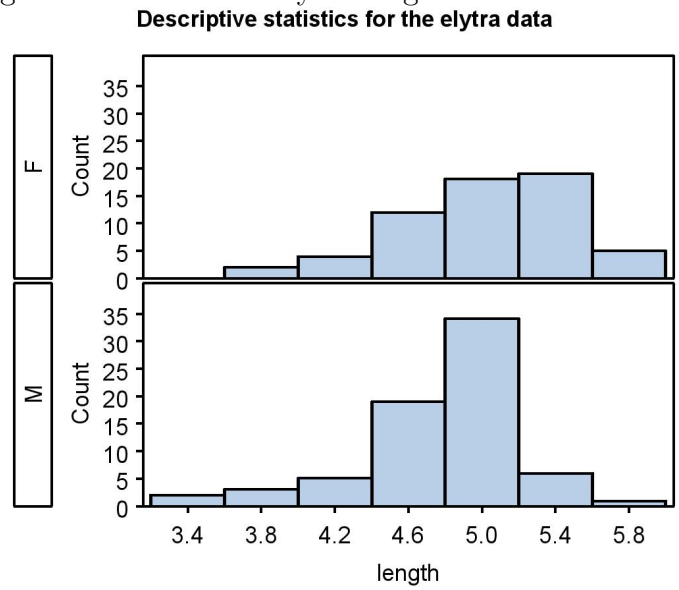
Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	87.66769	Pr > |t|	<.0001
Sign	M	35	Pr >= |M|	<.0001
Signed Rank	S	1242.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	5.80
99%	5.80
95%	5.20
90%	5.15
75% Q3	5.00
50% Median	4.80
25% Q1	4.50
10%	4.00
5%	3.80
1%	3.40
0% Min	3.40

Figure 3.1: *T. dubius* elytra length - females and males



3.2.8 Mode

The mode is defined to be the most frequent value in the data set, and is another statistic of location. The mode in itself does not have many applications in biology, but is commonly used to describe the shape of a frequency distribution for the sample (see above). For example, we describe a frequency distribution as being unimodal if it has a single peak, and bimodal if there are two peaks. Examining the SAS output listed above, we see that female *T. dubius* beetles have a mode of 5.2 mm, while the mode for males is 5.0 mm. Both distributions appear to be unimodal.

3.2.9 Skewness

Skewness is a measure of the symmetry of the frequency distribution. Distributions that show an extended left tail to the frequency distribution, as well as the pattern mode $>$ median $>$ mean, are said to be skewed to the left. Fig. 3.2 shows an example of a left-skewed frequency distribution for some

variable `y`. Conversely, distributions with an extended right tail and the pattern mean > median > mode are skewed to the right (Fig. 3.3). Skewness can be quantified by calculating the statistic g_1, given by the formula

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{s} \right)^3 . \tag{3.11}$$

The cubic terms here measure the asymmetry of the distribution. If the distribution is skewed to the left, with more values farther to the left than the right of $\bar{Y}$, there will tend to be large negative cubic terms, making $g_1 < 0$. Conversely, distributions skewed to the right will have large positive cubic terms and $g_1 > 0$. For distributions that are symmetrical we have $g_1 \approx 0$. For example, a frequency distribution for normally-distributed data would be symmetrical with $g_1 \approx 0$ (Fig. 3.4). For the elytra example, both male and female *T. dubius* have frequency distributions that appear skewed to the left, and also have negative g_1 values. Skewness is most often used as a description of the general shape of a distribution.

Figure 3.2: Frequency distribution that is skewed left ($g_1 < 0$).

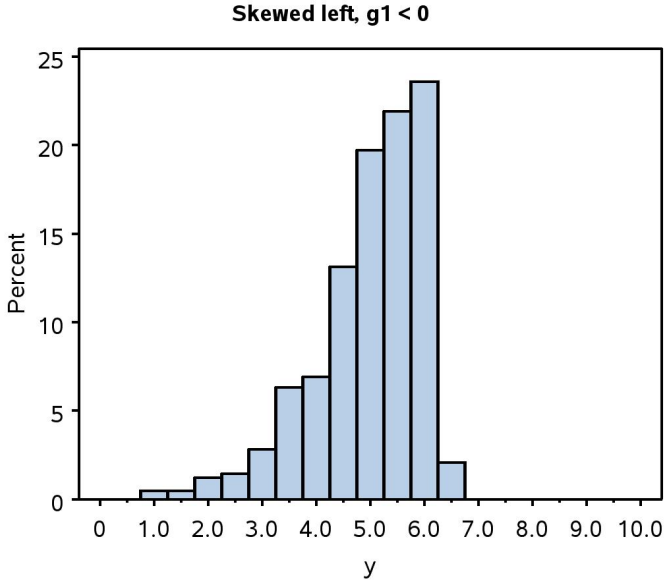


Figure 3.3: Frequency distribution that is skewed right ($g_1 > 0$).

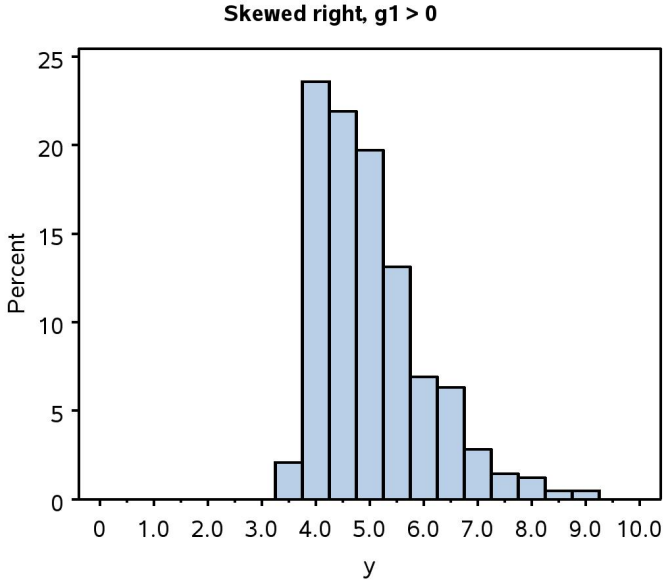
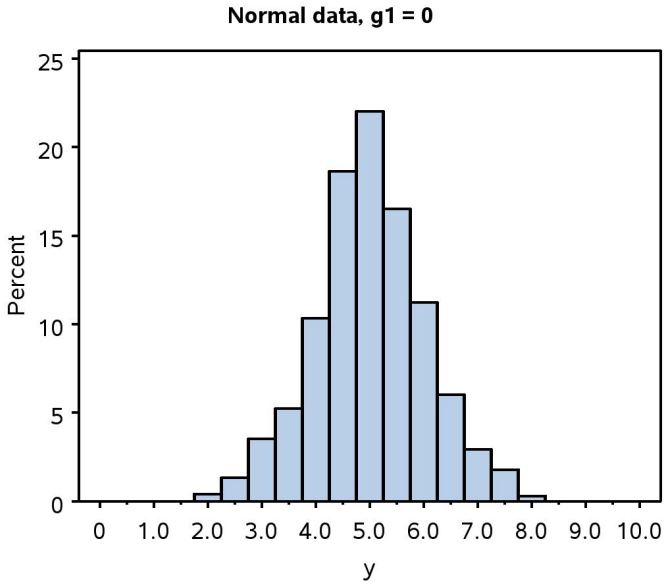


Figure 3.4: Frequency distribution for normal data ($g_1 \approx 0$).



3.2.10 Kurtosis

Kurtosis is a measure of how peaked or flat is a frequency distribution relative to the normal distribution. Distributions with a stronger central peak than the normal, and heavier left and right tails, are called leptokurtic (compare Fig. 3.5 and 3.6). Conversely, distributions with a weak peak and tails are called platykurtic (see Fig. 3.7 vs. 3.6). Kurtosis is quantified by calculating the statistic g_2:

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}. \qquad (3.12)$$

The behavior of the terms in g_2 is less intuitive than those in the skewness statistic g_1. In any event, distributions that are leptokurtic have values of $g_2 > 0$, while platykurtic ones have $g_2 < 0$, with $g_2 \approx 0$ for distributions resembling the normal. For the elytra example, male *T. dubius* have a leptokurtic distribution with $g_2 = 1.003$, and the frequency distribution shows a strong central peak with heavy tails. The value of $g_2 = 0.161$ is smaller for female *T. dubius*, suggesting a shape more similar to the normal distribution. Like skewness, kurtosis is used to describe the general shape of the distribution.

Figure 3.5: Frequency distribution that is leptokurtic ($g_2 > 0$).

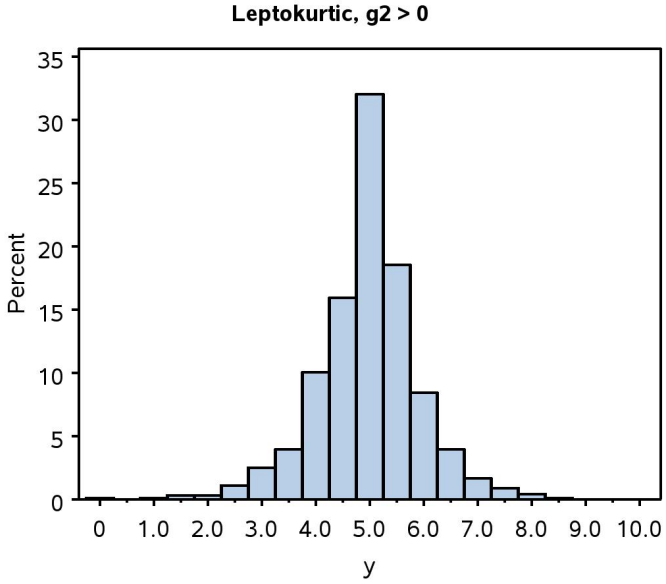


Figure 3.6: Frequency distribution for normal data ($g_2 \approx 0$).

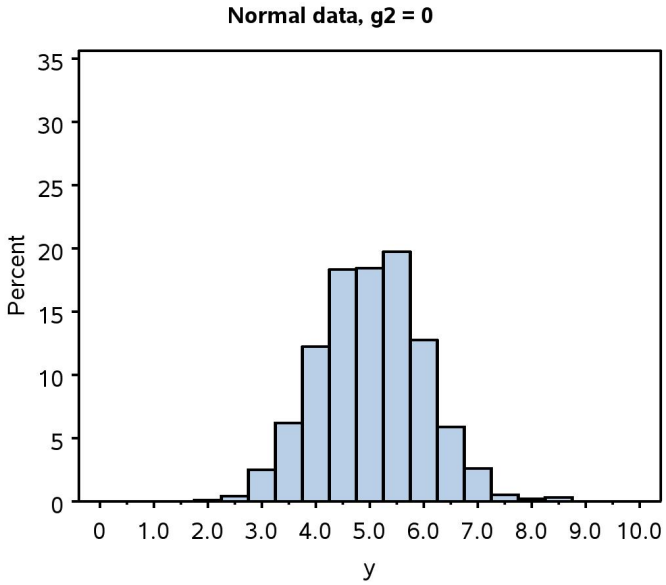
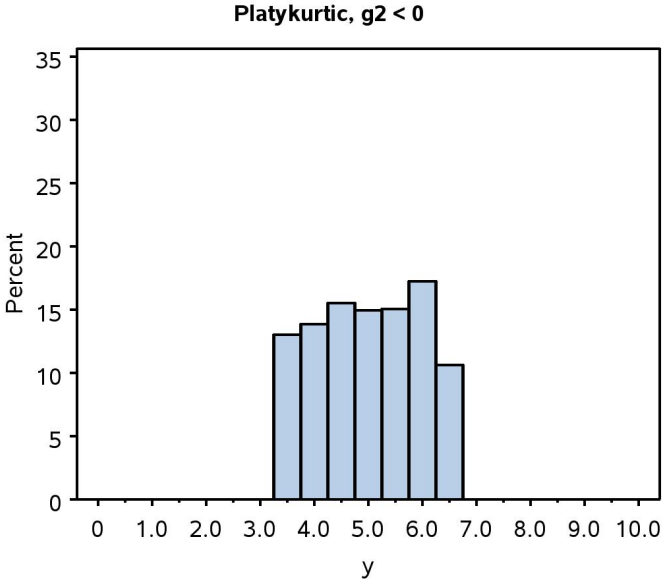


Figure 3.7: Frequency distribution that is platykurtic ($g_2 < 0$).



3.2.11 Development time - SAS demo

We now examine another data set involving the development time of *T. dubius* reared under laboratory conditions (Reeve et al. 2003). Two different development times were measured, the time from the first larval stage until the prepupal stage, and the prepupal to adult stage. The program used to analyze these data is listed below. The `input` line is different than our previous program, because there are two variables (`time_pp` and `time_adult`) to analyze for each insect listed, which occur in two columns. The `var` and `histogram` statements in `proc univariate` are similar, listing the two variables so that descriptive statistics and frequency distributions are generated for both.

Note the periods (. values) given in the data set - these indicate missing values to SAS. In this study, observations were missing usually because the insect died before reaching the adult stage, but missing values can also be used to indicate lost data. The full data set for this example is listed in Chapter 21, Section 21.2.

After running the program, we obtain output with statistics of location and dispersion as well as a frequency distribution, with a separate analysis for each variable. Clearly the larval-prepupal development time (`time_pp`) is shorter than the prepupal adult (`time_adult`) one ($\bar{Y} = 31.354$ vs. 75.353 days), and also shows less variability as indicated by the sample standard deviation ($s = 3.328$ vs. 26.347 days). Both variables appear to be skewed to the right, as indicated by positive values of g_1 as well as the result that mean > median > mode. Larval-prepupal development time shows little kurtosis ($g_2 = 0.047$), while prepupal-adult time apparently has a platykurtic distribution ($g_2 = -0.624$). This can also be observed in the frequency distribution for this variable, which is relatively flat in shape.

SAS Program

```
* descriptive_2.sas;
options pageno=1 linesize=80;
title 'Descriptive statistics for the development data';
data devel_time;
    input time_pp time_adult;
    datalines;
34  65
31  48
29   .
30  55
32  62

etc.

29   .
29 108
31 103
33   .
29  92
;
run;
* Print data set;
proc print data=devel_time;
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate plots data=devel_time;
    var time_pp time_adult;
    histogram time_pp time_adult / vscale=count wbarline=3 waxis=3 height=4;
run;
quit;
```

SAS Output

Descriptive statistics for the development data 1
13:44 Tuesday, May 18, 2010

Obs	time_pp	time_adult
1	34	65
2	31	48
3	29	.
4	30	55
5	32	62

etc.

Descriptive statistics for the development data 3
13:44 Tuesday, May 18, 2010

The UNIVARIATE Procedure
Variable: time_pp

Moments

N	96	Sum Weights	96
Mean	31.3541667	Sum Observations	3010
Std Deviation	3.32764866	Variance	11.0732456
Skewness	0.75038358	Kurtosis	0.04666776
Uncorrected SS	95428	Corrected SS	1051.95833
Coeff Variation	10.6130987	Std Error Mean	0.33962672

Basic Statistical Measures

Location		Variability	
Mean	31.35417	Std Deviation	3.32765
Median	31.00000	Variance	11.07325
Mode	30.00000	Range	14.00000
		Interquartile Range	5.00000

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value------

Student's t	t	92.31949	Pr > |t|	<.0001
Sign	M	48	Pr >= |M|	<.0001
Signed Rank	S	2328	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	41
99%	41
95%	39
90%	36
75% Q3	34
50% Median	31
25% Q1	29
10%	27
5%	27
1%	27
0% Min	27

Descriptive statistics for the development data 6
13:44 Tuesday, May 18, 2010

The UNIVARIATE Procedure
Variable: time_adult

Moments

N	68	Sum Weights	68
Mean	75.3529412	Sum Observations	5124
Std Deviation	26.3465791	Variance	694.14223
Skewness	0.51461555	Kurtosis	-0.6244048
Uncorrected SS	432616	Corrected SS	46507.5294
Coeff Variation	34.9642346	Std Error Mean	3.19499201

Basic Statistical Measures

Location		Variability	
Mean	75.35294	Std Deviation	26.34658
Median	68.00000	Variance	694.14223

Mode	42.00000	Range	105.00000
		Interquartile Range	46.50000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	23.5847	Pr > |t|	<.0001
Sign	M	34	Pr >= |M|	<.0001
Signed Rank	S	1173	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	147.0
99%	147.0
95%	116.0
90%	110.0
75% Q3	99.0
50% Median	68.0
25% Q1	52.5
10%	43.0
5%	42.0
1%	42.0
0% Min	42.0

Figure 3.8: Development time - larval to prepupal stage

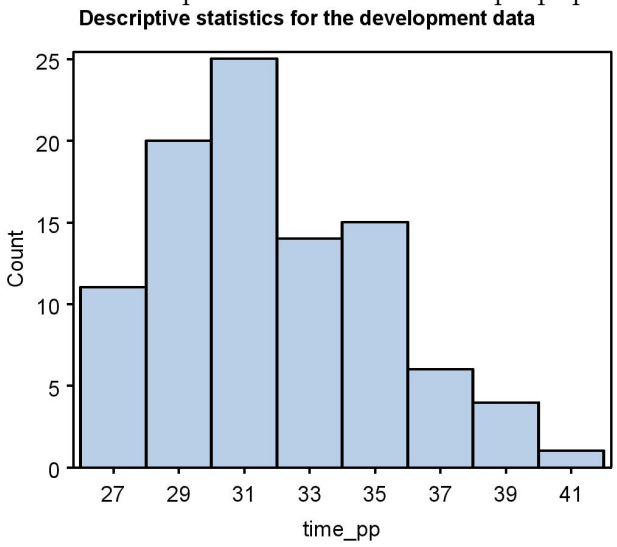
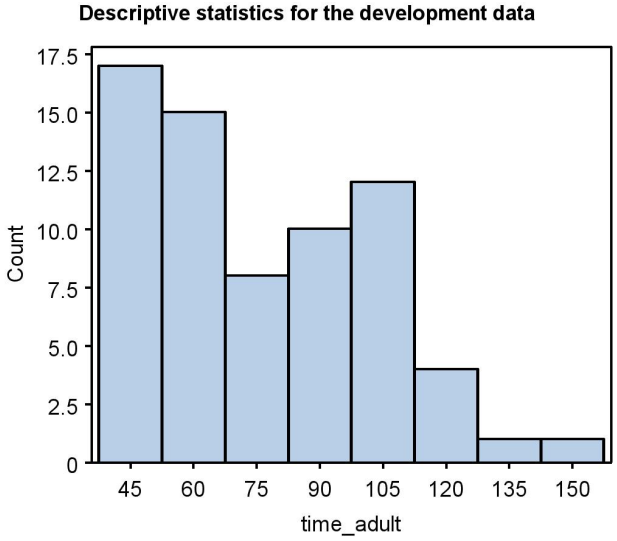


Figure 3.9: Development time - prepupal to adult stage



3.2.12 Frequency distributions for categorical data - SAS demo

The descriptive statistics we have developed so far are appropriate for continuous or discrete data. What about categorical data? One common way of summarizing categorical data is a frequency distribution, showing the number of occurrences in each category and possibly also their percentages. We can illustrate this process using the elytra data. There is one categorical variable in this data set, the sex of the beetle, and we might be interested in whether there were equal numbers of males and females. It also possible to derive categorical variables from the observations themselves. Suppose we classify a beetle as being 'small' if `length` is less than 5.0 mm, and 'large' otherwise. We can define this new variable within the SAS `data` set using an `if-then-else` statement. The code necessary to generate this new variable for the elytra data is shown below. It generates a new variable called `size` that takes the value `small` or `large` depending on the value of `length`.

```
* descriptive_freq.sas;
options pageno=1 linesize=80;
title 'Frequency distribution for the elytra data';
data elytra;
        input sex $ length;
        * Classify insects into two groups by size;
        if length < 5.0 then size="small"; else size="large";
        datalines;
M        4.9
F        5.2
M        4.9
F        4.2
F        5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
```

We can then generate a frequency distribution for both `sex` and `size` using `proc freq` (SAS Institute Inc. 2014b). The `tables sex*size` statement will generate a two-way table of frequencies, classifying each observation into one of four categories (female-large, female-small, male-large, male-small). See below.

```
* Frequency distribution;
proc freq data=elytra;
        table sex*size;
run;
```

The complete program and output are listed below. From the frequency table generated by `proc freq`, we see that there are more males than females in the data set, and more small vs. large insects. Female beetles have a greater proportion of large insects than males.

SAS Program

```
* descriptive_freq.sas;
options pageno=1 linesize=80;
title 'Frequency distribution for the elytra data';
data elytra;
        input sex $ length;
        * Classify insects into two groups by size;
        if length < 5.0 then size="small"; else size="large";
        datalines;
M        4.9
F        5.2
M        4.9
F        4.2
F        5.7

etc.

M        5.1
F        4.4
M        4.8
M        4.6
F        3.7
;
run;
* Print data set;
proc print data=elytra;
run;
* Frequency distribution;
proc freq data=elytra;
        table sex*size;
run;
quit;
```

SAS Output

Frequency distribution for the elytra data 1
09:37 Wednesday, August 18, 2010

Obs	sex	length	size
1	M	4.9	small
2	F	5.2	large
3	M	4.9	small
4	F	4.2	small
5	F	5.7	large

etc.

Frequency distribution for the elytra data 4
09:37 Wednesday, August 18, 2010

The FREQ Procedure

Table of sex by size

sex / size	large	small	Total
F	31	29	60
	23.85	22.31	46.15
	51.67	48.33	
	56.36	38.67	
M	24	46	70
	18.46	35.38	53.85
	34.29	65.71	
	43.64	61.33	
Total	55	75	130
	42.31	57.69	100.00

3.3 References

Berryman, A. A. (1988) *Dynamics of Forest Insect Populations: Patterns, Causes, Implications*. Plenum Press, New York, NY.

Lei, C.-H. & Armitage, K. B. (1980) Growth, development and body size of field and laboratory population of *Daphnia ambigua*. *Oikos* 35: 31-48.

Reeve, J. D., Rojas, M. G. & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

SAS Institute Inc. (2014a) *SAS 9.4 Language Reference: Concepts, Third Edition.* SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc. (2014b) *Base SAS 9.4 Procedures Guide: Statistical Procedures, Third Edition.* SAS Institute Inc., Cary, NC, USA.

3.4 Problems

1. For the data below, find the mean, median, variance, standard deviation and CV using the formulas for these quantities and a calculator. Show the steps in your calculations. Feel free to check your answers using SAS.

```
88.6  88.0  89.8  92.0  108.1  113.6  103.4  109.9  94.5  96.7  101.7
```

2. Ten adult females of the zooplankton species *Daphnia ambigua* were selected and their carapace length measured (μm) (Lei & Armitage 1980). The following data were obtained:

```
487 429 428 378 410 401 358 392 414 480
```

   Calculate the mean, median, variance, standard deviation, and CV for these data by hand. Show all your calculations. Check your answers using SAS.

3. A laboratory study was conducted on the development time of another bark beetle predator, *Temnochila virescens* (Coleoptera: Trogositidae). The numbers listed below are the larval development time (days) of 35 insects.

```
 73  65  58  54  78  57  90
103  59  52  73  67  67  53
 59  55  58  78  64  60  52
 96  68  81  76  77  57  79
 71  74  65  65  64  56  62
```

   (a) Use SAS to find the mean, median, mode, variance, standard deviation, and CV of these data, then plot a frequency distribution. Attach your program, output, and graph.

   (b) Examine the frequency distribution and skewness value (g_1) for these data. Do the data appear to be skewed, and if so in what direction? Explain your answer.