

February 22, 2009

Chapter 2: Descriptive Statistics

In order to determine what a set of data has to tell us, we must extract the information in the data. Typically a first step in a statistical analysis of data is to simply describe the data using descriptive statistics and *look* at the data using appropriate graphing techniques.

In this chapter we will introduce some basic descriptive statistics (means, medians, standard deviations, etc.) and some common graphical techniques for looking at the data.

We shall use SAS to obtain the desired statistics and graphs.

To illustrate ideas, we return to our previous example with pelican eggs:

Example. The PCB concentration (ppm) and thickness (mm) of 65 Anacapa pelican eggs (Risebrough 1972) were measured and the values are listed below:

452	0.14	204	0.28	261	0.34	199	0.46
184	0.19	89	0.28	150	0.34	216	0.46
315	0.20	320	0.28	143	0.35	236	0.47
115	0.20	138	0.29	229	0.35	206	0.49
139	0.21	198	0.29	132	0.36	237	0.49
177	0.22	191	0.29	175	0.36		
214	0.22	193	0.29	173	0.36		
356	0.22	316	0.29	220	0.37		
246	0.23	265	0.29	212	0.37		
177	0.23	122	0.30	236	0.37		
289	0.23	305	0.30	119	0.39		
166	0.23	203	0.30	144	0.39		
175	0.24	396	0.30	147	0.39		
296	0.25	250	0.30	171	0.40		
205	0.25	230	0.30	232	0.41		
260	0.26	214	0.30	216	0.41		
188	0.26	256	0.31	164	0.42		
208	0.26	46	0.31	185	0.42		
324	0.26	204	0.32	216	0.42		
109	0.27	218	0.34	87	0.44		

Measures of Location

What is the central or typical value for the eggshell thicknesses? What we would like is a measure of “central location”. If we plot the eggshell thicknesses on a number line, what is the central location of the thicknesses? The three most popular statistics for measuring location are the sample mean, the median, and the mode. The mode is simply the observation that occurs most frequently.

The Sample Mean. If we let μ denote the average pelican egg thickness for the entire population, we can estimate μ using the *sample mean* (or *arithmetic mean*) which is simply the average of the observations in the data set. Letting y_1, y_2, \dots, y_n denote the data values for the observed egg thicknesses, the sample mean is defined and denoted as:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (1)$$

The \sum symbol simply means to add up the values. For this particular example we have:

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{65} [y_1 + y_2 + y_3 + \cdots + y_{65}] \\ &= \frac{1}{65} [0.14 + 0.19 + 0.20 + \cdots + 0.49] \\ &= 0.316\end{aligned}$$

Computing \bar{y} by hand is quite tedious for a sample size of $n = 65$. Instead we shall illustrate how to use SAS to compute this for us in a moment. One way to think of the sample mean is that it is the “center of gravity” of all the data points. If the points are lined up on a number-line “seesaw”, the seesaw would balance at exactly the sample mean.

The Sample Median. Another very commonly used measure of location is the *sample median*. The median is simply the middle value of the data after it has been arranged in ascending order. In the pelican egg data set above note that the observations are listed in ascending order according to eggshell thickness. If there are an odd number n of observations, then there is a unique middle value, the $(n + 1)/2$ th value. In the pelican example, $n = 65$ is odd, so the median is the $(65 + 1)/2 = 33$ rd value which is 0.30. If the sample size n is even, then the median is taken as the average of the $n/2$ and the $n/2 + 1$ values, i.e. the two middle values.

Note that the sample median is an estimator of the population median which is the median value for the entire population.

Question: Which measure of location, the mean or median, should we use in practice? The answer depends on the type of data set under investigation. If the distribution of data points are *symmetric* about the mean, then the mean and median will be equal. A symmetric distribution is one where the relative position of data points are the same on each side of the mean. If the distribution of observations for the entire population are symmetric about the population mean, then samples obtained from the population should be approximately symmetric about the sample mean. In practice, if the population is symmetric about the mean, the sample mean and median will be close in value but not necessarily exactly equal to each other. This will become clearer once we plot some data sets later.

One often encounters *skewed* distributions with particular types of data, such as toxin levels in soil or water, home prices, or salaries. For strongly skewed distributions, the median tends to be a better measure of location than the mean. As an extreme example, consider the following hypothetical example:

Example. Suppose a small neighborhood consists of 5 modest homes and one very large mansion. The home prices (in thousands of dollars) are:

112, 115, 116, 120, 122, 564.

The mean home price is \$191,500 which is clearly not representative of the neighborhood because its value is greater than all the home prices except the mansion. The median for this example is $(116 + 120)/2 = 118$ (since we have an even number of observations). Clearly \$118,000 is a better measure of central location than the mean in this example.

Measures of Spread

In order to adequately describe a data set, the mean (or median) is generally not sufficient by itself. To see why, consider the following simple example: Suppose the annual amount of rain in a sample of 10 United States cities was obtained in the years 1900 and 2000 yielding the following values (in inches):

1900	:	34 38 39 40 40 40 41 41 43 45
2000	:	21 21 33 33 36 38 39 50 60 70

The mean amount of rainfall in the 10 cities for both years is 40.1 inches. However, the rainfall amounts in 2000 are much more spread out than in 1900. This means that in the year 2000 one is much more likely to have seen droughts and floods than in 1900, even though the mean amount of rainfall is the same for the two years. This example illustrates that the mean (or average) is not sufficient by itself to describe a population (or sample). What is needed is a statistic for measuring the “spread” in the data. A simple measure of spread is the *range* which is the difference between the largest and smallest observation. However, the most commonly used measure of spread is the *variance* and its square root, the *standard deviation*. Given a set of observations y_1, y_2, \dots, y_n with sample mean \bar{y} , the idea is to look at the deviations about the mean:

$$(y_i - \bar{y}).$$

However, the deviations always average out to be zero. Instead, we look at the average of the squared deviations and define the sample variance, denoted s^2 as follows:

$$\text{Sample Variance : } s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}. \quad (2)$$

The (positive) square root of the variance is called the *sample standard deviation*:

$$\text{Sample Standard Deviation : } s = \sqrt{s^2} = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)}. \quad (3)$$

To illustrate the computation, we compute the sample variance and standard deviation for the two sets of rainfalls. Letting s_1^2 and s_2^2 denote the sample variances for 1900 and 2000 we compute

$$\begin{aligned} s_1^2 &= \{(34 - 40.1)^2 + (38 - 40.1)^2 + (39 - 40.1)^2 + (40 - 40.1)^2 + (40 - 40.1)^2 + \\ &\quad (40 - 40.1)^2 + (41 - 40.1)^2 + (41 - 40.1)^2 + (43 - 40.1)^2 + (45 - 40.1)^2\} / (10 - 1) \\ &= 8.54 \end{aligned}$$

and

$$\begin{aligned} s_2^2 &= \{(21 - 40.1)^2 + (21 - 40.1)^2 + (33 - 40.1)^2 + (33 - 40.1)^2 + (36 - 40.1)^2 + \\ &\quad (38 - 40.1)^2 + (39 - 40.1)^2 + (50 - 40.1)^2 + (60 - 40.1)^2 + (70 - 40.1)^2\} / (10 - 1) \\ &= 248.99. \end{aligned}$$

The standard deviations for tests 1 and 2 are

$$s_1 = \sqrt{8.54} = 2.923$$

and

$$s_2 = \sqrt{248.99} = 15.779.$$

Note that the standard deviation of the rainfall amounts in the year 2000 is much larger than for the year 1900 indicating much more variability in rainfall amounts in the year 2000.

Just as the sample mean \bar{y} was an estimator for the population mean μ , the sample variance s^2 is an estimator for the population variance which is denoted by the Greek letter σ^2 (sigma-squared). Recall that the sample variance is the “average” of the squared deviations of the observations about the sample mean.

Question: Why do we divide by $n - 1$ then instead of n in (2)? The answer is that if we divide by n instead of $n - 1$, the estimator s^2 will be too small on average. In other words, dividing by n will tend to make s^2 systematically smaller than the true population variance σ^2 on average. This is known as **bias** in statistics. Often we prefer to use estimators that are unbiased. A biased estimator will tend to be systematically too large or too small. Dividing by $n - 1$ in (2) makes s^2 unbiased.

Note that the standard deviation (or variance) will be zero if and only if all the observations have exactly the same value (no variability).

Using SAS to Obtain Descriptive Statistics.

Ordinarily, statistical software is used to do the tedious computations needed to obtain the values of descriptive statistics. We shall illustrate how to do this using SAS. The appendix to these notes has more detail about specifics of using SAS. Generally in SAS, to obtain statistical output of interest, one uses one of several *Procedures* which are abbreviated “PROC” in SAS. In order to obtain simple statistics like the mean and standard deviation, one could use “PROC MEANS”. In the SAS program below, we use “PROC UNIVARIATE” which generates a large number of descriptive statistics, most of which we will ignore for now.

```

/*****
The PCB concentration (ppm) and thickness (mm) of
65 Anacapa pelican eggs (Risebrough 1972) were measured.
This SAS program will illustrate PROC UNIVARIATE for obtaining
some basic statistics.
*****/
data pelican;
infile 'c:\stt630\notes\pelican.dat';
input pcb thick;
run;
proc print data=pelican;
run;
proc univariate plot;
var thick;
run;

```

Here are some notes about the SAS program above:

- It is a good idea to put lots of comments in your SAS programs so that you will know what the

program is suppose to do. In SAS, comments are added by using “/*” and the “*/”. Anything that appears between these symbols is ignored by SAS.

- Each command line in SAS must end with a semi-colon “;”. Please remember this – I once spend a couple hours trying to debug a SAS program when I first started using SAS and my error was that I forgot a semi-colon.
- When you read a data set into SAS, we need to give it a name. The “data pelican;” statement creates an internal SAS data set and we have named it pelican.
- In order to analyze a data set, the program must read in the data. There are two ways to do this in SAS. In this program, we read in the data from an external file which in this case is called ‘pelican.dat’. The data is read in using the “infile” command. Following the infile command, you must give the location of the data set on your computer hard drive in single quotes. Note that data files often have the “.dat” extension so that one will know that the file is a data file. The other way of reading in data is to type the data lines directly into the program. We shall see examples of this later.
- The nextline of the program is the “input” statement. This statement gives names to each of the variables in the data set. The way data sets are organized is that *each row corresponds to an observation* and *each column corresponds to a variable*. In this data set there are two columns for the two variables which I have names “pcb” for PCB concentration and “thick” for eggshell thickness.
- After each command in a SAS program, it is a good idea to have a “run” statement which directs SAS to run the command. Thus, the run statement following the infile and input statements directs SAS to go ahead and read in the data.
- The first PROC statement in this program is the PROC PRINT statement which simply prints out the data in the output window. Note that we are specifying SAS to print out the pelican data set by writing “data=pelican”. This statement is not needed here because there is only one SAS data set in this program. In more complicated programs, one can work with several SAS data sets in a single program and it is useful to specify which one to apply a procedure to. If you do not specify a data set using the “data=”, then SAS automatically uses the most recently created data set.
- The main purpose of the above program is to obtain some basic descriptive statistics. This is done using PROC UNIVARIATE which produces a lot of output, most of which we are not interested in at the moment. However, below we see some of the pertinent output from PROC UNIVARIATE from the output window. The “plot” command in the proc univariate line instructs SAS to produce some plots. We discuss plots below.
- The “var thick” tells SAS to compute descriptive statistics for the variable egg thickness only. If we leave off the “var” statement, the SAS will automatically produce statistics for all the variables (i.e. each column in the data set).

Moments

N	65	Sum Weights	65
Mean	0.31630769	Sum Observations	20.56
Std Deviation	0.08047844	Variance	0.00647678

Skewness	0.31932748	Kurtosis	-0.501788
Uncorrected SS	6.9178	Corrected SS	0.41451385
Coeff Variation	25.4430857	Std Error Mean	0.00998212

Basic Statistical Measures

Location		Variability	
Mean	0.316308	Std Deviation	0.08048
Median	0.300000	Variance	0.00648
Mode	0.300000	Range	0.35000
		Interquartile Range	0.11000

From this SAS output, we see that the sample size is 65, the sample mean is 0.3163, the median is 0.30, and the sample standard deviation is 0.080478. Several other statistics are produced such as the skewness. Note that the standard error of the mean is the sample standard deviation divided by the square root of the sample size. The *coefficient of variation* is $100 \times (\text{standard deviation}/\text{mean})$

Linear Transformations When collecting data by taking measurements, the scale that the measurements are made is somewhat arbitrary: inches or centimeters, Fahrenheit or Celsius, etc. It is quite common to change the measurement scale when analyzing the data. Many transformations are linear. That is, suppose our original measurement is y_i and we change the scale by considering instead $x_i = a + by_i$ where a and $b \neq 0$ are constants. What happens to the mean and variance if we perform a linear transformation? Consider first adding a constant to each measurement $x_i = a + y_i$. For instance, suppose y_i represents the salary for an employee of a large company and the owner gives each employee a \$1000 raise. Then the new salaries are $x_i = a + y_i$ where $a = 1000$. The average (or mean) salary will increase by 1000. However, if we add a constant to each observation, does that change the spread? The answer is no. The variance of the x_i 's will be the same as the y_i 's. This is easy to demonstrate algebraically.

If you make a transformation by multiplying each observation by a non-zero constant b , $x_i = by_i$, then the sample mean will change by a factor of b and the sample variance changes by a factor of b^2 :

$$\bar{x} = b\bar{y} \quad \text{and} \quad s_x^2 = b^2 s_y^2$$

where s_x^2 is the sample variance for the x -measurements. Note that the sample standard deviation for the x_i 's is $|b|s_y$ (b could be negative but the standard deviation cannot be negative).

In general, if $x_i = a + by_i$, then the mean of the x_i 's is $\bar{x} = a + b\bar{y}$ and the standard deviation for the x_i 's is $s_x = |b|s_y$.

Graphical Methods

Descriptive statistics are very useful for understanding a data set. Complementing the use of descriptive statistics are visual aids – ways of graphing the data so one can get a picture of the distribution of the data points. The “shape” of the distribution can tell us a lot about the population of interest as we shall see with some examples. We shall introduce a few very common graphical methods.

Stem-and-leaf plots. In PROC UNIVARIATE in the SAS program above, because we specified “plot” in the program, SAS produced a *stem-and-leaf* plot and a *boxplot*. The advantage to the stem-and-leaf plot is that it is easy to draw by hand. In the stem-and-leaf plot, each data point is broken into two pieces: a stem and a leaf. For example, if one considers a data set consisting of test scores in a range of 0 to 100, the stem could be the 10’s digit and the leaf could be the units digit. There may be several observations with a common stem (say test scores in the 70’s). To draw the plot, begin by drawing a vertical line. On the left side, write out the stem values in order. On the right side, write down all the leaf values for each stem value (e.g. for the stem value of 7 corresponding to test scores in the 70’s, to the left of the vertical line, write down all the leaf values for each test score in the 70’s. The stem-and-leaf plot produced by SAS in proc univariate simply uses “0”’s for the leaf values, but one should use the actual leaf values.

Box plots. The box plot is drawn by simply drawing a box. The bottom and top of the box correspond to the lower quartile (i.e. the 25th percentile) and upper quartile (i.e. the 75th percentile) of the data sets respectively. A line is drawn through the middle of the box corresponding to the median of the data set (note that the median is the 50th percentile). Also, “+” sign is drawn in the box to indicate the sample mean. If the distribution is roughly symmetric the mean and median will be approximately equal in which case the line through the middle of the box and the + sign will coincide. If the distribution is skewed to the left or right, the mean will be pulled in the direction of skewness away from the median. Next, vertical lines are drawn from the top and bottom of the box – these lines are called “whiskers”. The whisker lines extend to the most extreme value in the data set within the range of

$$\text{upper quartile} + 1.5 \times [\text{IQR}],$$

and

$$\text{lower quartile} - 1.5 \times [\text{IQR}],$$

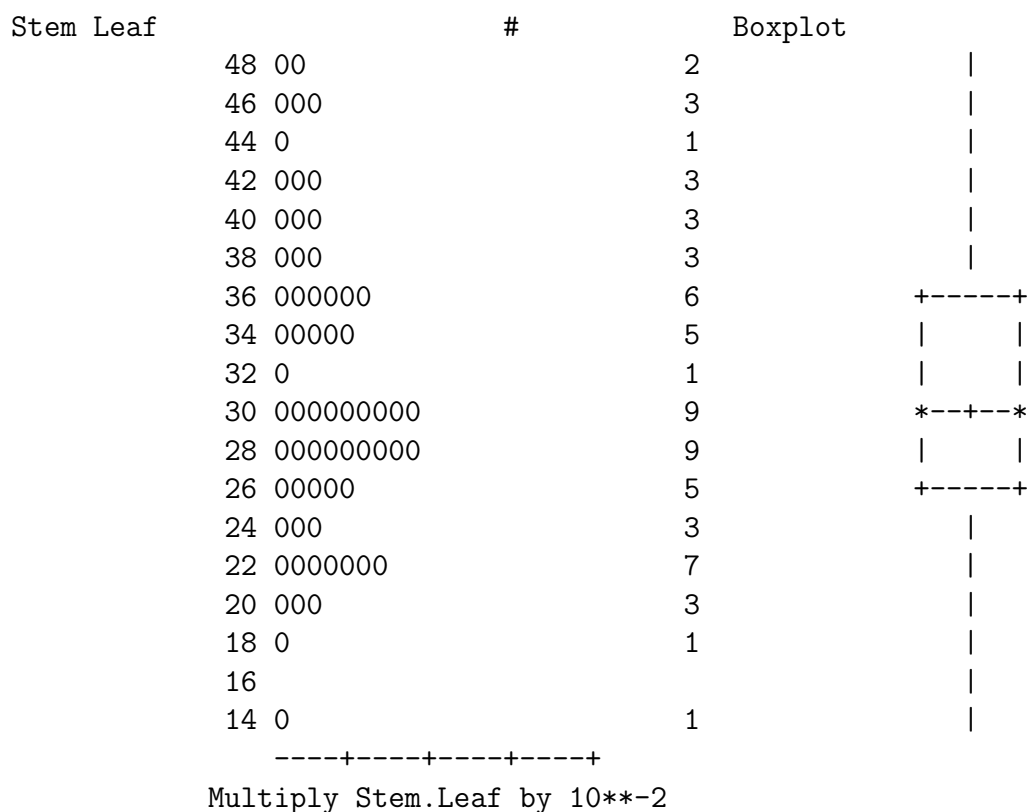
where

$$\text{IQR} = \text{Inter-Quartile Range} = \text{upper quartile} - \text{lower quartile}.$$

From the SAS output above, the IQR for the pelican eggshell thickness is 0.11. Observations that lie beyond the whisker lines are considered outlying values and these are plotted in the box plot beyond the whiskers. Any observation lying more than 3 times the IQR above or below the upper or lower quartiles are considered extreme outliers. One should examine these data points to see if they correspond to a typographical error. Sometimes extreme outliers occur because the observation may come from a different population than the one under consideration.

Box plots are particularly useful when comparing two or more distributions by drawing the box plots for the different distributions side-by-side. We will see examples of this later as well.

Below is the SAS stem-and-leaf and box plot produced from proc univariate in SAS:



In the box plot, the mean and median coincide. The stem-and-leaf plot shows that the shape of the distribution appears to be slightly skewed to the right (towards larger values) and *unimodal* – a single mode. The box plot does not show any observations as being outliers.

Histograms. One of the most useful plots for capturing the shape of a distribution is the histogram. The idea is to divide the observations into groups. In the test score example, one can create groups of scores as 90–100, 80–89, 70–79, etc. Draw a horizontal number line and mark the group boundaries on this number line. Next, draw rectangles for each group whose base is on the number line and whose height corresponds to either the frequency (i.e. number of observations) in the group or the relative frequency (i.e. the percentage of observations) in the group. The histogram on the next page, produced by SAS, is a relative frequency histogram. Again, we see a unimodal distribution of eggshell thicknesses for the pelicans that has a slight hint of being skewed to the right.

It is possible to write SAS code to obtain this high-resolution histogram. However, SAS for windows has a built-in menu driven way of producing these plots (as well as obtaining basic statistical output) without having to write any programs. To obtain this histogram in SAS, on the top toolbar, click with your mouse on “Solutions” followed by “Analysis” followed by “Analyst”:

Solutions → Analysis → Analyst.

A new window will open up. Under the file menu, choose “Open by SAS name”. Double-click on the “Work” file which contains all the SAS data sets in operation. Next, double-click on “pelican”. You will see the pelican data set open up in the Analyst window. At the top along the toolbar you will see several options. Choose “Graphs” and click on “histogram”. Click on “thick” followed by “Analysis” followed by “OK”. This will produce the histogram on the next page.

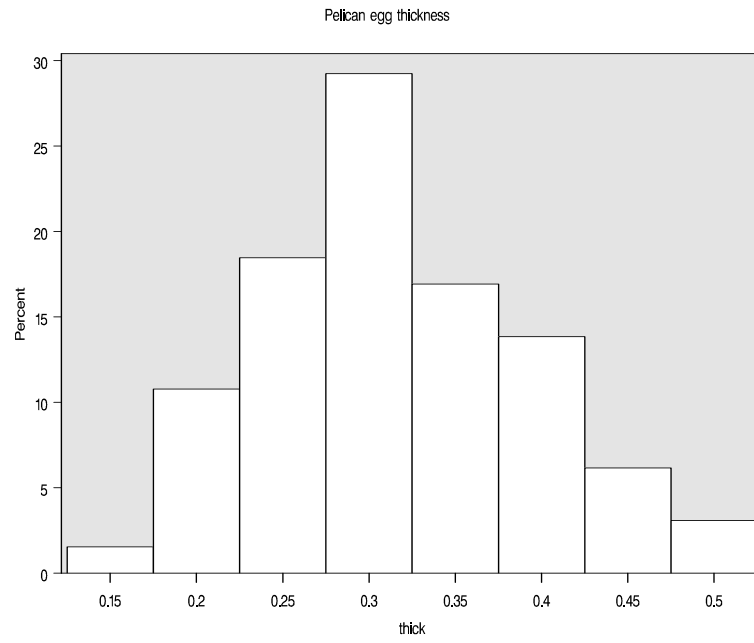


Figure 1: Histogram of pelican eggshell thicknesses, produced by SAS.

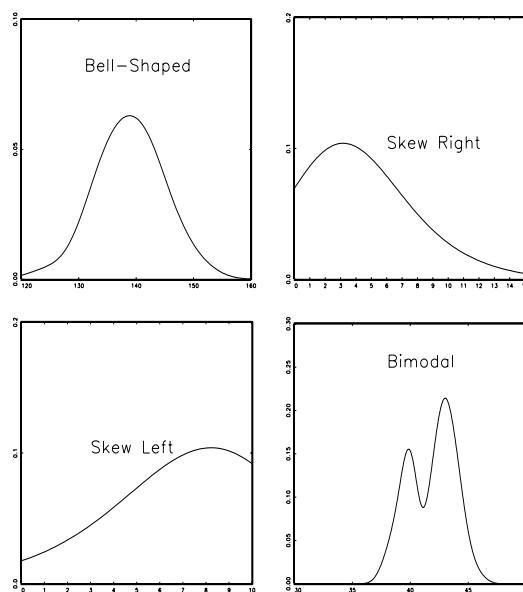


Figure 2: Four common shapes of data distributions: top-left: symmetric and bell-shaped (or mound-shaped), top-right: skewed to the right, bottom-left: skewed to the left, bottom-right: bimodal

Distribution Shapes. Part of the problem with constructing a histogram is determining the number of groups for dividing the data into. If the number of groups is too large, then there will not be enough observations in each group. If the number of groups is too small, then it will be difficult to get a picture of the shape of the distribution. If we could obtain larger and larger sample sizes from the population of interest, then we could create histograms with a larger number of groups over smaller and smaller regions. Imagine then that in the limit as the sample size becomes arbitrarily large, the histogram will become finer and finer and will look like a smooth curve. These smooth plots are known as density curves and they tell us the shape of the distribution.

Figure 2 shows four different shapes commonly encountered with real data sets. The bell-shaped distribution (top-left panel of Figure 2) was constructed based on facial measurements of a sample of adult men. This symmetric bell-shape distribution is actually quite common in many applications. The bell-shaped curves are often seen with data from homogeneous populations where the frequency of measurements on either side of the mean drop off symmetrically the further you move away from the mean.

A study on mumps was conducted on children and the amount of mumps antibodies in the blood was the variable measured. The data from this study produced a histogram with the shape shown in the plot of the bottom-right of Figure 2. The shape of this distribution is *bimodal*. Why is this? The answer is that some children had already had mumps and therefore had the mumps antibodies in their blood while other children had not yet had mumps. Bimodal distributions appear quite frequently in practice (e.g. data on males and females). Of course, one can encounter multi-modal distributions as well in practice.

The distributions plotted in the top-right and bottom-left panels of Figure 2 show skewed distributions. Data on home prices and salaries tend to be skewed to the right due to some very expensive homes and high incomes. If you look at test scores from an easy exam, then you would expect to see a distribution skewed to the left.

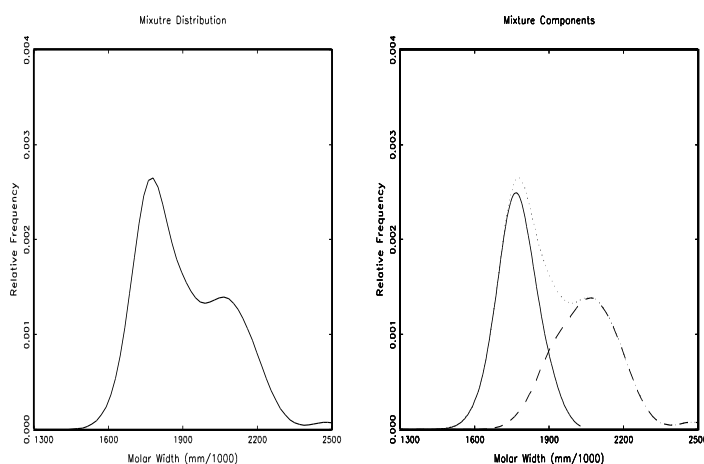


Figure 3: Mixture distribution of molar widths for two species of voles.

Mixtures. Sometimes a population may contain homogeneous subgroups. If we collect data from such a population, then the resulting histogram may reveal multiple modes as in the bottom-right plot in Figure 2. A study was conducted in order to distinguish closely related species of voles (Airoidi et al 1996). For example, the left panel of Figure 3 shows a smoothed histogram of a sample of voles where the variable measured was molar widths. This smoothed histogram shows a clear “bimodal” pattern. The reason why is that the sample contains two closely related species of vole: *microtus multiplex* and *microtus subterraneus*. The right panel of Figure 3 shows the individual smoothed histograms for each species. A distribution that consists of several homogeneous sub-populations is known as a *mixture* distribution. One of the statistical problems in this framework is to estimate the parameters (means, standard deviations, etc.) for the separate components of the mixture. The problem is often further complicated in cases where there is no information in the data set as to which sub-population a particular observation belongs.

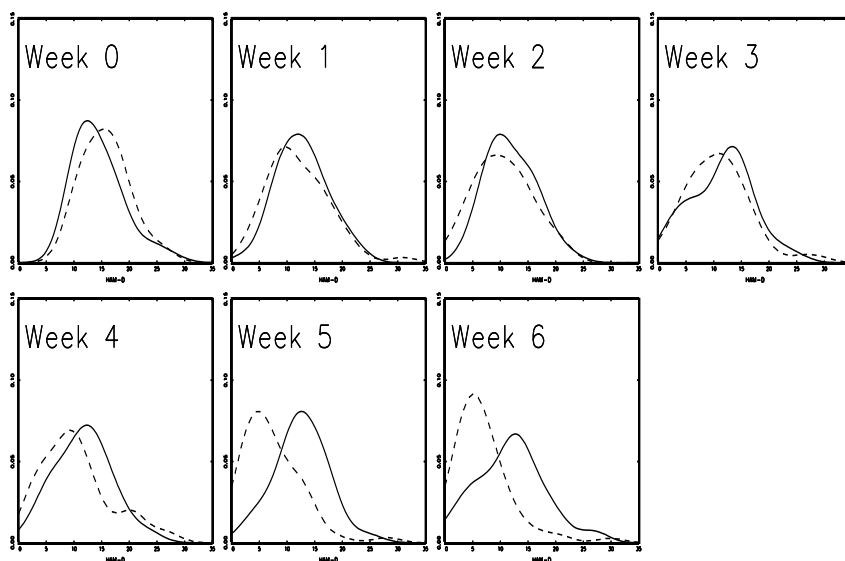


Figure 4: Smoothed histograms of weekly HAM-D scores for phenelzine subjects (dashed histogram) and placebo subjects (solid histogram).

Example. A clinical trial was conducted to study the effectiveness of phenelzine in treating depression. In order to evaluate the effectiveness of the drug, a control is needed in the study. If a patient gets better, can we claim that it is due to the drug, or due to a *placebo* effect, or due to spontaneous improvement? In this particular study, subjects were randomized to either receive the study drug phenelzine or a placebo. The subjects were then evaluated for six weeks. At each weekly examination, the patient's depression was recorded on the Hamilton depression (HAM-D) scale. Lower scores indicate lower levels of depression. Figure 4 shows smoothed histograms of the HAM-D distributions for patients on phenelzine (dashed curve) and placebo (solid curve) respectively. At baseline (week 0), the HAM-D distributions for patients on phenelzine and placebo both appear to be approximately bell-shaped. However, as the weeks go by, the distribution for phenelzine treated subjects moves to the left towards improvement. That is, the mean HAM-D is shifting to lower values each week. However, notice how the distribution becomes more and more skewed each week for phenelzine treated subjects. The reason this happens is that the distribution has a sharp boundary of zero on the left, the lowest score possible. Patients tend to gravitate towards this score during the course of treatment. When there is a sharp boundary for a variable of interest and data values fall near this boundary, the result is often a skewed distribution away from this boundary.

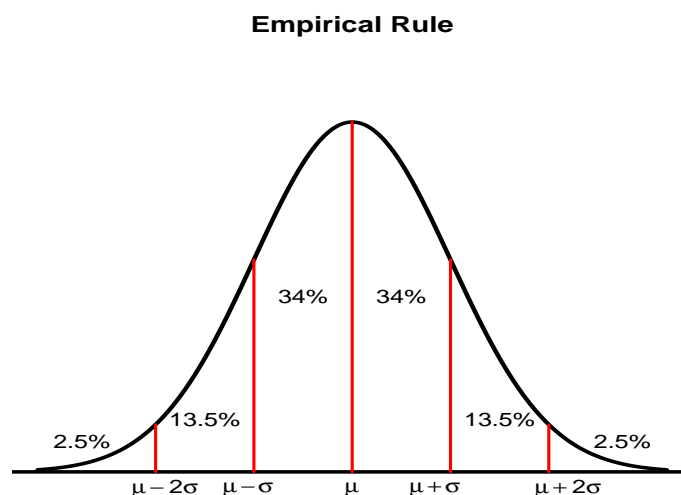


Figure 5: An illustration of the empirical rule

The Empirical Rule

We end this chapter with a very useful rule for interpreting standard deviations. Recall from Figure 2 that many populations exhibit a “bell-shape” when plotted (in histograms say). For these types of data sets, the information can typically be summarized by the mean and standard deviation without much loss of information. The following rule, known as the *empirical rule*, allows us to interpret the mean μ and standard deviation σ of a distribution:

- Approximately 68% of the observations will lie within 1 standard deviation of the mean: $\mu \pm \sigma$.
- Approximately 95% of the observations will lie within 2 standard deviations of the mean: $\mu \pm 2\sigma$.
- Approximately 99.7% of the observations lie within three σ of the mean. That is, practically all observations will lie within 3 standard deviations of the mean.

If 68% of the distribution lies within one standard deviation of the mean, then we can divide this 68% evenly in two ($68\%/2 = 34\%$) with 34% lying between $\mu - \sigma$ and μ . The other 34% lies between μ and $\mu + \sigma$. Further breakdown of the distribution is illustrated in Figure 5.

The empirical rule also holds for sample data possessing a bell-shaped by simply replacing μ by \bar{y} and σ by s in the above rule.

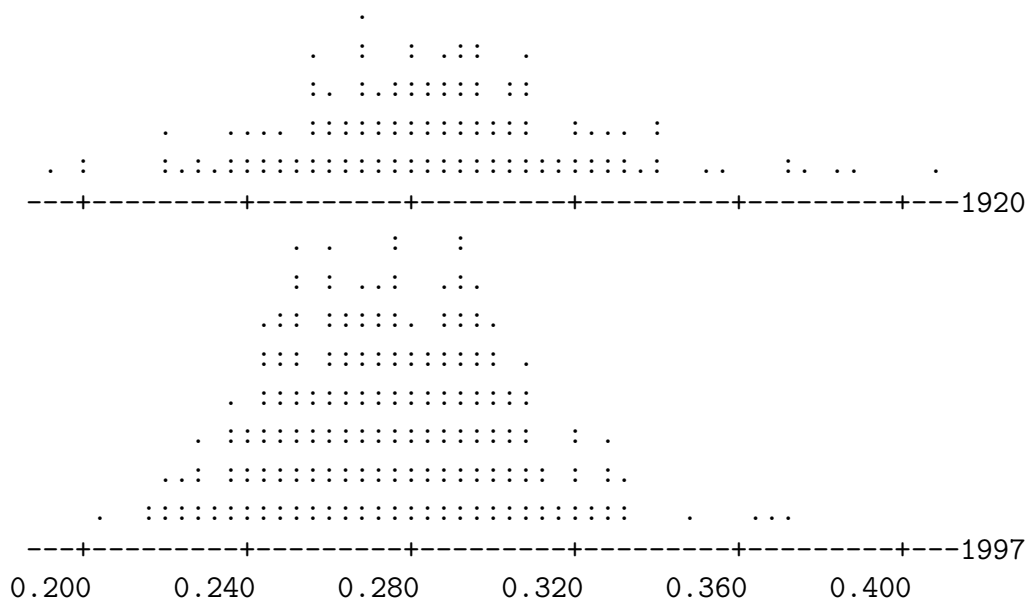
According to the empirical rule, practically the entire distribution (99.7%) will lie in a range of 6σ (from $\mu - 3\sigma$ to $\mu + 3\sigma$).

A useful way to interpret an observation is in terms of its *z-score* which is the relative standing of the observation in terms of the mean and standard deviation. If x is a data point, the *z-score* is computed by

standardizing the value by

$$\text{Population } z\text{-score} : z = \frac{x - \mu}{\sigma} \qquad \text{Sample } z\text{-score} : z = \frac{x - \bar{x}}{s}. \qquad (4)$$

Batting Averages Example. No major league baseball player has had a batting average of .400 or better since Ted Williams in 1941. A player’s batting average is basically the number of hits the player has divided by the number of times the player goes to bat. Many baseball fans have debated the reason for the extinction of .400 hitting in baseball for years. If we look at the data, the empirical rule demonstrates why .400 hitting is non-existent in major league baseball. Below are dotplots for league batting averages in the years 1920 and 1997 – the dots represent batting averages for all players with at least 250 plate appearances for the season. The dotplots show that the spread of batting averages in 1920 is greater than in 1997. The dotplots also show approximately bell-shaped batting average distributions.



The mean and standard deviation for league batting averages in 1920 are $\mu_1 = 0.28553$ and $\sigma_1 = 0.03769$ and for 1997, the mean and standard deviation are $\mu_2 = 0.27490$ and $\sigma_2 = 0.02874$. In 1920, a .400 batting average was about three standard deviations above the mean: $z_1 = (.400 - .28553)/0.03769 = 3.037$. Thus, a .400 batting average in 1920 is pretty extreme but it would not be too unusual to observe one player batting .400 or better and that is what happened in 1920. However, in 1997, a .400 batting average is about 4.35 standard deviations above the mean: $z_2 = (.400 - .27490)/0.02874$. According to the empirical rule, practically all players will have batting averages within three σ of the mean. In 1997, a .400 batting average was 4.35 standard deviations above the mean indicating that obtaining a .400 batting average that year would have been extraordinarily rare. One of the differences in the statistics between 1920 and 1997 is that the standard deviation of batting averages is smaller in 1997. In fact, the standard deviations of batting averages became smaller and smaller from the early 1900’s to the end of the 20th century while the overall league batting averages tended to fluctuate up and down. To understand why .400 hitting became extinct, one needs to figure out why the variability in batting averages (and thus the standard deviations)

declined as the years went by. Stephen Jay Gould provided a discussion of this in his book *Full House* as well as his explanation for the decreasing variability in batting averages.

Diversity Indices. An important class of statistics for measuring the ecological health of an area is given by diversity indices. Diversity indices are used for nominal scale data where observations are recorded depending upon which category they fall in (e.g. data on birds may record how many birds are observed as belonging to particular species of bird). Suppose you are examining the ecological health of a stream that is populated by 5 different species of fish. Let p_1, p_2, \dots, p_5 denote the proportion of fish that fall into the 5 categories of fish ($\sum p_i = 1$.) If $p_1 = 1$ and all the other p_i 's equal zero, then there is very little diversity of fish. Typically the ecological health of an area corresponds to high levels of diversity. If $p_1 = p_2 = \dots = p_5$, then there is an *even* spread of fish in the stream or high diversity. A couple of popular diversity indices are

$$\begin{aligned} \text{Shannon's Index:} &= - \sum_{i=1}^k p_i \log p_i \\ \text{Simpson's Index:} &= 1 - \sum_{i=1}^k p_i^2 \end{aligned}$$

where k = the number of categories.

1 Problems

SAS code and data can be found in the Appendix of this chapter.

1. $n = 7$ sediment samples are obtained from a lake yielding the following PCB concentrations (in ng/g dry wt.):

0.4, 0.7, 1.3, 2.8, 3.9, 6.0, 6.1.

- a) What is the sample mean PCB concentration?
 - b) What is the median PCB concentration?
 - c) What is the sample standard deviation from these measurements?
 - d) Suppose each of the $n = 7$ observations are re-scaled by multiply each by 100. What is the new sample mean and sample standard deviation for the re-scaled data?
2. Data from the Skyline 50k Ultra-Marathon (August 4, 2002) is contained in the data file "ultra.dat". For this assignment, run the SAS program "ultra.sas" to obtain basic univariate statistics for the ultra-marathon times. The goal is to use the output from this program to compare the performances of for men and women in the ultra-marathon race. For this problem, write a short report (no more than one page not including graphics) with the following parts:
 - a) Write an introductory paragraph.
 - b) Compare men and women in terms of their average times and their standard deviations.
 - c) Include a plot in your report that compares the times for men and women and comment on the plot.

- d) Write a concluding paragraph summarizing your results.
3. The systolic blood pressure was recorded (mm Hg) for 6 adult males between the ages of 30-40 while standing yielding the following data:

105, 124, 102, 114, 96, 106.

- a) (2 point) What is the population of interest?
- b) What is the variable?
- c) What is the sample mean from this data?
- d) What is the sample median?
- e) Compute the sample standard deviation (without using a computer)?
4. Run the *nortemp.sas* program and report the average pulse rates for men and women and their standard deviations. From these 130 observations, can we infer that in the population of healthy adults that women have a higher pulse rate on average than men? Just give your thoughts on this question (a sentence or two) without any formal statistical analysis (which we haven't covered yet).
5. Data was collected on soil samples in the Florida Everglades and the the amount of phosphorus in the soil samples in parts per million were measured. As is often the case with this type of data, the distribution of phosphorus is skewed to the right. The table below shows SAS output from PROC MEANS for the phosphorus data and the natural logarithm of the phosphorus data. A logarithm transformation is often used for skewed right data so that the resulting distribution becomes more "bell-shaped".

	n	mean	standard deviation	min	max
phosphorus	287	302.06	161.220	76.82	740.94
log(phosphorus)	287	5.56	0.568	4.342	6.608

Suppose we change the phosphorus data unit of measurement from parts per million to parts per billion by multiplying each phosphorus value by 1000. Let x_1, x_2, \dots, x_n denote the phosphorus data in parts per million. Then $y_i = 1000x_i$ is the resulting value in parts per billion. Answer the following questions:

- a) What will be the mean phosphorus level in parts per billion (ppb)? Show this algebraically.
- b) What is the standard deviation of phosphorus in ppb? Show this algebraically.
- c) What is the mean of the natural log-transformed phosphorus values in ppb? Show this algebraically.
- d) What is the standard deviation of the natural-log transformed values of phosphorus in ppb? Show this algebraically.
6. The SAS program *butterclams.sas* contains data on the ratio (length/width) for a sample of native butter clam (*Saxidomus giganteus*) shells from the Puget Sound. The lengths and widths were measured in centimeters. Run the program and do the following parts:
- a) What is the sample mean and median for the data?
- b) What is the sample standard deviation?

- c) Describe the “shape” of the distribution from the histogram.
- d) The boxplot identifies an observation as an outlier. What is the value of this observation? How many standard deviations is this observation from the mean?
7. (5 points) The distribution of weights for newborn full term babies is bell-shaped (not necessarily normal) with an average birth weight of 7.2 pounds and standard deviation 0.5 pounds. Approximately what percentage of full term babies have weights below 6.7 pounds? (Justify your answer.)
8. Suppose babies learn to walk on average at age 12 months with standard deviation 1.8 months and that the distribution of ages at which babies learn to walk is bell-shaped. A mother is concerned that her 13 month-old baby is not walking yet. Using the empirical rule, explain to the mother (in one or two sentences) why she should not be concerned.
9. The editor of Mad Magazine stated recently in the New York Times that the average age of Mad Magazine readers is 26 and the median age is 19. Which of the following do you think best describes the shape of the age distribution for readers of Mad Magazine?
- (a) symmetric (b) skewed right (c) uniform (d) discrete
10. Bicep girth (in cm) was measured in a sample of 6 adults yielding the following measurements:
- 32.5, 34.4, 33.4, 31.0, 32.0, 33.0
- a) Find the sample mean for these 6 measurements.
- b) Compute the sample variance for these 6 measurements.
- c) What is the median for these 6 measurements?
11. The volume of solution in 50 test tubes is measured and the average volume was found to be 5.2 ml with standard deviation 0.3 ml. Exactly 1.2 ml of additional solution is added to each test tube. Which of the following is the standard deviation for the volumes of solutions after adding the additional solution? (Circle one)
- (a) 0.3 ml (b) 1.5 ml (c) 0.3286335 ml (d) 0 ml (e) 7.4 ml (f) 0.36 ml

References

- Airoldi, J. P., Flury, B. and Salvioni, M., (1996), “Discrimination between two species of *Microtus* using both classified and unclassified observations. *Journal of Theoretical Biology*, **177**, p 247–262.
- Gould, S. J. (1996), *Full House : The Spread of Excellence from Plato to Darwin*, Harmony Books.

2 Appendix: SAS Programs and Data

Ultra-Marathon Data

```

/*****
Data from the Skyline 50k Ultra-Marathon (August 4, 2002)
column 1: hours,

```

```
column 2: minutes
column 3: seconds,
column 4: Code for sex 0=male, 1=female,
column 5 = age of runner
*****/
options ls=80;
data ultra;
input hours minutes seconds sex age;
time = hours*60 + minutes + seconds/60;
datalines;
3 59 15 0 34
4 05 08 0 31
4 09 56 0 36
4 13 55 0 44
4 23 33 0 22
4 28 14 0 46
4 28 59 0 51
4 29 50 0 36
4 33 14 0 41
4 33 21 1 27
4 36 04 0 42
4 41 46 1 29
4 44 20 0 40
4 46 08 0 44
4 49 40 0 47
4 51 07 0 54
4 51 15 1 29
4 51 37 0 42
4 52 16 0 48
4 52 47 0 58
4 52 47 0 38
4 53 31 0 42
4 54 28 0 45
4 59 35 1 39
4 59 59 0 33
5 00 59 0 48
5 03 31 0 48
5 04 00 0 35
5 05 19 0 28
5 06 04 0 55
5 06 53 1 38
5 07 09 0 50
5 07 36 0 44
5 08 04 0 49
5 08 21 0 42
5 09 05 0 39
5 09 13 0 58
```

5	11	00	1	49
5	11	27	0	28
5	11	38	0	45
5	14	07	0	38
5	14	07	0	34
5	14	41	0	40
5	16	17	0	47
5	17	59	0	56
5	18	28	0	30
5	18	57	0	39
5	19	01	1	35
5	20	50	0	54
5	21	08	0	42
5	21	37	0	56
5	21	42	0	42
5	23	00	0	48
5	23	32	0	48
5	23	55	1	36
5	24	06	1	39
5	25	19	0	43
5	25	29	0	53
5	25	50	0	34
5	26	05	0	55
5	27	06	1	34
5	29	45	0	28
5	32	03	1	47
5	33	23	0	42
5	34	06	0	36
5	34	48	0	48
5	37	53	0	39
5	39	29	1	38
5	40	40	1	37
5	41	13	0	29
5	42	57	0	50
5	46	16	0	27
5	46	39	1	48
5	46	39	0	39
5	47	32	0	60
5	48	29	0	50
5	49	42	0	51
5	50	15	0	40
5	50	26	0	47
5	50	43	1	31
5	51	04	0	51
5	51	15	0	48
5	52	28	1	38
5	54	40	0	31

5	55	57	0	32
5	58	41	0	38
5	58	49	0	45
5	59	59	0	35
6	00	43	1	56
6	01	26	1	40
6	03	11	0	47
6	03	56	0	48
6	04	07	0	48
6	04	15	0	34
6	04	29	1	26
6	05	05	0	41
6	05	18	1	39
6	05	19	1	43
6	05	20	0	53
6	06	31	0	31
6	06	37	0	35
6	06	44	1	45
6	09	56	0	49
6	10	35	0	49
6	12	13	0	41
6	12	20	1	61
6	12	56	1	52
6	12	59	0	62
6	13	14	0	64
6	13	18	1	44
6	14	40	1	43
6	15	25	0	47
6	15	32	0	43
6	17	26	0	55
6	17	27	0	33
6	17	32	0	48
6	18	39	0	45
6	18	40	0	53
6	19	41	0	50
6	19	50	1	32
6	19	50	1	37
6	19	50	0	37
6	22	27	1	29
6	22	29	0	29
6	22	34	1	56
6	23	56	0	56
6	24	34	1	41
6	24	35	0	65
6	25	46	1	44
6	25	46	1	52
6	26	47	1	47

6	27	17	0	55
6	27	36	1	44
6	28	06	1	61
6	28	39	0	51
6	28	42	0	60
6	29	05	1	52
6	31	10	1	24
6	31	10	0	30
6	31	13	1	45
6	31	13	1	50
6	35	00	0	64
6	35	26	0	46
6	35	47	0	39
6	36	09	0	63
6	36	09	0	53
6	36	09	0	48
6	38	46	0	49
6	38	53	0	40
6	38	53	1	35
6	39	01	1	55
6	39	01	1	51
6	39	27	0	54
6	40	24	0	52
6	40	25	0	57
6	41	25	1	50
6	42	08	0	52
6	42	36	0	56
6	45	10	0	43
6	46	15	1	37
6	46	16	0	39
6	47	42	1	55
6	48	06	0	46
6	48	22	0	34
6	50	12	0	51
6	50	17	1	41
6	50	58	0	34
6	51	00	0	38
6	52	25	0	33
6	53	04	0	65
6	53	17	0	59
6	57	17	0	56
6	57	37	0	45
6	58	05	0	58
6	58	50	0	51
6	59	22	1	61
6	59	55	0	54
7	00	54	0	48

```

7  01  40 1 60
7  02  34 0 40
7  06  35 0 55
7  08  23 0 56
7  11  46 1 44
7  16  48 1 38
7  20  08 1 55
7  21  59 1 57
7  21  59 0 35
7  23  21 0 57
7  29  08 1 50
7  35  26 1 79
7  35  31 0 64
7  41  29 0 61
7  44  03 0 59
7  55  19 1 45
7  55  44 1 44
;
run;
proc corr; var time age; run;
proc print;
run;
proc sort;
    by sex;
run;
proc means;
    by sex;
run;
proc univariate plot normal; * obtain summary statistics on running times;
    var time;
    by sex;
run;
proc univariate plot normal;
    var time;
run;
quit;

```

Butterclams.sas

```

/*****
The data below gives the ratio = length/width for a sample of
native butter clam (Saxidomus giganteus) shells from the Puget
Sound. The lengths and widths were measured in centimeters.
Data is from the Quantitative Environmental Learning Project
webpage: http://www.seattlecentral.edu/qelp/Data\_MathTopics.html#Linear
*****/
data butterclams;

```

input ratio;

datalines;

1.31

1.36

1.29

1.27

1.21

1.67

1.33

1.29

1.36

1.23

1.30

1.33

1.21

1.40

1.32

1.35

1.31

1.31

1.19

1.36

1.25

1.31

1.34

1.31

1.28

1.27

1.33

1.23

1.32

1.29

1.26

1.29

1.28

1.33

1.36

1.25

1.32

1.38

1.33

1.29

1.32

1.27

1.29

1.31

1.33

1.24
1.30
1.30
1.38
1.27
1.25
1.24
1.32
1.24
1.29
1.23
1.31
1.17
1.25
1.41
1.28
1.28
1.19
1.34
1.28
1.30
1.22
1.37
1.21
1.21
1.43
1.23
1.21
1.22
1.37
1.19
1.18
1.25
1.17
1.36
1.28
1.28
1.32
1.17
1.27
1.22
1.20
1.15
run;
proc univariate plot normal;
run;

Body Temperature and Pulse Data

```

/*****
SOURCE:
Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical
Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body
Temperature, and Other Legacies of Carl Reinhold August Wunderlich,"
_Journal of the American Medical Association_, 268, 1578-1580.

Column 1: Temperature (F)
Column 2: Gender (1 = male, 2 = female)
Column 3: Pulse rate
*****/
options linesize=76 nodate;
data bodytemp;
input temp gender pulse;
datalines;
96.3    1    70
96.7    1    71
96.9    1    74
97.0    1    80
97.1    1    73
97.1    1    75
97.1    1    82
97.2    1    64
97.3    1    69
97.4    1    70
97.4    1    68
97.4    1    72
97.4    1    78
97.5    1    70
97.5    1    75
97.6    1    74
97.6    1    69
97.6    1    73
97.7    1    77
97.8    1    58
97.8    1    73
97.8    1    65
97.8    1    74
97.9    1    76
97.9    1    72
98.0    1    78
98.0    1    71
98.0    1    74
98.0    1    67
98.0    1    64

```

98.0	1	78
98.1	1	73
98.1	1	67
98.2	1	66
98.2	1	64
98.2	1	71
98.2	1	72
98.3	1	86
98.3	1	72
98.4	1	68
98.4	1	70
98.4	1	82
98.4	1	84
98.5	1	68
98.5	1	71
98.6	1	77
98.6	1	78
98.6	1	83
98.6	1	66
98.6	1	70
98.6	1	82
98.7	1	73
98.7	1	78
98.8	1	78
98.8	1	81
98.8	1	78
98.9	1	80
99.0	1	75
99.0	1	79
99.0	1	81
99.1	1	71
99.2	1	83
99.3	1	63
99.4	1	70
99.5	1	75
96.4	2	69
96.7	2	62
96.8	2	75
97.2	2	66
97.2	2	68
97.4	2	57
97.6	2	61
97.7	2	84
97.7	2	61
97.8	2	77
97.8	2	62
97.8	2	71

97.9	2	68
97.9	2	69
97.9	2	79
98.0	2	76
98.0	2	87
98.0	2	78
98.0	2	73
98.0	2	89
98.1	2	81
98.2	2	73
98.2	2	64
98.2	2	65
98.2	2	73
98.2	2	69
98.2	2	57
98.3	2	79
98.3	2	78
98.3	2	80
98.4	2	79
98.4	2	81
98.4	2	73
98.4	2	74
98.4	2	84
98.5	2	83
98.6	2	82
98.6	2	85
98.6	2	86
98.6	2	77
98.7	2	72
98.7	2	79
98.7	2	59
98.7	2	64
98.7	2	65
98.7	2	82
98.8	2	64
98.8	2	70
98.8	2	83
98.8	2	89
98.8	2	69
98.8	2	73
98.8	2	84
98.9	2	76
99.0	2	79
99.0	2	81
99.1	2	80
99.1	2	74
99.2	2	77

```
99.2    2    66
99.3    2    68
99.4    2    77
99.9    2    79
100.0   2    78
100.8   2    77
;
run;
proc means;
  by gender;
run;
proc univariate normal plot;
var temp;
run;
```