Chapter 6

Interpreting Spatial Models[1]

This chapter aims to do two things. Part A focuses on how to estimate statistics, particularly the mean and standard deviation, from data that is only presented in summary form (like a frequency table or a histogram). Part B takes this one step further, by helping you connect two different ways of picturing data by relating histograms and boxplots. Both give a picture of how the data is spread out. The difference is that a boxplot takes the data and breaks it into four chunks with the same number of observations in each chunk, but with each chunk of data having a different length. Histograms are the opposite: each chunk has exactly the same length, but probably has different numbers of observations in it.

- *As a result of this chapter, students will learn*
  - √ Why summarized data cannot be used to compute an accurate mean or standard deviation
  - √ What a percentile is
  - √ What a cumulative distribution is
- *As a result of this chapter, students will be able to*
  - √ Estimate the mean from a set of summarized data
  - √ Estimate the standard deviation from a set of summarized data
  - √ Sketch a boxplot of the data underlying a histogram without having the data itself
  - √ Sketch a rough idea of a histogram of data based only on a boxplot of the data

6.1 Estimating Stats from Frequency Data

Many times we are presented with data, in newspapers, magazines, the Internet, or meetings, but these data are rarely presented in its entirety. After all, in many cases, there are thousands of observations of each variable. It is therefore more common to present summarized data in the form of tables or charts that show the number (or frequency) of observations that fall into a certain range (or bin). In the last chapter, we used this idea to create a graphical depiction of the data in the form of a histogram. But what if you are starting from the summarized data and what to know something about the original data itself?

For example, what if you wish to compute the mean of the data? This is the most frequently used measure of central tendency and is often used a model of the data. The way in which we compute this measure of central tendency is based on having all of the individual data points in the set of data. In a summarized table of data, though, we do not have the actual values to add up. One thing is certain; we cannot simply average the frequency counts, as this does nothing to account for the actual values of the data and the frequency counts are not (usually) even in the same units as the data itself. For example, in looking at the table below, we see data on salary distribution at a company. If we average the frequency counts (labeled "Number of Employees") we get 11.8, which means that if the distribution were uniform, there would be 11.8 employees in each salary range. But this number has units of number of people. The average salary must have units of dollars. Somehow, we must estimate the mean based on both the salary ranges and the number of observations in that range.

Salary Range	Number of Employees
\$200,000 - \$250,000	1
\$150,000 - \$199,999	2
\$100,000 - \$149,999	5
\$50,000 - \$99,999	13
\$0 - \$49,999	38

Unfortunately, as we'll discover, once you have only the summarized data, there is no way to get the actual mean of the original data. At best, you are estimating the mean, and your estimate has a great deal of possible error, depending on the size (width) of each bin into which the data has been summarized. These same ideas hold true for estimating the standard deviation of the data, especially since we must first estimate the mean in order to compute the deviations of each observation (or, in this case, each group of observations) from the mean.

And while it is true that in many cases we have the actual data and can compute the true mean of the data, this is often not true. Have you every filled in a customer satisfaction survey? Such surveys often collect demographic data, such as the age of the person filling in the form, but rarely do they ask you to write in your age. It is more common to check off a box marking a range where your age fits (for example, 31-40 years old). In situations like this, the data starts as a summarized frequency table; the company collecting the data never has the actual ages of each survey participant. So they must resort to estimating the mean if they need it for other calculations.

6.1.1 Definitions and Formulas

Summarized Data Summarized data is data not presented in raw form. Instead, the data has been grouped (or summarized) into categories. For example, rather than listing the salaries of all 250 employees at a company, a summarized presentation of this data might simply tell you the number of employees in each salary range, such as 10 employees making \$0 - \$20,000, 34 employees making \$20,001 to \$40,000 and so forth.

Weighted Average A weighted average is a type of mean where each item to be included in the average has a different weight depending on either its frequency or importance. One of the most common weighted averages is a student's GPA in college. Each class is assigned a value, based on the grade (usually a number from 0 - 4 quality points) and is assigned a weight based on the number of credit hours (3 for a three credit course, 4 for a 4 credit course, etc.) The overall GPA is then computed by weighting each grade (multiply the quality points by the weight [number of credit hours]), adding these weighted grades up, and dividing by the total number of credit hours (which is just the sum of all the weights). This means that a low grade in a high weight course (one with more credit hours) is more damaging than a low grade in a course with few credit hours. Another common use of weighted averages is to estimate the mean of a set of data given by a frequency table. In this case, the weight is determined by the frequency counts. For example, if 10% of a class scored 50 on an exam, 20% scored 60, 40% scored 70, 10% scored 80 and 20% scored 90, then the class average is

$$\frac{0.10(50) + 0.20(60) + 0.40(70) + 0.10(80) + 0.20(90)}{0.10 + 0.20 + 0.40 + 0.10 + 0.20} = \frac{71}{1} = 71.$$

More generally, if the data are given by x_i and the weights are given by w_i, the weighted average of the data is given by

$$\text{Weighted Average} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Weight Each item to be included in a weighted average is assigned a weight that identifies how much that item contributes to the overall average. The weight assigned to the ith piece of data is commonly denoted by the symbol w_i.

Estimated Mean When computing the mean of data that is given only by frequency tables, we cannot compute the actual mean since we only know the number of data points falling into each range of the frequency table. We must estimate the value for the data in each bin of the table, which results in an overall estimate of the mean. Estimated means from frequency data are computing by using the formula for the weighted average, with the weights given by the frequency counts (the number of pieces of data in the bin). The symbol $\bar{x}_{est}$ will be used to represent the estimated mean.

Estimated Standard Deviation Estimating the standard deviation from frequency data is similar to the process of estimating the mean, but involves a few more steps. First,

of course, we need to estimate the mean of the data. Now, if each of the data points (or central points of each bin in the data, more precisely) is denoted by x_i and the frequency (number of items in that bin of the frequency table) is given by w_i, then the estimated standard deviation is

$$S = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_{est})^2}{(\sum_{i=1}^{n} w_i) - 1}}$$

Note that the sum of all the weights

$$\sum_{i=1}^{n} w_i = w_1 + w_2 + \ldots + w_n = n = \text{Total number of data points.}$$

Sumproduct Another way to think of weighted averages is to think of lining up the data in one column and the weights in another column. By multiplying line-by-line, we find the contribution of each item to the weighted average. Adding these contributions results in calculating the top part of a weighted average. In Excel, the SUMPRODUCT function takes two lists (one is data, one is weights) and carries out this computation. This is equivalent to thinking of the items as a vector and the weights as a vector and computing the vector dot product or scalar product.

If the two lists are given by $x = (x_1, x_2, \ldots, x_n)$ and $w = (w_1, w_2, \ldots, w_n)$, the scalar product or SUMPRODUCT of these two lists is simply

$$\text{SUMPRODUCT}(x, w) = \sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n.$$

6.1.2 Worked Examples

Example 6.1. Estimating the mean
Suppose we have been presented with a table of information like that below which summarizes the salaries of the employees at a company. What is the average salary at your company (as determined by using the mean)?

Salary Range	Number of Employees
\$200,000 - \$250,000	1
\$150,000 - \$199,999	2
\$100,000 - \$149,999	5
\$50,000 - \$99,999	13
\$0 - \$49,999	38

We notice several problems immediately. First, we do not have the actual salaries of each employee; we only know a range of salaries. Second, each employee in each range could have a different salary within that range. How can we compensate for this?

The first problem is one of identifying a particular salary to represent each range. Common choices are the midpoint of the range and either endpoint. We will start with the midpoint and then repeat the analysis using the endpoints. The second problem is more interesting. Typically, we assume that all the observations in a given range of salaries are the same. Clearly, this is a poor assumption, but without it we cannot really get anywhere. This ambiguity in dealing with summarized data is why we can only claim to be estimating the mean of the data, not computing it. Based on these assumptions, we now have the following table of data to work with (after rounding the salaries off).

Salary	Number of Employees
\$225,000	1
\$175,000	2
\$125,000	6
\$75,000	13
\$25,000	38

So, is the average just (225,000 + 175,000 + 125,000 + 75,000 + 25,000)/5 = \$125,000? That would seem to be a bit high, wouldn't it, since only 8 employees make that much money and 51 employees are below that level. The problem with this kind of computation is that each of the salary ranges must be weighted. This means that we must put several copies of each salary into the calculation, one copy for each observation that matches that data value. Think about it this way, if we had a complete list of the salaries for computing the mean, it would look something like the table below.

225,000	75,000	75,000	25,000	25,000	25,000
175,000	75,000	75,000	25,000	25,000	25,000
175,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
75,000	75,000	25,000	25,000	25,000	

To add up all these salaries and compute the mean, we would need to include 2 copies of the \$175,000 salary, 6 copies of the \$125,000 salary and so on. Thus, we estimate the mean (measuring the data in thousands of dollars) as

$$\frac{225(1) + 175(2) + 125(6) + 75(13) + 25(38)}{1 + 2 + 6 + 13 + 38} = \frac{3,250}{60} = \$54.166.$$

Clearly this number is more reasonable for the salary. How can we estimate such means in general? First, we need to assign symbols to each quantity. Suppose we have N different groups of data. In the above example, we have 5 groups of data, so $N = 5$. Let the value of the i^{th} range be (where) and we will let the number of data points in the i^{th} range be given by . With these naming conventions, the data table above would look like this:

Salary	Number of Employees	Product
x_i	n_i	$x_i n_i$
$x_1 = \$225{,}000$	$n_1 = 1$	225,000
$x_2 = \$175{,}000$	$n_2 = 2$	350,000
$x_3 = \$125{,}000$	$n_3 = 6$	750,000
$x_4 = \$75{,}000$	$n_4 = 13$	975,000
$x_5 = \$25{,}000$	$n_5 = 38$	950,000
Total	60	3,300,000

Now, to compute the mean, we multiply each data value (the x's) by its weight (the n's) and add these up. Then we divide by the total number of observations (the sum of all the n's):

$$\bar{x}_{est} = \frac{n_1x_1 + n_2x_2 + n_3x_3 + n_4x_4 + n_5x_5}{n_1 + n_2 + n_3 + n_4 + n_5}$$

Using the sigma notation, this becomes much easier to write down:

$$\bar{x}_{est} = \frac{\sum_{i=1}^{n} n_i x_i}{\sum_{i=1}^{n} n_i}$$

Now, in Excel, the numerator can be computed as a "sum product" of two lists of numbers. One is the list of weights and the other is the list of the values associated with the data ranges. Using this, we can write the Excel version of the weighted average formula as

$$\bar{x}_{est} = \frac{\text{SUMPRODUCT(weights, values)}}{\text{SUM(weights)}}$$

Example 6.2. Effects of different midpoints
Using example the previous example, we can compute the average is several different ways, using different values for the data range. Below is a table showing the estimation of the mean for the salary data above, using the midpoint, left endpoint and right endpoint of each range of values. Keep in mind, though, that unless you have a very good reason for doing otherwise, you should probably use the midpoint to estimate the mean, since that will likely be a better representation of the data in any particular bin. Using the endpoints implies that the data within a particular bin is highly skewed, which might be the case, but would need justification.

Salary Range	Number of employees	Midpoint	Left	Right
\$200,000 - \$249,999	1	225,000	200,000	249,999
\$150,000 - \$199,999	2	175,000	150,000	199,999
\$100,000 - \$149,999	6	125,000	100,000	149,999
\$50,000 - \$99,999	13	75,000	50,000	99,999
\$0 - \$49,999	38	25,000	0	49,999
	Estimated Average	\$54,166.67	\$29,166.67	\$79,165.67

As you can see, the choice of where to place the value for each data range has a huge effect on the estimate of the mean. In fact, if each of the ranges has the same degree of spread (all of the ranges above cover \$49,999) the estimate of the mean from the left and right endpoints will differ by the spread (notice that \$79,165.67 - \$29,166.67 = \$49,999, the spread of the data in each of the ranges). Mathematically, this is easy to prove. Assume that each range has a spread of S. Then, if the left endpoints of the ranges are given by x_i, the right endpoints are given by $x_i + S$ and the estimate for the mean using the right endpoint will be (all sums are from $i = 1$ to $i = n$)

$$\bar{x}_{est} = \frac{\sum n_i(x_i + S)}{\sum n_i} = \frac{\sum n_i x_i + \sum n_i S}{\sum n_i} = \frac{\sum n_i x_i}{\sum n_i} + \frac{n_i S}{\sum n_i} = \frac{\sum n_i x_i}{\sum n_i} + S\frac{\sum n_i}{\sum n_i} = \frac{\sum n_i x_i}{\sum n_i} + S.$$

However, there is any easier way to see what happens using different estimates for the data points. Recall that if we add the same amount to every single data point it will shift the mean by that amount exactly. So if we add 10 to each data point, the mean will increase by 10.

Thus, with summarized data, we can never nail down an exact value for the mean. At best, we can estimate it to fall within a particular span of values that is closely tied to the spread each range of data covers.

Example 6.3. Averaging Averages
We can also use the previous examples to understand why we cannot average several averages: Each average must be weighted by the number of data points used to compute that average. For example, if we have two sections of a course being taught and the two sections take the exact same final examination, we might want to determine the overall average on the final exam before deciding what grades to give. If one class scored an average of 82 on the final exam and the other class scored an average of 75 on the final, we cannot simply say that the overall average is $(82 + 75)/2 = 78.5$ because the two classes may have very different numbers of students. The table below uses the correct method, weighted averages, to determine the overall average for different sizes of each class. Notice that the more students there are in the class with the high average, the closer the overall average is to that class's average score.

Class 1 Size (test ave = 82)	Class 2 Size (test ave = 75)	Overall Average
10	30	76.75
15	15	78.5
30	10	80.25
35	5	81.125

Also note that when averaging averages, we are not estimating the mean; in this case we are computing the actual mean of the combined data. Algebraically, we can see why. To compute the weighted average in the above case, the numerator will be $n_1\bar{x}_1 + n_2\bar{x}_2$, but when we multiply an average by the count of data points (the n) we are getting the actual total of all the data points, since

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N} \Longrightarrow N\bar{x} = \sum_{i=1}^{n} x_i.$$

So in this case, the numerator is the exact total of all the data points in each class, even though we do not have the individual scores for any particular student!

Example 6.4. Estimating Standard Deviation
Estimating the standard deviation for summarized data is not that much different from calculating the standard deviation normally. Recall that the formula for the sample standard deviation of a set of data given by $x_1, x_2, \ldots, x_n$ is simply

$$\sigma_{n-1} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

When we only have a summary of the data, though, we must estimate the data points and then weight the deviations based on the number of data points falling near that estimated value. Thus, the formula becomes

$$\sigma_{n-1} = \sqrt{\frac{\sum_{i=1}^{m} n_i(X_i - \bar{x}_{est})^2}{(\sum_{i=1}^{m} n_i) - 1}}.$$

where we have used the symbol X_i to refer to the estimate of the value of the data in the ith bin of the summarized data, n_i is the number of data points in that bin, and m is the total number of bins in the summary.

Using the data from the previous section, we can compute each piece of the formula and combine them into an estimate of the standard deviation in the salaries. We will use the midpoint estimate for both the mean and the standard deviation. Recall that the estimate for the mean salary was \$54,166.67.

So, the estimate for the standard deviation is

$$\sigma_{est} = \sqrt{\frac{1.26458E+11}{60-1}} = \$46,296.45.$$

Salary Range	Number of employees (n_i)	Midpoint (X_i)	Deviation ($X_i - \bar{x}_{est}$)	$n_i(X_i - \bar{x}_{est})^2$
\$200,000 - \$249,999	1	225,000	170,833	29184027778
\$150,000 - \$199,999	2	175,000	120,833	29201388889
\$100,000 - \$149,999	6	125,000	70,833	30104166667
\$50,000 - \$99,999	13	75,000	20,833	5642361111
\$0 - \$49,999	38	25,000	-29,167	32326388889
Sum	60			1.26458E+11

Computing the quartiles for this data is relatively easy. Since there are 60 data points, each quartile should contain 15 points ($60/4 = 15$). Thus, starting at the smallest value, the first quartile is between the 15th and 16th data points. This would place the first quartile in the first bin, between \$0 and \$49,999. The second quartile would fall between the 30th and 31st data points, also in the first bin. The third quartile would lie between the 45th and 46th data points, placing it inside the second bin, between \$50,000 and \$99,999. Our estimates of the statistics relating to this data are then gathered together below.

Mean	54,166.67
Standard deviation	46,296.45
Q1	25,000
Q2 = median	25,000
Q3	75,000

With summarized data, this is about as accurately as one can estimate the standard deviation and the quartiles, since better estimates would require narrowing down the bin widths so that there is not so much possible variation in scores inside a particular bin.

6.1.3 Exploration 6A: Data Summaries and Sensitivity

Open the data file C06 ExplorationA.xls. This file contains 1000 observations from five random distributions of data. Each data set contains values in the range of 0 to 100. In this exploration, you are going to investigate two aspects of the data. You will look at the error in estimating the mean and standard deviation of data from different distributions (symmetric, positively skewed, negatively skewed, bimodal and uniform) and you will explore how these errors are affected by the way the data is summarized.

1. To prepare the data for this exploration, we need to name the ranges of data. Highlight the cells in the column labeled "Symmetric", click in the name box and type "Symm" for the name of this set of cells. Repeat this process for each of the remaining columns of data, naming them, respectively, "PSkew", "NSkew", "BMod" and "Unif". This will make entering all our formulas much easier.

2. We are now going to compute the actual mean and standard deviation for each set of data and record the results in the table below for later analysis. To compute these statistics, either use StatPro's one-variable summary statistics routine or simply type =AVERAGE(Name of cell range) to compute the mean and =STDEV(Name of cell range) to compute the standard deviation.

3. Next, we'll create our own frequency table. First, in cell H1 enter the label "Left" and in I1 enter "Right". In cell H2 enter "0" and in cell I2 enter 10. In cell I3 enter 20. Now highlight I2:I3 and use the fill handle to drag down until you get to 100 (in cell I11). In cell H3 enter "=I2" and copy this formula down all the way to cell H11. This should create a column that counts from 0 to 90 by 10 in column H, and a column from 10 to 100 in column I. Finally, highlight all the cells in the range I2:I11 and name them "Bins".

4. In cells J1:N1 copy the names of the five data sets from cells B1:F1.

5. Now to compute the frequency data. Highlight cells J2:J11 and then just type the following formula "=FREQUENCY(Symm, bins)" then hit CTRL+SHIFT+ENTER at the same time to enter the array formula. You should see a frequency table appear. Now, repeat this for the remaining data sets, by typing the formula into the appropriate cell ranges and substituting the name of the data set for the "Symm" label in the formula.

6. Now that you have the frequency data, you can use the techniques of this section (and the computer How To Guide for this section) to estimate the mean and standard deviation for the data sets. Enter the results in the table below.

7. Now, go back and change the bin sizes. In cell I2:I11 enter the values 25, 50, 75, 100, 100, 100, 100, 100, 100, 100. (It looks odd, we know, but this way, you don't have to start over, and you should almost instantly see the new estimates for the mean and standard deviation, without having to do more work.)

8. Once you have recorded the results of your calculations in the table below, think about how our assumptions work when estimating the mean and standard deviation. For which types of data are these estimates most accurate? Why? For which are the estimates least accurate? Why? Keep in mind, these errors may seem small; typically the most error you get with these data is about 0.5 to 1.0, but that's about a 1% to 2% error in estimating these statistics! How does the bin width affect the accuracy of the estimates?

	Symmetric	Positively Skewed	Negatively Skewed	Bimodal	Uniform
Actual Mean					
Mean (bin width = 10)					
Mean (bin width = 25)					
Actual St Dev					
St Dev (bin width = 10)					
St Dev (bin width = 25)					

6.1.4 How To Guide

Estimating the mean in Excel

We will start by assuming that we want to enter the data from example 1 (page 166) and estimate its mean. The easiest way to do this is to structure the data table in Excel as shown below in cells A1:B6, with the minimum value of each salary bin (or whatever variable is used to summarize the data) in one column (marked "Low" in the example) and enter the maximum values of each bin (marked "High") in a separate column. This will allow us to enter these as numbers and to have Excel to calculate the middle of each bin for use in the calculations.

	A	B	C	D	E
1	Low	High	Number of Employees	Midpoint	Count*Mid
2	200000	250000	1	$225,000.00	225000
3	150000	199999	2	$174,999.50	349999
4	100000	149999	5	$124,999.50	624997.5
5	50000	99999	13	$ 74,999.50	974993.5
6	0	49999	38	$ 24,999.50	949981
7					
8		SUM	59		3124971
9					
10		Mean (estimated)			$ 52,965.61

Figure 6.1: Estimating the mean of a set of data given only by frequency counts.

Now, column D contains your estimate of the midpoint for each bin. In cell D2, enter

```
= (A2 + B2)/2
```

Now copy this formula down the column. Column E is just the individual frequency counts multiplied by the midpoint for the bin, so in E2, enter

```
= D2*C2
```

After copying this formula down, you will need to compute the sum of the frequency counts (in cell C8 in the example) and the sum of the frequency count*midpoint data (in cell E8 in the example). The estimate for the mean is then simply **=E8/C8**.

Estimating the standard deviation of the data in Excel

To estimate the standard deviation for the data shown in the previous example, we need to add two columns to the data. Column F will contain the deviation of each data bin from the estimated average. Column G will contain the frequency data times the square of the deviations. This is shown in the figure below.

To compute the deviations, enter the formula below in cell F2 and copy it down.

	A	B	C	D	E	F	G
1	Low	High	Number of Employees	Midpoint	Count*Mid	Deviation	Count*Dev^2
2	200000	250000	1	$225,000.00	225000	$172,034.39	$ 29,595,831,284.36
3	150000	199999	2	$174,999.50	349999	$122,033.89	$ 29,784,540,534.33
4	100000	149999	5	$124,999.50	624997.5	$ 72,033.89	$ 25,944,406,420.57
5	50000	99999	13	$ 74,999.50	974993.5	$ 22,033.89	$ 6,311,399,913.82
6	0	49999	38	$ 24,999.50	949981	$ (27,966.11)	$ 29,719,926,084.46
7							
8		SUM	59		3124971		1.21356E+11
9							
10		Mean (estimated)			$ 52,965.61		
11							
12					Std Dev (estimated)		$ 45,742.18

Figure 6.2: Estimating the standard deviation of a set of data given only by frequency counts.

```
= D2 - $E$10
```

To compute the frequency times the square of the deviations, enter the formula below in cell G2 and copy it down.

```
= C2*F2*F2
```

Now, sum up the weighted squares of the deviations by entering

```
= SUM(G2:G6)
```

in cell G8. To complete the calculation that estimates the standard deviation, we need to divide the sum of the squared deviations by the total number of observations (minus 1 for the loss of one degree of freedom) and take the square root. So we enter the following in cell G12 (Note that, in this formula, the parentheses are all necessary in order to enforce the proper rder of operations):

```
= SQRT(G8/(C8-1))
```

6.2 Two Perspectives are Better than One

Open any newspaper or magazine and you will come across graphs and representations of data that are supposed to help you make sense of some issue or help you decide whether to vote in favor of some proposition or not. You will eventually find yourself sitting in a meeting listening to a presentation with graphs and charts in it. You will probably have employees sending you reports with graphical representations of data designed to help you make a decision. However, it is relatively easy to manipulate your perceptions by presenting a particular graph. By choosing how to present the information, the writer can control the way you perceive the issue. This is true even when the writer is supposedly objective.

With a little work, though, you can look at a graph and mentally convert it to another type of graph. This will provide you with the flexibility of seeing data from multiple perspectives, gaining a much deeper insight into the way the data is structured. This, in turn, will help you make more informed decisions and will help you recognize when someone is trying to manipulate the presentation of the data toward a certain end. For this section, though, we will concentrate on the connections between boxplots and histograms, and we will develop ways to picture one graph when presented with the other type of graph.

Another key benefit to having this flexibility is that you can use a boxplot to help decide how to set up a histogram. Often, it is difficult to set up a useful histogram on the first try. Look back at the histogram of Beef N' Buns service times in Exploration 6A. If the student had first created the boxplot shown below, she might have had a better starting point for making the histogram.

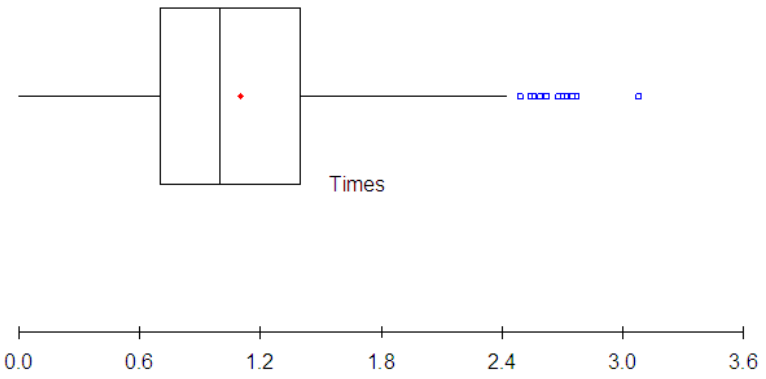

Figure 6.3: Boxplot of service times at Beef n' Buns.

Based on this, she might have set a minimum value of about 0.3 (about halfway between 0 and the first quartile). Then, using 10 bins for the histogram to cover the range from 0.3 to 2.4 would make the bin widths (2.4 - 0.3)/10 = 2.1/10 = 0.21 which is about 0.2 (round off to make nice bins in the graph). She could then add two bins (for the "¡= 0.3" and the open bin on the right side) making the histogram shown below. This graph clearly shows that the data is positively skewed, indicating that the bulk of the service times are below the mean service time (about 1.1 minutes, based on the boxplot).

This section will involve a lot of estimation and inferencing. Estimation involves making

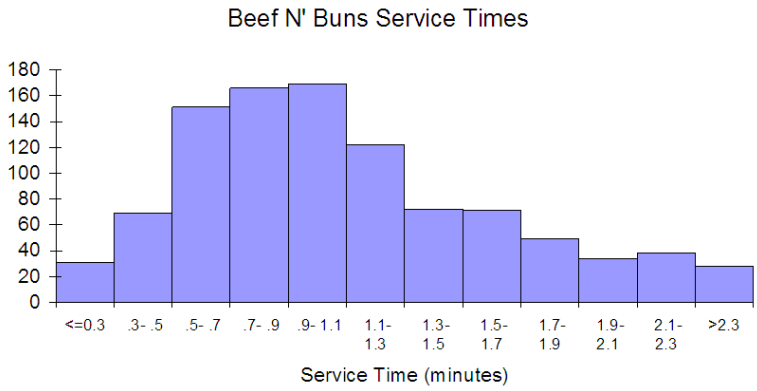

Figure 6.4: Histogram of service times at Beef n' Buns.

a rough guess at some quantity based on either scale or units. Inferencing involves drawing conclusions based on limited information. In order to inference, you will have to interpret the information you are given and "fill in the missing pieces" since you will not have complete information. As part of this process, notice that in the Beef N' Buns service times, reporting the "average service time" as 1.102 minutes (this is the mean) with a standard deviation of 0.542 minutes would misrepresent the situation. Since the data is positively skewed, we can see that most of the data falls to the left (below) the mean service time. In fact, the three largest bins of the histogram are to the left of the mean. This tells us that the mean may not be the best choice for representing the average service time. The median service time of 1 minute may be a better choice. Thus, we can use a combination of graphs to learn more about the data than we could learn from either graph individually. We might also infer that the reason the data is positively skewed has nothing to do with our service overall, but rather with specific orders. If certain orders are taking longer, but these orders do not occur that often, then we might see a few high service times (as high as 3 minutes from the boxplot!) These service times are clearly outliers, and they fall almost four standard deviations from the mean. We could even analyze the percentage of service times within one, two, and three standard deviations above and below the mean (a histogram of the z-scores for the service times would help) to determine whether we should be concerned at all with the service times at Beef N' Buns.

6.2.1 Definitions and Formulas

Percentiles Percentiles are similar to quartiles, except that the data is broken into one hundred pieces, rather than four. For comparison, the first quartile is the same as the twenty-fifth percentile, since one-fourth of 100 is 25. The median is the same as the 50th percentile, and the third quartile is the same as the 75th percentile. Percentiles are often used to break the data down even further than is possible with quartiles. The strict definition of the nth percentile is that it is the observation below which n% of the data falls. Thus, 90% of the data is less than the 90th percentile and 99% of the data is less than the 99th percentile.

Cumulative Distribution Cumulative distributions are similar to histograms. However, each bin in a cumulative distribution includes all of the observations in the bins to the left of the bin as well. Usually, the number of observations in each bin is expressed as a percentage of the total number of observations so that the right-most bin will have 100% of the observations in it.

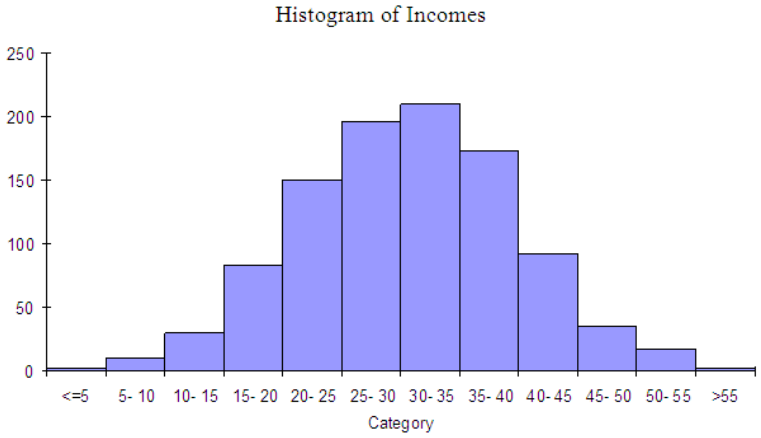

Figure 6.5: Histogram of Incomes.

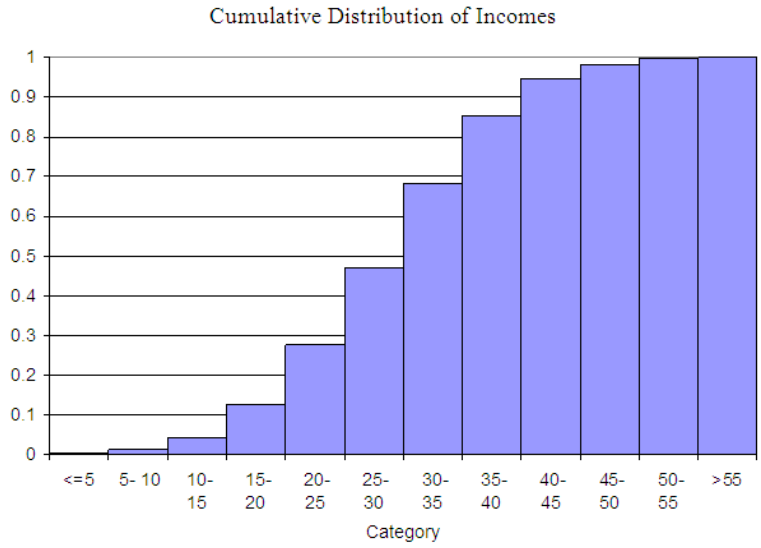

Figure 6.6: Cumulative Distribution of Incomes.

6.2.2 Worked Examples

Example 6.5. From histograms to cumulative distributions

Consider the data on family incomes in Country A from "P04P 02Incomes.xls". These are shown in a histogram below. There are 1,000 total observations in the data.

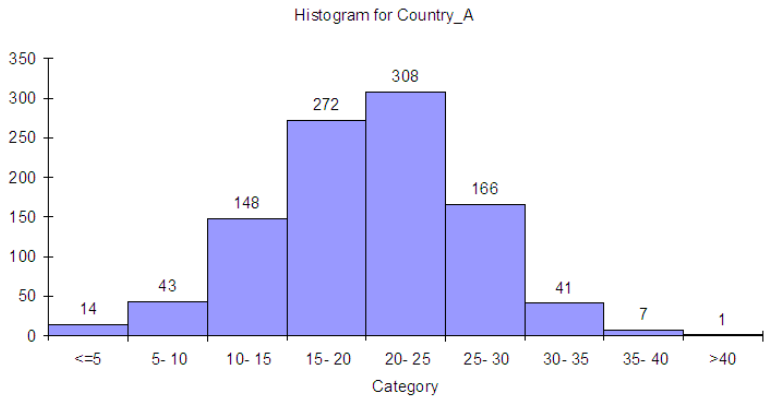

Figure 6.7: Histogram of incomes for 1,000 families in country A.

To convert this graph to a cumulative distribution, we simply start adding. In the first bin, labeled "¡= 5", we have a total of 14 observations. In the second bin, we have 43 observations. In the cumulative distribution, the second bin will have 14 + 43 for a total of 57 observations, since it includes all the bins to the left. The third bin of the cumulative distribution will have 148 + 57 = 205 observations. The fourth bin "15 - 20" will have 272 + 205 = 477. Continuing on, we get the totals shown in the graph below.

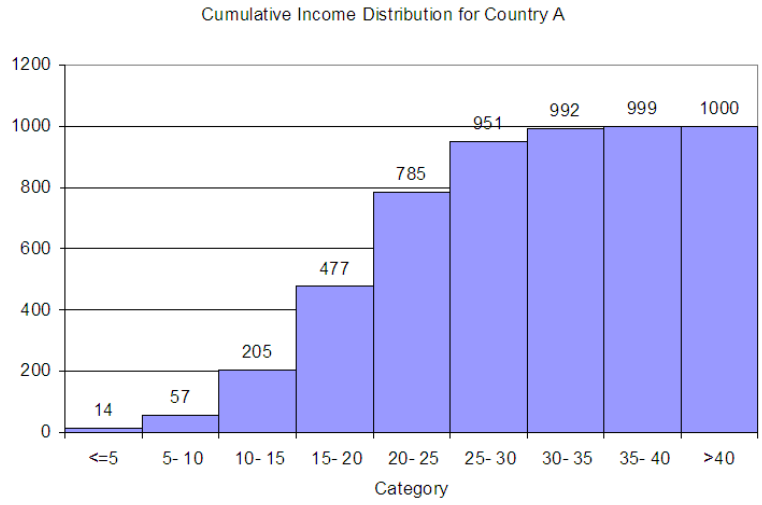

Figure 6.8: Cumulative histogram of incomes for 1,000 families in country A.

Example 6.6. Using the cumulative distribution to sketch a boxplot
Now, to generate a boxplot of the data above, we probably want the cumulative distribution graphed in terms of percentages (of the total number of observations), rather than total

amounts. This graph is shown below. Once we have this, it is relatively easy to determine in which bin each of the quartiles falls. Remember the first quartile is the same as the 25th percentile, so in the graph below, we know that the first quartile is somewhere in the bin marked "15 - 20". Since 20.5% of the data is to the left of this bin, we can probably guess that the first quartile will be close to the left side of the "15 - 20" bin. We can also find the median; 50% of the data is less than the median, so it must be in the fifth bin, marked "20 - 25". It is probably close to the left edge of this bin. Interestingly enough, the third quartile includes 75% of the data to its left, so it is also in the fifth bin, "20 - 25". Q3 is probably close to the right side of the bin. Based on these estimates, then, we can sketch a boxplot on the same scale axis as the histogram. We know where the minimum and maximum are, so we can also compute whether there are any outliers in the data. For this graph, the largest the IQR could be is 10, since Q1, the median and Q3 are in the fourth and fifth bins. Thus, anything further than 15 (three bins) from the end of either side of the box must be an outlier. (The histogram itself shown only one observation in the last bin, so it is the only outlier.)

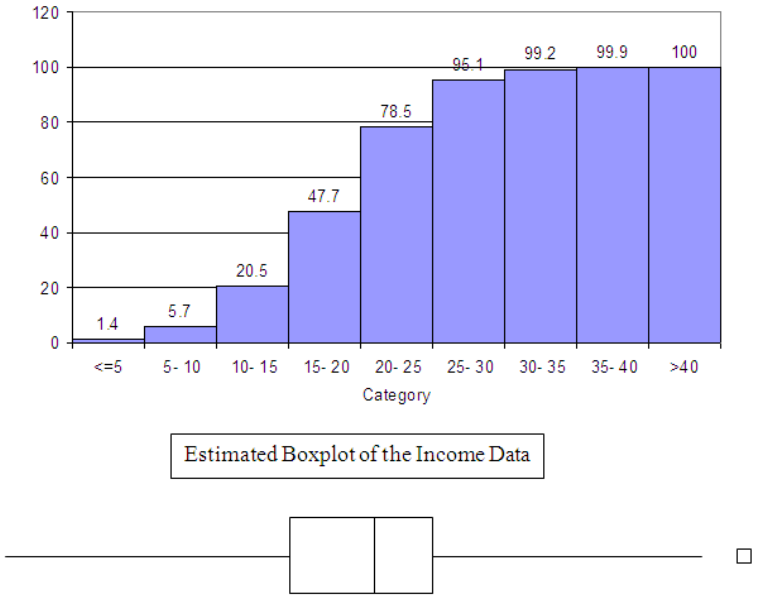

Figure 6.9: Cumulative distribution (as percents) and an estimated boxplot of family incomes.

Example 6.7. From a boxplot to a histogram

There are several ways to sketch a histogram from a given boxplot. One way is to reverse the process in the previous two examples. But there is a quicker way to sketch the histogram, based on the shape of the boxplot. Consider the graphs below, which show four basic histograms and their associated boxplots. All graphs are on the same 0 to 100 axis.

As you can see, the box of the boxplot falls in about the same place as the large bulk of the data. This means that you can start with a boxplot and sketch the bulk-part of

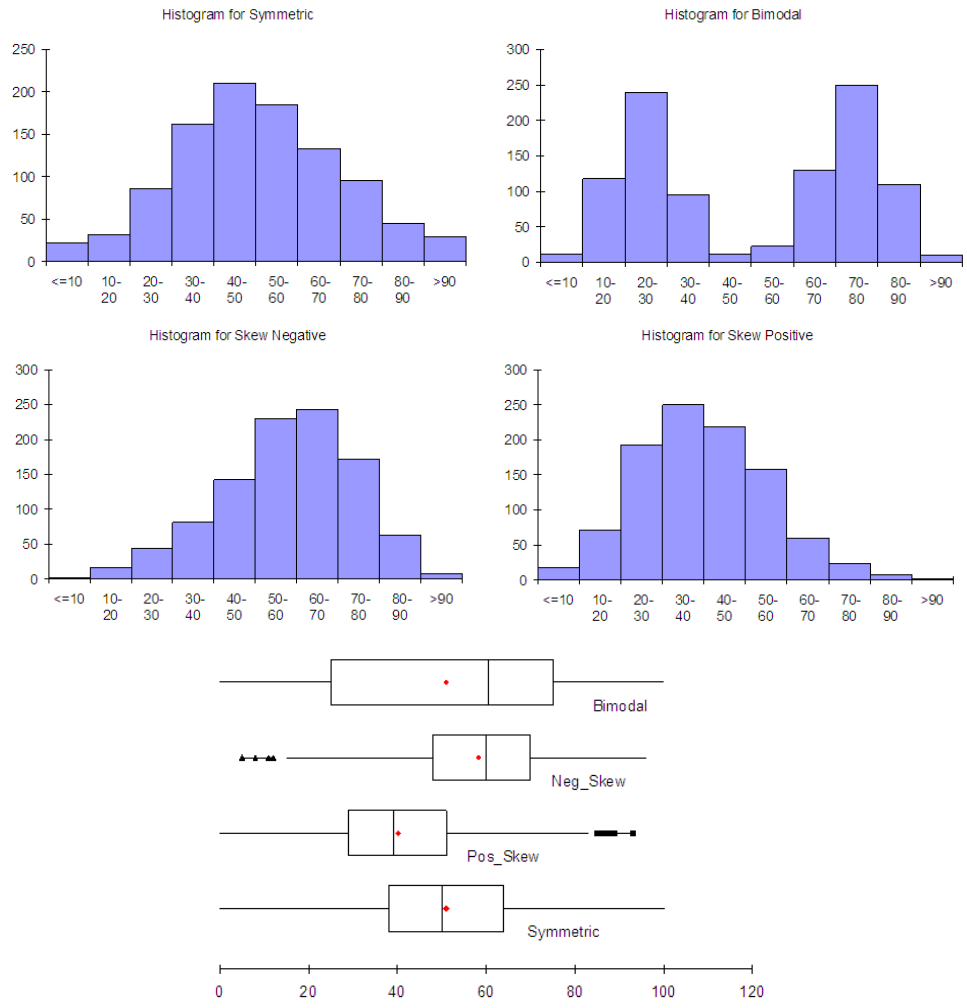

Figure 6.10: Four distributions shown as histograms and their corresponding boxplots.

the histogram where the box is if the box is fairly narrow (if it is a wide box, then the distribution is probably bimodal). Notice that the box is centered between the min and max for a symmetric distribution and for a bimodal distribution, except that a bimodal distribution has a larger spread, so the box is very long. In fact, for bimodal distributions, the box tries include both of the "humps" in the distribution. Notice that for skewed distributions, the box is still where the bulk of the data is, but it is offset from the center. For positively skewed distributions, the bulk is shifted to the left, and the mean is greater than the median. For negatively skewed distributions, the bulk is shifted to the right, and the mean is less than the median.

6.2.3 Exploration 6B: Stock Investment Decisions

You have just started working for a new company, Impressive Business Machinery. As part of the paperwork for your hiring, you have been asked to choose an investment stock for your retirement planning. Your employer offers you four choices and provides you with histograms (figure 6.11) of the daily returns for these stocks over the last 3 months. (You suspect that your employer is testing you, but you can't be sure.) For the near future, which of these stocks would you choose? Why would you choose that stock? How will you justify your decision to your family if it does not perform as well as expected?

1. Which stock did you choose? Why?

2. Discuss your ideas with a partner. Do you still agree that your original choice of stock was the best?

   (a) If your ideas have changed, what influenced those changes?

   (b) If your ideas have not changed, what strengthened them?

3. What makes the selection of a stock easy? What makes it difficult?

   It may be helpful to sketch the cumulative distribution and a boxplot for each of the stocks. Each graph contains 96 observations (about three months worth of data). It will also be helpful to rank the four stocks from highest to lowest in terms of both the mean and the standard deviation.

Statistic	Highest	Med-High	Med-Low	Lowest
Mean				
Standard Deviation				
Minimum				
First Quartile				
Median				
Third Quartile				
Maximum				

4. After sketching your graphs and completing your estimates, has your decision as to which stock to select changed? Why or why not?

5. Does your selection of a stock depend on what your investment goals are? In what way?

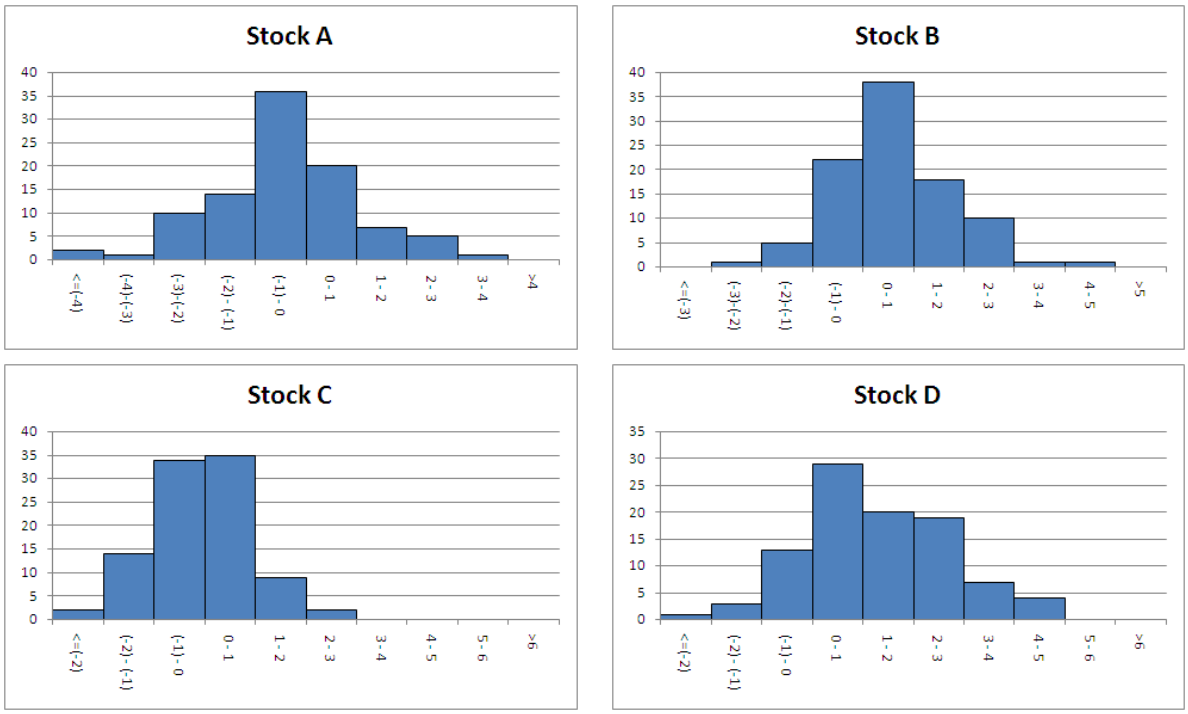

Figure 6.11: Graphs for Exploration Stock Investment Decisions

6.2.4 How To Guide

Percentiles in StatPro

Using StatPro, it is easy to calculate the 1st, 5th, 10th, 90th, 95th and 99th percentiles. Start by generating the "One variable summary statistics" (see chapter 3). At the bottom of the dialog box with the statistic options, select the percentiles that you want calculated.

Percentiles in Excel

To calculate percentiles in Excel, use the formula

```
=PERCENTILE(array of cells, percentile)
```

Note that percentile should be entered as a decimal number. Thus, for the 80% percentile, you should enter 0.80. For the 35th percentile, enter 0.35.

Cumulative Distributions in Excel (with help from StatPro)

First make a histogram of the actual data, using whatever minimum, number of categories, and category length makes sense. Notice that StatPro generates two worksheets. One will be called "Hist-(variable name)" and contains the actual histogram. The other sheet is called "Hist-(variable name)Data" and contains a frequency table of the data which is used to generate the histogram. To make the cumulative distribution graph, we will use the

frequency table and add two columns to it. We will then make a column chart of the data under the "Category" column and the data in our last column that we create.

Start by adding headings for each column into cells D3 ("Total") and E3 ("Percent").

To calculate the values for the "Total" column, enter the following formulas: In cell D4, enter "=C4". This looks up the first bin (C4). In cell D5, enter "=D4 + C5". This adds the next bin (C5) to the previous total (D4). Now, select cell D5 and double-click on the fill handle. This should complete the column. The total in the last row should be equal to the total number of observations in the data.

To calculate the values for the "Percent" column, enter the formula "=D4/(last total)" where (last total) is an absolute cell reference to the last cell in the "Total" column. If you have 12 categories, this cell will be cell D15. If you have a different number of categories, just enter the appropriate cell reference, but make sure it is an absolute reference.

Now we have all the data we need; all that's left is to make a bar graph of the data in columns B ("Category") and E ("Percent"). To do this, select all the cells with data in column B. Now, hold down the CONTROL (CTRL) key and select the data cells in row E. When you release the mouse button, all the data in both columns should be highlighted, but nothing else should be highlighted. Be sure to include the headings for each column as well, if you want the graph to automatically set up labels for the axes.

Now click on the "chart wizard" button on the toolbar. Select the "Column" type and choose the subtype in the upper left corner, the "clustered column" graph. Click "Next". At step 2 click "Next". At step 3, enter the titles you want and adjust the legend (probably you can remove the legend; it's not very helpful and clutters up the graph). To make the chart look like StatPro's histograms, click "As new sheet..." in step 4.

Now you can clean the chart up to make it look better. Right click on the grey area and select "Format plot area". Under "Area", select "None" and hit "OK". This will remove the background, making it easier to print and easier to read when you copy the chart into a Word document. Next, click on one of the columns in the graph and select "Format Data Series". Go to the last tab, marked "Options". Set the gap width to 0 (either type "0" in the box or use the arrow buttons) and hit "OK". You will be left with a graph that looks a lot like StatPro's!

Checking for Normality with StatPro and Histograms

When making histograms using StatPro, you might have noticed the "test normal fit" option in the lower left corner of the dialog box where you set up the categories.

If you click on this box, StatPro will add some information to your histogram and to the frequency table of the data. First off, you will notice that a dialog box appears over the histogram telling you whether the data could be from a normal distribution. To determine this, StatPro uses a p-value. We won't get into the way p-values are calculated, but the main idea is that if the p-value is close to 0, then the data is probably not from a normal distribution. Clicking "OK" will remove the dialog box and let you see the graph.

Notice that the graph has your data shown in blue, but there are other data outlined in between each of your columns. This extra information shows you what the heights of the bins should be if the data came from a perfect normal distribution. The bigger the difference between your data and the theoretical data, the lower the p-value.

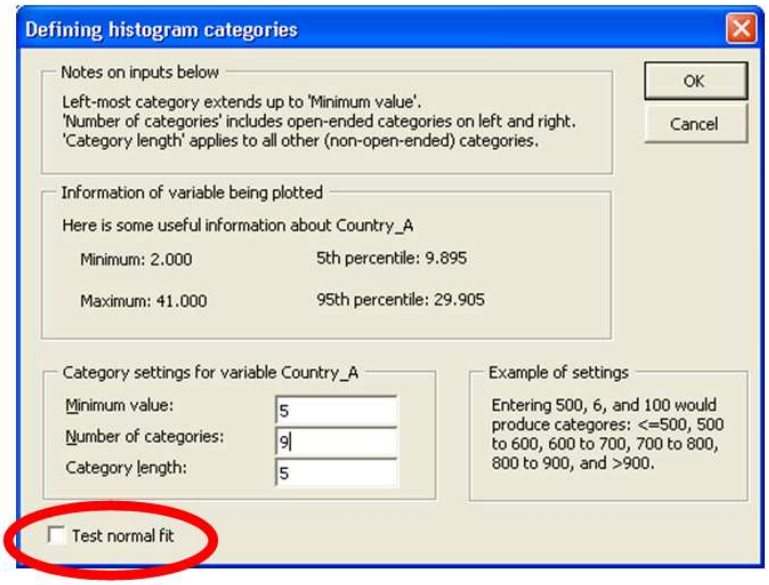

Figure 6.12: Testing a distribution for normalcy with StatPro's histogram routine.

On the frequency table, you will also notice additional information. Two new columns ("Normal" and "Distance measure") are in the table, and there is a "Test of Normal Fit" off to the side. The "Normal" column is used to graph the theoretical normal data. The other information is used to generate the p-value. Notice that there is no definite answer as to whether the data is or is not normal. This is because the data is a sample from a distribution; we don't have all the population information, so it's possible that the underlying population characteristic is normal, but your sample is slightly skewed.

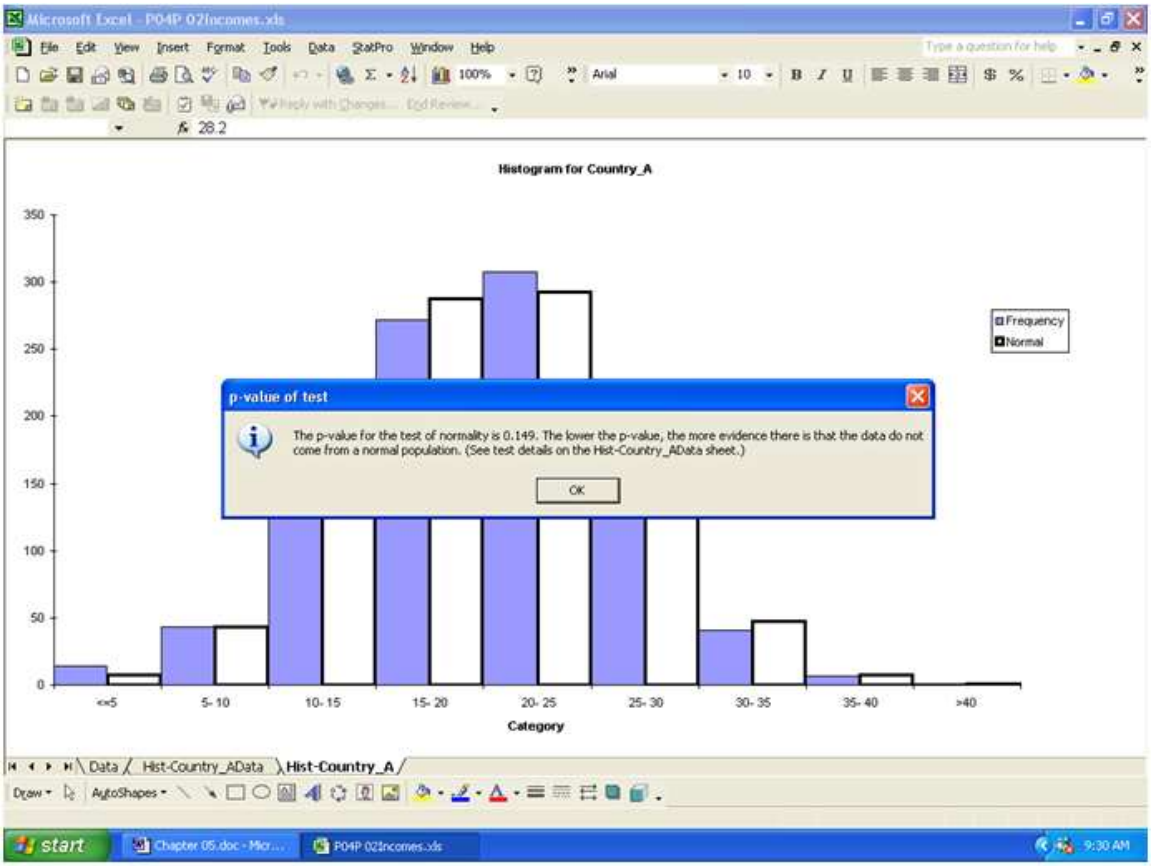

Figure 6.13: Results of testing for normalcy using StatPro.

6.3 Homework

6.3.1 Mechanics and Techniques Problems

6.1. Consider the summarized frequency data found in file "C06 Deliveries.xls". This shows data on the unloading times for trucks at StateEx, both broken down by truck type (Semi or Van) and overall.

1. Estimate the average unloading time for semis, vans, and overall.
2. Estimate the standard deviation of the loading times for semis, vans, and overall.
3. Create a histogram of the overall unloading times. Mark the location of the mean on the histogram and add six additional markers to show the location for one, two and three standard deviations both above and below the mean. Where do the average unloading times for the semis fall? What about the vans?

6.2. EverythingRUs is an extremely diversified company, manufacturing and distributing goods as well as providing a variety of services. The table below shows the mean monthly revenue and mean monthly cost for each sector of the company, along with the percentage each sector occupies in the overall revenue and cost structure. Use this information to estimate the mean monthly revenue and mean monthly cost for the entire company. All revenue and cost figures are in thousands of dollars.

Sector	Mean Monthly Revenue	% of Total Revenue	Mean Monthly Cost	% of Total Cost
Food services	$1,200	15%	$380	22%
Repair services	$2,460	18%	$115	6%
Security services	$875	11%	$219	10%
Health and beauty products	$1,620	14%	$652	17%
Automobile parts	$565	8%	$95	12%
Clothing	$3,218	13%	$1,897	15%
Medical supplies	$1,979	21%	$934	18%

6.3. Match the histograms below with their cumulative distributions shown in figure 6.14. The graphs in the left-hand column (labeled A - D) are histograms. The graphs in the right-hand column (labeled E - H) are cumulative distributions. Each histogram contains the same number of total observations. The cumulative distributions are given by percentage of total, rather than actual count.

6.4. Match the histograms (labeled A - D) below with the boxplot (labeled 1 - 4) in figure 6.15 that best matches the data.

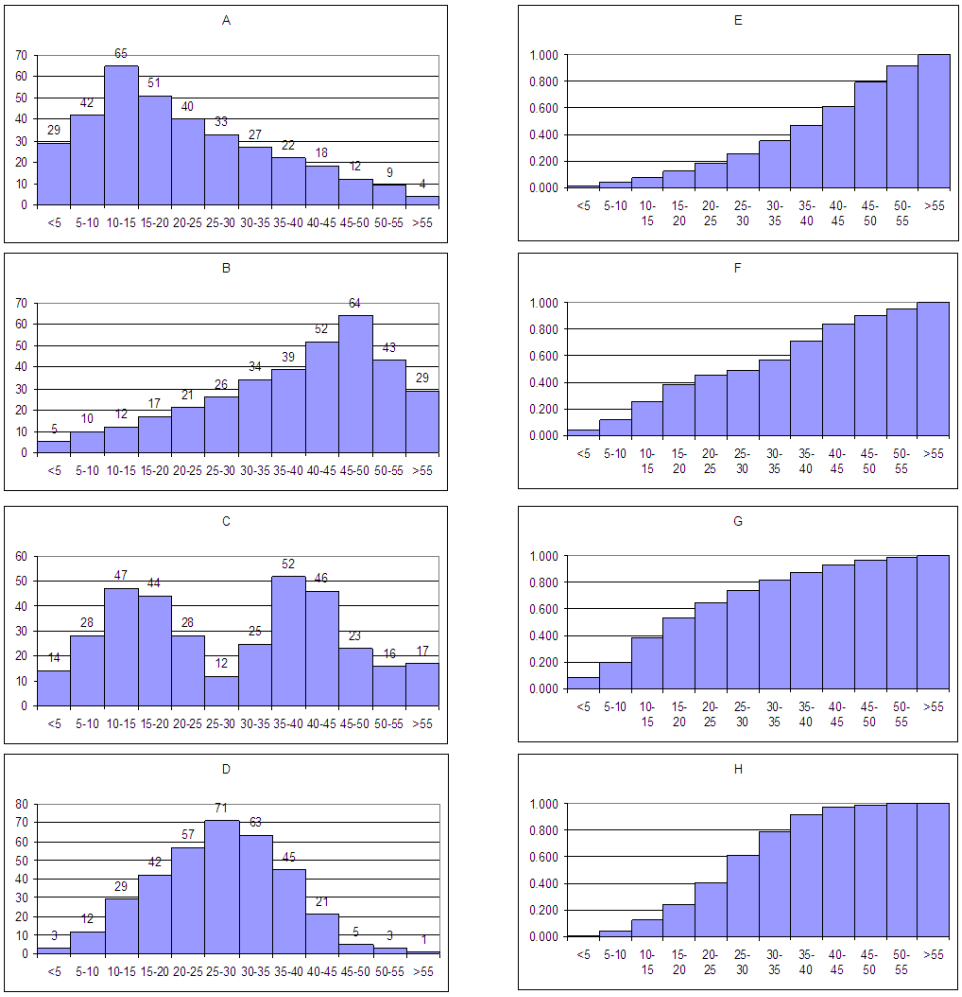

Figure 6.14: Match the histograms (A - D) with the cumulative distributions (E - H) in problem 3.

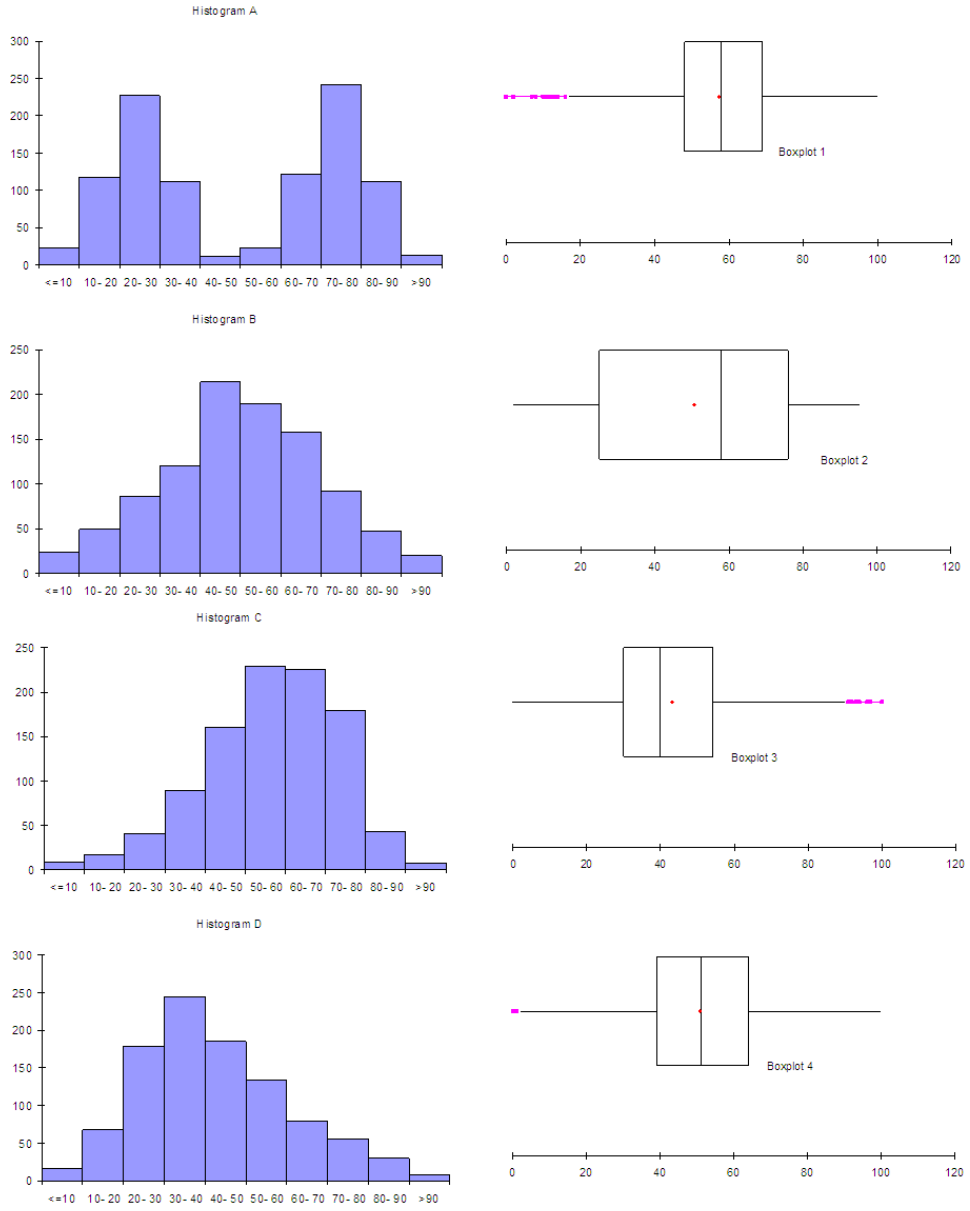

Figure 6.15: Match the histograms (A - D) with the cumulative distributions (E - H) in problem 4.

6.3.2 Application and Reasoning Problems

6.5. Two cumulative distributions are shown in figure 6.16. Describe the differences between the two underlying histograms from which these cumulative distributions were constructed.

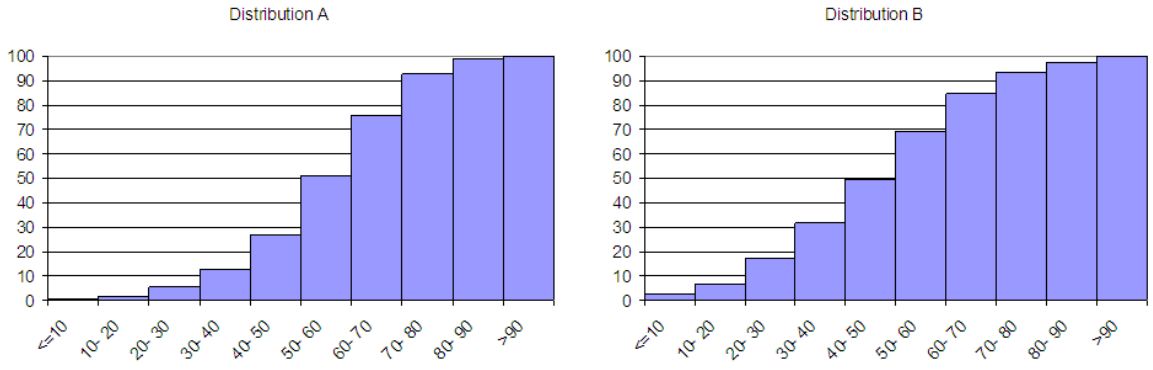

Figure 6.16: Compare these two distributions in problem 5.

6.6. During a recent meeting at your company, the group is examining the breakdown of customers, based on income (see the file "C06 Frequency.xls"). The coworker presenting the data claimed that the company's average customer has a mean income over $80,000. This coworker then proceeded to explain how this would impact the company. But you suspect something is not quite complete in your coworker's analysis, so at a break in the meeting, you modify the data file as shown to help you explore how the estimated mean and standard deviation change with different assumptions about the distribution of the data. At the top of the spreadsheet is a parameter labeled "Mid". This is a number between 0 and 1 (like 0.25) that represents how far from the left (as a percentage of the total bin size) you would like to position the "midpoint" of the bin for estimating the mean and standard deviation. The rest of the data table is set up similarly to the one shown in example 4 (page 170) to estimate the mean and standard deviation.

1. Make your own table with three columns to summarize your exploration of the data. The first column should contain values of the parameter "Mid". The second column should contain the estimated mean for that value of the parameter, and the third column should contain the estimated standard deviation. Use at least the following five values for the parameter: 0, 0.25, 0.50, 0.75, 1.0.

2. After looking at the table you have produced, what will you tell the rest of your coworkers in the meeting when you return from the break?

6.7. The graphs in figure 6.17 show boxplots of employee salaries at four local companies. Based on these boxplots, describe the shape of the histograms of the salaries. Estimate

values for the minimum, the number of categories, and the category length that would help create a decent histogram of the salary data.

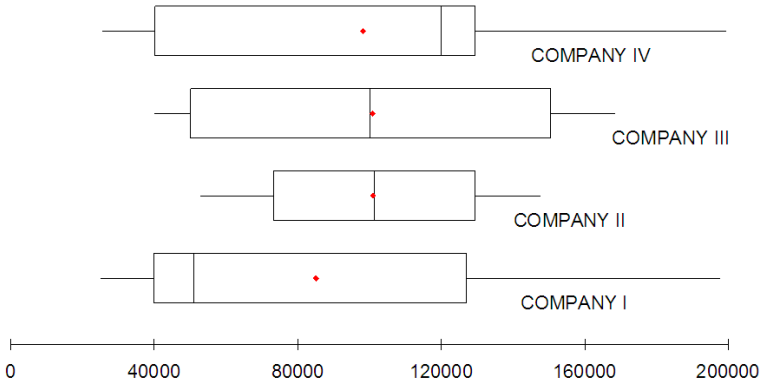

Figure 6.17: Boxplots of salaries at four different companies for problem 7.

6.3.3 Memo Problem

To: Financial Planning Services Department
From: John E. Cash
Date: May 18, 2008
Re: Portfolio development for clients

As you know, one of the many services offered by Oracular Consulting is financial planning. We have recently acquired two new clients. The financial planning department has pre-filtered the current market and provided a list of ten (10) stocks that have been performing well in recent months. The data on these stocks for the last three years is provided in the attached file, presented as frequency data based on the number of days the stock provided a particular daily return.

The two clients are quite different. Client A is young and energetic. She has a long time before retirement, and is willing to take risks in order to gain a lot. Client B is older and has much less time before retirement. He needs a stock portfolio that will provide income in his retirement, so he is not willing to accept a lot of risk, but he of course would like a steady return. Both clients recognize the need to diversify their portfolio in order to plan for the future.

Using the frequency data provided, fill out a chart like the one below to help present the data to the clients in a convenient, easy-to-compare format. Then put together two portfolios, one for each client, composed of 4 stocks, showing the percentage of the investment in each of the stocks in the portfolio. Justify your choices carefully, and provide the clients with both an estimated rate of return for their portfolio and a range of possible returns that they can reasonably expect. To estimate the range of likely returns, simply use the high and low expected values for each stock in the portfolio, weighted by the percentage of the investment in that stock.

	Stock 1	Stock 2	Stock 3
Mean			
Std. deviation			
Minimum			
Q1			
Median			
Q3			
Maximum			

Attachment: Data file "C06 StockPerformance.XLS"