

# Chapter 5

## Histograms<sup>1</sup>

In this chapter, we'll look at how we can use z-scores for each data point to abstract the notion of how spread out the data is. We can also use these to test whether the data appears to come from what is called a "normal distribution" which most randomly generated data should follow. This In part B we will then use z-scores to help us build a good graphical representation of the data with a histogram. This type of graph gives a more detailed picture of the observations of a single variable and helps to classify data into one of several types. This classification then makes it easier to draw conclusions from the data.

- *As a result of this chapter, students will learn*
  - ✓ How z-scores determine a relative ranking of observations
  - ✓ How z-scores allow for comparison of data that is in different units and of different sizes
  - ✓ What normally distributed data is and what the "rules of thumb" are for checking it
  - ✓ The difference between absolute and relative cell references
  - ✓ The characteristics of each of the classic histograms: uniform, symmetric, bi-modal, positively and negatively skewed
  - ✓ How to check the rules of thumb using a histogram of z-scores
  - ✓ The characteristics of good and bad histograms
  
- *As a result of this chapter, students will be able to*
  - ✓ Compute z-scores by hand or with Excel
  - ✓ Explain why the standard deviation formula is set up the way it is
  - ✓ Check the rules of thumb using an array formula in Excel
  - ✓ Stack and unstack data using the Data Utilities in StatPro
  - ✓ Read a histogram
  - ✓ Interpret the information in a histogram
  - ✓ Make a histogram of data either by hand or using StatPro
  - ✓ Improve on a badly made histogram in order to tease more information from it

---

<sup>1</sup>©2011 Kris H. Green and W. Allen Emerson

## 5.1 Getting the Data to Fit a Common Ruler

Well, now we have a tool for measuring the spread of a set of data. This measuring stick, the standard deviation, is a useful tool for looking at the entire set of data. Notice that we are slowly building up more information about the data as a whole: We started with thousands of observations. Then we reduced this down to a single statistic, the mean, which measures the typical data point. Next we added a little more information by using a boxplot, which really contains seven pieces of information (minimum, first quartile, median, mean, third quartile, maximum, outliers). Then we added the standard deviation to our arsenal, giving us quite a bit of information about the typical observations and how the rest of the data is spread out.

However, these tools really only help us look at a single variable in a set of data. The measuring stick for determining the spread of the data is different for every set of data. In essence, the standard deviation is a "ruler made of rubber"; it stretches to measure the spread of data that has a large range, and it contracts to measure the spread of data with a small range. What if we want to compare two sets of data? Better yet, what if we want to compare individual observations from two different variables in our data? How can we do this when all our tools are designed to change to fit the data? Is there no standard?

As a matter of fact, there is a standard ruler that applies to all data, regardless of its size or units. Each observation in a set of data can be converted to what is called a standard score, also known as a z-score. This converts all data to a dimensionless number on a common ruler. Once this is done, we can compare z-scores for observations from different variables and we can determine which observation is farther away from the mean in an absolute sense.

It is important to realize that the concept of a z-score is fundamentally different from the other statistics we have discussed so far. The mean, the standard deviation, and the statistics shown on a boxplot are descriptive statistics for an entire set of data. On the other hand, each observation has its own z-score; thus, z-scores are more individual. At the same time, a z-score is a comparative number. Z-scores show you how a particular observation compares to the entire data set. Essentially, a z-score is a number that tells you how many standard deviations (or fractions of a standard deviation) an observation is from the mean of the data.

### 5.1.1 Definitions and Formulas

**Z-scores or Standard Scores** The z-score for an observation is a dimensionless quantity that tells how many standard deviations the observation is from the mean. To compute the z-score for observation  $i$  of a variable  $x$ , we calculate:

$$z_i = \frac{x_i - \bar{x}}{S_x}$$

Z-scores indicate the signed distance (in standard deviations) between an observation and the mean. For example, a z-score of 0 indicates that the observation is equal to the mean, while a z-score of -1.5 indicates an observation between one and two standard deviations below the mean (because of the negative sign).

**Normally Distributed Data** Statistically speaking, characteristics of a population (such as height, weight, or salary) are what are called normally distributed data. This is data that is symmetrically spread around the mean according to the normal distribution. The normal distribution itself is a product of a complicated-looking formula, but the basic idea is that the data should satisfy certain rules of thumb (see below).

**Rules of Thumb** In normally distributed data, there are approximately 68% of the observations within 1 standard deviation of the mean, 95% of the observations within two standard deviations, and 99.7% within 3 standard deviations of the mean. Thus, most of the data is fairly close to the mean, with equal amounts being above and below. In terms of z-scores, the rules of thumb would say that

Z Scores	Percentage of Observations in that Range
-3 to -2	2.35%
-2 to -1	13.5%
-1 to 0	34%
0 to 1	34%
1 to 2	13.5%
2 to 3	2.35%
Total	99.7%

Thus, very few observations (0.3%) should have z-scores larger than 3 or less than -3 if the data is normally distributed. Keep in mind however, that unless you have a lot of data (several hundred observations) the rules of thumb may not be helpful for determining whether the data came from a normal distribution.

## 5.1.2 Worked Examples

### Example 5.1. Converting Observations to Z-scores

Let's return to the data from example 2 (page 67). Remember that the mean is 6 and the standard deviation (example 3 (page 68)) is approximately 1.852. To calculate the z-scores, we take the observation, subtract the mean, and divide the result by the standard deviation. A few of these are done for you. Fill in the rest of the table on your own.

$$z_1 = \frac{x_1 - \bar{x}}{S_x} = \frac{3 - 6}{1.852} = \frac{-3}{1.852} \approx -1.620 \quad z_{11} = \frac{x_{11} - \bar{x}}{S_x} = \frac{7 - 6}{1.852} = \frac{1}{1.852} \approx 0.540$$

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$x_i$	3	3	4	5	5	5	6	6	6	7	7	8	8	8	9
$x_i - \bar{x}$	-3	-3	-2	-1	-1	-1	0	0	0	1	1	2	2	2	3
$z_i$	-1.6										0.5				

From this, we see that the smallest observations in the data ( $x_1$  and  $x_2$ ) are between 1 and 2 standard deviations below the mean, since the z-scores for these observations are between -1 and -2. What do you predict will happen when you add all the z-scores up? Why? Is this generally true, or special for this set of data?

Is this data from a normal distribution? That's harder to answer. We have so few data points, that we will have a hard time matching the rules of thumb exactly. If you've calculated all the z-scores correctly then you should get 8 out of 15 observations with z-scores from -1 to +1. This accounts for  $8/15 = 53.33\%$  of the data, which is a little short of expectation. If this were a normal distribution, we would expect closer to 68% of  $15 = 0.68 * 15 = 10.2$  observations in this range. We have  $15/15 = 100\%$  of the data with z-scores between -2 and +2, which is slightly higher than the 95% expectation from the rules of thumb, but 95% of  $15 = 0.95 * 15 = 14.25$ , so we are close to the right number of observations. Overall, this data does not appear to be from a normal distribution. However, to really see whether it is from a normal distribution, we need to apply some other mathematical tools. It's entirely possible that this data really is normally distributed. We simply can't tell with the tools we currently have, but we can make a guess that the data is not normally distributed: Specifically, there are not enough observations near the mean to make it normal.

### Example 5.2. Converting from Z-scores to back to the Data

Notice that z-scores, regardless of the units of the original data, have no units themselves. This is because units cancel out. Suppose we have data measured in dollars. Thus, the units for  $x_i$ ,  $\bar{x}$ , the deviations from the mean, and  $\sigma_x$  are all in dollars. When we compute the z-scores, the units vanish:

$$z_i = \frac{\text{deviation}}{\text{standard deviation}} = \frac{\text{dollars}}{\text{dollars}} = \text{no units}$$

This is the power of the z-scores: they are dimensionless, so they can be used to compare data with completely different units and sizes. Everything is placed onto a standard measuring stick that is "one standard deviation long" no matter how big the standard deviation is for a given set of data.

But suppose all we know is that a particular observation has a z-score of 1.2. What is the original data value? We know it is 1.2 standard deviations above the mean, so if we take the mean and add 1.2 standard deviations, we'll have the original observation. So, in order to answer this question about a specific observation we need to know two things about the entire set of data: the mean and the standard deviation. So, if the data represents scores on a test in one class, and the class earned a mean score of 55 points with a standard deviation of 8 points, then a student with a standard score of 1.2 would have a real score of 55 points +  $1.2(8 \text{ points}) = 55 \text{ points} + 12 \text{ points} = 67 \text{ points}$ . If a student in another class had a z-score of 1.5, then we know that the second student did better *compared to his/her class* than the first student did, because the z-score is higher. Even if the second student is in a class with a lower mean and lower standard deviation, the second student performed better relative to his/her classmates than the first student did.

On standardized tests, your score is usually reported as a type of z-score, rather than a raw score. All you really know is where your score sits relative to the mean and spread of the entire set of test-takers.

### Example 5.3. Checking Rules of Thumb for the Sales Data

Does our sales data from the Cool Toys for Tots chain follow the rules of thumb for normally distributed data? (See the new version of the data in "C05 Tots.xls".) We can ask this

question from two different perspectives: by regional comparison and by comparison across the entire chain. Let's just look at the chain as a whole and include both the northeast and north central regions. First, we'll compute the z-scores for the stores in the chain in order to convert all the data to a common ruler. For this, we'll need the standard deviation and mean of the chain, rather than the individual means and standard deviations for each region. The final result is shown in the table below:

Store	Region	Sales	Sales Z
1	NC	\$668,694.31	2.0100
2	NC	\$515,539.13	1.1483
3	NC	\$313,879.39	0.0138
4	NC	<b>\$345,156.13</b>	0.1898
5	NC	\$245,182.96	-0.3727
6	NC	\$273,000.00	-0.2162
7	NC	\$135,000.00	-0.9926
8	NC	\$222,973.44	-0.4977
9	NC	\$161,632.85	-0.8428
10	NC	\$373,742.75	0.3506
11	NC	\$171,235.07	-0.7887
12	NC	\$215,000.00	-0.5425
13	NC	\$276,659.53	-0.1956
14	NC	\$302,689.11	-0.0492
15	NC	\$244,067.77	-0.3790
16	NC	\$193,000.00	-0.6663
17	NC	\$141,903.82	-0.9538
18	NC	\$393,047.98	0.4592
19	NC	\$507,595.76	1.1036
20	NE	\$95,643.20	-1.2140
21	NE	\$80,000.00	-1.3020
22	NE	\$543,779.27	1.3072
23	NE	\$499,883.07	1.0603
24	NE	\$173,461.46	-0.7762
25	NE	\$581,738.16	1.5208
26	NE	\$189,368.10	-0.6867
27	NE	\$485,344.87	0.9785
28	NE	\$122,256.49	-1.0643
29	NE	<b>\$370,026.87</b>	0.3297
30	NE	\$140,251.25	-0.9630
31	NE	\$314,737.79	0.0186
32	NE	\$134,896.35	-0.9932
33	NE	\$438,995.30	0.7177
34	NE	\$211,211.90	-0.5638
35	NE	\$818,405.93	2.8523

To check the rules of thumb, we need to determine how many stores fall into each of the breakdowns by using the z-scores.

- Between 0 and 1 standard deviation:

There are 25 stores out of 35. This is  $25/35 = 0.7143 = 71.43\%$ . This value is a little higher than the rule of thumb suggests, but not by much.

- Between 0 and 2 standard deviations:

There are 33 stores in this group, giving  $33/35 = 0.9429 = 94.29\%$ . This is very close to the rule of thumb.

- Between 0 and 3 standard deviations:

There are 35 stores, giving  $35/35 = 100\%$  of the stores in this range. This is slightly higher than the rule of thumb suggests.

Overall, this data is fairly close to satisfying the rules of thumb for being normally distributed. There seems to be one too many observations within one standard deviation of the mean, but that is generally acceptable. (Note:  $71.43\% - 68\% = 0.0343\%$  and  $0.0343\%$  of  $35 = 1.2$ .) With only 35 data points, these results are very close to what one would expect from normally distributed data. Is the data symmetric? If it is, there should be roughly the same number of stores above the mean as there are below the mean.

### 5.1.3 Exploration 5A: Cool Toys for Tots

Now we have all the tools we need to look at the various stores in the two sales regions of Cool Toys for Tots (examples 4 (page 70) and 3 (page 106)) in complete detail. Which individual stores in our chain are performing the best, relative to their regions? Which stores are performing worst in their regions? Which are performing best and worst overall? To analyze these questions, try computing z-scores for each store in two different ways: relative to each region and relative to the entire chain of stores. You can also analyze the data with and without any outliers.

The data file "C05 Tots.xls" contains a column of identifiers (store number), a categorical variable identifying the region of each store, and the sales figures for the stores (numerical variable).

Here are some questions to guide you in this exploration:

1. How can you compute the z-scores for each store?
2. How do you compute z-scores for each store relative to only the stores in its region? (Try looking at the How to Guide for information on "How to Stack and Unstack Data".)
3. How do you identify an outlier? Is it an outlier for its region or for the entire chain of stores?
4. Are there any stores whose relative performance in their region is not reflected when it is compared to the entire chain or vice versa?
5. Are these stores performing poorly with respect to (1) the entire chain or (2) their individual regions?

### 5.1.4 How To Guide

The following comments are based on the data file "C05 Homes.xls" which has data in cells A3:M278. The data records information about homes that sold in a three-month period in Rochester, NY. The variables are address (identifier), location (categorical), taxes (numerical), style (categorical), bath (numerical), bed (numerical), rooms (numerical), garage (numerical), year (numerical), acres (numerical), size (numerical), value (numerical), price (numerical).

#### Computing z-scores

To compute z-scores for the variable Price (cells M3:M278), we first need to compute the mean and standard deviation. In cell P1 enter =AVERAGE(M3:M278) to compute the average and in cell P2 enter =STDEV(M3:M278) to get the standard deviation. Now, in column N, enter "Z Score" in N3 and enter the formula below into N4

=(M4 - \$P\$1)/\$P\$2

All that is left is to copy the formula to the rest of column N (N5:N278).

#### Naming Cell Ranges

There is another way to refer to cell ranges (or individual cells) besides a cell reference. You can give the cells or cell ranges their own names and then use these names in formulas for computations. Figure 5.1 shows the C05 Homes.xls data with the data in the Price variable (column M) selected. To give this range of cells a name, we simply click on the "Name Box" to the left of the formula bar and type a name; in this case, we'll call the range of cells "Price". Note: there are no spaces allowed in the name box.

Now, in any formula in the worksheet we can use "Price" instead of the range M4:M278. Instead of typing the formula =AVERAGE(M4:M278), you could just type the formula =AVERAGE(Price). Notice that such references are always absolute. This has the benefit of making all the formulas more readable. You can see a list of all the named ranges in the workbook by clicking on the downward pointing triangle next to the name box. Notice that if you use any of StatPro's commands to manipulate your work, there will be a lot of ranges named for you, because that's what StatPro does first in order to simplify its formulas. If you want to get rid of these, StatPro has a command for this: go to "StatPro/Clean Up..." and select "Delete range names". StatPro will then wipe out all of these range names. This may free up some space and make the Excel file smaller.

#### How to Stack and Unstack Data Using StatPro

This is a handy feature if you have a lot of data with several categorical variables. You may want to analyze each group of data separately. For example, you may want to compute the average salary of males and females separately in order to compare them. If the data is not sorted to make this easy, the best approach is to unstack the data. This is a standard StatPro procedure, so the six steps discussed in chapter 3 apply:

1. Select the region of the worksheet that contains the data. This step is the same as it is described in chapter 3 for "Summary Statistics in StatPro".

The screenshot shows a Microsoft Excel spreadsheet titled "C05 Homes.xls [Compatibility Mode] - Microsoft Excel". The spreadsheet contains housing data for 275 homes. The columns are labeled: Address, Location, Taxes, Style, Bath, Bed, Rooms, Garage, Year, Acres, Size, Value, and Price. The 'Price' column is highlighted in yellow, and the name box on the left of the formula bar shows "Price".

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Housing data for 275 homes that sold in Rochester, NY in 2000																				
2																					
3		Address	Location	Taxes	Style	Bath	Bed	Rooms	Garage	Year	Acres	Size	Value	Price							
4	30	Walnut Park	14622	2936	2	2.0	4	8	1	1928	0.15	1952	85.0	97.0							
5	249	Cabot Rd	14626	2350	3	1.0	4	7	0	1942	0.11	1397	72.0	69.0							
6	199	Rosemont Drive	14617	3338	1	1.0	3	5	2	1955	0.021	1316	57.0	84.0							
7	471	Baker Rd	14580	5258	1	3.0	2	6	2	1983	1.83	2485	212.0	140.0							
8	449	Glendonan Terr	14580	5474	2	2.0	4	8	2	2001	0.41	2599	182.0	195.0							
9	1638	Glendonan Terr	14580	6117	2	2.0	4	8	2	2001	0.51	2595	217.0	213.0							
10	898	Ridge Rd Lot 28	14580	*	1	2.0	2	5	0	1985	0	1240	*	38.0							
11	1506	Chigwell Lane	14580	5950	2	2.0	5	9	2	2001	0.41	2635	196.0	299.0							
12	13	Countryside Rd	14450	5820	4	2.0	5	10	2	1960	1.1	3188	218.0	173.0							
13	17	Mattock Place	14534	5000	2	2.0	4	8	2	1980	0.41	2388	178.0	167.0							
14	266	Behvista Dr	14625	5410	2	2.0	5	9	2	1968	0.48	2252	168.0	140.0							
15	4	Beauclaire Mint	14450	12113	2	3.0	4	9	2	1992	1.24	3747	438.0	470.0							
16	18	Columbia Court	14450	5216	3	2.0	3	6	2	2000	0.26	2099	182.0	226.0							
17	201	Hazelwood Terr	14609	2100	2	2.0	4	8	1	1910	0.09	1729	51.0	50.0							
18	191	Trevor Court	14610	14117	5	3.0	4	9	2	1930	0.45	3775	385.0	440.0							
19	388	Cascade Place	14609	2546	1	1.0	3	6	2	1950	0.22	1176	71.0	66.0							
20	483	Basket Rd	14580	2416	1	1.0	3	5	2	1983	0.47	1056	83.0	64.0							
21	41	Brookview Dr	14617	3880	1	1.0	2	7	2	1954	0.52	1386	87.0	89.0							
22	135	Woodrow Avenue	14609	3056	2	1.0	3	6	2	1955	0.23	1350	80.0	83.0							
23	1874	Jackson Road	14526	3607	4	1.0	3	8	2	1959	0.34	1771	108.0	122.0							
24	55	Trevor Court	14610	10800	2	3.0	4	9	2	1934	0.3	2882	296.0	460.0							
25	65	Kenwick Drive	14623	2263	1	1.0	3	8	1	1965	0.22	2400	69.0	51.0							
26	252	Somershire Drive	14617	3729	2	1.0	3	6	1	1920	0.12	1614	91.0	105.0							
27	196	Bersford Road	14610	2904	5	1.0	2	5	1	1927	0.1	1523	88.0	100.0							
28	27	Cave Hollow Drive	14586	2219	4	1.0	3	6	1	1991	0.31	1600	73.0	100.0							
29	99	Windelin Drive	14467	3481	2	2.0	3	7	2	1993	0.38	1810	111.0	144.0							
30	54	Woodleaf	14534	4423	1	2.0	3	6	1	1987	0.46	1608	142.0	160.0							
31	78	Thistledown Drive	14617	3387	2	1.0	3	6	2	1947	0.16	1382	87.0	97.0							

Figure 5.1: Homes data showing the price variable selected and being assigned the name "Price" using the name box to the left of the formula bar, below the ribbon.

2. Select the StatPro routine to apply to the data. The routine for this is under "Data Utilities..." then select "Unstack Variables..."
3. Verify that the data region is correct. This does not happen for this routine.
4. Select the variables to which the routine will apply. First we select the code variable. This should be a categorical variable with a small number of categories. For example "Gender" might be a good choice, since there are only two options, Male and Female.
5. Fill in the details of the routine. Next select all the variables that you want unstacked. In this example, if you only wanted the salary data for males and females, just select salary. The routine will create new variables called "Salary male" and "Salary female". If you want several variables unstacked, the routine will create a new variable called "Old Variable Category #1" and then "Old variable Category #2", etc.
6. Select the placement for the output of the routine. Select whether you want the results (the new variables) to be placed in a cell next to the data, on a new worksheet, or in a particular cell.

## 5.2 Profiling Your Data

One of the most important questions a detective asks about those involved in a criminal investigation is "What did the suspect look like?" Without a physical description, detectives will have difficulty finding the suspect. Likewise, they ask about the suspect's habits and personality. Eventually, they build a profile of the suspect. Such profiles describe the suspect physically and psychologically. They are based on statistical analyses of criminals and are extremely helpful in locating the suspect before more crimes can be committed. In order for you to study data from a business setting, you will also need to develop a profile of the data. We have begun this in chapter three with a discussion of central tendency. In chapter four we described the way the data is spread out using various tools. Along the way we got a blurry picture of the data, the boxplot. Now it's time to sharpen the picture and get more detail. The best tool for this is called a histogram. It will help answer the question "What does your data look like?"

A histogram is basically a graph that breaks the observations of a single variable into intervals called bins. By counting the number of observations in each bin we can generate a frequency table of the data which is then turned into a type of bar chart, with one bar for each bin and the height of each bar indicating the number of observations it contains. Usually histograms have eight to twelve bins. This means that we get a more detailed picture of the data than from a boxplot. With each step, we get more information about the data to help us make decisions.

Representation	What it is	What it tells you
Raw Data	Many observations, lots of information	Hard to make sense out of
↓		
Averages	Single number (mean, median, or mode)	Tells what is "typical"
↓		
Boxplot	Seven pieces of information (min, Q1, median, mean, Q3, max, outliers)	Shows where the data is bunched together
↓		
Histogram	Ten to fourteen pieces of information usually (min, bin width, frequencies)	More detailed profile of the data

Most histograms can be classified into one of five types: uniform, symmetric, bimodal, positively skewed, or negatively skewed. Each type has certain characteristics that make it easy to recognize. Being able to classify the data as one of these types helps you analyze the data in much the same way that a good profile of a suspect tells the detectives a lot about how to catch him or her. In this section, you will learn to recognize each of these classic histograms and will learn what each one tells you about the data. As you learn how to make,

read, and interpret histograms, keep in mind that real data will never exactly look like any of the "perfect examples". Many times you will be required to make a judgment call as to which type of distribution the data fits.

Another important detail about histograms to remember: depending on what bins you use to make the histogram, the data may look different. It's a good idea to look at the data in several ways before drawing any conclusions.

### 5.2.1 Definitions and Formulas

**Frequency Table** Sometimes it may be useful to group the data together into subgroups (called bins, see below). To do this, you simply count how many observations fall into each bin. This count is called a frequency. When you have all of the observations placed into bins, the entire list of bins and frequencies is a frequency table for the data.

**Bins** A bin is one of the "boxes" in which data are placed to make a frequency table. Typically bins are all the same size or cover the same number of categories. For example: Ages of people could be divided into bins like 10-19, 20-29, 30-39, etc. You could also divide the ages into 0-19, 20-39, 40-59, etc. Each of these intervals is a bin into which observations are placed. Think of this as making a bunch of boxes, each labeled with a range of values. If an observation falls inside that range, place a counter into the box. When you have finished doing this for all the observations, you will have a frequency count for the data.

**Distribution** In the sense that we are referring to it in this text, distribution refers to the way the data is spread out or bunched together.

**Histogram** A histogram is a graphical representation of a frequency table. It shows the bins along the horizontal axis and has bars above each bin. The height of each bar represents the number of observations that fall in that bin. Histograms can be made using StatPro or by creating a frequency table and generating a bar graph.

**Skewness** Skewness measures how far the distribution of data is from being symmetric. The actual formula for skewness uses the z-scores of the data and is a little ugly:

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n z_i^3$$

compares the data to the mean. If most of the data is less than the mean, then the skewness will be negative. If most of the data is greater than the mean, then the skewness is positive. The reason for this behavior is the exponent of three: data points far from the mean (and thus having a large deviation and a large z-score) will affect the total more than points close to the mean. In a positively skewed data set, the smallest values are much closer to the mean than the largest values, so the large positive deviations are made even larger by cubing them. The opposite happens for negatively skewed data.

**Uniform Distribution** A uniform distribution (figure 5.2) has roughly the same number of observations in each bin. It looks almost flat, with each bin having almost the same height:

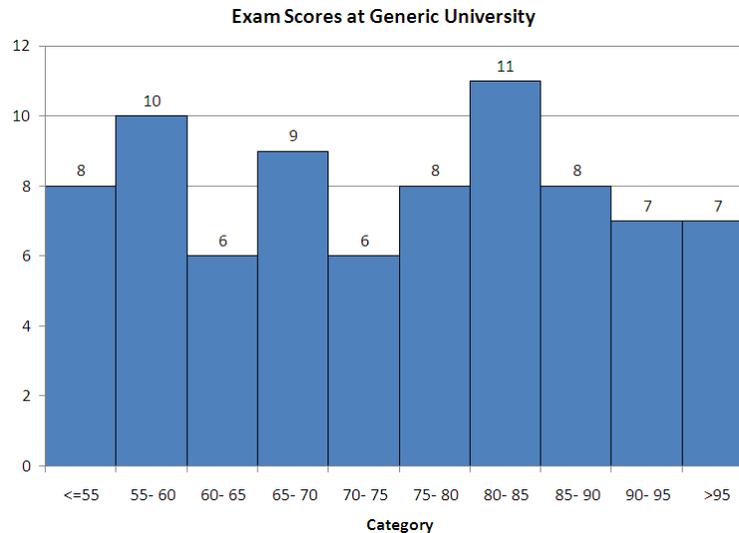


Figure 5.2: A histogram of uniform data.

**Symmetric Distribution** A symmetric distribution (figure 5.3) has equal amounts of data on each side of a central bin. As you move farther from the central bin in either direction, the same number of observations (approximately) can be found.

**Positively Skewed Distribution** A positively skewed distribution (figure 5.3) has more data on the left side of the mean. Typically, the skewness of such distributions is positive, and the median is less than the mean. The "tail" of the distribution points toward increasing values on the horizontal axis.

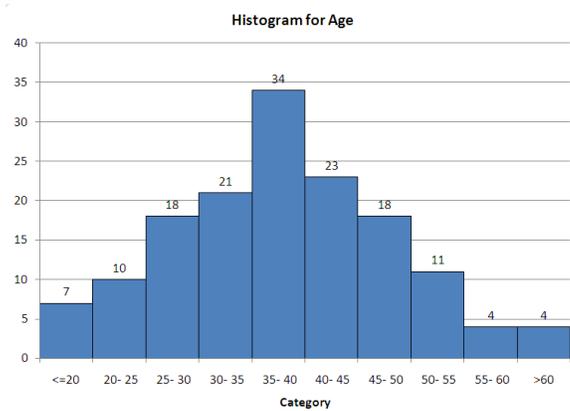
**Negatively Skewed Distribution** A negatively skewed distribution (figure 5.3) has more data on the right side of the mean. Typically, the skewness of such distributions is negative, and the median is more than the mean. The "tail" of the distribution points toward decreasing values on the horizontal axis.

**Bimodal Distribution** A bimodal distribution (figure 5.3) has two major peaks in it (there are two modes to the data, hence the term bi-modal). There is usually a gap between the peaks with fewer observations.

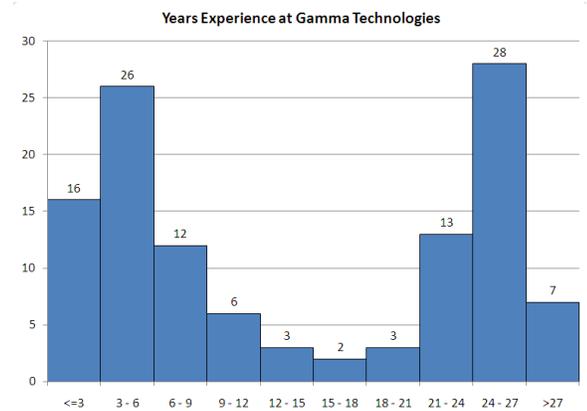
## 5.2.2 Worked Examples

### Example 5.4. Reading a Histogram

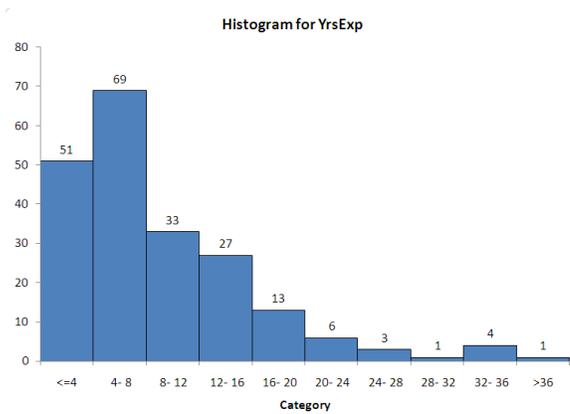
Look at the data shown in the histogram below. What can we learn about the data? First



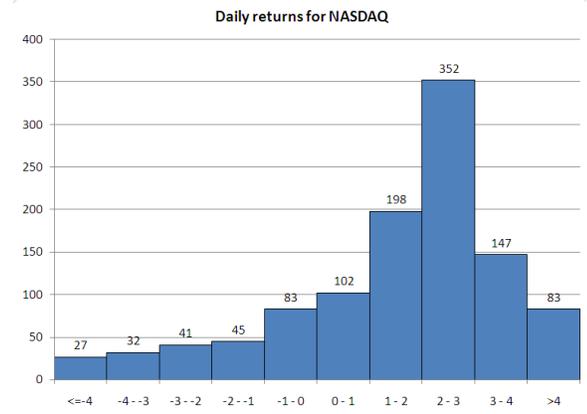
A histogram of symmetric data.



A histogram of bimodal data.



A histogram of positively, or right-skewed, data.



A histogram of negatively, or left-skewed, data.

Figure 5.3: Illustrations of the major types of distributions of data.

of all, notice the bins; they are labeled below the bars on the graph. Each bin is the same width as the others, with the exception of the two ends. These are open-ended so they can catch all the observations that are outside the main bulk of the data. When you make a histogram in StatPro, you control three things about the graph: the minimum value (which is really the upper end of the first bin), the number of bins (which includes the two ends!) and the width of each bin. Consider the histogram shown in figure 5.4, which is an example of a positively skewed histogram.

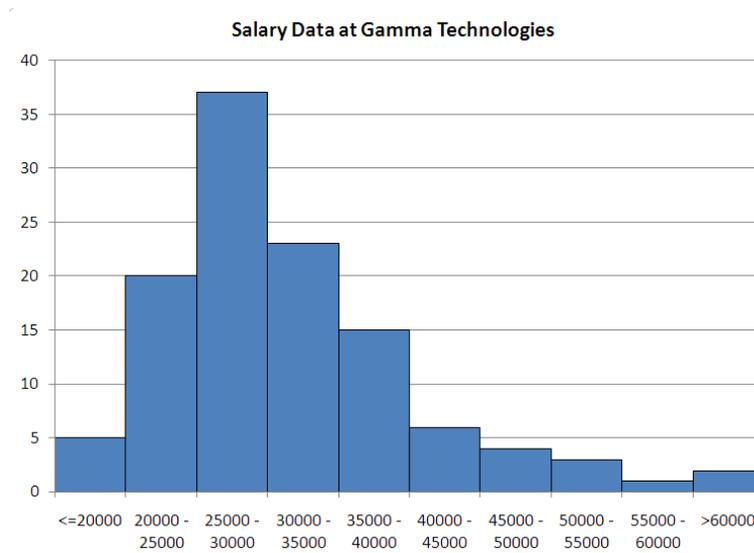


Figure 5.4: Salary distribution at Gamma Technologies, showing a distinct positive skewness.

It seems that whoever created this graph used the following settings to make the graph:

Minimum	20,000
Number of Categories	10
Category Width	5,000

Notice that even though the minimum is set at 20,000, the first bin contains all the observations less than this number. What else can we see? First off, we notice that there are a total of 116 observations in this data. The bin with the most observations is the 25,000 to 30,000 bin, containing  $37/116 = 31.90\%$  of the observations. Most of the observations fall in the bins between 20,000 and 35,000, with a total of  $20 + 37 + 23 = 80$  or about 68.97%. Because the data has such a long tail in the direction of increasing salaries, this graph is said to be positively skewed. This means that the mean is larger than (to the right of) the bulk of the data. Think of it like a teeter-totter. Try to find the point along the axis where the histogram would balance. Start in the 25,000 -30,000 bin. The bins on either side of it are roughly equal, so don't move the balance point. Now the next bins out ( $\leq 20,000$  and 35,000 - 40,000) are very unequal, with more of the weight on the right. This pulls the mean to the right of our starting point. All of the other bins are even more unbalanced, so the mean is pulled far to the right of the highest peak. All of this tells us that the measures

of "typical" for this data are a little skewed. If we report the mean, which is approximately \$34,000, then we are ignoring the skewness of the data. More than half of the data, 85 out of 116 points, is less than the mean, while only 31 points are greater than the mean. So, in what way is the mean a measure of "typical" for this data? On the other hand, the median is slightly less than \$30,000, with exactly half of the data above and below it. In the next section, chapter 6B, we'll look more closely at how to read off statistics from a histogram. As it turns out, almost all of the observations in the last three bins are outliers.

**Example 5.5. Using histograms to check rules of thumb**

By making a histogram of z-scores, we can check to see if the data is normally distributed. First, compute the mean and standard deviation of the data. Then create a column of z-scores for the data. Now, when you make the histogram, we make it with the following options in StatPro:

Minimum	-3
Number of Categories	8
Category Width	1

This will ensure that you have the following bins for the z-scores:

First Bin:	$\leq (-3)$
	-3 to -2
	-2 to -1
	-1 to 0
	0 to 1
	1 to 2
	2 to 3
Last bin:	$> 3$ .

The graph in figure 5.5 shows such a histogram for data taken from the stock market (daily returns of a particular stock for about two years).

Notice that since the mean has a z-score of zero, the mean of the data will always fall in the middle of a z-score histogram, between the fourth and fifth bins. Each bin is one standard deviation wide, so we can now compare the frequency counts of the data to the expected frequency counts from a normal distribution (see chapter 3B).

First, is the distribution symmetric? This graph is pretty close to being symmetric. The two central bins are close in height, as are the bins on either side of the central peak. The bin marking 2 to 3 standard deviations below the mean and the bin marking 2 to 3 standard deviations above the mean are about equal in height. The only parts that don't match are the ends. There should be very little data in these two bins anyway (less than 0.3%), and both are close to zero.

Second, does the rule of thumb hold for this data? Let's check it out.

- Within one standard deviation of the mean, there are about  $194 + 217 = 411$  observations of the variable return. This is  $419/552 = 74.46\%$  of the data. This is a little high when compared to the 68% rule of thumb.

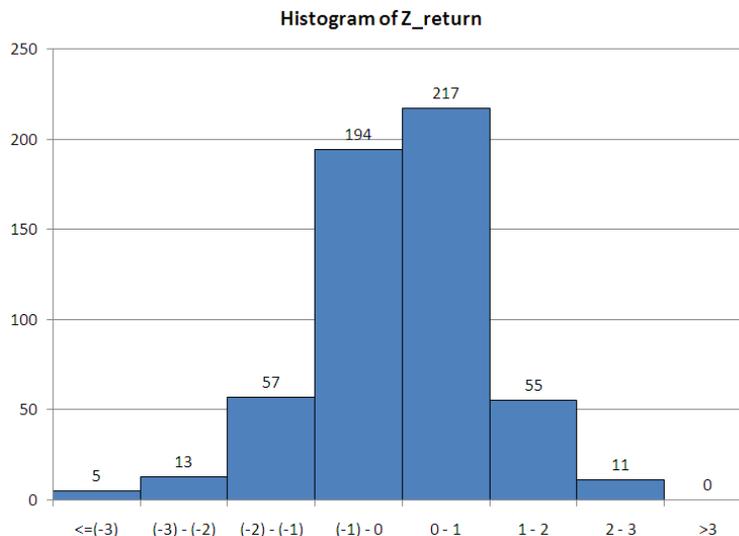


Figure 5.5: Histogram of z-scores for daily returns of a particular stock over a two-year period.

- Within two standard deviations of the mean, there are about  $57 + 194 + 217 + 55 = 523$  observations of the variable return. This accounts for  $523/552 = 94.75\%$  of the data. This is fairly close to the 95% rule of thumb.
- Within three standard deviations of the mean, there are about  $13 + 57 + 194 + 217 + 55 + 11 = 547$  observations of return. This accounts for  $547/552 = 99.09\%$  of the data. This is very close to the 99.7% rule of thumb.

Overall, this data is close to the rule of thumb, but seems to have too much data within one standard deviation of the mean, so we would probably conclude that this data is not from a normal distribution.

Now the question you've been waiting for: Why do we bother checking if the data is from a normal distribution? The main reason is related to the statement we made about normal distributions in chapter 3 (page 61): "Statistically speaking, characteristics of a population (such as height, weight, or salary) are what are called normally distributed data." Suppose that we conducted a customer satisfaction survey. Part of the survey would likely contain demographic data on the ages, incomes, and so forth of the customers. This is to ensure that the conclusions we draw about their satisfaction are accurate conclusions. We would expect that, if we properly sampled the customers, the demographic data would be normally distributed around some mean with some standard deviation. If this is not the case, then we are getting too much satisfaction data from some group or groups. This could completely invalidate our conclusions. A famous case of this type of mistake occurred with Literary Digest in 1936. The magazine surveyed its viewers about the upcoming presidential election. The survey overwhelmingly favored the Republican candidate, who lost by a landslide in the election. The magazine failed to consider that their audience was not a good sample of the entire U.S. population: their readers were mostly high-income families in an

era of economic depression where most families could not afford the money to subscribe to a literary magazine.

### **Example 5.6. The Good, the Bad, and the Ugly**

There are four ways for a histogram to "go bad". By this, we mean that the histogram does not tell you everything it could tell you about the data. Each of these four cases is described below. When making your histograms, if you see graphs like those below, you should try to correct the problem. There are three numbers you can manipulate when building histograms: the starting point (called the minimum value in StatPro), the number of bins (called the number of categories in StatPro) and the width of each bin (called the category length in StatPro). To fix the histogram, simply change one or more these values to adjust the shape. Try several combinations out before you settle on a particular graph. Use each graph you make to help choose better values for the next graph.

The five graphs in figure 5.6 (page 150) illustrate the ways that histogram-makers often "go wrong". The data for each graph is the same, and it represents the amount of household debt accumulated by various households in a recent survey. Notice how each graph makes the data look very different, possibly leading to misunderstandings about the underlying data.

#### Case 1. Too much data in the end bins

This is a classic problem, typically caused by miscalculating where the two ends of the distribution are. The reason it is such a problem is that the end bins are usually "open-ended". This means that they do not cover a specific interval. Instead, they are usually labeled with " $\geq$  some number" for the left end and " $\leq$  some number" for the right end. If too much data is in either of these bins, then you cannot really describe the distribution, because you do not know how far the data extends.

#### Case 2. Bins are too wide (lumpy)

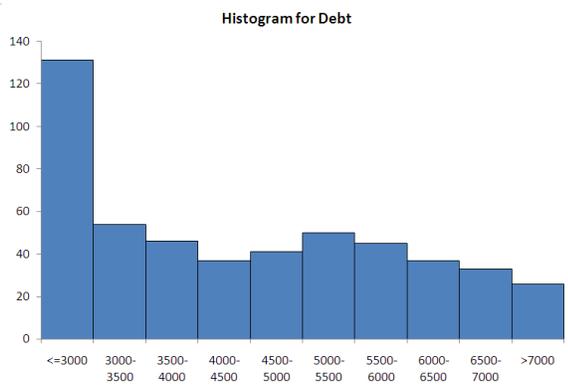
This problem is usually caused by making the bins too wide. This means that each bin covers a large range of observations and means that you can fit fewer bins into the range of the data. To see how many bins of a particular size will fit into a range of data, take the range of the data (maximum value minus the minimum value) and divide by the width of the bins. For example, if the data goes from 0 to 100, and you make each bin 25 units wide, then you will have all of the data in  $(100-0)/25 = 100/25 = 4$  bins! This will not provide you with much information.

#### Case 3. Too few observations in each bin (compared to total number of observations)

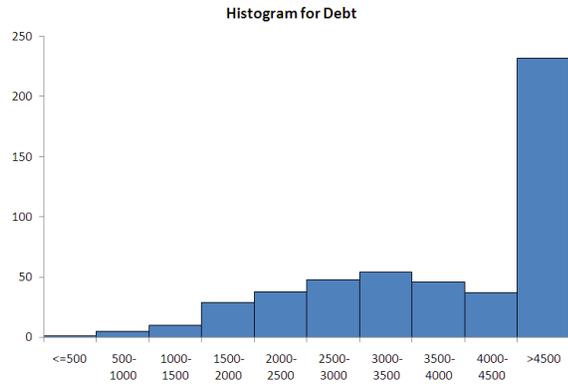
This problem is the opposite of case 2. This is caused by having so many bins that only one or two observations fall into each range. If you have 100 observations that range from 0 to 100, then it does not make sense to use 100 bins that are 1 unit wide. This isn't any better than the original data, since each bin will have, on average, only one observation. It's usually best to stick to eight to twelve bins for a histogram, unless there are many observations and the data needs more bins in order to see the detail. Typically, this is only needed with bimodal data.

#### Case 4. There are empty bins on the ends

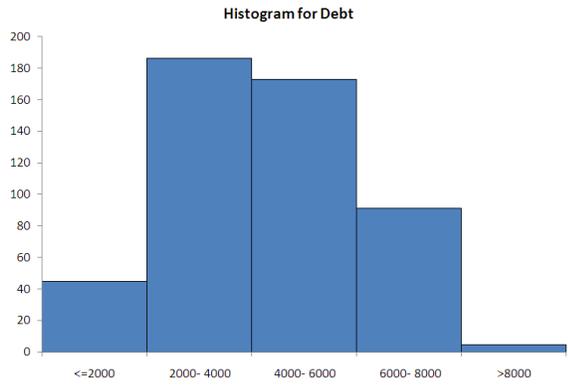
This problem is also caused by miscalculating the locations of the ends of the data. While it is not really a problem, this situation does lead to wasted space in your graph. In addition, the empty bins could be used to mislead someone about the true distribution of the data.



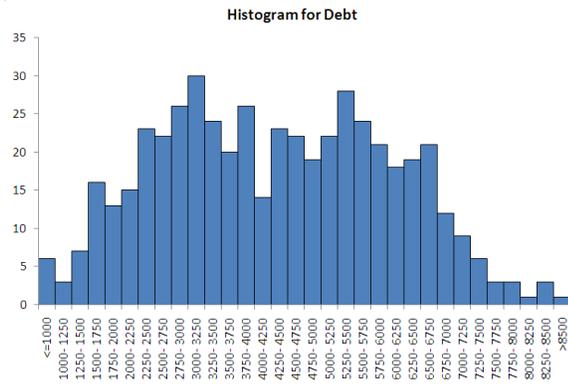
Case 1a: Too much data in first bin.



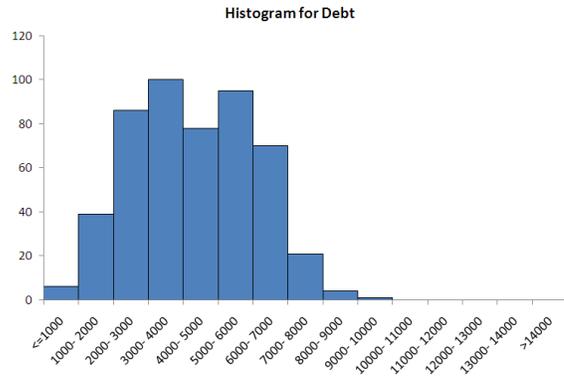
Case 1b: Too much data in last bin.



Case 2: Too lumpy.



Case 3: Too spread out.



Case 4: Too much wasted space

Figure 5.6: Illustrations of the major mistakes in displaying data with a histogram.

### 5.2.3 Exploration 5B: Beef n' Buns Service Times

The manager of a local fast-food restaurant is interested in improving the service provided to customers who use the restaurant's drive-up window. As a first step in this process, the manager asks his assistant to record the time (in minutes) it takes to serve 200 different customers at the final window in the facility's drive-up system. The given 200 customer service times are all observed during the busiest hour of the day for this fast-food operation. The data are in the file "C05 BeefNBuns.xls". Are shorter or longer service times more likely in this case?

**STUDENT analysis:** A student produces the graph shown and then states: "As the graph below shows, most of the service times are on the higher end of the graph, so we expect that there will be many customer complaints."

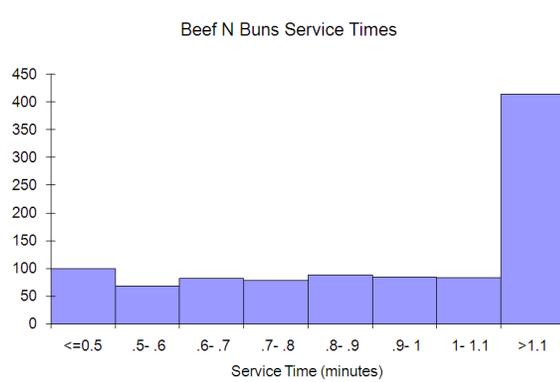


Figure 5.7: Sample histogram of service times at Beef n' Buns.

1. **OBSERVATIONS:** What does this graph tell you about the situation? How many service times were long? How many were short?
2. **INFERENCES:** What do we mean by "long service times" or "short service times"? What is wrong with the graph above? What doesn't it show?
3. **QUESTIONS:** What information that you need is left unsaid by the graph? What questions about the data do you have that a more accurate representation of the data might help you answer?
4. **HYPOTHESIS:** What will happen to the graph if we change  $X$ ?
5. **CONCLUSION:** Make an accurate sketch of what the new graph will look like.

### 5.2.4 How To Guide

The following guide uses the file C05 FamilyIncome.xls which contains the following variables, observed for many different families in a particular city: Family Size, Location (where in the city they live), Ownership (whether they own their home), First income, Second income, Monthly payment, Utilities, and Debt. This data occupies cells A1:I503 of the worksheet "Data".

#### Histograms in StatPro

Making histograms in StatPro is easy. The process follows the same basic structure as for all StatPro procedures.

1. Select a cell in the region of the worksheet that contains the data.
2. Select the StatPro routine to apply to the data. In this case, you should select "Charts/Histograms".
3. Verify that the data region is correct.
4. Select the variables to which the routine will apply. Choose any of the variables in the list. Notice that only numerical variables and categorical variables that are coded as numbers (like Likert data) are available. In this example, we have selected the Debt variable.
5. Fill in the details of the routine. For this routine, you will see a dialog box like the one below. It includes some information about the variable you selected, and asks for you to fill in three pieces of information: minimum, number of categories, and category width. We have selected the values shown in the figure; in reality, there is an art to selecting these. Usually you will need to try several different combinations to get a reasonably good representation of the data.
6. Select the placement for the output of the routine. As with all charts made through StatPro, you have no choice on this one. StatPro will automatically place the chart on a new worksheet called "Hist - (your variable name)". Notice that another worksheet is also created during this process. StatPro generates a frequency table of the data on a worksheet called "Hist - (your variable) Data". What StatPro does is to create a frequency table of the data using your settings for minimum, number of categories and category width, and then it creates a bar chart from this. By changing some of the features of the bar chart, the final graph looks like a histogram.

Notice that the first category in the histogram is actually a category less than what you enter as the minimum value in the dialog box in figure 5.8. So you need to think about this when choosing values for the minimum, number of categories and category length.

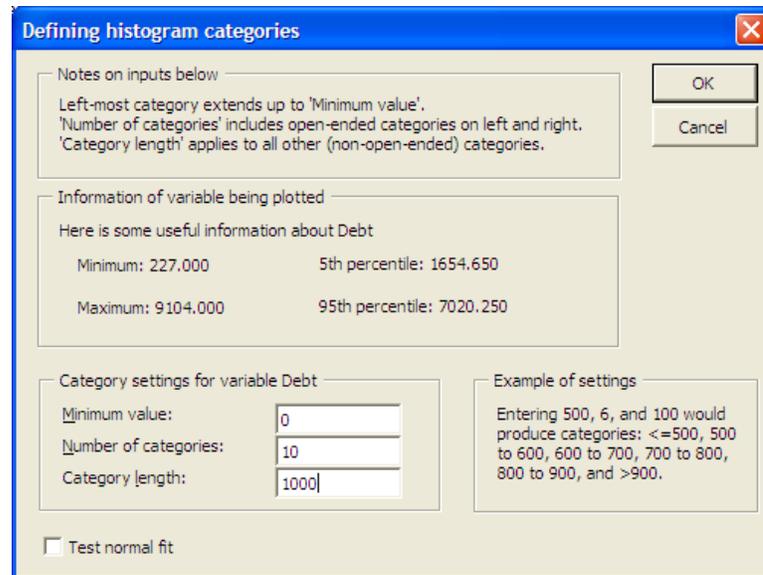


Figure 5.8: Histogram settings dialog box.

### Making histograms of z-scores with StatPro

Sometimes it can be difficult to select reasonable values for the settings to produce an informative histogram. One way to simplify this process is to make a histogram of z-scores, rather than a histogram of the actual data.

1. Compute the mean, the standard deviation.
2. Compute the z-scores for each piece of data.
3. Generate the histogram. When you create the histogram, use the -3 as the minimum, make 8 bins (explained below), and make the width 1 unit wide. For the z-scores of family debt in C05 FamilyIncome.xls, this produces the following histogram, to which we have added the frequency information (see below, "Adding Information to Histograms"). If you have done this correctly, the mean of the data should fall exactly in the center of the horizontal axis, at the zero point between the fourth and fifth bins. Each bin will be one standard deviation wide, easily showing you the number of observations that are within  $k$  standard deviations of the mean.

### Adding Information to Histograms

Rather than include a frequency table with every histogram, it is usually best to combine the two types of information into one graph. This will make it easier to read the graph and interpret the information that is presented. Adding this information is easy in Excel 2007.

1. Start with any histogram on the screen in Excel.
2. Click on the "Layout" ribbon to access the chart features.

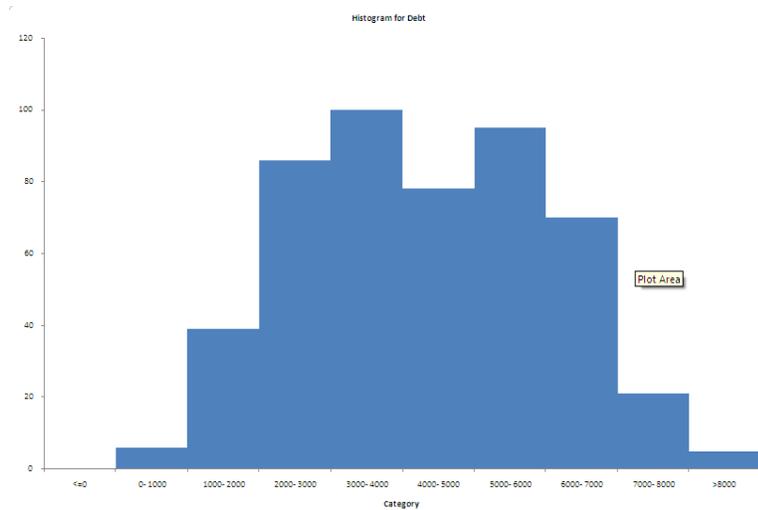


Figure 5.9: Histogram produced using the settings in figure 5.8 for the Debt variable in the file C05 FamilyIncome.xls.

3. You can then add the data table, data labels, or change just about any aspect of the chart. For the graph shown in figure 5.11 (page 156), we added data labels using the "outside end" option for their placement.

### Using array formulas to make a frequency table

Some of Excel's formulas (such as the "Frequency" formula) are array formulas. Array formulas work with an entire group of cells at once and produce multiple outputs at one time, each of which is placed in its own cell.

To enter an array formula, highlight all of the cells in which the formula should be calculated (this is the "array"). Type your formula (for example, you might want the frequencies for a set of data in a named range called "return" and the frequency ranges are in "bins" so =FREQUENCY(return, bins) would be the formula). Next, hit Control + Shift + Enter to enter the array formula in all the cells you highlighted at one time.

If you hit ENTER instead of CTRL+SHIFT+ENTER, the formula will only be entered in the first cell of the array, and you will need to start over.

If you make a mistake entering an array formula you must start over completely (if you try to type something else it gives you an error, saying "you cannot change part of an array formula"). To start over, highlight all the cells in the array. Hit "Delete" on the keyboard. Then begin the process over at the beginning (as above).

If you get trapped and cannot get out of editing a cell in an array formula or any other formula, hit the ESCAPE button on the keyboard. This will take you out of "edit mode" and restore the original contents of the cell that you were trying to change.

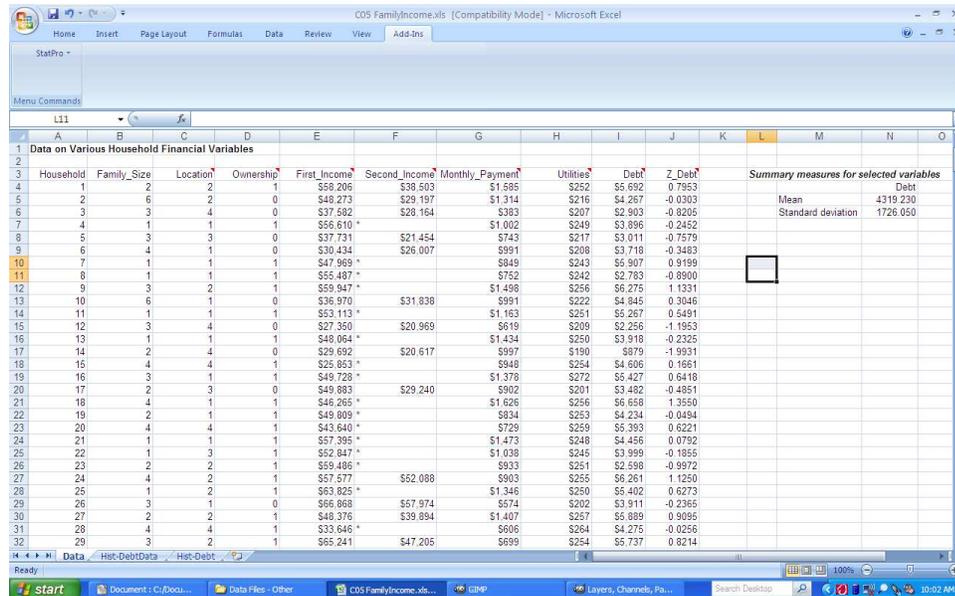


Figure 5.10: Family Income data, with z-scores for debt computed in column J.

### Creating a Histogram from a Frequency Table

1. First follow the steps above to make a frequency table. (But you don't have to use the mean and the standard deviation; you can pick any bin width and any starting point.)
2. Next, highlight the frequency table and click to activate the Insert ribbon.
3. Select the "column" type of chart and use the first subtype "clustered column".
4. Make any changes you want to the legend, data tables, and so forth. When you click "Next", you will have the option of making this a graph in the current sheet, or making it an entirely new sheet in the workbook. (StatPro's histogram feature automatically puts the graph on a new page.)
5. To make your graph look more like the ones that StatPro produces, place the cursor on one of the bars of the graph and RIGHT-CLICK. Select "Format Data Series..." from the bottom of the pop-up menu. Click on the "Series Options" tab and set the "Gap Width" to be "no gap". You can also outline the columns on the chart to stand out better by clicking on the "Border Color" tab, selecting "Solid Line" for the outline type and setting the color of the outline to black using the pull-down menu.

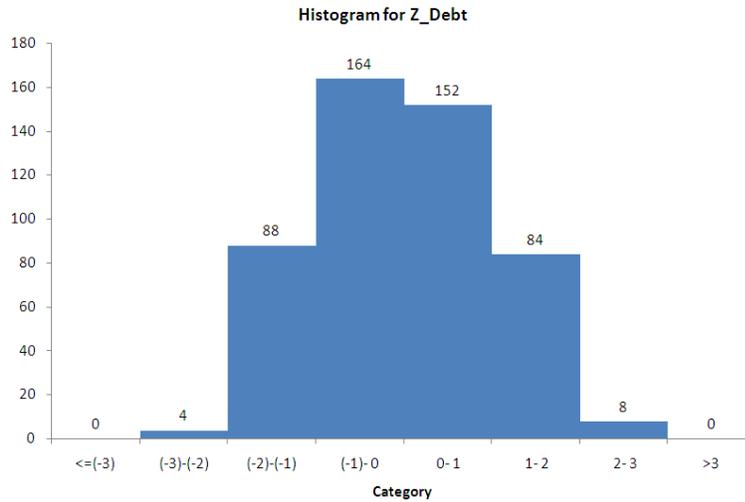


Figure 5.11: Histogram of z-scores for debt in the family income data.



Figure 5.12: The chart layout ribbon for a histogram (or a bar/column chart).

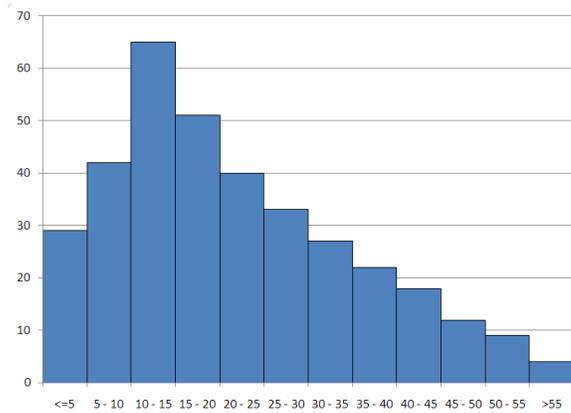
## 5.3 Homework

### 5.3.1 Mechanics and Techniques Problems

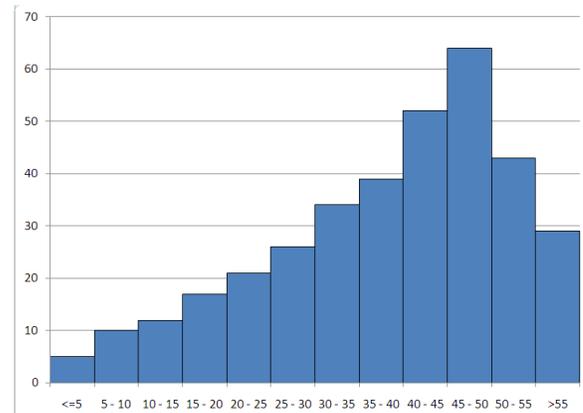
5.1. The data file "C05 Homes.xls" contains data on 275 homes that sold recently in the Rochester, NY area. Included in the data are observations of the location of the home, the annual taxes, the style of the home, the number of bedrooms, the number of bathrooms, the total number of rooms, the number of cars the garage will hold, the year in which the house was built, the lot size, the size of the home, the appraised value of the home, and the sale price.

1. Which of these variables are categorical? Which are numerical? For numerical variables, give the units and the range, for categorical variables describe the categories.
2. Make histogram of the appraised values and the selling price of the homes. Compare these distributions. In what way(s) are they similar? In what ways are they different? What does this tell you about the housing market in the greater Rochester area?

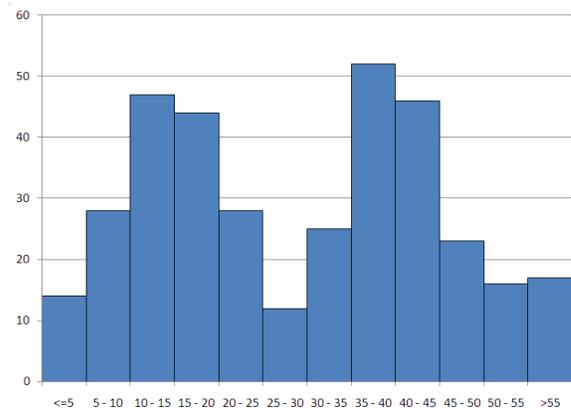
5.2. Consider the four histograms in figure 5.13 (labeled A - D). For each histogram, describe the shape.



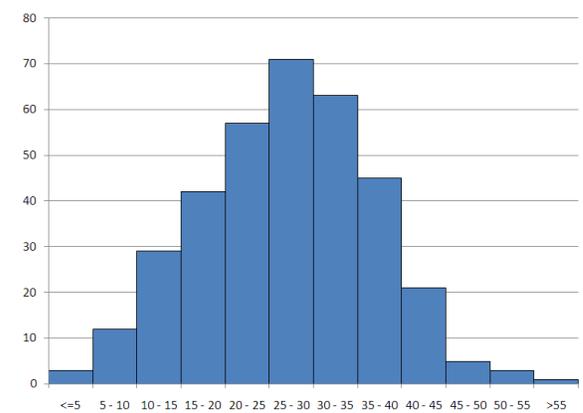
Histogram A.



Histogram B.



Histogram C.



Histogram D.

Figure 5.13: Histograms for Mechanics and Techniques problem 2.

5.3. Suppose you know that the mean of a set of data is 207.8 and the standard deviation is 43.2. If the median has a Z-score of  $-0.85$ , what is the median of this data?

5.4. Figure 5.14 shows two histograms. Histogram A represents the ages of people attending a recent event; histogram B represents the salary ranges of employees at a company. Each of these histograms could be improved in order to provide a better picture of the underlying data. For each, explain why the given histogram is less-than-ideal, and explain what you would do to improve it. In other words, how would you go about making a better histogram?

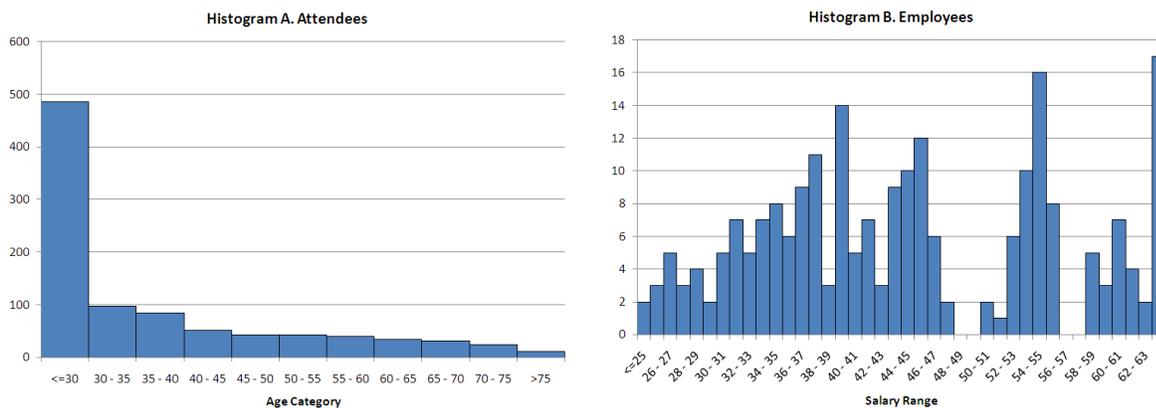


Figure 5.14: Histograms for Mechanics and Techniques problem 4.

### 5.3.2 Application and Reasoning Problems

5.5. The histograms in figure 5.15 (page 159) show flight arrival data for two airlines your company is considering signing a permanent contract with. Both airlines offer similar flight times, similar service, and travel to all of the same cities. For 1,000 flights on each airline last year, the data is a record of how far before or after the scheduled arrival time that the flights actually arrived. Negative times indicate that the flight landed ahead of schedule. Positive times indicate late-arriving flights. For each of these graphs, construct both a cumulative distribution graph and a boxplot. Based on these data and your graphs which airline would you be willing to pay more for? Explain your reasons.

5.6. Over Easy, a breakfast restaurant, has collected data on the number of customers the restaurant serves each day of the week for a full year. Assume these data are normally distributed with a mean of 135 and a standard deviation of 22. They want to use this information to determine staffing for the restaurant.

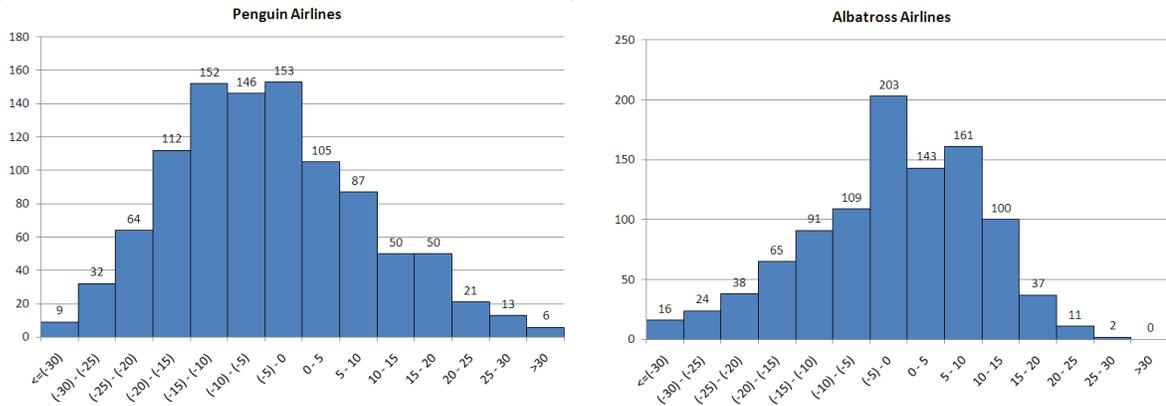


Figure 5.15: Histograms for Application and Reasoning problem 5.

- 68% of the time, the total number of customers will fall between what two numbers (approximately)? What about 95% of the time? What about 99.7% of the time?
- Suppose a normal server can handle about 35 people in one day (at this restaurant). For each of the situations in part (a), how many servers are needed?
- How many servers would you staff? Explain your answer.

5.7. You are the manager of a local grocery store and would like to collect data on the checkout times for customers in order to help inform staffing and the arrangement of your various express and standard checkout lines. You expect about 1000 customers each day and plan to measure the checkout time as being the total time elapsed from when the cashier starts scanning the first item to when the customer leaves with his/her groceries. Sketch a reasonable histogram of the checkout times, paying particular attention to the shape of the distribution and the units on the axes. Make sure that each axis is labeled and that each bin of the histogram is labeled with reasonable values for this situation. When you have a reasonable graph, write a short (2-3 sentence) paragraph explaining why you expect the checkout times to have the distribution you have drawn (left/right skewed, uniform, symmetric, bimodal).

- 5.8. Consider the salary distribution for a company named OutRun shown in figure 5.16.
- Write a brief description (2-3 sentences) explaining what you infer about the company from looking at this distribution of salaries.
  - Now, sketch a new histogram of these same data by making the bins twice as wide. In other words, you have exactly the same data and distribution, we are just going to draw the histogram with half as many bins (each cover twice the range of salaries).

3. What does your new histogram (part b) seem to suggest about the salary distribution at the company? Write 2-3 sentences.
4. Based on your two graphs and analysis, how useful is the mean salary in representing the company? What other measures might be more appropriate?

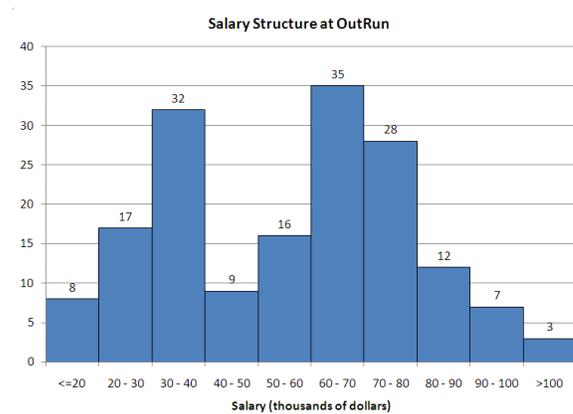


Figure 5.16: Histograms for Application and Reasoning problem 8.

5.9. Consider the histogram below, made by one of your co-workers to represent the age distribution of customers at your company. This person claims that the graph supports the idea that most of their customers are neither young nor old and do not have children at home, which will help you decide when and where to do your advertising. Answer the following questions.

1. What does this graph tell you about your customers? What percentage of your customers is young? What percentage is old? What do you even mean by "young" or "old" in this case?
2. Why do you think your coworker claims that most of your customers do not have children at home? Do you agree? What would help you make a more informed decision about this?
3. How would you improve the graph in order to get more information without needing more data? What might the graph look like then? Sketch three possible versions of this data if the histogram were created differently (say, with more bins or different starting points).

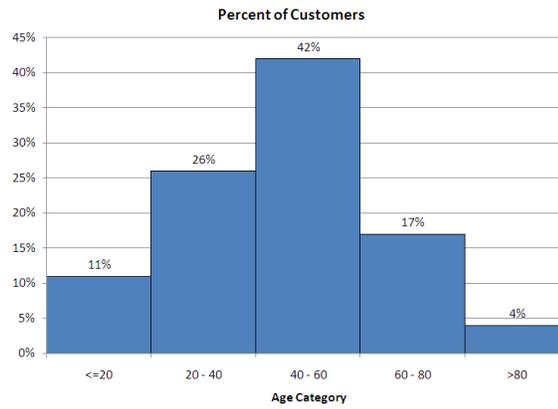


Figure 5.17: Histograms for Application and Reasoning problem 9.

### 5.3.3 Memo Problem

To: Beef n' Buns Store Managers  
From: Chad R. Chez, Regional Manager of Beef n' Buns  
Date: May 18, 2008  
Re: Response to poor service

As you know, we have been analyzing the customer service data that each of you has compiled and submitted to me. What we have found from an analysis of the data is that service times seem to be dependent on two things:

1. Whether the order is placed from the breakfast menu in the morning or from the lunch menu during the day
2. Whether the order is placed at the counter or at the drive through

The attached data file contains the service times from your restaurant and has been reworked to reflect these venues (a venue is a combination of a menu type and a location for the order). We need a detailed description of the service times in each venue. Analyze the patterns of the service times in all the venues and state whether these seem reasonable or whether they might reveal possible problems giving rise to customer complaints. It will be critical to this ongoing study that you first identify the venue or venues that may require more investigation. Once you have identified these venues you need to propose possible reasons for the problems that will lead us to further data collection. What additional data do we need in order to get at the underlying problems with customer service?

**Attachment:** Data file "C05 BeefNBuns 2.XLS"