

A Little Stats Won't Hurt You

Nathaniel Derby, Statis Pro Data Analytics, Seattle, WA

ABSTRACT

This paper gives an introduction to some basic but critically important concepts of statistics and data analysis for the SAS programmer who pulls or manipulates data, but who might not understand what goes into a proper data analysis. We first introduce some basic ideas of descriptive statistics for one-variable data, and then expand those ideas into many variables. We then introduce the idea of statistical significance, and then conclude with how all these ideas can be used to answer questions about the data. Examples and SAS[®] code are provided.

KEYWORDS: Descriptive statistics, Statistical significance, SAS.

All SAS code used in this paper is downloadable from <http://nderby.org/docs/RUG09-IntroStats.sas>.

INTRODUCTION: WHAT CAN STATISTICAL METHODS TELL US?

Statistical methods can be used to describe data and then extract information from them. They can be used to test hypotheses from the data ("Is X correlated with Y ?").¹ They can be used to extrapolate trends for forecasts, or to look into the past and quantify what happened ("What is the effect of X on Y ?"). But before any of these more complex methods can be used, the first step is always to *look at the data* in many different ways. The idea is to effectively describe (or *summarize*) the data, and then look for "interesting" features. Note, however, that the precise meaning of that word depends on both the data and the research goal in question; one person's interesting feature is another person's nuisance. For example,

- In the preliminary steps of an analysis, we care about whether each data point is valid: "Should this data point be included in the later analysis or not?"
- In the intermediate stages of an analysis, after we have verified the data, we care about estimating the effect of a given treatment: "What is the effect of the treatment?"
- In more advanced stages of an analysis, we may focus on how robust the effect estimates are: "If we change some of the data slightly, do we get about the same effect of the treatment?"

Moreover, we want our data to make sense within our given context. A basic strategy is to look for *data irregularities*, which could be errors or something interesting. A common maxim in statistics is "if it looks interesting, it's probably wrong." However, if a data irregularity isn't wrong, it may be an interesting feature that might not have been found had a statistical method not been done. A good strategy is to think about what caused such an irregularity and to investigate it.

In this paper, we will illustrate these ideas with a few examples. These methods comprise *exploratory data analysis* (EDA), which involves looking at the data in a few different ways to let us see what the data are telling us, and which questions we should consider asking. Though many of these methods may seem simple, they are prerequisites for more advanced statistical methods we may have heard of, such as ANOVA (PROC ANOVA, PROC GLM), linear regression (PROC REG), logistic regression (PROC LOGISTIC), and ARIMA (PROC ARIMA). As such, *exploratory data analysis is essential*, and must be done before using more advanced statistical techniques.² These more complex statistical techniques all quantify the results we uncover with an EDA. For example, if an EDA shows us some evidence that X has some kind of an effect on Y , we can use a more complex model like linear regression (PROC REG) to give us our best estimate of that effect. If an EDA weren't performed first, we wouldn't know which two variables X and Y to look at.

For simplicity, this paper will focus on *univariate* methods, where we look at one variable at a time. We will compare two data sets by comparing their univariate characteristics, but we will not look at methods involving interactions between two variables (e.g., two-dimensional scatterplots).

¹Note that correlation alone does not imply causality. As an example, cities with larger police forces tend to also have more crime. Does it follow that police presence causes crime?

²The major exception to this rule is for many techniques in data mining, for which complex statistical methods are applied to data without first looking at the data. This is done out of necessity; the data sets are huge, and there are simply too many variables to allow for doing an EDA.

EXPLORATORY DATA ANALYSIS (EDA)

The main idea of *exploratory data analysis* is, as its name implies, to explore the data. For the univariate case (looking at one variable at a time), this means looking at data in ways designed to easily discern the *distribution* of the data (i.e., how it is *distributed*, or spread out). This mainly involves two methods:

- *Data Visualization*: Graphical techniques which allow us to quickly and easily see general trends in the data.
- *Descriptive Statistics*: Statistical measures which summarize various characteristics of the data distribution.

Throughout this paper, we will illustrate examples of both classes of the above methods with six data sets of systolic blood pressure measurements of patients (in millimeters of mercury (mmHg)), each having 50 observations. For example, this is our first data set:

113	122	106	131	130	112	132	122	114	117
108	103	117	120	117	126	116	128	124	123
126	143	118	110	103	119	136	109	113	116
127	97	144	108	121	128	115	115	124	115
120	98	115	107	131	126	112	118	121	126

These data sets will simply be named `data1` through `data6`, and will have the numeric variables `bp` (shown above) and `group` (equal to a number between 1 and 6).

Looking at these raw data points above makes it difficult to discern any characteristics about their distribution. Furthermore, the problem would be much worse if we had 500 or 5000 data points. There is simply too much information here. However, the two classes of methods above will alleviate this problem.

ONE-DIMENSIONAL SCATTERPLOTS AND BUBBLE PLOTS

An easy and logical first step at looking at the data is to simply make a one-dimensional *scatterplot* or *bubble plot*, as shown in Figures 1(a)-(c).³ We make a basic scatterplot with the SAS/GRAPH[®] package as such:⁴

```
PROC GPLOT data=data1;
  PLOT group*bp;
RUN;
```

The above code produces Figure 1(a), which looks very plain. Furthermore, it is misleading, since it shows only 30 values out of our data set of 50 observations. This is because our data set only has 30 *distinct* values; some values are repeated multiple times, and this scatterplot doesn't show us which values are repeated or how many times they are repeated. Using a tip from Kucera (1996), we can create an auxiliary data set which contains counts of each distinct value, and then use that with PROC GPLOT to create a scatterplot which shows us multiple values, using color coding of the scatterplot points. We can also customize the output format for visual clarity, as shown in Figure 1(b):

```
PROC FREQ data=data1 noprint;
  TABLES group*bp / out=data1stats ( keep = group bp count );
RUN;

goptions ftitle='Times/bold' ftext='Times';
axis1 label= ( 'mmHg' ) order= ( 90 to 150 by 10 ) minor= ( number=4 ) value= ( height=1.2 );
axis2 label= ( ' ' ) value= ( height=1.2 );
title 'Systolic Blood Pressure';

PROC GPLOT data=data1stats;
  PLOT group*bp=count / haxis=axis1 vaxis=axis2;
RUN;
```

Using the key at the bottom of the plot in Figure 1(b), we can see which values on the scatterplot are repeated, and how many times. However, this is not a convenient depiction of the data, since colors do not have a natural order (e.g., it may be difficult to remember that a blue point represents twice as many values as a red one). Furthermore, the color coding is meaningless if viewed from a black-and-white copy. For both of these reasons, a *bubble plot* is often preferable to a scatterplot. In a bubble plot, each point is marked with a circle (or “bubble”) whose radius is proportional to how many repeats of that value are in the data set.

³Most uses of a scatterplot involve two dimensions, where two-dimensional data points (x,y) are plotted onto an x - y set of axes. Here we are essentially doing the same, except that we are plotting multiple values of x with one value of y (equal to 1 in this case).

⁴These results are from ODS PDF on SAS 9.1.3. Because of ODS restructuring on SAS 9.2, these results will look slightly different on that platform. As with any ODS PDF output, these examples can be generated by first using `options papersize="letter" orientation=landscape;` and then placing the example code between the statements `ods pdf file="&outputroot\outputx.pdf";` and `ods pdf close;`. Note that the code for this and other examples can be done without the SAS/GRAPH package simply by removing the `G` (e.g., PROC PLOT rather than PROC GPLOT).

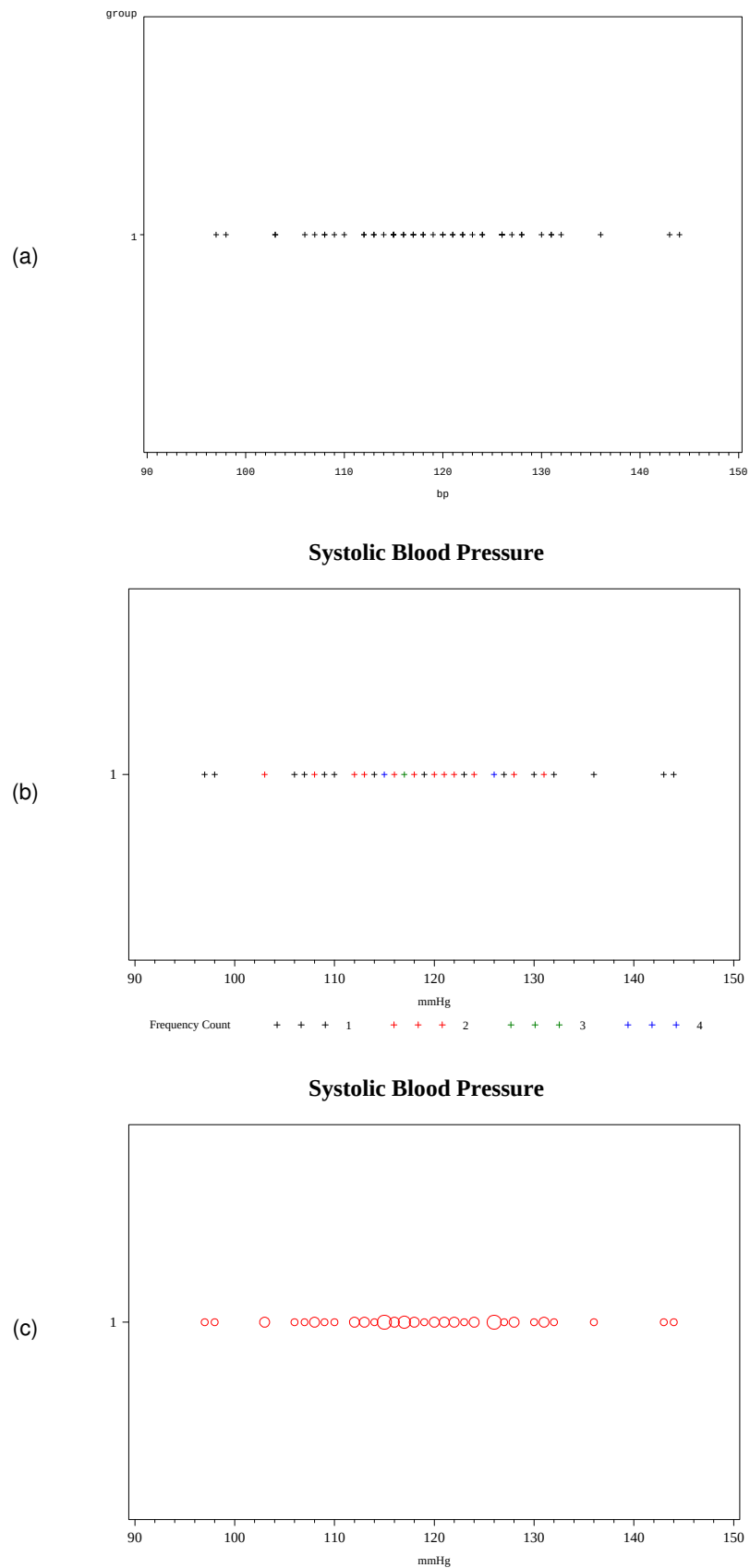


Figure 1: A one-dimensional graphical representation of systolic blood pressure measurements of 50 patients using SAS/GRAPH, as a default-formatted scatterplot (a), a custom-formatted scatterplot (b), and bubble plot (c).

This is done using the `BUBBLE` rather than the `PLOT` statement, where we make the bubbles red as shown in Figure 1(c):

```
PROC GPLOT data=data1stats;
  BUBBLE group*bp=count / haxis=axis1 vaxis=axis2 bcolor=red;
RUN;
```

Note that if we didn't have repeated values, a scatterplot as shown in Figure 1(a) would be sufficient for our purposes.

The scatterplot and bubble plot simply give a first glance of the data, without giving any quantitative information. However, they are both designed to effectively convey the main characteristics of the data in a visual manner. In the bubble plot in Figure 1(c), we see that most of the values are between 108 mmHg and 132 mmHg. Values below 100 mmHg or above 140 mmHg are rare, and as such might be called *outliers*. This term will be clarified later on in this paper, but for now we will use this term loosely to simply mean data points that are outside the range that contains most of the data points.

Beyond finding outliers and a range (or number of different ranges) where most of the data lie, a scatterplot or bubble plot does not have many uses. Still, just for these two uses, these two plots are extremely useful, as they show us information that we might not otherwise see, which could be important.

HISTOGRAMS

A *histogram* is simply a chart of frequency or percentage counts for different ranges of the data. We can do this with `PROC GCHART` as such, producing Figure 2(a):

```
PROC GCHART DATA=data1;
  VBAR bp;
RUN;
```

As before, we would like to add some custom formatting to make it look a little more readable, producing Figure 2(b):

```
goptions ftitle='Times/bold' ftext='Times';
axis1 label= ( 'Interval Midpoint (mmHg)' height=1.2 ) offset=( 4, 4 ) value=( height=1.2 );
axis2 label=( angle=90 height=1.2 'Frequency' ) order=( 0 to 20 by 5 ) minor=( number=4 ) value=( height=1.2 );
title 'Systolic Blood Pressure';

PROC GCHART DATA=data1;
  VBAR bp / maxis=axis1 raxis=axis2 width=4 space=2;
RUN;
```

Above, in the `axis1` statement, the `offset` option places spaces between the left-/rightmost frequency bars and the left/right edges of the graph. Here we see a chart of seven bars whose heights are proportional to the frequency counts for their designated variable intervals. For instance, the first bar tells us that our data set has two data points with blood pressure in an interval centered at 96 mmHg. However, this is still not ideal, since we don't know explicitly what this interval is. Mathematically, the upper limit of this interval is halfway between 96 and the midpoint of the next interval, 104, which is equal to 100 (mmHg).⁵ For a quick glance at this data, this could be fine. However, it might be more useful to show these intervals explicitly. We can do this by modifying the `axis1` statement and adding a `midpoints=` option, to make sure that SAS doesn't change the underlying midpoints (and to remind ourselves of where those midpoints are), resulting in Figure 2(c).⁶

```
axis1 label=( 'mmHg' ) value=( height=1.2 '92 - 100' '100 - 108' '108 - 116' '116 - 124' '124 - 132'
  '132 - 140' '140 - 148' ) offset=( 4, 4 );

PROC GCHART DATA=data1;
  VBAR bp / maxis=axis1 raxis=axis2 width=4 space=2 midpoints = 96 to 144 by 8;
RUN;
```

In the above code, we explicitly set the midpoints of our intervals to be

96 104 112 120 128 136 144

with the labels

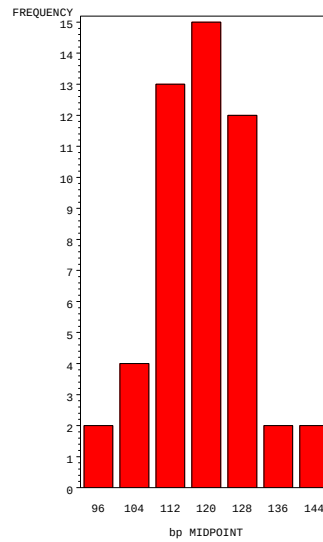
92 - 100 100 - 108 108 - 116 116 - 124 124 - 132 132 - 140 140 - 148

This makes our output just a little more readable (and thus informative).

⁵A data value that lies exactly on the limit (100 in this case) will be counted in the higher range (the interval centered at 104). As such, the resulting histogram could be misleading if there are many values on the limits. In such a case, it would be better to set limits to values that are not included in the data set – such as fractional values (like 100.5) for this particular data set.

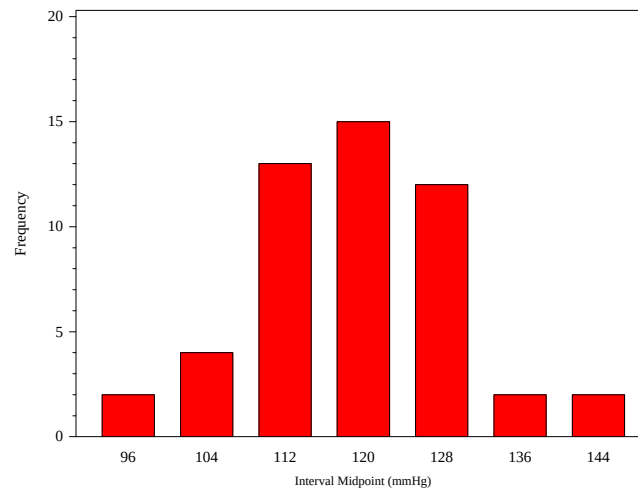
⁶The `axis2` statement is the same as in the previous code, so it is omitted.

(a)



Systolic Blood Pressure

(b)



Systolic Blood Pressure

(c)

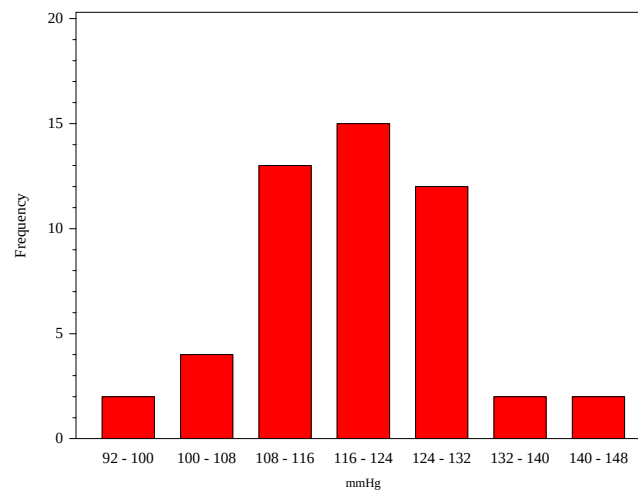


Figure 2: Histograms of systolic blood pressure measurements of 50 patients, with default (a) and custom (b) formatting, and with explicit ranges (c).

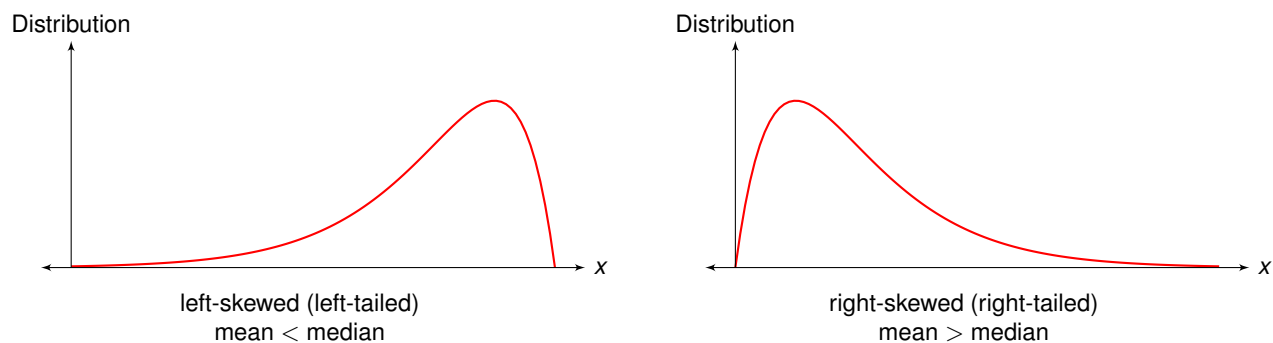


Figure 3: Skewness.

Looking at Figure 2(c), we see that the data appear to be centered at around 120 mmHg, and that the “bulk” of the data appears to be within 8 or 10 mmHg of 120 mmHg. That is, we are answering two questions of interest about the data:

- *What is the central value of the data?* This is typically the most important question about any data set, which might follow our intuition. We usually want to know the average (expected) weight of a group of patients, or the average effect of a treatment, or the average amount of a given drug administered. While we will quantify the term “average” shortly, the idea is that we would like to know the central value.
- *How spread out are the data?* This is typically the second most important question about a given data set. Indeed, having a central value is almost meaningless if we don’t also know how spread out the data are. Given that the patient blood pressures in our data set are centered around 120 mmHg, it makes quite a difference if the “bulk” of the data is within 10 mmHg of 120 mmHg versus within 20 mmHg of 120 mmHg.

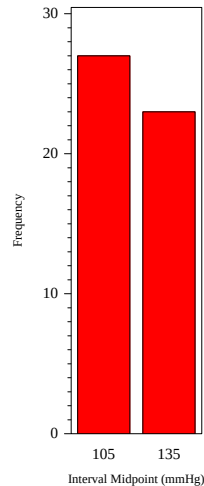
By using an odd number of intervals in Figure 2(c), it is easier (at least in this case) to see an estimate of the central value and spread. We can quantify both of these concepts:

- *What is the central value of the data?* We can address this in three ways:
 - The *mean* of a data set is the arithmetic average, which means that we take their sum and divide it by the number of data points.
 - The *median* of a data set is the “middle value,” meaning that 50% of the data is below this value. Arithmetically, this means that we order the data points. If we have an odd number of data points, we take the observation at the middle number (e.g., 3 is the middle number of 5). If we have an even number of data points, the median is the arithmetic average of the $\frac{n}{2}$ th and the $\frac{n}{2} + 1$ th observation.
It is often useful to look at both the mean and the median. If they are equal, we say that the data is *symmetric*, meaning that there is just as much data above the mean as there is below it. Otherwise, we say that the data is *skewed*, or *long-tailed*, as shown in Figure 3. If the mean is greater than the median, the “bulk” of the data is to the left, and there is a long tail on the right, and vice-versa. The direction of the longer tail is the same as the direction of the skewness. Therefore, a right-skewed distribution is sometimes informally called *right-tailed*. In our histogram in Figure 2(c), the data appears to be slightly right-skewed (i.e., the bulk of the data is to the left), but this might be misleading, as we’ll comment on shortly.
 - The *mode* of a data set is the value or the category that is a maximum (or “high point”) in the distribution or the histogram. In our histogram in Figure 2(c), the mode is the category “116 - 124 mmHg”. Here we have only one mode, but more than one mode is possible. Practically, this means that there is only one value around which the data is clustered.
- *How spread out are the data?* We can address this in several ways:
 - The *standard deviation* gives a rough estimate of the average distance from the mean.⁷ In a bell-shaped (or *normal*) distribution, about 65% of the data lie within one standard deviation of the mean.
 - The *minimum* is simply the smallest value in a given data set.
 - The *25th percentile* is a data point for which 25% of the data set lies below this value. This is also called the *1st quartile*.
 - The *75th percentile* is a data point for which 75% of the data set lies below this value. This is also called the *3rd quartile*.

⁷More specifically, it is the square root of the mean squared distance from the mean. It is calculated this way so that a negative distance is counted the same as a positive one. More details about this are given on page 16.

Systolic Blood Pressure, 2 Intervals

(a)



Systolic Blood Pressure, 15 Intervals

(b)

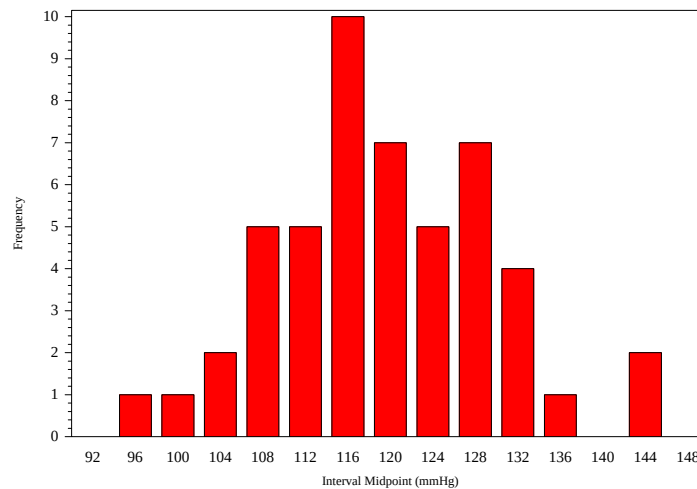


Figure 4: Histogram of systolic blood pressure measurements of 50 patients, with 2 intervals (a) and 15 intervals (b).

- The *interquartile range* is the difference between the 75th and 25th percentiles.
- The *maximum* is the largest value in a given data set.

Note that the median is the same as the 50th percentile. Furthermore, as the standard deviation increases, the difference between the 75th and the 25th percentile also increases.

The shape of the distribution is often very important. For example, there is a huge difference between a right-skewed blood pressure distribution (with more patients with lower blood pressure) than a left-skewed one (with more patients with higher blood pressure).

While our histogram in Figure 2(c) appears to give us a useful estimate of the distribution of our data, care must be given to make sure that we don't have too many or too few intervals (also known as *levels* or *classes*). The number of intervals can be set with the `levels` option in `PROC GCHART` as follows:

```
axis1 label=( 'Interval Midpoint (mmHg)' height=1.2 ) value=( height=1.2 );
axis2 label=( angle=90 'Frequency' height=1.2 ) value=( height=1.2 );
title 'Systolic Blood Pressure, xx Intervals';

PROC GCHART data=datal;
  VBAR bp / maxis=axis1 raxis=axis2 levels = xx;
RUN;
```

Doing this for 2 and 15 intervals gives us Figures 4(a) and (b), respectively. Here we see that with 2 intervals, we don't have enough intervals to show any useful information about the distribution. On the other hand, with 15 intervals, we get a misleading estimate of our distribution, since with so many intervals (in relation to the number of data points), we have many intervals with very small frequencies. In such a situation, a difference of one or two data points will look significant. For example, in Figure 4(b), it looks like our data has four modes (high points): At the 7th, 10th and 14th intervals. Actually, the data has one mode; it just looks like there are more than one because of the paucity of data in many of these intervals. This illustrates a general rule as well expressed by Cabrera and McDougall (2002, p. 84):

Too few intervals and important distributional features will be missed; too many and artificial features of distribution will be introduced by the display itself.

Practically, Six Sigma proposes the following guide as a rough rule of thumb (Stamatis, 2003, p. 55):

Number of Data Points	Number of Intervals
31 to 50	5 to 7
50 to 100	6 to 10
100 to 250	7 to 12
over 250	10 to 20

For more information about customizing histograms with SAS, see Watts (2008).

Overall, because of variations in histograms resulting from different numbers of intervals (or even from shifting the intervals, which is not illustrated here), inferences we get from histograms might be misleading. For more robust results which are less prone to spurious conclusions, we turn to the box plot.

BOX PLOTS

For more precise descriptions of the data distribution, we can create a *box plot*.⁸ This is simply a graphical representation of the six most common descriptive statistics of the data, as defined in the previous section:

minimum, 25th percentile, mean (50th percentile), 75th percentile, maximum, mean.

The first five of these numbers, when used together, comprise a *five-number summary* of the data.

The SAS code needed to create a box plot is analogous to the PROC GGPLOT statement for our scatterplot, but changing the order of the variables `bp` and `group`. This creates a one-dimensional bubble plot that is oriented vertically rather than horizontally, as shown in Figure 5(a):

```
goptions ftitle='Times/bold' ftext='Times';
symbol1 c=red;
axis1 label=( angle=90 'mmHg' height=1.2 ) order=( 90 to 150 by 10 ) minor=( number=3 ) value=( height=1.2 );
axis2 label=( ' ' ) value=( height=1.2 );
title 'Systolic Blood Pressure';

PROC GGPLOT data=datalstats;
  BUBBLE bp*group=count / vaxis=axis1 haxis=axis2 bcolor=red;
RUN;
```

While we can create a box plot using a minor variation of the above PROC GGPLOT code, it is easier to use PROC BOXPLOT as follows, creating Figure 5(b):⁹

```
axis1 label=( 'mmHg' height=1.2 ) order=( 90 to 150 by 10 ) minor=( number=3 ) value=( height=1.2 );
symbol1;

PROC BOXPLOT data=datal;
  PLOT bp*group / vaxis=axis1 haxis=axis2;
RUN;
```

⁸Also known as a *box-and-whiskers plot*.

⁹To make *almost* the same output as Figure 5(b) by using PROC GGPLOT, use the code above used to create Figure 5(a), but change the `symbol1` statement to `symbol1 i=boxt co=red bwidth=10`. This creates the desired boxplot, but without the black cross designating the mean value. This can be added to the graph via an `annotate` data set, but that seems overly complicated, given that PROC BOXPLOT does it without the `annotate` data set. This is one of a few differences between PROC GGPLOT and PROC BOXPLOT when making box plots; see Adams (2008) for details.

Careful comparisons of Figures 5(a) and 5(b) show slight differences between their respective dimensions. This is simply due to various default setting differences between PROC GGPLOT and PROC BOXPLOT.

SAS appears unable to make horizontal box plots (at least with PROC GGPLOT or PROC BOXPLOT).

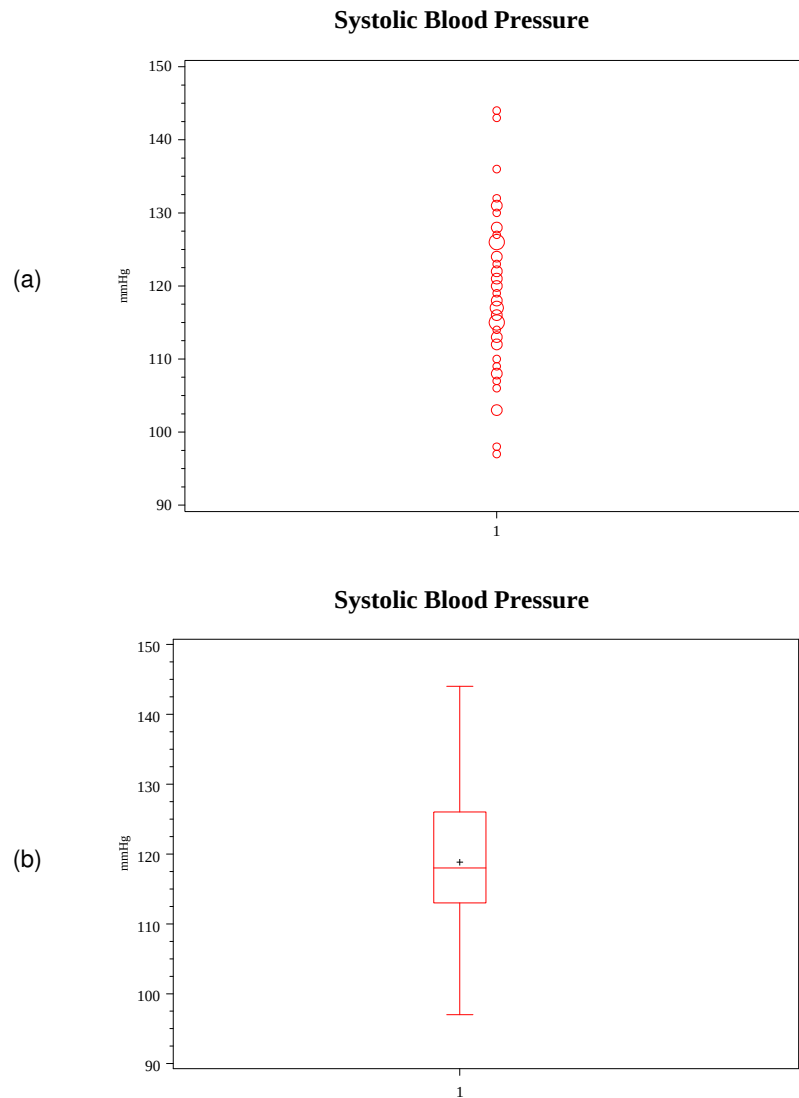


Figure 5: Vertical scatterplot (a) and box plot (b) of systolic blood pressure measurements of 50 patients.

In Figure 5(b), the highest and lowest horizontal lines represent the maximum and minimum values, respectively. The “box” has three horizontal lines in it: The highest and lowest ones represent the 75th and 25th percentiles, respectively, while the middle one represents the median. As such, the height of the box represents the interquartile range. The black cross represents the mean value.

Comparing Figures 5(a) and 5(b), we see two different representations of the same data set. They both show that the “bulk” of the data is between 108 mmHg and 132 mmHg (as we saw in the scatterplot/bubble plot section), except that with the box plot this is more precisely quantified. That is, in the scatterplot in Figure 5(a), it is difficult to see precisely where the 25th and 75th percentiles might be – whereas this is very clear in the box plot.

When looking at histograms of our data set, we had tentative evidence that the data were slightly right-skewed (or right-tailed). This was tentative because, as we saw with the histograms of 2 and 15 levels, the shape of a histogram can be unduly influenced by the number of intervals, or by shifting the intervals. However, we know from Figure 3 that a right-skewed distribution has its mean greater than its median. With our box plot in Figure 5(b), we see that that is indeed the case, so we can now definitively conclude that our data set is right-skewed.

In summary,

A box plot is more reliable than a histogram in illustrating percentiles or skewness.

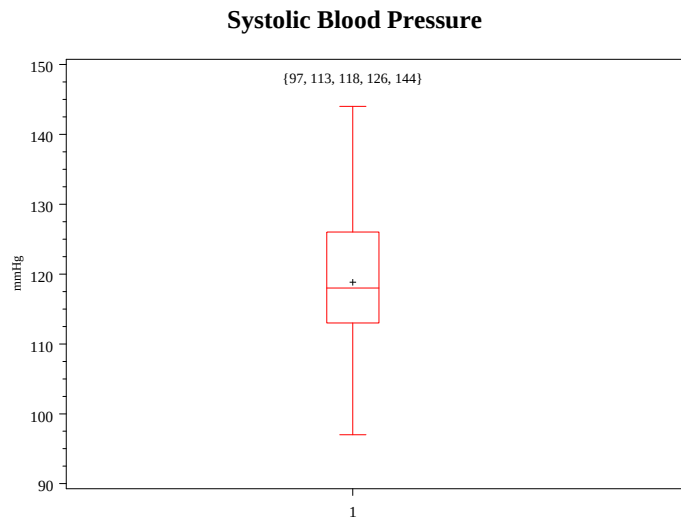


Figure 6: Box plot of systolic blood pressure measurements of 50 patients, annotated to include the five-number summary above the data plot.

This is because a box plot incorporates *summary statistics*, which are statistical measures (such as percentiles, or a sample mean) which describe the data. For illustrative purposes, it might be useful to write the five-number summary over the box plot with an `annotate data set`:¹⁰

```
PROC UNIVARIATE noprint data=data1;
  VAR bp;
  BY group;
  OUTPUT min=min mean=mean q1=q1 median=med q3=q3 max=max out=stats;
RUN;

DATA annol;
  SET stats;
  FORMAT function $8. text $50.;
  RETAIN when 'a';
  function = 'label';
  text = '{'||trim( left( put( min, 5. ) ) )||', '||trim( left( put( q1, 5. ) ) )||', '||
    trim( left( put( med, 5. ) ) )||', '||trim( left( put( q3, 5. ) ) )||', '||trim( left( put( max, 5. ) ) )||'}';
  position = '2';
  xsys = '2';
  ysys = '3';
  x = 1;
  y = 85;
  size = 1.1;
  OUTPUT;
RUN;

PROC BOXPLOT data=data1;
  PLOT bp*group / vaxis=axis1 haxis=axis2 annotate=annol;
RUN;
```

Here we use `PROC UNIVARIATE` to compute these statistics and put them into a data set (`stats`), and then use that data set to create the `annotate data set`, which is then used in our box plot graph. The result is shown in Figure 6. Here we see that our five-number summary is

{ 97, 113, 118, 126, 144 }.

¹⁰An *annotate data set* is a data set with a special structure, used to add a symbol or data to an existing SAS graph made with the SAS/GRAPH package. For more information, see Carpenter (2006).

COMPARING MULTIPLE DATA SETS

Scatterplots, histograms, box plots and summary statistics are effective for summary purposes, and they can be particularly effective for comparing multiple data sets. Suppose we have data sets of blood pressure from three different groups of patients (groups A, B and C), where group A is the data we've been analyzing up to now. We will compare these three data sets with box plots and the five-number summary:

```
DATA data123;
  SET data1 data2a data3;
RUN;

PROC UNIVARIATE data=data123 noprint;
  VAR bp;
  BY group;
  OUTPUT min=min mean=mean q1=q1 median=med q3=q3 max = max out=stats;
RUN;

DATA anno123;
  SET stats;
  FORMAT function $8. text $50.;
  RETAIN when 'a';
  function = 'label';
  text = '{||trim( left( put( min, 5. ) ) )||', '||trim( left( put( q1, 5. ) ) )||', '||
    trim( left( put( med, 5. ) ) )||', '||trim( left( put( q3, 5. ) ) )||', '||trim( left( put( max, 5. ) ) )||}';
  position = '2';
  xsys = '2';
  ysys = '3';
  x = group;
  y = 85;
  size = 1.1;
  OUTPUT;
RUN;

axis1 label=( 'mmHg' ) minor=( number=4 ) value=( height=1.2 );
axis2 label=( ' ' ) order=( 1 to 3 by 1 ) value=( height=1.2 'Group A' 'Group B' 'Group C' ) minor=none;

PROC BOXPLOT data=data123;
  PLOT bp*group / vaxis=axis1 haxis=axis2 annotate=anno123;
RUN;
```

The output, shown in Figure 7(a), shows an obvious data irregularity of some kind in Group B. Indeed, this is the real value of exploratory data analysis: To quickly and easily find data irregularities, just as we are seeing here. Certainly there is something to comment about with Group C as well, but Group B is the bigger problem. This analysis shows not only that the minimum is equal to zero, but that the data is heavily skewed toward that value as well. That is, the mean is so far below the median (making it left-skewed) than it is actually closer to the 25th percentile than to the median.

When finding a data irregularity such as this, a good first step is to ask if this makes sense. Without even looking at the raw data, we know from the minimal value in the five-number summary that we have at least one data point with a blood pressure of 0 mmHg. Clearly this is possible only if either a given patient is deceased, or if there is some kind of data error. Assuming that we are not including deceased patients in our data set (which in certain situations might not be the case!), we can assume that this is a data error which was corrected the next day, thus producing Figure 7(b) when we re-run the code. We can make a couple observations from this revised box plot:

- All three groups appear to have about the same central values, at least compared with their spreads. That is, the median values (118, 119 and 120) and mean values (the locations of the black crosses) appear to be about the same.
- These three groups have very different spreads. Both the overall ranges (maximum minus the minimum) and interquartile ranges (75th percentile minus the 25th percentile) are highest for group B and lowest for group C.

Again, we can ask ourselves if this makes sense, or if it's a sign of some kind of data irregularity. For the purposes of this paper, we can assume that it does make sense; perhaps group B is composed of patients of many different ages, races, geographic areas, income levels and states of health; whereas group C is composed of healthy middle-class white males between 40-50 years old; and group A is between these two extremes. Regardless of the plausibility of this explanation, these general principles of considering whether the data behavior makes sense still stand.

For another example, we now turn to Figure 8(a) to consider data from groups D, E and F, using similar SAS code to that of the previous example. Here we see a data irregularity for group E, except that unlike group B in the previous example, it's oriented upward (right-skewed) this time. Does this make sense?

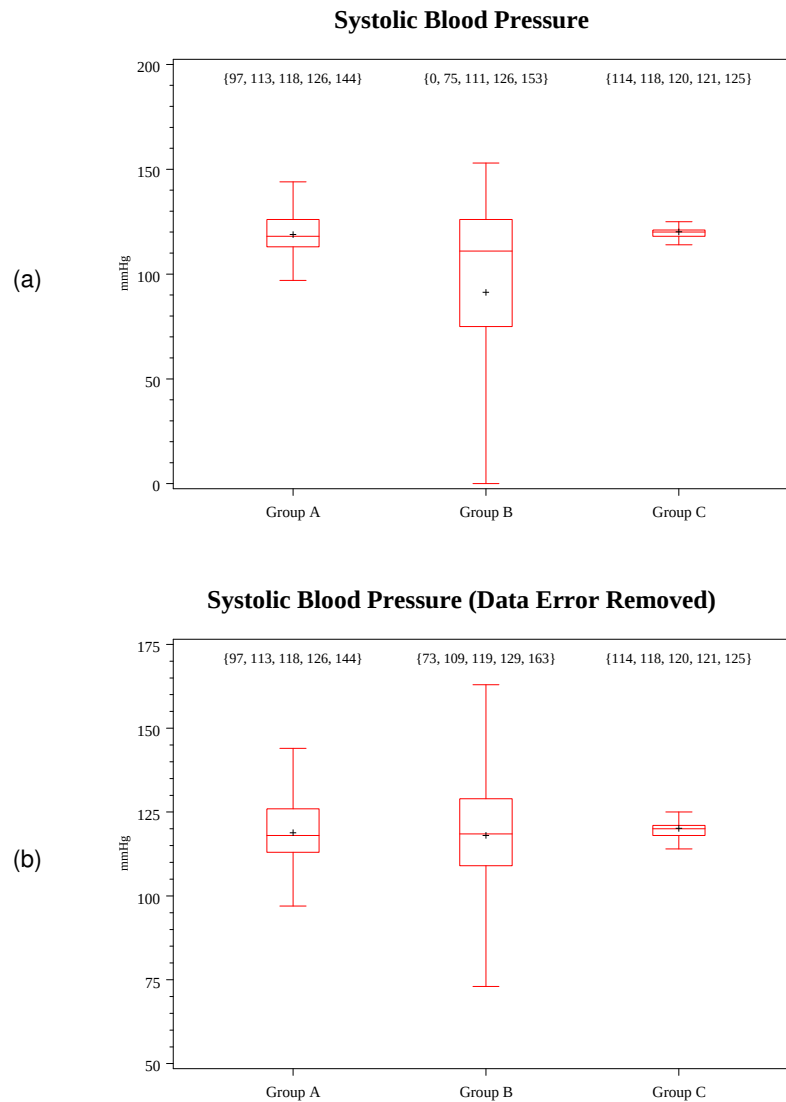


Figure 7: Box plots and five-number summaries of of systolic blood pressure measurements from three groups of patients, with (a) and without (b) a data error for group B.

Before we go any further, we may want to try the `boxstyle=schematic` option with our boxplot:

```
PROC BOXPLOT data=data456a;
  PLOT bp*group / vaxis=axis1 haxis=axis2 annotate=anno456a boxstyle=schematic;
RUN;
```

This changes the output slightly. As noted earlier, the interquartile range (IQR) is defined as the 75th percentile minus the 25th percentile. In a box plot, this is represented as the length of the “box”. All data points either

- below the 25th percentile - $1.5 \times \text{IQR}$, or
- above the 75th percentile + $1.5 \times \text{IQR}$

are designated as *outliers*.¹¹ When the `boxstyle=schematic` option is invoked in `PROC BOXPLOT`, all outliers are shown as isolated points, and the “whiskers” of the box plot represent the minimum and maximum values within $1.5 \times \text{IQR}$ of the 25th and 75th percentiles, respectively. Applied to our current example, this gives us Figure 8(b), where we see that we have five outliers with values above 180 mmHg (and a sixth with a value just below 180 mmHg). This is an improvement over Figure 8(a), since we can clearly see that the right skewness of Group D is being driven by these six outliers (There is an outlier in group D as well, but that is less important).

¹¹Mathematically, there are various definitions of an outlier; this is the one chosen for the `boxstyle=schematic` option of `PROC BOXPLOT`. For most purposes, this is sufficient.

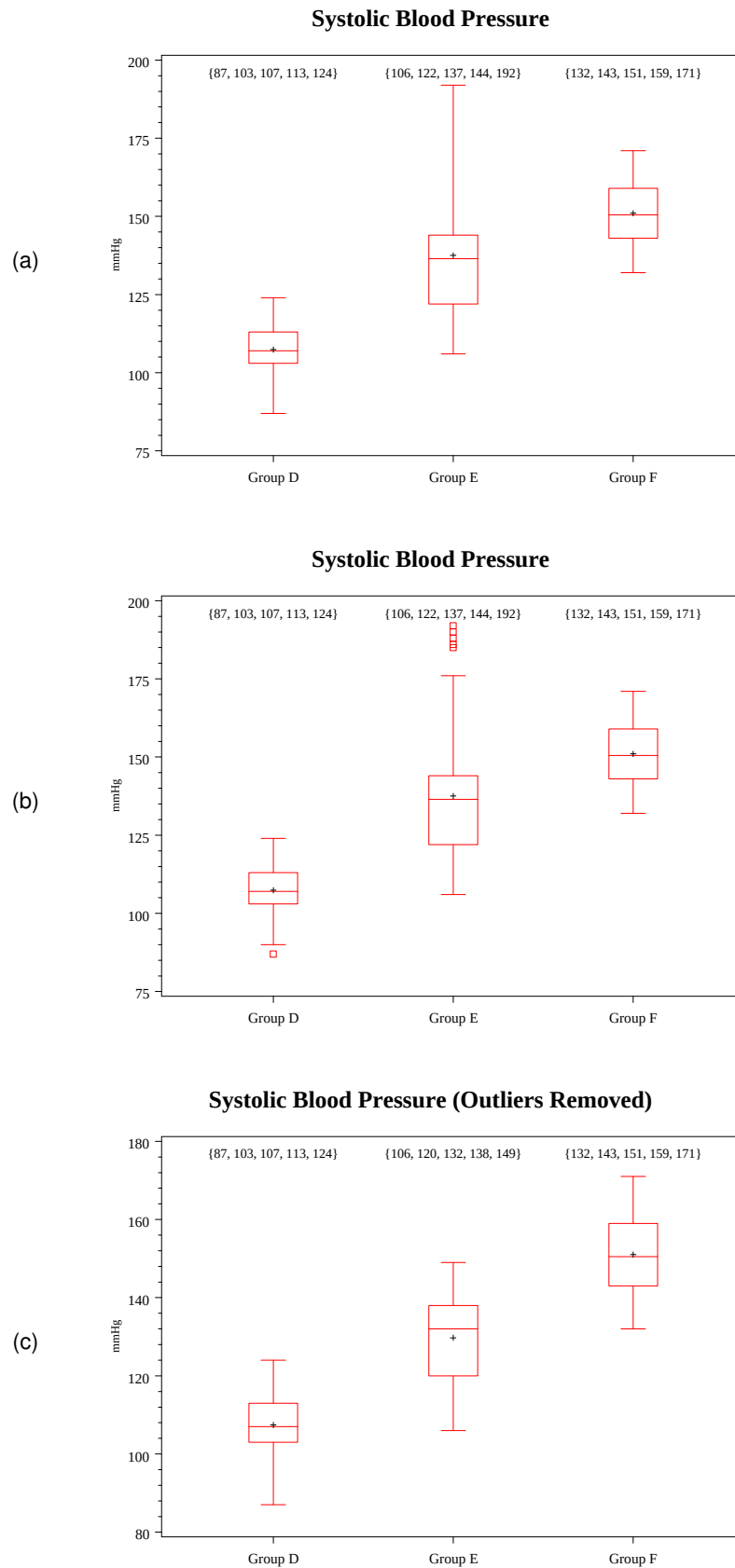


Figure 8: Box plots and five-number summaries of of systolic blood pressure measurements from three groups of patients, with (a, b [with the `boxstyle=schematic` option]), and without (c) outliers for group E.

Now that we have identified the outliers, we once again ask if this makes sense. Are these outliers errors, or legitimate data points? If they are legitimate, what is the explanation behind them (i.e., why are they outliers?). Let's assume that they are indeed legitimate data points. For example, perhaps this group of people includes six who have extremely high blood pressure and are at extreme risk of a heart attack. The next question is what to do with these outliers. Do we remove them from our analysis or keep them in there?

For guidance on this question, we can run the analysis without these outliers, producing Figure 8(c). Here we see that the distribution for Group E is much different. Indeed, our data have a much smaller range, and it is now left-skewed rather than right-skewed. This illustrates a general tendency: By definition, outliers are a small number of values that are (far) outside the range for all the rest of the data points. As such, including them in the data analysis can distort the general tendencies we are looking for in the data. On the other hand, if they are legitimate data points, they are part of the distribution and should not be ignored!

There are merits to both sides of the argument of whether or not to include them in the analysis, but it is common practice to remove them from the analysis, which we shall do in this example. However, whenever we do remove the outliers, **we must make sure to report them somewhere in the final analysis**. That is, we cannot simply discard them! In most cases, it is enough to mention it in a sentence or a footnote in the final report. But to either not report them or hide them (e.g., in the end notes or the appendix), is very deceptive! Be sure to always mention something about outliers that have been removed from the analysis.

Assuming our final analysis is shown in Figure 8(c), we can make some observations about the central tendencies and spreads:

- All three groups appear to have about the same spreads. That is, the overall ranges (maximum minus the minimum) and interquartile ranges (75th percentile minus the 25th percentile) are about the same for the three groups.
- These three groups have very different central values, at least compared with their spreads. That is, both the mean and median for group D are about equal to the minimum for group E, whose mean and median are about equal to the minimum for group F.

Again, the questions to ask is if these observations make sense. What could explain this behavior? Are the people in group D generally healthier than those in group E? Do they have different ages, or demographics? These questions are left for researchers and are outside the realm of this paper. But this illustrates the value of box plots, and of exploratory data analysis in general: They help us find the right questions to ask about the data.

DATA SUMMARIES: DEMYSTIFYING PROC UNIVARIATE

Up to now, we have been deliberately ignoring `PROC UNIVARIATE`. There is a reason for this: **`PROC UNIVARIATE` has much, much more information than we would usually need** for any one data summary. As an illustration, let's look at Figure 9, which is the `PROC UNIVARIATE` output when applied to our first data set (as shown in Figure 1). The purpose of a data summary (like the five-number summary) is to describe a large set of data with just a few numbers. With this goal in mind, note that for this example,

`PROC UNIVARIATE` is using 46 numbers to summarize 50 data points.

As such, it would appear that `PROC UNIVARIATE` is not an effective data summary. Furthermore, this would be true even if we were analyzing a data set with 500 or 5000 values; indeed, a 46-number data summary is not very effective.

However, `PROC UNIVARIATE` was not meant to be a data summary *per se*. Rather, it was designed to give (nearly) all possible statistics for a data summary that a user would ever want. Keep in mind, however, that just because a statistic is listed on `PROC UNIVARIATE` does not mean that we need to use it! Indeed, some of the output such as kurtosis, standard error (of the) mean, or the signed rank test are rather esoteric measures which are of interest only to very few data analysts.

Nonetheless, for completeness, we will detail all parts of the output here, starting from the bottom:

- `Extreme Observations` are simply the five lowest and highest observations, and their observation numbers (i.e., where they are in the data set).
- `Quantiles` are just percentiles (the words are synonyms). Thus, the 75th percentile is 126, meaning that 75% of the blood pressure measurements are below the value of 126 mmHg. However, there are some different ways to define a percentile value. The output (`Definition 5`) means that we define the percentiles just as we did the median on page 6. That is, we order the data points and calculate the percentile rank. For instance, for 50 data points, the 25th percentile rank is $0.25 \times 50 = 12.5$.
 - If the percentile rank is a whole number j , then the percentile is the number in the j^{th} position when the values are in sequence. For instance, with 50 data points, 50% of 50 is 25, so the percentile is the 25th number in the ordered sequence of all data points.

Moments			
N	50	Sum Weights	50
Mean	118.84	Sum Observations	5942
Std Deviation	10.14861	Variance	102.994286
Skewness	0.19322593	Kurtosis	0.26180882
Uncorrected SS	711194	Corrected SS	5046.72
Coeff Variation	8.53972571	Std Error Mean	1.4352302

Basic Statistical Measures			
Location		Variability	
Mean	118.8400	Std Deviation	10.14861
Median	118.0000	Variance	102.99429
Mode	115.0000	Range	47.00000
		Interquartile Range	13.00000

NOTE: The mode displayed is the smallest of 2 modes with a count of 4.

Tests for Location: Mu0=0			
Test	-Statistic-	----p Value-----	
Student's t	t 82.80205	Pr > t	<.0001
Sign	M 25	Pr >= M	<.0001
Signed Rank	S 637.5	Pr >= S	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	144.0
99%	144.0
95%	136.0
90%	131.0
75% Q3	126.0
50% Median	118.0
25% Q1	113.0
10%	106.5
5%	103.0
1%	97.0
0% Min	97.0

Extreme Observations			
----Lowest----		----Highest---	
Value	Obs	Value	Obs
97	32	131	45
98	42	132	7
103	25	136	27
103	12	143	22
106	3	144	33

Figure 9: PROC UNIVARIATE output for the blood pressure values of the original data set.

- If the percentile rank is a fractional number $j + g$, where j is a whole number and g is a fractional number, then the percentile is the average of the numbers in the j^{th} and $(j + 1)^{\text{st}}$ positions when the values are in sequence. For instance, with 50 data points, 25% of 50 is 12.5, so the percentile is $0.5 \times (x_{12} + x_{13})$, where x_j is the j^{th} number in the ordered sequence of all data points.

There are other definitions that also make sense. See the SAS help file “The UNIVARIATE Procedure: Calculating Percentiles” for more information. For general exploratory data analysis, however, the percentile definition is not of major importance.

From here, we then go to the top of the output:

- **N** is the total number of data points in our data set.
- **Sum Weights** gives the total sum of all our data point weights. This is only important when we weight some data points more than others; otherwise, it is always equal to the total number of data points, **N**.
- **Mean** is our sample mean, as we have seen before.
- **Sum Observations** is simply the sum of the variable observations in question over all the data points. This is sometimes used for further calculations (such as the mean, which is equal to $\frac{\text{Sum Observations}}{N}$) and can be used to quickly check some other calculations from PROC UNIVARIATE.

$$\text{Sum of Observations} = \sum_{i=1}^N x_i, \quad x_i = i^{\text{th}} \text{ observed value.}$$

- `Std Deviation` is our sample *standard deviation*, as we briefly mentioned before. This gives us an estimate of our average deviation from the sample mean, and is the square root of the *sample variance*:

$$\text{Standard Deviation} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}, \quad x_i = i^{\text{th}} \text{ observed value, } \bar{x} = \text{sample mean.}$$

- `Variance` is our sample *variance*, which gives an estimate of our average squared distance from the mean:¹²

$$\text{Variance} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2, \quad x_i = i^{\text{th}} \text{ observed value, } \bar{x} = \text{sample mean.}$$

The distance is squared so that we equally count positive and negative distances (i.e., whether a given value is above or below the mean).

- `Skewness` gives us a numerical estimate of how *skewed* our data is. As shown in Figure 3, a negative skew is a left skew, while a positive skew is a right skew. The absolute value gives the degree to which a sample distribution is skewed. A skewness of zero means the sample data is symmetric, so that the mean equals the median.
- `Kurtosis` gives us a numerical estimate of how “peaked” our data distribution is at its mode (if the distribution is unimodal). A higher value indicates a higher peak and thinner tails, while a lower value indicates a lower peak and fatter tails.
- `Uncorrected SS` gives us an *uncorrected sum of squares*, which is the quantity

$$\text{Uncorrected Sum of Squares} = \sum_{i=1}^N x_i^2, \quad x_i = i^{\text{th}} \text{ observed value.}$$

It is simply an intermediate calculation which may be of interest.

- `Corrected SS` gives us an *uncorrected sum of squares*, which is the quantity

$$\text{Corrected Sum of Squares} = \sum_{i=1}^N (x_i - \bar{x})^2, \quad x_i = i^{\text{th}} \text{ observed value, } \bar{x} = \text{sample mean.}$$

Again, it is an intermediate calculation.

- `Coeff Variation` is the *coefficient of variation*, which is simply the ratio of $100 \times$ the sample standard deviation to the standard mean:

$$\text{Coefficient of Variation} = \frac{100 \times \text{Std Deviation}}{\text{Mean}}.$$

It is used as a scaled version of the spread, which is sometimes useful (i.e., a spread of 100 is large if the mean is 1, but very small if the mean is 10,000).

- `Std Error Mean` is the *standard error of the mean* and is equal to

$$\text{Standard Error of the Mean} = \frac{\text{Std Deviation}}{\sqrt{N}}.$$

This is the estimated standard deviation of the distribution for the actual value of the mean (i.e., not the estimated value). This will be explained more below, in the discussion of hypothesis tests.

All the quantities under the heading `Basic Statistical Measures`, (`Mean`, `Median`, `Mode`, `Std Deviation`, `Variance`, `Range` and `Interquartile Range`) have been explained either above or in earlier sections. Note, however, the note in our output in Figure 9 that there were two modes, each with a count of four. This simply means that there were two values that each had four observations. If there were one value with three observations, that would be the new (unique) mode.

This leaves the section entitled `Tests for Location: Mu0=0`. This will be explained in the next section.

Lastly, be aware that SAS procedures like `PROC UNIVARIATE` often give more decimal places than are needed, which can sometimes give a misleading level of accuracy. Here, `PROC UNIVARIATE` states that the mean is 118.84 mmHg, but that number comes from only 50 data points, so this isn't actually accurate to two decimal places. That is, 118.84 gives an exact mean of this particular group of 50 patients, but if these 50 patients are a sample from a population of 5000 patients, it would be misleading to state that 118.84 is an estimate of the mean of the population of 5000 patients. In a case like this, where we have taken a very small sample (1% of the total data in this case), stating that the mean is 118 usually suffices; adding the decimal places would give a misleading sense of accuracy.

Overall, we see by inspection that all the other data in our `PROC UNIVARIATE` output in Figure 9 basically matches our results from the scatterplots, bubble plot, histograms and box plots, as they should.

¹²The sum of squared distances is divided by the quantity $N-1$ rather than by N to account for mathematically expected bias that would otherwise result.

HYPOTHESIS TESTS AND STATISTICAL SIGNIFICANCE

As briefly mentioned above, all the statistical techniques described up to now have computed *sample* quantities. For instance, the sample mean is the mean of our sample of 50 observations. The assumption is that our data set of 50 observations comprises a small sample of a (possibly infinite) number of observations that we are not observing. However, we assume that our sample is representative of the population of all possible observations (i.e., we don't have a biased sample), so that our inferences from our sample can be applied to the population.

Using our example, we assume that each data set of interest rates is a sample from interest rates of all loans in a given geographical area.

A *hypothesis test* is a test on our data of whether a theoretical (unobservable) quantity (such as the mean of the underlying data) is significantly different from some other value. In Figure 9, our `PROC UNIVARIATE` tells us that our sample mean (i.e., the mean from our sample of 50 observations) is 118.84. But what we often *really* care about is the theoretical mean, not the sample one.

For many uses, a standard question to ask is: Is the (unobservable) theoretical value significantly different from zero?

By *significant*, we mean *statistically significant*, meaning that we account for the spread of the data. Generally, the larger the spread (i.e., the more volatile the data), the less reliable our estimates will be, including our estimate of the sample mean. Therefore, determining whether our sample data point is significantly different than zero takes into account both our sample estimates of the mean (is the sample mean far from zero?) and the standard deviation (is the spread small enough to count out zero?).

This is a hypothesis test, and this particular one is so common that sample output from it is included in `PROC UNIVARIATE`.

There are many, many kinds of hypothesis tests, but `PROC UNIVARIATE` lists outcomes from three of them. Each of them calculates a *test statistic* for our hypothesis that the mean value is equal to zero:

- Student's *t* is the *student's t-test*, or simply the *t-test*. It calculates the quantity

$$\text{Test statistic} = \frac{\bar{x}}{SE(\bar{x})} = \frac{\text{Mean}}{\text{Std Error Mean}}$$

and matches it against the student's *t* distribution.

- *Sign* is the *sign test* to test whether the median is significantly different from zero. The test statistic is the average of number n^+ of values greater than zero and the number n^- of values less than zero:

$$\text{Test Statistic} = \frac{n^+ + n^-}{2}.$$

- *Signed Rank* is the *Wilcoxon signed rank test*, where the test statistic is a complex calculation derived from ranks of the values.

Which of these tests to use depends on which mathematical assumptions can be considered valid from the data. For more information on any of these tests or their underlying assumptions, see the SAS help file "The `UNIVARIATE` Procedure: Tests for Location," or Kanji (1999, pp. 17,78,80). For convenience, `PROC UNIVARIATE` lists all three of them, so that the user can glance at the results without first assessing the mathematical assumptions in question.

For each of these tests, the test statistic is matched against a certain theoretical distribution, and a nonzero *p-value* is computed. This value gives the probability that the estimated quantity is equal to zero. We can reject this hypothesis (and thus conclude that the theoretical mean is nonzero) if this *p-value* is smaller than a given number (usually 0.05).

In the `PROC UNIVARIATE` output shown in Figure 9, the `Tests for Location: Mu0=0` section gives a table of test statistics and *p-values* for each of these three tests described above. For our data, the *p-values* are all very small (less than 0.0001), and thus indicate that our mean values are significantly different from zero (which is to be expected, since every single data value was positive).

CONCLUSIONS

This paper presents some of the most commonly used tools of exploratory data analysis. These methods give us data visualization and summarization techniques which are simple, yet very effective for quickly estimating the central tendency, spread, and other characteristics of the data distribution. These methods help us detect data irregularities which might be errors, outliers, or some interesting aspect of the data (depending on the context). Using these methods can also help us compare two or more different data sets and assess their differences. Lastly, these methods give us indicators of what should be further analyzed with more complex statistical methods.

For more information about any of the statistical ideas in this paper, good references are Gonick and Smith (1993) and Siegel and Morgan (1996). For more information about using statistical techniques in SAS, see UCLA (2009).

REFERENCES

- Adams, R. (2008), Box plots in SAS: UNIVARIATE, BOXPLOT, or GPLOT?, *Proceedings of the Twenty-First Northeast SAS Users Group Conference*.
<http://nesug.org/proceedings/nesug08/np/np16.pdf>
- Cabrera, J. and McDougall, A. (2002), *Statistical Consulting*, Springer-Verlag, New York.
- Carpenter, A. L. (2006), Data driven annotations: An introduction to SAS/GRAPH's annotate facility, *Proceedings of the Thirty-First SAS Users Group International Conference*, paper 108-31.
<http://www2.sas.com/proceedings/sugi31/108-31.pdf>
- Gonick, L. and Smith, W. (1993), *The Cartoon Guide to Statistics*, HarperCollins Publishers, New York.
- Kanji, G. K. (1999), *100 Statistical Tests*, Sage Publications, London.
- Kucera, F. E. (1996), More informative scatter plots – adding a third dimension with bubbles, *Proceedings of the Twenty-First SAS Users Group International Conference*, paper 069a-21.
<http://www.lexjansen.com/sugi/sugi21/cc/069a-21.pdf>
- Siegel, A. F. and Morgan, C. J. (1996), *Statistics and Data Analysis: An Introduction*, second edn, John Wiley and Sons, Inc., New York.
- Stamatis, D. H. (2003), *Six Sigma and Beyond: Statistical Process Control, Volume IV*, CRC Press, Boca Raton, FL.
- UCLA (2009), Resources to help you learn and use SAS, Academic Technology Services: Statistical Consulting Group.
<http://www.ats.ucla.edu/stat/sas/>
- Watts, P. (2008), Using SAS software to generate textbook style histograms, *Proceedings of the Twenty-First Northeast SAS Users Group Conference*.
<http://nesug.org/proceedings/nesug08/np/np03.pdf>

ACKNOWLEDGMENTS

I thank Lisa Eckler for originally giving me the idea for the first version of this paper and inviting me to present it at the 2009 SAS Global Forum. I further thank MaryAnne Hope, Gwendolyn Brophy and W. Wilson Will for helping me expand upon those ideas and change the focus from a financial to a medical application for this second version of this paper.

I furthermore thank my former professors and employers for making me realize that basic statistical ideas are far from trivial. Lastly, and most importantly, I thank Charles for his patience and support.

CONTACT INFORMATION

Comments and questions are valued and encouraged. Contact the author:

Nathaniel Derby
Statis Pro Data Analytics
815 First Ave., Suite 287
Seattle, WA 98104-1404
206-973-2403
nderby@sprodata.com
<http://nderby.org>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.