SEGMENT 1: Exploratory data analysis

Introduction

Statistics is the science of collecting, analyzing, displaying and interpreting data. The first question we should ask is: *Why learn statistics?*

The NY Times published an article in 2009 titled " For Today's Graduate, Just One Word: Statistics" on the increasing need for statistics and statisticians. Read it at http://www.nytimes.com/2009/08/06/technology/06stats.html.

Hal Varian - Chief Economist at Google – said that "*the sexy job in the next 10 years will be statisticians".* Watch this short video of his Keynote Presentation at the 2008 Almaden Institute "Innovating with Information": http://www.youtube.com/watch?v=D4FQsYTbLoI

Top reasons to be a statistician

- Estimating parameters is easier than dealing with real life.
- Statisticians are significant
- I always wanted to learn the entire Greek alphabet.
- The probability a statistician major will get a job is > .9999.
- If I flunk out I can always transfer to Engineering.
- You never have to be right - only close.
- We're normal and everyone else is skewed.
- The regression line looks better than the unemployment line.
- No one knows what we do so we are always right

A fairly recent Wall Street Journal article (August 4th 2011) reports that "because of an increasing stream of data from the Web and other electronic sources, many companies are seeking professionals who can make sense of the numbers through the growing practice of data analytics…As the use of analytics grows quickly, companies will need employees who understand the data. A study from McKinsey & Co. found that by 2018, the U.S. will face a shortage of 1.5 million professionals who can use data to shape business decisions."

We are confronted with quantitative information and statistics daily. By knowing how to analyze data and understand statistics, you can be a *better decision maker*. Data analysis has become a critical tool to support good business decisions and improve performance. This is a principle that is adopted by many successful companies. Wal-Mart, for instance, analyzes purchase data to learn about their customers' needs and predict future sale trends. Results of data analysis are used for inventory management (Rahul Asthana, 2006)[1].

Data analysis is not only important for business applications, but is also used in a variety of fields, in network applications statistical methods are used to detect malicious attacks and unauthorized uses of computer accounts by analyzing unusual patterns in the server log data, web logs are analyzed to learn about customer behavior, computer usage data are analyzed to learn about users' needs. All medical

[1] Rahul Asthana, "Crossing the Analytics Chasm" in *Business Intelligence Journal,* Vol. 11, n.1, 2006 http://www.tdwi.org/Publications/BIJournal/display.aspx?ID=7892

discoveries are based on experimental studies – statistical methods are used to design and analyze the experiments and to test the researcher's hypotheses.

QUESTION: Can you think of a recent study in the news that reported some statistics? What was the field of the study? What statistics the study reported? What were the study conclusions?

As cited in the *NY Times* article above, "We're rapidly entering a world where everything can be monitored and measured," said Erik Brynjolfsson, an economist and director of the Massachusetts Institute of Technology's Center for Digital Business. "But the big problem is going to be the ability of humans to use, analyze and make sense of the data."

In this module you will explore the initial steps of a correct data analysis. These are the steps that will help us "make sense of the data":

1) Examine the dataset or spreadsheet and identify the variables in the dataset
2) Create simple displays of the data to visualize the overall pattern and deviations from the pattern
3) Compute summary statistics to measure center and spread of the data, and possible deviations from the pattern.

STEP1: Identify variables and cases in dataset

Data contain information about cases or individuals that are considered in the study.
Example: Purchase data contain information about store customers that are regarded as the "individuals" in the study. Network data contain information about logins to a certain server.

Data are organized in variables. A **variable** is a characteristic or an attribute that takes different values for the population of interest.
A **quantitative** variable (continuous) takes numerical values.

- ❖ Ex.: Height, Weight, Age, Income, Measurements

A **qualitative/categorical** variable classifies an individual into categories or groups.

- ❖ Ex. : Sex, Religion, Occupation, Age (in classes e.g. 10-20, 20-30, 30-40)

When we analyze a new dataset, we should ask the following questions:

- ❖ What was the purpose of the study?
- ❖ Who or what was observed? How many cases are in the data?
- ❖ What and how many variables are observed in the study? How were the variables measured?

Before we start any analysis, it is important to understand the variables that were collected. It is always good practice to create a data dictionary that contains variable names, their definitions and units of measurement.

Example: The following table shows a portion of a survey about Facebook:

Obs.	**Hours online/ week**	**Log-ins/Day**	**Friends**	**Sex**	**Age**	**Profession**
1	5	3	45	Female	30	Teacher

2	7	4	140	Male	21	Student
3	3	2	35	Female	25	Student
4	4	3	68	Female	23	Student
5	5	2	83	Male	27	Sales

What are the variables in the dataset?
The Facebook dataset has 6 variables: Weekly Hours, Log-Ins per day, Number of friends and Age are quantitative; Sex (Male, Female) and Profession (Student, Teacher, Sales, Management, Other) are qualitative.

STEP2: Display distributions with graphs
The **distribution** of a variable tells us what values it takes and how often it takes those values. Distributions vary depending on the variable type. The distribution of a categorical or qualitative variable lists the count or percentage of observed cases in each category. The distribution of a quantitative variable represents the percentage of observations that fall specified intervals.

Bar graphs or pie charts are used to represent distributions of qualitative variables. In a bar chart, the heights of the bars represent the percentage of cases in each category. Pie charts help us visualize each observed category in relation to the whole. It can only be used if all categories can be included.

Example: The graphs below compare the age distribution of WiFi Internet users observed during a study on Internet users. Notice that in this example, AGE is stored as a categorical variable with three categories: 18-34 years old, 35-54 years old and 55 years old or older.

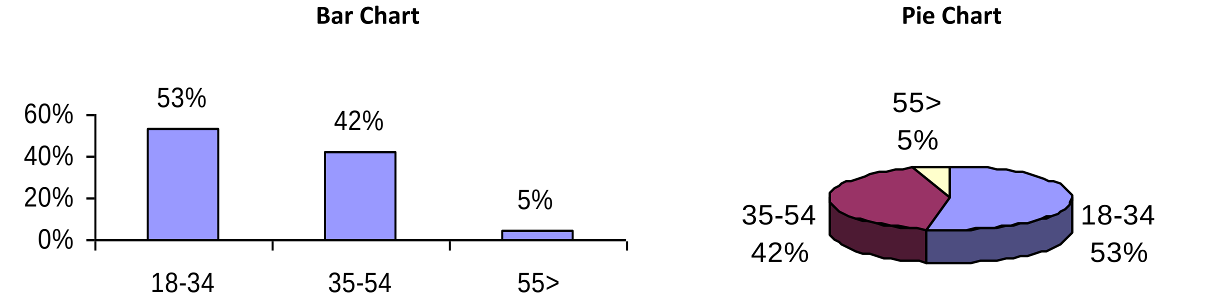


Figure 1: Age distribution of WiFi Internet users

The distribution of a **quantitative variable** is displayed using a **histogram**. A histogram breaks the variable values into classes or intervals and displays the percentage of observed cases that fall into each interval.

*Example: (*Table 1.10 page 31 in IPS, 6th ed.) The histograms in Figure 2 show the distributions for city and highway gas mileage for "two-seater" or "mini-compact" cars as reported in 2004 by the Environmental Protection Agency. The data are reported in Table 1.10 page 31 in the 6th edition of the course textbook.

The two histograms show the overall distribution of the observed gas mileage values among the 34 sampled vehicles. The graphs also show the presence of an extreme value. A car had gas mileage close to 60 Mpg in the city. This value corresponds to the hybrid Honda Insight, and can be considered an **outlier** since it is so strikingly different from the other values.

Histograms can also be used to compare related distributions. Just be careful – plots must have X-axes on the same scale. This example shows that gas mileage is higher if car is driven on the highway.

How do we build a histogram?

1. Divide range of data into intervals.
2. Count how many values fall into each interval.
3. Draw bar over each interval with height = count (or proportion) of observations in each interval.

Statistical software computes the steps above and creates histograms. For instance, in SPSS histograms are created under "Graphs > Legacy Dialogs > Histogram…" – a video tutorial showing you how to create histograms in SPSS is posted on the course website. Excel has a Histogram function in the Data Analysis toolbox.

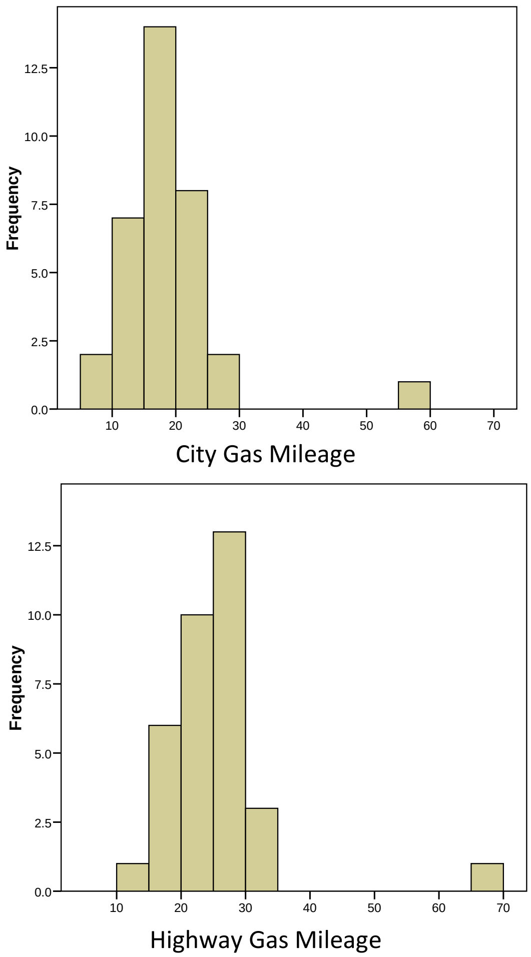


Figure 2: 2004 Gas mileage for two-seater and mini-compact cars

Properties of a histogram:

- A histogram represents percent by area. The area of each block represents the percentages of cases in the corresponding class interval.
- The total area under a histogram is 100%
- There is no fixed choice for the number of classes in a histogram. Typically statistical software will choose the class intervals for you, but you can modify them.
  - If class intervals are too small, the histogram will have spikes;
  - If class intervals are too large, some information will be missed.

Remarks about statistical graphics:

Good statistical graphs should display and communicate information with clarity and precision. Graph axes should be properly labeled, and scales should be chosen to display the actual pattern. One common mistake is to use limited axis ranges that make small differences look bigger. For instance, the two graphs in Figure 1 show the distribution of company earnings from 1998 to 2006 in thousand dollars. At first glance, the first graph seems to suggest that the company earnings had significant increase after 2003.

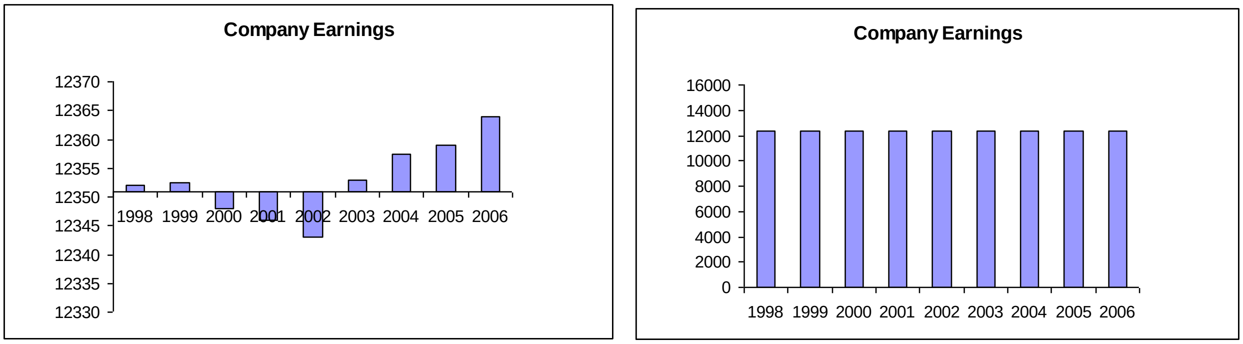


Figure 3: Distribution of company earnings

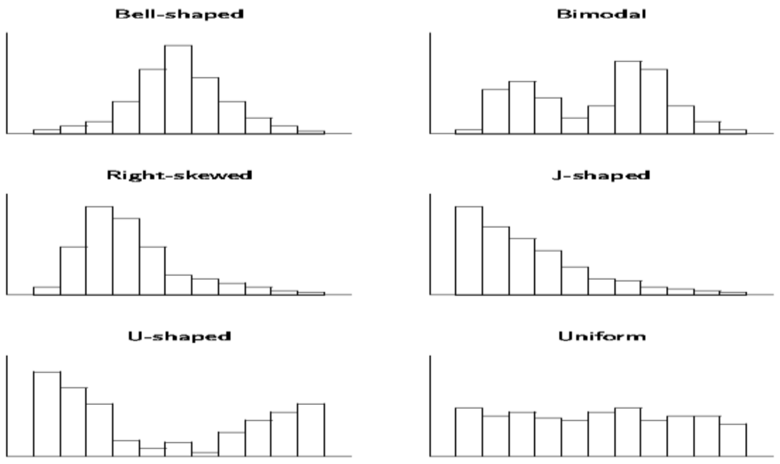
However, this would be a lie. The Y-axis has a very small range. Actual differences in earnings between years are quite small, less than $20,000 dollars. The second graph reports results on a full scale and shows that company earnings do not vary significantly across the years.

Many cases of bad graphical displays can be found at http://www.math.yorku.ca/SCS/Gallery/ - check the "thumb-down" links.

Examining distributions

In any graph of data, we should look for the overall pattern and for striking deviations from that pattern. Distributions can be categorized according to their shape:

- **Symmetric:** if we draw a line through center, the picture on one side would be mirror image of the picture on other side. *Example*: bell-shaped data set.
- **Unimodal:** single prominent peak
- **Bimodal:** two prominent peaks
- **Right-Skewed:** *higher* values more spread out than lower values – the right tail (larger values) is much longer than the left tail (lower values)
- **Left-Skewed:** *lower* values more spread out and higher ones tend to be clumped - the left tail (lower values) is much longer than the right tail (larger values)



STEP 3: Describe distributions with numbers

In the previous section we discussed distributions and learned how to visualize them using histograms and bar charts. The next step is to compute a set of metrics describing the center and spread of the observed distribution.

- **Measures for center:**
    - **Mean** or Average (better for symmetric distributions)
    - **Median** is the value such that 50% of observations fall below it (better for skewed distributions).
- **Measures for spread:**
    - **Standard deviation** (To be used only in conjunction with the sample average)
    - **Quartiles** (Often used in conjunction with the median)

Measuring the center of a distribution

The *mean* of a set of numbers {x_1, x_2, …,x_n} is defined as $\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$. It should be used only for symmetric distributions, since its value is affected by the presence of outliers or extreme values. For instance, a large outlier pulls the average to the right (see example later in this section)

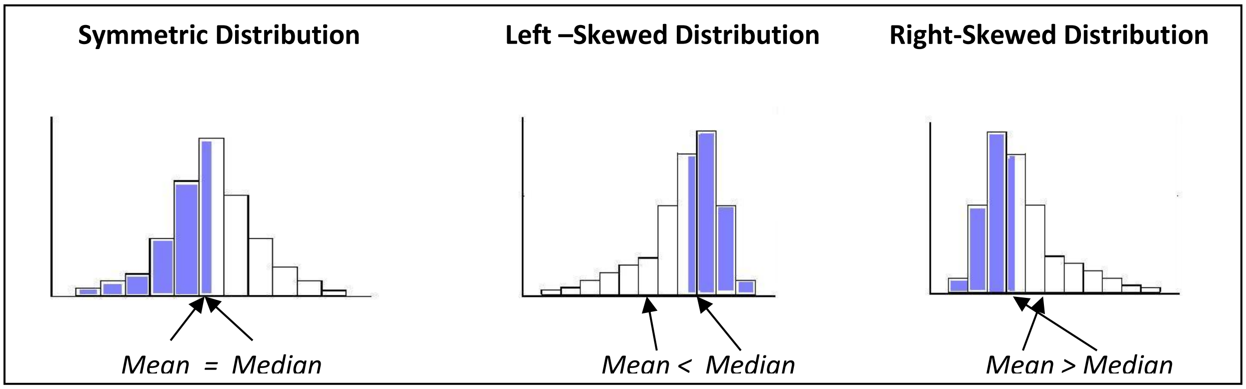
The *median* is the midpoint of a list of ordered numbers or in other words is the point for which 50% of the observations fall below it. The median is computed as follows:

1. Sort all the observations in order of size from smallest to largest
2. If the number of observations n is odd, the median M is the center observation in the ordered list; I.e. M=(n+1)/2-th obs.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list.

The median is resistant to outliers and its value won't change much if a few outliers are present.

Mean vs Median

For symmetric distributions, mean and median are close together. For skewed distributions the mean is pulled toward the longer tail since it is affected by fewer extreme values.



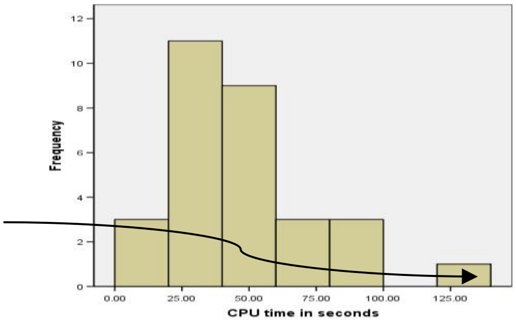
Thus the mean is larger than the median for right-skewed distributions, and smaller than the median for left-skewed distributions. (See SPSS example on descriptive statistics)

Example 1: To evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time for n=30 randomly chosen jobs (in sec.):

70 36 43 69 82 48 34 62 35 15 59 35 22 56 139
46 37 42 30 55 56 36 82 38 89 54 25 24 9 19

The *mean* CPU time of all jobs is (70 + 26+ 43 + … +9 +19)/30 = 48.233 sec.
To compute the median, we sort the CPU data in increasing order:
9 15 19 22 24 25 30 34 35 35 36 36 37 38 **42** **43** 46 48 54 55 56 56 59 62 69 60 82 82 89 139
The *median* is the list midpoint (42+43)/2 = 42.5 sec.
Question: Why are the mean and median so different?
The distribution of CPU times is right skewed and the mean is affected by the extreme value.

The CPU time dataset contains an outlier equal to 139 seconds. If we removed the outlier, the mean of the new dataset will be equal to 45.10 seconds and the median will be equal to 42 sec. Only the mean changes significantly from 48 sec. to 45.10 sec. The median is NOT affected by outliers.

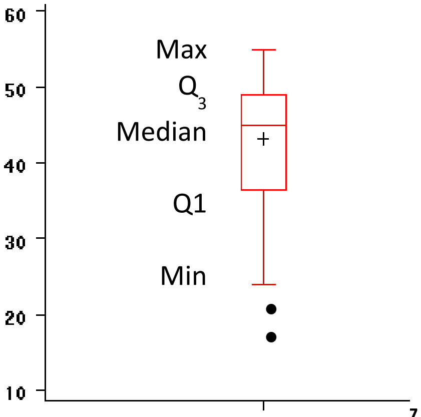


Remarks: The mean is a good measure for the center of a symmetric distribution. The median is a resistant measure (i.e. not affected by outliers) and should be used for skewed distributions. Its value is affected only slightly by the presence of extreme observations, no matter how large these observations are.

Measuring the spread of a distribution

Percentiles and Quartiles:
The *p-th percentile* of a distribution is the smallest value that is greater than *p* percent of the observations. For instance if your height falls in the 80th percentile, that means that you are taller than 80% of the people in your age/sex class.
Two commonly used percentiles are *the first and third quartiles*. The first quartile Q1 is equal to the 25th percentile, and is defined as the point that falls above 25% of the data. The third quartile Q3 is the 75th percentile and is defined as the point that falls above 75% of the data (or equivalently that falls below 25% of the data).
The *five-number summary* of a set of observations consists of the min, first quartile, median, third quartile and max. It is used to summarize the center and spread of skewed distributions.

The 1.5 IQR rule to detect outliers
The Interquartile range IQR is the distance between first and third quartiles: IQR = Q_3 - Q_1
An observation can be a possible outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile. Thus data values



larger than (Q_3 + 1.5 x IQR) or smaller than (Q_1 - 1.5 x IQR) are potential outliers.
Box plots are used to visualize the five-number summary on a graph. They are particularly effective in comparing distributions of sets of data. The central box spans the quartiles, the line in the box marks the median and the lines outside the box extend to the min and max excluding the outliers. Outliers flagged by the 1.5 IQR rule are marked by “dots” on the plot. The “plus” symbol denotes the average.

Standard Deviation

The standard deviation is a commonly used measure of spread and measures how far observations are from the average. It is defined as $s = \sqrt{\dfrac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}}$

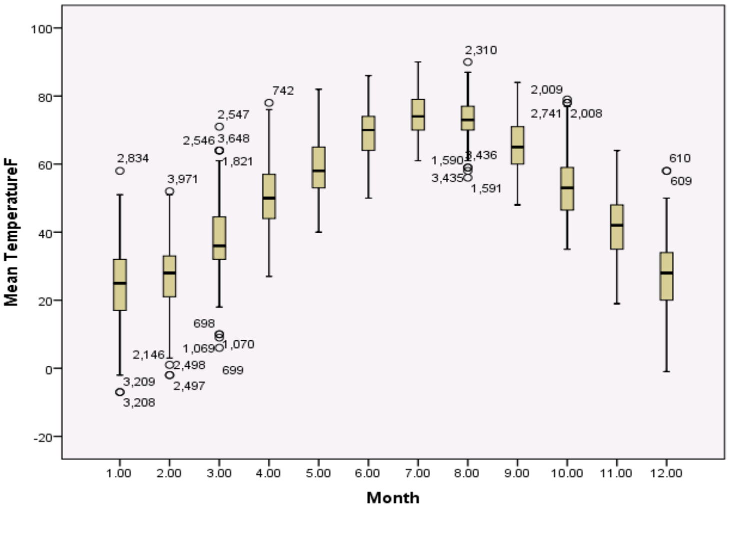
It is a meaningful measure only when it is used in conjunction with the mean.
Standard deviation is equal to zero only when all observations fall on the same value. It is measured on the same scale of the observations. It is not resistant to outliers. Therefore the presence of a few extreme values can make the standard deviation larger.

Comparing distributions of groups of observations

It is often interesting to compare groups of observations. Is it rainier in June or September? Are any months that are particularly rainy? Both histograms and boxplots can be used to compare the distributions of groups.

Boxplots are particularly effective to display groups of observations as they display groups of observations side-by-side. Although they hide some details, boxplots display overall summary information. So we can easily see which group has the largest median, or the largest IQR, etc…

This example shows the boxplots of daily mean temperatures recorded in Chicago from 2000 to 2011. The boxplots display the data by month. The graph was created using SPSS, the circles represent outliers identified using the 1.5 IQR rule and the numbers next to the circles denote the observation numbers. Thus for instance observation #3,208 is an outlier for the month of January.



The graph shows that the hottest month is July, and the coldest months are December, January and February. It is also interesting to note that the range of temperatures varies from month to month, with summer months having less variation in daily temperatures than winter months.

SPSS FUNCTIONS FOR THIS MODULE

Computing descriptive statistics in SPSS

Descriptive statistics are computed in SPSS by two functions listed under the “Analyze > Descriptive Statistics ” menu. The two functions are

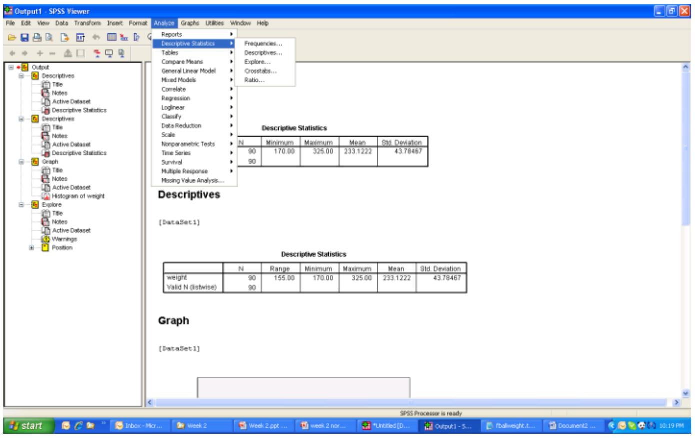
a. “Descriptive…” that computes only basic statistics, and

b. “Explore …” that computes several statistics and produces statistical graphs including boxplots and histograms.

I recommend using “Explore…” since it produces more useful statistics.

In summary:

1. Import data in SPSS and view them in the Data View table.
2. Select: **Analyze > Descriptive Statistics > Descriptive …** under the top menu to compute simple descriptive statistics (such as average, st.dev., min, max,range)
3. Select: **Analyze > Descriptive Statistics > Explore …** to compute a large number of descriptive statistics and to draw histograms and normal probability plots.



Descriptive Statistics

N	Minimum	Maximum	Mean	Std. Deviation
90	170.00	325.00	233.1222	43.78467
90				

Descriptives

```
[DataSet1]
```

Descriptive Statistics

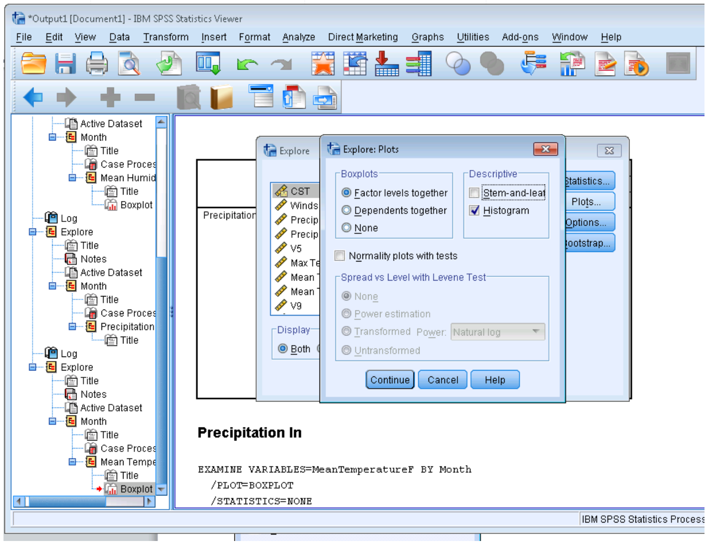
	N	Range	Minimum	Maximum	Mean	Std. Deviation
weight	90	155.00	170.00	325.00	233.1222	43.78467
Valid N (listwise)	90					

Graph

```
[DataSet1]
```

(see SPSS video tutorial on our course page or the SPSS video tutorial at http://www.as.ysu.edu/~chang/SPSS/SPSSmain.htm)

Creating histograms or boxplots in SPSS Histograms can be created in SPSS using the “Explore …” function found under **Analyze > Descriptive Statistics > Explore …** . Click on the “Plots” button and select “histogram” .



Explore: Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
□ Stem-and-leaf
☑ Histogram

□ Normality plots with tests

Spread vs Level with Levene Test
● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

Continue Cancel Help

Display
● Both

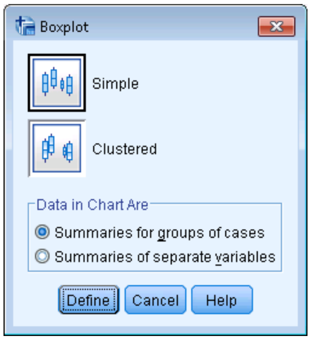
Precipitation In

```
EXAMINE VARIABLES=MeanTemperatureF BY Month
  /PLOT=BOXPLOT
  /STATISTICS=NONE
```

Alternatively, you can create a histogram using the function under **Graphs>Legacy Dialogs> Histogram.**

For boxplots use:

1) Select **Graphs>Legacy Dialogs> Boxplots**
2) Select "Simple" in the Boxplots window and click Define



Boxplot

Simple

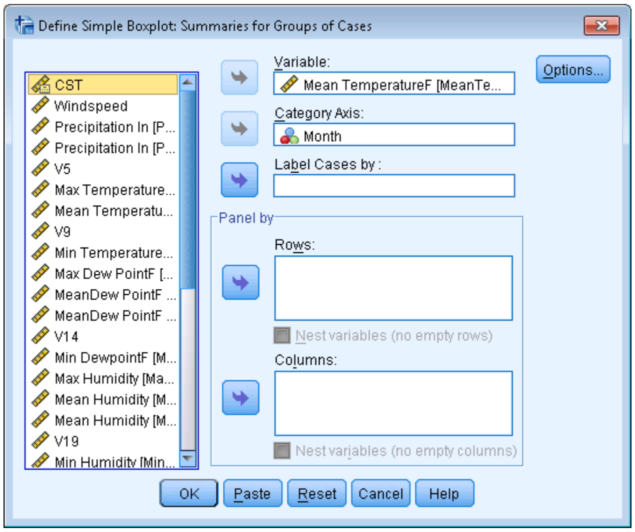
Clustered

Data in Chart Are

● Summaries for groups of cases

○ Summaries of separate variables

3) The variable containing the data for the boxplots should be moved into the "Variable" field and the grouping variable should be placed in the "Category Axis" field. Then select OK.



Define Simple Boxplot: Summaries for Groups of Cases

Variable: Mean TemperatureF [MeanTe...

Category Axis: Month

Label Cases by:

Panel by

Rows:

□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

See the SPSS video tutorial page for short videos on how to create histograms and boxplots.