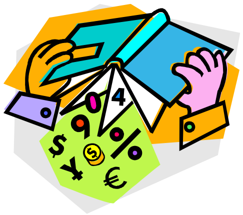
http://wiki.stat.ucla.edu/socr/index.php/SOCR_Courses_2008_Thomson_ECON261

DESCRIPTIVE STATISTICS

PART II DESCRIBING YOUR DATA USING NUMERICAL MEASURES



Grace S. Thomson

BUSINESS RESEARCH AND DESCRIPTIVE STATISTICS

PART II DESCRIBING YOUR DATA USING NUMERICAL MEASURES

This chapter contains 3 main topics related to available techniques to describe and interpret statistic data, using numerical measures of center, location, and variation.

1. Measures of Center and Location
2. Measures of Variation
3. Describing and comparing measures

Let me summarize what you will find in this chapter: Remember when you learned about nominal and ordinal data? Now we are going to use these concepts to understand what type of measurement is suitable to describe data.

Our first concept is the difference between a **parameter and a statistic.** When you are measuring data from the entire population, you are calculating a **parameter,** whereas when measuring data from a sample, you are calculating a **statistic** (Lind, 2005). It is important to keep these 2 concepts in mind all the time, because you will see them repeatedly through our class.

Types of Measurements

In statistics there are basically 2 types of measurements: a) measures of location and b) measures of variation

This book addresses seven measures of location and six measures of variation. At the end of the chapter you will learn how to integrate these measures in five indicators to reach conclusions about the data. Let's start summarizing them in the following tables:

Measures of location

1. Measure of Central Tendency

a. Population Mean

b. Population /Sample

c. Median

d. Mode

2. Other Measures of location

e. Weighted Mean → population/ sample

f. Percentiles

g. Quartiles

Measures of variation

1. Range
2. Interquartile Range
3. Population Variance
4. Sample Variance
5. Population Deviation
6. Sample Deviation

Mean and standard deviation combined

1. Coefficient of Variation for Population
2. Coefficient of Variation for Sample
3. Empirical rule
4. Tchebysheff's Theorem
5. Standardized Data Value

Measures of Location

Mean

Let’s say that you need to express the average annual income of your 1000 customers. That number is called the **mean,** and since you compute it from the totality of your customers you will be calculating a **population mean**. If you take only a sample you will be using a s**ample mean**.

Procedure: Very simple, divide the sum of the values by the number of values in the data. Let’s use this example to understand the concept. Following there is a record of revenues in dollars per month from a retail store.

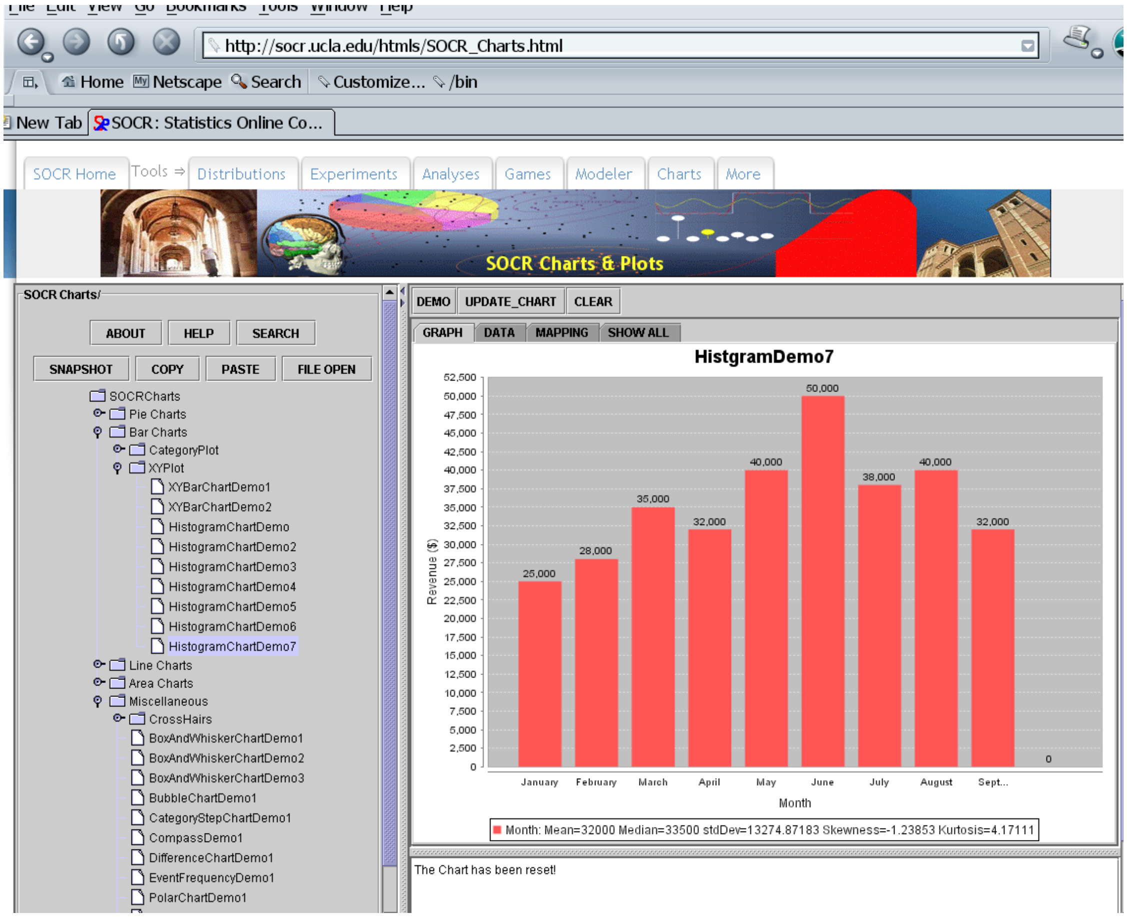
Table 1

Monthly Revenues for a Retail Store

January-September 2xx7

Months	**Revenue in dollars**
January	25,000
February	28,000
March	35,000
April	32,000
May	40,000
June	50,000
July	38,000
August	40,000
September	32,000
Total	320,000
n=	9
Mean=	320,000/9 = 35,556

Notice how we divide the total revenue for the nine months by the number of months. If this is all the data we have $35,556 is the mean or average of the population. The formula we use in statistics is:



🟥 Month: Mean=32000 Median=33500 stdDev=13274.87183 Skewness=-1.23853 Kurtosis=4.17111

The Chart has been reset!

$$\mu = \frac{\sum_{i=1}^{N} Xi}{N}$$

Population Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n}$$

Sample mean

Where Σ means sum, and the notation under and over it, means that the sum operates from the observation number one to the last observation (N). Xi represents all the observations i that our problem has. N represents the count of observations of our problem.

Notice I have cited two formulae: One is for the population mean and the other for the sample mean. The population mean̦ μ) formula has all capitalized characters while the sample mean $\bar{x}$ has all the characters in lower case.

So using our example, your population mean is $35,556 if we consider all the months in the list. But what if you choose a sample of the revenues of 3 random months? Let's say that you chose March, June and September to compute the average:

Table 1

Monthly Revenues for a Retail Store

March- September 2xx7

Months	**Revenue in dollars**
March	35,000
June	50,000
September	32,000
Total	117,000
n=	3
Mean=	117,000/3 = 39,000

Notice that the mean is now $39,000 because we chose 3 of the highest monthly revenue, by coincidence.

I have good news for you, using MSExcel makes it easy to compute the mean, as easy as 1-2-3. The formula to compute the mean is "=average(range)".

Median

However if what you are more interested in finding out is their mid-point income, you need to calculate the **MEDIAN.** It will give you the number for which at least half of the data are at least as large as the data value, and at least half of the data are as small as or smaller than that data value.

Procedure: Simply, arrange data in numerical order from smallest to largest (data array), and locate the value halfway from either end, that's your median. To locate this number divide the number of observations plus 1 by 2, like this (N+1)/2

Here a quick example: If you have a sample of 21 customers with their billing information and you want to know the median amount of billing in your portfolio, you will arrange all your customers from the lowest to the highest amount and then locate the client who is in the 11th position -since (N + 1) /2= 11. The amount of billing that this customer had is the median of your portfolio. If his billing amount is $50,000 in a year, $50,000 is the median, which means that 50% of your portfolio has billings above it and 50% of your portfolio has billings under it.

If your portfolio contained 20 customers, the median would be located between the 10th and 11th position since (20+1)/2= 10.5 and you would need to compute an average between those two middle numbers.

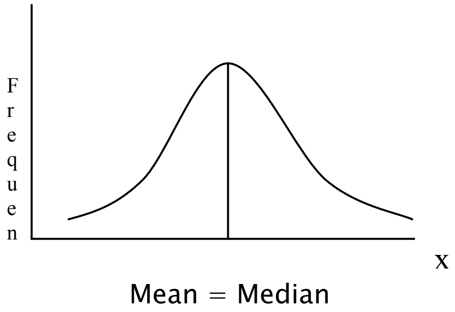
Mode

If you are interested in the most repeated annual income among your potential customers, that number is called the **MODE.**

Procedure: Lay out your information and identify the most frequent value in the list. That is the mode. Some data sets have two or more modes in which case it's said that the sample or population is multimodal; others have no repeated numbers, so no mode for that data set. Now, be careful because the mode is given by the repeated number, not by the repetitions. So if in your customer portfolio $70,000 is repeated 6 times, $70,000 is the mode of your portfolio, not 6; 6 is the indication of the number of customers who have amount.

Skewness and Symmetry

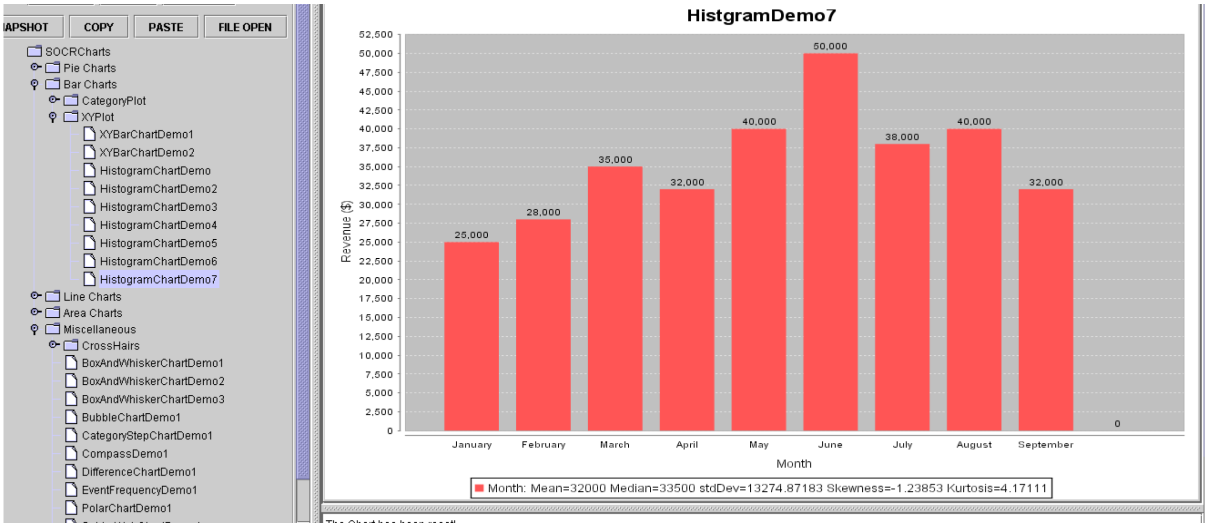
Now, let's take a look at other 2 important concepts –**Skewed and Symmetric distributions.** Data sets are **symmetric** when their values are evenly spread around the center, and to confirm this, median and mean must be equal. Take a look at the following curve:



When this doesn't happen, data might be **left-skewed distributed** →the mean is smaller than (to the left of) the median. Or **right-skewed distributed** → the mean is larger than (to the right of) the median.

Here is another concept to remember: The mean can be highly affected by extreme values. If one of the observations has very low or very high values, it affects the mean ma king it lower or higher, respectively.

Notice the Mean, Median, Skewness and Kurtosis measures on bottom of most SOCR Charts (http://socr.ucla.edu/htmls/SOCR_Charts.html)



🟥 Month: Mean=32000 Median=33500 stdDev=13274.87183 Skewness=-1.23853 Kurtosis=4.17111

Other measures of location

Weighted mean is a measure of location used when there is a relative importance of each value in the data. It's also called mean for grouped data.

Procedure: Collect the data and assign weights to each observation, multiply each weight by the data value and sum them. Sum the weights, too. Then divide the first sum by the sum of weights and you'll have the weighted mean:

You can compute weighted mean for populations and samples. We will go over this with an example in class.

$$\mu_w = \frac{\sum Xifi}{\sum fi}$$

Weighted Mean for a population

$$\overline{x}_w = \frac{\sum xifi}{\sum fi}$$

Weighted Mean for a sample

The difference between these 2 formulas is simply the source of information and the symbol for the weighted mean (μ) or X bar.

Percentiles

It's a measure of position expressed in percentage up to 100%. It divides the data in two segments: At p% a value is as large or larger than that p% and smaller than the remaining (100-p%) . e.g. If you are in the 90th percentile of your class, it means that your score is as high or higher than 90% of the class, and lower than 10% of the class. So, that's good, the higher the percentile, the better.

Procedure: Sort data from low to high, then assign a location indicator from 1 to **n** to each data value. Apply the formula for percentiles to locate the percentile you are interested in:

$$i = \frac{P}{100}(n+1)$$

P= desired value
n= number of values in data set

e.g. If there are 20 students in your class, and you are in the 90th percentile of the class based on the grades, by replacing the *90* in the formula you will find out that you are in position 18.90 However decimals don't make much sense for a location, so we need to interpolate to locate the exact position of your grade. If the grade in position 18 is 98, and the grade in position 19 is 98.5, the interpolation would result in a score of **98.45** (98 + 0.90*(98.5 – 98)], that's your positional score.

Quartiles

Works similarly to the percentile, with the difference the percentage divides the data set in four equal-sized groups. There is a relationship between quartiles and percentiles:

Table 3

Relationship between quartiles and percentiles

1st. quartile→	25th percentile
2nd. quartile→	50th percentile
	MEDIAN
3rd. quartile→	75th percentile
4th. Quartile→	100th percentile

Notice that the 2nd quartile, or 50th percentile is the same as the *median*.

Quartiles operate with the same formula for percentiles.

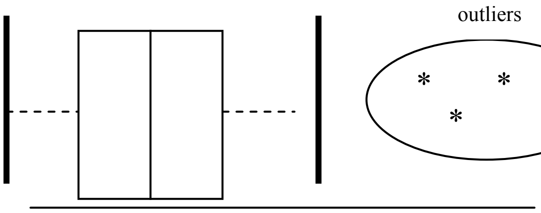
$$i = \frac{P}{100}(n+1)$$

P= desired value
n= number of values in data set

Box and Whiskers plot

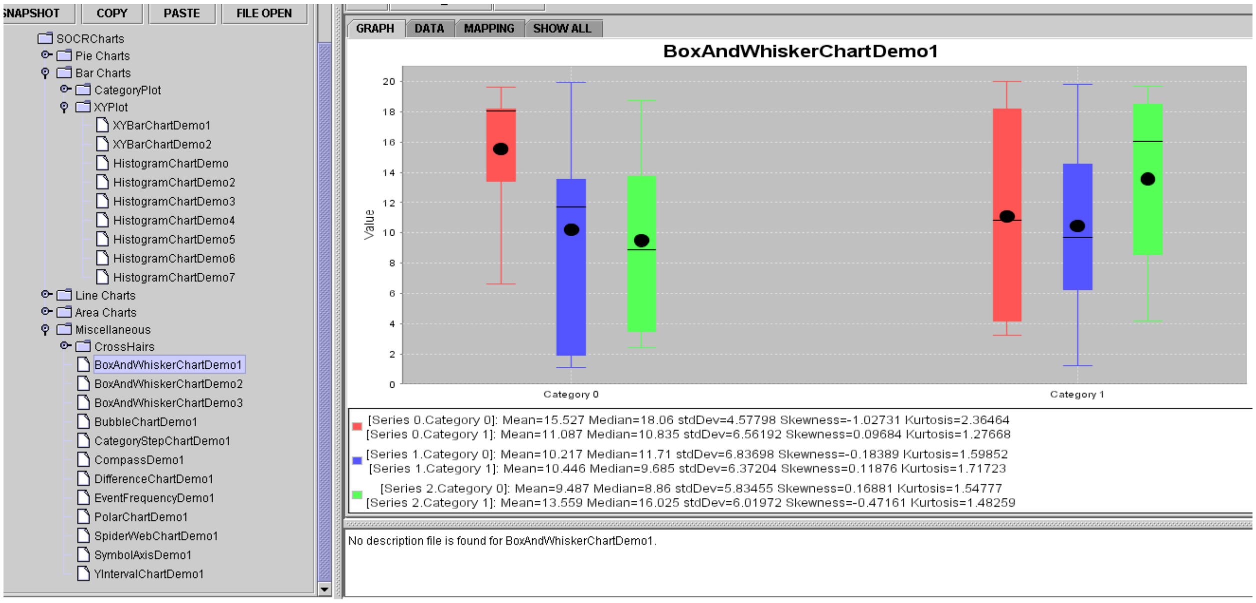
This is a descriptive tool that allows graphic observation of the distribution of the data. Any value outside the limits of this box is considered an outlier.

Procedure: Sort data from lower to high. Calculate Q1(1st. quartile), Q2 (2nd. quartile), Q3 (3rd. quartile) and **build a box** with ends located at Q1 and Q3. A **vertical line** through the box is placed at the median (Q2). Limits are set up at each side, by calculating **the interquartile range** (IQR = Q3-Q1) and multiplying it by 1.5 times. Dashed lines (Whiskers) are drawn within these limits. Numbers outside these limits are marked with an **asterisk (*)**



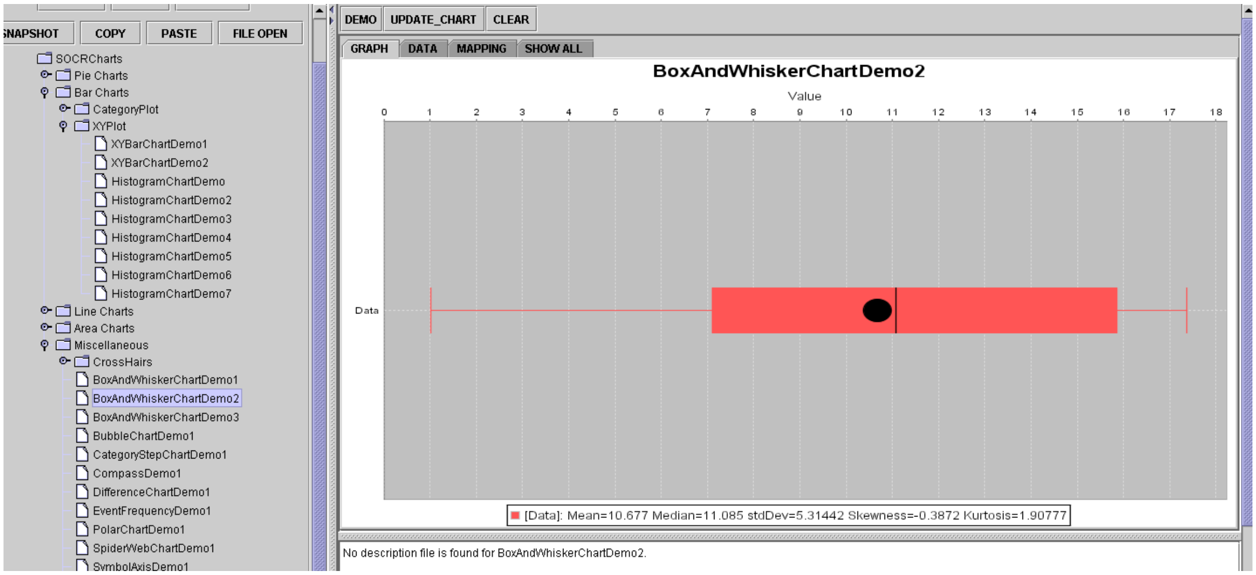
Using SOCR to Get Box Plots

Go to SOCR Charts (http://socr.ucla.edu/htmls/SOCR_Charts.html) and select one of the BoxAndWhisker’s Plots ← Miscellaneous:



```
🟥 [Series 0.Category 0]: Mean=15.527 Median=18.06 stdDev=4.57798 Skewness=-1.02731 Kurtosis=2.36464
   [Series 0.Category 1]: Mean=11.087 Median=10.835 stdDev=6.56192 Skewness=0.09684 Kurtosis=1.27668
🟦 [Series 1.Category 0]: Mean=10.217 Median=11.71 stdDev=6.83698 Skewness=-0.18389 Kurtosis=1.59852
   [Series 1.Category 1]: Mean=10.446 Median=9.685 stdDev=6.37204 Skewness=0.11876 Kurtosis=1.71723
🟩 [Series 2.Category 0]: Mean=9.487 Median=8.86 stdDev=5.83455 Skewness=0.16881 Kurtosis=1.54777
   [Series 2.Category 1]: Mean=13.559 Median=16.025 stdDev=6.01972 Skewness=-0.47161 Kurtosis=1.48259
```

No description file is found for BoxAndWhiskerChartDemo1.



🟥 [Data]: Mean=10.677 Median=11.085 stdDev=5.31442 Skewness=-0.3872 Kurtosis=1.90777

No description file is found for BoxAndWhiskerChartDemo2.

Measures of variation

If all the data are not the same value you have got **VARIATION,** isn't it an easy concept? Sometimes 2 data sets may have the same mean, but variation (or behavior) of their observations is different making one set more stable than other.

When measuring variation you may use any of the following 6 measures:

Range

Difference between maximum and minimum value in a data set:

R = Maximum value – Minimum value

It's useful when we want to have an idea of what is the general composition of the data, and how apart our maximum and minimum is.

Interquartile Range

Difference between 3rd and 1st quartile. It's not affected by extreme values, more efficient than Range.

IQR = Third Quartile – First Quartile

This range measures the information grouped within 25% and 75% of the data set, leaving out the data above and below the limits. It's more accurate than the range, but still presents an important weakness: None of these two formulae use all the data for computations. To overcome this difficulty the following measures were created:

Population Variance

This is one of the most common measures of variation in Statistics. Many of the concepts that we will learn in the future regarding probabilities and hypothesis tests, rely on the accurate computation of the variance.

Variance is the average of the squared variations from the mean. As the formula suggests below, it's necessary to compute the difference between each value and the mean, then square that difference and finally add up all this variations and divide them by the total number of observations.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(xi-\mu)^2}{N}$$

$$\boxed{\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$$

Population Standard Deviation

Square root of Variance, explains how spread out a distribution is, and it's very useful to make comparisons between data sets with the same mean. If distributions have the same mean, the one with the largest standard deviation has the greatest relative spread.

$$\boxed{\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(xi-\mu)^2}{N}}}$$

Sample Variance

The formula is similar to population variance, but notice that the denominator is n-1. The source of the information is a sample and not the entire population. Notice also that the variables are written in lower case and the mean is expressed by x bar and not μ.

$$s^2 = \frac{\sum_{i=1}^{N} (xi - \overline{x})^2}{n-1}$$

$\overline{x} =$ sample mean
n = sample size
s^2 = sample variance

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Sample Standard Deviation

Square root of Sample variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N} (xi - \overline{x})^2}{n-1}}$$

$\overline{x} =$ sample mean
n = sample size
s^2 = sample variance

Using Excel to compute measures of location and variation

Most SOCR Charts (http://socr.ucla.edu/htmls/SOCR_Charts.html) compute the main measures of centrality and variation. Now that you have learned the operational part of computing measures of location and variation, we will take a quick look at a tool provided by Excel to help us in this process. It's the Data Analysis option. We can request a *Summary Statistics* report using the following commands:

a. Open Excel

b. Click on TOOLS menu

c. Click on DATA ANALYSIS

d. Click on DESCRIPTIVE STATISTICS

e. Follow the prompts and select the range with the data you want to input

f. Click on SUMMARY STATISTICS

g. Click OK

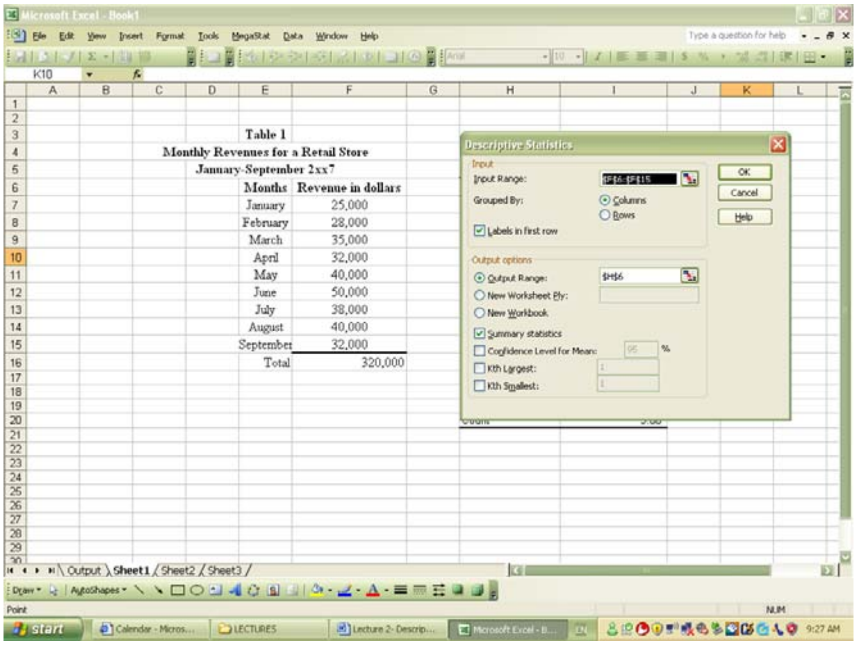


Table 1

Monthly Revenues for a Retail Store

January-September 2xx7

Months	Revenue in dollars
January	25,000
February	28,000
March	35,000
April	32,000
May	40,000
June	50,000
July	38,000
August	40,000
September	32,000
Total	320,000

Descriptive Statistics

Input

Input Range: $F$6:$F$15

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $H$6

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

A table with all the information about: Mean, Median, Mode, Standard Deviation, Sample Variance, Range, Minimum, Maximum, Sum, Count will appear. You will be able to compare samples, populations and make a more informed decision about the variation or stability of the data set.

Revenue in dollars	
Mean	35,555.56
Standard Error	2,495.06
Median	35,000.00
Mode	32,000.00
Standard Deviation	7,485.17
Sample Variance	56,027,777.78
Kurtosis	0.58
Skewness	0.58
Range	25,000.00
Minimum	25,000.00
Maximum	50,000.00
Sum	320,000.00
Count	9.00

Note: When inputting the information in the Input range cell, make sure that you include only quantitative data. The software will warn you that you can't input qualitative data. In the case of our exercise, the months listed on the table are qualitative data.

We can also use MEGASTAT to compute the descriptive statistics; proceed as follows:

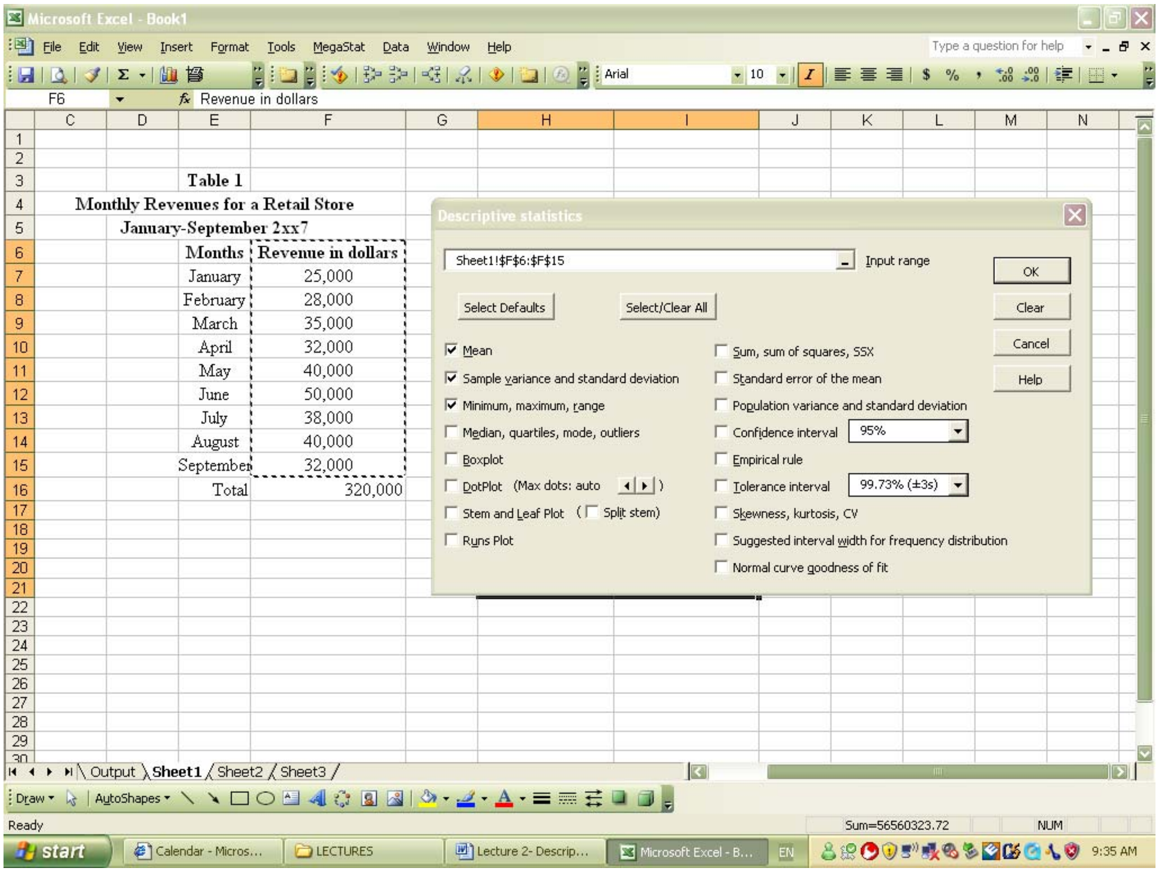
a. Click on MEGASTAT from the menu options

b. Click on Descriptive Statistics

c. Input the range on the window

d. Choose the measurement tools you need: Sample mean, variance, percentiles, box and whisker plots, etc.

e. Click ok



F6: Revenue in dollars

Table 1

Monthly Revenues for a Retail Store

January-September 2xx7

Months	Revenue in dollars
January	25,000
February	28,000
March	35,000
April	32,000
May	40,000
June	50,000
July	38,000
August	40,000
September	32,000
Total	320,000

Descriptive statistics

Input range: Sheet1!$F$6:$F$15

☑ Mean
☑ Sample variance and standard deviation
☑ Minimum, maximum, range
☐ Median, quartiles, mode, outliers
☐ Boxplot
☐ DotPlot (Max dots: auto)
☐ Stem and Leaf Plot (☐ Split stem)
☐ Runs Plot
☐ Sum, sum of squares, SSX
☐ Standard error of the mean
☐ Population variance and standard deviation
☐ Confidence interval 95%
☐ Empirical rule
☐ Tolerance interval 99.73% (±3s)
☐ Skewness, kurtosis, CV
☐ Suggested interval width for frequency distribution
☐ Normal curve goodness of fit

The following report is prepared by MEGASTAT, scroll down and see how many of these you can recognize:

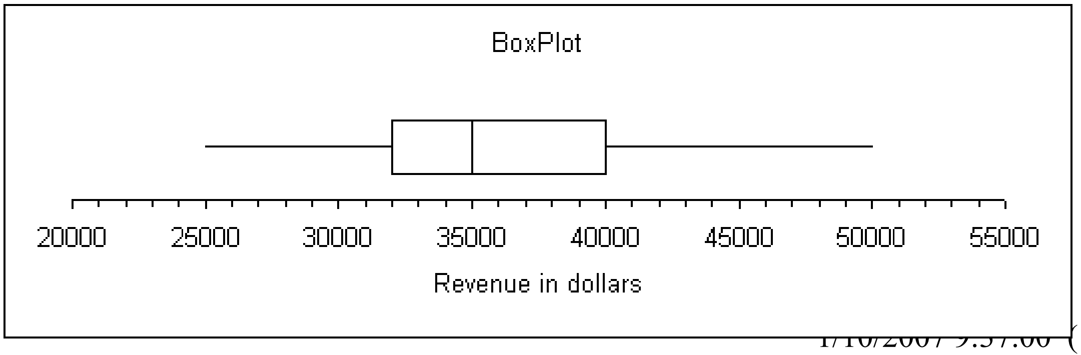
Descriptive statistics

	Revenue in dollars
count	9
mean	35,555.56
sample variance	56,027,777.78
sample standard deviation	7,485.17
minimum	25000
maximum	50000
range	25000
population variance	49,802,469.14
population standard deviation	7,057.09
empirical rule	
mean - 1s	28,070.39
mean + 1s	43,040.73
percent in interval (68.26%)	66.7%
mean - 2s	20,585.21
mean + 2s	50,525.90
percent in interval (95.44%)	100.0%
mean - 3s	13,100.04
mean + 3s	58,011.07
percent in interval (99.73%)	100.0%
1st quartile	32,000.00
median	35,000.00
3rd quartile	40,000.00
interquartile range	8,000.00
mode	32,000.00
low extremes	0
low outliers	0
high outliers	0
high extremes	0

Stem and Leaf plot for	Revenue in dollars
stem unit =	10000
leaf unit =	1000

Frequency	Stem	Leaf

2	2	5 8
4	3	2 2 5 8
2	4	0 0
1	5	0
9		



1/10/2007 9:37:00 (1)

Notice that in this report there are some new measurement tools: The empirical rule and the stem and leaf table. Read more about the empirical rule below.

Combining measurement tools

In this section we will combine the measurements of location and the measurement of variation and use it for applications in business.

Coefficient of Variation (CV)

$CV = \sigma/\mu\ (100)$
Population CV

$CV = s/\bar{x}\ (100)$
Population CV

This indicator combines the standard deviation and the mean in a very useful measure that provides information about variation of data sets, when their means are different. There is a coefficient of variation for population and for samples; their only difference is the type of standard deviation used.

The Empirical Rule

Combines information about (μ) and (σ), to explain approximately how much information in your data set is contained within a specific range. This is a very useful indicator for decision makers, because it identifies the outliers or extreme elements of our data. Refer to the table below. The table says that in a normal distribution of values, 68% of the observations will be contained in a range of one standard deviation from the mean. 95% of the observations are within 2 standard deviations from the mean, and virtually all the data values should be within 3 standard deviations from the mean.

Table 4

The empirical Rule

$\mu \pm 1\sigma$	Contains approx. 68% of the values
$\mu \pm 2\sigma$	Contains approx. 95% of the values
$\mu \pm 3\sigma$	Contains virtually all of the data values

Note: Frequency distribution must be bell-shaped and symmetric to apply this rule.

Let’s use the information from the MEGASTAT report.

```
empirical rule
  mean - 1s              28,070.39
  mean + 1s              43,040.73
  percent in interval
(68.26%)                 66.7%
  mean - 2s              20,585.21
  mean + 2s              50,525.90
  percent in interval
(95.44%)                 100.0%
  mean - 3s              13,100.04
  mean + 3s              58,011.07
  percent in interval
(99.73%)                 100.0%
```

According to this report, 66.7% of the data is located within 1 standard deviation from the mean, this is between 28,000 and 43,000; 100% of the data is within 2 standard deviations from the mean, this is between 20,600 and 50,500 and 100% of the data is within 3 standard deviations from the mean, this is between 13,100 and 58,000. This implies that there are not outliers in this data set, because one hundred percent of the data is included within 3 standard deviations.

Now, how do you use this knowledge? Let's say that next month you have a customer with a billing amount of $62,000, he is definitely an outlier in this distribution, because he is over the 3 standard deviations from the mean.

Tchebysheff's Theorem

Very similar to the Empirical rule, with the only difference that frequency distributions doesn't need to be bell-shaped and symmetric to apply this rule.

The table below states the ranges of validity of Tchebysheff's theorem.

Table 5

Tchebysheff's Theorem

$\mu \pm 1\sigma$	Contains approx. 0% of the values
$\mu \pm 2\sigma$	Contains approx. 75% of the values
$\mu \pm 3\sigma$	Contains approx. 89% of the values

Standardized Data Values

The standardization of data values is a procedure that we will use intensively in the following chapters and it allows comparisons between data sets with completely different data scales (e.g. prices for an article expressed in dollars vs. prices expressed in pesos; scores based on 100 points vs. scores based on 20 points).

To standardize a value is to express the value in terms of the number of standard deviations from the mean. Also called z values:

The formula to be used to standardize is very simple:

Standardized population data

$$z = \frac{x - \mu}{\sigma}$$

X= original data value

μ = population mean

σ= population standard deviation

Z= standard score (number of standard deviation x is from μ)

A standard value Z is expressed in terms of standard deviations. So for example

Standardized sample data

$$z = \frac{x - \bar{x}}{s}$$

How do we use the concept? Let's say that you are comparing the billing portfolio of branch 1 and branch 2, and you want to analyze which branch has more dispersion of data. A very good way to do it is by computing Z values, and then compare them.

Always remember to practice the suggested problems in each section of the chapter.

Practice, practice, practice!! Statistics is so useful and these first 3 chapters are the cornerstone of the rest of your class.

SOCR Tools for Exploratory Data Analysis (EDA):

The links below provide additional help and instructions on how to use SOCR for EDA:

wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_Histogram_Graphs
wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_BoxPlot
wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_BarCharts_CategoryPlot
wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_PieChart
wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_IndexChart
wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_StatisticalLineChart

See you in class!