**Module 1: Data and Variation**

# What is Statistics?

Statistics, the discipline, is the art and science of extracting useful information from data.[[1]](#footnote-1)

Three phases

Problem analysis, linking appropriate data to substantive questions

Exploratory Data Analysis - discovery (found it!)

summarization/description/exploration

Confirmatory Data Analysis - skepticism (really?)

confirmation/inference/assessment

Data analyst as detective (detects but isn't fooled)

Statistics helps us *make decisions* in the presence of uncertainty and variation.

# What are Data?

A collection of numbers, labels, or symbols, and the context of those values

Often, a sequence of measurements on a process (a time series)

The time it takes to get to school each morning

The closing price of a stock each week

The sales volume at Amazon.com each month

Often, a subset of a larger group (a sample from a population)

The ages of the students in this class

The preferences of potential automobile purchasers visiting a showroom

The undergraduate GPA scores of the students in the front row

Data are often arranged in tables, with *cases or observations* along the rows and *variables* defining columns. Data can be numerical (e.g. prices) or labels (e.g. colors).

# Key Feature: Variation

A common feature**—**virtually all data exhibit variation. The values found in a column vary from one case to another.

A principle goal of statistics is to describe and understand the implications of variation.

# Discovery with Statistical Tools

Just looking at lists of raw numbers reveals little, especially for large data sets.

Finding a revealing view of the data is the key to effective statistical analysis.

Different types of displays: graphical and numerical.

Goal is to focus attention on essential features of data.

Separate reproducible patterns from random, coincidental features.

Relevance in decision making

Summarization can be very useful when using data to form decisions.

Avoid distraction from extraneous features of data.

To motivate and illustrate such methods, we’ll study three data sets:

• GMAT scores (review histograms, summary statistics)  
 • Microsoft stock (checking for normality)  
 • Executive salaries (outliers, skewness and logarithms)

# GMAT Scores

The analysis of data depends on the nature of your question. Managers have different questions and so will often approach a task differently.

Consider the questions that might be asked about GMAT scores at Wharton.

Student:

Where do I place in the class of Wharton MBA students?

Vice-Dean:

How does the entering class compare to those at other programs?

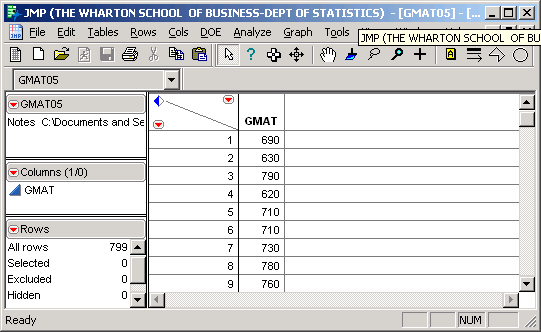
How have scores trended over time?

Effective statistical summaries can make it easy to answer both types of questions much more precisely and concisely than from the raw scores themselves. We arrange data into a table.

The *rows* in the table go by various names such as observations, cases, or even subjects. The table collects values of various attributes of the observations.

The *columns* in the table are called “variables.” Each column holds the values of some attribute of the observations that comprise the rows of the table.

The first column of the Excel file *GMAT05.XLS* consists of the GMAT test scores of 799 members of the Wharton MBA, Class of 2005. That’s one numerical variable[[2]](#footnote-2) with 799 observations, one row for each of these students. Let’s open this file with JMP. [[3]](#footnote-3)



What do you learn by looking directly at this column of 799 numbers?

JMP provides graphical and numerical summaries[[4]](#footnote-4) of the 799 observations of the variable named GMAT.



On the left are two graphical summaries of the data**—**a boxplot and a histogram. On the right are numerical summaries**—**various quantiles and moments.

Let’s consider these one at a time and link the numerical summaries to the picture.

# Histogram

A *histogram* is a visual display that reveals the location, dispersion, and shape of the data distribution.



The basic idea is that the area of each rectangle is proportional to the number of values (frequency) in each interval or bin.[[5]](#footnote-5)

Although programs such as JMP automatically pick a reasonable number of intervals for the histogram, it is sometimes useful to consider an alternative number of intervals.[[6]](#footnote-6)

What is gained and lost by changing the number of intervals?

General shape characteristics: unimodal, symmetry, left or right skewness.

# Moments

Moments are average quantities of interest.[[7]](#footnote-7)

The mean and standard deviation (SD) are the two most basic and commonly used moments. Both statistics have the same measurement units as the data.

The (*sample*)[[8]](#footnote-8) *mean*, denoted by , is defined as





where *x*1,…, *xn* stands for the *n* data values[[9]](#footnote-9) of the variable *x*. Subscripts identify the row in the data table. Note that is just a simple average of the data values.



From the JMP summary for the GMAT data, the mean = 713.7 points.



What aspect of the data is captured by the mean?

The mean is the “balancing point” on the histogram. A piece of wood of constant thickness shaped like the histogram would balance at the mean.[[10]](#footnote-10)

The (*sample*) *variance*, denoted by *s*2, is defined as

.

Note that *s*2 is essentially the average[[11]](#footnote-11) of the *n* squared deviations from the mean, which are represented above by .

The (*sample*) *standard deviation*, denoted by *s*, is defined as



which is the square root of the variance.

From the JMP summary for the GMAT data, the standard deviation *s* = 36.43 points.

From this we can deduce that the variance *s*2 = 1327.36 points2.

Since one can easily obtain the standard deviation from the variance, and vice-versa, it suffices to report only one of them.[[12]](#footnote-12)

Which of the following two sets of five values will have a larger variance and standard deviation? Why?

-2, -1, 0, 1, 2 or -4, -2, 0, 2, 4

What aspect of the data is captured by the variance and the standard deviation?

Notice the units of the standard deviation and the variance differ because of the squaring.

Which of the variance and standard deviation seems more understandable?

# Quantiles

Quantiles are derived from ordered data values. The *k% quantile* (also called the *kth percentile*) is a value that separates the bottom *k* percent of the ordered data from the top.

Quantiles can be used both to capture the location of the data (instead of the mean), as well as measure the spread of data (instead of the standard deviation).

The most common quantiles are the median (50th percentile), the first quartile Q1 (25th percentile), and the third quartile Q3 (75th percentile). How do these divide up the area on a histogram?

Which is a more appropriate measure of the center of data**,** the mean or the median?

For a right skewed distribution, which is larger**,** the mean or the median?

An alternative to the standard deviation as a measure of spread is the *interquartile range*, defined as IQR = Q3 – Q1, which for the GMAT data is 740 - 690 = 50 points.[[13]](#footnote-13)

What aspect of data is captured by the IQR?

# Boxplots

A *boxplot* is a visual summary of the data based on quantiles. A boxplot makes it easy to relate the quantiles to the histogram.

The main features of a boxplot identify the median, Q1, and Q3.

These are displayed as follows[[14]](#footnote-14)



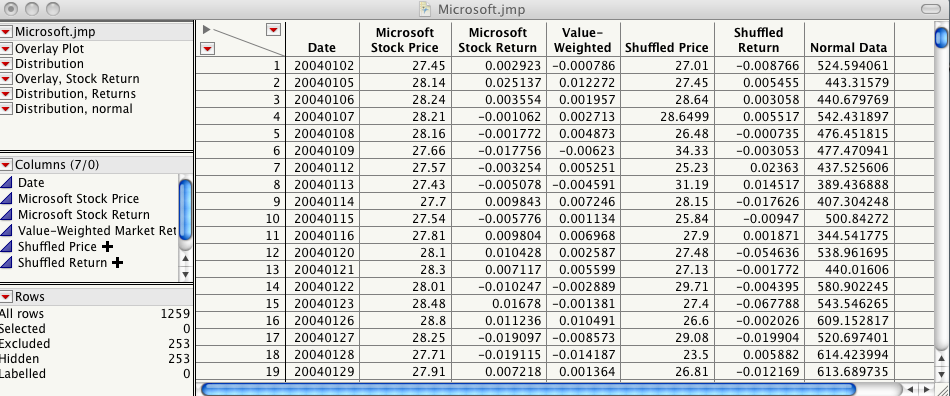
Remark: JMP calls this an “outlier boxplot.” JMP also provides a “quantile boxplot” which is similar but does not highlight the presence of outliers.[[15]](#footnote-15)

# Prices of Microsoft Stock[[16]](#footnote-16)

Question: What is the risk associated with owning a stock like that of Microsoft?

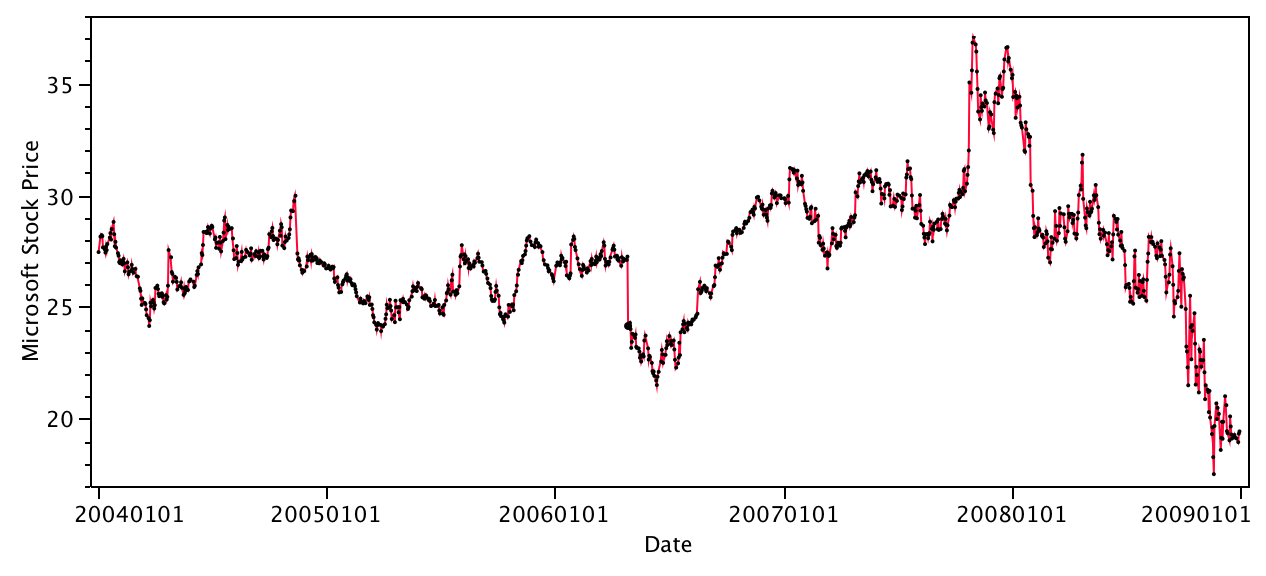
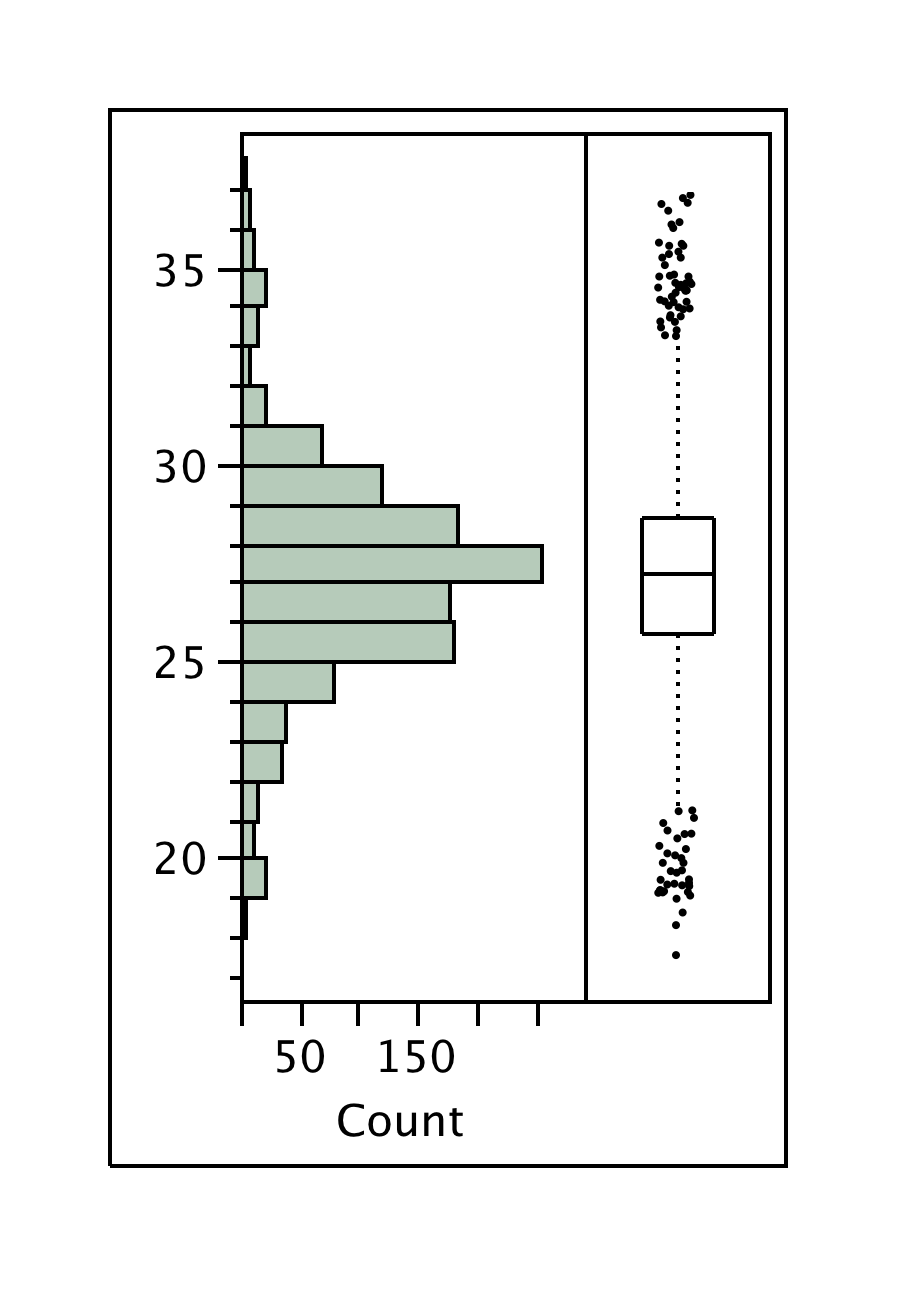
The Excel file *Microsoft.XLS* contains daily closing prices from 2004 through 2008 on a share of Microsoft stock.

Let’s use the JMP command Open to see these data and study the price of this stock.



Because the prices form a time series (*i.e*., a sequence of observations that are ordered in time), we begin by plotting the prices over time.[[17]](#footnote-17)

Daily Prices of Microsoft, 2004-2008

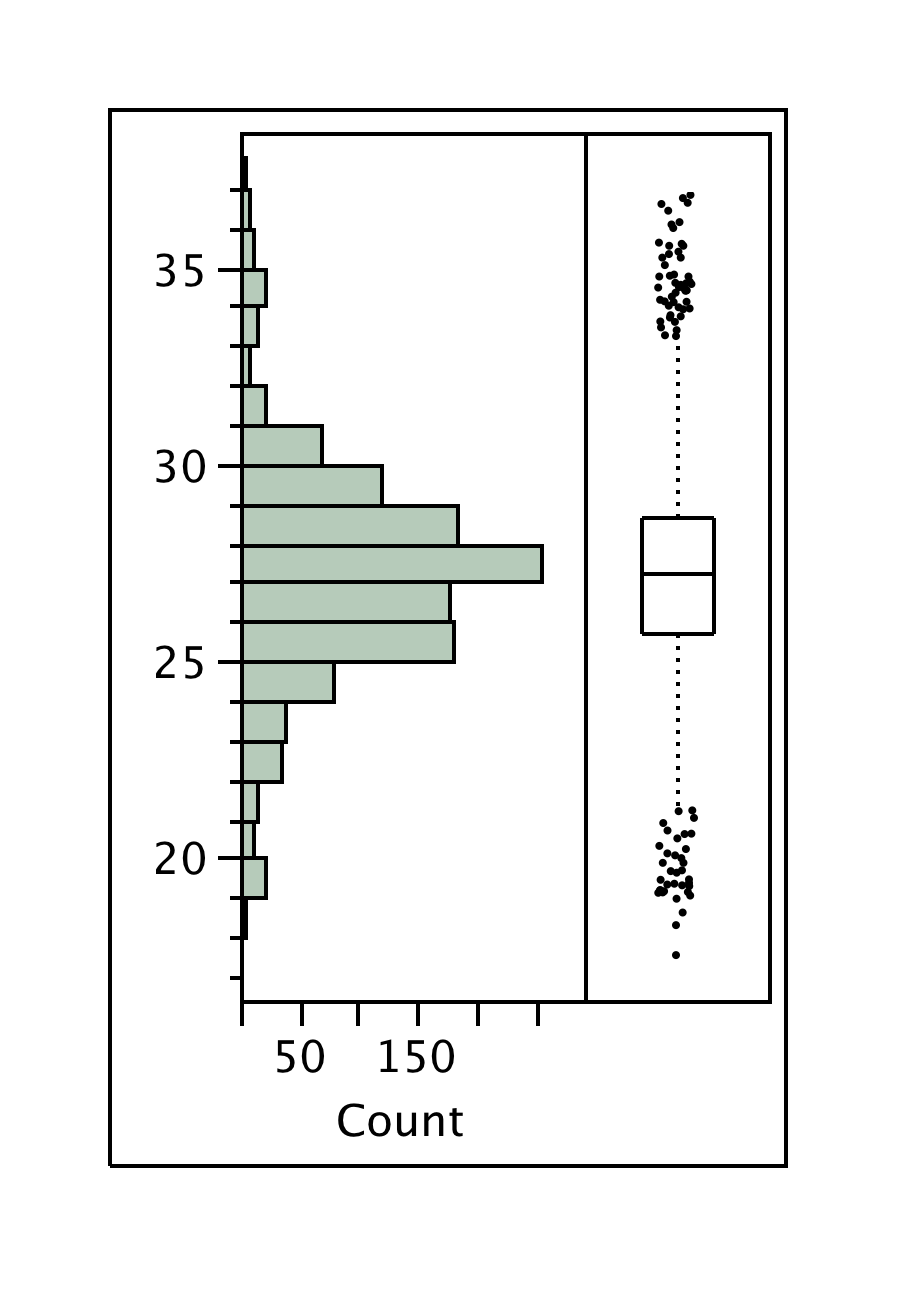
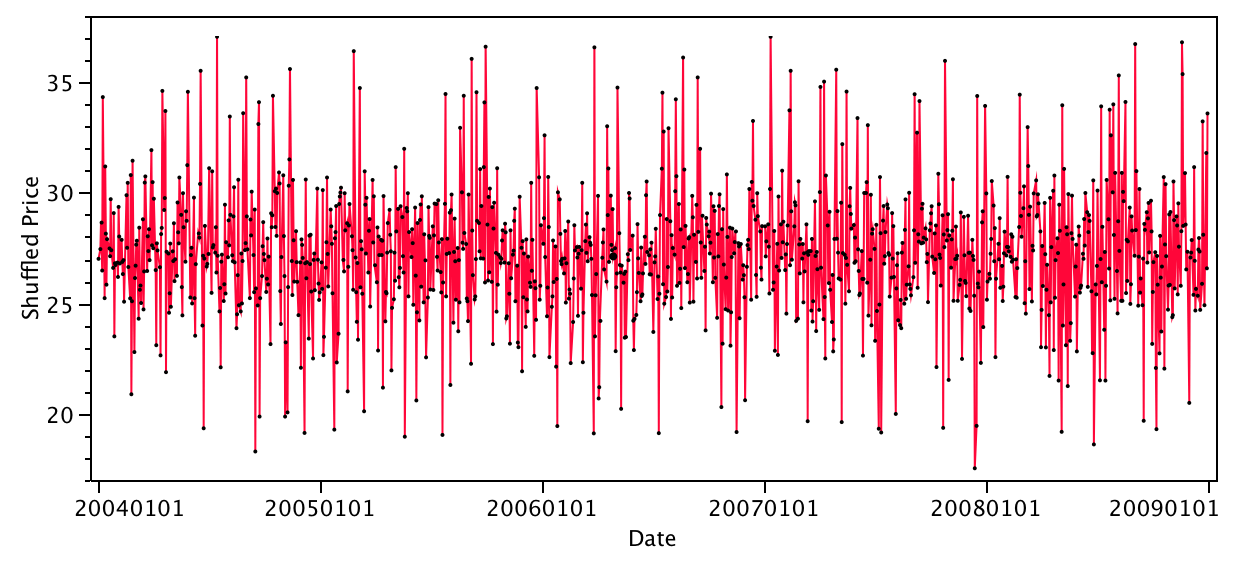
 

Note the meandering behavior of this time series.[[18]](#footnote-18)

Successive draws do not stray too far from each other. This is a manifestation of sequential *dependence*. The price each day *depends* on the price the previous day.

To see what this data would look like without the sequential dependence, let’s randomly shuffle the order of Price.[[19]](#footnote-19)

A time series plot of Shuffled Price looks like

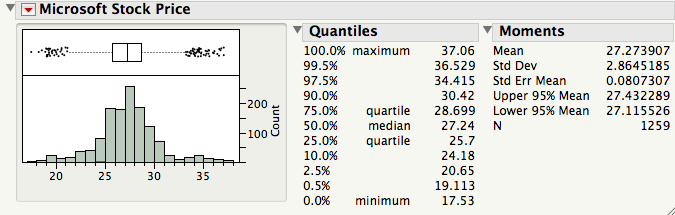


Note the very different appearance of this plot compared to the previous one – the meandering has disappeared – but the histogram is the same!

This lack of the meandering pattern is typical of a sequence of *independent* random draws from a population (i.e., a random sample). What we see is *constant level*, *constant variance* and a *lack of meandering*.

Because the sequential patterns of *Price* and *Shuffled Price* are so different**,** it is implausible the stock prices are a sequence of random draws from a population.

Suppose that instead of looking at time series plot, you only considered the summary



This summary conceals the meandering behavior!

If you only saw this summary, you wouldn’t be able to recover the sequential dependence that is such a dominant feature of the time series plot.

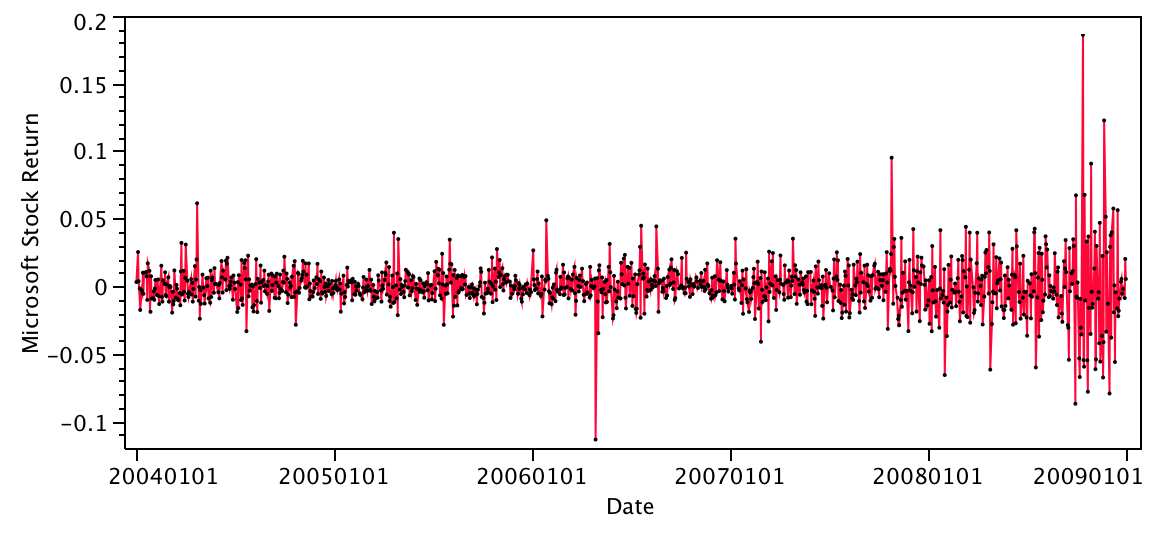
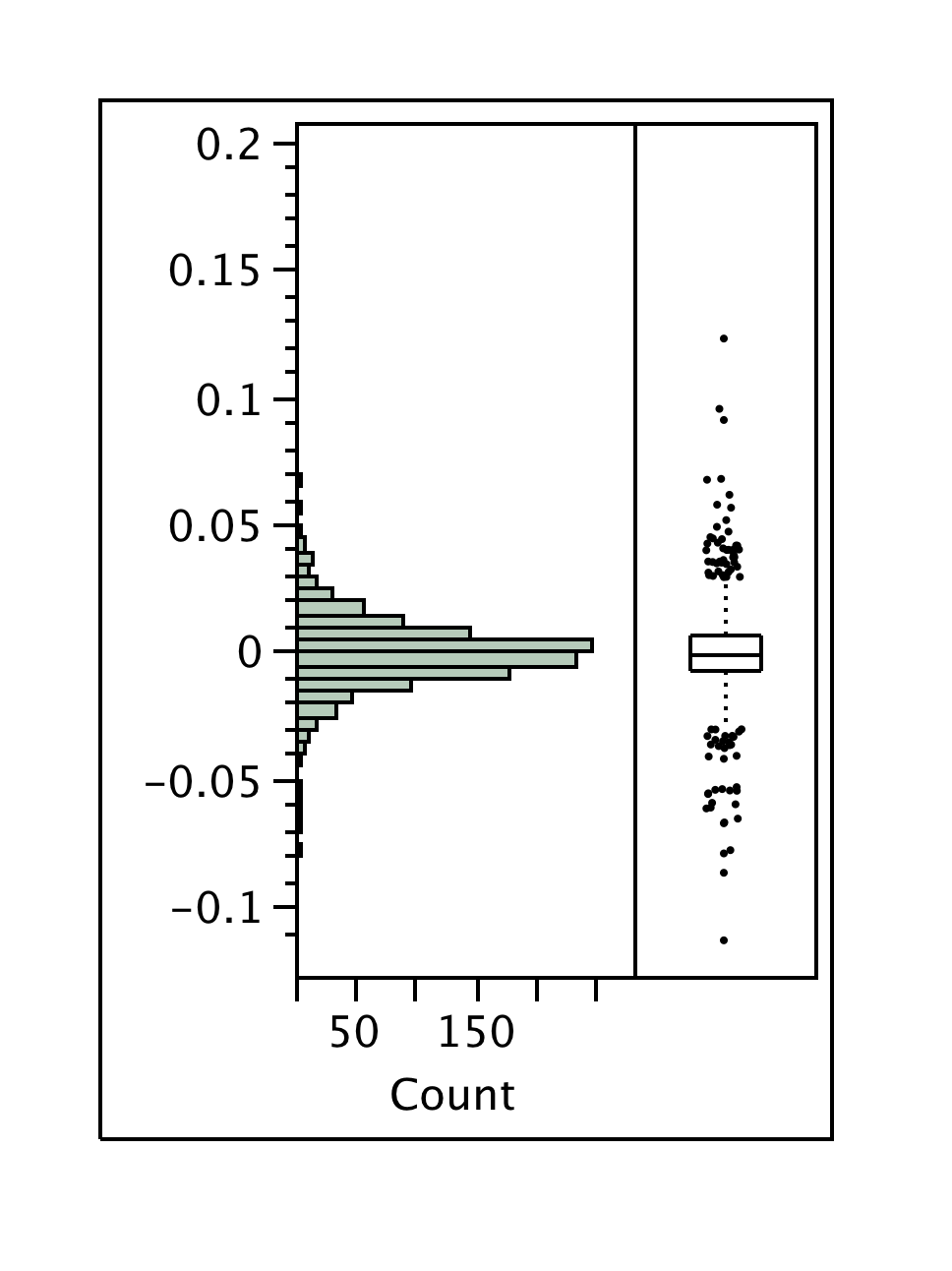
What does a histogram suggest to estimate the next value? Its accuracy?

As a summary of the variation in a time series, a histogram is misleading if the time series is dependent. A histogram does not distinguish between data without patterns like those on p. 1-15 (after scrambling the order) and data with smooth, meandering patterns like those in the sequence plot of prices on p. 1-14.

Letting *pt* denote the price of Microsoft at time *t*, a different perspective is obtained by focusing on the successive daily relative changes or returns.[[20]](#footnote-20)

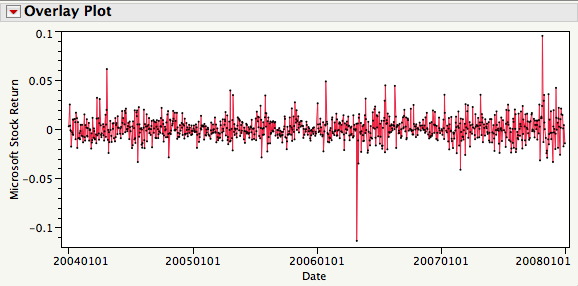


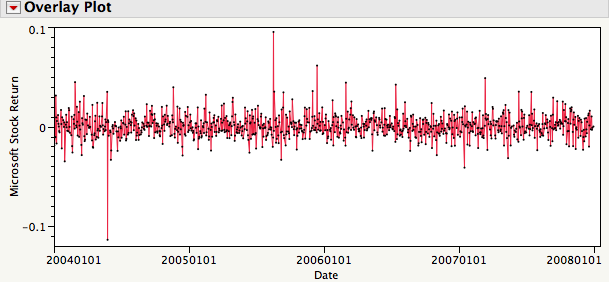
The time series plot of *Returns* looks much more like a sequence of independent draws from a population. The meandering pattern disappears, and we instead see the huge increase in volatility (variation) in 2008 and several conspicuous outliers.

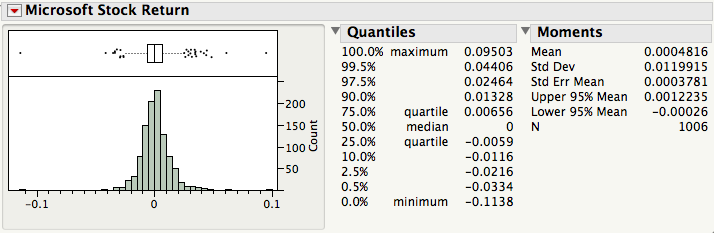
Let’s focus for the moment on the period from 2004 through 2007. One of these plots shows the actual returns, and the other shows the returns after shuffling the order.

Can you tell which is which?





Unless we can predict the timing of the outlying events, the histogram of this time series is not hiding much from us.[[21]](#footnote-21)



The histogram of the returns during these years shows that the data have a “bell-shaped” distribution with a scattering of outliers.

This is what you would probably see when sampling[[22]](#footnote-22) from a *normal population*.

This bell-shaped tendency is also evident, though perhaps not so clearly, in the histogram of the GMAT scores and many other phenomena.



# The Normal Distribution

The *normal distribution* is the prototype of “bell-shaped” distributions.

This ideal often approximates the shape of histograms of many data sets that occur naturally.

A picture of a normal distribution:



The normal distribution may be thought of as a smooth histogram of a very large data set.

The mean and standard deviation of a normal distribution are denoted by *μ* and *σ* respectively.

Important characteristics of a normal distribution are:

Bell-shaped and *symmetric* about its mean *μ.*

About 68% (actually .6827)[[23]](#footnote-23) of the area lies within (*μ – σ* , *μ + σ* ).

About 95% (actually .9545) of the area lies within (*μ – 2σ* , *μ + 2σ* ).

Almost 99.7% (actually .9973) of the area lies within (*μ – 3σ* , *μ + 3σ* ).

The standard deviation σ is a natural unit of distance for the normal distribution.

# Normality of Data

Using the refined histogram interpretation, the relative area under the curve over an interval can be associated with both the

Relative frequency of values in the interval, and the

Probability that a randomly drawn observation falls within the interval.

Normality describes how data concentrate near the mean. Let’s see how well a normal distribution describes the returns on Microsoft during 2004 through 2007. Here, ≈ .00048 and *s* ≈ .01199 so that:



≈ (-0.01151, 0.01247)



≈ (-0.02350, 0.02446)



≈ (-0.03549, 0.03645)



For each of these intervals, it turns out that:

789/1,006 ≈ 78.4% of the observations fall within *s* of .



963/1,006 ≈ 95.7% of the observations fall within 2*s* of .



997/1,006 ≈ 99.1% of the observations fall within 3*s* of .



Key point

This *approximate* agreement between the normal distribution and data is sometimes called the *empirical rule*. Since the empirical rule approximates the distribution of these returns, we can concisely summarize the 1,006 returns observed in Microsoft stock with just a mean and SD. The empirical rule translates these two summary statistics into statements about where the values concentrate.

A cool thing about normally distributed data is that the mean and standard deviation summarize all we may need to know about the dataset.

Why this bell-shape and not another?

Fact: If the data values arise as the sums of many small random effects, then a histogram of the data will resemble the normal distribution. The mathematical formalization of this fact is the Central Limit Theorem (CLT) (to be discussed in Module 6).

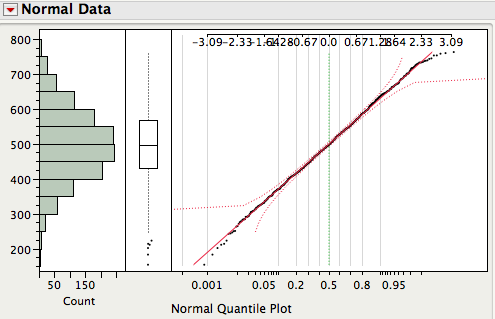
# Normal Quantile Plot

Normality turns out to be such a useful assumption that we need a diagnostic to check for normality that works better than looking at the histogram. Many of our diagnostics come in the form of plots that help us *judge* how well the data match our assumptions.

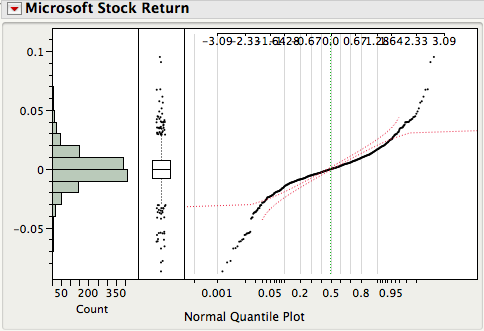
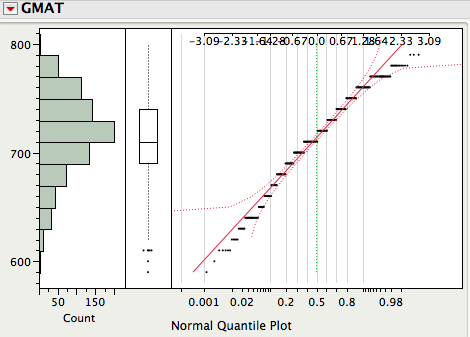
To judge if a histogram of data is close to the ideal normal distribution**—**so that the data can be treated as a sample from a normal population**—**use a *normal quantile plot*.[[24]](#footnote-24)

How to use it:

If the histogram is close to the ideal normal shape, then *all* the points will lie between the two dashed lines on either side of the diagonal.[[25]](#footnote-25)



What conclusion would you draw about the Microsoft returns that include 2008 (left) or the GMAT scores? What’s causing those flat “steps”?[[26]](#footnote-26)

Key point

Many statistical inferences rely on normal distributions.

Use a normal quantile plot to check for normality rather than assume normality.

Why the normal quantile plot works:

The *y*-axis of the normal quantile plot shows the data.

The *x*-axis shows the distance from *μ* in units of *σ* along the top scale and the corresponding percentiles associated with each value along the bottom scale.

With the x-axis defined in this way, a “perfect” normal sample would fall along the diagonal line from the lower left to the upper right corner of the plot.

# Executive Compensation Data

Not all data is normally distributed. Some data is not even close to normally distributed.

The JMP file *CEO\_Comp\_03.JMP* contains the annual compensation of 1,495 executives in 2003.[[27]](#footnote-27) The summary of TotalComp (total compensation) is



Someone made nearly $75,000,000! Who was it?[[28]](#footnote-28)

Extreme values, often called *outliers*, dominate the output and make it difficult to see the rest of the distribution.

Let’s exclude the five largest values and see what happens.[[29]](#footnote-29)



That doesn’t help much. The compensations of a relatively small number of executives continue to dominate the plot.

# The Effect of Outliers on Summary Statistics

Even though these 5 executives make up only 1/3 of 1% of the cases, the mean dropped from $4.63 million to $4.45 million and the standard deviation dropped from $6.2 million to $5.5 million.

These five large salaries are very influential. Note that the mean and, especially, the SD change dramatically when adding or removing points that are far away from the others. In contrast, note that the median and the interquartile range change very little.

Now look what happens if we remove all cases for which TotalComp > $10 million.[[30]](#footnote-30) What’s going on?



# Transforming Data

An alternative way of looking at the big picture is to transform the data to a different scale. For these data, let’s consider the log transformation:

y = log10 x

If we use base 10, then the logs of the compensations essentially count the number of digits (minus 1 since, for example, log10 100 = 2 and log10 1000 = 3).

Note that log(y/x) = log y – log x so that multiplicative increases in the original scale correspond to additive increases in the log scale.[[31]](#footnote-31) Thus

percentage changes in the original scale become additive changes in the log scale.

Question: Which is larger

[log $11,000 – log $10,000] or [log $1,001,000 – log $1,000,000] ?

For the CEO\_Comp\_03 data, let’s create the new variable[[32]](#footnote-32)

Log10TotalComp = log10 TotalComp

A summary of these logged compensation values is revealing:



Wow! The transformation has revealed an unusually small extreme value. Who is it?

Let’s exclude this point to get a better picture of the distribution of Log10TotalComp.



By transforming, we can see the variation that distinguishes the compensation of the majority of the executives rather than just singling out those that make much more than the rest.

The log transformation has done this by pulling in the large values and stretched out the small values. It has transformed right skewed data into more normal data.

Is any information lost by transforming the data?

Is the average of the logs equal to the log of the average?

Given one, can you find the other?

# Comparison

Questions of interest

Are *top* executives in some industries more highly paid than those in others?

Are *typical* executives in some more highly paid than in others?

Let’s first obtain JMP summaries of the two variables: Industry and Log10TotalComp.[[33]](#footnote-33)



Our goal is to compare compensation among industries. Can we tell from these summaries if executives in some industries are higher paid that those in other industries?

What about the relationship *between* the category of the industry and total compensation?

To get some sense of the relationship between them, we can highlight industries such as financial or IT.[[34]](#footnote-34) *Plot linking* in JMP highlights the associated values of Log10TotalComp.[[35]](#footnote-35)





Does it look like financial executives are paid differently than IT executives?

Are the differences small, happening just for one or two, or more general?

Suppose we “raw” compensation rather than compensation on the log scale.

Which question would this help us answer?

# Comparison Boxplots

Another useful approach for comparisons here is to look at side-by-side *comparison boxplots*.[[36]](#footnote-36)

Comparison boxplots provide a plot that captures differences that we might discover from plot linking in a single, static image.

Each boxplot summarizes compensation within an industry. The plots are shown parallel to each other on the same scale. What can we say about the executive compensation differences across these industries?



What would be a better way to label the *y*-axis?

# Statistics and Variation

In many ways, statistics can be seen as the study of variation in data.

Basically, we would like to “explain” the variation in data. Why? If we can identify what causes the data to vary, we might be able to control this variation to our benefit or use the insights to improve predictions.

At a minimum, we would like to understand why our data vary. The reasons for variation are what we’ll call the “sources of variation.”

For the executive compensation data we just studied, to what extent is the salary variation explained by industry? Is it major?

Statistics 603 emphasizes the concept of variation. Statistics 621 emphasizes explaining that variation, finding explanations for why data vary. Those explanations are useful for planning and forecasting.

# Take Away Review

Statistical summaries of data, both graphical (e.g., histogram and boxplot) and numerical (e.g., mean and standard deviation).

Normality and the shape of a histogram.

Identified by a mean and standard deviation. Use of normal quantile plot.

Transforming data to remove skewness.

Using statistical methods to understand sources of variation.

Comparison boxplots may reveal underlying sources of variation such as variation due to grouping.

# Next Module

Probability models for random variation.

1. Any numerical summary of data, such as an average, is called a statistic. [↑](#footnote-ref-1)
2. JMP identifies numerical data by the blue triangle in the small window at the left. Data can also name categories. JMP identifies categorical variables with a small red bar chart. Chapter 2 of SF discusses types of variables and Chapter 3 describes methods to summarize and display categorical data. [↑](#footnote-ref-2)
3. Get a copy of the data file GMAT.XLS from WebCafé. Open this file with the JMP command File > Open. Select file type as Excel files (\*.XLS). Since the column is labeled GMAT in the first row of the Excel file, we select “Always enforce Excel Row 1 as a label” in the JMP dialogue box. By saving this file as GMAT05.JMP, you can then open it by simply clicking on the GMAT05.JMP file. [↑](#footnote-ref-3)
4. Use Analyze > Distribution with GMAT selected as the Y column. To get the horizontal layout shown next, right-click on GMAT and then select Display Options > Horizontal Layout. Note sometimes the histogram won’t initially have contiguous rectangles. To fix this, you need to click on the hand tool (which changes the cursor to look like a hand) and then drag the “hand” on the histogram. [↑](#footnote-ref-4)
5. Histograms illustrate the Area Principle of statistical graphics. See SF, page 35. [↑](#footnote-ref-5)
6. In JMP, this can be done *interactively* with the hand tool. [↑](#footnote-ref-6)
7. Why are these things called moments? In physics, moments of inertia describe physical properties of an object**—**where it balances or how hard it is to spin. It turns out that statistical moments capture analogous features of data. [↑](#footnote-ref-7)
8. What’s a sample? For now, a sample simply refers to a collection of values. Later on, we’ll define samples and populations more carefully. [↑](#footnote-ref-8)
9. We typically use *n* to stand for the number of observations of a variable. [↑](#footnote-ref-9)
10. Remember the connection between statistics and physics? The mean is just the first moment of inertia of the histogram, the balancing point if we think of the histogram as a solid thing sitting on a see-saw. [↑](#footnote-ref-10)
11. Why divide by (*n* - 1) rather than *n*? An honest answer requires details that we will not reach until later. For now, think of the variance as the average squared deviation. [↑](#footnote-ref-11)
12. You might wonder why we define both *s* and *s*2 since they are redundant. As we’ll see later, there are contexts where each turns out to be of direct interest. [↑](#footnote-ref-12)
13. Some business schools summarize the GMAT scores of incoming students in the form of the 20% and 80% values. That range is almost the interquartile range, the gap between the 25% score and the 75% score. [↑](#footnote-ref-13)
14. See Section 4.3 of SF (page 60) for further details of boxplots and their relationship to the histogram. [↑](#footnote-ref-14)
15. JMP’s quantile boxplot shows the box as done in the usual boxplot. In place of outliers, however, the quantile boxplot joins the box to the maximum and minimum of the data and draws small lines at the 10% and 90% quantiles. See the JMP Help for further details. [↑](#footnote-ref-15)
16. See the case study “Financial Time Series” in SF, beginning page 134, for more discussion and another example of stock prices with more details on manipulating this type of data. [↑](#footnote-ref-16)
17. This time series plot was obtained with the command Graph > Overlay Plot. Select Price as Y and Date as X. After the Overlay Plot appears, click on the red triangle and select “Connect Thru Missing”. [↑](#footnote-ref-17)
18. A more precise characterization of meandering is positive autocorrelation, which we will discuss towards the end of STAT 621. For now, it is most helpful to simply recognize meandering behavior as the data “following itself,” much like a meandering river. [↑](#footnote-ref-18)
19. To do this in JMP, select Cols > New Column and name the new variable Shuffled Price. Next, select Formula under New Property. In the Formula Editor Window, select Price under Table Columns, select Row > Subscript under Functions, select Random > ColShuffle under Functions, and click OK. Again use Overlay Plot to create the time series plot. Note the value of connecting the points in order to see the variation. See the visual test for association in section 6.2 of SF, p 107. [↑](#footnote-ref-19)
20. Calculation of returns often requires other adjustments, such as to take account of stock splits and dividends. See the case “Financial Time Series”, beginning on page 134 in SF. [↑](#footnote-ref-20)
21. The histogram does hide the timing of the events. But when the observations of a variable are independent, can we really anticipate the timing of the next large value? [↑](#footnote-ref-21)
22. When the data is a random sample from a population, the shape of the histogram tends to mimic the shape of the population distribution. We’ll discuss random sampling in more depth later. [↑](#footnote-ref-22)
23. See the discussion of the empirical rule in section 4.4 of SF, starting on p. 62. [↑](#footnote-ref-23)
24. To generate a normal quantile plot, you first need to be looking at a histogram. Click on the red triangle in the title bar next to the name of the variable; select Normal Quantile Plot. Select Fit Distribution > Normal to overlay a normal curve on the histogram. See section 12.4 of SF, page 272 for discussion of this plot. [↑](#footnote-ref-24)
25. More precisely, 95% of the time, a random sample from a normal distribution will lie between the two dashed lines. [↑](#footnote-ref-25)
26. Some versions of JMP change the labeling of normal quantile plots, reversing the values placed along the top *x*-axis and the bottom *x*-axis. Whichever version you use, the plot works the same way. If the data remain inside the limits, the data are close to normally distributed. [↑](#footnote-ref-26)
27. We obtained these data from Wharton Research Data Service (WRDS). The data table contains quite a few other variables; some are categorical. Some variables have numeric values (e.g., TotalComp) and other variables (e.g., Industry) have character values. JMP procedures often treat these types differently. These data appear in SF, page 142. [↑](#footnote-ref-27)
28. To identify points by a variable, highlight the variable’s column and then select Cols > Label/Unlabel. Try it here with the variable CEO and then notice what happens when you point to the extreme value. [↑](#footnote-ref-28)
29. To exclude these points, first select them or their corresponding rows in the data table. Then select Exclude/Unexclude from the Rows menu. [↑](#footnote-ref-29)
30. To select these points, use Rows > Row Selection > Select Where, enter the condition “TotalComp is greater than 10,000,000.” Then choose Exclude/Unexclude from the Rows menu to exclude these cases from the analysis. You can also use Rows > Data Filter. [↑](#footnote-ref-30)
31. This is true for any base. [↑](#footnote-ref-31)
32. To do this in JMP, select Cols > New Column, name the new variable Log10TotalComp, and select Formula under New Property. In the Formula Editor Window, select Transcendental Functions and then Log10. Select TotalComp under Table Columns and click OK. [↑](#footnote-ref-32)
33. Again use Analyze > Distribution. Note that because Industry is a categorical variable (a column with text data that identify groups of observations), JMP automatically provides the summaries in terms of counts and proportions. [↑](#footnote-ref-33)
34. The plots on the left of these images are bar charts, counting the number of companies in each industry. A bar chart is used to summarize the variation in a categorical variable, such as the one telling us the industry for each CEO. The charts on the right are histograms. [↑](#footnote-ref-34)
35. The JMP software shows a bar chart for categorical data such as the industry codes, rather than a histogram. [↑](#footnote-ref-35)
36. The following plot is obtained using Analyze > Fit Y by X, selecting Log10TotalComp as Y and Industry as X. To show the boxplots, click the red triangle and select Boxplots under Display Options. The widths are proportional to the frequencies. The display shows quantile boxplots (see page 1-12). [↑](#footnote-ref-36)