**Module 1: Introduction – Data and Variation**

# What is Statistics?

Statistics, the discipline, is the art and science of extracting useful information from data.[[1]](#footnote-2)

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.

# 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.

# An Illustrative Example of a Statistical Analysis

Good analyses begin with a question such as, “How should I invest my money?” Once we have a question in mind, we can focus on finding *relevant* data.

Consider the choice between two stocks, say eBay and OSI Pharmaceuticals Inc. (OSIP).[[2]](#footnote-3) To make our choices more realistic, we’ll pretend that, based on the data available in 2003, we have to choose one of these two going forward.

Our data are daily stock prices and returns for these two companies in 2003. We arrange this data into a table.

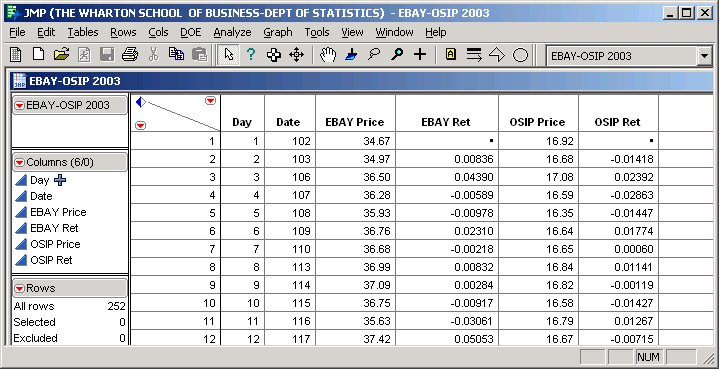
The *rows* in the table go by various names such as observations, cases, or even subjects. The table collects measured values of various attributes of the observations.

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 JMP software organizes and presents the data in a table.

We begin by clicking on the JMP file *EBAY-OSIP 2003.JMP* which contains daily prices and returns from 2003.[[3]](#footnote-4) The variables associated with the columns are listed on the left and repeated in the column headings.

The rows in this example denote trading days during 2003. With time series, the rows always denote the time of the measurement.



In the *EBAY-OSIP 2003.JMP* data, there are \_\_\_\_\_ observations on \_\_\_\_\_ variables.

JMP provides many useful interactive graphics. Let’s first use it to produce the following

time series plots of the daily prices.[[4]](#footnote-5)

When data form a time series, it’s best to look first for the effects of the passage of time by looking at a sequence plot. Think about the impression conveyed by this plot. What’s your initial impression of these stocks after looking at this graph?



Which of these investments would have been better to have held throughout 2003?

Could this plot be modified to make it more useful? How?

A much more challenging question is: based on this data which of eBay or OSIP would be a better investment for the future?

To answer this question, we look at the daily returns rather than the daily prices. Returns are essentially percentage changes without multiplying by 100.



As we will see later in the course, returns typically produce a much more stable process and so are more useful for prediction. It is much *simpler* to describe and summarize the variation in the returns rather than the variation in the original series.

JMP is easily used to obtain graphical comparisons of the two sets of returns. To facilitate comparisons, the pair of histograms on the right shows a common scale for both variables.[[5]](#footnote-6)

Histograms - show bars that represent counts of returns in an interval.

Boxplots - summarize the data further, drawing lines at certain ordered values.

Shape - what remains after you remove the value labeling on the axes.

How do the returns on the two assets compare? Which is riskier to hold?

JMP provides numerical summaries of the two sets of returns. These summaries *quantify* the features seen in the previous histograms.



Mean –the average of the returns.

St Dev – a measure of the variation of the returns.

How do the returns on the two assets compare? What sort of trade-off has to be made?

Is this type of trade-off an exception, or typical when comparing investments?

We have seen that compared to eBay, OSIP yielded a higher return on average, but was also more volatile.

As we will see in Modules 4 and 5, variation diminishes the long run value of an asset by an amount that we can quantify.

We will also see that a portfolio of partial investments in eBay and OSIP would provide a better expected long run return than a 100% investment in either one.

# But Don’t Worry

The main purpose of this example is to whet your appetite for what we will cover.

The methods and techniques illustrated here will all be carefully defined and motivated throughout the course.

We begin with graphical and numerical summary statistics in the next Module.

1. Any numerical summary of data, such as an average, is called a statistic. [↑](#footnote-ref-2)
2. eBay operates an online market place. OSIP discovers, develops, and markets anti-cancer products. [↑](#footnote-ref-3)
3. These data were downloaded from the Center for Research in Security Prices (CRSP) database on Wharton Research Data Services (WRDS) and then stored in the JMP file. The returns have been carefully adjusted for dividend payments and stock splits. [↑](#footnote-ref-4)
4. If you want to reproduce this figure, follow the menu commands: Graph > Overlay Plot and fill in the dialog. An option lets you define separate axes. [↑](#footnote-ref-5)
5. Use the Analyze > Distribution command to build histograms with JMP. The option of using the pop-up menu forces the plots to use a common scale. [↑](#footnote-ref-6)