1. INTRODUCTION

"Statistics is data analysis, together with everything you need to do data analysis." --John Tukey

"What is real, and what is an illusion?" --The Moody Blues

"Understand Variation." --W. Edwards Deming

**Our Main Goals:**

- To understand variation, which blurs the message in the data.
- To learn to assess risk.
- To learn to distinguish real patterns from illusory ones, using statistical inference.
- To analyze data, especially regression data, using statistical inference to draw conclusions (learn) from these data.

Statistics plays a key role in Business. A few examples:

**Accounting:** Measurement of Earnings Surprise  
**Economics:** Economic Forecasting; Seasonal Adjustment; Rational Expectations; Prediction Markets  
**Finance:** Portfolio Theory; CAPM; Predictive Regressions  
**Info. Systems:** Machine Learning; Data Mining  
**Management:** Psychometrics  
**Marketing:** Focus Groups, Market Research  
**Operations:** Quality Control; Supply Chain Management

Let's analyze some data, and then describe some basic concepts.

MarketReturn is the daily excess value-weighted return on all NYSE, AMEX and NASDAQ stocks. (Percent per day.)

The risk-free rate (one-month T-bill) has been subtracted.

Look at Minitab's Descriptive Statistics Graphical Summary.

- Market returns are not all the same: there is variation.

We can measure variation with the **standard deviation**.
Standard Deviations

1) For data $x_1, \ldots, x_n$ get sample mean $\bar{x} = (x_1 + \cdots + x_n) / n$ (measures the typical value).

2) Sample variance = Average squared deviation from mean
   
   $$s^2 = \frac{1}{n-1} \left\{ (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \right\}$$
   
   • $n-1 = \text{Degrees of Freedom:}$
     #Deviations needed to recover full data set.

3) Sample standard deviation $s = \sqrt{s^2}$ measures fluctuation (spread) around the mean.

Outliers

Outliers are atypical data points. (“Black Swan” effect.)

If a given event (e.g., catastrophic failure of an oil well blowout prevention device, “500-year” flood, product failure, etc.) has never happened before, some may assume that it won’t happen in the future: do black swans exist if we have never seen them?

The market returns may appear to follow a normal distribution (bell curve) but only if we ignore the outliers. And it’s very hard for an investor to avoid, say, the 10 worst days in market history.

The sample mean and sample standard deviation are very sensitive to outliers.
Probability

To manage risk, we try to calculate probabilities for future events. “Low-probability” events are not impossible.

Not paying attention to quality control and/or risk may save money in the short run, but eventually “Murphy’s Law” kicks in.

Probability is also a key ingredient in statistical inference. Our intuition about probability may need sharpening.

[Let's Make a Deal Demo]

http://www.stat.tamu.edu/~west/applets/LetsMakeaDeal.html

Populations and Samples

We will often assume that the data are a random sample from a larger population.

The incomes of \( n = 50 \) randomly-selected Stern MBA students would be a random sample. \( (n \) is the sample size).

A population may be finite (incomes of all Stern MBA students) or infinite (stock returns, as described by a probability distribution).

Descriptive measures of a population are called parameters.

Eg: Population Mean (\( \mu \)), Population Standard Deviation (\( \sigma \)).

Parameters are typically not observable.

Statistical Inference, Estimation

Descriptive measures of a sample are called statistics.
Eg: Sample mean (\( \bar{x} \)), sample standard deviation (\( s \)).

The value of a statistic will fluctuate from sample to sample. We must take account of this variability in order to draw valid conclusions from the given data set.

Statistical Inference is the process of drawing conclusions about populations based on random samples. (Learning from data).

Eg: If there is no global warming, then why did the seven hottest years on record occur in the most recent decade?

Eg: If 51% of a random sample of 2500 voters favor the Democrat And 49% favor the Republican, why is the election too close to call?

To do statistical inference, we think of a statistic as an estimate of a parameter: \( \bar{x} \) estimates \( \mu \), and \( s \) estimates \( \sigma \).

Then we ask how such estimates would behave under random sampling.

Here are some key properties, assuming random sampling.

The sample variance neither underestimates nor overestimates the population variance \( \sigma^2 \) on the average, that is:

- The sample variance is unbiased for \( \sigma^2 \).

(Another reason to use Degrees of Freedom = \( n - 1 \).)

- The sample mean is unbiased for \( \mu \).