

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Preliminaries

Syllabus

Required materials

Class Notes (full copy can be purchased from Wharton Reprographics)

BBS Casebook

JMP-IN 5.1.2 software (with manual/stat book)

Optional materials

Freedman, Pisani & Purves (FPP) - thoughtful background

Hildebrand, Ott & Gray (HOG) - formulas, calculation, worked examples

All information and other course material (including Class Notes) are posted on WebCafe.

Feedback

Assignments and quizzes - but no grades (yet!)

Office hours

Teaching assistants

Schedule and contact information is available in WebCafe

JMP practice sessions

Classroom Expectations - Concert rules!

Class starts on time

Sit according to the seating chart

Late entry or reentry is severely discouraged

Name tents displayed

All phones and electronic devices turned off

Course Organization

Part I Variation

Module 1 – Introduction: Data and Variation

Module 2 – Statistical Summaries of Data

Module 3 – Sources of Variation

Part II Models of Variation

Module 4 – Probability Models

Module 5 – Variance and the Volatility of Investments

Module 6 – Covariance and Portfolios

Part III Inference

Module 7 – Sampling, Sampling Distributions and Standard Errors

Module 8 – Confidence Intervals

Module 9 – Statistical Hypothesis Testing

Course emphasizes interpretation rather than computation. JMP software will do the tedious calculation.

Stat 603 provides the crucial foundation for Stat 621, the fall term core statistics course.

Other statistics courses for you to consider:

Stat 608 Waiver preparation

Stat 622 Continuation of Stat 621

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Module 1

Introduction: Data and Variation

What is Statistics?

Statistics, the discipline¹, is the art and science of extracting useful information from data

Two parts:

1) Exploratory Data Analysis - discovery (found it!)
summarization/description/exploration

2) 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, variation.

¹ Any numerical summary of data, such as an average, is called a statistic.

What are Data?

Basically, 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 GM each week
- ❖ the sales volume at Amazon.com each month

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

- ❖ the ages of the students in this class
- ❖ the preferences of potential automobile purchasers visiting a showroom
- ❖ the GMAT scores of the students in the front row

Key Feature: Variation

A common feature: virtually all data exhibit variation. The values are not all the same. Even though each value for age is measured in years, the measurement is likely to be different for each person.

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

An Illustrative Example of a Statistical Analysis

All 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 OSIP.² To make our choices more realistic, we’ll pretend we have to choose one of these two based on the data available in 2003. Our choice is to pick one of these going forward.

The data that we have measure the daily values of stock in these two companies in 2003. We will 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.

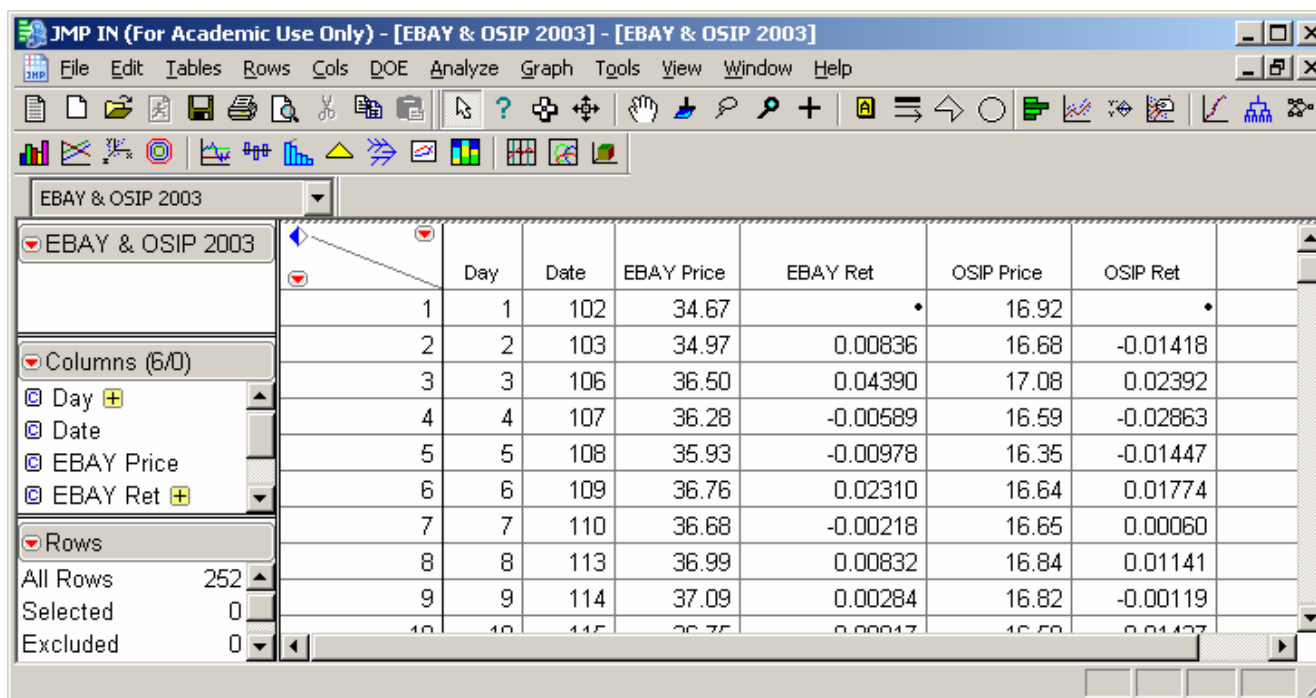
The *columns* in the table are called “variables.” Each column holds the value of some attribute of the observations that define the rows of the table.

The JMP software organizes and presents the data in a table.

² EBAY operates an online market place. OSIP discovers, develops and markets anti-cancer products.

We begin by clicking on the JMP file “EBAY & OSIP 2003.JMP”, which contains daily prices and returns from 2003.³ The variables are listed at the left and repeated in the column headings.

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



The screenshot shows the JMP IN software window titled "JMP IN (For Academic Use Only) - [EBAY & OSIP 2003] - [EBAY & OSIP 2003]". The main window displays a data table with the following columns: Day, Date, EBAY Price, EBAY Ret, OSIP Price, and OSIP Ret. The left sidebar shows the variable list under "Columns (6/0)" and "Rows". The "Rows" section indicates 252 All Rows, 0 Selected, and 0 Excluded.

Day	Date	EBAY Price	EBAY Ret	OSIP Price	OSIP Ret
1	102	34.67		16.92	
2	103	34.97	0.00836	16.68	-0.01418
3	106	36.50	0.04390	17.08	0.02392
4	107	36.28	-0.00589	16.59	-0.02863
5	108	35.93	-0.00978	16.35	-0.01447
6	109	36.76	0.02310	16.64	0.01774
7	110	36.68	-0.00218	16.65	0.00060
8	113	36.99	0.00832	16.84	0.01141
9	114	37.09	0.00284	16.82	-0.00119
10	115	36.75	0.00017	16.59	-0.01437

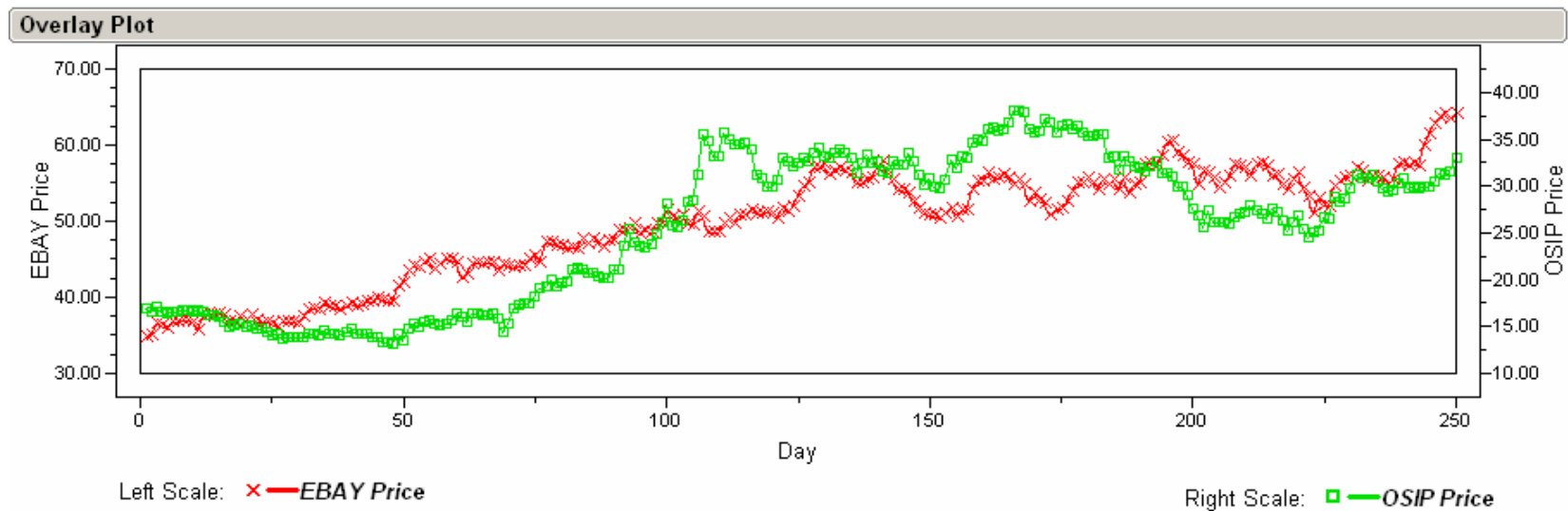
In the EBAY & OSIP 2003.JMP data, there are observations on variables.

³ These data were downloaded from the CRSP database on WRDS, and then stored in the JMP file.

JMP provides many useful interactive graphics. Let's first use it to produce the following time series plots of the daily prices.

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.

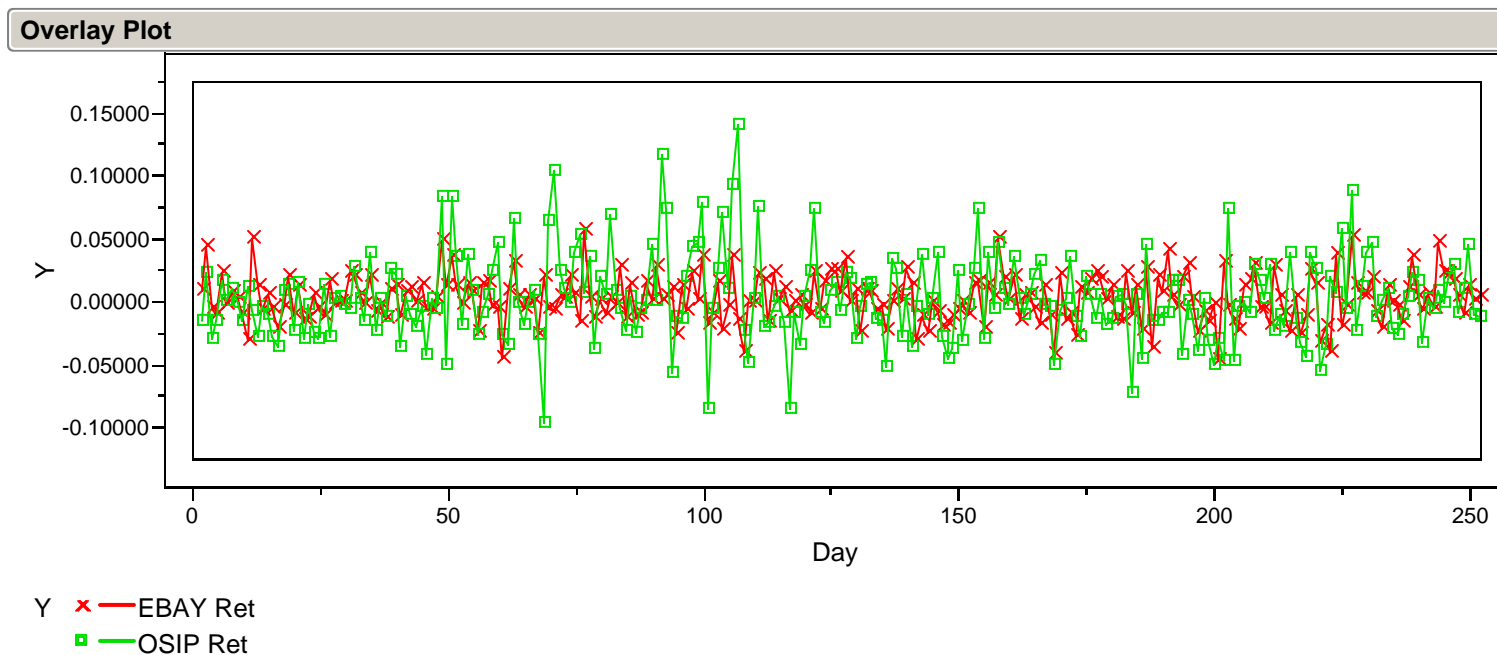
Would this plot be more useful with a common axis for both series?



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

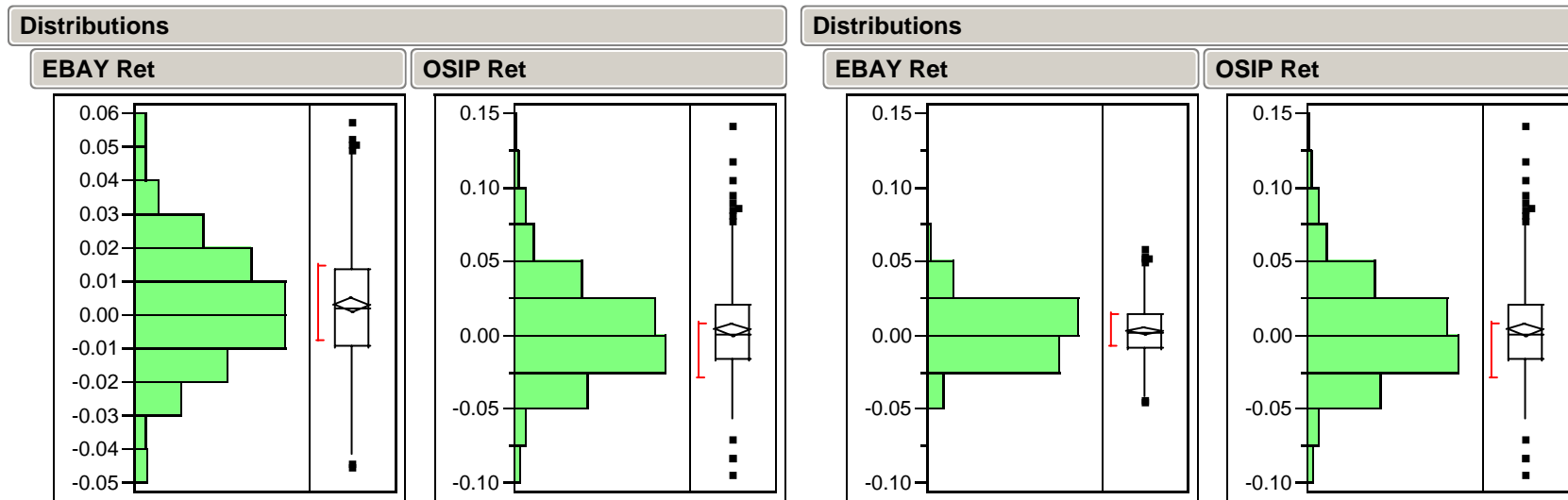
A much more challenging question is, based on this data which of EBAY or OSIP would be better investment for the future?

To answer such a question, we look at the net returns rather than the prices.



As we will see later in the course, the returns are 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 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 use a common scale for both variables.



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?

In Finance, the variation in the returns measures the *risk* of owning an asset.

JMP is also easily used to obtain numerical summaries of the two sets of returns. These summaries can be used to *quantify* the differences seen in the previous histograms.

Distributions					
EBAY Ret			OSIP Ret		
Quantiles			Quantiles		
100.0%	maximum	0.0572	100.0%	maximum	0.1405
99.5%		0.0558	99.5%		0.1343
97.5%		0.0463	97.5%		0.0843
90.0%		0.0250	90.0%		0.0453
75.0%	quartile	0.0135	75.0%	quartile	0.0202
50.0%	median	0.0015	50.0%	median	0.0000
25.0%	quartile	-0.0097	25.0%	quartile	-0.0162
10.0%		-0.0206	10.0%		-0.0341
2.5%		-0.0352	2.5%		-0.0537
0.5%		-0.0456	0.5%		-0.0921
0.0%	minimum	-0.0459	0.0%	minimum	-0.0950
Moments			Moments		
Mean		0.0026496	Mean		0.0031388
Std Dev		0.0183532	Std Dev		0.0339757
Std Err Mean		0.0011584	Std Err Mean		0.0021445
upper 95% Mean		0.0049311	upper 95% Mean		0.0073624
lower 95% Mean		0.000368	lower 95% Mean		-0.001085
N		251	N		251

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?

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 5 and 6, such 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.

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Module 2
Statistical Summaries of Data

Why summarize data?

Summarization can be very useful in forming a decision. In order to make use of data, we can often make a better choice if we can summarize the information.

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

Graphical and numerical statistical summaries narrow focus to essential features of data.

To motivate and illustrate such methods, we'll study three data sets using JMP.

GMAT Scores

(BBS, p. 9)¹

The analysis of data depend on the nature of your question. Managers have different questions, and so will often need to approach the task differently.

For this example, consider the sorts of questions that might be asked about GMAT scores at Wharton.

Student

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

Vice-Dean

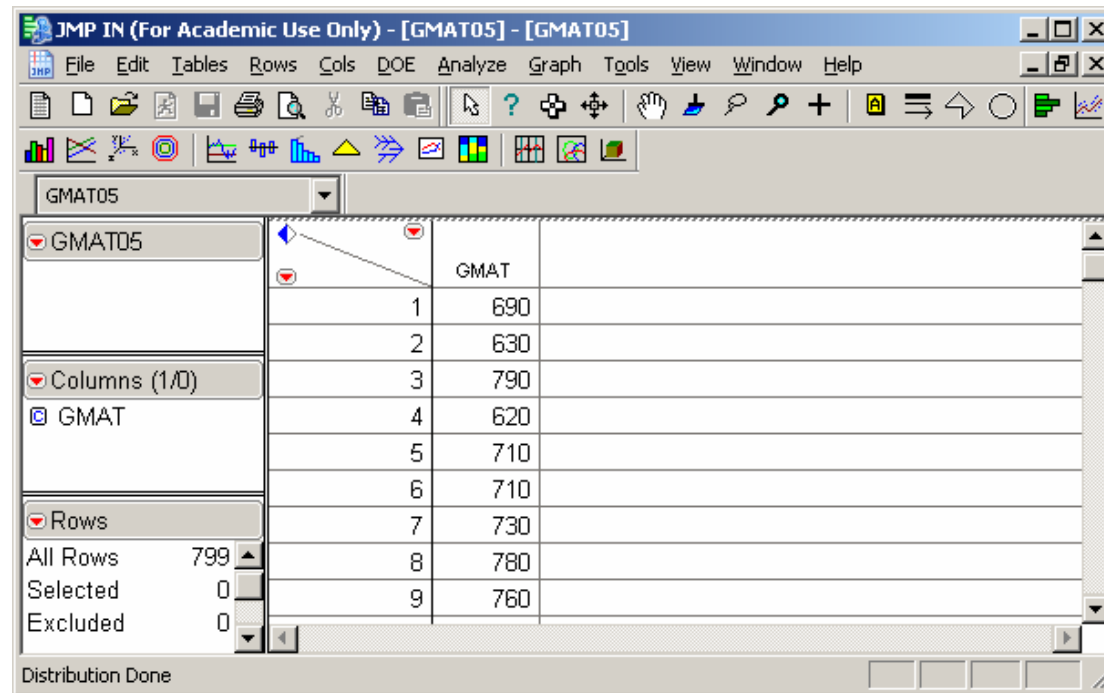
How does the entering class compare to those at other programs?

How have scores trended over time?

Effective statistical summarizes of data like these can make it easy to answer both types of questions, much more precisely and concisely than from the raw scores themselves.

¹ Many of the examples that we use in these notes appear with further discussion in the Basic Business Statistics casebook. In some cases, the casebook uses the same data that we consider in class, and in others (like this one), the casebook considers a similar data set.

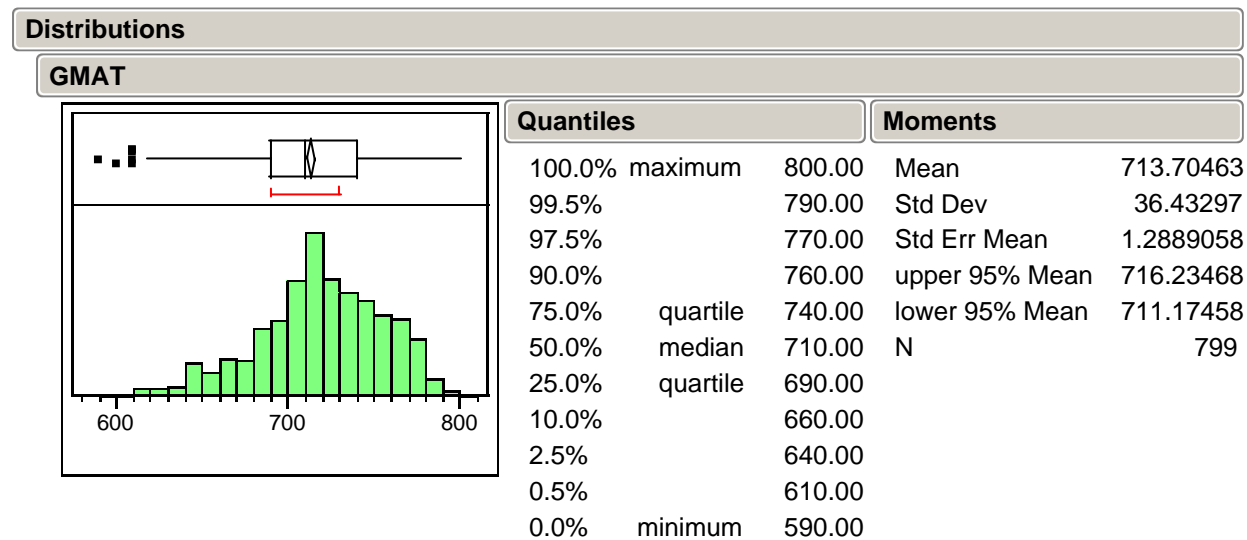
The first column of the Excel file GMAT05.xls consists of the GMAT test scores of 799 members of the Wharton MBA Class of 2005. That's one variable with 799 observations, one on each of these students. Let's open this file with JMP²



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

² Open GMAT.xls with the JMP command File > Open (select file type Excel files (*.xls)). By saving this file as GMAT05.JMP, you can then open it by simply clicking on the GMAT05.JMP file.

JMP provides graphical and numerical summaries of the 799 observations of the variable named GMAT.³



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 summarizes to the picture.

³ Use Analyze > Distribution with GMAT selected as the Y column. To get the horizontal layout shown next, right click on GMAT and then 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" downwards on the histogram.

Moments

Moments⁴ are average quantities of interest.

The mean and standard deviation are the two most basic and commonly used moments.

The mean is the average of the data values for a variable, namely

$$mean = \frac{\text{sum of all values}}{n}$$

where we typically use n to stand for the number of values.

From the JMP summary for the GMAT data, the mean =

What aspect of the data is captured by the mean?

⁴ 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.

The variance is the average squared deviation from the mean. It is calculated as follows:⁵

1. For each observation x , calculate $(x - \text{mean})$, the deviation from the mean of the data
2. Now square each deviation: $(x - \text{mean}) * (x - \text{mean})$
3. Add them and divide by⁶ $(n - 1)$

$$\text{variance} = \frac{\text{sum of all squared deviations}}{n - 1}$$

For the GMAT data, the variance =

What aspect of the data is captured by the variance?

What units ought to be attached to the variance?

⁵ Even though Excel has a built-in formula for the variance, it might be a good idea to work these calculations out directly once or twice to appreciate all of the calculations that the computer software will be doing. Done directly, you'll need several columns for the calculations.

⁶ Why divide by $(n - 1)$ rather than n ? An honest answer requires some details that we will not come to until later. For now, just think of the variance as the average squared deviation.

The standard deviation (st dev or SD) is the square root of the variance,

$$st\ dev = \sqrt{variance}$$

From the JMP summary for the GMAT data, the st dev =

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

The standard deviation is often more understandable. Why is this so?

Mini-quiz: Which of the following variables will have a larger variance and standard deviation? Why?

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

Statistical Notation

Suppose we have n data values of the variable x . We use the notation

$$x_1, \dots, x_n$$

to stand for the n data values. The subscript or index shown with the name of the variable, $x_{i \leftarrow}$, identifies the i th measurement of the variable.⁷

For example, if we let x_1, \dots, x_n stand for the GMAT data, then $x_1 =$, $x_2 =$,
 $x_{799} =$, and $n =$

For any function $f(x)$, summation using this notation (and the Greek Letter Σ) is symbolically abbreviated as

$$\sum_{i=1}^n f(x_i) = f(x_1) + f(x_2) + \dots + f(x_n)$$

⁷ The subscript letter is not always i . For example, the letter t is often used for time series data. Thus x_t would be the observed value at time t .

Notation and Formulas for Sample Statistics

The (Sample⁸) Mean

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

The (Sample) Variance

$$s^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The (Sample) Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

⁸ We'll have a lot more to say about samples and populations later in Lecture 7. For the moment, a sample is another name for a set of data, such as the collection of GMAT scores.

Quantiles

Quantiles are derived from the ordered data values. The k% quantile (also called the kth percentile) is a value that separates the bottom k% 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 commonly used quantiles

median =	percentile
first quartile Q1 =	percentile
third quartile Q3 =	percentile

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

An alternative to the standard deviation as a measure of spread is the Interquartile Range, defined as $IQR = Q3 - Q1$ ⁹

What aspect of data is captured by the IQR?

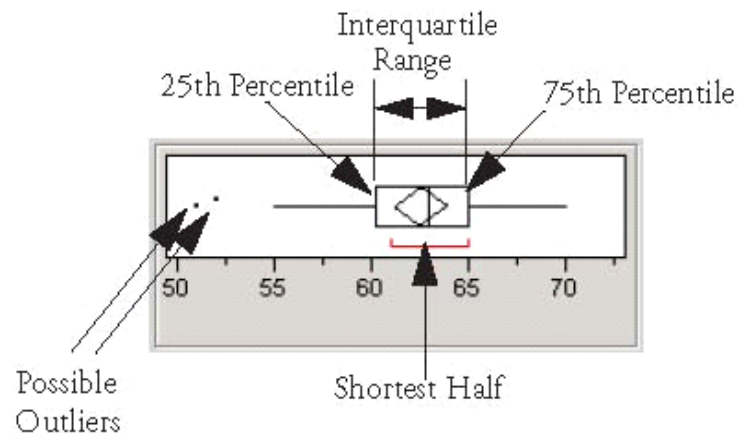
⁹ Some schools summarize the range of GMAT scores in the form of the 20% and 80% values. That range is almost the Interquartile range, which is just the gap between the 25% score and the 75% score.

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 (BBS, p.12)



Remark: JMP calls this an “outlier boxplot”. JMP also provides a “quantile boxplot” which is similar.

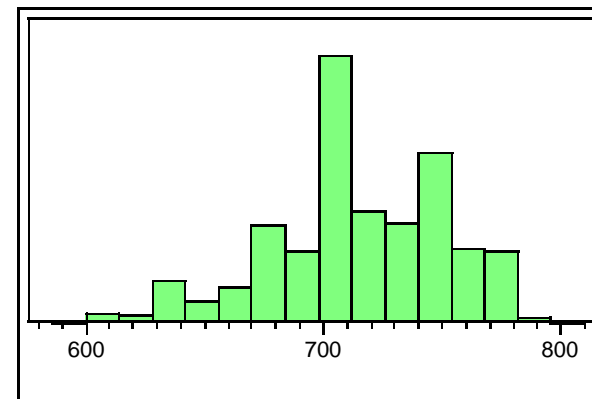
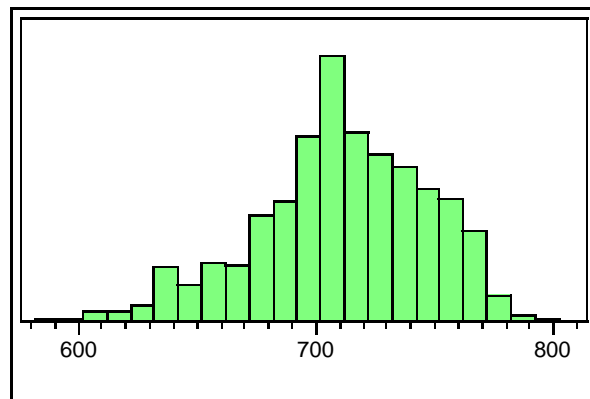
Histogram

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.

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.¹⁰

What is gained and lost by decreasing the number of intervals? Are the modes found in the histogram on the right meaningful, or just artifacts of a subjective choice?



¹⁰ In JMP, this can be done *interactively* with the hand tool.

You can learn a lot about a variable by looking at its histogram:

- The median is the point where half the area is to the left and half to the right.
- The mean of the data is the “balancing point” on the histogram. A piece of wood of constant thickness shaped like the histogram would balance at the mean.¹¹
- Skewed distributions are asymmetric: the mean will not equal the median.
- If the data is skewed left, then $\text{median} > \text{mean}$.
- If the data is skewed right, then $\text{median} < \text{mean}$.

What does a histogram “hide”, particularly when looking at the variation in a time series such as the returns in Module 1? When is it OK to hide such details?

¹¹ 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.

Prices of GM Stock

(BBS, p. 23)

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

The Excel file GM.xls contains daily closing prices on a share of GM stock. The worksheet GM92 contains 1992-93 data and GM87 contains 1987-88 data.

Let's use the JMP command Open to select GM92 and study the variable Price.

JMP IN (For Academic Use Only) - GM92

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

GM92

GM92

	Day	Month	Year	Price	RelChange	Time	RiskFree
1	2	1	1992	31	0.0735931	92.0028	31
2	3	1	1992	32.5	0.0483871	92.0056	31.005828
3	6	1	1992	33.5	0.0307692	92.0139	31.0116571
4	7	1	1992	33	-0.0149254	92.0167	31.0174873
5	8	1	1992	32.125	-0.0265152	92.0194	31.0233186
6	9	1	1992	32.125	0	92.0222	31.029151
7	10	1	1992	31.75	-0.0116732	92.025	31.0349844
8	13	1	1992	31	-0.0236221	92.0333	31.040819
9	14	1	1992	32	0.0322581	92.0361	31.0466547
10	15	1	1992	32.125	0.00390625	92.0389	31.0524915
11	16	1	1992	34.375	0.0700389	92.0417	31.0583293
12	17	1	1992	33.875	-0.0145455	92.0444	31.0641683
13	20	1	1992	33.875	0	92.0528	31.0700084

Columns (7/0)

- Day
- Month
- Year
- Price
- RelChange
- Time

Rows

All Rows 507

Selected 0

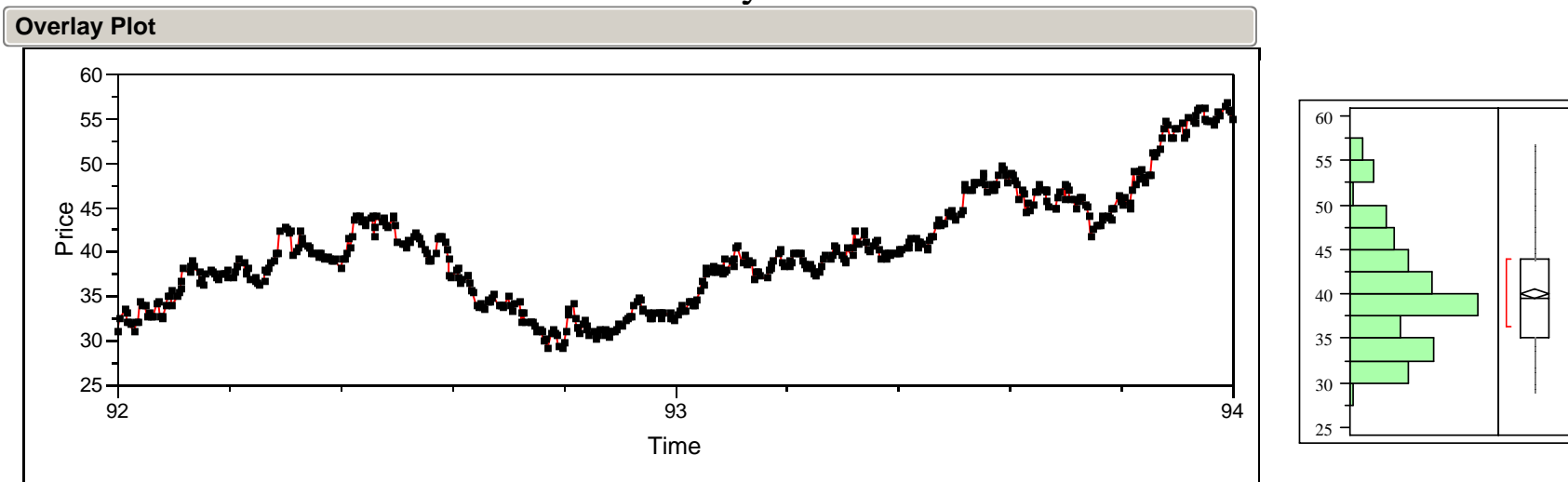
Excluded 0

Hidden 0

Ready

Because the prices form a time series (i.e. a sequence of observations that are ordered in time), we begin with a time series plot (BBS, p. 24-25).¹²

GM Daily Price 1992-93



Note the meandering behavior¹³ of this time series. Successive draws do not stray too far from each other.

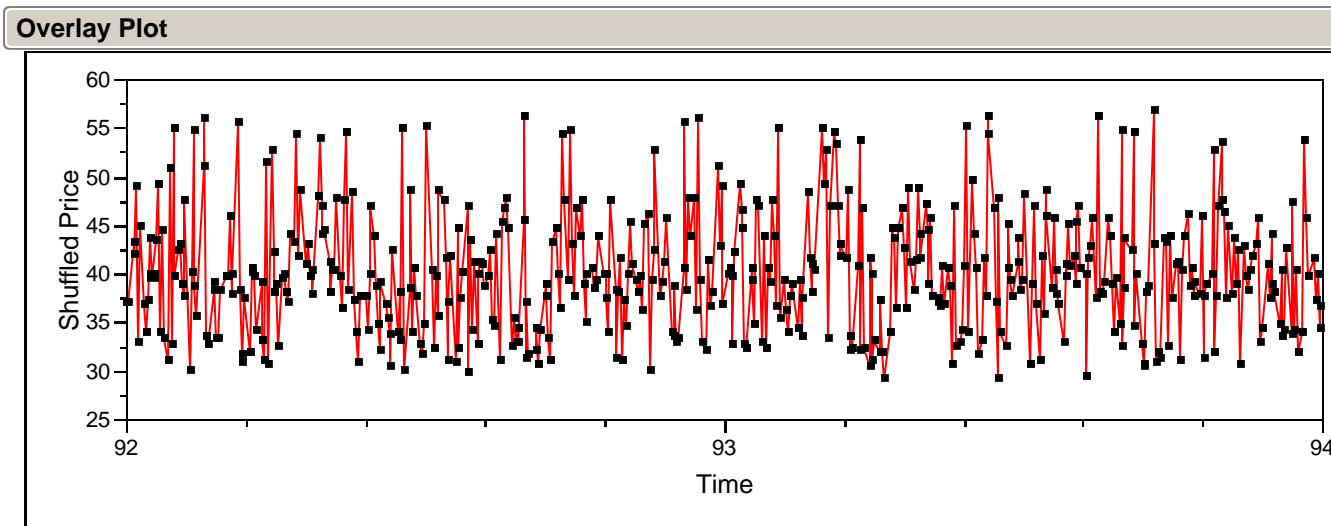
This is a manifestation of sequential *dependence*. Does the histogram provide an adequate summary of a variable with this sort of dependence? What does it hide?

¹² This time series plot was obtained with the command Overlay Plot. Select Price as Y and Time as X and right click on Overlay Plot bar to select Y Options > Connect Points.

¹³ 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.

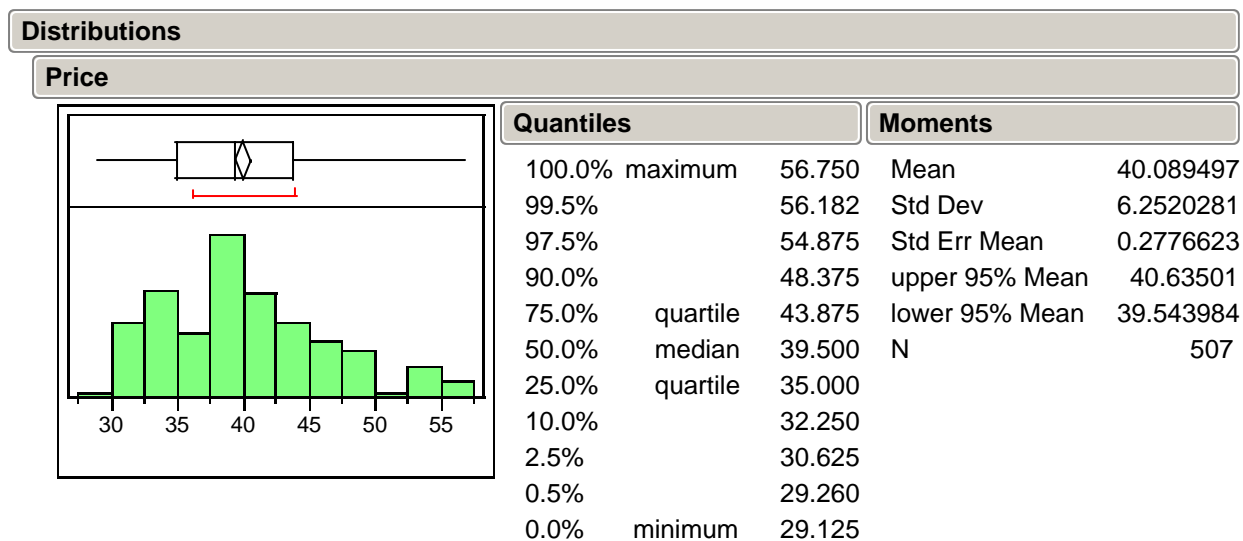
A sequence of independent random draws from a population, i.e. a random sample, would not manifest such meandering behavior. The ordering would not affect our understanding of the variance of the data.

To get some insight into what a sequence of independent draws might look like, let's randomly shuffle the order of Price.¹⁴ A time series plot of Shuffled Price looks like



The sequential variation patterns of Price and Shuffled Price are so different, it is implausible that the Price values are a sequence of independent draws from a population.

¹⁴ 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.



Note that *both* the histogram and summary statistics conceal the meandering behavior.

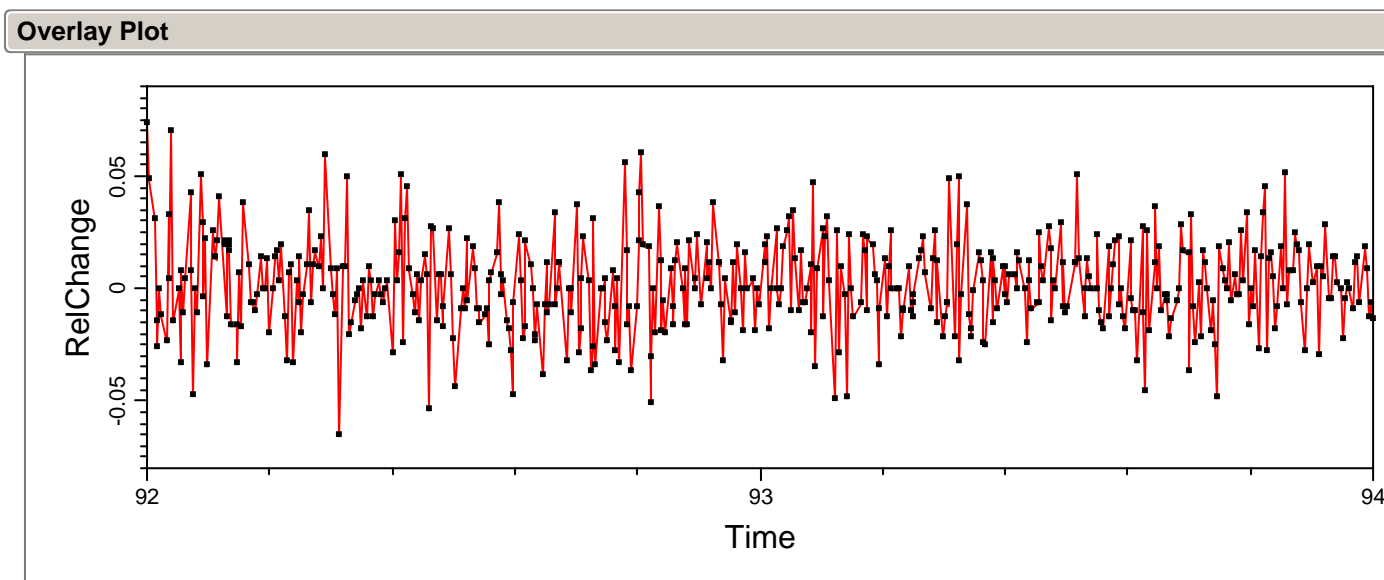
If you only saw this output, you wouldn't be able to recover the sequential dependence that is such a dominant feature of the time series plot.

As a summary of the variation, a histogram is misleading. This histogram does not distinguish between a plot of data without patterns like that on page 2-16 (after scrambling the order), and the smooth, meandering patterns seen in the sequence plot of the returns on page 2-15.

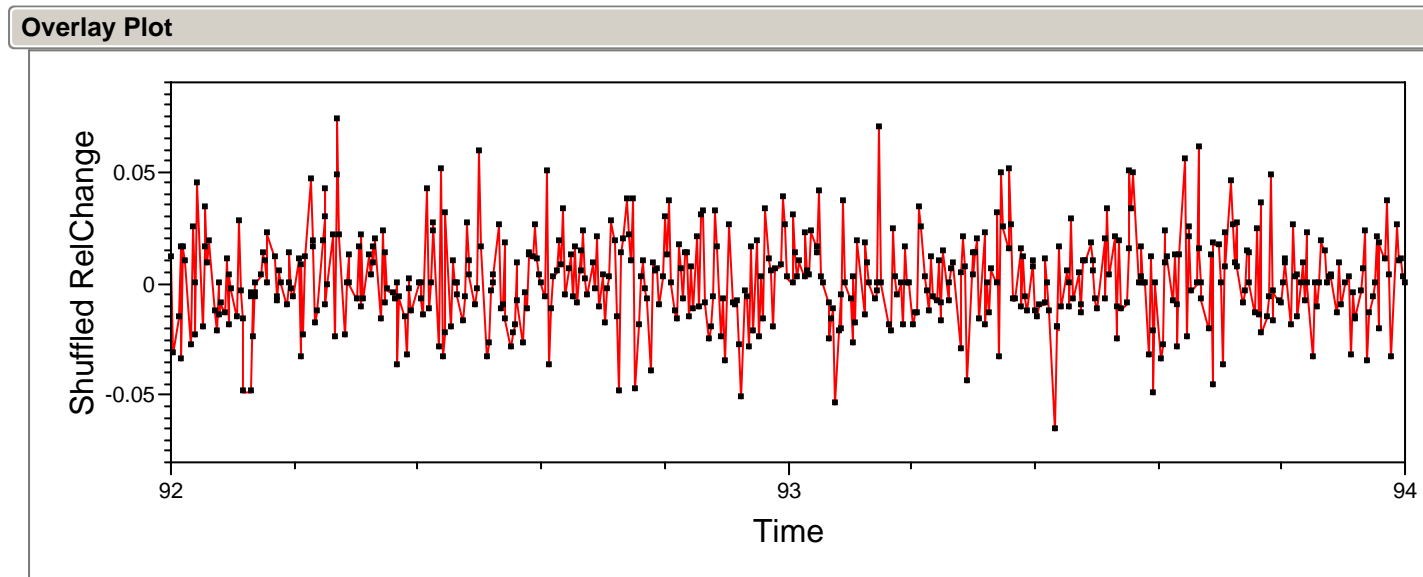
Letting p_t denote the price of GM at time t , a different perspective is obtained by focusing on the successive daily relative changes or net returns (BBS, p 26-27)

$$\frac{p_t - p_{t-1}}{p_{t-1}}$$

The time series plot of RelChange now looks much more like a sequence of independent draws from a population.

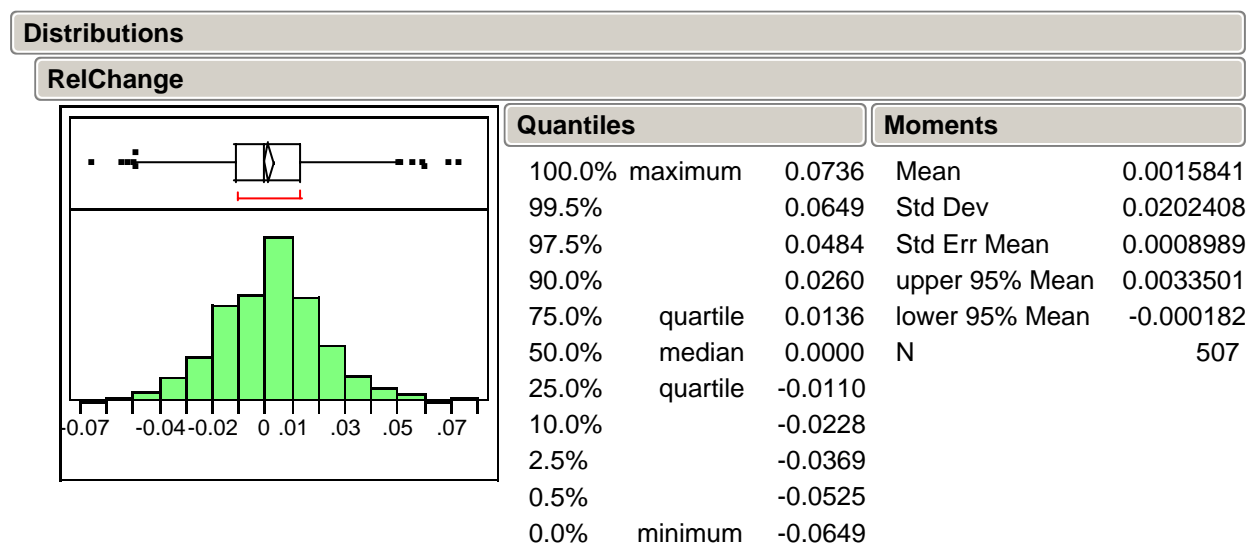


This is confirmed by the following time series plot of the randomly shuffled values of RelChange.



Note the similarity of the variation patterns of RelChange and Shuffled RelChange.

And since RelChange lacks the sequential dependence seen in Price, the histogram is not hiding much from us.¹⁵

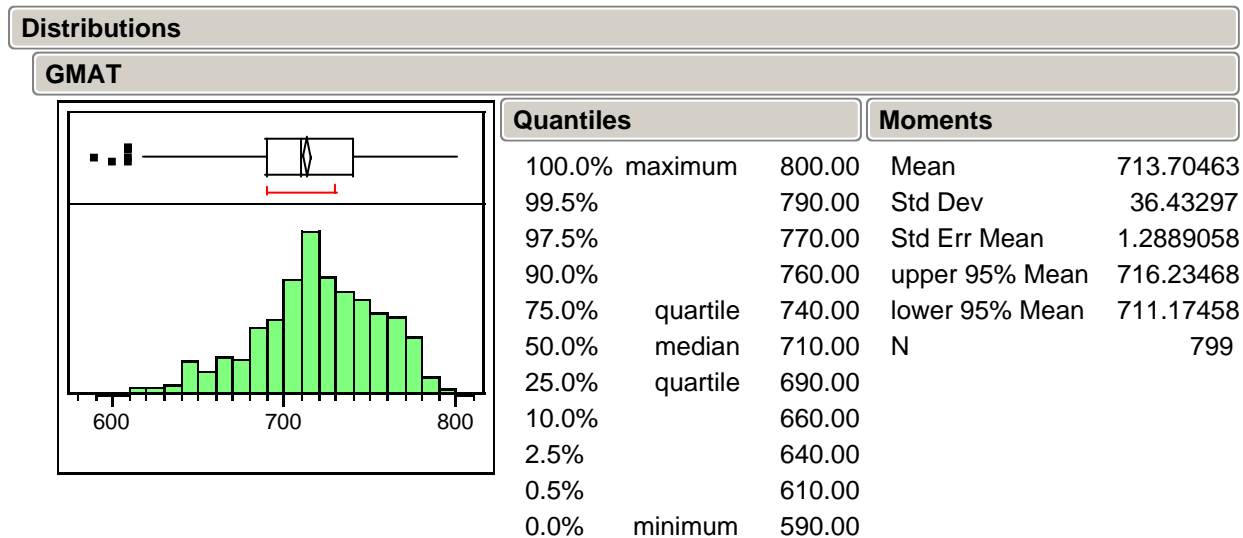


The GM92 histogram shows that the data appears to have a “bell-shaped” distribution in the middle. This is what you would probably see when sampling¹⁶ from a *normal population*.

¹⁵ Of course, 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?

¹⁶ 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).

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

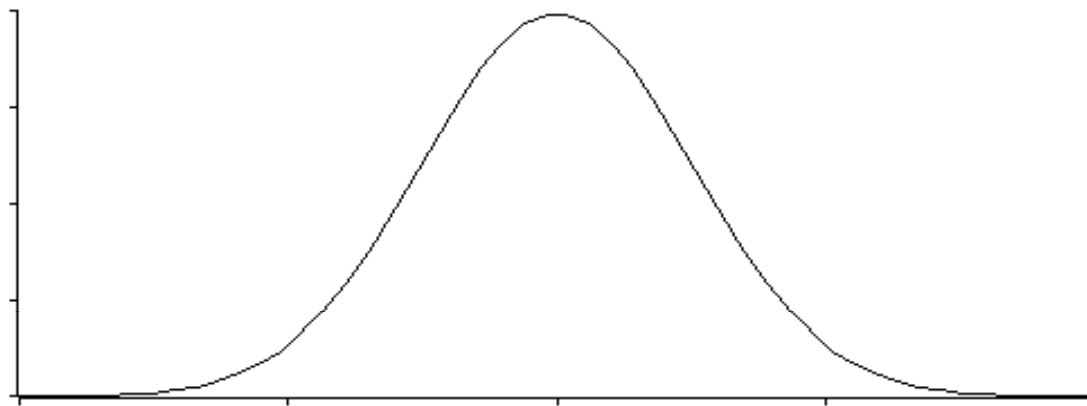


The Normal Distribution

The normal distribution is the ideal “bell-shaped” distribution.

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 refined histogram of a very, 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) of the area lies within $(\mu - \sigma, \mu + \sigma)$
- About 95% (actually .9545) of the area lies within $(\mu - 2\sigma, \mu + 2\sigma)$
- Almost 99.7% (actually .9973) of the area lies within $(\mu - 3\sigma, \mu + 3\sigma)$

Note that the standard deviation σ is a natural unit of distance for the normal distribution. These values are also tabled in the text, BBS p. 17.

Normality of Data

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

The relative frequency of values in the interval

The probability that a randomly drawn observation falls within the interval

Let us see how well the normal distribution approximates the GM92 data. Here, $\bar{x} \approx .00158$ and $s \approx .02024$ so that

$$(\bar{x} - s, \bar{x} + s) \approx (-0.0187, 0.0218)$$

$$(\bar{x} - 2s, \bar{x} + 2s) \approx (-0.0389, 0.0421)$$

$$(\bar{x} - 3s, \bar{x} + 3s) \approx (-0.0591, 0.0623)$$

For each of these intervals, it turns out that

$366/507 \approx 72.19\%$ of the observations fall within s of \bar{x}

$481/507 \approx 94.9\%$ of the observations fall within $2s$ of \bar{x}

$504/507 \approx 99.4\%$ of the observations fall within $3s$ of \bar{x}

Key point

This approximate agreement between the normal distribution and data is sometimes called the *empirical rule*. Since the empirical rule seems to hold here, we can concisely summarize the 507 relative changes observed in GM92 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.

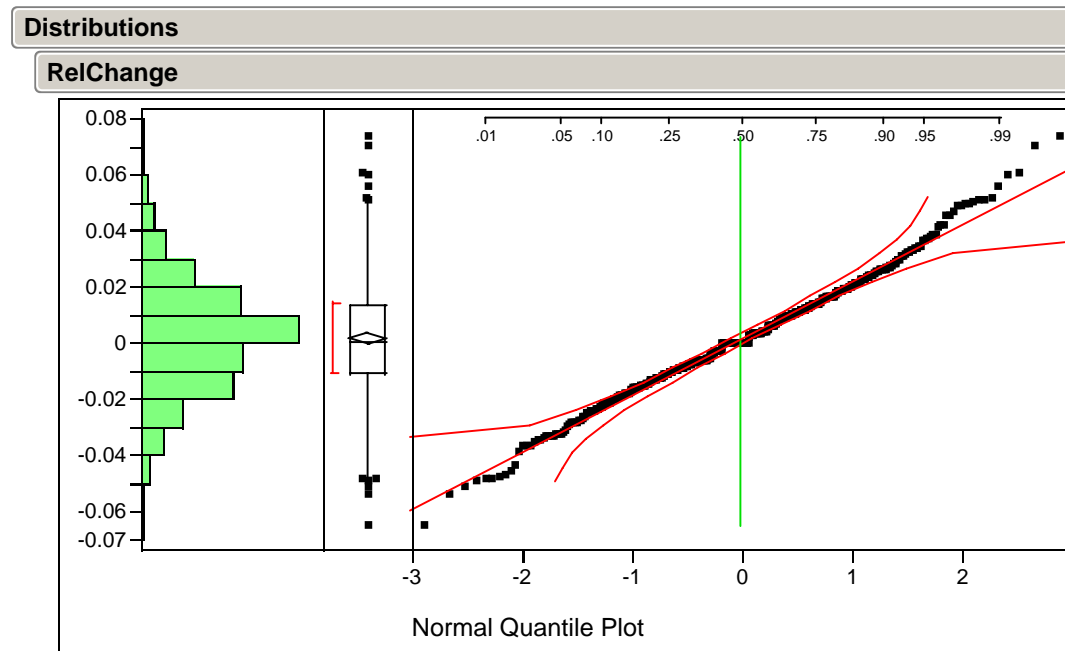
Why this bell-shape and not another?

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 is the Central Limit Theorem (CLT), to be discussed in Module 7).

Normal Quantile Plot

Normality is a powerful assumption; we need a diagnostic to check for normality that works better than just looking at the histogram. Most of our diagnostics come in the form of a plot that requires us to *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.¹⁷



¹⁷ To generate a quantile plot in JMP, you first need to be looking at a histogram. Then use right click on the RelChange bar and select Normal Quantile Plot. Select Fit Distribution > Normal to get the normal overlay curve on the Histogram.

How it works:

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 (BBS, p. 18-19).

What conclusion would you draw about the GM92 data (BBS, p 28)?

Why it works:

The y-axis of the normal quantile plot shows the values of the variable.

The x-axis shows the distance from μ in units of σ for the ideal normal quantile associated with each value.

Under this rescaling of the x-axis, a “perfect” normal sample would fall along the straight line $y = x$.

What causes the small anomaly around 0 in the GM92 normal quantile plot? Did you notice this feature of the data in other plots, or did it only become apparent now?

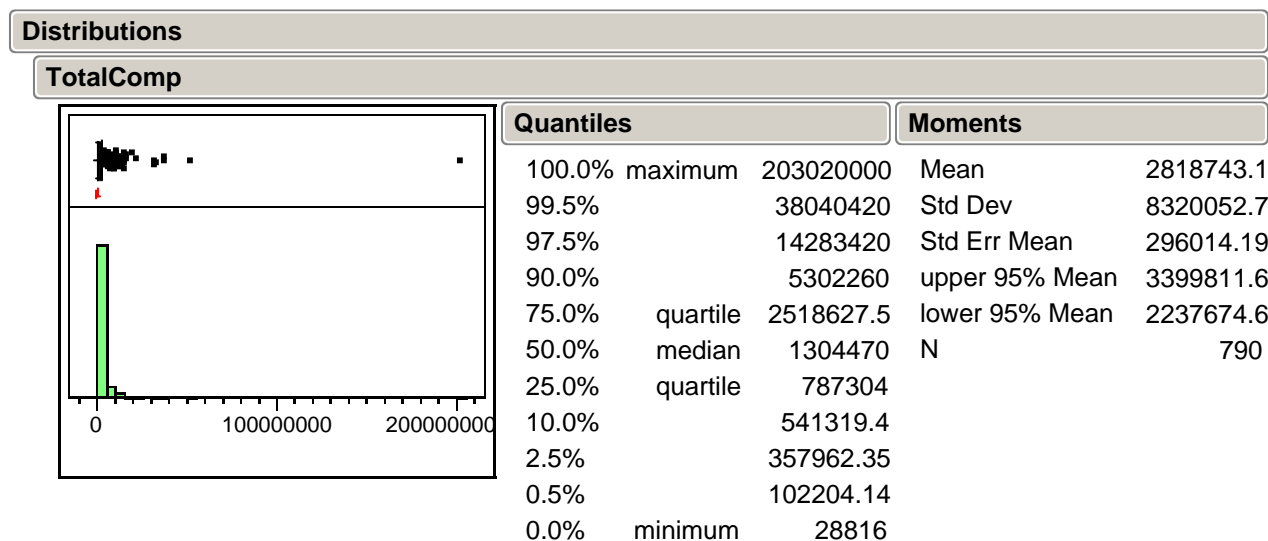
Forbes Executive Compensation Data

(BBS p.34)

Not all data is normally distributed.

The Excel file Forbes94.xls contains the annual executive compensation of 790 executives as reported in the May 23, 1994 issue of Forbes.

The JMP summary of TotalComp (total compensation) is obtained as¹⁸

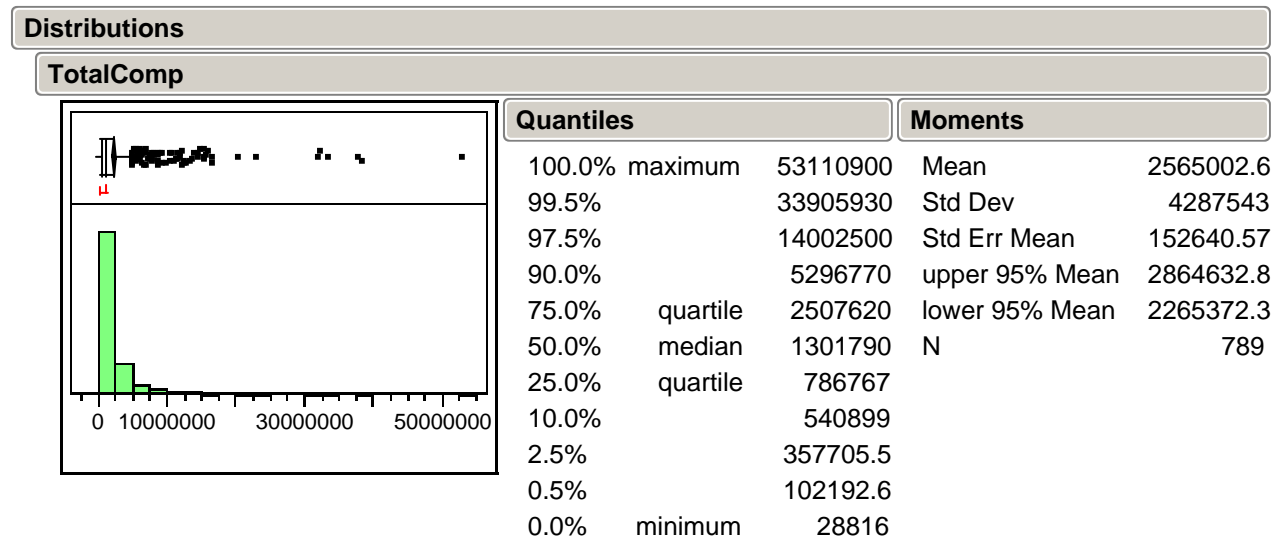


Someone made \$203,020,000! Who was it? (BBS, p.35)

¹⁸ When we open Forbes94.xls with JMP, note that some variables such as TotalComp have numeric values and some have label values such as WideIndustry. JMP denotes the former with a “C” for continuous and the latter with an “N” for Nominal. As we will later see, JMP procedures often treat each of these types in different ways.

This extreme value dominates the output and makes it difficult to see the rest of the distribution.

Let's exclude this maximum value and see what happens.¹⁹



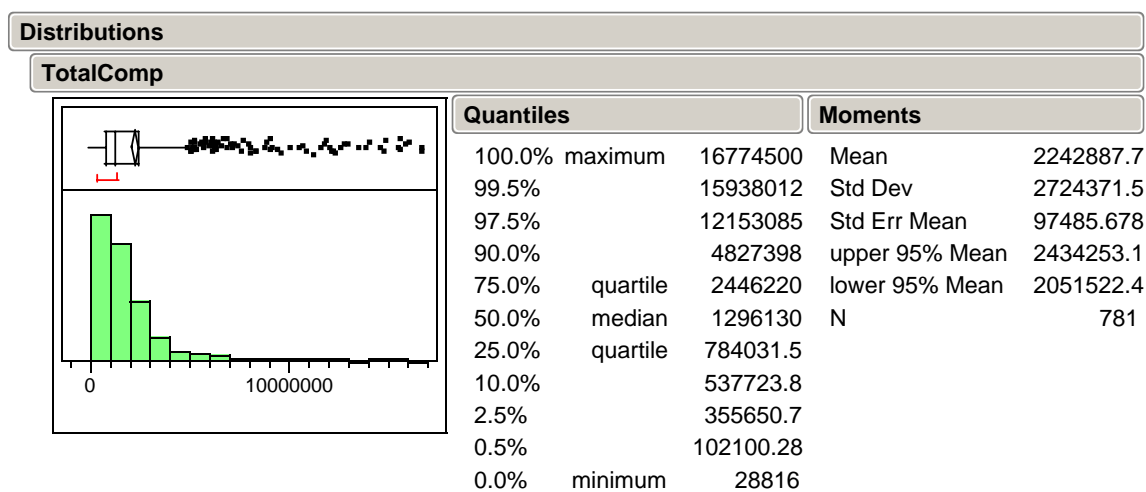
That didn't seem to help much. A small number of compensations continue to dominate the plot. (BBS p.36-37)

¹⁹ To exclude this point, first select it or its corresponding row in the data table. Then select Exclude/Unexclude from the rows menu.

Note, however, that the mean dropped from \$2.82 million to \$2.57 million and the standard deviation was nearly halved to \$4.3 million. The one large salary was very influential upon these summaries.

The size of the effects illustrates how the mean and especially the standard deviation can change dramatically by adding or removing a single point that is far away from the others. The median, in contrast, will not change much.

Let's remove the next largest 8 salaries²⁰ and see how things change...



What's happening?

²⁰ To select more than one point, hold down the shift key when selecting from the graph, or hold down the ctrl key when selecting rows from the data table. Then select Exclude/Unexclude from the rows menu.

Transforming the Data

An alternative way of looking at the big picture is to transform the data to a different scale. In this case, let's consider the log transformation

$$y = \log_{10} x$$

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

Note that $\log(y/x) = \log y - \log x$ so that ratio increases in the original scale correspond to additive increases on the log scale²¹. Thus

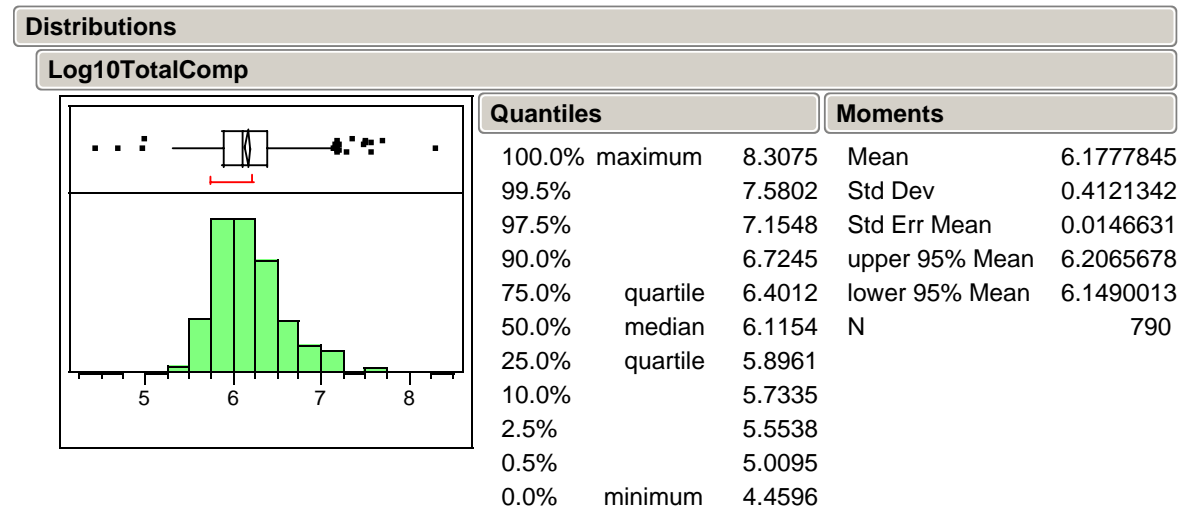
changes on a log scale are related to percentage changes.

Question: Which is larger,

$$[\log \$11,000 - \log \$10,000] \quad \text{or} \quad [\log \$1,001,000 - \log \$1,000,000] ?$$

²¹ This is true for any base.

For the Forbes data, let's create the new variable²² $\text{Log10TotalComp} = \log_{10} \text{TotalComp}$
A summary of these log of the compensation values:



By transforming to the log scale, we can more easily get a picture of the distribution of compensation values.

Is any information lost by transforming the data?

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

²² 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.

Take-Away Review

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

Normality and the shape of a histogram

Characterized by a mean and standard deviation

Use of normal quantile plot

Transforming data to remove skewness

Next Module

Using statistical methods to understand sources of variation.

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

**Module 3
Sources of Variation**

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 is causing the data to vary, we might be able to control this variation to our benefit.

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.”

Back to the Forbes Compensation Data

(BBS, p.45)

A question of interest: Are executives in some industries more highly paid than those in others? Do we want to find the ‘stars’ or identify industries that pay higher salaries on average?

We’ll use FrbSubt.jmp which restricts the data to a subset of 10 industry categories

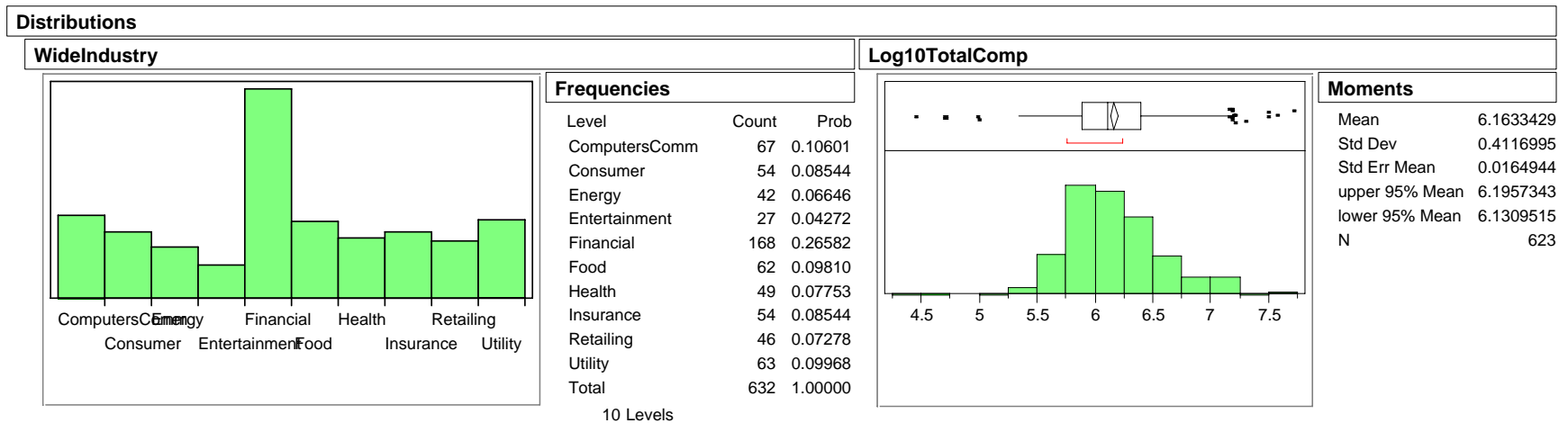
Let’s first obtain summaries of the two variables WideIndustry and Log10TotalComp¹

Goal is to compare compensation among industries.

JMP summaries reveal the variation of each of these variables.

Should we use the “raw” compensation, or use compensation on the log scale?
Which question are we trying to answer?

¹ Again use Analyze > Distribution. Note that because WideIndustry 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

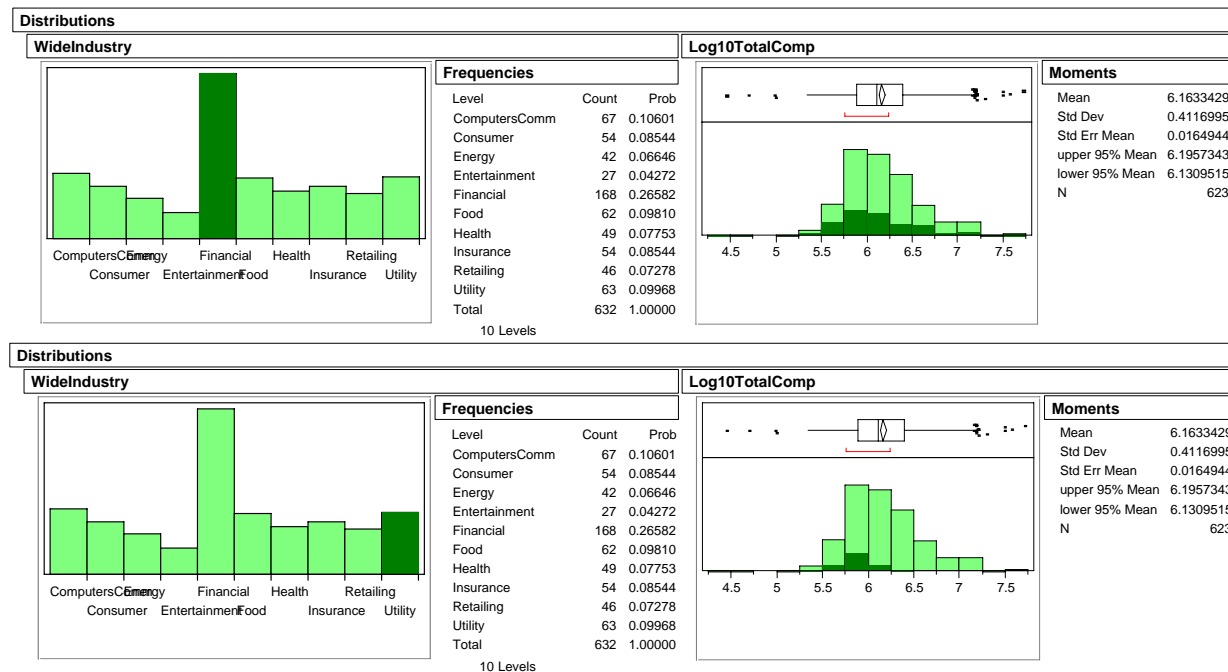


What about the relationship *between* industry category and total compensation?

Can we tell from these summaries if executives in some industries consistently more highly paid than those in other industries?

Plot Linking

To get some sense of the relationship between them, we can highlight industries such as Financial or Utility.² *Plot linking* in JMP highlights the associated values of Log10TotalComp.



Does it look like Financial or Utility executives are paid differently? Are the differences small, happening just for one or two, or more general?

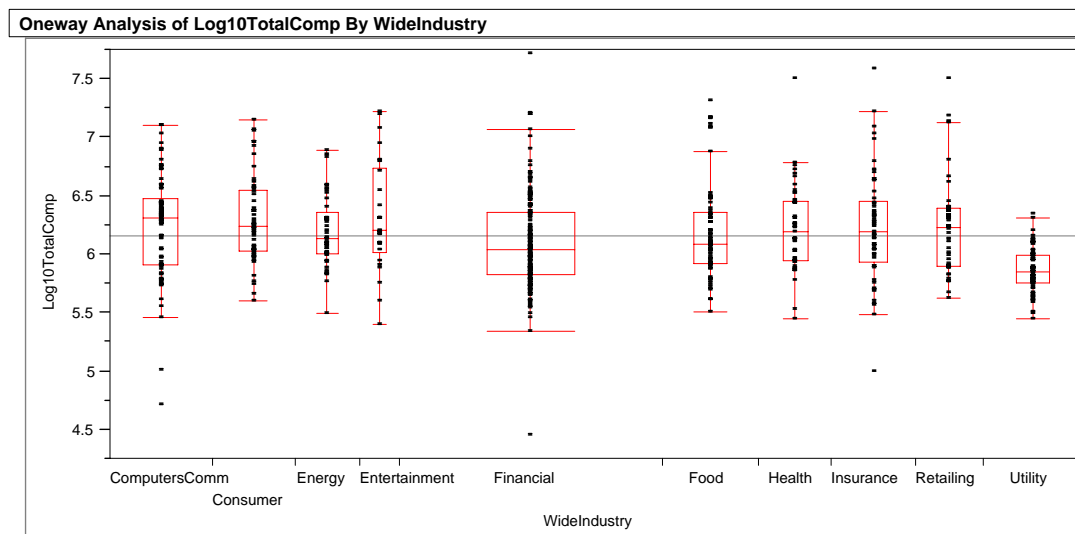
² The plots of the left of these images are bar charts, counting the number of companies in each industry. A bar chart is used to summarize the variation in a categorical variable, such as the one telling us the industry for each CEO. The charts on the right are histograms.

Comparison Boxplots

Another useful approach for comparison is to look at *comparison boxplots*.

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.³ What can we say about the executive compensation differences across these industries? (BBS, p.48)



³ The following plot is obtained using Analyze > Fit Y by X, selecting Log10TotalComp as Y and WideIndustry as X. To show the boxplots, right click on the title bar and select Boxplots under Display Options.

Predicting Airline Passenger Demand

(BBS, p.50)

A question of interest: How might one go about predicting future demand in a growth industry?

By understanding the variation in demand, we can not only

- (a) get a useful prediction of future demand,

we can also

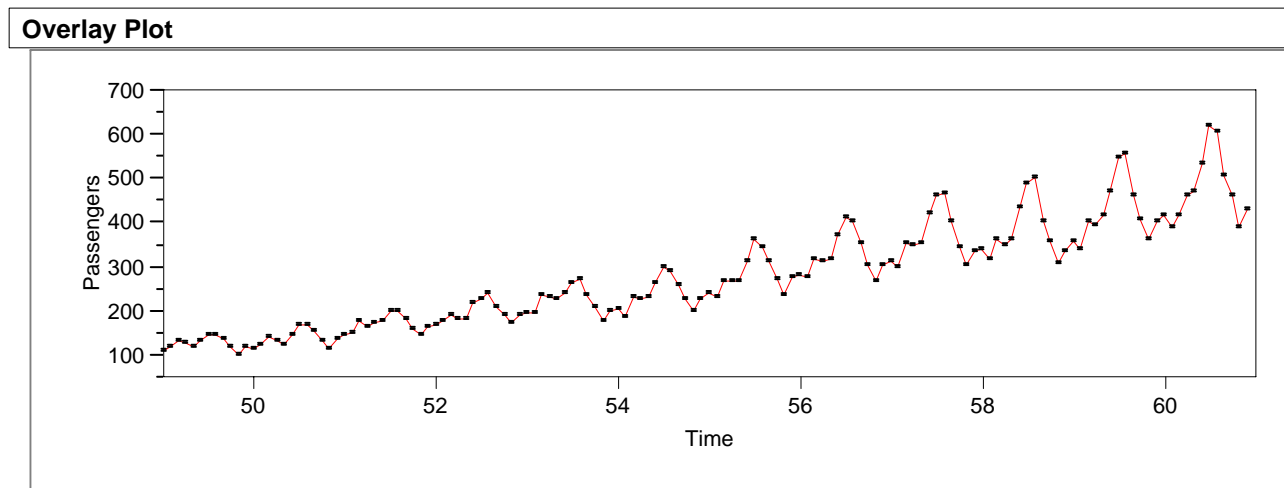
- (b) provide a range that indicates the accuracy of our prediction.

The file IntlAir.jmp contains monthly passenger data from 1949 to 1960, a period of rapid growth.

To anticipate demand, let's use this data to predict the number of passengers in January 1961.

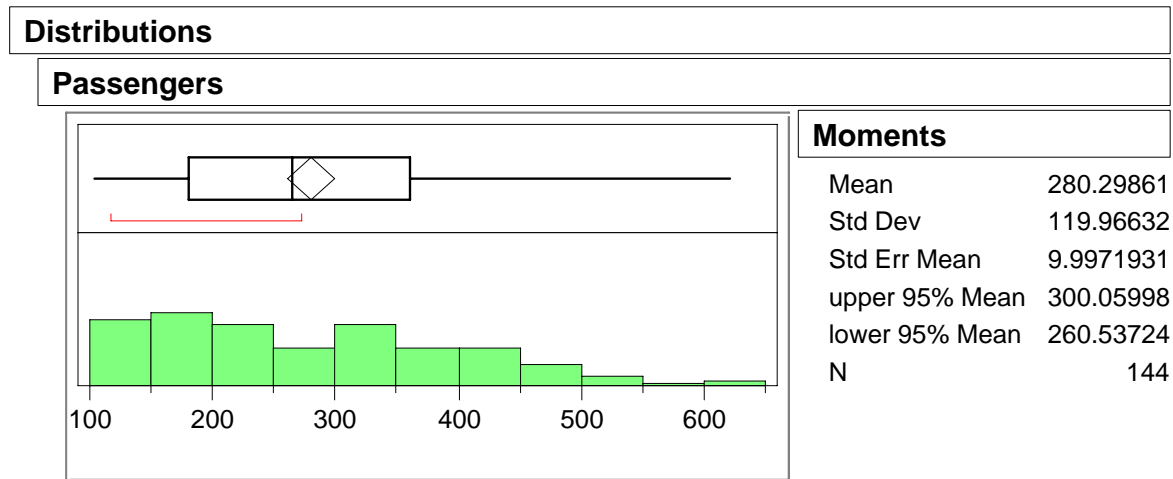
Since this is time series data, we'll start by looking at a time series plot just as we did with the GM data in Module 2.

Does the data look like a random sample, or does it show the meandering appearance of dependence?



These data show a complex pattern. How would you describe the variation in the passenger level over time? What explains the changes in this series?

Here is the summary of the distribution of Passengers



Should we use the mean of about 280,000 per month to predict January 1961, or does this histogram and the accompanying numerical summary hide (ignore) too much?

What offers a better approach?

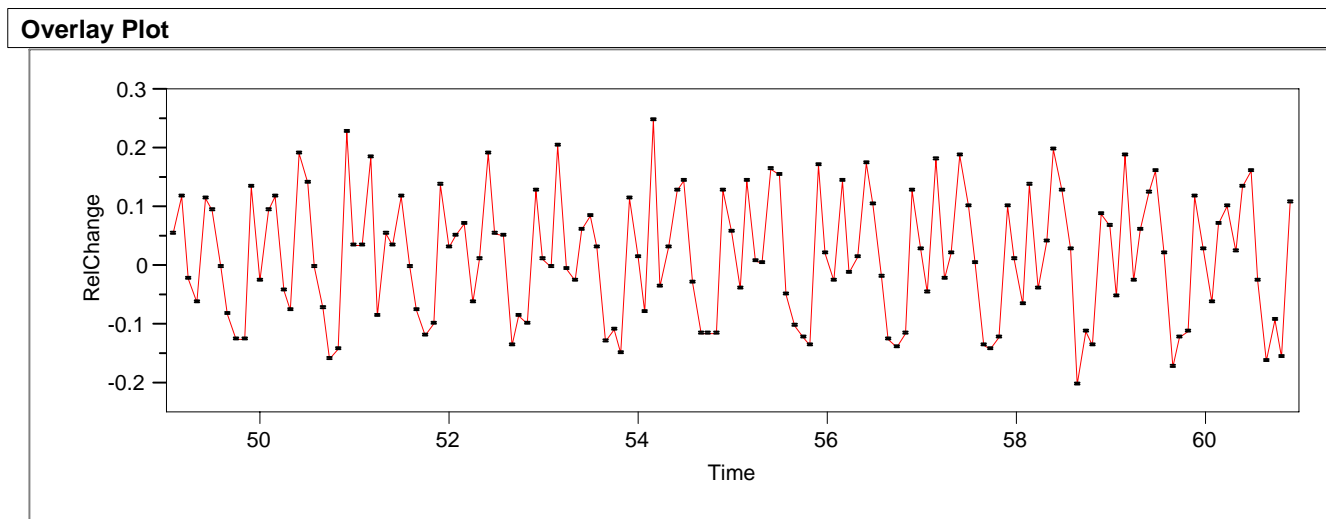
Working with Relative Changes

Let's instead look at the successive monthly relative changes, just as we did with the time series that gave prices of GM stock data,

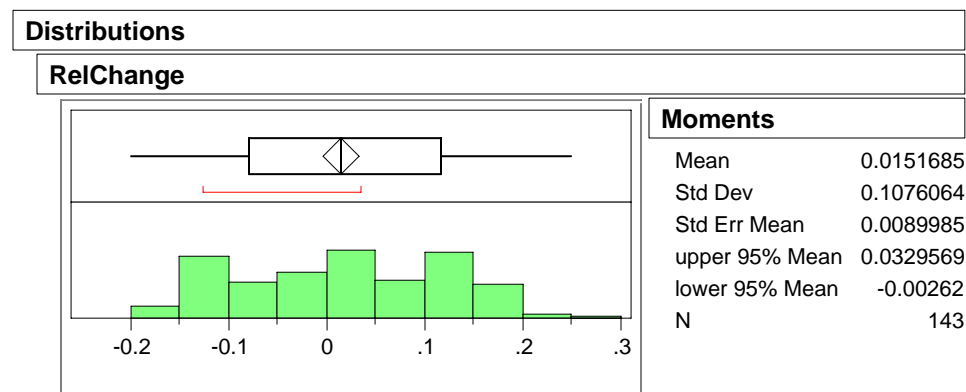
$$\frac{p_t - p_{t-1}}{p_{t-1}}$$

where p_t is the value of Passengers in month t .

Here is the corresponding time series plot (BBS, p.52)



along with the summary of the distribution of the relative changes.



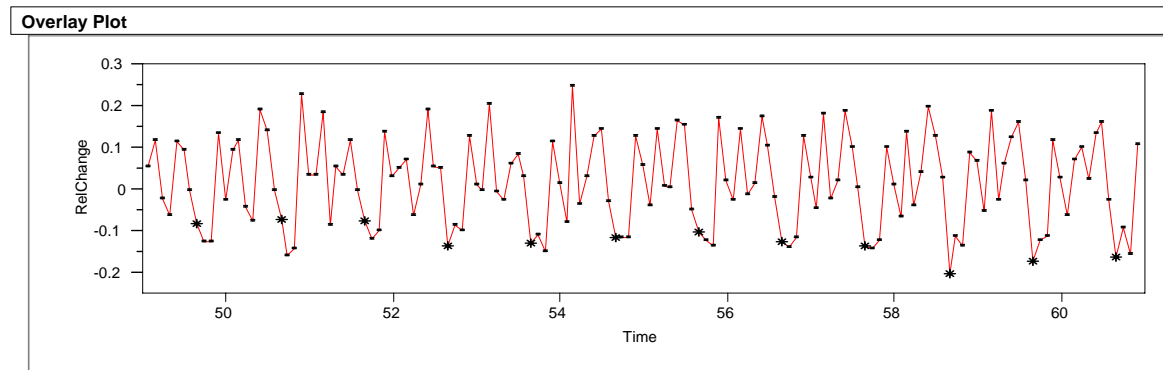
What happened to the upward trend?

What happened to the increasing variation?

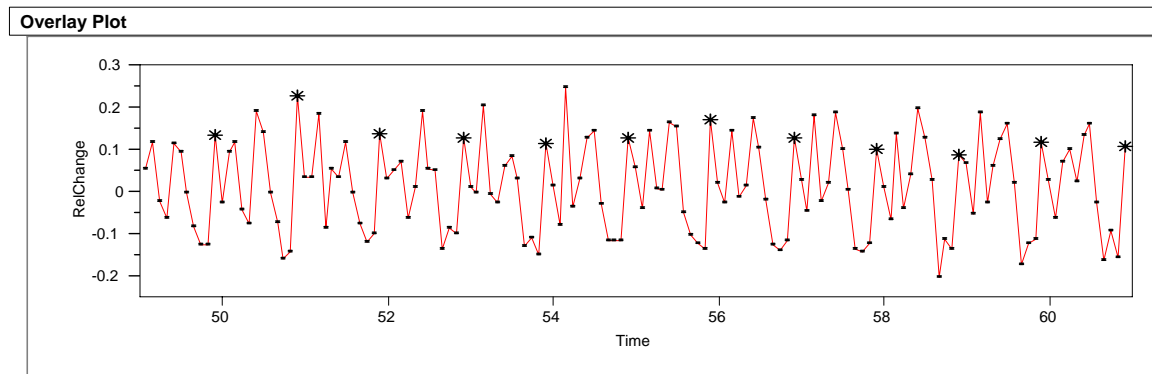
At first glance, it appears that the mean relative change, about 1.5%, is about the same throughout.

Thus, we might consider predicting the January, 1961 value by $432,000 \times 1.015 \approx 438,000$, where 432 is the December, 1960 value. This certainly is better than the mean of 280,000, but the sequence plot on page 3-9 suggests this estimate will not be very precise.

We can do even better! Let's mark the September values (BBS p.52-53)



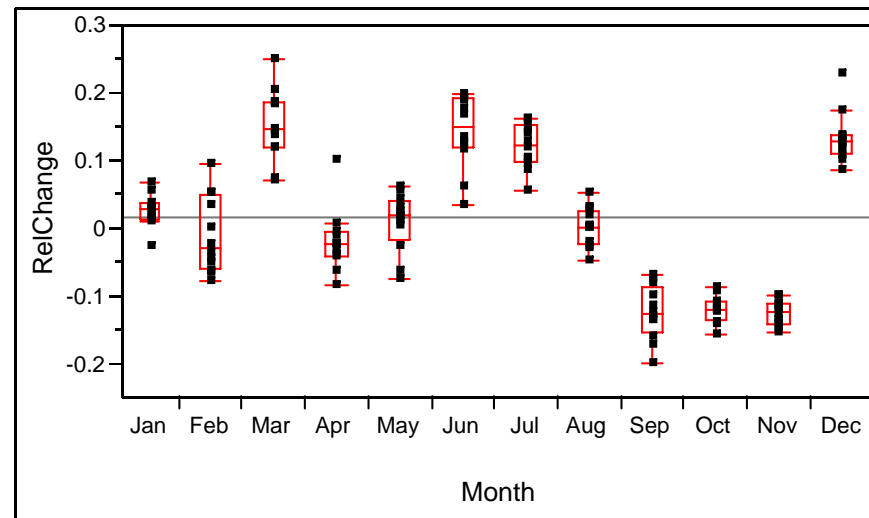
and then mark the December values



What's going on?

Seasonality

A more complete picture of this month-to-month variation, called *seasonality*, is given by comparison boxplots organized by month (BBS, p.53)



Note that months March, June, July, and December tend to be systematically high and months September through November are systematically low. This seasonal effect drives the replicating pattern that we observed in the original time series plot.

The mean change in January is about 2.6% leading to a more appealing estimate of demand for the next January, $432,000 \times 1.026 \approx 443,000$. How accurate is this prediction?

Might we have missed something? Could there be further patterns?

Monitoring an Automotive Manufacturing Process

(BBS, p.55)

A major impact of statistics in business has been in the development of methods for quality control, with applications ranging from manufacturing (this example) to the service industry (as in monitoring call centers).

A question of interest: Is the process currently used to manufacture motor shafts consistently producing satisfactory components?

To study the motor shaft production process at an engine-building plant, 5 shafts were sampled and measured each weekday for 4 weeks.

Standards required for engine assembly require the diameter to fall between 810 and 820 thousandths of an inch. These are called *acceptance limits*.

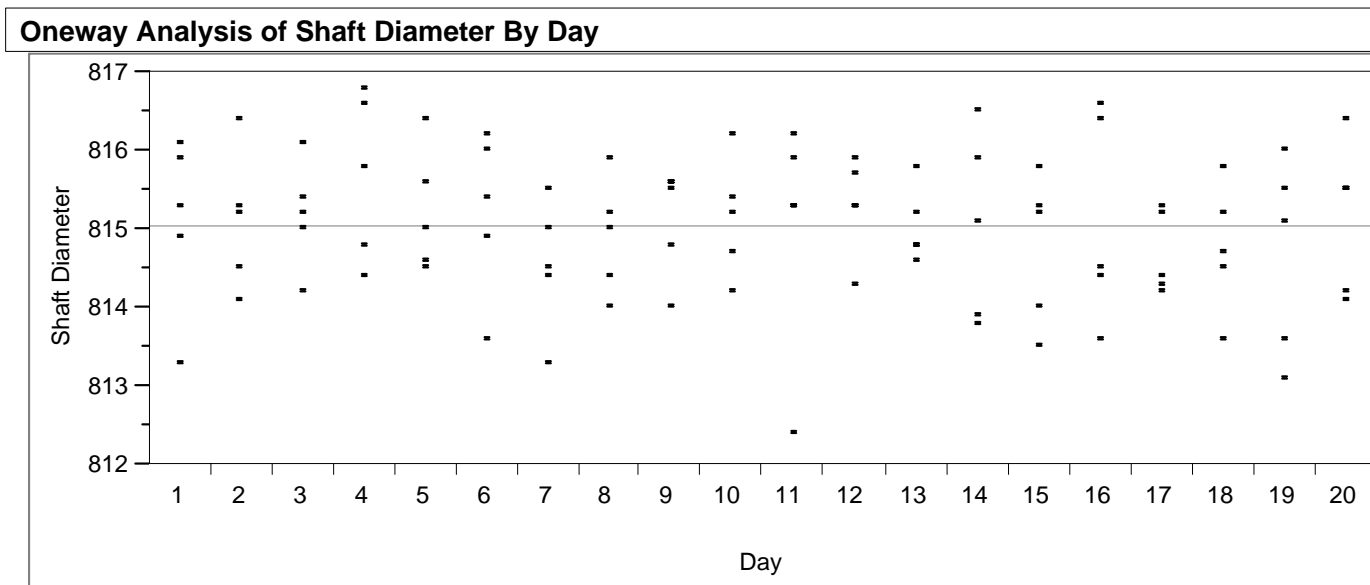
We would like to know if the process is

In Control – behaves like a sequence of independent draws from one population

Capable – output falls within acceptance limits

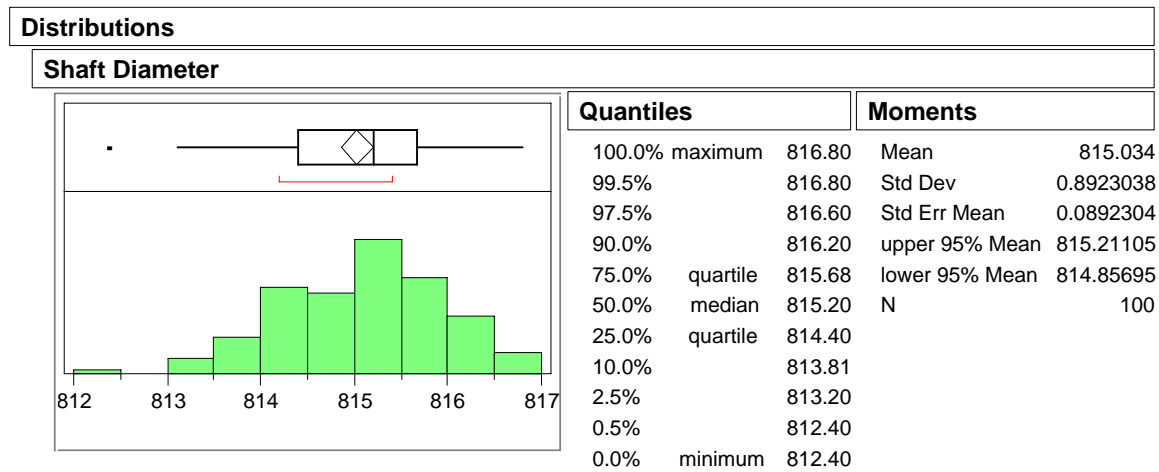
A process can be in control, but not capable. Similarly, a process can be capable, but not in control.

The data for these 100 shafts is in ShaftDia.jmp; a larger set of 400 is in ShaftXtr.jmp
A times series plot of the 100 shaft diameters (with a center line at the mean) reveals no systematic trend, thereby suggesting that the process is in control⁴



⁴ We'll discuss more sophisticated checks for the presence of a systematic trend later in the course.

Here's the summary of the distribution of Shaft Diameter (BBS, p.56)



All the measured shafts fall well within the acceptance limits 810 and 820.

The process is capable.

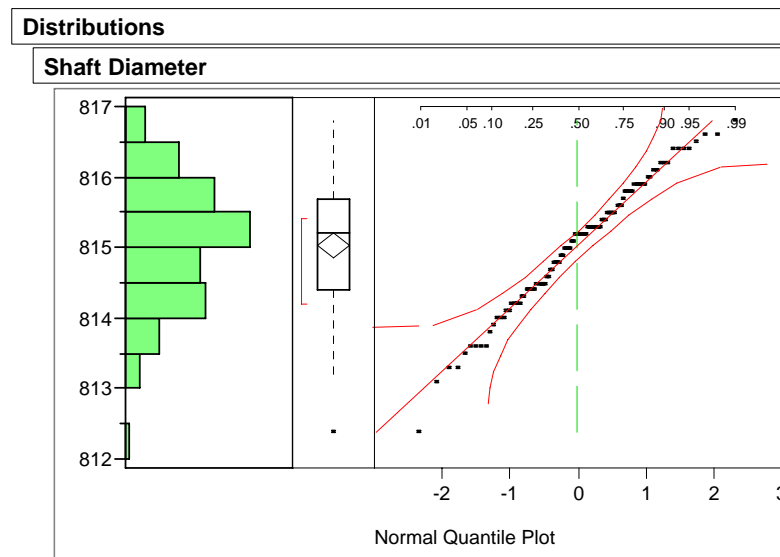
Going Further: What else can we learn about the process? In particular, can we say anything about the chances for this process producing a defective shaft?

A broader goal is not simply to check these 100 shafts, but rather to make statements about future shafts produced by the underlying manufacturing process.

For example, suppose we would like to estimate the probability that the diameter of some future shaft falls outside the acceptance interval (810, 820).

To make such inferences, we need to assume some shape for the population. We can start by using the normal quantile plot to see if it is reasonable to treat the population (i.e., the total process output) as normal (BBS, p.57)

Does the following normal quantile plot lend support to such an assumption?



Suppose we are willing to assume that the process remains in control and that the population is normal with mean 815 and standard deviation .89, (where did these numbers come from? See BBS, p.57)

Then the probability that a future shaft diameter will be outside the interval (810,820) is less than .000000005.⁵

Here's a more recognizable calculation...

Under our assumptions, what is the probability that a future shaft diameter will be outside the interval

$$(815 - 3(.89), 815 + 3(.89)) = (812.33, 817.67)?$$

This interval's endpoints are called 3-sigma control limits and are often used to monitor such a process.

⁵ 810 and 820 are more than 5.6 standard deviations from the mean.

A popular strategy is to deem the process as out of control and needing some corrective action when a value falls outside these 3-sigma limits.

How often will such a strategy incorrectly deem the process to be out of control?
Some details of one relevant calculation are given in BBS, p.58-59.

To lower this probability, wider control limits could be used. What is the tradeoff of such a strategy?

CAREFUL!!! Do not confuse control limits with acceptance limits. They are not the same thing!!!

Take-Away Review

A key objective is to learn from data about the underlying sources of variation, such as

- Variation due to clustering or grouping.
- Variation determined by systematic effects over time.
- Variation due to chance

Plots and summary statistics help us understand variation.

Normality allows us to offer a probability for some future event, assuming that the future looks like the past. The normal distribution is our first encounter with a *statistical model*.

Next Module

Probability and models for random variation.

**Department of Statistics
The Wharton School
University of Pennsylvania**

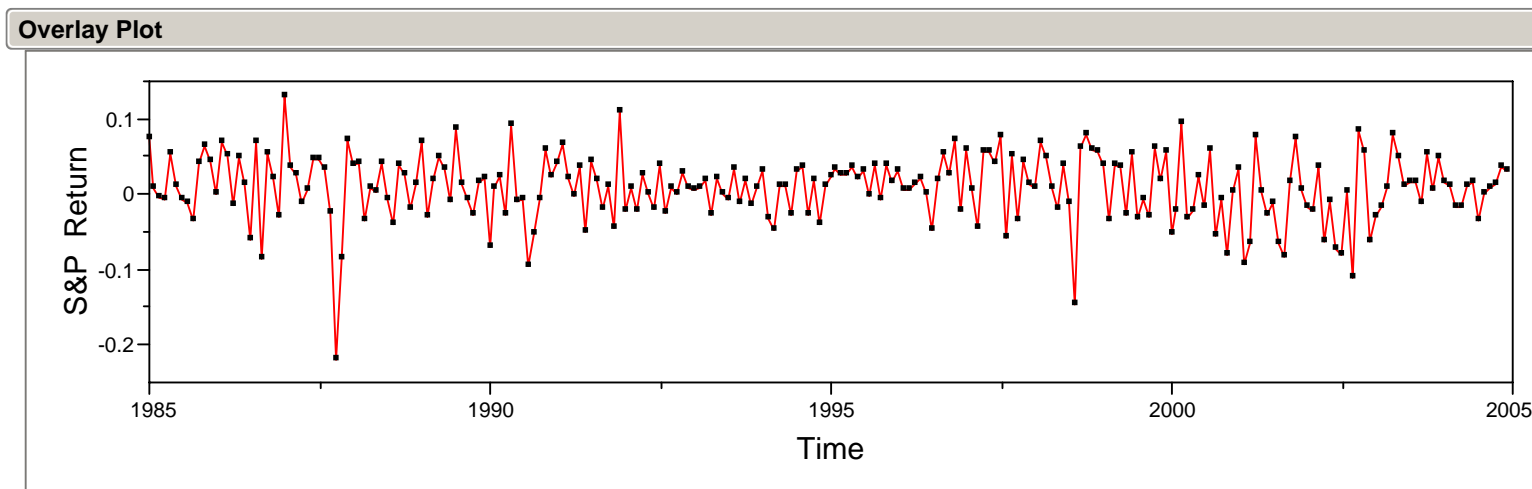
STAT 603

August 2006

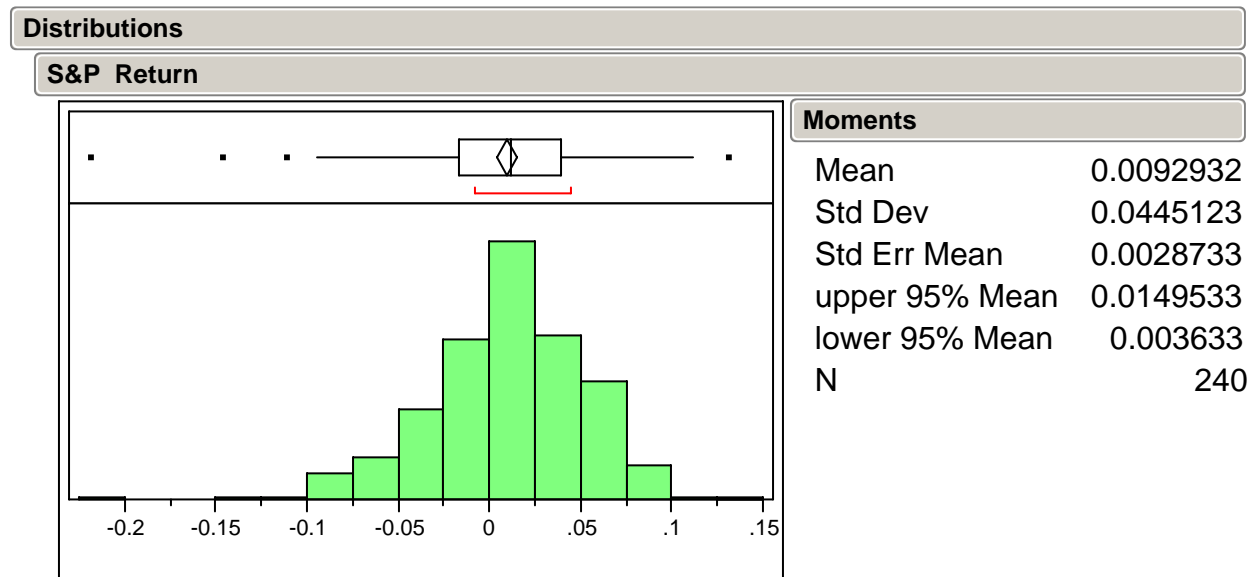
**Module 4
Probability Models**

Application: Pricing a Digital Option

Many investors put a large part of their assets in stocks, such as purchasing an index fund. The history of the stock market in the US reveals that these investors face risks. Although the market has headed up for most of its history, some months have shown very large declines. S&P85-04.jmp contains the following monthly returns on the S&P index. Do you recognize some of these dates?



We can summarize these data with a histogram that shows the distribution of the monthly returns over these years.



The average monthly return is .93% with standard deviation 4.45%. What do these mean?

Investors with large positions in the stock market might want to avoid some of these risks. Rather than selling their stocks (and paying all of those fees), investors can purchase a type of insurance in the form of an *option*.

We'll consider a very special type of option known as a *digital option*. These can be used to construct all sort of more complicated options known as derivatives.

The question of interest

To insure yourself against a large drop in the S&P index, how much should you expect to pay for a digital option that pays \$1 if the S&P index drops by more than 10% next month?

The price for one option is not so interesting, but consider an investor who wants to buy 1,000,000 of these options.

What's a good price for the option from the point of view of the investor? Of the bank that offers the option?

The complicating factor

We need to anticipate what might happen in a future month. All we have to go on is what has happened in the past. The idea that we'll use to think carefully about such problems is known as a random variable.

Random Variables

A *random variable* (rv) (e.g. X , Y , Z , etc.) represents the uncertain numerical outcome of an experiment or process that has yet to occur.

Each possible outcome of a random variable is associated with a number between 0 and 1. This number is called the *probability* of the outcome.

Useful Notation: If X is a random variable, then

$P(X = x)$ is the probability that X will take on the value x

$P(x_1 \leq X \leq x_2)$ is the probability that X will take on a value in the interval $[x_1, x_2]$

The Long Run Manifestation of Probability: Intuitively, $P(X = x)$ is the proportion of times $X = x$ over an “infinitely” long sequence of repetitions of the experiment.

In this sense, a graph of $P(X = x)$ is the histogram of the outcomes in this long sequence.

Example: Toss a Fair Coin

Random variable: $X = 1$ if heads $X = 0$ if tails

Probabilities: $P(X = 1) = 1/2$ $P(X = 0) = 1/2$

Example: Play Roulette

38 slots labeled 00, 0, 1, 2, ..., 36

1-36 colored Red (if even) and Black (if odd)

For \$1, place a bet on Red

Pays you \$2 if randomly selected slot is Red

Random variable: $R = \text{Net winnings}$ Possible outcomes: -1 and 1

Probabilities: $P(R = -1) = 20/38$, $P(R = 1) = 18/38$

Example: Play the Lottery

For \$1, buy a lottery ticket numbered from 000 to 999.

Pays you \$500 if ticket matches randomly selected number.

Random variable: $D = \text{Net winnings}$

Possible outcomes: -1 and 499

Probabilities: $P(D = -1) = 999/1000$, $P(D = 499) = 1/1000$

Example: Pick a Chip

Randomly draw a chip from a bowl containing 10,000 chips with these values:

5000	\$1 chips
3000	\$5 chips
1000	\$10 chips
1000	\$20 chips

Random variable: $X = \text{value of the drawn chip}$ Possible Outcomes: 1, 5, 10, 20

Probabilities: $P(X = 1) = 5/10$, $P(X = 5) = 3/10$, $P(X = 10) = 1/10$, $P(X = 20) = 1/10$

For convenience, we often write $p(x) = P(X = x)$, so that

$$p(1) = 5/10 \quad p(5) = 3/10 \quad p(10) = 1/10 \quad p(20) = 1/10$$

The function $p(x)$ is called the *probability distribution* of X

Note that:

$$p(x) = 0 \quad \text{for } x \neq 1, 5, 10, 20$$

$$p(1) + p(5) + p(10) + p(20) = 1$$

The probability of any *event* concerning X can be computed from $p(x)$. For example,

$$P(X \leq 5) = 5/10 + 3/10 = 8/10$$

$$P(5 \leq X \leq 10) = 3/10 + 1/10 = 4/10$$

$$P(X \text{ is an even number}) = 1/10 + 1/10 = 2/10$$

Let's draw a histogram of the population of chips in the bowl:

How does this shape compare to the shape of the probability distribution?

This bowl of chips can be thought of as a population from which the chip was drawn

Useful Interpretation: The outcome of random variable with probability distribution $p(x)$ *can always be interpreted as* a random selection from a $p(x)$ shaped population distribution

Let's now consider computing the mean of the chip histogram above.

The Expected Value of a Random Variable

Suppose the N possible outcomes of a rv X are

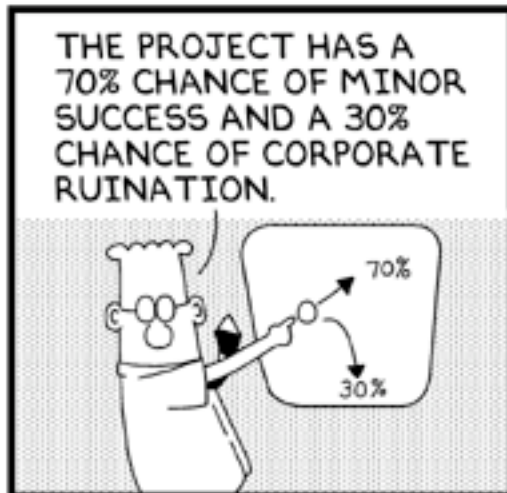
$$x_1, x_2, \dots, x_N.$$

The *expected value* of X , denoted by $E[X]$, is defined as a weighted sum of the possible outcomes of X ,

$$E[X] = x_1 p(x_1) + x_2 p(x_2) + \dots + x_N p(x_N) = \sum_{i=1}^N x_i p(x_i)$$

Sometimes the expected value of X is called the *mean* of X and is denoted by the Greek letter μ . This is the same as the average value in the population. For example, for the chips we find the average value is

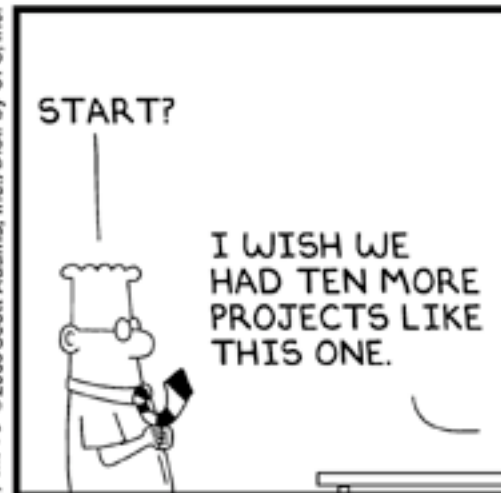
$$\begin{aligned} \frac{1 \times 5,000 + 5 \times 3,000 + 10 \times 1,000 + 20 \times 1,000}{10,000} &= 1 \times \left(\frac{5}{10} \right) + 5 \times \left(\frac{3}{10} \right) + 10 \times \left(\frac{1}{10} \right) + 20 \times \left(\frac{1}{10} \right) \\ &= 1p(1) + 5p(5) + 10p(10) + 20p(20) \\ &= E[X] \\ &= \mu \end{aligned}$$



www.dilbert.com scottadams@aol.com



7-23-95 © 2005 Scott Adams, Inc./Dist. by UFS, Inc.

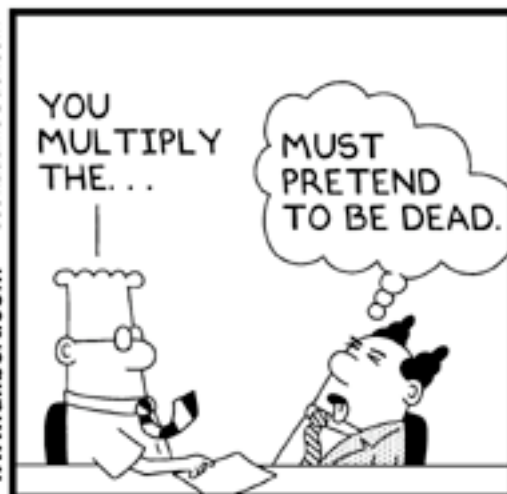


© Scott Adams, Inc./Dist. by UFS, Inc.

1



www.dilbert.com scottadams@aol.com



7-23-95 © 2005 Scott Adams, Inc./Dist. by UFS, Inc.



© Scott Adams, Inc./Dist. by UFS, Inc.

¹ Alas, Dilbert is not a statistician. You want the expected value of the project, not an outcome!

Examples

Example: R = Winnings in Roulette

$$E[R] = -1 p(-1) + 1 p(1) = -20/38 + 18/38 = -1/19$$

Example: D = Winnings in Daily Numbers Lottery

$$E[D] = -1 p(-1) + 499 p(499) = -999/1000 + 499/1000 = -0.5$$

Example: X = value of the drawn chip

$$\mu = 1 p(1) + 5 p(5) + 10 p(10) + 20 p(20) = 5/10 + 15/10 + 10/10 + 20/10 = 5$$

Interpretations of the expected value or mean

Probability-weighted average of possible outcomes

Center of the probability distribution of X

Mean of the population from which the rv is drawn

Long-run average over “infinitely” many repetitions of X

“Fair value” of a gamble

Jargon

μ is sometimes called the *population mean*

to distinguish it from \bar{x} , the *sample mean*.

Variance and Standard Deviation of a Random Variable

As with our descriptions of a histogram, it's not enough to know the mean value. We also need a summary of the variation in a random variable as well.

Suppose the possible outcomes of a rv X are

$$x_1, x_2, \dots, x_N.$$

Then the *variance* of X , denoted $\text{Var}[X]$ or σ^2 , is

$$\text{Var}[X] = \sigma^2 = (x_1 - \mu)^2 p(x_1) + \dots + (x_N - \mu)^2 p(x_N) = \sum_{i=1}^N (x_i - \mu)^2 p(x_i)$$

An equivalent and slicker definition is

$$\text{Var}[X] = \sigma^2 = E[(X - \mu)^2]$$

The *standard deviation* of X , denoted $\text{SD}[X]$ or σ , is the square root of the variance

Examples

Example: R = Net winnings in Roulette

$$\text{Var}[R] = (-1 - (-1/19))^2 p(-1) + (1 - (-1/19))^2 p(1) = 0.9972$$

$$\text{SD}[R] = 0.9986$$

Example: D = Net winnings in Daily Numbers Lottery

$$\text{Var}[D] = (-1 - (-0.5))^2 p(-1) + (499 - (-0.5))^2 p(499) = 249.75$$

$$\text{SD}[D] = 15.80$$

Example: X = value of the drawn chip

$$\sigma^2 = (1-5)^2 p(1) + (5-5)^2 p(5) + (10-5)^2 p(10) + (20-5)^2 p(20) = 33$$

$$\sigma = 5.744$$

Interpretations of the variance

Probability-weighted average squared deviation

Dispersion of the probability distribution of X about μ

Variance of the population from which the rv is drawn

Long run average squared deviation over “infinitely” many repetitions of X

“Risk” of a gamble (more coming in Module 5)

Jargon

σ^2 and σ are sometimes called the *population variance* and the *population standard deviation*

to distinguish them from s^2 and s , the *sample variance* and the *sample standard deviation*.

Discrete versus Continuous Random Variables

The random variables considered so far are called *discrete*, because we can count the possible outcomes (a discrete set).

As we saw, the probability distributions of discrete random variables are described by a function $p(x)$ that assigns masses of probability to the distinct outcomes.

However, some random variables such as the temperature tomorrow or the return on the S&P index next month are instead called *continuous* because (at least in principle) their set of possible outcomes is an interval (a continuous set).

The probability distribution of a continuous random variable is described with the area under a curve $f(x)$, called a density

Useful Interpretation: The outcome of random variable with density $f(x)$ ***can be interpreted as*** a random selection from an $f(x)$ shaped population distribution

Main idea: $P(x_1 \leq X \leq x_2) = \text{area under } f(x) \text{ between } x_1 \text{ and } x_2$

An intuitive way to conceptualize μ and σ^2 for a continuous random variable X is as the mean and variance of the population described by $f(x)$.

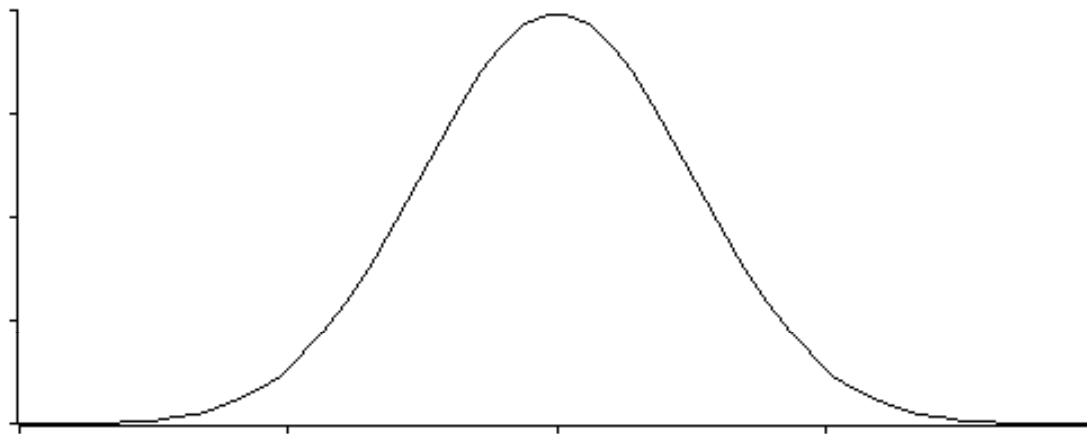
Example: A Normal Random Variable

X = value drawn randomly from a normal population with mean μ and st dev σ .

Often abbreviated as $X \sim N(\mu, \sigma^2)$.

Is X continuous or discrete?

Density: $f(x) \left(= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2 / (2\sigma^2)} \right)$ *if you would like to know*



Some probabilities:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68\%$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\%$$

$$P(\mu - \sigma \leq X \leq \mu + 2\sigma) \approx 81.5\%$$

$$P(-\infty < X < \infty) = 1$$

More generally, normal probabilities for any given values of μ and σ can be computed using the Normal Distribution probability formula in JMP.

For example², when $X \sim N(0.5, 1.8^2)$,

$$P(X \leq 0.8) = .566 \quad \text{and} \quad P(0.8 \leq X \leq 1.4) = .125$$

² In JMP, the probability $P(X < a)$ when $X \sim N(\mu, \sigma^2)$ is obtained with the JMP formula `Normal Distribution(a, μ , σ)`. The calculation here is illustrated in the file `Norm Prob.jmp`.

Where calculus comes into play - if you would like to know!

Formulas for calculating probabilities for a continuous random variable X , that is, for calculating areas under $f(x)$ can be obtained using

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

Formulas for the mean μ and variance σ^2 for a continuous random variable X are obtained using calculus as

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \quad \text{and} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Example

For the normal random variable $X \sim N(\mu, \sigma^2)$, it can be shown that

$$\mu = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx \quad \text{and} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

Application: Pricing a Digital Option

Let X = the return on the S&P index next month

Suppose you had strong enough evidence to *assume* that X was a normal random variable with mean .0093 and standard deviation .0445

These values for the mean and SD match the month-to-month performance of the stock market from January 1985 to December 2004.

Let's now return to the our question on pg 4-3: How much should you pay for a digital option that pays \$1 if the S&P drops by more than 10% next month?

We can express the value of such an option by defining a new, discrete random variable Y whose values are determined by X , with

$$\begin{aligned} Y &= \$1 && \text{if } X < -.10 \\ &= \$0 && \text{otherwise} \end{aligned}$$

A fair price for the option would then be³ $E[Y] = P[Y = 1] = P[X < -.10] = .0070$

³ The calculation of this probability is illustrated in the file Option Prob.jmp.

The Normal Distribution as a Model

The value of $E[Y]$ in the previous example depends strongly on $P[X < -.10] = .0070$ which we obtained by assuming that X would be a draw from a particular normal distribution.

An outline of the steps used to make such an assumption:

- 1) Verify that the behavior of the time series of past monthly S&P returns is consistent with assuming a sequence of independent draws from the same population
- 2) Verify that the histogram and a normal quantile plot of the past monthly returns is consistent with assuming a normal population
- 3) Estimate μ and σ^2 of the population by the values of \bar{x} and s^2 for the past monthly returns

Without such a distributional assumption, a “common sense” procedure might be to instead estimate $P[X < -.10]$ by the proportion of prior months that the S&P fell below $-.10$.

What is the drawback of such an approach?

Compared to this simplistic procedure, the normal assumption allows us to use the data to extrapolate more efficiently. Careful - we can get very different answers using other models!

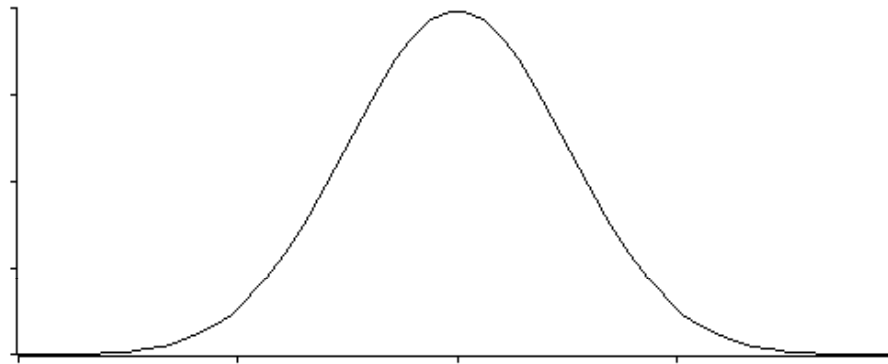
Although the normal distributions approximate many distributions that naturally occur, important alternatives may also be considered. Some are hard to tell from the normal.

The (Student) t Distributions

The t distributions are a family of continuous probability distributions that resemble normal distributions.

They are symmetric and bell-shaped, but have heavier or fatter tails.

Because of their heavier tails, extreme observations are more likely with draws from a t distribution.



Each t distribution is identified by 3 parameters μ , σ^2 , and a shape parameter df (degrees of freedom). (df is always positive). As df gets larger, the tails get lighter and the t distributions more closely approximate the normal distributions.

t distribution probabilities can be obtained by the probability function t distribution on the JMP calculator.

If we treat X as having a t-distribution that is matched to the market during the 1985-2004 period, we obtain

$$P(X < -.10) \approx .05$$

Why does this probability differ so much (it's much larger) from the previous normal probability calculation?

Quantile plots can also be used as a diagnostic check for t distribution assumptions.⁴

Key point

In this example, we have two different models for the data. Both describe the data reasonably well, so that the data are not clear about which model is better.

The model that we use matters: we get very different prices for the options.

In cases like this, we need to be sensitive to the consequences of our assumptions.

We may never know which model is “right”, but we should be aware of the alternatives.

⁴ These alternative quantile plots are not in the version of JMP we are using. The plots would look like normal quantile plots, but the reference model would be something other than the normal, such as one of the t-distributions.

Take-Away Review

Random variables provide a useful notation for describing problems that have uncertain outcomes.

The outcomes of a random variable can be associated with probabilities.

The expected value is the average of the outcomes, weighted by probabilities. The variance is the average deviation, weighted by probabilities.

If we match a random variable to features of a sample, we can use that random variable to model the underlying process.

But, the results we obtain are sensitive to the model that we use. Check the assumptions of your model.

Next Module

Random variables and returns on investments.

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Module 5

Variance and the Volatility of Investments

Random Variables in Finance

Random variables are used in Finance to represent returns on investments.

Variance of such random variables measures the volatility (risk) of the investment.

Example

Let Green, Red and White denote three hypothetical investments with these probability distributions for their annual gross returns¹ R

Die value	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6
Green	0.8	0.9	1.1	1.1	1.2	1.4
Red	0.06	0.2	1	3	3	3
White	0.95	1	1	1	1	1.1

The expected values and standard deviations of the annual percentage changes are

Investment	Expected R	StDev of R
Green	1.083	.20
Red	1.710	1.32
White	1.008	.04

Which of these investments is most appealing? What tradeoffs did you consider?

¹ The value of an asset at the end of a year is the product of the gross return and the value at the start of the year. If you start with S dollars and finish with F dollars, then the gross return is $R = F/S$. We denote gross returns by “big R ” to distinguish them from “little r ” returns (pg 2-17). These types of returns are related by $R = 1 + r$. “Little r ” returns, sometimes called net returns, become percentage changes when multiplied by 100.

A Simulation Experiment

Suppose you begin with a \$1000 investment in each of Green, Red and White

The outcome from rolling three dice determines the annual outcome of the investment of matching color

The value of the investment value changes according to the gross return given in the appropriate column of the table.

For example, suppose that on the first roll of all three dice, you obtain

(Green 2) (Red 5) (White 3)

Then the values of the investments after the first year are

Green: $\$1000 \cdot 0.9 = \900

Red: $\$1000 \cdot 3 = \3000

White: $\$1000 \cdot 1 = \1000

Die value	1	2	3	4	5	6
Green	0.8	0.9	1.1	1.1	1.2	1.4
Red	0.06	0.2	1	3	3	3
White	0.95	1	1	1	1	1.1

Suppose the second roll gives

(Green 4) (Red 2) (White 6)

By compounding, you get

Green: $\$900 \cdot 1.1 = \990

Red: $\$3000 \cdot 0.2 = \600

White: $\$1000 \cdot 1.1 = \1100

Die value	1	2	3	4	5	6
Green	0.8	0.9	1.1	1.1	1.2	1.4
Red	0.06	0.2	1	3	3	3
White	0.95	1	1	1	1	1.1

Note that Green went down by 10% and then up by 10%, but ended up losing value.

Why does that happen?

A Class Experiment

Form teams of 3 to 4 students.

Start with \$1000 in each investment as above and carry out the simulation for 20 years of returns. Each roll of all three dice represents one year.

Roles for team members

- “Nature” rolls the dice

- “Market” finds the dice and records outcome

- “Accountant” keeps track of what happens

- Others manage and keep the rest making progress.

Record the sequence of results on the results form as shown on the next page

We've filled in the first two rounds to match the previous two outcomes to illustrate the calculations. Use the outcomes from rolling the dice to determine what happens to your investments.

Round	Green	Red	White	
Starting value	\$1000	\$1000	\$1000	
gross return ₁	0.9	3	1	
value ₁	900	3000	1000	
gross return ₂	1.1	0.2	1.1	
value ₂	990	600	1100	
gross return ₃				
value ₃				
gross return ₄				

What happened? Are you surprised?

A Hybrid Investment

Consider a fourth investment which puts half in Red and half in White – call it Pink. The gross return on Pink is just the average of the gross return in each round on Red and White. So, you can find out what happens to Pink without needing to roll the dice further.

For example, for the first round illustrated above, the gross return on Pink is the average of the gross return on Red and White, namely $(3+1)/2 = 2$.

$$\text{Pink: } \$1000 \cdot 2 = \$2000$$

In the second round, the gross return on Pink would be $(0.2 + 1.1)/2 = 0.65$ yielding

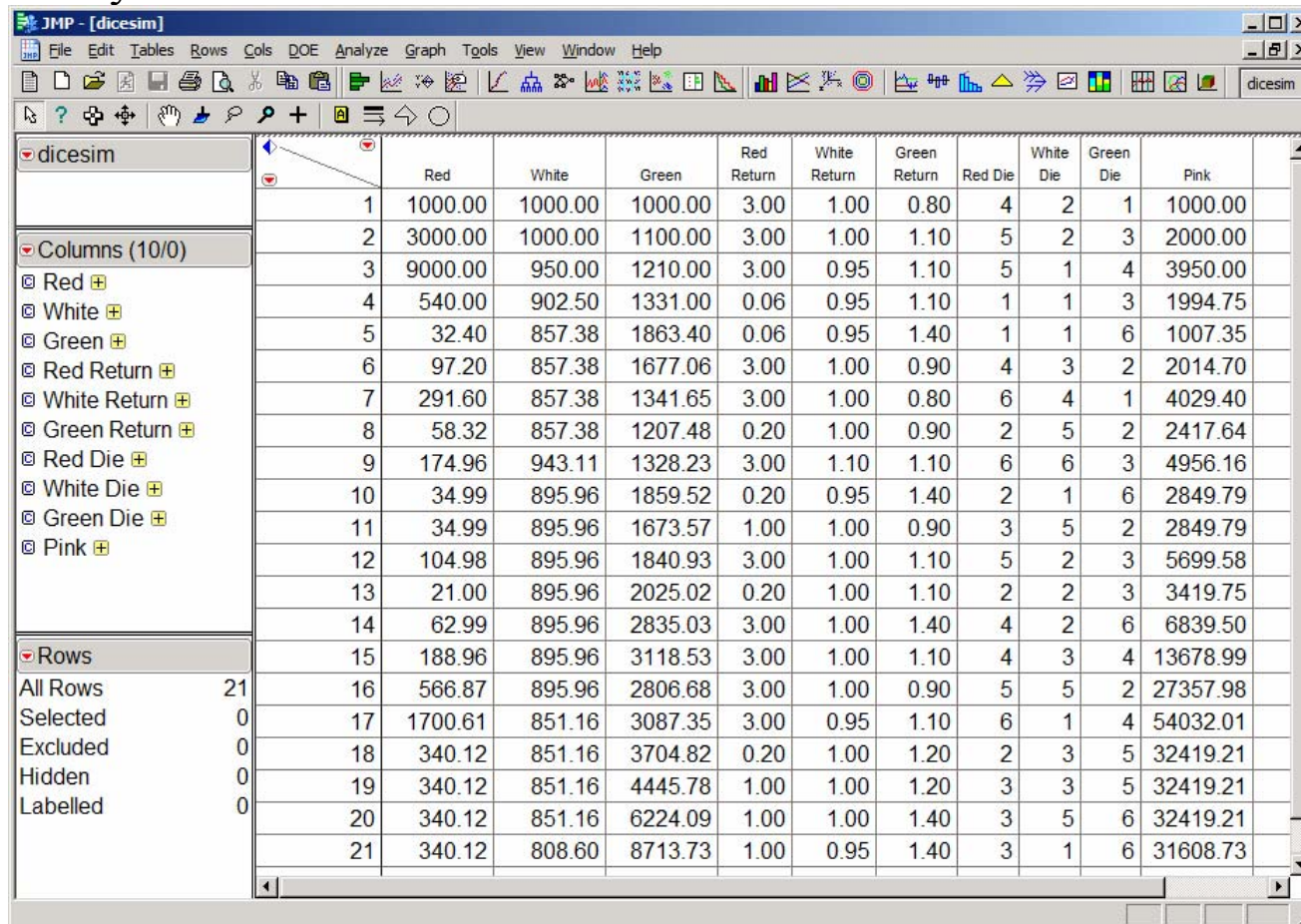
$$\text{Pink: } \$2000 \cdot 0.65 = \$1300$$

How does Pink fare in your simulation?

No more dice tossing. Just compute the gross returns using the information recorded on your data sheet.

Performing the Simulation on a Computer

The file dicesim.jmp is set up to perform the simulation in JMP. Opening the file and adding 20 rows yields:



	Red	White	Green	Red Return	White Return	Green Return	Red Die	White Die	Green Die	Pink
1	1000.00	1000.00	1000.00	3.00	1.00	0.80	4	2	1	1000.00
2	3000.00	1000.00	1100.00	3.00	1.00	1.10	5	2	3	2000.00
3	9000.00	950.00	1210.00	3.00	0.95	1.10	5	1	4	3950.00
4	540.00	902.50	1331.00	0.06	0.95	1.10	1	1	3	1994.75
5	32.40	857.38	1863.40	0.06	0.95	1.40	1	1	6	1007.35
6	97.20	857.38	1677.06	3.00	1.00	0.90	4	3	2	2014.70
7	291.60	857.38	1341.65	3.00	1.00	0.80	6	4	1	4029.40
8	58.32	857.38	1207.48	0.20	1.00	0.90	2	5	2	2417.64
9	174.96	943.11	1328.23	3.00	1.10	1.10	6	6	3	4956.16
10	34.99	895.96	1859.52	0.20	0.95	1.40	2	1	6	2849.79
11	34.99	895.96	1673.57	1.00	1.00	0.90	3	5	2	2849.79
12	104.98	895.96	1840.93	3.00	1.00	1.10	5	2	3	5699.58
13	21.00	895.96	2025.02	0.20	1.00	1.10	2	2	3	3419.75
14	62.99	895.96	2835.03	3.00	1.00	1.40	4	2	6	6839.50
15	188.96	895.96	3118.53	3.00	1.00	1.10	4	3	4	13678.99
16	566.87	895.96	2806.68	3.00	1.00	0.90	5	5	2	27357.98
17	1700.61	851.16	3087.35	3.00	0.95	1.10	6	1	4	54032.01
18	340.12	851.16	3704.82	0.20	1.00	1.20	2	3	5	32419.21
19	340.12	851.16	4445.78	1.00	1.00	1.20	3	3	5	32419.21
20	340.12	851.16	6224.09	1.00	1.00	1.40	3	5	6	32419.21
21	340.12	808.60	8713.73	1.00	0.95	1.40	3	1	6	31608.73

Essentially, the computer “rolls the dice” in the three columns Red Die, White Die and Green Die.²

Note the persistent behavior of each of these four investments, especially if you let the computer do a few more rounds. Which investment consistently wins?

Variances and Volatility Drag

Let’s turn to understanding the simulation results

Example: Your starting salary was \$100,000. You received a salary increase of 10% and then a salary reduction of 10%. What is your current salary?

What would happen to your salary if this up/down bounce was repeated over and over?

Volatility hurts by eating away at the average rate of return

² To recalculate these, click on the Apply button in the Formula Editor Window for each die column.

As we will see below, it turns out that

$$\text{Long-run gross return} \approx \text{Expected gross return} - \text{Variance}/2$$

The quantity $\text{Variance}/2$ is called “volatility drag”

Applied to annual returns on Green, Red and White, we obtain (with more digits shown)³

Investment	Mean	StDev	Var	Mean–Var/2
Green	1.083	.195	.038	1.064
Red	1.710	1.32	1.755	0.833
White	1.008	.045	.002	1.007

It is now clear why Red and White are both losers!

But why did Pink do so well?

³ For example, to find the expected return on Green, we computed $E[\text{Green}] = (1/6)0.8 + (1/6)0.9 + (2/6)1.1 + (1/6)1.2 + (1/6)1.4 = 1.083$ as shown in this summary. Similar calculations produce the variances.

Analyzing a Mixed Investment

Pink is itself a random variable with a probability distribution. Using our previous formulas, we can directly calculate

Investment	Mean	StDev	Var	Mean – Var/2
Pink	1.359	0.662	.439	1.140

Wow! Even though Red is a big loser and White is pretty poor, mixing the two losers yields Pink, a big winner.

What's going on?

We also need some help to find these means and variances. Direct calculation of the mean and standard deviation is really tedious:

$$\begin{aligned} E[Pink] &= \frac{1}{36} \left(\frac{0.06 + 0.95}{2} \right) + \frac{1}{36} \left(\frac{0.06 + 1}{2} \right) + \cdots + \frac{1}{36} \left(\frac{3 + 1.1}{2} \right) \\ &= 1.359 \end{aligned}$$

An Easier Way to Calculate E(Pink) and Var(Pink)

As we will see in Module 6, we can obtain E(Pink) and Var(Pink) using the formulas

$$E(\text{Pink}) = .5 E(\text{Red}) + .5 E(\text{White}) = 1.359$$

$$\text{Var}(\text{Pink}) = .5^2 \text{Var}(\text{Red}) + .5^2 \text{Var}(\text{White}) = .439$$

Note that the first formula makes perfect sense: the average of averages is itself an average.⁴

⁴ OK, we agree, the second one is not so obvious. We'll save it for Module 6.

Justification for the Volatility Drag Adjustment – if you would like to know!

Suppose our initial wealth at time 0 is W_0 and the gross return from period $t-1$ to period t is R_t .

Our wealth at the end of year 1 is $W_1 = W_0 R_1$

In general, our wealth at the end of year T is

$$W_T = W_{T-1} R_T = \dots = W_0 R_1 R_2 \dots R_T$$

Believe it or not, it is easier to take logs so that we can work with sums instead of products:

$$\begin{aligned} \log W_T &= \log W_0 + \log R_1 + \log R_2 + \dots + \log R_T \\ &= \log W_0 + T \text{ avg } \log R_t \\ &\approx \log W_0 + T E[\log R_t] \end{aligned}$$

As the number of time points increases, the average of $\log R_t$ gets closer to the expected log return. This is another name for the long-run growth rate. Here is where the Volatility Drag enters. Calculus produces the approximation

$$\log(1+x) \approx x - x^2/2$$

which works well when x is close to zero.

As long as the average gross returns are close to 1, we obtain (using the “little r ” form for writing the gross return as $R_t = 1 + r_t$)

$$\begin{aligned} E \log R_t &= E \log (1+r_t) \approx E (r_t - r_t^2/2) \\ &\approx E (r_t) - \text{Var}(r_t)/2 \end{aligned}$$

and by “unlogging” (taking the exponential) we get – hold on to your seats –

$$\begin{aligned} W_T &\approx W_0 \times \exp(T E[\log R_t]) \\ &\approx W_0 \times [\exp(E (r_t) - \text{Var}(r_t)/2)]^T \\ &\approx W_0 \times [E R_t - \text{Var}(R_t)/2]^T \end{aligned}$$

This approximation to the wealth W_T shows that the rate of growth per time period is close to the expected return *minus* the volatility drag.

Whew! But relax, you’ll see these ideas again in Finance.

Background

The choices of the means and variances for two of these investments come from things that you can buy. Green matches the long-run historical performance of the US stock market, as reflected by the value-weighted index since 1925, adjusted for inflation.

We calibrated White to approximate the historical performance of 30-day Treasury Bills, also adjusted for inflation.

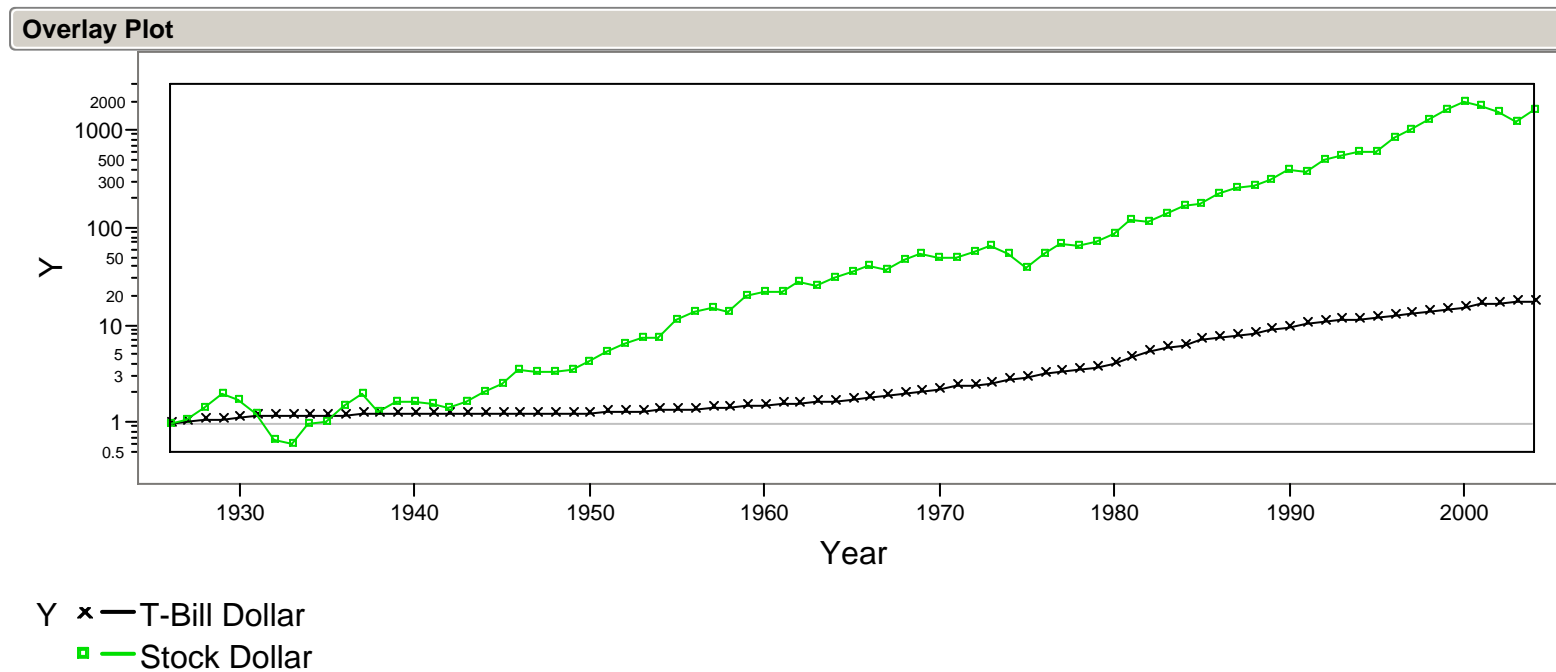
Notice how closely the means and standard deviations of Green and White match those of the annual gross returns of the Stocks and T-Bills

	Stocks	T-Bills
Mean	1.0877	1.0073
Std Dev	0.205	0.0404
Variance	0.042	0.0016
N	78	78

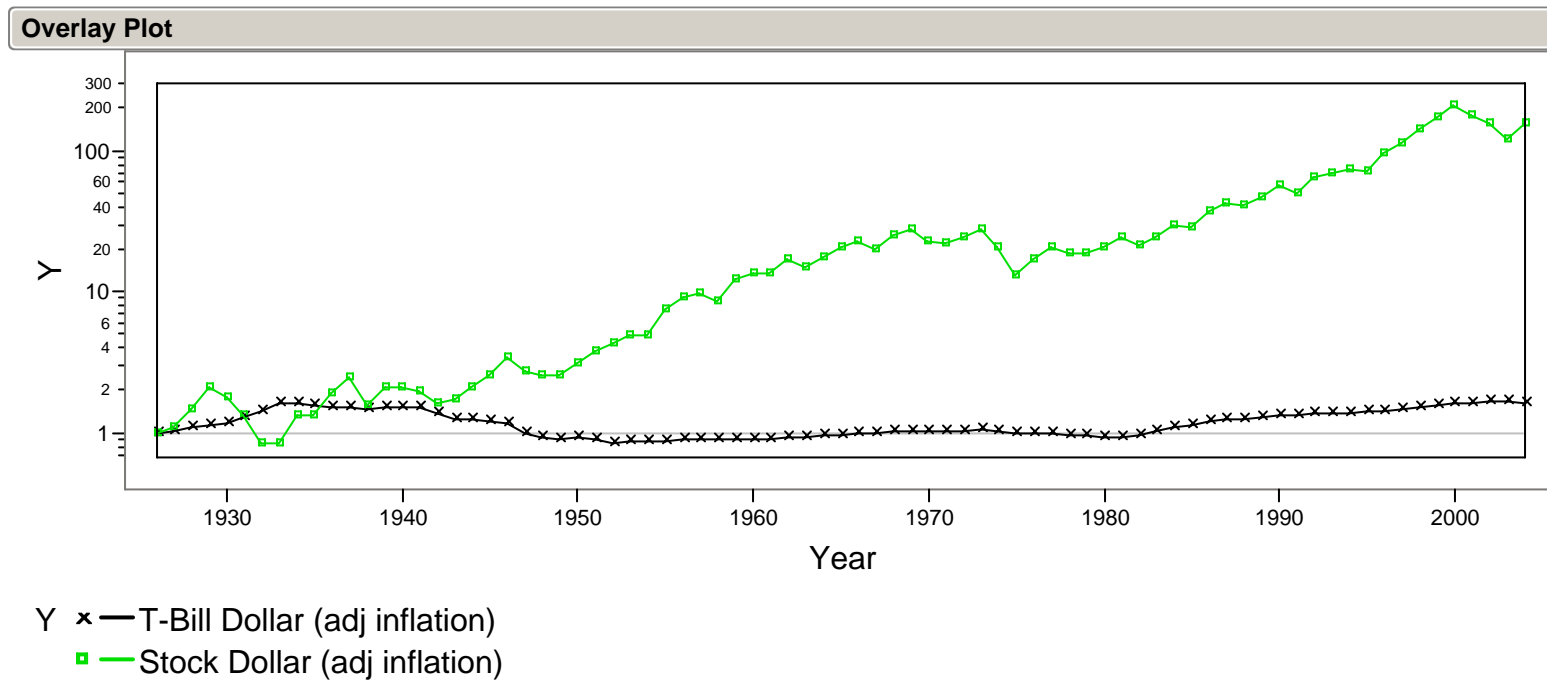
We made up Red. That's too bad!

The data for these indexes is contained in annual_markets.jmp.

The following Overlay Plot shows the performance through 2004 of these two indexes on the log scale (BBS, p.188-191)



The following plot (using the same scale), shows the indexes adjusted for inflation. Note that T-bills no longer appear to be so “risk-free”.



Take-Away Review

Random variables are used to model the returns on investments in finance.

The long-term value of an investment depends upon its average rate of return and its volatility.

Volatility is another name for the variance of the return.

Variance eats away the return via the volatility drag.

Mixing investments is a means to achieving better investments with higher long-term returns.

Next Module

Portfolios and covariance: how to mix investments to reduce risk while preserving returns. The simulation has two artificial aspects:

(1) We made up Red.

(2) The returns on each investment are independent of each other.

The next module shows how to make portfolios like Pink from real stock returns.

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Module 6

Covariance and Portfolios

Dependence and Independence

In many situations, we are interested in the simultaneous outcomes of two or more random variables.

For a randomly chosen person, we could define X = height and Y = weight

For a randomly selected store selling a particular product, X = price and Y = sales

For next January, X = return on Disney and Y = return on McDonalds

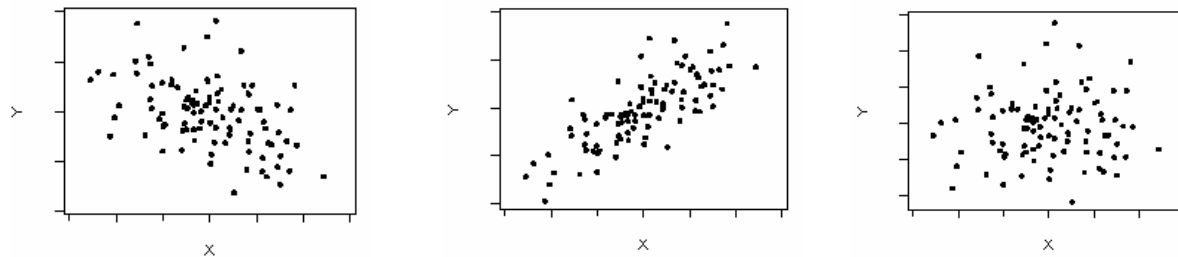
We can think of each of these pairs as a draw from a population of (x,y) values.

Definition: If knowing the outcome of X changes the probability distribution of Y (or vice-versa), then the rv's X and Y are said to be *dependent*. Otherwise, X and Y are said to be *independent*.

Are the above pairs of rv's dependent? How?

Linear Association

Suppose that we repeat the process that attaches values to X and Y, and then plot the joint outcomes. The plot might resemble one of the following:



Note that X and Y vary together along the diagonals in the first two plots.

Which of these would you associate with the previous X,Y pairs?

The tendency to cluster along the diagonals (seen in the first two plots) is called *linear association*

Does linear association imply dependence?

Does dependence imply linear association?

Covariance and Correlation

Covariance and correlation measure linear association between two sets of measurements that we denote by the n pairs of observations $(x_1, y_1), \dots, (x_n, y_n)$.

Examples:

For n individuals, x_i = height, y_i = weight

For n days, x_i = return on Disney, y_i = return on McDonalds

Covariance

The sample covariance between these two sets of measurements is defined as the average cross product

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Notice in the product, pairs with both values above (or below) the mean contribute positive values to the sum. Pairs with one value above the mean and another below contribute negative values to the sum.

Bad news: Covariance depends on the units in which X and Y are measured.

Correlation

The sample correlation between X and Y is defined as

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\text{sample covariance}(x, y)}{(\text{sample SD of } x)(\text{sample SD of } y)}$$

Good news: r_{XY} is unit-free. It measures linear association.

r_{XY} has the same sign as s_{XY} but satisfies $-1 \leq r_{XY} \leq 1$.

$s_{XY} > 0$ and $r_{XY} > 0$ indicate positive linear association.

$s_{XY} < 0$ and $r_{XY} < 0$ indicate negative linear association.

$s_{XY} = 0$ and $r_{XY} = 0$ indicate no linear association.

$r_{XY} = 1$ (or -1) indicates perfect positive (or negative) linear association.

What are the signs of s_{XY} and r_{XY} in the three plots on page 6-2?

Population Parameters

We have sample means and population means, sample variances and population variances. We also have population covariances and correlations. They are obtained as follows.

Let μ_X , μ_Y , σ_X , σ_Y denote the population means and standard deviations of X and Y , respectively. Then, analogous to the sample value (which is the average product)

$$\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Intuition: (When you see “E”, think “on average.”)

When $(X - \mu_X)$ and $(Y - \mu_Y)$ have the same sign, the contribution of $[(X - \mu_X)(Y - \mu_Y)]$ to $\text{Cov}(X,Y)$ is positive; otherwise it is negative.

Similarly, the population correlation is

$$\text{Corr}(X,Y) = \rho_{xy} = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Correlation and Independence

$\text{Cov}(X,Y) \neq 0$ or $\rho_{XY} \neq 0$ implies that X and Y are dependent

$\rho_{XY} \neq 0$ implies Dependence

X and Y independent implies $\text{Cov}(X,Y) = 0$ and $\rho_{XY} = 0$

Independence implies $\rho_{XY} = 0$

Careful!

$\text{Cov}(X,Y) = 0$ or $\rho_{XY} = 0$ does *not* imply independence. Why?

Think of a plot of data that has $\rho_{XY} = 0$ but is not independent. Think: dependence means knowing the position along the x-axis helps you to predict the position along the y-axis.

Weighted Sums of Random Variables

A weighted sum of random variables X and Y is

$$a X + b Y$$

where a and b stand for fixed numbers that are not random.

Jargon - Weighted sums such as these are often called linear combinations

Key application:

Portfolios of investments are weighted averages with the random variables denoting the returns on the component investments.

Example: From Module 5, Pink = .5 Red + .5 White

We will now see that

$E(a X + b Y)$ can be easily expressed in terms of $E(X)$ and $E(Y)$

$\text{Var}(a X + b Y)$ can be easily expressed in terms of $\text{Var}(X)$, $\text{Var}(Y)$ and $\text{Cov}(X, Y)$

FACT #1

For any weighted sum of random variables

$$E(a X + b Y) = a E(X) + b E(Y)$$

In Module 5 (pg 5-15), we saw a special case of this formula

$$\begin{aligned} E(\text{Pink}) &= E(.5 \text{ Red} + .5 \text{ White}) \\ &= .5 E(\text{Red}) + .5 E(\text{White}) \\ &= .5 (1.71) + .5 (1.008) = 1.359 \end{aligned}$$

FACT #2

For any weighted sum of independent random variables (like Pink)

$$\text{Var}(a X + b Y) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$

FACT #3

For *any* weighted sum of random variables, such as real stocks that are dependent,

$$\text{Var}(a X + b Y) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y)$$

In Module 5 (pg 5-15), we saw a special case of this formula

$$\begin{aligned}\text{Var}(\text{Pink}) &= \text{Var}(.5 \text{ Red} + .5 \text{ White}) \\ &= .5^2 \text{Var}(\text{Red}) + .5^2 \text{Var}(\text{White}) + 2(.5)(.5)\text{Cov}(\text{Red}, \text{White}) \\ &= .5^2 (1.755) + .5^2 (.002) = .439\end{aligned}$$

Some useful special cases of these facts:

$$E(a X) = a E(X), \quad \text{Var} (a X) = a^2 \text{Var}(X), \quad \text{SD}(a X) = |a| \text{SD} (X)$$

For independent X and Y, the variance of a sum is the sum of the variances

$$\text{Var} (X + Y) = \text{Var} (X) + \text{Var} (Y), \quad \text{SD}(X + Y) = \sqrt{\text{Var}(X) + \text{Var}(Y)}$$

The same applies to differences of independent random variables¹

$$\text{Var} (X - Y) = \text{Var} (X) + \text{Var} (Y), \quad \text{SD}(X - Y) = \sqrt{\text{Var}(X) + \text{Var}(Y)}$$

¹ It seems natural that the variance of a sum might be the sum of the variances. But why should the variance of the difference also be the sum of the variances? It might help to think of $\text{Var}(X-Y)$ as $\text{Var}(X+(-Y)) = \text{Var}(X) + \text{Var}(-Y)$. Changing the sign of Y does not change its variance, $\text{Var}(Y) = \text{Var}(-Y)$.

The Mean and Variance of Some Real Portfolios

Let's now move beyond the Pink portfolio and consider portfolios of real stocks. Two new aspects need to be considered:

- 1) Returns on the individual investments are typically not independent of each other
- 2) The probability distributions of the returns on the individual investments are unknown

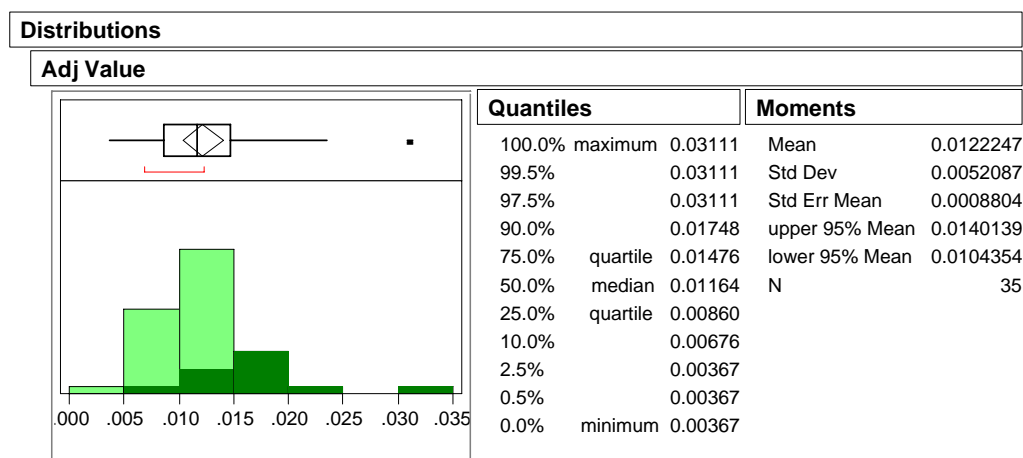
Thus

- 1) We cannot assume independence
- 2) The unknown characteristics must be estimated from data

Stock Market Data Files

StockReturns.jmp contains the monthly (net) returns on the stocks of 35 companies from 1975-1999. StockReturnsSummary.jmp contains a summary of these returns. These data were obtained from the Center for Research in Security Prices (CRSP), which is available to you on WRDS. Another analysis of stocks appears in BBS (p. 197-204).

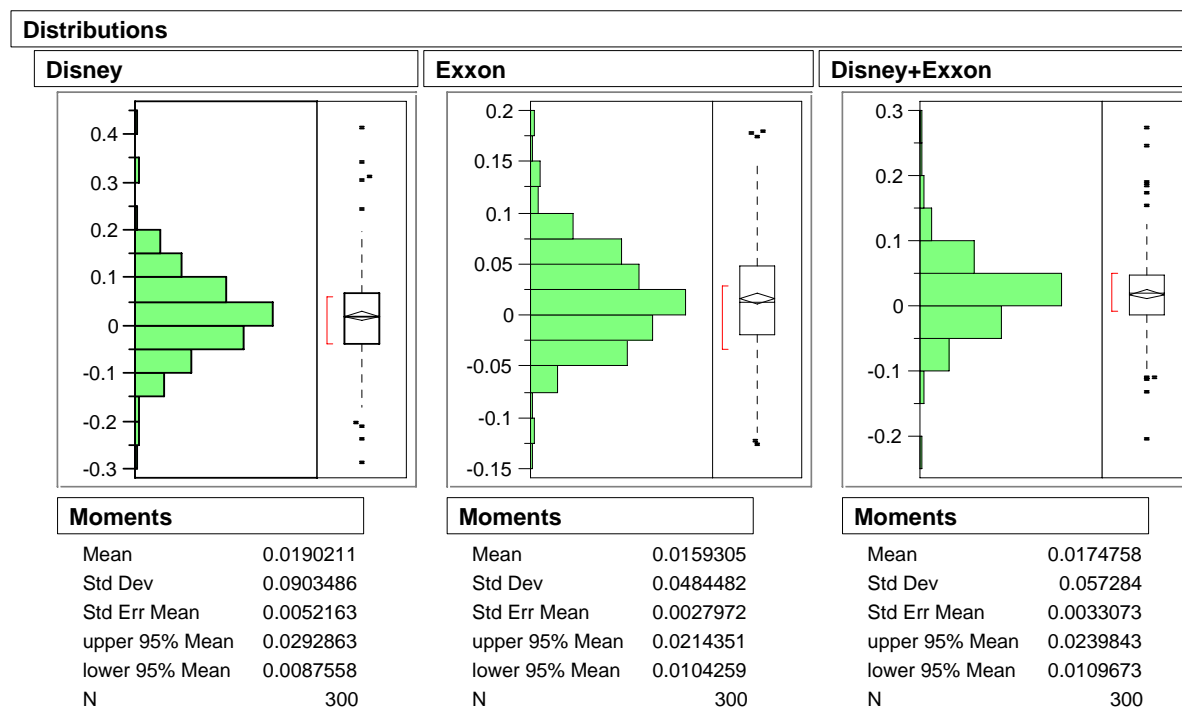
The 35 companies were chosen as the dominant companies in 7 industries in December 1974. Only 22 firms have data that spanned the entire 300 month period. The following output summarizes the volatility-adjusted returns for all 35 stocks. The highlighted stocks lasted less than 200 months during 1975-1999.



What happened to the firms that dropped out of the sample?

Of the stocks that survived the whole 25 years (300 months), Disney had the highest volatility-adjusted return 0.0149. Exxon was a close second with adjusted return 0.0148.

Let's combine these into the equally-weighted portfolio² $\text{DisExx} = .5 \text{ Disney} + .5 \text{ Exxon}$



² As in the dice simulation, this portfolio rebalances after each month so that half of the value of the portfolio is kept in Disney, and the other half in Exxon.

From these summaries³ we obtain the volatility-adjusted returns⁴ as follows:

Investment	Average	Variance	Avg – Var/2
Disney	.0190	.0082	.0149
Exxon	.0159	.0023	.0148
DisExx	.0174	.0033	.0158

Once again, a portfolio offers an improvement in long-term gains over investing 100% in either of the individual investments.

Looking for higher returns, why should we use an equally weighted mixture of these two stocks?

Why these companies?

Why these weights (50/50)?

Why only two?

³ We can get the variance by simply squaring the SD that appears in the output. You can also ask JMP to give you the variance as well. You do this by adding “more moments” to the output of the description command.

⁴ The volatility adjusted return formula for gross returns from Module 5 (pg 5-13) also holds for net returns.

Role of Covariance

For the equally weighted portfolio of Disney and Exxon stocks, the return on the portfolio is just the average of the returns on the two stocks

$$\text{Avg}(\text{DisExx}) = .01774 = .5(.0190) + .5(.0159)$$

but for the variance we find

$$\text{Var}(\text{DisExx}) = .0033 \neq .5^2(.0082) + .5^2(.0023)$$

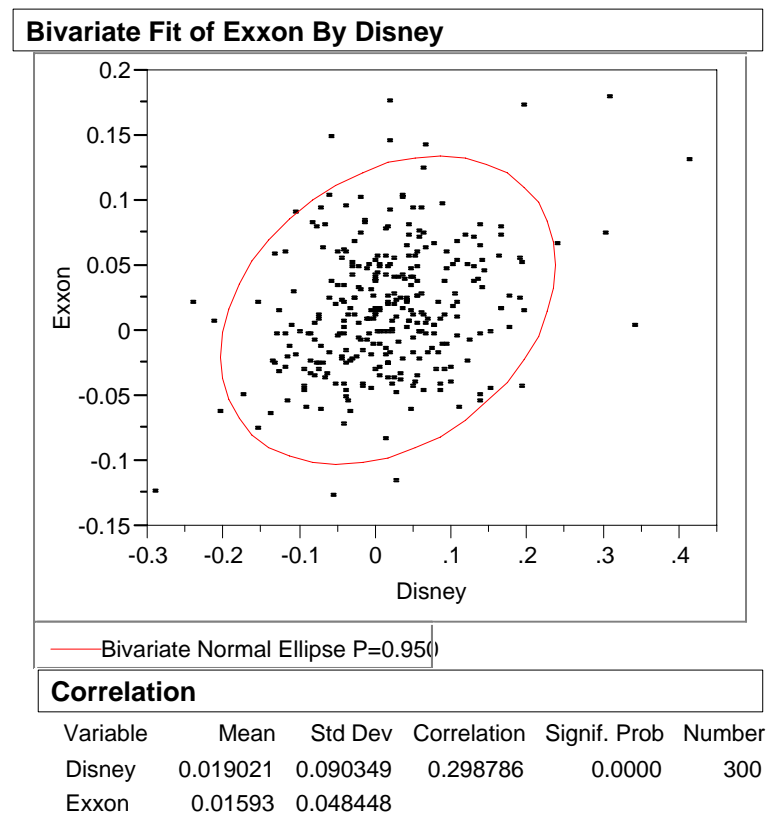
which was the formula we used for Pink. What's missing?

Using the complete formula for the variance of a weighted sum (Fact #3, page 6-9), the value of $\text{Var}(\text{DisExx})$ implies

$$.0033 = .5^2(.0082) + .5^2(.0023) + 2(.5)(.5)\text{Cov}(\text{Dis}, \text{Exx})$$

Solving for the Cov term, we can calculate that the covariance between Disney and Exxon is $\text{Cov}(\text{Dis}, \text{Exx}) = .00131$. What would be a preferable covariance value?

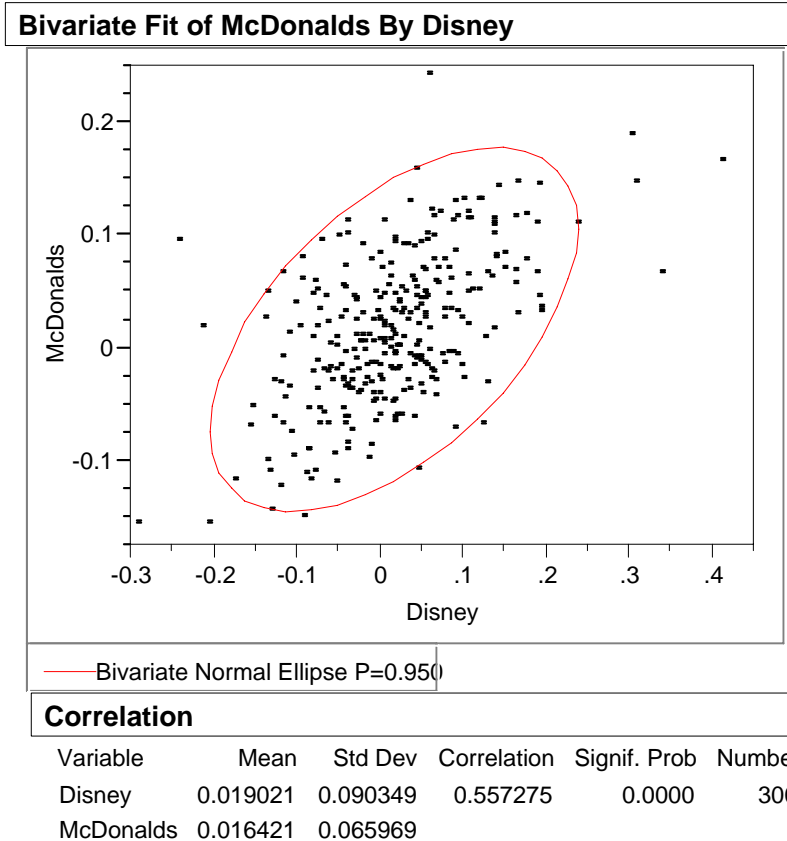
We can see⁵ that returns on Disney and Exxon are related, though only weakly.



Note that the sample correlation between Disney and Exxon is .298
(= .00131/(.090)(.0484)).

⁵ This output is obtained using Analyze > Fit Y by X with Exxon as Y and Disney as X, and selecting Density Ellipse .95. The more “tilted” the football shape of the ellipse, the more linearly dependent the pair of measurements.

Let's now consider pairing Disney with McDonalds which had the third largest volatility adjusted return, at .0142.



Compared to Exxon, McDonalds has a larger average return .0164, but McDonalds also has a larger variance .0044 and a higher correlation .557 with Disney. The plot confirms the stronger positive relationship (notice the more flattened elliptical shape).

For the equally-weighted portfolio of Disney and McDonalds we obtain

Investment	Average	Variance	Avg – Var/2
Disney	.0190	.0082	.0149
McDonalds	.0164	.0044	.0142
DisMcD	.0177	.0048	.0153

DisMcD does not outperform DisExx. Why?

Constructing an “Optimal” Portfolio

The pairwise portfolios that we have considered thus far put equal weight on the two stocks in the portfolio.

More generally, for two investments with returns X and Y , we could consider portfolios of the form

$$Z = p X + (1-p) Y$$

where p is the proportion⁶ of the total invested in X .

Given the means, variances, and covariance between X and Y , we can use our previous facts to obtain

$$E(Z) = p E(X) + (1-p) E(Y)$$

$$\text{Var}(Z) = p^2 \text{Var}(X) + (1-p)^2 \text{Var}(Y) + 2p(1-p) \text{Cov}(X, Y)$$

⁶ Ordinarily p is chosen between 0 and 1. However, negative values could be used to represent short selling.

An “optimal” portfolio can then be obtained by choosing the value of p which maximizes the volatility adjusted return, $E(Z) - \text{Var}(Z)/2$.

This problem can be solved with the Solver in Excel.⁷

For example, the optimal Disney-Exxon portfolio is .52 Disney + .48 Exxon yielding a volatility adjusted return of .0158 (only a small improvement).

The optimal Disney-McDonalds portfolio is .62 Disney + .38 McDonalds yielding a volatility adjusted return of .0154 (again only a small improvement).

⁷ The Excel files portfolio1.xls and portfolio2.xls are set up to do the Solver optimization for the Disney-McDonalds and the Disney-Exxon portfolios. Just select Solver from the Tools menu.

Larger Portfolios

More importantly, why restrict portfolios to only two investments?

More generally, for K investments X_1, \dots, X_K and weights p_1, \dots, p_K (that sum to 1), the general form of a portfolio is given by (another linear combination, just bigger)

$$Z = \sum_{i=1}^K p_i X_i$$

The mean and variance of Z are given by the formulas

$$E(Z) = \sum_{i=1}^K p_i E(X_i)$$
$$\text{Var}(Z) = \left(\sum_{i=1}^K p_i^2 \text{Var}(X_i) \right) + \left(\sum_{i \neq j} p_i p_j \text{Cov}(X_i, X_j) \right)$$

(Remain calm – these formulas are just straightforward generalizations of the two variable formulas⁸)

⁸ At least remain calm for now. You'll see them again in Finance. They *really* like these formulas!

As before, an “optimal” portfolio might then be obtained by choosing the values of p_1, \dots, p_K which maximize the volatility adjusted return $E(Z) - \text{Var}(Z)/2$.

Another strategy that is sometimes considered fixes the “risk” of the portfolio (its variance) and then chooses the values mixture weights p_1, \dots, p_K that maximize the expected return.

For example, using the Solver in Excel we find that for a portfolio of a subset of our 35 stocks

Var(Z)	Maximum E(Z)	E(Z) - Var(Z)/2
0.00135	0.0139	0.013225
0.00140	0.0149	0.014200
0.00150	0.0161	0.015350
0.00200	0.0191	0.018100
0.00300	0.0225	0.021000
0.00400	0.0250	0.023000
0.00500	0.0271	0.024600

Note how increased volatility allows for increased expected returns.⁹

⁹ At some point, however, accepting more risk will not lead to higher volatility adjusted returns. You’ll see more of this concept when you study the “efficient frontier” in Finance.

For our previous two stock portfolios we found

Investment	Average	Variance	Avg – Var/2
DisMcD	.0177	.0048	.0153
DisExx	.0174	.0033	.0158

How do the larger portfolios compare to these?

Caveat!

The means, variances and covariances may change in the future. Remember Long Term Capital Management?

Take-Away Review

Portfolios can be represented as weighted sums of random variables, allowing us to study methods for optimizing the performance of a portfolio.

Finding the variance of such a portfolio requires that we know the covariances among the different investments.

Covariance and correlation measure the strength of linear association between pairs of random variables.

$\text{Cov}(X, Y) \neq 0$ implies dependence, but
 $\text{Cov}(X, Y) = 0$ *does not* imply independence.

Covariance is more useful for working with portfolios, but correlation will become more useful in Stat 621.

Next Module

Samples and populations are key ideas in inferential statistics, and we next take a closer look at these concepts.

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Module 7

Sampling, Sampling Distributions and Standard Errors

The Population-Sample Paradigm

Treat the observed data as a sample from a population

Use sample characteristics to make inferences about population characteristics

Some Managerial Examples where sampling is useful

Electrical supplier estimates the proportion of defectives in a large batch

Retail firm estimates market share of a product

Personnel manager estimates variability of hourly wages across an industry

Target Population versus Sampling Frame

Sample	Target Pop'n	Sampling Frame
100 Incomes	US Incomes	
Political Poll	Actual voters	
Mall Taste Test	Soup consumers	
CNN Poll of Callers	US opinions	
Sample 10 Goats	All Goats	

Sampling bias is a mismatch between the target population and the sampling frame

Typical causes of sampling bias: self selection, non-response, incentives to answer, interviewer characteristics, formulation of questions, sensitivity of questions
(see BBS, p.108)

Hypothetical Populations

Suppose a genetic scientist at an agricultural company harvests 200 oranges, the first of a new variety. Are these a sample from a population of interest?

From which populations might the following be considered a sample

The 100 motor shaft diameters in ShaftDia.jmp (Module 3)

The 507 daily returns in GM92.jmp (Module 2)

A bag of M&M candies

(Simple) Random Sampling

Random Sampling - Every possible subset of a given size has an equal chance of being drawn.

Can be obtained as a sequence of individual random draws from the population¹

Sampling without replacement - items can only be selected once

Sampling with replacement - items can be selected repeatedly

When will sampling with replacement be virtually the same as sampling without replacement?

Random sampling should be done with a device that provides random selection.
Careful! Haphazard \neq Random.

Other Sampling Designs - systematic sampling, stratified sampling, cluster sampling, multistage sampling

¹A sequence of independent draws from specified probability distributions can be obtained with JMP. From the list of “random” functions, consider choices such as random uniform or random normal.

iid Sampling

We shall be especially interested in simple random samples obtained by sampling a population of conceptually infinite size.²

Real populations have finite size, but it's often reasonable to treat them as infinite when the size of the sample is small relative to the size of the population.

In this case, the data x_1, \dots, x_n can be thought of as

- (1) n independent realizations of random variables
- (2) all with the same probability distribution

This common distribution is that of the population

² An alternative artificial way to avoid the complications of sampling from a finite population is to assume that we sample with replacement. If we put them back, the result of one case does not influence other cases.

Such samples are called *iid samples*:

iid = independent and identically distributed

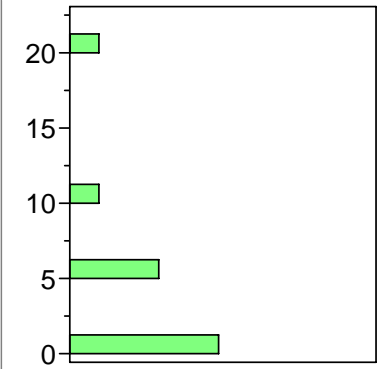
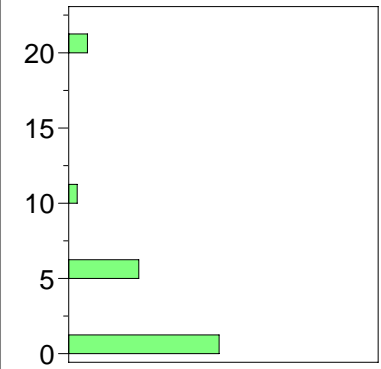
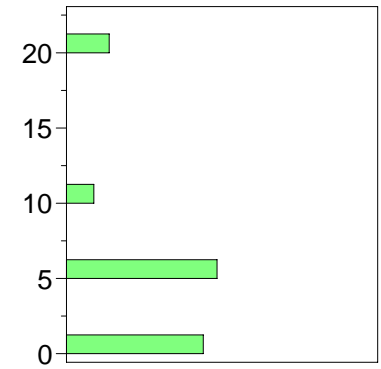
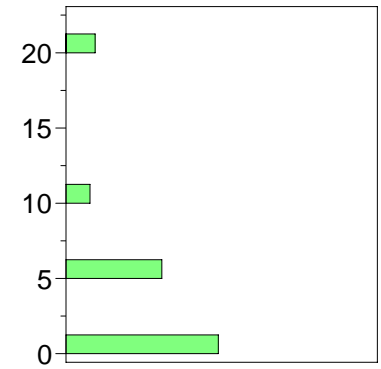
Notation: We'll use $x_1, \dots, x_n \text{ iid} \sim N(\mu, \sigma^2)$ to denote the data of an iid sample from the normal distribution $N(\mu, \sigma^2)$.

What aspects of the GM 92 daily returns (pg 2-18 and BBS p.27-28) support an assumption that

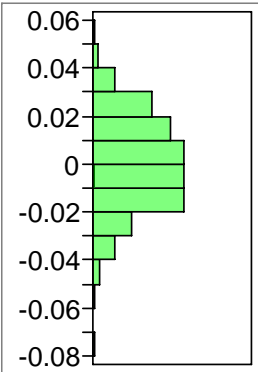
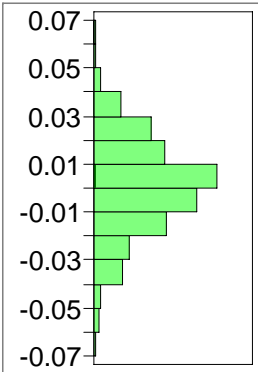
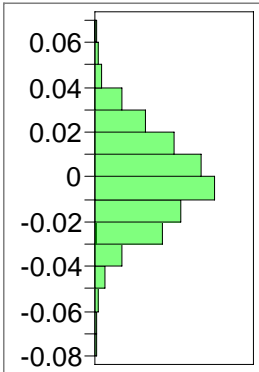
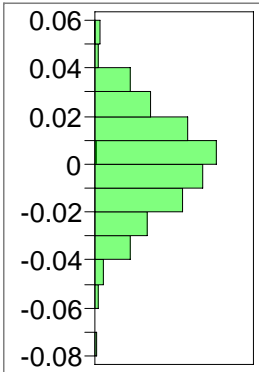
$$x_1, \dots, x_{507} \text{ iid} \sim N(\mu, \sigma^2)?$$

Key benefit: Statistical theory shows that characteristics of iid samples tend to emulate population characteristics.

Example 1: Simulating 50 iid chip selections (with Chipsim.jmp)

Distributions			
Pop'n	Sample 1	Sample 2	Sample 3
			
Moments	Moments	Moments	Moments
Mean 5	Mean 4	Mean 5.68	Mean 4.9
Std Dev 5.8028846	Std Dev 5.3069073	Std Dev 5.9434066	Std Dev 5.7578695
Std Err Mean 0.8206518	Std Err Mean 0.75051	Std Err Mean 0.8405246	Std Err Mean 0.8142857
upper 95% Mean 6.6491615	upper 95% Mean 5.5082064	upper 95% Mean 7.3690975	upper 95% Mean 6.5363684
lower 95% Mean 3.3508385	lower 95% Mean 2.4917936	lower 95% Mean 3.9909025	lower 95% Mean 3.2636316
N 50	N 50	N 50	N 50

Example 2: Simulating x_1, \dots, x_{507} iid $\sim N(0, .02^2)$ (with Normsim.jmp)

Distributions							
Sample 1		Sample 2		Sample 3		Sample 4	
							
Moments		Moments		Moments		Moments	
Mean	0.0005489	Mean	0.0007621	Mean	-0.001571	Mean	-0.000283
Std Dev	0.0196285	Std Dev	0.0205766	Std Dev	0.0207258	Std Dev	0.0204095
Std Err Mean	0.0008717	Std Err Mean	0.0009138	Std Err Mean	0.0009205	Std Err Mean	0.0009064
upper 95% Mean	0.0022615	upper 95% Mean	0.0025575	upper 95% Mean	0.0002372	upper 95% Mean	0.001498
lower 95% Mean	-0.001164	lower 95% Mean	-0.001033	lower 95% Mean	-0.00338	lower 95% Mean	-0.002064
N	507	N	507	N	507	N	507

Sample Estimates of Population Parameters

Simulation examples are artificial because we know population features such as μ , σ^2

In real problems, these population features – called *parameters* – are not known and must be *estimated* from data.

The sample statistics \bar{x} , s^2 and s are typically used to estimate μ , σ^2 and σ

For the GM92 returns data we computed $\bar{x} = .00158$ and $s = .0202$.

Based on our normal simulation results, does it seem plausible that $\mu = 0$ and $\sigma = .02$ in the GM92 population?

A Class Experiment

Organize into teams of 2 or 3 students.

Every team will receive a bag of M&M candies.

Is it reasonable to treat the contents of your bag as an iid sample from a population?
Which population?

Estimate the population proportion of blue M&M's using only the information in your sample.

Will every team come up with the same estimate? Why not?

What range of values do you think will be found by the class?

Note that a sample proportion is a special case of \bar{x} . Why?

The Sampling Distribution of a Statistic

As previous examples show, sample estimates, such as \bar{x} or s , do not match μ and σ and vary from sample to sample. Once we admit that we would have gotten a different value if we had gotten a different sample, we need to describe just how different the result might have been.

To quantify this *sample-to-sample variation*, we introduce two new populations:

The *population of samples* – the distribution of all possible samples of size n that could be drawn from the original population.

The *population of values of the sample statistic*– the distribution of all possible values of the sample statistic (one for each sample).

Definition: The population of sample statistic values is called the *sampling distribution of the statistic*.

Example: For the class M&M experiment

What is the population of samples? What is the population of sample statistic values?

How do the class samples and estimates relate to these populations?

The Sampling Distribution of \bar{X}

Astonishing Fact: For x_1, \dots, x_n iid from any population with mean μ and standard deviation σ , the sampling distribution of \bar{x}

- a) is approximately normal when n is large, and so is essentially determined by the mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$
- b) has mean $\mu_{\bar{x}} = \mu$
- c) has standard deviation $\sigma_{\bar{x}} = (\sigma / \sqrt{n})$

We can use this fact to quantify the amount of sampling variation of \bar{x} from the information in just *one sample*.

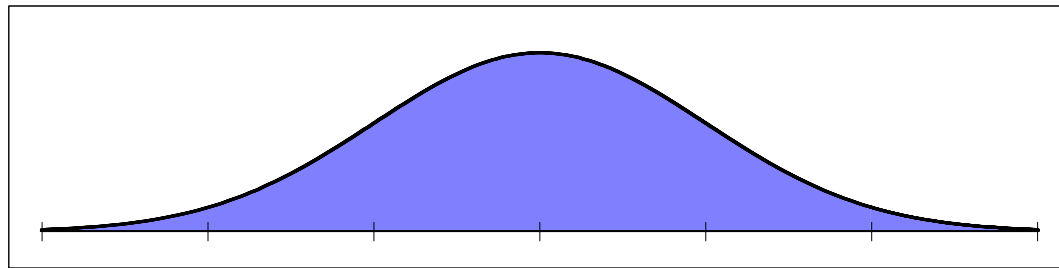
Remarks:

This is known as the Central Limit Theorem (CLT). For practical purposes, normality can be assumed when $n \geq 15$.

When the original population is exactly normal, then the sampling distribution of \bar{x} will also be exactly normal

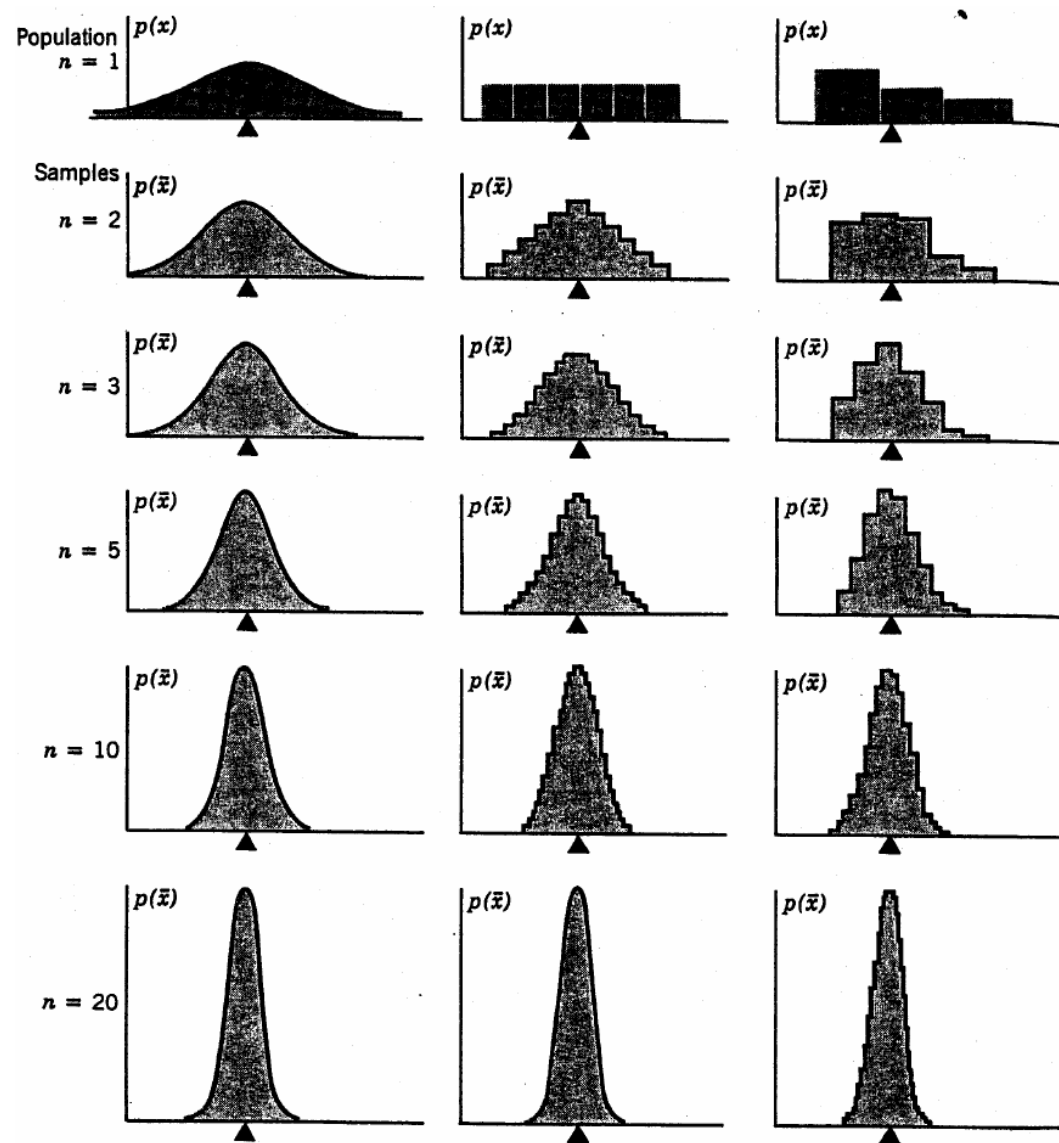
The Astonishing Fact says that the sampling distribution of the sample mean \bar{x} is always approximately $N(\quad, \quad)$.

Pictorially:



The Astonishing Fact in Action!

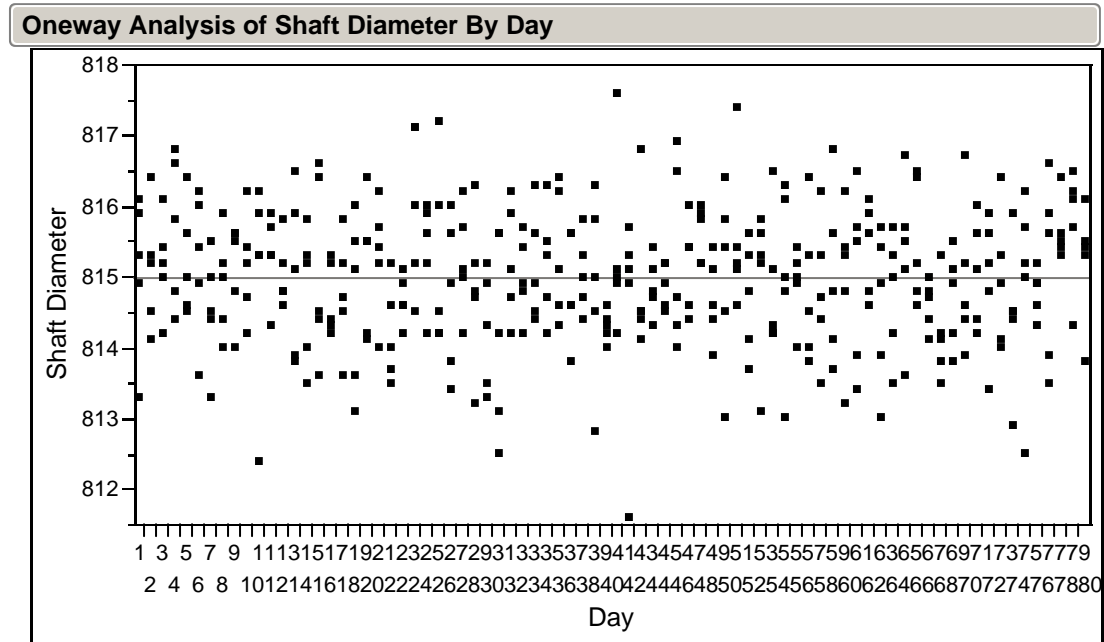
The following slide shows the exact sampling distribution of \bar{x} for three different populations and various values of n . Regardless of the shape of the population, the sampling distribution of the average gets closer and closer to a normal distribution.



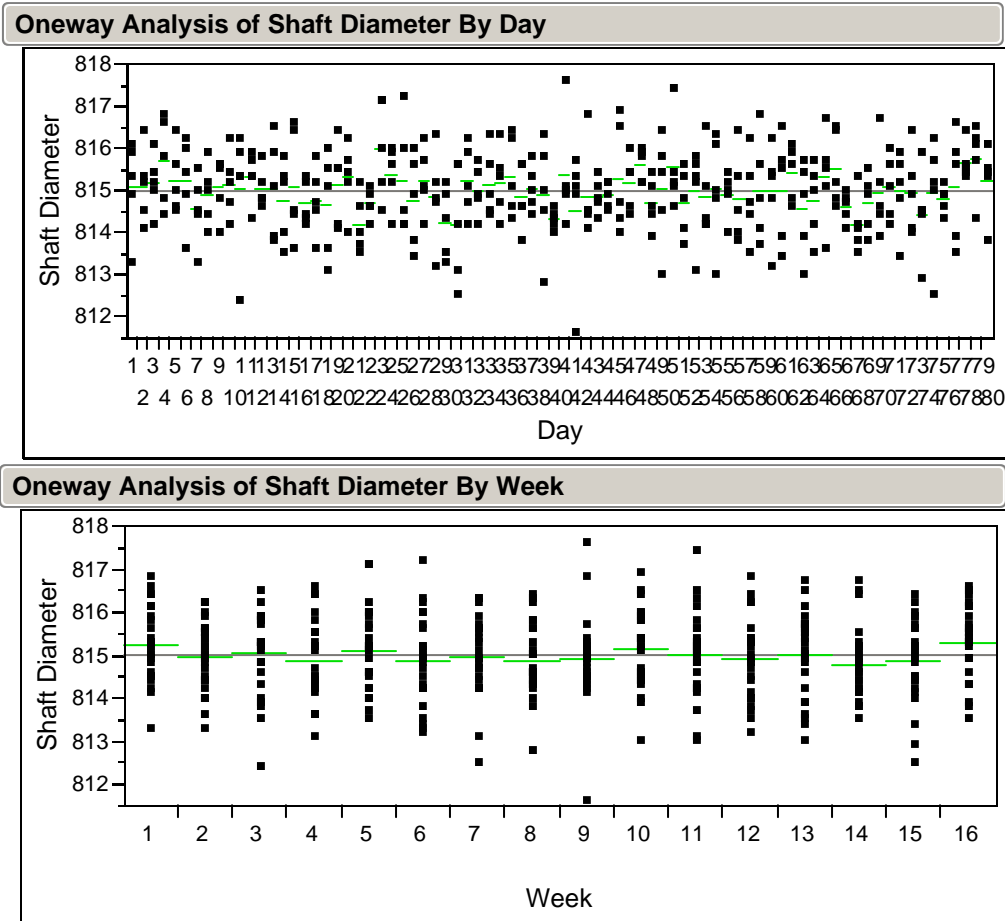
Motor Shaft Data

(BBS, p.68)

A quality control application provides another chance for us to see the sample-to-sample variation of the average. The file ShaftXtr.jmp contains 400 observations on the motor shaft diameters. Five observations were taken per day for 16 weeks.

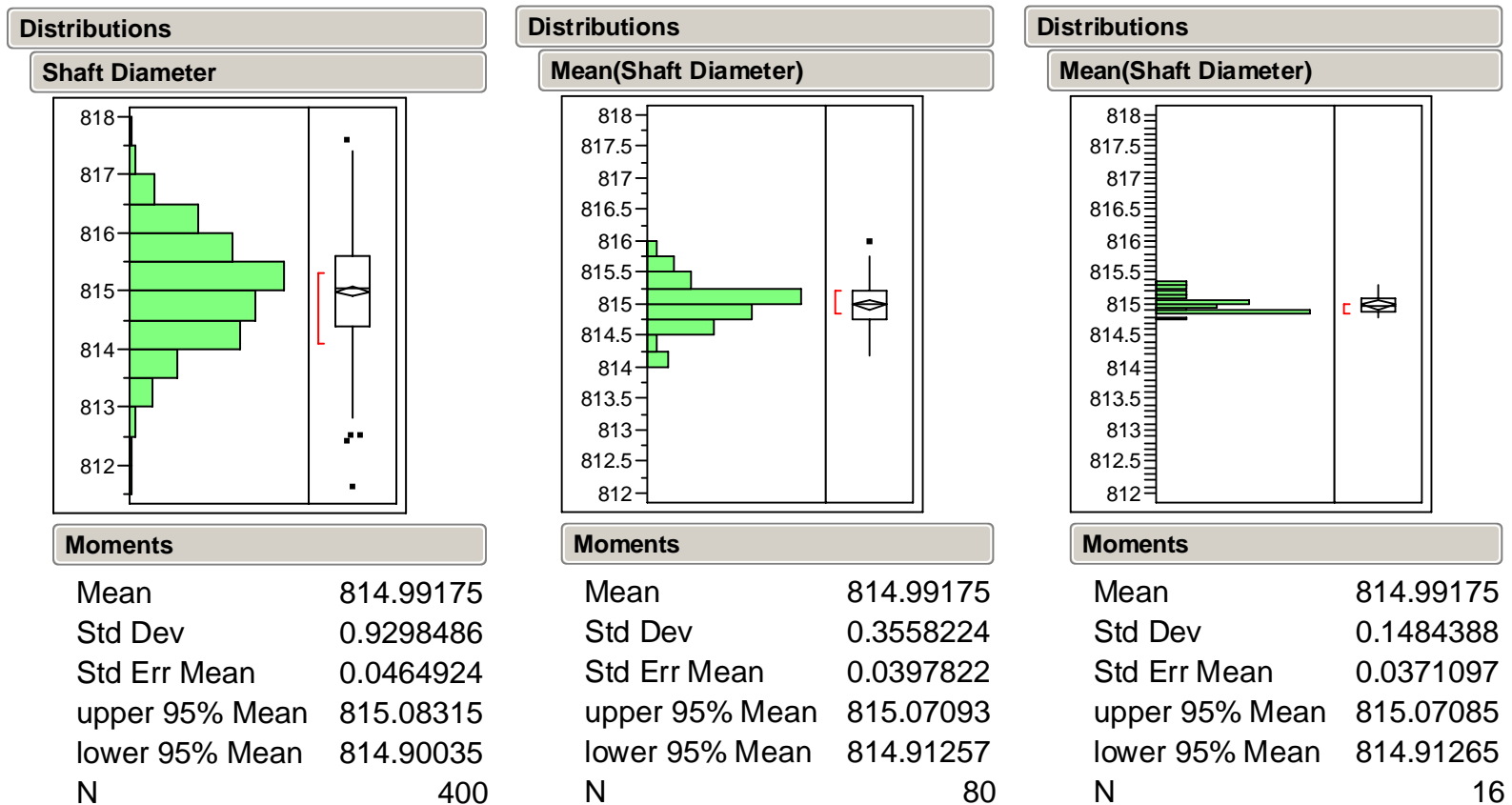


Look what happens when we instead plot daily means and weekly means



The individual diameters can be considered as sample means for samples of size $n = 1$.
 The daily averages can be considered as sample means for samples of size $n = 5$
 The weekly averages can be considered as sample means for samples of size $n = 25$

Summaries of the individual, daily and weekly means (BBS, p.69)



What is the difference between N and n here?

Notice that the three StdDev values here get smaller as n gets larger

The Standard Error of \bar{X}

The reason that we care about the sampling variation of \bar{x} is that we can use this property of the statistic to find out how close we are to the population mean. The astonishing fact about the sampling distribution of \bar{x} enables us to make general statements like

$$P(\mu - 2\sigma/\sqrt{n} \leq \bar{x} \leq \mu + 2\sigma/\sqrt{n}) \approx$$

What does this convey about the proximity of \bar{x} to μ ?

Because $\sigma_{\bar{x}} = (\sigma / \sqrt{n})$ is unknown in practice, it is usually estimated by

$$s_{\bar{x}} = (s / \sqrt{n}),$$

which is called the *standard error* of \bar{x}

$s_{\bar{x}}$ is an estimate of the standard deviation of the sampling distribution of \bar{x}

$s_{\bar{x}}$ should be reported along with \bar{x} to indicate its precision

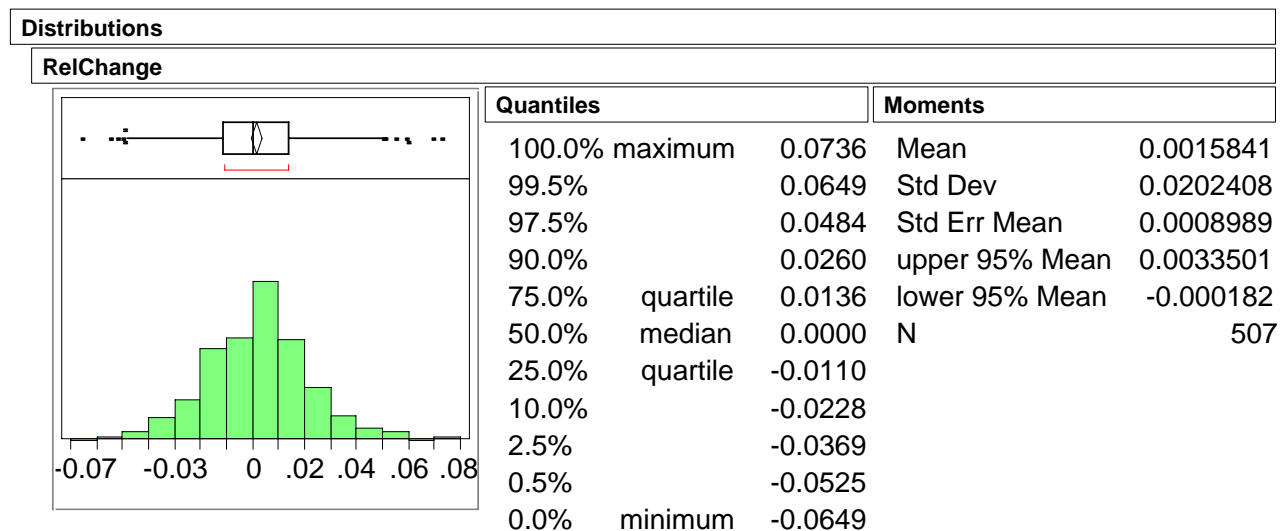
Example: Returns on GM stock

For the GM92 data we estimated μ by $\bar{x} = .00158$, and estimated σ by $s = .0202$, with $n = 507$ observations.

Here, the standard error of \bar{x} is $.0202/\sqrt{507} = .00090$

What does this standard error convey?

The previous JMP output for the GM92 returns reports the standard error of \bar{x}



The Standard Error of the M&M Estimates

We saw before that the standard deviation of the class estimates was

Bad news: We needed all the class samples to calculate this number. This is useless when we only have one sample.

Good news: An effective alternative is the standard error $s_{\bar{x}} = (s / \sqrt{n})$, and it can be calculated from a single sample!

Using one bag of M&M's, we obtain³

$$\bar{x} = \quad \text{and} \quad s / \sqrt{n} =$$

Is this close to our class findings?

³ An easily computed approximation to s in this context is $\sqrt{\bar{x}(1-\bar{x})}$

Take-Away Review

The information in iid samples allows us to

Estimate population parameters like as μ , σ^2 by the corresponding statistics based on the sample, \bar{x} and s .

As well as to

Estimate how close the sample statistics are likely to come to the corresponding population parameters. The key ingredient is the standard error of the statistic.

The standard error of \bar{x} is s/\sqrt{n} and can be estimated from the same sample used to calculate \bar{x} . It estimates σ/\sqrt{n} , the standard deviation of the sampling distribution of \bar{x} .

Next Module

When we combine standard error with the implications of the Central Limit Theorem, we can make profound statements about features of the population with *confidence intervals*.

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Module 8
Confidence Intervals

A 95% Confidence Interval for μ

A convenient way to convey the precision of a statistical estimate is to report a range of “probable” values of the parameter. This is done by reporting a *confidence interval*.

When x_1, \dots, x_n is an iid sample from a population

$$\left(\bar{x} - 2s / \sqrt{n}, \bar{x} + 2s / \sqrt{n} \right)$$

is called an (approximate¹) 95% confidence interval (CI) for the population mean μ

The “95%” part of this is known as the *confidence level* of the interval.

¹ Approximate because we use 2 and estimate σ , the population SD, by s , the sample SD..

Example

For the stock prices of General Motors (GM92), $\bar{x} = .0016$ and $s = .0202$ and $n = 507$

The resulting (approximate) 95% CI for μ is

$$(.0016 - 2(.0202)/\sqrt{507}, .0016 + 2(.0202)/\sqrt{507}) = (-.0002, .0034)$$

Key property: 95% CIs contain the population mean μ “about 95% of the time.”

To see why, recall from page 7-8 that

$$P\left(\mu - 2\sigma/\sqrt{n} \leq \bar{x} \leq \mu + 2\sigma/\sqrt{n}\right) \approx 95\%$$

so if we estimate σ by s , then

$$P\left(\mu - 2s/\sqrt{n} \leq \bar{x} \leq \mu + 2s/\sqrt{n}\right) \approx 95\%$$

Thus, \bar{x} is within $2s/\sqrt{n}$ of μ for about 95 % of samples.

Exact Confidence Intervals For μ

When x_1, \dots, x_n iid $\sim N(\mu, \sigma^2)$, exact confidence intervals for μ take the form

$$\left(\bar{x} - t s / \sqrt{n}, \bar{x} + t s / \sqrt{n} \right)$$

where t is a predetermined constant that depends on the sample size n and the desired confidence level (again, this is usually 95%).

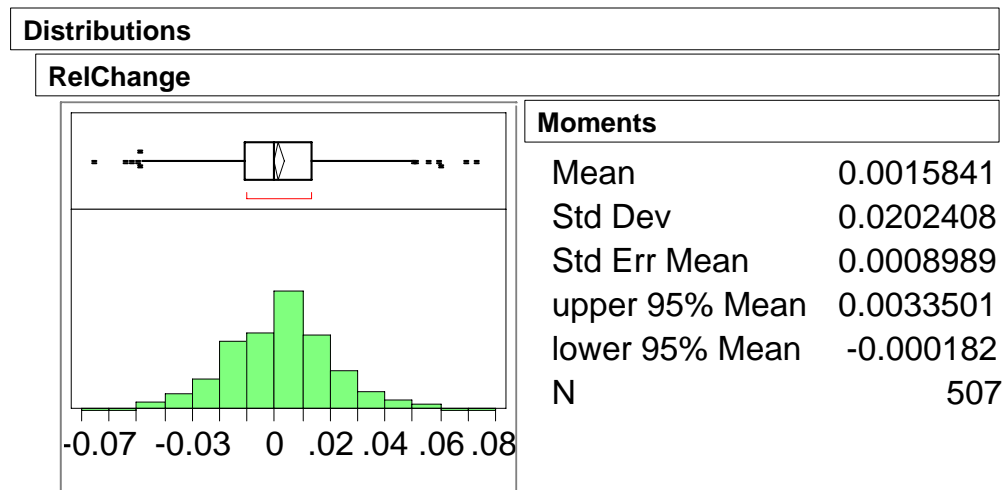
For example, if $n = 20$, we can see the “price” of not knowing σ and having to estimate it from the data. The confidence interval is longer.

$t = 2.09$ yields a 95% CI and $t = 1.72$ yields a 90% CI.

Good news: JMP provides² exact 95% CI limits for μ

² Using Analyze > Distribution, it is listed in the Moments output.

Example: The 95% CI for μ for the GM92 returns



The 95% CI for μ here uses the exact procedure. How does it compare with the approximate interval $(-.0002, .0034)$ on page 8-2?

The " ± 2 standard error" Rule of Thumb

Unless n is small (≤ 30) and precise confidence is needed, the approximate 95% CI bounds given by $\bar{x} \pm 2s / \sqrt{n}$ are fine.

Confidence Intervals are not Tolerance Intervals!

CAREFUL! When x_1, \dots, x_n are iid $\sim N(\mu, \sigma^2)$, the (approximate) 95% tolerance interval³ for a future draw from this population is defined as

$$(\bar{x} - 2s, \bar{x} + 2s)$$

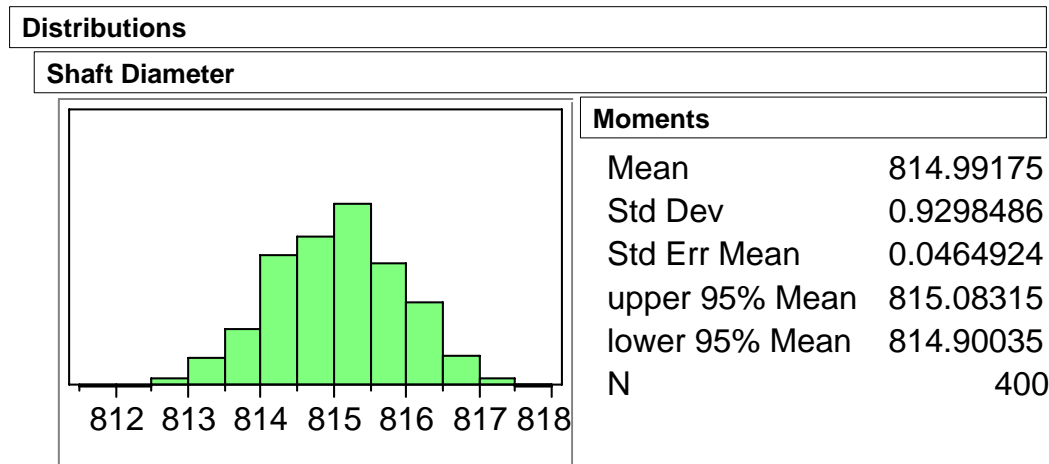
This is not the same as the (approximate) 95% CI for the population mean μ

$$\left(\bar{x} - 2s / \sqrt{n}, \bar{x} + 2s / \sqrt{n} \right)$$

The tolerance interval is wider because individual values are more variable than averages.

³ Tolerance intervals are a special case of prediction intervals that we'll see in Stat 621.

Example: Let's again revisit the 400 observations of motor shaft diameters in ShaftXtr.jmp. (BBS, p.97)



A 95% CI for the mean of the shaft-making process is (814.90, 815.08).

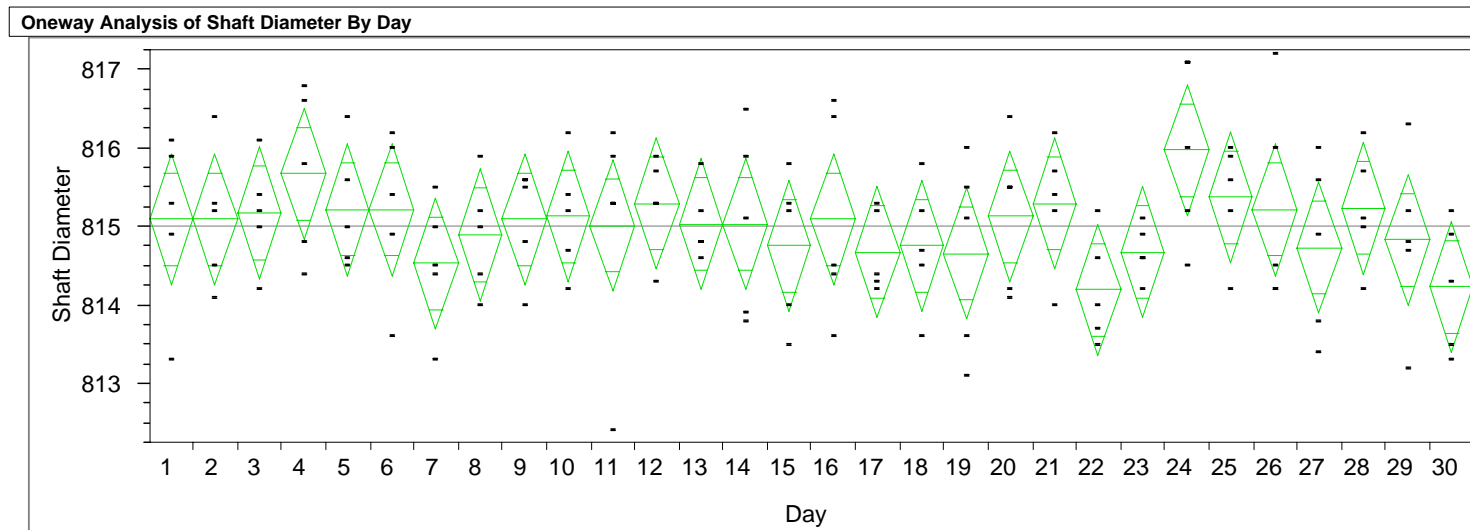
Assuming the process remains in control (recall control charts, as shown on BBS p.75), an approximate 95% tolerance interval for the diameter of a future shaft is

$$(814.99 - 2(.93), 814.99 + 2(.93)) = (813.13, 816.85).$$

Which interval is wider?

Why?

Suppose that instead of one overall confidence interval, we wanted separate 95% CIs for each day. JMP easily provides these as⁴ (BBS, p.101)



The center of each diamond is the daily sample mean and the tips of each diamond are the 95% confidence limits for the true daily mean.

Note that 95% CI for day 24 does not cover the overall mean. Is this worrisome?

⁴ Use the Fit Y by X command, right click on the title bar, selecting Display Options and then Mean Diamonds. You can only get this display if the X variable in the plot is ordinal or nominal. We only consider 30 days here so that the plot fits on the page.

Choosing a Needed Sample Size n

The width of the 95% CI for the mean of the motor shaft process is (using all of the data)
 $815.1 - 814.9 = 0.2$ (page 8-6)

What would happen to the width of the 95% CI for the mean if the sample size n was larger?

Using the ± 2 standard error rule of thumb, the width of a 95% CI for the mean is

$$\text{Width} = 4s / \sqrt{n}$$

Suppose we doubled the sample size from n to $2n$. What would happen to the width of the interval?

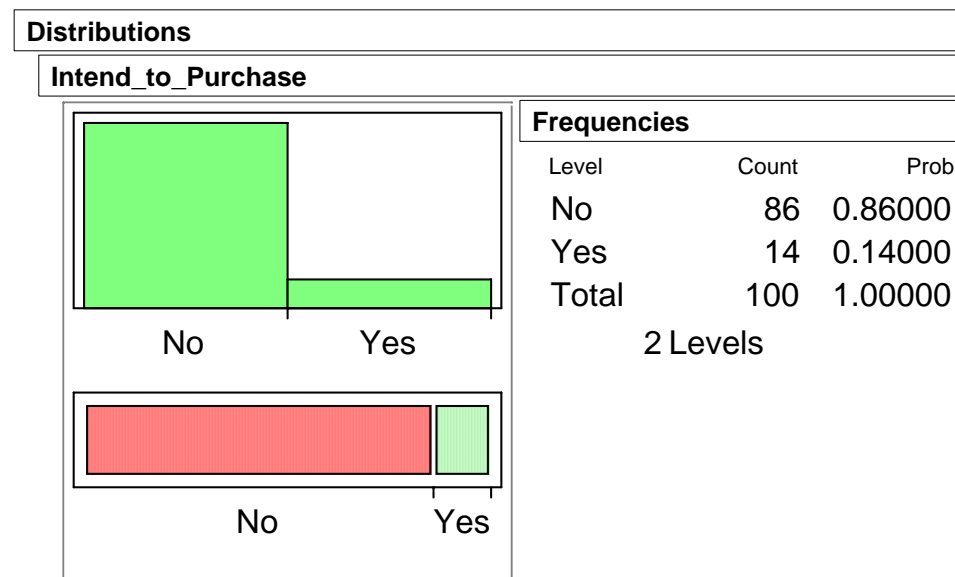
Suppose it was of interest to obtain a 95% CI width less than a particular width, say MaxWidth. If a preliminary estimate for s , call it s_{GUESS} , was available, how large should n be chosen to obtain $\text{Width} \leq \text{MaxWidth}$?

An Intent to Purchase Survey

(BBS, p.104)

A manufacturer of consumer electronics would like to know how many households intend to purchase a computer next year. The file CompPur.jmp contains the yes or no responses of 100 households.

The JMP summary⁵ for the categorical yes-no variable Intend_to_Purchase



⁵ Using Analyze > Distribution, the output for a categorical variable takes this form.

Management hopes that the proportion is at least 25% in order to justify sales projections. Does the survey dash their hopes?

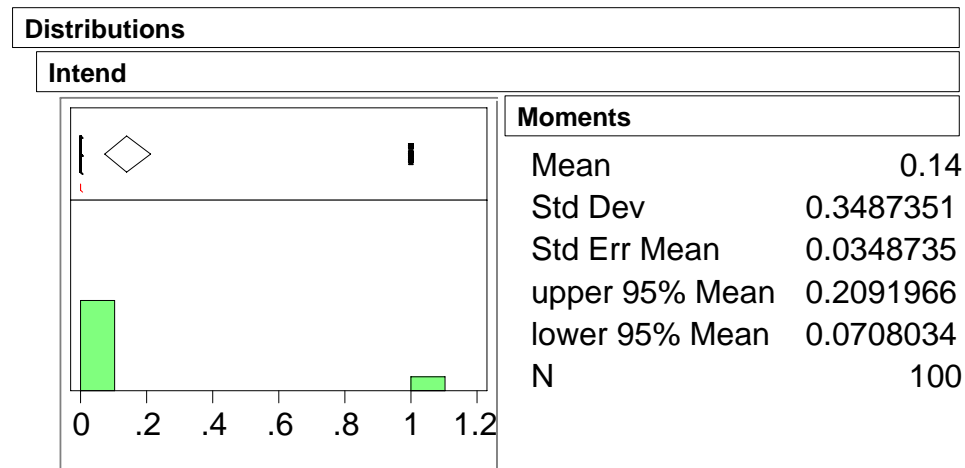
Assuming the survey has truly sampled the population of interest, the issue here is:

Can the difference between 14% and 25% reasonably be attributed to chance (i.e. to sampling variation)?

To answer this question with a confidence interval, we transform the variable Intend_to_Purchase to a numerical variable Intend:

Intend = 0	if	Intend_to_Purchase = No
Intend = 1	if	Intend_to_Purchase = Yes

Because Intend is numerical, the JMP summary provides moments and the 95% CI for the (unknown) population proportion



Note that the sample proportion 14% is the mean \bar{x} of Intend.

Now do management's hopes seem reasonable?

Take-Away Review

A confidence interval is a range of plausible values for the population parameter. The confidence level of the interval indicates just how plausible this range is.

Most of the confidence intervals that we'll use are 95% confidence intervals.

To form a 95% confidence interval for the population mean, we can use the exact JMP interval or the approximate interval

$$\left(\bar{x} - 2s / \sqrt{n}, \bar{x} + 2s / \sqrt{n} \right)$$

This form is typical of most of the CIs we will use in this course and Stat 621, namely

$$(\text{estimated value} \pm 2 \text{ SE}(\text{estimated value}))$$

Next Module

If a value lies outside of the confidence interval, our analysis today suggests that “it’s not plausible” at some level of confidence. But once a value lies out of the interval, that’s about all we can say.

Statistical tests take this idea further and directly measure the reasonableness of a hypothesized population value.

**Department of Statistics
The Wharton School
University of Pennsylvania**

STAT 603

August 2006

Module 9
Statistical Hypothesis Testing

Setting up a Hypothesis Test

Recall the intent to purchase example from Module 8

On the basis of a sample, management needs to decide whether to reject the claim that 25% of the relevant population intends to purchase a computer.

The general problem of statistical hypothesis testing concerns using data to make a decision. In this setting, the decision is whether a hypothesis of interest should be rejected.

Jargon:

The hypothesis of interest is called the *null hypothesis* and is denoted H_0 .

A second hypothesis, denoted H_a , is often considered as an alternative to H_0 .¹

In the intent to purchase example, these two hypotheses can be expressed as

$$H_0: \mu = .25 \quad \text{versus} \quad H_a: \mu \neq .25$$

where μ denotes the unknown true population proportion.

Don't take such hypotheses too literally. If $\mu = .2500001$, it make sense to retain H_0 because remains a rather good description of the population. Instead, think of the null here as saying "It's reasonable to treat the mean of the population as .25."

¹ Some textbooks denote the alternative hypothesis as H_1 rather than H_a . Conventions adopted for the notation of hypothesis testing vary from author to author.

The One-Sample t Test

Suppose we have a random sample x_1, \dots, x_n from a population with unknown mean μ . A common hypothesis that is often considered for this setup is²

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_a: \mu \neq \mu_0$$

Note that the intent to purchase hypotheses is the special case with $\mu_0 = .25$

The key quantity for testing these hypotheses

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

is called a t statistic or a t ratio.

The t statistic measures the distance between \bar{x} and μ_0 in units of standard errors

Intuition: A large t statistic implies that the data are implausible if the null hypothesis is true, and we interpret this as evidence against H_0 .

² Hypotheses of this form are sometimes called “two-sided” hypotheses to distinguish them from hypotheses of the form $H_0: \mu \geq \mu_0$ vs $H_a: \mu < \mu_0$ or $H_0: \mu \leq \mu_0$ vs $H_a: \mu > \mu_0$, which are called “one-sided”. Such refinements of the hypotheses are not often used in practice.

Example: Testing the intent to purchase hypotheses³

$$H_0: \mu = .25 \quad \text{versus} \quad H_a: \mu \neq .25$$

Distributions	
Intend	
Test Mean=value	
Hypothesized Value	0.25
Actual Estimate	0.14
df	99
Std Dev	0.34874
t Test	
Test Statistic	-3.1543
Prob > t	0.0021
Prob > t	0.9989
Prob < t	0.0011

The estimate of the unknown proportion μ is the sample proportion $\bar{x} = .14$

In units of standard errors, how far is .14 from the hypothesized value $\mu_0 = .25$?

³ For the data in CompPur.jmp, apply Analyze > Distribution to the column *Intend*, select Test Mean after clicking on the title bar, and enter .25 for the Hypothesized Mean to obtain this output.

The p-value: How extreme is enough to reject H_0 ?

Key Issue: How large should t be in magnitude (positive or negative) in order to convince us to reject H_0 and declare a “statistically significant” result?

To answer this question in the previous example, JMP provides the quantity:

$$\text{Prob} > |t| = .0021$$

which is the probability of observing a t-statistic more extreme than -3.15 if in fact $H_0: \mu = .25$ were true.⁴

Thus, if μ were .25, a $|t|$ value larger than 3.15 would occur only 0.21% of the time!

The quantity $\text{Prob} > |t| = .0021$ is called a *p-value*, and it measures the rarity of the data when H_0 is true.

⁴ For testing the hypotheses $H_0: \mu \leq .25$ versus $H_a: \mu > .25$, JMP reports the p-value of $\bar{x} = .14$ as $\text{Prob} > t = .9989$. For testing the hypotheses $H_0: \mu \geq .25$ versus $H_a: \mu < .25$, JMP reports the p-value of $\bar{x} = .14$ as $\text{Prob} < t = .0011$.

“Small” p-values indicate one of two things: Either H_0 is false or else something very unusual has happened.

Faced with these two choices, statistical practice is to reject H_0 when the p-value is “small” enough.

The “Official” Rules for Testing H_0 versus H_a

Procedure to test a pair of hypotheses is:

- 1) Pick a value α called the *significance level* (traditionally $\alpha = .05$ or $.01$)
- 2) If the p-value $\leq \alpha$, reject H_0 and report the result as ***statistically significant*** (at the α level of significance)
- 3) Otherwise, if the p-value $> \alpha$, the result is said to be *not statistically significant*

Should the intent to purchase hypothesis $H_0: \mu = .25$ be rejected at the .05 level of significance? At the .01 level of significance?

Example: GM92 data

Suppose you want to test the following pair of hypotheses about returns on GM stock using this two years of data

$$H_0: \mu = 0 \quad \text{versus} \quad H_a: \mu \neq 0$$

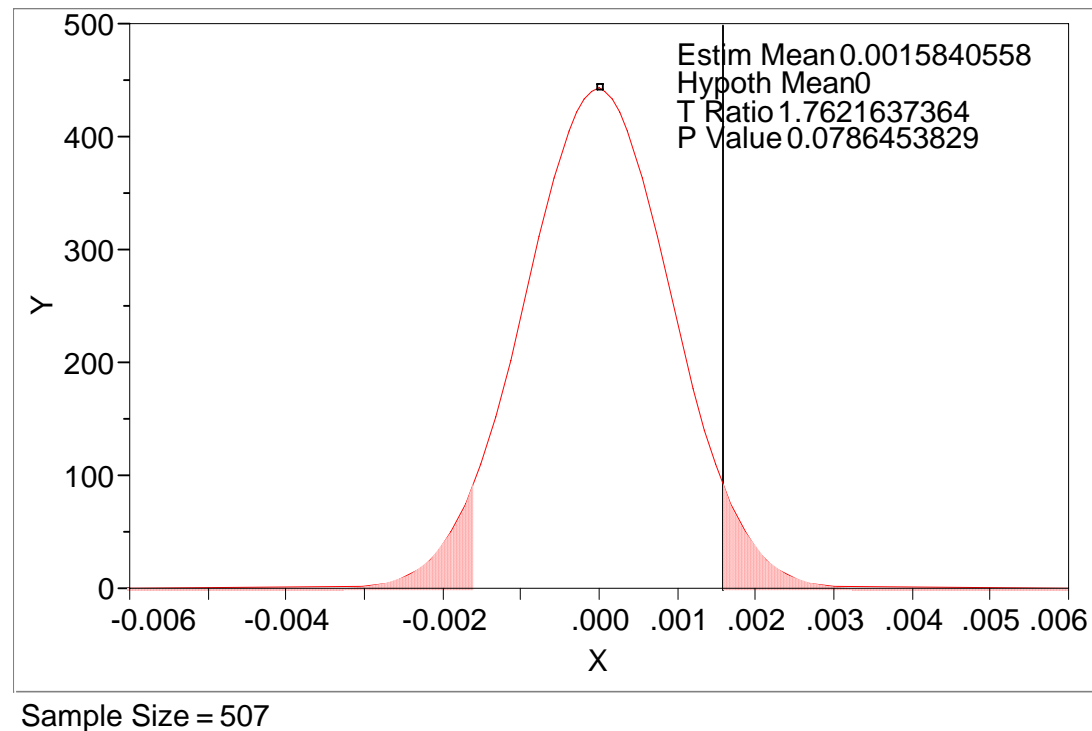
where μ is the unknown mean of the population of GM daily returns. JMP yields⁵

Distributions	
RelChange	
Test Mean=value	
Hypothesized Value	0
Actual Estimate	0.00158
df	506
Std Dev	0.02024
t Test	
Test Statistic	1.7622
Prob > t	0.0786
Prob > t	0.0393
Prob < t	0.9607

Can H_0 be rejected at the .05 level of significance?

⁵ Again, use the Test Mean command with Analyze > Distribution.

The following graph shows the probability represented by p-value here. Notice the role of the assumptions in this graph.⁶



⁶ JMP optionally produces this plot when you test a hypothesis about the mean of a population. The view shown here is just one part of an animation; see the animation options offered by the popup menu that appears at the bottom of the test output.

The "2 standard error" Rule of Thumb

For practical purposes, a good *approximate* test of

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_a: \mu \neq \mu_0$$

is to

reject $H_0: \mu = \mu_0$ at the $\alpha=.05$ level when $|t \text{ statistic}| \geq 2$

i.e. reject H_0 when \bar{x} is more than 2 standard errors away from the hypothesized value μ_0 .

Comparison to Confidence Intervals

The traditional choice of the significance level of a test is $\alpha = 5\%$. Also, the traditional choice of the level of confidence of a confidence interval is 95% (= 100% - 5%).

This correspondence⁷ is not accidental!

An alternative way to test $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ at the $\alpha = .05$ level of significance uses a confidence interval:

Reject $H_0: \mu = \mu_0$ when μ_0 lies outside the 95% CI for μ

For example, for the intent to purchase example, the 95% CI for the population proportion is (.07, .21).

Because the conjectured value 0.25 does not lie inside the confidence interval, we can reject $H_0: \mu = .25$ at the $\alpha = .05$ level since $\mu_0 = .25$ is not contained in (.07, .21).

This seems reasonable since a confidence interval is the set of plausible values for μ given the data.

⁷ The correspondence described here works for “two-sided” hypotheses. See BBS, 102-103.

Statistical Significance versus Practical Significance

t statistics measure discrepancies from H_0 in units of standard error.

As we saw from the astonishing fact, the standard error that sits in the denominator of a t statistic gets smaller and smaller as n gets larger.

Thus, any departure from the null hypothesis H_0 , no matter how small, will eventually be significant when the sample size grows large enough.

BE CAREFUL!

Sometimes, such departures are trivially small and of little practical significance.

You should not conclude that a result is important just because it is statistically significant!

Testing Other Statistical Hypotheses

Many other hypotheses are of interest in statistical analysis.

For example, useful null hypotheses make claims about other features of populations. Hypothesis testing is not restricted to statements about mean values alone.

H_0 : Population is normal

H_0 : Population correlation is 0

H_0 : Two populations are identical

H_0 : Time series is iid

In all of these settings, conventional statistical practice proceeds as follows:

- 1) Calculate a p-value for the disparity between the null hypothesis H_0 and the data
- 2) If $p\text{-value} \leq \alpha$, reject H_0 and report the result as statistically significant (at the α level of significance)
- 3) If $p\text{-value} > \alpha$, the result is not statistically significant

We will describe some of these tests in more detail in Stat 621. See you in September!