The History of Science has suffered greatly from the use by teachers of second-hand material, and the consequent obliteration of the circumstances and the intellectual atmosphere in which the great discoveries of the past were made. A first-hand study is always instructive, and often full of surprises.

Ronald A. Fisher

Our world, our life, our destiny, are dominated by uncertainty; this is perhaps the only statement we may assert without uncertainty.

Bruno de Finetti

If this [probability] calculus be condemned, then the whole of the sciences must also be condemned.

Hans Hahn

Those who ignore Statistics are condemned to reinvent it.

Bradley Efron

All models are wrong, but some are useful.

George E. P. Box

TEXTBOOK
Lind, Marchal and Wathen's "Statistical Techniques in Business & Economics (12th, 13th or 14th editions)" plays a supporting role in this class, particularly for students who find handouts either too superficial or need additional examples and explanations to any given subject. The book contains several examples and solved problems.

STATISTICAL PACKAGES
Most of the computations in the classroom examples are simple enough to be performed by a scientific calculator and/or Excel. Several of the computation and plots that appear in the lecture notes were obtained from MINITAB, R, Excel or MegaStat for Excel. MegaStat for Excel is a set of routines that can be easily "added-in" by Microsoft Excel. It comes with Lind, Marchal and Wathen's textbook. However, Excel by itself will be enough for most of our computations.

HOMEWORK ASSIGNMENTS
Four to six homework sets will be assigned, each of which is invariably due one week after it has been handed out.

GRADE POINT AVERAGE, FINAL NUMBER GRADE and LETTER GRADE
The University of Chicago Graduate School of Business mandates a maximum (not minimum!) class grade point average (GPA) of 3.33. The overall class average will be used to set the class grade point average (GPA). It is important that the highest class GPA be less than (or equal to) 3.33.

The final number grade (FNG) will be the weighted average of i) homework assignments average (HWA), ii) the midterm exam (MT) and iii) the final exam (FI). The weights are 20%, 30% and 50%, respectively. For example, suppose that your grades on HW1, HW2, HW3 and HW4 are 90, 80, 70 and 60, respectively. Then your homework average grade (HWA) is the average of HWA = \( \frac{90 + 80 + 70 + 60}{4} = \frac{280}{4} = 70 \). The letter grades I use are A, A-, B+, B, B-, C, D (lowest grading pass) and F (fail).

CALCULATOR, CHEAT SHEET AND REQUESTS FOR RE-GRADING
Bring your own calculator to all exams. For the midterm exam, a two-page (one sheet) "cheat sheet" is allowed. For the final exam, a four-page (two sheets) "cheat sheet" is allowed. All requests for re-grading of exams must be made in writing and must clearly state the basis of the request.
Main topics

Exploratory data analysis

Probability

Statistical inference and hypothesis testing

Simple and multiple linear regression

Univariate Exploratory Data Analysis

1. Graphical summaries of the data
   1.1 Dot plot
   1.2 Histogram
   1.3 Time series plot
2. Numerical descriptive measures
   2.1 Measures of central tendency
      2.1.1 The sample mean
      2.1.2 The median
   2.2 Measures of dispersion
      2.2.1 The sample variance
      2.2.2 The sample standard deviation
   2.3 Measure of asymmetric: skewness
   2.4 Measure of extremely: kurtosis
3. Boxplot
Summary of the lecture

- In this class you will learn how to graph small sets of quantitative observations: dotplot.
- Large sets of quantitative observations: histogram.
- Observations that are collected as time evolves: time-series plot.
- You also will learn how to construct a boxplot, which can be prove useful when comparing observations from several samples.
- Even though graphs are extremely useful and relatively simple to draw, in many situations numerical summaries are required, for instance as input into other systems.
- We will also talk about measures of central tendency (mean and median), measures of dispersion (variance, standard deviation), measures of asymmetry (skewness), measures of extremity (kurtosis).
- We will also discuss the empirical rule that says that roughly 68% of the observations in any sample should fall within one sample standard deviation around the sample mean and 95% should fall within two sample standard deviations around the sample mean.

Book material

- Chapter 1
  Types of statistics (pages 6-7 (12 & 13)* ) and types of variables (pages 8-9 (12 & 13)).
- Chapter 2
  Frequency distributions and Histogram (pages 25-33 (12), 22-37 (13)).
- Chapter 3
  Sample mean (page 58 (12 & 13)) and sample median (page 62 (12& 13))
  Measures of dispersion (pages 71-77 (12), 71-80 (13))
  Empirical rule (page 80 (12), 82 (13)).
- Chapter 4
  Dotplots (pages 97-98 (12), 99-100 (13))
  Boxplots (pages 108-111 (12), 110-113 (13)).
  Skewness (pages 114-117 (12), 113-117 (13)).

*Numbers in parentheses refer to the book edition

1. Graphical Summaries of the Data

Two key ideas

Exploratory (descriptive) issues:
Look at the data (sample).
Understand its structure without generalizing.

Inference issues:
Use data (sample) to generalize results to a larger population of interest.
Example

**Problem:** How many of 100,000 voters (population) prefer A over B? We can’t ask them all!

**Solution:** Ask a sample of 500 voters.

Summarize, describe the data: 300 voters for A (A = 1), 200 for B (B = 0). We will learn how to generalize to the population. For now, we just learn how to analyze (describe) the data.

Let us look at some data. Data are the statistician’s raw material, the numbers that we use to interpret reality.

All statistical problems involve either the collection, description and analysis of data, or thinking about the collection, description and analysis of data.

There are many aspects of data. Data may be:

- **univariate** (one variable per case) or
- **multivariate** (more than one variable per case).

There are also different types of data:

- **discrete** (transactions in a given day) and
- **continuous** (SP500)

The Canadian Return Data

Here is a specific data set (or sample). We have 107 monthly returns on a broad based portfolio of Canadian assets (more on portfolios later).

<table>
<thead>
<tr>
<th>canada</th>
<th>0.07</th>
<th>0.05</th>
<th>0.02</th>
<th>-0.04</th>
<th>0.08</th>
<th>-0.02</th>
<th>-0.05</th>
<th>0.02</th>
<th>0.03</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00</td>
<td>0.03</td>
<td>0.08</td>
<td>-0.03</td>
<td>0.04</td>
<td>-0.03</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.04</td>
<td>0.06</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.09</td>
<td>0.11</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>-0.06</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>-0.04</td>
<td>0.04</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>-0.09</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>-0.04</td>
<td>0.05</td>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>-0.06</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>-0.08</td>
<td>-0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>-0.04</td>
<td>0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
</tr>
</tbody>
</table>

**Interpret:** Each number corresponds to a month. They are given in time order (go across columns first). Our first observation is .07. In the first month, the return was .07, in the 11th .03.
We are interested in ways to summarize or “see” the data. The previous table was very unclear. To display the returns we can use a simple graphical tool: the dot plot. For each number simply place a dot above the corresponding point on the number line.

**1.1 The dot plot**

Interpret:
The returns are centered or located at about .01.
The spread or variation in the returns is huge.

Notice that the data has a nice mound or bell shape. There is a central peak and right and left “tails” that die off roughly symmetrically.

Some data does not have the mound shape. Daily volume of trades in the cattle pit.

It is skewed to the right or positively skewed.
We also have data on countries other than Canada. Let us compare Canada with Japan.

It really helps to get things on the same scale. How is Japan different from Canada?

**Mutual fund data**

Let us use the dot plot to compare returns on some other kinds of assets.

We will look at returns on different **mutual funds** such as the equally weighted market and T-bills.

The equally weighted market represents returns on a portfolio where you spread your money out equally over a wide variety of stocks.
Data on 4 different kinds of returns:
- Dreyfus growth fund
- Putman income fund (Note that each dot is now 2 points)
- Equally weighted market
- T-bills (each dot is 7 points here. This is the risk free asset)

The beer data
ribeer: the number of beers male MBA students claim they can drink without getting drunk
ribeerf: same for females

Character Dotplot: We call a point like this an outlier

The histogram
Sometimes the dot plot can look rather jumpy. The histogram gives us a smoother picture of the data. The height of each bar tells us how many observations are in the corresponding interval.

3 women have a number of beers between 1.5 and 4.5.
3 women have a number of beers in the interval (1.5, 4.5).
Here is the histogram of the Canadian returns.

The number of bars you use affects how “smooth” the picture looks.

1.3 The time series plot

We just looked at two kinds of data:
1) the return data
2) the number of beers

For the return data, each number corresponds to a month.
For the beer data, each number corresponds to a person.

The return data has an important feature that the beer data does not have.

*It has an order!*

There is a first one, a second one, and ....

A sequence of observations taken over time is often called a **time series**.

We could have daily data (temperature), annual data (inflation), quarterly data (inflation, GDP) and so on.

For time series data, the **time series plot** is an important way to look at the data.
Time series plot of the Canadian returns:

On the vertical axis we have returns.
On the horizontal axis we have "time".

Do you see a pattern?

Time series plot of Daily volume of trades in the cattle pit:

On the vertical axis we have volumes.
On the horizontal axis we have days.

Do you see a pattern?

Monthly US beer production.

Now, do you see a pattern?
Australia: monthly production of beer. megalitres. April 1956 - Aug 1995

Two components: a seasonal (annual) cycle plus an increasing trend from 100 to 175, then a constant trend for the second half of the time series.

2. Numerical Descriptive Measures

We have looked at graphs.

Suppose we are now interested in having numerical summaries of the data rather than graphical representations.

We have seen that two important features of any data set are:

1) how spread out the data is, and

2) the central or typical value of the data set

In this part of the notes we will describe methods to summarize a data set numerically.

First, we will introduce measures of central tendency to determine the “center” of a distribution of data values, or possibly the “most typical” data value.

Measures of central tendency include: the mean and the median.

Second, we will discuss measures of dispersion, such as the sample standard deviation and the sample variance.
2.1 Measures of Central Tendency

2.1.1 The sample mean

Suppose we collect $n$ pieces of data. We need some way of describing the data. We write

$$X_1, X_2, X_3, \ldots, X_n$$

the first number

the last number, $n$ is the number of numbers, or the "number of observations." You may also hear it referred to as the "sample size."

They are the values that we observe.

Here, $x$ is just a name for the set of numbers, we could just as easily use $y$ (or Buddy).

$$\bar{x}$$

$$X_1 \rightarrow 5 \quad n=5$$

$$X_3 \rightarrow 8$$

$$2$$

Sometimes the order of the observations means something. In our return data the first observation corresponds to the first time period. Sometimes it does not. In our beer data we just have a list of numbers, each of which corresponds to a student.

The sample mean is just the average of the numbers "$x$":

$$\bar{x} = \frac{\text{sum}}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

We often use the $\bar{x}$ symbol to denote the mean of the numbers $x$.

We call it "$x$ bar".
Here is a more compact way to write the same thing...

Consider

\[ x_1 + x_2 + \cdots + x_n \]

We use a shorthand for it (it is just notation):

\[ \sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \]

This is summation notation

Using summation notation we have:

**The sample mean:**

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

**Graphical interpretation of the sample mean**

Let us go back to our standard dot plots

In some sense, the men claim to drink more. To summarize this we can compute the average value for both men and women. (I deleted the outlier, I do not believe him!).
On average women claim they can drink 4.2 beers. Men claim they can drink 7.8 beers.

In the picture, I think of the mean as the “center” of the data.

Let us compare the means of the Canadian and Japanese returns.

Mean of Canada = 0.0090654
Mean of Japan = 0.0023364

This is a big difference.

It was hard to see this difference in the dot plots (page 14) because the difference is small compared to the variation.

More on summation notation (take this as an aside)

Let us look at summation in more detail.

\[ \sum_{i=1}^{n} x_i \]

means that for each value of i, from 1 to n, we add to the sum the value indicated, in this case \( x_i \).

add in this value for each i
To understand how it works let us consider some examples.

Think of each row as an observation on both $x$ and $y$.

<table>
<thead>
<tr>
<th>year</th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.07</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>0.03</td>
</tr>
</tbody>
</table>

To make things concrete, think of each row as corresponding to a year and let $x$ and $y$ be annual returns on two different assets.

In year 1 asset “$x$” had return 7%.
In year 4 asset “$y$” had return 3%.

\[
\sum_{i=1}^{n} x_i = \sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 \\
\text{compute } x \text{ bar.} \\
= 0.07 + 0.06 + 0.04 + 0.03 \\
= 0.2 \\
\bar{x} = \frac{0.2}{4} = 0.05 \\
\text{compute } y \text{ bar.}
\]

For each value of $i$, we can add in anything we want:

\[
\sum_{i=1}^{n} (x_i - \bar{x}) =
\]

\[
\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) =
\]
2.1.2 The median

After ordering the data, the median is the middle value of the data.

If there is an even number of data points, the median is the average of the two middle values.

Example

1,2,3,4,5 Median = 3
1,1,2,3,4,5 Median = (2+3)/2 = 2.5

Mean versus median

Although both the mean and the median are good measures of the center of a distribution of measurements, the median is less sensitive to extreme values.

The median is not affected by extreme values since the numerical values of the measurements are not used in its computation.

Example

1,2,3,4,5 Mean: 3 Median: 3
1,2,3,4,100 Mean: 22 Median: 3

2.2 Measures of Dispersion

The mean and the median give us information about the central tendency of a set of observations, but they shed no light on the dispersion, or spread of the data.

Example: Which data set is more variable?

5,5,5,5,5 Mean: 5
1,3,5,8,8 Mean: 5

Do you only care about the average return on a mutual fund or you need a measure of risk, too?

Here is one …
2.2.1 The Sample Variance

The y numbers are more spread out than the x numbers. We want a numerical measure of variation or spread. The basic idea is to view variability in terms of distance between each measurement and the mean.

\[ x_i - \bar{x} \]

We cannot just look at the distance between each measurement and the mean. We need an overall measure of how big the differences are (i.e., just one number like in the case of the mean).

Also, we cannot just sum the individual distances because the negative distances cancel out with the positive ones giving zero always (Why?).

We average the squared distances and define

\[ \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]
So, the **sample variance** of the x data is defined to be:

\[
S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

We use \( n-1 \) instead of \( n \) for technical reasons that will be discussed later.

Think of it as the average squared distance of the observations from the mean.

**Questions**

1) What is the smallest value a variance can be?

2) What are the units of the variance?

It is helpful to have a measure of spread which is in the original units. The sample variance is **not** in the original units. We now introduce a measure of dispersion that solves this problem: **the sample standard deviation**

**2.2.2 The sample standard deviation**

It is defined as the square root of the sample variance (easy).

\[
S_x = \sqrt{S_x^2}
\]

The units of the standard deviation are the same as those of the original data.
Example 1 (numerical)

Assume as before:

\[
Y - \bar{Y} = 0.04, -0.02, 0.02, -0.04
\]
\[
X - \bar{X} = 0.02, 0.01, 0.01, 0.02
\]

\[
S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

\[
= \frac{1}{3} (0.04^2 + (-0.02)^2 + 0.02^2 + (-0.04)^2)
\]

\[
= \frac{1}{3} (0.016 + 0.0004 + 0.0004 + 0.0016)
\]

\[
= \frac{0.004}{3} = 0.00133
\]

\[
S_y = \sqrt{0.00133} = 0.0365
\]

The sample standard deviation for the y data is bigger than that for the x data. This numerically captures the fact that y has "more variation" about its mean than x.

Example 2 (graphical)

The standard deviations measure the fact that there is more spread in the Japanese returns.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>canada</td>
<td>107</td>
<td>0.00907</td>
<td>0.03833</td>
</tr>
<tr>
<td>japan</td>
<td>107</td>
<td>0.00214</td>
<td>0.07368</td>
</tr>
</tbody>
</table>
2.3 Measure of asymmetry: Skewness

Measures asymmetry of a distribution.

Symmetric data has zero skewness.

Negatively skewness (the left tail is longer – mean < median)
Occurs when the values to the left of (less than) the mean are fewer but farther from the mean than are values to the right of the mean.

Positively skewness (the right tail is longer – mean > median)
Example: investment returns -5%, -10%, -15%, 30%
People like bets with positive skewness.
Willing to accept low, or even negative, expected returns when an asset exhibits positive skewness.

2.4 Measure of extremity: Kurtosis

Measures the degree to which exceptional values occur more frequently (high kurtosis) or less frequently (low kurtosis)

A reference distribution is the normal distribution, whose kurtosis is three.

High kurtosis results in exceptional values that are called “fat tails.”
Fat tails indicate a higher percentage of very low and very high returns than would be expected with a normal distribution.

Low kurtosis results in “thin tails” and a wide middle with more values close to the average than there would be in a normal distribution, and tails are thinner than there would be in a normal distribution.
Volume data

Kurtosis: historical facts

- Kurtosis was used by Karl Pearson in 1905 in "Das Fehlergesetz und seine Verallgemeinerungen durch Ficher und Pearson: A Rejoinder," Biometrika, 4, 109-212, in the phrase "the degree of kurtosis." He states therein that he has used the term previously (OED). According to the OED and to Schwartzman, the term is based on the Greek meaning a bulging, convexity.
- He introduced the terms leptokurtic, platykurtic, and mesokurtic, writing in Biometrika (1905), 5, 173: "Given two frequency distributions which have the same variability as measured by the standard deviation, they may be relatively more or less flat-topped than the normal curve. If more flat-topped I term them platykurtic, if less flat-topped leptokurtic, and if equally flat-topped mesokurtic" (OED2).
- In his "Errors of Routine Analysis," Biometrika, 19, (1927), p. 160 Student provided a mnemonic:

**Compressing skewness and excess kurtosis**

Excess kurtosis is kurtosis minus 3. Excel computes excess kurtosis.

\[
\text{skewness} = \frac{n}{(n-1)(n-2)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^3
\]

\[
\text{excess kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}
\]
Volume data: kurtosis and outliers

Same kurtosis, different skewness

Same skewness, different kurtosis

10 largest obs.

11.25794
11.26239
11.43341
11.48154
11.52330
11.94644
12.10322
12.33747
12.75935
15.32864
10.13302
10.98134
11.38262
11.73549
11.77891
12.84776
14.80519
15.38212
21.74778
35.23782
**Kurtosis and standard deviation**

Left histogram: higher variability.
Left histogram: lower kurtosis or thinner tails.
Bottom curves: left tail behavior of both histograms.

**Same mean, variance, skewness**

**Different kurtosis**

In both cases, mean=0, variance=3 and skewness=0.

Excess kurtosis is 0.054 for the thin-tail distribution (black).
Excess kurtosis is 65.18 for the fat-tail distribution (red).

<table>
<thead>
<tr>
<th>Percentage of observations below cutoff</th>
<th>Red</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10</td>
<td>0.1064</td>
<td>0.0000</td>
</tr>
<tr>
<td>-9</td>
<td>0.1448</td>
<td>0.0000</td>
</tr>
<tr>
<td>-8</td>
<td>0.2038</td>
<td>0.0005</td>
</tr>
<tr>
<td>-7</td>
<td>0.2993</td>
<td>0.0065</td>
</tr>
<tr>
<td>-6</td>
<td>0.4636</td>
<td>0.0571</td>
</tr>
<tr>
<td>-5</td>
<td>0.7686</td>
<td>0.3571</td>
</tr>
<tr>
<td>-4</td>
<td>1.4004</td>
<td>1.6004</td>
</tr>
<tr>
<td>-3</td>
<td>2.8334</td>
<td>5.1393</td>
</tr>
<tr>
<td>-2</td>
<td>6.9663</td>
<td>11.8255</td>
</tr>
<tr>
<td>-1</td>
<td>19.5501</td>
<td>19.4970</td>
</tr>
</tbody>
</table>

**2.5 Quantiles**

**Quartiles**: divide the data into 4 equal parts.
Q1 = Median of the first half of the data
Q2 = Median
Q3 = Median of the second half of the data

IQ = Interquartile range
IQ = Q3 - Q1

**Deciles**: divide the data into 10 equal parts.
**Percentiles**: divide the data into 100 equal parts.
2.6 The Empirical Rule

We now have two numerical summaries for the data

\[ \bar{x}, s_x \]

where the data is how variable the data is

The mean is pretty easy to interpret (some sort of "center" of the data).

We know that the bigger \( s_x \) is, the more variable the data is, but how do we really interpret this number?

What is a big \( s_x \), what is a small one?

The empirical rule will help us understand \( s_x \) and relate the summaries back to the dot plot (or the histogram).

**Empirical Rule**

For "mound shaped data":

Approximately 68% of the data is in the interval

\[ (\bar{x} - S_x, \bar{x} + S_x) = \bar{x} \pm S_x \]

Approximately 95% of the data is in the interval

\[ (\bar{x} - 2S_x, \bar{x} + 2S_x) = \bar{x} \pm 2S_x \]

Let us see this with the Canadian returns

\[ \bar{x} = 0.00907 \]

\[ s_x = 0.03833 \]

The empirical rule says that roughly 95% of the observations are between the dashed lines and roughly 68% between the dotted lines.

Looks reasonable.
Same thing viewed from the perspective of the time series plot.

5% outside would be about 5 points.

There are 4 points outside, which is pretty close.

3. BOXPLOT

1-2-2-3-3-4-4-4-4-4-4-4-5-5-5.5-7

1.6 is the smallest observation greater than $Q_1 - 1.5*IQ$
3.5 is the largest observation lower than $Q_3 + 1.5*IQ$

Step by step illustration

Data: 65 69 70 63 63 72 63 60 69 66 71 73 70 65 74 69 69 87
Sort: 60 63 63 63 65 66 69 69 69 70 70 71 72 73 74 87

$Q_1 = \Box$
$Q_2 = \Box$
$Q_3 = \Box$

$IQ = \Box$
$1.5*IQ = \Box$
$Q_1 - 1.5*IQ = \Box$
$Q_3 + 1.5*IQ = \Box$
Solution

Sort: 60 63 63 65 66 69 69 69 70 70 71 72 73 74 87

Q1 = 65
Q2 = 69
Q3 = 71

IQ = Q3 - Q1 = 71 - 65 = 6
1.5*IQ = 9
Q1 - 1.5*IQ = 65 - 9 = 56
Q3 + 1.5*IQ = 71 + 9 = 80

Example: European returns

Example: Annual salary (in thousands of dollars)
Example: SP500 components

1st row: ordered by skewness
2nd row: ordered by kurtosis

S&P500: kurtosis and skewness
Skewness and logarithm of excess kurtosis for the S&P500 components.

S&P500: Components with fattest and thinnest tails
The bottom graphs are excess kurtosis computed over time.
Example: Number of siblings - MBA students
Data collected from Business Stats students on January 10th 2009 (41000-85):
0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 4 5 5

Xbar = 1.73
Median = 1.50
Var = 1.087
St.dev.=1.042
Q1 = 1.00
Q3 = 2.00
Skewness = 1.616
Excess kurtosis = 3.093

Example: Number of siblings – Boston College

Source: Statistical Methods for Health Care Research (5th edition) by Barbara H. Murray. Publisher: Lippincott, Williams & Wilkins.
Xbar = 2.022013
St.Dev. = 1.640233
Skewness = 2.165848
Excess kurtosis = 8.029811
Q1 = 1
Q2 = 2
Q3 = 3

Example: Average daily temperature in Rio de Janeiro, Durham and Chicago

Seasonality is more pronounced in Durham and Chicago.
Variability is also higher in Durham and Chicago.
Longer winters in Chicago (really???)
The time series were smoothed by replacing each observation by the average of 21 neighboring days, 10 to the left and 10 to the right of the observation. Smoothing the time series helps to highlight the short-term patterns.

The time series were smoothed by replacing each observation by the average of 364 neighboring days, 182 to the left and 182 to the right of the observation. Smoothing the time series helps to highlight the long-term patterns.

Monthly behavior

Rio: variability seems to be constant throughout the year.

Durham, Chicago: variability seems to be higher during colder months than during warmer months.

Dot: medians vertical bar: Q1 to Q3
January behavior

Rio is the warmest place in January (it is summer there!)
Even Durham is much warmer than Chicago (what am I doing here?)
Temperature in Chicago is the most variable.

Example: Highest temperatures in the USA

<table>
<thead>
<tr>
<th>State</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAWAII</td>
<td>37.8</td>
</tr>
<tr>
<td>ALABAMA</td>
<td>64.6</td>
</tr>
<tr>
<td>ALASKA</td>
<td>37.8</td>
</tr>
<tr>
<td>ALABAMA</td>
<td>44.4</td>
</tr>
<tr>
<td>ALASKA</td>
<td>47.8</td>
</tr>
<tr>
<td>ALASKA</td>
<td>37.8</td>
</tr>
<tr>
<td>ALABAMA</td>
<td>44.4</td>
</tr>
<tr>
<td>ALABAMA</td>
<td>47.8</td>
</tr>
<tr>
<td>RHODE-ISLAND</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
<tr>
<td>TENNESSEE</td>
<td>40.6</td>
</tr>
</tbody>
</table>
| TENNESSE
Example: US 2004 unemployment rates

Mean (\(x_{bar}\)) = 5.2078431
Variance = 1.1691373
Standard deviation (s) = 1.0812665
Q1 = 4.6 (Georgia)
Q2 = 5.2 (Rhode Island)
Q3 = 5.7 (New Mexico)
Skewness = 0.4798145
Kurtosis = 0.3317919

Empirical rule

- \([x_{bar} - 1s; x_{bar} + 1s]\) = \([4.13; 6.289110]\) 72.55%
- \([x_{bar} - 2s; x_{bar} + 2s]\) = \([3.05; 7.370376]\) 94.12%
- \([x_{bar} - 3s; x_{bar} + 3s]\) = \([1.96; 8.451643]\) 100.00%
### Multivariate Exploratory Data Analysis

1. How to relate two things
2. Correlations and covariances
3. Linearly related variables
   3.1 Mean and variance of a linear function
   3.2 Linear combinations
   3.3 Mean and variance of a linear combination: 2 inputs
   3.4 Mean and variance of a linear combination: 3 inputs
   3.5 Mean and variance of a linear combination: k inputs
4. Portfolio example
5. Simple linear regression

### Summary of the lecture

In this class you will learn how to
- Relate two sets of variables: sample linear correlation coefficient
- Compute sample mean, variance and standard deviation of linear combinations of variables
- Study the practical example of portfolio allocation

**Book**

Skewness (pages 114-117 (12)*, 113-117 (13))
What is correlation analysis? (pages 429-435 (12), 458-465 (13))

*Number in parenthesis refers to the book edition

### Example: Comparing international stock returns

July 10, 1987 until December 31, 1997 (2733 days) - Amsterdam (EOE), Frankfurt (DAX), Paris (CAC40), London (FTSE100), Hong Kong (Hang Seng) Tokyo (Nikkei), Singapore (Singapore All Shares), New York (S&P500).
It is considered good to have a large mean return and a small standard deviation.
1. How to Relate Two Things

The mean and standard deviation help us summarize a bunch of numbers which are measurements of just one thing (one variable).

A fundamental and totally different question is how one thing relates to another.

In this section of the notes we look at scatter plots and how covariance and correlation can be used to summarize them.

When examining two things (variables) at the time, the scatter plot will be our main graphical tool whereas covariance and correlation will be our main numerical summaries.

### Example

Is the number of beers you can drink related to your weight?

<table>
<thead>
<tr>
<th>number</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.0</td>
<td>192</td>
</tr>
<tr>
<td>12.0</td>
<td>160</td>
</tr>
<tr>
<td>5.0</td>
<td>155</td>
</tr>
<tr>
<td>5.0</td>
<td>120</td>
</tr>
<tr>
<td>7.0</td>
<td>150</td>
</tr>
<tr>
<td>13.0</td>
<td>175</td>
</tr>
<tr>
<td>4.0</td>
<td>100</td>
</tr>
<tr>
<td>12.0</td>
<td>165</td>
</tr>
<tr>
<td>12.0</td>
<td>150</td>
</tr>
</tbody>
</table>

Now we think of each pair of numbers as an observation. Each person has two numbers associated with him/her, # beers and weight. Each pair corresponds to a point on the plot.

### Example

International stock returns: Amsterdam and Frankfurt.

Each point corresponds to a day.
In general we have observations 
\[(x_i, y_i)\]  

and each point on the plot corresponds to an observation.

Our data looks like:

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.0</td>
<td>192</td>
<td>1</td>
</tr>
<tr>
<td>12.0</td>
<td>160</td>
<td>2</td>
</tr>
<tr>
<td>5.0</td>
<td>155</td>
<td>3</td>
</tr>
<tr>
<td>5.0</td>
<td>120</td>
<td>4</td>
</tr>
<tr>
<td>7.0</td>
<td>150</td>
<td>5</td>
</tr>
<tr>
<td>13.0</td>
<td>175</td>
<td>6</td>
</tr>
<tr>
<td>4.0</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>12.0</td>
<td>165</td>
<td>8</td>
</tr>
</tbody>
</table>

The plot enables us to see the relationship between \(x\) and \(y\).

Even more, the relationship looks linear in that it looks like we could draw a line through the plot to capture the pattern.

Covariance and correlation summarize how strong a linear relationship there is between two variables.

In our first example weight and # beers were two variables. In our second example our two variables were two kinds of returns.

In general, we think of the two variables as \(x\) and \(y\).

2. Covariance and Correlation

In both examples it does look like there is a relationship.

Even more, the relationship looks linear in that it looks like we could draw a line through the plot to capture the pattern.

Covariance and correlation summarize how strong a linear relationship there is between two variables.

In our first example weight and # beers were two variables. In our second example our two variables were two kinds of returns.

In general, we think of the two variables as \(x\) and \(y\).

Historical note

1885: Sir Francis Galton: studying the heights of children versus the heights of parents.

There’s a regression-back-to-the-mean effect: If your parents are on average higher than the average, you’ll regress back to the average.

1888: Co-relation: slope of the least-squares regression line for data in standardized (by median and quartile range) form

1896: Karl Pearson, product moment definition

The misuse of correlation has multiplied faster than the proper use of it!
The sample covariance between \( x \) and \( y \):

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

The sample correlation between \( x \) and \( y \):

\[
r_{xy} = \frac{s_{xy}}{s_x s_y}
\]

So, the correlation is just the covariance divided by the two standard deviations.

We will get some intuition about these formulae, but first let us see them in action. How do they summarize data for us? Let us start with the correlation.

**Correlation**, the facts of life:

\[-1 \leq r_{xy} \leq 1\]

The closer \( r \) is to 1 the stronger the linear relationship is with a positive slope. When one goes up, the other tends to go up.

The closer \( r \) is to -1 the stronger the linear relationship is with a negative slope. When one goes up, the other tends to go down.

The correlations corresponding to the two scatter plots we looked at are:

- Correlation of amsterdam and frankfurt = 0.677
- Correlation of nbeer and weight = 0.692

The larger correlation between nbeer and weight indicates that the linear relationship is stronger.

Let us look at some more examples.
Correlation of $y_1$ and $x_1 = 0.019$

Correlation of $y_2$ and $x_2 = 0.995$

Correlation of $y_3$ and $x_3 = 0.586$

Correlation of $y_4$ and $x_4 = -0.982$

Correlation of $y_5$ and $x_5 = 0.210$

The correlation only measures linear relationships (here the value is small but there is a strong nonlinear relationship between $y_5$ and $x_5$.)
Example: The country data

Which countries go up and down together?
I have data on 23 countries.
That would be a lot of plots!

Example: International stock returns

To summarize, we can compute all pair wise correlations:

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Frankfurt</th>
<th>Paris</th>
<th>London</th>
<th>HongKong</th>
<th>Tokyo</th>
<th>Singapore</th>
<th>NewYork</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frankfurt</td>
<td>0.676</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>0.365</td>
<td>0.393</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>London</td>
<td>0.487</td>
<td>0.424</td>
<td>0.290</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HongKong</td>
<td>0.406</td>
<td>0.284</td>
<td>0.177</td>
<td>0.619</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tokyo</td>
<td>0.662</td>
<td>0.607</td>
<td>0.325</td>
<td>0.545</td>
<td>0.480</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singapore</td>
<td>0.357</td>
<td>0.371</td>
<td>0.482</td>
<td>0.248</td>
<td>0.174</td>
<td>0.292</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>NewYork</td>
<td>0.284</td>
<td>0.295</td>
<td>0.302</td>
<td>0.150</td>
<td>0.243</td>
<td>0.298</td>
<td>0.299</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Why compute both covariance and correlation?

The four covariances are around 0.5......
3 Linearly Related Variables

We have studied data sets that display some kind of relation with each other (the mutual fund returns and the market returns, for instance).

Sometimes there is an exact linear relation between variables:

$$y = c_0 + c_1 x$$

Can we say something about the sample mean of $y$ if all we know is the sample mean of $x$ (and vice versa)?

Can we say something about the sample standard deviation of $y$ if all we know is the sample standard deviation of $x$ (and vice versa)?

We will answer these questions in the sequel.

Example

Suppose we have daily temperatures (in Celsius degree) in Rio de Janeiro from January 1st, 1995 to December, 11th 2008.

We also know that the sample mean and the sample variance for the daily temperature for this period are 24.24C and 2.78C.

What in the hell are Celsius degree?

Don't panic!!!!

$$F = \frac{2}{5} + 1.8C$$
The variable $y$ is a linear function of the variable $x$ if:

$$y = c_0 + c_1 x$$

$c_0$: the intercept
$c_1$: the slope

We think of the $c$’s as constants (fixed numbers) while $x$ and $y$ vary.

3.1 Mean and variance of a linear function

How are the mean and variance of $y$ related to those of $x$?

Let us look at our temperature example. It is not a coincident that $32 + 1.8 \times 24.235 = 75.622$ and that $1.8 \times 2.782 = 5.008$

Suppose

$$y = c_0 + c_1 x$$

Then,

$$\bar{y} = c_0 + c_1 \bar{x}$$

$$s_y^2 = c_1^2 s_x^2$$

$$s_y = |c_1| s_x$$

Recall that $|x|$ is the absolute value of $x$. For instance, $|-5| = 5$ and $|10| = 10$
When a variable $y$ is linearly related to several others, we call it a **linear combination**.

$$y = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_k x_k$$

$y$ is a linear combination of the $x$'s.  
$c_i$ is the coefficient of $x_i$.

### 3.2. Linear combinations

We may want a variable to be related to several others instead of just one. We will assume that $Y$ is a function of $X, \ldots$ rather than just a function of $X$.

**Example: house pricing**

<table>
<thead>
<tr>
<th>Home Nbhd Offer</th>
<th>SqFt</th>
<th>Brick</th>
<th>Bed</th>
<th>Bath</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1790</td>
<td>No</td>
<td>2</td>
<td>114300</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2030</td>
<td>No</td>
<td>4</td>
<td>114200</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1740</td>
<td>No</td>
<td>3</td>
<td>114800</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1980</td>
<td>No</td>
<td>3</td>
<td>94700</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2130</td>
<td>No</td>
<td>3</td>
<td>119800</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1780</td>
<td>No</td>
<td>3</td>
<td>114600</td>
</tr>
</tbody>
</table>

We will see later, when studying multiple linear regression, that the price can be modeled as a linear combination of the other variables.

The following formula relates the expected sales price of a house (Price) to its size (SqFt), number of bedrooms (Bed) and number of bathrooms (Bath):

$$\text{Price} = -5640.83 + 35.64 \times \text{SqFt} + 10459.93 \times \text{Bed} + 13546.13 \times \text{Bath}$$

### Example: Portfolio allocation

Let us use country returns and suppose that we had put 0.5 into USA and 0.5 into Hong Kong, i.e.

$$\text{port} = 0.5 \times \text{honkong} + 0.5 \times \text{usa}$$

What would our returns have been?

<table>
<thead>
<tr>
<th>honkong</th>
<th>usa</th>
<th>port</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.04</td>
<td>0.030</td>
</tr>
<tr>
<td>0.06</td>
<td>-0.03</td>
<td>0.015</td>
</tr>
<tr>
<td>0.02</td>
<td>0.01</td>
<td>0.015</td>
</tr>
<tr>
<td>-0.03</td>
<td>0.01</td>
<td>-0.010</td>
</tr>
<tr>
<td>0.08</td>
<td>0.05</td>
<td>0.065</td>
</tr>
</tbody>
</table>

For each month, we get the portfolio return as $\frac{1}{2}\times \text{honkong} + \frac{1}{2}\times \text{usa}$.安庆
How do the returns on this portfolio compare with those of Hong Kong and USA?

It looks like the mean for my portfolio is right in between the means of USA and Hong Kong. What about the standard deviation?

Let us try a portfolio with three stocks.

Let us go short on Canada (i.e., we borrow Canada to invest in the other stocks), i.e.

\[ \text{port} = -0.5 \cdot \text{canada} + 1.0 \cdot \text{usa} + 0.5 \cdot \text{honkong} \]

Clearly, forming portfolios is an interesting thing to do!

Basic question: why would we form portfolios?

Maybe the portfolio has a nice mean and variance (i.e. nice “average return” and nice “risk”)

There are some basic formulae that relate the mean and standard deviation of a linear combination to the means, variances and covariances of the input variables.

We can apply these formulae to understand how the mean and variance of a portfolio depend on the input assets. These formulae constitute the basic part of the tool-kit of those who really understand finance.
3.3. Mean and variance of a linear combination: 2 inputs

First, we consider the case where we have only two inputs.

2 inputs:

Suppose \( y = c_0 + c_1x_1 + c_2x_2 \)

Then,

\[
\bar{y} = c_0 + c_1\bar{x}_1 + c_2\bar{x}_2
\]

\[
\sigma_y^2 = c_1^2\sigma_{x_1}^2 + c_2^2\sigma_{x_2}^2 + 2c_1c_2\rho_{x_1x_2}
\]

Example: Portfolio means

\( \text{Port} = 0.5*\text{hongkong} + 0.5*\text{usa} \)

<table>
<thead>
<tr>
<th></th>
<th>usa</th>
<th>port</th>
</tr>
</thead>
<tbody>
<tr>
<td>hongkong</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>usa</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>-0.03</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>-0.03</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

For each month, we get the portfolio return as \( \frac{1}{2}*\text{hongkong} + \frac{1}{2}*\text{usa} \).

The mean returns on USA, and Hong Kong are 0.01346, and 0.02103

The mean return on Port is

\[0.5*0.01346 + 0.5*0.02103 = 0.01724\]

Let us do the same exercise for the variance:

Covariance matrix

<table>
<thead>
<tr>
<th></th>
<th>hongkong</th>
<th>usa</th>
</tr>
</thead>
<tbody>
<tr>
<td>hongkong</td>
<td>0.00521</td>
<td>0.00103</td>
</tr>
<tr>
<td>usa</td>
<td>0.00103</td>
<td>0.00111</td>
</tr>
</tbody>
</table>

The diagonals are variances, The off diagonals are Covariances.

As before, we apply the formula:

\[
\text{Var}(\text{Port}) = (0.5)^2(0.5)*0.00521 + (0.5)^2(0.5)*0.00111 + 2*(0.5)^2(0.5)^2*0.001
\]

\[= 0.25*0.00521 + 0.25*0.00111 + 0.5*0.001 = 0.0021.\]
Let us do it one more time:

\[
\text{Port} = 0.25 \times \text{usa} + 0.75 \times \text{hongkong}
\]

\[
\text{Var}(\text{Port}) =
\]
\[
(0.25)^2(0.25)^20.00111 +
(0.75)^2(0.75)^20.00521 +
(2)^2(0.25)(0.75)(0.00103)
\]
\[
= 0.0033
\]

3.4. Mean and variance of a linear combination: 3 inputs

Second, we consider the case where we have three inputs.

3 inputs:

Suppose

\[
Y = c_0 + c_1X_1 + c_2X_2 + c_3X_3
\]

Then,

\[
\bar{Y} = c_0 + c_1\bar{X}_1 + c_2\bar{X}_2 + c_3\bar{X}_3
\]

\[
s_Y^2 = c_1^2s_{X_1}^2 + c_2^2s_{X_2}^2 + c_3^2s_{X_3}^2 + 2(c_1c_2s_{X_1X_2} + c_1c_3s_{X_1X_3} + c_2c_3s_{X_2X_3})
\]

Example: Portfolio based on fidel, eqmrkt and windsor funds.

\[
\text{port} = 0.1 \times \text{fidel} + 0.4 \times \text{eqmrkt} + 0.5 \times \text{windsor}
\]

Covariance matrix

\begin{array}{ccc}
\text{fidel} & \text{eqmrkt} & \text{windsor} \\
0.003202 & 0.004700 & \\
0.003190 & 0.002990 & 0.0023658
\end{array}

\[
\text{Var(port)} = (0.1)^2(0.1)^20.003202 +
(0.4)^2(0.4)^20.0047 +
(0.5)^2(0.5)^20.0023658 +
2(0.1)(0.4)(0.00319) + (0.1)(0.5)(0.00299) + (0.4)(0.5)(0.00299) = 0.0030876
\]
3.5. Mean and variance of a linear combination: k inputs

K inputs: Suppose

\[ y = c_0 + c_1x_1 + c_2x_2 + \cdots + c_kx_k \]

then,

\[ \bar{y} = c_0 + c_1\bar{x}_1 + c_2\bar{x}_2 + \cdots + c_k\bar{x}_k \]

\[ s^2_y = c_1^2s^2_{x_1} + c_2^2s^2_{x_2} + \cdots + c_k^2s^2_{x_k} + 2\sum_{i<j}c_ic_js_{x_i}x_j \]

4. Portfolio example

Cut from a Finance Textbook:

Fama (1976) has illustrated this result empirically.\(^{15}\) His results are shown in Fig. 6.18. He randomly selected 50 securities listed on the New York Stock Exchange and calculated their standard deviations using monthly data from July 1963 to June 1968. Then a single security was selected randomly. Its standard deviation of return was around 18%. Next, this security was combined with another (also randomly selected) to form an equally weighted portfolio of two securities. The standard deviation fell to around 7.2%. Step by step more securities were randomly added to the portfolio until all 50 securities were included. Almost all of the diversification was obtained after the first 10-15 securities were randomly selected. In addition the portfolio stan-

\(^{15}\) See Fama (1976), Foundations of Finance, pp. 213-214.
**Dow Jones components:** January 1997 to December 2006 - 2516 observations

You will learn everything about the minimum variance portfolio in an Investments course. For now, just keep in mind it is a portfolio whose variance is smaller than other portfolios.

<table>
<thead>
<tr>
<th>Company</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALCOA</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERICAN EXPRESS</td>
<td>0.0747</td>
</tr>
<tr>
<td>Becton Dickinson</td>
<td>0.0568</td>
</tr>
<tr>
<td>CitiGroup</td>
<td>0.0701</td>
</tr>
<tr>
<td>Du Pont (Du Pont)</td>
<td>0.0502</td>
</tr>
<tr>
<td>General Electric</td>
<td>0.0546</td>
</tr>
<tr>
<td>General Motors</td>
<td>0.0393</td>
</tr>
<tr>
<td>Home Depot</td>
<td>0.0523</td>
</tr>
<tr>
<td>Honeywell</td>
<td>0.0546</td>
</tr>
<tr>
<td>IBM</td>
<td>0.0422</td>
</tr>
<tr>
<td>Intel</td>
<td>0.0393</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>0.0512</td>
</tr>
<tr>
<td>JP Morgan Chase</td>
<td>0.0558</td>
</tr>
<tr>
<td>McDonalds</td>
<td>0.0512</td>
</tr>
<tr>
<td>INTC</td>
<td>0.0393</td>
</tr>
<tr>
<td>AMERCO</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMAT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ABB</td>
<td>0.0471</td>
</tr>
<tr>
<td>ADP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AEP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIG</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIXT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ALCOA</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERICAN EXPRESS</td>
<td>0.0471</td>
</tr>
<tr>
<td>Becton Dickinson</td>
<td>0.0471</td>
</tr>
<tr>
<td>CitiGroup</td>
<td>0.0471</td>
</tr>
<tr>
<td>Du Pont (Du Pont)</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Electric</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Motors</td>
<td>0.0471</td>
</tr>
<tr>
<td>Home Depot</td>
<td>0.0471</td>
</tr>
<tr>
<td>Honeywell</td>
<td>0.0471</td>
</tr>
<tr>
<td>IBM</td>
<td>0.0471</td>
</tr>
<tr>
<td>Intel</td>
<td>0.0471</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>0.0471</td>
</tr>
<tr>
<td>JP Morgan Chase</td>
<td>0.0471</td>
</tr>
<tr>
<td>McDonalds</td>
<td>0.0471</td>
</tr>
<tr>
<td>INTC</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERCO</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMAT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ABB</td>
<td>0.0471</td>
</tr>
<tr>
<td>ADP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AEP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIG</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIXT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ALCOA</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERICAN EXPRESS</td>
<td>0.0471</td>
</tr>
<tr>
<td>Becton Dickinson</td>
<td>0.0471</td>
</tr>
<tr>
<td>CitiGroup</td>
<td>0.0471</td>
</tr>
<tr>
<td>Du Pont (Du Pont)</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Electric</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Motors</td>
<td>0.0471</td>
</tr>
<tr>
<td>Home Depot</td>
<td>0.0471</td>
</tr>
<tr>
<td>Honeywell</td>
<td>0.0471</td>
</tr>
<tr>
<td>IBM</td>
<td>0.0471</td>
</tr>
<tr>
<td>Intel</td>
<td>0.0471</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>0.0471</td>
</tr>
<tr>
<td>JP Morgan Chase</td>
<td>0.0471</td>
</tr>
<tr>
<td>McDonalds</td>
<td>0.0471</td>
</tr>
<tr>
<td>INTC</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERCO</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMAT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ABB</td>
<td>0.0471</td>
</tr>
<tr>
<td>ADP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AEP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIG</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIXT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ALCOA</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERICAN EXPRESS</td>
<td>0.0471</td>
</tr>
<tr>
<td>Becton Dickinson</td>
<td>0.0471</td>
</tr>
<tr>
<td>CitiGroup</td>
<td>0.0471</td>
</tr>
<tr>
<td>Du Pont (Du Pont)</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Electric</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Motors</td>
<td>0.0471</td>
</tr>
<tr>
<td>Home Depot</td>
<td>0.0471</td>
</tr>
<tr>
<td>Honeywell</td>
<td>0.0471</td>
</tr>
<tr>
<td>IBM</td>
<td>0.0471</td>
</tr>
<tr>
<td>Intel</td>
<td>0.0471</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>0.0471</td>
</tr>
<tr>
<td>JP Morgan Chase</td>
<td>0.0471</td>
</tr>
<tr>
<td>McDonalds</td>
<td>0.0471</td>
</tr>
<tr>
<td>INTC</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERCO</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMAT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ABB</td>
<td>0.0471</td>
</tr>
<tr>
<td>ADP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AEP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIG</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIXT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ALCOA</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERICAN EXPRESS</td>
<td>0.0471</td>
</tr>
<tr>
<td>Becton Dickinson</td>
<td>0.0471</td>
</tr>
<tr>
<td>CitiGroup</td>
<td>0.0471</td>
</tr>
<tr>
<td>Du Pont (Du Pont)</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Electric</td>
<td>0.0471</td>
</tr>
<tr>
<td>General Motors</td>
<td>0.0471</td>
</tr>
<tr>
<td>Home Depot</td>
<td>0.0471</td>
</tr>
<tr>
<td>Honeywell</td>
<td>0.0471</td>
</tr>
<tr>
<td>IBM</td>
<td>0.0471</td>
</tr>
<tr>
<td>Intel</td>
<td>0.0471</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>0.0471</td>
</tr>
<tr>
<td>JP Morgan Chase</td>
<td>0.0471</td>
</tr>
<tr>
<td>McDonalds</td>
<td>0.0471</td>
</tr>
<tr>
<td>INTC</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMERCO</td>
<td>0.0471</td>
</tr>
<tr>
<td>AMAT</td>
<td>0.0471</td>
</tr>
<tr>
<td>ABB</td>
<td>0.0471</td>
</tr>
<tr>
<td>ADP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AEP</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIG</td>
<td>0.0471</td>
</tr>
<tr>
<td>AIXT</td>
<td>0.0471</td>
</tr>
</tbody>
</table>

**Weights for the minimum variance portfolio**

| Mean       | 0.050 |
| Variance   | 0.921 |
| Skewness   | -0.161 |
| Kurtosis   | 3.056 |
| Sharpe     | 0.960 |
| Beta       | 0.960 |
| Alpha      | 0.960 |
| Volatility | 0.960 |
Portfolios based on the 30 Dowjones components

Blue dot: Minimum variance portfolio.
Red line: Minimum variance portfolio for a given mean return target.
Positive and negative weights are allowed, as long as they add up to 1.

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEWLETT-PACKARD</td>
<td>0.03</td>
</tr>
<tr>
<td>IBM</td>
<td>0.12</td>
</tr>
<tr>
<td>INTEL CORP</td>
<td>0.17</td>
</tr>
<tr>
<td>MICROSOFT CORP</td>
<td>0.09</td>
</tr>
<tr>
<td>JOHNSON&amp;JOHNSON</td>
<td>0.09</td>
</tr>
<tr>
<td>MERCK &amp; CO</td>
<td>0.27</td>
</tr>
<tr>
<td>PFIZER</td>
<td>0.06</td>
</tr>
<tr>
<td>PROCTER &amp; GAMBLE</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Portfolios based on 8 Dowjones components

Blue dot: Minimum variance portfolio.
Red line: Minimum variance portfolio for a given mean return target.
Positive and negative weights are allowed, as long as they add up to 1.
Black dots: Several randomly selected portfolios with weights between 0 and 1 and adding up to 1.
5. Simple Linear Regression

This is data on 128 homes.

\[ x = \text{size (square feet)} \]
\[ y = \text{price (dollars)} \]
Covariance matrix

\[ \begin{array}{cc}
\text{SqFt} & \text{Price} \\
\text{SqFt} & 44762.89 & 3143533 \\
\text{Price} & 3143533.22 & 721930821 \\
\end{array} \]

Hard to say what “721930821” means.

Correlation matrix

\[ \begin{array}{cc}
\text{SqFt} & \text{Price} \\
\text{SqFt} & 1.0000000 & 0.5529822 \\
\text{Price} & 0.5529822 & 1.0000000 \\
\end{array} \]

That is better!

Size and Price are clearly linearly correlated!

But what is the equation of the line you would draw through the data?

Linear regression fits a line to the plot.

When I "run a regression" I get values for the intercept and the slope

\[ \text{PRICE} = \text{intercept} + \text{slope}\times \text{SIZE} \]

\[ \text{PRICE} = -10091.13 + 70.23\times \text{SIZE} \]

Here is the scatter plot with the line drawn through it.

Looks reasonable!
5.1. Regression and Prediction

Suppose you had a house and you knew the size = 2000 but you do not know the price.

How could you use regression to guess or "predict" the price?

Just plug the size into the equation of the line:

\[
\text{estimated price} = -10091.13 + 70.23 \times 2000 \\
= 130368.9
\]

Correlation and covariance are "symmetric". The covariance between \( y \) and \( x \) is the same thing as the covariance between \( x \) and \( y \).

Regression is not symmetric.

We regress \( y \) on \( x \).

\( y \): dependent variable
\( x \): independent variable.

We say that "\( y \) depends on \( x \)."

In our example \( y \)=price depends on \( x \)=size.

Basic Probability

1. Probability and Random Variables
2. Bivariate Random Variables
3. The Marginal Distribution
4. The Conditional Distribution
5. Independence
6. Computing Joints from Conditionals and Marginals
Summary of the lecture

In this lecture we will enter the realm of statistical modeling. However, in order to set the stage for more complex scenarios, such as estimation, hypothesis testing and linear regression, we must introduce the notation, the jargon of probability. We begin by

• Defining probability and presenting properties;
• Discrete random variables where the outcomes are countable, such as number of votes for candidate A per county, number of children per family, and number of collisions monthly claimed in a certain insurance company;
• Bivariate random variables by contingency tables: For instance, should salary level have 4 categories (low, medium, high, extreme) and happiness have 3 categories (unhappy, indifferent, happy), then one could argue that there are 8 joint levels of salary by happiness in a 4 by 3 contingency table;
• Marginal distributions: Looking at the margins of a table;
• Conditional distributions: looking at a column/row of a table.

Book material

• Chapter 5:
  Probability, experiment, outcome and event (141-142 (12), 140-141 (13))
  Events mutually exclusive (143 (12), 142 (13))
  Events collectively exhaustive (page 144 (12), 143 (13))
  Classical probability (143 (12), 142 (13))
  Empirical probability (144 (12), 143 (13))
  Subjective probability (145 (12), 144 (13))
  Rules for computing probabilities (147-154 (12), 174-155 (13))
  Contingency tables (155-157 (12), 156-158 (13))
• Chapter 6
  Discrete random variable (184 (12 &13))

In this section of the course we learn about random variables and probability.

This is a very important topic that gets used in a variety of situations.

In order to think about many real world problems we have to face the fact that we are uncertain about some important aspects of the situation.
Monty Hall Problem

http://www.youtube.com/watch?v=mhlc7peGlGg

Birthday Problem

1. Probability and Random Variables

Example 1: S7P500 ups and downs in 2008

69 days
33 ups (33 1’s)
36 downs (36 0’s)

<table>
<thead>
<tr>
<th>Date</th>
<th>SP500</th>
<th>Up/Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1/2008</td>
<td>1447.86</td>
<td>1</td>
</tr>
<tr>
<td>1/2/2008</td>
<td>1447.36</td>
<td>0</td>
</tr>
<tr>
<td>1/3/2008</td>
<td>1513.61</td>
<td>36</td>
</tr>
<tr>
<td>1/4/2008</td>
<td>1416.30</td>
<td>0</td>
</tr>
<tr>
<td>1/5/2008</td>
<td>1499.99</td>
<td>0</td>
</tr>
<tr>
<td>1/6/2008</td>
<td>1469.11</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The average tells us the percentage of days that resulted in a positive SP500 return.

\[ \frac{0 + 0 + 1 + 0 + 1 + \ldots + 0 + 0 + 1 + 0 + \ldots}{69} = \frac{33(1) + 36(0)}{69} = 0.478 \]

48% of the days resulted in a positive return.
What will happen the next day?

Let $X$ denote the outcome. Then $X$ is either 0 or 1.

$X$ is a numerical quantity about which we are uncertain.

Random Variable: We do not know what $X$ will be, but we do know that it will be either 1 or 0 with certain probabilities.

What are these probabilities?

Tough questions! 47.8% is simply a rough estimate of the actual chance that SP500 is up in a given day. It is a rough estimate because it is based only on a very recent past, which may or may not represent the TRUE process driving the SP500 movement.

Example 2: Tossing a “fair” coin

Let us see a (much simpler) example where we are more comfortable assessing these probabilities

They are $Pr(X=1)=0.5$ and $Pr(X=0)=0.5$.

The probability of a 1 is 0.5.

The probability of a 0 is 0.5.

What does it mean?

The two possible outcomes are equally likely (by the very nature of a coin).

Over the long run, if we tossed the coin over and over again, we expect a 1 (or, equivalently, a zero) 50% of the time.

Probability as the long-run frequency

How often it happens

That is, if we toss the coin $n$ times with $n$ really big and

- $n_1$ is the number of 1's
- $n_0$ is the number of 0's

then,

$$\frac{n_1}{n} \approx .5 \quad \frac{n_0}{n} \approx .5$$
10 tosses
Of course, if we toss a coin 10 times we do not necessarily expect to get exactly 5 heads and 5 tails.

1,000 tosses
If we toss it 1000 times we expect the proportion to work out in the long run.

10,000 tosses
For “all” the tosses we expect to get 50% heads. For some, we could get something different. The closer to “all” we get, the more likely it is that the observed fraction will be close to .5.

The larger the “sample size” the closer the observed frequency of heads is to true probability of 50%.

Example 3: defects
Suppose we are making computer chips.
We record 1 if defective 0 if good.

Mean of defects = 0.12000
12% are defective
Suppose we are about to make the next chip.

What will happen?

We will get either a one (a defective part) or a zero (a good part) with some probabilities.

Again, we think of $Y$ as an uncertain quantity (a random variable) with two possible outcomes, 1 and 0 (defective and good) having probabilities:

$$\Pr(Y = 1) = ? \quad \Pr(Y = 0) = 1 - ?$$

Important

Unlike the coin example, it is not obvious what to use for the probabilities here (why?).

In our sample we have 12% defectives.

Does that mean that the probability of a defective = .12?

Of course, NOT!

Later in the course we will think of the sample frequency as an estimate of the true probability.

So, we might estimate probabilities:

$$\Pr(Y = 1) = .12 \quad \Pr(Y = 0) = .88$$

But, we could be wrong!
Example 4: tossing 3 coins simultaneously

Suppose we toss three coins.

Let H: head and T: tail.
Then, the eight possible outcomes are
HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.

Let X denote the number of heads (it is a random variable). X has three possible outcomes: 0, 1, 2 or 3.

Tree diagram

Definition of discrete random variable

A discrete Random Variable is a numerical quantity we are unsure about. We quantify our uncertainty by:

1. Listing the numbers it could turn out to be, i.e., the possible outcomes.
2. Assigning to each number a probability. Probabilities are numbers between 0 and 1 and sum up to 1.

The word “discrete” refers to the fact that we just have a list of outcomes. Later we will study continuous random variables where “any” outcome is possible.
For the random variable denoted by $X$, we often use $x$ to denote a possible outcome.

**Example**

<table>
<thead>
<tr>
<th>$\Pr(X=x)$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>0.50</td>
<td>1</td>
</tr>
<tr>
<td>0.25</td>
<td>2</td>
</tr>
</tbody>
</table>

This table gives the probability distribution of the random variable $X$.

Each probability tells us how often the corresponding outcome happens.

**Interpret.** 25% of the time we get 2 heads.

**Important:** a probability distribution is a list of probabilities, one for each outcome.

**Notation**

We use various notations for the probability that the random variable $X$ takes on the value (outcome) $x$:

$$\Pr(X = x), \Pr(x), p_X(x), p(x)$$

These all mean the same thing.

With $p(x)$ it must be understood from the context that you are talking about the outcome $x$ of the random variable $X$.

**Example 5:**

Suppose we toss a die, let $z$ denote the outcome:

<table>
<thead>
<tr>
<th>$z$</th>
<th>$p(z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>5</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{6}$</td>
</tr>
</tbody>
</table>
Note: To get the probability that any one of a bunch of outcomes occurs we sum up their probabilities.

\[ P(a < X < b) = \sum_{x = a}^{b} p(x) \]

Example 5 (cont.)
Suppose you role a die.
Let \( X \) be the number.
\( P(2 < X < 5) = P(X = 3) + P(X = 4) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} \).

Example 6:
Suppose we toss two dice.
Let \( Y \) denote the sum.

<table>
<thead>
<tr>
<th>( y )</th>
<th>( p(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>3</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>5</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>6</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>7</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>8</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>9</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>10</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>11</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>12</td>
<td>( \frac{1}{4} )</td>
</tr>
</tbody>
</table>

What is the probability of getting more than 8?
\( \Pr(Y > 8) = \Pr(Y = 9) + \Pr(Y = 10) + \Pr(Y = 11) + \Pr(Y = 12) \)

Example 7: Investing in an asset
Suppose you are considering investing in an asset.
Let \( R \) denote the return next month.
We think of \( R \) as a random variable.
We do not know what the return will be (it is random) but we assume we know what the possible outcomes and probabilities are.
In other words, we are truly modeling a future event.

<table>
<thead>
<tr>
<th>( r )</th>
<th>( 0.05 )</th>
<th>( 0.10 )</th>
<th>( 0.15 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Pr(R = r) )</td>
<td>( 0.1 )</td>
<td>( 0.5 )</td>
<td>( 0.4 )</td>
</tr>
</tbody>
</table>

The probability that the return will be greater than 0.05 is 0.9.
Graphing discrete random variables

We can use a graph to see the probability distribution of a random variable. Simply plot p(y) versus y:

Example 8: Y = the sum of two dice.

<table>
<thead>
<tr>
<th>y</th>
<th>p(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>\frac{1}{36}</td>
</tr>
<tr>
<td>3</td>
<td>\frac{2}{36}</td>
</tr>
<tr>
<td>4</td>
<td>\frac{3}{36}</td>
</tr>
<tr>
<td>5</td>
<td>\frac{4}{36}</td>
</tr>
<tr>
<td>6</td>
<td>\frac{5}{36}</td>
</tr>
<tr>
<td>7</td>
<td>\frac{6}{36}</td>
</tr>
<tr>
<td>8</td>
<td>\frac{5}{36}</td>
</tr>
<tr>
<td>9</td>
<td>\frac{4}{36}</td>
</tr>
<tr>
<td>10</td>
<td>\frac{3}{36}</td>
</tr>
<tr>
<td>11</td>
<td>\frac{2}{36}</td>
</tr>
<tr>
<td>12</td>
<td>\frac{1}{36}</td>
</tr>
</tbody>
</table>

Example 9: Y = the sum of three dice

The Bernoulli distribution

The situation where something happens or not and we want to talk about the probability of it happening is our most basic scenario.

To describe this situation we use a random variable which is 1 if something happens and 0 otherwise and probability ("it happens") = p.

Such a random variable is said to have the Bernoulli distribution.

Notation: Y ~ Bernoulli(p) means P(Y=1)=p, P(Y=0)=1-p

Example 10: Toss a coin. X=1 if head, 0 else.

Then,

X ~ Bernoulli(0.5).
The random variable $X$ has the Bernoulli distribution with parameter $p$ (between 0 and 1) if

$$\Pr(X = 1) = p$$
$$\Pr(X = 0) = 1 - p$$

In general, we think of $X=1$ as the thing happens and $X=0$ as the thing does not happen.

**Something to think about**

The word random variable refers to the outcome before it happens.

A random variable describes what we think will happen.

After we have an outcome (say, after we toss a coin), the obtained value is sometimes called a draw from the common distribution (it is a data point or an observation from the sample).

The Bernoulli distribution is named after Jakob Bernoulli, who was born in Basel, Switzerland on December 27, 1654 and lived until August 16, 1705. He is one of the eight prominent mathematicians in the Bernoulli family.

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erhard Weigel</td>
<td>Universität Leipzig</td>
<td>1650</td>
</tr>
<tr>
<td>Gottfried Leibniz</td>
<td>Universität Altdorf</td>
<td>1666</td>
</tr>
<tr>
<td>Jakob Bernoulli</td>
<td>Universit&quot;at Basel</td>
<td>1694</td>
</tr>
<tr>
<td>Leibniz</td>
<td>École Polytechnique</td>
<td></td>
</tr>
<tr>
<td>Joseph Louis Lagrange</td>
<td>École Polytechnique</td>
<td></td>
</tr>
<tr>
<td>Simon Denis Poisson</td>
<td>École Polytechnique</td>
<td></td>
</tr>
<tr>
<td>Jakob Erhard Weigel</td>
<td>Universität Leipzig</td>
<td>1650</td>
</tr>
<tr>
<td>Robert Augustus</td>
<td>Virc University</td>
<td>1650</td>
</tr>
<tr>
<td>Hubert Anton Weissen</td>
<td>Universität Leipzig</td>
<td>1694</td>
</tr>
<tr>
<td>Jakob Erhard Weigel</td>
<td>Universität Zürich</td>
<td>1694</td>
</tr>
<tr>
<td>Gottfried Wilhelm</td>
<td>Universität Basel</td>
<td>1694</td>
</tr>
<tr>
<td>Jean Louis LAGRANGE</td>
<td>École Polytechnique</td>
<td>1694</td>
</tr>
<tr>
<td>Jacques BERNOUlli</td>
<td>École Polytechnique</td>
<td>1694</td>
</tr>
<tr>
<td>J. J. Leibniz</td>
<td>Universität Altdorf</td>
<td>1666</td>
</tr>
<tr>
<td>J. J. Leibniz</td>
<td>Universität Leipzig</td>
<td>1650</td>
</tr>
<tr>
<td>Joseph Louis Lagrange</td>
<td>École Polytechnique</td>
<td>1694</td>
</tr>
<tr>
<td>Simon Denis Poisson</td>
<td>École Polytechnique</td>
<td>1694</td>
</tr>
</tbody>
</table>

180
2. Bivariate Discrete Random Variables

Let X be the return on the nasdaq.
Let Y be the return on the djia.
We can think of both as random variables.

Could there be a relationship? If one "turns out big," will the other tend to be big as well?

We need probability to describe what both turn out to be.


We give the bivariate probability distribution of a pair of random variables by:

1. Listing out all the possible pairs of values that they could take on.
2. For each pair we give a probability.
   The sum of the probabilities over all pairs = 1.

Example 10: SP&500 and Dowjones ups and downs in 2008

Let X=1 if SP&500 is up and X=0 if it is down
Let Y=1 if DOW is up and Y=0 if it is down

Then, the joint distribution of X and Y is given by this table

<table>
<thead>
<tr>
<th>(x,y)</th>
<th>p(x,y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>0.478</td>
</tr>
<tr>
<td>(0,1)</td>
<td>0.072</td>
</tr>
<tr>
<td>(1,0)</td>
<td>0.044</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0.406</td>
</tr>
</tbody>
</table>

We simply list out all possibilities for the pairs and give each one a probability.
**Example 11: Tossing two coins**

Let $X$ be the result of tossing a coin ($1=H$, $0=T$). Let $Y$ be the result from a second coin toss.

<table>
<thead>
<tr>
<th>$(x,y)$</th>
<th>$p(x,y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(0,0)$</td>
<td>0.25</td>
</tr>
<tr>
<td>$(0,1)$</td>
<td>0.25</td>
</tr>
<tr>
<td>$(1,0)$</td>
<td>0.25</td>
</tr>
<tr>
<td>$(1,1)$</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Then, the joint distribution of $X$ and $Y$ is given by this table.

We simply list out all possibilities for the pairs and give each one a probability.

**Notation:**

$$p(x,y) = \Pr(X = x \text{ and } Y = y)$$

As before, we might also write

$$p_{xy}(x,y)$$

The joint bivariate distribution of $X$ and $Y$ is specified by the numbers $p(x,y)$ for all possible $x$ and $y$ (for all possible pairs).

The distribution is discrete in that there is just a list (a finite number) of possible $(x,y)$ pairs.

**Note:** An alternative way to display the probabilities is:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.478</td>
<td>0.044</td>
</tr>
<tr>
<td>1</td>
<td>0.072</td>
<td>0.406</td>
</tr>
</tbody>
</table>

We have a two way table where each spot in the table corresponds to a possible $(x,y)$ pair. At each spot we give the probability of the corresponding pair.
Example 12: Investing in 2 assets

Let X and Y be returns on two different assets.

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.10</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Y</td>
<td>0.03</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.30</td>
</tr>
</tbody>
</table>

What does this table say about the relationship between X and Y?

What is the probability that they are equal?

Probability means the same thing as in the univariate case

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.10</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Y</td>
<td>0.03</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.30</td>
</tr>
</tbody>
</table>

We expect to see the pair (x,y)=(10%,10%) 0.30 of the time.

3. The Marginal Distribution

The joint distribution of X and Y tells us what we expect to happen for both of them.

From this, we should be able to figure out what happens for one of them.

That is, we should be able to get

\[ p_X(x) \quad \text{and} \quad p_Y(y) \]

from

\[ p_{XY}(x,y) \]
Example 12 (cont.)

\[
\begin{array}{ccc}
X & 5\% & 10\% & 15\% \\
5\% & 0.10 & 0.07 & 0.07 \\
10\% & 0.03 & 0.30 & 0.03 \\
15\% & 0.05 & 0.05 & 0.30 \\
\end{array}
\]

\[Y\]

What is \( p_X(5\%) \)?

\[
p_X(5\%) = p_{X,Y}(5\%,5\%) + p_{X,Y}(5\%,10\%) + p_{X,Y}(5\%,15\%)
= 0.10 + 0.03 + 0.05 = 0.18
\]

The marginal distributions

Given the joint distribution of \( X \) and \( Y \) defined by

\[ p_{X,Y}(x,y) \]

the marginal (individual) distributions of \( X \) and \( Y \) are given by,

\[ p_x(x) = \sum_{y} p_{X,Y}(x,y) \]
\[ p_y(y) = \sum_{x} p_{X,Y}(x,y) \]

Example 12 (cont.)

Let us write out the marginal distributions (or, based on a common jargon, the marginals) using our standard two way table.

\[
\begin{array}{ccc}
X & 5\% & 10\% & 15\% \\
5\% & 0.10 & 0.07 & 0.07 \\
10\% & 0.03 & 0.30 & 0.03 \\
15\% & 0.05 & 0.05 & 0.30 \\
\end{array}
\]

\[
p_x(x) \quad p_{y}(y) \quad p_{x}(x) \quad 0.18 \quad 0.42 \quad 0.40 \\
\]
### Example 13:
In 1971 the Gallup company estimated the following joint probability distribution for $Y=$happiness and $X=$ income (at 4 levels).

<table>
<thead>
<tr>
<th>Salary (X)</th>
<th>Happiness (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>0.03 0.12 0.07</td>
</tr>
<tr>
<td>7.5</td>
<td>0.02 0.13 0.11</td>
</tr>
<tr>
<td>12.5</td>
<td>0.01 0.13 0.14</td>
</tr>
<tr>
<td>17.5</td>
<td>0.01 0.09 0.14</td>
</tr>
</tbody>
</table>

### Review questions

1) What is the chance a randomly chosen person is rich and happy? (Easy)

2) What is the chance that a person is rich? (Easy)

3) What is the chance that a person is happy? (Easy)

4) Given you know they are rich, what is the chance they are happy? (Yikes…)

To see what the answer to the fourth question is, let us first rephrase it.

Out of the people that are rich, what percent are also happy?

$$\text{percent rich and happy} = \frac{0.14}{0.01 + 0.09 + 0.14}$$
Out of the times $X=x$, what fraction also has $Y=y$?

In general, for random variables $X$ and $Y$, we ask what is $\Pr(Y=y)$ given we know $X=x$.

The conditional distribution

Given discrete random variables $X$ and $Y$ with associated probabilities $p_{XY}(x,y)$, the probability that $Y=y$ given $X=x$ is denoted by $\Pr(Y=y|X=x) = p_{Y|X}(y|x)$, and

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$
For a fixed $x$, the numbers $p(y|x)$ (for the various possible $y$) give the conditional distribution of $Y$ given $X=x$.

Of course,

$$p_{XY}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

**Notation**

$Y|X = x$ is sometimes used as a symbol for the conditional probability distribution of $Y$ given $X=x$.

**Recall:** probability distributions are lists of probabilities (one probability for every possible outcome).

**Example 13 (cont.)**

The conditional distribution of $Y$ given $X = 17.5$ is

| $y$ | $Pr(y|x=17.5)$ |
|-----|----------------|
| 0   | $0.01/0.24 = 0.0416$ |
| 1   | $0.09/0.24 = 0.375$  |
| 2   | $0.14/0.24 = 0.5833$ |

Note that the conditional probabilities have to sum up to 1. Note, also, that given $x=17.5$ the first three rows of the table become irrelevant.

Let us compare the marginal distribution of $Y$ to the conditional distribution of $Y$ given $X = 17.5$:

What do you notice?

| $y$ | $p(y)$ | $y$ | $p(y|17.5)$ |
|-----|--------|-----|-------------|
| 0   | 0.07   | 0   | $0.01/0.24 = 0.0416$ |
| 1   | 0.47   | 1   | $0.09/0.24 = 0.375$  |
| 2   | 0.46   | 2   | $0.14/0.24 = 0.5833$ |

Learning that $X=17.5$ changes what you expect $Y$ to be (the probabilities are different).

**Important:** conditional probability shows us how to change our ideas about what we expect given information.
Example 13 (cont.)
What is the distribution of $X$ given $Y=0$?

| $X$   | $p(x|Y=0)$ |
|------|------------|
| 2.5  | 3/7        |
| 7.5  | 2/7        |
| 12.5 | 1/7        |
| 17.5 | 1/7        |

Example 12 (cont.)

<table>
<thead>
<tr>
<th>$X$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>$p_x(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>0.10</td>
<td>0.07</td>
<td>0.07</td>
<td>0.24</td>
</tr>
<tr>
<td>10%</td>
<td>0.03</td>
<td>0.30</td>
<td>0.03</td>
<td>0.36</td>
</tr>
<tr>
<td>15%</td>
<td>0.05</td>
<td>0.05</td>
<td>0.30</td>
<td>0.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$Y$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>$p_y(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>0.56</td>
<td>0.17</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>0.175</td>
<td>0.075</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>15%</td>
<td>0.175</td>
<td>0.075</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

Example: The Monty Hall Problem

**Action 1:** do not swap unopened doors

$Pr(win \mid action \ 1)=1/3$

**Action 2:** swap unopened doors

$Pr(win) = Pr(win \mid goat \ behind \ selected \ door)Pr(goad \ behind \ selected \ door) + Pr(win \mid car \ behind \ selected \ door)Pr(car \ behind \ selected \ door)$

$= (1)(2/3) + (0)(1/3) = 2/3$

Therefore,

$Pr(win \mid action \ 2)=2/3$
5. Independence

In our happiness/money example, knowing how much money a person has changes your expectations about how happy the person is.

This makes us think that happiness and money have something to do with each other (i.e., they are not independent).

Learning $X=x$ changed our probabilities for $Y$.

There was information in $X=x$ about $Y$.

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>.25</td>
<td>.25</td>
</tr>
</tbody>
</table>

What is the distribution of $Y$ given $X=0$?

| $y$ | $p(y|0)$ |
|-----|----------|
| 0   | $0.25/0.5 = 0.5$ |
| 1   | $0.25/0.5 = 0.5$ |

What is the distribution of $Y$ given $X=1$?

| $y$ | $p(y|1)$ |
|-----|----------|
| 0   | $0.25/0.5 = 0.5$ |
| 1   | $0.25/0.5 = 0.5$ |

What is the marginal $p(y)$?

<table>
<thead>
<tr>
<th>$y$</th>
<th>$p(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$0.25 + 0.25 = 0.5$</td>
</tr>
<tr>
<td>1</td>
<td>$0.25 + 0.25 = 0.5$</td>
</tr>
</tbody>
</table>

All three of $p(y|0)$, $p(y|1)$, and $p(y)$ are the same!

What does this mean?

What you expect for $Y$ does not depend on what you know about $X$.

There is no information in $X$ about $Y$. They have nothing to do with each other.
If X is the toss of the first coin, and Y is the toss of the second coin, this makes sense.

Knowing whether the first coin is 0 or 1 (tails or heads) does not affect what you expect for the next one.

When two things have nothing to do with each other we say that they are independent.

**Independence**

Let X and Y be discrete random variables.

If \( p_{Y|X}(y|x) = p_Y(y) \) for all \( x,y \)

we say the random variables are independent.

**Another (equivalent) definition of Independence**

Suppose X and Y are independent. Then,

\[
p_Y(y) = p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}
\]

So,

\[
p_{XY}(x,y) = p_Y(y)p_X(x)
\]

The joint is the product of the marginals (this is the standard textbook definition).

**Example**

<table>
<thead>
<tr>
<th>( X )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The two coins again:
Example 12 (cont.)

What is $Y \mid X=5\%$?

<table>
<thead>
<tr>
<th>$y$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(y\mid X=5%)$</td>
<td>0.56</td>
<td>0.17</td>
<td>0.28</td>
</tr>
</tbody>
</table>

What is $Y \mid X=15\%$?

<table>
<thead>
<tr>
<th>$y$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(y\mid X=15%)$</td>
<td>0.175</td>
<td>0.075</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Clearly, $X$ and $Y$ are not independent in this example.

Example 14:

The Gallup Organization did a nationwide poll asking the following question:

The Supreme Court has ruled that a woman may go to a doctor to end pregnancy at any time during the first 4 months of pregnancy, do you favor or oppose this ruling?

<table>
<thead>
<tr>
<th></th>
<th>Favor</th>
<th>Opposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.27</td>
<td>0.21</td>
</tr>
<tr>
<td>Female</td>
<td>0.24</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Are they independent?

Solution:

$Pr(male) = 0.48$ and $Pr(female) = 0.52$

$Pr(favor) = 0.51$ and $Pr(opposed) = 0.49$

Therefore,

$P(favor\mid male)=Pr(favor, male)/Pr(male)=0.27/0.48=0.5625$

$P(favor\mid female)=Pr(favor, female)/Pr(female)=0.24/0.52=0.4615$

Since $P(favor\mid male)$ and $P(favor\mid female)$ are not the same, it follows that gender and view towards pregnancy are not independent.
Example 15:

Same Gallup poll as before, now with people classified by political views (Leftist or rightist).

<table>
<thead>
<tr>
<th></th>
<th>Favor</th>
<th>Opposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>0.459</td>
<td>0.441</td>
</tr>
<tr>
<td>Right</td>
<td>0.051</td>
<td>0.049</td>
</tr>
</tbody>
</table>

Are they independent?

Solution:

\[
\begin{align*}
\Pr(\text{left}) &= 0.9 \quad \text{and} \quad \Pr(\text{right}) = 0.1 \\
\Pr(\text{favor}) &= 0.51 \quad \text{and} \quad \Pr(\text{opposed}) = 0.49 \\
\Pr(\text{left})\Pr(\text{favor}) &= (0.9)(0.51) = 0.459 = \Pr(\text{left, favor}) \\
\Pr(\text{left})\Pr(\text{opposed}) &= (0.9)(0.49) = 0.441 = \Pr(\text{left, opposed}) \\
\Pr(\text{right})\Pr(\text{favor}) &= (0.1)(0.51) = 0.051 = \Pr(\text{right, favor}) \\
\Pr(\text{right})\Pr(\text{opposed}) &= (0.1)(0.49) = 0.049 = \Pr(\text{right, opposed})
\end{align*}
\]

Conclusion: Since the joint equals the product of the marginals, political view and view towards pregnancy are independent.

6. Computing joints from conditionals and marginals

Remember the definition of conditional probability? Here it is:

\[
p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}
\]

We can rewrite the previous formula as follows:

\[
p_{XY}(x,y) = p_Y(y)p_{X|Y}(x|y)
\]
Interpret. We are simply saying:

\[ p_{XY}(x,y) = p_y(y) \times p_{X|Y}(x|y) \]

Alternatively (but equivalently),

How often we get \( x \) and \( y \) equals how often we get \( y \) times the fraction of those times we then get \( x \).

This is a very straightforward way to interpret and compute joint probabilities. Let us see a couple of examples:

**Example 16:**

Suppose you have 10 voters. 5 are Republicans, 5 are Democrats. You randomly pick two. This would be a random sample of two voters from 10.

What is the probability of two Republicans?

Think of randomly picking the first, and then the second.

Probability that both are Republicans = probability the first is a Republican times the probability the second is a Republican given that the first is \( \frac{5}{10} \times \frac{4}{9} = \frac{2}{9} \)

Is the outcome for the second chosen voter independent of the outcome for the first?

Suppose we have 5 million Democrats and 5 million Republicans.

Is the outcome for the second chosen voter independent of the outcome for the first in this case?
Example 17: Birthday problem

The birthday problem asks whether any of the persons in this classroom have a matching birthday with any of the others — not one in particular.

In a list of 50 people, for example, comparing the birthday of the first person on the list to the others allows 49 chances for a matching birthday.

Comparing every person to all of the others allows 1225 distinct chances: in a group of 50 people there are 50×49/2 = 1225 pairs.

To compute the approximate probability that in a room of \( n \) people, at least two have the same birthday, we disregard variations in the distribution, such as leap years, twins, seasonal or weekday variations, and assume that the 365 possible birthdays are equally likely.

Real-life birthday distributions are not uniform since not all dates are equally likely.

It is easier to first calculate the probability \( p(n) \) that all \( n \) birthdays are different. If \( n \leq 365 \), it is

\[
p(n) = \frac{365}{365} \cdot \frac{364}{365} \cdot \ldots \cdot \frac{365-n+1}{365}
\]

because the second person cannot have the same birthday as the first (364/365), the third cannot have the same birthday as the first two (363/365), etc.

The event of at least two of the \( n \) persons having the same birthday is complementary to all \( n \) birthdays being different. Therefore, its probability \( p(n) \) is

\[
p(n) = 1 - p(n)
\]

The following table shows the probability for some other values of \( n \):

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>11.7%</td>
</tr>
<tr>
<td>20</td>
<td>41.1%</td>
</tr>
<tr>
<td>30</td>
<td>70.6%</td>
</tr>
<tr>
<td>50</td>
<td>97.0%</td>
</tr>
<tr>
<td>57</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

A reasonable approximation is

\[
p(n) = 1 - p(n) \approx 1 - e^{-n(n-1)/[2 \times 365]}
\]

where \( e = 2.72 \).
Example 18: Inverse Probability

X = 1 : patient is ill
X = 0 : patient is not ill

Doctor’s expert opinion : Pr(X=1) = 0.05

Clinical trial characteristics

T=1 : test indicates patient is ill
T=0 : test indicates patient is not ill

Pr(T=1|X=1) = 0.90
Pr(T=0|X=0) = 0.80

It is easy to see that

Pr(T=1,X=1) = Pr(T=1|X=1)Pr(X=1) = (0.9)(0.05) = 0.045
Pr(T=1,X=0) = Pr(T=1|X=0)Pr(X=0) = (0.2)(0.95) = 0.190
Pr(T=0,X=1) = Pr(T=0|X=1)Pr(X=1) = (0.1)(0.05) = 0.005
Pr(T=0,X=0) = Pr(T=0|X=0)Pr(X=0) = (0.8)(0.95) = 0.760

The joint distribution of X and Y and the marginal distributions of X and Y are:

<table>
<thead>
<tr>
<th>T</th>
<th>0</th>
<th>1</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.760</td>
<td>0.190</td>
<td>0.950</td>
</tr>
<tr>
<td>1</td>
<td>0.005</td>
<td>0.045</td>
<td>0.050</td>
</tr>
</tbody>
</table>

P(T) | 0.765 | 0.235 | 1.000

Therefore,

Pr(X=1 | T=1) = Pr(X=1,T=1)/Pr(T=1)
          = 0.045/0.235
          = 0.1915

Pr(X=1 | T=0) = Pr(X=1,T=0)/Pr(T=0)
          = 0.005/0.765
          = 0.0065
More on Probability

1. Continuous Distributions
2. The Normal Distribution
3. The Cumulative Distribution Function
4. Expectation as a Long Run Average
5. Expected Value and Variance of Continuous RV’s
6. Random Variables and Formulas
7. Covariance/correlation for pairs of random variables
8. Independence and correlation

Summary of the lecture

In this lecture we will learn about

- Continuous distributions, such as the famous normal distributions,
- How to compute probabilities under continuous distributions,
- Independent and identically distributed (i.i.d.) draws: random sample,
- How to related actual data to the normal model: model fitting,
- How to compute means, variances, covariances of functions of random variables
- The binomial distribution to model the number of times a particular characteristic appears in your sample
- The famous (or infamous) Central Limit Theorem (C.L.T.)

Book material

- The family of normal distributions (pages 217-221 (12), 227-241 (13))
- The standard normal distribution (pages 219-223 (12), 229-233 (13))
- Finding areas under the normal curve (pages 224-228 (12), 234-238 (13))
- The mean, variance, and standard deviation of a probability distribution (184-187 (12), 185-187 (13))
- Sampling distribution of the sample mean (pages 209-211 (12), 210-213 (13))
- The Central limit theorem (pages 263-268 (12), 274-286 (13))
1. Continuous distributions

Example: Suppose we have a machine that cuts cloth. When pieces are cut, there are remnants.

We believe that the length of a remnant could be anything between 0 and 0.5 inches and, any value in the interval is equally likely.

The machine is about to cut, leaving a new remnant. The length of the remnant is a number we are unsure about, so it is a random variable.

How do we describe our beliefs?

Example: 99.87% of S&P500 returns falls in the range -4.922 and 4.925, with 1st, 2nd and 3rd quartiles given by -0.404, 0.0368 and 0.0448, respectively.

These are examples of continuous random variables.

The random variables can take on any value in an interval.

In both examples each value in the interval is equally likely.
We can not list out the possible values and give each a probability. Instead we give the probability of intervals. Instead of

\[ \Pr(X=x) = 0.1 \]

we have

\[ \Pr(a<X<b) = 0.1 \]

Example (cont.): 14391 distinct returns out of 14665 days. Therefore, 0.007% is the approximate probability that a future return equals any of the previous 14391 returns.

Probability density function
One convenient way to specify the probability of any interval is with the probability density function (pdf).

The probability of an interval is the area under the pdf.

In this example values closer to 0 are more likely.

For this random variable the probability that it is in the interval \( [0,2] \) is 0.477. (47.7% of the time it will fall in this interval).
Note: The area under the entire curve must be 1 (Why?)

Here is another p.d.f:

Most of the probability is concentrated in 1 to 15, but you could get a value much bigger. This kind of distribution is called skewed to the right.

For a continuous random variable X, the probability of the interval (a,b), denoted by

\[ \Pr(a < X < b) \]

is the area under the probability density function from a to b.

2. The Normal Distribution

This pdf describes the standard normal distribution.

We often use Z to denote the RV which has this pdf.

Note: any value in \(-\infty \rightarrow \infty\) is “possible”.
Properties of the standard normal

\[ P(-1 < Z < 1) = 0.6826895 \]
\[ P(-2 < Z < 2) = 0.9544997 \]
\[ P(-3 < Z < 3) = 0.9973002 \]
\[ P(-4 < Z < 4) = 0.9999367 \]
\[ P(-5 < Z < 5) = 0.9999994 \]

Also

\[ P(-1.96 < Z < 1.96) = 0.9500042 \]

In these notes I will usually act as if 1.96 = 2.

The standard normal is not of much use by itself. How often would you use that pdf to describe a quantity of interest?

When we say "the normal distribution", we really mean a family of distributions all of which have the same "shape" as the standard normal.

If \( X \) is a normal random variable we write:

\[ X \sim N(\mu, \sigma^2) \]

\[ X \sim N(\mu, \sigma^2) \] means \( X \) has this pdf:

\[ P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95 \]
\[ P(\mu - \sigma < X < \mu + \sigma) = 0.68 \]
The normal family has two parameters:

\( \mu \): where the curve is centered

\( \sigma \): how spread out the curve is

\( Z \sim N(0,1) \)
\( Y \sim N(1,0.5^2) \)
\( X \sim N(0,4) \)

Z, X, and Y are all "normally distributed".

Interpretation of \( \mu \) and \( \sigma \)

We will see that

\( \mu \) is the “mean”
\( \sigma \) is the "standard deviation"
\( \sigma^2 \) is the "variance"

of the normal random variable.

But we have not yet defined the mean and variance of a continuous random variable.

I will use these names right away, but explain what they mean later.

Interpreting the normal

\( X \sim N(\mu, \sigma^2) \):

There is a 95% chance \( X \) is in the interval \( \mu \pm 2\sigma \)

\( \mu \): what you think will happen

\( \pm 2\sigma \): how wrong you could be

\( \mu \): where the curve is

\( \sigma \): how spread out the curve is
Example:
You believe the return next month on a certain mutual fund, denoted by R, can be described by

\[ R \sim N(0.1, 0.01) \]

Normality allows us to say that there is a 95% probability that R will be in the interval (-0.1, 0.3)

Why?
\[ 0.1 - 2\times(0.1) = -0.1 \]
\[ 0.1 + 2\times(0.1) = 0.3 \]

3. The Cumulative Distribution Function

The cumulative distribution function (c.d.f.) is just another way (besides the p.d.f.) to specify the probability of intervals for a continuous random variable.

Definition: For a random variable X the c.d.f., which we denote by F, is defined by

\[ F(x) = \Pr(X \leq x) \]

which is the area to the left of x.

Example: c.d.f. of the standard normal distribution.
The c.d.f. is handy for computing the probabilities of intervals.

\[ P(a < X < b) = P(X < b) - P(X < a) = F(b) - F(a) \]

Example:
For Z (standard normal), we have:
\[ P(-1 < X < 1) = F(1) - F(-1) = 84.13\% - 15.87\% = 68.26\% \]

Example (cont.):

<table>
<thead>
<tr>
<th>x</th>
<th>F(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>0.00000029</td>
</tr>
<tr>
<td>-4</td>
<td>0.00003167</td>
</tr>
<tr>
<td>-3</td>
<td>0.00134990</td>
</tr>
<tr>
<td>-2</td>
<td>0.02275013</td>
</tr>
<tr>
<td>-1</td>
<td>0.15865525</td>
</tr>
<tr>
<td>0</td>
<td>0.50000000</td>
</tr>
<tr>
<td>1</td>
<td>0.84134475</td>
</tr>
<tr>
<td>2</td>
<td>0.97724987</td>
</tr>
<tr>
<td>3</td>
<td>0.99865010</td>
</tr>
<tr>
<td>4</td>
<td>0.99996833</td>
</tr>
<tr>
<td>5</td>
<td>0.99999971</td>
</tr>
</tbody>
</table>

Note: for x big enough, F(x) must get close to 1. For x small enough, F(x) must get close to 0.

The probability of an interval is the jump in the c.d.f. over that interval.
**Example: S&P500 and NASDAQ**

The 14665 daily returns were used to compute the empirical c.d.f. for the S&P500 returns. Sample mean=0.0368 and sample variance=0.7291.

Comparing the empirical c.d.f. of S&P500 returns with the normal model with mean 0.0368 and variance 0.7291.

A closer look between -2 and 2: The normal model IS NOT a good model for the SP500 returns.
NASDAQ composite returns from 2000-2008
Mean return = -0.02369155
Standard deviation = 1.945645
Skewness = 0.3094729
Excess kurtosis = 4.687585
Sample size = 2262

I propose the "other" model as an alternative to the normal model. The "other" model fits the data "better" than the normal model both in the center and the tails of the empirical distribution.

The normal model gives negligible probability to nasdaq returns above 6.
The "other" model mimics the empirical quantiles all the way to returns equal to 10.

Empirical c.d.f. versus models

A closer look at the right tail

<table>
<thead>
<tr>
<th>MODEL</th>
<th>DATA</th>
<th>NORMAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extreme Years</td>
<td>Prob.</td>
<td>Years</td>
</tr>
<tr>
<td>4.386</td>
<td>98%</td>
<td>0.2</td>
</tr>
<tr>
<td>5.526</td>
<td>99%</td>
<td>0.4</td>
</tr>
<tr>
<td>10.231</td>
<td>99.9%</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Prob. = Probability of the right tail
Years = expected number of years until rare event.
Robert Rubin was Bill Clinton's treasury secretary. He has worked at the top of Goldman Sachs and Citigroup. But he made arguably the single most influential decision of his career on September 11, when he asked David Viniar, Goldman's chief financial officer, to go on live television and promise that the public should not fear the company's financial health.

From the economist archive

Almost as damaging is the belief that markets have made “value at risk” (VaR) calculations, a measure of the potential losses of a portfolio. This is supposed to show whether banks and other financial entities are being safely run. Regulators use VaR calculations to work out whether capital banks need to put aside for a rainy day. But the calculations are flawed. The mistake was to turn an isolated event into a known “tail risk”. Think of the tails of a range of possible daily losses and gains from any distribution. Most of the time you gain a little or lose a little. Occasionally you gain or lose a lot. Very rarely you win or lose a fortune. If you plot those daily investments on a graph, you get the familiar bell-shaped curve of a normal distribution.

Poetry in Brownian motion

Brownian motion. In fact, share-price movements are more violent than that. Over the years the “quants” have found ways to cope with this—better ways to deal with, as it were, quirks in the prices of fruit and fruit salad. For a start, you can concentrate on the short-run volatility of prices, which in some ways tends to behave more like the Brownian motion that Black imagined. The quants can introduce sudden jumps or tweak their models to match actual share-price movements more closely. Mr Derman, now a professor at New York’s Columbia University and a partner at Prisma Capital Partners, a fund of hedge funds, did some of his best-known work modelling what is called the “volatility smile”—an anomaly in options markets that first appeared after the 1987 stockmarket crash when investors would pay extra for protection against another imminent fall in share prices. The fixes can make models complex and hard to compute that the markets leave it behind. The idea behind quantitative finance is to manage risk. You make money by taking known risks and hedging the truth or elegance, just a way of capturing the behaviour of a market and of linking an unobservable or illiquid price to prices in traded markets. The limit to these models is “how close to reality do they get?”

Almost as damaging is the belief that markets have made “value at risk” (VaR) calculations, a measure of the potential losses of a portfolio. This is supposed to show whether banks and other financial entities are being safely run. Regulators use VaR calculations to work out whether capital banks need to put aside for a rainy day. But the calculations are flawed. The mistake was to turn an isolated event into a known “tail risk”. Think of the tails of a range of possible daily losses and gains from any distribution. Most of the time you gain a little or lose a little. Occasionally you gain or lose a lot. Very rarely you win or lose a fortune. If you plot those daily investments on a graph, you get the familiar bell-shaped curve of a normal distribution. Most of the time you gain or lose a little. Occasionally you gain or lose a lot. Very rarely you win or lose a fortune. If you plot those daily investments on a graph, you get the familiar bell-shaped curve of a normal distribution.
Example:
Let R denote the return on our portfolio next month. We do not know what R will be. Let us assume we can describe what we think it will be by:

\[ R \sim N(0.01, 0.04^2) \]

What is the probability of a negative return?
In excel we use:

\[ =NORMDIST(0,0.01,0.04,TRUE) \]

And then the cell will be: 0.4013

\[ P(R<0) = F(0) = 0.4013 \]

What is the probability of a return between 0 and 0.05?

\[ =NORMDIST(0.05,0.01,0.04,TRUE) = .8413 \]

\[ P(0<R<0.05) = 0.84 - 0.4 = 0.44 \]

4. Introducing Expectation via Long Run Average

We have seen that one interpretation of probability is "long run frequency".

At right is the result of tossing a coin 5000 times.

After each toss we compute the fraction of heads so far.

Eventually, it settles down to 0.5.
We can interpret probability as the long run frequency from i.i.d. draws.

We can also interpret expectation (or expected value) as the long run average from i.i.d. draws.

Example: Tossing a pair of coins 10 times
Each time we record the number of heads.
1 0 2 1 0 1 2 0 2 0
Mean of $x = \frac{4(0) + 3(1) + 3(2))}{10} = 0.9$

Question: what is the sample mean of the number of heads?

Now suppose we toss the pair of coins 1000 times:

What is the sample mean?
Number of heads: 0 1 2
Frequencies : 241 507 252
Therefore, the sample mean is 1.011
What should the mean be?
Let $n_0, n_1, n_2$ be the number of 0's, 1's and 2's.
Then, the average would be
\[
\bar{x} = \frac{n_0 \times 0 + n_1 \times 1 + n_2 \times 2}{n}
\]
which is the same as
\[
\frac{n_0}{n} \times 0 + \frac{n_1}{n} \times 1 + \frac{n_2}{n} \times 2
\]
Now note that the values are i.i.d draws from the
TRUE PROBABILITY DISTRIBUTION.

<table>
<thead>
<tr>
<th>$\Pr(x)$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>0.50</td>
<td>1</td>
</tr>
<tr>
<td>0.25</td>
<td>2</td>
</tr>
</tbody>
</table>

So, for $n$ large, we should have
\[
\frac{n_0}{n} = 0.25 \quad \frac{n_1}{n} = 0.5 \quad \frac{n_2}{n} = 0.25
\]
Hence, the average should be about
\[
0.25(0) + 0.5(1) + 0.25(2) = 1.00
\]
but this is the expected value of the random variable $X$. 
The actual sample mean is:

\[0.241 \times 0 + 0.507 \times 1 + 0.252 \times 2 = 1.011\]

Hence, with a very, very, very, ..., large number of tosses we would expect the sample mean (the empirical mean of the numbers) to be very close to 1 (the expected value).

To summarize, we can think of the expected value, which in this case is equal to

\[p_X(0) \times 0 + p_X(1) \times 1 + p_X(2) \times 2 = 1\]

as the long run average (sample mean) of i.i.d draws.

---

Expected value and variance of a discrete r.v.

If \(X\) is a discrete random variable that takes values \(x_1, x_2, \ldots, x_n\)

Then, the Expected value of \(X\), or simply expectation of \(X\) is given by

\[E(X) = x_1 \Pr(X=x_1) + x_2 \Pr(X=x_2) + \ldots + x_n \Pr(X=x_n)\]

Similarly, the variance of \(X\) is given by

\[V(X) = (x_1 - E(X))^2 \Pr(X=x_1) + \ldots + (x_n - E(X))^2 \Pr(X=x_n)\]
Example:
Toss two coins over and over. As before, count number of heads.

average is 0.974.
average of \((x-1)^2\) is 0.51.

If \(X\) is number of heads from one toss of two coins:

\[
\text{Var}(X) = 0.25(0-1)^2 + 0.5(1-1)^2 + 0.25(2-1)^2 = 0.5
\]

Thus, for "large samples" the quantities we talked about for samples should be similar to the quantities we talked about for random variables:

\[
\text{Var}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
\[
= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]

If we really believe we are taking i.i.d. draws!!

5. Expected Value and Variance of Continuous RV's

If \(X\) is a continuous random variable with p.d.f. \(p(x)\) then

\[
E(X) = \int x p(x) \, dx
\]

The variance is

\[
\text{Var}(X) = E((X - \mu)^2) = \int (x - \mu)^2 p(x) \, dx
\]
If you know calculus that's fairly intuitive.
If you don't, it is completely incomprehensible.

**Good news:**

Intuitively, whether $X$ is discrete or continuous, we can always think of $E(X)$ as

$$E(X) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

for i.i.d. $X_i$ all having the same distribution as $X$.

Same for the variance.

---

**Example:**

500 i.i.d. draws from $N(0,1)$.

What is $E(Z)$?

Not so obvious:

Var($Z$)=1.

---

**For $Z \sim N(0,1)$**

Expectation: $E(Z)=0$

Variance: $\text{Var}(Z)=1$
Example: Modeling the number of heads (X) when tossing 5 fair coins.

TRUE MODEL

<table>
<thead>
<tr>
<th>X</th>
<th>Pr(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.03125</td>
</tr>
<tr>
<td>1</td>
<td>0.15625</td>
</tr>
<tr>
<td>2</td>
<td>0.31250</td>
</tr>
<tr>
<td>3</td>
<td>0.31250</td>
</tr>
<tr>
<td>4</td>
<td>0.15625</td>
</tr>
<tr>
<td>5</td>
<td>0.03125</td>
</tr>
</tbody>
</table>

I asked 150 students to toss 5 coins and counting the number of heads.

2 3 3 2 1 1 2 4 4 2 2 2 3 2 2
4 3 3 4 2 3 1 1 2 3 3 3 3 2
1 1 3 1 2 3 3 4 2 3 3 1 2 2 4
2 2 2 2 4 4 2 3 3 2 4 2 0 3
3 3 3 2 2 3 3 2 3 4 2 3 3 2
3 3 2 0 1 4 3 1 2 4 2 2 2 2 2
2 1 3 2 2 3 3 4 2 2 2 3 3 2
1 2 4 2 1 2 2 3 3 1 2 3 2
1 2 3 3 5 1 1 4 4 3 3 2 4 1 3
2 3 2 3 2 1 2 1 1 3 2 2 3 2 3

Sample mean = 2.42667
Sample standard deviation = 0.936556

Observed      True model
x   Frequency   P(x)     Pr(x)
0   0.01333     0.03125
1   0.13333     0.15625
2   0.40000     0.31250
3   0.32667     0.31250
4   0.12000     0.15625
5   0.00667     0.03125

Sample mean = 2.42667
Sample standard deviation = 0.93656
True mean = 2.5
True standard deviation = 1.12
Example: Modeling the number of heads (X) when tossing 10 fair coins.

I asked the same 150 students to toss 10 coins and counting the number of heads.

6 8 6 4 5 4 4 8 5 5 6 4 5 3
6 4 4 4 4 4 7 7 7 9 5 6 7 2 5
3 3 4 6 2 3 3 6 6 3 3 4 5 6
7 6 6 8 5 6 3 6 5 6 6 6 8 3 5
3 4 7 6 5 7 6 7 6 8 5 7 4 6 7
3 4 7 8 5 4 5 4 5 3 6 4 7 4 7
7 6 5 3 8 4 5 4 6 5 6 6 9 7 7
7 6 6 3 6 8 2 8 4 5 5 6 5 7 4
6 7 7 6 5 6 4 3 6 5 4 5 6 6 5
5 4 5 4 4 7 5 4 5 1 4 6 6 8 3

Estimated mean = 5.287
Estimated standard deviation = 1.586

True mean = 5.0 (Estimated = 5.287)
True standard deviation = 1.581 (Estimated = 1.586)
6. Random Variables and Formulas

We use mathematical formulas to express relationships between variables.

Even though a random variable is not a variable in the usual sense, we can still use formulas to express relationships.

We will develop formulas for linear combinations of random variables that are analogous to the ones we had for samples!

Example: A contractor estimates the probabilities for the time (in number of days) required to complete a certain type of job as follows:

<table>
<thead>
<tr>
<th>t</th>
<th>Pr(T=t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>0.30</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Let T denote the time.

Review question: what is the probability that a project will take less than 3 days to complete?

The longer it takes to complete the job, the greater the cost.

There is a fixed cost of 20000 and an additional 2000 per day.

Let C denote the cost.

Then,

\[ C = 20000 + 2000T \]

Before the project is completed, both T and C are unknown and hence random variables.

Whatever T and C turn out to be, they will satisfy the equation.
Mean and Variance of a Linear Function

Let $Y$ and $X$ be random variables such that

$$Y = c_0 + c_1X$$

Then,

$$E(Y) = c_0 + c_1E(X)$$
$$\text{Var}(Y) = c_1^2 \text{Var}(X)$$

$$\sigma_Y = |c_1| \sigma_X$$

These formulas mirror what we had for sample means and variances.

Intuitively, we get the same sort of result because the quantities for RV's can be thought of as long run averages.

The intuition and rules are the same for continuous and discrete RV's !!!

**Careful !!**

*While we have stressed the analogies, the mean (expectation) of an RV is not the same thing as the mean of a sample.*

---

Example (cont.):

Recall our time to project completion example.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$p(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>0.30</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
</tr>
</tbody>
</table>

The expected value of time is $E(T) = 3.2$.

The variance of time is $\text{Var}(T) = 1.06$.

$C = 20000 + 2000T$

Since $C$ is a linear function of $T$, we easily get its mean and variance from those of $T$:

$E(C) = 20000 + 2000E(T) = 20000+2000(3.2) = 26,400$

$\text{Var}(C) = 2000^2 \text{Var}(T) = 4,240,000$

$s_C = \sqrt{424000} = 2000 \times \sqrt{1.06} = 2,059$
An important example: the non-standard normal

Suppose $Z \sim N(0,1)$
If $X = \mu + \sigma Z$, then it can be shown that $X \sim N(\mu, \sigma^2)$
\[
E(X) = \mu + \sigma E(Z) = \mu.
\]
\[
Var(X) = \sigma^2 \text{Var}(Z) = \sigma^2.
\]
For $X \sim N(\mu, \sigma^2)$
\[
E(X) = \mu, \quad \text{Var}(X) = \sigma^2
\]

7. Covariance/correlation for pairs of random variables

Suppose we have a pair of random variables $(X,Y)$.
Also, suppose we know what their bivariate probability distribution is.
A meaningful question to ask is: are $X$ and $Y$ related (independent)?
In this subsection we will define the covariance and correlation between two random variables to summarize their linear relationship.

For discrete random variables we have a (relatively) simple formula:
The covariance between bivariate discrete random variables $X$ and $Y$ is given by:
\[
\text{cov}(X,Y) = \sigma_{XY} = \sum_{x,y} p(x,y)(x - \mu_x)(y - \mu_y)
\]
Example:

\[
\begin{array}{c|cc}
\mu_X &= 0.1 & \mu_Y = 0.1 \\
\sigma_X &= 0.05 & \sigma_Y = 0.05 \\
X & 0.05 & 0.15 \\
Y & 0.05 & 0.10 & 0.15 & 0.40 \\
\end{array}
\]

cov(X,Y) = \sigma_{XY} = 0.4(0.05-0.1)(0.05-0.1) + 0.1(0.05-0.1)(0.15-0.1) + 0.1(0.15-0.1)(0.05-0.1) + 0.4(0.15-0.1)(0.15-0.1) = 0.0015

Intuition: we have an 80% chance that X and Y are both above the mean or both below the mean together.

The correlation between random variables (discrete or continuous) is

\[
\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}
\]

\(\rho\): the basic facts

\[-1 \leq \rho \leq 1\]

If \(\rho\) is close to 1, then it means there is a line, with positive slope, such that (X,Y) is likely to fall close to it.

If \(\rho\) is close to -1, same thing, but the line has a negative slope.
The correlation is:

\[ \rho_{XY} = \frac{0.015/((.05 \cdot .05))}{0.6} \]

**Example (cont.):**

<table>
<thead>
<tr>
<th>X</th>
<th>.05</th>
<th>.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>.05</td>
<td>.4</td>
</tr>
</tbody>
</table>

The correlation is:

\[ \rho_{XY} = \frac{0.015/(.05 \cdot .05)}{0.6} = 0.6 \]

**Example:**

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.25</td>
</tr>
</tbody>
</table>

Let us compute the covariance:

\[ .25(-.5)(-.5) + .25(-.5)(.5) + .25(.5)(-.5) + .25(.5)(.5) = 0 \]

The covariance is 0 and so is the correlation: not surprising, right?

**For continuous random variables:**

\[ \text{Cov}(X,Y) = \int \int (x - \mu_x)(y - \mu_y) f(x,y) \, dx \, dy \]

Or, the long run average:

\[ \sigma_{xy} = \text{cov}(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_x)(Y_i - \mu_y) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \]

(for large \(n\))

where \( (X,Y) = 1, 2, 3, \ldots, n \) are a large number of i.i.d draws from the bivariate distribution of \((X,Y)\).

As earlier in the case of the expected value and variance, the theoretical covariance can be interpreted as the long run sample covariance.
8. Independence and correlation

Suppose two random variables are independent.
That means they have nothing to do with each other.
That means they have nothing to do with each other linearly.
That means the correlation is 0.

\[ \text{cov}(X, Y) = 0 \]

The converse is not necessarily true.

\[ \text{cov}(X, Y) = 0 \]

DOES NOT necessarily mean they are independent.

Example: Zero correlation DOES NOT imply independence

\[
\begin{array}{c|ccc|c}
X & -1 & 0 & 1 & \text{p}_y(Y) \\
\hline
P(X=0,Y=0)=0 & 0.10 & 0.15 & 0.10 & 0.35 \\
P(X=0):P(Y=0)=0.09 & 0.00 & 0.15 & 0.10 & 0.30 \\
& 0.10 & 0.15 & 0.35 & 0.35 \\
Y & -1 & 0 & 1 & \text{p}_x(X) \\
\hline
\text{p}_x(X) & 0.35 & 0.30 & 0.35 & 1.00 \\
\end{array}
\]

\[ \text{COV}(X,Y) = (-1)(-1)(0.1)+(-1)(1)(0.1)+(1)(-1)(0.1)+(1)(1)(0.1)+0, \]
so \( X \) and \( Y \) are uncorrelated.

INDEPENDENCE IS STRONGER THAN UNCORRELATION

Statistical inference

0. I.I.D. draws from the normal distribution
1. The Binomial Distribution
2. The Central Limit Theorem
3. Estimating \( p \), population and sample values
4. The sampling distribution of the estimator
5. Confidence interval for \( p \)

BOOK:
Point estimates and confidence intervals  (283–296 (12), 294–308 (13))
A confidence interval for a proportion  (297–298 (12), 309–312 (13))
0. **I.I.D Draws from the Normal Distribution**

We want to use the normal distribution to model data in the real world.

Surprisingly often, data looks like i.i.d. draws from a normal distribution.

---

**Note:** We can have i.i.d. draws from any distribution.

By writing

\[ X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2) \text{ i.i.d.} \]

we mean that each random variable \( X \) will be an independent draw from the same normal distribution.

We have not formally defined independence for continuous distributions, but our intuition is the same as before!

Each draw is has no effect on the others, and the same normal distribution describes what we think each variable will turn out to be.

---

**What do i.i.d. normal draws look like?**

We can simulate i.i.d draws from the normal distribution.

Here are 300 "draws" simulated from the standard normal distribution.

There is no pattern, they look "random"
Same with lines drawn in at $\mu=0$, +/- 1 and +/- 2.

In the long run, 95% will be between +2 and −2.

*Do you remember the empirical rule?*

---

**Draws from a normal other than the standard one.**

These are 200 i.i.d. draws from $\mathcal{N}(5,4)$, i.e. a normal distribution with mean 5, variance 4 and, therefore, standard deviation 2.

---

Here is the histogram of 5000 draws from the standard normal:

The height of each bar tells us the number of observations in the interval.

All the intervals have the same width.
For a large number of iid draws, the observed percent in an interval should be close to the probability.

For the density the area is the probability of the interval.

For the histogram the area is the observed percent in the interval.

In large samples these are close. See next page.

The histogram of a "large" number of i.i.d draws from any distribution should look like the p.d.f.

We look, once again, at the Canadian returns data. We have monthly returns from Feb '88 to Dec '96.

No apparent pattern!
Conclude: The returns look like i.i.d. normal draws!

Example: non-normal data

Not all data looks normal…

Daily volume of trades in the Cattle pit.

Skewed to the right.

Example: dependent data

…and not all time series look independent.

Dow Jones

Lake Level

Beer Production
1. The Binomial Distribution

Suppose you are about to make three parts. The parts are iid Bernoulli(p), where 1 means a good part and 0 means a defective.

Let $X_i$ denote the outcome for part $i$, $i=1,2,3$.

$X_1, X_2, X_3 \sim \text{Bernoulli}(p)$ iid.

How many parts will be good?

Let $Y$ denote the number of good parts.

$Y = X_1 + X_2 + X_3$

What is the distribution of $Y$?

Suppose we make $n$ parts, so $X_i \sim \text{Bernoulli}(p)$, i.i.d.

Let $Y = X_1 + X_2 + \cdots + X_n$ and $Z = \frac{Y}{n}$

$Y$: number of successes

$Z$: proportion of successes

Then

$E(Y) = np$ and $\text{Var}(Y) = np(1-p)$

$E(Z) = p$ and $\text{Var}(Z) = \frac{p(1-p)}{n}$
It can be shown that the probability distribution of \( Y \) is
\[
P(Y = y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \quad y = 0, 1, 2, \ldots, n
\]

\( n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 \)

\( n \) “trials” each of which results in a success or a failure.

Each trial is independent of the others.

On each trial we have the same chance \( p \) of “success”.

**The number of successes is Binomial\((n,p)\)**
- \( n \): number of trials.
- \( p \): probability of success on each trial.

**Example:** Below we plotted \( y \) vs \( p(y) \) for the binomial with \( n=10 \) and \( p=0.1, 0.2, 0.5, 0.95 \). The \( p=0.5 \) distribution looks symmetric, while the others are skewed.

**Example:**

Suppose the next 20 returns on an asset are modeled as i.i.d.
\[ X_1, \ldots, X_{20} \sim N(0.1, 0.01). \]

Let \( S \) denote the number of positive returns out of the next 20. What is the mean and variance of \( S \)?

**Solution:**

Probability of success \( p = \Pr(X>0) = 0.8413 \)

Therefore, \( S \sim \text{Binomial}(20, 0.8413) \)

\[ E(S) = 20 \cdot 0.8413 = 16.826 \]

\[ V(S) = 20 \cdot 0.8413 \cdot 0.1587 = 2.6703 \]

\( \text{Stdev. } S = 1.6341 \)

---

106
Notes:

The term success refers to what is being counted.

For example if the probability of a defect is 0.1, then the number of defects in a sample of size \( n \) is Binomial\( (n, 0.1) \), where a success means a defect.

If we count good ones, then it is Binomial\( (n, 0.9) \).

In terms of the underlying Bernoulli, you can make either of the two possible outcomes correspond to 1 (and the other to 0).

Bernoulli\( (p) \) is the same as Binomial\( (1, p) \).

Two easy special cases are:

\[
P(Y = n) = p^n \\
P(Y = 0) = (1 - p)^n
\]

Example:

Suppose the probability of a defect is 0.01 and you make 100 parts.

What the probability they are all good?

\((0.99)^{100} = 0.366\)

2. The Central Limit Theorem

Example: Suppose you are repeatedly making a part and 1 means defective, 0 else.

Let \( X_i \) corresponds to the \( i \)-th part.

Assume the model \( X_i \sim \text{Bernoulli}(p) \) i.i.d.

Suppose you are about to make \( n \) parts and are interested in

\[ Y = X_1 + X_2 + \ldots + X_n \]

the total number of defective parts out of the \( n \).
What is the distribution of Y?

It is a $Y \sim \text{Binomial}(n,p)$, but we already knew that!

There is a probability result (the central limit theorem) that says that we can get a simple approximate answer by using a normal with the mean equal to the mean of $Y$ and variance equal to the variance of $Y$.

We already know that $E(Y)=np$ and $V(Y)=np(1-p)$. Therefore,

$$Y \sim N(np,np(1-p)) \text{ approximately}$$

The bigger $n$ is, the better the normal approximation to the binomial.

Example: From the pictures it can be seen that as we increase the number of random variables $n$, the distribution of $Y$ gets closer and closer to a normal distribution with the same mean and variance as the binomial.

Example: Suppose defects are i.i.d. Bernoulli(0.1). You are about to make 100 parts.

We know that number of defects, $Y$, is $\text{Binomial}(100,0.1)$

Let us use the normal approximation, first.

$$E(Y) = np = 100 \times 0.1 = 10$$

$$V(Y) = np(1-p) = 100 \times 0.1 \times 0.9 = 9$$

$Y$ is approximately $N(10,3^2)$

Based on the normal approximation, there is a 95% chance that the number of defects is in the interval:

$$10 \pm 6 = [4,16]$$
**Exact** answer based on the true binomial probabilities.

**BINOMDIST**(number_s, trials, probability_s, cumulative)

*Number_s* is the number of successes in trials.

*Trials* is the number of independent trials.

*Probability_s* is the probability of success on each trial.

*Cumulative* if TRUE, then BINOMDIST returns c.d.f.; if FALSE, then BINOMDIST returns the p.d.f.

\[
P(4 \leq Y \leq 16) = P(Y = 4) + P(Y = 5) + \cdots + P(Y = 15) + P(Y = 16) \\
= 0.015875 + 0.033866 + \cdots + 0.032682 + 0.019292 \\
= 0.971565
\]

From just the mean and the standard deviation (using the central limit theorem and the normal approximation) we get a pretty good idea of what is likely to happen.

In general, if the distribution looks roughly normal shaped, you can try to approximate it with a normal curve having the same mean and variance.

---


**Example:**

Front page of Chicago Tribune, 1/14/2004:

"700 likely Illinois voters in the November general election were polled".

"48% would not like to see Bush re-elected."

"The survey has an error margin of 4 percentage points among general election voters.."

**What do these figures mean?**

---

Suppose we have a large population of voters. Each will vote either democratic or republican.

We would like to know the proportion that will vote democratic.

Doesn’t this scenario seem appropriate to the famous **Bernoulli model**?

**We can’t ask them all. Too costly!**

If we ask a sample of **some** of them, how much do we know about all of them?
We will take a random sample of voters and use the *sample proportion* of democrats as a guess or estimate of the *true proportion* in the whole population. The sample proportion is called an *estimator*.

The resulting (after we have the sample) actual value or guess is called the *estimate*.

\[ p : \text{proportion of democrats in the population.} \]
\[ \hat{p} : \text{proportion of democrats in the sample.} \]

Before we take the sample \( \hat{p} \) is a random variable.

We wonder how close \( \hat{p} \) will be to \( p \).

After we take the sample, the resulting sample proportion \( \hat{p} \) is just a number, it is just our estimate of \( p \).

4. The Sampling Distribution of the Estimator

Well, we have our plan.

*What are our chances?*

After we have our sample we are either close or not.

Before we have the sample we can think about what the properties of our estimator are.

*How wrong could we be?*
To get a feeling for the properties of our estimator, we see what it will do given a value for $p$.

Of course, the whole point is that we don't know $p$, but we can understand what we are doing by asking "given a value for $p$, how would we do?".

What if $p=0.5$ and $n=700$, then what kind of estimate could I get?

Conjecture: If I knew $p=0.5$ and $n=700$, then I would be surprised (even be willing to bet against!) if there were less than 300 or more than 500 successes!

Given $p$, we know the distribution of our estimator. Let $X_i = 1$ if the $i^{th}$ sampled voter is a dem, 0 if repub. Let $Y$ denote the number of democrats in the sample.

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{Y}{n} \sim B(n,p)$$

This is called the sampling distribution of the estimator.

Remember: We know the distribution of $\hat{p}$ because we are taking a random sample from a large population of size $N$, where $N$ is much, much, much larger than the sample size $n$, ie. $n << N$.

Don't confuse the probability distribution of $\hat{p}$ with how the 1's and 0's are "distributed" in the population.

The distribution of 1's and 0's in the population is summarized by the unknown proportion $p$.

Notice that the probability distribution of $\hat{p}$ when $n=100$, for instance, is not the same as the probability distribution of $\hat{p}$ when $n=1000$. 
We can compute the mean and variance of our estimator to summarize its properties:

\[
E(\hat{p}) = E\left( \frac{Y}{n} \right) = \frac{1}{n} E(Y) = \frac{np}{n} = p
\]

The estimator is unbiased.

Our estimate can turn out to be too big or too small, but it has no tendency to be wrong.

**Question**

Suppose instead of asking 700 randomly chosen people, you asked 700 friends.

Would the proportion of democratic voters in that sample be an unbiased estimate of the population proportion?

Not too useful by itself.

But we can combine it with the central limit theorem to get:

\[
\text{Var}(\hat{p}) = \text{Var}\left( \frac{Y}{n} \right) = \frac{1}{n^2} \text{Var}(Y) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}
\]
A simple approximate sampling distribution is:

\[ \hat{p} \sim N \left( p, \frac{p(1-p)}{n} \right) \]

This gives us a simple way of thinking about what kind of estimate our estimator is likely to give us!!

Example: Suppose we have a coin and we are not certain whether the coin is fair. We run an experiment: each one of us (me + 149 students) flip the coin 10 times and record the proportion of 1’s (1=head and 0=tail).

The 150 proportions are:

0.4 0.3 0.2 0.4 0.3 0.2 0.4 0.3 0.2 0.3 0.5 0.3 0.4 0.5 0.5 0.3 0.2 0.3 0.4 0.5 0.2 0.5 0.3 0.3 0.4 0.5 0.3 0.5 0.2 0.5 0.3 0.4 0.6 0.3 0.4 0.3 0.4
Let us suppose that now me and 1499 students toss the coin 10 times each.

Let us suppose that now the same 1500 persons toss the coin 100 times each.

Information accumulation: None of the 1500 persons obtained less than 20 or more than 53 heads when tossing the coin 100 times.

The true proportion of heads is \textbf{35\%}!
Since the true proportion of heads is $p = 0.35$, we can check how good the normal approximation is.

$$\hat{p} \sim N \left( 0.35, \frac{0.35(1 - 0.35)}{100} \right) = N(0.35, 0.047697^2)$$

The approximate probabilities (under normality) are

- $Pr(20 \text{ heads or less}) = 0.08308472\%$
- $Pr(53 \text{ heads or more}) = 0.008038164\%$

The true probabilities are

- $Pr(20 \text{ heads or less}) = 0.07836153\%$
- $Pr(53 \text{ heads or more}) = 0.007757356\%$

Example:

Let $p = 0.6$ and $n = 200$.

Then $\mu = np = 0.6 \times 200 = 120$, $\sigma = \sqrt{np(1-p)} = \sqrt{200 \times 0.6 \times 0.4} = 7.746$.

The normal curve tells us what kinds of estimates we could get if we about to take a sample of size $n=200$ and the true population $p = 0.6$. 
Notice that the bigger \( n \) is, the better our chances are!!

In general this is what we expect \( \hat{p} \) to be like:

\[
\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)
\]

\[
\frac{\hat{p} + p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)
\]

Example (cont.)

Sample size: \( n=100 \)

True proportion: \( p=0.35 \)

Estimated proportion: \( \hat{p} \sim \mathcal{N}(0.35,0.002275) \)

The approximate 95% probability interval for \( \hat{p} \) is \((0.35-2*0.047697 ; 0.35+2*0.047697) = (0.255; 0.445)\).

Example (cont.)

Sample size: \( n=200 \)

True proportion: \( p=0.6 \)

Estimated proportion: \( \hat{p} \sim \mathcal{N}(0.6,0.0012) \)

The approximate 95% probability interval for \( \hat{p} \) is \((0.6-2*0.0346 ; 0.6+2*0.0346) = (0.531; 0.669)\).

5. Confidence Interval for \( p \)

Well, that’s all very well, but we still don’t have an answer to our real question:

Given the data, how do we feel about \( p \)?

The confidence interval is the classic solution.

It builds directly on all that we have done.
Confidence Interval for \( p \):

**How different is our estimate from \( p \)?**

\[
P( p - 2 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < \hat{p} < p + 2 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}) = 0.95
\]

\[\hat{p} = p \pm 2 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\]

Since we don’t know \( p \), we just plug in the estimate for the standard deviation. *This is wrong, but we hope not too wrong!*

The difference between the sample and population proportions is approximately:

\[
2 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

Example (cont.):

Front page of Chicago Tribune, 1/14/2004:

“700 likely Illinois voters in the November general election were polled”. \(\hat{p} = 0.48\) (abuse of notation!)

“48% would not like to see Bush re-elected.”

“The survey has an error margin of four percentage points among general election voters...”

\[
2 \sqrt{\frac{0.48 * (1 - 0.48)}{700}} = 0.038
\]
So the difference between our estimate of 0.48 and the unknown true value is about 0.038.

The 95% confidence interval for the true \( p \) is

\[
0.48 \pm 0.038
\]

“estimate +/- error”

Interval: \((0.442 ; 0.518)\)

Is that a big interval?

If the election is tomorrow and we want to know the winner it is big.

If the election is three months away and last month Bush was at 70% approval then the interval is small enough to tell us things have really changed.

Do our estimates of \( p \) always pan out?

**Example:** Leading up to a democratic primary in Wisconsin, a poll of 600 showed Kerry with 53\% ± 4\% and Edwards with 16\% ± 4\%. The actual results a few days later were Kerry 40\% and Edwards 34\%.

**Example:** Results are based on telephone interviews with 1,002 national adults, aged 18 and older, conducted Feb. 9-12, 2004. For results based on the total sample of national adults, one can say with 95\% confidence that the margin of sampling error is ±3 percentage points.

In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.

In practice, getting a random sample, or, more generally, a sample that is not biased towards some particular subset, can be tough!!
Example: Dowjones (6/18/1929 to 2/6/2009)

A total of 20100 days.

Below are the proportions of positive returns for consecutive samples of size 50 days, so 402 samples.

0.50 0.60 0.38 0.56 0.53 0.38 0.48 0.44 0.38 0.46 0.42 0.38 0.36 0.46 0.38 0.36 0.46 0.58 0.46 0.48 0.44 0.62

0.50 0.52 0.50 0.54 0.44 0.44 0.54 0.58 0.56 0.58 0.64 0.66 0.50 0.56 0.54 0.58 0.60 0.52 0.54 0.52 0.52 0.42

0.46 0.42 0.52 0.46 0.38 0.52 0.46 0.48 0.50 0.50 0.62 0.56 0.56 0.46 0.46 0.60 0.58 0.62 0.46 0.40 0.44 0.46

A total of 10431 days (out of 20100) with positive returns.

Therefore, \( \hat{p} = 0.5189552 \).

The difference between the sample and population proportions is approximately

\[ 2\sqrt{\hat{p}(1-\hat{p})/50} \approx 0.1413197. \]

The approximate 95% confidence interval for the true \( p \) is

\[ (0.5189552 - 0.1413197; 0.5189552 + 0.1413197) \]

or

\[ (0.3776;0.6603). \]

Note: There are several hidden and strong assumptions here.

1. The proportionality of positive returns is i.i.d. (Independent and Identically Distributed). Later in this class we will see statistically whether a sequence of measurements is i.i.d.

2. We use the term standard error to denote the estimate of a standard deviation.

Before you get the sample, you have an (approximate) 95% chance the true value will be in the confidence interval. After you get the data and compute the interval it is either in there or not.

We call the interval a “confidence interval” rather than a probability interval to emphasize this difference.

The “root n” in the formula precisely captures the fact that with larger samples we know more!!
Question: How much do I know about the parameter?

Answer: Confidence interval small: I know a lot. Confidence interval big: I know little.

Example: Suppose $\hat{p} = 0.2$ and $n=100$.
Standard error: $s.e. = 0.04$

Suppose $\hat{p} = 0.2$ and $n=10,000$. (n went up by a factor of 100)
Standard error: $s.e. = 0.004$ (s.e. went down by 1/10)

If I want to half the s.e., I have to increase the sample size by a factor of 4!

This is the “the tragedy of root n”.

Example: How many observations should you collect to guarantee that, on average, the difference between the true $p$ and the estimated $\hat{p}$, namely $\hat{p}$, is less than 0.01?

What you want is to find $n$ such that

$$2\sqrt{p(1-p)/n} < 0.01$$

or

$$n > 40000\cdot p(1-p).$$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3600</td>
</tr>
<tr>
<td>0.3</td>
<td>8400</td>
</tr>
</tbody>
</table>
| 0.5  | 10000| **<=** A conservative decision maker would probably choose $n$ around 10000
| 0.6  | 9600 |
| 0.8  | 6400 |
If now you wanted the different between \( p \) and \( \hat{p} \) to be, on average, less than 0.04 (like in example 1)? Again, you want to find \( n \) such that \( 2\sqrt{p(1-p)/n} < 0.04 \) or \( n > 2500p(1-p) \).

<table>
<thead>
<tr>
<th>( p )</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>225</td>
<td>525</td>
<td>625</td>
<td>600</td>
<td>400</td>
</tr>
</tbody>
</table>

### Hypothesis testing

1. Hypothesis testing
2. P-values.
3. Confidence intervals, tests, and p-values in general.

---

**Example:** Suppose we have an important manufacturing process. The manager claims that the defect rate is 10%.

What does this mean?

If defects are i.i.d. Bernoulli with \( p = 0.1 \), then in the long run we will have 10% defective.

We want to test the claim or hypothesis that \( p = 0.1 \).
Experiment 1:
Suppose we make 5 parts and 1 of the parts is defective.
The estimated defect rate is 0.2.
What does that tell us about $p=0.1$?

Experiment 2:
Suppose we make 20 parts and 4 of the parts is defective.
The estimated defect rate is 0.2.
What does that tell us about $p=0.1$?

Experiment 3:
Suppose we make 1000 parts and 200 parts are defective.
The estimated defect rate is 0.2.
What does that tell us about $p=0.1$?

Experiment 1:
If we get 1 out of 5, then we have 20% defective.
This is \textit{highly probable} if $p=0.1$.
In fact, the chance of 1 out of 5 is 32.8% when $p=0.1$.
So, it \textit{seems hard to reject the claim}.

Experiment 2:
If we get 4 out of 20, then we have 20% defective.
That is \textit{somewhat likely} if $p=0.1$.
In fact, the chance of 4 out of 20 is 8.98% when $p=0.1$.
So, it \textit{seems hard to reject the claim}.

Experiment 3:
If we get 200 out of 1000, then we have 20% defective.
That is \textit{highly unlikely} if $p=0.1$.
In fact, the chance of 150 or more out of 1000 is negligible.
So, we are \textit{likely to reject the claim}.
Under the hypothesis that $p=0.1$, the data is

Experiment 1: Highly probable $\Rightarrow$ 32.80%
Experiment 2: Somewhat likely $\Rightarrow$ 8.98%
Experiment 3: Very unlikely $\Rightarrow$ 0.00%

Basic Intuition (and strategy)

If the outcome of an experiment is very unlikely under the tested hypothesis, then the data provides evidence to reject the hypothesis.

Clearly, we have to trust the data!

Now that we have the intuition, let us be more formal

Example: Suppose we have the data below where $n=100$ and 18% are defective. We want to test whether $p=0.1$?
The question that we are interested in is:
Can we get $\hat{p} = 0.18$ if $p = 0.1$?

To put it differently:
Is it possible to obtain 18% defects out of 100 observations, if the true defect rate is 10%.

Or, again, is the difference between 18% and 10% so big that it could not happen just "by chance"?

The flip side of the coin:
If $p=0.1$, what kind of value can we expect for $\hat{p}$?

Recall that, under the hypothesis that $p=0.1$, it follows that

\[
\hat{p} = N\left( p, \frac{p(1-p)}{n} \right) = N\left( 0.1, \frac{0.1(1-0.1)}{100} \right) = N(0.1, 0.03^2)
\]

If $p=0.1$, then the possible values of $\hat{p}$ will be (approximately) normal with mean 0.1 and variance 0.03^2.

If $p=0.1$, then

$\hat{p} = N(0.1, 0.03^2)$

There is a very small probability of getting a value as big as 0.18 (which is what we obtain from our specific sample).

It is very unlikely to obtain a value that big given that $p=0.1$.

Since we trust what we see (the estimated value from the data) we infer that a distribution with $p=0.1$ is not likely to be the generating one.

We should probably reject the claim.
It is easy to see that 0.18 is roughly 2.7 standard deviations to the right of 0.1:

\[
\frac{0.18 - 0.1}{0.03} = 2.67
\]

In other words, obtaining 0.18 from a normal distribution with mean 0.1 and variance 0.03² is the same as obtaining 2.67 from a normal distribution with mean 0 and variance 1 (the standard normal).

2.67 is pretty unlikely. It is reasonable to reject the claim.

Basic Logic:

If the null hypothesis \( p = p^0 \) is true then,

\[
\frac{\hat{p} - p^0}{\sqrt{p^0(1 - p^0)/n}}
\]

should look like a draw from the standard normal distribution!!

We have outlined the main intuition of what we do. But we really want to be precise. We now describe a precise rule to assess (test) the validity of a hypothesis.

To test the null hypothesis

\[ H_0: p = p^0 \]

against the alternative

\[ H_a: p \neq p^0 \]

We reject \( H_0 \) at the 5% level if

\[
\frac{\hat{p} - p^0}{\sqrt{p^0(1 - p^0)/n}} > 2
\]

Otherwise, we fail to reject \( H_0 \).
**Note (1)**

The quantity \( \frac{\hat{p} - p^*}{\sqrt{\frac{p^*(1 - p^*)}{n}}} \) is called the **test statistic**.

The numerator is simply the difference between the estimated \( \hat{p} \) and the conjectured \( p^* \).

We are truly comparing the **estimated** \( p \) and the **conjectured** \( p^* \) taking statistical uncertainty into account.

When the **estimated** \( p \) is more than 2 standard errors away from the **conjectured** \( p^* \), then we reject the null hypothesis at the 5% level.

**Note (2)**

If we do not reject, we **do not say** that we accept.

We say that we **fail to reject**.

This is because if we do not reject we have not proven that the null is true, we just **do not have enough evidence to reject it**.

**Note (3)**

The **level** has the interpretation:

\[ \Pr(\text{reject } H_0 \mid H_0 \text{ is true}) = 0.05 \]

**Decision rule:**

Reject \( H_0 \) whenever the **test statistics** is bigger than 2 or smaller than \(-2\).

If \( H_0 \) is **true**, then 5% of the time, on average, the above decision rule will be a mistake.
Example: Let us check the claim that \( H_0: \) the daily closing price of GE in 2008 is just as likely to go up as down.

Model: Assume that, day to day, it is i.i.d Bernoulli (\( p \)) whether the price of GE goes up or not. Record a 1 if it goes down and a 0 if it goes up. Then, \( p \) is the probability that the stock goes down. **We want to test** \( H_0: p = 0.5. \)

Data summary:
It went down 133 days out of 252 days.
It went up 119 days out of 252 days.

The estimated \( p \) is \( \frac{133}{252} = 0.52778 \)

The test statistic is
\[
\frac{\hat{p} - p^*}{\sqrt{\frac{p^*(1-p^*)}{n}}} = \frac{0.52778 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{252}}} = 0.8819876
\]

Since 0.8819876 is in the interval (-2,2), we **DO NOT** have strong evidence to reject \( H_0. \) **We fail to reject** \( H_0. \)

2. p-values

Example: Suppose that an i.i.d. sample of size \( n = 100 \) is taken from a Bernoulli(\( p \)) model, for some unknown value \( p \) (just like with the previous GE example). **We want to test** \( H_0: p = 0.2. \)

Case I: Suppose the data produces \( \hat{p} = 0.278. \)
Test statistic: \( \frac{0.278-0.2}{\sqrt{0.2*0.8/100}} = 1.95. \)

Case II: Suppose the data produces \( \hat{p} = 0.282. \)
Test statistic: \( \frac{0.282-0.2}{\sqrt{0.2*0.8/100}} = 2.05. \)

**Not very interesting decision rule:**
Failing to reject \( H_0 \) in Case I and Rejecting \( H_0 \) in Case II.
The **evidence** is only a little different, but we act **totally differently!!**
Remember our basic idea: Reject if what we see is unlikely given the hypothesis.

The standard normal tells us what kind of test statistic we should get if the null hypothesis is true.

The farther out in the tail the test statistic is, the more we want to reject!!

Rather than picking a cutoff, the p-value measures how far out in the tail the test stat is.

Null hypothesis $H_0: \hat{p} = p^0$

$$\text{test statistic} = \frac{\hat{p} - p^0}{\sqrt{p^0(1-p^0)/n}}$$

The $p$-value for $H_0$ is defined as

$$p\text{-value} = 1 - P(Z<|\text{test statistic}|)$$

where $Z \sim N(0,1)$.

$p$-value is the probability of getting a test statistic as far out or farther than the one we got.

Example:

Suppose the test statistic = 1. What is the $p$-value?

Suppose the test statistic = 2. What is the $p$-value?

Suppose the test statistic = 3. What is the $p$-value?

Suppose the test statistic = 4. What is the $p$-value?
Suppose the test statistic = 1. What is the p-value? 
0.3173105

Suppose the test statistic = 2. What is the p-value? 
0.04550026

Suppose the test statistic = 3. What is the p-value? 
0.002699796

Suppose the test statistic = 4. What is the p-value? 
0.00006334248

Here is a table of test statistics and p-values.

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.000000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.677075</td>
</tr>
<tr>
<td>1.0</td>
<td>0.317310</td>
</tr>
<tr>
<td>1.5</td>
<td>0.136414</td>
</tr>
<tr>
<td>2.0</td>
<td>0.045500</td>
</tr>
<tr>
<td>2.5</td>
<td>0.012419</td>
</tr>
<tr>
<td>3.0</td>
<td>0.002700</td>
</tr>
<tr>
<td>3.5</td>
<td>0.000465</td>
</tr>
<tr>
<td>4.0</td>
<td>0.000053</td>
</tr>
<tr>
<td>4.5</td>
<td>0.000077</td>
</tr>
<tr>
<td>5.0</td>
<td>0.000021</td>
</tr>
<tr>
<td>5.5</td>
<td>0.000003</td>
</tr>
<tr>
<td>6.0</td>
<td>0.000000</td>
</tr>
<tr>
<td>6.5</td>
<td>0.000000</td>
</tr>
<tr>
<td>7.0</td>
<td>0.000000</td>
</tr>
<tr>
<td>7.5</td>
<td>0.000000</td>
</tr>
<tr>
<td>8.0</td>
<td>0.000000</td>
</tr>
<tr>
<td>8.5</td>
<td>0.000000</td>
</tr>
<tr>
<td>9.0</td>
<td>0.000000</td>
</tr>
<tr>
<td>9.5</td>
<td>0.000000</td>
</tr>
<tr>
<td>10.0</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

The p-value is just a measure of how "far out" the test statistic is.

Example (cont.):
Null hypothesis: p=0.1. Sample size: n=100 parts. Sample proportion of defective: 0.18. Test statistic: (0.18-0.1)/0.03 = 0.666667.
The p-value is 0.007660761. Strong data evidence against the null hypothesis.

Example (cont.):
Null hypothesis: p=0.5. Sample size: n=252 days. Sample proportion of downs: 0.52778. Test statistic: (0.52778-0.5)/0.03149704 = 0.8819876.
The p-value is 0.3777835. Lack of data evidence against the null hypothesis.
Rejection and the p-value

If the test statistic is less than 2 (in absolute value) then the p-value is greater than 0.05.

If the test statistic is greater than 2 (in absolute value) then the p-value is less than 0.05.

If you want to accept/reject you can just look at the p-value.

But the p-value tells you much more.

The p-value tells you about the strength of the data evidence against a particular hypothesis.

To test the null hypothesis at level 0.05, we reject if the p-value is less than 0.05.

To test the null hypothesis at level \( \alpha \), we reject if the p-value is less than \( \alpha \).

**SMALL P-VALUE**

**BIG TEST STATISTIC**

**REJECT**

3. Confidence Intervals, Tests, and p-values in General

We have discussed confidence intervals for two parameters:

**NORMAL**

\( m \), the mean of i.i.d. normal observations

**BERNOULLI**

\( p \), the probability of 1, for i.i.d. Bernoulli observations
More generally, we could have a parameter which we could call $q$.

$q$ represents a true feature of the process or population under study.

Given a sample we obtain an estimate of $q$, say $\hat{q}$.

Here are some examples:

Let $\hat{q}$ denote an estimate of $q$.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$\hat{\theta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = E(X)$</td>
<td>$\bar{x}$</td>
</tr>
<tr>
<td>The probability of success</td>
<td>$p$</td>
</tr>
<tr>
<td>The standard deviation</td>
<td>$\sigma$</td>
</tr>
</tbody>
</table>

We think of each sample quantity as an estimate of the corresponding 'population' quantity (assuming our observations are i.i.d).

Confidence Intervals

Because of the variation inherent in our data, we know our estimates could be wrong.

How wrong can we be?

The standard error tells us.

In general, we have (at least approximately, by the central limit theorem, for a sufficient number of observations) a 95% chance that the true value will be within 2 standard errors of the estimate.
In general: \[ \hat{\theta} \pm 2\text{se}(\hat{\theta}) \]

\[ \hat{X} \pm 2\text{se}(\hat{X}) \quad \text{se}(\hat{X}) = \frac{S_x}{\sqrt{n}} \]

Bernoulli p: \[ \hat{p} \pm 2\text{se}(\hat{p}) \quad \text{se}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} \]

Now that we have the basic idea, we can look at confidence intervals for any quantity without necessarily knowing the details (i.e., the formula per se).

**Example:** We can get a confidence interval for \( s \) in the i.i.d. normal model!!

**Results for one-sample analysis for Canada**

<table>
<thead>
<tr>
<th>Summary measures</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>107</td>
</tr>
<tr>
<td>Sample mean</td>
<td>0.009</td>
</tr>
<tr>
<td>Sample standard deviation</td>
<td>0.038</td>
</tr>
</tbody>
</table>

**Confidence interval for mean**

- Confidence level: 95.0%
- Std error of mean: 0.004
- Degrees of freedom: 106
- Lower limit: 0.002
- Upper limit: 0.016

We don’t know how the confidence interval for \( s \) is computed!!

**Confidence interval for standard deviation**

- Confidence level: 95.0%
- Sample standard deviation: 0.038
- Degrees of freedom: 106
- Lower limit: 0.034
- Upper limit: 0.044

We’re not going into the details anymore!!

**Hypothesis Tests**

Here someone has some hypothesis about the real world.

Given the data we ask:

Could this data have arisen if the hypothesis is true?

**The p-value provides an answer for us.**

A small p-value means something weird happened if the hypothesis were true. We reject the hypothesis!

In particular, if the p-value < \( a \), we reject at level \( a \)!
Example: Assuming Canadian returns are i.i.d. normal, we can test the null hypothesis that \( H_0: \mu = \mu_0 \).

Results for one-sample analysis for Canada

Summary measures
- Sample size: 107
- Sample mean: 0.099
- Sample standard deviation: 0.038

Test of mean=0 versus two-tailed alternative
- Hypothesized mean: 0.090
- Sample mean: 0.099
- Std error of mean: 0.034
- Degrees of freedom: 106
- t-test statistic: 2.447
- p-value: 0.016

Here is the p-value for \( H_0: \mu = 0 \).
We reject at level 5%.

Again, even though we don’t know the details of the test, we have some sense of how to interpret it.

But,

it only means something if we understand what hypothesis is being tested!!

The calculation of the p-value assumes iid returns!!

If the returns are not iid, it is garbage!!

You don’t have to understand the details of the test, you do have to understand the modeling assumptions that underlie it!!

Example: There is a test for whether a sequence looks like it is i.i.d.!!

Runs Test Results for Canada

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of obs</td>
<td>107</td>
</tr>
<tr>
<td>Number above cutoff</td>
<td>61</td>
</tr>
<tr>
<td>Number below cutoff</td>
<td>46</td>
</tr>
<tr>
<td>Number of runs</td>
<td>60</td>
</tr>
<tr>
<td>E(R)</td>
<td>53.449</td>
</tr>
<tr>
<td>Stdev(R)</td>
<td>5.045</td>
</tr>
<tr>
<td>Z-value</td>
<td>1.298</td>
</tr>
<tr>
<td>p-value (2-tailed)</td>
<td>0.194</td>
</tr>
</tbody>
</table>

Null hypothesis: Ho: data are i.i.d.

The p-value is 0.2
Fail to reject!!
Example: Daily volume of shares traded.

Null hypothesis:
Ho: data are i.i.d.

Summary
In general, given a model we compute a confidence interval as estimate +/- 2 standard errors.

In general we can assess a hypothesis by the p-value. Small p-value => reject.

The standard errors and p-values are computed given the basic assumptions of the model. To use them properly, you must understand what these are !!
Simple Linear Regression

1. The Simple Linear Regression Model
2. Estimates and Plug-in Prediction
3. Confidence Intervals and Hypothesis Tests
4. Fits, resids, and R-squared

Book material
- What is correlation analysis and drawing the line of regression (pages 429-445 (12), 458-477 (13))
- Assumptions underlying linear regression (pages 449-450 (12), 480-482 (13))
- The standard error of estimate Confidence and prediction intervals (pages 446-448 and 451-454 (12), 477-480 and 482-486 (13))
- The relationships among the coefficient of correlation, the coefficient of determination, and the standard error of estimate (pages 457-459 (12), 489-491 (13))

1. The Simple Linear Regression Model

price vs size from the housing data we looked at before.

Two numeric variables.

We want to build a formal probability model for the variables.

price: thousands of dollars
size: thousands of square feet

![Graph showing correlation between price and size]
Do you remember conditional probabilities?

Regression looks at the conditional distribution of Y given X.

Instead of coming up with a story for the joint p(x,y), regression just talks about p(y|x):

Given that I know x, what will y be?

Example 1:

Given I know that x = 6'5" (height), what will y (weight) be?

Why regression is so popular?

Lots of reasons but two would be:

(i) Sometimes you know x and just need to predict y, as in the house price data;

(ii) As we discussed before, the conditional distribution is an excellent way to think about the relationship between two variables.

What kind of model should we use?

In the housing data, the "overall linear relationship" is striking.

Given x, y is approximately a linear function of x.

\[ y = \text{linear function of } x + \text{error} \]
The Simple Linear Regression Model

\[ y_i = \alpha + \beta x_i + \epsilon_i \]
\[ \epsilon_i \sim N(0, \sigma^2) \quad \text{iid} \]

We need the normal distribution to describe what kinds of errors we might get!!!

How far \( y_i \) is from the line \( \alpha + \beta x_i \) ?

Here is a picture of our model.

How do we get \( y_i \) from \( x_i \)? "true" linear relationship

Of course, the model is "behind the curtain", all we see are the data.

We'll have to estimate or guess the "true" model parameters from the data.
The role of $s$

We need $s$ in the model to describe how close the relationship is to linear, how big the errors are.

Another way to think about the model

\[ Y = \alpha + \beta x + \epsilon \]
\[ \epsilon \sim N(0, \sigma^2) \]
\[ \epsilon \text{ independent of } X \]

is,

\[ Y | x \sim N(\alpha + \beta x, \sigma^2) \]

since given $x$, $Y$ is just the normal $\epsilon$ plus the constant $a + bx$. 

Given the model, and $x$, what do you think $Y$ will be?

Your guess:

\[ \alpha + \beta x \]

How wrong could you be?

\[ \pm 2\sigma \]

\[ Y = \alpha + \beta x \pm 2\sigma \]

Of course we don't know the $b$'s and $s$ so we have to estimate them!!
2. Estimates and Plug-in Prediction

Example 2:
Here is the output from the regression of price on size

\[
\begin{array}{ll}
\text{Estimate} & \text{SE} \\
a & 0.8500 \\
b & 0.2500 \\
\end{array}
\]

\(a\) and \(b\) are our estimate of \(\alpha\) and \(\beta\), respectively. \(s_u\) is our estimate of \(s\).

Now we think of the fitted regression line as an estimate of the true line.

If the fitted line is

\[y = a + bx\]

then \(a\) is our estimate of \(\alpha\) and \(b\) is our estimate of \(\beta\).

"StErr of Est" is our estimate of \(s\).

We'll denote this by \(s_u\).

We may give the formulas for the estimators later!

If we plug in our estimates for the true values then a "plug-in" predictive interval given \(x\) is:

\[y = a + bx \pm 2s_u\]

Suppose we know \(x = 2.2\).

\(a + bx = 144.41\)

\(2s_u = 44.95\)

interval for \(y = 144.41 \pm 44.95\)
summary:

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>s</td>
<td>s_a</td>
</tr>
</tbody>
</table>

plug-in predictive interval given a value for x:

\[ a + bx \pm 2s_a \]

3. Confidence Intervals and Hypothesis Tests

I randomly picked 10 of the houses out of our data set.

With just those 10 observations, I get the solid line as my estimated line.

The dashed line uses all the data.

Which line would you rather use to predict?

With more data we expect we have a better chance that our estimates will be close to the true (or "population" values).

The "true line" is the one that "generalizes" to the size and price of future houses, not just the ones in our current data.

How big is our error?

We have standard errors and confidence intervals for our estimates of the true slope and intercept.
Let \( s_a \) denote the standard error associated with the estimate \( a \).
Let \( s_b \) denote the standard error associated with the estimate \( b \).

\[
\text{Notation: it might make more sense to use } se(b) \text{ instead of } s_{b}, \text{ but I am following the book.}
\]

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Standard error</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5895</td>
<td>0.5895</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>0.9673</td>
<td>0.9673</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

95% confidence interval for \( a \):
\[
a = \text{estimate } \pm \text{tval } * s_{a}
\]
\[
t\text{val = } TINV(0.05, n - 2) \quad \text{(in excel)}
\]

95% confidence interval for \( b \):
\[
b = \text{estimate } \pm \text{tval } * s_{b}
\]
\[
t\text{val = } TINV(0.05, n - 2) \quad \text{(in excel)}
\]

If \( n \) is bigger than 30 or so, tval is about 2.

Example 2 (cont.)
For the housing data the 95% confidence interval for the slope is:
\[
70.23 +/- 2(9.43) = 70.23 +/- 18.86 = (51.4, 89.1)
\]
big !! (what are the units?)
With only 10 observations $b = 135.50$ and $s_b = 49.77$.

Note how much bigger the standard error is than with all 128 observations!!

$$135.5 \pm 2.3\times 50 = (20.5, 250.5)$$

really big!!

Note:

It the confidence interval for slope and intercept are big the plug-in predictive interval can be misleading!!

There are ways to correct for plugging in estimates but we won’t cover them.

The predictive interval just gets bigger!!
Hypothesis tests on coefficients:

To test the null hypothesis

\[ H_0 : \alpha = \alpha^0 \quad \text{vs.} \quad H_a : \alpha \neq \alpha^0 \]

We reject at level .05 if

\[ |t| = \left| \frac{\hat{\alpha} - \alpha^0}{s_{\alpha}} \right| > t_{\text{val}} \]

\[ t_{\text{val}} = \text{TINV}(0.05, n - 2) \]

Otherwise, we fail to reject.

Intuitively, we reject if estimate is more than 2 se's away from proposed value.

\[ t \text{ is the "t statistic" reject if the t statistic is bigger than 2 !!} \]

Same for slope:

To test the null hypothesis

\[ H_0 : \beta = \beta^0 \quad \text{vs.} \quad H_a : \beta \neq \beta^0 \]

We reject at level .05 if

\[ |t| = \left| \frac{\hat{\beta} - \beta^0}{s_{\beta}} \right| > t_{\text{val}} \]

\[ t_{\text{val}} = \text{TINV}(0.05, n - 2) \]

Otherwise, we fail to reject.

Intuitively, we reject if estimate is more than 2 se's away from proposed value.

Note:

the hypothesis: \( H_0 : b = 0 \)

is often tested.

Why?

\[ Y \mid x \sim N(\alpha + \beta x, \sigma^2) \]

If the slope \( b = 0 \), then the conditional distribution of \( Y \) does not depend on \( x \) \( \Rightarrow \) they are independent!

(under the assumptions of our model)
Example 2 (cont.)

Stats packages automatically print out the t-statistics for testing whether the intercept=0 and whether the slope=0.

To test b=0, the t-statistic is \( \frac{b-0}{s_b} = \frac{70.2263}{9.4265} = 7.45 \)

We reject the null at level 5% because the t-stat is bigger than 2 (in absolute value).

---

p-values

Most regression packages automatically print out the p-values for the hypotheses that the intercept=0 and that the slope is 0.

That's the p-value column in the StatPro output.

Is the intercept 0?, p-value = .59, fail to reject

Is the slope 0?, p-value = .0000, reject

---

Note:

For \( n \) greater than about 30, the t-stat can be interpreted as a z-value. Thus we can compute the p-value.

For the intercept:

\[
2 \times (\text{the standard normal cdf at -.53}) = .596
\]

which is the p-value given by the package.
Example 3: The market model

In finance, a popular model is to regress stock returns against returns on some market index, such as the S&P 500.

The slope of the regression line, referred to as “beta”, is a measure of how sensitive a stock is to movements in the market.

Usually, a beta less than 1 means the stock is less risky than the market, equal to 1 same risk as the market and greater than 1, riskier than the market.

We will examine the market model for the stock General Electric, using the S&P 500 as a proxy for the market.

Three years of monthly data give 36 observations.

Regression output:

The regression equation is

\[ \text{ge} = 0.00301 + 1.20 \text{sp500} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.003013</td>
<td>0.006229</td>
<td>0.48</td>
<td>0.632</td>
</tr>
<tr>
<td>sp500</td>
<td>1.1995</td>
<td>0.1895</td>
<td>6.33</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 0.03454 \quad \text{R-sq = 54.1\%} \quad \text{R-sq(adj) = 52.7\%} \]

We can test the hypothesis that the slope is zero: that is, are GE returns related to the market?
The test statistic is

\[ t = \frac{b - 0}{s_b} = \frac{1.2}{0.1895} = 6.33 \]

and

\[ t_{\text{val}} = 2.03 \]

so we reject the null hypothesis at level .05. We could have looked at the p-value (which is smaller than .05) and said the same thing right away.

We now test the hypothesis that GE has the same risk as the market: that is, the slope equals 1.

The t statistic is:

\[ t = \frac{1.1995 - 1}{0.1895} = 1.055 \]

Now, 1.055 is less than 2.03 so we fail to reject.

What is the p-value ??

What is the 95% confidence interval for the GE beta?

\[ 1.2 \pm 2(0.2) = [0.8, 1.6] \]

**Question:** what does this interval tell us about our level of certainty about the beta for GE?
4. Fits, resids, and R-squared

Our model is:

\[ Y = \alpha + \beta x + \epsilon \]

We think of each \((x_i, y_i)\) as having been generated by

\[ Y_i = \alpha + \beta x_i + \epsilon_i \]

part of \(y\) that depends on \(x\)
part of \(y\) that has nothing to do with \(x\)

It turns out to be useful to estimate these two parts for each observation in our sample. For each \((x_i, y_i)\) in the data:

\[ \alpha + \beta x_i = a + bx_i \]
\[ e_i = y_i - (\alpha + \beta x_i) = y_i - (a + bx_i) = e_i \]

have,

\[ \hat{y}_i = a + b x_i, \quad e_i = y_i - \hat{y}_i \]
\[ y_i = \hat{y}_i + e_i \]

\( \hat{y}_i \): fitted value for \(i^{th}\) observation.
\( e_i \): residual for \(i^{th}\) observation.

Fits and resids for the housing data.
Regression chooses a, b so that:
\[ E(e) = 0, \text{cor}(e, x) = 0 \]

Intuition:
Model: \( E(e) = 0, \text{cor}(x, e) = 0 \)
\[ \Rightarrow \text{make sample quantities exactly so:} \]
- Residual line off line vs x
- Residual line vs y
- Regression line too big

Note:
\[ \text{cor}(e, x) = 0 \Rightarrow \text{cor}(e, a + bx) = 0 \]
\[ \Rightarrow \text{cor}(e, \hat{y}) = 0 \]

Have:
\[ y_i = \hat{y}_i + e_i \]
\[ \text{cor}(e, \hat{y}) = 0 \Rightarrow \bar{e} = 0 \]

\[ y_i = \hat{y}_i + e_i \]
\[ \Rightarrow \bar{y} = \bar{\hat{y}} + \bar{e} = \bar{\hat{y}} \]
because residuals have 0 sample average
\[ s^2_y = s^2_\hat{y} + s^2_e \]
because residuals and fits have 0 sample correlation.
\[ \Rightarrow \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^{n} e_i^2 \]
total variation in y = variation explained by x + unexplained variation
R-squared

\[
R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}
\]

\[
= 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}
\]

0 ≤ R^2 ≤ 1 the closer R-squared is to 1, the better the fit.

Note:

R^2 is also equal to the square of the correlation between y and x.

Table of correlations

<table>
<thead>
<tr>
<th></th>
<th>SigPl</th>
<th>Price</th>
<th>Fitted Values</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>SigPl</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Price</td>
<td>1.00</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Fitted Values</td>
<td>1.00</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Residuals</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
</tbody>
</table>

.553^2 = 0.305809

Note: \( \text{cor}(y,x) = \text{cor}(y,\hat{y}) \)

\( R^2 = \) the square of the correlation between y and the fits !!
Example 4

Housing data from a different neighborhood.

price: thousands of dollars
size: thousands of square feet

The correlation is .974.

R² = .974² = 0.948676

Regression output:

The regression equation is

price = 5.76 + 14.8 size

Predictor Coef Stdev t-ratio p
Constant 5.776 1.633 3.53 0.003
size 14.8159 0.8829 16.78 0.000

s = 2.210 R-sq = 94.9% R-sq(adj) = 94.6%

Analysis of Variance

SOURCE DF SS MS F p
Regression 1 1374.7 1374.7 281.58 0.000
Error 15 73.2 4.9
Total 16 1447.9

Fit Std.err.Fit 95% C.I. 95% P.I.
38.358 0.669 (36.932, 39.783) (33.436, 43.279)

For any x, the plug-in predictive interval has error

+/− 2s_e = +/-4.4 thousands of dollars: big!!!

Even though R² is big, we still have a lot of predictive uncertainty !!!

I think people over-emphasize R².
I like s_e !!!
Multiple Linear Regression

1. The Multiple Linear Regression Model
2. Estimates and Plug-in Prediction
3. Confidence Intervals and Hypothesis Tests
4. Fits, resids, R-squared, and the overall F-test
5. Categorical Explanatory Variables: Dummy Variables

Book material

• What is correlation analysis and drawing the line of regression (pages 429-445 (12), 458-477 (13))
• Assumptions underlying linear regression (pages 449-450 (12), 480-482 (13))
• The standard error of estimate Confidence and prediction intervals (pages 446-448 and 451-454 (12), 477-480 and 482-486 (13))
• The relationships among the coefficient of correlation, the coefficient of determination, and the standard error of estimate (pages 457-459 (12), 489-491 (13))
• Multiple regression analysis (pages 475-483 (12), 512-519 (13))

1. The Multiple Linear Regression Model

The plug-in predictive interval for the price of a house given its size is quite large.

How can we improve this?

If we know more about a house, we should have a better idea of its price !!
Our data has more variables than just size and price:
The first 7 rows are:

<table>
<thead>
<tr>
<th>Home</th>
<th>Size</th>
<th>Bathrooms</th>
<th>Price</th>
<th>(price and size /1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000</td>
<td>3</td>
<td>15000</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>1500</td>
<td>4</td>
<td>20000</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>2000</td>
<td>2</td>
<td>15800</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>1800</td>
<td>3</td>
<td>17000</td>
<td>1.7</td>
</tr>
<tr>
<td>5</td>
<td>2100</td>
<td>4</td>
<td>25000</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>1900</td>
<td>2</td>
<td>20000</td>
<td>2.0</td>
</tr>
<tr>
<td>7</td>
<td>1600</td>
<td>3</td>
<td>15800</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Suppose we know the number of bedrooms and bathrooms a house has as well as its size, then what would our prediction for price be?

---

The Multiple Linear Regression Model

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{iid}$$

$y$ is a linear combination of the $x$ variables + error.

The error works exactly the same way as in simple linear reg!! We assume the $\epsilon$ are independent of all the $x$'s.

---

Another way to think about the model

$$Y | x = (x_1, x_2, \ldots, x_k) \sim \mathcal{N}(\mu_x, \sigma^2)$$

$$\mu_x = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$Y$ is normal with the mean depending on the $x$'s through a linear combination.
If we model price as depending on size, nbed, nbath, then we have:

\[ \text{Price}_i = \alpha + \beta_1 \text{nbed}_i + \beta_2 \text{nbath}_i + \beta_3 \text{size}_i + \varepsilon_i \]

Given data, we have estimates of \( \alpha, \beta_1, \beta_2, \) and \( \varepsilon_i. \)

\( a \) is our estimate of \( \alpha. \)
\( b_1 \) is our estimate of \( \beta_1. \)
\( s_e \) is our estimate of \( \varepsilon. \)

2. Estimates and Plug-in Prediction

Here is the output from the regression of price on size (SqFt), nbed (Bedrooms) and nbath (Bathrooms):

**Summary measures**
- Multiple-R: 0.6635
- R- squared: 0.6458
- Adj R- squared: 0.6425
- Std. of Est: 40.3065

**ANOVA Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unexplained</td>
<td>124</td>
<td>11594.2206</td>
<td>414.5890</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

**Regression coefficients**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-ratio</th>
<th>p-value</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.9408</td>
<td>17.2240</td>
<td>-0.0570</td>
<td>0.9553</td>
<td>-36.6932</td>
</tr>
<tr>
<td>nbed</td>
<td>10.4358</td>
<td>2.4130</td>
<td>4.3410</td>
<td>0.0000</td>
<td>10.2042</td>
</tr>
<tr>
<td>nbath</td>
<td>13.1461</td>
<td>2.2197</td>
<td>5.9710</td>
<td>0.0000</td>
<td>9.9641</td>
</tr>
<tr>
<td>size</td>
<td>30.9427</td>
<td>10.9573</td>
<td>3.4103</td>
<td>0.0011</td>
<td>14.5262</td>
</tr>
</tbody>
</table>

So, for example, \( b_2 = 13.5461 \)

Our estimated relationship is:

\[ \text{Price} = -5.64 + 10.46^*\text{nbed} + 13.55^*\text{nbath} + 35.64^*\text{size} \]

Interpret:

With size, and nbath held fixed, adding one bedroom adds 10.460 thousands of dollars.

With nbed and nbath held fixed, 1 square foot increases the price $36.
Suppose a house had size = 2.2, 3 bedrooms and 2 bathrooms.

What is your (estimated) idea of the price?

\[-5.64 + 10.46*3 + 13.55*2 + 35.64*2.2 = 131.248\]

\[2s_a=40.72\]

\[131.248 +/- 40.72\]

This is our multiple regression plug-in predictive interval.

The error is still estimated to be +/- 2s_a!

Note:

When we regressed price on size the coefficient was about 70.

Now the coefficient for size is about 36.

Without nbath and nbed in the regression, an increase in size can be associated with an increase in nbath and nbed in the background.

If all I know is that one house is a lot bigger than another I might expect the bigger house to have more beds and baths!

With nbath and nbed held fixed, the effect of size is smaller.

Note:

With just size, our predictive +/- was

\[2\times22.467 = 44.952\]

With nbath and nbed added to the model the +/- is

\[2\times20.36 = 40.72\]

The additional information makes our prediction more precise (but not a whole lot in the case, we still need some "better x's").
3. Confidence Intervals and Hypothesis Tests

95% confidence interval for $a$:

$$a \pm t_{\text{val}} \cdot s_a$$

tval = TINV(.05, n-k-1) (in excel)

95% confidence interval for $b$:

$$b \pm t_{\text{val}} \cdot s_b$$

tval = TINV(.05, n-k-1) (in excel)

(recall the $k$ is the number of x's)

Example:

StatPro prints out all the confidence intervals.

---

4. Results of multiple regression for profitbasis

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressed</td>
<td>9</td>
<td>1939.26</td>
<td>215.48</td>
<td>41.30</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Regression coefficients:

- Constant: $-3.9466$, $t = -2.06$, $p = 0.082$, Lower limit = $-9.896$, Upper limit = $2.005$
- b1: $1.251$, $t = 1.34$, $p = 0.20$, Lower limit = $-0.79$, Upper limit = $3.26$
- b2: $4.412$, $t = 2.45$, $p = 0.02$, Lower limit = $1.67$, Upper limit = $7.15$
- b3: $-3.9466$, $t = -2.06$, $p = 0.082$, Lower limit = $-9.896$, Upper limit = $2.005$

eg $s_b = 4.22$

the interval for $b_2$ is $13.57 \pm 2(4.22)$

Hypothesis tests on coefficients:

To test the null hypothesis

$$H_0: \alpha = \alpha^0 \text{ vs. } H_a: \alpha \neq \alpha^0$$

We reject at level .05 if

$$|t| > t_{\text{val}} \text{ where, } t = \frac{a - \alpha^0}{s_a}$$

tval = TINV(.05, n-k-1)

Otherwise, we fail to reject.

Intuitively, we reject if estimate is more than 2 se's away from proposed value.

The t statistic is the "t statistic" reject if the t statistic is bigger than 2!!
Same for slope:

To test the null hypothesis

\[ H_0 : \beta_i = \beta_i^* \quad \text{vs.} \quad H_a : \beta_i \neq \beta_i^* \]

**We reject** at level .05 if

\[ |t| > t_{\text{val}} \quad \text{where, } t = \frac{b_i - \beta_i^*}{s_{b_i}} \]

\[ t_{\text{val}} = TINV(.05, n - k - 1) \]

Otherwise, **we fail to reject**.

Intuitively, we reject if estimate is more than 2 se's away from proposed value.

---

**Example**

Packages automatically print out the t-statistics for testing whether the intercept=0 and whether each slope=0 as well as the associated p-values.

**Results of multiple regression for aname:**

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>B</th>
<th>SE</th>
<th>t</th>
<th>p-value</th>
<th>Lower limit</th>
<th>Upper limit</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>4.354</td>
<td>0.304</td>
<td>14.32</td>
<td>0.000</td>
<td>3.709</td>
<td>4.999</td>
<td>3.709</td>
<td>4.999</td>
</tr>
<tr>
<td>Aname</td>
<td>0.594</td>
<td>0.023</td>
<td>26.43</td>
<td>0.000</td>
<td>0.549</td>
<td>0.640</td>
<td>0.549</td>
<td>0.640</td>
</tr>
</tbody>
</table>

**Fits, resids, and R-squared**

In multiple regression the fit is:

\[ \hat{y}_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} \]

"the part of y related to the x’s"

as before, the residual is the part left over:

\[ e_i = y_i - \hat{y}_i \]
In multiple regression, the residuals have sample mean 0 and are uncorrelated with each of the x's and the fitted values:

\[ y_i = \hat{y}_i + e_i \]

estimated x part of y

estimated part of y

that has nothing to do with x's

This is the plot of the residuals from the multiple regression of price on size, nbath, nbed vs the fitted values. We see the 0 correlation.

The correlation is also 0, for each of the x's.

\[ \text{cor}(\hat{y}, e) = 0, \quad \bar{e} = 0 \]

So, just as with one x we have:

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 \]

total variation in y = variation explained by x + unexplained variation.
R-squared

\[
R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}
\]

\[
= 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}
\]

\[0 \leq R^2 \leq 1\]  the closer R-squared is to 1, the better the fit.

In our housing example:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>Std. Error</th>
<th>t-Value</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>49.430</td>
<td>17.648</td>
<td>2.772</td>
<td>0.006</td>
<td>25.139</td>
<td>73.721</td>
<td>25.139</td>
</tr>
<tr>
<td>Area</td>
<td>1.186</td>
<td>0.142</td>
<td>8.377</td>
<td>0.000</td>
<td>1.008</td>
<td>1.364</td>
<td>1.008</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>1.935</td>
<td>0.835</td>
<td>2.335</td>
<td>0.023</td>
<td>-0.797</td>
<td>4.667</td>
<td>-0.797</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>-3.979</td>
<td>1.513</td>
<td>-2.628</td>
<td>0.012</td>
<td>-7.128</td>
<td>-0.830</td>
<td>-7.128</td>
</tr>
</tbody>
</table>

\[R^2 = \frac{40301}{40301 + 51384} = 0.439\]

\[R^2 = \text{ explained variance} = \sum (\hat{y}_i - \bar{y})^2 / \sum (y_i - \bar{y})^2\]

\[R^2 = \text{ total variance} = \sum (y_i - \bar{y})^2 / \sum (y_i - \bar{y})^2\]

\[R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}\]

\[0 \leq R^2 \leq 1\]

\[\text{closer R-squared is to 1, the better the fit.}\]

\[\text{Regression finds the linear combination of the x's which is most correlated with y.}\]

\[\text{with just size, the correlation between fits and y was 0.553}\]

\[\text{R}^2\text{ is also the square of the correlation between the fitted values and y:}\]

\[\text{Regression finds the linear combination of the x's which is most correlated with y.}\]

\[\text{cor}(\hat{y}, y) = 0.663\]

\[0.663^2 = 0.439569\]
The "Multiple R" is the correlation between y and the fits

\[
\text{cor}(\hat{y}, y) = 0.663
\]

**Summary measures**
- Multiple R: 0.6632
- R-Square: 0.4396
- Adj R-Square: 0.4396
- S=3 of Est: 20.3665

**ANOVA Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explained</td>
<td>3</td>
<td>45000.9677</td>
<td>15000.2559</td>
<td>32.4192</td>
<td>0.0003</td>
</tr>
<tr>
<td>Unexplained</td>
<td>124</td>
<td>51594.2296</td>
<td>414.3985</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Regression coefficients**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5.6658</td>
<td>12.3964</td>
<td>-0.2378</td>
<td>0.7703</td>
<td>-56.6658</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>10.0399</td>
<td>2.8213</td>
<td>3.5915</td>
<td>0.0005</td>
<td>4.5955</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>11.5389</td>
<td>2.8173</td>
<td>3.2715</td>
<td>0.0017</td>
<td>5.1992</td>
</tr>
<tr>
<td>Square</td>
<td>20.3437</td>
<td>10.6073</td>
<td>3.3413</td>
<td>0.0011</td>
<td>14.5292</td>
</tr>
</tbody>
</table>

The overall F-test

The p-value beside "F" if testing the null hypothesis:

\[ H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \] (all the slopes are 0)

We reject the null, at least some of the slopes are not 0.

5. Categorical Explanatory Variables: Dummy Variables

Here, again, is the first 7 rows of our housing data:

<table>
<thead>
<tr>
<th>Name</th>
<th>Neighborhood</th>
<th>Bedrooms</th>
<th>Bathroom</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>9800</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10000</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>10200</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>10000</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>10200</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>10200</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>10200</td>
</tr>
</tbody>
</table>

Does whether a house is brick or not affect the price of the house?

This is a categorical variable.

How can we use multiple regression with categorical x’s ??!!

What about the neighborhood? (location, location, location !!!)
Adding a Binary Categorical $x$

To add "brick" as an explanatory variable in our regression we create the dummy variable which is 1 if the house is brick and 0 otherwise:

<table>
<thead>
<tr>
<th>House</th>
<th>Year</th>
<th>Area</th>
<th>Sqft</th>
<th>Brick</th>
<th>Type</th>
<th>Size</th>
<th>Brick dummy</th>
<th>Price</th>
<th>Income</th>
<th>Stories</th>
<th>Rooms</th>
<th>Bathrooms</th>
<th>Fireplace</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
</tbody>
</table>

Note:

I created the dummy by using the excel formula:

=IF(Brick="Yes",1,0)

but we'll see that StatPro has a nice utility for creating dummies.

As a simple first example, let's regress price on size and brick.

Here is our model:

$$\text{Price}_i = \alpha + \beta_1 \text{size}_i + \beta_2 \text{brickdum}_i + \epsilon_i$$

How do you interpret $\beta_2$?
What is the expected price of a brick house given the size?

$$E(\text{Price} | \text{size} = s, \text{brick}) = \alpha + \beta_s + \beta_2$$

What is the expected price of a non-brick house given the size?

$$E(\text{Price} | \text{size} = s, \text{nonbrick}) = \alpha + \beta_s$$

$b_2$ is the expected difference in price between a brick and non-brick house.

Note:

You could also create a dummy which was 1 if a house was non brick and 0 if brick.

That would be fine, but the meaning of $b_2$ which change.

You can't put both dummies in though because given one, the information in the other is redundant.

Let's try it !!

Result of multiple regression for price:

<table>
<thead>
<tr>
<th>Summary measure</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>$F$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_1$</td>
<td>0.54</td>
<td>0.53</td>
<td>12.78</td>
<td>0.000</td>
</tr>
<tr>
<td>$b_2$</td>
<td>0.46</td>
<td>0.45</td>
<td>9.12</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Regression coefficients

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_1$</td>
<td>23.4</td>
<td>4.96</td>
<td>0.000</td>
</tr>
<tr>
<td>$b_2$</td>
<td>2.35</td>
<td>0.75</td>
<td>0.45</td>
</tr>
</tbody>
</table>

$+/−2\text{se} = 39.3$, this is the best we've done!

what is the brick effect:

$$23.4 +/- 2(3.7) = 23.4 +/- 7.4$$
We can see the effect of the dummy by plotting the fitted values vs size.

The upper line is for the brick houses and the lower line is for the non-brick houses.

We can interpret $b_2$ as a shift in the intercept.

Notice that our model assumes that the price difference between a brick and non-brick house does not depend on the size!

The two variables do not "interact".

Sometimes we expect variables to interact.

Now let's add brick to the regression of price on size, nbath, and nbed:

\[
\begin{align*}
\text{Price} & \sim \beta_0 + \beta_1 \text{Size} + \beta_2 \text{Brick} + \beta_3 \text{Nbath} + \beta_4 \text{Nbed} + e
\end{align*}
\]

Adding brick seems to be a good idea!!
I created one dummy for each the neighborhoods.

<table>
<thead>
<tr>
<th>Home</th>
<th>Size</th>
<th>Rooms</th>
<th>Bath</th>
<th>Bad</th>
<th>Bedrooms</th>
<th>Bathrooms</th>
<th>Price</th>
<th>Neighborhood 1</th>
<th>Neighborhood 2</th>
<th>Neighborhood 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2000</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1500</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1800</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1700</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1900</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1800</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1700</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

eg. Nbhd_1 indicates if the house is in neighborhood 1 or not.

Now we add any two of the three dummies. Given any two, the information in the third is redundant.

Let's first do price on size and neighborhood:

\[ \text{Price}_i = \alpha + \beta_1 \text{size}_i + \beta_2 \text{N1}_i + \beta_3 \text{N2}_i + \epsilon_i \]

where now I've use N1 to denote the dummy for neighborhood 1 and same for 2.

\[ \text{Price}_i = \alpha + \beta_1 \text{size}_i + \beta_2 \text{N1}_i + \beta_3 \text{N2}_i + \epsilon_i \]

\[ \text{E(Price} | \text{size} = s, \text{neighborhood3}) = \alpha + \beta_1 s \]
\[ \text{E(Price} | \text{size} = s, \text{neighborhood2}) = \alpha + \beta_1 s + \beta_2 \]
\[ \text{E(Price} | \text{size} = s, \text{neighborhood1}) = \alpha + \beta_1 s + \beta_2 \]

\[ b_1: \text{difference between hood 2 and hood 3} \]
\[ b_2: \text{difference between hood 1 and hood 3} \]

The neighborhood corresponding to the dummy we leave out becomes the "base case" we compare to.
Let's try it!

Results of multiple regression for priathouse

Summary measures:

<table>
<thead>
<tr>
<th>Estimate</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th>Estimate</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Again we are assuming that size and neighborhood do not interact.

Here is fits vs size.

Which line corresponds to which neighborhood?

Where do you want to live?

ok, let's try price on size, nbnd, nbath, brick, and neighborhood.

Results of multiple regression for priceathouse

Summary measures:

<table>
<thead>
<tr>
<th>Estimate</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th>Estimate</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+/- 2se = 30.52 !!!

+/- 2es = 24 !!!
Dropping bedrooms did not increase $s_e$ or decrease R-Square so no need to bother with it.

Regression finds a linear combination of the variables that is like $y$. 

price vs size: 

price vs combination of size, nbath, brick, nbhd
The residuals are the part of $y$ not related to the $x$'s.

summary: adding a Categorical $x$

In general to add a categorical $x$, you can create dummies, one for each possible category (or level as we sometimes call it).

Use all but one of the dummies.

It does not matter which one you drop for the fit, but the interpretation of the coefficients will depend on which one you choose to drop.

Topics in Regression

1. Residuals as Diagnostics
2. Transformations as Cures
3. Logistic Regression
4. Understanding Multicollinearity
5. Autoregressive Models
6. Financial Time Series
1. Residuals as Diagnostics

Example 1: Here is the regression output for four different data sets. In each case we have just one x.

<table>
<thead>
<tr>
<th>DATASET 1</th>
<th>The regression equation is y1 = 3.00 + 0.500 x1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor</td>
<td>Coef</td>
</tr>
<tr>
<td>Constant</td>
<td>3.000</td>
</tr>
<tr>
<td>x1</td>
<td>0.5001</td>
</tr>
<tr>
<td>s</td>
<td>1.237</td>
</tr>
<tr>
<td>R-sq</td>
<td>66.7%</td>
</tr>
<tr>
<td>R-sq(adj)</td>
<td>62.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DATASET 2</th>
<th>The regression equation is y2 = 3.00 + 0.500 x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor</td>
<td>Coef</td>
</tr>
<tr>
<td>Constant</td>
<td>3.001</td>
</tr>
<tr>
<td>x2</td>
<td>0.5000</td>
</tr>
<tr>
<td>s</td>
<td>1.237</td>
</tr>
<tr>
<td>R-sq</td>
<td>66.6%</td>
</tr>
<tr>
<td>R-sq(adj)</td>
<td>62.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DATASET 3</th>
<th>The regression equation is y3 = 3.00 + 0.500 x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor</td>
<td>Coef</td>
</tr>
<tr>
<td>Constant</td>
<td>3.002</td>
</tr>
<tr>
<td>x3</td>
<td>0.4997</td>
</tr>
<tr>
<td>s</td>
<td>1.236</td>
</tr>
<tr>
<td>R-sq</td>
<td>66.6%</td>
</tr>
<tr>
<td>R-sq(adj)</td>
<td>62.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DATASET 4</th>
<th>The regression equation is y4 = 3.00 + 0.500 x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor</td>
<td>Coef</td>
</tr>
<tr>
<td>Constant</td>
<td>3.002</td>
</tr>
<tr>
<td>x4</td>
<td>0.4999</td>
</tr>
<tr>
<td>s</td>
<td>1.236</td>
</tr>
<tr>
<td>R-sq</td>
<td>66.7%</td>
</tr>
<tr>
<td>R-sq(adj)</td>
<td>63.0%</td>
</tr>
</tbody>
</table>

In each case the output is identical. Whatever decision you are trying to make (e.g., prediction) would be the same!!

Moral of the Story

Only in the first case does the plot suggest that the simple linear regression model is a good way to think about the data.

In the other cases a blind use of the model would lead to bad decisions.

QUESTION:
So, how do you tell if the model is “a good way to think about your data”?

Plot the data!
ANOTHER QUESTION: With more than one $x$, how do we "plot" the data? How can we diagnose a problem with the regression model?

Basic idea: If the model is right then

$$ e_i = \epsilon_i - N(0, \sigma^2) \text{ independent of the } x's !!! $$

The residuals should look i.i.d. normal;
The residuals should be unrelated to the $x$'s.

To see how this works, we'll first use one $x$ for simplicity. But the real problem is multiple regression (with one $x$ you can just plot $y$ vs $x$).

Example 2: nonlinear regression

Example 3: outliers

Example 4: heteroskedasticity
In each example we can see something wrong or peculiar!!

**Example 2:**
Failure of basic assumption of linear relationship.

**Example 3:**
A funny point, an outlier.

**Example 4:**
The variance of errors increases with x, we have nonconstant variance: "heteroskedasticity".

Our model assumes "homoskedasticity", i.e. a constant variance.

In multiple regression we plot the resids vs each x. There should be nothing funny!!

Since the fits are a function of the x’s, we also plot the resids vs the fits and again there should be no relationship.

In principle, the resids should be unrelated to any function of the x’s, but in practice we just do individual x’s and the fits.

Note: now you know why most regression packages/softwares, such as excel, give you the option of making these plots!

**Example 5**
Here are resids vs fitted from house price on size, nbed, and nbath.

Looks pretty good!

Is there an outlier?

*This plot is a good thing!!*
This is a plot of \( y_n = \frac{e^{\epsilon_n}}{\sigma} \sim N(0,1) \) vs the fits.

If the model is right these standardized residuals should look like iid standard normal draws independent of the x's (and hence the fits).

Is -2.64 unusual? 20 times I simulated 128 iid standard normals. Each time I picked off the smallest one.

The smallest of 128 could easily be -2.6 if the model were true.

2. Transformations as Cures

Ok, suppose you find a problem. What can you do about it?

If you find an outlier you should investigate! Why is it weird??

If you find nonlinearity or heteroskedasticity you can sometimes "fix it" by using transformations.

We'll look at the two most common transformations: Logarithms and polynomials.
2.1 The Log Transformation

Suppose we have this relationship:

\[ Y = cx^2(1 + r) \]

Here \((1+r)\) is a multiplicative error.
\(r\) is percentage error.

Often we see this, the size of the error is a percentage of the expected response.

This would lead to heteroskedasticity.

Take the log:

\[ Y = cx^2(1 + r) \]

\[ \log(Y) = \log(c) + \beta \log(x) + \log(1 + r) \]

\[ = \alpha + \beta \log(x) + \epsilon \]

where \(\alpha = \log(c)\) and \(\epsilon = \log(1+r)\).

*We can regress the log of \(y\) on the log of \(x\)!!*

Obviously, taking the log turns these nonlinear relationships into linear ones in terms of the transformed variables.

It also takes a multiplicative (percentage error) and turns it into the additive error of the regression model.

In practice, logging \(y\) is often a good cure for heteroskedasticity.
Suppose now the relationship is:

\[ Y = ce^{x(1+r)} \]

\[ \log(Y) = \log(c) + \beta x + \log(1+r) = \alpha + \beta x + \epsilon \]

Here we regress log of y on x.

In practice you can just log y or y and some of the x's.

Don't log a dummy variable!!.

---

**Example 6**

Goal: relate the brain weight of a model to its body weight.

Each observation corresponds to a mammal.

y: brain weight (grams)

x: body weight (grams)

Each observation corresponds to a mammal.

Does additive error make sense?

---

logy vs logx

Looks pretty nice!!
The big residual is the chinchilla.

Very few people know that the chinchilla is a master race of supreme intelligence.

No.

The book I got this from had chinchilla at 64 grams instead of 6.4 grams (which I found in another book).

The next biggest positive residual is man.

2.2 Polynomials

Example 7: each observation corresponds to a service call.

x: number of units serviced
y: time to complete
The usual linear model,

\[ Y = \alpha + \beta x + \epsilon \quad (y = \text{linear} + \text{error}) \]

does not look like a great idea.

We’ll try:

\[ Y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon \quad (y = \text{quadratic} + \text{error}) \]

a multiple regression where one x is the square of the other !!

Just create a new column with the squares of the old x column:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>X^2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here is the output:
Fits vs x.

\[ y = -9.75 + 22.22x - 0.59x^2 \]

To make a prediction, plug in \( x \) and \( x^2 \).

Residuals versus fitted values

not bad!

In general our model

\[ y = \text{polynomial} + \text{error} \]

For example with two \( x \)'s we might have:

\[ Y = \alpha + \beta_1x_1 + \beta_2x_1^2 + \beta_3x_2 + \beta_4x_2^2 + \beta_5x_1x_2 + \varepsilon \]

With many \( x \)'s you can see that there are a lot of possibilities.

Note that the product term give us interaction. It is no longer true that the effect of changing one \( x \) does not depend on the value of the others.
Example 8

The housing data again.

\[ y: \text{price} \]

\[ x_1: \text{size} \]

\[ x_2: \text{dummy for neighborhood 1} \]

\[ x_3: \text{dummy for neighborhood 2} \]

model:

\[ Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \varepsilon \]

interpret:

\[ E(Y \mid \text{neighborhood}) = \alpha + \beta_1 x_1 + \beta_3 x_3 \]

\[ = (\alpha + \beta_3) + (\beta_1 + \beta_3) x_1 \]

It makes no sense to square or log a dummy !!!

Fits vs size.

Now we see that lines don't have to be parallel !

But it does not seem that there is much interaction.

On the other hand the lower slope for the "worst" neighborhood makes sense !!

here is the regression output:

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-ratio</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>50.0667</td>
<td>0.9954</td>
<td>50.13</td>
<td>0.0000</td>
<td>59.0667</td>
</tr>
<tr>
<td>x1</td>
<td>1.0012</td>
<td>0.0023</td>
<td>43.56</td>
<td>0.0000</td>
<td>1.0012</td>
</tr>
<tr>
<td>x2</td>
<td>1.0031</td>
<td>0.0036</td>
<td>27.89</td>
<td>0.0000</td>
<td>1.0031</td>
</tr>
<tr>
<td>x3</td>
<td>1.0052</td>
<td>0.0058</td>
<td>17.43</td>
<td>0.0000</td>
<td>1.0052</td>
</tr>
<tr>
<td>x1x3</td>
<td>1.0073</td>
<td>0.0106</td>
<td>95.42</td>
<td>0.0000</td>
<td>1.0073</td>
</tr>
</tbody>
</table>

what happens if you throw out each variable with t-statistic less than 2?
3. Logistic Regression

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>race</th>
<th>Reg</th>
<th>Inc</th>
<th>poli</th>
<th>party</th>
<th>public</th>
<th>attit</th>
<th>news</th>
<th>ender</th>
<th>friend</th>
<th>simp</th>
<th>foot</th>
<th>cola</th>
<th>juice</th>
<th>cigs</th>
<th>antq</th>
<th>town</th>
<th>shop</th>
<th>hull</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>63</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We want to relate football viewing to demographics.

Linear regression:

- relate numeric y to numeric x's.
- If you have a categorical x, you use dummies.

**Now we have a (binary) categorical y !!!!**

- It does not make sense to think of y as a linear combination + error !!
- As usual, we will represent y as a 0-1 dummy.

**The Logit Model**

Now we want a model for

\[ Y|x \]

where Y is 0 or 1.

Given x, what is the distribution of Y?

\[ Y|x \sim \text{Bernoulli}(p). \]

We need p to depend on x.

(just like m did in regression)
**p as a function of x**

Two steps:

(i)

x only affects y through a linear combination of the x's.

Let,

\[ \eta = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \]

we assume that \( h \) captures everything the x's have to say about Y !!

(ii)

p is a function of \( \eta \).

We can't have \( p = h \) because we need to have \( p \) between 0 and 1!

We let,

\[ p = F(\eta) \]

\[ F(\eta) = \frac{e^\eta}{1 + e^\eta} \]

What does \( F(\eta) = \frac{e^\eta}{1 + e^\eta} \) look like?

Notice that F takes on values between 0 and 1.

Bigger \( h \) means bigger F means bigger p.
That is,

\[ Y | x_1, x_2, \ldots, x_k \sim \text{Bernoulli}(p) \]
\[ p = F(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \]

Given data, most packages will give you estimates of the b’s and standard errors.

Let’s try it.

Example 9: Football on Age

\[ h = -0.8101 - 0.0285 \text{age} \]
\[ p_{\text{football}} = \exp(h)/(1+\exp(h)) \]

As age increase from 20 to 80, h decreases from -1.2 to -3.6, p decreases from 0.22 to 0.03.

An older person has a smaller h, and then a smaller p.
This plot is the one that really summarizes our estimated relationship:

\[ p(\text{football}|\text{age}) \]

Confidence intervals and hypothesis tests

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
<th>t-value</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.007</td>
<td>-0.007</td>
<td>0.011</td>
<td>0.000</td>
<td>-0.114</td>
<td>0.114</td>
<td>0.999</td>
</tr>
<tr>
<td>age</td>
<td>0.028</td>
<td>0.007</td>
<td>0.014</td>
<td>0.040</td>
<td>-0.042</td>
<td>-0.014</td>
<td>0.005</td>
</tr>
</tbody>
</table>

CI for age = estimate +/- 2se

\[ \text{CI} = -0.0285 +/- 2(0.007) = (-0.0422, -0.0148) \]

It's not easy to interpret these coefficients.

To test whether the coefficient is 0:

\[ \frac{b - 0}{\sigma_b} = \frac{-0.0285 - 0}{0.007} = -4.972 \]

If the null were true, this should look like a draw from the standard normal. We reject \( b = 0 \). Again, the small p-value also means reject.
Example 10: Football on age and sex

Just as with linear regression, we create a dummy for sex: sex_1: 1 if male, 0 otherwise.

Regression coefficients:

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Err.</th>
<th>t value</th>
<th>p-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex_1</td>
<td>0.044</td>
<td>2.266</td>
<td>0.0301</td>
<td>0.014</td>
<td>0.075</td>
</tr>
</tbody>
</table>

Since the coefficient for sex_1 is positive, a man has a larger h, and hence a large prob.

It seems the both coefficients are clearly different from 0.

4. Multicolinearity

Suppose we are regressing a Y on x's and the x's are highly correlated.

What happens to the standard errors?

\[ s_{b_i} = \frac{s_b}{\sqrt{\text{SSE}}} \]

this will be small !!!!

Which makes the standard error large.

What happens to the t statistic for testing the coefficient = 0?
Example 11: We have one Y and two X's.

Plot x1 vs x2.
They are highly correlated.
There is very little variation in one x not associated with variation in the other.

How can you tell if a change in Y was caused by a change in X1 or X2 when they always change together!!! They never do anything on their own!!!

The regression equation is
\[ y = 0.130 + 1.33 \times x_1 - 0.14 \times x_2 \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.1304</td>
<td>0.1504</td>
<td>0.87</td>
<td>0.390</td>
</tr>
<tr>
<td>x1</td>
<td>1.334</td>
<td>1.090</td>
<td>1.22</td>
<td>0.227</td>
</tr>
<tr>
<td>x2</td>
<td>-0.140</td>
<td>1.114</td>
<td>-0.13</td>
<td>0.900</td>
</tr>
</tbody>
</table>

\[ s = 1.030 \quad R^2 = 60.9\% \quad R^2(adj) = 59.2\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>77.506</td>
<td>38.753</td>
<td>36.53</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>47</td>
<td>49.856</td>
<td>1.061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>127.362</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notice that the overall F is very significant but neither t is!!!

Clearly, if we regress Y on each X one at a time the t values for the slopes will be big!!!

Clearly, Y is related to the X's (the big F).

But it is very difficult to estimate the two multiple regression coefficients because the X's are so closely linearly related (the small t's).
**Multicolinearity:**

When the x’s are highly correlated it may be that there is not enough variation in some of the x’s which is unrelated to the other x’s to be able to estimated their slopes well.

We get large standard errors and hence small t’s so we would fail to reject the null that the true slope is 0.

Here is an important example where “fail to reject” does not mean accept. If we get a small t because of multicolinearity it just means we cannot estimate the slope well so we don’t know that it is not 0.

Before you run a regression check all the correlations between your x’s.

If they are high, multicolinearity may be a problem.

---

**Dealing with the Problem of Multicolinearity**

Basically multicolinearity means there is not enough information in the data to estimate the separate slopes.

The basic solution is to get more data with less correlation amongst the x’s.

In experimental design we choose the x’s so that the correlation is low (0 usually).

Sometimes people throw out some x’s or combine some x’s into an average.

---

**5. Autoregressive models**

The mean July level of lake Michigan in number of feet above sea level in excess of 570

One numeric variable, measured over time (annually).

Is it iid??
If \( Y_t \) denotes level at year \( t \), then iid means:
\[
p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2)\cdots p(y_n)
\]
in particular,
\[
p(y_{t+1} | y_t, y_{t-1}, \ldots) = p(y_{t+1} | y_t)
\]
Now we wonder if maybe, for example,
\[
p(y_{t+1} | y_t, y_{t-1}, \ldots) = p(y_{t+1} | y_t)
\]
What happens next, is related to what happened before.

**Autocorrelation**

Let's see if \( y_t \) and \( y_{t-1} \) are related. We can do this by lagging the series.

<table>
<thead>
<tr>
<th>Year</th>
<th>Level</th>
<th>Level Lagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>1770</td>
<td>10.96</td>
<td>10.96</td>
</tr>
<tr>
<td>1771</td>
<td>11.98</td>
<td>10.96</td>
</tr>
<tr>
<td>1772</td>
<td>11.35</td>
<td>10.96</td>
</tr>
<tr>
<td>1773</td>
<td>11.14</td>
<td>11.35</td>
</tr>
<tr>
<td>1774</td>
<td>12.70</td>
<td>11.14</td>
</tr>
</tbody>
</table>

The second column is simply the previous value of the first. It is the first lagged once.

.......

Each row is \( (y_t, y_{t-1}) \).

*they are clearly related !!!*

Now we can plot this year's lake level against last year's to see if they are related.

Note that we are assuming that the nature of the relationship between successive years does not change over time.
How about this year and two years ago:

<table>
<thead>
<tr>
<th>Year</th>
<th>Level</th>
<th>Level_Lag2</th>
<th>Level_Lag3</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10.85</td>
<td>12.36</td>
<td>11.86</td>
</tr>
<tr>
<td>12</td>
<td>12.18</td>
<td>10.37</td>
<td>12.82</td>
</tr>
<tr>
<td>9</td>
<td>9.79</td>
<td>10.97</td>
<td>10.95</td>
</tr>
<tr>
<td>10</td>
<td>10.39</td>
<td>10.82</td>
<td>9.79</td>
</tr>
<tr>
<td>10</td>
<td>10.42</td>
<td>9.79</td>
<td>10.95</td>
</tr>
</tbody>
</table>

The second lag give us \((y_t, y_{t-2})\) pairs.

This year, is related to two years ago.

We can summarize the relationships with autocorrelations:

<table>
<thead>
<tr>
<th>Table of correlations</th>
<th>Level</th>
<th>Level_Lag2</th>
<th>Level_Lag3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Level</td>
<td>Level_Lag2</td>
<td>Level_Lag3</td>
</tr>
<tr>
<td>Level</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Level_Lag2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Level_Lag3</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Level this year is correlated .839 with level last year, and .632 with level two years ago.

Autocorrelation is the correlation between values of a variable and past values of the same variable.
The standard error is \( \frac{1}{\sqrt{T}} \)

where \( T \) is the number of observations.

Our lake data has 98 observations so the standard error is about .1

An autocorrelation bigger than \( \frac{2}{\sqrt{T}} \)

is considered "significant".

It is traditional to plot the autocorrelations:

This plot is called the ACF.

This year’s lake level is related to that of past years but the strength of the relationship diminishes with the lag.

Suppose data were iid.

What should the ACF look like?
I simulated 100 iid $N(0,1)$

The acf:
none are bigger than $2/\sqrt{100} = .2$

The AR(1) Model

Ok suppose the acf indicates dependence. We need a model to describe it.

In the case
$$p(y_{t+1} \mid y_t, y_{t-1}, \ldots) = p(y_{t+1} \mid y_t)$$

we often try:
$$y_t = \alpha + \beta y_{t-1} + \epsilon_t$$

where $\epsilon_t$ is independent of the past $= (y_{t-1}, y_{t-2}, \ldots)$

$$y_t = \alpha + \beta y_{t-1} + \epsilon_t$$

the part of Y predictable from the past

the new part of y unpredictable from the past

We often assume:
$$\epsilon_t \sim N(0, \sigma^2) \text{ iid}$$
How do we estimate the parameters?
Simply run an autoregression:

If this year's level is 11, what is your prediction for next year's level?

\[ y = 1.467 + 0.8364(11) \pm 2(0.72) \]

\[ = 10.67 \pm 1.44 \]

Does the model fit the data, that is, capture all the dependent structure?
If the model is right, the residuals should look like iid normal draws.
Here is the acf of the resids:

No evidence of dependent structure in the resids!!

The AR(p) Model

There is no guarantee the AR(1) model work capture the dependence in the data.
The current value may be related to more than just the previous one.
We can try the AR(p) model:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \epsilon_t$$

Trend Plus error model

Another popular time series model is the trend model:

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$
6. Financial Time Series

- Many time series applications involve time price series.
- Data: \( Y_1, Y_2, \ldots, Y_{t+1}, \ldots \) where \( t \) indexes the day, month, year, or any "time" interval.
- Key idea: today's price has information about tomorrow's or \( Y_{t-1} \) related to \( P_{t+1} \) and hence is not independent.

- Trends:
  How do we determine whether a series has a trend in it or not? Remember: one of the key biases is that people confuse a realised trend (from one sample) with the existence of a "real" trend.

Consider a price series \( P_t \).

1. Expect no change: \( P_{t+1} = P_t + \epsilon_t \) where \( E(\epsilon_t) = 0 \) and so \( E(P_{t+1}|P_t) = P_t \). I expect tomorrow's price to be the same as today. This is a simple random walk model.

2. A Trend: \( P_{t+1} = \mu + P_t + \epsilon_t \) Here \( E(P_{t+1}|P_t) = \mu + P_t \) \( \mu \) is the daily trend. If \( \mu > 0 \) there's a tendency to increase and if \( \mu < 0 \) to decrease. Don't forget the error term \( \epsilon_t \) means that the sample path (realisation) won't always go up or down. This is called a random walk with drift.

Mean Reversion

Mean Reversion involves a regression type model of the form

\[
P_{t+1} = \mu + \beta P_t + \epsilon_t
\]

where \( |\beta| < 1 \). The long run average is given by

\[
P = \mu + \beta P \text{ or } P = \frac{\mu}{1 - \beta}
\]

Whenever the series is above this long run average there's a tendency for the series to mean-revert to its long run average. This is known as an autoregressive model of order one, AR(1).
How to Analyse Financial Data

- Should we care whether the series are levels, differences or returns?
- Returns are defined as $\frac{P_{t+1}}{P_t}$ and log-returns as $\ln \left( \frac{P_{t+1}}{P_t} \right)$.
  In most cases you want to understand the return, $R_t$, process
  \[ R_t = \mu + \sigma B_t \]
  where $B_t$ is a Brownian motion. All that means is that $B_t$ has a $N(0,t)$ distribution.

Stationarity

- A series is stationary if
  \[ E(P_t) \text{ and } V(P_t) \]
  are finite and constant.
- Is a random walk stationary?
  \[ P_{t+1} = P_t + \epsilon_t \text{ and } E(P_t) = 0, \text{Var}(P_t) = \sigma^2 t \]
  why?

Daily, Weekly, Monthly Vol

StatFact: A 15% return with a 10% volatility per annum translates into a 93% probability of making money.
why? $p = \Phi \left( \frac{15}{10} \right) = 0.93$

- Effect of Time:
  On a narrow time scale (one second) this translates to a probability of only 50.02%.
  Key Fact: $\mu_t = \mu t$ and $\sigma_t = \sigma \sqrt{t}$
  why? expectations and variances add.
- Hence $p_t = \Phi \left( \frac{\mu t}{\sigma} \right)$
There are also two types of financial volatilities:

**Historical Volatility**
These are volatility estimates arrived at from looking at the historical path of prices and using a model (maybe time-varying) to estimate the future path of volatility;

**Implied Volatility**
These come from exchange based market measures explaining the market's current perception about what average future volatility will look like. VIX and VXN indices for the S&P500 and NASDAQ indices, respectively.

---

**Time-Varying Volatility**

- Let's first make volatility time-vary $\sigma_t$ so the price evolution looks like
  \[ P_{t+1} = P_t + \sigma_t \epsilon_t, \]
- What properties of volatility do we believe in?
  1. Is it related to yesterday's movement?
  2. What if yesterday was a large down versus a large up?
  3. Is volatility mean-reverting?

---

**GARCH**

- Generalized Autoregressive Conditional Heteroscedastic (GARCH)
- Let $\epsilon_t^2$ be yesterday's squared residual.
  \[ \sigma_{t+1}^2 = \alpha + \beta \sigma_t^2 + \gamma \epsilon_t^2 \]
  How about an asymmetry effect?
  \[ \log \sigma_{t+1}^2 = \alpha + \beta \log \sigma_t^2 + \gamma \epsilon_t - \rho |\epsilon_t| \]
  Lots of our related models, ARCH, ...

---

**There are also two types of financial volatilities:**

**Historical Volatility**
These are volatility estimates arrived at from looking at the historical path of prices and using a model (maybe time-varying) to estimate the future path of volatility;

**Implied Volatility**
These come from exchange based market measures explaining the market's current perception about what average future volatility will look like. VIX and VXN indices for the S&P500 and NASDAQ indices, respectively.
More about VIX

VIX is based on the Black-Scholes option pricing model to calculate implied volatilities for a number of stock options.

VIX is constructed using the S&P 500 index.

VIX is expressed as an annual percentage. A VIX of 15, for example, means the market is expecting a 15% change in price over the next year.
BUSINESS STATISTICS

Exploratory Data Analysis
Looking for clues and patterns in order to select better models.

Probability
The language/metric of uncertainty.

Statistical Inference and Hypothesis Testing
From deductions to inductions.

Regression Analysis
Pretty neat way of modeling conditional dependences.

THANK YOU!