Business Statistics 41000
Lecture 1

Mladen Kolar

University of Chicago, Booth School of Business

April 3 and 5, 2014

What is Statistics?

Data
Real world

1. This is the raw information
2. Computerization => Lots of data!

Models
Capture Uncertainty

1. Explain the real world
2. Predict outcomes
3. Test competing ideas

Statistics is the link between the models and the data.

1. Estimation: Which model is the "best" model?
2. Testing: If a given model is "right" then the real world data should look consistent with the models predictions. We can rule out some models using this idea.

Course structure

1. We begin with a discussion of summarizing real world data (means, variances, covariances etc…).
2. We next introduce models. Since the real world is uncertain our models will involve uncertainty/probability.
3. We then combine the data with the models to discuss estimation and testing. We begin with very simple models and move to more complex models.

1.Descriptive Statistics

Visualization

Dotplot and Histogram

1.1 Dotplot and Histogram

Let's look at some *returns* data.

Simple Returns

The return on an asset is the percentage increase in wealth invested in the asset over a given time period.

If you invest B at the beginning of the time period you get E = B+rB=(1+r)B at the end of the time period, where r is the return.

Clearly, given E and B we can calculate r: r = (E-B)/B

The Canadian Returns Data

Here are 107 *monthly* returns on a broadbased portfolio of Canadian assets (more on portfolios later).

```
canada
    0.07    0.05    0.02   -0.04    0.08   -0.02   -0.05    0.02    0.03
    0.00    0.03    0.08   -0.03    0.01    0.03    0.01    0.02    0.08
    0.02   -0.02    0.00    0.01    0.02   -0.09    0.00    0.01   -0.07
    0.07    0.00    0.02   -0.05   -0.04   -0.03    0.03    0.04    0.00
    0.07    0.00    0.01    0.04   -0.02    0.02    0.01   -0.03    0.05
   -0.02    0.00    0.01   -0.01   -0.05   -0.01    0.01    0.00    0.02
   -0.02   -0.07    0.03   -0.04    0.03   -0.02    0.06    0.03    0.04
    0.01   -0.01   -0.01    0.01   -0.05    0.09   -0.02    0.05    0.06
   -0.05   -0.04   -0.01    0.01   -0.06    0.05    0.06    0.02   -0.01
   -0.06    0.02   -0.05    0.06    0.04    0.02    0.04    0.02    0.02
    0.00    0.00   -0.01    0.04    0.01    0.05   -0.01    0.02    0.04
    0.02   -0.03   -0.03    0.05    0.04    0.08    0.07   -0.03
```

Each number corresponds to a month.

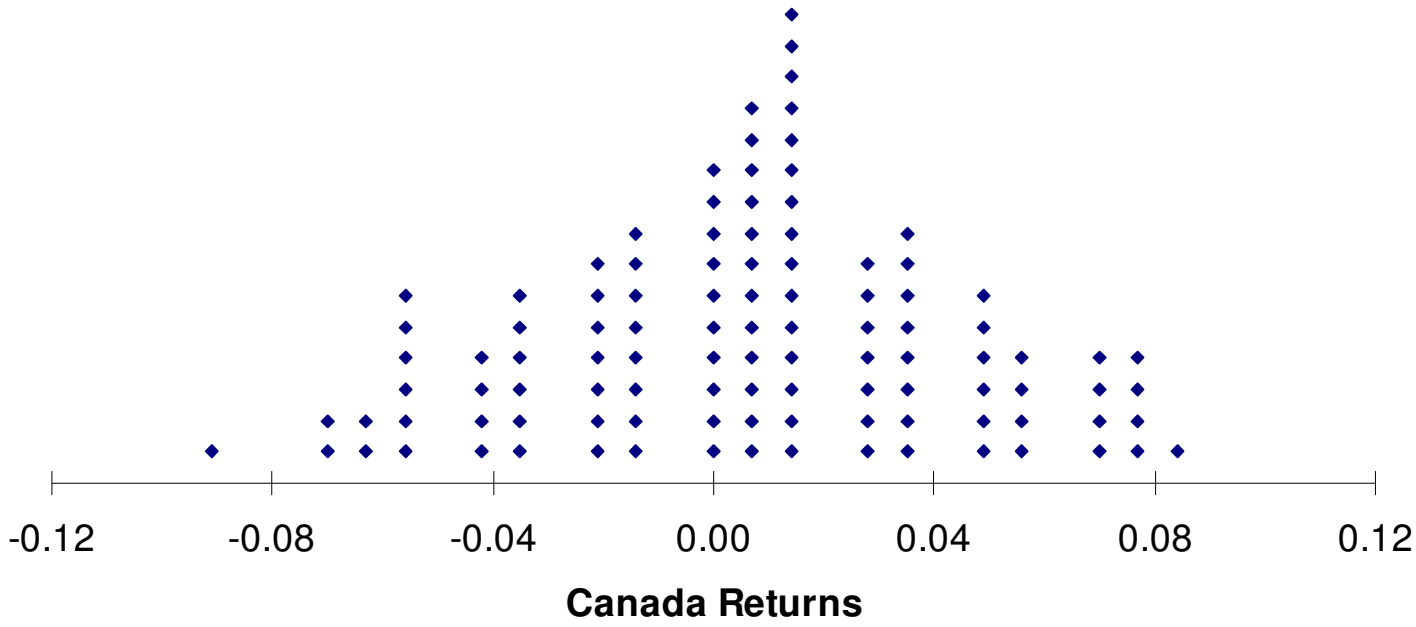
They are given in time order (go across rows first).

Our first *observation* is .07.

In the first month, the return was .07, in the 11th, .03.

The Dotplot

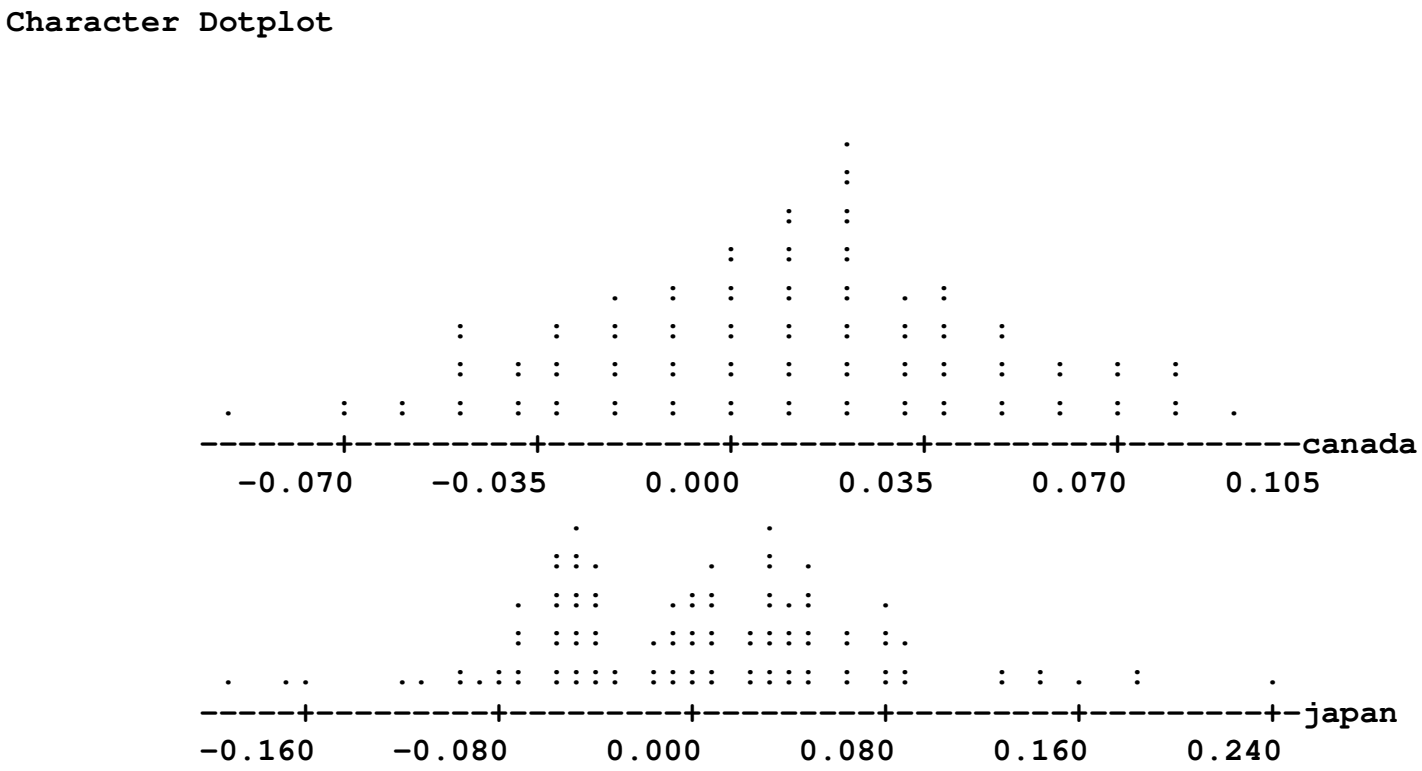
To display the returns we'll first use a dotplot.
For each number simply place a dot above the corresponding point on the number line.



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

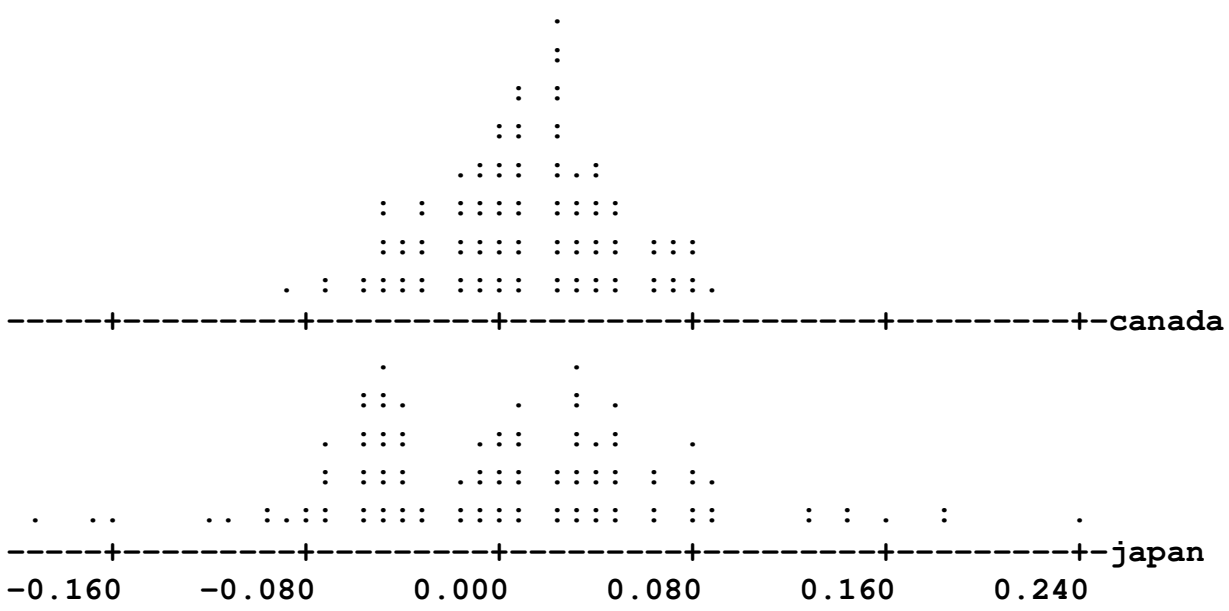
We also have data on countries other than Canada.

Let’s compare Canada with Japan.



It really helps to get things on the same scale.

How is Japan different from Canada?



Mutual Funds Data

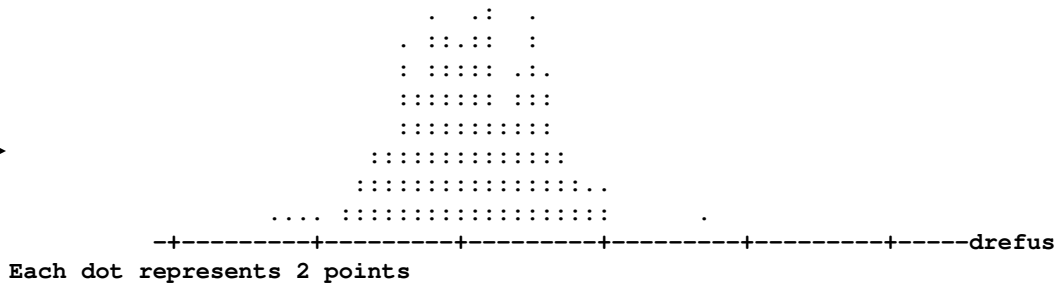
Let's use the dotplot to compare returns on some other kinds of assets.

We'll look at returns on two different mutual funds, the equally weighted market, and T-bills.

The equally weighted market is returns on a *portfolio* where you spread you money out equally over a wide variety of stocks.

Data on 4 different kinds of returns:

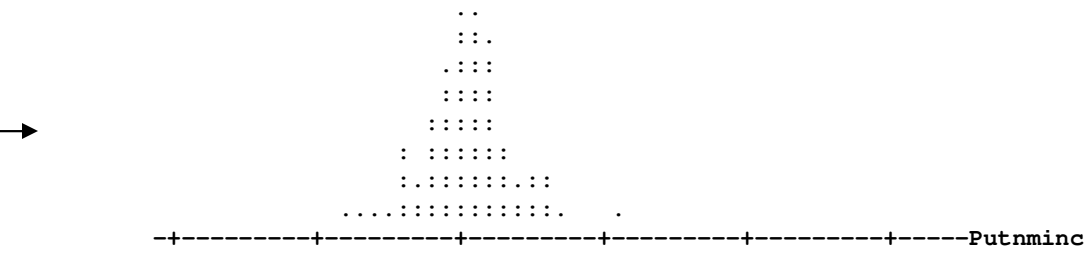
Dreyfus growth fund



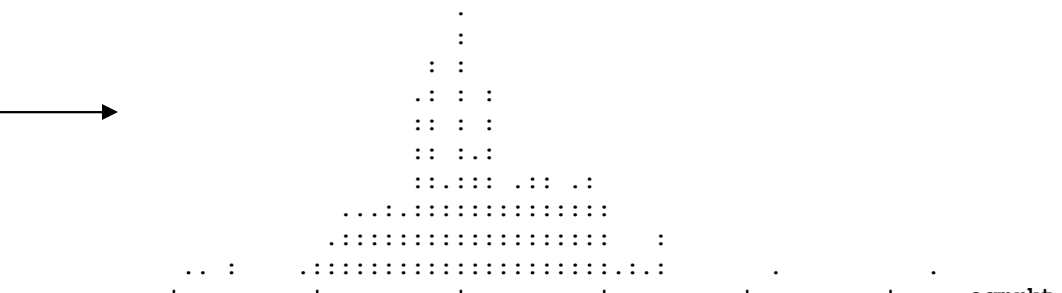
Each dot represents 2 points

Putman income fund

(*note that each dot is now 2 points)*



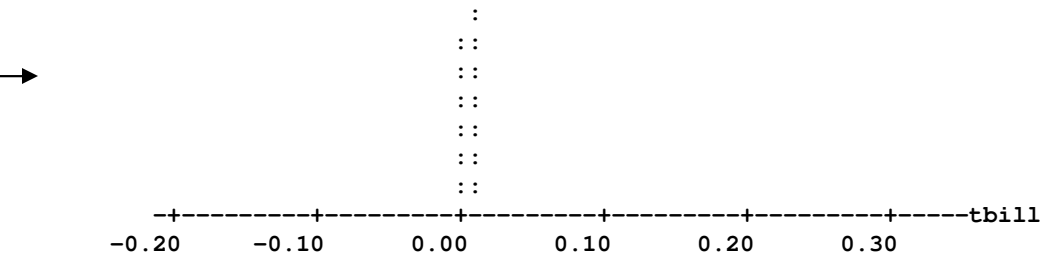
Equally weighted market



Each dot represents 7 points

T-bills

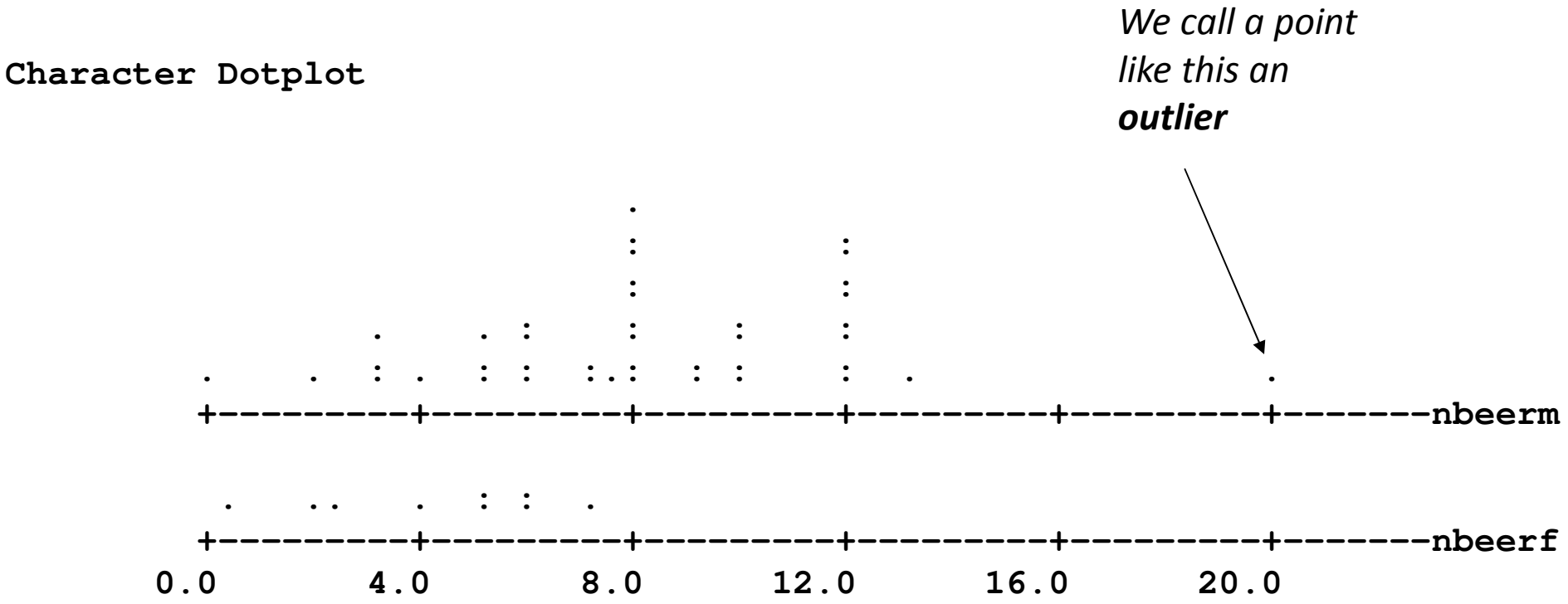
(the risk free asset)



Beer Data

nbeerm: the number of beers male mba students claim they can drink without getting drunk

nbeerf: same for females



Generally the males claim more,
their numbers are centered or located at larger values.

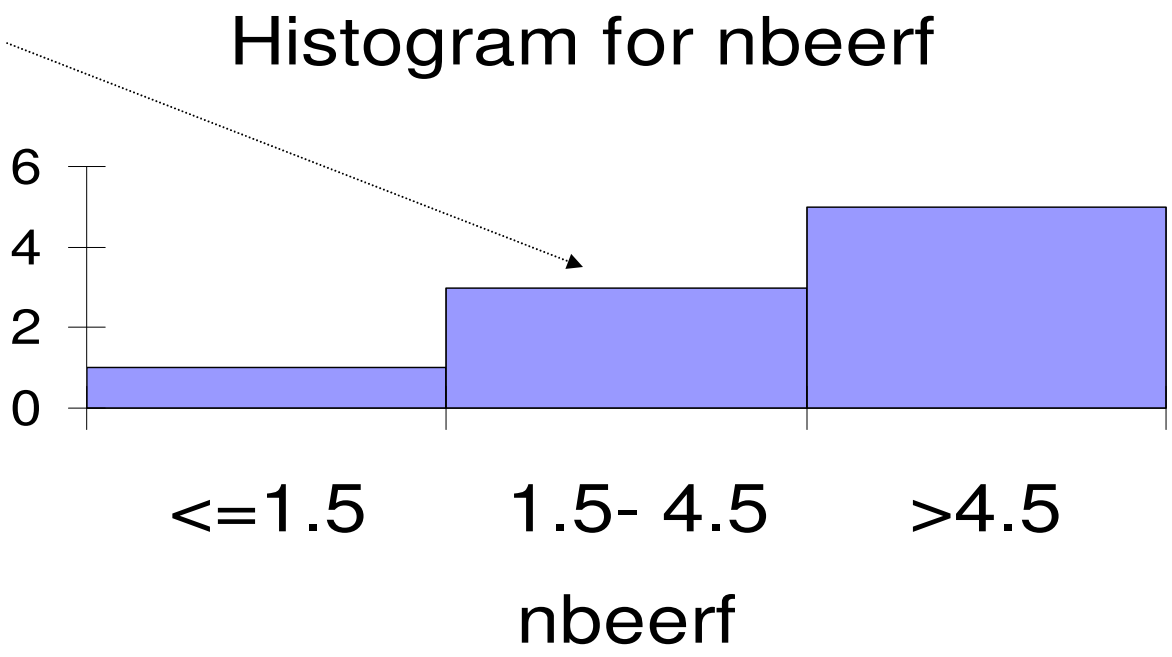
The Histogram

Sometimes the dotplot 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.

```
nbeerf
     4.0     2.0     5.0     6.0     0.5     7.0     6.0     2.5     5.0
```

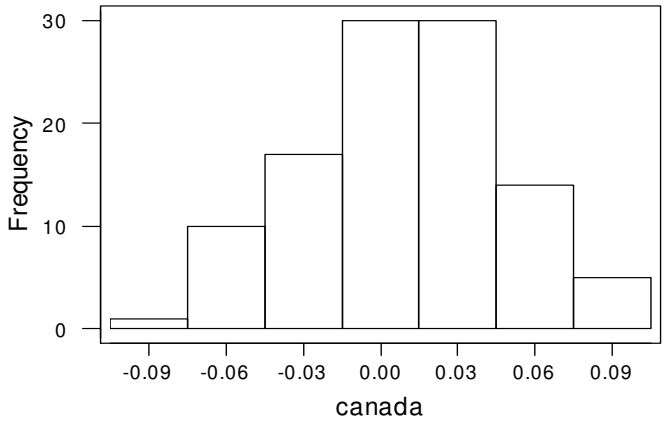
3 women have number of beers between 1.5 and 4.5.

3 women have number of beers in the *interval* (1.5,4.5).

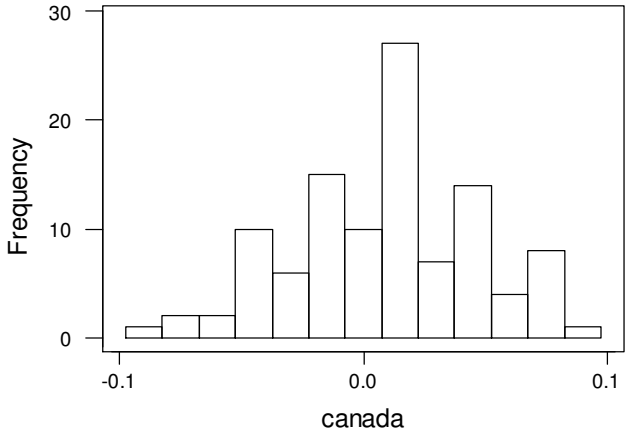


The choice of bin width determines the degree of resolution. A large bin With is very smooth (flat) with no detail.
A small bin has more detail, but maybe too much.

Here is the histogram of the Canadian returns.



Here it is again with the bin width half the size of the one above.



There is no special way of determining the bin width. I use my eye.

The Time Series Plot

1.2 The Time Series Plot

We just looked at two data kinds of data.

the returns data

the number of beers

For the returns data, each number corresponds to a month.
For the beers data, each number corresponds to a person.

The returns data has an important feature the beer data does not.

It has an order!

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

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

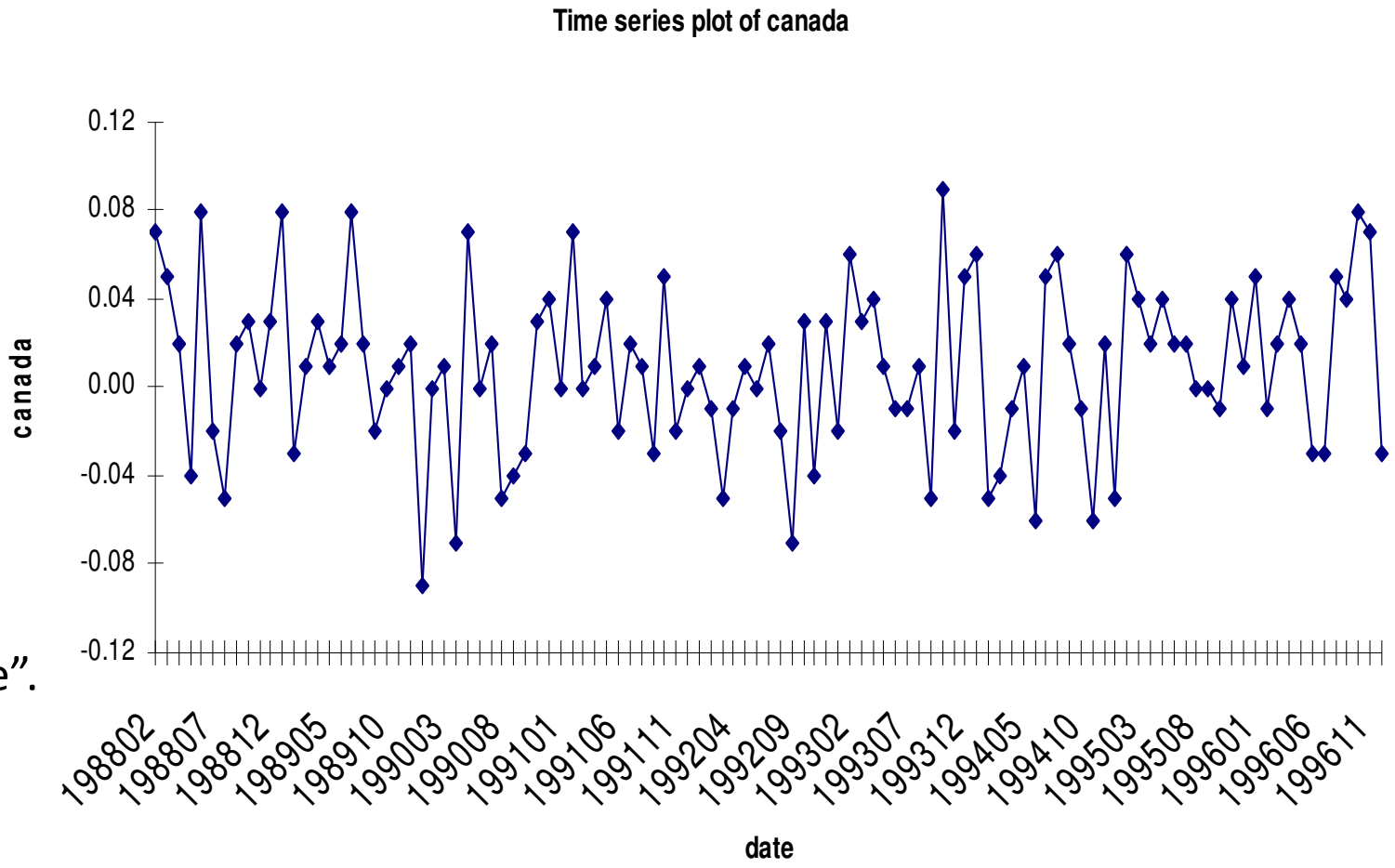
We could have daily data (closing price stock price)
annual data (inflation)
quarterly data (unemployment)
Trade by trade transaction prices.

For time series data, the time series plot is an important way to look at our data.

The returns data,
time series plot of the Canadian returns:

On the vertical axis we have the return.

On the horizontal axis we have "time".

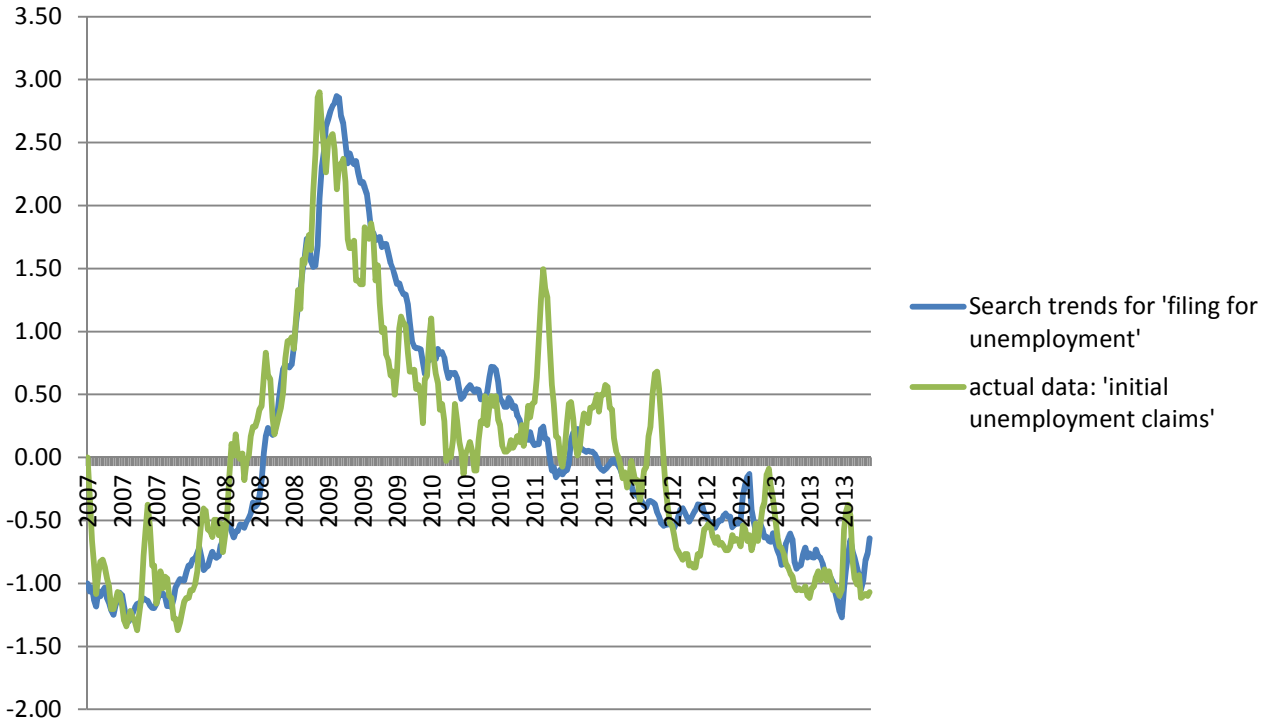


Do you see a pattern ?

Its also interesting to plot multiple time series on the same plot for comparison.

Google Trends: This shows you the ups-and-downs of the public's interest in a particular topic using how often we search for it.

Google Correlate: Reverse engineers the problem!



Summarizing a Single Numeric Variable

1.3 Summarizing a Single Numeric Variable

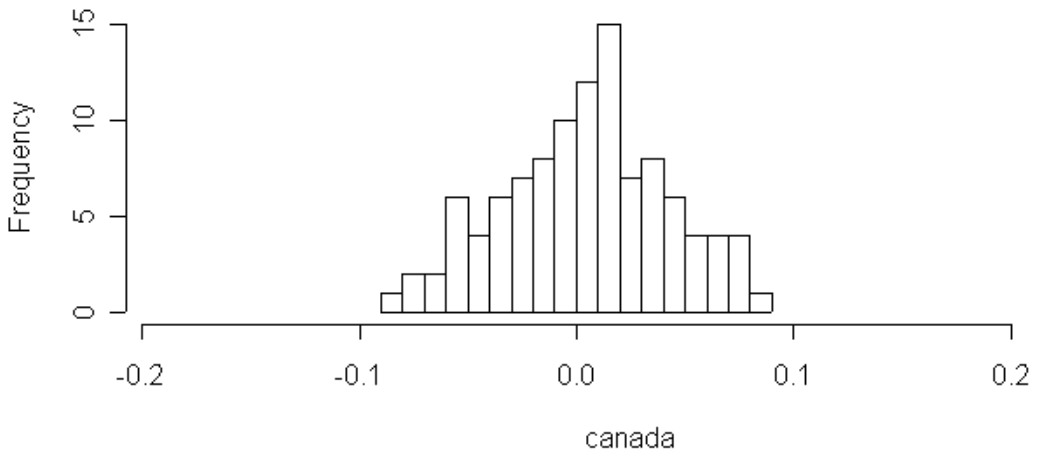
We have looked at graphs. Suppose we are now interested in having numerical summaries of the data rather than graphical representations.

Two important features of any numeric variable are:

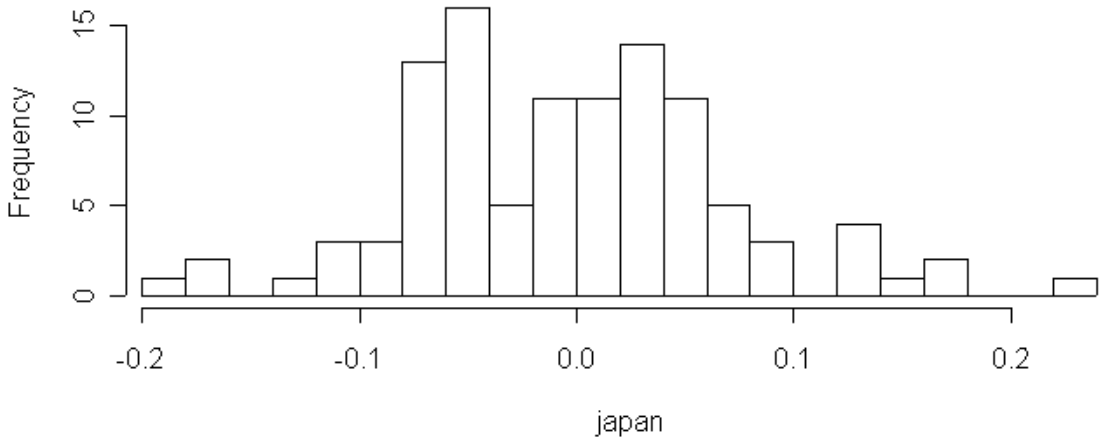
1) What is a typical or average value?

2) How spread out or ‘variable’ are the values?

Monthly returns on Canadian portfolio and Japanese portfolio.



They seem to be centered roughly at the same place but Japan has more spread.

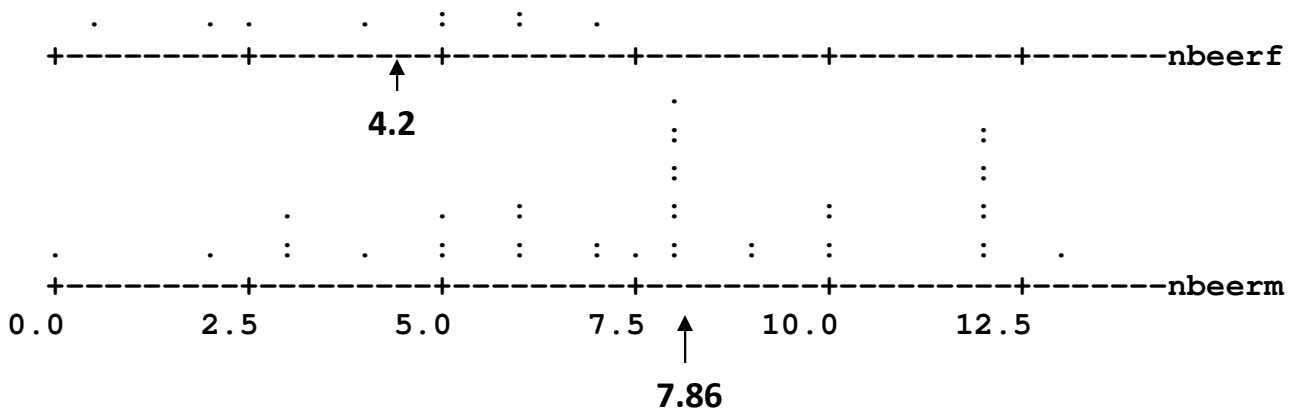


How can we summarize this?

```
=AVERAGE(nbeerf)

    4.2222

=AVERAGE(nbeerm)

    7.8625
```

"On average women claim they can drink 4.2 beers".

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



Let’s compare the means of the Canadian and Japanese returns.

```
AVERAGE(canada)

0.0090654

AVERAGE(japan)

0.0023364
```

This is a big difference.

It was hard to see this difference in the dotplots because the difference is small compared to the variation.

The median is another measure of central tendency

- Arrange the data in ascending order. After this ranking, the middle number in the data set is the median. If there are an even number of data points then the median is the average of the two center numbers.

- Example 1 4 7 8 10: Median=7
- Example 1 4 5 8: Median=4.5

- Unlike the mean, the median is not affected by outliers.

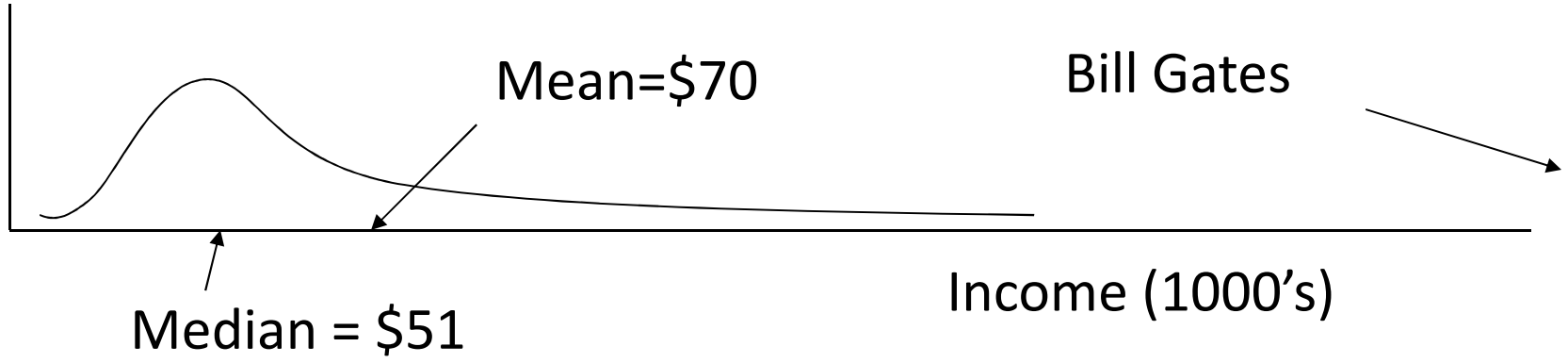
- Consider the two data sets:

  1 4 7 8 10: Median=7 Mean=6

  1 4 7 8 100: Median=7 Mean=24

Mean and Median Comparison

- The median is often nice to report when there are extreme values in the data
- Think about US household income (2012).



Summation Notation

1.4 Summation Notation

We are going to be doing a lot of adding up and averaging.

To talk about this in general we need some special notation.

First of all, we denote a general set of numbers by:

$$x_1, x_2, x_3, \ldots x_n$$

the first number

the last number of data points, or the "sample size".

In general the average of the numbers “x” is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

We call it “x bar”.

Let's look summation in more detail.

$$\sum_{i=1}^{n} x_i$$

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

add in this value for each i

Example: Suppose we have data x and y with n=4.

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

x	y	year
0.07	0.11	1
0.06	0.05	2
0.04	0.09	3
0.03	0.03	4

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

x	y	year
0.07	0.11	1
0.06	0.05	2
0.04	0.09	3
0.03	0.03	4

$$\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{4}\left[.07 + .06 + .04 + .03\right] = .05$$

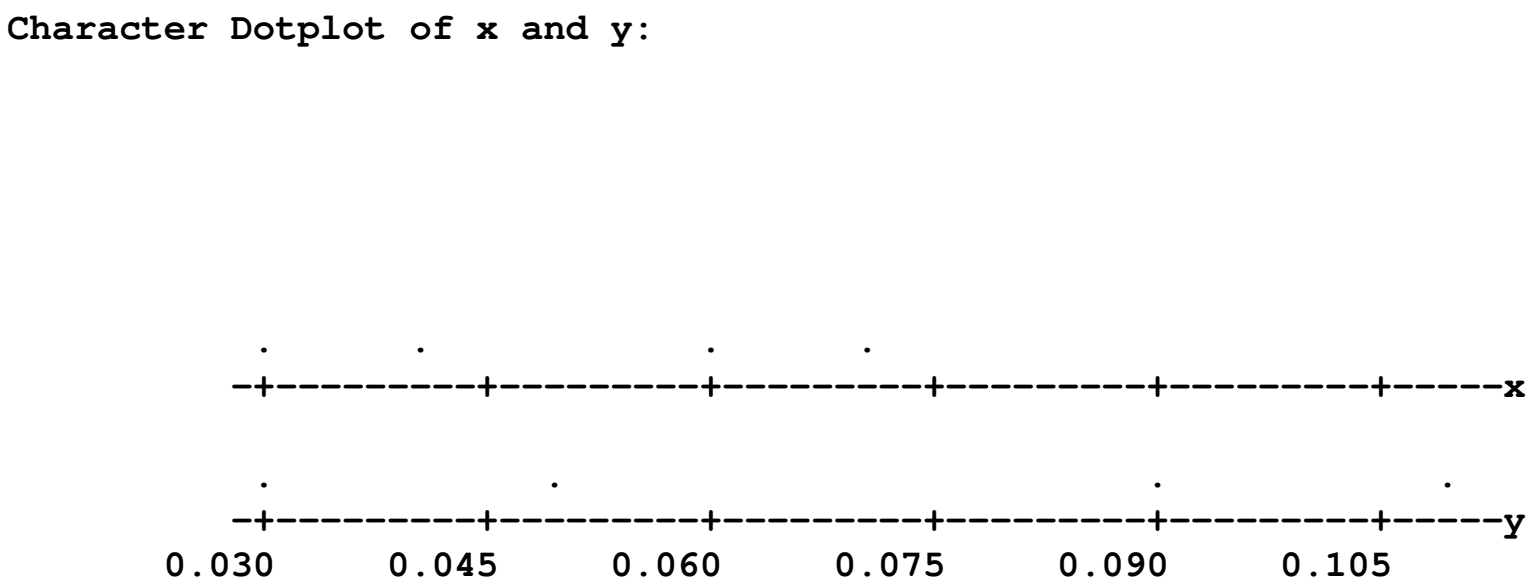
$$\bar{y} = \frac{1}{4}\sum_{i=1}^{4} y_i = \frac{1}{4}\left[.11 + .05 + .09 + .03\right] = .07$$

$$\frac{1}{4}\sum_{i=1}^{4}\left|x_i - \bar{x}\right| = \frac{1}{4}\left[\left|.07 - .05\right| + \left|.06 - .05\right| + \left|.04 - .05\right| + \left|.03 - .05\right|\right]$$

$$= \frac{1}{4}\left[.02 + .01 + .01 + .02\right] = .015$$

Sample Variance and Standard Deviation

1.5 The Sample Variance and Standard Deviation



The y numbers are more *spread out* than the x numbers.

We want a numerical measure of variation or spread.

To examine the variation of the data x, we look at:

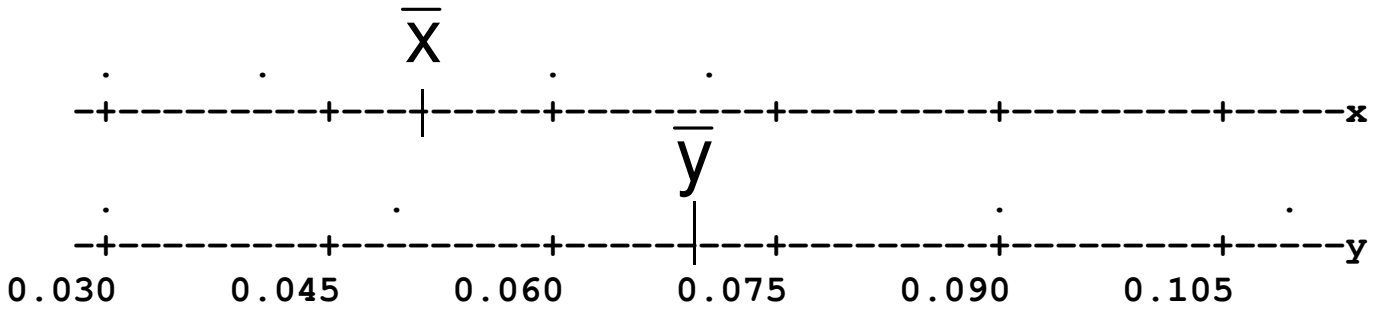
$$x_i - \overline{x}$$

A lot of variation means these numbers are big.

In our example,

x	$(x-\bar{x})$	y	$(y-\bar{y})$
0.07	0.02	0.11	0.04
0.06	0.01	0.05	-0.02
0.04	-0.01	0.09	0.02
0.03	-0.02	0.03	-0.04

Character Dotplot



Now we need an overall measure of how big the differences are.

We can't just sum them because the negative distances cancel out the positive ones.

We average the squared distances.

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

For technical reasons (we'll see why later) the sample variance of the data x is defined to be:

Sample Variance:

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

With large n, there is little difference between dividing by n and (n-1). *Think of it as the average squared distance from the mean*.

What is the smallest value a variance can be?

What are the units of the variance ??

It will be helpful to have a measure of spread which is in the original units.

The sample standard deviation is:

Sample Standard Deviation:

$$s_x = \sqrt{s_x^2}$$

The units of the standard deviation are the same as those of the original data.

Let's try these formulas on our example.

x	$(x-\bar{x})$	y	$(y-\bar{y})$
0.07	0.02	0.11	0.04
0.06	0.01	0.05	-0.02
0.04	-0.01	0.09	0.02
0.03	-0.02	0.03	-0.04

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \frac{1}{3}\left(.04^2 + (-.02)^2 + .02^2 + (-.04)^2\right)$$

$$= \frac{.004}{3} = .00133$$

$$s_y = \sqrt{.00133} = .0365$$

The sample standard deviation for the y data is bigger than that of the x data.

This numerically captures the fact that y has "more variation" about it's mean than x.

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

$$= \frac{1}{3}\left(.02^2 + .01^2 + (-.01)^2 + (-.02)^2\right)$$

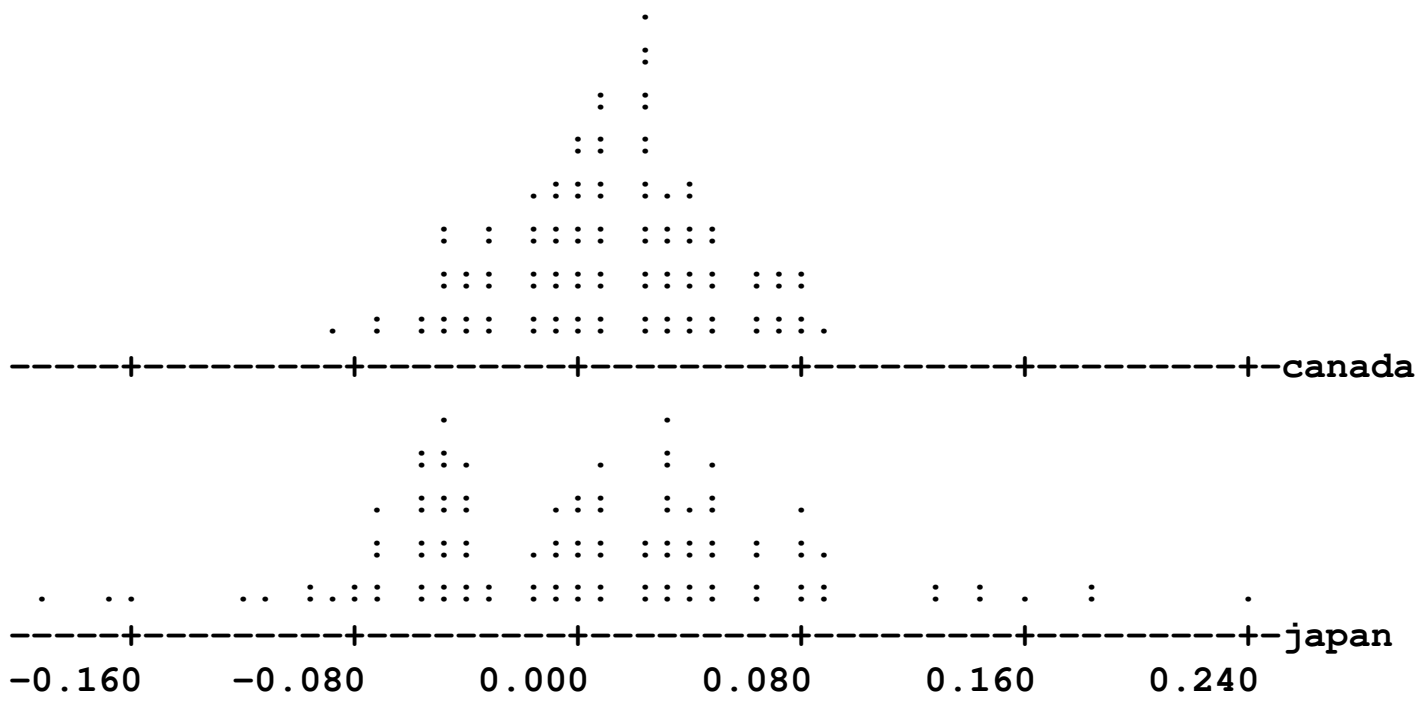
$$= \frac{.001}{3} = .000333$$

$$s_x = \sqrt{.00033} = .01826$$

Returns Example

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

Character Dotplot



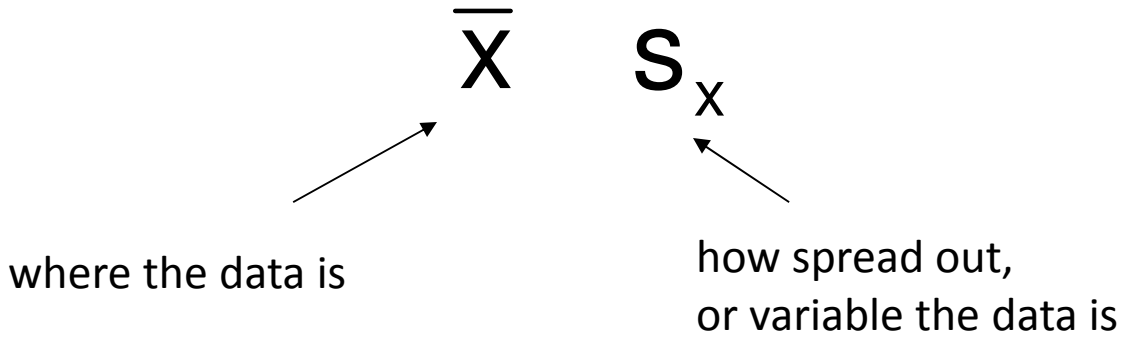
Summary measures for selected variables

	canada	japan
Count	107.00	107.00
Mean	0.009	0.002
Standard deviation	0.038	0.074

The Empirical Rule

1.6 The Empirical Rule

We now have the two summaries



The mean is pretty easy to understand.

What are the units?

We know that the bigger s_x is, the more variable the data is,
but how do we interpret the number ?
What is a big s_x , what is a small one ?
What are the units of s_x ?

The empirical rule will help us understand s_x and relate the summaries back to the dotplot (or 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

$$(\bar{x} - 2s_x, \bar{x} + 2s_x) = \bar{x} \pm 2s_x$$

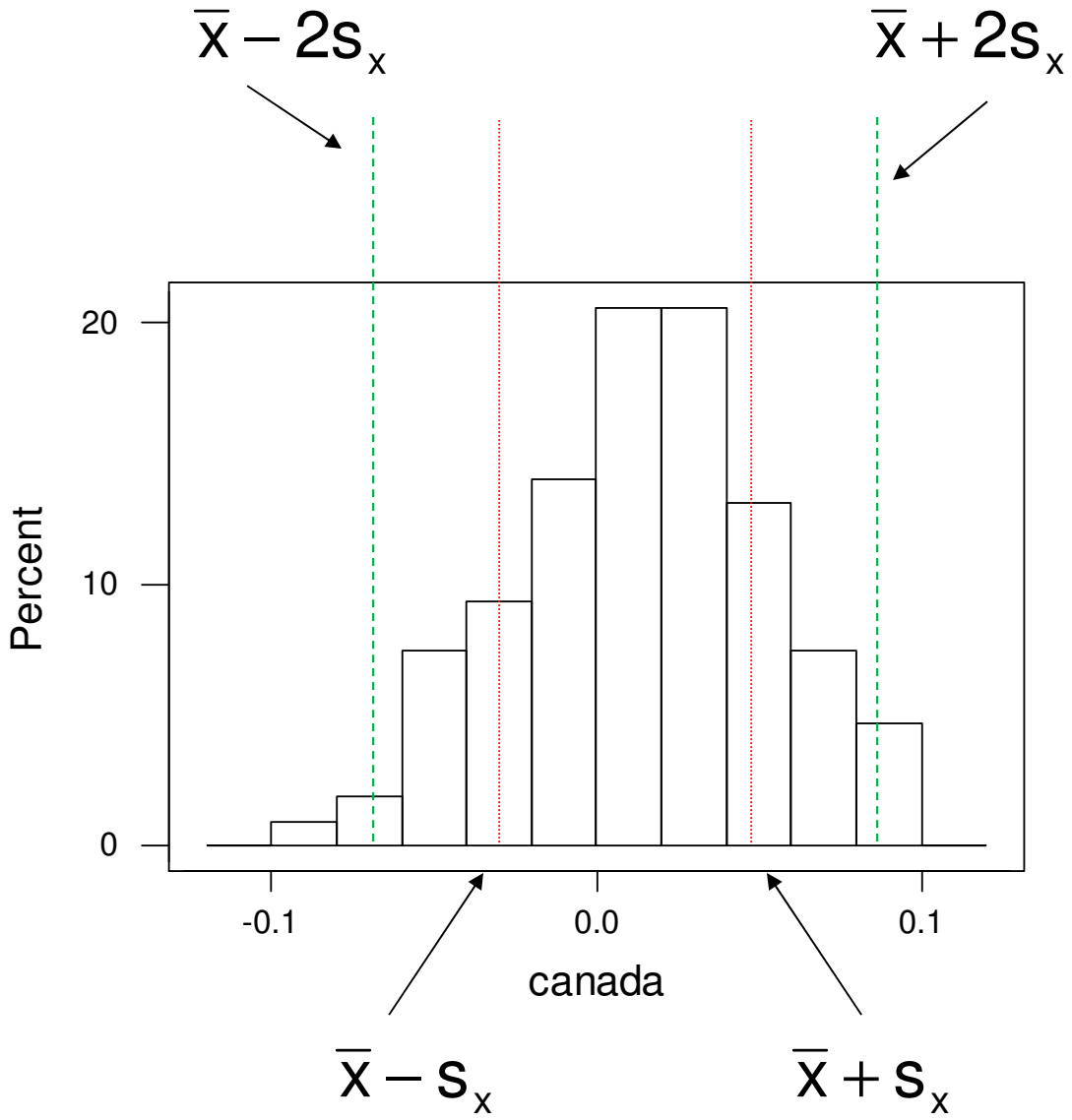
Let's see this with the Canadian returns.

$\bar{x} = .00907$

$s_x = .03833$

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

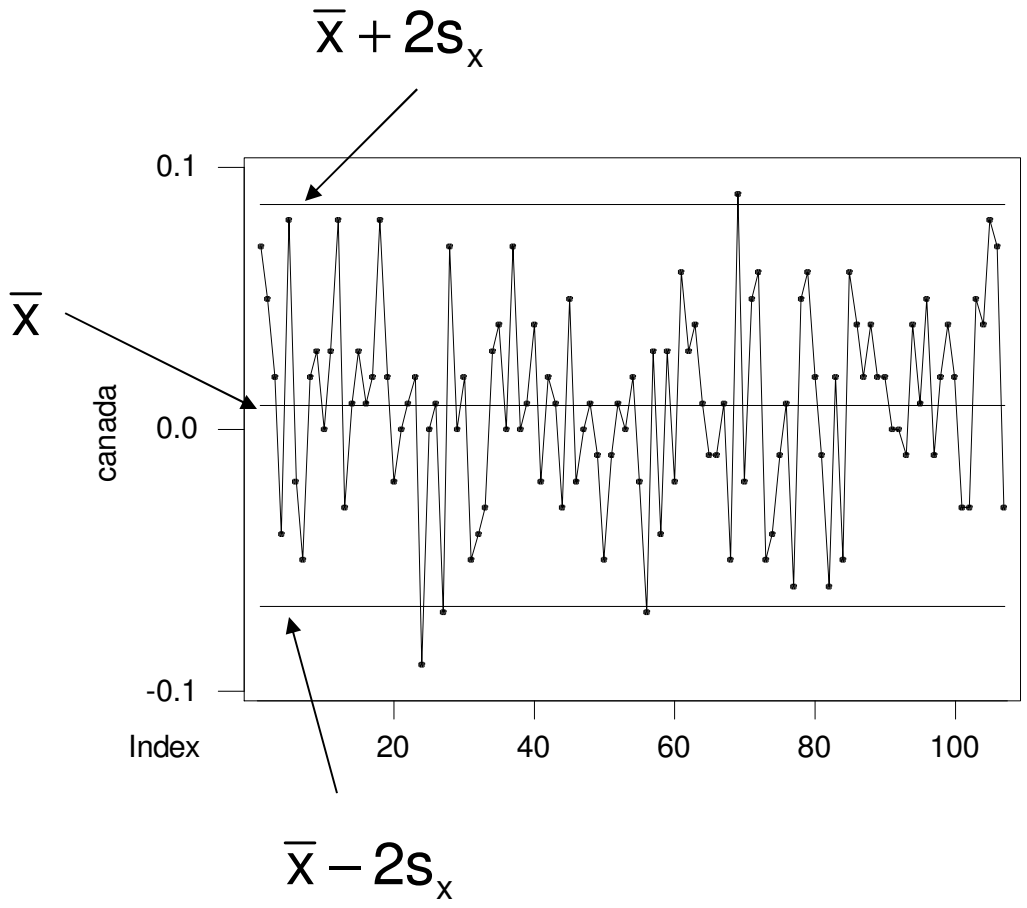
Looks reasonable.



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

n=108, so 5% outside would be about 5 points.

There are 4 points outside, which is pretty close.



Comparing Mutual Funds

Let's use means and sd's to compare mutual funds.
For 9 different assets we compute the mean and sd.
Then plot mean vs sd.

The assets are:

- Drefus (growth)
- Fidelity Trend fund (growth)
- Keystone Speculative fund (max capital gain)
- Putnam Income Fund (growth)
- Scudder Income
- Windsor Fund (growth)
- Equally Weighted Market
- Value weighted marktet
- tbill rate

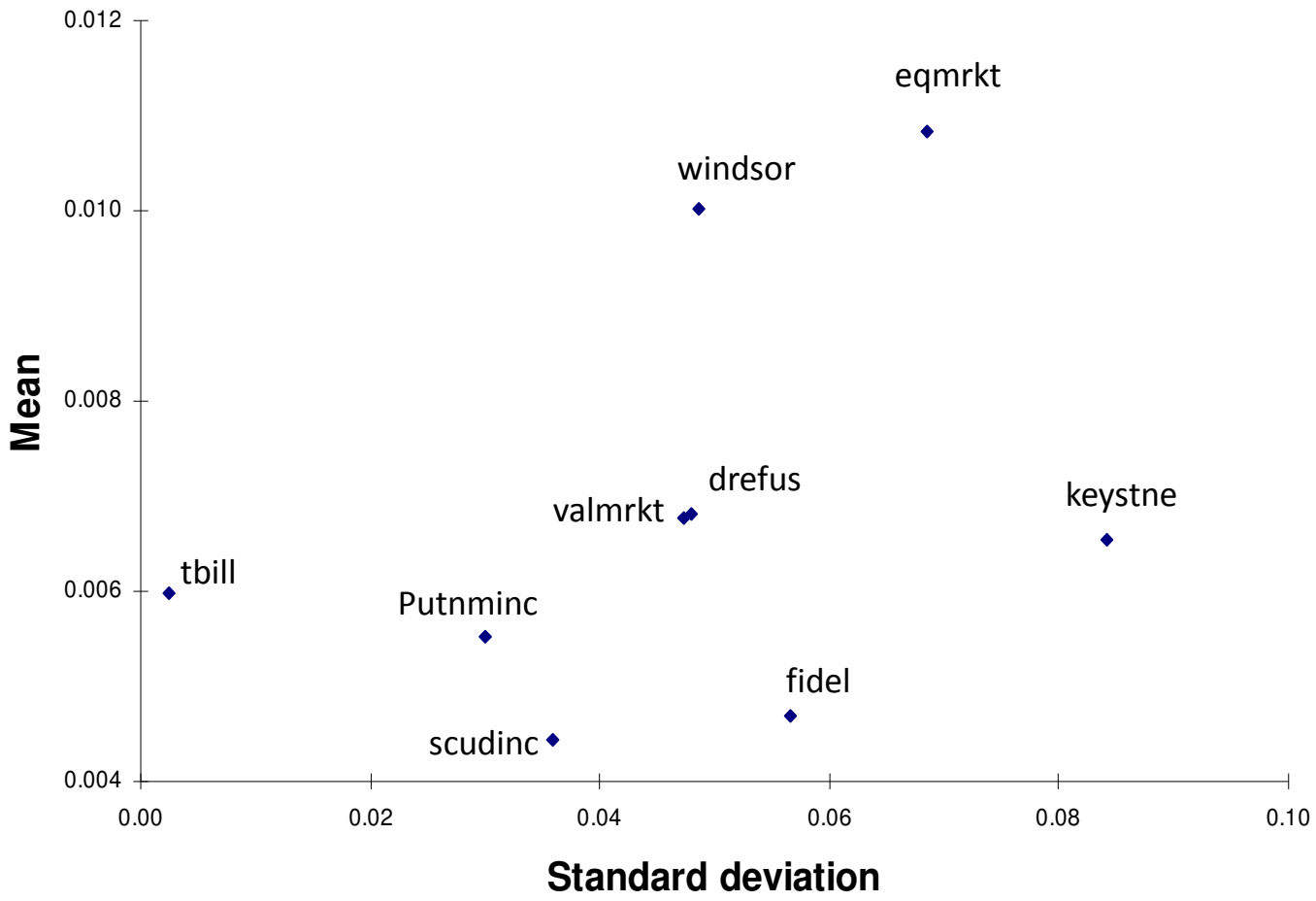
Summary measures for selected variables

	drefus	fidel	keystne	Putinc	scudinc	windsr	eqmrkt	valmrkt	tbill
Count	180	180	180	180	180	180	180	180	180
Mean	0.007	0.005	0.007	0.006	0.004	0.010	0.011	0.007	0.006
Standard deviation	0.047	0.057	0.084	0.030	0.036	0.049	0.069	0.048	0.003

It is considered good to have a large mean return and a small standard deviation.

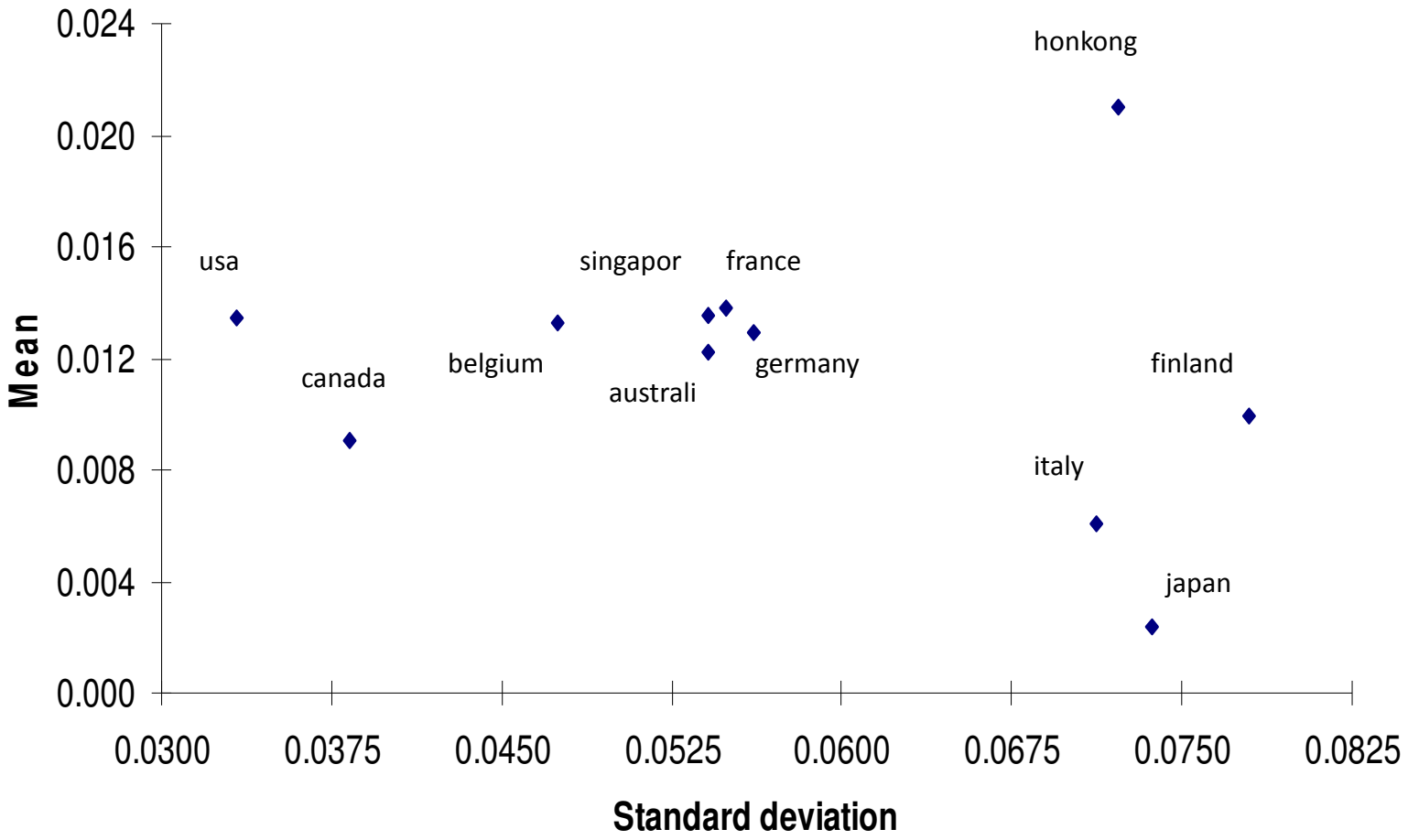
How does this graph relate back to the dotplots we saw before ??

If you're a fund manager, where do you want to be on this plot?



Let's compare some countries:

Based on monthly returns from '88 to '96



Covariance and Correlation

1.7 Covariance and Correlation

The mean and sd help us summarize a bunch of numbers which are measurements of just one thing.

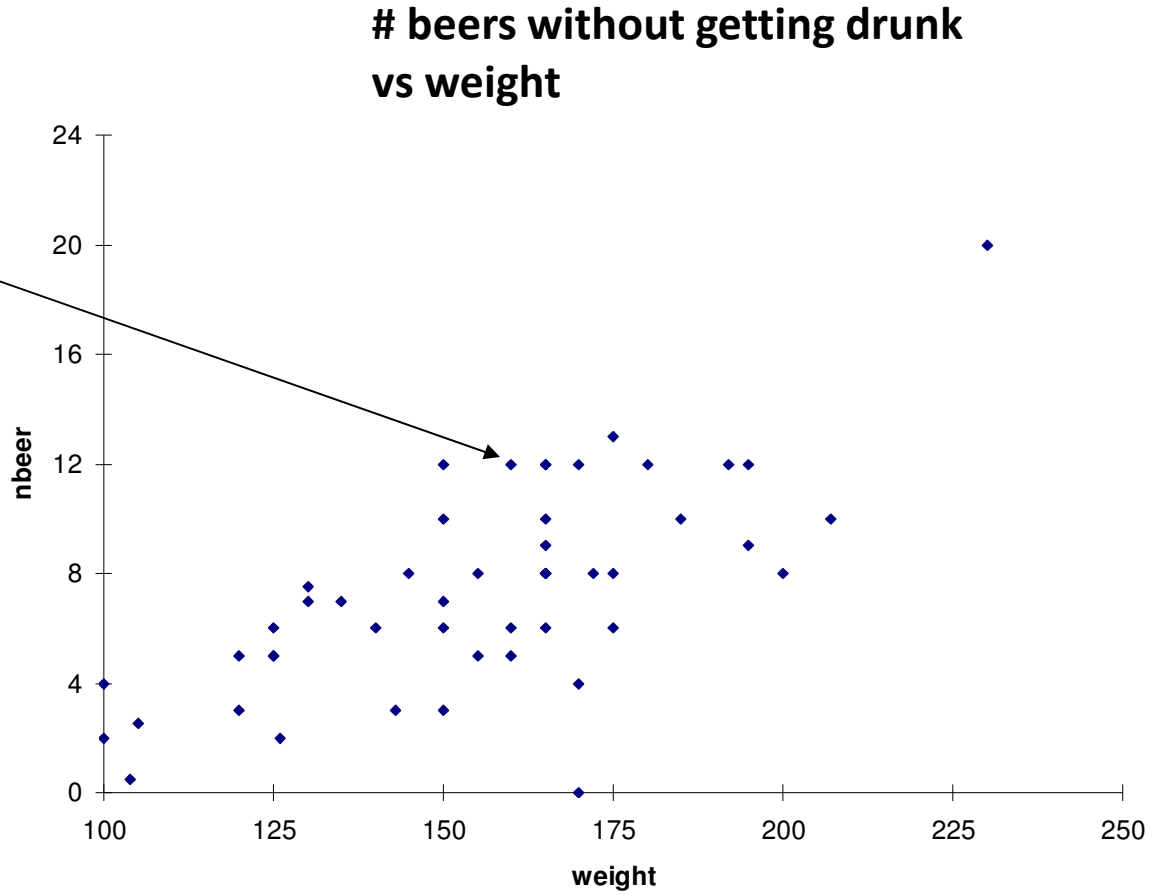
A fundamental and totally different question is how one thing relates to another.

We just used a scatterplot to look at two things: the mean and sd of different assets.

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

Example

nbeer	*weight*	*i*
12.0	192	1
12.0	160	2
5.0	155	3
5.0	120	4
7.0	150	5
13.0	175	6
4.0	100	7
12.0	165	8
12.0	165	9
12.0	150	10
.	.	.
.	.	.
.	.	.



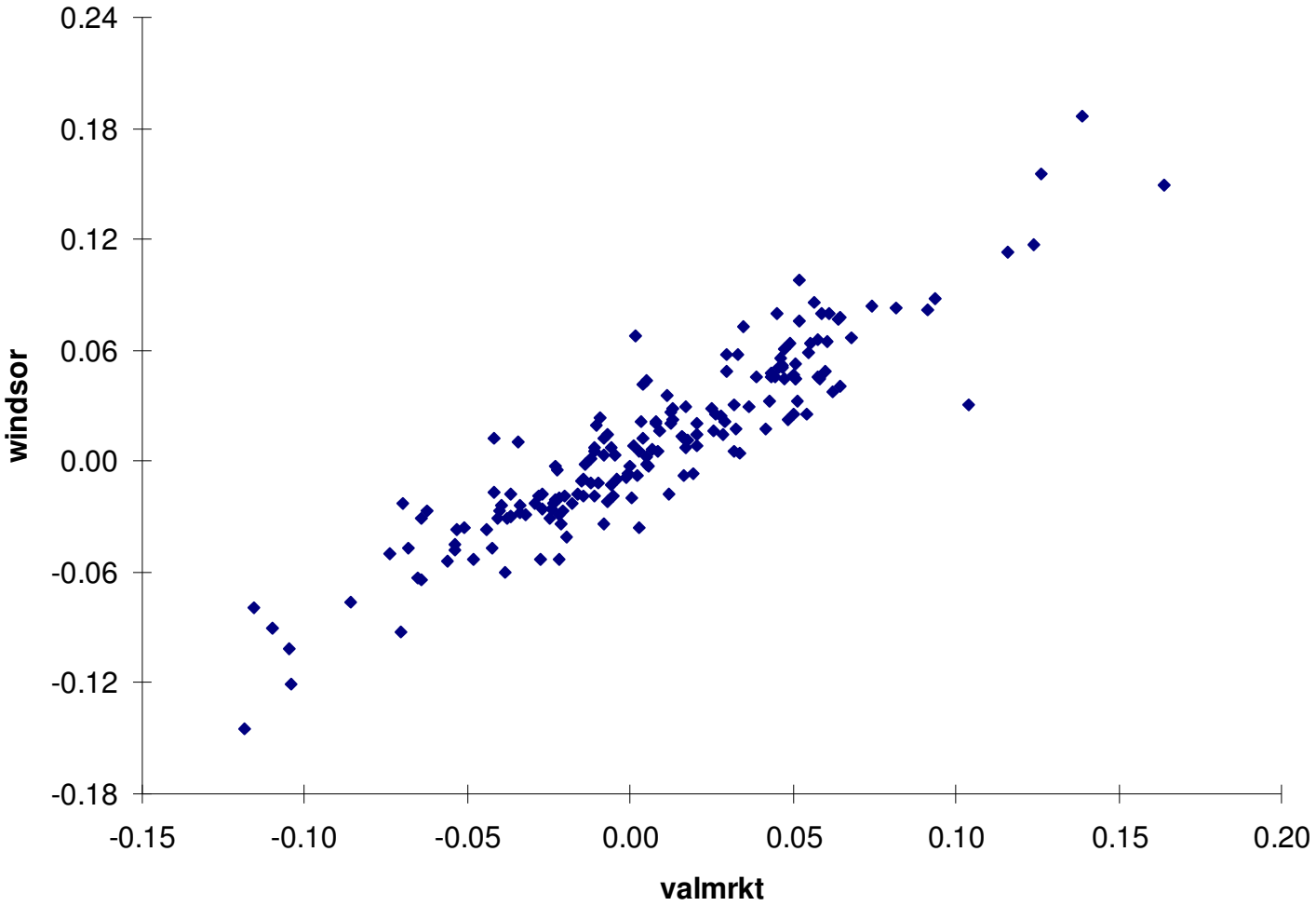
Now we think of each *pair* of numbers as an observation. Each pair corresponds to a person.

Each person has two numbers associated with him/her, #beers and weight.

Example

Are returns on a mutual fund related to market returns?

Each point corresponds to a month.



In general we have observations

(x_i, y_i) ← the ith observation is a pair of numbers

and each point on the plot corresponds to an observation.

Our data looks like:

x	y	i
12.0	192	1
12.0	160	2
5.0	155	3
5.0	120	4
7.0	150	5
13.0	175	6
4.0	100	7
12.0	165	8

.....

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

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.

Let me first introduce covariance and correlation then we will consider the interpretation and use of each.

The sample covariance between x and y is:

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

The sample correlation between x and y is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

so the correlation is just the covariance divided by the two standard deviations.

We'll get some intuition about these formulas, but first let's see them in action. How do they summarize data for us ? Correlation FACTS:

- $-1 \leq r_{xy} \leq 1$

- The closer r is to 1 the stronger the linear relationship is with a positive slope. That is, the largest values of y tend to be associated with largest value of x (and vice-versa).

- The closer r is to -1 the stronger the linear relationship is with a negative slope. That is, large values of y tend to be associated with *small* value of x (and vice-versa).

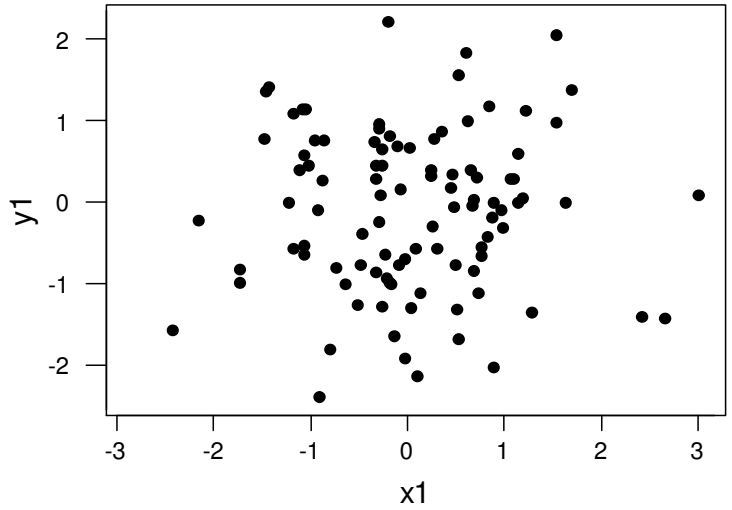
The correlations corresponding to the two scatterplots we looked at are:

```
Correlation of valmrkt and windsor = 0.923

Correlation of nbeer and weight = 0.692
```

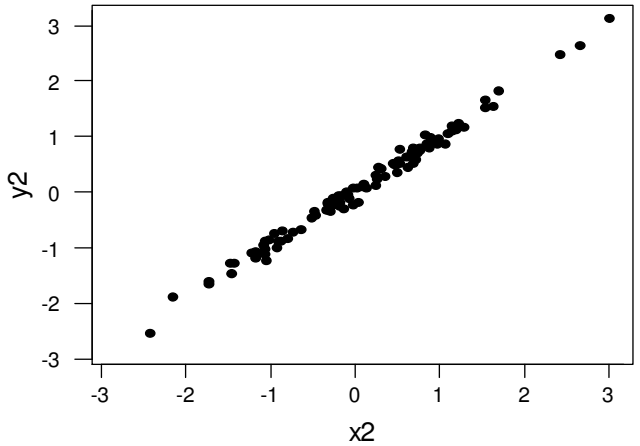
The larger correlation between valmrkt and windsor indicates that the linear relationship is stronger.

Let's look at some more examples.

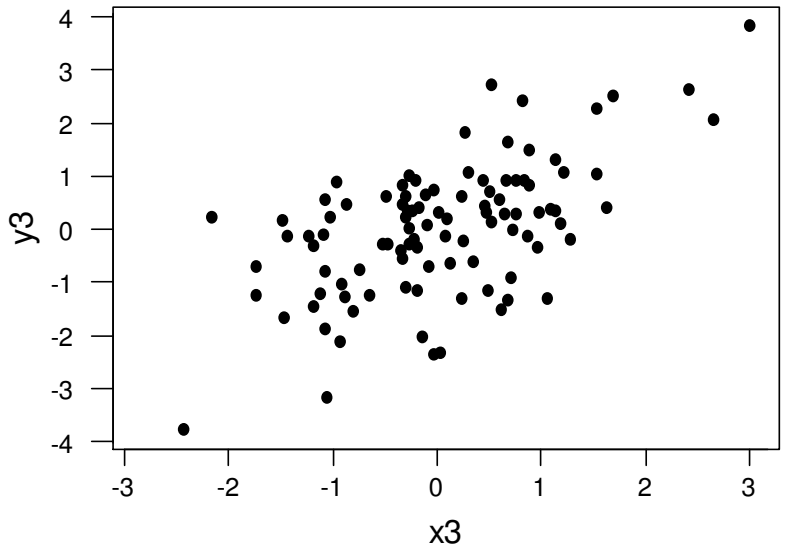
Correlation of
y1 and x1 = 0.019



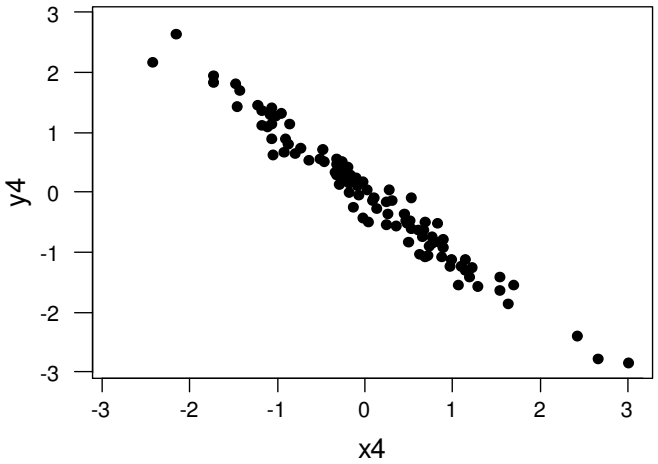
Correlation of
y2 and x2 = 0.995



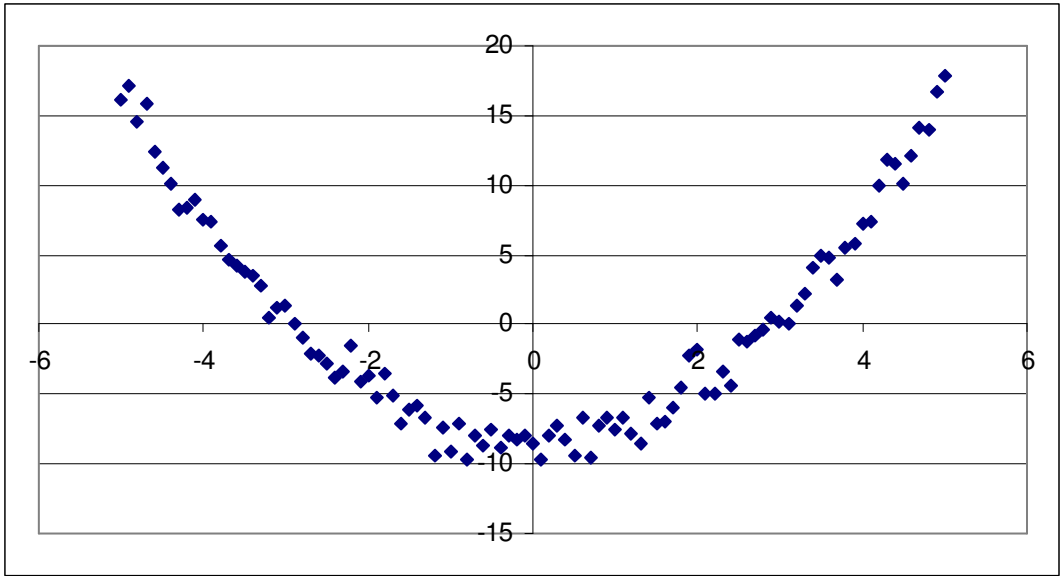
Correlation of
y3 and x3 = 0.586



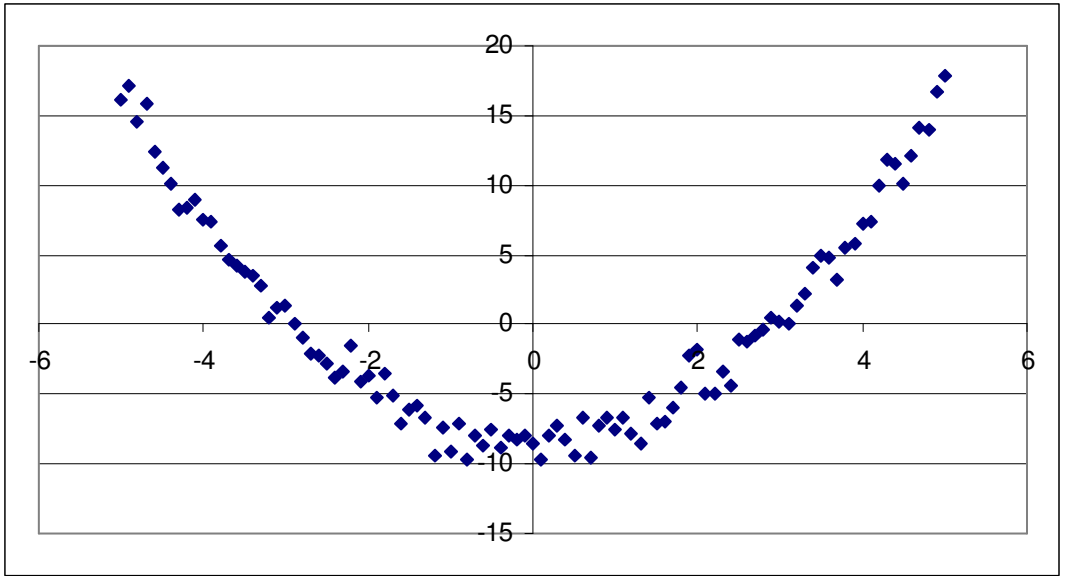
Correlation of
y4 and x4 = −0.982



What's the correlation here???



Correlation of x and y = 0.020 (basically 0)



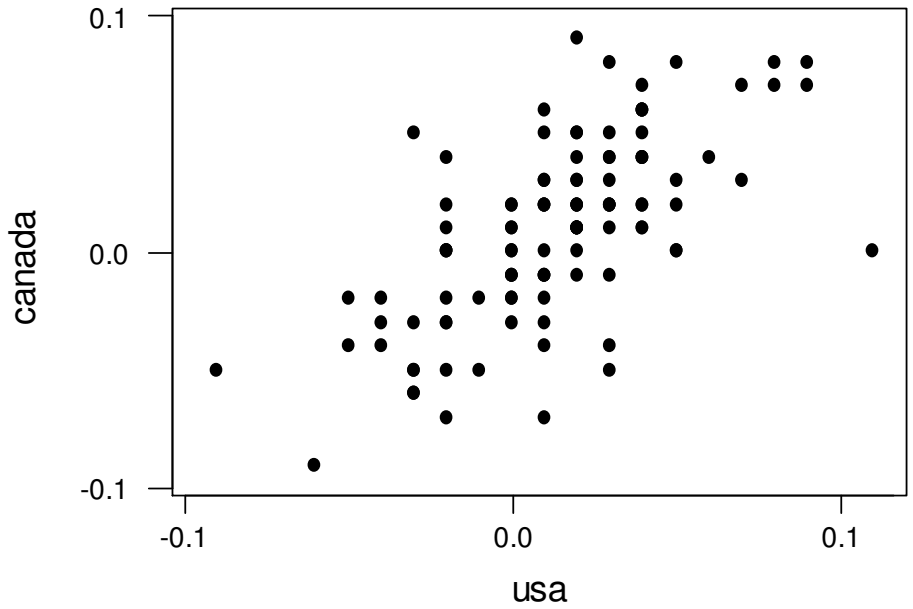
Correlation only measures the *linear* relationship.

Example: The Countries Data

Which countries go up and down together ?

I have data on 23 countries.

That would be a lot of plots !!!

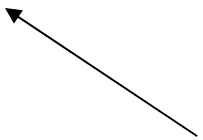


To summarize we can compute all pairwise correlations.

	australi	belgium	canada	finalnd	france	germany	honkong	italy
belgium	0.189							
canada	0.507	0.357						
finalnd	0.387	0.183	0.386					
france	0.275	0.734	0.342	0.176				
germany	0.226	0.691	0.302	0.304	0.709			
honkong	0.334	0.301	0.558	0.355	0.359	0.339		
italy	0.159	0.367	0.334	0.389	0.352	0.465	0.261	
japan	0.25	0.418	0.271	0.307	0.421	0.318	0.219	0.426
usa	0.360	0.429	0.651	0.264	0.501	0.372	0.429	0.240
singapor	0.409	0.355	0.478	0.391	0.408	0.467	0.647	0.416

	japan	usa
usa	0.246	
singapor	0.407	0.473

why is this blank ?



This got “wrapped” around

Understanding the Cov and Corr Formulas

Why are the formulas for cov and corr capturing the relationship?

To get a feeling for this, let's go back to the simple example and compute the correlation.

x	y
0.07	0.11
0.06	0.05
0.04	0.09
0.03	0.03

First let's compute the covariance:

x	$(x-\bar{x})$	y	$(y-\bar{y})$
0.07	0.02	0.11	0.04
0.06	0.01	0.05	-0.02
0.04	-0.01	0.09	0.02
0.03	-0.02	0.03	-0.04

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

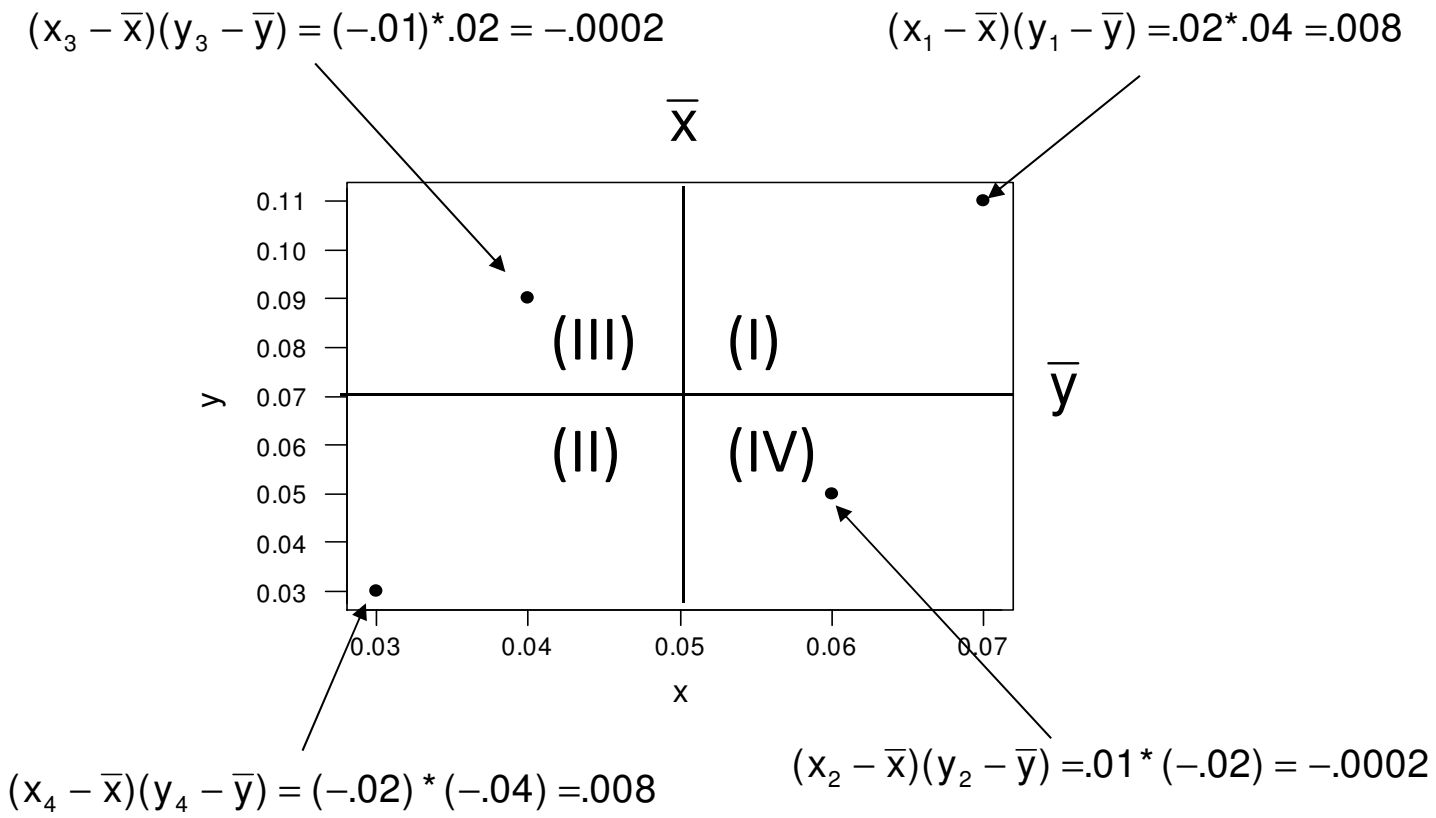
$$\frac{1}{3}((.07-.05)(.11-.07)+(.06-.05)(.05-.07)+(.04-.05)(.09-.07)+(.03-.05)(.03-.07))$$

$$=\frac{1}{3}(.02*.04+.01*(-.02)+(-.1)*.02+(-.02)*(-.04))$$

$$=\frac{1}{3}(.0008-.0002-.0002+.0008)=\frac{1}{3}(.0012)=.0004$$

= .0004

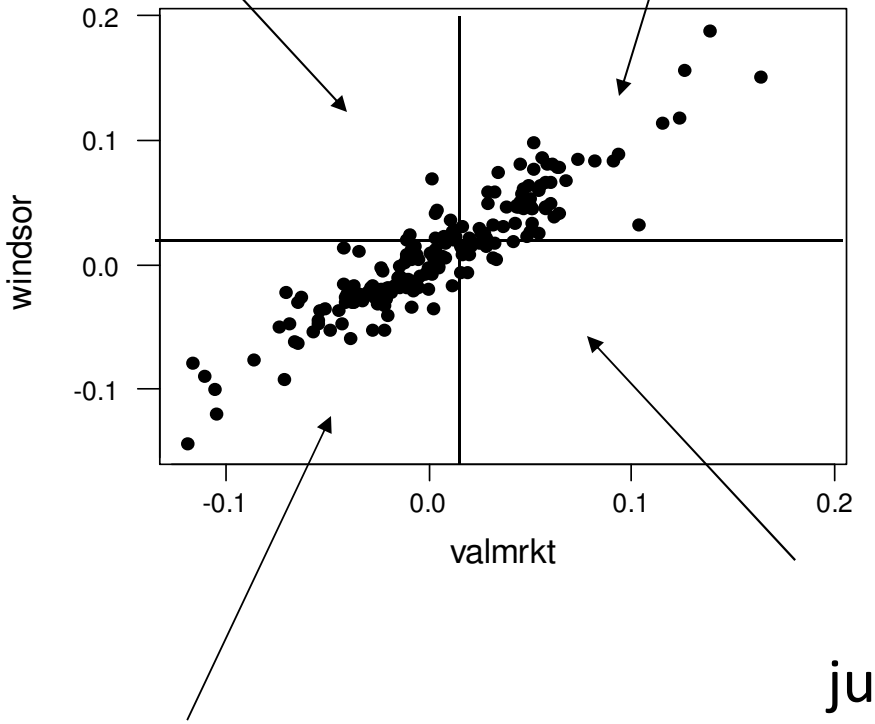
Each of the 4 points makes a contribution to the sum.
Let's see which point does what.



Points in (I) have both x and y bigger than their mean
so we get a positive contribution.
Points in (II) have both less than their mean so we get a positive contribution.
In (III) and (IV) one of x and y is less than its mean and the other is greater
so we get a negative contribution.
The farther out the point is, the bigger the contribution.

just a few
relatively small
contributions

Lots of positive contributions



Lots of positive contributions

just a few
relatively small
contributions

So,

A *positive covariance* means that when one variable
is above it's average the other one tends to be as well.
They move up and down together.

A *negative covariance* means that when one is up
the other tends to be down.
They move in opposite direction.

to finish the example,

$$r_{xy} = \frac{.0004}{(.0365)(.0183)} = .6$$

The division by the standard deviations standardizes the covariance so that the correlation is always between +/- 1.

What are the units of the covariance?

What are the units of the correlation ?

Linear Functions

1.8 Linear Functions

In this section of the notes we study the situation where two variables are *exactly* linear related.

That is if I give you the value of x you know exactly What the value of y must be.

Example

Suppose we have these temps in Celsius and Fahrenheit.

cel	fahr
10	50
15	59
20	68
25	77
40	104
30	86
50	122
70	158

How are the F values related to the C values?

$$F = 32 + (9/5)C$$

1.9 Mean and Variance of a Linear Function

Suppose y is a linear of function of x.

How are the mean and variance (standard deviation) of y related to those of x ?

Let's look at our temperature example.

Suppose we first multiply by (9/5) and then add 32.

Information on the Worksheet

Summary measures for selected variables

	cel	fahr
Count	8.000	8.000

Create a new column, 'mul', in Excel

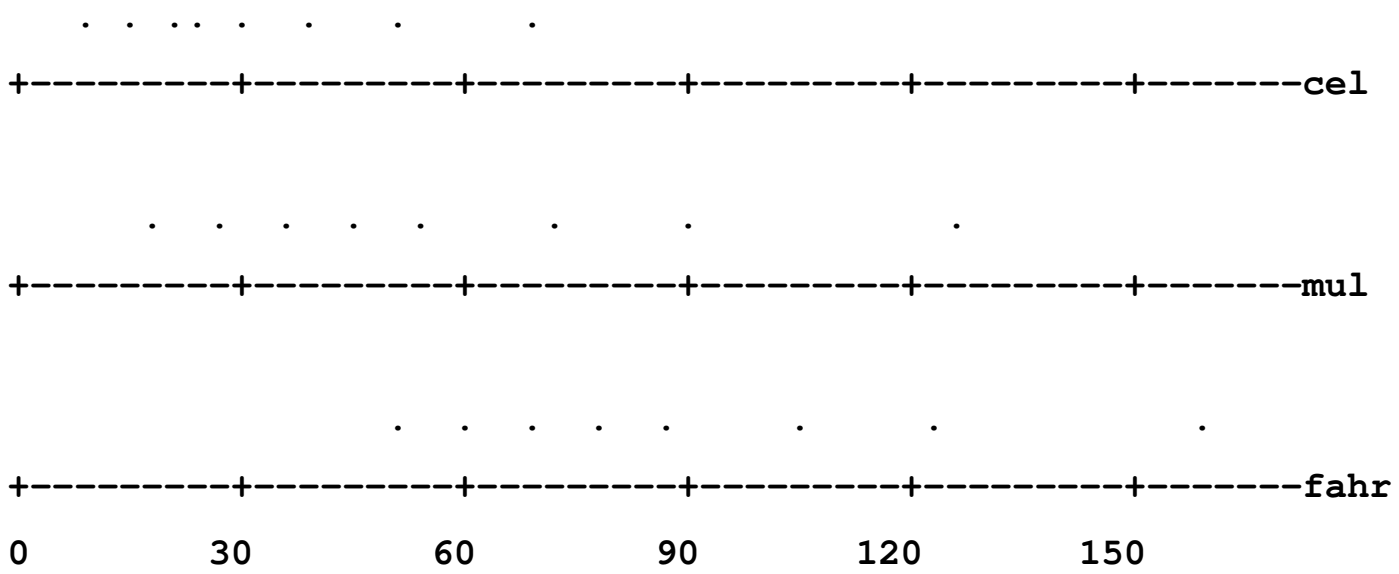
=(9/5)*cel

Let’s get some intuition about what happens when we multiply and add numbers to a data set to obtain a new data set.

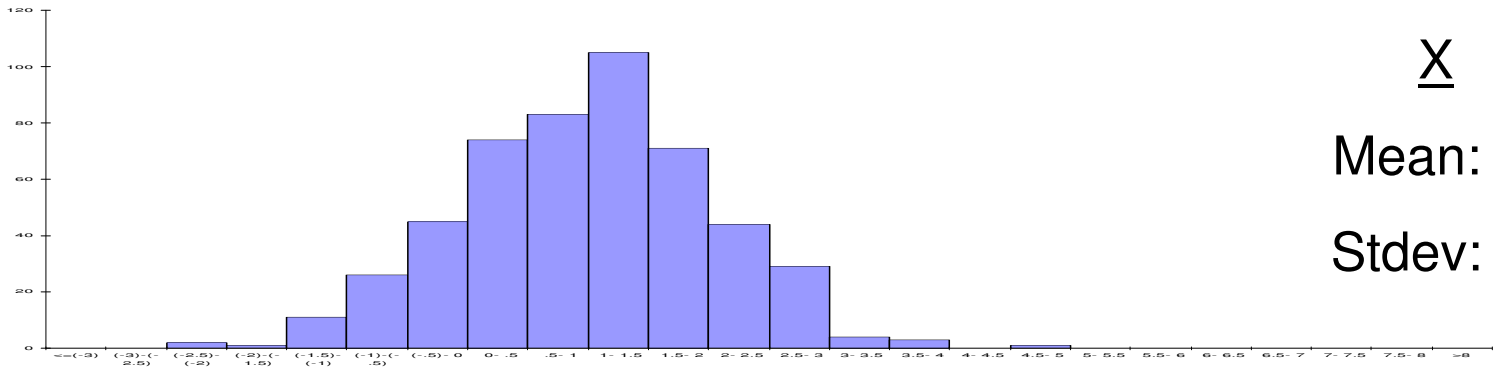
Celsius

Mul=(9/5)C

F = 32 + (9/5)C



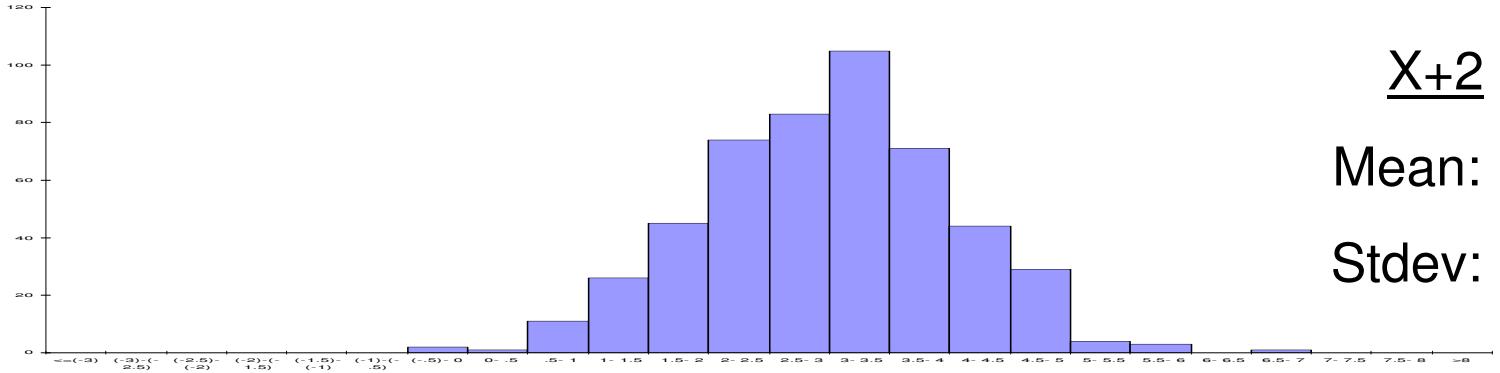
	cel	fahr	mul
Mean	32.500	90.500	58.500
Standard deviation	20.000	36.000	36.000



X

Mean: 1

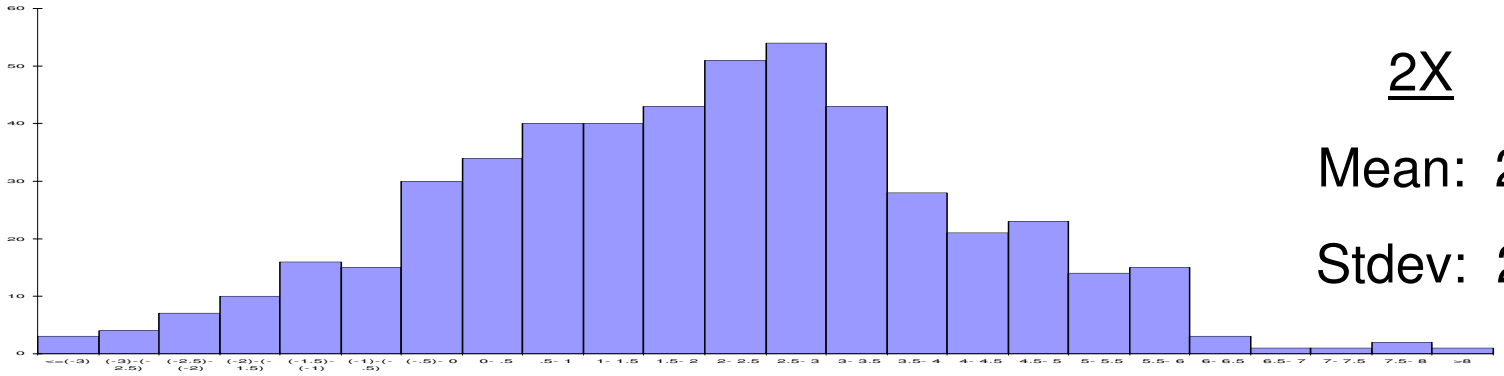
Stdev: 1



X+2

Mean: 3

Stdev: 1



2X

Mean: 2

Stdev: 2

Suppose

$$y = c_0 + c_1 x$$

then,

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

$$s_y = |c_1| s_x$$

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

Example

- We convert from fractions to percent by taking $y = 100x$

What is mean of y compared to $\overline{x}$?

What is the standard deviation of y compared to s_x ?

- We convert from $ to 1000's of $ by taking $y = \frac{1}{1000}x$

Now what are the mean and standard deviation of y?

Why?

$$y_i = c_0 + c_1 x_i$$

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

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} (c_0 + c_1 x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} c_0 + \frac{1}{n}\sum_{i=1}^{n} c_1 x_i$$

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

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

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (c_0 + c_1 x_i - (c_0 + c_1 \bar{x}))^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} (\cancel{c_0} + c_1 x_i - \cancel{c_0} - c_1 \bar{x})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} c_1^2 (x_i - \bar{x})^2 \quad = \quad c_1^2 s_x^2$$

Why are these formulas useful?

We could always just type everything into a spreadsheet and use spreadsheet functions to get the answers.

Really, though, the reason for these formulas will become apparent when we study probability, statistical inference, and regression. ***You cannot understand statistics or regression without a solid understanding of linear relationships.***

In other words, yes, I recognize these formulas are probably the least fun part of the course (and considering this is basic stats, that's saying something). But you absolutely *must* know them.

Example

Suppose x has mean 100 and standard deviation 10.

What are the mean, standard deviation and variance of:

(i) $y = 2x$? ($c_0=0$, $c_1=2$)

(ii) $y = 5+x$? ($c_0=5$, $c_1=1$)

(iii) $y = 5-2x$? ($c_0=5$, $c_1= -2$)

Linear Combinations and Portfolios

1.10 Linear Combinations

When a variable y is linearly related to several others, we call it a *linear combination*.

$$y = c_0 + c_1x_1 + c_2x_2 + \ldots c_kx_k$$

y is a linear combination of the x's.
c_i is the coefficient of x_i.
We have k data sets (notation has different meaning here).

1.11 Portfolios and Linear Combinations

Suppose you have $100 to invest.

Let x_1 be the return on asset 1.
If $x_1 = .1$, and you put all your money into asset 1
you will have $110 at the end of the period.

Let x_2 be the return on asset 2.
If $x_2 = .15$, and you put all your money into asset 2
you will have $115 at the end of the period.

Suppose you put 1/2 your money into 1 and 1/2 into 2.
What will happen ?

At the end of the period you will have

E=E1+E2

where E1 is the end of period wealth from your investment in asset 1 and E2 is the end of period wealth from your investment in asset 2.

$$E=\underbrace{(1+.1)*.5*100}_{E1}+\underbrace{(1+.15)*.5*100}_{E2}$$

Hence the return on the total investment of $100 is

$$E=(1+r)*B=(1+\underbrace{.5*.1+.5*.15}_{r})*100$$

r=.5*.1 + .5*.15=.125.

To generalize, let w_1 be the fraction of your wealth you invest in asset 1.
Let w_2 be the fraction of your wealth you invest in asset 2.
Let B be the wealth.

The w's are called the portfolio weights. What are the restrictions on the weights?

Then at the end of the period you have:

$$E = (1+x_1)w_1B + (1+x_2)w_2B = \left[(w_1 + x_1w_1) + (w_2 + x_2w_2)\right]B$$

$$= \left[1 + \underbrace{x_1w_1 + x_2w_2}_{r_p}\right]B$$

Hence the portfolio return is,

$$r_p = w_1X_1 + w_2X_2$$

Suppose we have k assets.

The return on the ith asset is x_i.

Put w_i fraction of your wealth into asset i.

Your portfolio is determined by the portfolio weights w_i.

Then the return on the portfolio is:

$$R_p = \sum_{i=1}^{k} w_i x_i$$

What is return on the equally weighted portfolio?

Example (the country data again)

Let's use our country data and suppose that we had put .5 into usa and .5 into Hongkong.

What would our returns have been ?

In Excel, create a new column, entitled 'port':

```
= .5*(honkong) + .5*(usa)
```

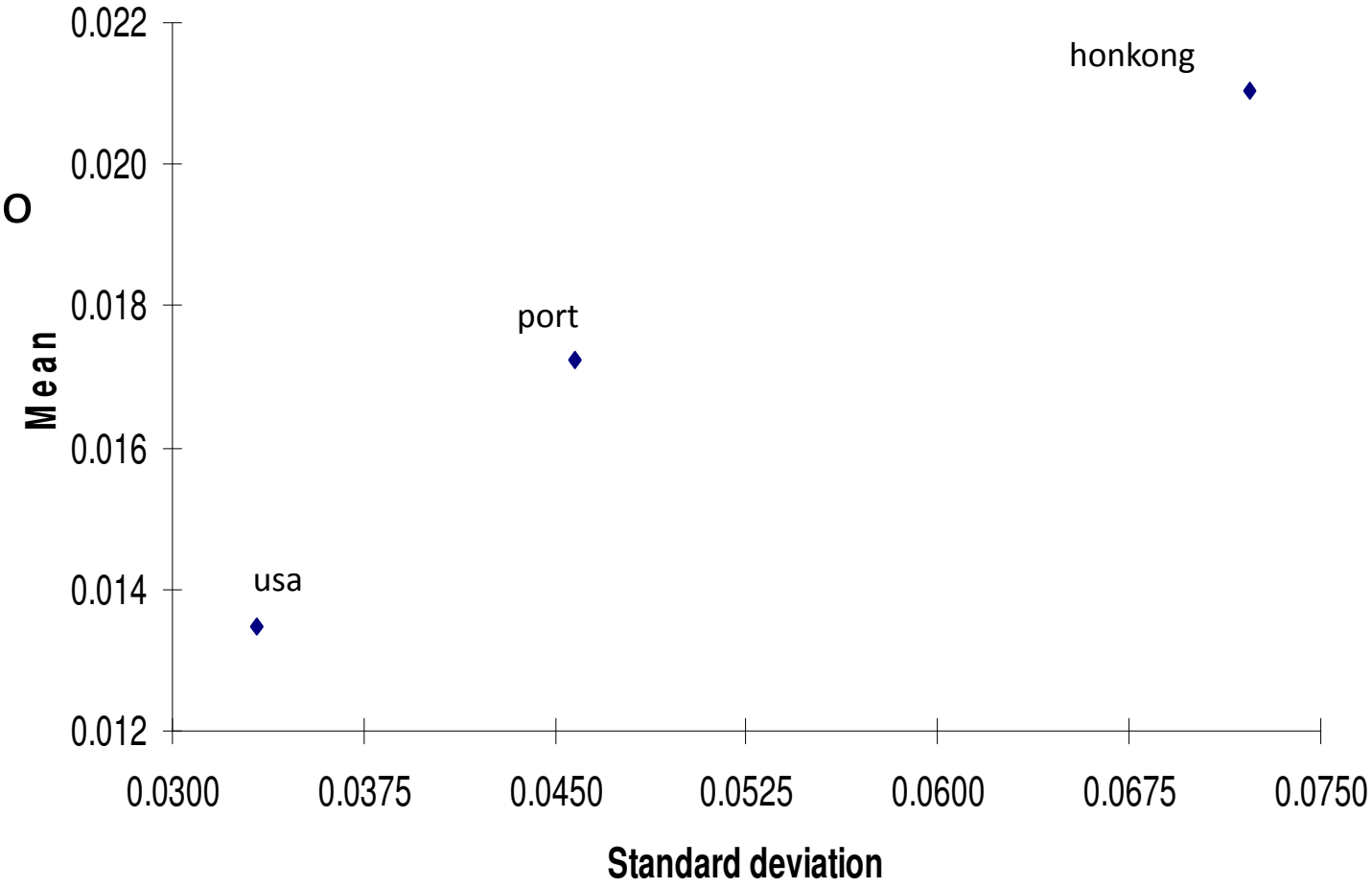
honkong	usa	port
0.02	0.04	0.030
0.06	-0.03	0.015
0.02	0.01	0.015
-0.03	0.01	-0.010
0.08	0.05	0.065
.....		

for each month, we get the portfolio return as 1/2 hongkong + 1/2 usa.

How do the returns on this portfolio compare with those of hongkong and usa?

It looks like the mean for my portfolio is right in between the means of usa and hongkong.

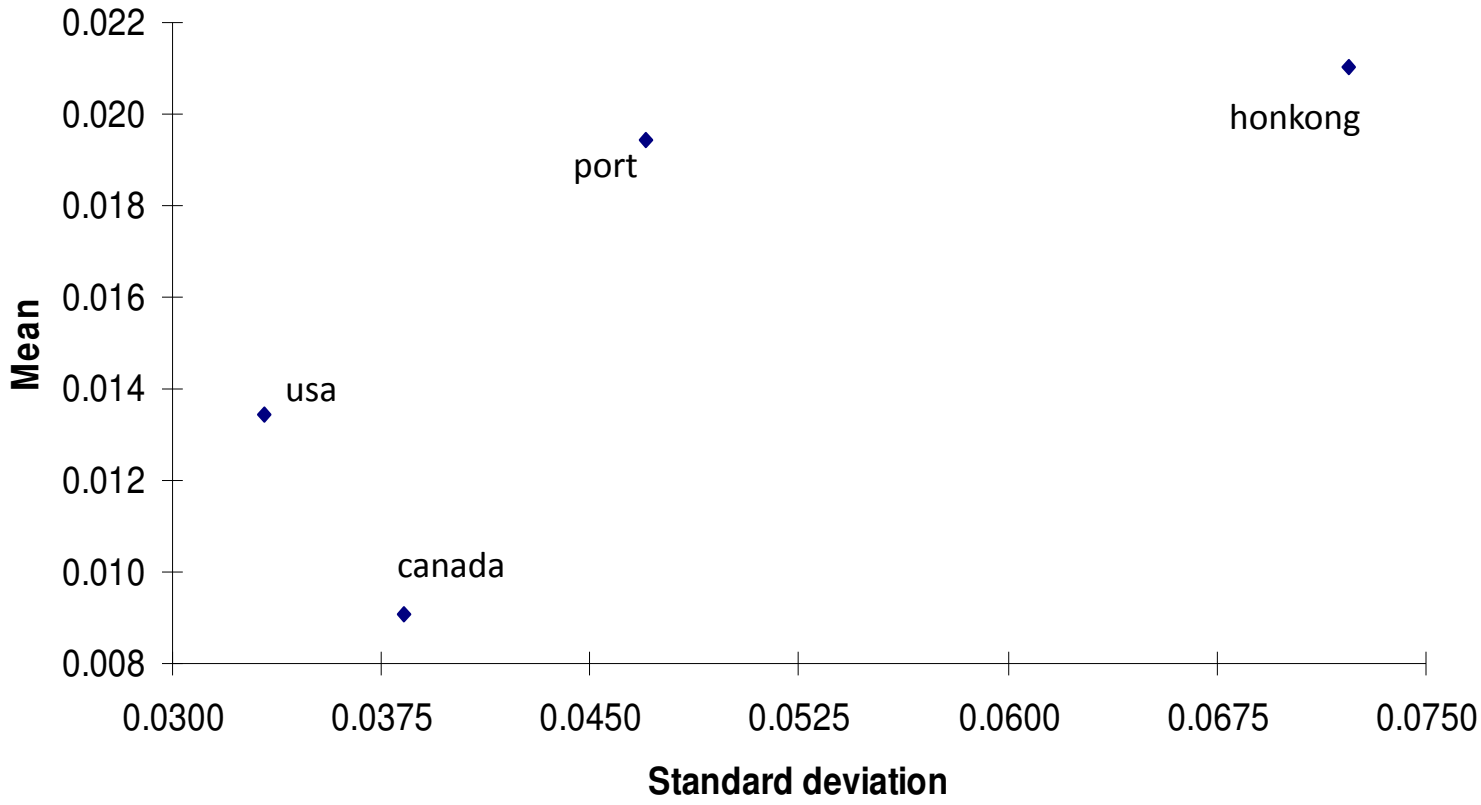
What about the sd?



Let's try a portfolio with three stocks.
The weights must add up to one, but they can be negative, this is called going short in the asset.

```
= -.5*(canada)+ usa +.5*(honkong)
```

Clearly, forming portfolios is an interesting thing to do!!



Why would we form a portfolio?

Maybe the portfolio has a nice mean and variance.

It would be nice to be able to figure out what the mean and variance are of any portfolio as a function of the portfolio weights.

We need some basic equations to relate means and variances of the assets to means and variances of the different possible portfolios obtained by different weights.

There are some basic formulas that relate the mean sd of a linear combination to the means, variances, *and covariances* of the input x variables.

1.12 Mean and Variance of a Linear Combination

First we consider the case where we have two inputs.

Suppose

$$y = c_0 + c_1 x_1 + c_2 x_2$$

then,

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

$$s_y^2 = c_1^2 s_{x_1}^2 + c_2^2 s_{x_2}^2 + 2 c_1 c_2 s_{x_1 x_2}$$

Example

```
= .5*(honkong) + .5*(usa)
```

honkong	usa	port
0.02	0.04	0.030
0.06	−0.03	0.015
0.02	0.01	0.015
−0.03	0.01	−0.010
0.08	0.05	0.065
.....		

for each month, we
get the portfolio return
as 1/2 hongkong + 1/2 usa.

The mean returns on port, usa, and honkong are
.01724, .01346, and .02103.

.01724 = .5*.01346+.5*.02103

off diagonals are covariances

Table of covariances (variances on the diagonal)

	honkong	usa	port
honkong	0.00521497		
usa	0.00103037	0.00110774	
port	0.00312267	0.00106906	0.00209586

.0021
= (.5)*(.5)*.00521 + (.5)*(.5)*.00111 + 2*(.5)*(.5)*.001
= .25*.00521 + .25*.00111 + .5*.001

Let's do one more:

```
= .25*(usa)+.75*(honkong)
```

Table of covariances (variances on the diagonal)

	honkong	usa	port
honkong	0.00521497		
usa	0.00103037	0.00110774	
port	0.00416882	0.00104972	0.00338905

.0033 =
(.25)*(.25)*.00111 + (.75)*(.75)*.0052+(2)*(.25)*(.75)*(.00103)

- The variance of the portfolio depends upon the variances of the individual assets as well as the covariance.
- The covariance tells us something about the relationship between the returns on the individual assets.
- Why should this relationship matter?
- We'll look at several examples to get a feel.

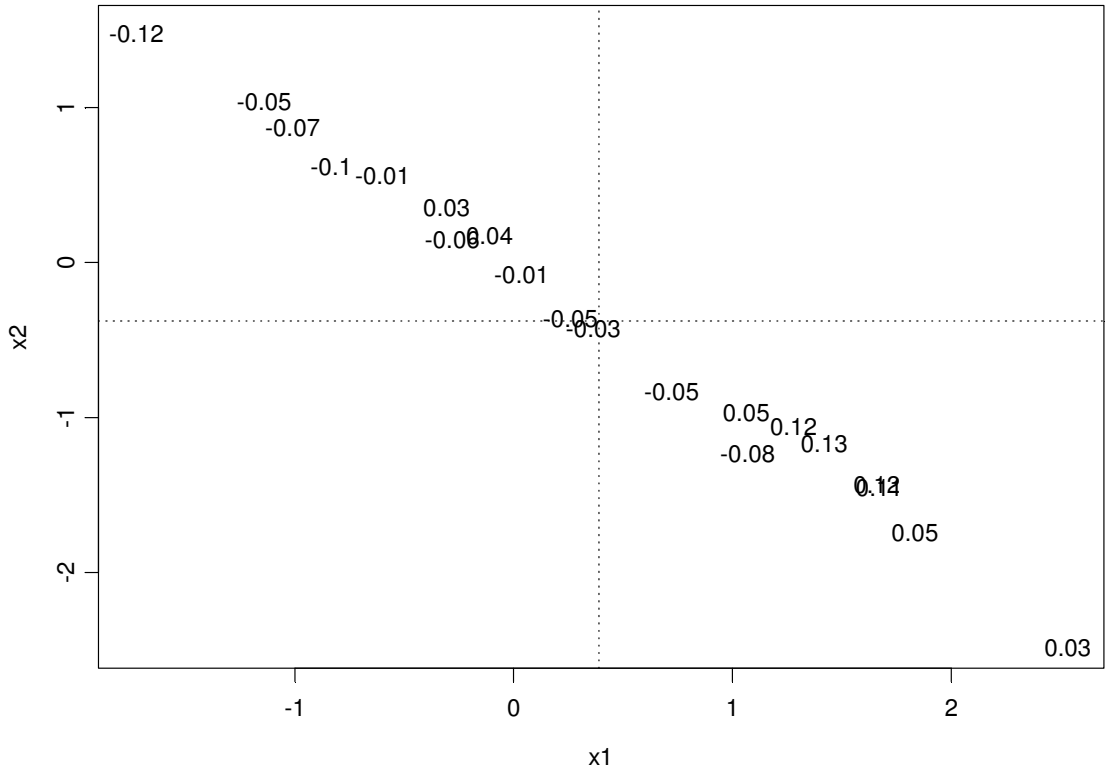
Example

$y = .5x_1 + .5\ x_2$

At each point we plot the value of y.

The variances and covariance are:

```
       x1         x2
x1  1.334636
x2 -1.208679  1.106238
```



The variance of y is

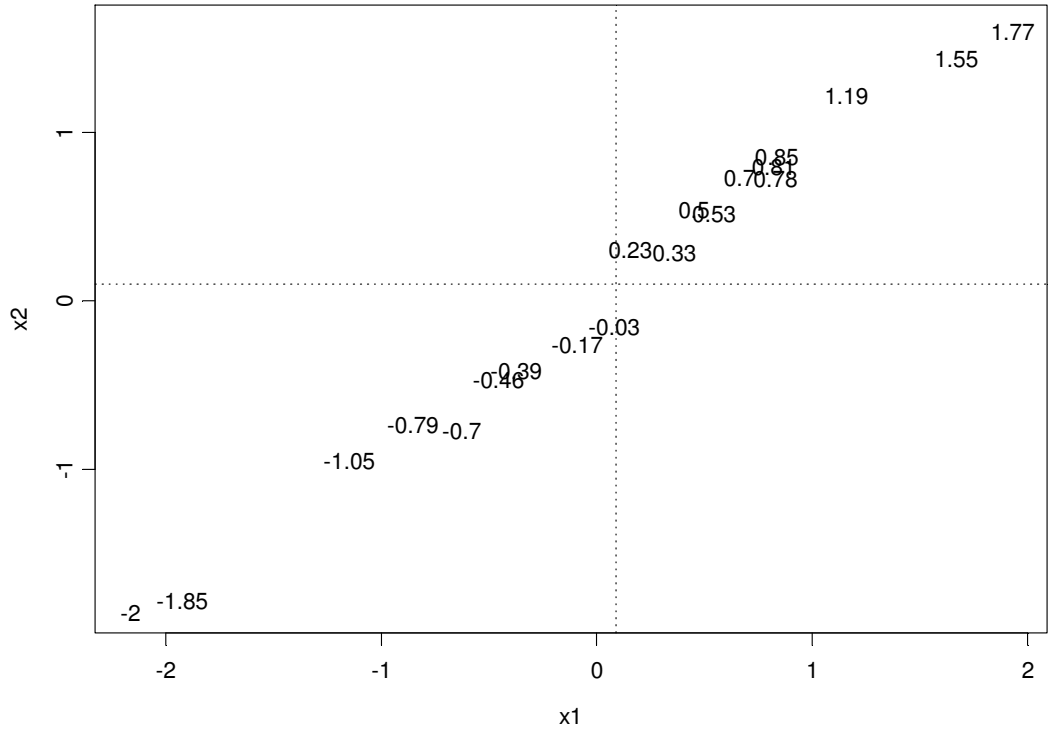
```
0.0058105 = .5*.5*1.3346 + .5*.5*1.106 +2*.5*.5*(–1.208679)
```

Why is the variance of y so much smaller than those of the x's ?

Example

$y = .5x_1 + .5\ x_2$

At each point we plot the value of y.



The variances and covariance are:

```
       x1          x2
x1 1.158167
x2 1.046490 0.9609463
```

The variance of y is

1.053 = .5*.5*1.158 + .5*.5*.961 + 2*.5*.5*1.0465

Why is the variance of y similar to those of the x's ?

Example

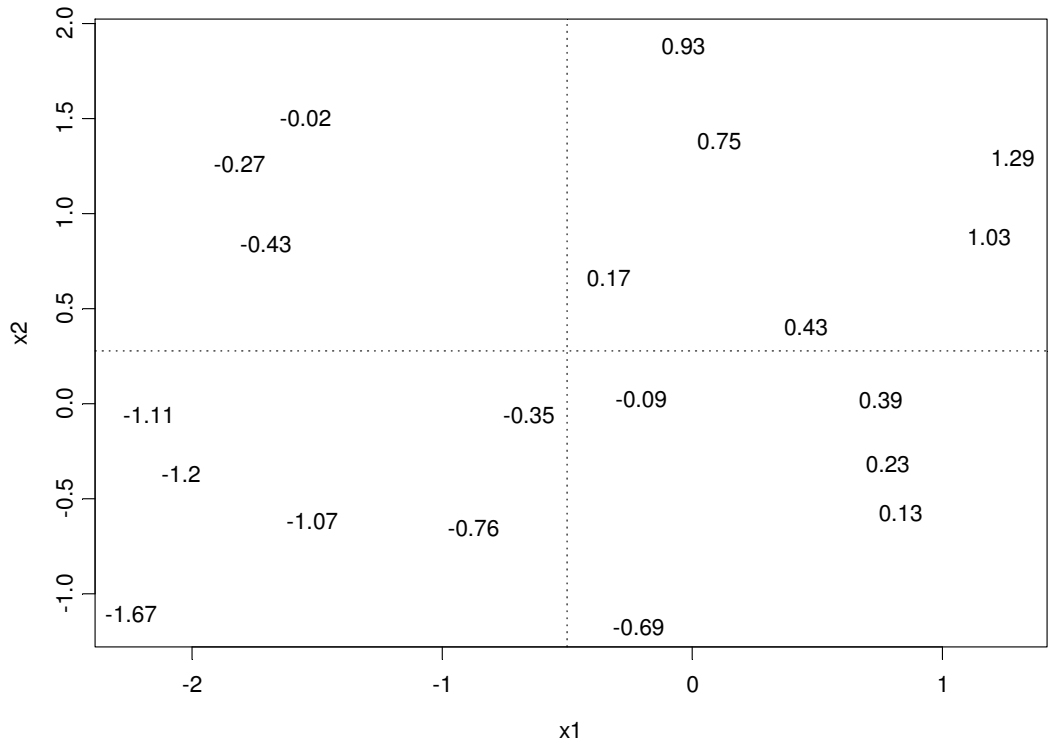
$y = .5x_1 + .5\ x_2$

At each point we plot the value of y.

The variances and covariance are:

```
      x1        x2
x1 1.3870537
x2 0.1976187 0.8247886
```

The variance of y is



The dashed lines are drawn at the mean of x_1 and x_2.

$0.65175 = .5*.5*1.387 + .5*.5*.8248 + 2*.5*.5*.1976$

Why is the variance of y less than that of x_1 and x_2 ?

Suppose

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3 + \cdots + c_k x_k$$

then,

$$\bar{y} = c_0 + c_1 \bar{x}_1 + c_2 \bar{x}_2 + c_3 \bar{x}_3 + \cdots + c_k \bar{x}_k$$

$$s_y^2 = c_1^2 s_{x_1}^2 + c_2^2 s_{x_2}^2 + \ldots + c_k^2 s_{x_k}^2$$

$$+ 2 \left\{ \begin{array}{l} \text{all the possible combinations} \\ \text{of covariance terms} \end{array} \right\}$$

Example

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3$$

$$\bar{y} = c_0 + c_1 \bar{x}_1 + c_2 \bar{x}_2 + c_3 \bar{x}_3$$

$$s_y^2 = c_1^2 s_{x_1}^2 + c_2^2 s_{x_2}^2 + c_3^2 s_{x_3}^2 + 2\left[c_1 c_2 s_{x_1 x_2} + c_1 c_3 s_{x_1 x_3} + c_3 c_2 s_{x_3 x_2}\right]$$

Example

```
= .1*(fidel)+.4*(eqmrkt)+.5*(windsor)
```

Table of covariances (variances on the diagonal)

	fidel	windsor	eqmrkt	port
fidel	0.00320210			
windsor	0.00241087	0.00236580		
eqmrkt	0.00319150	0.00298922	0.00470021	
port	0.00280224	0.00261967	0.00369384	0.00306760

.0030676 = (.1)*(.1)*.003202 + (.4)*(.4)*.0047 + (.5)*(.5)*.0023658
+2*((.1)*(.4)*.00319 + (.1)*(.5)*.00241+(.4)*(.5)*.00299)

Example

By forming portfolios we can reduce variance of returns !!!

This means people should hold portfolios !!!!

Cut from a Finance Textbook:

Fama [1976] has illustrated this result empirically.[11] 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 11%. 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-

[11] See Fama [1976], *Foundations of Finance*, pp. 253–254.

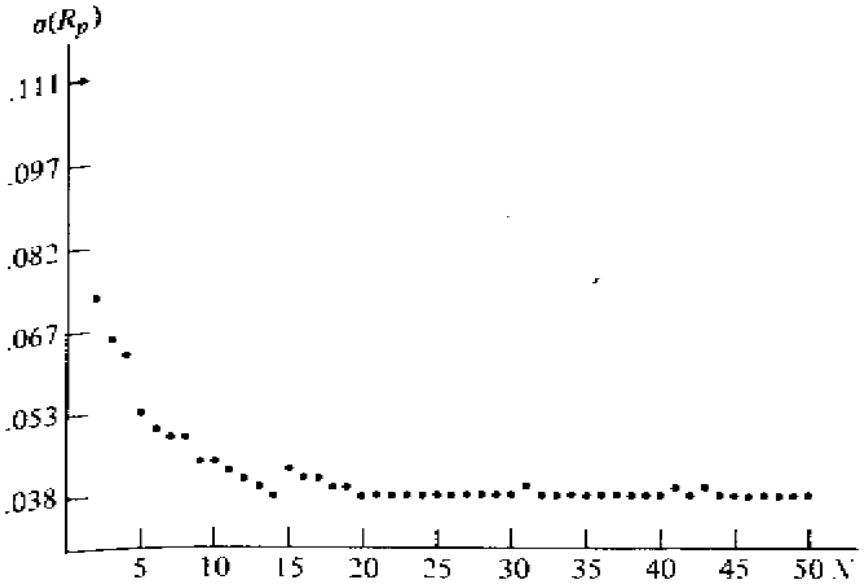


Figure 6.18
The standard deviation of portfolio return as a function of the number of securities in the portfolio. (From Fama, E. F., *Foundations of Finance*, reprinted with permission of the author.)

dard deviation quickly approached a limit which is roughly equal to the average covariance of all securities. One of the practical implications is that most of the benefits of diversification (given a random portfolio selection strategy) can be achieved with fewer than 15 stocks.