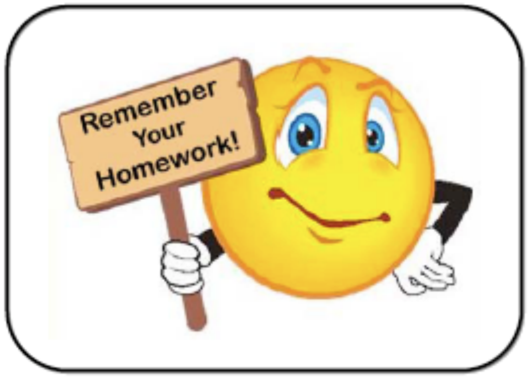
MATH E-3: Lecture 10

Quantitative Reasoning: Practical Math

April 19, 2016

http://promptguides.com/boston/attractions/harvard_museum_natural_history.htm
http://thumbnails.thecrimson.com.s3.amazonaws.com/photos/2009/12/13/202539_1224682.jpg.800x533_q95_crop-smart_upscale.jpg

Homework



© Freeology.com

- **Assignment 8** Excel extra credit – grades are available
- **Assignment 9** is April 23
- **Assignment 10** will be posted tomorrow

Homework Resources

SECTIONS:

On campus – Tuesday, 5:30 pm, Sever 104

Online – Wednesday, 7:30 pm (ET), "Conversations"

Math Question Center:

https://www.extension.harvard.edu/resources-policies/resources/math-question-center

Final Exam Information

- **DATE:** Tuesday, May 10, 2016

- **TIME**: 7:40 pm – 9:40 pm (ET)*

  *Students living outside the 6 New England states will have a 24 hour window in which to complete the exam. You must arrange, with a proctor, any two hour period between May 10, 7:40 pm (ET) and May 11, 7:40 pm (ET) in which to complete your exam.

- **LOCATION**:
  - If you live within the 6 New England states: Maxwell Dworkin G115
  - If you live outside the 6 New England states: You must arrange for a proctor:

http://www.extension.harvard.edu/resources-policies/exams-grades-transcripts/exams-online-courses

- Proctor questions should be directed to: distance_exams@dcemail.harvard.edu or call (617) 495-0977 Monday through Friday, 9 am to 5 pm eastern time.

Final Exam Review Section

- When: Tuesday, May 3* – 7:40 pm (ET)
    - *no class meeting

- Where: Online, via Canvas Conferences

- Review session video and slides will be posted the next day

Math in the News…

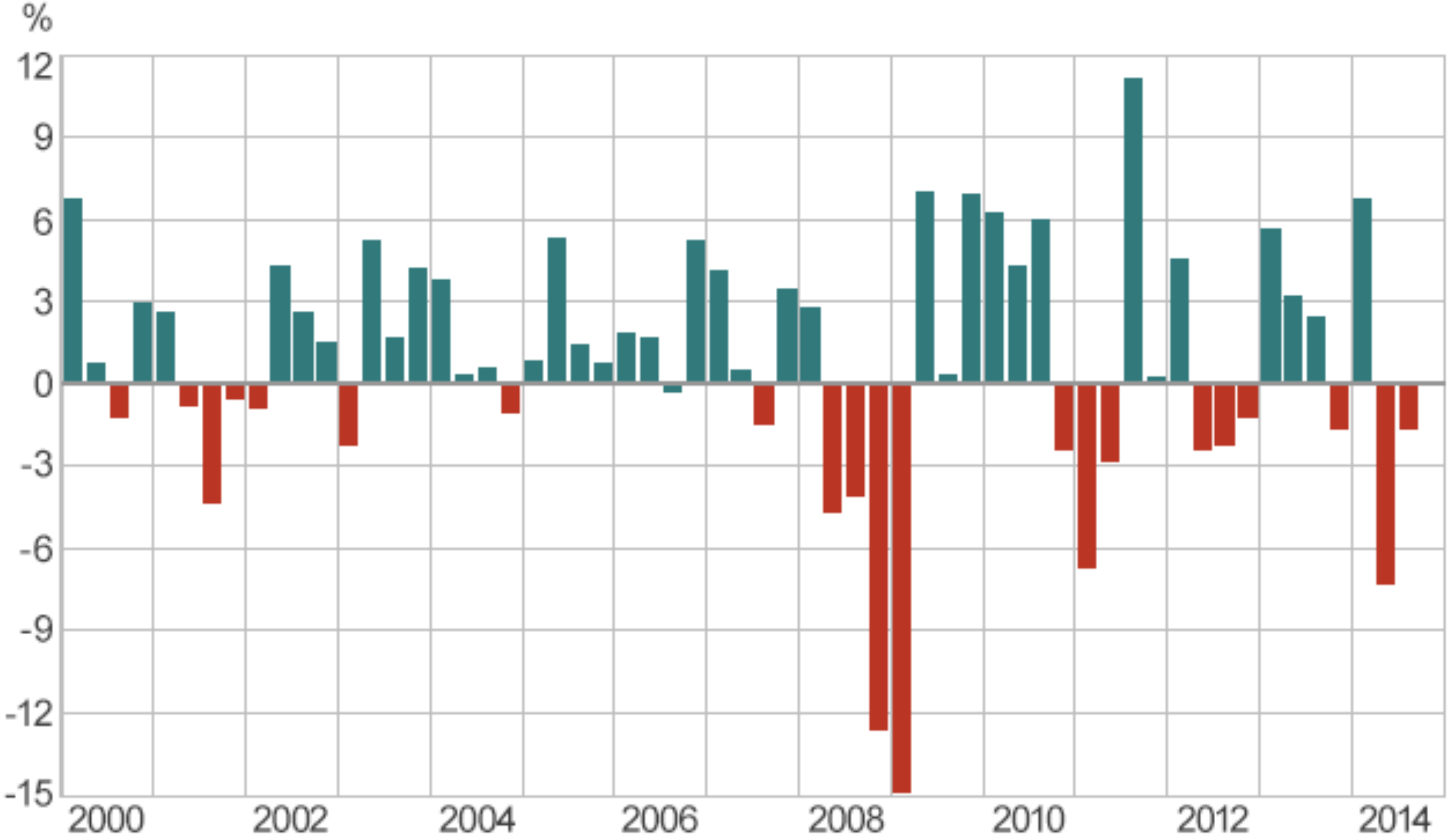
Definition: The textbook definition of a recession is when GDP growth is negative for two consecutive quarters or more.

http://useconomy.about.com/od/glossary/g/recession.htm

So which highly developed country announced in the fall of 2014 that it is in a recession?

Japan's economy since 2000

Change in quarterly GDP figures (annualised)



Source: Japan Cabinet Office

Topics relating to Straight Lines

Some “new” terms …

Interpolation

Extrapolation

Regression

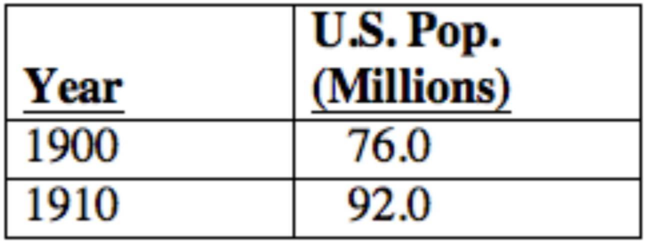
Correlation

Causation

a) Interpolation

- Finding a value between two known values

- E.g.



Year	U.S. Pop. (Millions)
1900	76.0
1910	92.0

- What was the population in the year 1904?
- How accurate will our answer be?
- What assumption(s) are we making?

Interpolation cont.

The population grew by 16 million in this 10-year period.

What was the average annual growth rate?

What are we assuming?

Under this assumption, the population in 1904 would be:

Interpolation cont.

The population grew by 16 million in this 10-year period.

What was the average annual growth rate?   **16m/10 = 1.6m/year**

What are we assuming?

Under this assumption, the population in 1904 would be:

Interpolation cont.

The population grew by 16 million in this 10-year period.

What was the average annual growth rate? **16m/10 = 1.6m/year**

What are we assuming? **Constant or linear growth**

Under this assumption, the population in 1904 would be:

Interpolation cont.

The population grew by 16 million in this 10-year period.

What was the average annual growth rate?   16m/10 = 1.6m/year

What are we assuming? Constant or linear growth

Under this assumption, the population in 1904 would be:

Population in 1900 + 4 years of 1.6 million/year growth

= 76.0m + 4 * 1.6m = **82.4 million in 1904.**

How good was our estimate?

- We estimated 82.4 million in 1904.
- It turns out that the actual population was 81.8 million. (How do we know that?)
- So we were fairly close. Interpolation generally is reasonably accurate, assuming nothing “weird” is happening . . .

NOTE…

Remember that earlier slide:

“The population grew by 16 million in this 10-year period.

What was the average annual growth rate?   16m/10 = 1.6m/year”

Does this “1.6m/year” remind you of anything? (hint, hint)

NOTE…

Remember that earlier slide:

“The population grew by 16 million in this 10-year period.

What was the average annual growth rate?   16m/10 = 1.6m/year”

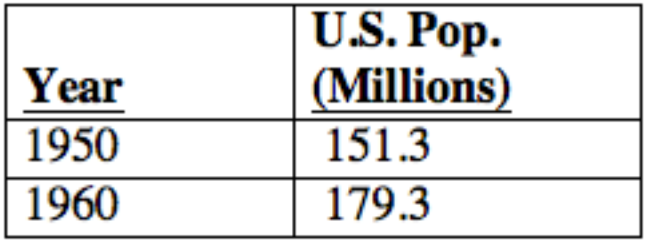
Does this “1.6m/year” remind you of anything? (hint, hint)

It’s the slope!!

b) Extrapolation

- Finding a value <u>outside</u> two known values

- E.g.



Year	U.S. Pop. (Millions)
1950	151.3
1960	179.3

- What was the population in the year 1976?
- How accurate will our answer be?
- What assumption(s) are we making?

Extrapolation cont.

The population grew by (179.3 – 151.3) = 28 million in this 10-year period.

What was the average annual growth rate?

What are we assuming?

Under this assumption, the population in 1976 would be:

Extrapolation cont.

The population grew by (179.3 – 151.3) = 28 million in this 10-year period.

What was the average annual growth rate? **28m/10 = 2.8m/year**

What are we assuming?

Under this assumption, the population in 1976 would be:

Extrapolation cont.

The population grew by (179.3 – 151.3) = 28 million in this 10-year period.

What was the average annual growth rate? **28m/10 = 2.8m/year**

What are we assuming? **Constant or linear growth** (note: the population is growing at a faster rate than in the early 1900s).

Under this assumption, the population in 1976 would be:

Extrapolation cont.

The population grew by (179.3 – 151.3) = 28 million in this 10-year period.

What was the average annual growth rate?   28m/10 = 2.8m/year

What are we assuming? **Constant or linear growth** (note: the population is growing at a faster rate than in the early 1900s).

Under this assumption, the population in 1976 would be:

Population in 1960 + 16 years of 2.8 million/year growth

= 179.3m + 16 * 2.8m = 179.3m + 44.8m = **224.1 million in 1976.**

How good was our estimate?

- We estimated 224.1 million in 1976.

- It turns out that the actual population in 1976 was 214.3 million. (From www.census.gov/main/www/cen 2000)

- So we were off by about 10 million. Extrapolation generally is less accurate than interpolation; and the further you are from "known data," the less accurate it gets.

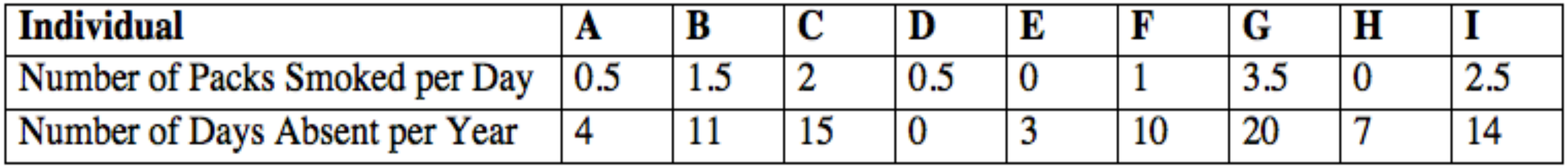
c) Linear Regression

- So far we have been only using two points at a time.

- Often we have several points, which do not all lie on a straight line, but will nevertheless show a more or less clear trend. We can use linear regression in order to interpolate and/or extrapolate other values.

Linear Regression, cont.

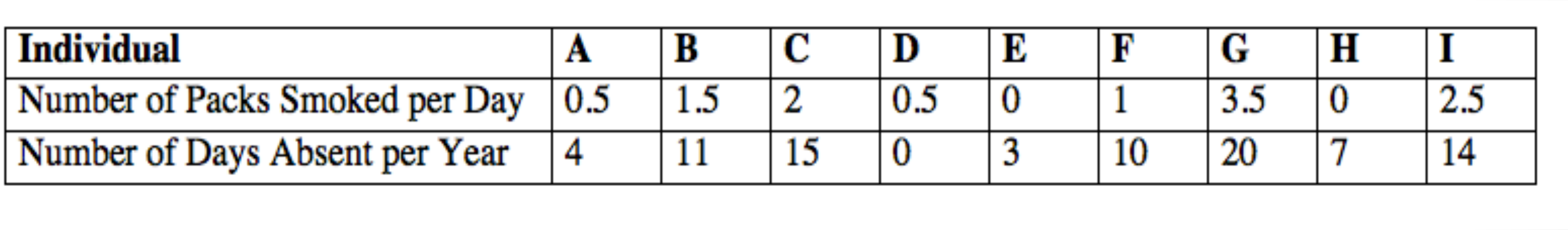
- Look at the data below. What initial observations can you make?



Individual	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Number of Packs Smoked per Day	0.5	1.5	2	0.5	0	1	3.5	0	2.5
Number of Days Absent per Year	4	11	15	0	3	10	20	7	14

Linear Regression, cont.

- Look at the data below. What initial observations can you make?



Individual	A	B	C	D	E	F	G	H	I
Number of Packs Smoked per Day	0.5	1.5	2	0.5	0	1	3.5	0	2.5
Number of Days Absent per Year	4	11	15	0	3	10	20	7	14

- "In general," more smoking tends to go with more days absent . . .

- But it's not completely straightforward: compare A and D, or E and H.

Let's graph the data . . .

- What type of graph would be most appropriate?

- Which variable should go on which axis, and why?

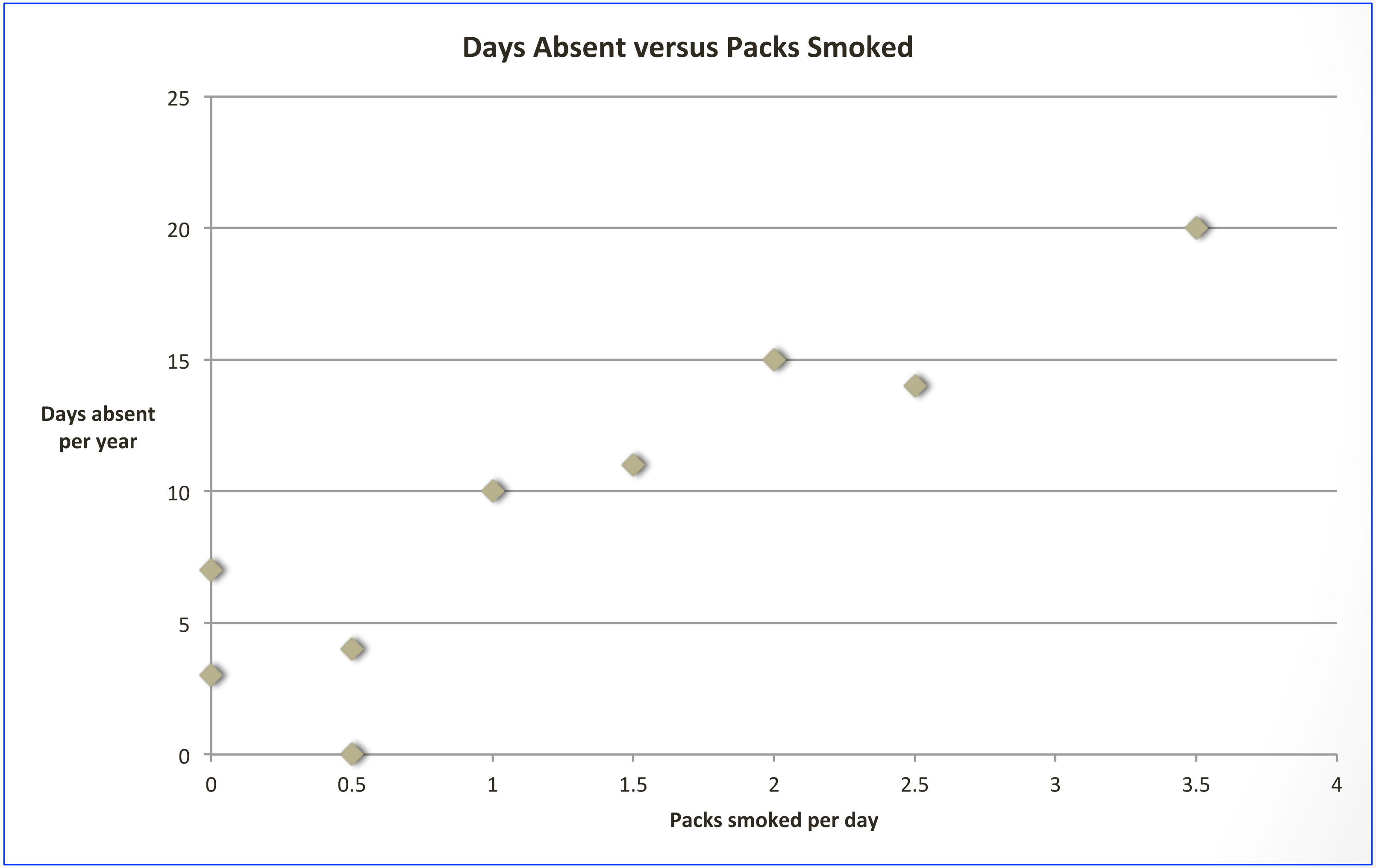
Let's graph the data . . .

- What type of graph would be most appropriate?
- **Scatterplot**
- Which variable should go on which axis, and why?

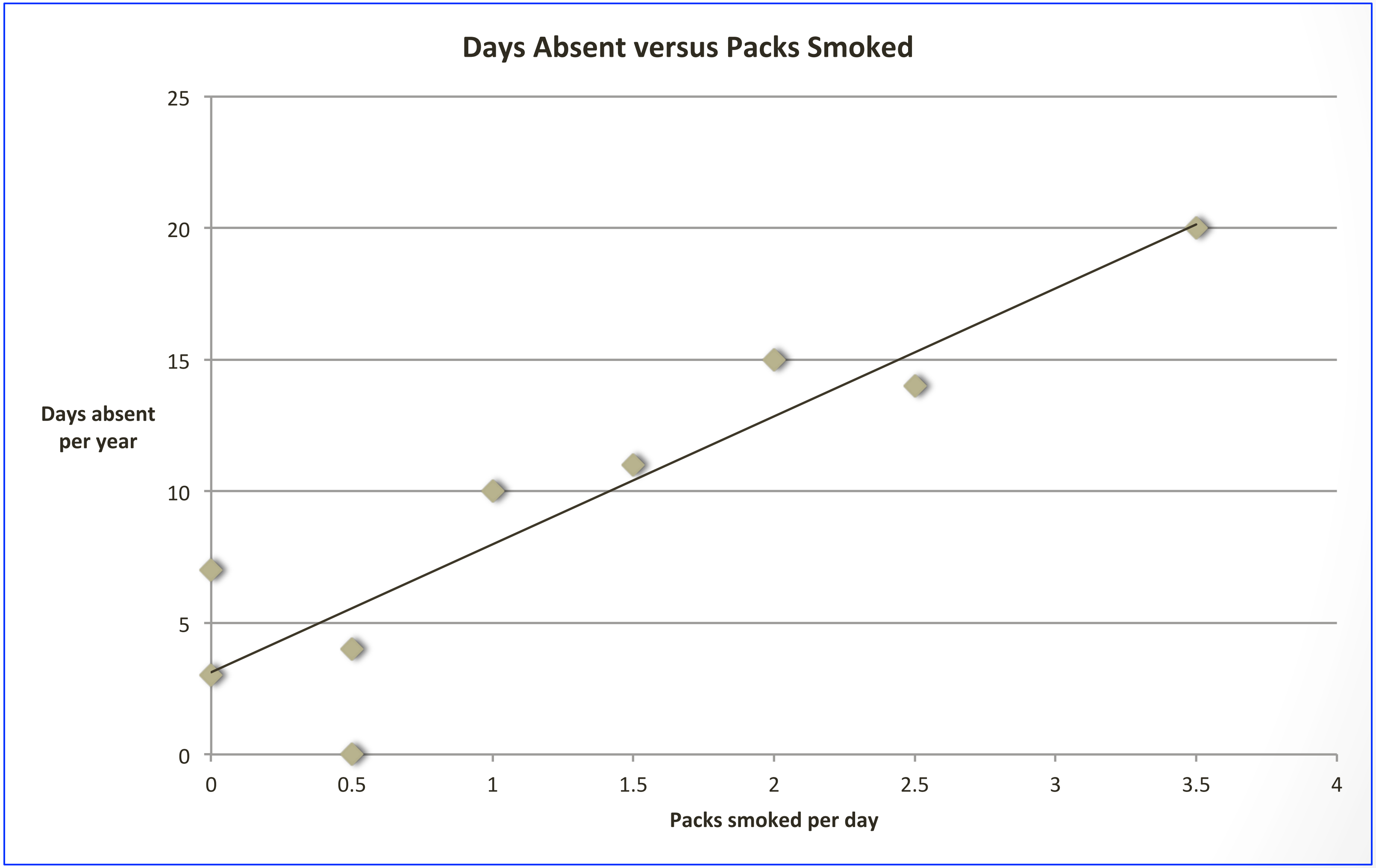
Let's graph the data . . .

- What type of graph would be most appropriate?
- **Scatterplot**
- Which variable should go on which axis, and why?
- **Cigarettes smoked on the horizontal or x-axis; this is considered the "independent variable"; the assumption being that somehow smoking has an effect on the number of days absent, rather than the other way around.**

A Scatterplot makes the trend clearer



Add a “regression line”



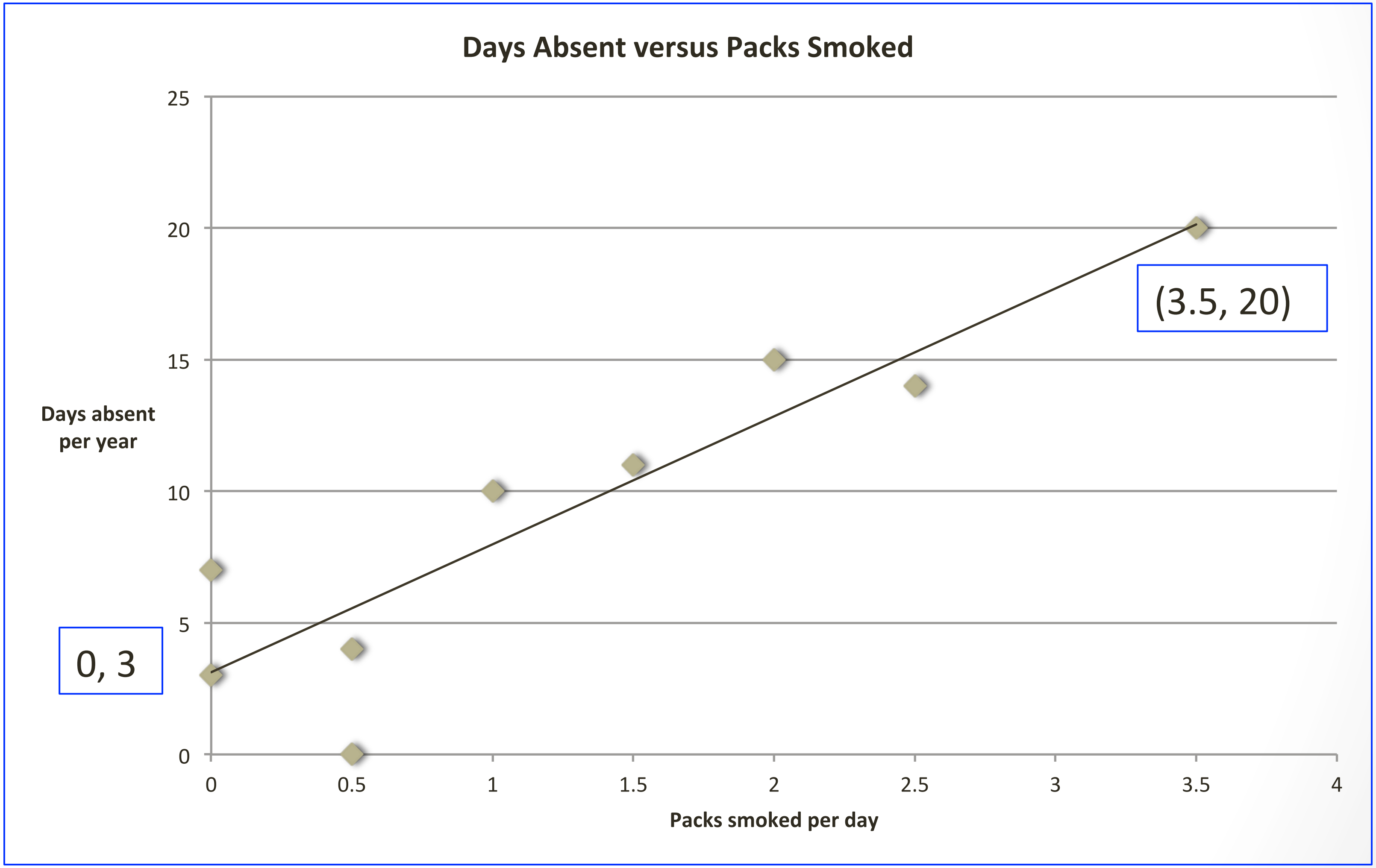
The Regression Line

- Should reflect the "trend" of the data
- Should go "through the middle" of the data
- Should be straight! (this is linear regression . . .)
- Does not have to go through any of the original points, but may do so
- Will be different depending on who is drawing it

The Regression Line, cont.

- To “make life easy”:
- Choose the y-intercept in advance (that’s your first point)
- Choose one other point, preferably but not necessarily one of the original points; but it MUST be a point on the line!!
- Calculate the slope,
- Use the slope and the y-intercept to create the equation of your regression line

Back to our regression line:



(3.5, 20)

0, 3

Our regression line:

Has a y-intercept of (0, 3).

Also goes through the point (3.5, 20)

The slope is:

So the equation of our regression of line is:

Our regression line:

Has a y-intercept of (0, 3).

Also goes through the point (3.5, 20)

The slope is: (20 – 3)/(3.5 – 0) = 17/3.5 = **4.86 (2 d.p.)**

So the equation of our regression of line is:

Our regression line:

Has a y-intercept of (0, 3).

Also goes through the point (3.5, 20)

The slope is: (20 – 3)/(3.5 – 0) = 17/3.5 = **4.86 (2 d.p.)**

So the equation of our regression line is:

$$Y = 4.86x + 3$$

We can use our regression line:

a) To interpolate:

If someone smokes 3/4 pack per day, what do we predict?

We can use our regression line:

a) To interpolate:

If someone smokes 3/4 pack per day, what do we predict?

Plug in x = 0.75 to the equation Y = 4.86x + 3

We get Y = 4.86 * 0.75 + 3 = **6.65 days absent.**

We can use our regression line:

b) To extrapolate:

If someone smokes 5 packs per day, what do we predict?

We can use our regression line:

b) To extrapolate:

If someone smokes 5 packs per day, what do we predict?

Plug in x = 5 to the equation Y = 4.86x + 3

We get Y = 4.86 * 5 + 3 = **27.3 days absent.**

Some caviar, I mean caveats . . .

- Remember, neither interpolation nor extrapolation are guarantees

- Extrapolation is generally less reliable than interpolation

- We have not proved a cause-and-effect relationship between smoking and days absent. Why not?

Some caveats . . .

- Remember, neither interpolation nor extrapolation are guarantees
- Extrapolation is generally less reliable than interpolation
- We have not proved a cause-and-effect relationship between smoking and days absent. Why not?
- Other possible factors: caring for a sick relative, an appointment with a child's teacher, sickness not related to smoking, etc.

d) Correlation (we use the symbol "r")

Intuitively, correlation is an indication that the two variables are "connected" in some way; in this case, as smoking increases, so does number of days absent, suggesting a connection between them.

Correlation can be measured; in this course, we'll "guestimate" it as follows:

First, correlation can only be **between -1 and +1. We say**

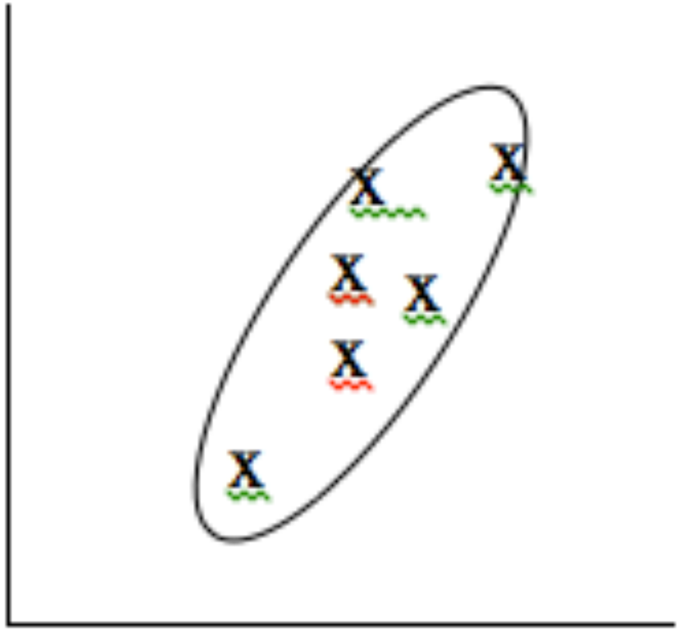
$$\mathbf{-1 \le r \le +1}$$

A <u>negative</u> correlation indicates that the regression line is sloping downwards, while a <u>positive</u> correlation indicates an upward sloping regression line.

Correlation, cont.

- A correlation of +1 would be a "perfect positive correlation," meaning all the data points would lie perfectly on the regression line. This is a pretty rare situation.
- Similarly, a correlation of -1 would be a "perfect negative correlation," again with all the points on the line, and again fairly rare.
- Most correlations are in between -1 and +1.

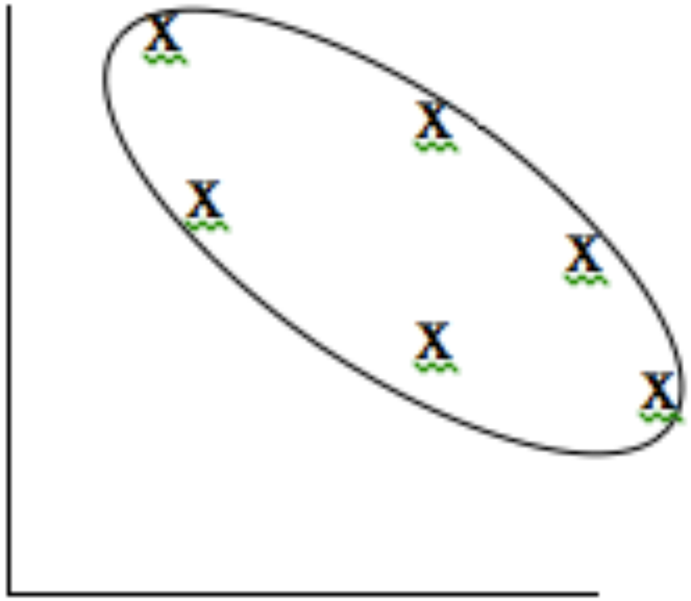
Examples of correlation



strong positive correlation

$r \cong .6$ or $.7$

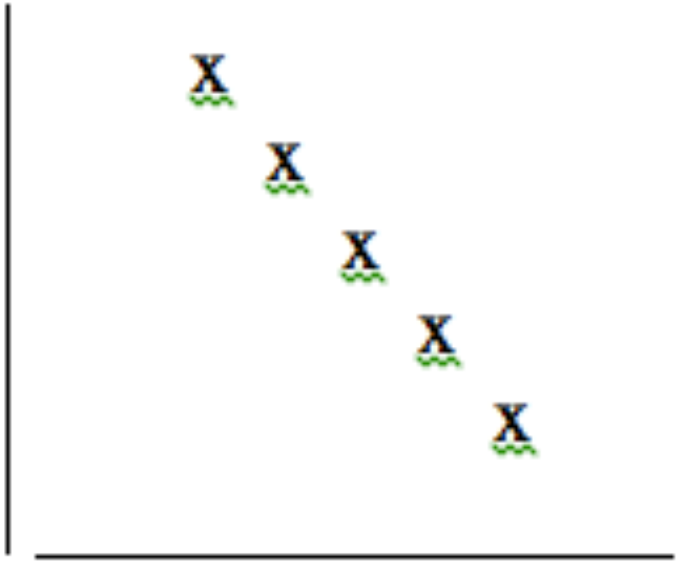
Examples of correlation



moderately weak negative correlation

r ≅ –.3 or – .4

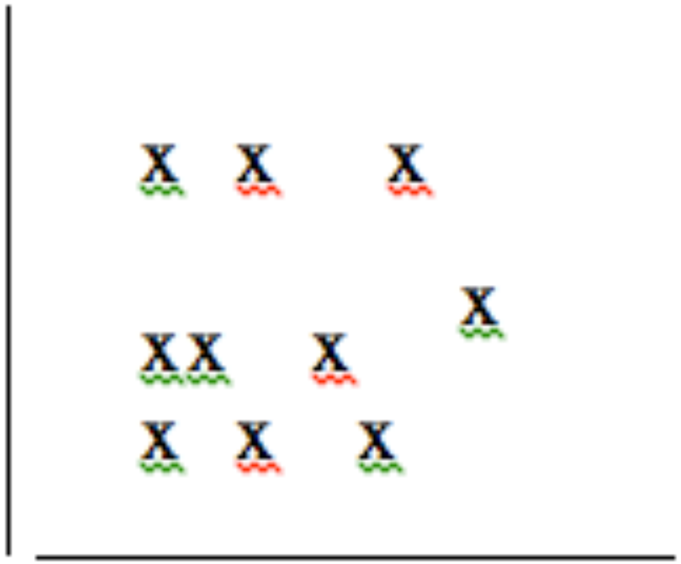
Examples of correlation



perfect negative correlation
r = – 1

What would a correlation of 0 look like?

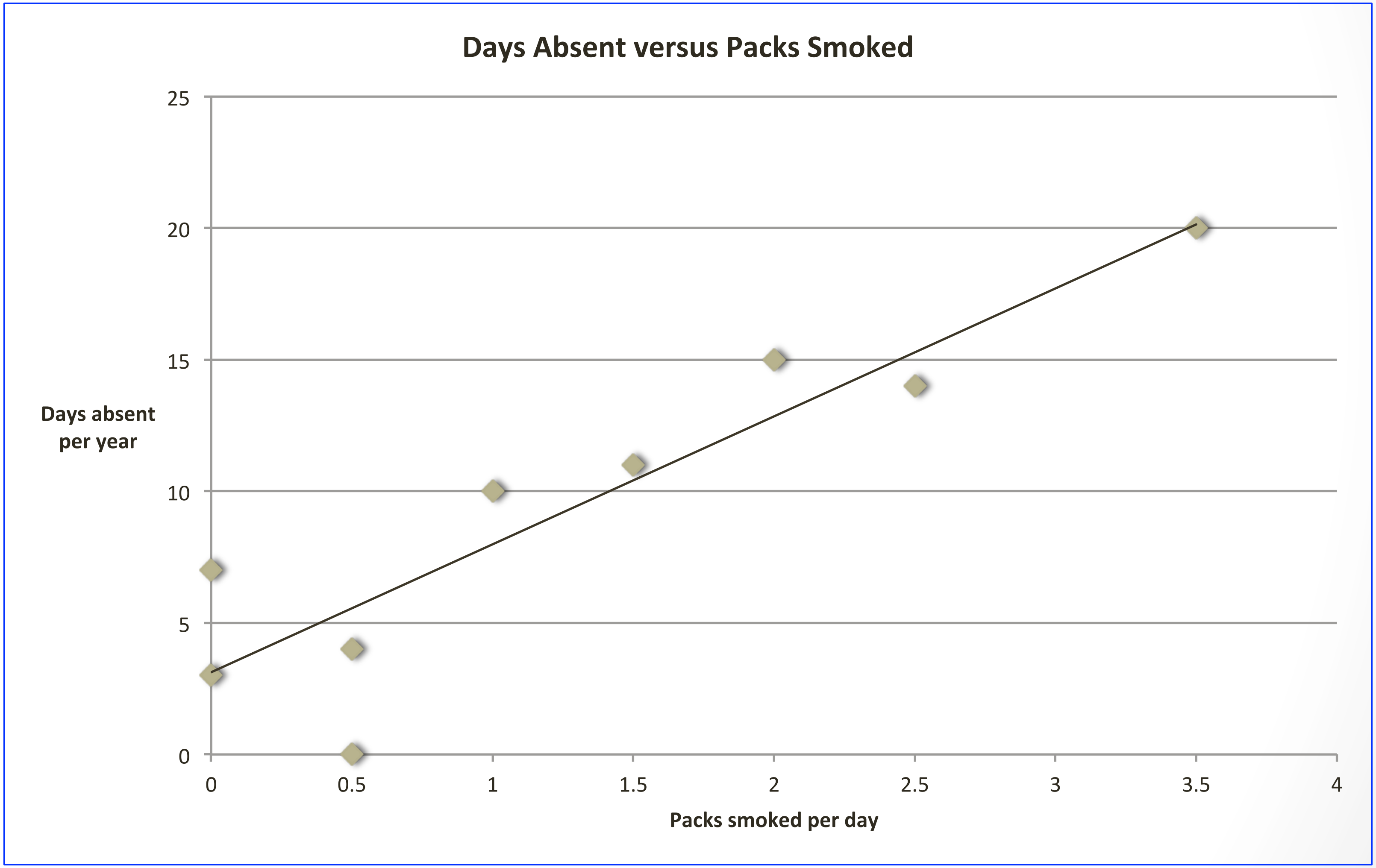
What would a correlation of 0 look like?



no apparent correlation

r = 0

Our previous example . . .



Choose the most likely correlation figure, and explain why . . .

–1 –.7 –.3 0 .3 .7 1

Choose the most likely correlation figure, and explain why . . .

–1 –.7 –.3 0 .3 <u>**.7**</u> +1

- Upward sloping regression line, so correlation must be <u>positive</u>.

- The data are not perfectly on the line, so it cannot be +1.

- The data are reasonably close to the line, so .7 looks more likely than .3. (in fact Excel gives r as closer to 0.9)

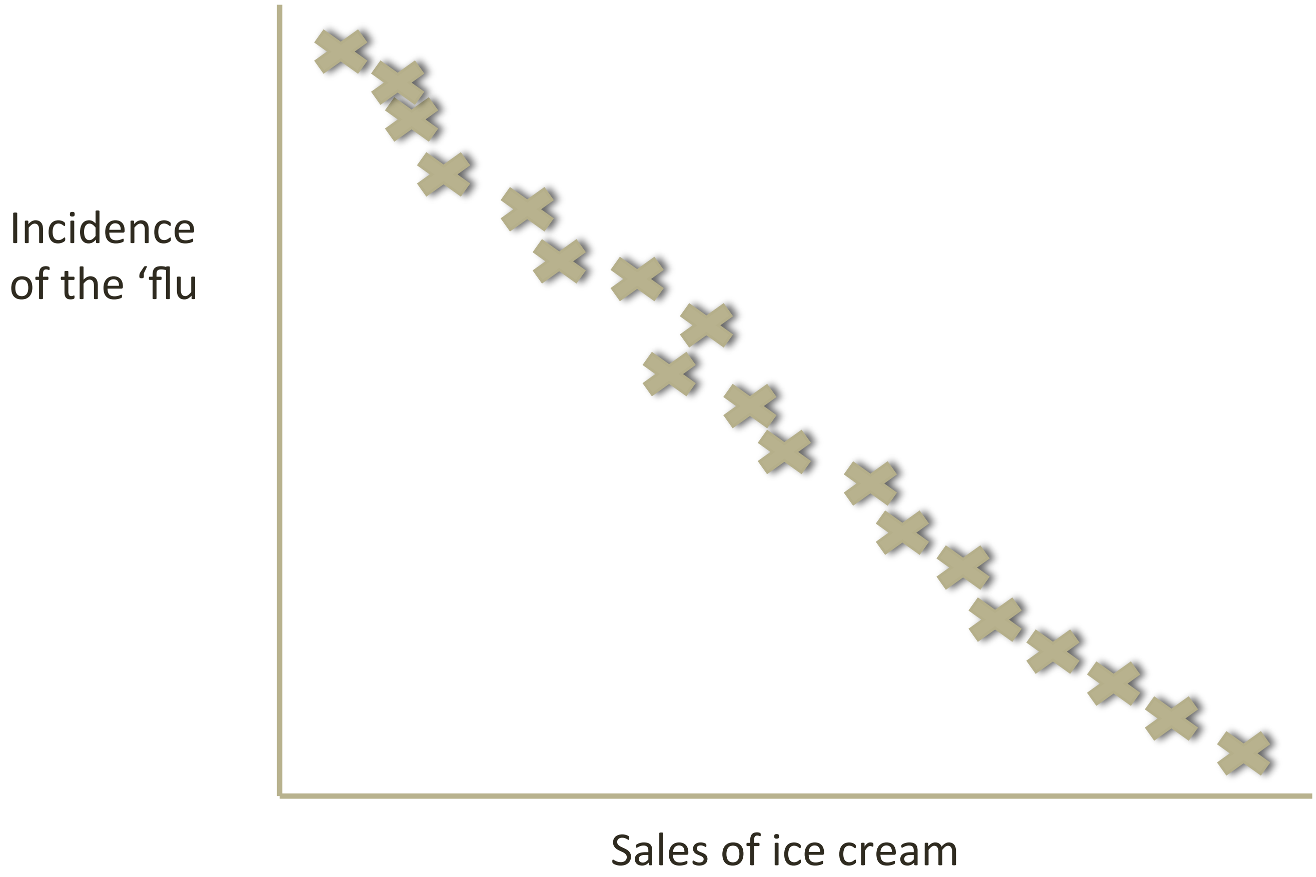
- We would say there is “moderately strong positive correlation.”

However . . .

- Although there is a fairly strong positive correlation, meaning that as smoking increases, so does the number of days absent, we cannot categorically state that there is a cause-and-effect relationship between these two variables, based just on these figures.

- In fact, look at this example, and see what you think about cause and effect versus correlation . . .

Sales of ice cream and the 'flu

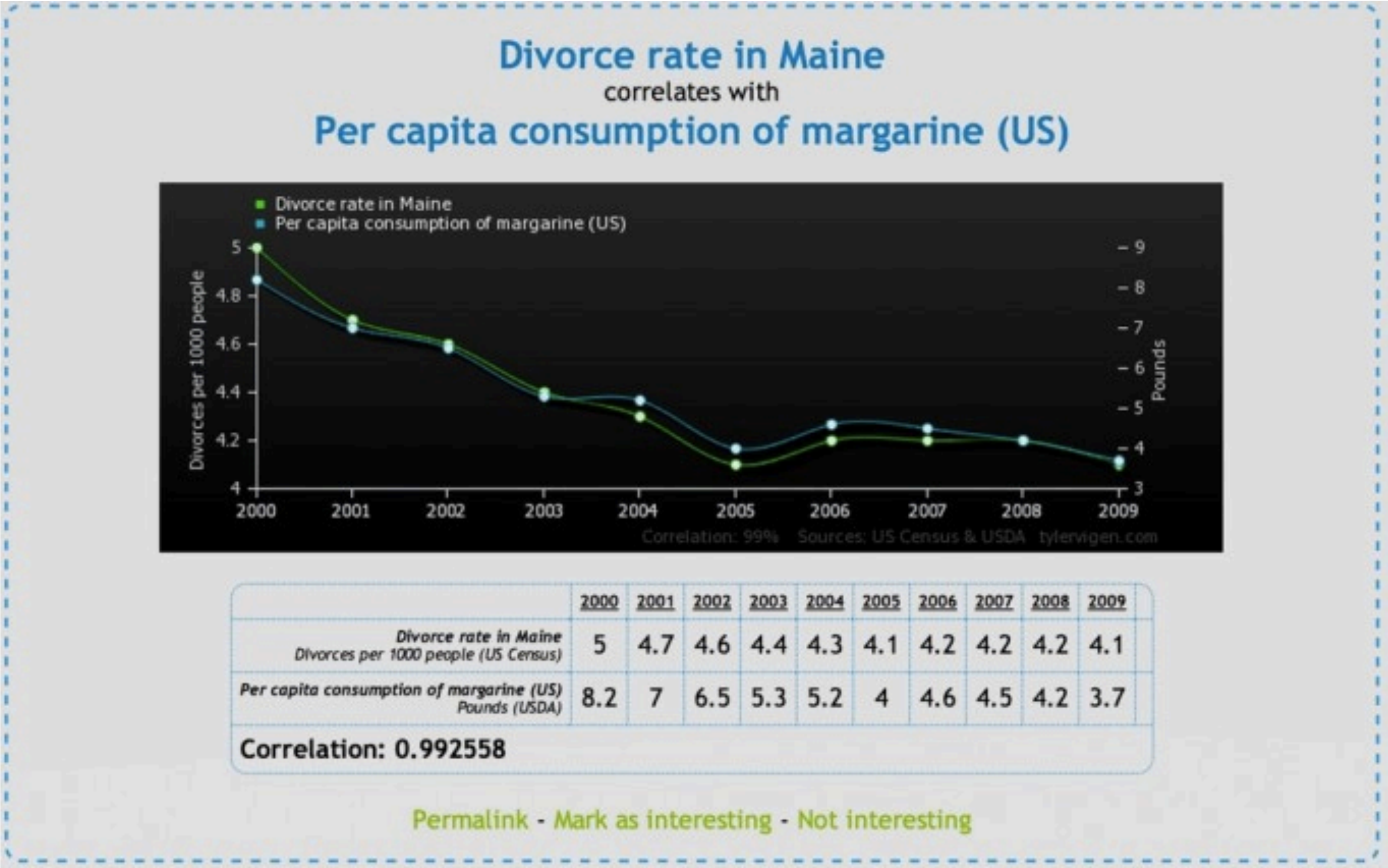


Have we just discovered a new cure for the flu?!

Have we just discovered a new cure for the flu?!

- Probably not . . .
- When is ice cream sold the most?
- When is the ‘flu the most prevalent?
- So what’s going on?
- “other factors” . . .

Another one . . .



Divorce rate in Maine

correlates with

Per capita consumption of margarine (US)

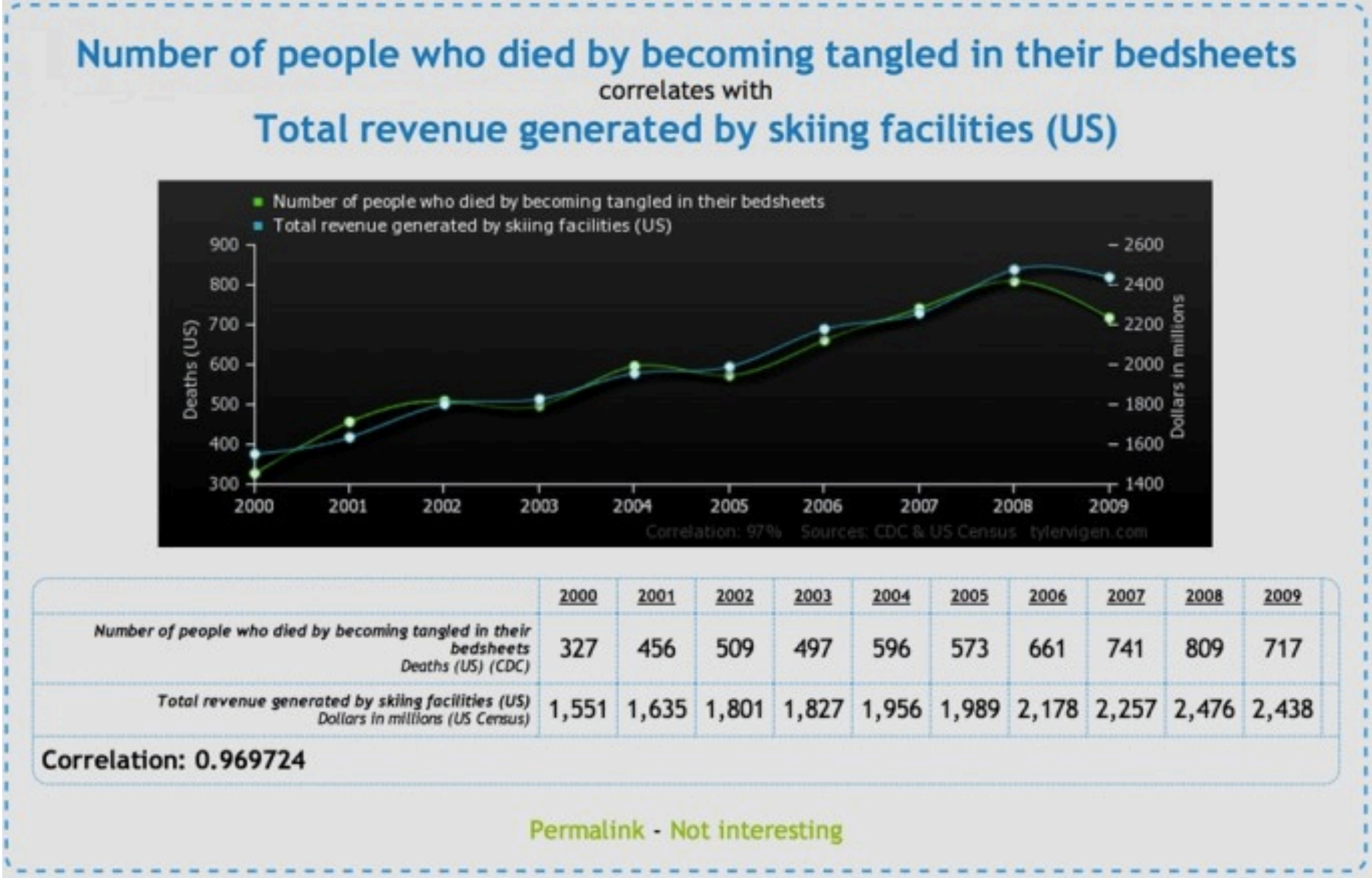
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Permalink - Mark as interesting - Not interesting

http://twentytwowords.com/funny-graphs-show-correlation-between-completely-unrelated-stats-9-pictures/

And another . . .



Number of people who died by becoming tangled in their bedsheets

correlates with

Total revenue generated by skiing facilities (US)

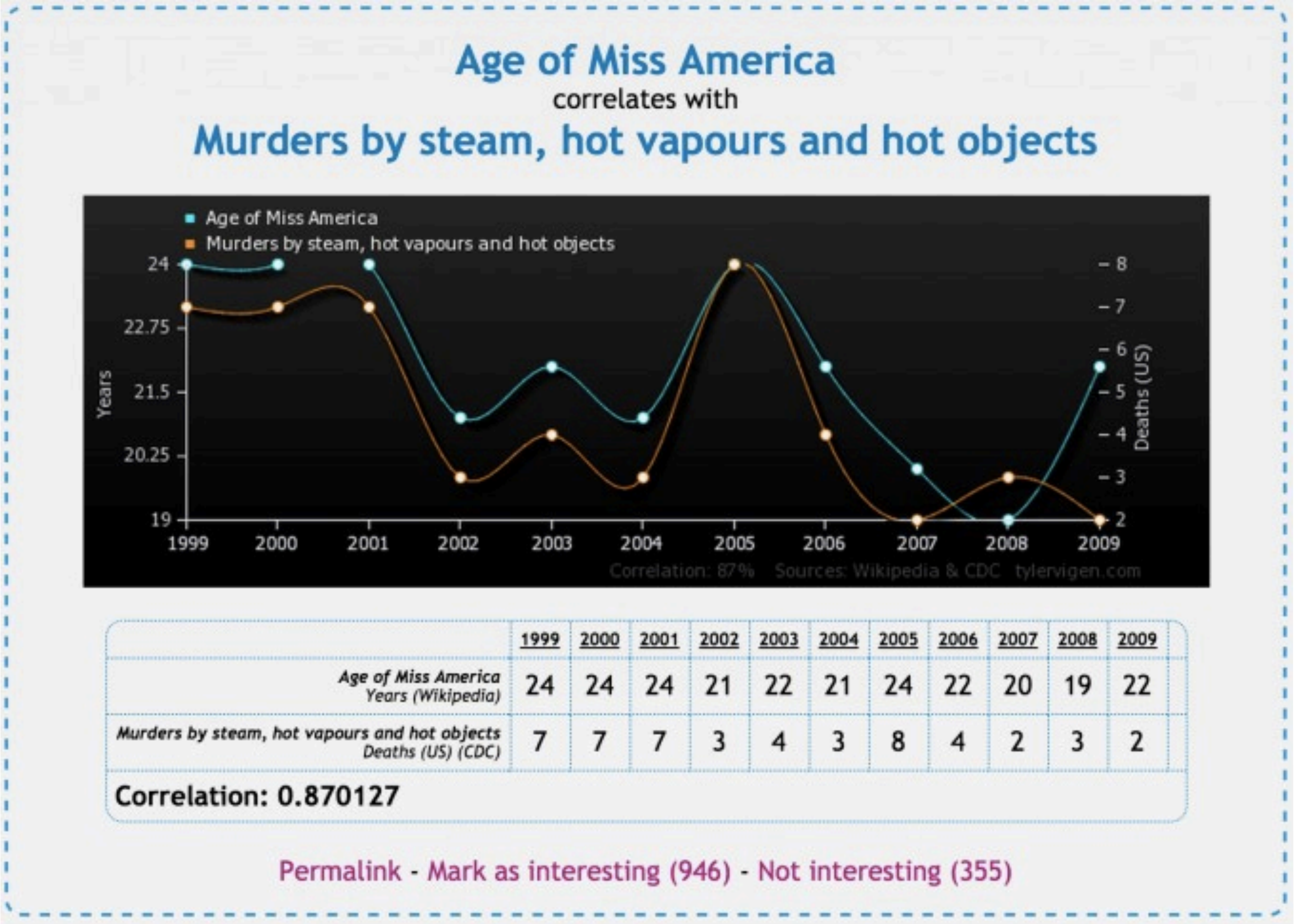
Correlation: 97%   Sources: CDC & US Census   tylervigen.com

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717
Total revenue generated by skiing facilities (US) *Dollars in millions (US Census)*	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Correlation: 0.969724

Permalink - Not interesting

One more . . .



Age of Miss America

correlates with

Murders by steam, hot vapours and hot objects

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Age of Miss America *Years (Wikipedia)*	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects *Deaths (US) (CDC)*	7	7	7	3	4	3	8	4	2	3	2

Correlation: 0.870127

Permalink - Mark as interesting (946) - Not interesting (355)

One moral of this story:

Correlation

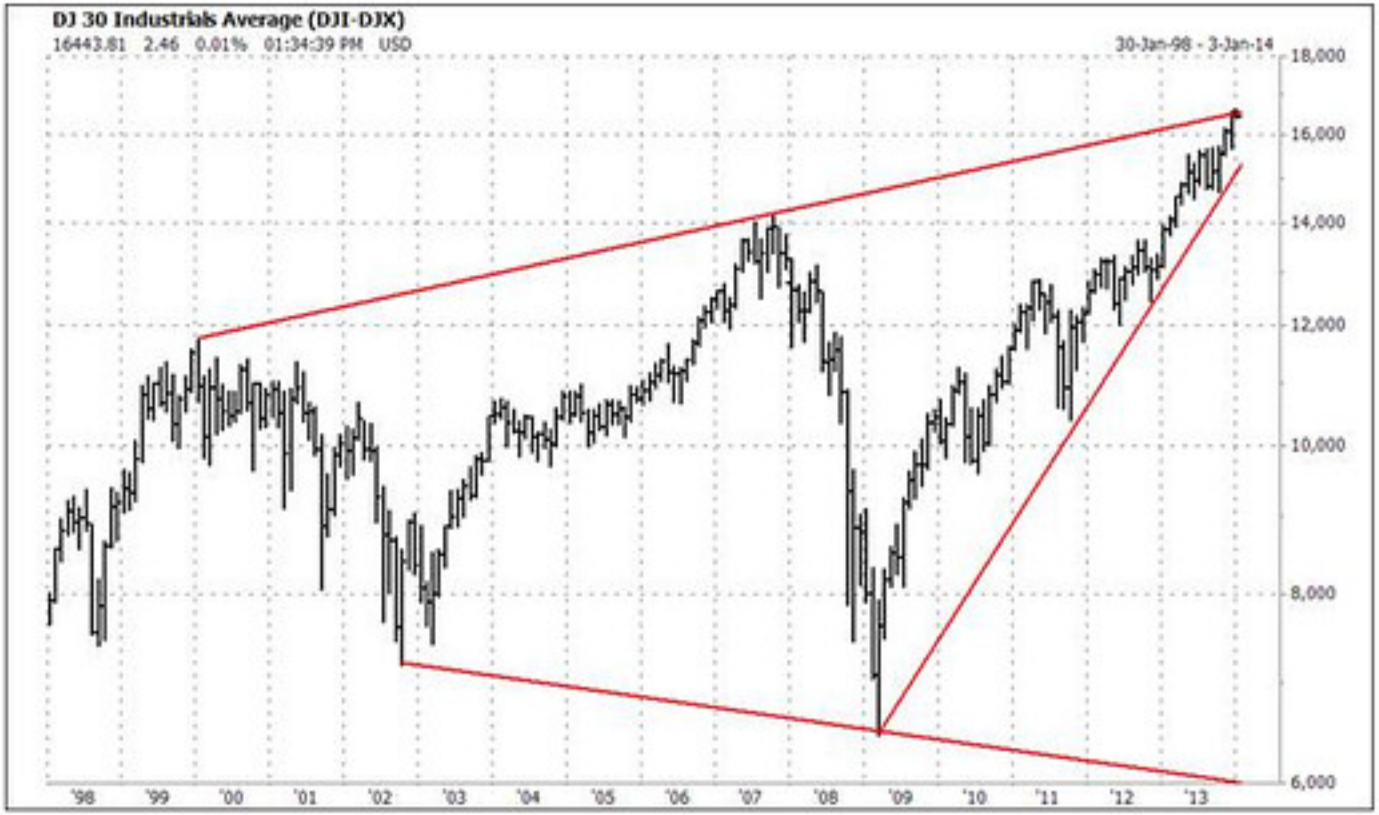
does not necessarily mean

Causation

Another moral . . .

Extrapolation doesn't guarantee anything

Which way is the trend going?



http://blogs.wsj.com/moneybeat/2014/01/03/the-bearish-call-to-end-all-bearish-calls/?mod=WSJ_hpp_MIDDLENexttoWhatsNewsFifth

Mark Twain and Extrapolation

Twain begins his indication of frustration with those who tried to use extrapolation to predict anything about the river by citing statistics that showed the lower Mississippi River had "shortened itself two hundred and forty-two miles" over a period of one hundred seventy-six years. Twain creates an average rate of change based on this distance and time ("a trifle over a mile and a third per year"), then uses extrapolation to draw the ridiculous conclusion that

just a million years ago next November, the Lower Mississippi was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing-pole.