The Elements of Probability and Statistics

E. Bruce Pitman

The University at Buffalo

CCR Workshop – June 27, 2017

Basic Premise of Statistics

One can group statistical ideas into a few groupings

- Aggregation
- Likelihood
- Regression
- Variation

We will see examples of these notions throughout.

Definitions

The probability P of an event A is the fraction of possible favorable outcomes – the number of favorable outcomes f divided by the total number of possible outcomes n.

$$P(A) = f/n$$

Definitions 2

The probability of drawing an ace from a standard deck of cards is 4/52.
What this really means is that if I draw one card from a deck and see whether or not it is an ace and then replace the card and shuffle well, and repeat this exercise infinitely many times, then over the long haul I will find an ace 7.69% of the time.

Definitions 3

However, suppose you are playing a game in which all outcomes are equally likely (e.g., rolling dice), and you are on a losing streak. You might commit the Gamblers Fallacy if you believe your losing streak makes it more likely that you'll roll the numbers you want on the next roll (because you're "due"). The truth is that your odds don't change; you start over with each roll.

Definitions 4

A sample space is the set of all possible outcomes of an experiment.
An event is a specific outcome of an experiment.
So if you are rolling a die, the sample space is the set $A = \{1, 2, 3, 4, 5, 6\}$. The event 'roll an odd number' is the set $E = \{1, 3, 5\}$.

One perspective

Counting

Say we have a sample of n items

$$\text{numbers} = \{1, 3, 3, 4, 5, 6, 7, 7, 7, 9\}$$

How to measure "average"?

Say we have a sample of n items

$$\text{numbers} = \{1, 3, 3, 4, 5, 6, 7, 7, 7, 9\}$$

How to measure "average"?

$$\text{mean } \overline{x} = \frac{sum}{number} = \frac{1+3+3+4+5+6+7+7+7+9}{10} = \frac{52}{10} = 5.2$$

Say we have a sample of n items

$$\text{numbers} = \{1, 3, 3, 4, 5, 6, 7, 7, 7, 9\}$$

How to measure "average"?

$$\text{mean } \overline{x} = \frac{sum}{number} = \frac{1+3+3+4+5+6+7+7+7+9}{10} = \frac{52}{10} = 5.2$$

$$\text{median} = \text{middle value} = 5.5$$

Say we have a sample of n items

$$\text{numbers} = \{1, 3, 3, 4, 5, 6, 7, 7, 7, 9\}$$

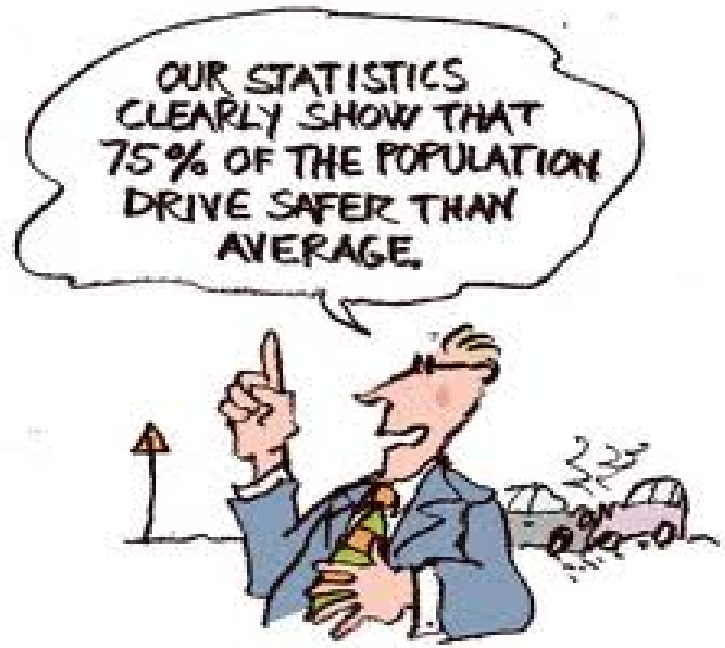
How to measure "average"?

mean $\overline{x} = \frac{sum}{number} = \frac{1+3+3+4+5+6+7+7+7+9}{10} = \frac{52}{10} = 5.2$

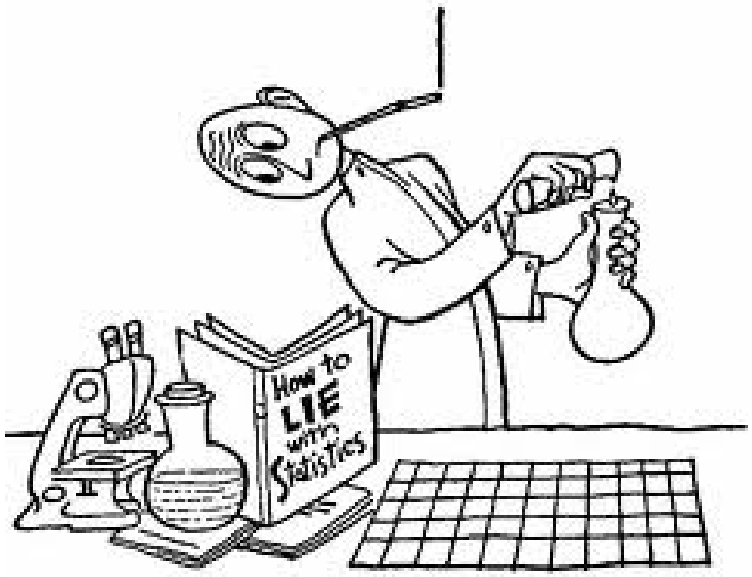
median = middle value = 5.5

mode is the item that occurs with the highest frequency = 7

What does it mean?



For a different sample, say $\mathrm{morenumbers} = \{1, 5, 7, 8, 9\}$, the mean $= 6$ and the median $= 7$



Independent items

Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur; that is, the events have no influence on each other.
If events A and B are independent, then the probability of A and B happening is just the product $p(A)p(B)$.
If events are not independent, the joint probability is not the product – but we don't know what it is without more information.

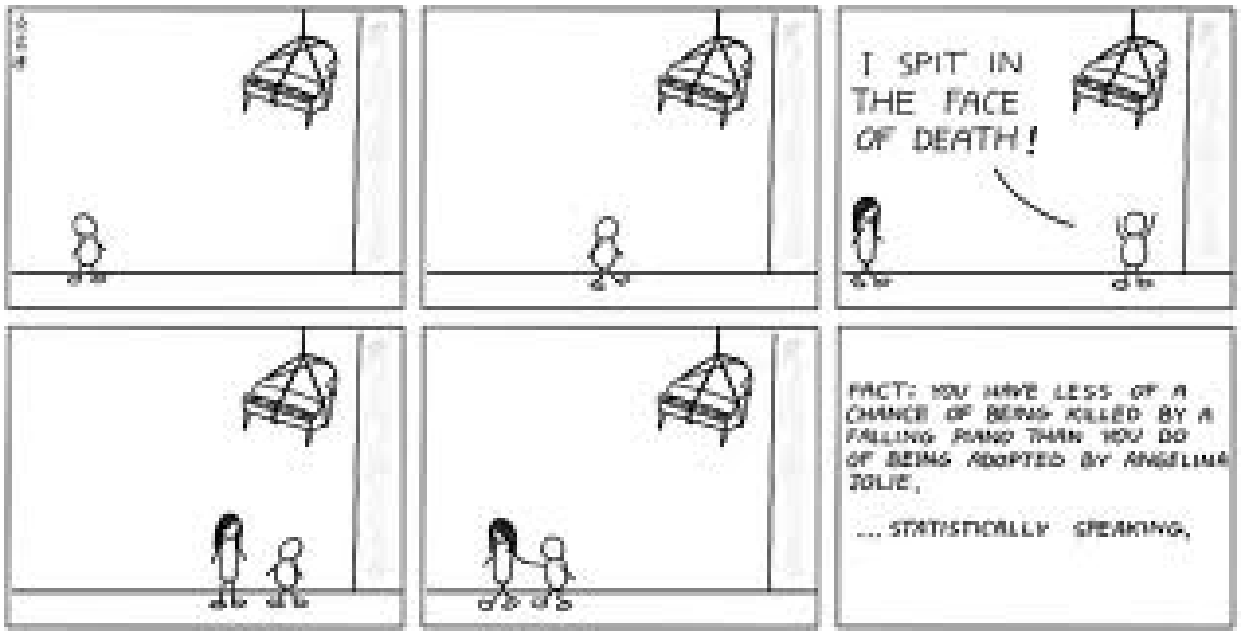
So, for instance, whether it rains tomorrow, and whether or not tomorrow is a Friday, are - really - independent.

Independent items again

The total number of points the Bills score in a season and the number of points the Dolphins score are independent.

But the total number of points the Bills score in a season and the event "Tyrod is the starting quarterback in December" are dependent.

So now we know some statistics



Variance

Variance is how wide of a spread is present in the data

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

For morenumbers, we have

$$\sigma^2 = \frac{1}{4}[(1-6)^2+(5-6)^2+(7-6)^2+(8-6)^2+(9-6)^2]$$

$$= \frac{1}{4}[25+1+1+4+9] = \frac{1}{4}(40) = 10$$

The standard deviation is the square root of the variance $= \sqrt{\sigma^2}$
For our data, $\sigma = \sqrt{10}$

A story about variance

1997 Grand Forks floods. Billions of damage That winters snowfall especially heavy and the potential for flood well known.

A story about variance

1997 Grand Forks floods. Billions of damage That winters snowfall especially heavy and the potential for flood well known.
Two months before the spring melt started, NWS predicted the Red River would crest at 49.

A story about variance

1997 Grand Forks floods. Billions of damage That winters snowfall especially heavy and the potential for flood well known.
Two months before the spring melt started, NWS predicted the Red River would crest at 49.
Levees around the river were built for 51 flood.

A story about variance

1997 Grand Forks floods. Billions of damage That winters snowfall especially heavy and the potential for flood well known.
Two months before the spring melt started, NWS predicted the Red River would crest at 49.
Levees around the river were built for 51 flood.
54 actual flood maximum.

A story about variance

1997 Grand Forks floods. Billions of damage That winters snowfall especially heavy and the potential for flood well known.
Two months before the spring melt started, NWS predicted the Red River would crest at 49.
Levees around the river were built for 51 flood.
54 actual flood maximum.
But NWS prediction was +/- 9. They did not want to talk about the variance because they were afraid prople would not believe them if notions of uncertainty were discussed.
There was a 1-in-3 chance the river would overtop the 51 levee.

Quartiles

Divide the sample set into quarters and plot the marks of first and third quartile.

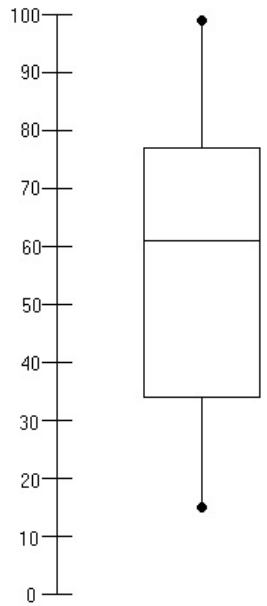
Think of it this way - the first quartile is the median of the items **below** the median. The third quartile is median of the items **above** the median.

One could do a similar thing in tenths.

Box plot

A diagram that show the maximum, minimum, the first and third quartiles, and the median.

$\text{boxnumbers} = \{12, 20, 35, 38, 45, 60, 62, 70, 78, 90, 99\}$

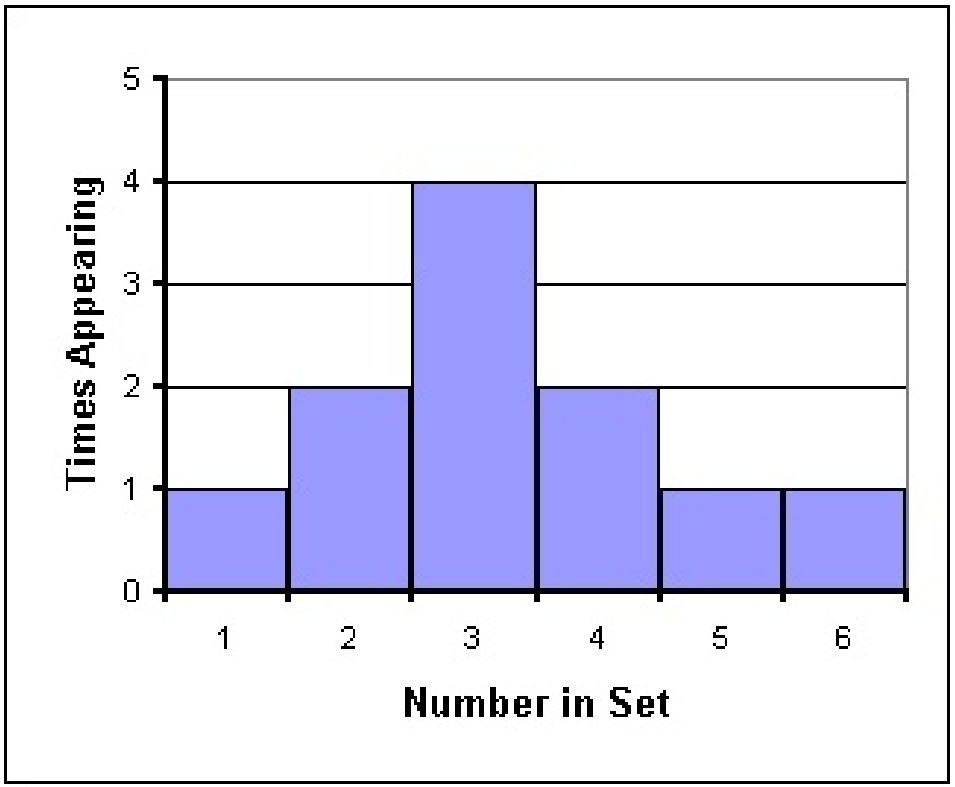


If there are a whole lot of numbers and the max/min are way out, you sometimes mark the 10% and 90% points.

Histogram

Plot items using rectangles to represent the number of items within a range of values.

$\text{agenumbers} = \{1, 2, 2, 3, 3, 3, 3, 4, 4, 5, 6\}$



Histograms

You need to decide how many rectangles to include. Too few doesn't give a sense of numbers and frequency.

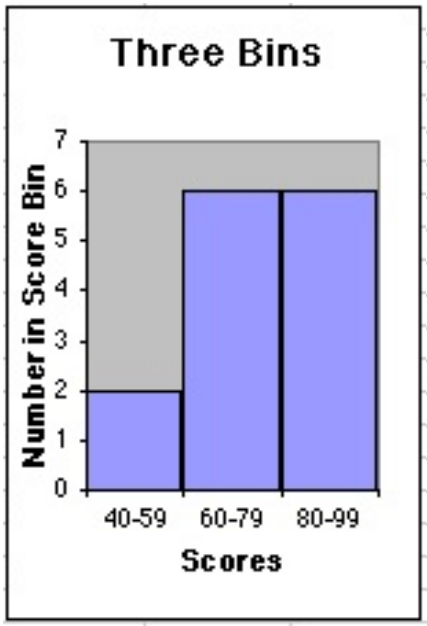
Too many is both more work and often doesn't give a good frequency reading – most of the time the numbers who appear are unique.

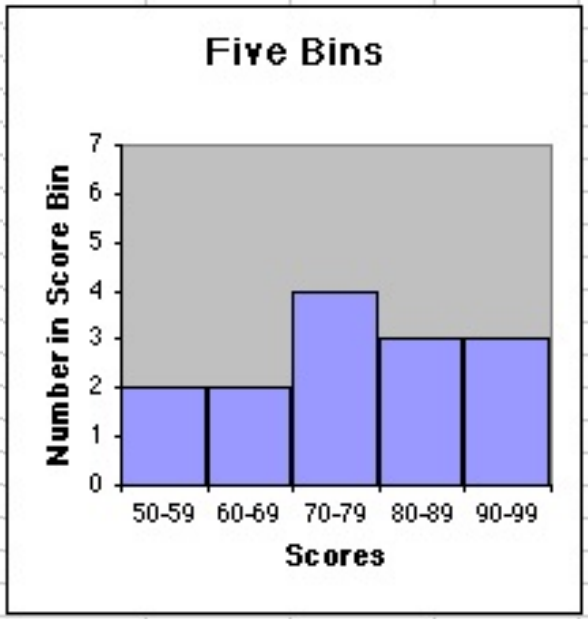
This is referred to as "binning".

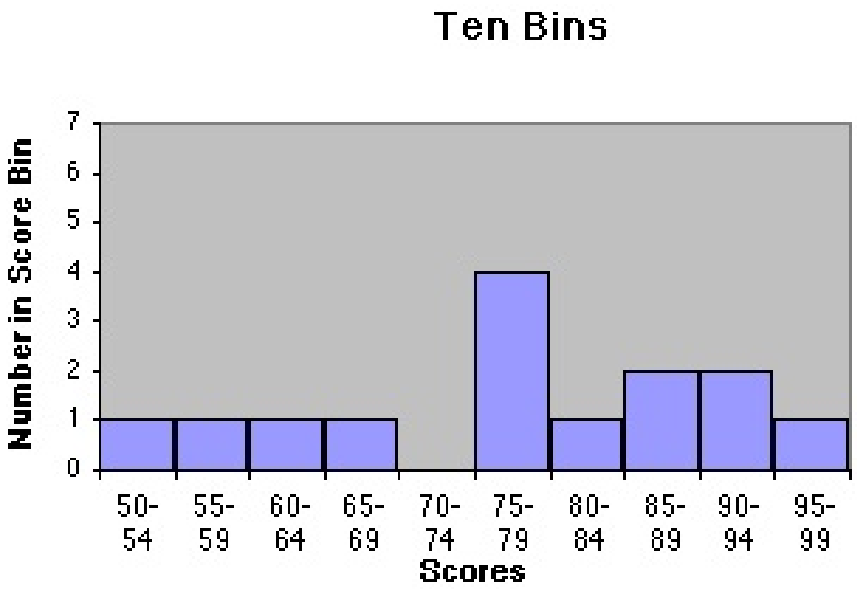
More histograms

Name	Grade
Bullwinkle	84
Rocky	91
Bugs	75
Daffy	68
Wylie	98
Mickey	78
Minnie	77
Lucy	86
Linus	94
Charlie	64
Patty	59
Donald	54
Sam	89
Taz	76

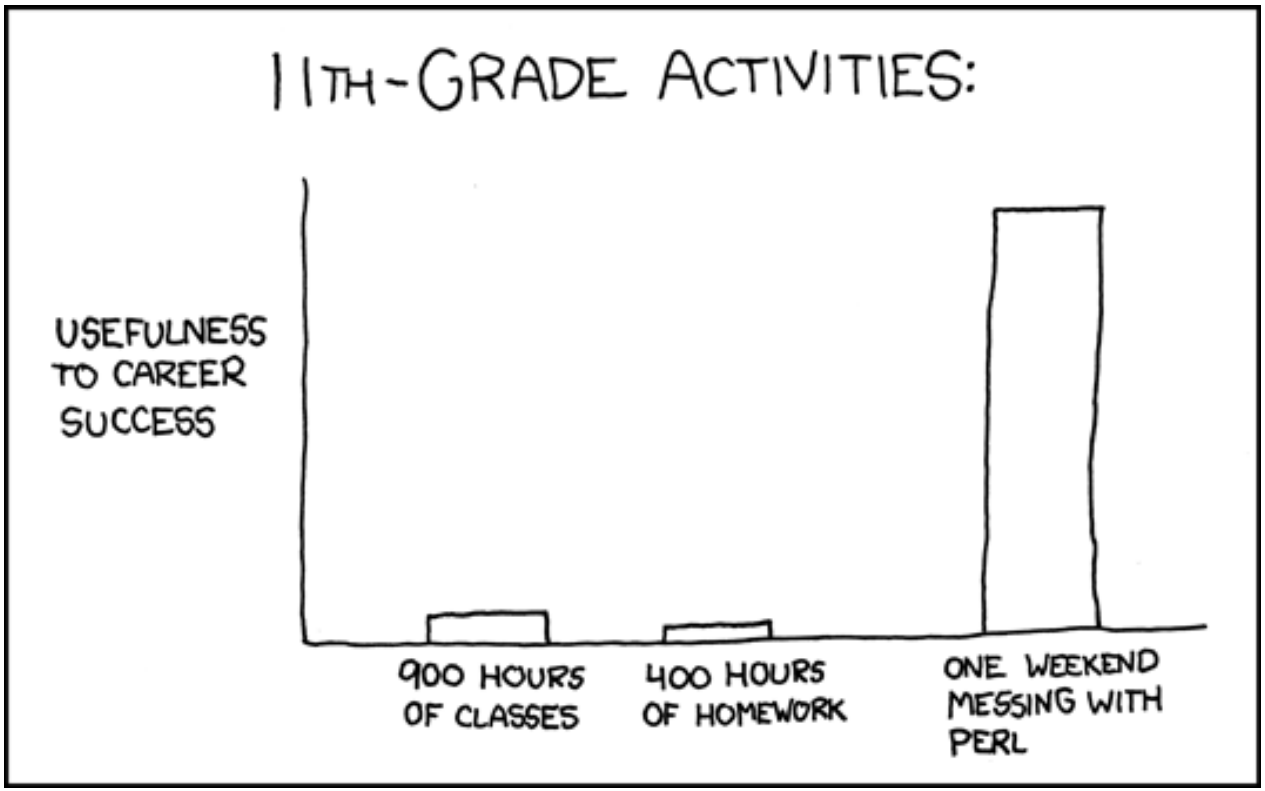
More histograms-2





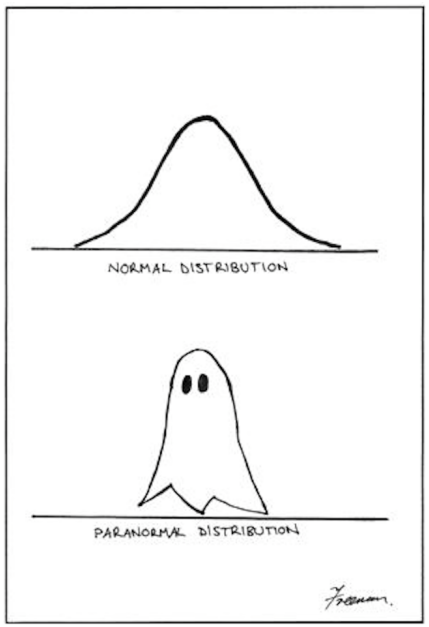


More histograms-3



Normal Distribution

What happens then when there are infinitely many bins - one for every real number? You get a probability distribution – a function. The grandaddy of distributions is the "Normal distribution". The normal distribution is the bell-shaped distribution you have probably seen.



Normal Distribution-2

The normal distribution has several properties that are useful in practice.

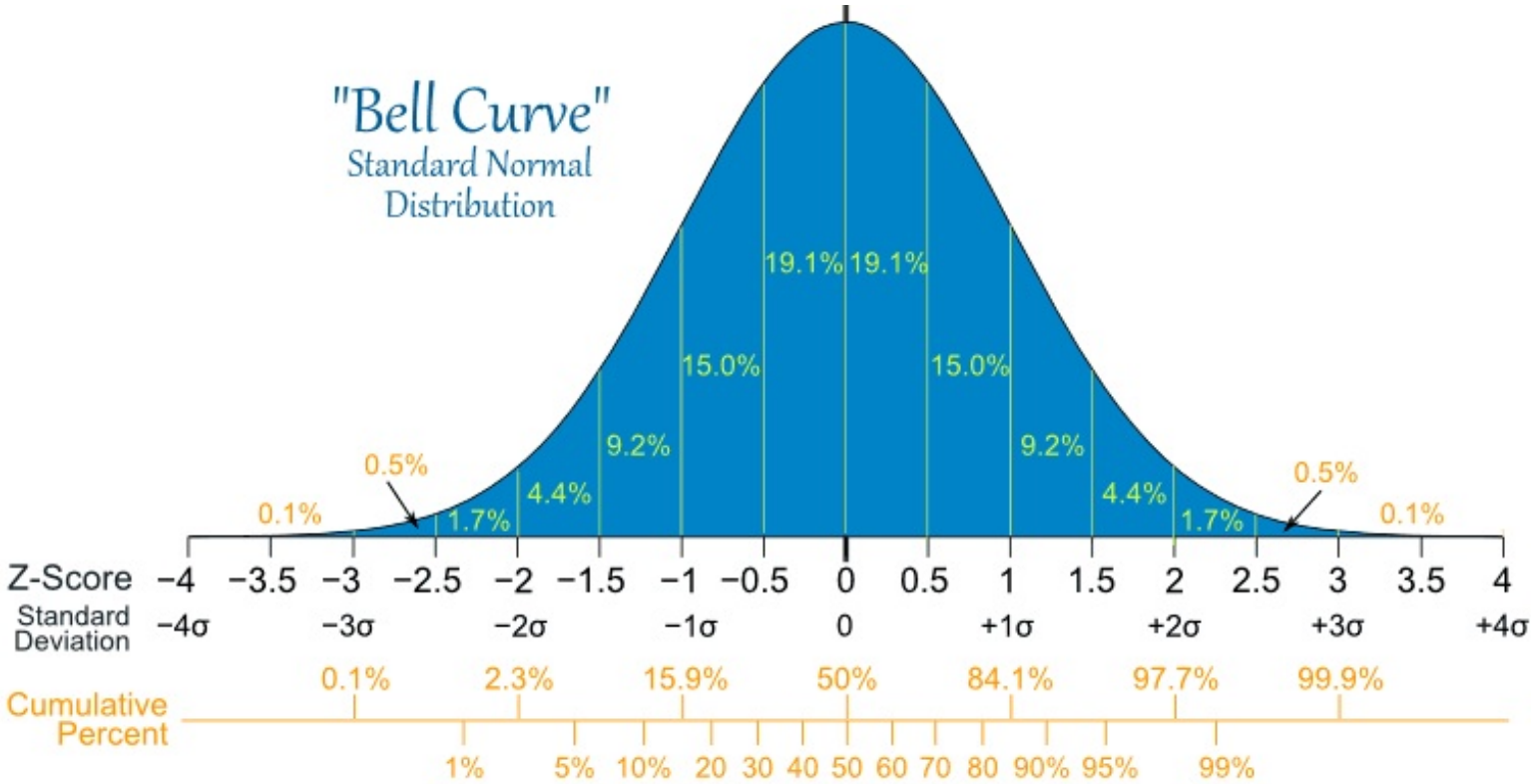
The mean is the highest value.

The distribution is symmetric about its mean.

Within ± 1 standard deviation (i.e. $\pm 1\sigma$) of the mean is 68% of the mass of the distribution, and 95% within $\pm 2\sigma$.

The first and third quartiles are at about $\pm .67\sigma$.

Normal Distribution-3



Other distributions

Log-normal, Chi, and there are others.
Useful in many applications.
But they (usually) don't have the nice properties above. Although, for example, the log-nomal is always positive.

Conditional probability

If a probability represent the odds of something happening – say the odds of event A happening, or $p(A)$, then the "conditional probability" is the odds of something happening given additional information – the odds of A given that B has occurred, or $p(A|B)$.

Lots of words but the idea is simple. A conditional probability accounts for additional information that informs the odds.

So the odds of rolling a 3 with a single die is $1/6$. What is the conditional probability of rolling a 3 given the roll was odd? It is $1/3$. That is, $A = \{1, 3, 5\}$, so $P(3|\text{odd}) = 1/3$.

Conditional probability-2

Formally the conditional probability is given as

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Notice how this works for rolling a 3. Since rolling a 3 and rolling an odd number are dependent, the probability of rolling a 3 AND odd is the same as rolling a 3.

$$p(3|\text{odd}) = \frac{p(3 \cap \text{odd})}{p(\text{odd})} = \frac{1/6}{1/2} = 1/3$$

Exercise in counting

A fair coin is flipped three times. What is the probability of at least one head? Given that the first flip came up tails, what is the probability of at least one head?

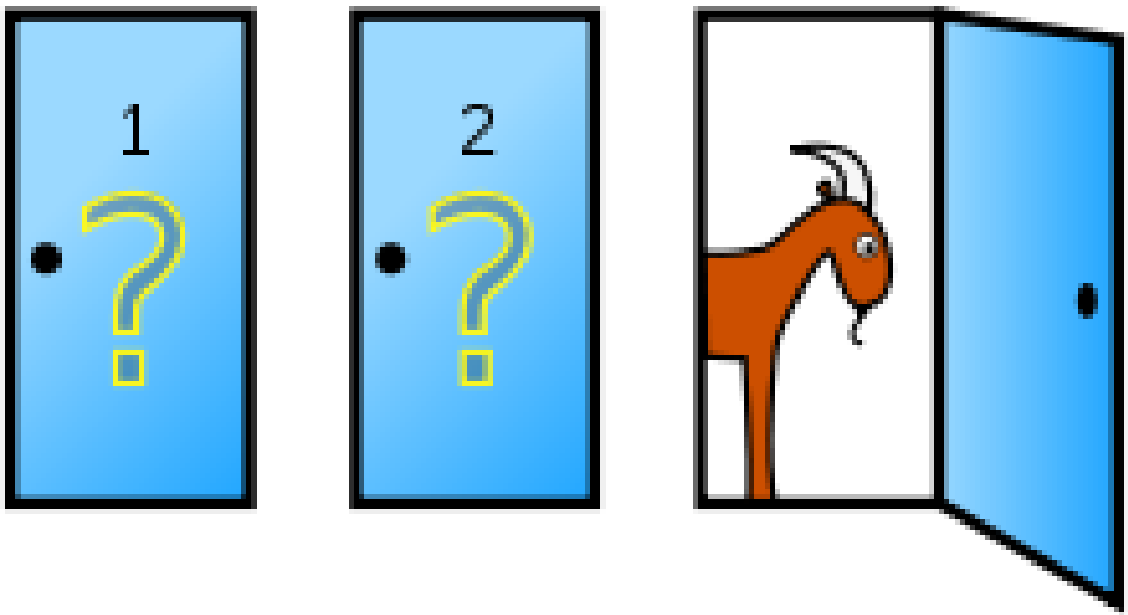
Exercise in counting-2

The sample space
$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.
The event "at least 1 head" is
$E1 = \{HHH, HHT, HTH, THH, HTT, THT, TTH\}$ which has a 7/8 probability.
The event "first flip tails" is $E2 = \{THH, THT, TTH, TTT\}$. So
$P(\text{at least one head}|\text{first flip tails}) = 3/4$.

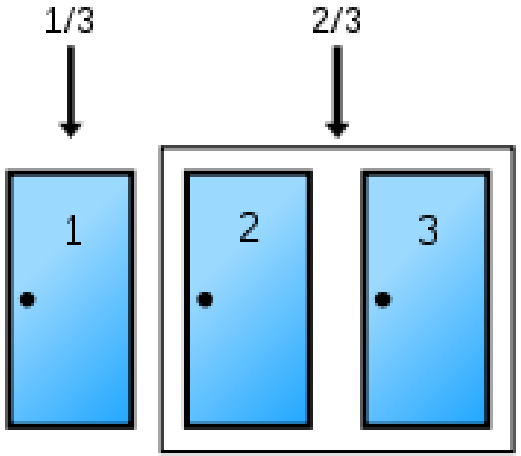
Sounds too good to be true

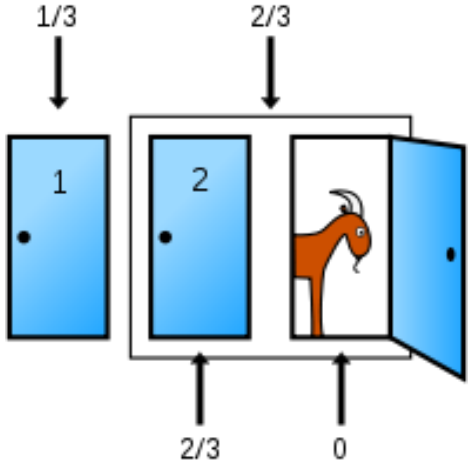
Monty Hall problem

Monty lets you choose a door from 3 possibilities. Behind one of the doors is a new car, behind the other two are goats. You choose door 1. Monty reveals what is behind one of the remaining doors (he knows where the car is), and asks you Do you want to switch doors or stick with your original choice?
What should you do?



Monty solution





Monty solutuon-2

Door 1	Door 2	Door 3	Result if stay with 1	Result if switch
car	goat	goat	car	goat
goat	car	goat	goat	car
goat	goat	car	goat	car

Bayes Theorem

From the definition of conditional probability

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

But it is equally true that we can reverse the roles of A and B

$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

Solving for the intersection we have

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Conditional probability again

An easy and more private test for the AIDS antibodies was developed, particularly for use in Africa, where men won't to go clinics to be tested. This new test could be administered at home. If you carried the antibodies, the test would confirm 99.1% of the time with a 0.9% false negative. The test gave a true negative 99.6%, and false positive 0.4%.
So the question is: If you administered the test, and it came back positive, how likely is it that you actually have AIDS?

Conditional probability again-2

Well in the US, the incidence of AIDS is about .5% or so.
If you think about it, the fact that the test gives false positives about 0.4% of the time (so 4 for every 1000 people tested) should give pause when 5 of every 1000 people tested actually carry the antibodies.
The detailed arithmetic calculates

$$\begin{aligned} P(D|+) &= \frac{P(+|D) * P(D)}{P(+|D) * P(D) + P(+|N) * P(N)} \\ &= \frac{0.991 * 0.005}{0.991 * 0.005 + 0.004 * .995} \\ &= 0.55456 \end{aligned}$$

i.e. about a 55.5% chance you have AIDS and 44.5% chance you don't.
Why such a low probability?

Conditional probability again-3

In Natal, South Africa, the incidence of AIDS is about 40%. Repeating the same calculation, but with this incidence rate one has

$$\begin{aligned} P(D|+) &= \frac{P(+|D) * P(D)}{P(+|D) * P(D) + P(+|N) * P(N)} \\ &= \frac{0.991 * 0.4}{0.991 * 0.4 + .004 * 0.6} \\ &= 0.994 \end{aligned}$$

i.e. 99.4% chance you DO have AIDS

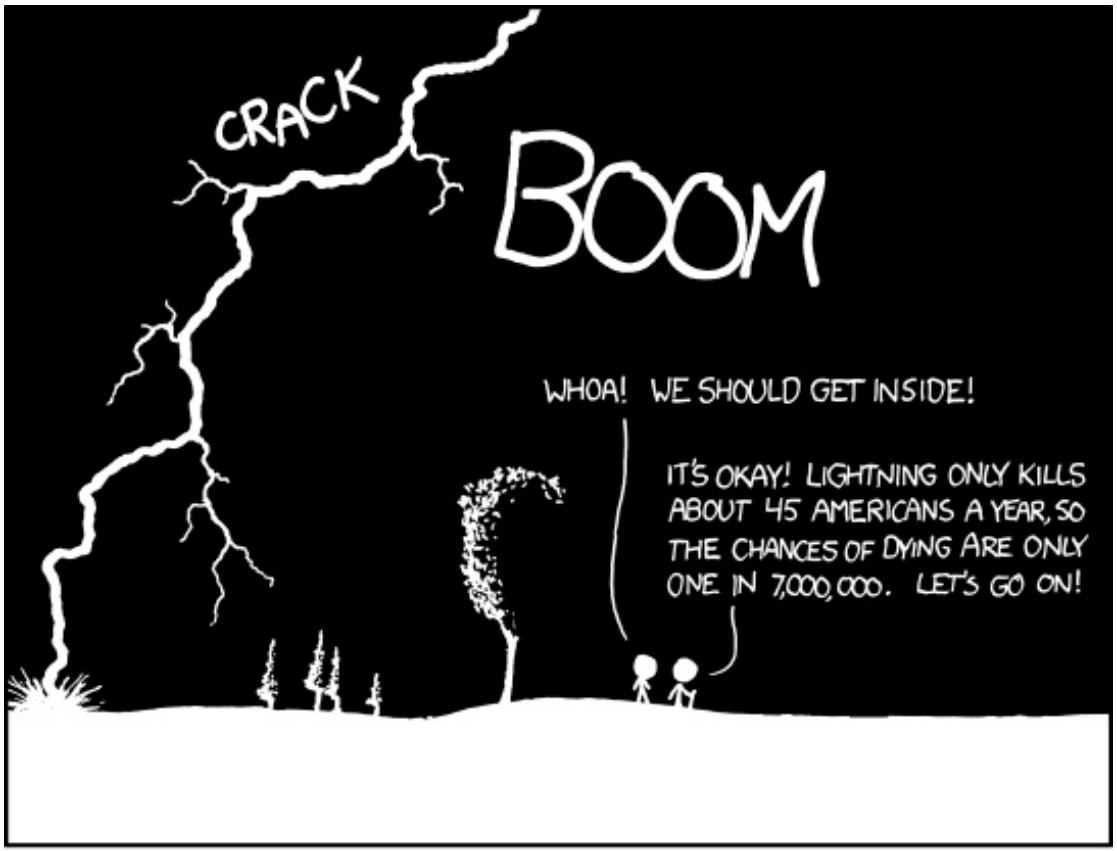
Conditional probability again-4

OK so what is really going on?

Say we have 1000 people in Africa. In a random sample 400 of them have AIDS, and virtually all of them will test positive. An additional 4 of the 600 who are healthy will also test positive (false positive). So if you are one of the 404 people who test positive, your odds are 400/404 that in fact you have AIDS.

In the US, of 1000 people only 5 will have AIDS, and 4 others will test positive. So if you test positive, your odds of having AIDS are 5/9 or about 55%.

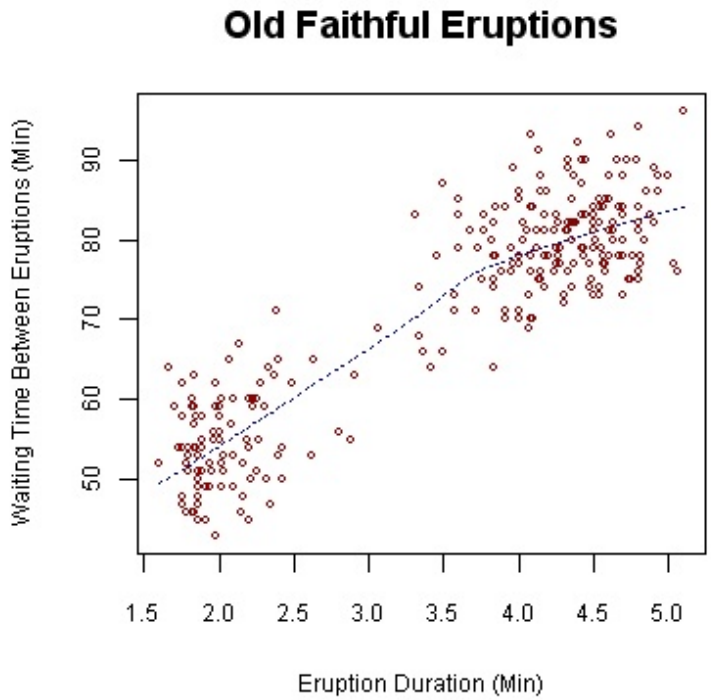
So now we know more statistics



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

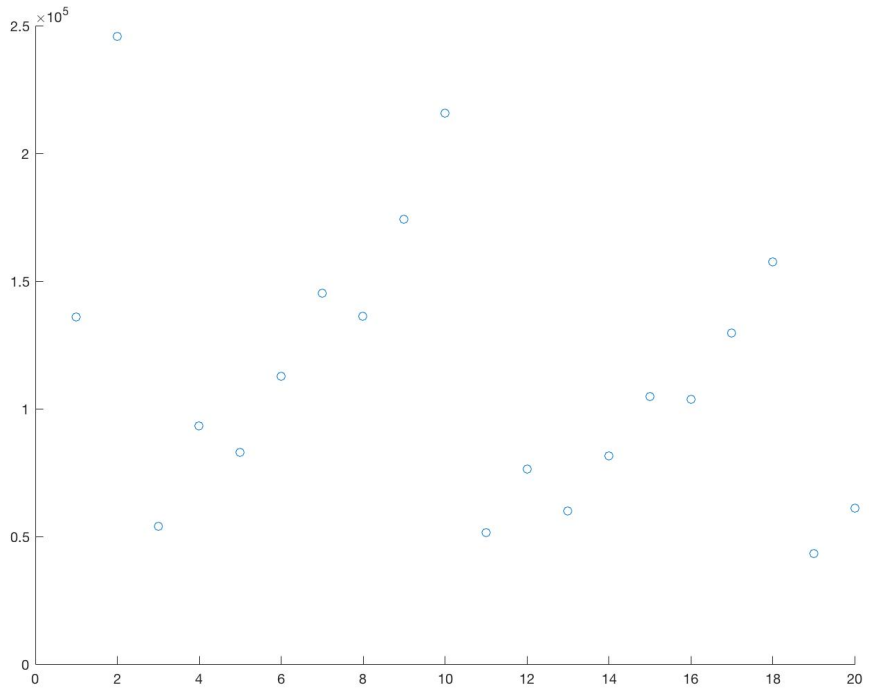
Scatter plot

A scatter plot is a diagram showing two variables of a dataset.



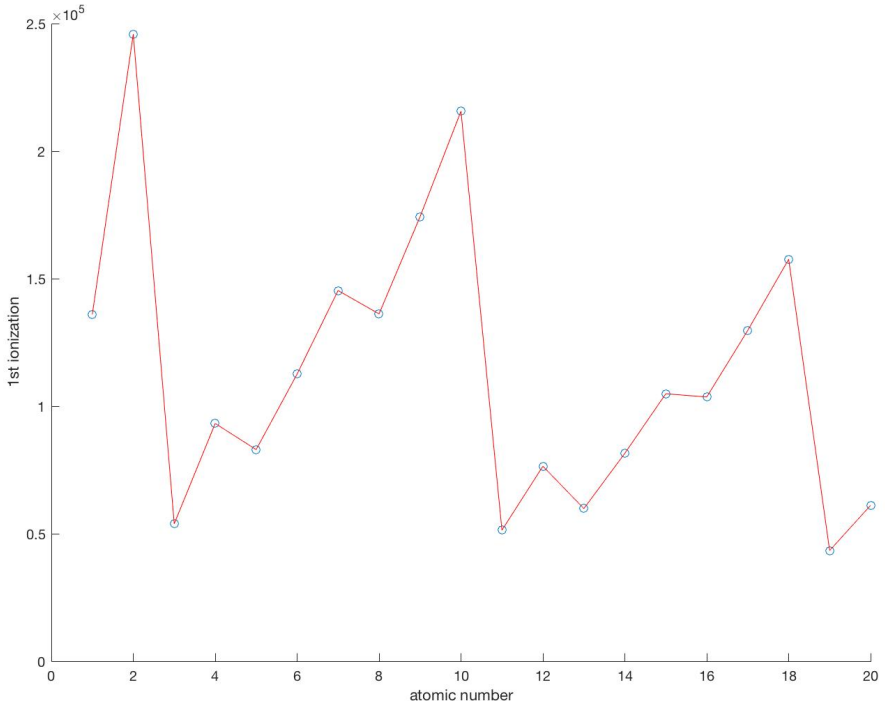
More scatter plot

If I just give you the points on a graph, what do you see?



More scatter plot

If I add labels and a fitting curve, does it take on more meaning for you?



Correlation and causation

In statistics, dependence refers to any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationships involving dependence.

Because two things are correlated does not mean one causes the other.

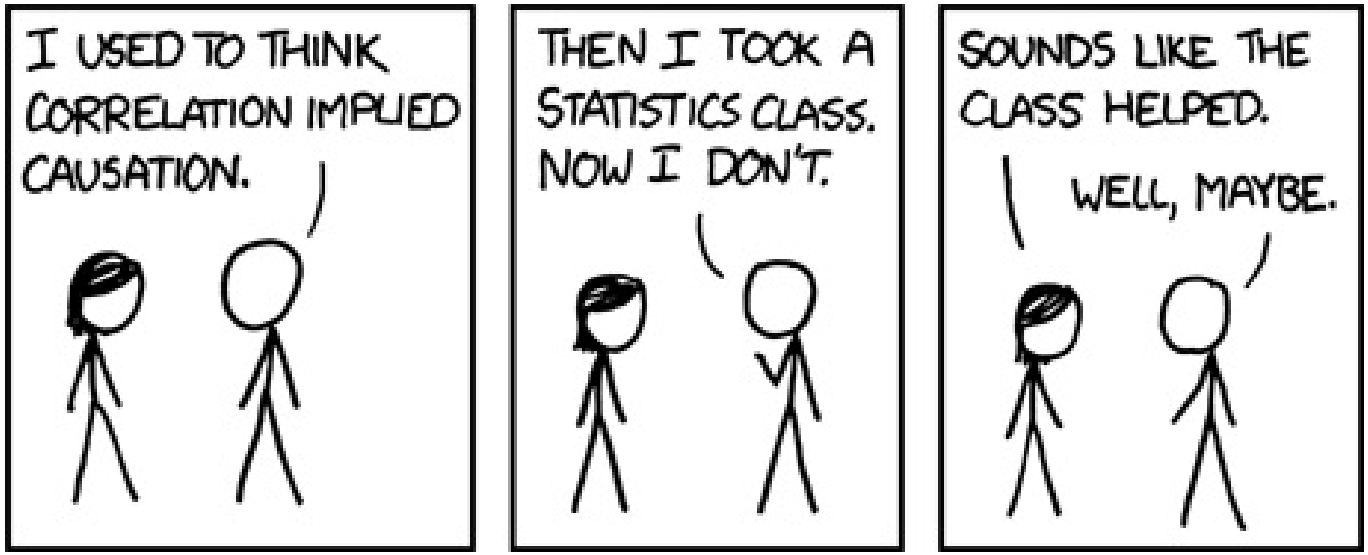
A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health, or does good health lead to good mood, or both? Or does some other factor underlie both? In other words, a correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

Measure of correlation

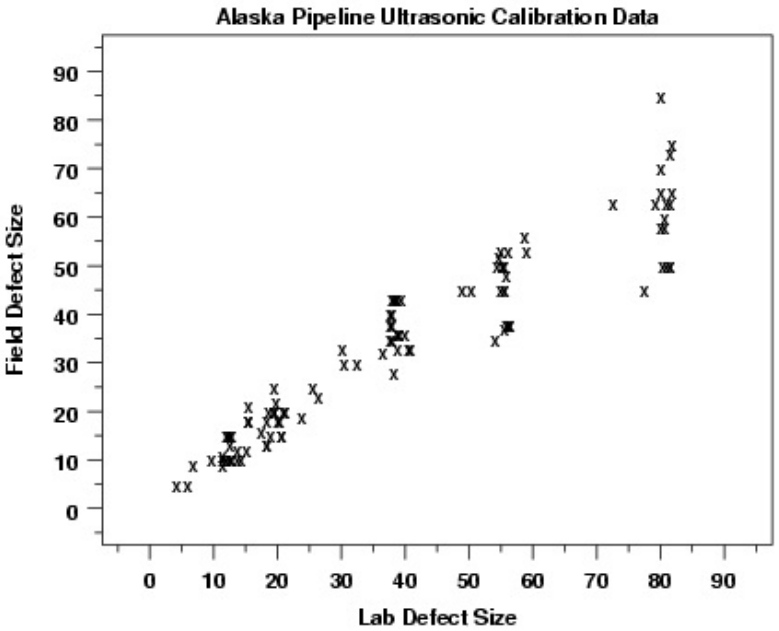
There is a definition of a coefficient of correlation

$$\rho(X, Y) = \frac{E[(X - \overline{x})(Y - \overline{y})]}{\sigma_X \sigma_Y}$$

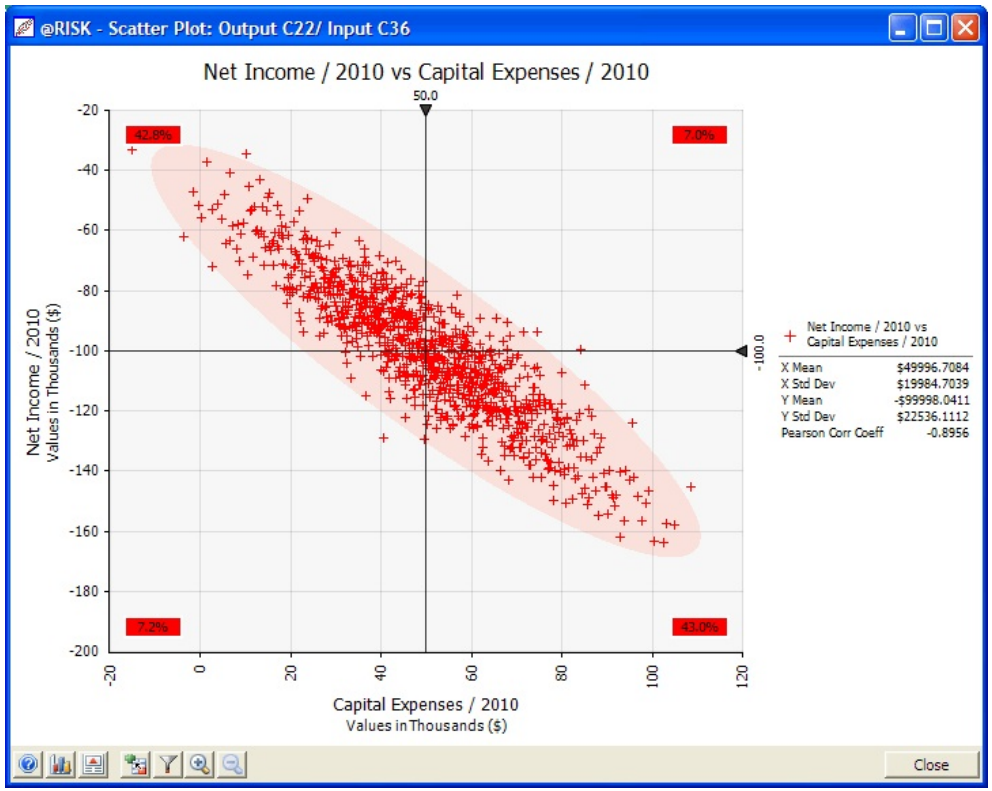
Correlation and causation



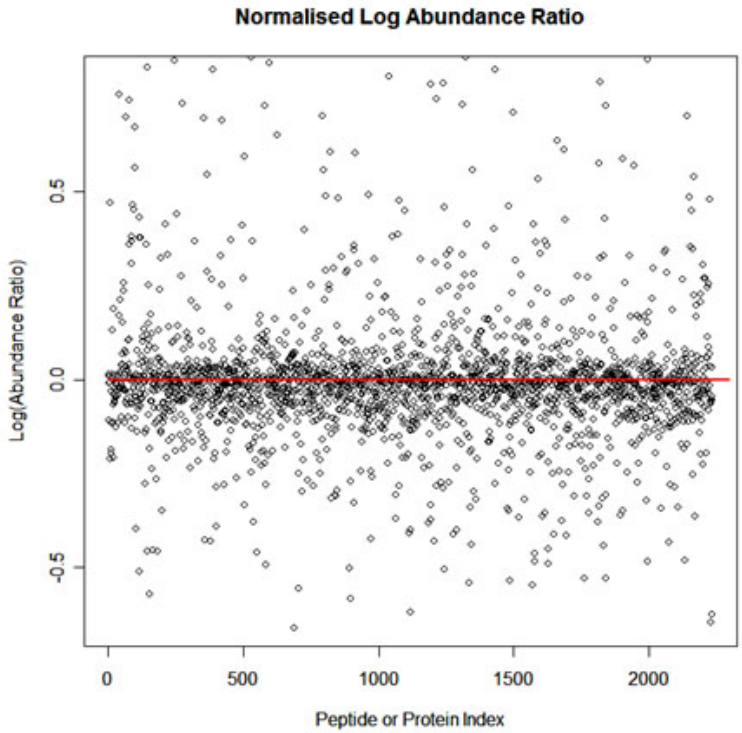
Positive, negative, and no correlation



Positive, negative, and no correlation-2



Positive, negative, and no correlation-3



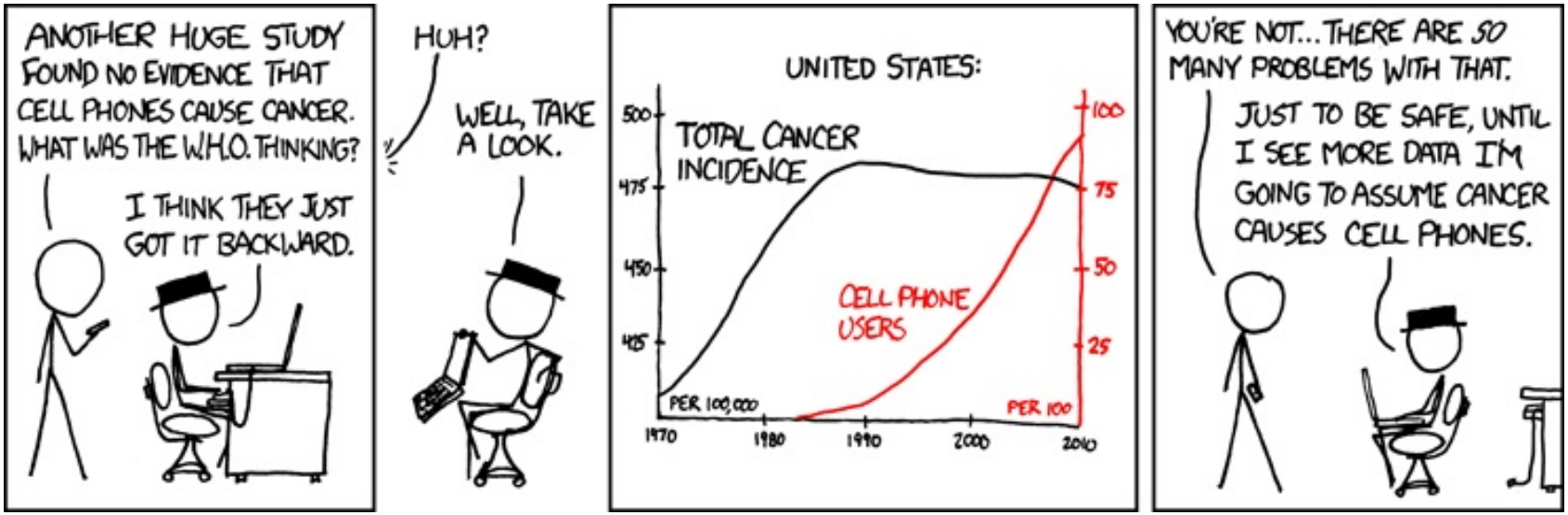
Correlation and causation

In a widely studied example, numerous epidemiological studies showed that women taking combined hormone replacement therapy (HRT) also had a lower-than-average incidence of coronary heart disease (CHD), leading doctors to propose that HRT was protective against CHD. But randomized controlled trials showed that HRT caused a small but statistically significant increase in risk of CHD. Re-analysis of the original data showed that women undertaking HRT were more likely to be from higher socio-economic groups, and thus had better-than-average diet and exercise regimens. The use of HRT and decreased incidence of coronary heart disease were coincident effects of a common cause (i.e. the benefits associated with a higher socio-economic status), rather than cause and effect, as had been supposed.

Correlation and causation-3

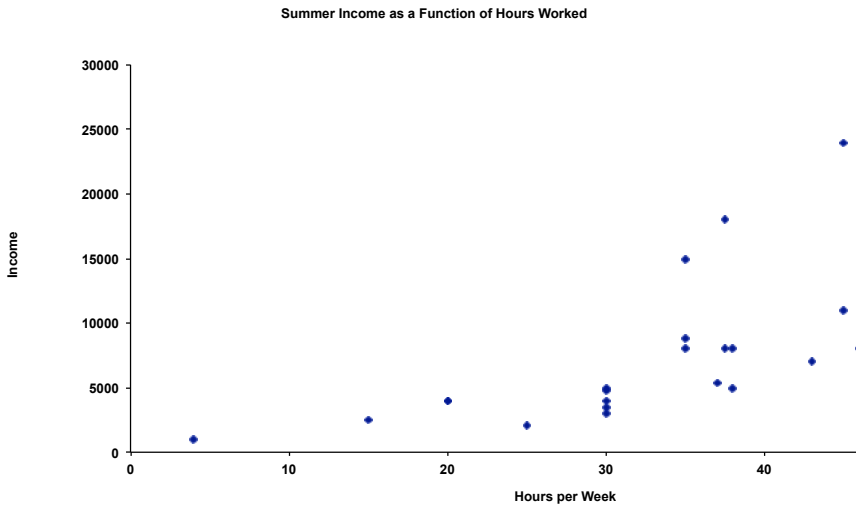
As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream consumption causes drowning.

Correlation and causation-4

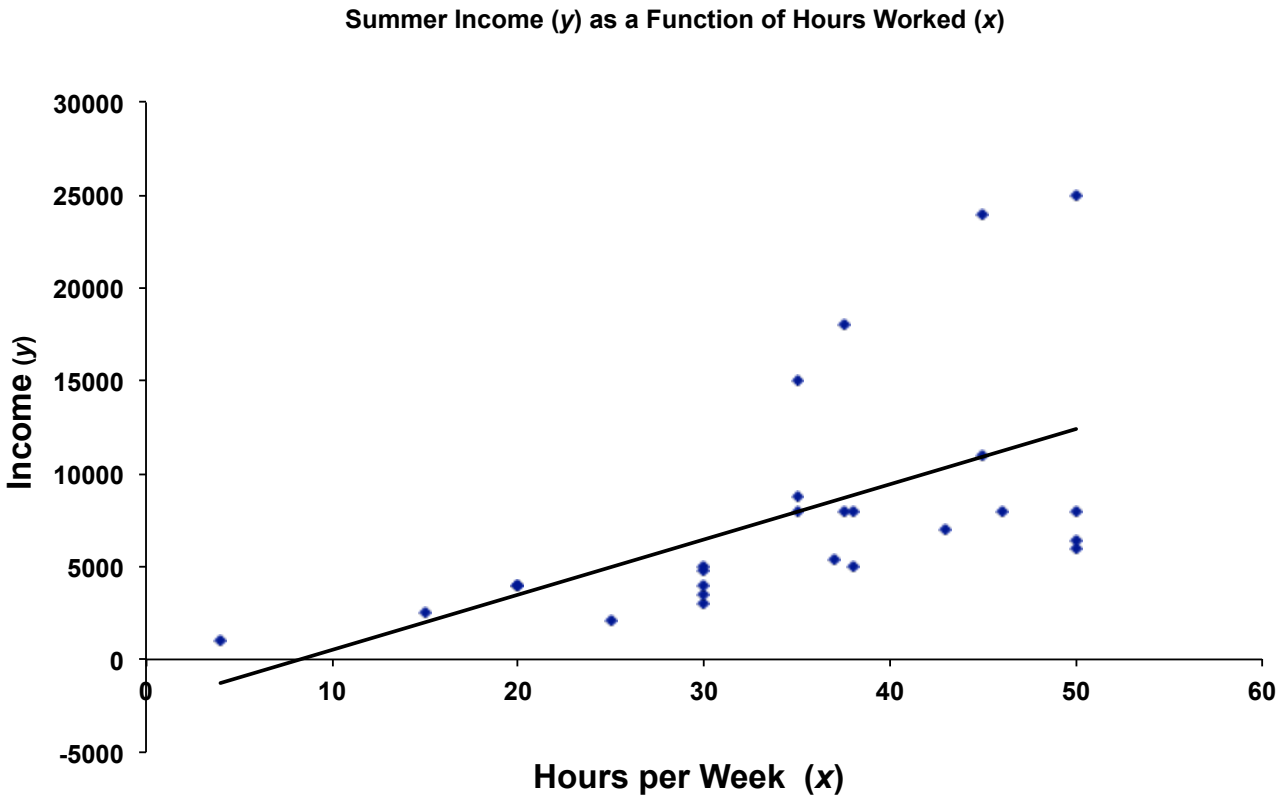


Regression

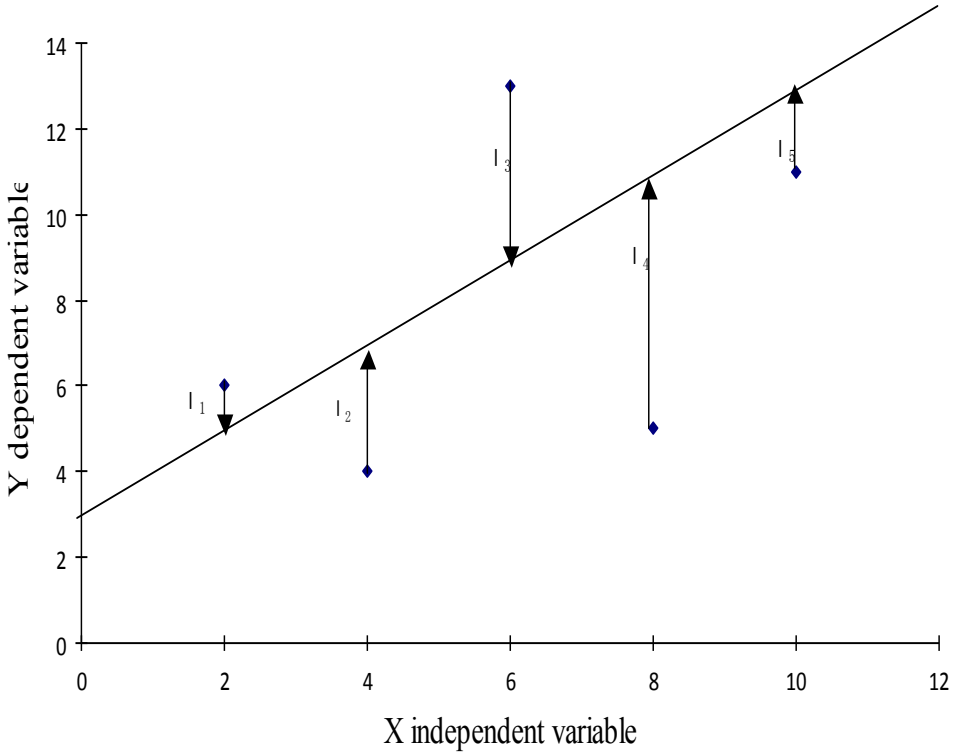
Given the data in, say, a scatterplot, can you draw a line (or some other curve) that fits the data well?



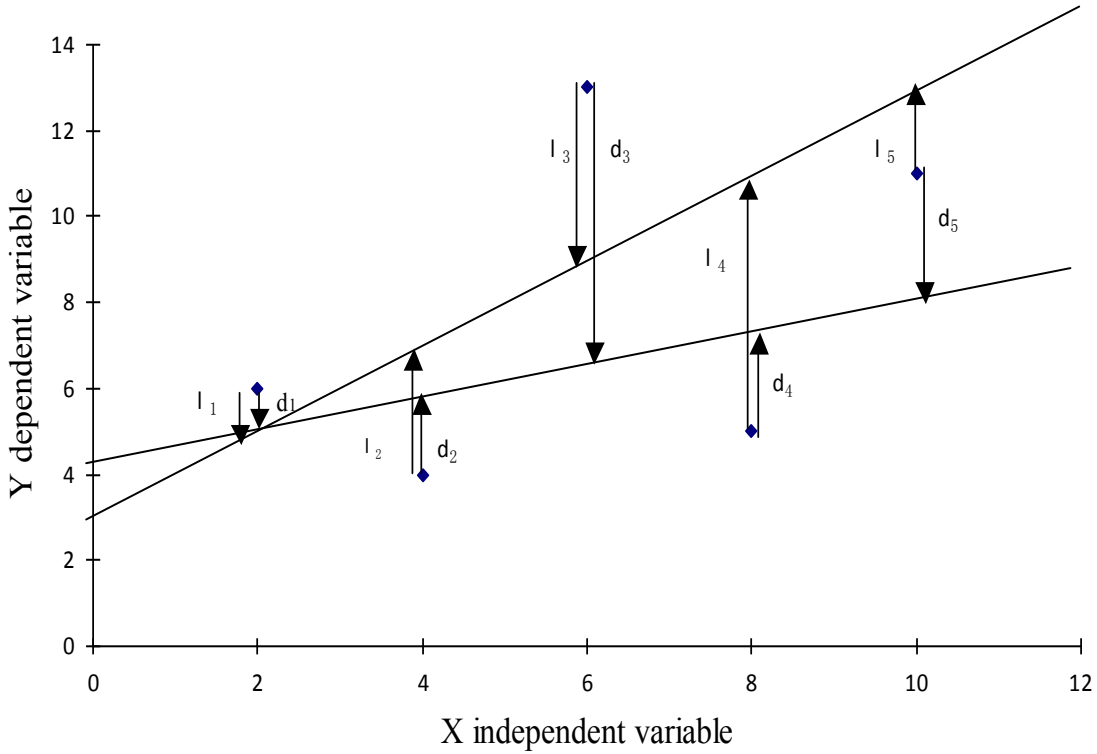
Regression 2



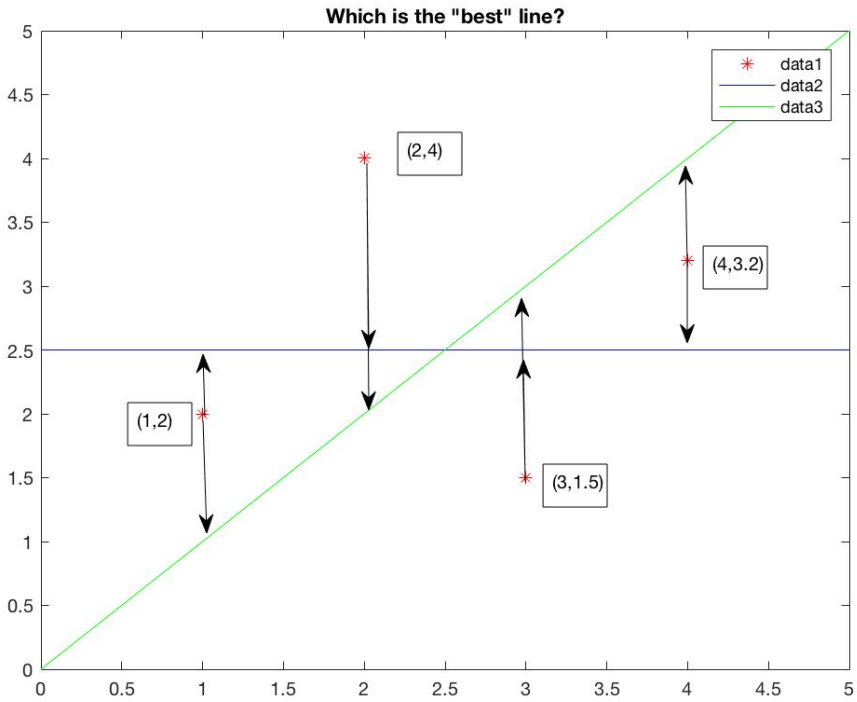
How to tell which is the "best" line?



How to tell which is the "best" line?



Regression 4

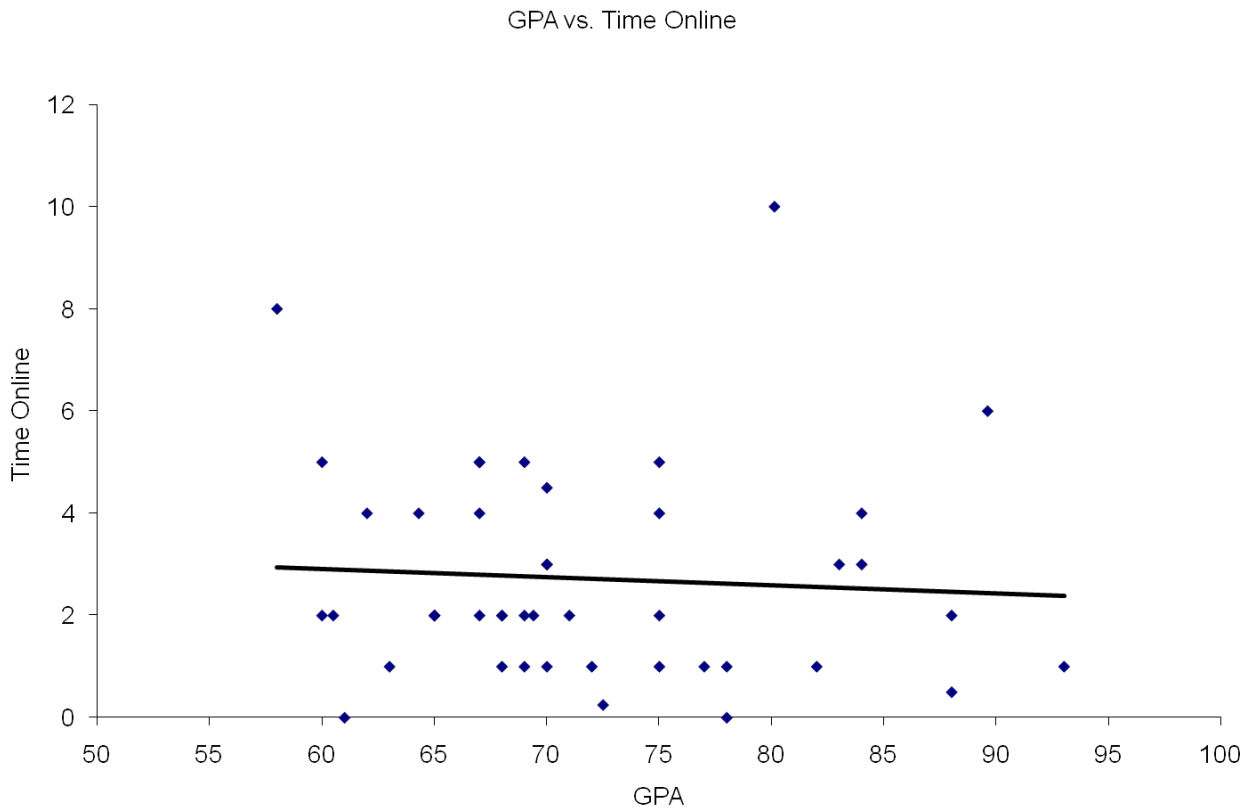


The best fit goes back to the correlation coefficient

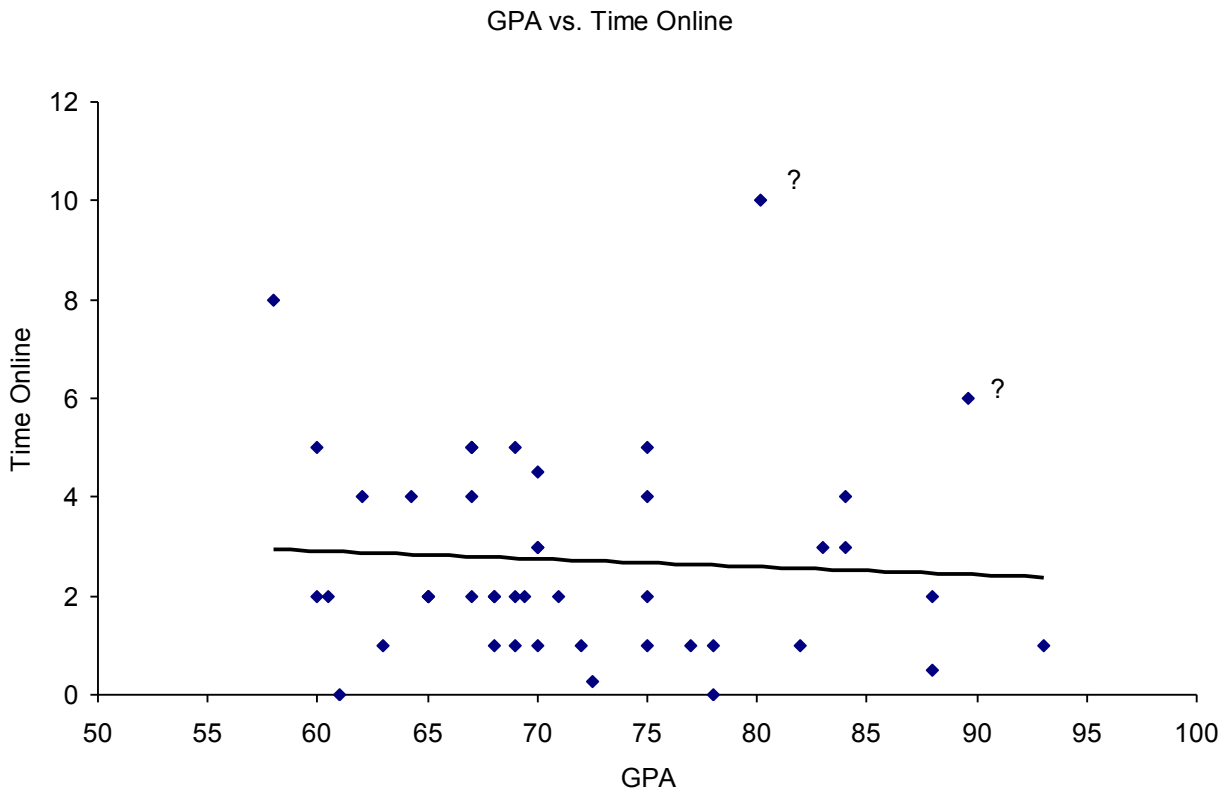
$$\rho(X, Y) = \frac{E[(X - \bar{x})(Y - \bar{y})]}{\sigma_X \sigma_Y}$$

The idea is to minimize this (actually the square of this)

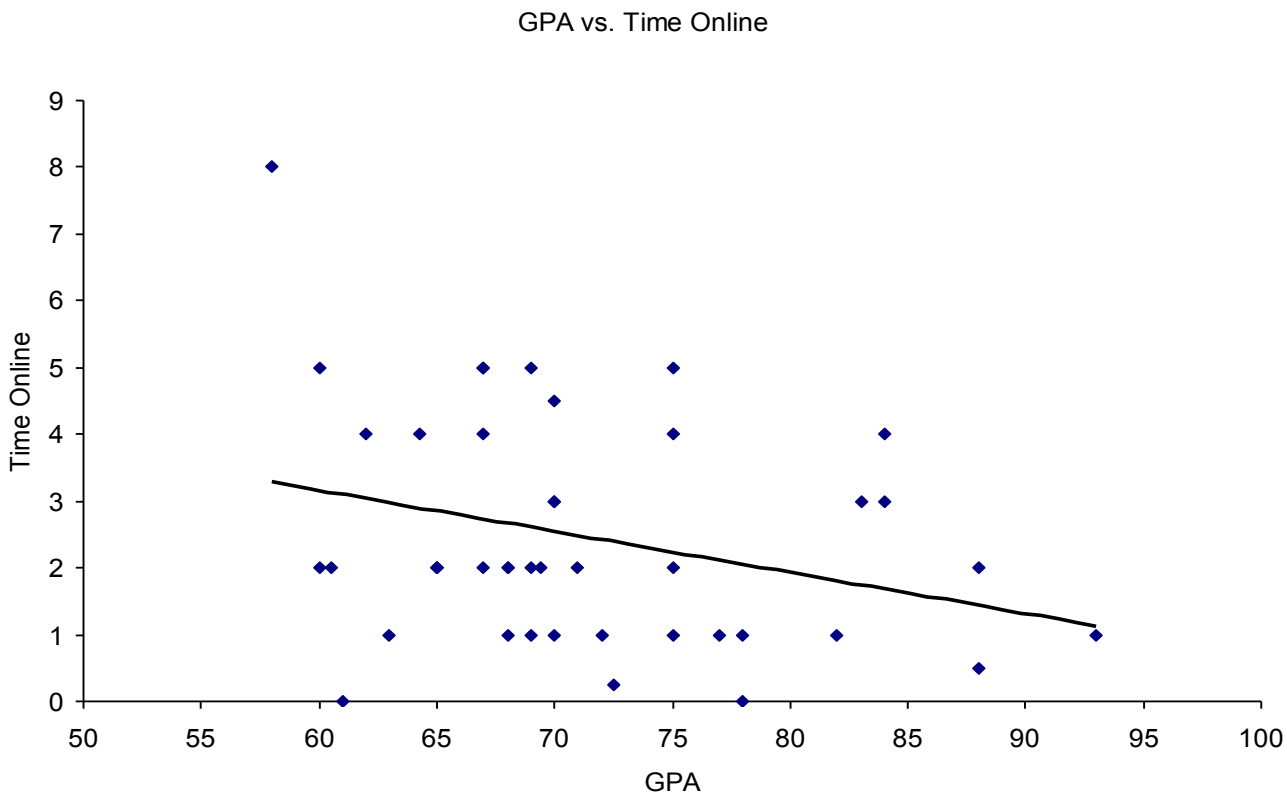
Outliers


GPA vs. Time Online
Time Online
GPA
0
2
4
6
8
10
12
50
55
60
65
70
75
80
85
90
95
100

Outliers - 2


GPA vs. Time Online
Time Online
GPA
0
2
4
6
8
10
12
50
55
60
65
70
75
80
85
90
95
100
?
?

Outlier - 3


GPA vs. Time Online
Time Online
GPA
0
1
2
3
4
5
6
7
8
9
50
55
60
65
70
75
80
85
90
95
100

Outliers - 4

But you always want to look carefully and ask Is this datum telling me something important?
For example, it could signify a low probability but high impact event.

Other Statistical Predictions

If you have a large sample size, you can make predictions about behavior in the aggregate.
For example, Amazon uses your own buying habits and those of others like you, to give you recommendations for other purchases.
The Federal Reserve makes predictions on the growth of the economy.

Statistical Predictions - 2

A different kind of prediction arises from statistical transitions. For example, data on home ownership shows that, if you live in a single family home currently, odds are 95% that you will be in a single family home next year. On the other hand, if you live in an apartment/multi-household setting currently, odds are 15% that you will live in a single family unit next year. We have enough data to be able to make a claim like this.

Statistical Predictions - 3

We can create an array that explains the situation, showing the current status on the left, and your status next year from the top.

$$\left(\begin{array}{c|cc} & \text{single} & \text{multi} \\ \hline \text{single} & 0.95 & 0.05 \\ \text{multi} & 0.15 & 0.85 \end{array}\right)$$

You can repeat this for subsequent years too.

Statistical Predictions - 4

Be wary of the variability in your predictions.
The average weather in WNY is 53° and a 40% chance of precipitation.
The varibility – that is, variance – matters!

Now that you are an expert