Ideas of Probability
Basic Statistics
Spread of items
Distribution
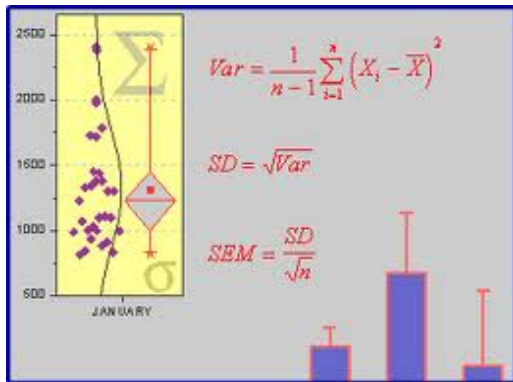More Probability
Making Predictions

The Elements of Probability and Statistics

E. Bruce Pitman

The University at Buffalo

CCR Workshop – July 3, 2013

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Outline

1. Ideas of Probability

2. Basic Statistics

3. Spread of items

4. Distribution

5. More Probability

6. Making Predictions

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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 outcomes n.

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

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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 commit the Gamblers Fallacy if you believe your losing streak makes it more likely that youll roll the numbers you want on the next roll (because youre "due"). The truth is that your odds don't change; you start over with each roll.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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 $E = \{1, 3, 5\}$.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

One perspective

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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"?

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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"?

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

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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"?

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

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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"?

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

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

More counting

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

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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 "EJ Manuel is the starting quarterback in December" are dependent.

So now we know some statistics

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions
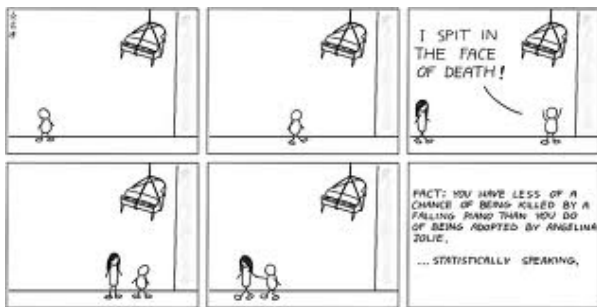
Variance

Variance is how wide of a spread is pesent 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, $\sigma = \sqrt{10}$

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Box plot

A diagram that show the maximum, minimum, the first and third quartiles, and the median.
$\text{boxnumbers} = \{12, 20, 35, 38, 45, 60, 62, 70, 78, 90, 99\}$
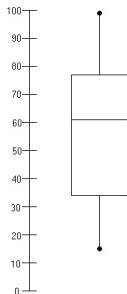
If there are a whole lot of numbers and the max/min are way out, you sometimes mark the 10% and 90% points.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Histogram

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

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

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions
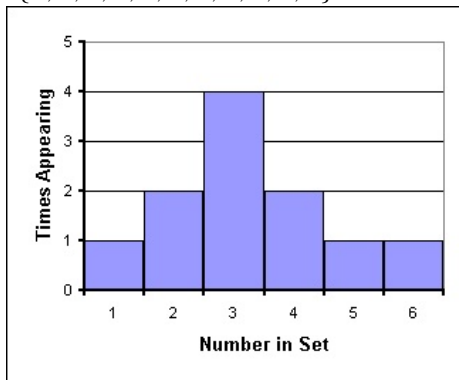
Histograms

You need to decide how many rectangles to include. Too few doesn't give a sense of numbers and frequency.
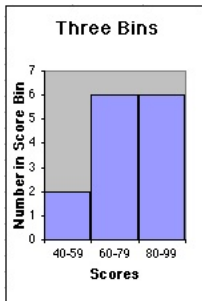
Too many is both more work and often doesn't give a good frequency reading – most of the time the numbers who appear are unique.
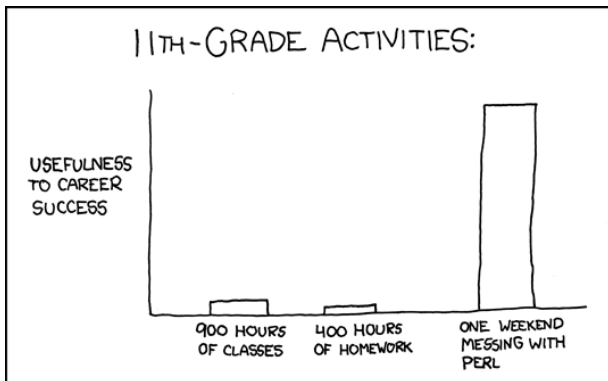
This is referred to as "binning".

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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
Donld	54
Sam	89

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

More histograms-2

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

More histograms-3

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions
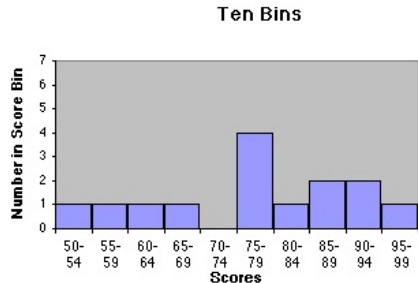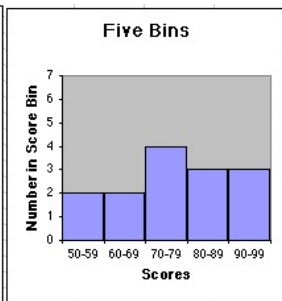
Normal Distribution

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.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions
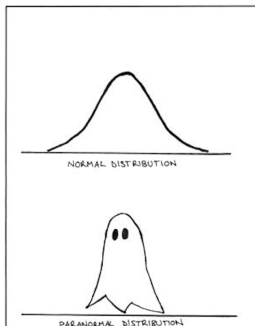
Normal Distribution-2

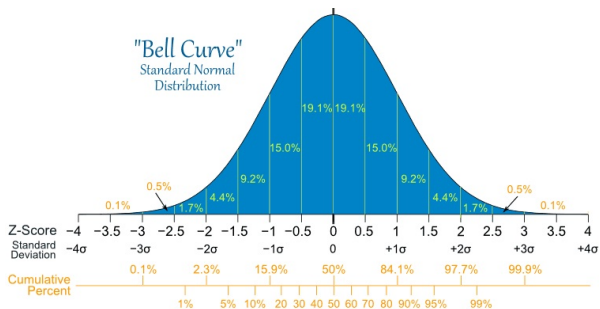The normal distribution has several properties that are useful in practice.

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 $\pm .67\sigma$.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Normal Distribution-3

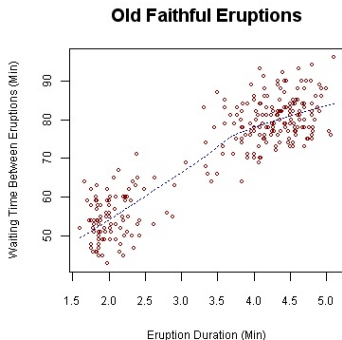## Other distributions

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.)

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Scatter plot

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

Old Faithful Eruptions

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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 $P(3|\text{odd}) = 1/3$.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

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

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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?

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Exercise in counting-2

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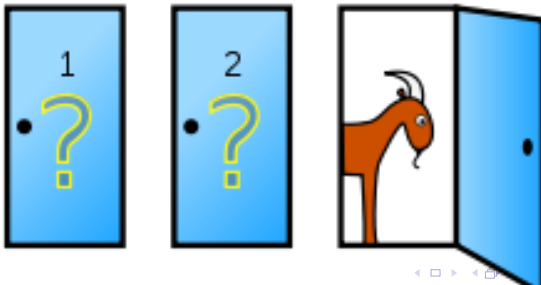$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.
The event "at least 1 head" is
$E1 = \{HHH, HHT, HTH, THH, HTT, THT, TTH\}$ which has a
7/8 probability.
The event "first flip tails" is $E2 = \{THH, THT, TTH, TTT\}$. So
$P(\text{at least one head}|\text{first flip tails}) = 3/4$.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
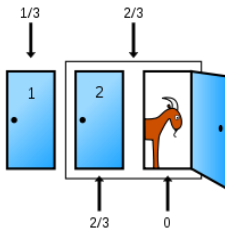**More Probability**
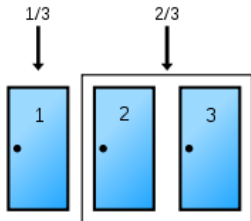Making Predictions

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 a door. 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?

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Monty solution

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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	got	car

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Conditional probability again

One half percent of the population has a particular disease. A test
is developed for the disease. The test gives a false positive 3% of
the time and a false negative 2% of the time.

(a) What is the probability that Joe (a random person) tests
positive?

(b) Joe just got the bad news that the test came back positive;
what is the probability that Joe has the disease?

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Conditional probability again-2

Let D be the event that Joe has the disease, ND that he doesn't.
Let T be the event that Joe's test comes back positive. We are
told that $p(D) = 0.005$, since $1/2\%$ of the population has the
disease, and Joe is just an average guy. We are also told that
$p(T|D) = .98$ since 2% of the time a person having the disease is
missed (false negative). We are told that $p(T|ND) = .03$ since
there are 3% false positives.
(a) We want to compute $p(T)$. We do so by conditioning on
whether or not Joe has the disease:

$$
\begin{aligned}
p(T) &= p(T|D)p(D) + p(T|ND)p(ND) \\
&= (.98)(.005) + (.03)(.995) = 0.03475
\end{aligned}
$$

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Conditional probability-3

(b) We want to compute

$$
\begin{aligned}
p(D|T) &= \frac{p(D \cap T)}{p(T)} = \frac{p(T|D)p(D)}{p(T|D)p(D) + p(T|ND)p(ND)} \\
&= \frac{(.98)(.005)}{(.98)(.005) + (.03)(.995)} = .14
\end{aligned}
$$

What a relief! There is only a 14% chance Joe has the disease,
even though the test came back positive! The issue here is that
the false-positive and false-negative percentages are in fact high,
relative to the occurrence of the disease.

Why such a low chance that Joe is ill?

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Conditional probability -4

The background rate is very low – half of a percent. And the corresponding false positive rate is (relatively) large.

Consider that happens is the background rate is 5%? Repeating the calculation

$$
\begin{aligned}
p(D|T) &= \frac{p(D \cap T)}{p(T)} = \frac{p(T|D)p(D)}{p(T|D)p(D) + p(T|ND)p(ND)} \\
&= \frac{(.98)(.05)}{(.98)(.05) + (.03)(.95)} = .63
\end{aligned}
$$

A huge change!!

Ideas of Probability
Basic Statistics
Spread of items
Distribution
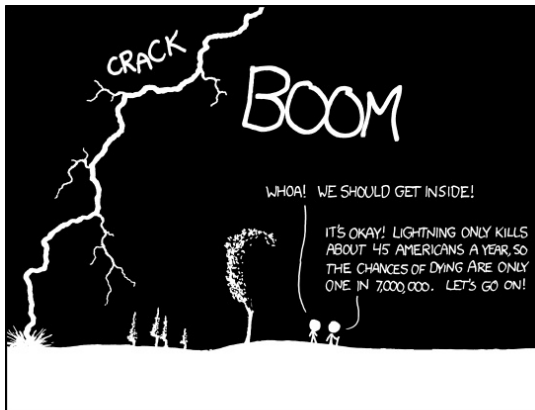**More Probability**
Making Predictions

Conditional probability -5

A different way of thinking about what is happening:

- Out of a thousand people, 5 will be ill
- Of these 5, all of them (just about) will get a positive test
- But 30 other people will also test positive for the disease (3%)
- So 35 people out one thousand test positive...but only $1/7$ of them are ill!!

Notice that in the alternative scenario with a 5% background rate, you would see 50 out of a thousand who are ill and will test positive, in addition to the 30 false positives – thus the 63%.

So now we know more statistics

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

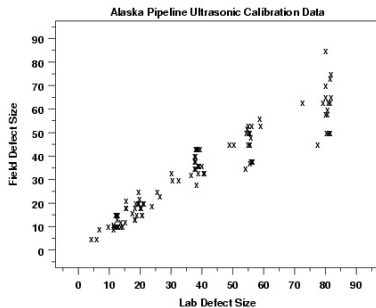Because two things are correlated does not mean one causes the other.

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.

Ideas of Probability
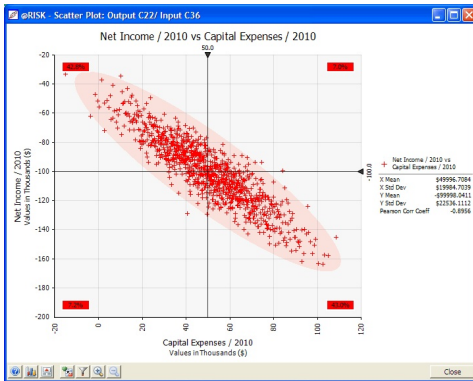Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Measure of correlation

There is a definition of a coefficient of correlation

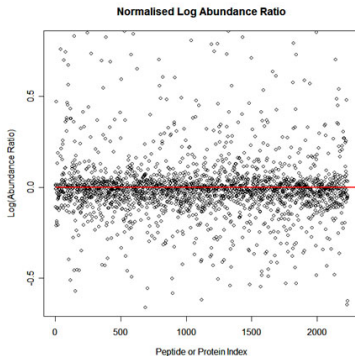$$\rho(X, Y) = \frac{E[(X - \bar{x})(Y - \bar{y})]}{\sigma_X \sigma_Y}$$

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Positive, negative, and no correlation

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Positive, negative, and no correlation-2

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Positive, negative, and no correlation-3

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
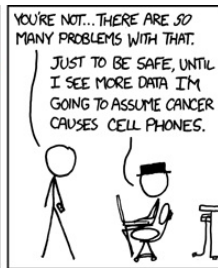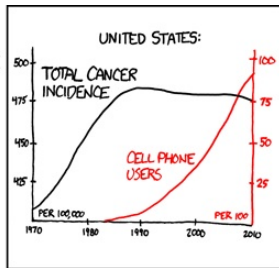Making Predictions

Correlation and causation-2

In a widely studied example, numerous epidemiological studies showed that women who were 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 data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic groups, with 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 socioeconomic status), rather than cause and effect, as had been supposed.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Correlation and causation-3

As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream consumption causes drowning.
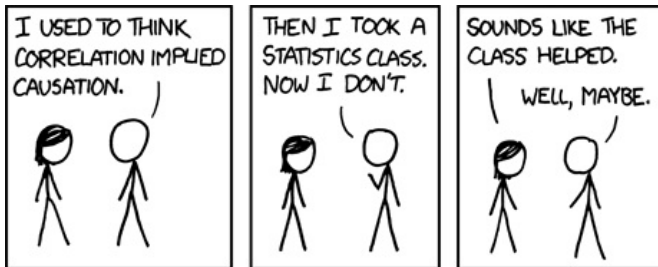
Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Correlation and causation-4

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Correlation and causation-5

Feeding children a diet rich in fish could prevent violent and
anti-social behaviour in their teens, according to research to be
announced this week which suggests the root causes of crime may
be biological rather than social. The study raises major questions
over the extent to which criminals exercise free will, as well as
fuelling fresh debate over whether simple childhood interventions
might be more use in preventing crime than blaming parents or
organising draconian crackdowns on crime.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Correlation and causation-6

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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.

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

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!

Ideas of Probability
Basic Statistics
Spread of items
Distribution
More Probability
Making Predictions

Now that you are an expert