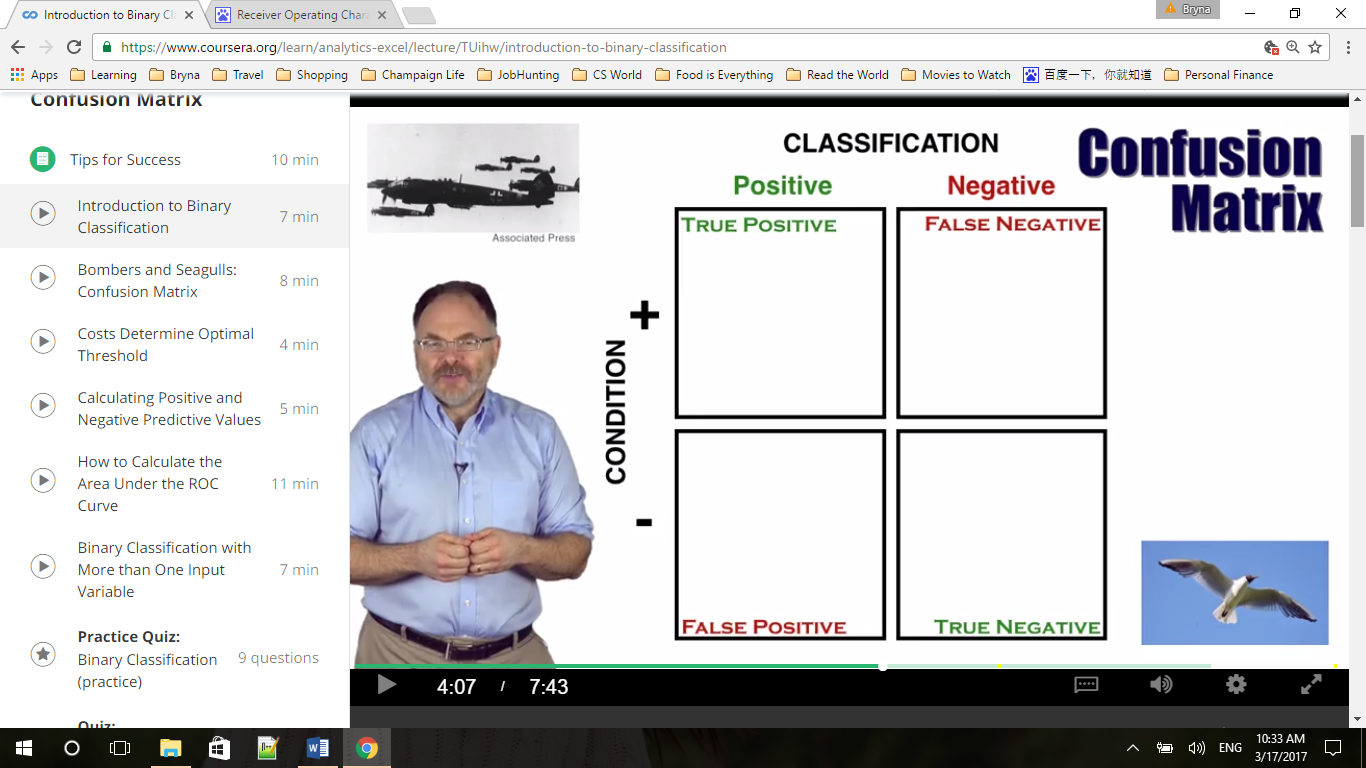
# Week2 Binary Classification

## Introduction to Binary Classification

False positive, false negative =>great costs

ROC Curve( Receiver Operating Characteristic Curve): estimate of the relative cost of the two different kinds of mistakes, maximize the area under the ROC curve=> discriminate signal from noise



Note that a false positive rate is the total number of false positive classifications, divided by total number of seagulls. The true positive rate is the total number of true positive classification, divided by the total number of bombers.

FPR = FP/(FP+TN) 假正类率

TPR = TP/(TP+FN) 真正类率

TNP = TN/(FP+TN) 真负类率 1-FPR

Keeping the scoring method constant but changing the threshold leads to different values for the confusion matrix.

Area under the Curve of ROC (AUC ROC)

## Bombers and Seagulls: Confusion Matrix

Appendix: 2.1

First Use - Exploring the relationship between Classification Errors, the ROC Curve and the Area under the Curve (AUC)

Interact with the Bombers and Seagulls spreadsheet to see how changing the number of classification errors changes the false positive and true positive rates, the shape of the resulting Receiver Operating Characteristic Curve, and the overall Area Under the Curve (AUC).

On the Spreadsheet, ranked scores corresponding to the size of radar images are given in Column C, rows 33 to 52.

The actual condition – Bomber or Seagull – that goes with each score is given in Column D, rows 33 to 52.

This spreadsheet is a calculator that measures the number of incorrect and correct positive classifications at each threshold to generate a false positive rate – Column H, rows 32 to 52 – and a true positive rate – Column I, rows 32 to 52.

At each threshold dividing positive from negative classification, the false positive rate and true positive rate provide an (x,y) ordered pair. These ordered pairs, when shown on a chart and linked together, form the ROC Curve.

These ordered pairs are shown in Columns D and E, rows 6 to 26 – and are displayed on the chart at columns K-N.

The total area under the resulting ROC curve - the very important AUC Metric - is given in cell I28.

Try an experiment yourself – what if the event with radar score “83” had turned out to be seagull and not a bomber? This change would result in one more false positive classification and one fewer true positive classification for every threshold of 83 or below.

How would this impact the performance metrics for the radar? Replace the condition “1” in cell D37 with a “0.” Note that almost all the false positive and true positive rates at different thresholds change. You can observe how the shape of the ROC Curve on the chart changes, and how the area under the curve is reduced from .824 to .806.

Next, try changing the event with radar score 97 from a seagull to a bomber – by replacing the “0” in cell D33 with a “1.” This should improve the area under the curve from .824 to .906.

Second Use - exploring how changing the input “Costs per Classification error” changes the overall cost, and changes the optimum (lowest-cost) threshold

The Bombers and Seagulls Spreadsheet can also be used to observe how, keeping the threshold constant, changing the costs per false negative classification and per false positive classification changes the total cost of a classification system. It can also lead to a different threshold becoming the lowest-cost threshold.

Spreadsheet cell L30 contains the cost per False Negative (FN) – recall that this means failure to signal an alarm when bombers attack.

Cell N30 contains the cost per False Positive (FP) – the cost of responding to a false alarm.

The total costs over all 20 events, at each different possible threshold for binary classification, are given in Column O, rows 32 to 52.

The minimum total cost – the cost if selecting the “optimum” threshold – is given in cell O54.

To identify the optimum threshold using this spreadsheet, find the row in Column O that has the same total cost as the minimum cost given in row 54. Cell P54 shows the lowest average cost per event - the total cost, divided by the total number of events.

At the default settings of 10 million pounds per FN and 4 million pounds per false positive, the minimum total cost - 20 million pounds – is found in cell O40, which uses the Excel “min” function on the list of totals.

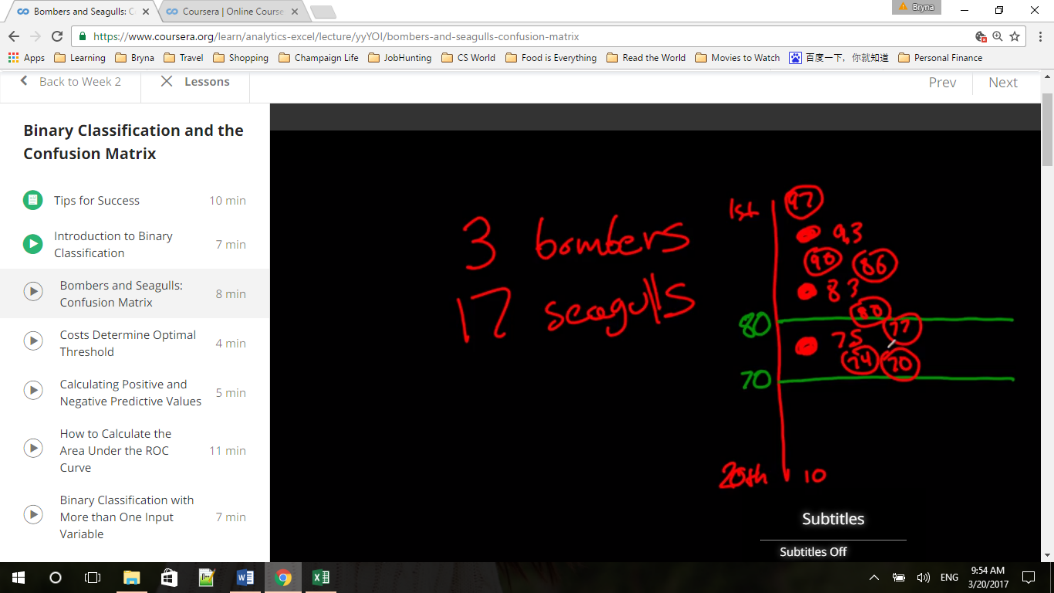
Now look across the spreadsheet to column C in the same row. Cell C40 shows that the optimum threshold at the default costs levels classifies 75 and above as “positive” and 74 and below as “negative.”

The total cost of 20 million pounds is due to 5 False Positive errors at 4 million pounds each, and 0 False Negatives errors at 10 million pounds each. The classification errors can be seen in Column D, rows 33, 35, 36, 38 and 39.

Now, try changing the cost per FN in cell L30 from 10 million pounds to 5 million pounds. Keep the cost per FP the same for now.

The new minimum total cost displayed in cell O54 is now 14 million pounds. Note that the minimum is no longer at 040. The new minimum-cost is found in cell O34. This corresponds to the threshold in cell C34. The new optimum threshold classifies 93 and above and positive and 90 and below as negative.

--------------------------------------------------------------------------------------------------------------------------------

The total cost of 14 million pounds is now due to one False Positive error (cell D33) at 4 million pounds and two False Negative errors (at cell D37 and D40) at 5 million pounds each.

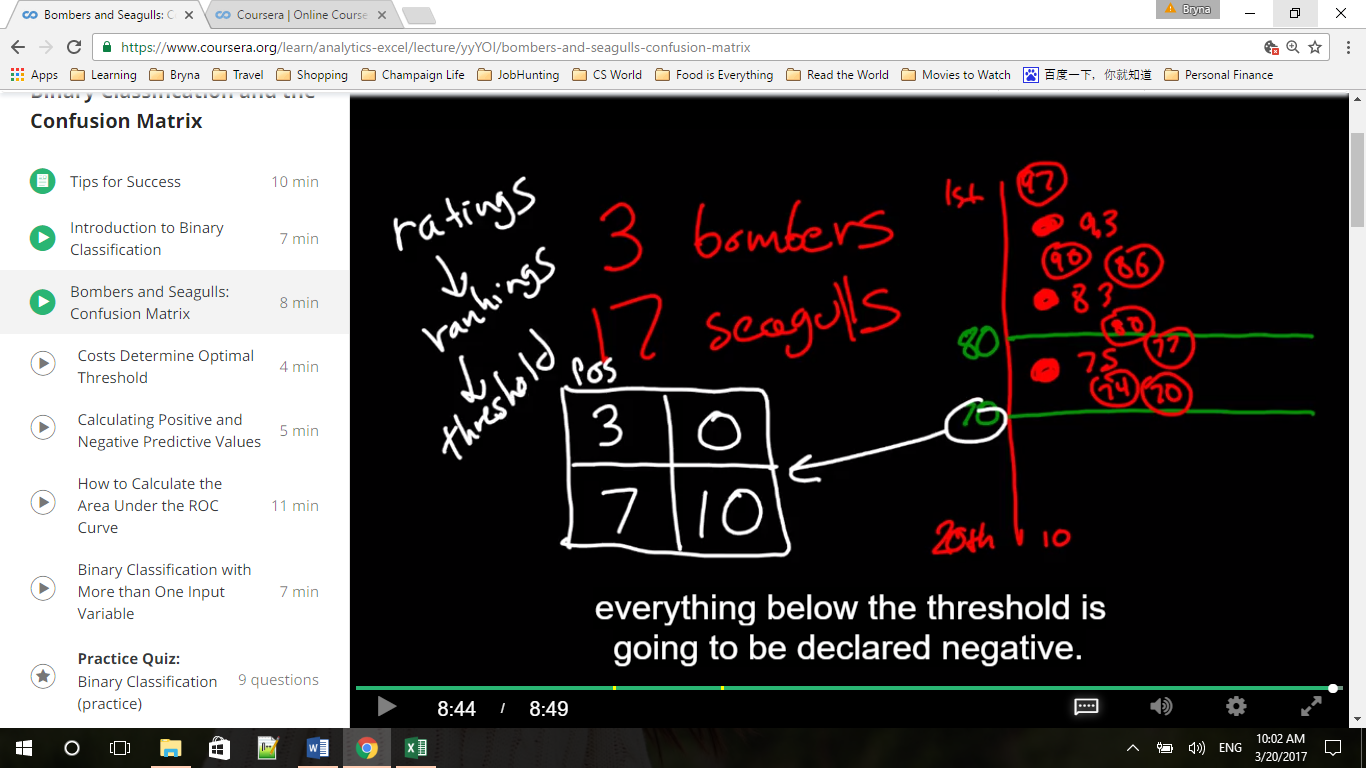
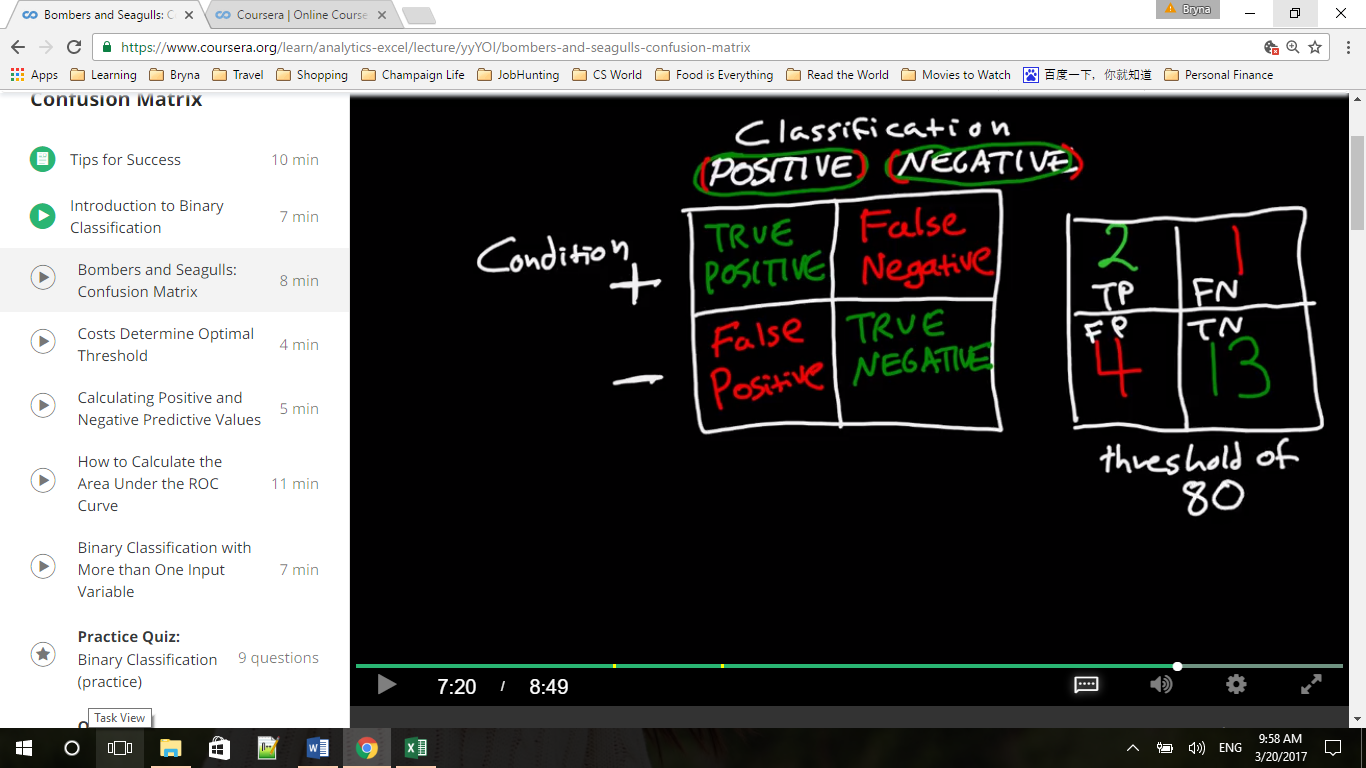
Threshold => determine how many false alarms we have

80=> 4 false alarms, 97, 90, 86, 80; 2 true positive

70=> 7 false alarms, 97, 90, 86, 80, 77, 74, 70; all 3 true positive

Not radar but radar and a threshold => determine how many values go into each of these boxes

## Costs Determine Optimal Threshold



Threshold 80: FPR = FP/(FP+TN)= FP/Neg = 4/17 =

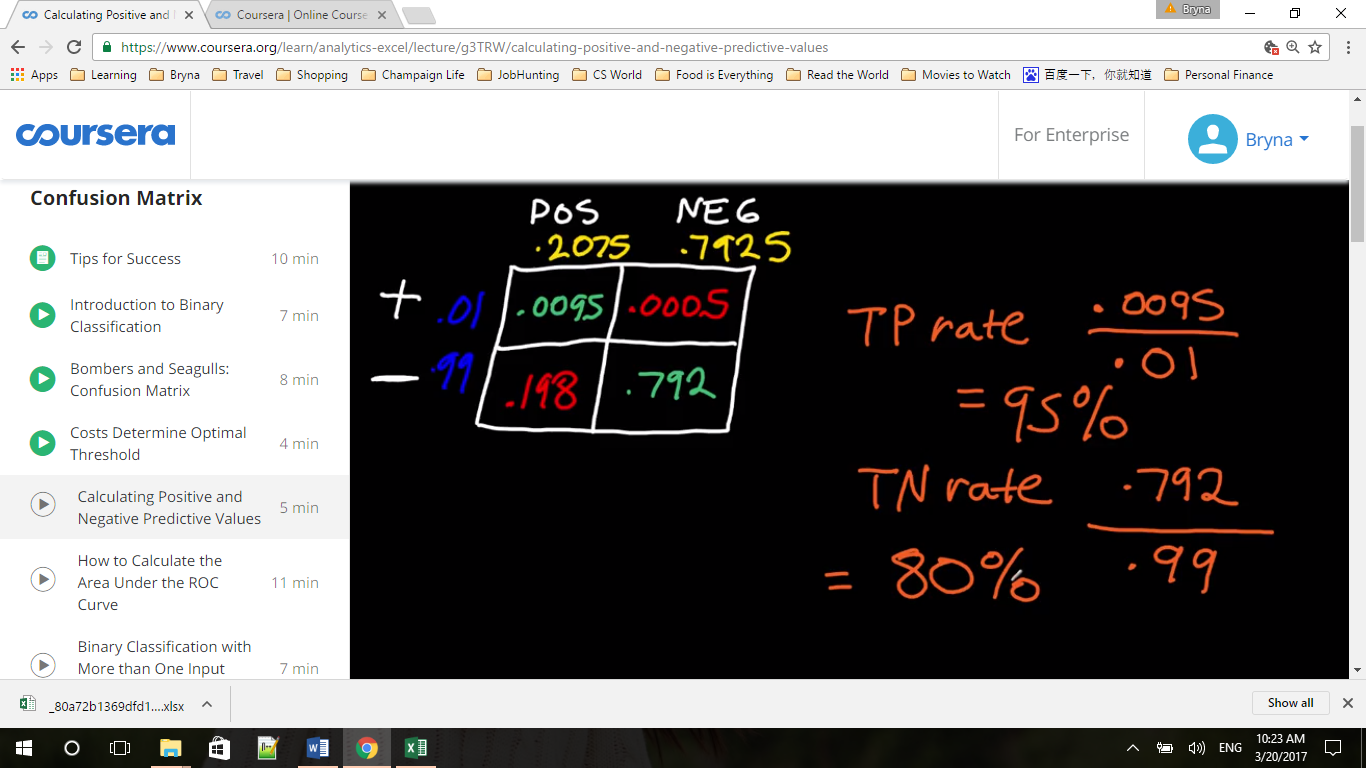
TPR = TP/(TP+FN) = TP/Pos = 2/3 =

Threshold 70: FPR = FP/(FP+TN)= FP/Neg = 7/17 =

TPR = TP/(TP+FN) = TP/Pos = 3/3 = 1

X = FPR, Y = TPR

## Calculating Positive and Negative Predictive Values



A true positive rate of 95% is conditional probability of having a having a positive test if I have cancer. And true negative rate is the p of having a negative test if I do not cancer.

What we want to know here is the latter: the conditional probability that we have the cancer if I have a positive test, or I don’t have cancer if I have a negative test?

P(POS TEST / + ) VS P( + / POS TEST)

P(NEG TEST/ - ) VS P( - / NEG TEST)

前者是说，我有病能测出来的准确率，后者是我测出来有病而我实际有病的准确率，不一样

计算方法：前者TP除以所有实际的True（上面两格）,后者TP除以测出来的Pos(左侧两格)

The latter is called positive predictive value (green yellow )

= 0.095/0.2075 = 4.58%, if I received a positive test, I have a 4.58% chance of having cancer

negative predictive value = 0.792/0.7925 = 99.937%, means that if the test turns out negative, I have 99.937% chance of not having cancer, 0.063% chance of having cancer

--------------------------------------------------------------------------------------------------------------------------------

Appendix 2.3

This spreadsheet works just the same way as the example shown in the Bombers and Seagulls Spreadsheet (and the video) but has 10,000 rows of data instead of 20. It is designed to provide a realistic simulation of the cost-benefit assumptions that must be made to set the classification threshold for a medical diagnostic product.

The ranked scores in Column A are the level of a certain protein as measured by the diagnostic test.

The true Condition for each protein level score is given in Column C [Cancer =1, No Cancer = 0].

A threshold for positive classification can be set between any two protein levels. For each threshold given in Column A, the resulting number of False Negative classification errors is given in the same row of Column H, and the False Positive classification errors in the same row of Column F.

This spreadsheet is designed to allow you to observe how changing cost inputs impacts both (a) the overall costs of using a cancer diagnostic test at each threshold, and (2) what threshold should be chosen as optimal (minimizing cost).

The cells to input the assumed costs of classification errors are Cell G3 for cost per False Negative (missing a case of cancer) and H3 for cost per False Positive (a false alarm). Total costs at each threshold are given in Column K.

The minimum total cost, and minimum cost per event (per diagnostic classification reported), are displayed in cells K4 and L4, and the optimum threshold – the lowest protein level score that should be classified positive - is displayed in cell M4.

At the default costs of $50,000 per False Negative error and $500 per False Positive error, the minimum cost per event is $119.90, and the optimal threshold for positive classification is 16551.930.

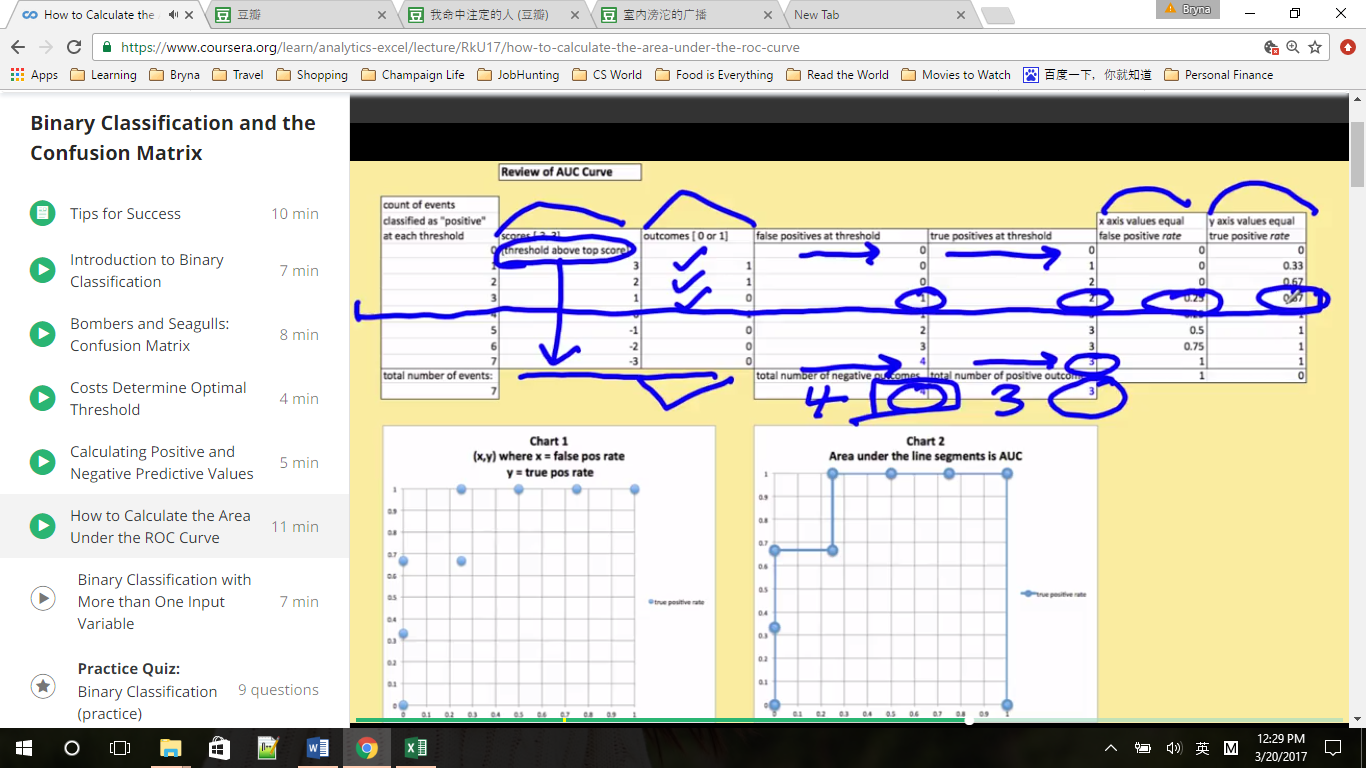
Try changing the inputs yourself. If you keep the cost per False Negative the same, but raise the cost per False Positive, would you in general expect the new optimum threshold to be higher (fewer total positive classifications) or lower (more total positive classifications)?

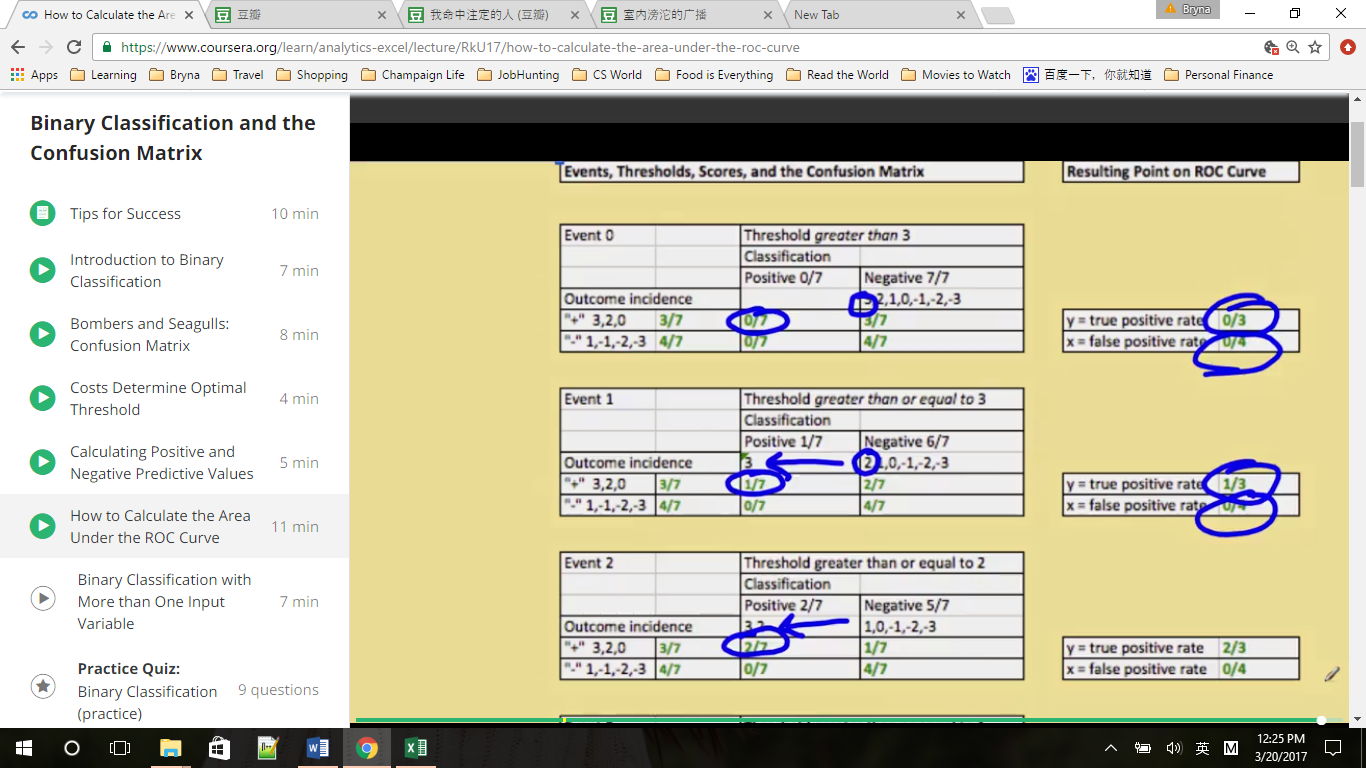
(In general, increasing the cost per FN while keeping the cost per FP constant will cause the cost-minimizing threshold score to: Increasing the cost per False Negative while keeping the cost per False Positive the same will move the threshold towards fewer False Negatives (and toward more False Positives). Then the question becomes "to reduce the number of False Negatives and increase the number of False positives, does the threshold move up (increase) or down (decrease)? decrease)

Change the cost per False Positive to $1500 and you will see that the optimum threshold is higher. The threshold for the first Positive classification moves from 16551.930 [item ranked 2094] to 16824.137 [item ranked 1822].

Similarly, if you reset the cost per FP at $500 and raise the cost per FN to from $50,000 to $500,000, the optimal threshold is lower – it falls to 13307.537 [item ranked 5338]

## How to Calculate the Area Under the ROC Curve





--------------------------------------------------------------------------------------------------------------------------------

Appendix 2.4

Input any value between 0 and 1 for the probability of True Positive classifications – the joint probability labeled letter e in the confusion matrix – into cell G9.

Next, enter any valid values (that is, any values less than or equal to the value entered into cell G9) for the condition incidence – letter a - (cell E9) and the “classification incidence” (also called “test incidence”) – letter c - (cell G7).

The Excel Spreadsheet will output:

-Values for the remaining five cells of the Confusion Matrix, designated by letters b, d, f, g, and h.

-Values for the eight primary performance metrics used to evaluate binary classification systems, Column F, rows 36-39 and 41-44.

-One point on the ROC Curve corresponding to this particular Confusion Matrix. Recall that the x-axis coordinate equals the False Positive Rate – output in Cell N37 – and the y axis coordinate equals the true positive rate – output in Cell O37.

Assume the Confusion Matrix shows the performance of a model that tries to predict which visitors to an Automobile dealership will buy a car from that dealer.

Experiment with how a change in one or more inputs impacts the outputs. For example, change the probability that an event will be a True Positive [Cell G9] to .18, without changing the condition incidence or test incidence.

Question: What is the conditional probability that a visitor classified as Positive by the predictive model (the “test”) will buy a car (the condition)?

Answer: 90%. The answer is the “Positive Predictive Value” – also written as the conditional probability p ("+" | Test POS) – as is found in cell F41, is 0.9.

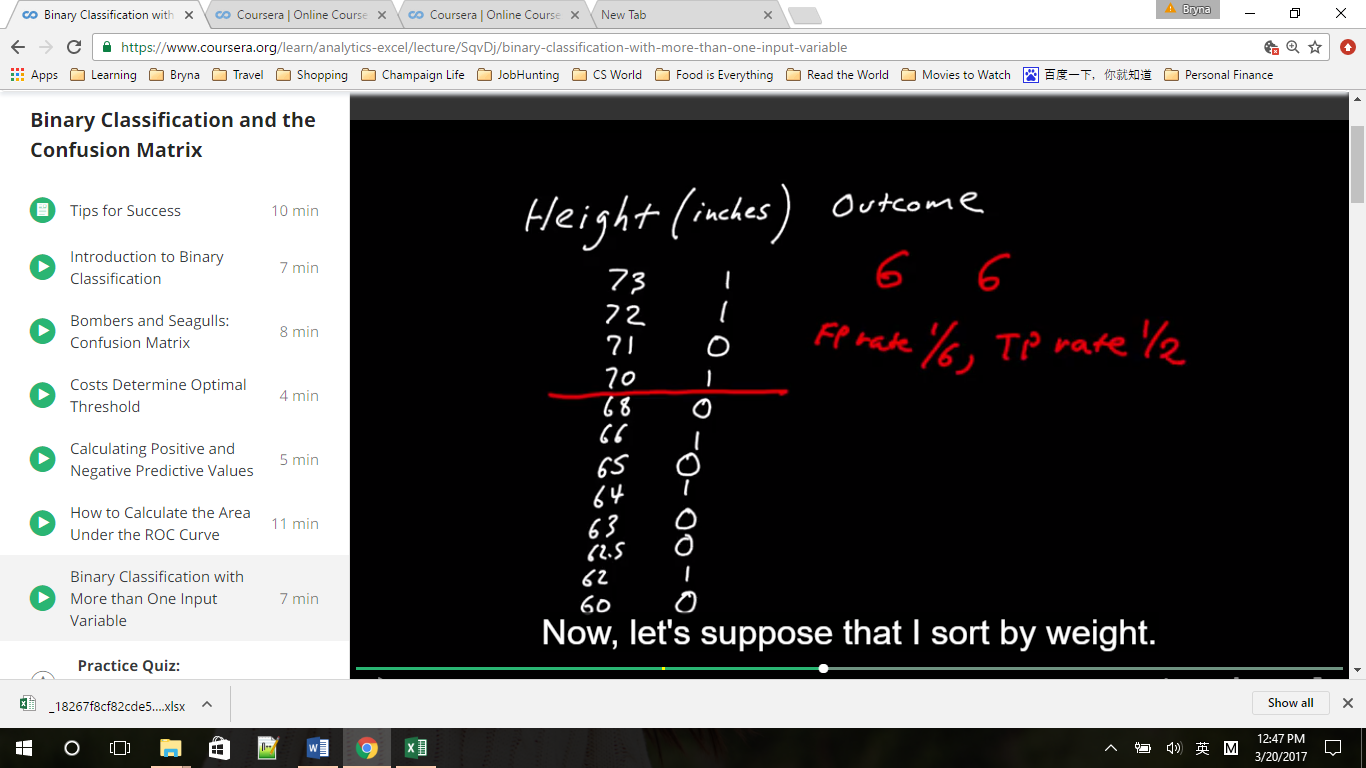
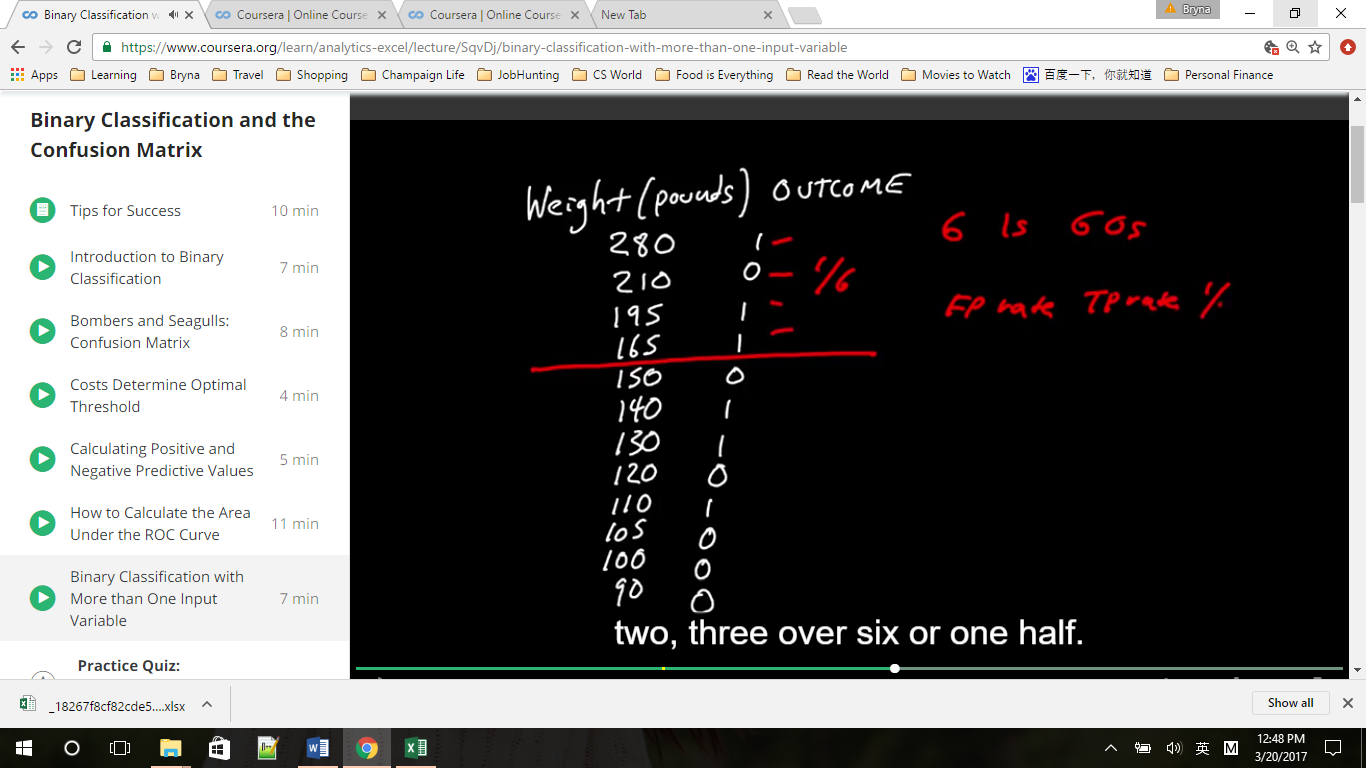
Question: What point on the ROC Curve summarizes this confusion matrix performance? Answer:

Answer: (.03, .6). The value 3% is the False Positive rate – also written as the conditional probability p(Test POS | "-") – found in Cells N37 and F38. The value 60% is the true positive rate – also known as the conditional probability p(Test POS | "+") - Given in cells O37 and F36.

Question: What is the probability that a visitor classified as “Negative” by the predictive model will not buy a car from that dealer?

Answer: 85%. This value is the Negative Predictive Value, also written as the conditional probability p("-" | Test NEG) – found in Cell F44.

## Binary Classification with More than One Input Variable

When you combine variables that have very different scales, first standardized them to figure out how much relative weight you should assign to each one. If we just add height and weight together, the weight swamp the results, nearly the same outcome based on weight only.

Why standardizing the data? To treat each of the two input variables as equally important

--------------------------------------------------------------------------------------------------------------------------------

Appendix 2.6

This Spreadsheet can be used to practice sorting data so that it is ordered by the “score” used for predictive purposes. Copy cells B18 to G30 and “paste special: values and number formats.” Past into the upper left hand corner of a new spreadsheet. Then select the right-most Column – containing the sum of standardized heights and weights – and choose Data/Sort/Descending – when asked, choose “Expand the selection.” You will now have created a ranked set of scores and outcomes, just like what is already provided in the Bombers and Seagulls and Cancer Diagnostics Spreadsheets. Repeat the process but sort on height alone, on weight alone, or on a combination of height and weight not previously standardized (Column G, rows 2 – 13) to compare the performance of the various scores when used for binary classification.

Assignment Questions:

Can a change in classification threshold change a diagnostic test's True Positive Rate? Use logic - no need to calculate any numbers.

* Yes. The number of Condition Positives is constant. But by moving one or more positive outcomes above or below the threshold the number of True Positive classifications changes. This changes the True Positive Rate, which is the ratio of True Positive classifications to total Condition Positives.

“Condition Incidence" is the portion of a population that actually has the Condition being studied. Can a change in threshold change the Condition incidence? Use logic - no need to calculate any numbers.

* No. Moving the threshold changes only classifications, not actual Conditions

Does the change in threshold change the test’s “classification incidence” (also called “test incidence”)? Use logic - no need to calculate any numbers.

* Yes. Moving the threshold down increases the proportion of outcomes classified as positive. Moving the threshold up reduces the proportion of outcomes classified as positive.

Does the change in threshold change the test's Area under the ROC Curve? Use logic - no need to calculate any numbers.

* No. Points on the ROC curve represent the false positive rate and true positive rate at each possible threshold. Changing the threshold signifies a different point on the ROC Curve, but does not change the overall shape of the curve.

# Week3 Information Measures

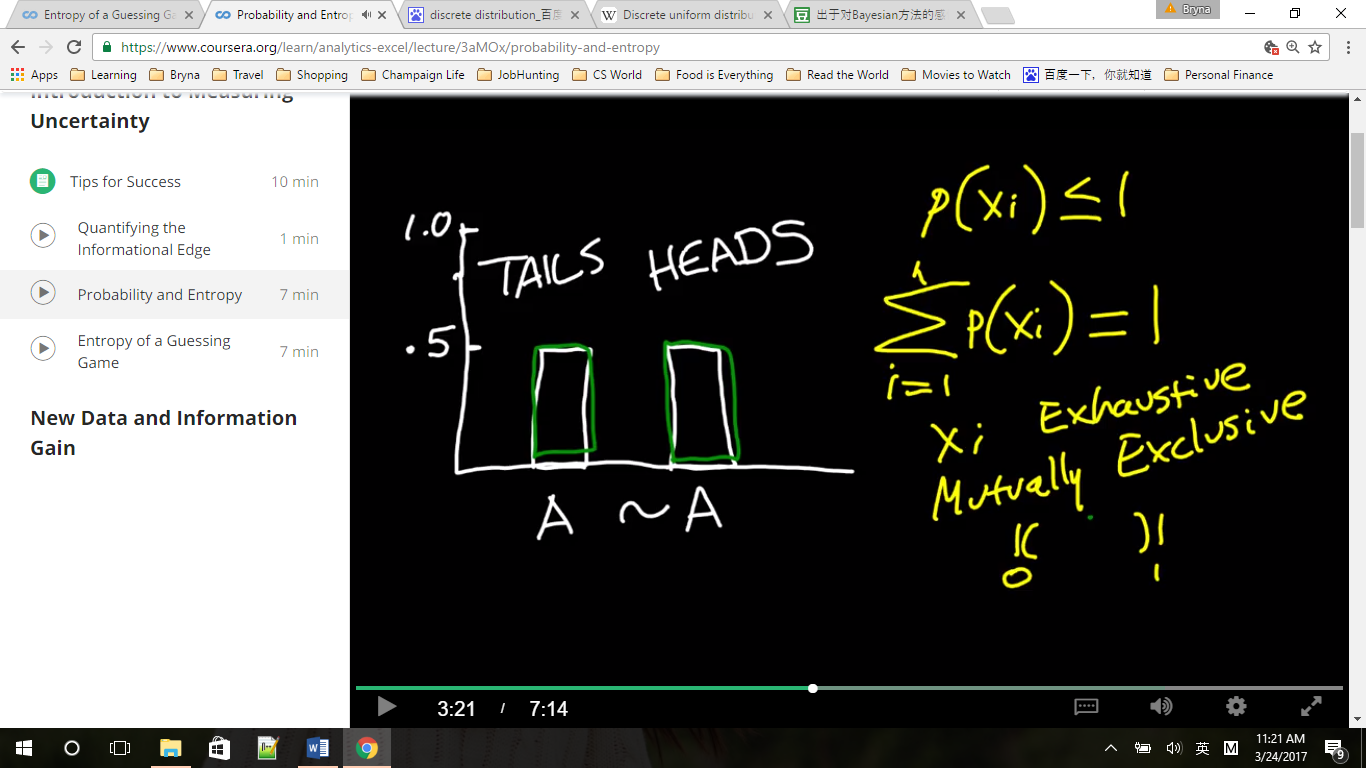
## Introduction to Measuring Uncertainty

### Probability and Entropy

Bayesian logical data analysis, P(X) is a measure of the degree of belief about the truth of the statement

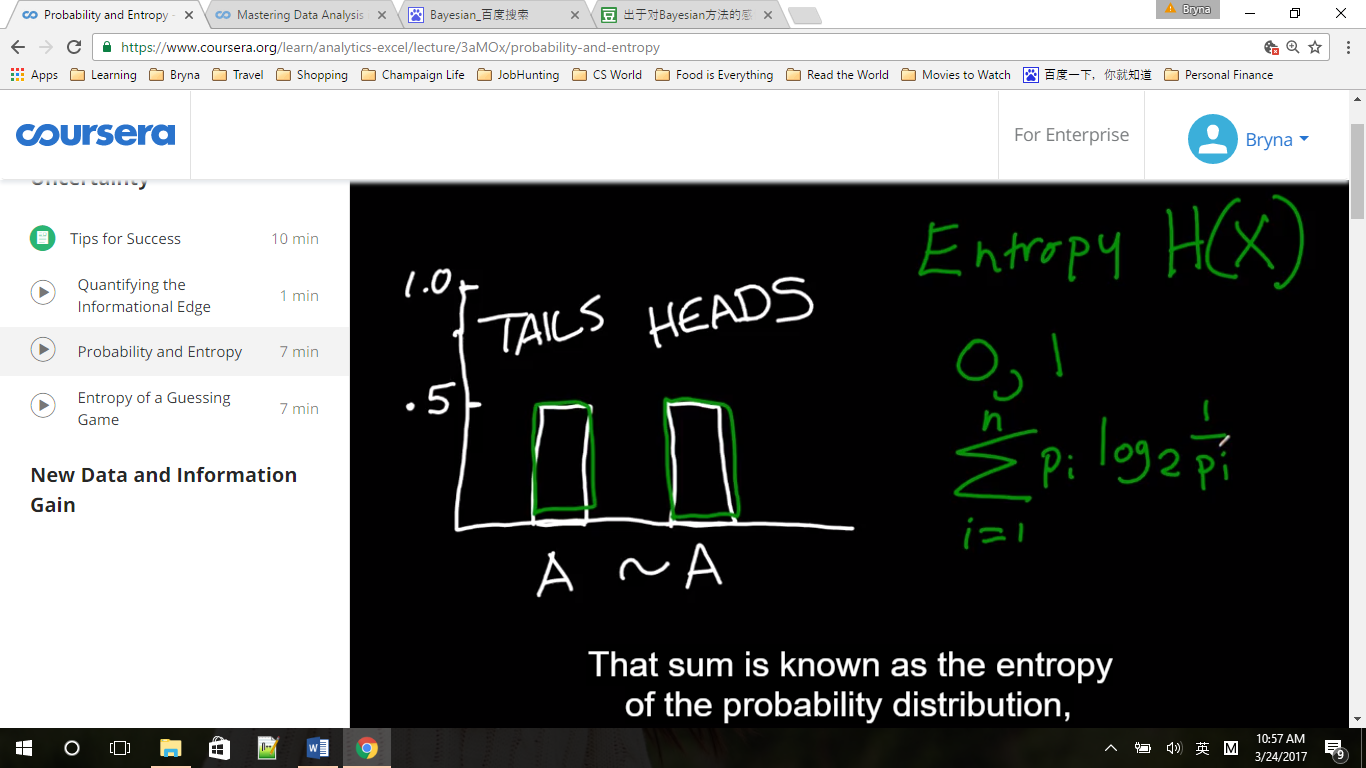
P(A) = 1, the statement is true with certainty ; P(A) = 0, the statement is false with certainty

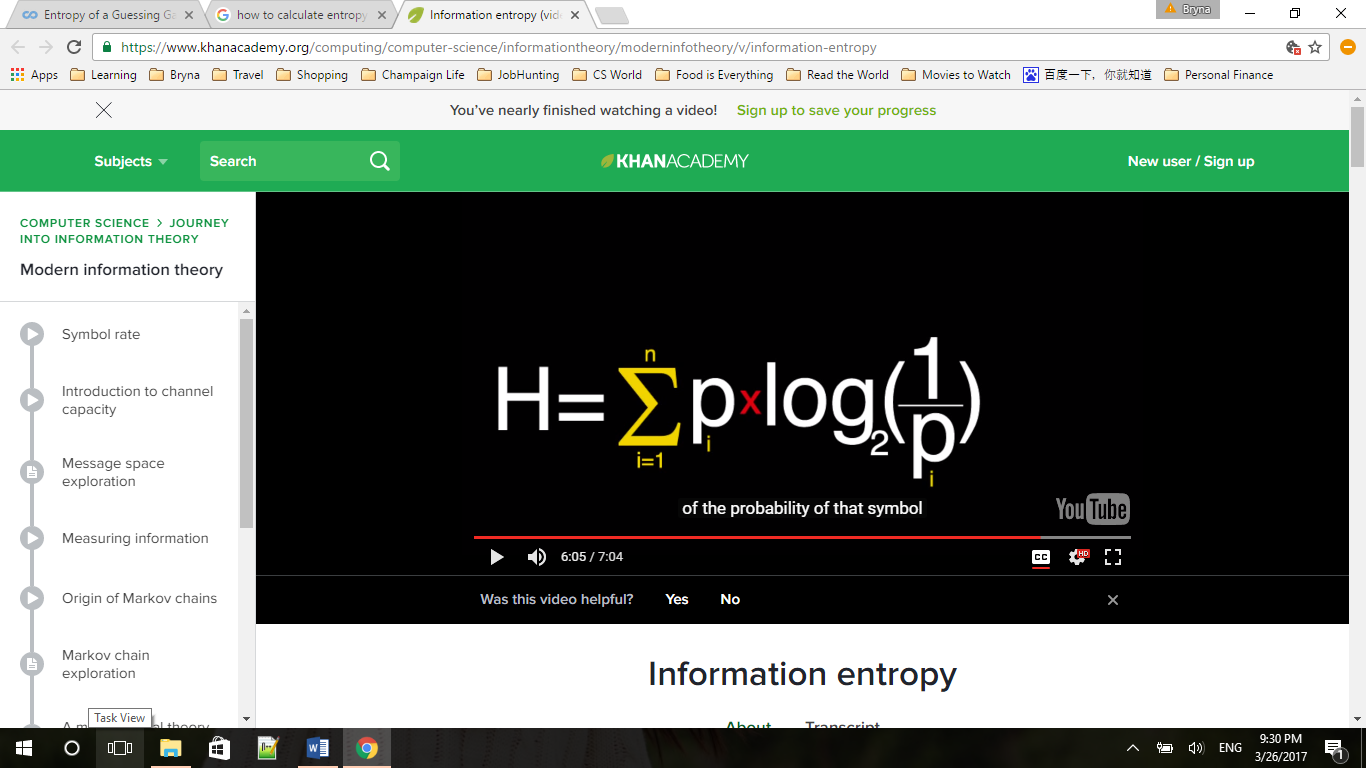
Probability distributions: collections of individual probabilities that sum to one



Exclusive, tails or heads; exhaustive, no third probability

H(X) is a measure of the uncertainty of the whole probability distribution measured in bits of information



 from Khan Academy

Entropy is the sum of each probability times the log to the base 2 of one over that probability

If we have a uniform distribution (均匀分布) with n possible outcomes, H(X) = log2n, also the maximum of entropy, other distribution will have a smaller H than a discrete distribution

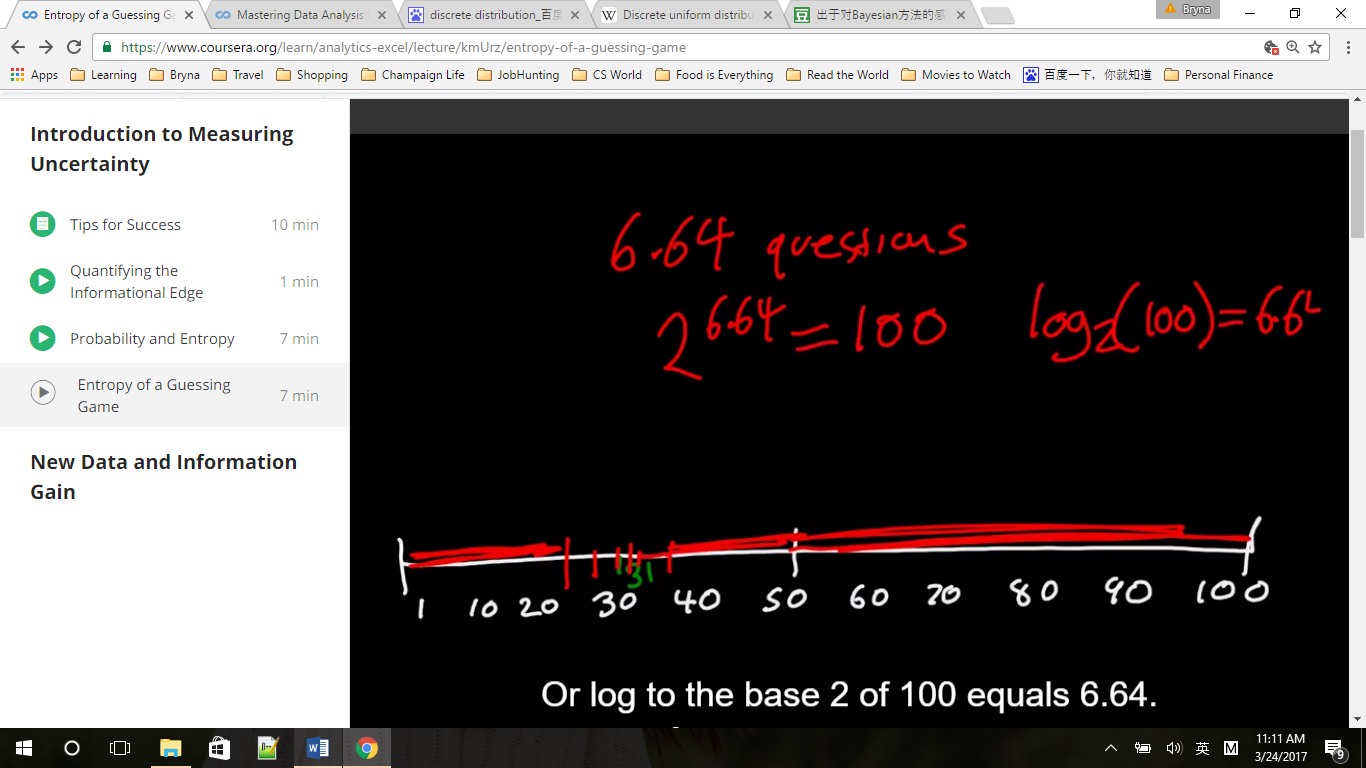
(for example, tail or head, p =1/2, H(X) = 1/2\*log2(1/(1/2)) +1/2\*log2(1/(1/2)) = log22 = 1

Or p=1/6, H(X) = 6\*( 1/6\*log2(1/(1/6))) = log26

After we’ve thrown the coin, P(Tails) = 1, H(X) = 0

The more uncertainty we have, the bigger it gets, minimum at zero

### Entropy of a Guessing Game



Now consider a guessing game where my friend picks a number between 1 and 100. And I need to guess what it is, asking only simple yes or no questions of the form, is your number bigger than x?

It will take me six point six four questions, on average, to get the correct answer.

Log2100 = 6.64, log210000 = 13.29 which is the entropy of the 10000 possibility game, also means at the beginning of the game my uncertainty can be quantified as 13.29 bits of uncertainty

The entropy of the joint probability H(e,f,g,h) = H(X,Y)

Conditional probability => conditional entropy

e/a, true positive rate, the conditional probability of having a positive test, if you have the condition

* Conditional entropy: the entropy of X conditioned on Y is written as H(X|Y).

在信息论中，条件熵描述了在已知第二个随机变量 Y 的值的前提下，随机变量X 的信息熵还有多少

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Classification/Test Y | |
|  |  |  | Pos | Neg |
|  |  |  | c | d |
| Outcome X | + | a | e | f |
|  | - | b | g | h |

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/bxI3iYK-EeWQ7RJdmuvcCw_b1665aa48e93aca948d644597029c645_JointEntropyFormula.bmp?expiry=1490745600000&hmac=BfJ9tahWIvYDGKs6cCm7Flzh5eboHoIFZ8Pd5MjIk4Y

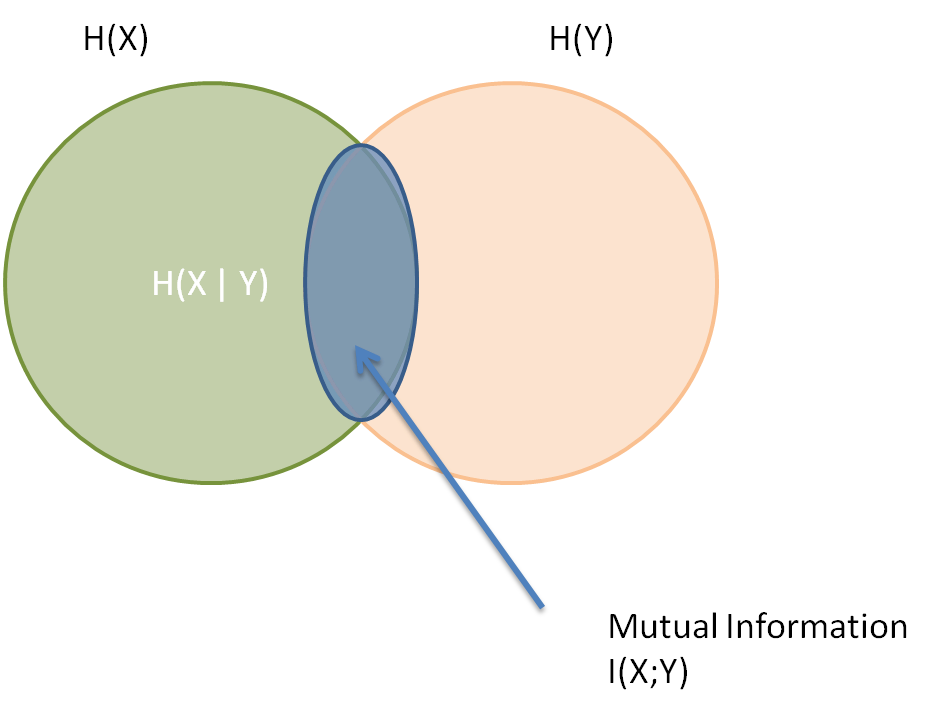
H(X|Y) = P(Y=Y1)H(X|Y=Y1) + P(Y=Y2)H(X|Y=Y2)

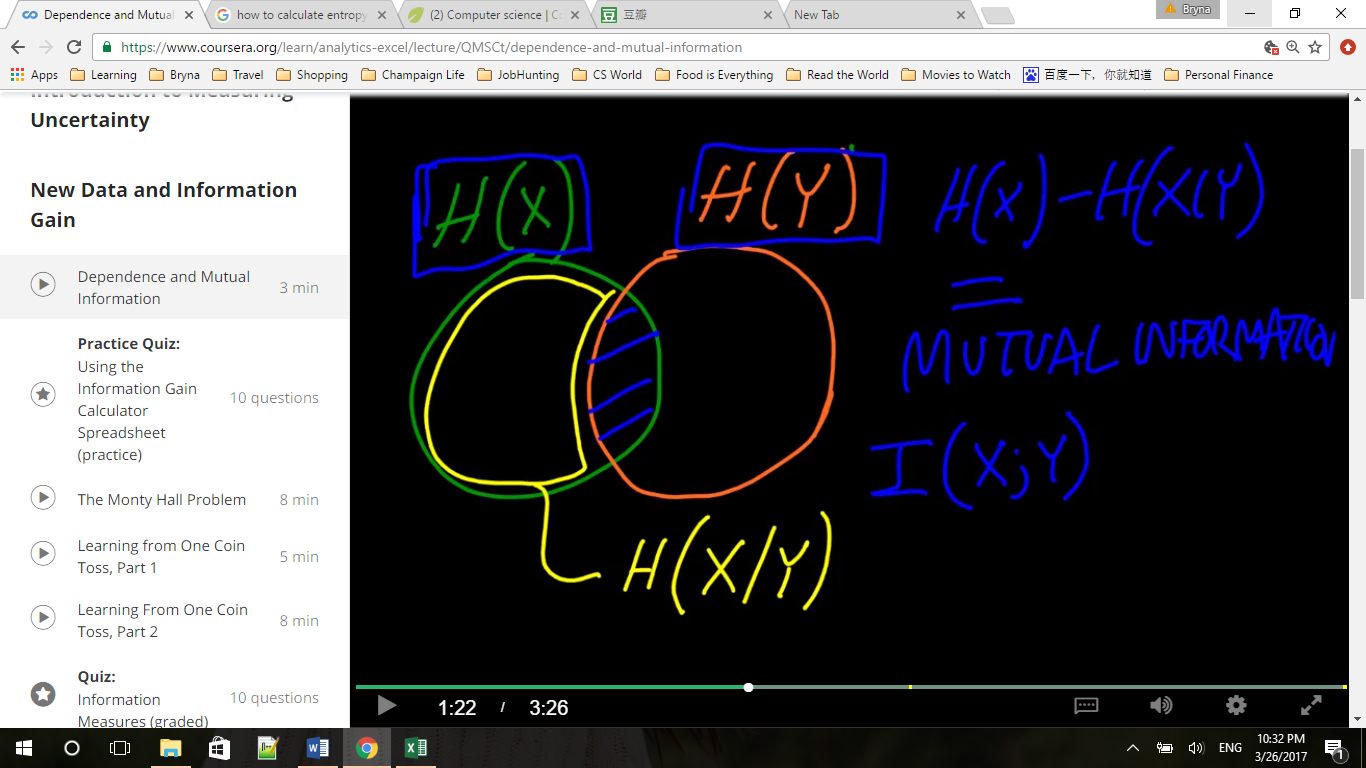
= c\*H(e/c,g/c)+d\*H(f/d,h/d)

## New Data and Information Gain

### Dependence and Mutual Information

We measure the mutual information by measuring the uncertainty we start with regarding x, that's H(x), and then subtracting out the uncertainty that remains for x after we learn y



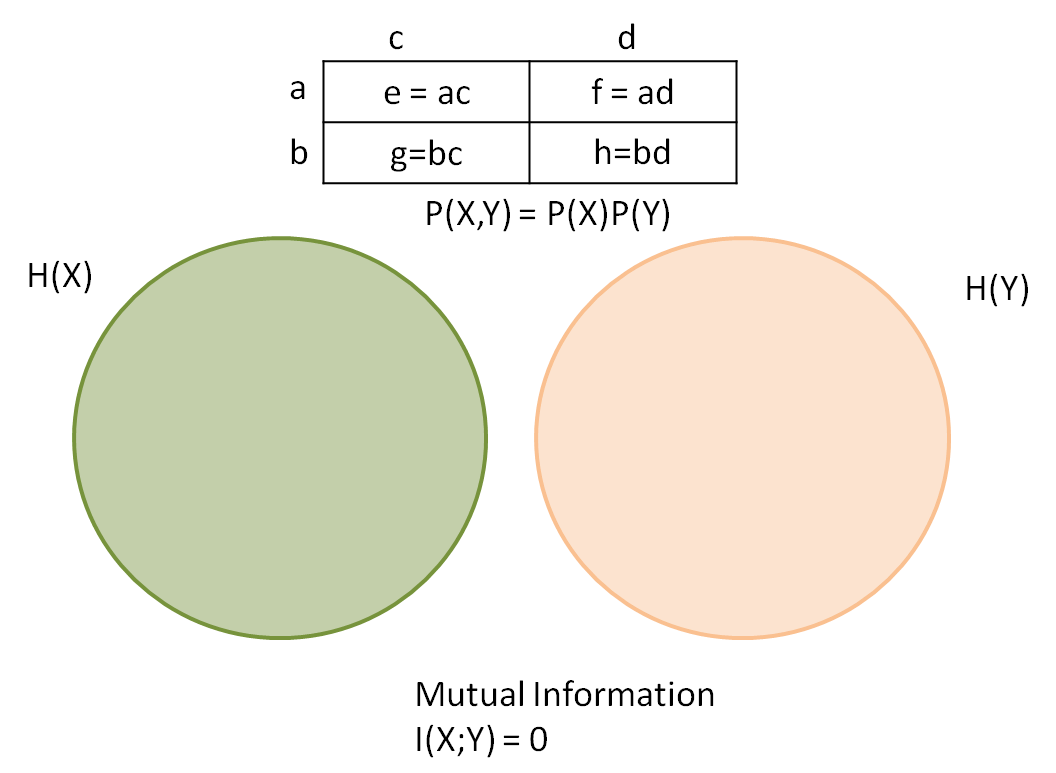


* Independence and dependence of probability distribution

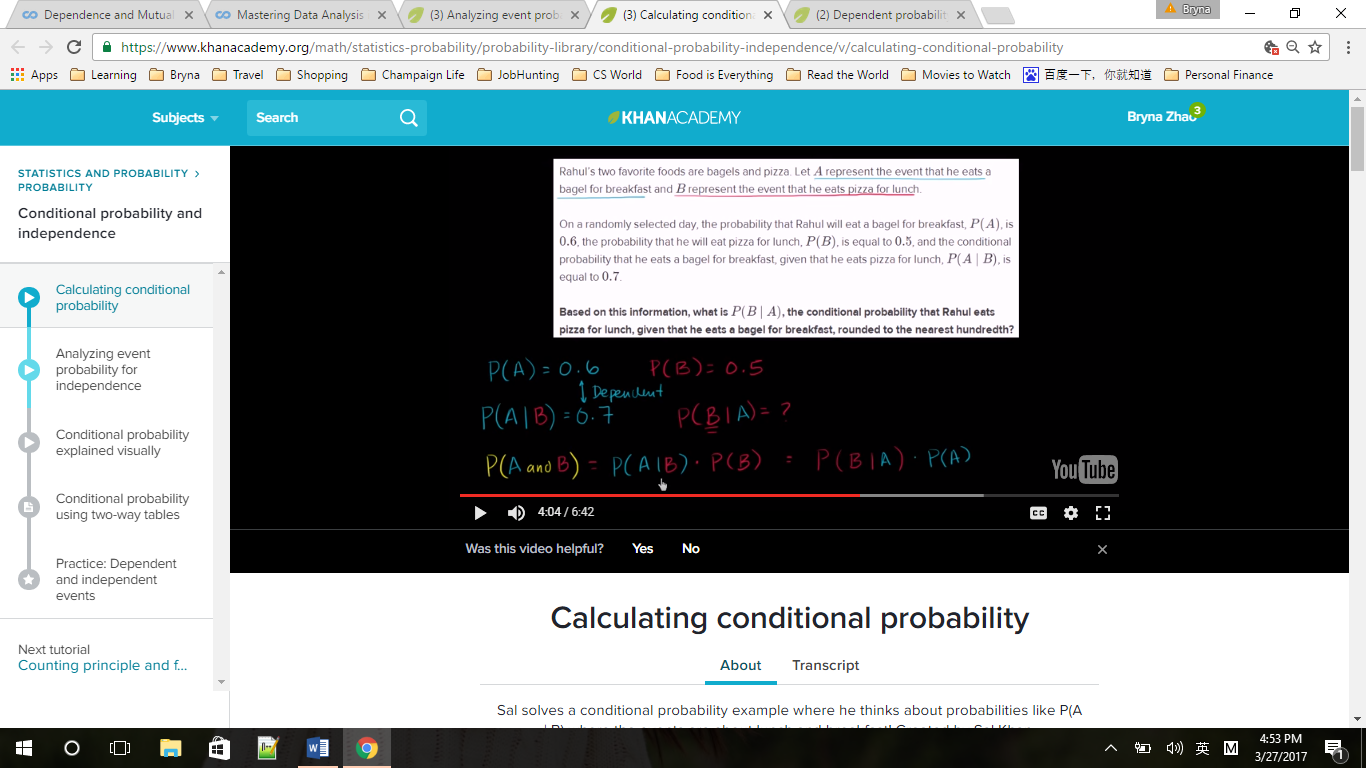
Independent: joint distribution = product distribution, P(X,Y)= P(X)P(Y)

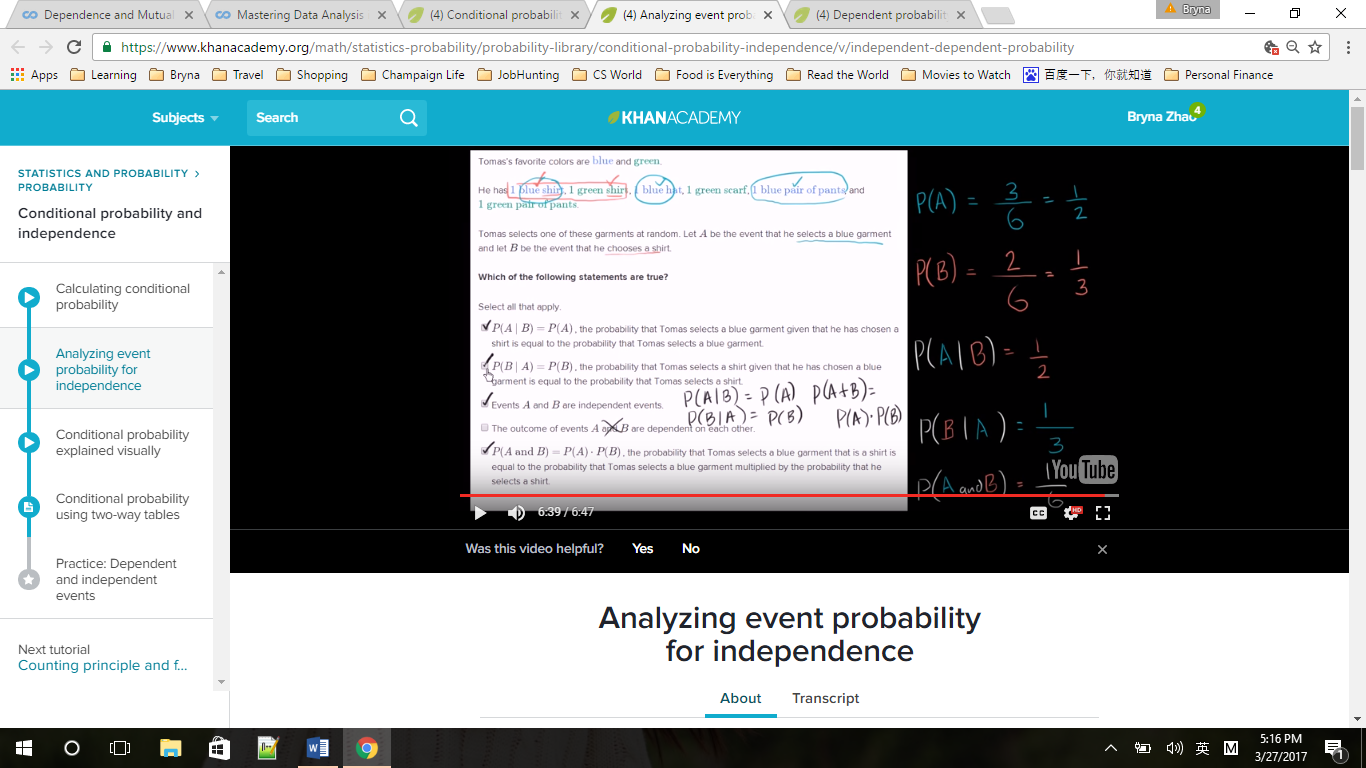


Correction: h = bd



If dependent, P(X,Y)!= P(X)P(Y), I(X;Y)>0;





### Information Gain Calculator

Imagine a massive factory assembly line that manufactures computer microchips. The factory produces tens of millions of chips per year. Defective chips have in the past been discarded after electronic testing, which is very accurate, but slow and expensive. Based on the “gold standard” of electronic testing, 20%(a) of chips are found to be defective and are discarded.

If it were possible to identify and discard some or all of the defective chips without the need for electronic testing, chip production costs would be much lower. Factory Management installs an optical scanner on the assembly line that photographs the chips while they are moving down the assembly line. The scanner software identifies shapes that vary by more than a specified amount from a standard template. The scanner can forecast which chips are defective without slowing down the assembly line.

However, optical scanning is not perfect – there will always be some false positive and false negative classifications. Consider the default values given in the spreadsheet. When the optical scanner is installed, and is set to classify the top 30% (c)of measured shape variations as positive for defects [cell G7] the test’s false positive rate is 25% [cell F38] and its false negative rate is 50% [cell F37].

Uncertainty about whether a chip is defective or not, before it is scanned, can be measured: H(.2,.8), H(X) = .7219 bits [cell O9].

Uncertainty about whether a chips is defective or not, after it is scanned and classified, is H(X|Y) = .3\*H(1/3, 2/3) + .7\*H(.14, .86) = .6896 bits [cell L45].

The reduction in uncertainty between the two is the “Mutual Information” between X and Y, written I(X;Y). This is H(X)-H(X|Y) = .7219 - .6896 = .0323 bits [Cell P17].

Taking the ratio of mutual information obtained to entropy before scanning, .0323/.7219, gives a 4.47% Reduction in Uncertainty; also called the Percentage Information Gain [P.I.G.].

Changing Spreadsheet Inputs

This Spreadsheet allows you to create any valid Confusion Matrix. Input three values:

Condition Incidence (labeled a) [cell E9],

Test Incidence (labeled c) [cell G7], and

Joint Probability that a classification will be a True Positive (labelled e) [cell G9].

\*The Spreadsheet automatically calculates all other Confusion Matrix values.

Try for yourself what happens to the performance metrics when the defect-forecasting software in the optical scanner is improved so that only 15% of chips are classified positive [cell G7], while the probability of a true positive classification remains constant at 10%.

The false positive rate falls from 25% to 6% [cell F38] while the false negative rate is unchanged.

The new uncertainty about whether a chips is defective or not, after it is scanned and classified, is less - .5819 bits [cell L45].

The new “Mutual Information” between X and Y, is more - .7219 - .5819 = .14 bits [Cell P17].

The new P.I.G is also higher - 19.39% [cell O11].

Determining Dependence or Independence of two Probability Distributions

The spreadsheet demonstrates multiple ways to calculate the Mutual Information between two binary probability distributions, including comparing their Joint Distribution [cells L28 to S 28] and their Product Distribution [cells L 29 to S 29].

When the joint distribution and the product distribution of two probability distributions X and Y are identical, then knowledge of X provides no reduction in uncertainty about Y, and vice versa. When the Mutual information between the two distributions is zero. This state is called “independence.”

Try it for yourself: enter the product of a and c, (p(X = +)\*p(Y = POS)) = (.2)(.3) = .06 into the cell labeled e [G9].

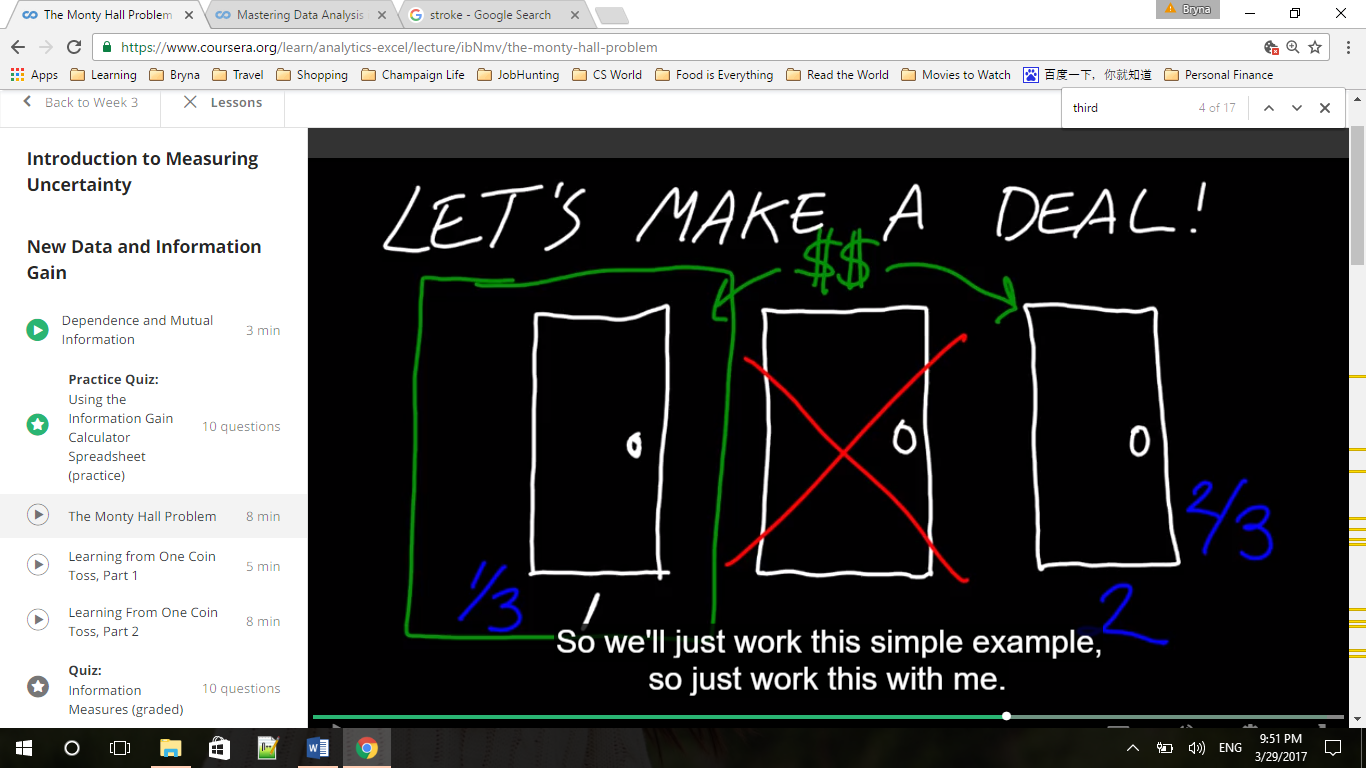
Note first that the three other joint distributions in the Confusion Matrix [cells L28 to S28] all become equal to their product distributions [cells L29 to S29]. When every value in a joint distribution p(X,Y) equals the product distribution p(X)p(Y), the Mutual Information and Percentage Information Gain equal zero.

In addition, because there is no intersection I(X;Y) between the regions H(X) and H(Y) in the Venn Diagram [Columns W-AH] , the sum H(X) [cell L17] plus H(Y) [cell L21] equals H(X,Y), 1.6032 bits [cell L26].

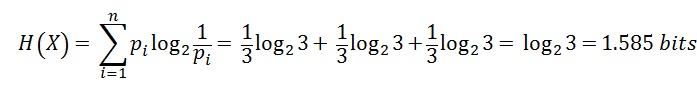
### The Monty Hall Problem

The probability that the prize is behind door one remains at one-third just like it was when that door was first chosen

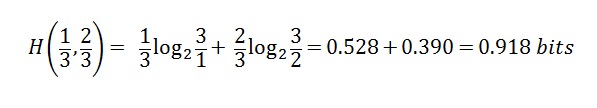
Because Monty Hall is very constrained in what he can do. He cannot open the door that has been chosen by the contestant and you cannot open the door that's got the valuable prize behind it, two-thirds of the time Monty is signaling where the prize is, we just don't know which two-thirds



P(X)= (1/3, 1/3, 1/3)



H of X given Y, H(X|Y)=

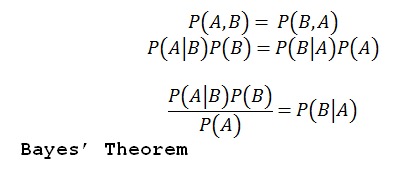


https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/ntA_MYUAEeWIQwrDOWOJxQ_cd975502ad63b4924f6245b0ccaed3ae_MontyEntropy3.bmp?expiry=1491004800000&hmac=SIyVOPaN59oZJ7Lsl0RDkzPoSRSw3DnNmzmZiISs6UE

Monty has provided us with 0.667 bits of information by opening one door.

P.I.G.= 0.667/1.585=0.421

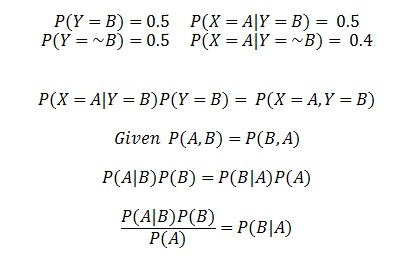
### Learning from one coin toss



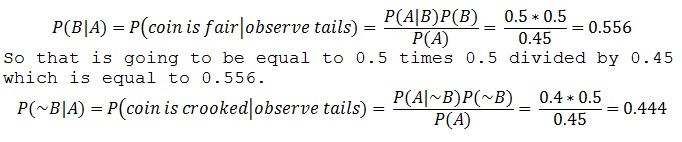
Probability distributions are carriers of information

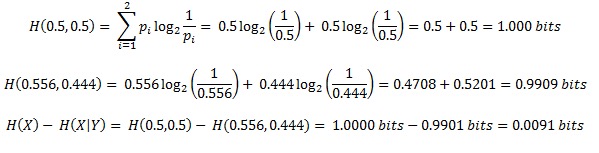
Inverse probability problem:

Probability of B given A is the value that we want to know. The probability that the coin in fair given that the coin came up tails.



https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/z83lFIUWEeWDLw5sAOXdIQ_d1c7f26f1d542d4c4ef1f889cd9b9503_Bayes3.bmp?expiry=1491004800000&hmac=vfExEfTg9swl58B3FUZnZo-8uRfdxeiFD4JilNnYEjI





Quiz:

B stands for “the coin is fair”, ~B stands for “the coin is crooked”. The p(heads | B) = 0.5, andp(heads | ~B) = 0.4. Your friend tells you that he often tests people to see if they can guess whether he is using the fair coin or the crooked coin, but that he is careful to use the crooked coin 70% of the time. He tosses the coin once and it comes up heads.

What is your new best estimate of the probability that the coin he just tossed is fair?

P(B)=0.3,P(~B)=0.7

P(A|B)=0.5,P(A|~B)=0.4

P(A)=P(A|B)\*P(B)+P(A|~B)\*P(~B)=0.5\*0.3+0.4\*0.7=0.43

P(B|A)=(P(A|B)\*P(B))/P(A)=0.349

Suppose you are given either a fair dice or an unfair dice (6-sided). You have no basis for considering either dice more likely before you roll it and observe an outcome. For the fair dice, the chance of observing “3” is 1/6. For the unfair dice, the chance of observing “3” is 1/3. After rolling the unknown dice, you observe the outcome to be 3.

P(A|B)=1/6, P(A|~B)=1/3,P(B)=P(~B)=1/2

P(A)= P(A|B)\*P(B)+P(A|~B)\*P(~B)=1/6\*1/2+1/3\*1/2=1/4

P(B|A)=1/3

What is the new probability that the die you rolled is fair?

Quiz: consider the Monty Hall problem, but with 5 doors to choose from instead of 3. You pick door #1, and Monty opens 2 of the other 4 doors. How many bits of information are communicated to you by Monty when you observe which two doors he opens?

Initial distribution is .2, .2, .2, .2, .2.

Distribution after learning from Monty is .2, .4, .4.

H(1/5,1/5,1/5,1/5,1/5)= log2(5)=2.32

H(1/5,2/5,2/5)=1/5\*log2(5)+ 4/5\*log2(5/2)=1.52

# Week4 Linear Regression

## Introduction to Parametric Models

1.1 Introducing the Gaussian

A good model includes both a signal and noise

Gaussian Probability Density Function: the most common model for noise or uncertainty used in data analysis

## Unpacking Linear Regression