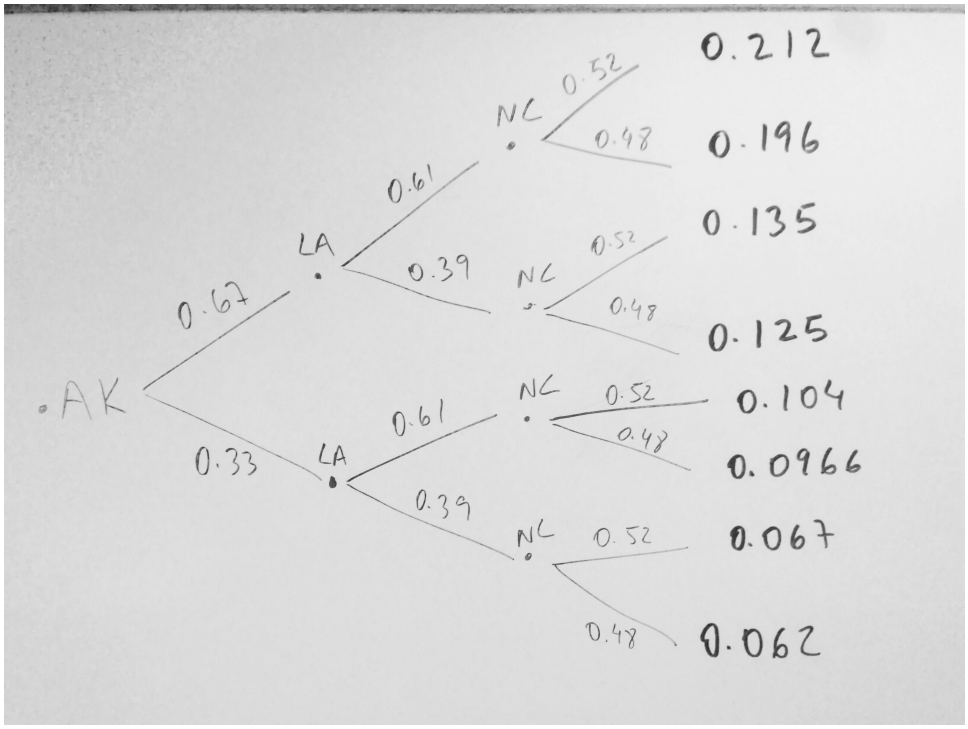
Homework Assignment # 1

Tengyuan Liang
Business Statistics
Booth School of Business

Problem 1: Senate Probabilities (Election 2014)

According to some analysis the control of the senate in the upcoming elections will be determined by the races in 3 states: Arkansas, Louisiana and North Carolina where 3 democratic incumbents face very competitive opponents. Based on predictions by experts at the NY Times, the Republicans have the following probabilities of wining each of these races: Arkansas 67%, Louisiana 61% and North Carolina 52%.

1. To win control of the senate, Republicans need to win at least two of these races. Based on the numbers above, what is the probability of the Republicans taking control of the senate?



So, republicans need to win at least 2 seats... therefore the probability they take over the senate equals 0.212 + 0.196 + 0.135 + 0.104 = 0.647

2. The betting markets are currently trading at a 80% probability for the Republicans to control the senate. How does your answer from the question above compare to this number? Can you explain why you are seeing a difference? (*Hint:* Did you have to make any assumption to answer the first questions)?

   We assumed independence across the 3 states. My guess is that those 3 events are not really independent of each other... for example, given that LA went to republicans on election night that should impact the conditional probability of NC going republican, right?

Problem 2

Suppose a person is randomly drawn from a large population and then tested for a disease.
Let $D = 1$ if the person has the disease and 0 otherwise.
Let $T = 1$ if the person tests positive and 0 otherwise.

Suppose
$P(D = 0) = .99$.
$P(T = 1 \mid D = 0) = .01$.
$P(T = 1 \mid D = 1) = .97$.

(a) Draw the diagram depicting the marginal of D and the conditional of $T \mid D$. (you know, the one that branches as you go left to right).

(b) Give the joint distribution of D and T in the two way table format.

(c) What is $P(D = 1 \mid T = 1)$?

(b)

```
      T0     T1
D0 0.9801 0.0099
D1 0.0003 0.0097
```

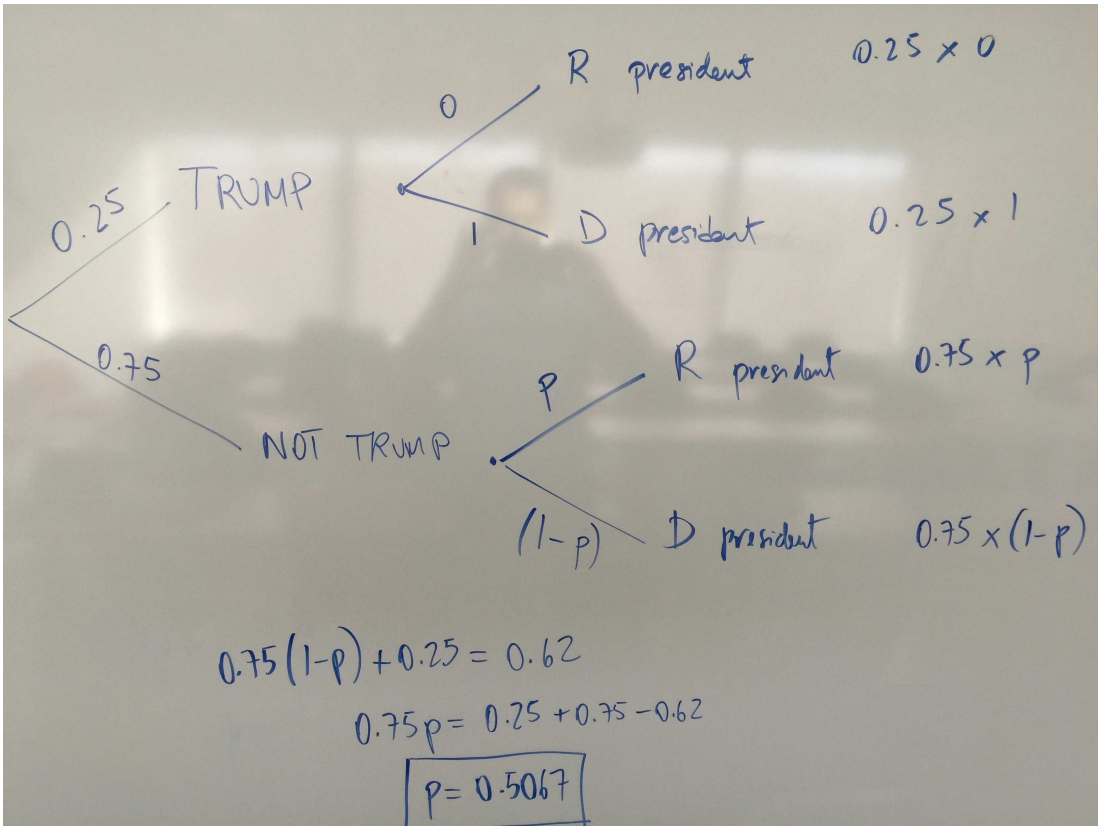
```
> .99*.99
[1] 0.9801
> .99*.01
[1] 0.0099
> .01*.03
[1] 3e-04
> .01*.97
[1] 0.0097
```

(c)
$P(D = 1 \mid T = 1) = P(D = 1, T = 1)/P(T = 1) = .0097/(.0097 + .0099) = 0.494898$.

Problem 3

Based on betting markets the probability of Donald Trump being the Republican nominee is 25%. The same markets have the probability that the next President will be a Democrat at 62%.

Assume that if Trump is the nominee he has no chance of becoming the President... so, if the nominee is someone NOT Donald Trump, what is the probability of a Republican becoming the President?



Problem 4

Here's a simplified look at a spam filter algorithm...

We are worried about the term *"Nigerian general"* and our IT team has figured that $pr(\text{"Nigerian general"}|\text{junk mail}) = 0.20$ and $pr(\text{"Nigerian general"}|\text{NOT junk mail}) = 0.001$
In addition they figured that half of our emails is junk.

1. What is the marginal probability of seeing *"Nigerian general"* in a message?
   In other words, what is the $pr(\text{"Nigerian general"})$?

2. If the spam filter always classify a message containing *"Nigerian general"* as junk, how often will it make a mistake?
   In other words, what is the $pr(\text{NOT junk mail}|\text{"Nigerian general"})$?

Let's first build the joint distribution table:

	Junk	Not Junk
Nigerian general	0.2×0.5	0.001×0.5
Not NG	0.8×0.5	0.999×0.5

	Junk	Not Junk
Nigerian general	0.1	0.0005
Not NG	0.4	0.4995

1. Therefore, the $pr(\text{"Nigerian general"}) = 0.1 + 0.0005 = 0.1005$

2. $pr(\text{NOT junk mail}|\text{"Nigerian general"}) = \frac{pr(\text{NOT junk mail and "Nigerian general"})}{pr(\text{"Nigerian general"})}$

$pr(\text{NOT junk mail}|\text{"Nigerian general"}) = \frac{0.0005}{0.10005} = 0.005$

i.e., only 0.5% of the emails will be wrongly flagged as junk.

Problem 5

After finishing your MBA and becoming a consultant you will be flying for meetings regularly! Say you'll be traveling routinely to Boston, Orlando, Philadelphia and San Diego... Also, you like to accumulate miles with both Delta and US Airways and you are trying to decide which airline will minimize potential delays. After a quick look on-line you find in the U.S. Bureau of Transportation Statistics the following probability table describing the delays of these two airlines:

	Delta	US Airways
Delayed	20%	22%
On Time	80%	78%

Is this enough information for you to make a decision? If not, can you explain a possible scenario in which choosing Delta doesn't make sense?
The table above suggests that the two airlines perform similarly with a small advantage to Delta with 80% on time arrivals relative to 78% of US Airways.
Before I make a decision, however, I should think about whether or not there are lurking variables (other variables) that really allow me to understand how the airlines are behaving. For example, imagine if US Airways flies into airports with bad weather more often than Delta... would that potentially change your view on this?

Just to go further on this example let's look at the on time arrival data by destination...

	Delta	US Airways
Boston	80.1%	81.7%
Orlando	80.5%	84.5%
Philadelphia	70.5%	74.3%
San Diego	84.2%	85.4%

Now, should you pick Delta or US Airways? What if I tell you that because of its hub US Airways flies more frequently to Philadelphia, does that potentially explains the results from the first table?