Bayesian Analysis

ORIGIN := 0

In medicine, the predictive value of patient symptoms or of a clinical test is often analyzed using Bayesian analysis. Let A represent symptoms presented by a patient or results of a clinical test. Let one or more B's represent patient condition(s) such as a disease. Using Bayes rule, the conditional probability $P(A|B_i)$ can be estimated from the portion of patients with known condition(s) B_i showing positive test results $P(B_i|A)$. Total probablility $P(B_i)$ for the condition(s) can be estimated from the population at large. Bayes' Rule allows the researcher to modify prior probilities from the population at large with specific information about the sensitivity and specificity of each test. This is powerful stuff!

Clinical Terminology often used with Bayes' Rule - following Rosner 2006:

(see also: http://en.wikipedia.org/wiki/Sensitivity_and_specificity):

P(B|A) / P(B|notA) — **Relative Risk**

$P(B_i|A)$ — **Predictive value positive of the test (PV^+)** — **Bayesian Result (Posterior)**

$P(\sim B_i|notA)$ — **Predictive value negative of the test (PV^-)**

$P(A|B_i)$ — ***Sensitivity*** **of the symptoms or test**

$P(notA|notB_i)$ — ***Specificity*** **of the symptoms or test**

$(notA|B_i)$ — **false negative for the symptoms or test**

$(A|notB_i)$ — **false positive for the symptoms or test**

Associated Truth Table:

		True state (Prior)		
		+	-	
Test outcome	+	*Sensitivity*	false +	test +
	-	false -	*Specificity*	test -
		true +	true -	

Bayes' Rule:

The point of this procedure for two events A & B is to estimate one conditional probability P(B|A) from the other conditional probability P(A|B) and one total probablity P(B).

$$P\left(\frac{B}{A}\right) = \frac{P\left(\frac{A}{B}\right)\cdot P(B)}{P\left(\frac{A}{B}\right)\cdot P(B) + P\left(\frac{A}{notB}\right)\cdot P(notB)}$$

< Of course, as above, the defined roles of A & B here can be reversed.

P (B|A) = (Joint Probability · Prior Probability) / Total Row Probability

Example:

Test for Cancer:

Suppose your patient apprehensively awaits the results of a test for a specific type of cancer. The test is a very good one with a sensitivity of 99% and specificity of 95%. The incidence of this cancer in the general population is 6%. Given that the patient tests positive for the test, do you advise immediate surgery?

$\text{Sensitivity} := 0.99$

$\text{Specificity} := 0.95$

$\text{FalsePos} := 1 - \text{Specificity}$ $\qquad \text{FalsePos} = 0.05$

$\text{PriorProb} := 0.06$

$\text{TotalPos} := (0.99 \cdot 0.06) + (0.05 \cdot 0.94)$ $\qquad \text{TotalPos} = 0.1064$

		Cancer		
		+	-	
Test outcome	+	*0.99*	0.05	0.1064
	-	0.01	0.95	0.8936
		0.06	0.94	

$$\text{PredValue} := \frac{\text{Sensitivity} \cdot \text{PriorProb}}{\text{TotalPos}} \qquad \text{PredValue} = 0.55827 \qquad \frac{0.99 \cdot 0.06}{(0.99 \cdot 0.06) + (0.05 \cdot 0.94)} = 0.55827$$

As you can see, even though the test is very good, because the incidence of this cancer in the general population is quite low, the frequency of false positives nearly equal the frequency of true positives for this test.

Expected frequency of false positives = $(\text{FalsePos}) \cdot (1 - \text{PriorProb}) = 0.047$

Expected frequency of true positives = $\text{Sensitivity} \cdot \text{PriorProb} = 0.059$

Note: the denominator records the total frequency of expected positive results either true or false.

$\text{TotalPos} := \text{Sensitivity} \cdot \text{PriorProb} + (\text{FalsePos}) \cdot (1 - \text{PriorProb})$ $\qquad \text{TotalPos} = 0.1064$

So, what to tell the patient? The test in itself is not conclusive. It is only slightly more likely that the patient has cancer than not. Of course, the patient will not want to hear this! Other factors and perhaps other disgnostic procedures will need to be considered...

Bayes Rule as Proportional probabilities:

The denominator in Bayes rule (TotalPos in the example above) is often a constant estimated from the literature directly without explicitly calculating false positives. In this event, Bayes Rule can be restated as follows using proportionality constant k:

$$P\left(\frac{B}{A}\right) = k \cdot P\left(\frac{A}{B}\right) \cdot P(B)$$

What's important is finding from some outside source for the conditional probabilities P(A|B), also known as *likelihoods*.

Example:

Lung Cancer Death due to Smoking:

To estimate death rates associated with smoking, one can either conduct an expensive survey to measure this directly or use available data in the literature and Bayes rule. According to the CDC, there were 158,248 deaths in the USA due to lung cancer in 2010, The US census counted population size in 2010 was 308,745,538. According to the American Lung Association website, smoking is responsible for 90% of lung cancer related deaths. The CDC estimates that 18.8% of adults in the US are smokers. What's the relative risk of smokers to non-smokers of dying from lung cancer?

		Lung Cancer Death		
		+	**-**	
Smokers	**+**	*0.9*		**0.188**
	-	0.1		**0.812**
		0.000513	**0.999487**	

Frequency of Lung Cancer Deaths:

$$\frac{158248}{308745538} = 0.000513$$

Smokers Risk:

$\text{sensitivityS} := 0.9$ **= P(smoker+ | LC death+)**

$\text{priorProb} := 0.000513$ **= P(LC death)**

$\text{totalPosS} := 0.188$ **= P(smoker+)**

PredValueS **= P(death+ |smoker+)**

$$\text{PredValueS} := \frac{\text{sensitivityS} \cdot \text{priorProb}}{\text{totalPosS}} \qquad \text{PredValueS} = 0.00246$$

< risk to smokers

Non-smokers Risk:

$\text{sensitivityN} := 0.1$ **= P(smoker- | death+)**

$\text{priorProb} := 0.000513$ **= P(deaths)**

$\text{totalPosN} := 1 - 0.188$ **= P(smoker-)** $\text{totalPosN} = 0.812$

PredValueN **= P(death+ | smoker-)**

$$\text{PredValueN} := \frac{\text{sensitivityN} \cdot \text{priorProb}}{\text{totalPosN}} \qquad \text{PredValueN} = 0.00006318$$

< risk to non-smokers

Relative Risk:

$$\frac{\text{PredValueS}}{\text{PredValueN}} = 38.872$$

< Smokers have almost 39 times the risk of dying of lung cancer than non-smokers.

Solving Bayes rule using a Tree Diagram:

Bayes Rule can also be stated in terms of intersection of probabilities as follows:

$$P\left(\frac{B}{A}\right) = \frac{P(A \wedge B)}{P(A)}$$

As a result, it is easy to visualize the problem in terms of a tree diagram.

Example:

This example and associated diagram comes from *www.milefoot.com/math/stat/prob-bayes.htm* analyzing a similar problem to above, but in an enlightening way using a tree diagram. I'll work the problem using Bayes rule and as intersection of probabilities using a tree diagram. Setup from Milefoot.com:

"Suppose 0.1% of the American population currently has lung cancer, that 90% of all lung cancer cases are smokers, and that 21% of those without lung cancer also smoke."

Using Bayes Rule:

$L := 0.001$ $\quad L = 0.001$ **= fraction with lung cancer in population P(L) = *priorProb***

$nL := 1 - L$ $\quad nL = 0.999$ **= fraction with no lung cancer P(nL)**

$SL := 0.9$ $\quad SL = 0.9$ **= fraction lung cancer patients are smokers P(S|L) = *sensitivityS***

$nSL := 1 - SL$ $\quad nSL = 0.1$ **= fraction lung cancer patients are non Smokers P(nS|L) = *sensitivityN***

$SnL := 0.21$ $\quad SnL = 0.21$ **= fraction non lung cancer patients are smokers P(S|nL) = *falsePosS***

$nSnL := 1 - SnL$ $\quad nSnL = 0.79$ **= fraction non lung cancer patients are non smokers P(nS|nL) = *falsePosN***

Risk to Smokers: P(L|S):

$$LS := \frac{SL \cdot L}{SL \cdot L + SnL \cdot nL} \qquad LS = 0.004272$$

$$LS := \frac{0.9 \cdot 0.001}{0.21069} \qquad LS = 0.004272$$

		Lung Cancer		
		+	-	
Smokers	+	*0.9*	0.21	0.21069
	-	0.1	0.79	0.78931
		0.001	0.999	

Risk to Non-smokers: P(L|nS):

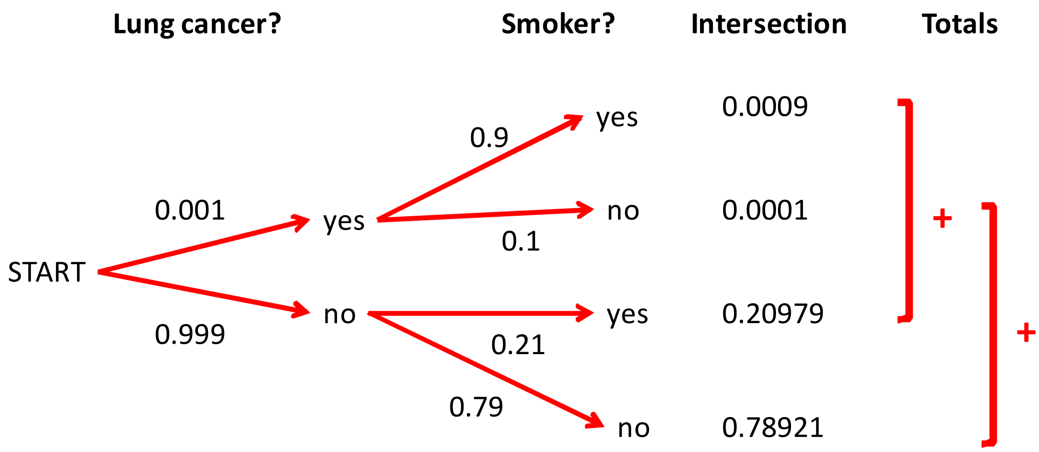
$$LnS := \frac{nSL \cdot L}{nSL \cdot L + nSnL \cdot nL} \qquad LnS = 0.0001267$$

$$LnS := \frac{0.1 \cdot 0.001}{0.78931} \qquad LnS = 0.0001267$$

Relative Risk:

$$\frac{LS}{LnS} = 33.717$$

Using a Tree Diagram:



Risk to Smokers: P(L|S):

$$P\left(\frac{L}{S}\right) = \frac{P(L \wedge S)}{P(S)}$$

< intersection determined by following a single path

< All smokers

$$LS := \frac{0.001 \cdot 0.9}{0.0009 + 0.20979} \qquad LS = 0.004272$$

Risk to Non-smokers:

$$P\left(\frac{L}{nS}\right) = \frac{P(L \wedge nS)}{P(nS)}$$

< intersection determined by following a single path

< All non-smokers

$$LnS := \frac{0.001 \cdot 0.1}{0.0001 + 0.78921} \qquad LnS = 0.0001267$$

Relative Risk:

$$\frac{LS}{LnS} = 33.717$$

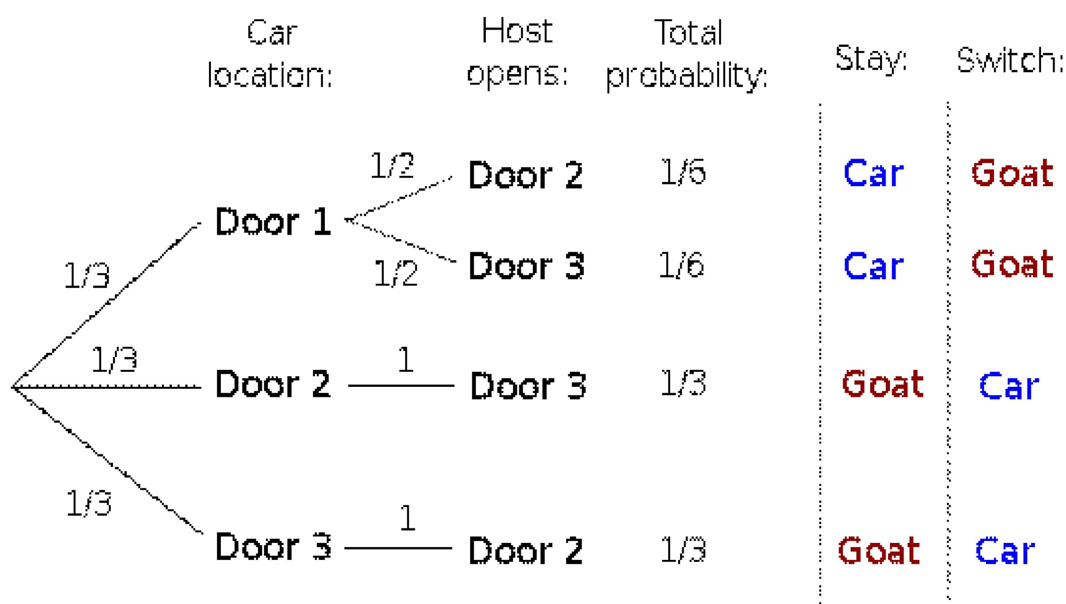
The Monty Hall Problem

This classic problem is a lot of fun because it stirs up a lot of controversy! Monty Hall was the original host of the perennial game show *Let's make a Deal*. Although not exactly following the show's rules, the problem goes like this:

A contestant is shown 3 doors. One door hides a shiny new *car*, whereas the other two doors only *goats*. The contestant initially chooses a door. Monty then reveals what's behind one of the other doors - *goats!* He then invites the contestant to *stay* with his first choice or *switch* to the remaining door.
What should the contestant do?

In this problem, the initial choice involves choosing one door out of three with equal probabilities. Thus P(B) = 1/3.

To calculate P(A|B) involves considering all possibilities where the *car* might be and given knowledge of this what door Monty chooses to open. If he has a choice of doors, he opens one with equal probability to the other.



from Wikipedia at: http://en.wikipedia.org/wiki/Monty_Hall_problem

Assuming the contestant initially chose door 1 and Monty opens door 3, there are only two viable options:
If the *car* is behind door 1, Monty had choice of doors 2 or 3 each with liklihood 1/2.
If the car is behind door 2, Monty can only open door 3, with likelihood 1.

So:

P(B|A) = K·(1/2)·(1/3) = K·(1/6) if the contestant *stays* with door 1, but

P(B|A) = K·(1)·(1/3) = K·(1/3) if the contestant *switches* to door 2.

Allowing the contestant to choose a different door at the start, yields different paths but the same results.

Thus switching yields *twice the probability of winning the car* because Monty gave us additional information by opening that door! That's Bayesian...

Check out the simulation in section 4.6 to see if this works! If you remain unconvinced, I particularly like the argument about using a full deck of 52 cards as opposed to the simulation with only three.