Understanding Probability

Consider the following statements. How do you interpret "probability" in each one of those? And how is it computed?

- Coin Toss – Probability of Head is ½
- Weather – Probability of thunderstorm tomorrow is 25%
- Cricket– India has only a 80% chance of a win when Virat as a captain

Probability vs Statistics

- Probability – Predict the likelihood of a future event
- Statistics – Analyze the past events

Questions addressed -

- Probability – What will happen in a given ideal world?
- Statistics – How ideal is the world?

Probability - Applications

8 National Vital Statistics Reports, Vol. 54, No. 14, April 19, 2006

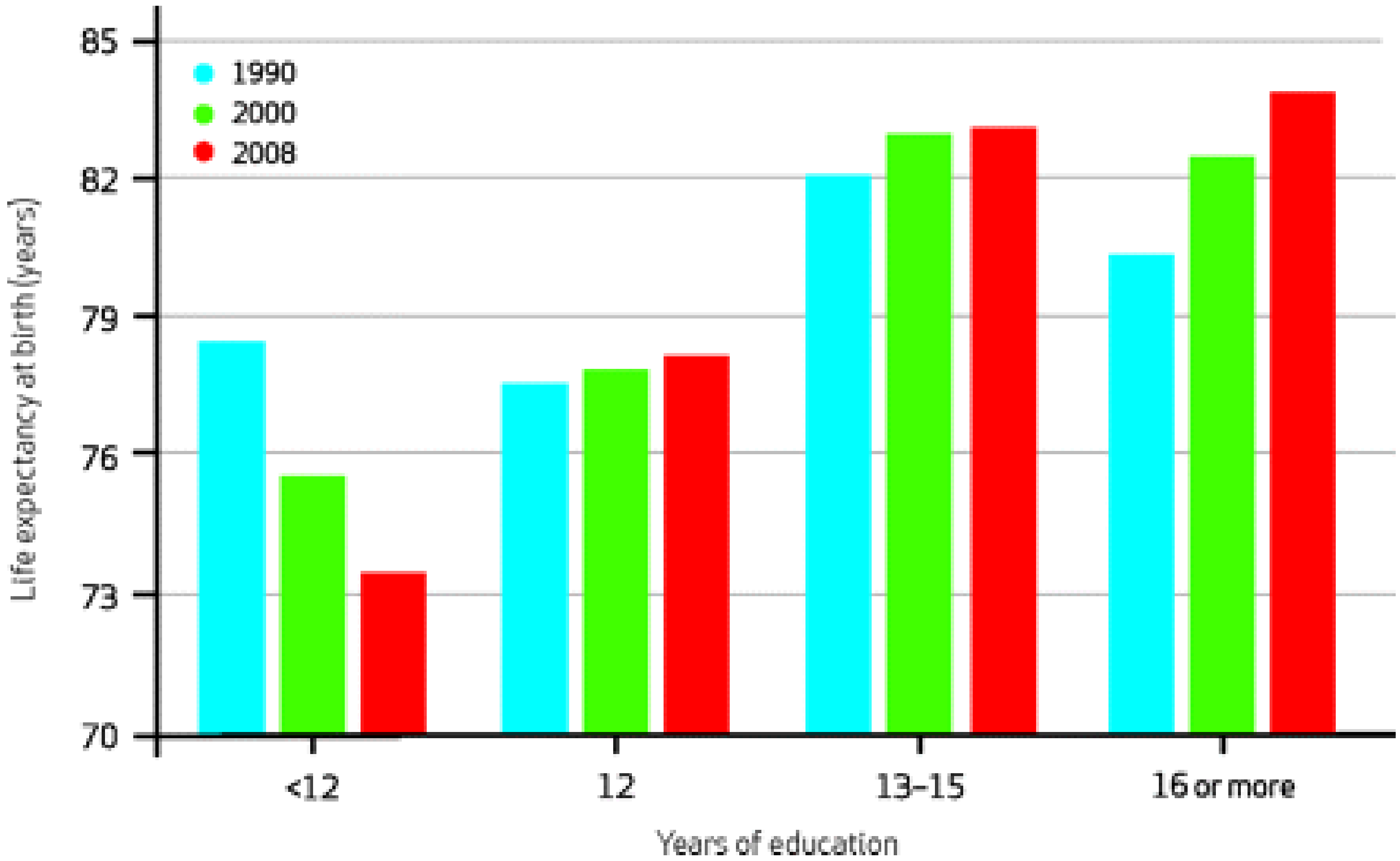
Table 1. Life table for the total population: United States, 2003

Click here for spreadhse

Age	Probability of dying between ages x to $x+1$	Number surviving to age x	Number dying between ages x to $x+1$	Person-years lived between ages x to $x+1$	Total number of person-years lived above age x	Expecta of lif at age
	$q(_x)$	$l(_x)$	$d(_x)$	$L(_x)$	$T(_x)$	$e(_x)$
0–1	0.006865	100,000	687	99,394	7,743,016	77.4
1–2	0.000469	99,313	47	99,290	7,643,622	77.0
2–3	0.000337	99,267	33	99,250	7,544,332	76.0
3–4	0.000254	99,233	25	99,221	7,445,082	75.0
4–5	0.000194	99,208	19	99,199	7,345,861	74.0
5–6	0.000177	99,189	18	99,180	7,246,663	73.1
6–7	0.000160	99,171	16	99,163	7,147,482	72.1

Insurance industry uses probabilities in actuarial tables for setting premiums and coverages.

Probability - Applications



Probability - Applications

- Gaming industry – Establish charges and payoffs
- Manufacturing/Aerospace – Prevent major breakdowns
- Business – Deciding on a business proposal based on probability of success vs cost
- Risk Evaluation – Scenario analysis

Assigning Probabilities

Classical Method – *A priori or Theoretical*

Probability can be determined prior to conducting any experiment.

$$P(E) = \frac{\#\ of\ \ outcomes\ in\ which\ the\ even\ occurs}{total\ possible\ \#\ of\ \ outcomes}$$

Example: Tossing of a fair die

Computing A priori Probability

Find the probability of pulling a yellow marble from a bag of 3 yellow, 2 red, 3 green and 1 blue marbles

$$P(yellow) = \frac{No\ of\ yellow\ marbles}{Total\ number\ of\ marbles}$$

$$= \frac{3}{9}$$

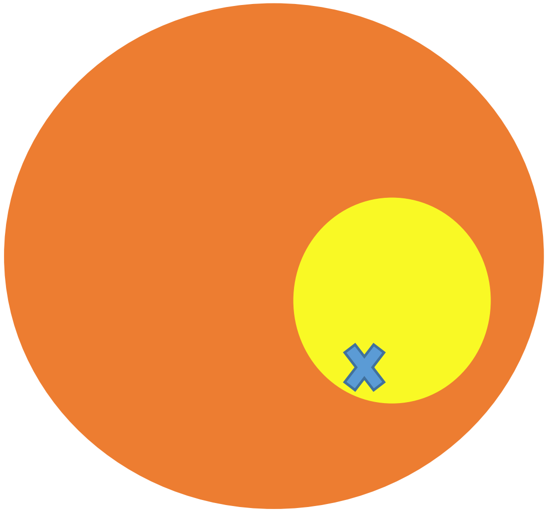
Computing probability

There are two concentric circles, The circumference of a circle is 36π. Contained in that circle is a smaller circle with an area of 16π. A point is selected at random from inside the larger circle. What is probability that the point also in the same circle.

$$\text{Area of smaller circle} = 16\,\pi$$

$$\text{Area of larger circle} \;= \pi \;*\; \left(\frac{36\pi}{2\pi}\right)^2$$

$$= 324\,\pi$$



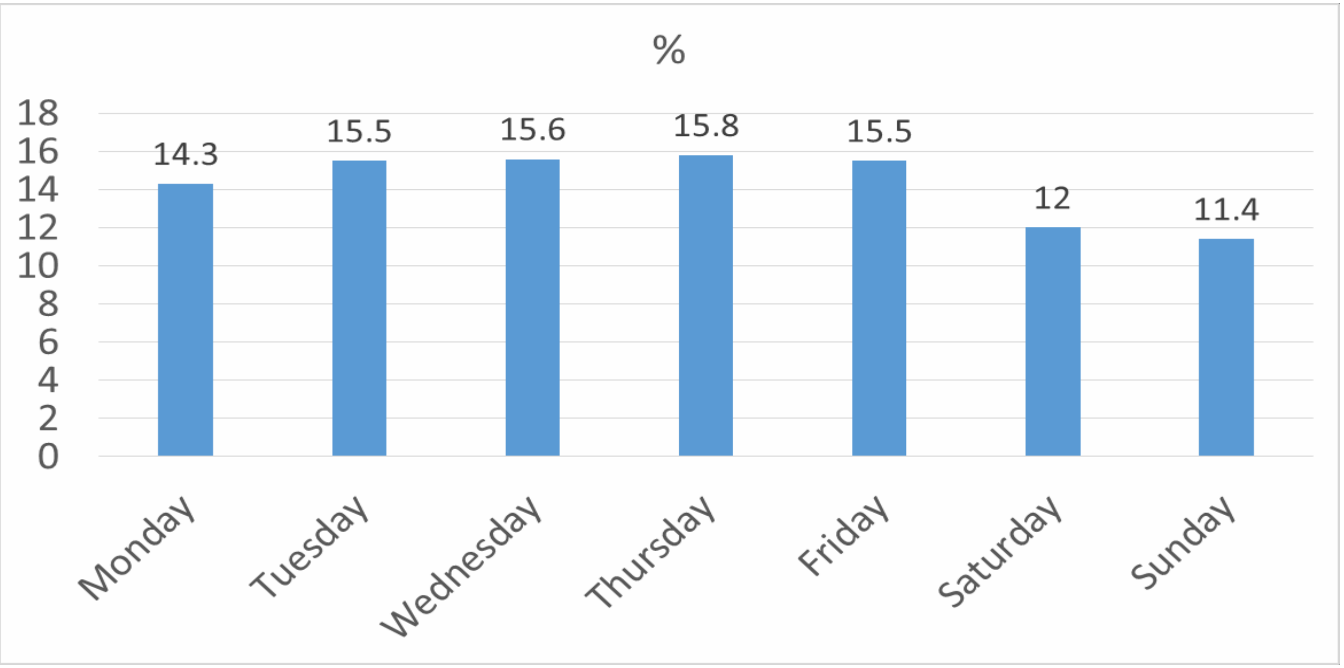
$$P(point\ in\ small\ circle\ ) = \frac{Area\ of\ Large\ circle}{Area\ of\ small\ circle}$$

$$= 16\pi/324\pi$$

Assigning Probabilities

What is the probability of a baby being born on a Wednesday?

A-priori probability = $\frac{1}{7} = 14.3\%$



Data from "Risks of Stillbirth and Early Neonatal Death by Day of Week", by Zhong-Cheng Luo, Shiliang Liu, Russell Wilkins, and Michael S. Kramer, for the Fetal and Infant Health Study Group of the Canadian Perinatal Surveillance System. Data of 3,239,972 births in Canada between 1985 and 1998. The reported percentages do not add up to 100% due to rounding.

Assigning Probabilities

Empirical Method – *A posteriori or Frequentist*

Probability can be determined post conducting a thought experiment.

$$P(E) = \frac{\#\ of\ times\ an\ event\ occured}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occured}$$

Example: Tossing of a weighted die…well!, even a fair die.

The larger the number of experiments, the better the approximation.

This is the most used method in statistical inference.

Assigning Probabilities

Subjective Method

Based on feelings, insights, knowledge, etc. of a person.

What is the probability of India winning the upcoming World cup 2019?

Probability - Terminology

Sample Space – Set of all possible outcomes, denoted S.

Example:

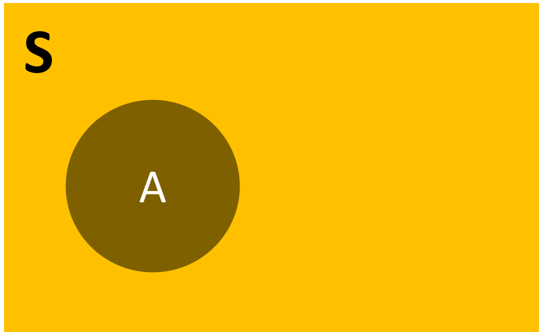
After 2 coin tosses, the set of all possible outcomes are {HH, HT, TH, TT}

Event – A subset of the sample space.

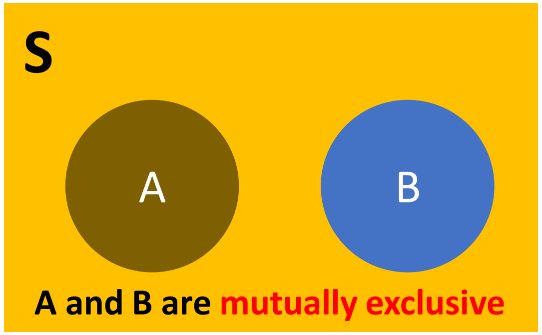
An Event of interest might be - HH

Probability - Rules

$$P(s) = 1$$



$$0 \leq P(A) \leq 1$$

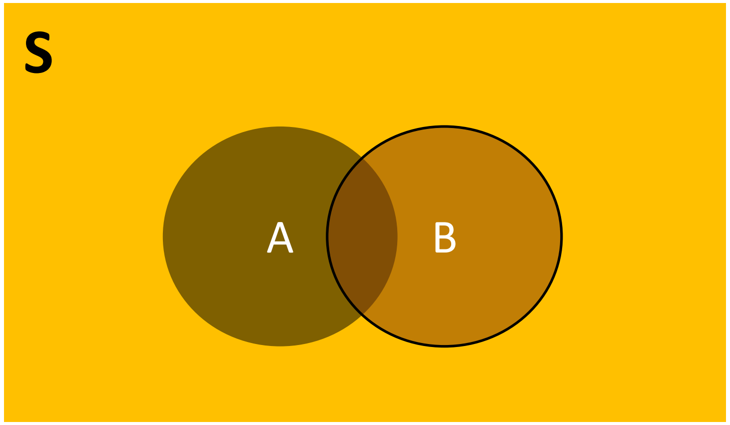


$$P(A \; or \; B) = P(A) + P(B)$$

Area of the rectangle denotes sample space, and since probability is associated with area, it cannot be negative.

Mutually Exclusive – If event A happens, event B cannot.

Probabilities Rules



$$P(A \, or \, B) = P(A) + P(B) \; - P(A \, and \, B)$$

Example

- Event A – Customers who default on loans
- Event B – Customers who are High Net Worth Individuals

Probability - Rules

Independent Events – Outcome of event B is not dependent on the outcome of event A.

Probability of customer B defaulting on the loan is not dependent on default (or otherwise) by customer A.

$$P(A\ and\ B) = P(A) * P(B)$$

If the probability of getting an *easy call is 0.7, what is the* probability that the next 3 calls will be *easy?* .

$$P(easy_1\ and\ easy_2\ and\ easy_3) = 0.7^3 = 0.343$$

Probability Question

A basketball team is down by 2 points with only a few seconds remaining in the game. Given that:

- Chance of making a 2-point shot to tie the game = 50%
- Chance of winning in overtime = 50%
- Chance of making a 3-point shot to win the game = 30%

What should the coach do: go for 2point or 3-point shot?
What are the assumptions, if any?

Probability - Types

Customer-Id	Customer Name	Age	Default
846596	Srikanth	28	Yes
846597	Raghu	25	No
846598	Ramya	24	No
…	…	…	…

Probability - Types

Contingency table summarizing 2 variables, *Loan Default* and *Age:*

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Probability - Types

Convert it into probabilities

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability - Types

Marginal Probability

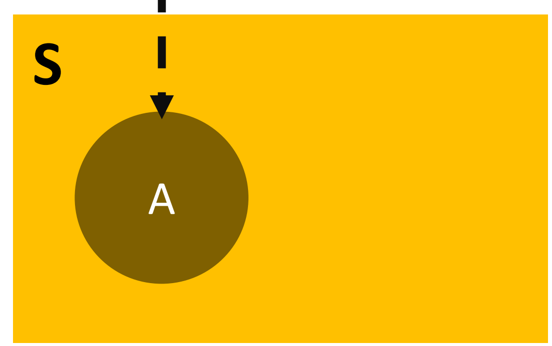
		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Marginal Probability

Probability describing a single attribute

P(Middle) = 0.690

P(old) = 0.008



Probability - Types

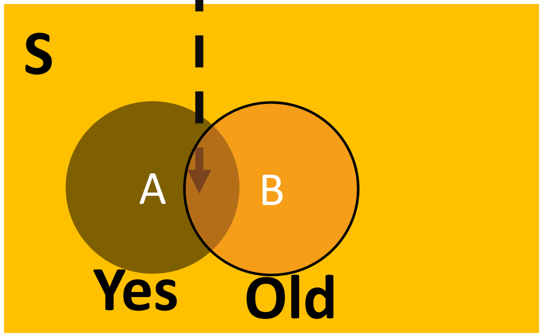
Joint Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Joint Probability

Probability describing a combination of attribute

P(Yes and old) = 0.003



Probability - Types

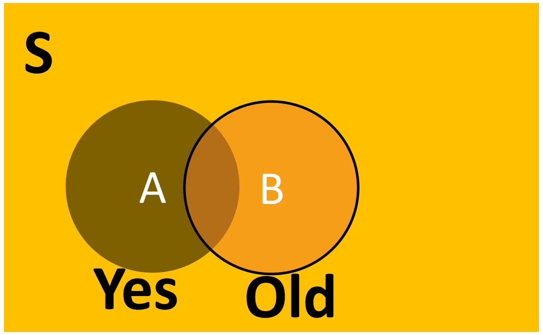
Union Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

P(Yes or old) = P(Yes) + P(old) – P(Yes and old)

= 0.184 + 0.008 – 0.003

= 0.189

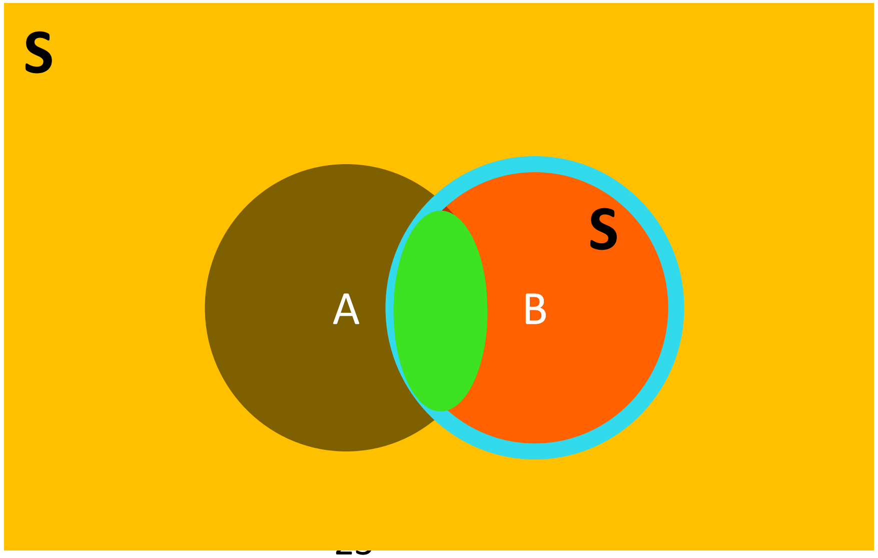


Probability - Types

Conditional Probability

- Probability of *A occurring* ***given that*** B has occurred.
- The sample space is restricted to a single row or column.
- This makes rest of the sample space irrelevant.

Probability, i.e., $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$



Probability - Types

Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

What is the probability that a person will not default on the loan payment **given she is middle-aged?**

Probability, i.e., $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$

P(No | Middle-Aged) = 0.586/0.690 = 0.85

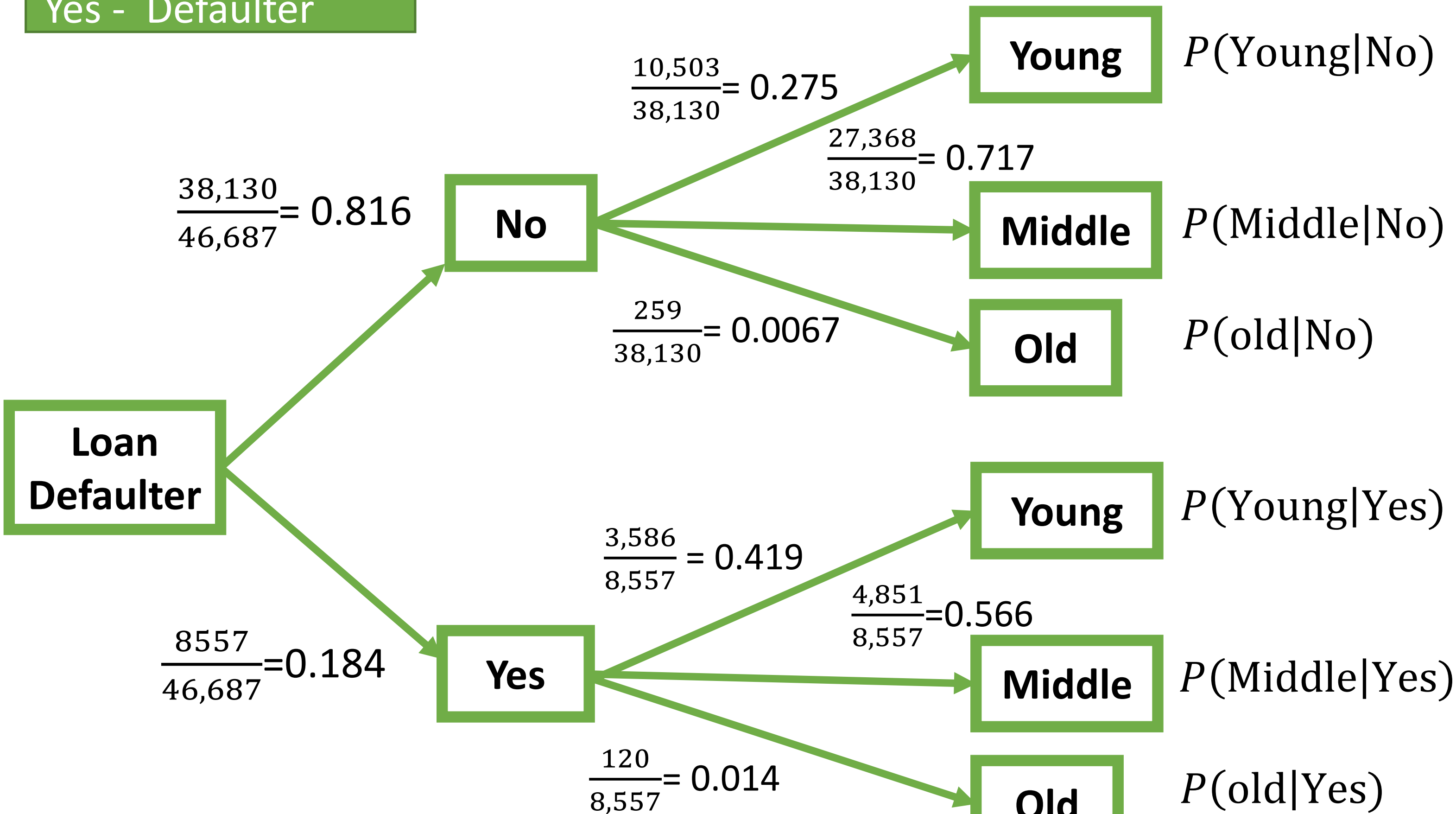
Note that this is the ratio of **Joint Probability to Marginal**

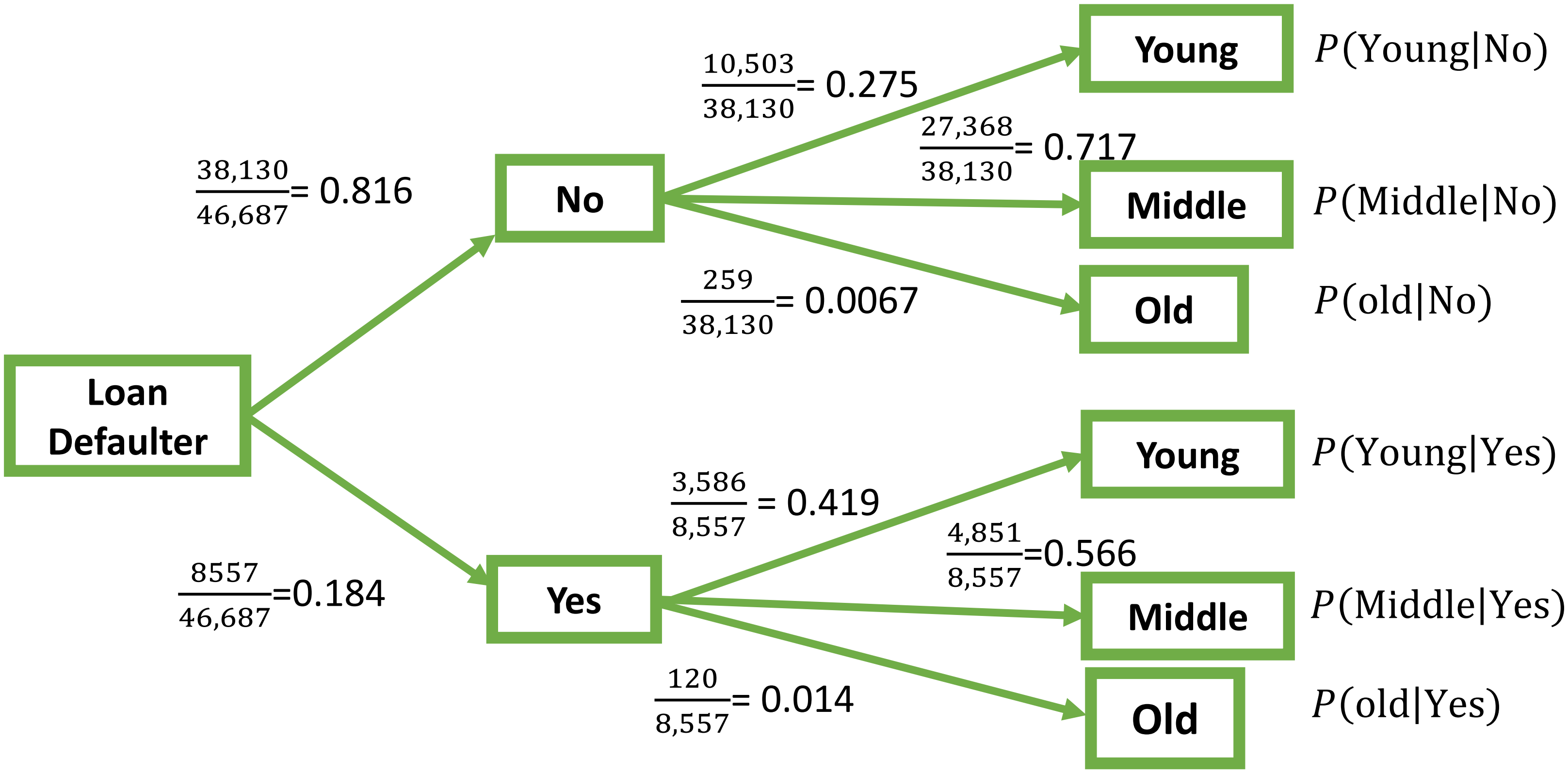
P(Middle-Aged | No) = $\frac{0.586}{0.816}$ = 0.72 (Order Matters)

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Age			
Young	Middle-aged	Old	Total
0.225	0.586	0.005	0.816
0.077	0.104	0.003	0.184
0.302	0.690	0.008	1.000

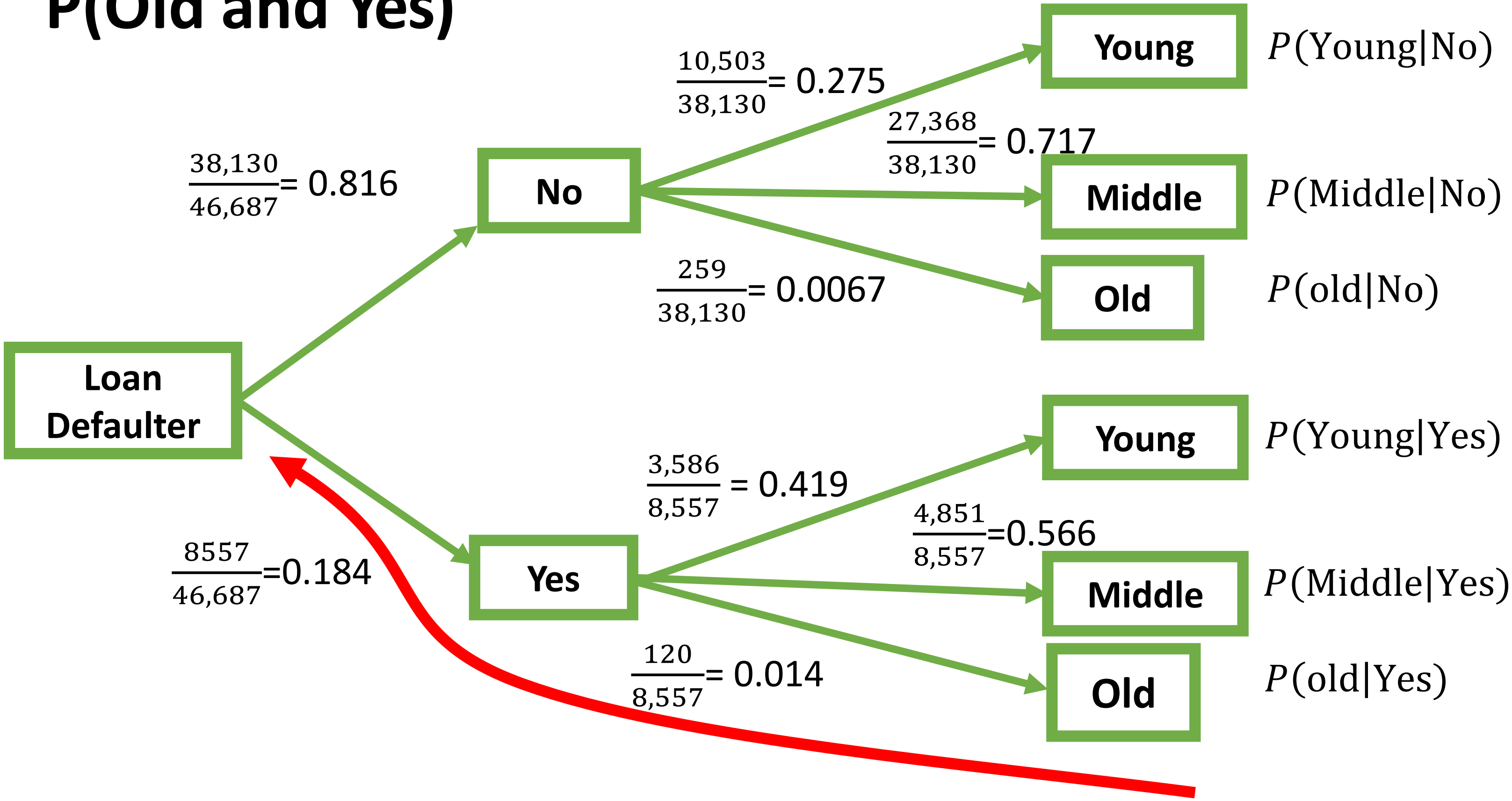
No – Non-defaulter
Yes - Defaulter





- P(Old and Yes)
- P(Yes and Old)
- P(Old)
- P(Yes)
- P(Old | Yes)
- P(Yes | Old)
- P(Young | No)

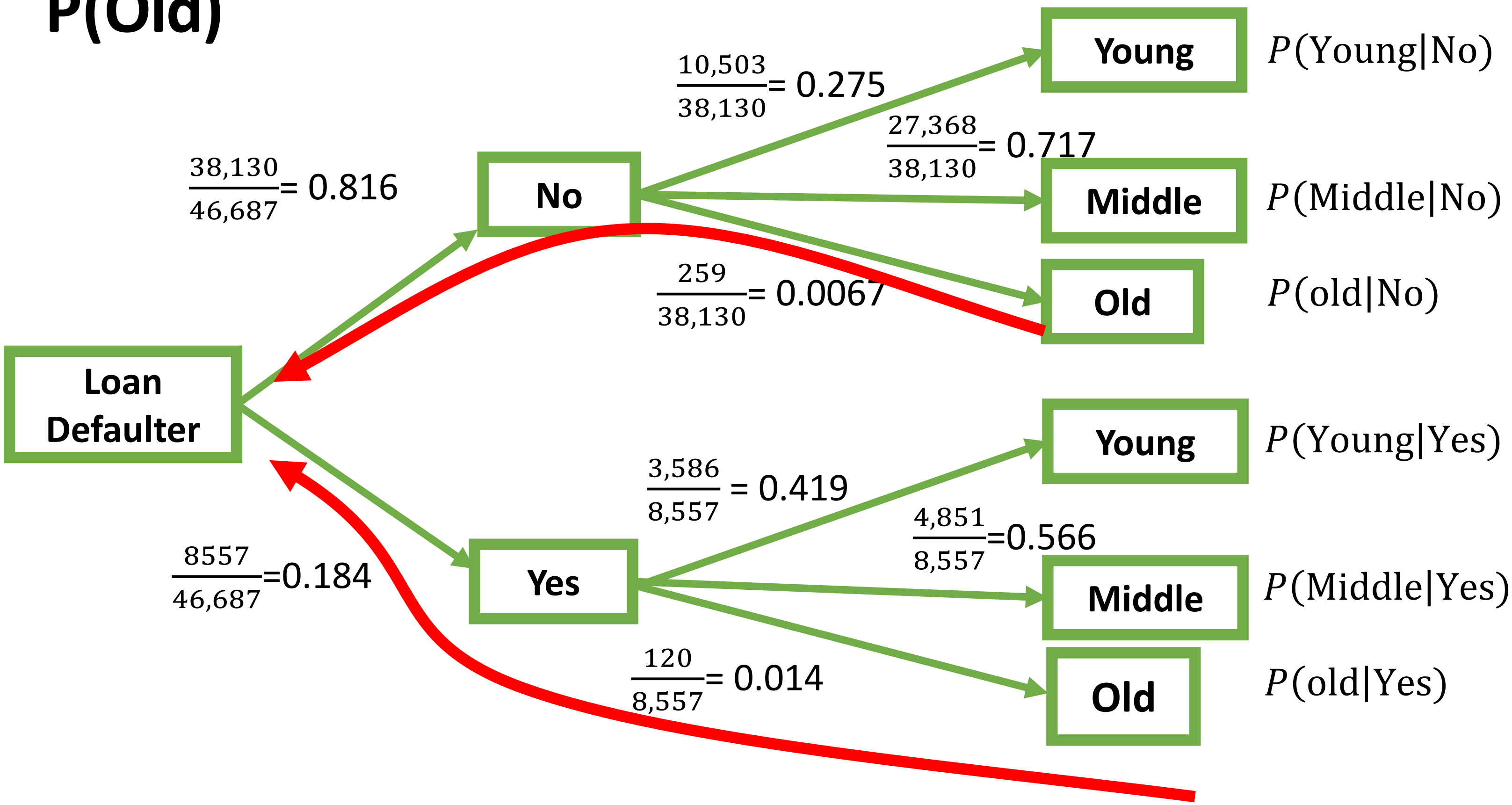
P(Old and Yes)



$$P(\text{Old} \mid \text{Yes}) = \frac{P(\text{Old and Yes})}{P(\text{Yes})}$$

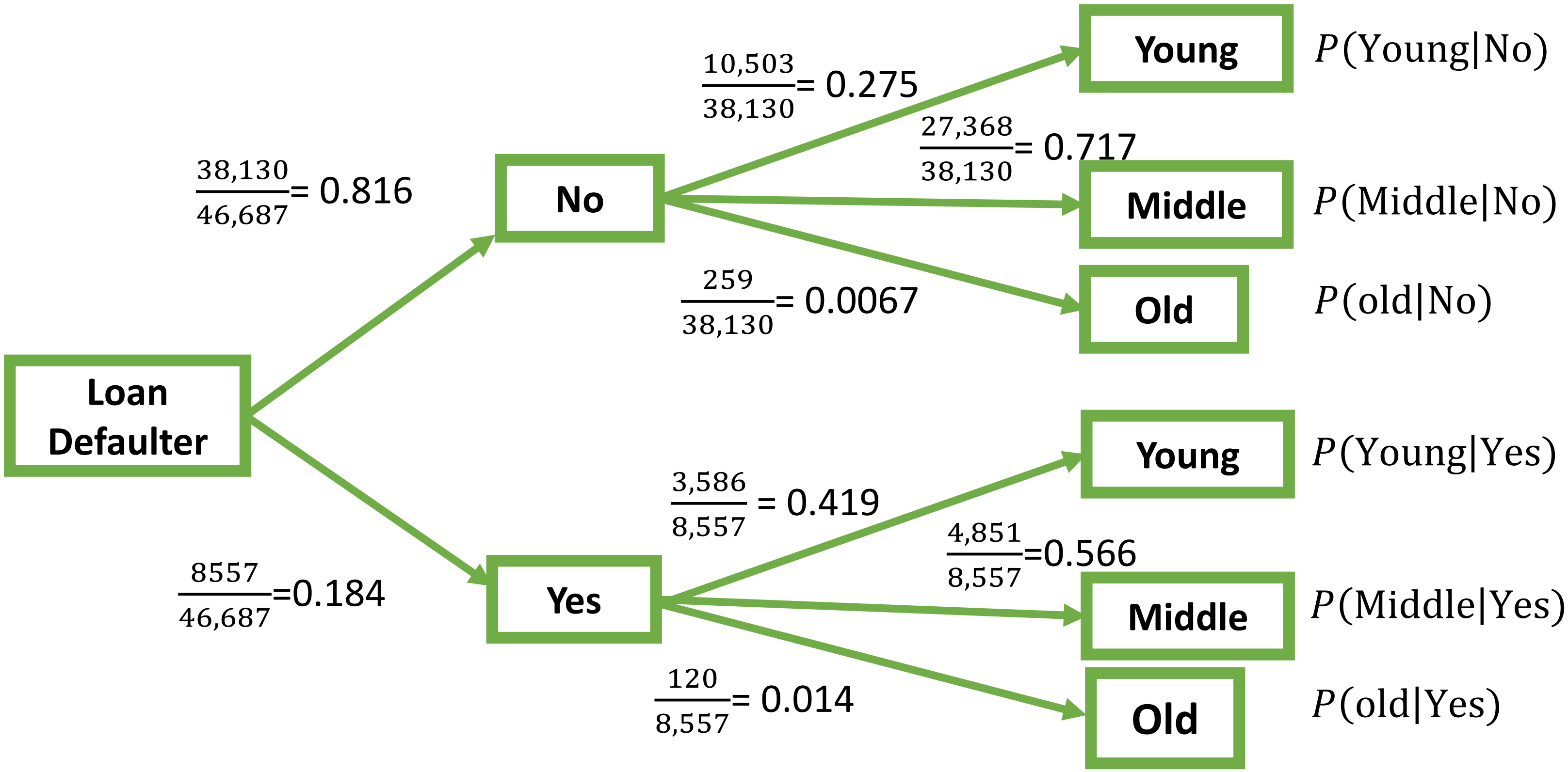
$$P(\text{Old and Yes}) = \text{P}(\text{Old} \mid \text{Yes}) * \text{P}(\text{Yes}) = 0.014 * 0.184$$

P(Old)



P(Old) $= P(\text{Old and Yes}) * P(\text{Old and No})$

$$P(Old) = 0.014 * 0.184 + 0.0067 * 0.816$$



- P(Old and Yes)
- P(Yes and Old)
- P(Old)
- P(Yes)
- P(Old | Yes)
- P(Yes | Old)
- P(Young | No)

Probability - Types

Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \Rightarrow P(A \text{ and } B) = P(B) \ * \ P(A|B)$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \Rightarrow P(A \text{ and } B) = P(A) \ * \ P(B|A)$$

Equating, we get

$$P(B) \ * \ P(A|B) = P(A) \ * \ P(B|A)$$

$$\therefore \ P(A|B) = \frac{P(A) \ * \ P(B|A)}{P(B)}$$

Probability - Types

$$\therefore \mathrm{P(A|B)} = \frac{\mathrm{P(A)} * \mathrm{P(B|A)}}{\mathrm{P(B)}}$$

In loan defaulters older people make up only 1.4%. Now the probability that someone defaults on a loan is 0.184, Find the probability default on loan knowing that he is old person. Older people make up only 0.8%.

Ans:

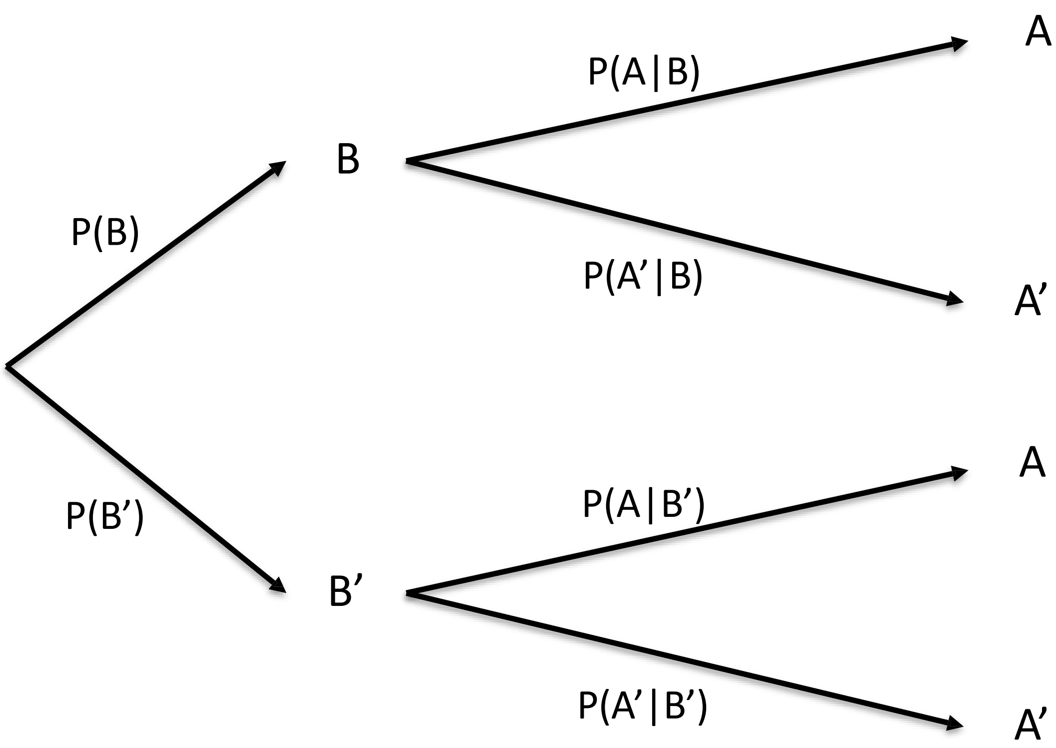
$$P(\mathrm{Old|Yes}) = 0.014$$

$$P(Old) = 0.008$$

$$P(Yes) = 0.184$$

$$\mathrm{P(Yes|Old)} = \frac{\mathrm{P(Yes)} * \mathrm{P(Old|Yes)}}{\mathrm{P(Old)}} = \frac{0.184 * 0.014}{0.008} = 0.32$$

Generalized Probability Tree



P(B and A) = P(B) * P(A|B)

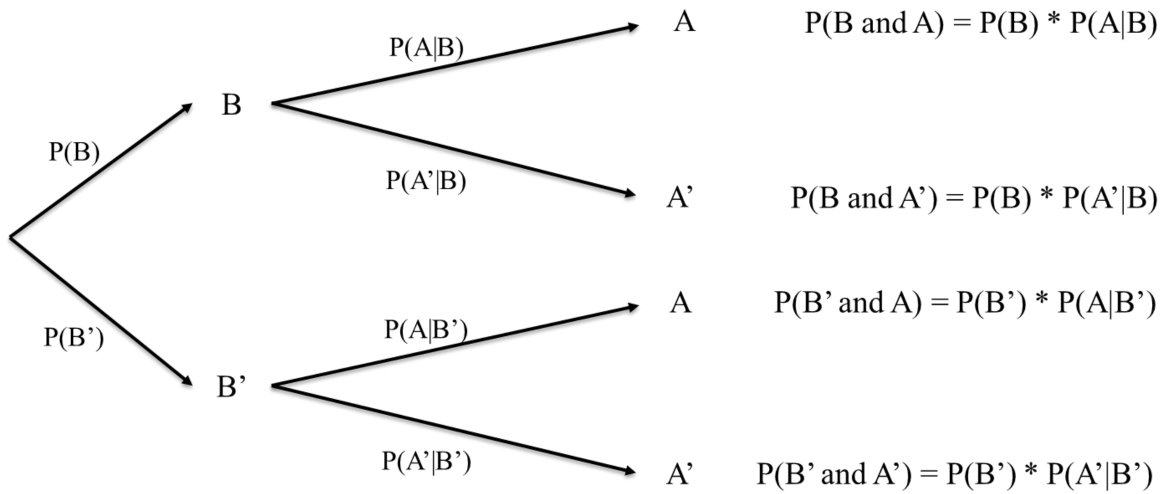
P(B and A') = P(B) * P(A'|B)

P(B' and A) = P(B') * P(A|B')

P(B' and A') = P(B') * P(A'|B')

State each probability in English; note B' means "not B".

Conditional Probability → Bayes Theorem



$$P(B|A) = \frac{P(B) * P(A|B)}{P(A)} = \frac{P(A|B) * P(B)}{P(A|B) * P(B) + P(A|not\,B\,) * P(not\ B)}$$

Note B' means "not B"

Bayes' Theorem

Case – Clinical trials

Epidemiologists claim that probability of breast cancer among Caucasian women in their mid-50s is 0.005. An established test identified people who had breast cancer and those that were healthy. A new mammography test in clinical trials has a probability of 0.85 for detecting cancer correctly. In women without breast cancer, it has a chance of 0.925 for a negative result. If a 55-year-old Caucasian woman tests positive for breast cancer, what is the probability that she in fact has breast cancer?

Case – Clinical trails

P(Cancer) = 0.005

P(Test positive | Cancer) = 0.85

P(Test negative | No Cancer) = 0.925

P(Cancer | Test positive) = ?

$$P(cancer \mid +ve) = \frac{P(cancer) * P(+ve \mid cancer)}{P(+ve \mid cancer) * P(cancer) + P(+ve \mid no\ cancer) * P(no\ cancer)}$$

$$= \frac{0.005 * 0.85}{0.85 * 0.005 + 0.075 * 0.995}$$

$$= 0.054$$

Draw a probability table and a Probability tree for the above case.

Bayes' Theorem ⇒Spam filtering

Apache SpamAssassin™

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word "free" appears in 20% of the mails marked as spam, i.e., P(Free | Spam) = 0.20. Assuming 0.1% of non-spam mail includes the word "free" and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word "free" appears in it.

Bayes' Theorem

P(Spam) = 0.50

P(Free | Spam) = 0.20 (*aka* Prior Probability)

P(Free | No spam) = 0.001

P(Spam | Free) = ? (*aka* Posterior or Revised Probability)

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free|Spam) * P(Spam) + P(Free|No\ spam) * P(No\ spam)}$$

$$= \frac{0.5 * 0.2}{0.2 * 0.5 + 0.001 * 0.5} = \frac{0.1}{0.1005} = 0.995$$

This helps the spam filter automatically classify the messages a
spam.

How Good is Your Classification

Confusion Matrix

Spam filtering		Predicted		Total
		Positive	Negative	
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Negative	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

Spam filtering		Predicted Positive	Predicted Negative	Total
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

$$Recall\ (sensitivity) = \frac{952}{1478} = 0.644$$

$$\text{Precision} = \frac{952}{1119} = 0.851$$

$$Accuracy \ = \frac{952 + 3025}{952 + 3025 + 526 + 167} = \frac{3977}{4670} = 0.852$$

$$Spcificity = \frac{3025}{3025 + 167} = 0.948$$

$$F_1 = 2 * \frac{Precision * Recall}{Preceision + Recall} = \frac{2 * 0.851 * 0.644}{0.851 + 0.644} = 0.733$$

Which measure(s) is/are more important?

Confusion Matrix

Cancer		Predicted: Positive	Predicted: Negative	Total
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

$$Recall\ (sensitivity) = \frac{952}{1478} = 0.644$$

$$Precesion = \frac{952}{1119} = 0.851$$

$$Accuracy\ = \frac{952 + 3025}{952 + 3025 + 526 + 167} = \frac{3977}{4670} = 0.852$$

$$Spcificity = \frac{3025}{3025 + 167} = 0.948$$

$$F_1 = 2\ * \frac{Precision * Recall}{Preceision + Recall} = \frac{2\ * 0.851\ * 0.644}{0.851 + 0.644} = 0.733$$

Which measure(s) is/are more important?

Confusion Matrix – Interview Question

You have been tasked to build a classifier for cancer diagnosis. It is of high importance that patients without cancer should RARELY be diagnosed as positive (even if some patients with cancer are diagnosed wrongly as negative).

Which of the following classification models would you prefer?

(Assuming: Positive = Cancer present, Negative = Cancer absent)

Options:

- True Positive Rate [which is = True Positive / Actual Positive]
- True Negative Rate [which is = True Negative / Actual Negative]
- Precision [which is = True Positive / Predicted Positive]
- Total Accuracy [which is = (True Positive + True Negative) / Total Population]