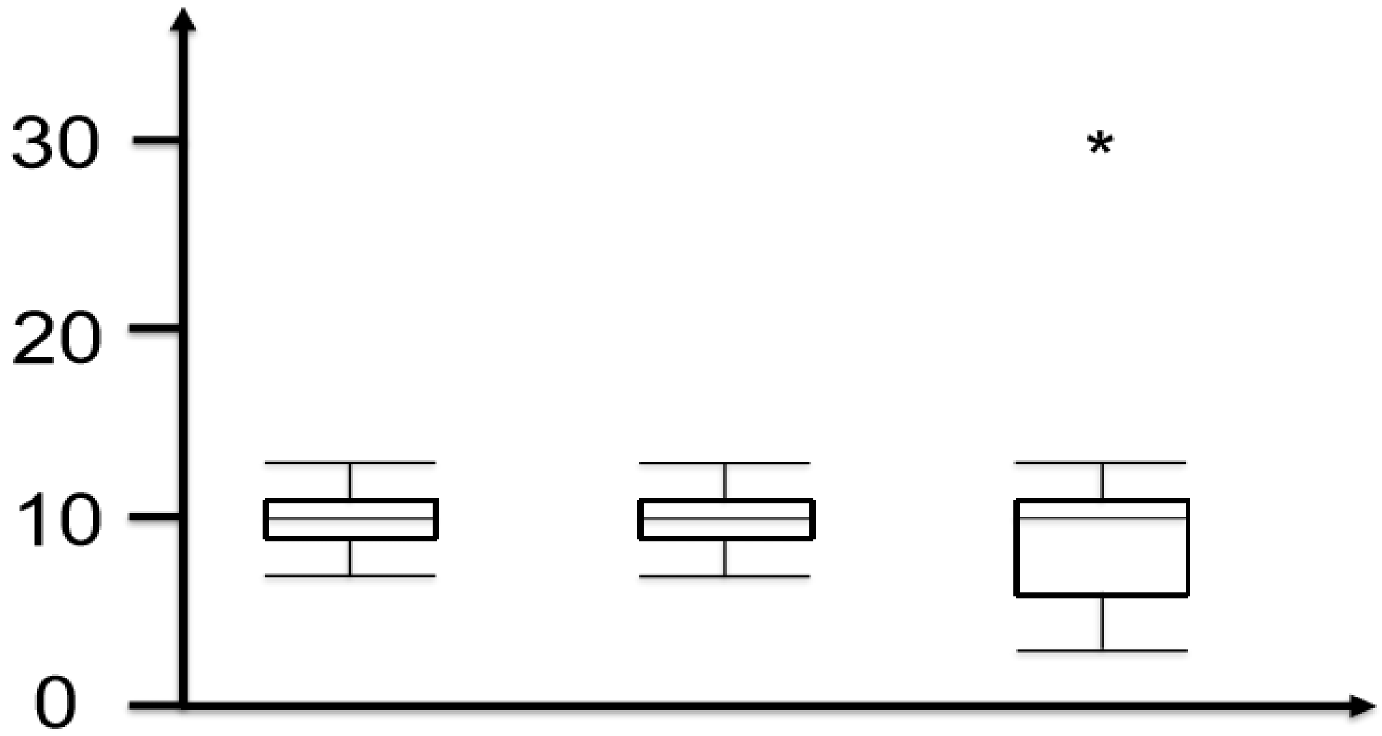
FOUNDATION OF PROBABILITY

Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

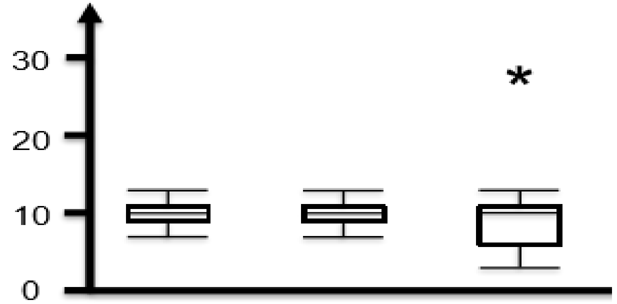
Box and whisker diagram or Box plot



Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

Box and whisker diagram or Box plot



Tukey fences

Name	Formula	Player 1	Player 2	Player 3
Upper Hinge	75th Percentile	11	11	11
Lower Hinge	25th Percentile	9	9	6
H-Spread	Upper Hinge - Lower Hinge (IQR)	2	2	5
Step	1.5 x H-Spread (1.5*IQR)	3	3	7.5
Upper Inner Fence	Upper Hinge + 1 Step (75th percentile + 1.5*IQR)	14	14	18.5
Lower Inner Fence	Lower Hinge - 1 Step (25th percentile - 1.5*IQR)	6	6	-1.5
Upper Outer Fence	Upper Hinge + 2 Steps (75th percentile + 3*IQR)	17	17	26
Lower Outer Fence	Lower Hinge - 2 Steps (25th percentile - 3*IQR)	3	3	-9
Upper Adjacent	Largest value below Upper Inner Fence	13	13	13
Lower Adjacent	Smallest value above Lower Inner Fence	7	7	3
Outside Value (Outliers)	A value beyond an Inner Fence but not beyond an Outer Fence			
Far Out Value (Extreme Values)	A value beyond an Outer Fence			30

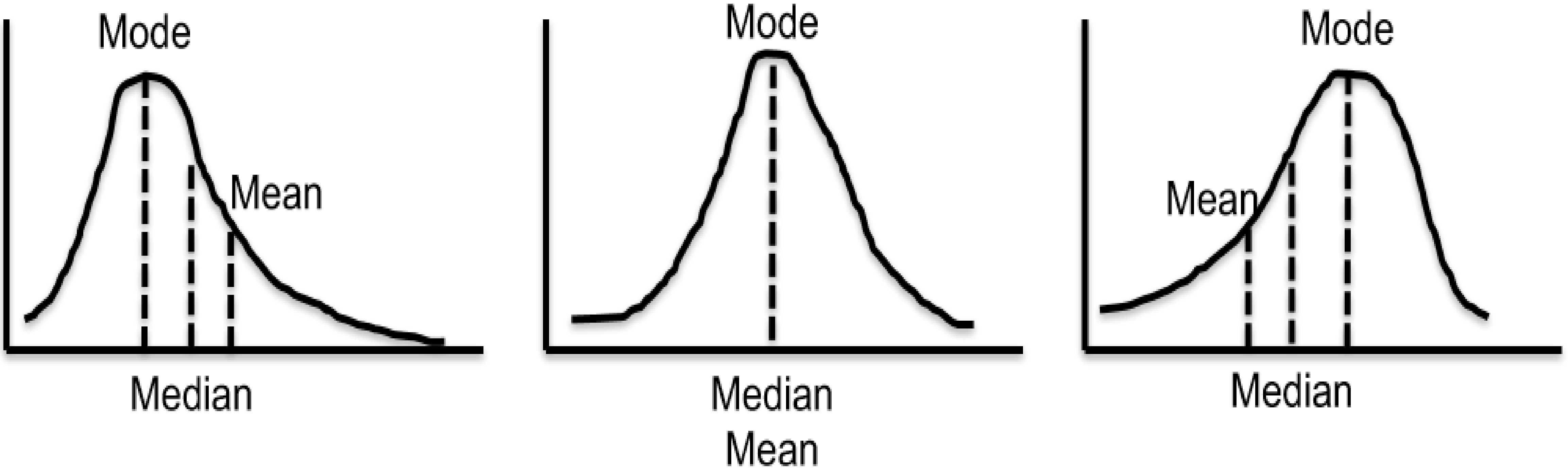
Data Types – Recent Interview Question

A sample of 400 Bangalore households is selected and several variables are recorded. Which of the following statements is correct?

- Socioeconomic status (recorded as "low income", "middle income", or "high income") is nominal level data
- The number of people living in a household is a discrete variable
- The primary language spoken in the household is ordinal level data (recorded as "Kannada", "Tamil", etc)

The Central Tendencies

Identify where the MODE, MEDIAN and MEAN lie in the below distributions.



Measures of Spread – Recent Interview Question

The spread of the data in a dataset could be studied using ________

- Interquartile range
- Variance
- Standard Deviation
- Range (max-min)
- All of the above

Measures of Spread – Recent Interview Question

Given the numbers are 68, 83, 58, 84, 100, 64, the second quartile is:

- 74.5
- 75.5
- 75
- 74

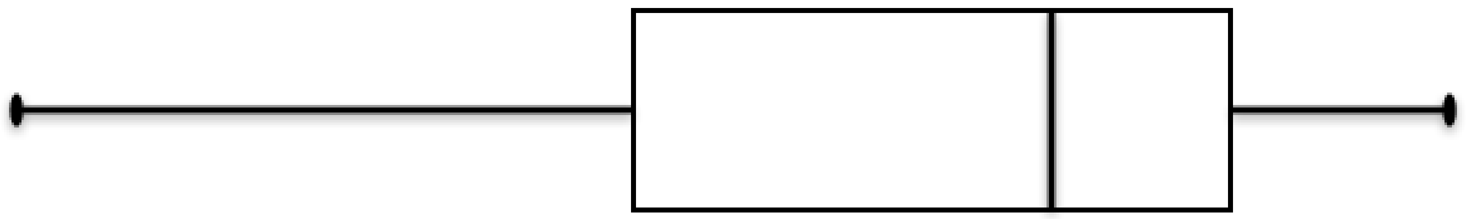
Measures of Spread – Recent Interview Question

Which of the following plot is used to analyze interquartile range

- Scatterplot
- Histogram
- Lineplot
- Boxplot
- All of the above

Measures of Spread – Recent Interview Question

What term would best describe the shape of the given boxplot?



- Symmetric
- Skewed with right tail
- Skewed with left tail
- Normal

Measures of Spread (Dispersion)

Just as Quartiles divide data into 4 equal parts, Deciles divide it into 10 equal parts and Percentiles into 100 equal parts.

Given the above, find the 25^{th}, 50^{th}, 75^{th} and the 90^{th} percentiles for the top 16 global marketing sectors for advertising spending for a recent year according to *Advertising Age*. Also, find Q2, 5^{th} decile and IQR. Data in next slide.

Sector	Ad spending (in $ million)
Automotive	22195
Personal Care	19526
Entertainment and Media	9538
Food	7793
Drugs	7707
Electronics	4023
Soft Drinks	3916
Retail	3576
Restaurants	3553
Cleaners	3571
Computers	3247
Telephone	2448
Financial	2433
Beer, Wine and Liquor	2050
Candy	1137
Toys	699

PROBABILITY BASICS

Probability vs Statistics

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

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

Probability vs Statistics

Probability is the basis of inferential statistics.

Probability -Applications

Gaming industry –Establish charges and payoffs

HR –Does a company have biased hiring policies?

Manufacturing/Aerospace –Prevent major breakdowns

Assigning Probabilities

Classical Method – *A priori* or Theoretical

Probability can be determined prior to conducting any experiment.

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

Example: Tossing of a fair die

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Probability can be determined post conducting a thought experiment.

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

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 rain tomorrow

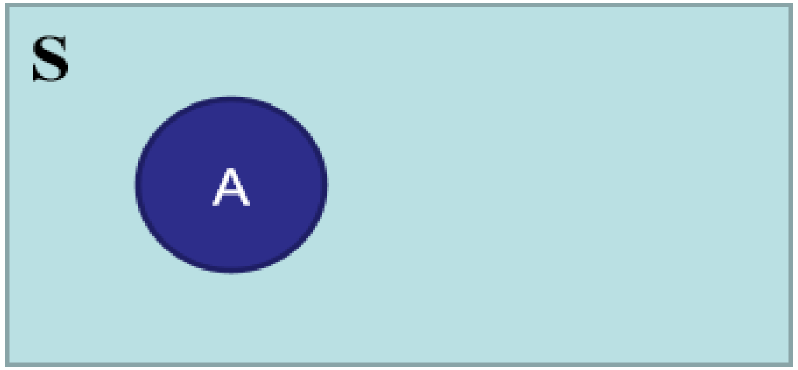
Probability -Terminology

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

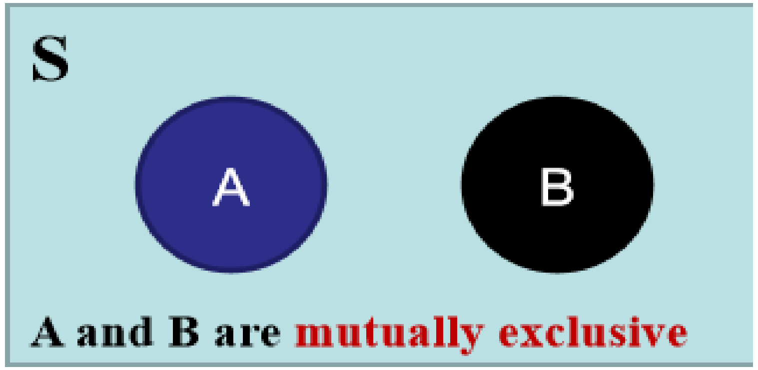
Event –A subset of the sample space

Probability - Rules

$$P(S) = 1$$



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

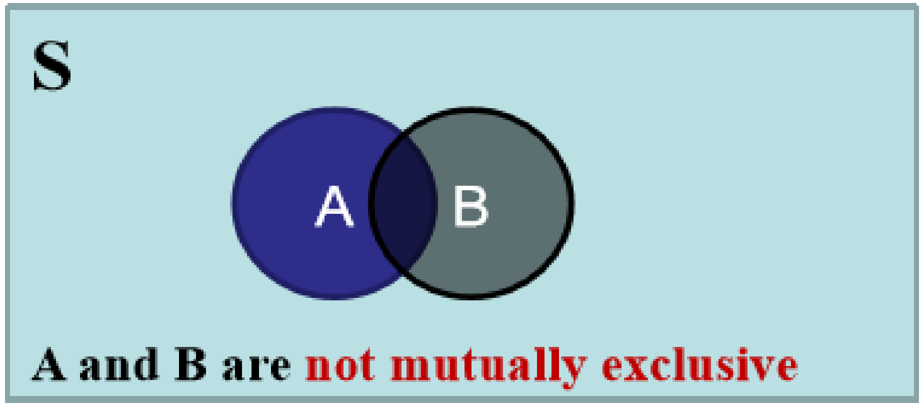


$$P(A \text{ or } B) = P(A) + P(B)$$

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.

Probability - Rules



$$P(A \text{ } or \text{ } B) = P(A) + P(B) - P(A \text{ } and \text{ } B)$$

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 - Types

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

		Age			
		Young	**Middle-aged**	**Old**	**Total**
Loan Default	**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 Default	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

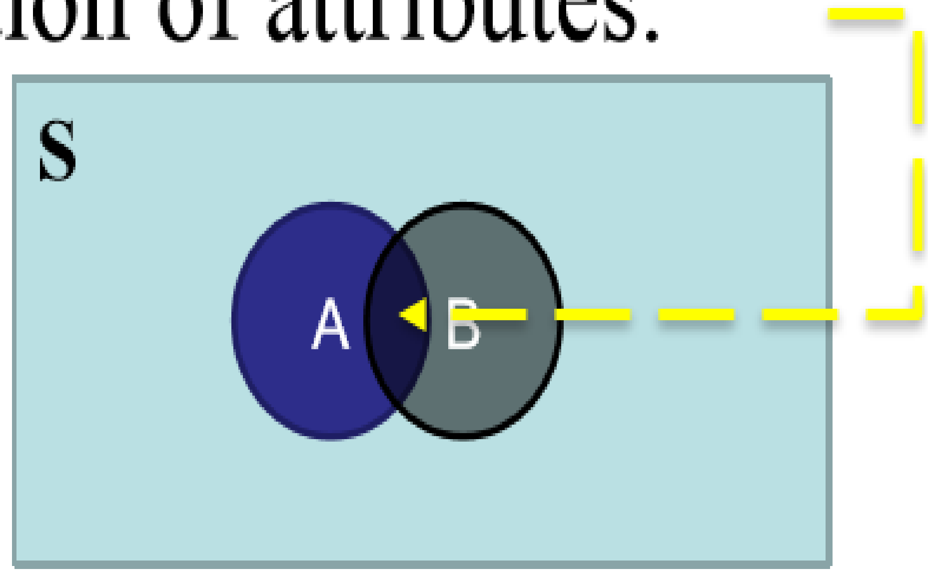
Probability - Types

Joint Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	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 describing a combination of attributes.

P(Yes **and** Young) = 0.077

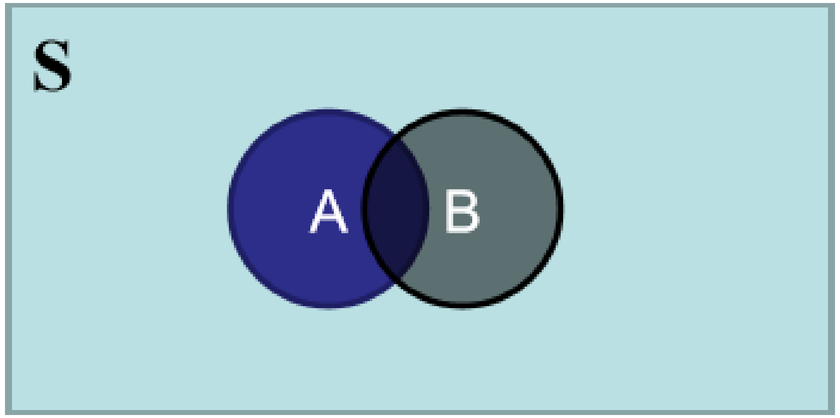


Probability - Types

Union Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	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** Young) = P(Yes) + P (Young) – P(Yes and Young) = 0.184 + 0.302 – 0.077 = 0.409



Probability - Types

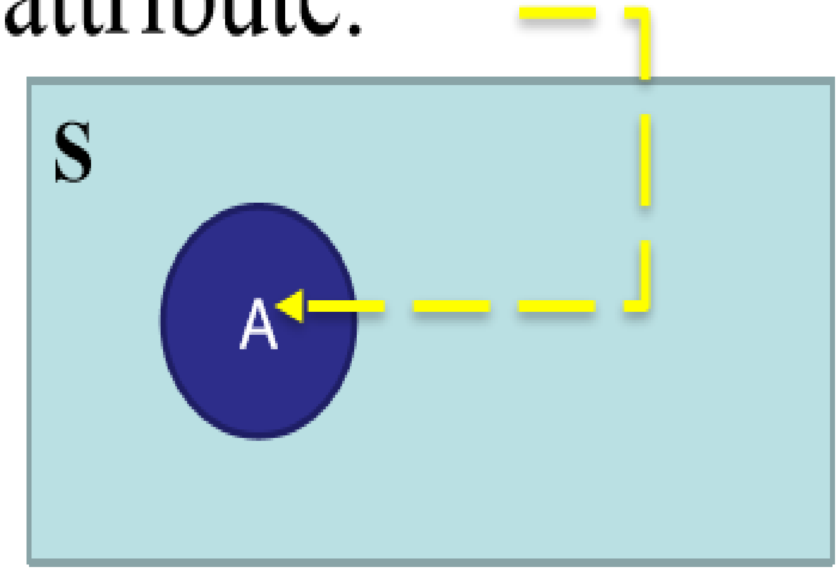
Marginal Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	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 describing a single attribute.

P(No) = 0.816

P(Old) = 0.008



Probability - Types

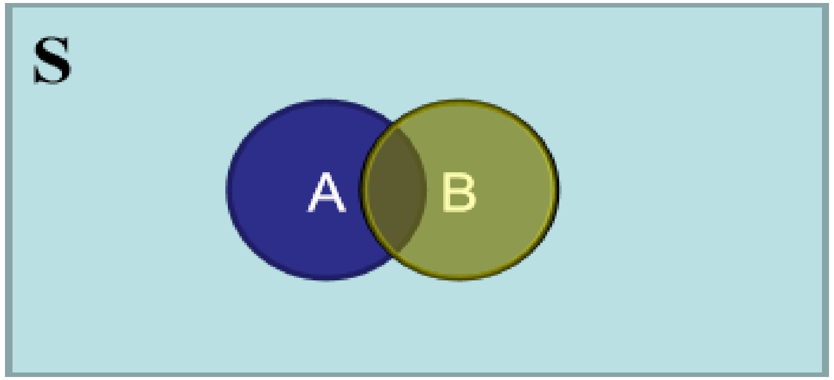
Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	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 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 - Types

Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	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?

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

Note that this is the ratio of **Joint Probability** to **Marginal Probability**, i.e., $P(A|B) = \dfrac{P(A \text{ and } B)}{P(B)}$

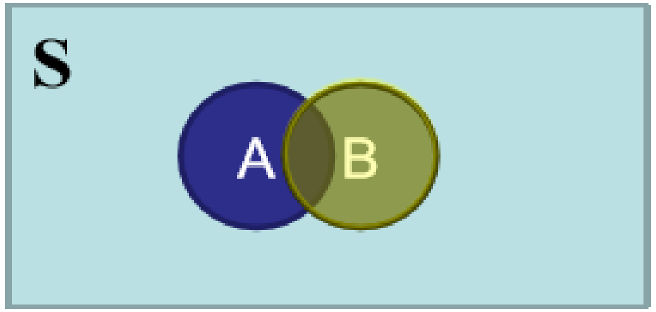
P(Middle-Aged | No) = 0.586/0.816 = 0.72 (<u>Order Matters</u>)

Probability - Types

Conditional Probability – Visualizing using Probability Tables and Venn Diagrams

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

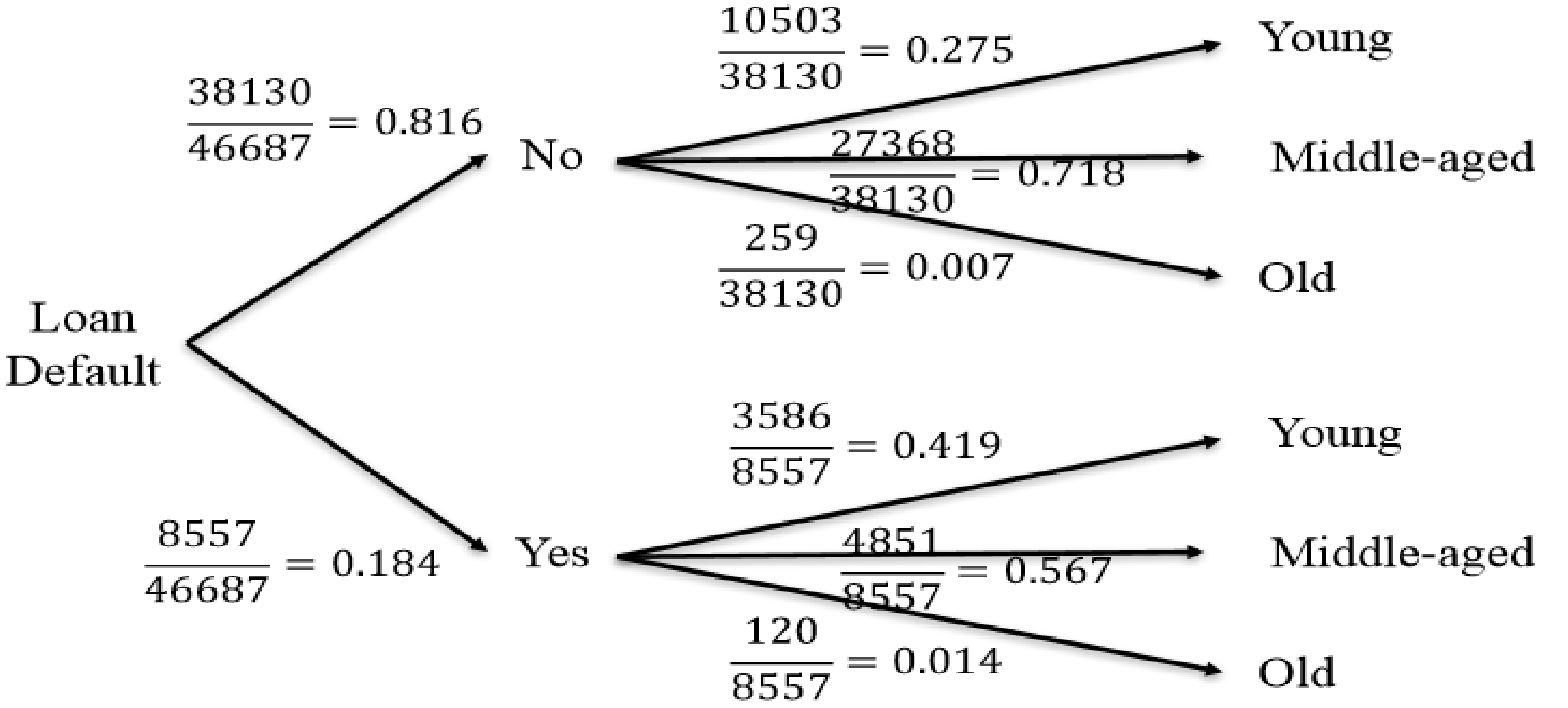
		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000



Probability - Types

Conditional Probability – Visualizing using Probability Trees

		Age (Numbers)				Age (Probabilities)			
		Young	Middle-aged	Old	Total	Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130	0.225	0.586	0.005	0.816
	Yes	3,586	4,851	120	8,557	0.077	0.104	0.003	0.184
	Total	14,089	32,219	379	46,687	0.302	0.690	0.008	1.000



Find

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

Probability - Types

Attention Check

Identify the type of probability in each of the below cases:

1. P(Old and Yes)
2. P(Yes and Old)
3. P(Old)
4. P(Yes)
5. P(Old | Yes)
6. P(Yes | Old)
7. P(Young | No)
8. P(Middle-aged or No)
9. P(Old or Young)

		Age (Probabilities)			
		Young	**Middle-aged**	**Old**	**Total**
Loan Default	**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**

1 and 2: **Joint**; 3 and 4: **Marginal**; 5, 6 and 7: **Conditional**; 8 and 9: **Union**

Probability - Types

Conditional Probability

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

Similarly

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

Equating, we get

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

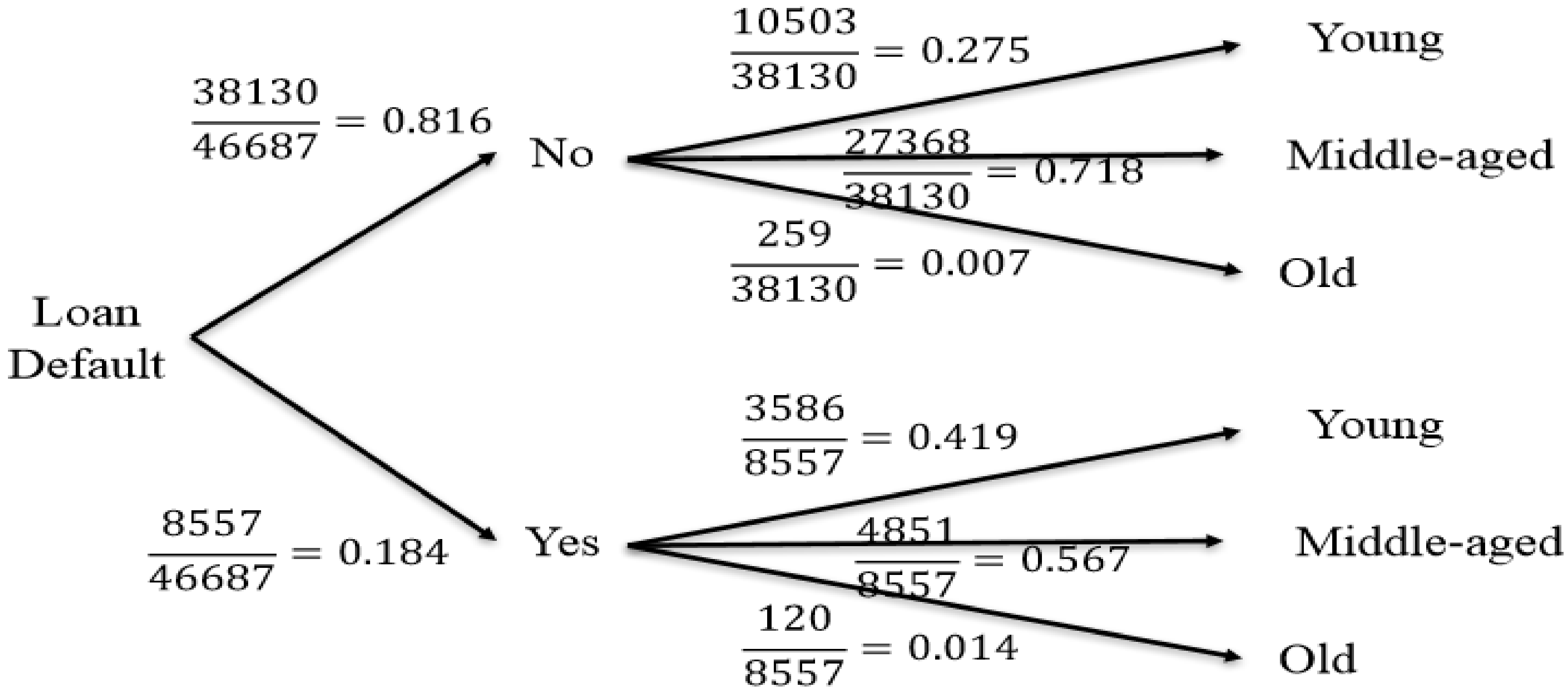
$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

Probability - Types

Conditional Probability – Visualizing using Probability Trees

		Age (Probabilities)			
		Young	**Middle-aged**	**Old**	**Total**
Loan Default	**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

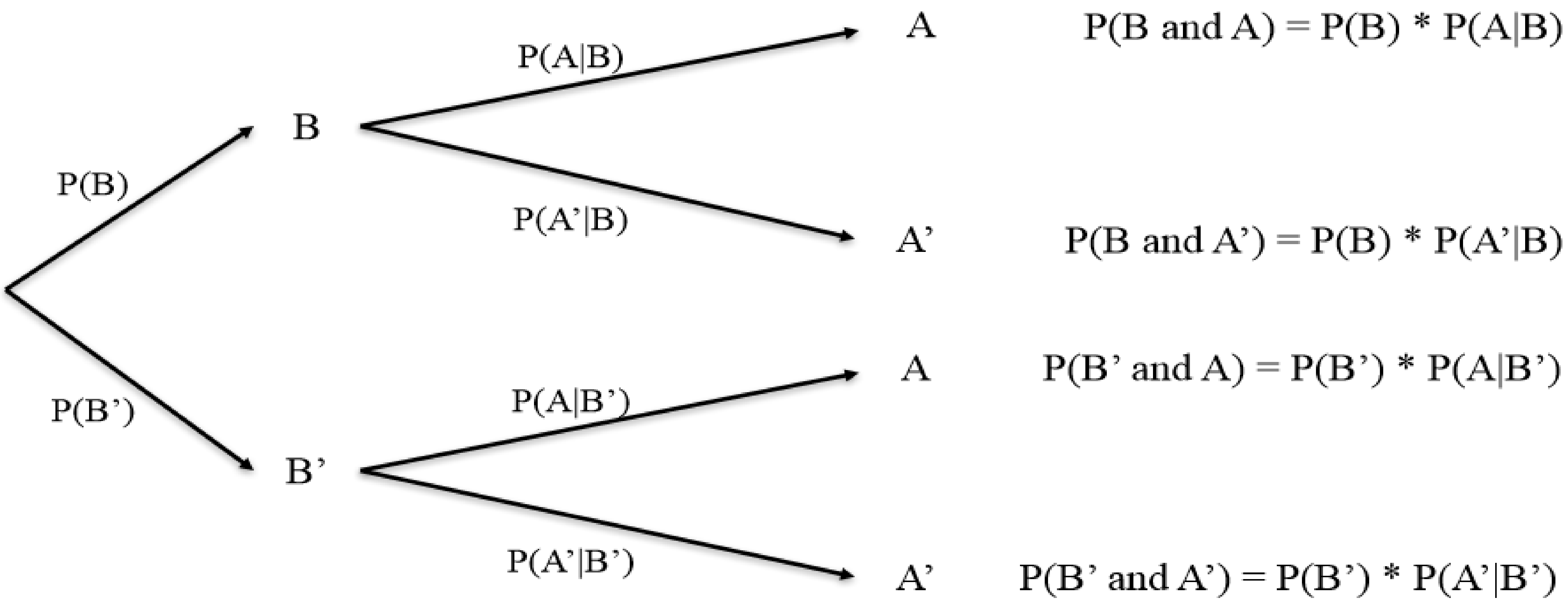
$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$



Now find P(Yes | Old)

Probability - Types

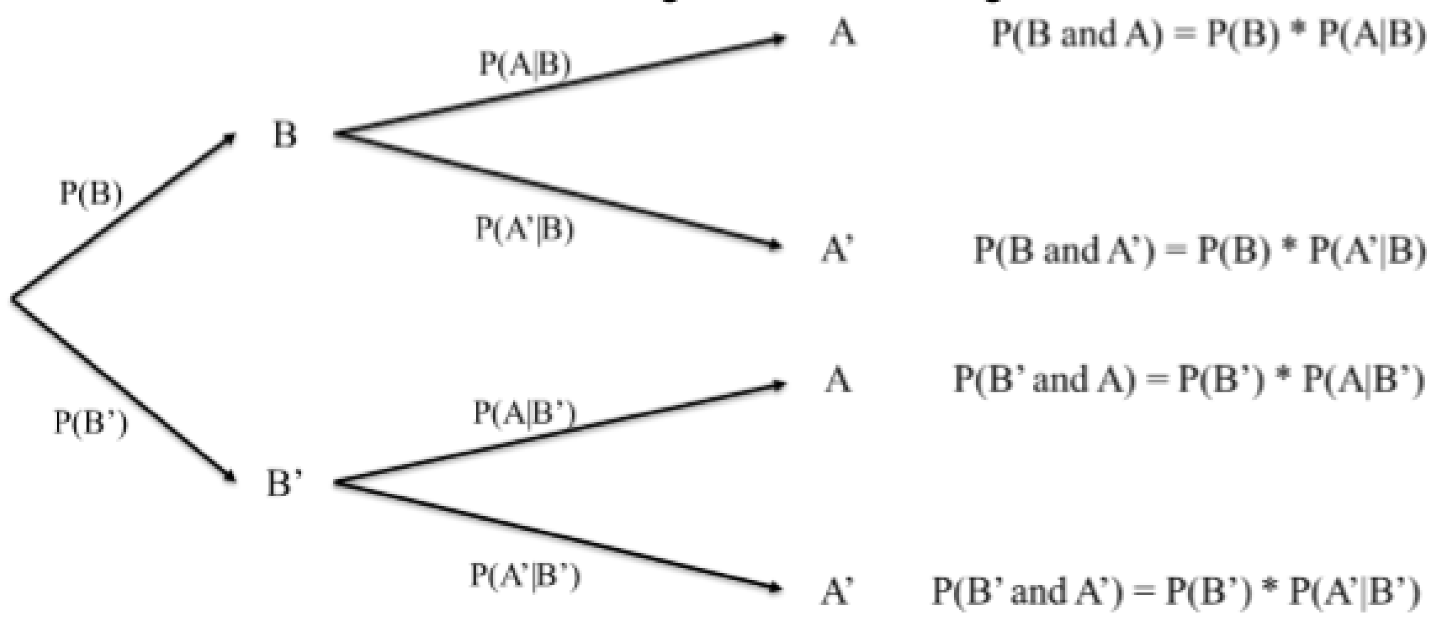
Generalized Probability Tree



State each probability in English; note B’ means “not B”.

Probability - Types

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

Bayes' Theorem allows you to find reverse probabilities, and to allow **revision of original probabilities** with new information.

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?

Bayes' Theorem

Case – Clinical trials

P(Cancer) = 0.005

P(Test positive | Cancer) = 0.85 (*aka* Prior Probability)

P(Test negative | No cancer) = 0.925

P(Cancer | Test positive) = ? (*aka* Posterior or Revised Probability)

$$P(Cancer|Test +) = \frac{P(Cancer) * P(Test + |Cancer)}{P(Test + |Cancer) * P(Cancer) + P(Test + |No\ cancer) * P(No\ cancer)}$$

$$= \frac{0.005 * 0.85}{0.85 * 0.005 + 0.075 * 0.995} = \frac{0.00425}{0.078875} = 0.054$$

Homework

Draw a Probability Table and a Probability Tree for the above case.

Bayes' Theorem

Case – Spam filtering

Apache SpamAssassin™

Latest News

2015-04-30: SpamAssassin 3.4.1 has been released! Highlights include:

- improved automation to help combat spammers that are abusing new top level dc
- tweaks to the SPF support to block more spoofed emails;
- increased character set normalization to make rules easier to develop and stop sp
- continued refinement to the native IPv6 support; and
- improved Bayesian classification with better debugging and attachment hashing.

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

Case – Spam filtering

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 as spam.