Fundamentals of Data Analytics
Lecture 01. Probability

Instructional Team

VEF ACADEMY

About this Course

- Probability
- Statistics
- Hands-on programming skills
- Meet your instructors & classmates

Vinh Dang
PhD., Data Scientist
Trusting Social

Thuy Nguyen
B.Sc., Data Analyst
Trusting Social

Sang Nguyen
M.Sc., Data Scientist
FE Credit

Huy Pham
PharmB, R&D Officer
OPC

Welcome to FDA

- ✓ Basic Probability and Statistics
- ✓ Introduction to Python Programming Language
- ✓ Real-Life Case Studies
- ✓ Networking
- ✓ Preparing for Advanced course DA / ML

- ? DA / ML / DS / BI / AI / 4th IR
- ? R Programming / SPSS / …
- ⇒ Discussing with Instructional team & classmates (Piazza…)

Content of Lecture

- Counting Rules
- Sample Space, Event
- Independent Event
- Conditional Probability
- Bayes' Theorem

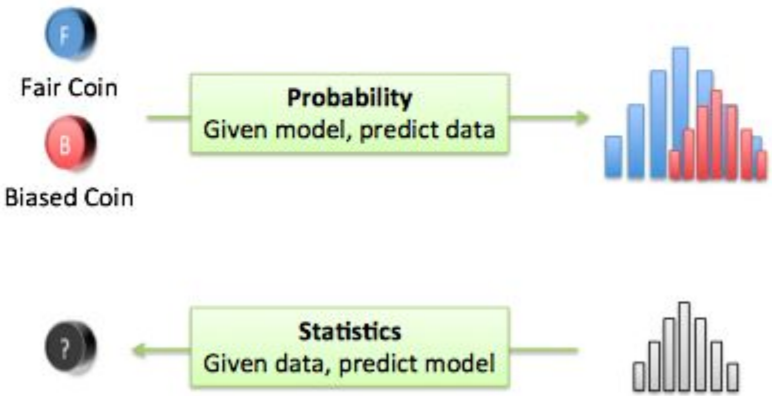
Motivation Example

Blaise Pascal (1623 - 1662)

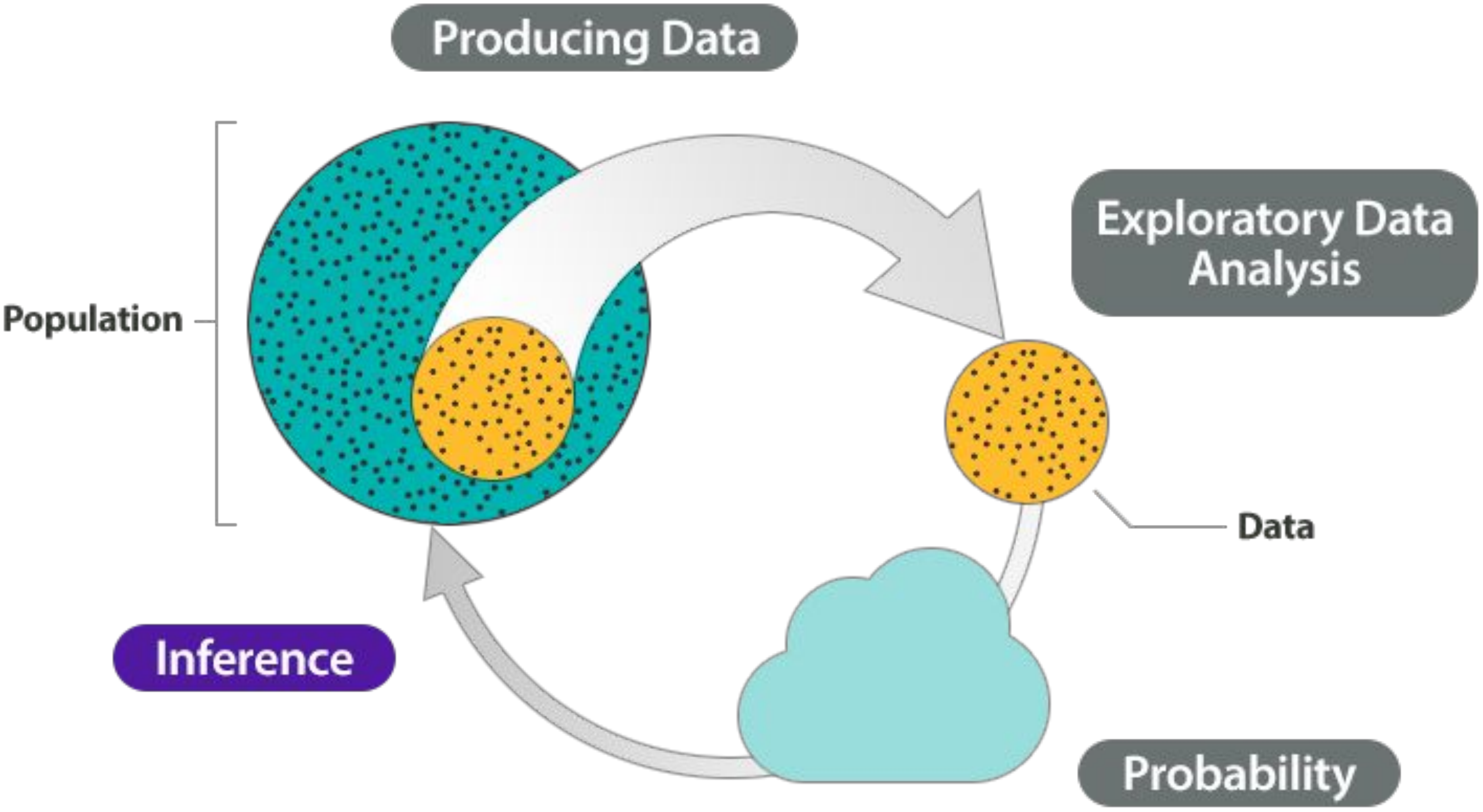
Pierre de Fermat (1601 - 1665)

Who first get to 3 will win the game and take all money.

INTRODUCTION



Probability & Statistics

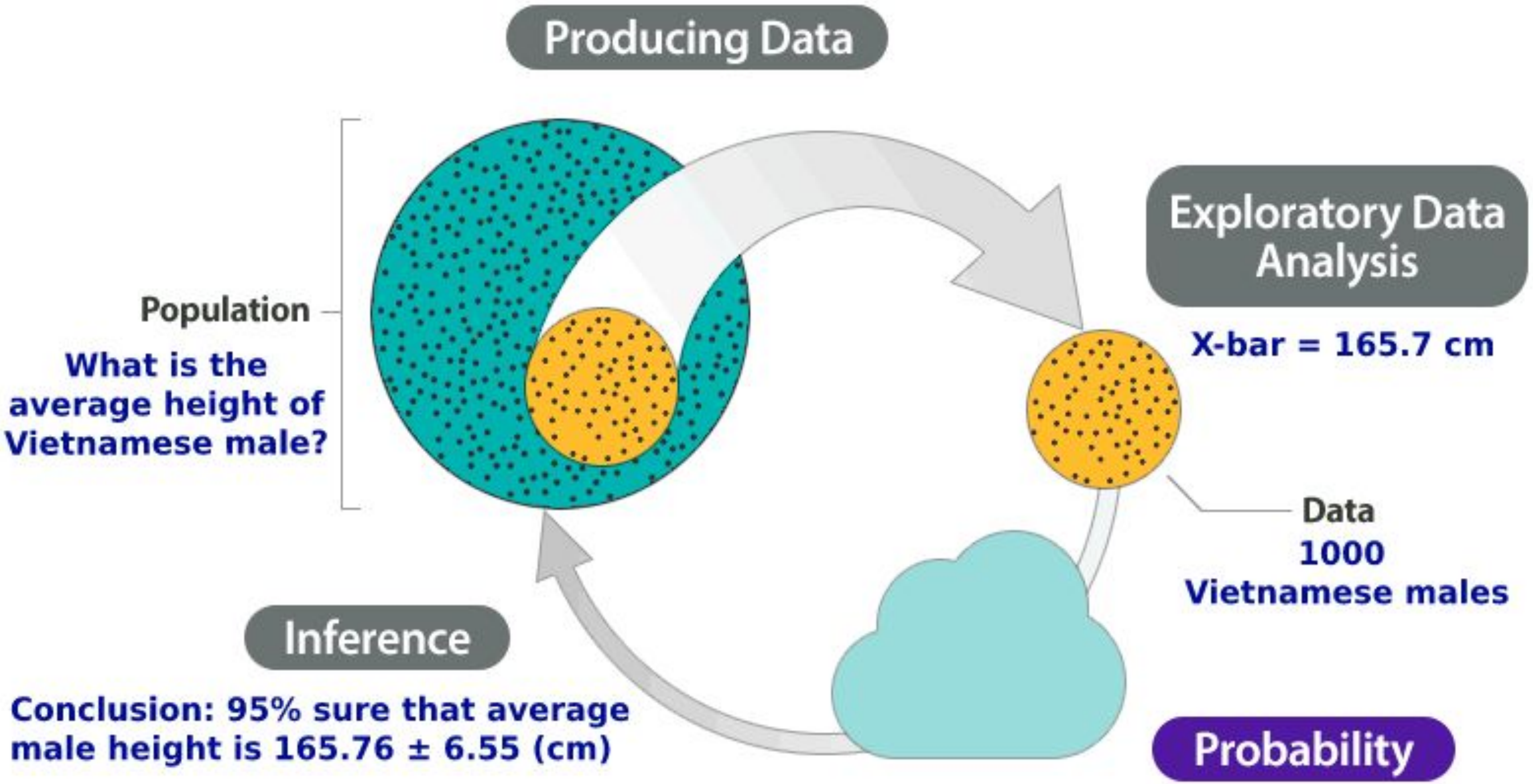


Example

What is average height of Vietnamese males?

1. **Produce Data:** Determine what to measure, then collect the data.
   → Selected 1000 of male adults at random.
   → Measured and collected the height
2. **Explore the Data:** Analyze and summarize the data.
   → In the sample, the average height is 165.7 cm.
3. **Draw a Conclusion:** Use the data, probability, and statistical inference
   → Draw a conclusion about the population.

Probability & Statistics



COUNTING

Counting rules

Rule of counting

Event A can occur in $\mathbf{n_1}$ ways & Event B can occur in $\mathbf{n_2}$ ways
$\Longrightarrow$ Events A and B can occur in $\mathbf{n_1 \times n_2}$ ways.
In general, the number of ways that **m** events can occur is $\mathbf{n_1 \times n_2 \times \ldots \times n_m}$.

Example:
How many unique stock-keeping unit (SKU) labels can a chain of hardware stores create by using **two letters** (ranging from AA to ZZ) followed by **four numbers** (digits 0 through 9)?

Solution:
26 x 26 x 10 x 10 x 10 x 10 = 6,760,000

Counting rules

Factorials

The number of unique ways that **n items** can be arranged in a particular order is **n!**

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$$

Example:
A home appliance service truck must make three stops (A, B, C). In how many ways could the three stops be arranged?

Solution:
$3! = 3 \times 2 \times 1 = 6$
That is {ABC, ACB, BAC, BCA, CAB, CBA}

Counting rules

Permutations

The number of possible **permutations** of **n** items taken **r** in a particular order is

$$nPr = \frac{n!}{(n-r)!}$$

Example:

Five home appliance customers (A, B, C, D, E) need service calls, but the field technician can service only three of them before noon. The order in which they are serviced is important (to the customers, anyway) so each possible arrangement of three service calls is different. The dispatcher must assign the sequence. How many possible permutation?

Solution:

$$nPr = \frac{5!}{(5-3)!} = \frac{120}{2} = 60$$

Counting rules

Combinations

A **combination** is a collection of **r items** chosen at random without replacement **from n items** where the order of the selected items is not important.

The number of possible combinations of **r** items chosen from **n** items is

$$nCr = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

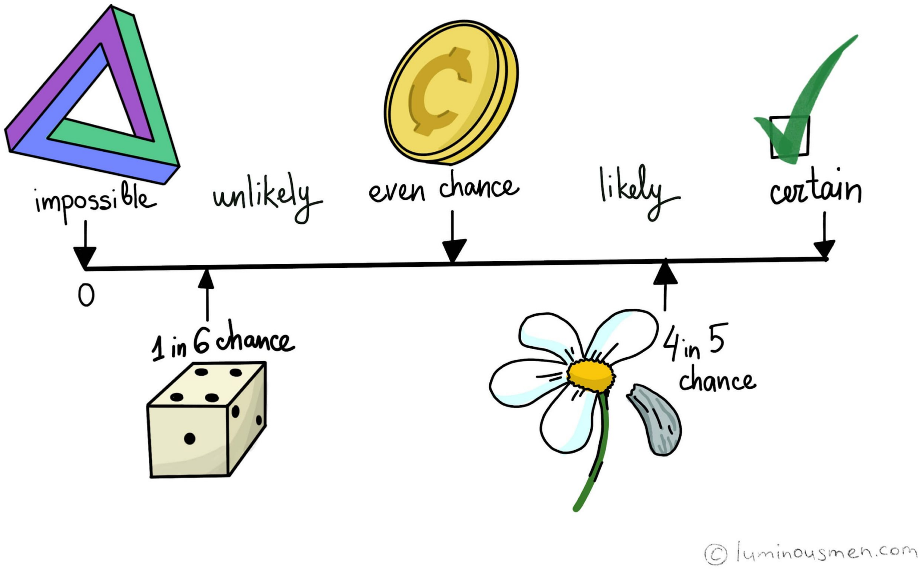
Example:

Suppose that five customers (A, B, C, D, E) need service calls and the maintenance worker can only service three of them this morning. The customers don't care when they are serviced as long as it's before noon, so the dispatcher does not care who is serviced first, second, or third. How many possible combinations?

Solutions:

$$nCr = \binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{120}{12} = 10$$

PROBABILITY



Sample Spaces & Events

Definition

The **Sample spaces S** is the set of possible outcomes of an experiment

Sample outcomes / Realizations are the points ω in the **Sample spaces**

Events (E) are subsets of **Sample spaces**

Example:

- If we toss a coin twice then **S = {HH, HT, TH, TT}**
  Event that the 1st coin is heads is **A = {HH, HT}**

- If we toss a coin forever then the **S** is the infinite set

  $$S = \{\omega = (\omega_1, \omega_2, \omega_3, \dots), \omega_i \in \{H,T\}\}$$

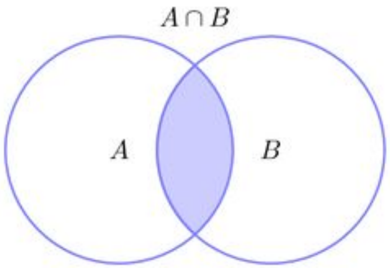
  Event that first head appears on the third toss

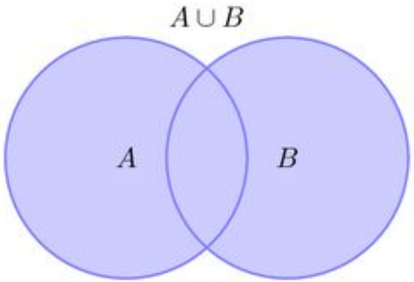
  $$E = \{(\omega_1, \omega_2, \omega_3, \dots): \omega_1 = T, \omega_2 = T, \omega_3 = H, \omega_i \in \{H,T\} \text{ for } i > 3\}$$

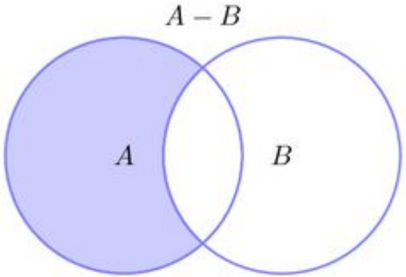
$$E \subseteq S$$

Sample Spaces & Events

S	Sample space
ω	Outcome
A or E,...	Event (Subset of S)
|A|	number of points in A (if A is finite)
A^C	Complement of A (not A)
A ∪ B	Union of A and B
A ∩ B	Intersection of A and B
A - B	Set difference (points in A that are not in B)
A ⊂ B	Set inclusion
∅	Null Event







Sample Spaces & Events

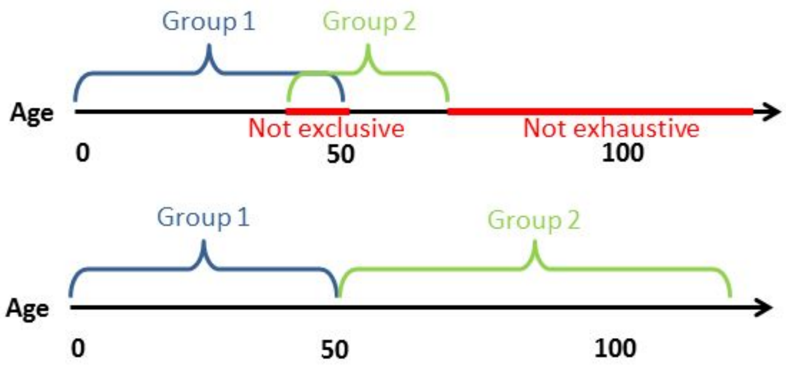
Mutually Exclusive Events

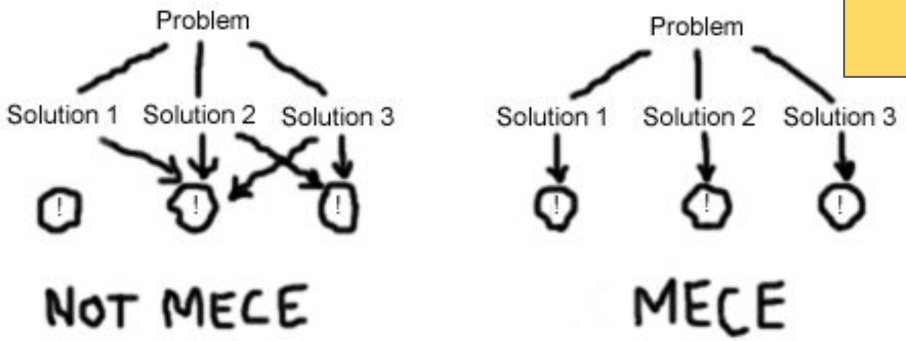
$A_1, A_2, \ldots$ are **disjoint** or are **mutually exclusive** if $A_i \cap A_j = \varnothing$ whenever $i \neq j$

Collectively Exhaustive Events

$A_1, A_2, \ldots$ are **collectively exhaustive** if $\cup_{i=1}^{\infty} A_i = S$

MECE?





Probability

Probability

The **probability** of an event is a number that measures the relative likelihood that the event will occur.

Axioms of Probability

- $P(A) \geq 0$ for every A
- $P(S) = 1$ (S is **Sample space)**
- If A_1, A_2, … are **disjoint/mutually exclusive** then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Views of Probability

Approach	How Assigned?	Example
Empirical	Estimated from observed outcome frequency	There is a 2 percent chance of twins in a randomly chosen birth.
Classical	Known a *priori* by the nature of the experiment	There is a 50 percent chance of heads on a coin flip.
Subjective	Based on informed opinion or judgment	There is a 60 percent chance that Toronto will bid for the 2024 Winter Olympics.

How Assigned? ⟶ Empirical Approach

Empirical approach

- Collecting empirical data through observations or experiments
- The number of observations is **n**
- The frequency of observed outcomes is **f**

$\Rightarrow$ The estimated probability is **f/n**

Example:

An company interviewed 280 production workers before hiring 70 of them.

Let H = event that a randomly chosen interviewee is hired $\Rightarrow P(H) = f/n = 70/280 = 0.25$

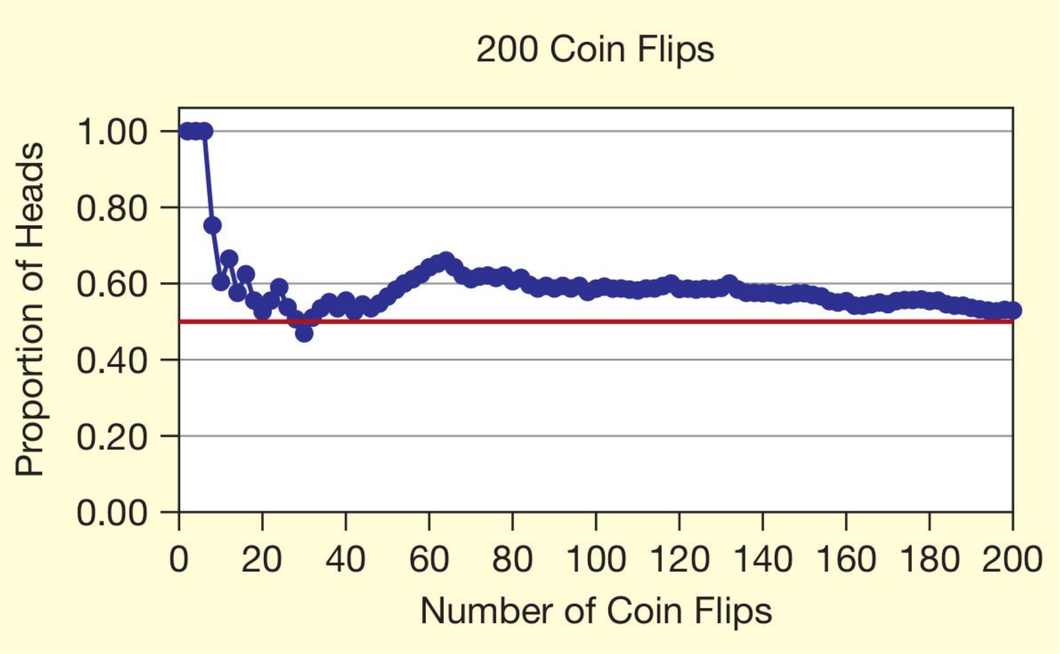
Law of LARGE number

As the number of trials increases, any empirical probability approaches its theoretical limit.

How Assigned? → Empirical Approach

Law of LARGE number

As the number of trials increases, any empirical probability approaches its theoretical limit.



How Assigned? → Empirical Approach

CASE STUDY: Practical Actuaries Issues

Actuaries help companies calculate payout rates on life insurance, pension plans, and health care plans by estimating the empirical probabilities

Actuaries created the tables that guide IRA withdrawal rates for individuals from age 70 to 99. Here are a few challenges that actuaries face:

1. Is n “large enough” to say that f/n has become a good approximation to the probability of the event of interest? (Data collection costs money, and decisions must be made)
2. Was the experiment repeated identically? (Subtle variations may exist in the experimental conditions and data collection procedures)
3. Is the underlying process stable over time? (For example, default rates on 2007 student loans may not apply in 2017, due to changes in attitudes and interest rates)

How Assigned? → Classical Approach

Classical approach

In classical approach, we do not actually have to perform an experiment because the nature of the process allows us to envision the entire sample space.
→ We can use deduction to determine P(A).

Example:
In the two-dice experiment, there are 36 possible outcomes.

H = rolling a seven

$$P(H) = \frac{\text{number of possible outcomes with 7 dots}}{\text{number of outcomes in sample space}} = \frac{6}{36} = 0.167$$

A priori: the process of assigning probabilities before we actually observe the event or try an experiment

How Assigned? → Subjective Approach

Subjective approach

A subjective probability reflects someone's informed judgment about the likelihood of an event when there is no repeatable random experiment.

Example:

- What is the probability that a new truck product program will show a return on investment of at least 10 percent?
- What is the probability that the price of Ford's stock will rise within the next 30 days?

Notes:

In such cases, we rely on personal judgment or expert opinion. However, such a judgment is not random because it is typically based on experience with similar events and knowledge of the underlying causal processes.

Interpretations of Probability

"Frequencies" approach	"Degrees of beliefs" approach
P(A) is the long run proportion of times that A is true in repetitions.	P(A) measures an observer's strength of belief that A is true, or uncertainty of A

E.g: The probability that a coin will land heads is 0.5

If we flip the coin many times, we expect it to land heads about half the time.	The coin is equally likely to land heads or tails on the next toss

Properties of Probability

Properties of Probability

- $P(\varnothing) = 0$
- $A \subset B \Rightarrow P(A) \leq P(B)$
- $0 \leq P(A) \leq 1$
- $P(A^C) = 1 - P(A)$
- $A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Independent Events

Definition

Two events A and B are independent if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$$

A set of events $\{A_i\}$ is independent if

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

Example: Tossing a fair dice

Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\} \Rightarrow A \cap B = \{2, 4\}$

$P(AB) = 2/6 = 1/3$ and $P(A)\,P(B) = (1/2) \times (2/3) = 1/3$

$\Rightarrow P(AB) = P(A)\,P(B) \Rightarrow$ A and B are independent

Contingency Tables

Definition

A **contingency table** is a cross-tabulation of frequencies into rows and columns.

Example: Tuition cost versus five-year net salary gains for MBA degree recipients at 67 top-tier graduate schools of business

	Salary Gain			
Tuition	*Small (S_1)* *Under \$50K*	*Medium (S_2)* *\$50K–\$100K*	*Large (S_3)* *\$100K+*	*Row Total*
Low (T_1) *Under \$40K*	5	10	1	16
Medium (T_2) *\$40K–\$50K*	7	11	1	19
High (T_3) *\$50K+*	5	12	15	32
Column Total	17	33	17	67

Contingency Tables

Calculation From Contingency Tables

- Marginal Probability
- Joint probability
- Conditional Probability
- Independence
- Relative Frequencies

Contingency Tables

Marginal Probability

The marginal probability of an event is a relative frequency that is found by dividing a row or column total by the total sample size.

	Salary Gain			
Tuition	*Small (S_1)*	*Medium (S_2)*	*Large (S_3)*	*Row Total*
Low (T_1)	5	10	1	16
Medium (T_2)	7	11	1	19
High (T_3)	5	12	15	32
Column Total	17	33	17	67

The marginal probability of a medium salary gain is $P(S_2) = 33/67 = 0.4925$

The marginal probability of low tuition is $P(T_1) = 16/67 = 0.2388$

Contingency Tables

Joint Probability

Each of cells is used to calculate a **joint probability** representing the intersection of two events.

Tuition	Salary Gain: Small (S_1)	Medium (S_2)	Large (S_3)	Row Total
Low (T_1)	5	10	1	16
Medium (T_2)	7	11	1	19
High (T_3)	5	12	15	32
Column Total	17	33	17	67

The joint probability that the school has low tuition (T_1) and has large salary gains (S_3)
is $P(T_1 \cap S_3) = 1/67 = 0.0149$

Contingency Tables

Conditional Probability

Conditional probabilities may be found by restricting ourselves to a single row or column (the condition).

	Salary Gain			
Tuition	*Small (S_1)*	*Medium (S_2)*	*Large (S_3)*	*Row Total*
Low (T_1)	5	10	1	16
Medium (T_2)	7	11	1	19
High (T_3)	5	12	15	32
Column Total	17	33	17	67

The conditional probability that salary gains are small (S_1) given that the MBA tuition is large (T_3) is $P(S_1 \mid T_3) = 5/32 = 0.1563$

Contingency Tables

Independence

To check whether events in a contingency table are independent, we can look at conditional probabilities.

Example: Is large salary gain (S_3) independent of low tuition (T_1) ?

Method 1: No, because

$P(S_3)\,P(T_1) = (17/67)(16/67) = 0.0606$

$P(S_3 \cap T_1) = 1/67 = 0.0149$

$\Rightarrow P(S_3)\,P(T_1) \neq P(S_3 \cap T_1)$

Method 2: No, because

$P(S_3 \mid T_1) = 1/16 = 0.0625 \neq P(S_3) = 17/67 = 0.2537$

Contingency Tables

Relative frequency

To facilitate probability calculations, we can divide each cell frequency f_{ij} by the total sample size to get the relative frequencies f_{ij} / n

	Salary Gains			
Tuition	*Small* (S_1)	*Medium* (S_2)	*Large* (S_3)	*Row Total*
Low (T_1)	.0746	.1493	.0149	.2388
Medium (T_2)	.1045	.1642	.0149	.2836
High (T_3)	.0746	.1791	.2239	.4776
Column Total	.2537	.4926	.2537	1.0000

Contingency Tables

Confusion matrix

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)

Tree Diagrams

Definition

Events and probabilities can be displayed in the form of a **tree diagram** or **decision tree** to help visualize all possible outcomes.

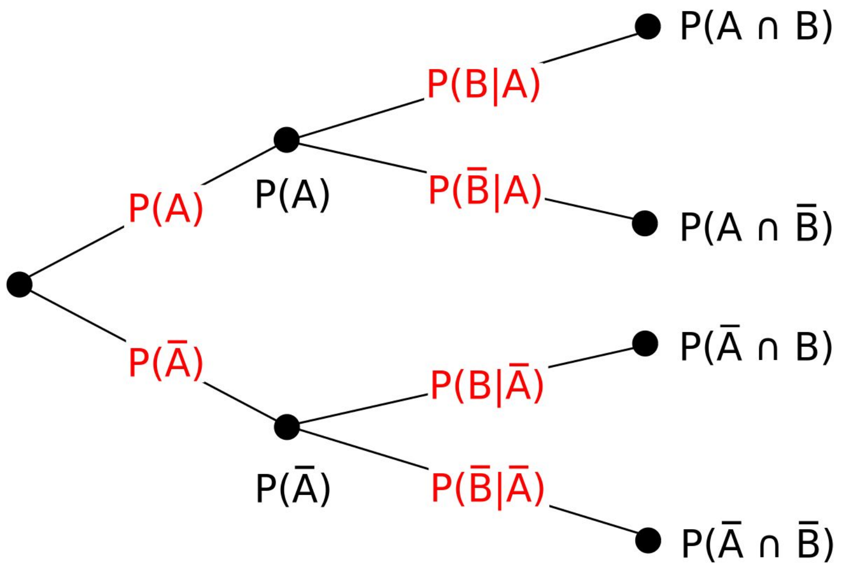
How to build a tree diagram?

(1) Make the Contingency Table

(2) Calculate the *conditional probabilities*.

(3) Calculate the *joint probabilities* from *conditional probabilities*.

$$P(A \cap B) = P(B)P(A \mid B)$$



Tree Diagrams

Step 1. Make the Contingency Table

	Fund Type		
Expense Ratio	*Bond Fund (B)*	*Stock Fund (S)*	*Row Total*
Low (L)	11	3	14
Medium (M)	7	9	16
High (H)	3	11	14
Column Total	21	23	44

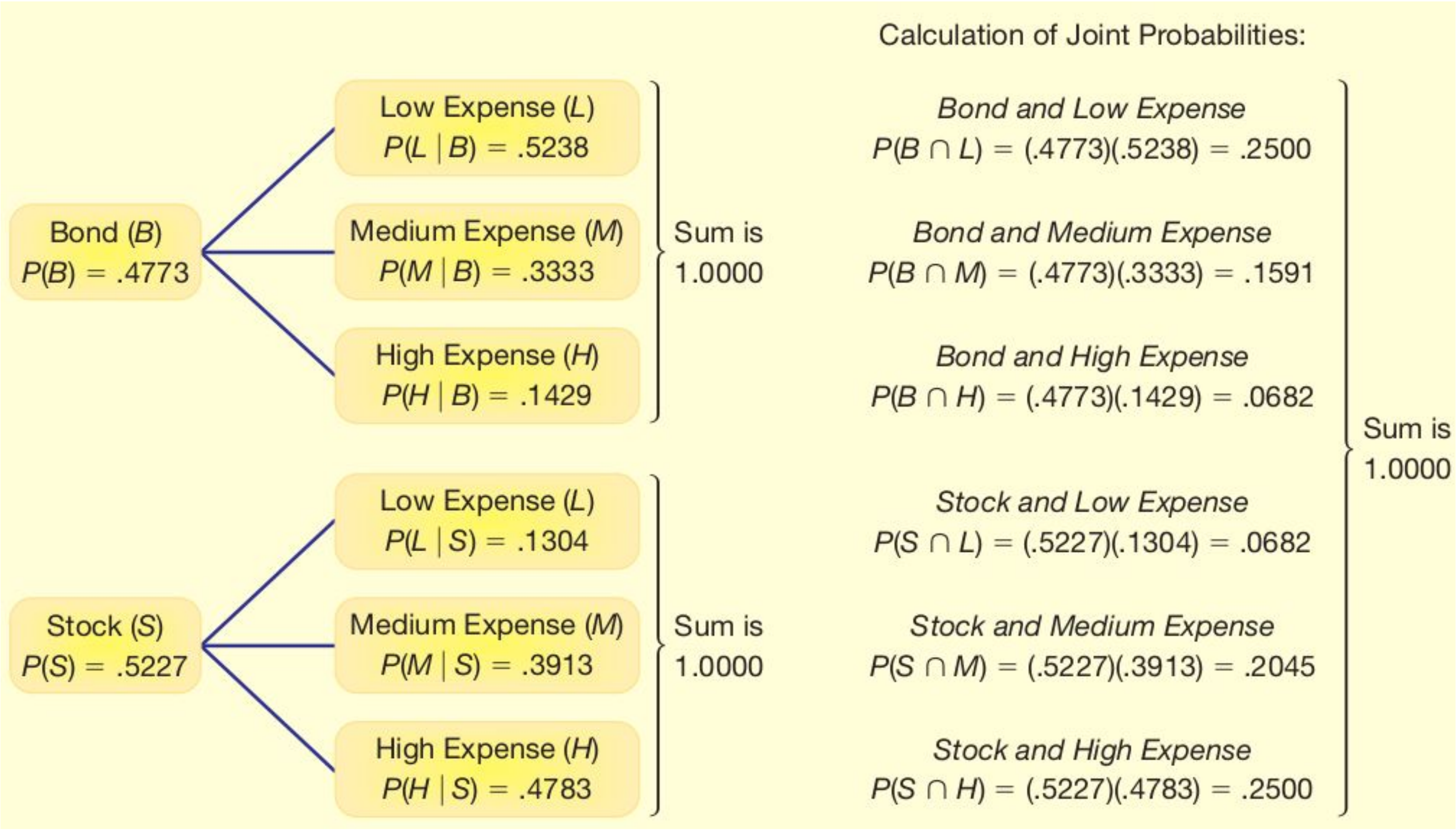
Tree Diagrams

Step 2. Calculate the *conditional probabilities*.

	Fund Type	
Expense Ratio	*Bond Fund (B)*	*Stock Fund (S)*
Low (L)	$= P(L \mid B) = 11/21 = .5238$	$= P(L \mid S) = 3/23 = .1304$
Medium (M)	$= P(M \mid B) = 7/21 = .3333$	$= P(M \mid S) = 9/23 = .3913$
High (H)	$= P(H \mid B) = 3/21 = .1429$	$= P(H \mid S) = 11/23 = .4783$
Column Total	1.0000	1.0000

Tree Diagrams

Step 3. Calculate the *joint probabilities* from the *conditional probabilities*.



Tree Diagram

Example (*Product Launching Plan*)

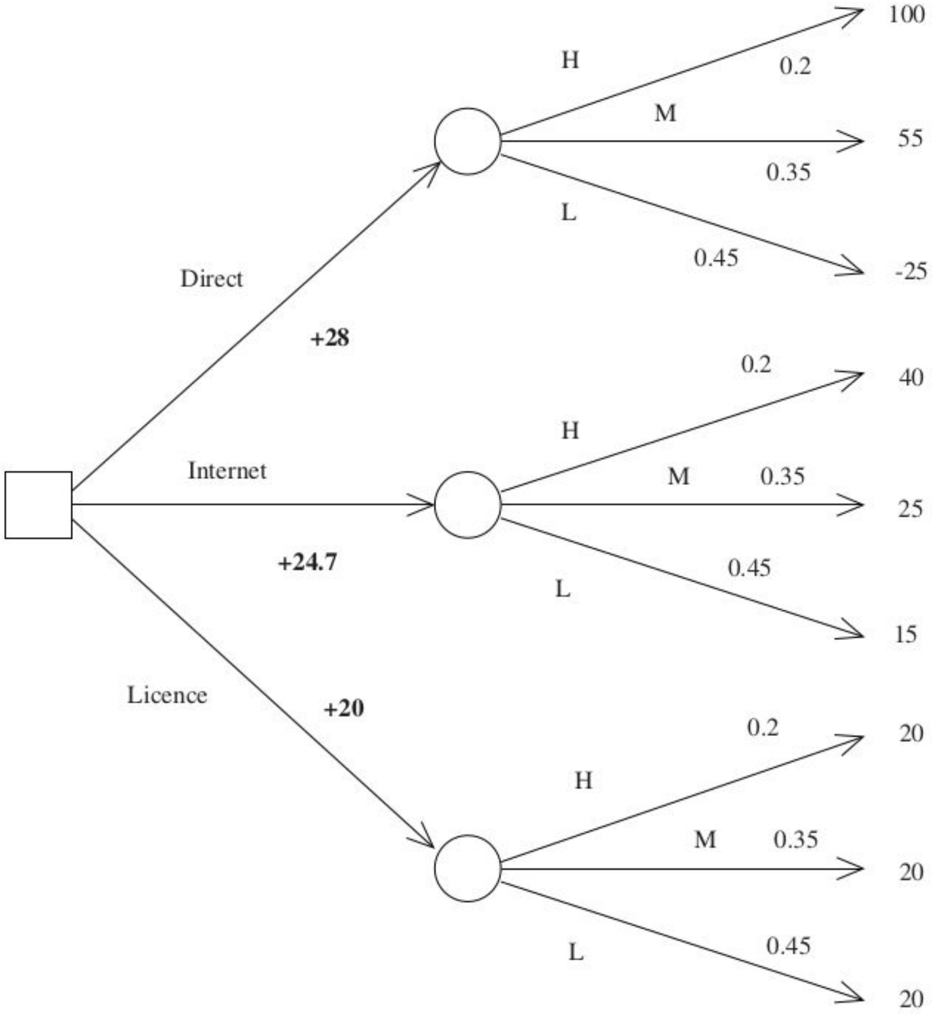
① A small technology company wish to launch a new and innovative product to the market. There are 3 options: ***Direct approach, Internet only*** *or* ***License****.*

② By Market research, the demand for the product can be classed into three categories: ***high, medium,*** or ***low*** with probabilities of 0.2, 0.35 and 0.45 respectively.

③ The likely profits to be earned in each plan are in the table

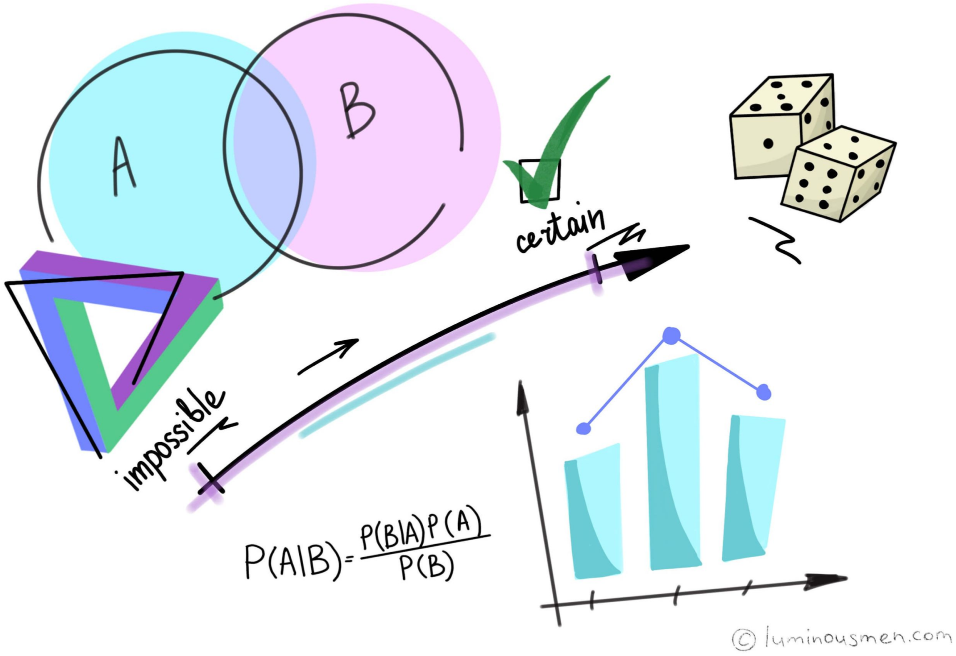
	High	**Medium**	**Low**
Direct	100	55	-25
Internet	46	25	15
License	20	20	20

How should the company launch the product ?



Direct
+28
Internet
+24.7
Licence
+20
H
M
L
0.2
0.35
0.45
100
55
-25
H
M
L
0.2
0.35
0.45
40
25
15
H
M
L
0.2
0.35
0.45
20
20
20

CONDITIONAL PROBABILITY & BAYES' THEOREM



Conditional Probability

Conditional Probability

If P(B) > 0 then the **conditional probability** of A given B is

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

Example:

Of the population age 16–21 and not in college:

- 13.50% are unemployed (U)
- 29.05% are high school dropouts (D)
- 5.32% are unemployed high school dropouts(U∩D)

→ The probability of an unemployed youth given that the person dropped out:

$$P(U|D) = \frac{P(U \cap D)}{P(D)} = \frac{0.0532}{0.2905} = 0.1831 = 18.31\%$$

The **conditional probability** of being unemployed is greater than the **unconditional probability** of being unemployed

→ In other words, knowing that someone is a high school dropout alters the probability that the person is unemployed.

minousmen.com

Bayes's Theorem

Theorem

Let A and B be event:

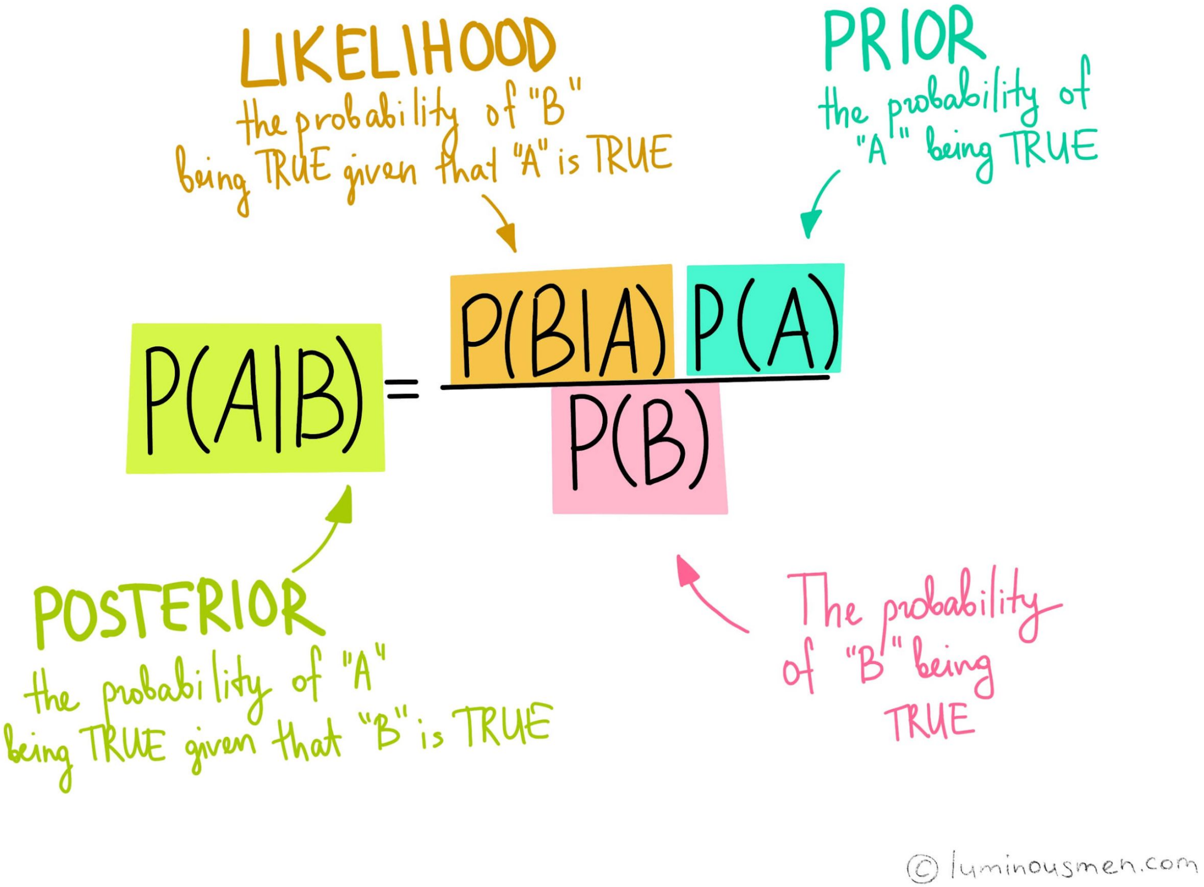
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

General form

If event B to have as many **mutually exclusive** and **collectively exhaustive** categories (B_1 , B_2 , ... , B_n)

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n)}$$

Bayes’s Theorem



Bayes' Theorem

Example: Rare Disease detection

A medical test for a rare disease D has outcomes (+) and (−).

Suppose you go for a test and get a positive.

What is the probability you have the disease?

	D	D^C
(+)	0.009	0.099
(−)	0.001	0.891

Most people choose P(+|D)=0.009/(0.009 + 0.001) =0.9 = **90%**

However, the correct answer is

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)}$$

With :

P(+|D) = 0.009 / (0.009+0.001) = 0.9

P(D) = (0.009 + 0.001) / (0.009 +0.001 + 0.099 + 0.891) = 0.01

P(+) = (0.009 + 0.099) / (0.009 + 0.099 + 0.001 + 0.891) = 0.108

→ **P(D|+) = 0.9 x 0.01 / 0.108 = 0.083 = 8.3%**

Bayes' Theorem

Example: Email Filter

A: The email contains the word "*free*"

B_1: "spam"

B_2: "low priority"

B_3: "high priority"

	B_1	B_2	B_3
$P(A\|B_i)$	0.90	0.01	0.01
$P(B_i)$	0.70	0.20	0.10

From previous experience, we can determine $P(A|B_i)$, $P(B_i)$

$\Rightarrow$ What is the probability that an email is spam containing a word "free"?

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)}$$

$$= \frac{0.9 \times 0.7}{0.90 \times 0.70 + 0.01 \times 0.20 + 0.01 \times 0.10} = 0.995$$

Programming

- A Crash Course in Python: https://nbviewer.jupyter.org/gist/rpmuller/5920182
- Programming tutorial: https://colab.research.google.com/drive/1jYKDeW74dULPRxJ7me-qtxbt_Gq8idw0
- Python tutorial: https://github.com/jerry-git/learn-python3

Reference

1. Doane, David P., and Lori E. Seward - *Applied statistics in business and economics*
2. Wasserman, Larry - *All of statistics: a concise course in statistical inference*
3. https://luminousmen.com/
4. http://www.mas.ncl.ac.uk/~ndah6/teaching/MAS1403/notes_chapter6.pdf
5. lumenlearning.com

End of Lecture 01

- What you have learned
  - Counting Rules
  - Sample Space, Event
  - Independent Event
  - Conditional Probability
  - Bayes' Theorem
- Questions?

Exercise for discussing

- *Ignoring leap years, and assuming birthdays are equally likely to be any day of the year, what is the chance of a tie in birthdays among the students in this class?*
- *In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?*
- *A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?*
- *How can you generate a random number between 1 - 7 with only a die?*