

EE2211 Introduction to Machine Learning

Lecture 3

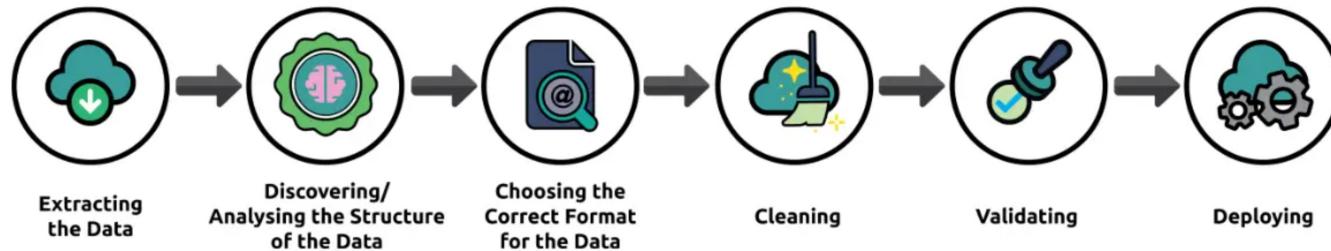
Wang Xinchao
xinchao@nus.edu.sg

Course Contents

- Introduction and Preliminaries (Xinchao)
 - Introduction
 - Data Engineering
 - Introduction to Linear Algebra, Probability and Statistics
- Fundamental Machine Learning Algorithms I (Yueming)
 - Systems of linear equations
 - Least squares, Linear regression
 - Ridge regression, Polynomial regression
- Fundamental Machine Learning Algorithms II (Yueming)
 - Over-fitting, bias/variance trade-off
 - Optimization, Gradient descent
 - Decision Trees, Random Forest
- Performance and More Algorithms (Xinchao)
 - Performance Issues
 - K-means Clustering
 - Neural Networks

Summary of Lec 2

- Types of data
 - NOIR
- Data wrangling and cleaning



- Data integrity and visualization
 - Integrity: Design
 - Visualization: Graphical Representation

Outline

- (Very Gentle) Introduction to Linear Algebra
 - Prof. Yueming's part will follow up
- Causality and Simpson's paradox
 - Understanding at intuitive level is sufficient
- Random Variable, Bayes' Rule

(Very Gentle) Introduction to Linear Algebra



- A **scalar** is a simple numerical value, like 15 or -3.25
 - Focus on **real** numbers
- **Variables** or **constants** that take scalar values are denoted by an *italic* letter, like x or a

Notations, Vectors, Matrices

$$\rightarrow \begin{bmatrix} 1 \\ 2 \end{bmatrix} \neq \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

- A **vector** is an **ordered list** of scalar values
 - Denoted by a **bold character**, e.g. \mathbf{x} or \mathbf{a}
- In many books, vectors are written column-wise:

$$\mathbf{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -2 \\ 5 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

\downarrow
column vector

- The three vectors above are **two-dimensional**, or have **two elements**

Notations, Vectors, Matrices

(element)

- We denote an **entry** or **attribute** of a vector as an italic value with an index, e.g. $a^{(j)}$ or $x^{(j)}$.
 - The index j denotes a specific dimension of the vector, the position of an attribute in the list

$$\mathbf{a} = \begin{bmatrix} a^{(1)} \\ a^{(2)} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \quad \text{or more commonly} \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

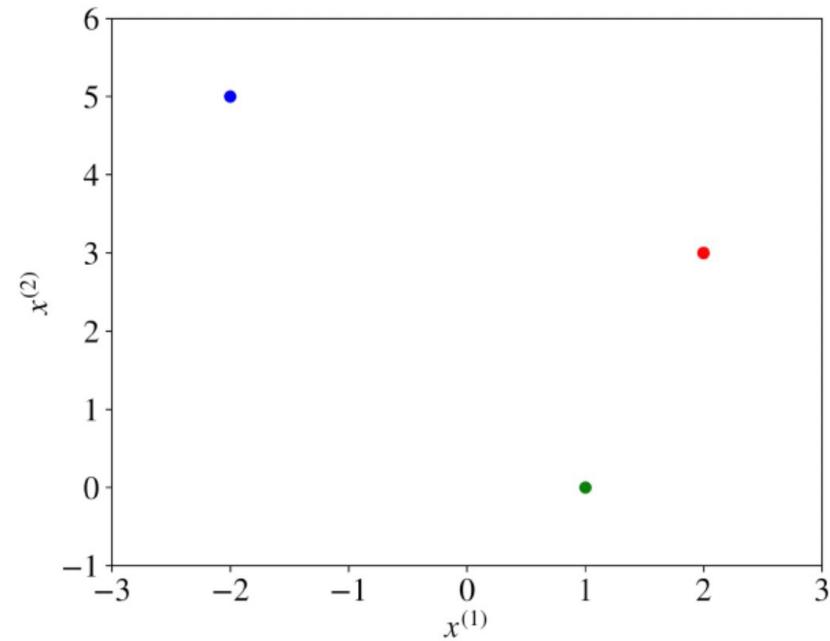
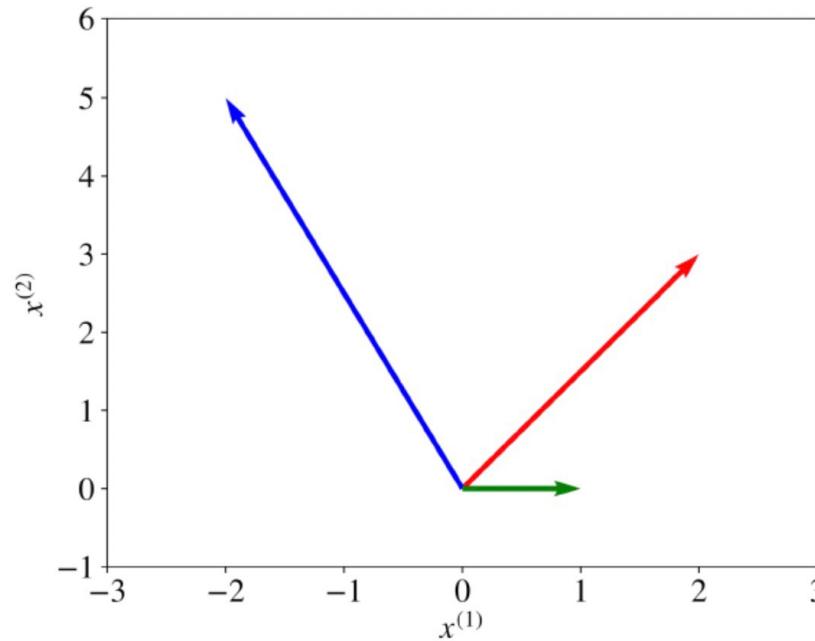
- Note:

- $x^{(j)}$ is not to be confused with the power operation, e.g., x^2 (squared)
- Square of an indexed attribute of a vector is denoted as $(x^{(j)})^2$.

Notations, Vectors, Matrices

- **Vectors** can be visualized as, in a multi-dimensional space,
 - arrows that point to some directions, or
 - points

Illustrations of three two-dimensional vectors, $\mathbf{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} -2 \\ 5 \end{bmatrix}$, and $\mathbf{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$



Notations, Vectors, Matrices

- A **matrix** is a rectangular array of numbers arranged in rows and columns

- Denoted with bold capital letters, such as **X** or **W**
- An example of a matrix with two rows and three columns:

	Feature 1	Feature 2	Feature 3	
x ₁₁	2	4	-3	sample 1
x ₂₁	21	-6	-1	sample 2

- A **set** is an unordered collection of unique elements

- When an element x belongs to a set S , we write $x \in S$.
- A special set denoted **R** includes all real numbers from minus infinity to plus infinity

$$S = \{1, 3, 5, 7, 9\} \\ = \{1, 5, 7, 9, 3\} \rightarrow \text{same (unordered)}$$

- Note:

- For elements in matrix **X**, we shall use the indexing $x_{1,1}$ where the first and second indices indicate the row and the column position.
- Usually, for input data, rows represent samples and columns represent features

Notations, Vectors, Matrices

- **Capital Sigma:** the **summation** over a collection $\{x_1, x_2, x_3, x_4, \dots, x_m\}$ is denoted by:

$$\sum_{i=1}^m x_i = x_1 + x_2 + \dots + x_{m-1} + x_m$$

\uparrow
 starting index
 $\overbrace{}$
 ending index

- **Capital Pi:** the **product** over a collection $\{x_1, x_2, x_3, x_4, \dots, x_m\}$ is denoted by:

$$\prod_{i=1}^m x_i = x_1 \cdot x_2 \cdot \dots \cdot x_{m-1} \cdot x_m$$

Systems of Linear Equations

Linear dependence and independence

- A collection of d -vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ (with $m \geq 1$) is called **linearly dependent** if

$$\beta_1 \mathbf{x}_1 + \cdots + \beta_m \mathbf{x}_m = 0$$

holds for some β_1, \dots, β_m that are not all zero.

E.g. $x_1 + 2x_2 = 0$

Since B_1 and $B_2 = 1, 2 \neq 0$,
it's linearly dependent

- A collection of d -vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ (with $m \geq 1$) is called **linearly independent** if it is not linearly dependent, which means that

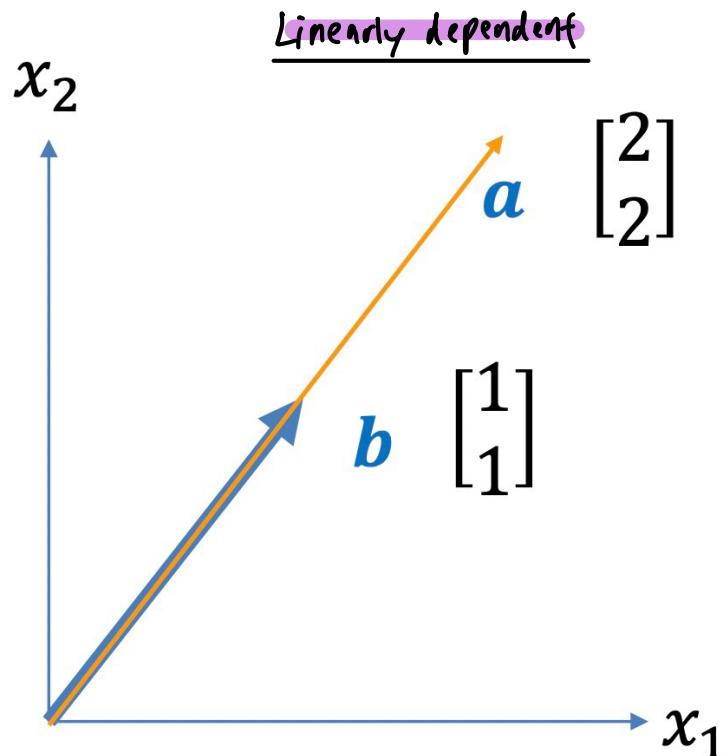
$$\beta_1 \mathbf{x}_1 + \cdots + \beta_m \mathbf{x}_m = 0$$

only holds for $\beta_1 = \cdots = \beta_m = 0$.

Note: If all rows or columns of a square matrix \mathbf{X} are linearly independent, then \mathbf{X} is invertible. \rightarrow full rank

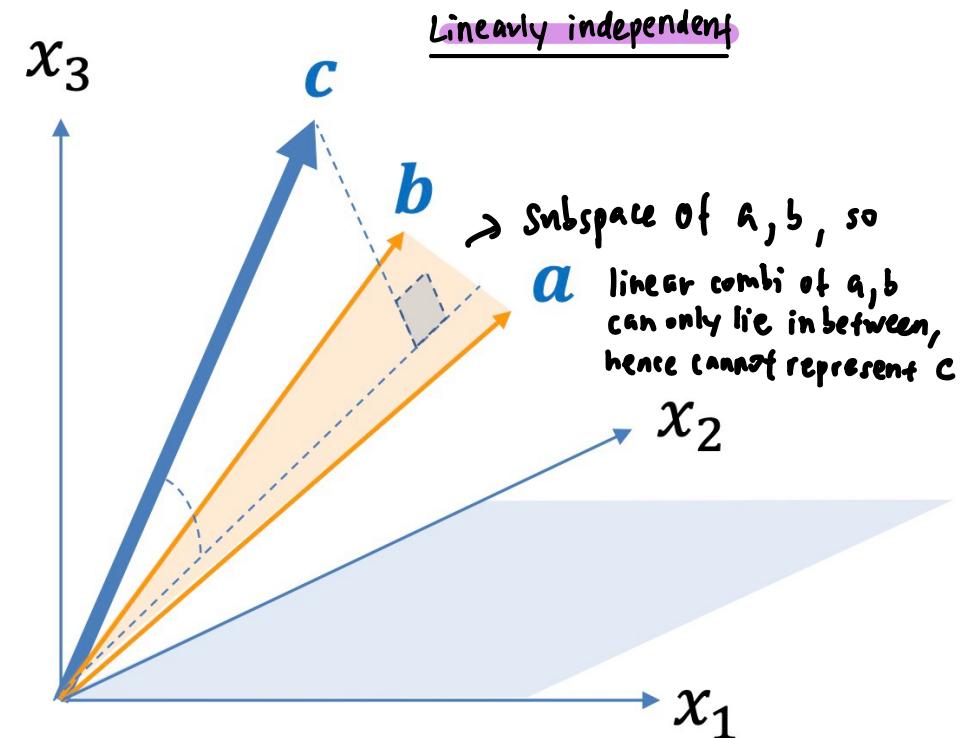
Systems of Linear Equations

Geometry of dependency and independency



$$\beta_1 a + \beta_2 b = 0$$

$$\beta_1 = 1 \quad \beta_2 = 2 = 0$$



$$\beta_1 a + \beta_2 b \neq \beta_3 c \therefore \beta_1 = \beta_2 = \beta_3 = 0$$

for eqtn to hold

Systems of Linear Equations

These equations can be written compactly in **matrix-vector notation**:

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

↑ ↑
 given parameter vector
 (unknown)

Where

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,d} \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

↓
 m × d
 sample features

dx1
 m × 1

Note:

- The **data matrix** $\mathbf{X} \in \mathcal{R}^{m \times d}$ and the **target vector** $\mathbf{y} \in \mathcal{R}^m$ are given
- The **unknown vector of parameters** $\mathbf{w} \in \mathcal{R}^d$ is to be learnt
- The $\text{rank}(\mathbf{X})$ corresponds to the maximal number of linearly independent columns/rows of \mathbf{X} .

Exercises

- The principled way for computing rank is to do Echelon Form
 - <https://stattrek.com/matrix-algebra/echelon-transform.aspx#MatrixA>
- For small-size matrices, however, the rank is in many cases easy to estimate
- What is the rank of

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \xrightarrow{\text{rref}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{rank = 2}$$

$$\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix} \xrightarrow{\text{rref}} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{rank = 1}$$

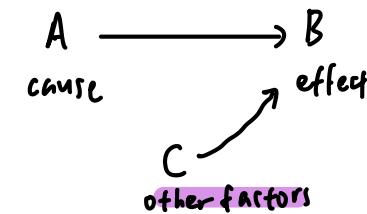
$$\begin{bmatrix} 1 & -2 & 3 \\ 0 & -3 & 3 \\ 1 & 1 & 0 \end{bmatrix} \xrightarrow{\text{rref}} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{rank = 2}$$

Outline

- (Very Gentle) Introduction to Linear Algebra
- Causality and Simpson's paradox
- Random Variable, Bayes' Rule

Causality

- Causality, or causation is:
 - The influence by which one event or process (i.e., **cause**) contributes to another (i.e. **effect**),
 - The **cause** is partly responsible for the **effect**, and the **effect** is partly dependent on the **cause**
- Causality relates to an extremely very wide domain of subjects: philosophy, science, management, humanity.
- Causality research is extremely complex
 - Researcher can **never be completely certain** that there are **no other factors** influencing the causal relationship,
 - In most cases, we can only say “**probably**” causal.



Causality

- (Probable) causal relations or non-causal?
 - New web design implemented ? Web page traffic increased **C Cprobable**
 - Your height and weight ? Gets A in EE2211 **NC C independent**
 - Uploaded new app store images ? Downloads increased by 2X **C Cprobable**
 - One works hard and attends lectures/tutorials ? Gets A in EE2211 **C Cprobable**
 - Your favorite color ? Your GPA in NUS **NC C independent**

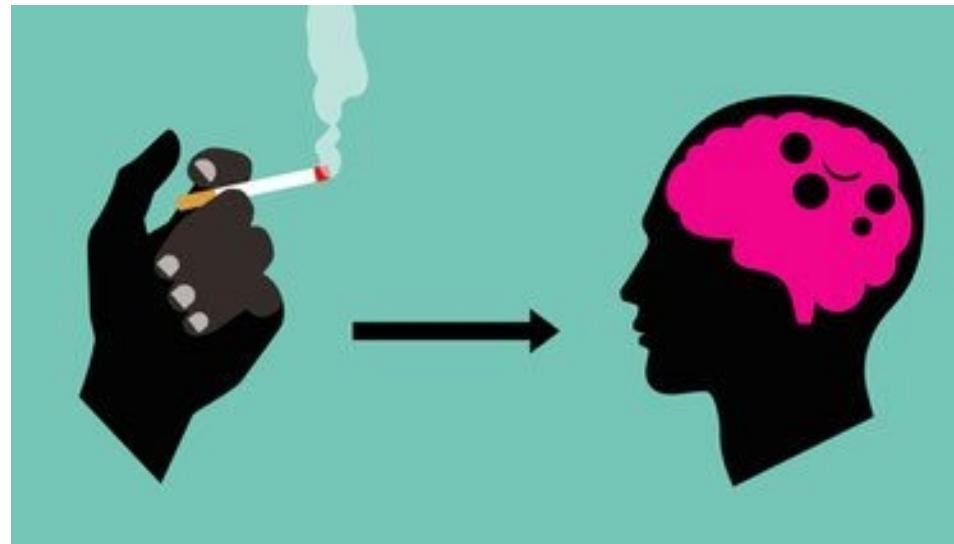
Causality

- One popular way to causal data analysis is **Randomized Controlled Trial (RCT)** *****
 - A study ~~design that randomly assigns~~ participants into an experimental group or a control group.
 - As the study is conducted, the ~~only expected difference~~ between two groups ~~is the outcome variable being studied.~~
- Example:
 - To decide whether smoking and lung cancer has a causal relation, we put participants into experimental group (people who smoke) and control group (people who don't smoke), and check whether they develop lung cancer eventually.
- RCT is sometimes infeasible to conduct, and also has moral issues.

Causality is a statistical relationship

- Decades of data show a clear causal relationship between smoking and cancer.
- If one smokes, it is a sure thing that his/her risk of cancer will increase.

- But it is not a sure thing that one will get cancer.
- The relationship is not deterministic.

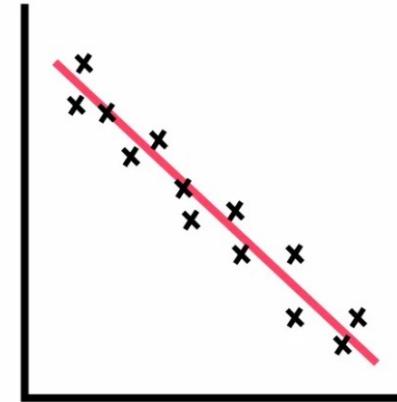


Correlation (vs Causality)

- In statistics, **correlation** is any **statistical relationship**, whether causal or not, between two random variables.
- Correlations are useful because they can indicate a **predictive relationship** that can be exploited in practice.



Positive
Correlation

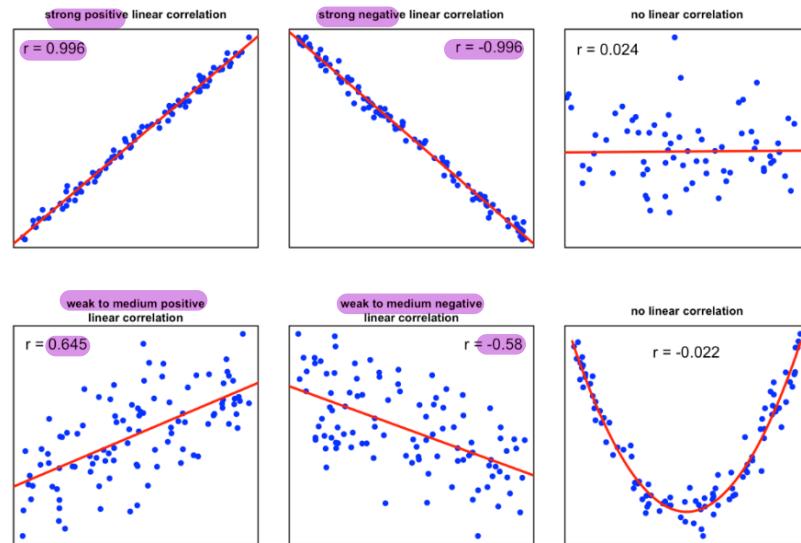


Negative
Correlation

Correlation (vs Causality)

- Linear correlation coefficient, r , which is also known as the Pearson Coefficient. → *degree of correlation (strength of linear relationship between 1 and -1)*

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y},$$



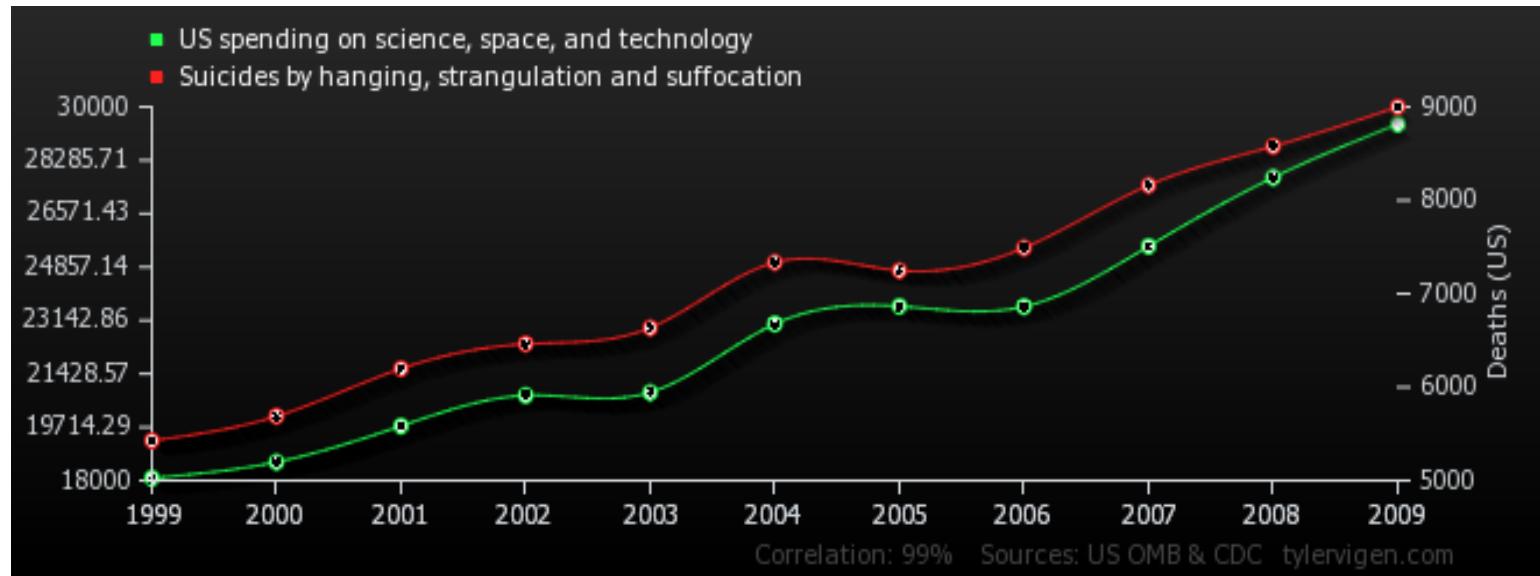
Strong linear relationship	$r > 0.9$
Medium linear relationship	$0.7 < r \leq 0.9$
Weak linear relationship	$0.5 < r \leq 0.7$
No or doubtful linear relationship	$0 < r \leq 0.5$

The same holds for negative values.

1. <https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Descriptive-Statistics/Measures-of-Relation-Between-Variables/Correlation/index.html>

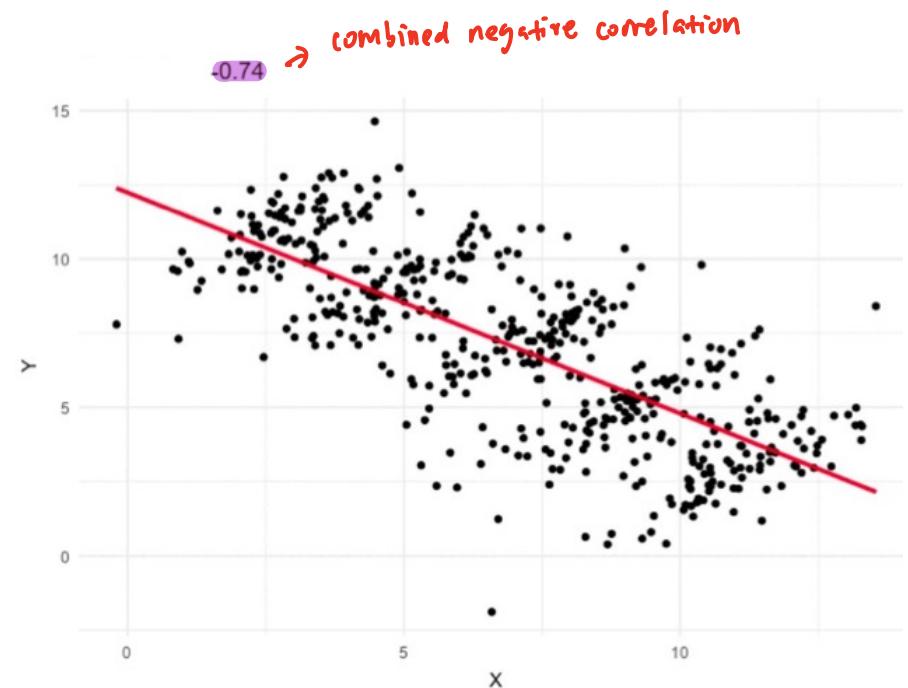
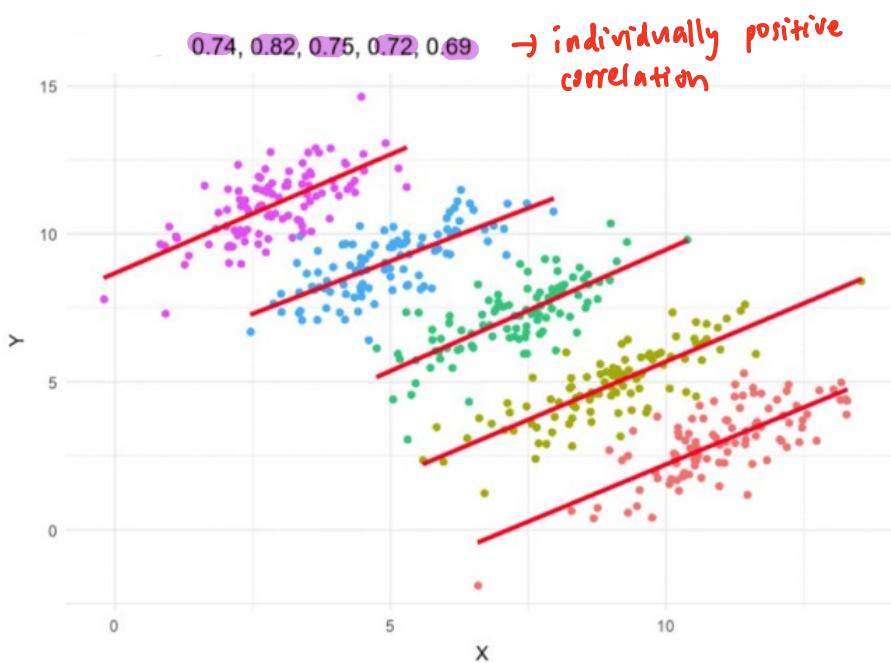
Correlation does not imply causation!

- Some great examples of correlations that can be calculated but are clearly not causally related appear at <http://tylervigen.com/>



Simpson's paradox

- Simpson's paradox is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.



The same set of samples!

Example

- Batting Average in professional baseball game
- Two well-known players, Derek Jeter and David Justice

Batter \ Year	1995		1996		Combined	
Batter						
Derek Jeter	12/48	.250	183/582	.314	195/630	.310
David Justice	104/411	.253	45/140	.321	149/551	.270

#of wins #of games

Outline

- (Very Gentle) Introduction to Linear Algebra
- Causality and Simpson's paradox
- Random Variable, Bayes' Rule

Probability

- We describe a *random experiment* by describing its procedure and observations of its *outcomes*.
- *Outcomes* are mutual exclusive in the sense that only one outcome occurs in a specific trial of the random experiment.
 - This also means an outcome is not decomposable.
 - All unique outcomes form a *sample space*.
- A subset of sample space S , denoted as A , is an *event* in a random experiment $A \subset S$, that is meaningful to an application.
 - Example of an event: faces with numbers no greater than 3

events is a subset of sample space and
a set of outcomes

\downarrow
event is subset of sample space



event



Axioms of Probability

Assuming events $A \subseteq S$ and $B \subseteq S$, the probabilities of events related with and must satisfy,

1. $Pr(A) \geq 0$
2. $Pr(S) = 1$
3. If $A \cap B = \emptyset$, then $Pr(A \cup B) = Pr(A) + Pr(B)$
*otherwise, $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

[https://en.wikipedia.org/wiki/Union_\(set_theory\)](https://en.wikipedia.org/wiki/Union_(set_theory))

[https://en.wikipedia.org/wiki/Intersection_\(set_theory\)](https://en.wikipedia.org/wiki/Intersection_(set_theory))

Random Variable

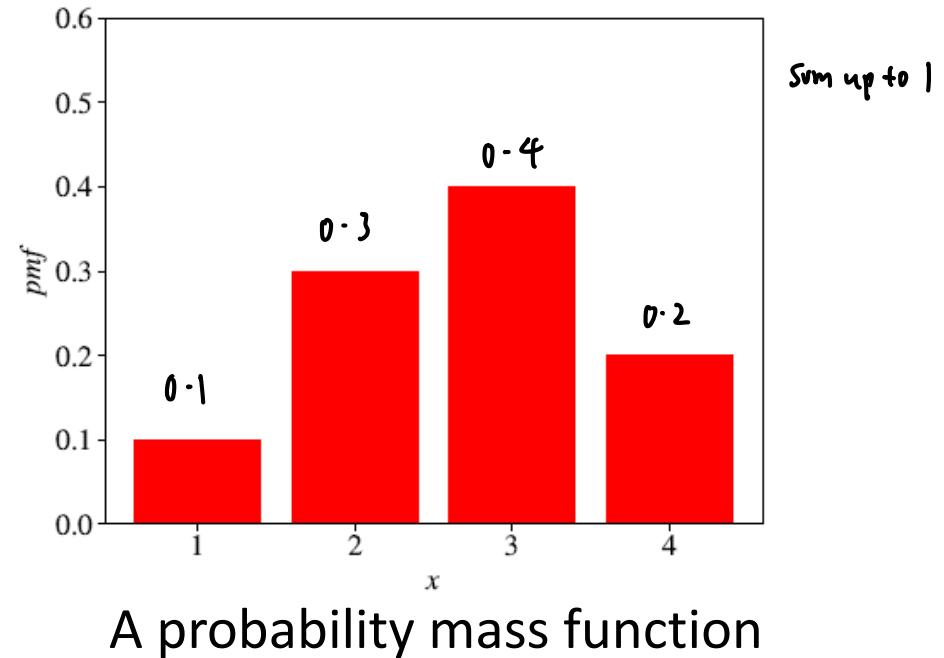
- A **random variable**, usually written as an *italic* capital letter, like X , is a variable whose ~~possible values are numerical outcomes of a random event.~~
- There are two types of random variables: **discrete** and **continuous**.

Notations

- Some books used $P(\cdot)$ and $p(\cdot)$ to distinguish between the probability of discrete random variable and the probability of continuous random variables respectively.
- We shall use $Pr(\cdot)$ for both the above cases

Discrete random variable

- A **discrete random variable (DRV)** takes on only a countable number of distinct values such as **red**, **orange**, **blue** or 1, 2, 3.
- The **probability distribution** of a discrete random variable is described by a list of probabilities associated with each of its possible values.
- This list of probabilities is called a **probability mass function (pmf)**.
 - Like a histogram, except that here the **probabilities sum to 1**



Discrete random variable

- Let a **discrete** random variable X have k possible values $\{x_i\}_{i=1}^k$.

↗ weighted average

- The **expectation** of X denoted as $E(x)$ is given by,

$$\begin{aligned} E(x) &\stackrel{\text{def}}{=} \sum_{i=1}^k [x_i \cdot \Pr(X = x_i)] \\ &= x_1 \cdot \Pr(X = x_1) + x_2 \cdot \Pr(X = x_2) + \dots + x_k \cdot \Pr(X = x_k) \end{aligned}$$

where $\Pr(X = x_i)$ is the probability that X has the value x_i according to the **pmf**.

- The expectation of a random variable is also called the **mean, average** or **expected value** and is frequently denoted with the letter μ .

Discrete random variable

- Another important statistic is the **standard deviation**, defined as,

$$\sigma \stackrel{\text{def}}{=} \sqrt{E[(X - \mu)^2]} .$$

- Variance**, denoted as σ^2 or $\text{var}(X)$, is defined as,

$$\sigma^2 = E[(X - \mu)^2]$$

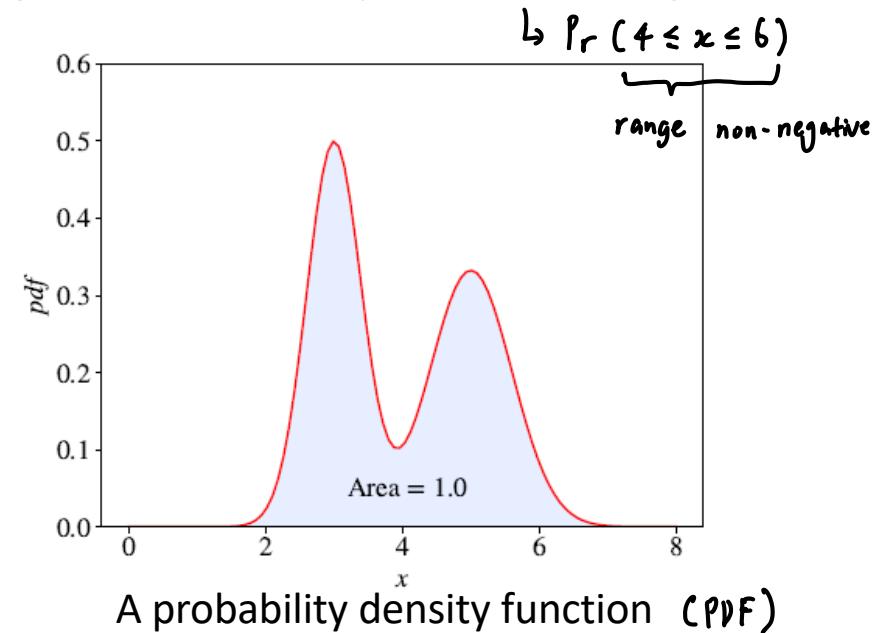
- For a **discrete random variable**, the standard deviation is given by

$$\sigma = \sqrt{\Pr(X = x_1)(x_1 - \mu)^2 + \cdots + \Pr(X = x_k)(x_k - \mu)^2}$$

where $\mu = E(X)$.

Continuous random variable

- A **continuous random variable (CRV)** takes an infinite number of possible values in some interval.
 - Examples include height, weight, and time. E.g. height : 1.723.... m
 - The number of values of a continuous random variable X is infinite, the probability $\Pr(X = c)$ for any c is 0.
 - Therefore, instead of the list of probabilities, the probability distribution of a CRV (a continuous probability distribution) is described by a **probability density function (pdf)**.
 - The pdf is a function whose range is **nonnegative** and the **area under the curve is equal to 1**.

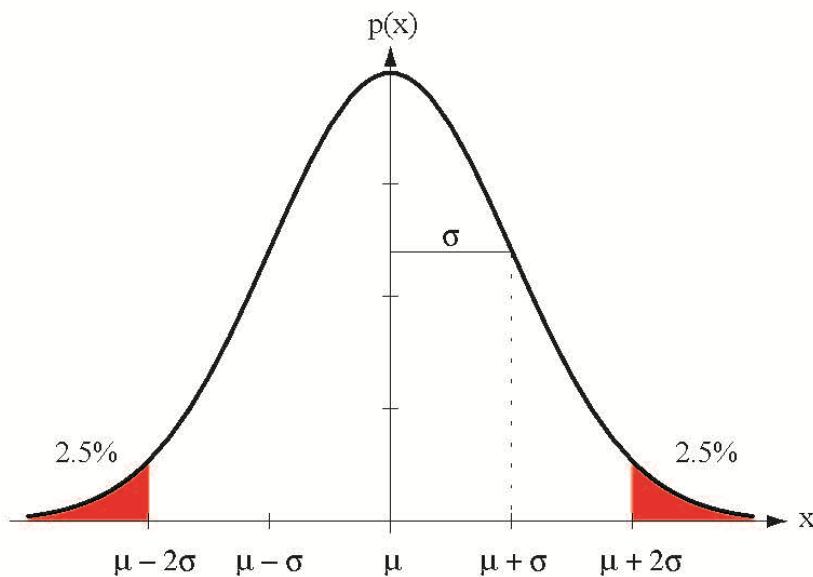


Continuous random variable

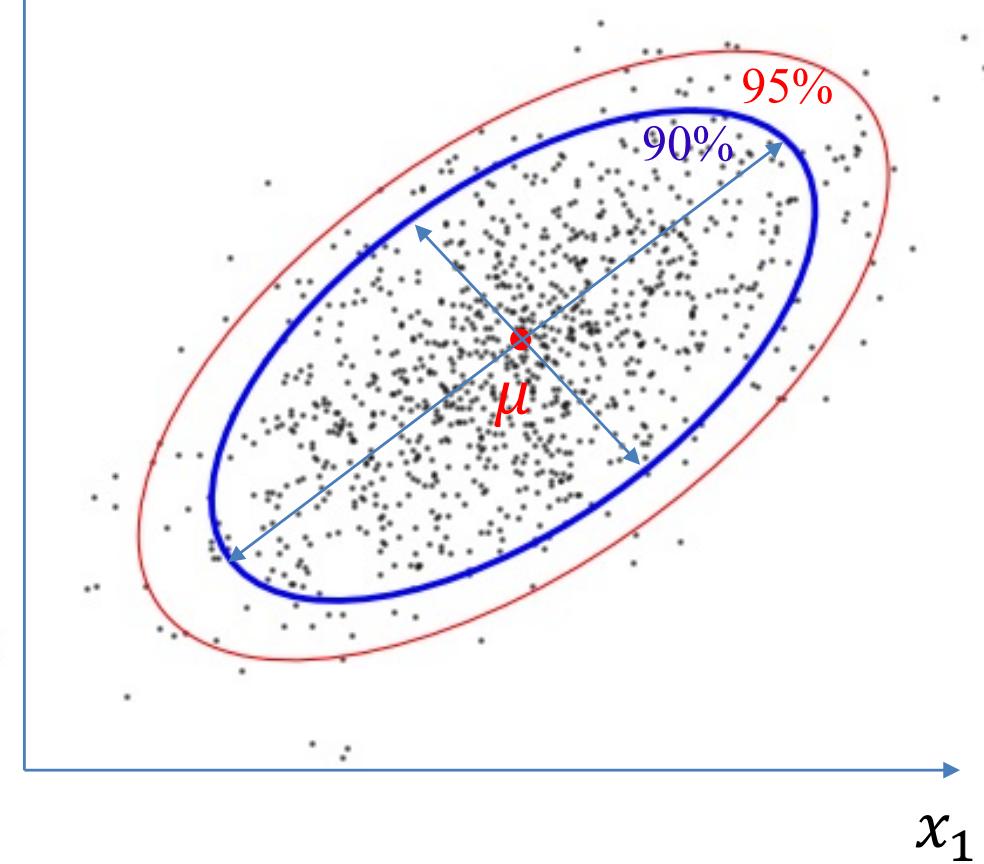
- The **expectation** of a continuous random variable X is given by $E[x] \stackrel{\text{def}}{=} \int_R x f_X(x) dx$ [$\int \dots dx \rightarrow \Sigma$] (intuitively) where f_X is the **pdf** of the variable X and \int_R is the integral of function $x f_X$.
- The **variance** of a continuous random variable X is given by $\sigma^2 \stackrel{\text{def}}{=} \int_R (X - \mu)^2 f_X(x) dx$
- Integral is an equivalent of the summation over all values of the function when the function has a continuous domain.
- It equals the area under the curve of the function.
- The property of the pdf that the area under its curve is 1 mathematically means that $\int_R f_X(x) dx = 1$

Mean and Standard Deviation of a Gaussian Distribution

1 Dimension



2 Dimension



Example 1

- **Independent random variables**
- Consider tossing a fair coin twice, what is the probability of having (H,H)? Assuming a coin has two sides, H=head and T=Tail
 - $\Pr(x=H, y=H) = \Pr(x=H)\Pr(y=H) = (1/2)(1/2) = 1/4$

Example 2

- **Dependent random variables**
- Given 2 balls with different colors (**Red** and Black), what is the probability of first drawing B and then **R**? Assuming we are drawing the balls **without replacement**.
- The space of outcomes of taking two balls sequentially without replacement:

B-R, | R-B

– Thus having B-R is $1/2$.

$\Pr(y|x)$ → Probability of y given x

- Mathematically:

$$-\Pr(x=B, y=R) = \underbrace{\Pr(y=R | x=B)}_{\text{Conditional Probability}} \Pr(x=B) = 1 \times (1/2) = 1/2$$

joint probability

↑
no replacement, hence
 $\Pr(y=R | x=B) = 1$

either red/black first so $1/2$

Example 3

- **Dependent random variables**
- Given 3 balls with different colors (R,G,B), and we draw 2 balls. What is the probability of first having B and then G, if we draw **without replacement**?
- The space of outcomes of taking two balls **sequentially without replacement**:

R-G | G-B | B-R
 R-B | G-R | B-G

$\Pr(y|x)$ → Probability of y given x

Thus, $\Pr(y=G, x=B) = 1/6$

- Mathematically:

$$\begin{aligned}
 \Pr(y=G, x=B) &= \Pr(y=G | x=B) \Pr(x=B) \\
 &= (1/2) \times (1/3) \\
 &= 1/6
 \end{aligned}$$

$\nearrow \frac{1}{2}$ → ball without replacing ball
 $\nwarrow \frac{2}{6} = \frac{1}{3}$

Two Basic Rules

- **Sum Rule**

$$\Pr(X = x) = \sum_Y \Pr(X = x, Y = y_i)$$

joint probability
probability of x *Y* *probability of x and y*
↓ *↓* *↓*

← Allows to change the # of variables within the bracket

- **Product Rule**

$$\Pr(X = x, Y = y) = \underbrace{\Pr(Y = y|X = x)}_{\text{conditional probability}} \underbrace{P(X = x)}_{\text{probability of } x}$$

joint probability *summed over all the possible values of y*
↓ *↓*

Bayes' Rule

- The conditional probability $\Pr(Y = y|X = x)$ is the probability of the random variable Y to have a specific value y , given that another random variable X has a specific value of x .
- The **Bayes' Rule** (also known as the **Bayes' Theorem**):

$$\Pr(Y = y|X = x) = \frac{\text{likelihood}}{\text{posterior}} \cdot \frac{\text{prior}}{\text{evidence}}$$
$$\Pr(Y = y|X = x) = \frac{\Pr(X = x|Y = y) \Pr(Y = y)}{\Pr(X = x)}$$

Example

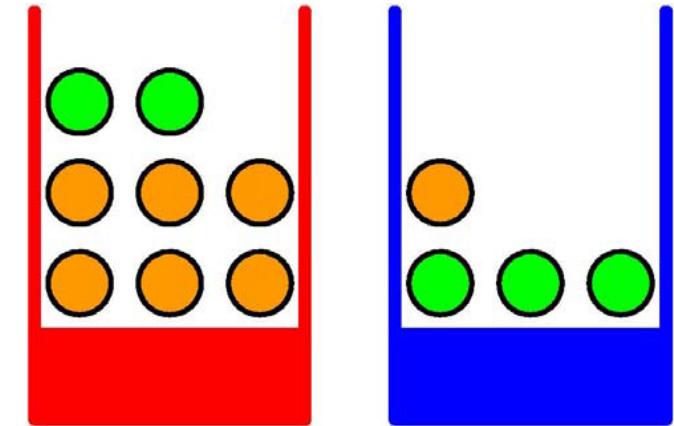
$$\Pr(F=0) = \Pr(F=0, B=r) + \Pr(F=0, B=b)$$

$$\Pr(F=0 | B=b) \cdot \Pr(B=b)$$

- Drawing a sample of fruit from a box

- First pick a box, and then draw a sample of fruit from it
- B : variable for Box, can be r (red) or b (blue)
- F : variable for Fruit, can be o (orange) or a (apple)

- $\Pr(B=r) = 0.4, \Pr(B=b) = 1 - 0.4$ prior
 - $\Pr(F=o | B=r) = 0.75$ likelihood
 - $\Pr(F=o) = 0.45$ evidence
- ↳ $0.4 \times \left(\frac{6}{8}\right) + 0.6 \times \left(\frac{1}{4}\right)$


 0.6

$$\begin{aligned} \Pr(B=r | F=o) &= \Pr(F=o | B=r) * \Pr(B=r) / \Pr(F=o) \\ &= 0.75 * 0.4 / 0.45 = 2/3 \end{aligned}$$

posterior

$$* \Pr(B=b | F=o) = 1 - \frac{2}{3} = \frac{1}{3}$$

(since probability of red box chosen given no. of orange is $\frac{2}{3}$, and there is only red box and blue box, it must be $\frac{1}{3}$)

Summary

- (Very Gentle) Introduction to Linear Algebra
- Causality and Simpson's paradox
- Random Variable, Bayes' Rule

Practice Question

(Type of Question to Expect in Exams)

Suppose the random variable X has the following probability mass function (pmf) listed in the table below. k is unknown.

(Independent)

X	1	2	3	4	5
$\Pr[X]$	0.1	0.05	0.05	0.6	k

What is the probability that X takes a value of odd numbers? (0.35)

$S = \text{Even and odd numbers}$

$$\Pr(\text{Even}) = 0.05 + 0.6 = 0.65$$

$$\Pr(S) = 1$$

$$\begin{aligned}\Pr(\text{odd}) &= 1 - 0.65 \\ &= 0.35\end{aligned}$$

$$\Pr(1) = 0.1$$

$$\Pr(2) = 0.05$$

$$\Pr(3) = 0.05$$

$$\Pr(4) = 0.6$$

$$\Pr(5) = k$$

$$\Pr(S) = 1 - 0.8 = 0.2$$

$$\therefore k = 0.2$$

