EE2211 Introduction to Machine Learning

Lecture 3

Wang Xinchao
xinchao@nus.edu.sg

Course Contents

1

- Introduction and Preliminaries (Xinchao)
  - Introduction
  - Data Engineering
  - Introduction to Linear Algebra, Probability and Statistics

2
- Fundamental Machine Learning Algorithms I (Yueming)
  - Systems of linear equations
  - Least squares, Linear regression
  - Ridge regression, Polynomial regression

3
- Fundamental Machine Learning Algorithms II (Yueming)
  - Over-fitting, bias/variance trade-off
  - Optimization, Gradient descent
  - Decision Trees, Random Forest

4
- 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

- A **vector** is an ordered list of scalar values
  – Denoted by a **bold character**, e.g. **x** or **a**

- In many books, vectors are written column-wise:

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

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

Notations, Vectors, Matrices

- 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

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

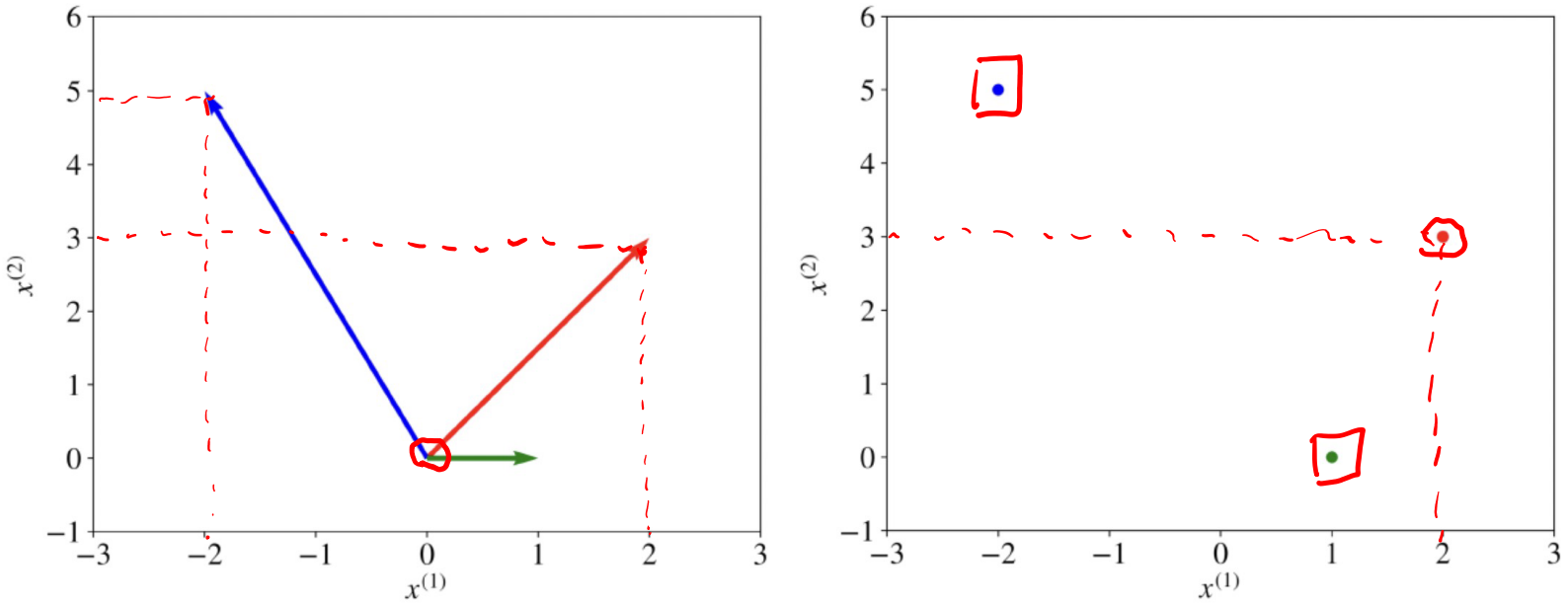
- Note:
  - $x^{(j)}$ is not to be confused with the power operation, e.g., x^2 (squared)
  - Square of an indexed attribute of a vector is denoted as $(x^{(j)})^2$.

$2^2 = 4$

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:

$$\mathbf{X} = \begin{bmatrix} 2 & 4 & -3 \\ 21 & -6 & -1 \end{bmatrix}$$

- A **set** is an unordered collection of unique elements
  - When an element x belongs to a set $\boldsymbol{S}$, we write $x \in \boldsymbol{S}$.
  - A special set denoted **R** includes all real numbers from minus infinity to plus infinity

- 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 $\{x_1, x_2, x_3, x_4, \ldots, x_m\}$ is denoted by:

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

ending index

starting index

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

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

Systems of Linear Equations

Linear dependence and independence

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

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

holds for some $\beta_1, \ldots, \beta_m$ that are **not all zero**.

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

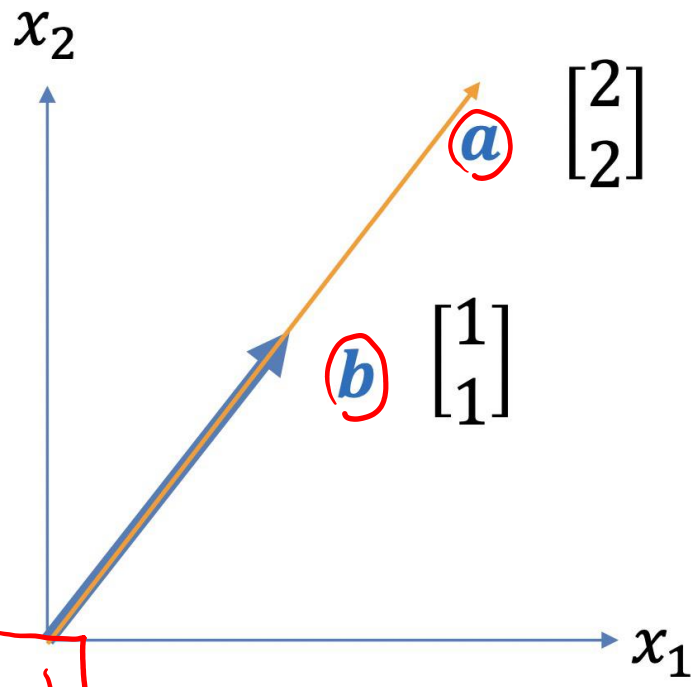
$$\beta_1 \mathbf{x}_1 + \cdots + \beta_m \mathbf{x}_m = 0$$

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

Note: If all rows or columns of a square matrix **X** are **linearly independent**, then **X** is **invertible.**

Systems of Linear Equations

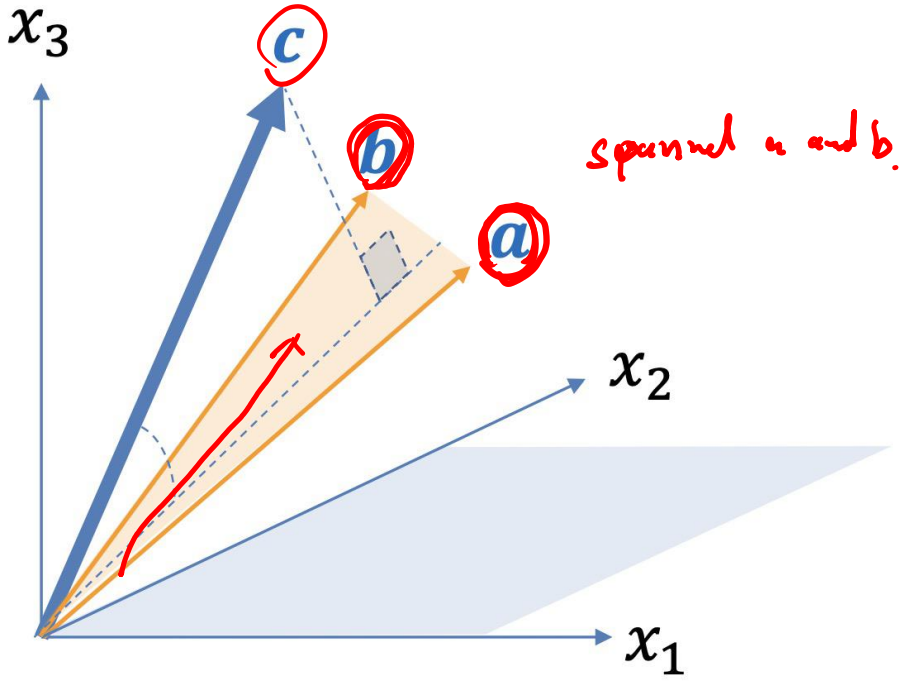
Geometry of dependency and independency



$\beta_1 = 1$
$\beta_2 = -2$

$\beta_1 \boldsymbol{a} + \beta_2 \mathbf{b} = 0$

$1 \cdot \begin{bmatrix}2\\2\end{bmatrix} + \beta_2 \cdot \begin{bmatrix}1\\1\end{bmatrix}$



spanned a and b.

$\beta_1 \boldsymbol{a} + \beta_2 \boldsymbol{b} \neq \beta_3 \mathbf{c}$

$\beta_1 a + \beta_2 b + \beta_3 c = 0$ X

Systems of Linear Equations

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

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

Where

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

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 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

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

rank 2

$$\begin{bmatrix} 1 \\ 100 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 100 \end{bmatrix}$$

rank = 1

$$\begin{bmatrix} 1 & -2 & 3 \\ 0 & -3 & 3 \\ 1 & 1 & 0 \end{bmatrix}$$

rank = 2 ~~* 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.

A causes B

I can't be sure there is no other factors C. DEF contributing to B.

- 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 Yes
  - Your height and weight **?** Gets A in EE2211 No
  - Uploaded new app store images **?** Downloads increased by 2X Yes
  - One works hard and attends lectures/tutorials **?** Gets A in EE2211 Yes
  - Your favorite color **?** Your GPA in NUS No

Causality

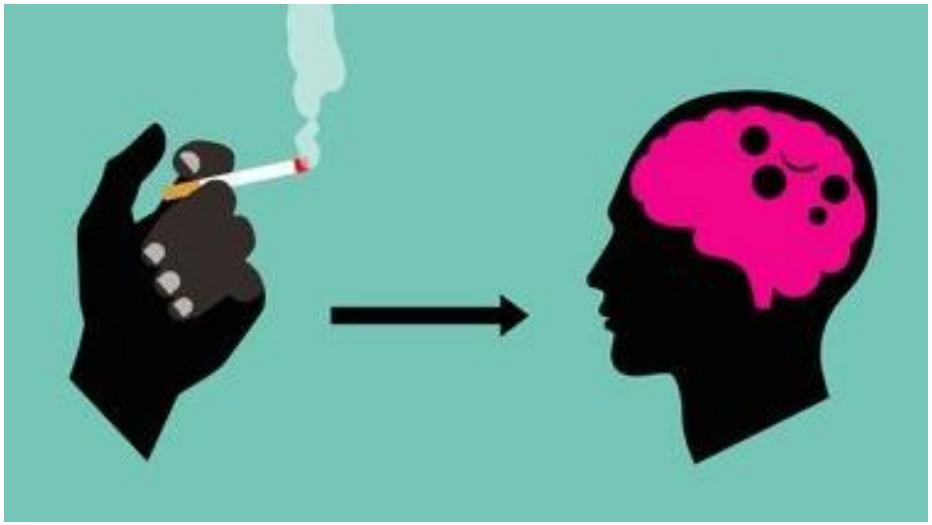
- One popular way to causal data analysis is **Randomized Controlled Trial (RCT)**
  - A study design that randomly assigns participants into an *experimental* group or a *control* group.
  - As the study is conducted, the only expected difference between two groups is the outcome variable being studied.

- 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.

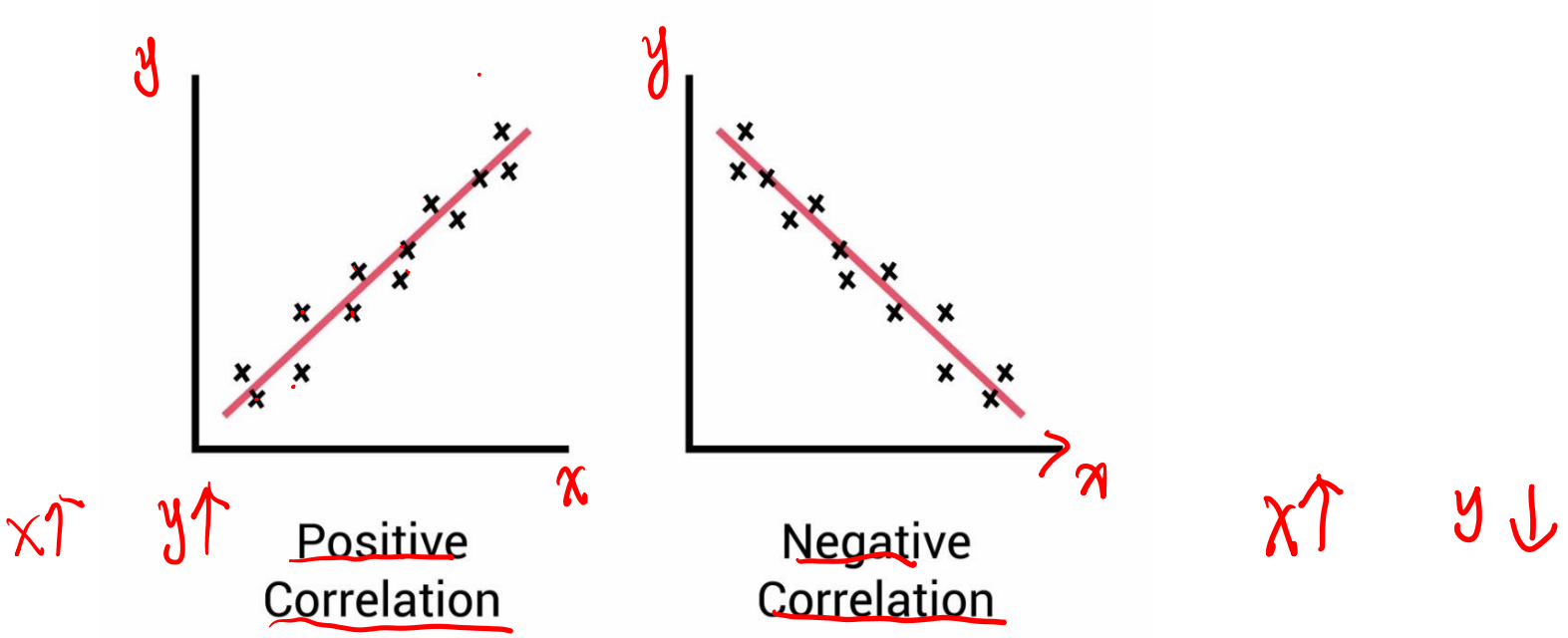
Causality is a statistical relationship

- 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.



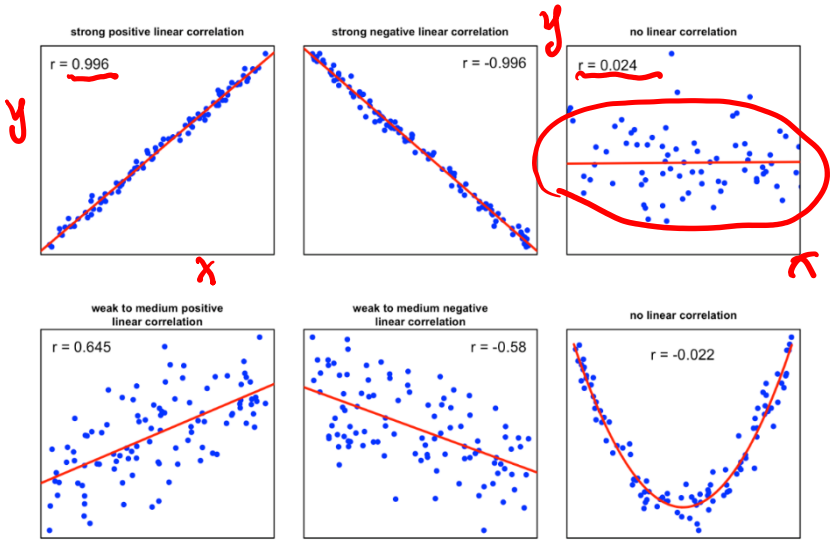
Correlation (vs Causality)

- Linear correlation coefficient, r, which is also known as the Pearson Coefficient.

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

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	$0 < r \leq 0.5$

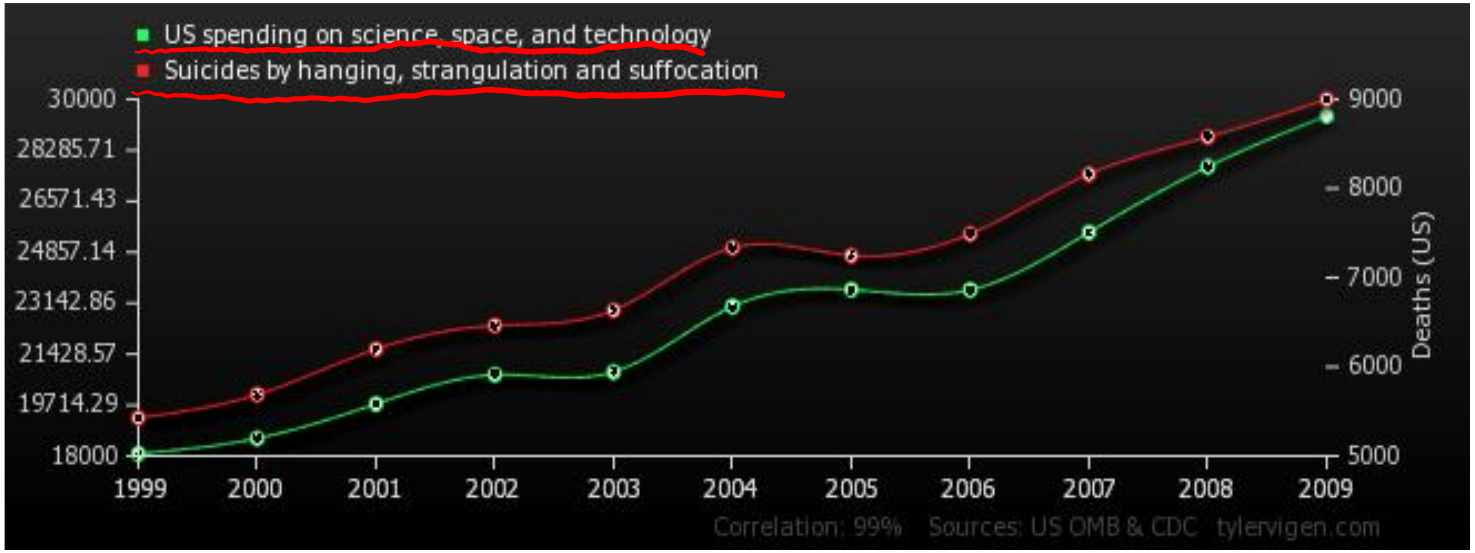
The same holds for negative values.



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

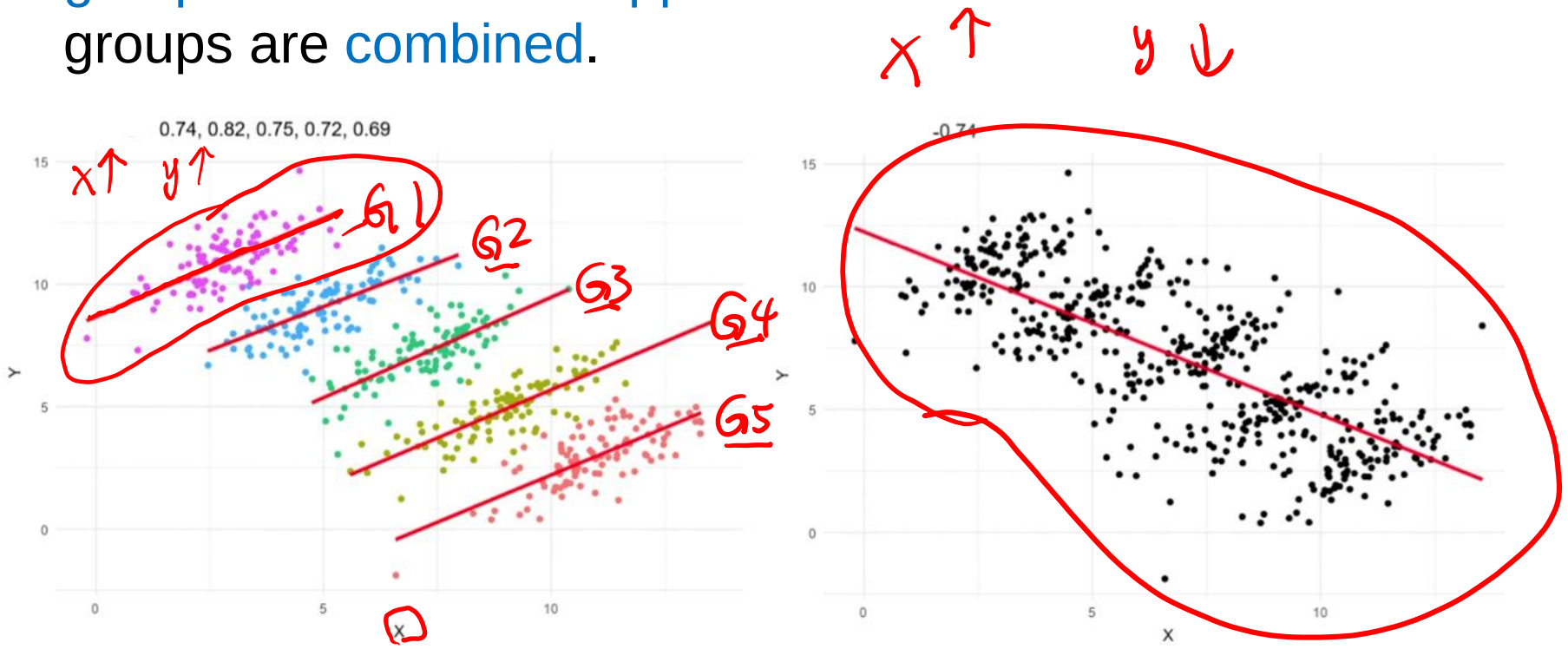
Correlation does not imply causation!

- 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	
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

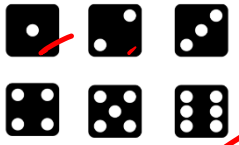
Outline

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

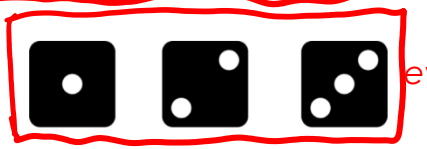
Probability

- 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

outcome



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/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.

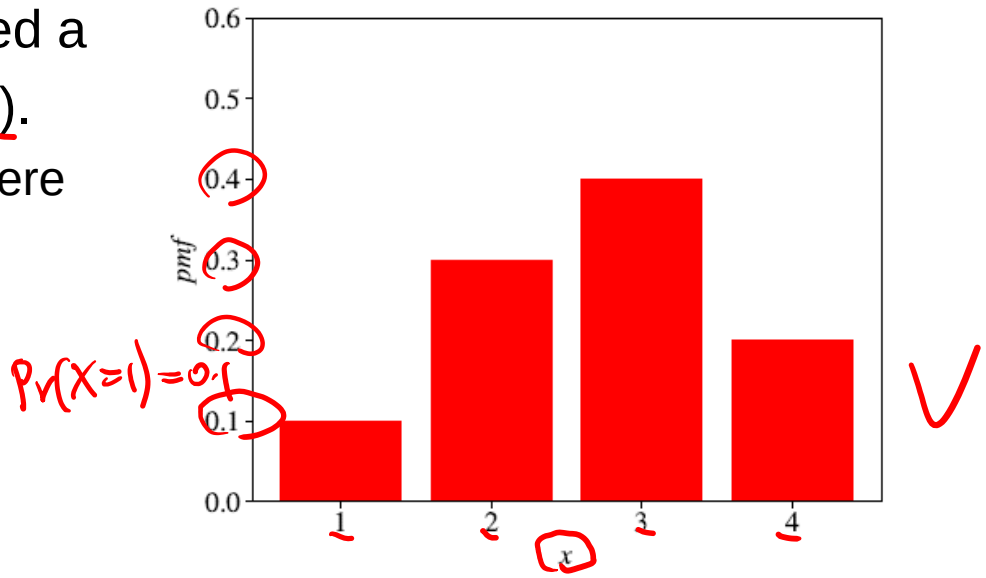
- We shall use $Pr(\cdot)$ for both the above cases

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



A probability mass function

Discrete random variable

- Let a **discrete** random variable X have k possible values $\{x_i\}_{i=1}^{k}$.
- The **expectation** of X denoted as $E(x)$ is given by,

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

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 $var(X)$, is defined as,

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

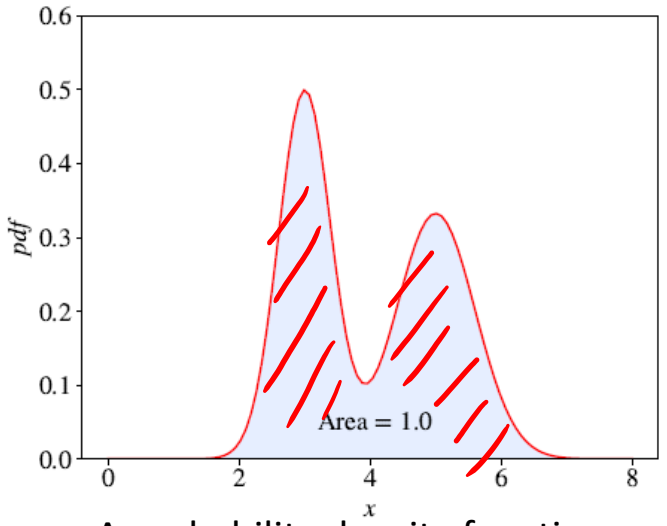
- For a **discrete random variable**, the standard deviation is given by

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

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.
  - 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**.**



A probability density function

Continuous random variable

- The **expectation** of a continuous random variable X is given by $E[x] \stackrel{\text{def}}{=} \int_R x\, f_X(x)\, \mathrm{d}x$ where f_X is the pdf of the variable X and $\int_R$ is the integral of function $x\, f_X$.

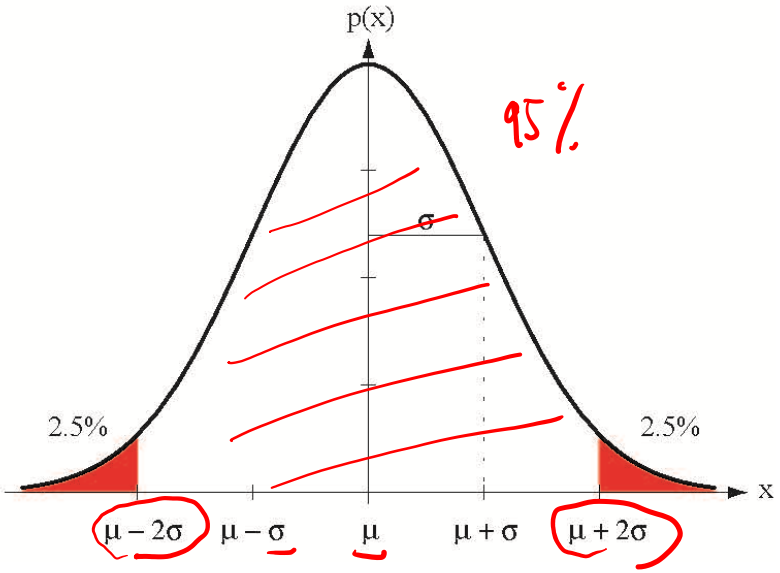
- The **variance** of a continuous random variable X is given by $\sigma^2 \stackrel{\text{def}}{=} \int_R (X-\mu)^2 f_X(x)\, \mathrm{d}x$

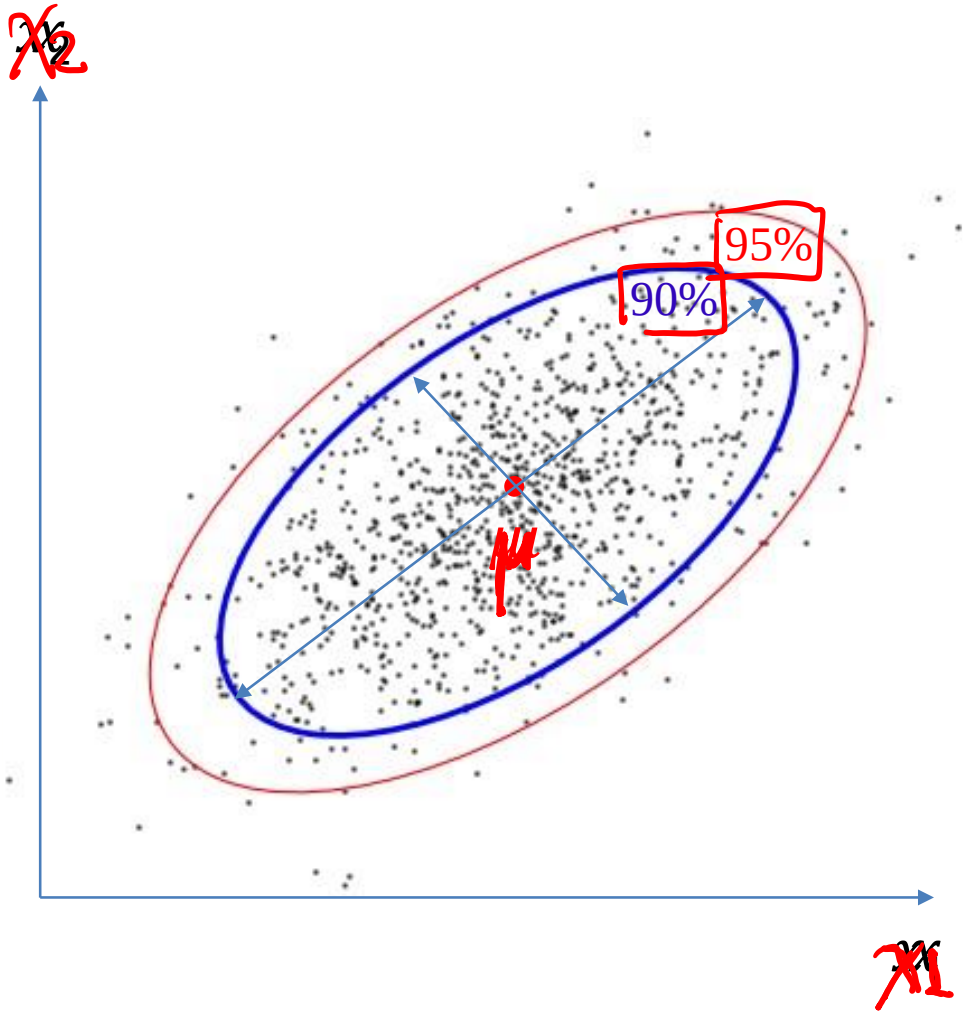
- Integral is an equivalent of the summation over all values of the function when the function has a continuous domain.

- 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 $\int_R f_X(x)\, \mathrm{d}x = 1$

Mean and Standard Deviation of a Gaussian Distribution





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.

I don't put back the first ball

- The space of outcomes of taking two balls sequentially without replacement:

  B–R, R–B

  – Thus having B-R is 1/2

- Mathematically:

  – $\Pr(x=B, y=R) = \Pr(y=R \mid x=B)\Pr(x=B) = 1 \times (1/2) = 1/2$

  given

  joint

  Conditional Probability

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

  Thus, Pr(y=G, x=B) = 1/6
- Mathematically:

$$\Pr(y=G, x=B) = \Pr(y=G \mid x=B)\Pr(x=B)$$
$$= (1/2) \times (1/3)$$
$$= 1/6$$

Two Basic Rules

- Sum Rule

$$\Pr(X = x) = \sum_{Y} \Pr(X = x, Y = y_i)$$

- Product Rule

Joint Prob. Conditional Prob.

$$\Pr(X = x, Y = y) = \Pr(Y = y | X = x)\, P(X = x)$$

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**):

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x|Y = y)\Pr(Y = y)}{\Pr(X = x)}$$

posterior — likelihood — prior — evidence

$$\Pr(Y|X) \cdot P(X) = P(XY) = P(X|Y) \cdot P(Y)$$

Example

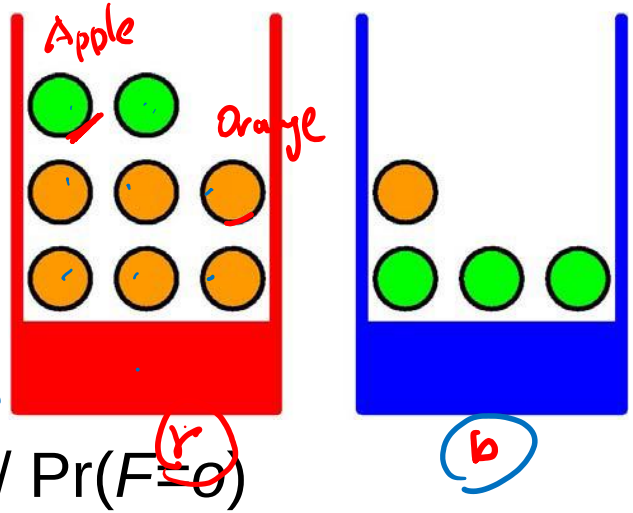
- Drawing a sample of fruit from a box
  - First pick a box, and then draw a sample of fruit from it
  - B: variable for Box, can be *r* (red) or *b* (blue)
  - F: variable for Fruit, can be *o* (orange) or *a* (apple)

- Pr(*B*=*r*)=0.4
- Pr(*F*=*o* | *B*=*r*)= 0.75
- Pr(*F*=*o*)= 0.45

prior

likelihood

evidence



- Pr(*B*=*r* | *F*=*o*) = Pr(*F*=*o* | *B*=*r*) * Pr(*B*=*r*) / Pr(*F*=*o*)

= 0.75*0.4/0.45 = 2/3

posterior

$1 - \frac{2}{3} = \frac{1}{3}$

$Pr(B=b \mid F=o) = ?$

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.

X	1	2	3	4	5
Pr[X]	0.1	0.05	0.05	0.6	*k* = 0.2

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

1) $0.1 + 0.05 + 0.2 = 0.35$

2) $1 - (0.05 + 0.6) = 0.35$