

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

January 5, 2023



Introduction to the Course

What can you expect to learn?

- Basics of Probability and Statistics
- Distributions
- Linear / Multiple Linear Regression
- Logistic Regression, SVM, ANN, Association Rules, Clustering
- Text Analytics

Course Logistics

- Book - No single text book for the course, various references will be provided
- TA - Shivani Jaiswal - shivani_j@ms.iitr.ac.in
- Attendance - Institute Rules will be enforced
- Evaluation
 - MTE - 25%
 - Assignments/in-class quiz - 10%
 - Group Assignment - 15%
 - Course Participation - 10%
 - ETE - 40%
- All course material will be posted on MS Teams platform

Learning Outcomes Session - 1

- Descriptive Statistics
- Concept of Probability
- Random Variable, Discrete and Continuous
- Expected Value, Variance, Correlation



Definitions of Statistics

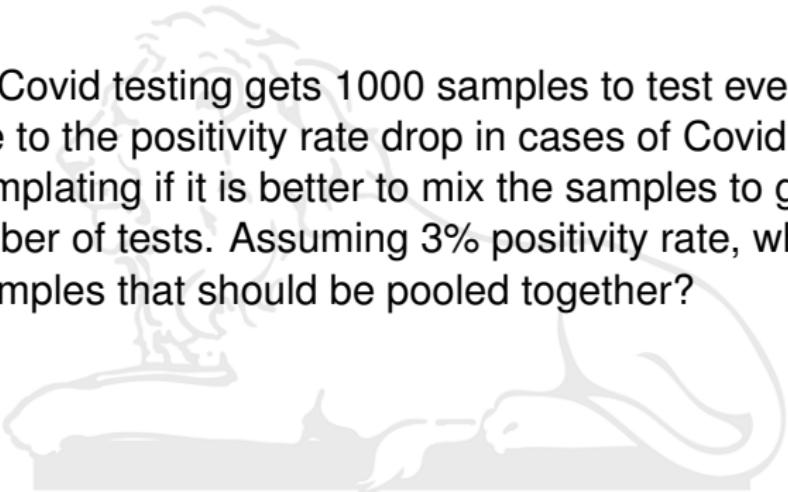
- ❑ Art of learning from data – Sheldon M. Ross, Introduction to Probability and Statistics for Engineers and Statisticians
- ❑ Statistics is a branch of mathematics working with data collection, organization, analysis, interpretation and presentation. - Wikipedia
- ❑ Statistics may be regarded as (i) the study of populations, (ii) as the study of variation, (iii) as the study of methods of the reduction of data. – Fisher, 1925

Descriptive Statistics

- ❑ Describing the data, Summarize the data, etc
- ❑ Using numbers, pictures etc.
- ❑ https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen
- ❑ https://www.ted.com/talks/hans_rosling_asia_s_rise_how_and_when

Pooled Testing Example - Warmup Problem

A LAB doing Covid testing gets 1000 samples to test everyday. However, due to the positivity rate drop in cases of Covid samples, the LAB is contemplating if it is better to mix the samples to get the result in lesser number of tests. Assuming 3% positivity rate, what is the number of samples that should be pooled together?



Summary Statistics

$$x_1, x_2, \dots, x_n$$

$$\mu = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\frac{\sum_{i=1}^n (x_i - \mu)^2}{n} = \sigma^2$$

- Measures of Central Tendency (mean, median, mode)
- Measures of Dispersion (Range, Variance)
- Chebyshev Inequality

10	8	0	0
10	9	5	10
10	10	10	10
10	11	15	10
10	12	20	20

(A) $\mu = 100$
 $\sigma^2 = 10000$

(B) $\mu = 100$
 $\sigma^2 = 100$

Summary Statistics(multiple data-sets)

Co-variance and Correlation

Mch
Gros
AL
MP
UP
UK

	2014	2022

Visually describing the data

Scatter Plot, Histogram

Need for visually describing the data

Anscombe's Quartet Counter Example

Box-Plot (in Python Example)

A Few More terminologies

- Cross-sectional Data
- Time Series Data
- Panel Data
- Qualitative Data
 - 1. Nominal
 - 2. Ordinal
- Quantitative Data
 - 1. Interval
 - 2. Ratio



Exit

- What is probability??
- Concept of Experiment, Sample Space, Events
- A number associated with each Sample Point $P(E_i)$
- Less than 1
- Sum of all probabilities = 1
- $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$
- Intersection of Events, Independent Events (Card Example)

Probability Concepts

- ❑ What is probability??
- ❑ Concept of Experiment, Sample Space, Events
- ❑ A number associated with each Sample Point $P(E_i)$
- ❑ Less than 1
- ❑ Sum of all probabilities = 1
- ❑ $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$
- ❑ Intersection of Events, Independent Events (Card Example)

Godbole's Problem



Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

January 17, 2023



Random Variable



- Definition - A mapping from sample space to real numbers
- Expectation, Variance, Correlation for Random Variables
- Examples - Throwing two dice, Sum of the two throws; Number of students who would come to 8am class

$$P(X=3) = ? \left(\frac{2}{36} \right)$$

Distribution of a Random Variable

- ❑ Let X be the random variable which can take values x_1, x_2, \dots
- ❑ The function $P(X = x_i) = f(x_i)$ is called the distribution (probability distribution) of X
- ❑ Joint Distribution of Random Variables

Basics about Random Variable

- ❑ A variable whose value depends on outcome of a random phenomenon
- ❑ A random variable is characterized by its distribution
- ❑ Sum of two or more different random variables is also a random variable, and thus will also have a distribution (we might not cover the mathematical tools required to find the distribution, but it is important to appreciate that it will have a distribution)
- ❑ Similarly, any other algebraic operation of two or more random variables also remain a random variable
- ❑ If X and Y are random variables, $X + Y$, $X - Y$, XY , $\frac{X}{Y}$ are all random variables

Expectation, Variance and Covariance Definition



Definition of Expectation

1, 2, 3, 4, 5, 6

3.5

$$\left(\frac{1}{6}\right) \left(\frac{1}{6}\right) - \left(\frac{1}{6}\right)$$

$0.16 \quad 0.16 \quad \dots \quad 0.16 \quad 0.95$

The expected value of a discrete random variable is

$$\mu = E(X) = \sum_x x p_x(x)$$

$$6 + 4 + 5 + \dots$$

$$6000$$

$$= 3.5$$

Variance of a random variable X

Let $E(X) = \mu$ (The Greek letter "mu").

$$\text{Var}(X) = E \left(\underline{(X - \mu)^2} \right)$$

Definition of Covariance

$$E(XY) - \mu_x \mu_y$$

Let X and Y be jointly distributed random variables with $E(X) = \mu_x$ and $E(Y) = \mu_y$. The covariance between X and Y is

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

- You could think of $\text{Var}(X) = E[(X - \mu_x)^2]$ as $\text{Cov}(X, X)$.

$$\begin{aligned} & \rightarrow E(XY - \mu_x Y - \mu_y X + \mu_x \mu_y) \\ &= E(XY) - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y \end{aligned}$$

Examples $X_i = \begin{cases} 1, & \text{if } i^{\text{th}} \text{ student gets their phone} \\ 0, & \text{otherwise} \end{cases}$

"N" students

- ❑ Number of matching - I take your mobile phones and return the mobile phones randomly back. How many students get their own mobile phone back? (X = this random variable). Find $E(X)$ and $\text{Var}(X)$
- ❑ Waiting time to get r unique objects - N different objects in a box. In each step, take out one object at random and keep it back. Repeat this until you get r unique objects. X = no. of trials required. Find $E(X)$
- ❑ Largest number in n drawings. A box contains balls numbered $1, 2, \dots, N$. Let X be the largest number drawn in n drawings, (done with replacement). Find $E(X)$

Standard Distributions

- ❑ Discrete
 - 1. Binomial
 - 2. Poisson
 - 3. Geometric
- ❑ Continuous
 - 1. Normal
 - 2. Uniform
 - 3. Exponential

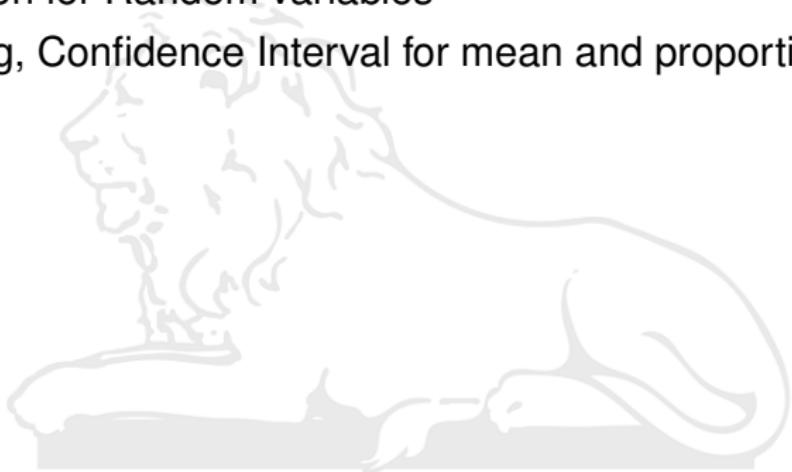


Simulation of Random numbers in Python



Learning Outcomes

- ❑ Standard discrete and Continuous Distributions
- ❑ Binomial, Poisson, Geometric, Normal, Uniform, Exponential
- ❑ Simulation for Random variables
- ❑ Sampling, Confidence Interval for mean and proportion



Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

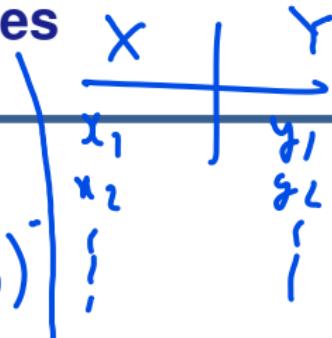
January 18, 2023



Independent Random Variables

$$P(X=x_i, Y=y_j)$$

$$= P(X=x_i) \cdot P(Y=y_j)$$



- If X and Y are independent random variables, then having any information about X doesn't change anything in distribution of Y (and vice versa).
- Check that for independent random variables - $E(XY) = E(X)E(Y)$.
- Is the reverse also true?

$$P(Y=2) = \frac{3}{9}$$

$$P(X=1) = \frac{2}{3}$$

		Y →		
		1	2	3
X ↓	0	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
	1	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$

$$P(X=1, Y=2) = \frac{2}{9}$$

Independent Random Variables

$$\left(\sum_i P(X=x_i) \cdot x_i \right) \left(\sum_j P(Y=y_j) \cdot y_j \right)$$

- If X and Y are independent random variables, then having any information about X doesn't change anything in distribution of Y (and vice versa)
- Check that for independent random variables - $E(XY) = E(X)E(Y)$.
- Is the reverse also true?

$$\sum_i \sum_j \underbrace{P(X=x_i) \cdot P(Y=y_j)}_{\text{Independent}} \cdot x_i \cdot y_j$$

$$\sum_i \sum_j P(X=x_i, Y=y_j) \underbrace{x_i y_j}_{\text{Independent}} = E(XY)$$

Independent Random Variables

$E(x) = 1$
 $E(y) = \frac{1}{2}$

Expt: Tossing a coin twice
 $X = \text{no. of heads in these 2 tosses}$

$E(xy) = \frac{1}{2}$

- If X and Y are independent random variables, then having any information about X doesn't change anything in distribution of Y (and vice versa)
- Check that for independent random variables - $E(XY) = E(X)E(Y)$.

Q Is the reverse also true?

$$Y = \begin{cases} 1, & \text{if both tosses are same} \\ 0, & \text{else} \end{cases}$$

$P(X=2, Y=0) = ??$

XY	2	1	0
2	1	0	0
1	0	0	0
0	0	1	0

Examples

'N'

- Number of matching - I take your mobile phones and return the mobile phones randomly back. How many students get their own mobile phone back? (X = this random variable). Find $E(X)$ and $\text{Var}(X)$!!
- Waiting time to get r unique objects - N different objects in a box. In each step, take out one object at random and keep it back. Repeat this until you get r unique objects. X = no. of trials required. Find $E(X)$
- Largest number in n drawings. A box contains balls numbered 1,2,...,N. Let X be the largest number drawn in n drawings, (done with replacement). Find $E(X)$

Standard Distributions

- Discrete
 - 1. Binomial
 - 2. Poisson
 - 3. Geometric
- Continuous
 - 1. Normal
 - 2. Uniform
 - 3. Exponential



Binomial Random Variable

$$E(X) = p, \quad \text{Var}(X) = p(1-p)$$

- ❑ Bernoulli Random Variable - Do an experiment once, probability of success = p . ($X = 1$, if success, 0 otherwise). Find $E(X)$ and $\text{Var}(X)$
- ❑ Binomial Random Variable - Repeat independent Bernoulli trials n times. Y = total number of successes in these n trials.
- ❑ Find $E(Y)$ and $\text{Var}(Y)$

Binomial Random Variable

$$Y = X_1 + X_2 + \dots + X_n$$

$$E(Y) = np, \quad \text{Var}(Y) = np(1-p)$$

- Bernoulli Random Variable - Do an experiment once, probability of success = p. ($X = 1$, if success, 0 otherwise). Find $E(X)$ and $\text{Var}(X)$
- Binomial Random Variable - Repeat independent Bernoulli trials n times. Y = total number of successes in these n trials.
- Find $E(Y)$ and $\text{Var}(Y)$

$$(0, 1, 2, \dots, n)$$

$$P(Y=g_n) = {}^n C_{g_n} p^{g_n} (1-p)^{n-g_n}$$

Check that $\sum_{n=0}^{\infty} g_n \cdot P(Y=g_n) = np$

Binomial Random Variable

- ❑ Bernoulli Random Variable - Do an experiment once, probability of success = p . ($X = 1$, if success, 0 otherwise). Find $E(X)$ and $\text{Var}(X)$
- ❑ Binomial Random Variable - Repeat independent Bernoulli trials n times. Y = total number of successes in these n trials.
- ❑ Find $E(Y)$ and $\text{Var}(Y)$

Poisson Random Variable

$$(e^{-\lambda} \frac{\lambda^i}{i!}) = P(X=i)$$

- Let the number of events happening in a given period of time be X
- If X follows the following probability distribution, we say that X follows Poisson distribution
- $P(X = i) = \frac{e^{-\lambda} \lambda^i}{i!}; i = 0, 1, 2, \dots$
- Find E(X) and Var(X)

$$E(X) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{i \lambda^i}{i!} = e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \lambda$$

Poisson Random Variable

$$\text{Var}(x) = E(x^2) - (E(x))^2$$
$$(\lambda^2 + \lambda) - \lambda^2 = \lambda$$

- Let the number of events happening in a given period of time be X
- If X follows the following probability distribution, we say that X follows Poisson distribution
- $P(X = i) = \frac{e^{-\lambda} \lambda^i}{i!}; i = 0, 1, 2, \dots$
- Find E(X) and Var(X)

$$\sum_{i=0}^{\infty} i^2 \left(\frac{e^{-\lambda} \lambda^i}{i!} \right)$$

$$\boxed{E(x) = \lambda}$$
$$\boxed{\text{Var}(x) = \lambda}$$

Random Sum of Random Numbers

$$E(c) = 1$$

$$C = \begin{cases} 0, & 0.1 \\ 1, & 0.8 \\ 2, & 0.1 \end{cases}$$

$$\text{Var}(c) = 0.2$$

- In a tea shop, the number of customers coming in a given day follows a Poisson distribution with parameter 500
- Each customer makes the purchase as per the following distribution - a.) No purchase with probability = 0.1, one cup of tea with probability = 0.8, two cups of tea with probability = 0.1
- Let X denote the number of tea cups sold in a day. What is $E(X)$ and $\text{Var}(X)$

$N \sim \text{Poisson}(500)$

$$E(x) = 500$$

$$\text{Var}(x) = 100$$

$$X = C_1 + C_2 + \dots + C_N$$

Simulation of Random numbers in Python



Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

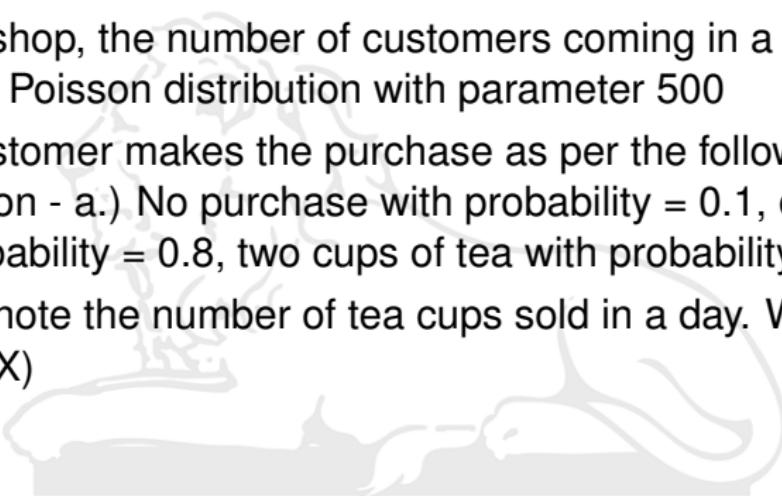
Department of Management Studies

January 24, 2023



Random Sum of Random Numbers

- ❑ In a tea shop, the number of customers coming in a given day follows a Poisson distribution with parameter 500
- ❑ Each customer makes the purchase as per the following distribution - a.) No purchase with probability = 0.1, one cup of tea with probability = 0.8, two cups of tea with probability = 0.1
- ❑ Let X denote the number of tea cups sold in a day. What is $E(X)$ and $\text{Var}(X)$



Conditional Expectation

$$P(X=2|Y=-1) = \frac{0.05}{0.2}$$

Joint Probability		X		
		2	3	5
Y	-1	0.05	0.05	0.1
	0	0.2	0.1	0.05
	2	0.01	0.02	0.05
	5	0.07	0.25	0.05

$$P(E(X|Y) = 3.75) = 0.2$$

What is the value of $E(X|Y)$?

Is it a function of Y ?

Is it a random variable?

What is $E(X)$?

Is it same as $E_y(E_x(X|Y))$

$$E(X|Y=-1) = 3.75$$

$$E(X|Y=0) = \underline{\hspace{2cm}}$$

$$E(X|Y=2) = \underline{\hspace{2cm}}$$

$$E(X|Y=5) = \underline{\hspace{2cm}}$$

Conditional Expectation

Joint Probability		X		
Y		2	3	5
	-1	0.05	0.05	0.1
	0	0.2	0.1	0.05
	2	0.01	0.02	0.05
	5	0.07	0.25	0.05

What is the value of $E(X|Y)$?

Is it a function of Y ?

Is it a random variable?

What is $E(X)$?

Is it same as $E_y(E_x(X|Y))$

Conditional Expectation

Joint Probability		X		
Y		2	3	5
	-1	0.05	0.05	0.1
	0	0.2	0.1	0.05
	2	0.01	0.02	0.05
	5	0.07	0.25	0.05

What is the value of $E(X|Y)$?

Is it a function of Y ?

Is it a random variable?

What is $E(X)$?

Is it same as $E_y(E_x(X|Y))$

Conditional Expectation

Joint Probability		X		
Y		2	3	5
	-1	0.05	0.05	0.1
	0	0.2	0.1	0.05
	2	0.01	0.02	0.05
	5	0.07	0.25	0.05

$$E_Y(E_X(X|Y)) = E(X)$$

What is the value of $E(X|Y)$?

Is it a function of Y ?

Is it a random variable?

What is $E(X)$?

Is it same as $E_y(E_x(X|Y))$

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??

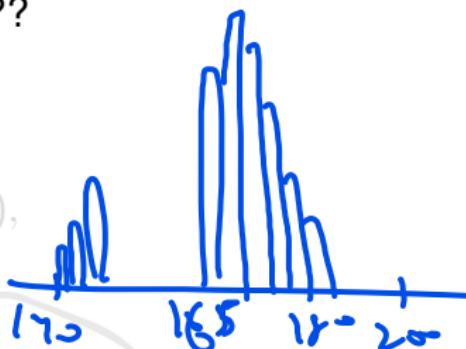
Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X \leq x),$$

or the *probability density function* (pdf):

$$\rho(x) dx = \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \rightarrow 0$, we find



$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^x \rho(t) dt.$$

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??

Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X \leq x),$$

or the *probability density function* (pdf):

$$\rho(x) dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \rightarrow 0$, we find

$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^x \rho(t) dt.$$

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??

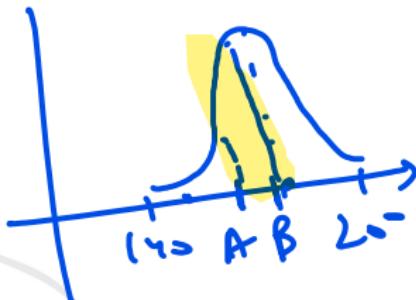
Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X \leq x),$$

or the *probability density function* (pdf):

$$\rho(x) dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \rightarrow 0$, we find



$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^x \rho(t) dt.$$

Expected Value

The expected value of a continuous random variable X is then defined by

$$E(X) = \int_{-\infty}^{\infty} x \rho(x) dx.$$

Note that by definition, $\int_{-\infty}^{\infty} \rho(x) dx = 1$. The expected value of X^2 is

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \rho(x) dx,$$

and the variance is again defined as $E(X^2) - (E(X))^2$.

Expected Value

The expected value of a continuous random variable X is then defined by

$$E(X) = \int_{-\infty}^{\infty} x \rho(x) dx.$$

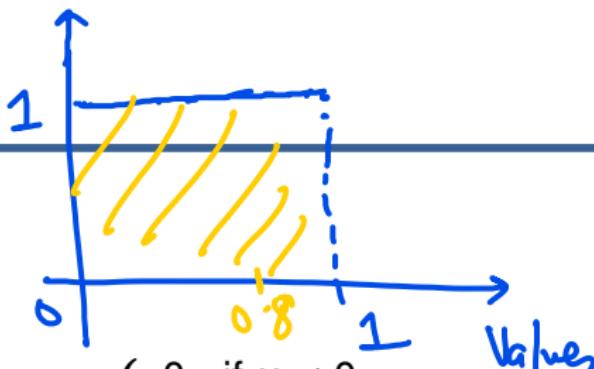
Note that by definition, $\int_{-\infty}^{\infty} \rho(x) dx = 1$. The expected value of X^2 is

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \rho(x) dx,$$

and the variance is again defined as $E(X^2) - (E(X))^2$.

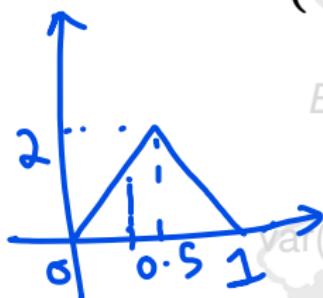
Uniform Distribution

- Density



Example: Uniform Distribution in $[0, 1]$.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$



$$\rho(x < 0.3)$$

$$E(X) = \int_{-\infty}^{\infty} x \rho(x) dx = \int_0^1 x dx = \frac{1}{2},$$

$$\text{Var}(X) = \int_0^1 x^2 dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

$$\rho(x > 0.8) = 0.2 = 0.18$$

Uniform Distribution

Example: Uniform Distribution in $[0, 1]$.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x\rho(x) dx = \int_0^1 x dx = \frac{1}{2},$$

$$\text{var}(X) = \int_0^1 x^2 dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Normal Distribution

Example: Normal (Gaussian) Distribution, Mean μ , Variance σ^2 .

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Simulation of Random numbers in Python



Practice Problem

Random Variable X follows uniform distribution from $[0,1]$, Random Variable Y follows same distribution and is independent of X.
What is the distribution of $X+Y$?

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

January 27, 2023



Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??

Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X \leq x),$$

or the *probability density function* (pdf):

$$\rho(x) dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \rightarrow 0$, we find

$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^x \rho(t) dt.$$

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??

Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X \leq x),$$

or the *probability density function* (pdf):

$$\rho(x) dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \rightarrow 0$, we find

$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^x \rho(t) dt.$$

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??

Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X \leq x),$$

or the *probability density function* (pdf):

$$\rho(x) dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \rightarrow 0$, we find

$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^x \rho(t) dt.$$

Expected Value



The expected value of a continuous random variable X is then defined by

$$E(X) = \int_{-\infty}^{\infty} x \rho(x) dx.$$

Note that by definition, $\int_{-\infty}^{\infty} \rho(x) dx = 1$. The expected value of X^2 is

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \rho(x) dx,$$

and the variance is again defined as $E(X^2) - (E(X))^2$.

Expected Value

The expected value of a continuous random variable X is then defined by

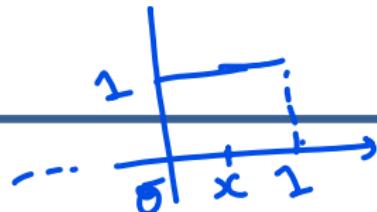
$$E(X) = \int_{-\infty}^{\infty} x \rho(x) dx.$$

Note that by definition, $\int_{-\infty}^{\infty} \rho(x) dx = 1$. The expected value of X^2 is

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \rho(x) dx,$$

and the variance is again defined as $E(X^2) - (E(X))^2$.

Uniform Distribution



X -

Example: Uniform Distribution in $[0, 1]$.

P(X ≤ x)

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$

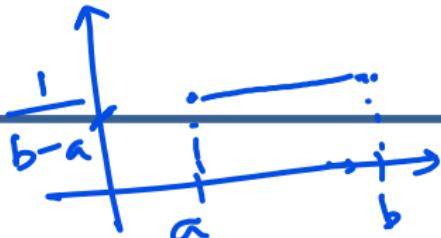
//

$$E(X) = \int_{-\infty}^{\infty} x \rho(x) dx = \int_0^1 x dx = \frac{1}{2},$$

$$\text{var}(X) = \int_0^1 x^2 dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Uniform Distribution

$[a, b]$



Example: Uniform Distribution in $[0, 1]$.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x\rho(x) dx = \int_0^1 x dx = \frac{1}{2},$$

$$\frac{a+b}{2}$$

$$\text{var}(X) = \int_0^1 x^2 dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

$$\frac{(b-a)^2}{12}$$

Practice Problem

$$Z = X + Y$$

(0, 2)

$$P(Z \leq z) = \frac{1}{2}z^2, 0 < z < 1$$

$$P(Z \leq z) = 1 - \frac{1}{2}(2-z)^2 \quad z > 1$$

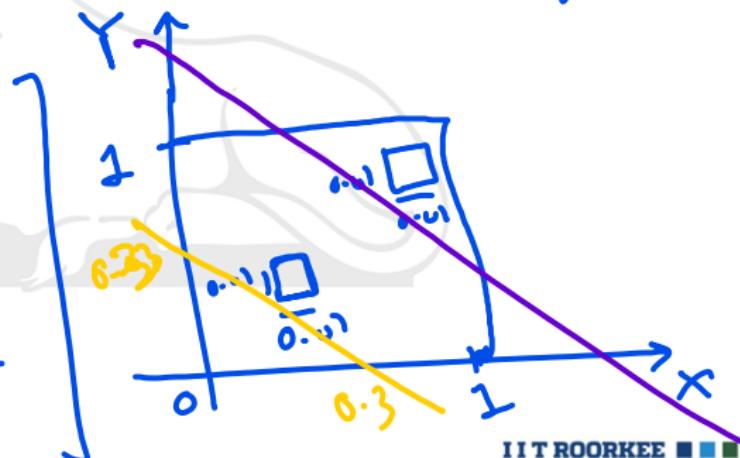


Random Variable X follows uniform distribution from [0, 1], Random Variable Y follows same distribution and is independent of X.

What is the distribution of X+Y?

$$f(z) = z; 0 < z \leq 1$$

$$= 2-z; 1 < z \leq 2$$



Normal Distribution

$X \sim \text{Normal}(10, 25)$

Eg. - $\mu = 10, \sigma^2 = 25$

Example: Normal (Gaussian) Distribution, Mean μ , Variance σ^2 .

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Why is this weird density called normal?

$$P(0 < X < 20) = ?? P(X < 20) - P(X < 0)$$

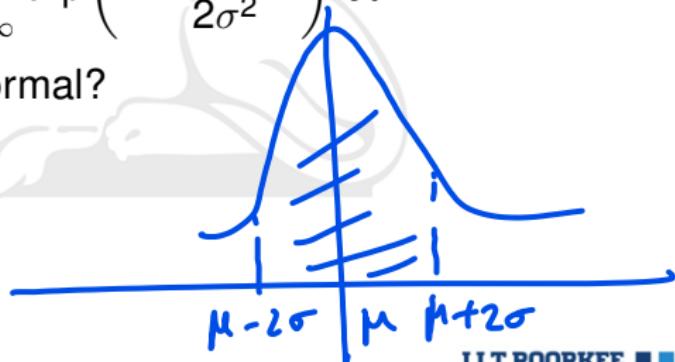
Normal Distribution

Example: Normal (Gaussian) Distribution, Mean μ , Variance σ^2 .

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \rightarrow$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Why is this weird density called normal?



Central Limit Theorem

Theorem

Let $\{X_k\}$ be a sequence of n mutually independent random variables having a common distribution, and mean (μ) and variance (σ^2) exists. Assuming, n to be large, the average of these random variables \bar{X} follows approximately normal distribution with

1. mean = μ
2. variance = $\frac{\sigma^2}{n}$

What is meant by large n ? Typically, $n \geq 30$

Central Limit Theorem - Special Case

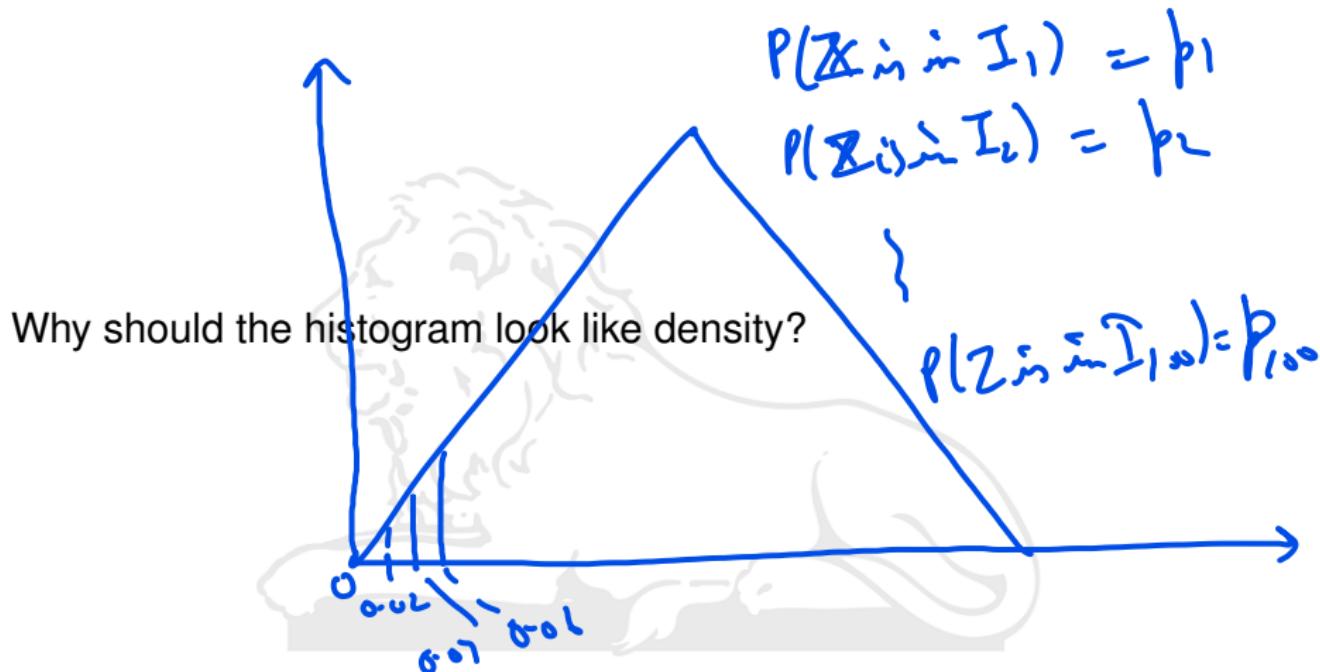
Theorem

If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean \bar{X} is approximately normal with

1. mean = μ
2. variance = $\frac{\sigma^2}{n}$

What is meant by large n ? Typically, $n \geq 30$

Simulation of Random numbers in Python



Simulation of Random numbers in Python

Inverse Transform Method??



Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

January 31, 2023



Uniform Distribution

Example: Uniform Distribution in $[0, 1]$.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x \rho(x) dx = \int_0^1 x dx = \frac{1}{2},$$

$$\text{var}(X) = \int_0^1 x^2 dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Uniform Distribution

Example: Uniform Distribution in $[0, 1]$.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x\rho(x) dx = \int_0^1 x dx = \frac{1}{2},$$

$$\text{var}(X) = \int_0^1 x^2 dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Practice Problem

$$P(Z \leq 0.7 | Y \in (0.3, 0.3 + dy)) = P(X + Y \leq 0.7 | Y \in \dots)$$
$$= P(X \leq 0.7 - 0.3)$$

Random Variable X follows uniform distribution from [0,1], Random Variable Y follows same distribution and is independent of X.

What is the distribution of X+Y?

Case-1: $z \leq 1$

$$(z - 0.3)$$

$$P(Z \leq z) = \sum_{y \in \dots} P(Z \leq z | Y \in (y, y + dy)). P(Y \in \dots)$$
$$\int_0^z (z-y) dy = (z^2/2)$$

Normal Distribution

Example: Normal (Gaussian) Distribution, Mean μ , Variance σ^2 .

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Why is this weird density called normal?

Normal Distribution

Example: Normal (Gaussian) Distribution, Mean μ , Variance σ^2 .

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Why is this weird density called normal?

Central Limit Theorem

Theorem

Let $\{X_k\}$ be a sequence of n mutually independent random variables having a common distribution, and mean (μ) and variance (σ^2) exists. Assuming, n to be large, the average of these random variables \bar{X} follows approximately normal distribution with

1. mean = μ
2. variance = $\frac{\sigma^2}{n}$

What is meant by large n ? Typically, $n \geq 30$

Central Limit Theorem - Special Case

Theorem

If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean \bar{X} is approximately normal with

1. mean = μ
2. variance = $\frac{\sigma^2}{n}$

What is meant by large n ? Typically, $n \geq 30$

Simulation of Random numbers in Python

$f(x)$ — PDF

$F(x)$ — CDF : $P(X \leq x)$

$Z \sim \text{unif}(0, 1)$

Inverse Transform Method??

$F^{-1}(x)$

$\tilde{F}^{-1}(z)$

$$P(F^{-1}(z) \leq z) = P(z \leq F(z)) = F(z)$$

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 2, 2023

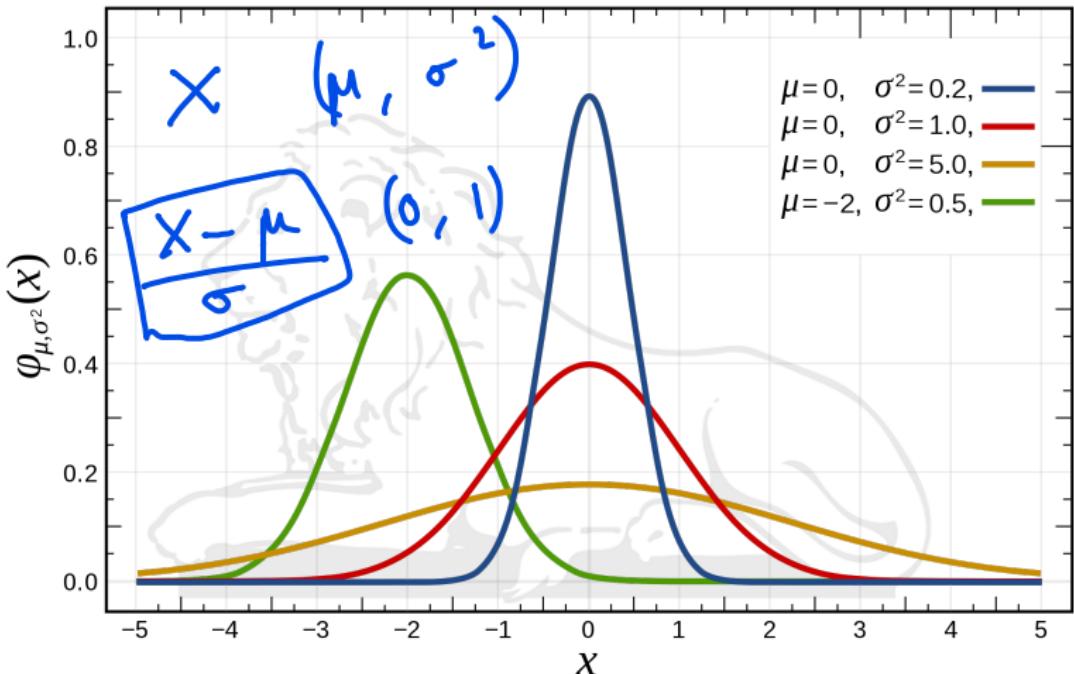


Basics about Random Variable

- ❑ A variable whose value depends on outcome of a random phenomenon
- ❑ A random variable is characterized by its distribution
- ❑ Sum of two or more different random variables is also a random variable, and thus will also have a distribution (we might not cover the mathematical tools required to find the distribution, but it is important to appreciate that it will have a distribution)
- ❑ Similarly, any other algebraic operation of two or more random variables also remain a random variable
- ❑ If X and Y are random variables, $X + Y$, $X - Y$, XY , $\frac{X}{Y}$ are all random variables

Standard Normal Distribution

A normal distribution with mean 0 and standard deviation as 1



Chi-square distribution

$$U_5 = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2$$

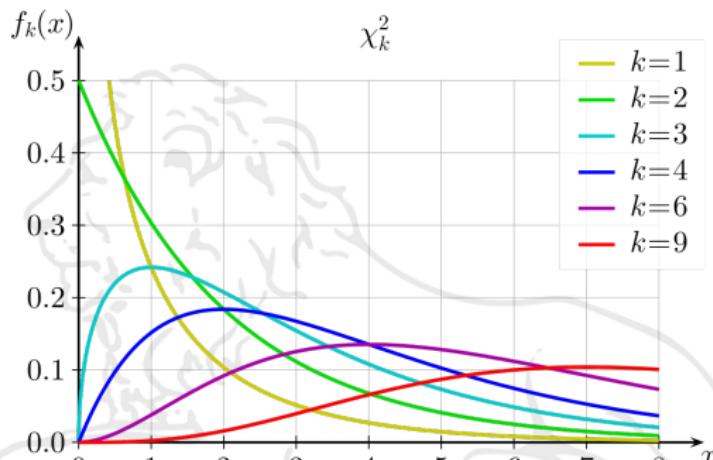
If Z_1, Z_2, \dots, Z_n are all independent and normally distributed with mean

0 and variance 1, then $U = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$

Alternatively, U is said to follow a χ^2 distribution with n degrees of freedom

$$\mathbb{E}(U_5) = ?? - 5$$

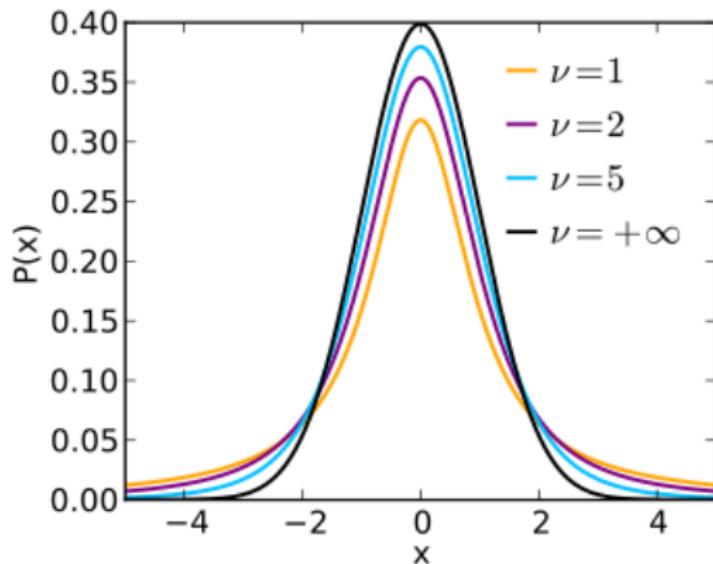
Chi-square distribution



Source - https://en.wikipedia.org/wiki/Chi-squared_distribution

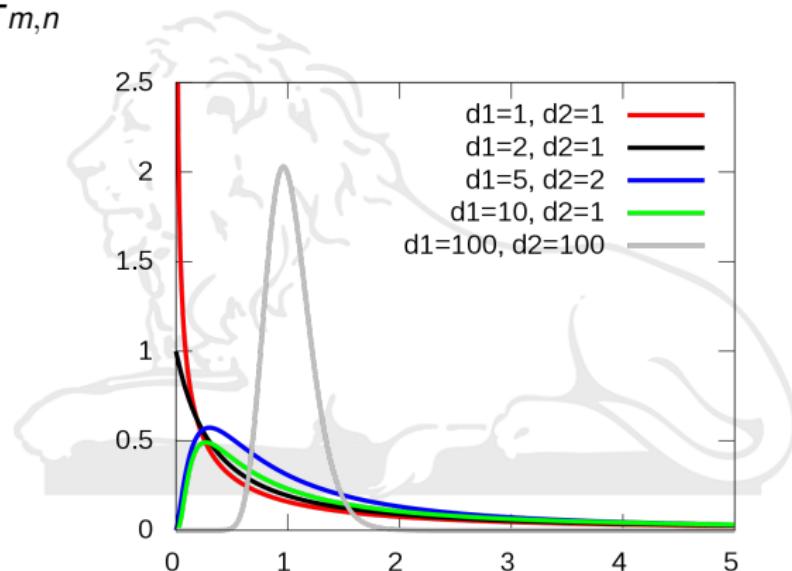
t-distribution

If $Z \sim N(0, 1)$ and $U \sim \chi_n^2$ are independent, then the distribution of $\frac{Z}{\sqrt{\frac{U}{n}}}$ is called the t-distribution with n degrees of freedom

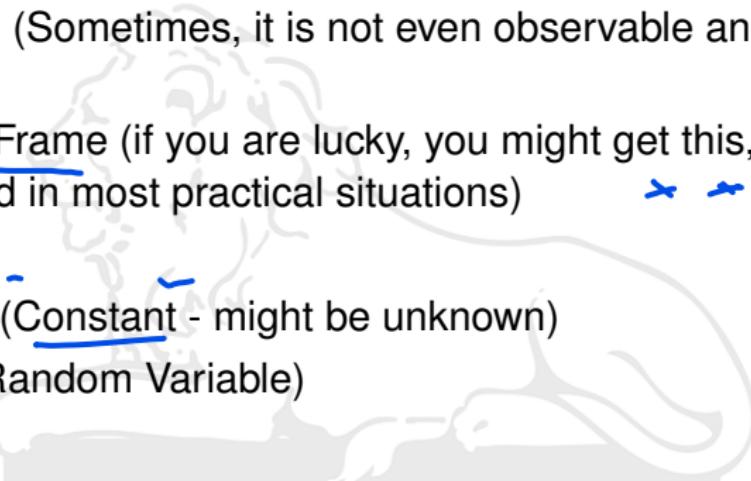


F distribution

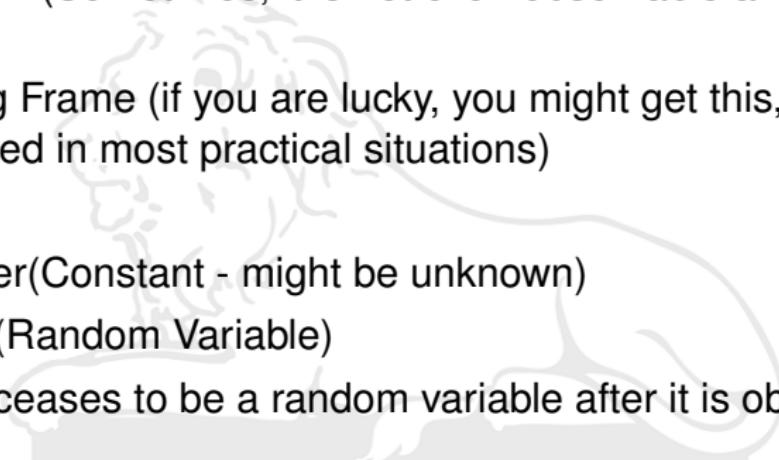
If $U \sim \chi_m^2$ and $V \sim \chi_n^2$ are independent, then the distribution of $\frac{U}{\frac{m}{n}V}$ is called the F-distribution with m and n degrees of freedom and is denoted by $F_{m,n}$



Basic Ideas of Sampling

- 1. Population (Sometimes, it is not even observable and only abstract)
- 2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations) 
- 3. Subject
- 4. Parameter(Constant - might be unknown)
- 5. Statistic (Random Variable)



Basic Ideas of Sampling

- 
1. Population (Sometimes, it is not even observable and only abstract)
 2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations)
 3. Subject
 4. Parameter(Constant - might be unknown)
 5. Statistic (Random Variable)
 6. Statistic ceases to be a random variable after it is observed

Types of Sampling

- Simple Random Sample With replacement
- Simple Random Sample without replacement
- Cluster Sampling
- Stratified Sampling

Potential Causes of Bias

- Convenience Sampling
- Volunteer Sampling
- Systematic Sampling
- Non-response Bias
- Response Bias

Does that mean we shouldn't use any of these types of sampling??

Potential Causes of Bias

- Convenience Sampling
- Volunteer Sampling
- Systematic Sampling
- Non-response Bias
- Response Bias

Does that mean we shouldn't use any of these types of sampling??
NO, one can use, but with caution. Make sure it is not leading to a systematic error

Sampling Using Python in Presence of Sampling Frame

```
import random  
from random import choices  
choices(samplingframe, k = sample size)
```

What after sampling?

- Ask, why did we do sampling? Objective is to learn about the population
- Statistical Inference - Learn about parameters from sample statistic
- Usually, the quantities of interest are mean and proportion in the population (depending on the context)
- We deal with them separately

Estimating from the statistic

- ❑ The rational behind estimating is expectation of statistic should be equal to the population parameter
- ❑ Biased and Unbiased Estimator
- ❑ Variance of estimate should be minimized to the extent possible

Errors in the Process of Estimation

- ❑ **Sampling Error** - Because we are only considering a subset of population, the point estimate is rarely exactly correct. Unavoidable error, but we can estimate the error and hence have some control over it
- ❑ **Non-sampling Error** - If there is bias in the observations, or sampling wasn't done properly. Can't be dealt with mathematically. Should be avoided

Estimating Population Mean from Sample Mean

N = 5000

n = 100

1. Let the true values in the population be $\underline{A_1}, \underline{A_2}, \underline{A_3}, \dots, \underline{A_N}$
2. Population mean is denoted by μ and equals $\frac{\sum_{i=1}^N A_i}{N}$
3. Also, population variance is denoted by σ^2 and equals $\frac{\sum_{i=1}^N (A_i - \mu)^2}{N}$
- with replacement -
4. Let the sample be a SRS of size n
5. Observations are $\underline{X_1}, \underline{X_2}, \dots, \underline{X_n}$
6. Sample mean is denoted by \bar{X} and defined as follows
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$
7. $E(\bar{X}) = \mu$

$$\boxed{\text{Var}(X_i) = \sigma^2} \quad \boxed{E(X_i) = \mu}$$

Estimating Population Variance from Sample Observations

- ❑ If the sampling scheme is WITH REPLACEMENT

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- ❑ If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N - 1}{N} \right) \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Why is estimating population variance important?

Estimating Population Variance from Sample Observations

- ❑ If the sampling scheme is WITH REPLACEMENT

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- ❑ If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N - 1}{N} \right) \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Why is estimating population variance important?

To get an idea about error in estimation of sample mean

Standard Error in Sample Mean

- ❑ If the sampling scheme is WITH REPLACEMENT

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- ❑ If the sampling scheme is WITHOUT REPLACEMENT

$$\text{Var}(\bar{X}) = \left(\frac{N-n}{N-1} \right) \frac{\sigma^2}{n}$$

- ❑ $\frac{N-n}{N-1}$ is called the finite population correction
- ❑ Typically, can be ignored if sampling fraction $\frac{n}{N} \leq 0.05$
- ❑ Standard deviation of \bar{X} is called the Standard error of the sample mean
- ❑ Do we know σ^2 ? What is the remedy??

Proof for $n - 1$ in the denominator

Proof of *sample variance is the unbiased estimator of population variance* (Sampling is done with replacement)

Let the true values in the population be $A_1, A_2, A_3, \dots, A_N$

Population mean is denoted by μ and equals $\frac{\sum_{i=1}^N A_i}{N}$

Also, population variance is denoted by σ^2 and equals $\frac{\sum_{i=1}^N (A_i - \mu)^2}{N}$

We will discuss the case when sampling is done WITH REPLACEMENT

Proof for $n - 1$ in the denominator

We now take a sample of size n (With replacement). Let the values that come in the sample are $X_1, X_2, X_3, \dots, X_n$

Our objective now is to use these n numbers to estimate population variance. The rational behind our approach is that the expectation of the estimate should be equal to the population variance.

We note that because the samples are taken with replacement, X_1, X_2, \dots, X_n are independent random variables.

We have seen that sample mean is an unbiased estimator of population mean. Sample mean is denoted by \bar{X} and defined as follows

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

We now claim that if we define the **sample variance**(denoted by s_X^2) as follows, it would be an unbiased estimator of population variance.

Proof for $n - 1$ in the denominator

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \text{ We need to show that } E(s_X^2) = \sigma^2$$

$$E(s_X^2) = E\left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}\right)$$

$$\implies E(s_X^2) = E\left(\frac{\sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2)}{n-1}\right)$$

$$\implies E(s_X^2) = E\left(\frac{\left(\sum_{i=1}^n X_i^2 - \sum_{i=1}^n 2X_i\bar{X} + \sum_{i=1}^n \bar{X}^2\right)}{n-1}\right)$$

$$\implies E(s_X^2) = E\left(\frac{\left(\sum_{i=1}^n X_i^2 - 2\bar{X}\sum_{i=1}^n X_i + n\bar{X}^2\right)}{n-1}\right)$$

$$\implies E(s_X^2) = E\left(\frac{\left(\sum_{i=1}^n X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2\right)}{n-1}\right)$$

Proof for $n - 1$ in the denominator

$$\begin{aligned}\implies E(s_X^2) &= E\left(\frac{\left(\sum_{i=1}^n X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2\right)}{n-1}\right) \\ \implies E(s_X^2) &= E\left(\frac{\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)}{n-1}\right) \\ \implies E(s_X^2) &= \frac{\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)}{n-1}\end{aligned}$$

Since, X_i can take any values from A_1, A_2, \dots, A_N each with probability $\frac{1}{N}$, X_i^2 can take any values from $A_1^2, A_2^2, \dots, A_N^2$ each with probability $\frac{1}{N}$,

$$E(X_i^2) = \frac{A_1^2 + A_2^2 + \cdots + A_N^2}{N}, \text{ for all } i.$$

Proof for $n - 1$ in the denominator

We can easily see the following result. As an exercise, it is recommended that you do this.

$$\sigma^2 = \frac{A_1^2 + A_2^2 + \cdots + A_N^2}{N} - \mu^2$$
$$\implies E(X_i^2) = \sigma^2 + \mu^2, \text{ for all } i.$$

$$\implies E(s_X^2) = \frac{n\sigma^2 + n\mu^2 - nE(\bar{X}^2)}{n-1} \quad (1)$$

We would now compute $E(\bar{X}^2)$ and substitute in the above equation (1) to see the result.

$$E(\bar{X}^2) = E\left(\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)^2\right)$$

$$\implies E(\bar{X}^2) =$$

$$E\left(\frac{X_1^2 + X_2^2 + \cdots + X_n^2 + 2X_1X_2 + 2X_1X_3 + \cdots + 2X_{n-1}X_n}{n^2}\right)$$

Proof for $n - 1$ in the denominator

$$\implies E(\bar{X}^2) =$$

$$\frac{E(X_1^2) + E(X_2^2) + \dots + E(X_n^2) + 2E(X_1 X_2) + 2E(X_1 X_3) + \dots + 2E(X_{n-1} X_n)}{n^2}$$

$$\implies E(\bar{X}^2) =$$

$$\frac{n\sigma^2 + n\mu^2 + 2E(X_1 X_2) + 2E(X_1 X_3) + \dots + 2E(X_{n-1} X_n)}{n^2}$$

To deal with terms like $E(X_j X_k)$, we will use a result about expectations of product of independent random variables.

If X and Y are independent random variables, the following result holds.

$$E(XY) = E(X)E(Y)$$

Proof for $n - 1$ in the denominator

Hence, $E(X_j X_k) = E(X_j)E(X_k) = (\mu)(\mu) = \mu^2$, for all j and k (because X_j and X_k are independent random variables when the sampling scheme is with replacement)

Also, there are $\frac{n(n-1)}{2}$ such terms. Thus,

$$\Rightarrow E(\bar{X}^2) = \frac{n\sigma^2 + n\mu^2 + 2\frac{n(n-1)}{2}\mu^2}{n^2}$$

$$\Rightarrow E(\bar{X}^2) = \frac{\sigma^2 + \mu^2 + (n-1)\mu^2}{n}$$

We now substitute the value of $E(\bar{X}^2)$ in equation (1) to obtain-

$$\Rightarrow E(s_X^2) = \frac{n\sigma^2 + n\mu^2 - \sigma^2 - \mu^2 - (n-1)\mu^2}{n-1}$$

$$\Rightarrow E(s_X^2) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

Thus, proved.

Remarks on the result

Remark - The variance of the numbers $X_1, X_2, X_3, \dots, X_n$ is equal to

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

s_X^2 is a quantity obtained by using the numbers $X_1, X_2, X_3, \dots, X_n$ in order to provide a unbiased estimator of σ^2

Central Limit Theorem

- ❑ Can we do better about our inference from sample mean??
- ❑ Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?

Central Limit Theorem

- ❑ Can we do better about our inference from sample mean??
- ❑ Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?
- ❑ Thank you Central Limit Theorem

Central Limit Theorem

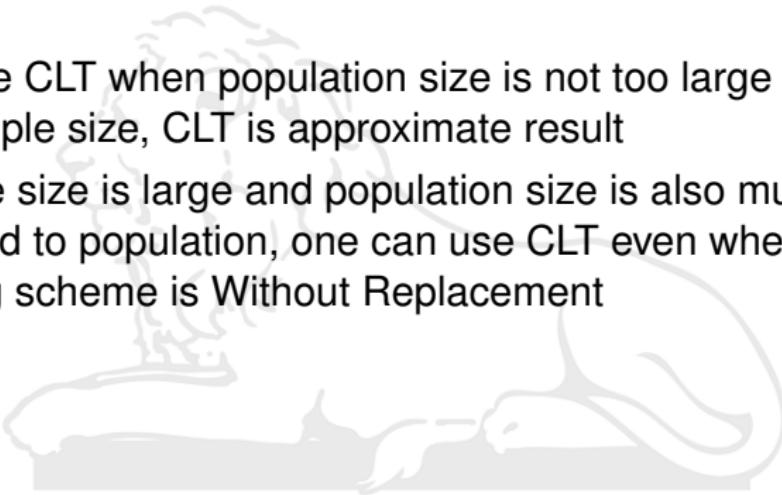
Theorem

If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean \bar{X} is approximately normal with

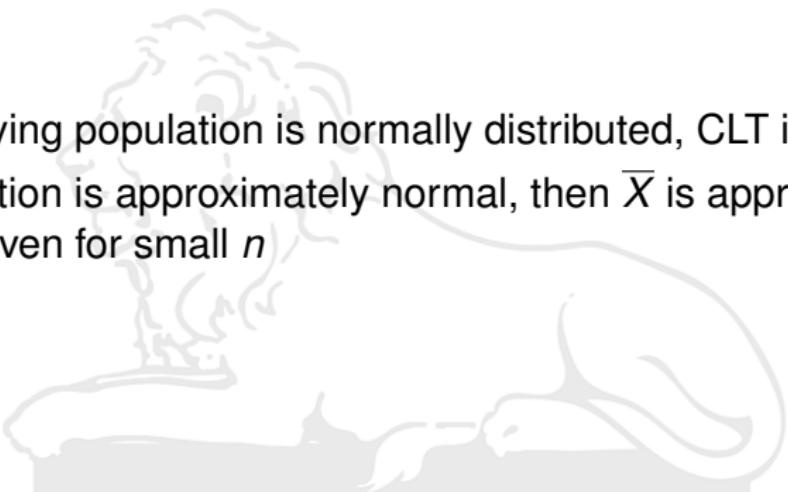
1. mean = μ
2. variance = $\frac{\sigma^2}{n}$

What is meant by large n ? Typically, $n \geq 30$

Comments about Central Limit Theorem

- 
1. Don't use CLT when population size is not too large compared with sample size, CLT is approximate result
 2. If sample size is large and population size is also much larger as compared to population, one can use CLT even when the sampling scheme is Without Replacement

CLT when population is normally distributed

- 
1. If underlying population is normally distributed, CLT is not required
 2. If population is approximately normal, then \bar{X} is approximately normal even for small n

Confidence Interval Ideas



Sample Proportion

- Sometimes, one is interested in estimating population proportion
- What is the proportion of IBM-312 students participants who like statistics?
- One can attempt the answer to this using sampling

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 3, 2023



Standard Normal Distribution

- ❑ is a normal distribution with mean = 0, variance = 1
- ❑ Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- ❑ Let $X \sim N(\mu, \sigma^2)$
- ❑ Hence, $X - \mu \sim N(0, \sigma^2)$
- ❑ Thus, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- ❑ Typically, Z is used to denote a standard normal distribution

Standard Normal Distribution

- ❑ is a normal distribution with mean = 0, variance = 1
- ❑ Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- ❑ Let $X \sim N(\mu, \sigma^2)$
- ❑ Hence, $X - \mu \sim N(0, \sigma^2)$
- ❑ Thus, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- ❑ Typically, Z is used to denote a standard normal distribution

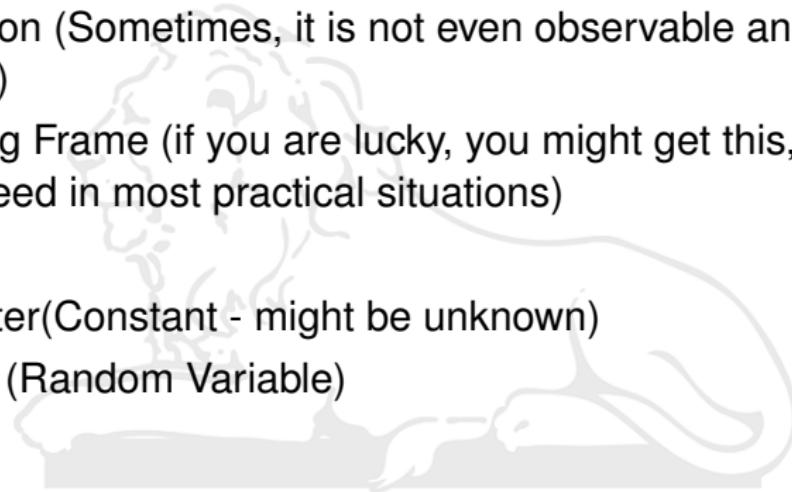
Standard Normal Distribution

- ❑ is a normal distribution with mean = 0, variance = 1
- ❑ Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- ❑ Let $X \sim N(\mu, \sigma^2)$
- ❑ Hence, $X - \mu \sim N(0, \sigma^2)$
- ❑ Thus, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- ❑ Typically, Z is used to denote a standard normal distribution

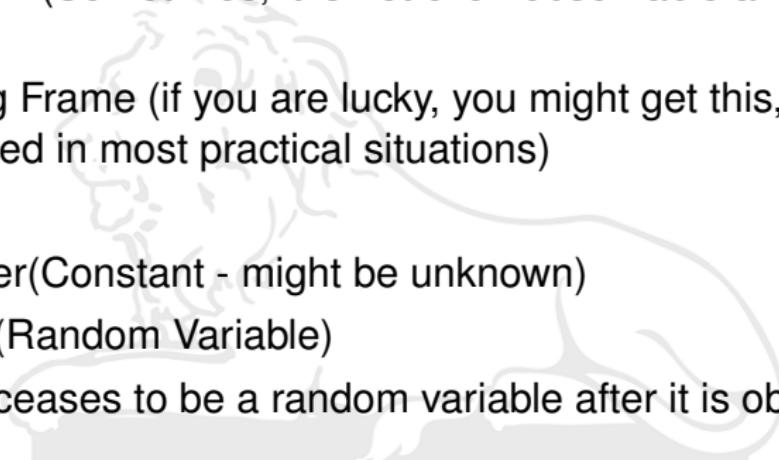
Standard Normal Distribution

- ❑ is a normal distribution with mean = 0, variance = 1
- ❑ Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- ❑ Let $X \sim N(\mu, \sigma^2)$
- ❑ Hence, $X - \mu \sim N(0, \sigma^2)$
- ❑ Thus, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- ❑ Typically, Z is used to denote a standard normal distribution

Basic Ideas of Sampling

- 
1. Population (Sometimes, it is not even observable and only abstract)
 2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations)
 3. Subject
 4. Parameter (Constant - might be unknown)
 5. Statistic (Random Variable)

Basic Ideas of Sampling

- 
1. Population (Sometimes, it is not even observable and only abstract)
 2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations)
 3. Subject
 4. Parameter (Constant - might be unknown)
 5. Statistic (Random Variable)
 6. Statistic ceases to be a random variable after it is observed

Types of Sampling

- Simple Random Sample With replacement
- Simple Random Sample without replacement
- Cluster Sampling
- Stratified Sampling

Potential Causes of Bias

- Convenience Sampling
- Volunteer Sampling
- Systematic Sampling
- Non-response Bias
- Response Bias

Does that mean we shouldn't use any of these types of sampling??

Potential Causes of Bias

- Convenience Sampling
- Volunteer Sampling
- Systematic Sampling
- Non-response Bias
- Response Bias

Does that mean we shouldn't use any of these types of sampling??
NO, one can use, but with caution. Make sure it is not leading to a systematic error

What after sampling?

- ❑ Ask, why did we do sampling? Objective is to learn about the population
- ❑ Statistical Inference - Learn about parameters from sample statistic
- ❑ Usually, the quantities of interest are mean and proportion in the population (depending on the context)
- ❑ We deal with them separately

Estimating from the statistic

- ❑ The rational behind estimating is expectation of statistic should be equal to the population parameter
- ❑ Biased and Unbiased Estimator
- ❑ Variance of estimate should be minimized to the extent possible

Errors in the Process of Estimation

- ❑ **Sampling Error** - Because we are only considering a subset of population, the point estimate is rarely exactly correct. Unavoidable error, but we can estimate the error and hence have some control over it
- ❑ **Non-sampling Error** - If there is bias in the observations, or sampling wasn't done properly. Can't be dealt with mathematically. Should be avoided

Estimating Population Mean from Sample Mean

1. Let the true values in the population be $A_1, A_2, A_3, \dots, A_N$
2. Population mean is denoted by μ and equals $\frac{\sum_{i=1}^N A_i}{N}$
3. Also, population variance is denoted by σ^2 and equals $\frac{\sum_{i=1}^N (A_i - \mu)^2}{N}$
4. Let the sample be a SRS of size n
5. Observations are X_1, X_2, \dots, X_n
6. Sample mean is denoted by \bar{X} and defined as follows
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$
7. $E(\bar{X}) = \mu$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Estimating Population Variance from Sample Observations

- ❑ If the sampling scheme is WITH REPLACEMENT

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- ❑ If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N - 1}{N} \right) \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Check $n - 1$ using a dataset

Why is estimating population variance important?

Estimating Population Variance from Sample Observations

- ❑ If the sampling scheme is WITH REPLACEMENT

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- ❑ If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N - 1}{N} \right) \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Check $n - 1$ using a dataset

Why is estimating population variance important?

To get an idea about error in estimation of sample mean

Standard Error in Sample Mean

- ❑ If the sampling scheme is WITH REPLACEMENT

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- ❑ If the sampling scheme is WITHOUT REPLACEMENT

$$\text{Var}(\bar{X}) = \left(\frac{N-n}{N-1} \right) \frac{\sigma^2}{n}$$

- ❑ $\frac{N-n}{N-1}$ is called the finite population correction
- ❑ Typically, can be ignored if sampling fraction $\frac{n}{N} \leq 0.05$
- ❑ Standard deviation of \bar{X} is called the Standard error of the sample mean
- ❑ Do we know σ^2 ? What is the remedy??

Central Limit Theorem

If n is well large,

$$\bar{X} \sim \text{Normal} \left(\mu, \frac{\sigma^2}{n} \right)$$

- Can we do better about our inference from sample mean??
- Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Central Limit Theorem

- ❑ Can we do better about our inference from sample mean??
- ❑ Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?
- ❑ Thank you Central Limit Theorem

Central Limit Theorem

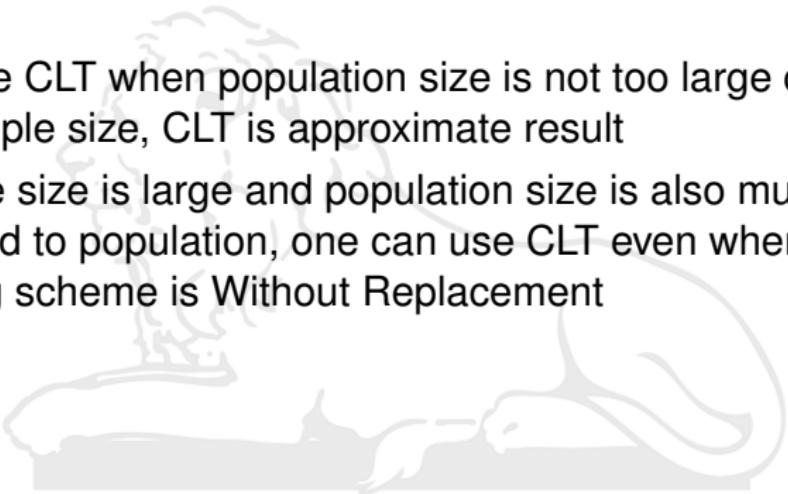
Theorem

If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean \bar{X} is approximately normal with

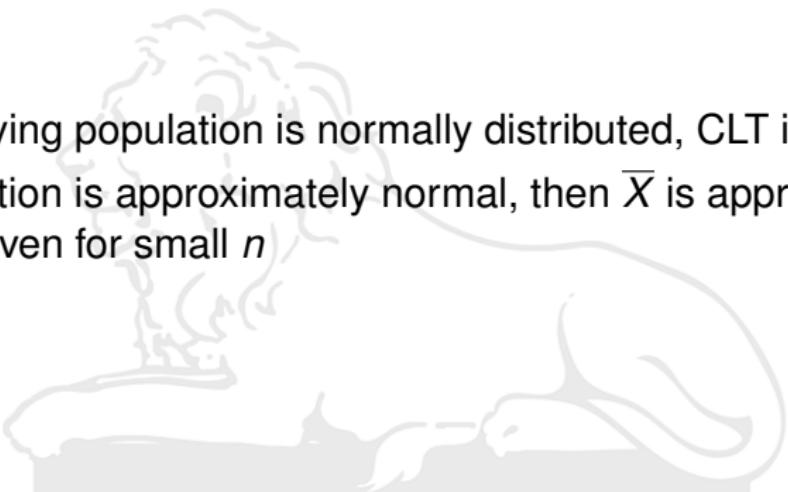
1. mean = μ
2. variance = $\frac{\sigma^2}{n}$

What is meant by large n ? Typically, $n \geq 30$

Comments about Central Limit Theorem

- 
1. Don't use CLT when population size is not too large compared with sample size, CLT is approximate result
 2. If sample size is large and population size is also much larger as compared to population, one can use CLT even when the sampling scheme is Without Replacement

CLT when population is normally distributed

- 
1. If underlying population is normally distributed, CLT is not required
 2. If population is approximately normal, then \bar{X} is approximately normal even for small n

Confidence Interval

$Y \sim \text{Normal}(\mu, \sigma^2)$

Back to Sample Mean \bar{X}

$$P(\mu - 2\sigma \leq Y \leq \mu + 2\sigma) =$$

1. \bar{X} is a random variable
2. Under certain conditions, large sample size, etc. We use CLT to get better idea about \bar{X}

3. Using properties of normal distribution, what can be said about

4. $P(\bar{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}})$

5. Is $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$? Why? Because CLT is approximate result.

Approximate

6. Thus, $P(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}) = 0.9544$

$$P\left(\bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 95.44\%$$

8. Magic here, we have created an interval for μ

9. This is nothing but the confidence interval

Confidence Interval

$$P\left(\frac{\bar{X} - 2\sigma}{\sqrt{n}} \leq \mu \leq \frac{\bar{X} + 2\sigma}{\sqrt{n}}\right) = 0.95447$$

Back to Sample Mean \bar{X}

1. $P\left(\bar{X}\text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is it not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\underline{\mu - 2\frac{\sigma}{\sqrt{n}}} \leq \bar{X} \leq \underline{\mu + 2\frac{\sigma}{\sqrt{n}}}\right) = 0.9544$

4. Or, by rearrangement of terms,

$$\left[\underline{\mu - 2\frac{\sigma}{\sqrt{n}}} \leq \bar{X} \quad \& \quad \bar{X} \leq \underline{\mu + 2\frac{\sigma}{\sqrt{n}}} \right]$$

5. Magic here we have created an interval for μ

6. This is nothing but the confidence interval

$$\left[\underline{\mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}} \quad ; \quad \bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \underline{\mu} \right]$$

Confidence Interval

Back to Sample Mean \bar{X}

1. $P\left(\bar{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is it not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
 $P\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

Back to Sample Mean \bar{X}

1. $P\left(\bar{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is it not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
$$P\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

$$\mu = 100.$$

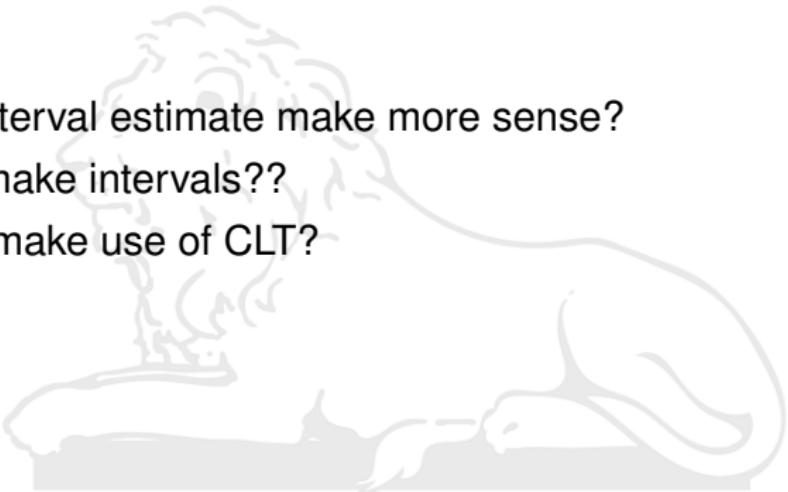
$$[25, 99]$$

$$[99, 140]$$

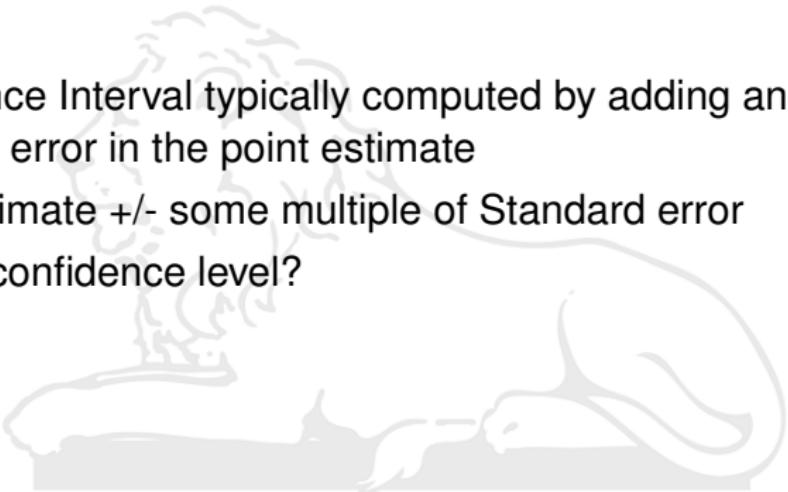
Back to Sample Mean \bar{X}

1. $P(\bar{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}})$
2. Is it not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
 $P\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

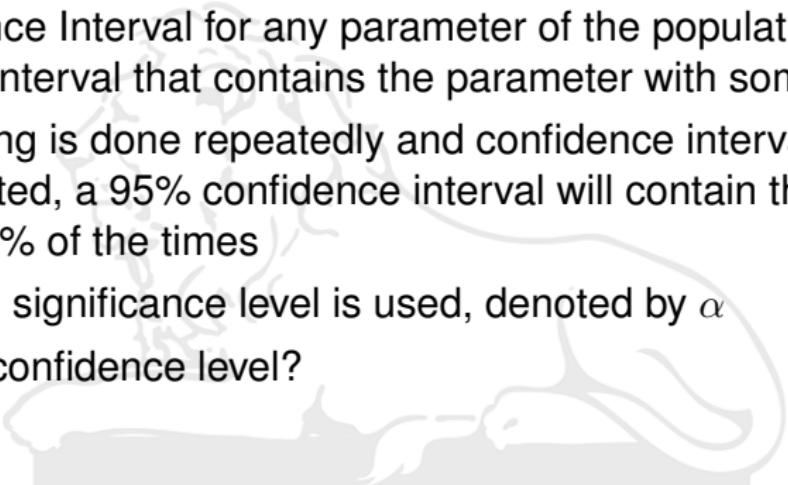
Confidence Interval

- 
1. Would interval estimate make more sense?
 2. How to make intervals??
 3. Can we make use of CLT?

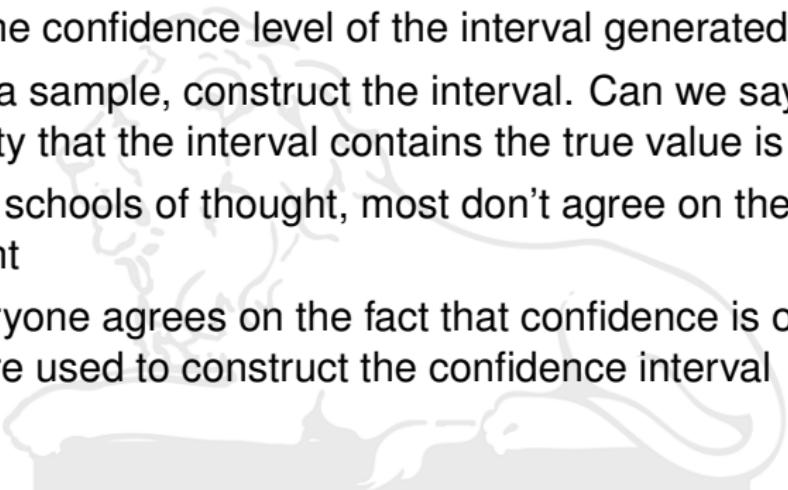
Confidence Interval

- 
1. Confidence Interval typically computed by adding and subtracting standard error in the point estimate
 2. Point estimate \pm some multiple of Standard error
 3. What is confidence level?

Confidence Interval Idea

- 
1. Confidence Interval for any parameter of the population is a random interval that contains the parameter with some probability
 2. If sampling is done repeatedly and confidence intervals are constructed, a 95% confidence interval will contain the values about 95% of the times
 3. Typically, significance level is used, denoted by α
 4. What is confidence level?

Confidence Interval Idea

- 
1. 95% is the confidence level of the interval generated
 2. We pick a sample, construct the interval. Can we say that the probability that the interval contains the true value is 0.95??
 3. Different schools of thought, most don't agree on the above made statement
 4. But, everyone agrees on the fact that confidence is on the procedure used to construct the confidence interval

Example of Confidence Interval

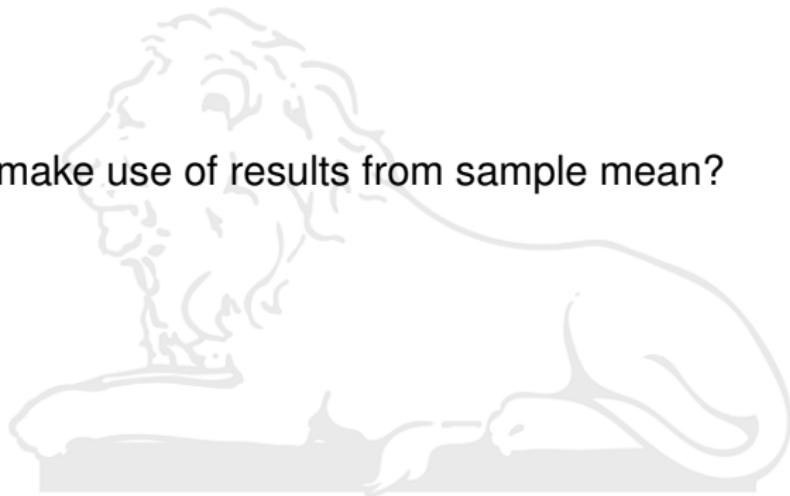
A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.

Sample Proportion

- Sometimes, one is interested in estimating population proportion
- What is the proportion of IBM312 students who like statistics?
- One can attempt the answer to this using sampling

Sample Proportion

- ❑ Can we make use of results from sample mean?



Sample Proportion

- Can we make use of results from sample mean?
- If the i^{th} respondent says YES, model it as $X_i = 1$
- If the i^{th} respondent says NO, model it as $X_i = 0$
- Denote by n_{YES} and n_{NO} are the responses in the sample of size n
- Denote by N_{YES} and N_{NO} are the actual values in the population of size N

Sampling Proportion

- We denote the estimate by \hat{p}
- The population proportion is denoted by p
- $\hat{p} = \frac{n_{YES}}{n}$
- $E(\hat{p}) = p$. Do we need to prove this??
- $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$. Why??
- Is p known?
- State CLT for sample proportion
- Additional conditions - $np \geq 10$ and $n(1-p) \geq 10$

Easier way for check unbiasedness of sample proportion

- ❑ We denote the estimate by \hat{p}
- ❑ The population proportion is denoted by p
- ❑ $\hat{p} = \frac{n_{YES}}{n}$
- ❑ What kind of random variable is n_{YES} ??
- ❑ n_{YES} is Binomial random variable with parameters p and n
- ❑ Hence, $E(\hat{p}) = E\left(\frac{n_{YES}}{n}\right) = \frac{E(n_{YES})}{n} = p$
- ❑ Also, $Var(\hat{p}) = Var\left(\frac{n_{YES}}{n}\right) = \frac{Var(n_{YES})}{n^2} = \frac{p(1-p)}{n}$
- ❑ But, we don't know p
- ❑ $Var(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{n - 1}$. Why??
- ❑ To provide an unbiased estimator of $Var(\hat{p})$

Easier way for check unbiasedness of sample proportion

- ❑ We denote the estimate by \hat{p}
- ❑ The population proportion is denoted by p
- ❑ $\hat{p} = \frac{n_{YES}}{n}$
- ❑ n_{YES} is Binomial random variable with parameters p and n
- ❑ Hence, $E(\hat{p}) = E\left(\frac{n_{YES}}{n}\right) = \frac{E(n_{YES})}{n} = p$
- ❑ Also, $Var(\hat{p}) = Var\left(\frac{n_{YES}}{n}\right) = \frac{Var(n_{YES})}{n^2} = \frac{p(1-p)}{n}$
- ❑ But, we don't know p
- ❑ $Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- ❑ To provide an unbiased estimator of $Var(\hat{p})$

Easier way for check unbiasedness of sample proportion

- ❑ We denote the estimate by \hat{p}
- ❑ The population proportion is denoted by p
- ❑ $\hat{p} = \frac{n_{YES}}{n}$
- ❑ n_{YES} is Binomial random variable with parameters p and n
- ❑ Hence, $E(\hat{p}) = E\left(\frac{n_{YES}}{n}\right) = \frac{E(n_{YES})}{n} = p$
- ❑ Also, $Var(\hat{p}) = Var\left(\frac{n_{YES}}{n}\right) = \frac{Var(n_{YES})}{n^2} = \frac{p(1-p)}{n}$
- ❑ But, we don't know p
- ❑ $Var(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{n - 1}$. Why??
- ❑ To provide an unbiased estimator of $Var(\hat{p})$

Easier way for check unbiasedness of sample proportion

- ❑ We denote the estimate by \hat{p}
- ❑ The population proportion is denoted by p
- ❑ $\hat{p} = \frac{n_{YES}}{n}$
- ❑ n_{YES} is Binomial random variable with parameters p and n
- ❑ Hence, $E(\hat{p}) = E\left(\frac{n_{YES}}{n}\right) = \frac{E(n_{YES})}{n} = p$
- ❑ Also, $Var(\hat{p}) = Var\left(\frac{n_{YES}}{n}\right) = \frac{Var(n_{YES})}{n^2} = \frac{p(1-p)}{n}$
- ❑ But, we don't know p
- ❑ $Var(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{n - 1}$. Why??
- ❑ To provide an unbiased estimator of $Var(\hat{p})$

Confidence Interval Discussions

- ❑ Can you also do similar calculations and make a confidence interval for Population proportion? (Hint - Use CLT and our remark that sample proportion can be given a similar treatment as sample mean)
- ❑ Khan Academy Video
<https://www.youtube.com/watch?v=bGALoCckICI>
- ❑ Which is bigger - 99% confidence interval or 95% confidence interval?

Summary of results for $100(1-\alpha)\%$ C.I.

n	σ^2	C.I. Type	Symmetric C.I.
Large	known	Approximate	$\left(\bar{X} - \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}\right)$
Large	unknown	Approximate	$\left(\bar{X} - \frac{Z_{\alpha/2}s}{\sqrt{n}}, \bar{X} + \frac{Z_{\alpha/2}s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ , s is sample standard deviation

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right)$

Table: C.I. for population proportion p , \hat{p} is sample proportion

Sample Size Determination

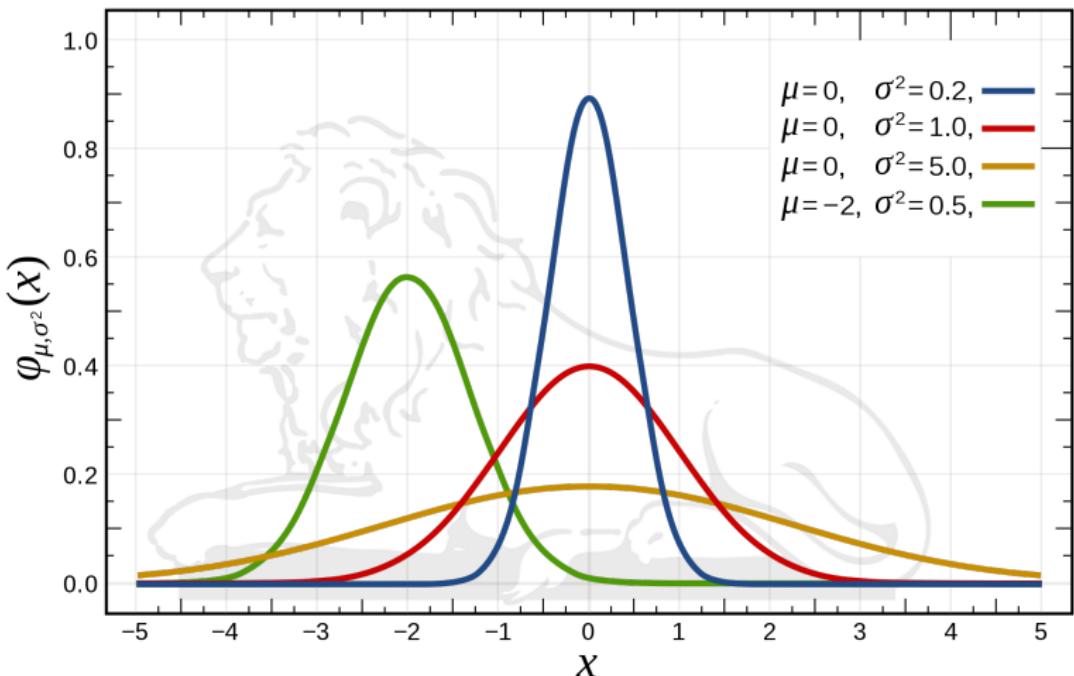
- ❑ A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.
- ❑ We want to repeat this study, how many students should you survey so that the 99% confidence interval's width is no more than 2 minutes?

Basics about Random Variable

- ❑ A variable whose value depends on outcome of a random phenomenon
- ❑ A random variable is characterized by its distribution
- ❑ Sum of two or more different random variables is also a random variable, and thus will also have a distribution (we might not cover the mathematical tools required to find the distribution, but it is important to appreciate that it will have a distribution)
- ❑ Similarly, any other algebraic operation of two or more random variables also remain a random variable
- ❑ If X and Y are random variables, $X + Y$, $X - Y$, XY , $\frac{X}{Y}$ are all random variables

Standard Normal Distribution

A normal distribution with mean 0 and standard deviation as 1



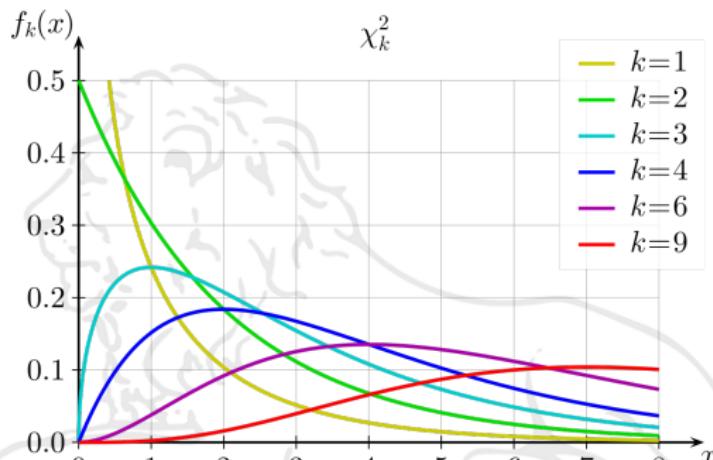
Chi-square distribution

If Z_1, Z_2, \dots, Z_n are all independent and normally distributed with mean

0 and variance 1, then $U = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$

Alternatively, U is said to follow a χ^2 distribution with n degrees of freedom

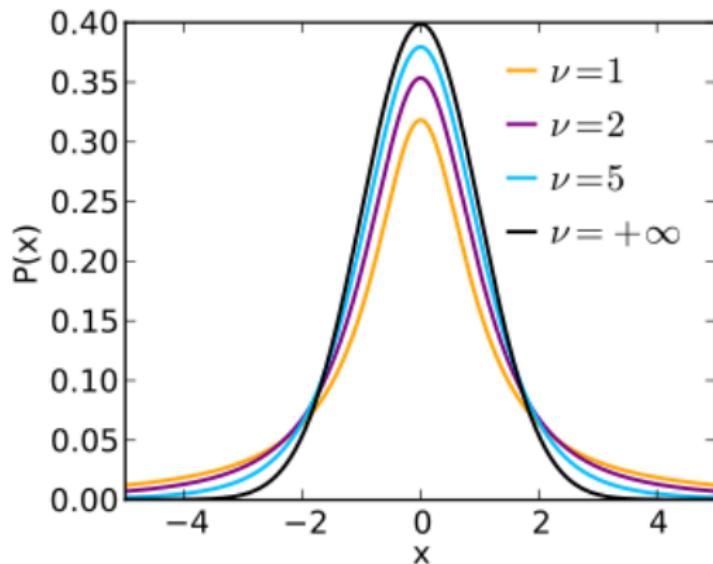
Chi-square distribution



Source - https://en.wikipedia.org/wiki/Chi-squared_distribution

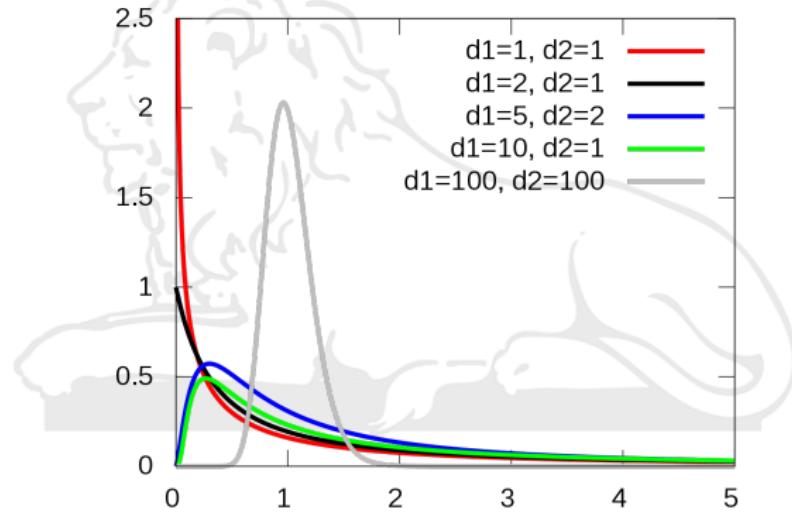
t-distribution

If $Z \sim N(0, 1)$ and $U \sim \chi_n^2$ are independent, then the distribution of $\frac{Z}{\sqrt{\frac{U}{n}}}$ is called the t-distribution with n degrees of freedom



F distribution

If $U \sim \chi_m^2$ and $V \sim \chi_n^2$ are independent, then the distribution of $\frac{U}{\frac{m}{n}V}$ is called the F-distribution with m and n degrees of freedom and is denoted by $F_{m,n}$



Application of distributions that we just saw

- Does having the idea of population distribution itself a useful information?
- If yes, how do we make use of it?
- Let us concern ourselves with sample mean
- Assume you have the information that the population distribution is normal.
- How do you use this??
- Is CLT required??

Case of normal population

- Sample mean is denoted by \bar{X} and defined as follows

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- If the sampling scheme is WITH REPLACEMENT, sample variance equals

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- It can be shown that -

1. \bar{X} and s_X^2 are independent

2. $\frac{(n - 1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$

- Can you guess the distribution of

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- If σ is unknown, we replace by s_X , but then the distribution ceases to be $N(0, 1)$. So what is it then??

- Distribution of $\frac{\bar{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$

- Hence, even for small sample size, we can make confidence

Case of normal population

- ❑ It can be shown that -
 1. \bar{X} and s_X^2 are independent
 2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$
- ❑ Can you guess the distribution of $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- ❑ If σ is unknown, we replace by s_X , but then the distribution ceases to be $N(0, 1)$. So what is it then??
- ❑ Distribution of $\frac{\bar{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- ❑ Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Case of normal population

- ❑ It can be shown that -
 1. \bar{X} and s_X^2 are independent
 2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$
- ❑ Can you guess the distribution of $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- ❑ If σ is unknown, we replace by s_X , but then the distribution ceases to be $N(0, 1)$. So what is it then??
- ❑ Distribution of $\frac{\bar{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- ❑ Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Moral of the story - results for $100(1-\alpha)\%$ C.I. for Mean

Population distribution	n	σ^2	C.I. Type	Symmetric C.I.
Any	Large	known	Approximate	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}\right)$
Any	Large	unknown	Approximate	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} s}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} s}{\sqrt{n}}\right)$
Normal	Any	known	Exact	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}\right)$
Normal	Any	unknown	Exact	$\left(\bar{X} - \frac{t_{n-1, \frac{\alpha}{2}} s}{\sqrt{n}}, \bar{X} + \frac{t_{n-1, \frac{\alpha}{2}} s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ , s is sample standard deviation

Typically, $t_{n-1, \frac{\alpha}{2}}$ is used only for small n , because for large n , $Z_{\frac{\alpha}{2}}$ gives a good approximation

Moral of the story - results for $100(1-\alpha)\%$ C.I. for Proportion

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n-1}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n-1}} \right)$

Table: C.I. for population proportion p , \hat{p} is sample proportion

For case of proportion, it is advised to use these formulae only when apart from n being large, $n\hat{p} \geq 10$ and also $n(1 - \hat{p}) \geq 10$

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 9, 2023



Recap and Today

Recap -

Confidence Interval for the case of Proportion
Sample Size Determination Ideas

Today -

Discussion on Sampling (Literary Digest Example)
Confidence Interval when Population distribution is normal
Hypothesis Testing

Summary of results for $100(1-\alpha)\%$ C.I.

n	σ^2	C.I. Type	Symmetric C.I.
Large	known	Approximate	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}\right)$
Large	unknown	Approximate	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} s}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ , s is sample standard deviation

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n - 1}}, \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n - 1}}\right)$

Table: C.I. for population proportion p , \hat{p} is sample proportion

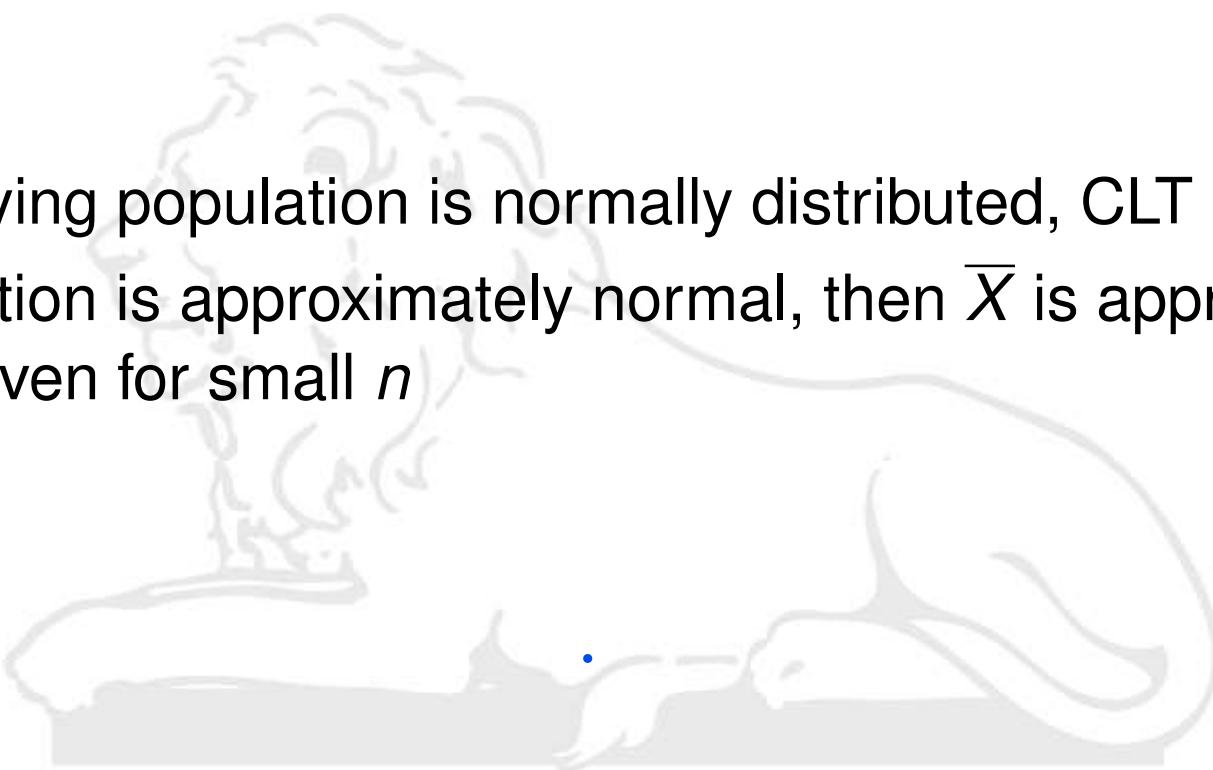
Sample Size Determination

- ❑ A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.
- ❑ We want to repeat this study, how many students should you survey so that the 99% confidence interval's width is no more than 2 minutes?

CLT when population is normally distributed

X

1. If underlying population is normally distributed, CLT is not required
2. If population is approximately normal, then \bar{X} is approximately normal even for small n

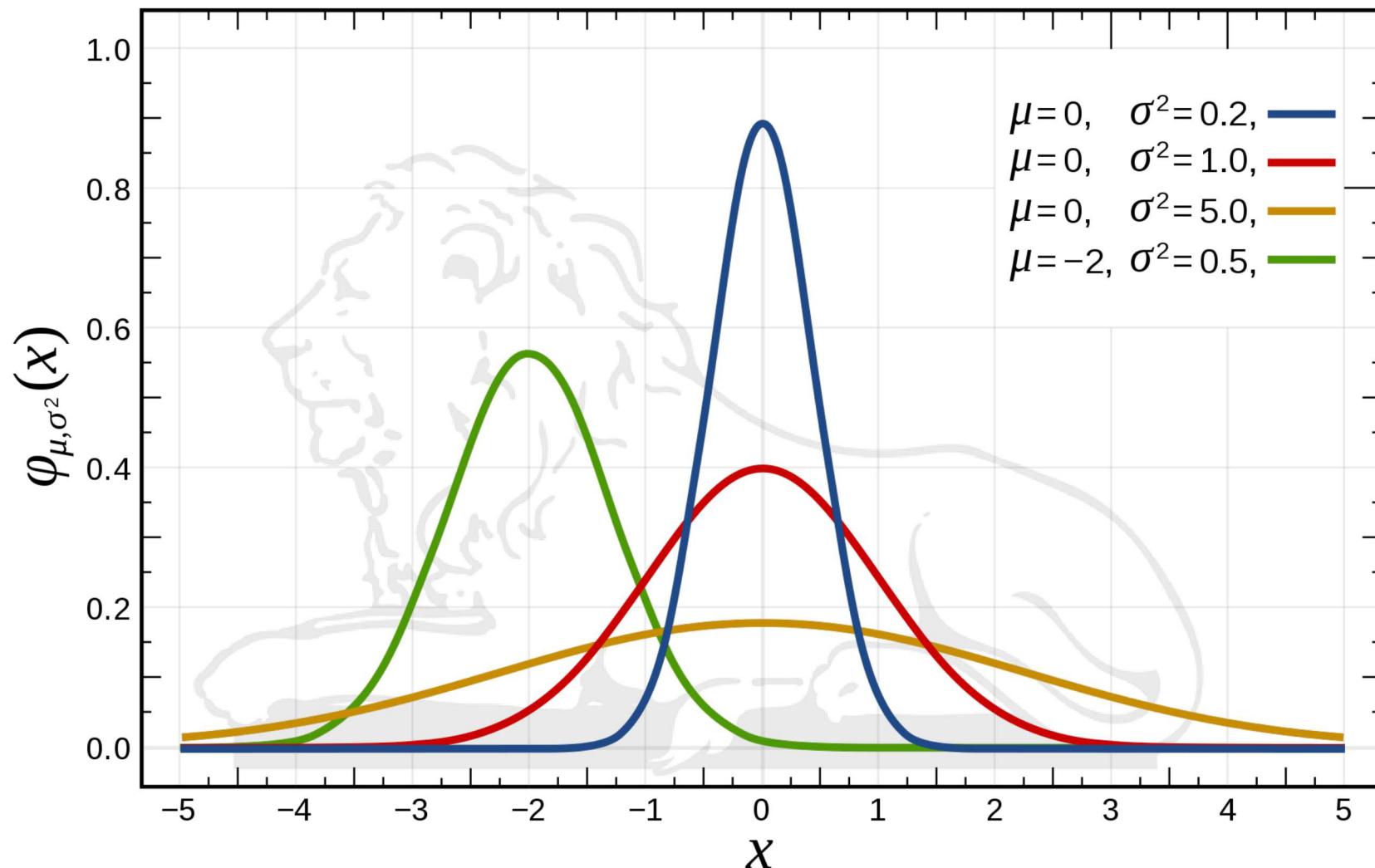


Basics about Random Variable - Recap

- ❑ A variable whose value depends on outcome of a random phenomenon
- ❑ A random variable is characterized by its distribution
- ❑ Sum of two or more different random variables is also a random variable, and thus will also have a distribution (we might not cover the mathematical tools required to find the distribution, but it is important to appreciate that it will have a distribution)
- ❑ Similarly, any other algebraic operation of two or more random variables also remain a random variable
- ❑ If X and Y are random variables, $X + Y$, $X - Y$, XY , $\frac{X}{Y}$ are all random variables

Standard Normal Distribution

A normal distribution with mean 0 and standard deviation as 1



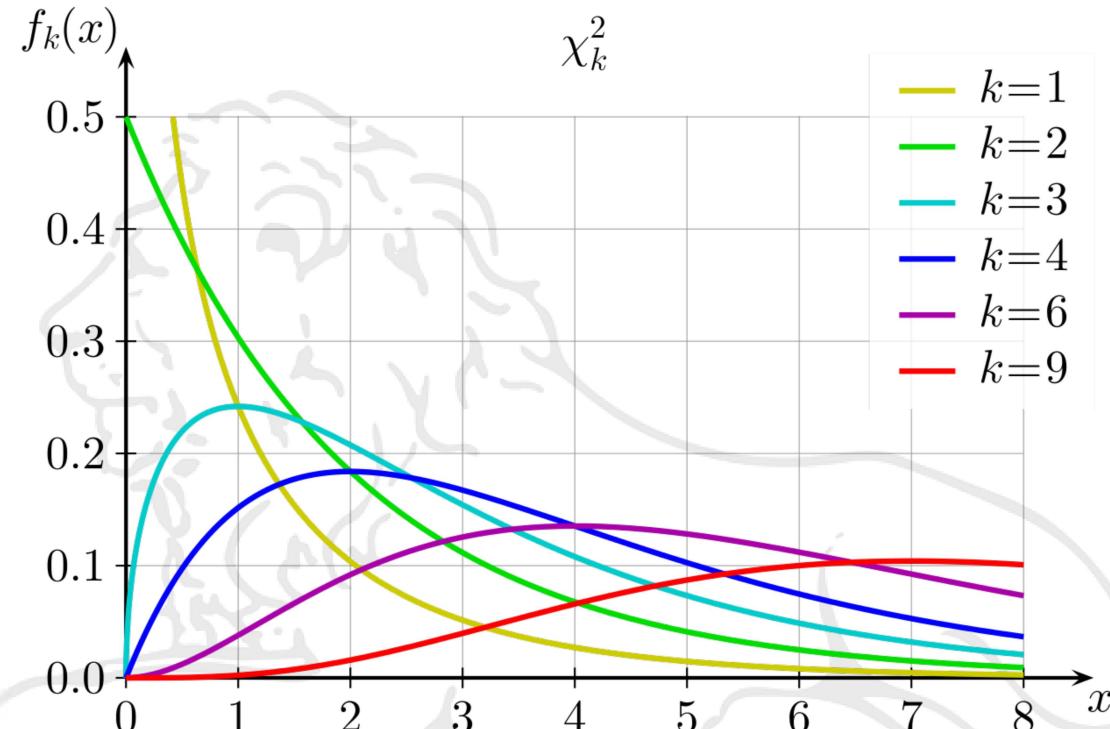
Chi-square distribution

If Z_1, Z_2, \dots, Z_n are all independent and normally distributed with mean

$$0 \text{ and variance } 1, \text{ then } U = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

Alternatively, U is said to follow a χ^2 distribution with n degrees of freedom

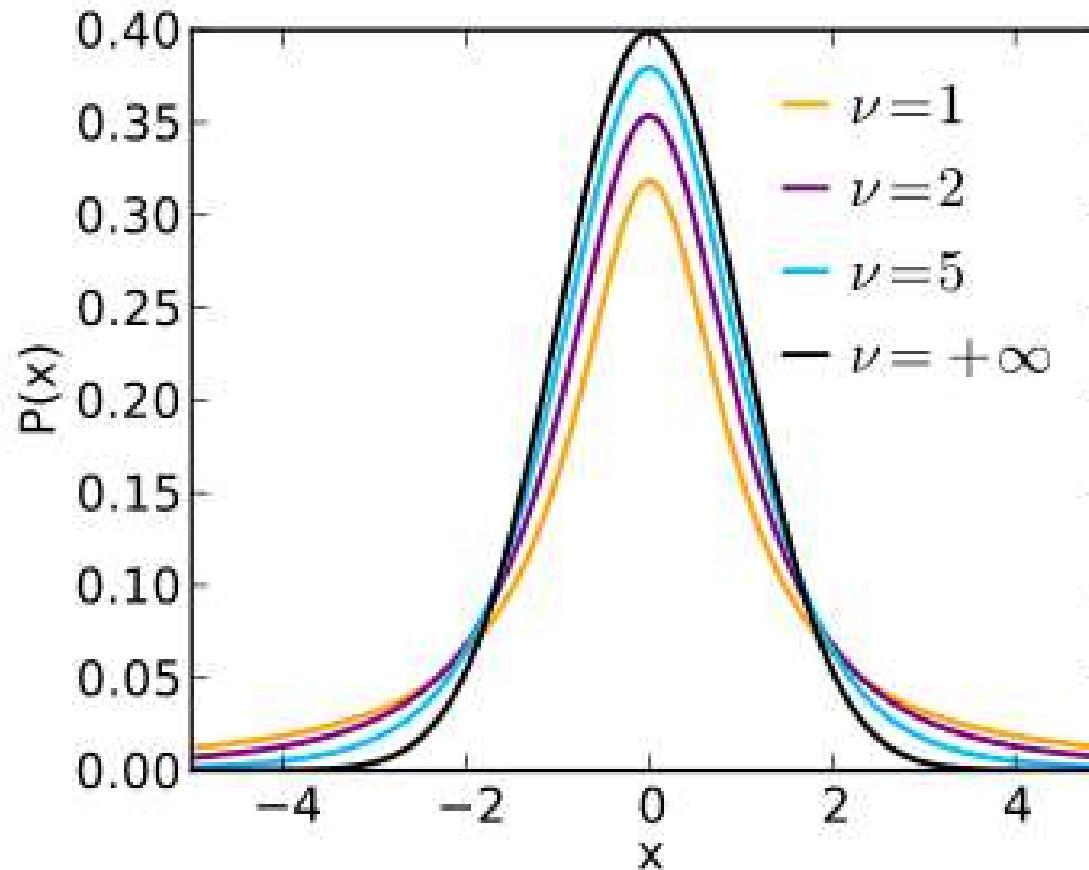
Chi-square distribution



Source - https://en.wikipedia.org/wiki/Chi-squared_distribution

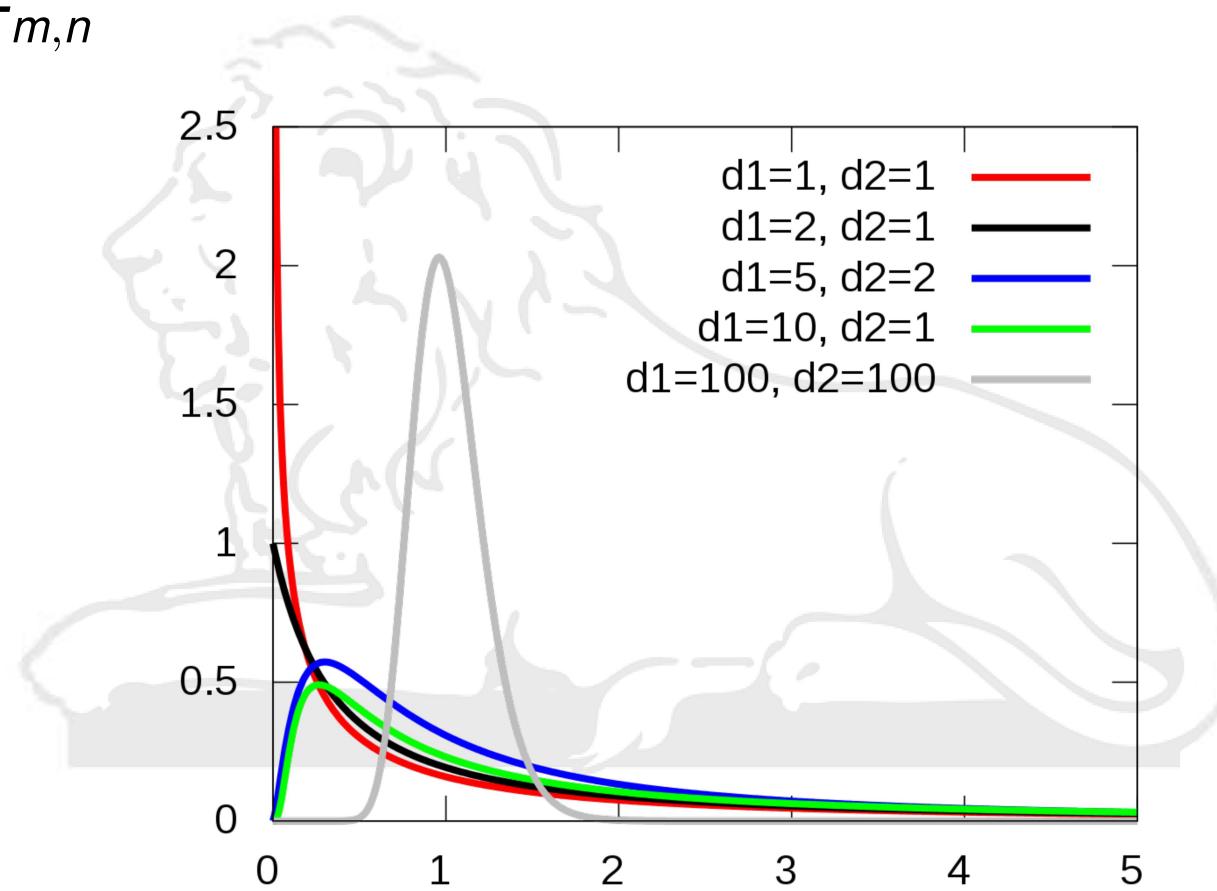
t-distribution

If $Z \sim N(0, 1)$ and $U \sim \chi_n^2$ are independent, then the distribution of $\frac{Z}{\sqrt{\frac{U}{n}}}$ is called the t-distribution with n degrees of freedom



F distribution

If $U \sim \chi_m^2$ and $V \sim \chi_n^2$ are independent, then the distribution of $\frac{U}{\frac{m}{n}}$ is called the F-distribution with m and n degrees of freedom and is denoted by $F_{m,n}$



Application of distributions that we just saw

- ❑ Does having the idea of population distribution itself a useful information?
- ❑ If yes, how do we make use of it?
- ❑ Let us concern ourselves with sample mean
- ❑ Assume you have the information that the population distribution is normal.
- ❑ How do you use this??
- ❑ Is CLT required??

Case of normal population

- Sample mean is denoted by \bar{X} and defined as follows

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- If the sampling scheme is WITH REPLACEMENT, sample variance equals

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- It can be shown that -

1. \bar{X} and s_X^2 are independent

2. $\frac{(n - 1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$

- Can you guess the distribution of

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- If σ is unknown, we replace by s_X , but then the distribution ceases to be $N(0, 1)$. So what is it then??

- Distribution of $\frac{\bar{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$

- Hence, even for small sample size, we can make confidence

Case of normal population

- It can be shown that -
 1. \bar{X} and s_X^2 are independent
 2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$
- Can you guess the distribution of $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X , but then the distribution ceases to be $N(0, 1)$. So what is it then??
- Distribution of $\frac{\bar{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Case of normal population

- It can be shown that -
 1. \bar{X} and s_X^2 are independent
 2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$
- Can you guess the distribution of $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X , but then the distribution ceases to be $N(0, 1)$. So what is it then??
- Distribution of $\frac{\bar{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Moral of the story - results for $100(1-\alpha)\%$ C.I. for Mean

Population distribution	n	σ^2	C.I. Type	Symmetric C.I.
Any	Large	known	Approximate	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}\right)$
Any	Large	unknown	Approximate	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} s}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} s}{\sqrt{n}}\right)$
Normal	Any	known	Exact	$\left(\bar{X} - \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}\right)$
Normal	Any	unknown	Exact	$\left(\bar{X} - \frac{t_{n-1, \frac{\alpha}{2}} s}{\sqrt{n}}, \bar{X} + \frac{t_{n-1, \frac{\alpha}{2}} s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ , s is sample standard deviation

Typically, $t_{n-1, \frac{\alpha}{2}}$ is used only for small n , because for large n , $Z_{\frac{\alpha}{2}}$ gives a good approximation

Moral of the story - results for $100(1-\alpha)\%$ C.I. for Proportion

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n - 1}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n - 1}} \right)$

Table: C.I. for population proportion p , \hat{p} is sample proportion

For case of proportion, it is advised to use these formulae only when apart from n being large, $n\hat{p} \geq 10$ and also $n(1 - \hat{p}) \geq 10$

Hypothesis Testing

(100,100) A:

$$\bar{X}_m = 72, \bar{X}_e = 73$$

" B: $\bar{X}_m = 72, \bar{X}_e = 71.8$

↳ C: $\bar{X}_m = 72, \bar{X}_e = 65$

- Hypothesis testing is a way of making statistical decisions using observed or experimental data
- Example** - Do students perform better if tests are given in the morning?
- Example** - Is the coin biased or not?
- Example** - Should the new book be adopted or not?

(10,10) D: $\bar{X}_m = 72, \bar{X}_e = 65$

Key Terms in Hypothesis Testing

- Null Hypothesis, usually denoted by H_0
- Alternative Hypothesis, usually denoted by H_1 or H_a
- Hypothesis testing is performed after data is available for a sample - X_1, X_2, \dots, X_n
- Decision of testing the hypothesis is whether to reject the null hypothesis or not to reject the null hypothesis
- Basis - from the value of a **statistic** T , which is calculated from the sample, called the **test statistic**
- Whenever the **null hypothesis is rejected, alternative hypothesis is accepted**

Key Terms in Hypothesis Testing

- Null Hypothesis, usually denoted by H_0
- Alternative Hypothesis, usually denoted by H_1 or H_a
- Hypothesis testing is performed after data is available for a sample - X_1, X_2, \dots, X_n
- Decision of testing the hypothesis is whether to reject the null hypothesis or not to reject the null hypothesis
- Basis - from the value of a **statistic** T , which is calculated from the sample, called the **test statistic**
- Whenever the **null hypothesis is rejected, alternative hypothesis is accepted**

Errors, Level and Power

	Null Rejected	Null not rejected
Null is true	Type 1 error	✓
Null is false	✓	Type 2 error

Table: Type 1 and Type 2 error

- Probability of type 1 error - level of the test or significance level of the test, denoted by α
- Probability of type 2 error usually denoted by β
- $1-\beta$ is called the power of the test

Remarks on α and β

- We control the type 1 error by our choice of level of significance α
- We then try to maximize power $1 - \beta$
- We cannot reduce the chances of both type 1 and type 2 error simultaneously

Example

$$(1000, 700) \rightarrow$$

$$(1000000, 700000) \rightarrow$$

BCCI was accused of using a biased coin in the toss for cricket match. A test was thus performed to check whether the given coin is biased towards heads. The coin was tossed 10 times.

1. Observed result - HHTHHHHHTH
2. Null Hypothesis $p = 0.5$
3. Alternative Hypothesis is $p > 0.5$

(It was decided that the null will be rejected if the number of heads is greater than 7?)

What is the level of the test?

What is the power of the test?

$$P(\text{Heads} \geq 8)$$

$$= {}^{10}C_8 (0.5)^8 (0.5)^2 + {}^{10}C_9 (0.5)^9 0.5^1 + 0.5^{10}$$

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a
- ❑ Philosophy of testing - Innocent until proven guilty
- ❑ What is meant by proven?? Depends on the criteria set by the decision maker.

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a
- ❑ Philosophy of testing - Innocent until proven guilty
- ❑ What is meant by proven?? Depends on the criteria set by the decision maker.

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a
- ❑ Philosophy of testing - Innocent until proven guilty
- ❑ What is meant by proven?? Depends on the criteria set by the decision maker.

Building Hypothesis - Examples

- Coin is biased towards heads
- Coin is not an unbiased coin
- Gravity Fitness Gym, Roorkee claims that if you join the gym, you will lose more than 5 kgs in one month on an average.
- BJP IT team claims that with the marketing strategy that they have adopted, more than 40% people who were voters of other parties will vote for BJP in the coming elections

Building Hypothesis - Example

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 14, 2023



Recap and Today

Recap -

Hypothesis Testing - Introduction

Today -

Examples

p-Value



Hypothesis Testing

- ❑ Hypothesis testing is a way of making statistical decisions using observed or experimental data
- ❑ **Example** - Do students perform better if tests are given in the morning?
- ❑ **Example** - Is the coin biased or not?
- ❑ **Example** - Should the new book be adopted or not?

Key Terms in Hypothesis Testing

- ❑ Null Hypothesis, usually denoted by H_0
- ❑ Alternative Hypothesis, usually denoted by H_1 or H_a
- ❑ Hypothesis testing is performed after data is available for a sample - X_1, X_2, \dots, X_n
- ❑ Decision of testing the hypothesis is whether to reject the null hypothesis or not to reject the null hypothesis
- ❑ Basis - from the value of a **statistic** T , which is calculated from the sample, called the **test statistic**
- ❑ Whenever the **null hypothesis** is rejected, **alternative hypothesis** is accepted

Key Terms in Hypothesis Testing

- ❑ Null Hypothesis, usually denoted by H_0
- ❑ Alternative Hypothesis, usually denoted by H_1 or H_a
- ❑ Hypothesis testing is performed after data is available for a sample - X_1, X_2, \dots, X_n
- ❑ Decision of testing the hypothesis is whether to reject the null hypothesis or not to reject the null hypothesis
- ❑ Basis - from the value of a **statistic** T , which is calculated from the sample, called the **test statistic**
- ❑ Whenever the **null hypothesis is rejected, alternative hypothesis is accepted**

Errors, Level and Power

	Null Rejected	Null not rejected
Null is true	Type 1 error	
Null is false		Type 2 error

Table: Type 1 and Type 2 error

- ❑ Probability of type 1 error - level of the test or significance level of the test, denoted by α
- ❑ Probability of type 2 error usually denoted by β
- ❑ $1-\beta$ is called the power of the test

Remarks on α and β

- ❑ We control the type 1 error by our choice of level of significance α
- ❑ We then try to maximize power $1 - \beta$
- ❑ We cannot reduce the chances of both type 1 and type 2 error simultaneously

Example

BCCI was accused of using a biased coin in the toss for cricket match. A test was thus performed to check whether the given coin is biased towards heads. The coin was tossed 10 times.

1. Observed result - HHTHHHHHTH
2. Null Hypothesis $p = 0.5$
3. Alternative Hypothesis is $p > 0.5$

It was decided that the null will be rejected if the number of heads is greater than 7?

What is the level of the test?

What is the power of the test?

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a
- ❑ Philosophy of testing - Innocent until proven guilty
- ❑ What is meant by proven?? Depends on the criteria set by the decision maker.

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a
- ❑ Philosophy of testing - Innocent until proven guilty
- ❑ What is meant by proven?? Depends on the criteria set by the decision maker.

Hypothesis Testing

- ❑ Null Hypothesis and Alternative Hypothesis
- ❑ Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- ❑ Alternative - Everything else comes into this hypothesis
- ❑ There should be no common possibility in null and alternative hypothesis
- ❑ There should be no possibility outside of H_0 and H_a
- ❑ Philosophy of testing - Innocent until proven guilty
- ❑ What is meant by proven?? Depends on the criteria set by the decision maker.

Building Hypothesis - Examples

$$H_0: p_n \leq 0.5$$

$$H_a: p_n > 0.5$$

$$H_0: p_n = 0.5$$

$$H_a: p_n \neq 0.5$$

- Coin is biased towards heads
- Coin is not an unbiased coin
- Gravity Fitness Gym, Roorkee claims that if you join the gym, you will lose more than 5 kgs in one month on an average.
- BJP IT team claims that with the marketing strategy that they have adopted, more than 40% people who were voters of other parties will vote for BJP in the coming elections

$$H_0: w \leq 5$$

$$H_a: w > 5$$

$$H_0: p_n \leq 0.4$$

$$H_a: p_n > 0.4$$

Building Hypothesis - Example

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

$$H_0: h \leq 18$$

$$H_a: h > 18$$

Building Hypothesis - Example

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

x_1, x_2, \dots

x_{100}

$$\bar{x} = 18.2$$

There is a perception that IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

$$H_0: h \leq 18$$

$$H_a: h > 18$$

One tail versus two tail test

In one-tailed test, we can reject null hypothesis only on one side of the hypothesized value of population parameter

In two-tailed test, we can reject null hypothesis on either side of the hypothesized value of population parameter

P-value

- ❑ p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- ❑ If the p-value is low, the null hypothesis should go
- ❑ In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- ❑ p-value for the shop sales problem? How to make a decision??

P-value

- ❑ p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- ❑ If the p-value is low, the null hypothesis should go
- ❑ In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- ❑ p-value for the shop sales problem? How to make a decision??

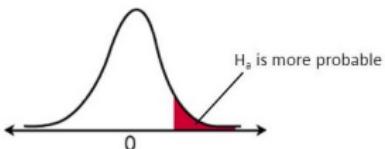
P-value

- ❑ p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- ❑ If the p-value is low, the null hypothesis should go
- ❑ In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- ❑ p-value for the shop sales problem? How to make a decision??

P-value

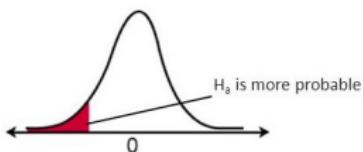
- ❑ p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- ❑ If the p-value is low, the null hypothesis should go
- ❑ In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- ❑ p -value for the shop-sales problem? How to make a decision??

One Tail and Two Tail Test



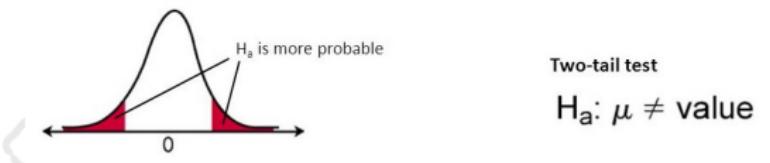
Right-tail test

$$H_a: \mu > \text{value}$$



Left-tail test

$$H_a: \mu < \text{value}$$



Two-tail test

$$H_a: \mu \neq \text{value}$$

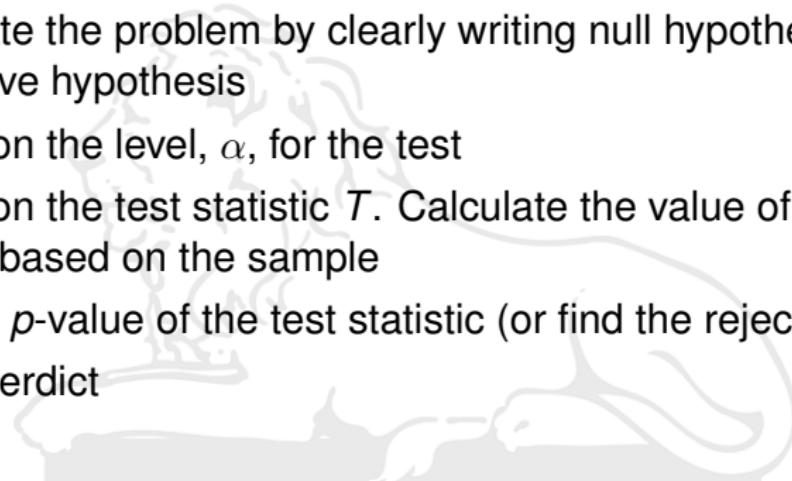
Source - <https://www.fromthegenesis.com/difference-between-one-tail-test-and-two-tail-test/>

Hypothesis testing for population mean

- ❑ Case when σ is known
- ❑ Case when σ is unknown



Steps in Hypothesis Testing

- 
1. Formulate the problem by clearly writing null hypothesis and alternative hypothesis
 2. Decide on the level, α , for the test
 3. Decide on the test statistic T . Calculate the value of the test statistic based on the sample
 4. Find the p -value of the test statistic (or find the rejection criteria)
 5. Give a verdict

Regression

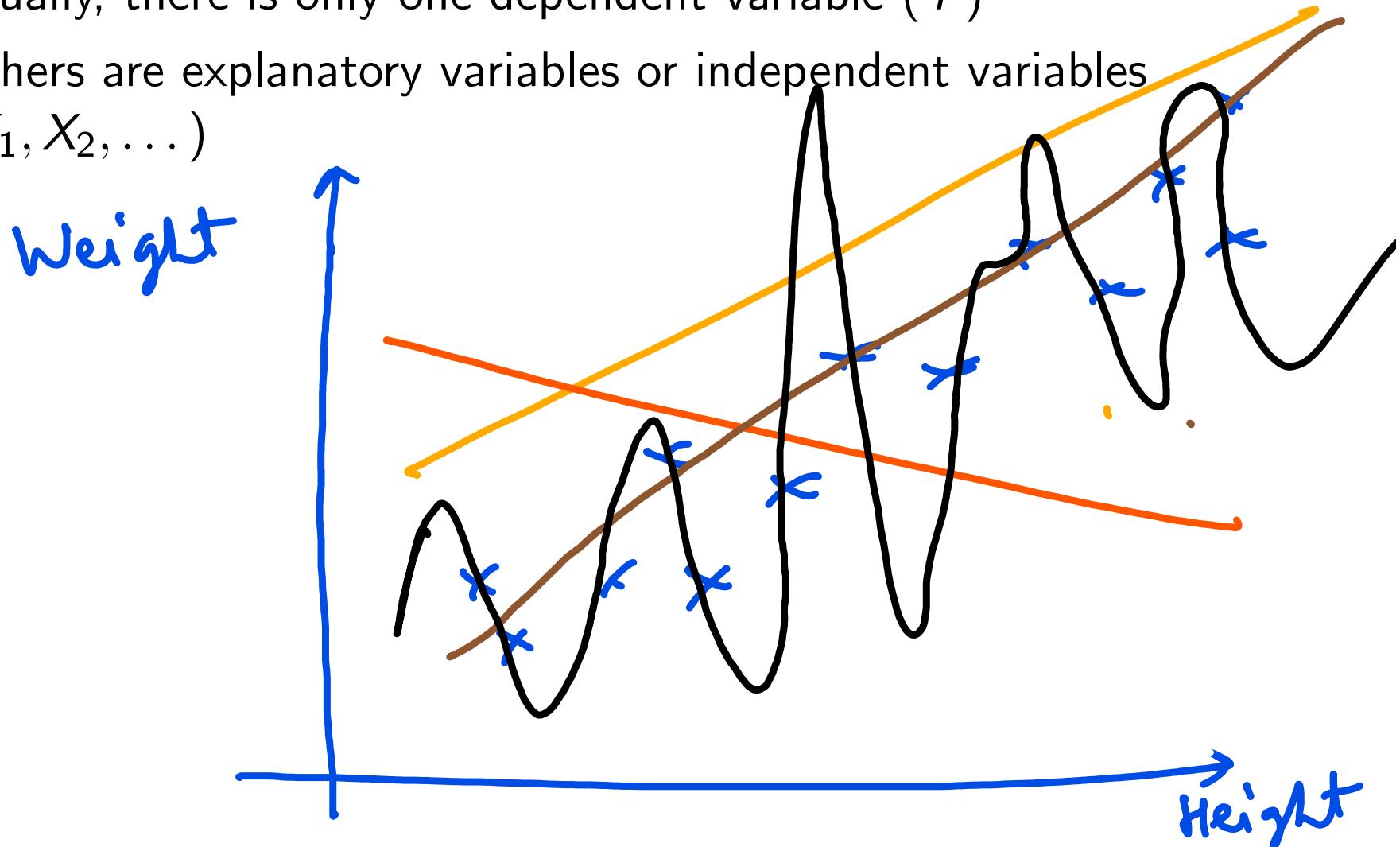
16 Feb 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Regression Analysis

- Looking for a relationship between a set of variables
- Usually, there is only one dependent variable (Y)
- Others are explanatory variables or independent variables (X_1, X_2, \dots)

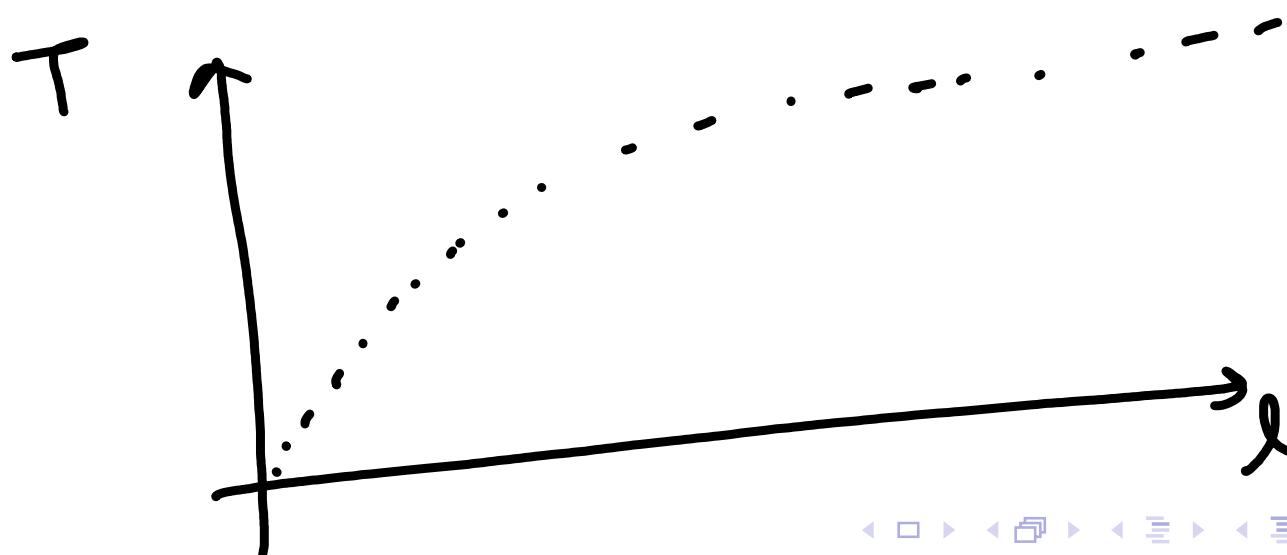


Regression Analysis

$$F = ma$$

$$T = 2\pi \sqrt{\frac{l}{g}}$$

- Looking for a relationship between a set of variables
- Usually, there is only one dependent variable (Y)
- Others are explanatory variables or independent variables (X_1, X_2, \dots)
- Can we assume a functional relationship between Y and the independent variables? $Y = f(X_1, X_2, \dots)$
- Usually, because of inherent nature of phenomena that we are trying to model, there is randomness and hence



Regression Analysis

- Looking for a relationship between a set of variables
- Usually, there is only one dependent variable (Y)
- Others are explanatory variables or independent variables (X_1, X_2, \dots)
- Can we assume a functional relationship between Y and the independent variables? $Y = f(X_1, X_2, \dots)$
- Usually, because of inherent nature of phenomena that we are trying to model, there is randomness and hence
- $Y = f(X_1, X_2, \dots) + \epsilon$
- ϵ is typically assumed to be a random variable with mean 0 and standard deviation σ
- Thus, $E(Y) = f(X_1, X_2, \dots)$

Simple Linear Regression

- If the assumed functional form is linear, we call it linear regression
- If the number of independent variables is one, we call it simple linear regression
- The linear form typically assumed is $Y = \alpha + \beta X + \epsilon$

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

Simple Linear Regression

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- β can be interpreted as average increase in Y for an unit increase in X
- α , in general, has no interpretation

Simple Linear Regression

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- α and β are population parameters, and hence are unknown
- Our task would be to estimate the values of α and β from the sample observations

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

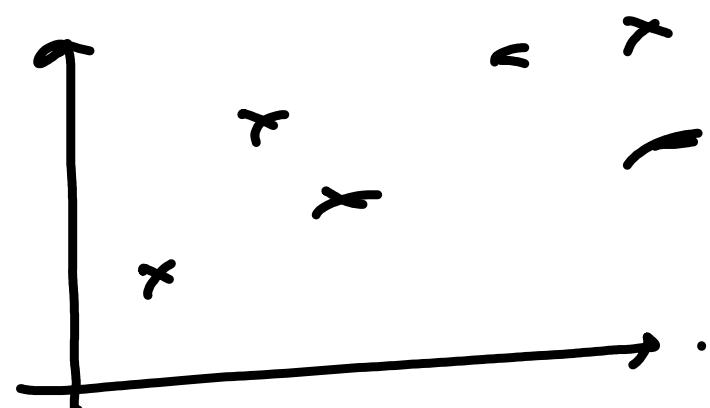
- α and β are population parameters, and hence are unknown
- Our task would be to estimate the values of α and β from the sample observations
- When the number of independent variables is just 1, we can observe the scatter plot to observe if linear relationship can be assumed between the variables
- If the scatter plot doesn't indicate that a linear relationship can be assumed, we should possibly drop the idea of simple linear regression, and do something more to understand the relationship between the variables

Simple Linear Regression

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- We would estimate the values of α and β from sample observations
- Denote by $\hat{\alpha}$ and $\hat{\beta}$ the estimates of α and β respectively
- Note that α and β uniquely determine the line
- Thus, given the data, we would determine $\hat{\alpha}$ and $\hat{\beta}$, which would uniquely determine a line
- Which line to fit??



Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- We would estimate the values of α and β from sample observations
- Denote by $\hat{\alpha}$ and $\hat{\beta}$ the estimates of α and β respectively
- Note that α and β uniquely determine the line
- Thus, given the data, we would determine $\hat{\alpha}$ and $\hat{\beta}$, which would uniquely determine a line
- The line which minimizes the sum of square of residuals

Simple Linear Regression

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- Given Data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Minimize: $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$
- Differentiate w.r.t $\hat{\alpha}$ and $\hat{\beta}$ and equate to zero
- We obtain 2 equations in 2 unknowns, which on solving give -
 - $\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
 - $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Simple Linear Regression - Estimation of Parameters

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- $\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
- To fully specify the model, one more parameter needs to be estimated, which is ??

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- $\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
- To fully specify the model, one more parameter needs to be estimated, which is σ
- σ is estimated using the standard deviation of residuals

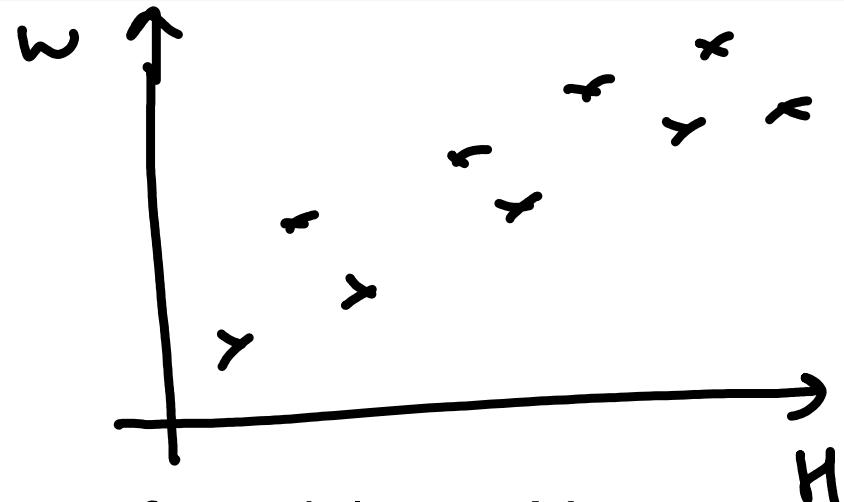
$$\bullet \hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n - 2}}$$

Goodness of fit - R^2

- Softwares will report a R^2 to you
- What does it mean??
- Gives an idea about what percentage of variability in Y is explained by the regression equation
- $SST = SSR + SSE$

$$\bullet \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

$$t \leftarrow \sum_{i=1}^n \left(\underbrace{y_i - \hat{y}_i}_{+} + \underbrace{\hat{y}_i - \bar{y}}_{+} \right)^2$$



Simple Linear Regression - Properties of Estimates of Parameters

Simple Linear Regression Model

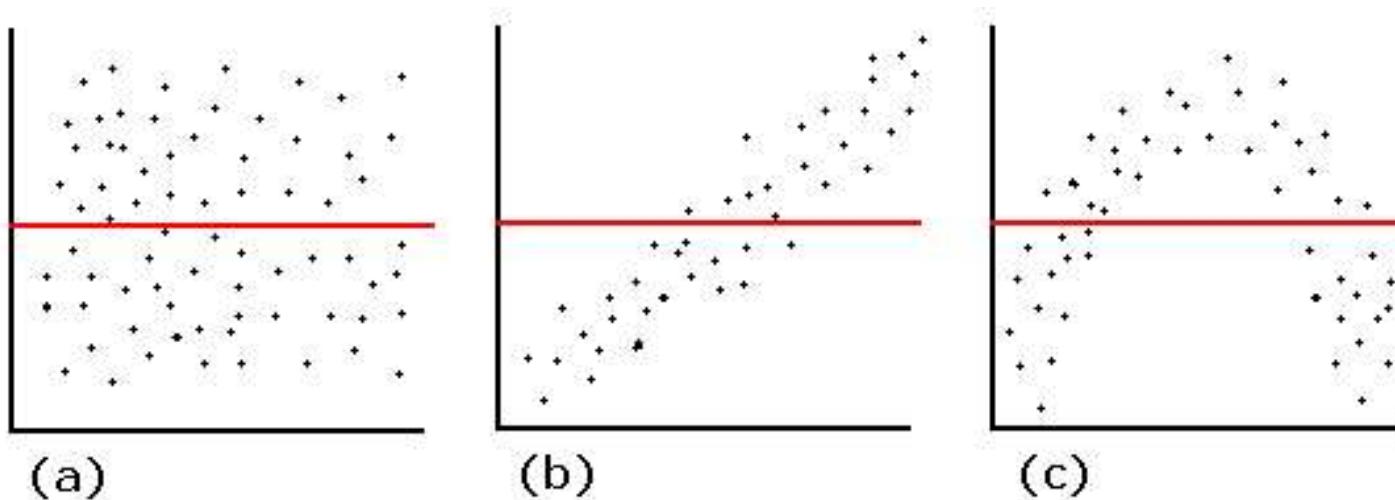
$$Y = \alpha + \beta X + \epsilon$$

- Sum of residuals is zero
- Residuals are uncorrelated with x_i 's
- It can also be shown that \hat{y}_i and e_i are uncorrelated
- $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$, since $y_i = \hat{y}_i + e_i$

Assumptions of Regression

- ϵ is a random variable that is normally distributed with mean 0 and s.d. σ
- Variance of ϵ is same for all values of x

Examples of Residual Plots



Source - <http://analyticspro.org/2016/03/05/r-tutorial-residual-analysis-for-regression/>

Multiple Linear Regression

We now have more than 1 independent variables. (say k)

Multiple Linear Regression Model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- Interpretation of β' s??
- How do you obtain α & β' s??
- Partial Differentiation to obtain $k + 1$ equations in $k + 1$ unknowns
- Example

Goodness of fit - R^2

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

$$\sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

- Softwares will report a R^2 to you
- What does it mean??
- Gives an idea about what percentage of variability in Y is explained by the regression equation
- $SST = SSR + SSE$
- $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$

To check:

$$\sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i - \bar{y}) (\hat{y}_i - \bar{y}) = 0$$

Simple Linear Regression - Properties of Estimates of Parameters

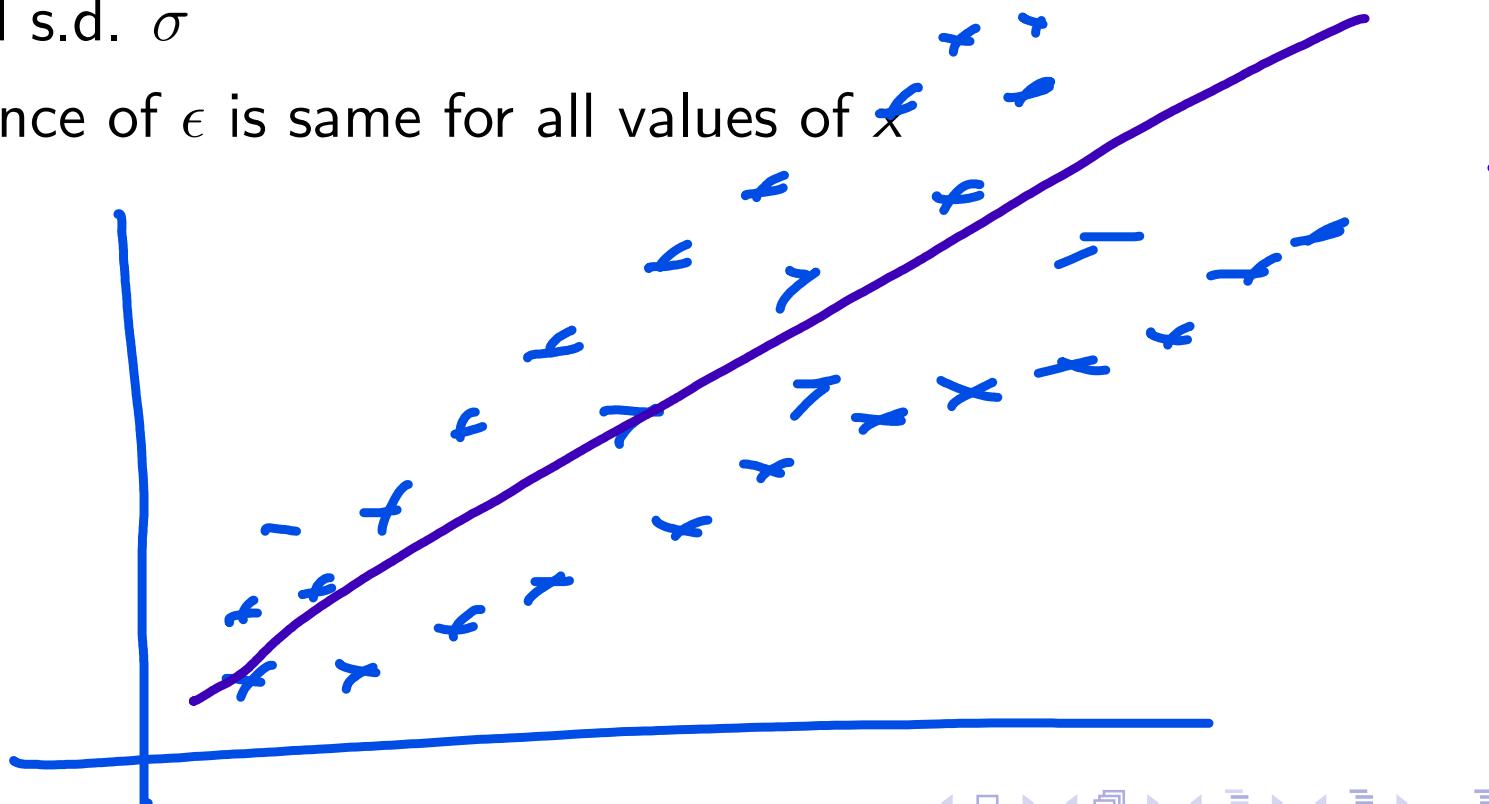
Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

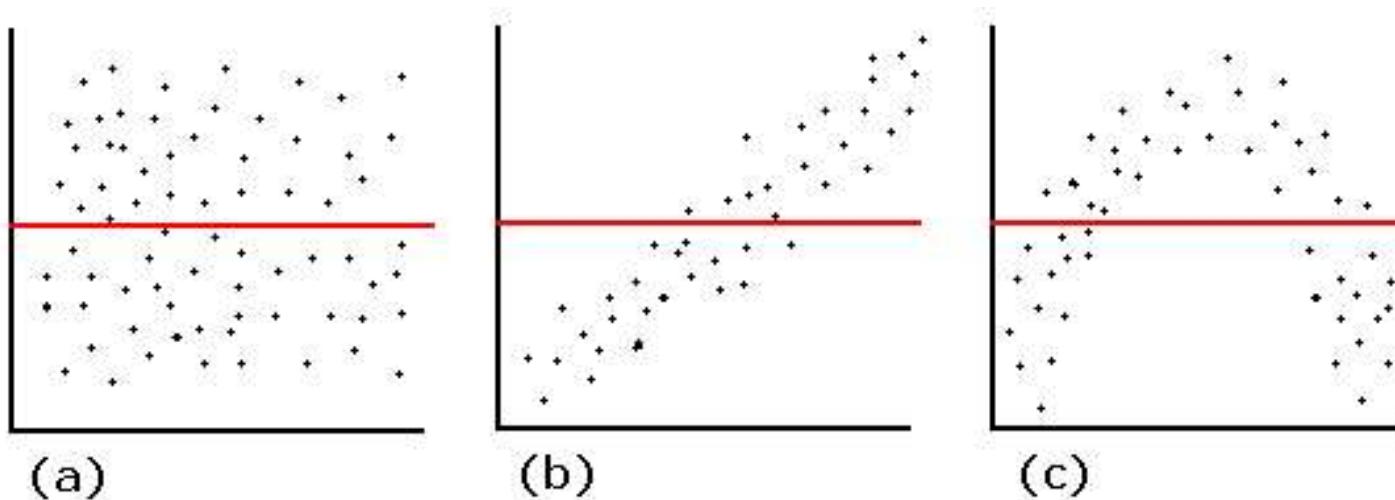
- Sum of residuals is zero
- Residuals are uncorrelated with x_i 's
- It can also be shown that \hat{y}_i and e_i are uncorrelated
- $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$, since $y_i = \hat{y}_i + e_i$

Assumptions of Regression

- ϵ is a random variable that is normally distributed with mean 0 and s.d. σ
- Variance of ϵ is same for all values of x



Examples of Residual Plots



Source - <http://analyticspro.org/2016/03/05/r-tutorial-residual-analysis-for-regression/>

Multiple Linear Regression

We now have more than 1 independent variables. (say k)

Multiple Linear Regression Model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- Interpretation of β' s??
- How do you obtain α & β' s??
- Partial Differentiation to obtain $k + 1$ equations in $k + 1$ unknowns
- Example

Some other aspects in Regression

- Adjusted R^2

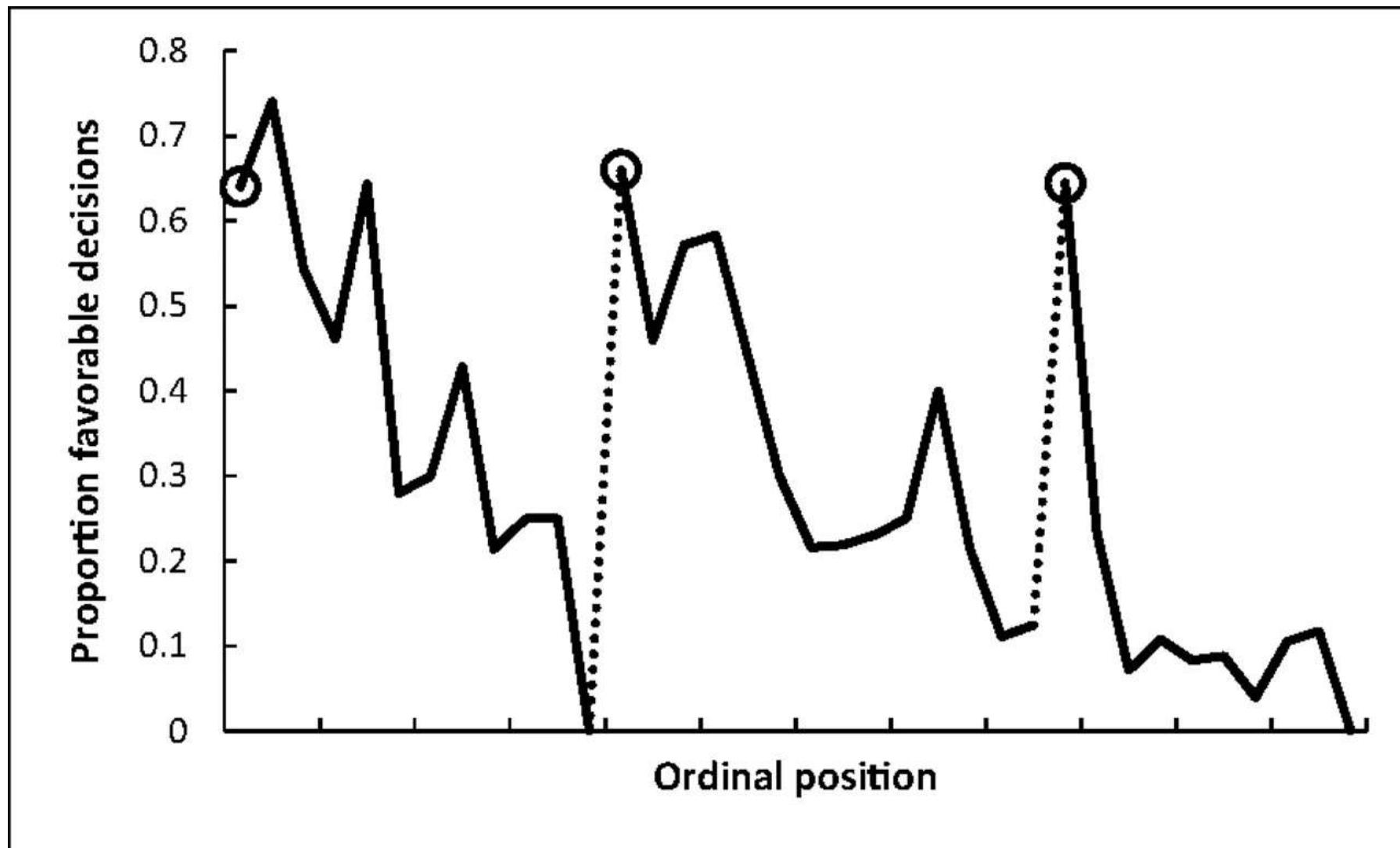
$$R_{adj}^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

- Outliers
- Multi-collinearity

Example - 1

US Election Result. A model uses 30 measurements of various economic, financial and societal quantities (inflation, GDP, crime rate, etc.) Model correctly predicts the winner of all elections 1928-2020. Can it be used to predict the results for 2024 elections?

Example - 2



Source - <https://www.pnas.org/doi/full/10.1073/pnas.1110910108/>

Logistic Regression

Example - A customer purchase decision for onion during her visits to a vegetable store are shown below in the table.

Visit Index	1	2	3	4	5	6	7	8	9	10
Price	2	2	2	2	2	3	3	3	3	4
Decision	Y	Y	Y	N	Y	Y	Y	N	N	Y

Visit Index	11	12	13	14	15	16	17	18	19	20
Price	4	4	4	4	4	5	5	5	5	5
Decision	N	N	Y	N	N	N	N	N	N	Y

Logistic Regression

$$(p_2 p_2 p_2(1-p_2)p_2)(p_3^2(1-p_3)^2)(p_4^2(1-p_4)^4)$$

Example - A customer purchase decision for onion during her visits to a vegetable store are shown below in the table.

$$p_5(1-p_5)^4$$

Visit Index	1	2	3	4	5	6	7	8	9	10
Price	2	2	2	2	2	3	3	3	3	4
Decision	Y	Y	Y	N	Y	Y	Y	N	N	Y

Visit Index	11	12	13	14	15	16	17	18	19	20
Price	4	4	4	4	4	5	5	5	5	5
Decision	N	N	Y	N	N	N	N	N	N	Y

When the price is 2.5, what is the probability that the person will make a purchase?

Example Contd-

When the price is 2.5, what is the probability that the person will make a purchase?

Price	No. of Visits	No. of Purchases	Probability of Purchase
2	5	4	0.8
3	4	2	0.5
4	6	2	0.33
5	5	1	0.2

Example Contd-

When the price is 6, what is the probability that the person will make a purchase?

Price	No. of Visits	No. of Purchases	Probability of Purchase
2	5	4	0.8
3	4	2	0.5
4	6	2	0.33
5	5	1	0.2

Objective - Find the probability of purchase as a function of price

$$P(\text{purchase}) = f(\text{Price})$$

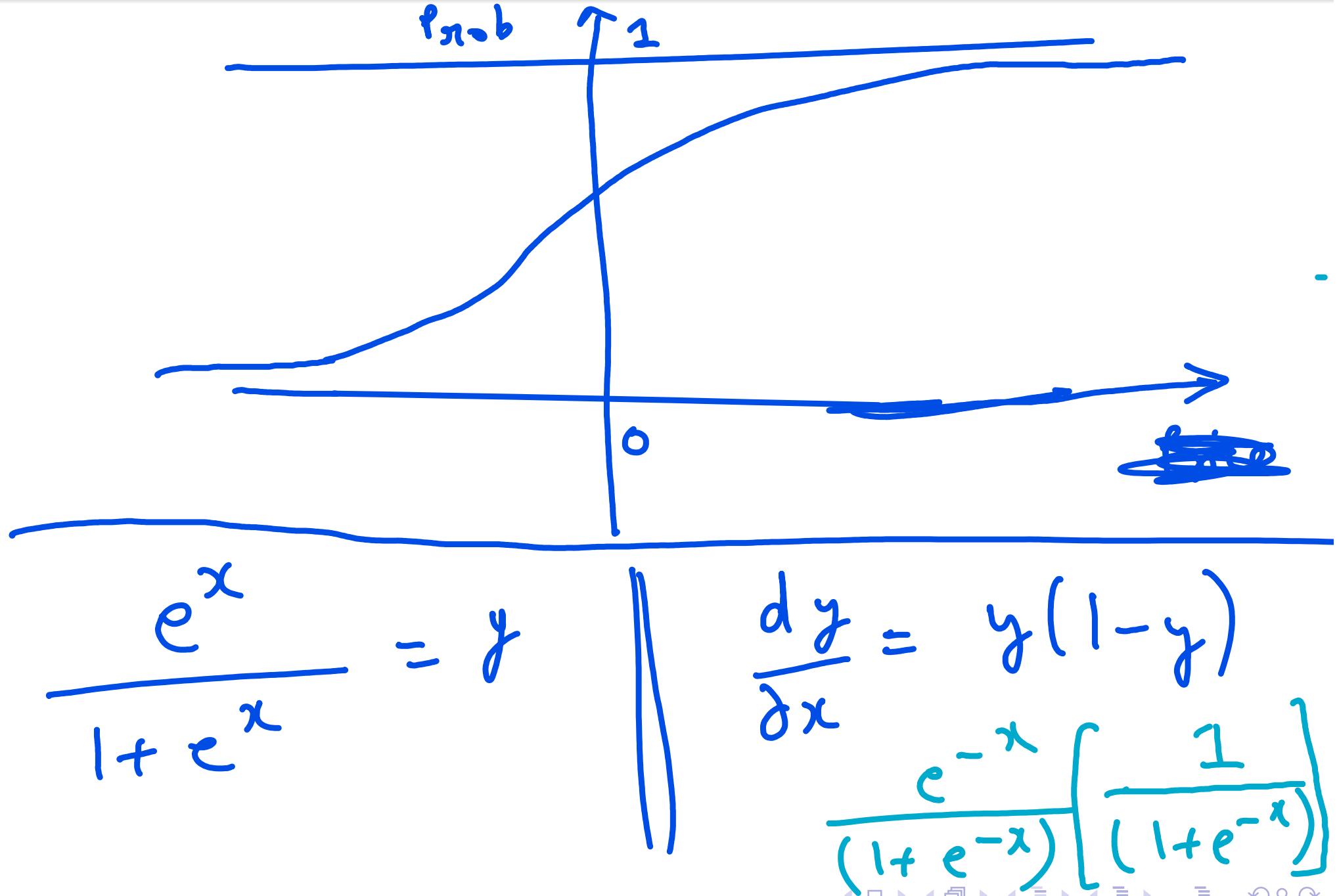
Can we make use of linear regression??

Attempt for a non-linear fit???

Limitations of Linear Regression

- Doesn't explicitly recognize 0-1 nature of the response
- The impact of change in price on probability of purchase decision is different at different levels of price
- Assumed Linear regression equation treats it as a constant

Logistic Regression - Main Ideas



Maximum Likelihood Principle - Ideas

Rain Prediction Model				
	Monday	Tuesday	Wednesday	Thursday
Model A	0.3	0.6	0.7	0.2
Model B	0.2	0.4	0.9	0.6
Rained?	N	Y	Y	N

Model C 0.1 0.6 0.6 0.3

Which Model is better??

Model D 0.1 0.9 0.9 0.1

Maximum Likelihood Principle - Example

Toss a coin 100 times

You get 70 Heads.

[HHTHTHTHTTHTHH-----TH]

What "prob" of Heads will maximize
the chances of observing what you
observed ??

$$p_n^{70} (1-p_n)^{30}$$

Max:

$$70 p_n^{69} (1-p_n)^{30} + (-30) p_n^{70} (1-p_n)^{29} = 0$$

Logistic Regression - Implementation in Python

Multinomial Regression

Multinomial Regression - Softmax Function

Softmax Function - Implementation in Python

Logistic Regression

Feb 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

- Multicollinearity Issues in Regression
- Logistic Regression
- VIF
- Logistic Regression - More discussion
- Softmax

Logistic Regression

Example - A customer purchase decision for onion during her visits to a vegetable store are shown below in the table.

Visit Index	1	2	3	4	5	6	7	8	9	10
Price	2	2	2	2	2	3	3	3	3	4
Decision	Y	Y	Y	N	Y	Y	Y	N	N	Y

Visit Index	11	12	13	14	15	16	17	18	19	20
Price	4	4	4	4	4	5	5	5	5	5
Decision	N	N	Y	N	N	N	N	N	N	Y

When the price is 2.5 or when the price is 6, what is the probability that the person will make a purchase?

Example Contd-

When the price is 2.5 (or 6), what is the probability that the person will make a purchase?

Price	No. of Visits	No. of Purchases	Probability of Purchase
2	5	4	0.8
3	4	2	0.5
4	6	2	0.33
5	5	1	0.2

Example Contd -

Example - A customer purchase decision for onion during her visits to a vegetable store are shown below in the table.

VI	1	2	3	4	5	6	7	8	9	10
P	2.1	2.05	1.98	2	2.2	3.1	3.06	3.02	3.2	4.1
D	Y	Y	Y	N	Y	Y	Y	N	N	Y

VI	11	12	13	14	15	16	17	18	19	20
P	4.07	4.12	4.14	3.98	4.08	5.1	4.99	5.2	5.15	4.98
D	N	N	Y	N	N	N	N	N	N	Y

When the price is 2.5 or when the price is 6, what is the probability that the person will make a purchase?

Maximum Likelihood Principle - Ideas

Rain Prediction Model				
	Monday	Tuesday	Wednesday	Thursday
Model A	0.3	0.6	0.7	0.2
Model B	0.2	0.4	0.9	0.6
Rained?	N	Y	Y	N

Which Model is better??

Logistic Regression - Main Ideas

$$\frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

$$P(\text{Yes}) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

$$P(\text{No}) = \frac{1}{1+e^z}$$

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$$\Pr[\text{Yes given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$

$$\Pr[\text{No given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$$\Pr[\text{Yes given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$

$$\Pr[\text{No given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Replace Yes with +1 and No with -1

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$y = 1$, if Yes, $y = -1$ if No

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$$y = 1, \text{ if Yes}, y = -1 \text{ if No}$$

$$\Pr[+1 \text{ given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$

$$\Pr[-1 \text{ given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

x_0	x_1	x_2	x_3	D
1	-	-	-	+1
1	-	-	-	-1
1	-	-	-	+1
1	-	-	-	-1

Or more compactly

$$\Pr[y \text{ given } x] = \frac{1}{1 + \exp(-y \times w \cdot x)}$$

Maximum Likelihood Principle

Probability of observing the dataset as per the model is

$$\prod_{i=1}^n \Pr [y_i \text{ given } x_i] = \prod_{i=1}^n \frac{1}{1 + \exp(-y_i w \cdot x_i)}$$

We are looking for w which will maximize this, or alternatively w which will minimize the expression below -

$$\min_w \sum_{i=1}^n \log (1 + \exp (-y_i w \cdot x_i))$$

or

$$\min \left(\overline{\left(\sum_{i=1}^n \log \left(\frac{1}{1 + e^{-y_i \vec{w} \cdot \vec{x}_i}} \right) \right)} \right)$$

Logistic Regression - Implementation in Python

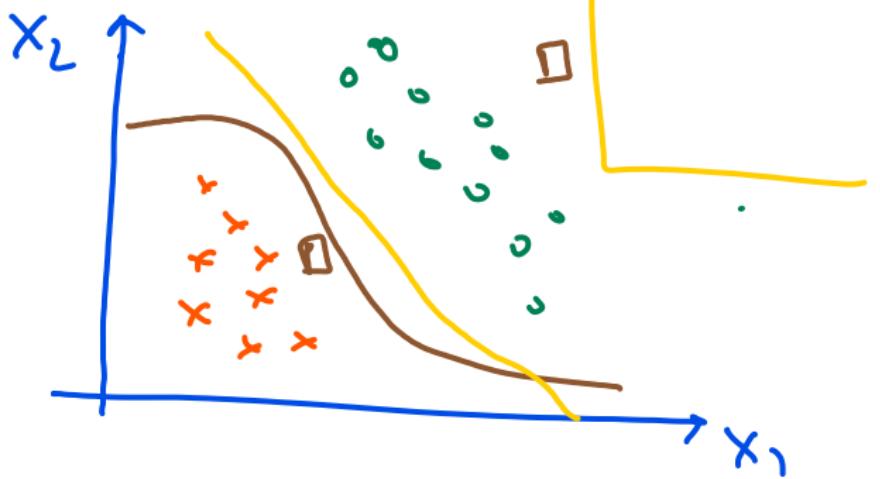
Equivalent Cutoff condition with threshold

$$\alpha, \beta_1, \beta_2$$

$$P(\text{Green}) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2)}}$$

$$\frac{1}{1 + e^{-(\cdot)}} > 0.5$$

Let us say we keep the threshold as 0.5. What is the decision surface?



Multi-class classification - Model

$$\omega_1 = [\alpha_1, \beta_{11}, \beta_{12}, \dots, \beta_{1k}]$$

$$\omega_2 = [\alpha_2, \beta_{21}, \beta_{22}, \dots, \beta_{2k}]$$

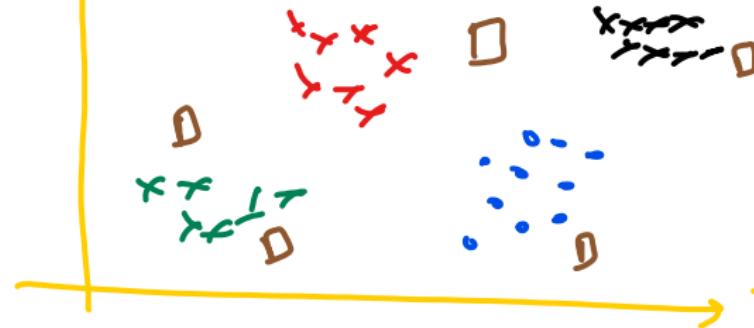
multiple classes in the data

$$y \in \{1, 2, 3, \dots, r\} \quad \omega_r = [\alpha_r, \beta_{r1}, \beta_{r2}, \dots, \beta_{rk}]$$

Instead of a single weight vector w , we consider r weight vectors

- $w_1, w_2, w_3, \dots, w_r$.

$$\Pr[\text{output } i \text{ given } x] = \frac{\exp(w_j \cdot x)}{\sum_{j=1}^r \exp(w_j \cdot x)}$$



Multi-class classification - Model

multiple classes in the data
 $y \in \{1, 2, 3, \dots, r\}$

$$\begin{bmatrix} 1, & 0, & -2 \\ + & 1 & 1 \\ e^1, & e^0, & e^{-2} \end{bmatrix}$$

$$\frac{e^1}{e^1 + e^0 + e^{-2}}$$

Instead of a single weight vector w , we consider r weight vectors $w_1, w_2, w_3, \dots, w_r$.

$$\Pr[\text{output } i \text{ given } x] = \frac{\exp(w_j \cdot x)}{\sum_{j=1}^r \exp(w_j \cdot x)}$$

This is called the Soft-Max function, it converts a given set of numbers to probabilities.

Multi-class classification

X_1	X_2	Y
10	11	A
12	13	B
14	15	C
16	17	D
18	19	A

$$\left(\frac{e^{\alpha_A + \beta_{A1}X_1 + \beta_{A2}X_2}}{e^{\alpha_A + \beta_{A1}X_1 + \beta_{A2}X_2} + e^{\alpha_B + \beta_{B1}X_1 + \beta_{B2}X_2} + e^{\alpha_C + \beta_{C1}X_1 + \beta_{C2}X_2}} \right)$$

$$\alpha_A, \beta_{A1}, \beta_{A2}$$

$$\alpha_B, \beta_{B1}, \beta_{B2}$$

$$\alpha_C, \beta_{C1}, \beta_{C2}$$

$$\square = \alpha_A + 10 \cdot \beta_{A1} + 11 \cdot \beta_{A2}$$

$$\square = \alpha_B + 10 \cdot \beta_{B1} + 11 \cdot \beta_{B2}$$

$$\square = \alpha_C + 10 \cdot \beta_{C1} + 11 \cdot \beta_{C2}$$

Softmax for 2 classes

Thank you for your attention

Logistic Regression

24 Feb 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

- VIF
- Logistic Regression - More discussion
- Softmax
- Logistic Regression - Even more discussion
- Softmax - More discussion
- Implementation of Logistic Regression/Softmax in Python
- Gradient Descent

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$$\Pr[\text{Yes given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$

$$\Pr[\text{No given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$$\Pr[\text{Yes given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$

$$\Pr[\text{No given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Replace Yes with +1 and No with -1

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$y = 1$, if Yes, $y = -1$ if No

Logistic Regression - Main Ideas

Let

$$w = [\alpha, \beta_1, \beta_2, \dots, \beta_k]$$

$$x = [1, x_1, x_2, x_3, \dots, x_k]$$

$y = 1$, if Yes, $y = -1$ if No

$$\Pr[+1 \text{ given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$

$$\Pr[-1 \text{ given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Or more compactly

$$\Pr[y \text{ given } x] = \frac{1}{1 + \exp(-y \times w \cdot x)}$$

Maximum Likelihood Principle

Probability of observing the dataset as per the model is

$$\prod_i \Pr [y_i \text{ given } x_i] = \prod_i \frac{1}{1 + \exp(-y_i w \cdot x_i)}$$

We are looking for w which will maximize this, or alternatively w which will minimize the expression below -

$$\min_w \sum_i \log (1 + \exp (-y_i w \cdot x_i))$$

Logistic Regression

A salesperson has visited 1000 customers to sell a book in a city. He has the data of several attributes of these customers (Income, education level, interest in reading, etc.). Finally, he has also maintained the record of who purchased the book or not. He uses logistic regression model to build a model for this data. Assume that the cost of the book is Rs. 500, selling price of the book is Rs. 3000, average cost of visiting a customer is Rs. 200.

The salesperson gets a new list of 10000 customers and the attributes used in building the model. For a person, the probability of purchasing the book comes out as 0.6. Should the salesperson visit this customer?

$$2300 \times 0.6 + (-200) \times 0.4 = 1300$$

Logistic Regression

$$2300p + (1-p)(-200) > 0 \quad | \boxed{p > 0.08}$$

A salesperson has visited 1000 customers to sell a book in a city. He has the data of several attributes of these customers (Income, education level, interest in reading, etc.). Finally, he has also maintained the record of who purchased the book or not. He uses logistic regression model to build a model for this data. Assume that the cost of the book is Rs. 500, selling price of the book is Rs. 3000, average cost of visiting a customer is Rs. 200.

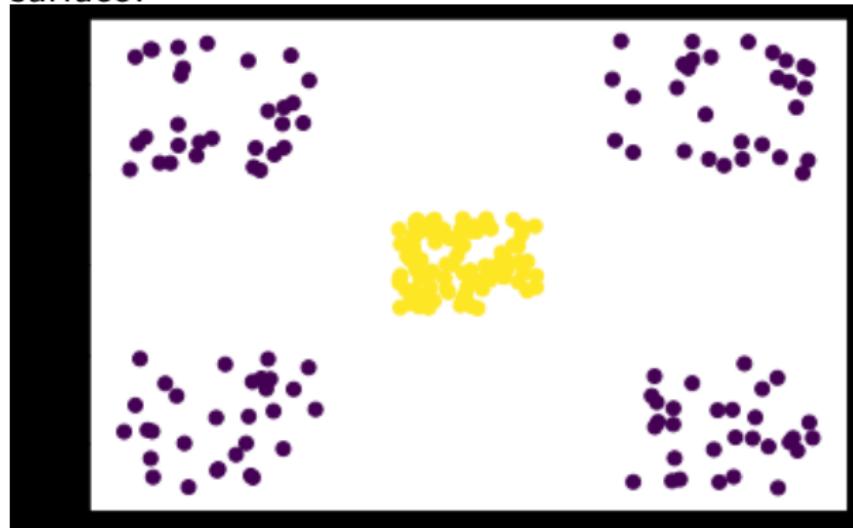
The salesperson gets a new list of 10000 customers and the attributes used in building the model. For a person, the probability of purchasing the book comes out as 0.6. Should the salesperson visit this customer?

What if the probability comes out as 0.4?

Logistic Regression - Implementation in Python

Equivalent Cutoff condition with threshold

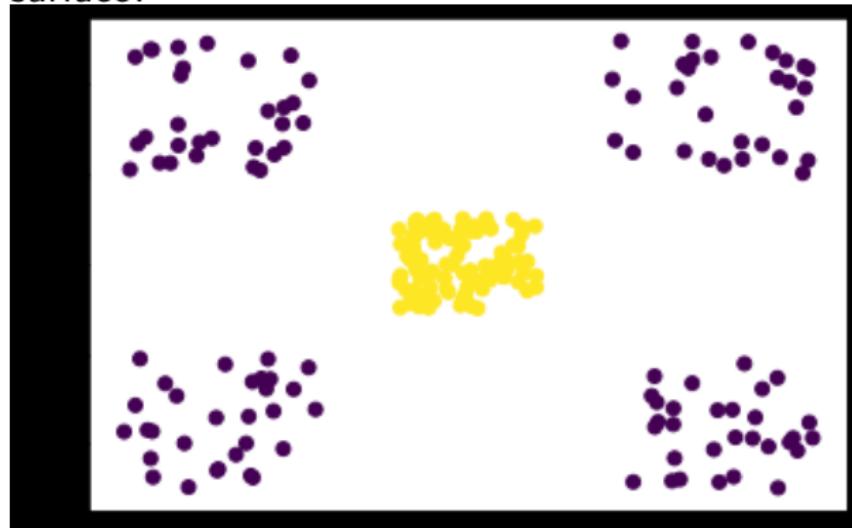
Let us say we keep the threshold as 0.5. What is the decision surface?



Decision Surface is a line (when data is 2-D, else a plane (or hyperplane))

Equivalent Cutoff condition with threshold

Let us say we keep the threshold as 0.5. What is the decision surface?



Decision Surface is a line (when data is 2-D, else a plane (or hyperplane))

For any threshold, logistic regression cannot be a good fit to this dataset because??

multiple classes in the data

$$y \in \{1, 2, 3, \dots, r\}$$

Instead of a single weight vector w , we consider r weight vectors $w_1, w_2, w_3, \dots, w_r$.

$$\Pr[\text{output } i \text{ given } x] = \frac{\exp(w_j \cdot x)}{\sum_{j=1}^r \exp(w_j \cdot x)}$$

Multi-class classification - Model

multiple classes in the data

$$y \in \{1, 2, 3, \dots, r\}$$

Instead of a single weight vector w , we consider r weight vectors $w_1, w_2, w_3, \dots, w_r$.

$$\Pr[\text{output } i \text{ given } x] = \frac{\exp(w_j \cdot x)}{\sum_{j=1}^r \exp(w_j \cdot x)}$$

This is called the Soft-Max function, it converts a given set of numbers to probabilities.

Multi-class classification

S.No	X1	X2	Class
1	0.12	0.11	Green
2	0.14	0.44	Blue
3	0.47	0.50	Red
4	0.43	0.14	Green
5	0.37	0.40	Blue
6	0.13	0.49	Blue
7	0.41	0.13	Blue
8	0.16	0.12	Red
9	0.31	0.38	Green
10	0.22	0.29	Red

Consider a model with

$$\alpha_g = 1, \beta_{g1} = 2, \beta_{g2} = 3$$

$$\alpha_b = 4, \beta_{b1} = -5, \beta_{b2} = 6$$

$$\alpha_r = 7, \beta_{r1} = 8, \beta_{r2} = -9$$

What is the likelihood?

Multi-class classification

Let p_{ij} be the probability that the i^{th} data point is of type j

Let's consider the first data point for which $X_1 = 0.12$, $X_2 = 0.11$

As per the model, the probability of it being green can be computed as -

$$p_{1g} = \frac{e^{1+2*0.12+3*0.11}}{e^{1+2*0.12+3*0.11} + e^{4+(-5)*0.12+6*0.11} + e^{7+8*0.12+(-9)*0.11}}$$

Similarly,

$$p_{1b} = \frac{e^{4+(-5)*0.12+6*0.11}}{e^{1+2*0.12+3*0.11} + e^{4+(-5)*0.12+6*0.11} + e^{7+8*0.12+(-9)*0.11}}$$

And,

$$p_{1r} = \frac{e^{7+8*0.12+(-9)*0.11}}{e^{1+2*0.12+3*0.11} + e^{4+(-5)*0.12+6*0.11} + e^{7+8*0.12+(-9)*0.11}}$$

Example - Contd.

A similar exercise can be performed on all the data points. We get the following result.

S.No	Green	Blue	Red	Observed Class
1	0.004265	0.051441	0.944294	Green
2	0.029431	0.830513	0.140056	Blue
3	0.047325	0.158705	0.79397	Red
4	0.001005	0.001515	0.99748	Green
5	0.027343	0.136793	0.835864	Blue
6	0.025891	0.910426	0.063683	Blue
7	0.001005	0.001691	0.997304	Blue
8	0.003846	0.036124	0.96003	Red
9	0.028342	0.203231	0.768427	Green
10	0.0173	0.177809	0.804891	Red

Thus, the likelihood of observing the data given the model is -

$$P_{1g} \cdot P_{2b} \cdot P_{3r} \cdot P_{4g} \cdot P_{5b} \cdot P_{6b} \cdot P_{7b} \cdot P_{8r} \cdot P_{9g} \cdot P_{10r}$$

Softmax for 2 classes

$$P(X = Y_0) = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}$$

$$P(X = Y_1) = \frac{e^{\alpha_Y + \beta_{1Y} x_1 + \beta_{2Y} x_2 + \dots + \beta_{kY} x_k}}{1 + e^{\alpha_Y + \beta_{1Y} x_1 + \beta_{2Y} x_2 + \dots + \beta_{kY} x_k}}$$

$$P(X = N_0) = \frac{e^{\alpha_N + \beta_{1N} x_1 + \dots + \beta_{kN} x_k}}{1 + e^{\alpha_N + \beta_{1N} x_1 + \dots + \beta_{kN} x_k}}$$

Gradient Descent

Thank you for your attention

SVM

14 March 2023

Sumit Kumar Yadav

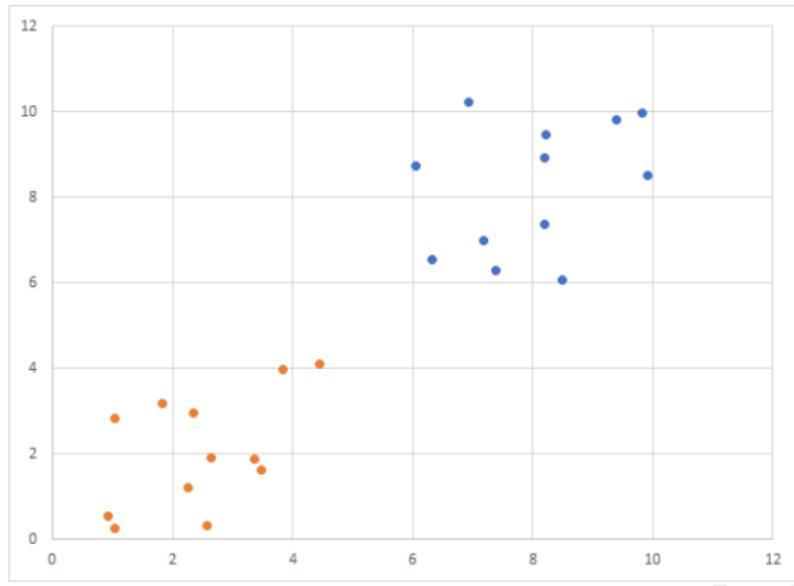
Department of Management Studies
Indian Institute of Technology, Roorkee

Recap and Today

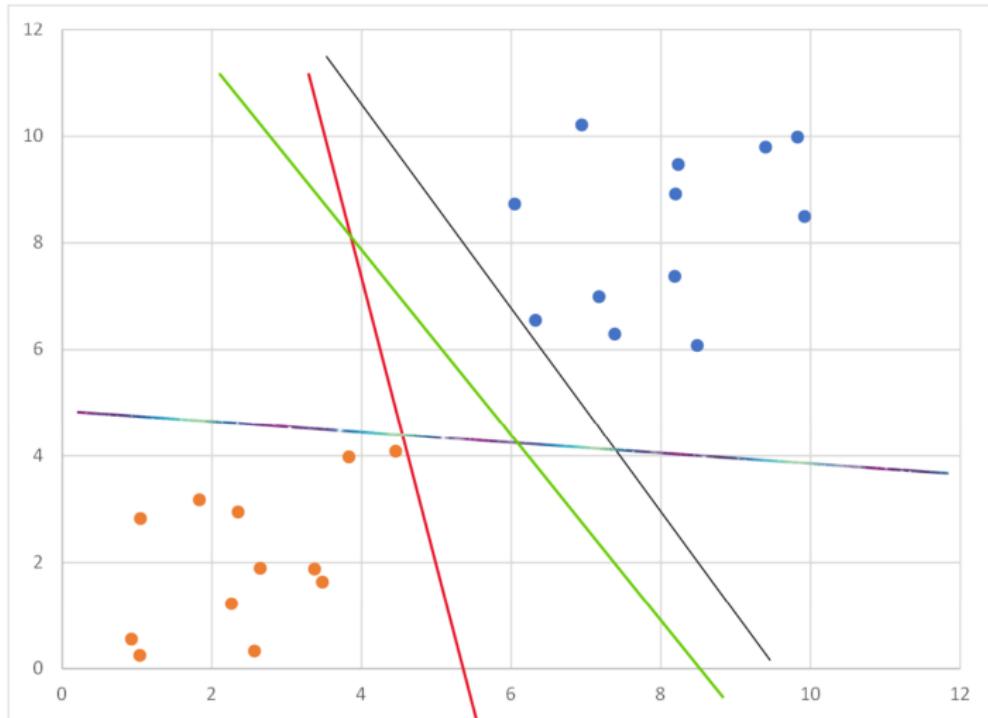
- SVM
- More of SVM

Support Vector Machine - Introduction

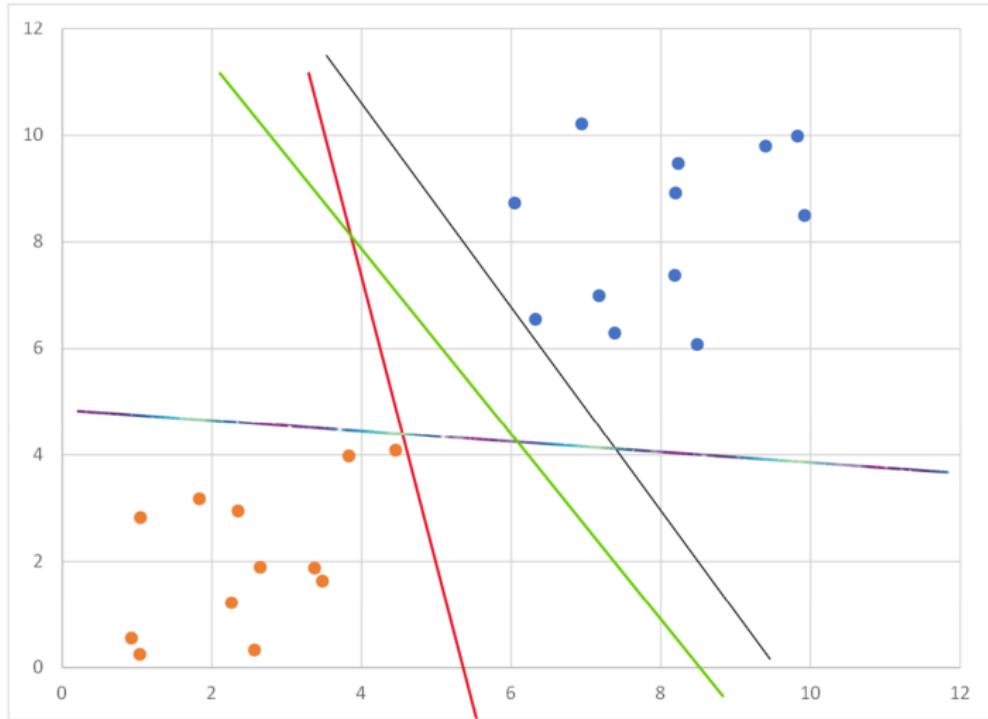
- Supervised Learning Algorithm for Classification
- Given N training points in which n_1 are of type A, n_2 are of type B, draw the **best** line(plane)
- To begin with, assume that the training points are linearly separable



Which is the best line?



Which is the best line?



What makes us think it is the green line? Can we make the ideas a bit more precise?

Notations

Let the data-set be denoted as -

S.No	X_1	X_2	Y (+1 or -1)
1	x_{11}	x_{12}	+1
2	x_{21}	x_{22}	-1
3	x_{31}	x_{32}	-1
.	.	.	.
.	.	.	.
.	.	.	.
N	x_{N1}	x_{N2}	+1

Notations

Let the data-set be denoted as -

S.No	X_1	X_2	Y (+1 or -1)
1	x_{11}	x_{12}	+1
2	x_{21}	x_{22}	-1
3	x_{31}	x_{32}	-1
.	.	.	.
.	.	.	.
.	.	.	.
N	x_{N1}	x_{N2}	+1

Let the equation of the line be $w_1x_1 + w_2x_2 + b = 0$

We need to determine w_1 , w_2 and b

Obtaining w_1 , w_2 and b

- Arbitrarily choose w_1 , w_2 and b such that
 $w_1x_{i1} + w_2x_{i2} + b > 0$ whenever $y_i = +1$
 $w_1x_{i1} + w_2x_{i2} + b < 0$ whenever $y_i = -1$
- Simply put, for all points, $y_i(w_1x_{i1} + w_2x_{i2} + b) > 0$
- Criteria - Consider a line. Find the distance of the line from all the training examples (or points). Look at the minimum of all these distances.

We are interested in the line for which this minimum distance is as large as possible.

- $\max_{(w_1, w_2, b)} \left(\min_{i=\{1,2,\dots,N\}} \frac{|w_1x_{i1} + w_2x_{i2} + b|}{\sqrt{w_1^2 + w_2^2}} \right)$
- The **max** in the equation is **maximize** and **min** in the equation is **minimum**

Optimization Problem

The optimization problem thus becomes -

$$\max_{(w_1, w_2, b)} \left(\min_{i=\{1,2,\dots,N\}} \frac{|w_1x_{i1} + w_2x_{i2} + b|}{\sqrt{w_1^2 + w_2^2}} \right)$$

The **max** in the equation is **maximize** and **min** in the equation is **minimum**

subject to the following **N** constraints - $y_i(w_1x_{i1} + w_2x_{i2} + b) > 0$

Optimization Problem

The optimization problem thus becomes -

$$\max_{(w_1, w_2, b)} \left(\frac{1}{\sqrt{w_1^2 + w_2^2}} \left[\min_{i=\{1,2,\dots,N\}} (|w_1 x_{i1} + w_2 x_{i2} + b|) \right] \right)$$

subject to the following **N** constraints -

$$y_i(w_1 x_{i1} + w_2 x_{i2} + b) > 0$$

Optimization Problem

The optimization problem thus becomes -

$$\max_{(w_1, w_2, b)} \left(\frac{1}{\sqrt{w_1^2 + w_2^2}} \left[\min_{i=\{1,2,\dots,N\}} (|w_1x_{i1} + w_2x_{i2} + b|) \right] \right)$$

subject to the following **N** constraints -

$$y_i(w_1x_{i1} + w_2x_{i2} + b) > 0$$

As scaling all w_1 , w_2 and b by the same factor (non-zero) doesn't change the line (or hyperplane), we will choose w_1 , w_2 and b such that -

$$\min_{i=\{1,2,\dots,N\}} (|w_1x_{i1} + w_2x_{i2} + b|) = 1$$

The **min** in the above equation is **minimum**

Optimization Problem

The optimization problem thus becomes -

$$\max_{(w_1, w_2, b)} \left(\frac{1}{\sqrt{w_1^2 + w_2^2}} \right)$$

subject to the following **2N** constraints -

$$y_i(w_1x_{i1} + w_2x_{i2} + b) > 0$$

$$\min_{i=\{1,2,\dots,N\}} (|w_1x_{i1} + w_2x_{i2} + b|) = 1$$

Optimization Problem

The optimization problem thus becomes -

$$\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right)$$

subject to the following **2N** constraints -

$$y_i(w_1x_{i1} + w_2x_{i2} + b) > 0$$

$$\min_{i=\{1,2,\dots,N\}} (|w_1x_{i1} + w_2x_{i2} + b|) = 1$$

Optimization Problem

The optimization problem thus becomes -

$$\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right)$$

subject to the following **2N** constraints -

$$y_i(w_1x_{i1} + w_2x_{i2} + b) > 0$$

$$\min_{i=\{1,2,\dots,N\}} (|w_1x_{i1} + w_2x_{i2} + b|) = 1$$

$$\min_{i=\{1,2,\dots,N\}} (|w_1x_{i1} + w_2x_{i2} + b|) = 1 \text{ implies -}$$

$$|w_1x_{i1} + w_2x_{i2} + b| \geq 1 \quad \forall i = \{1, 2, \dots, N\} \text{ or}$$

$$|y_i(w_1x_{i1} + w_2x_{i2} + b)| \geq 1 \quad \forall i = \{1, 2, \dots, N\} \text{ or}$$

$y_i(w_1x_{i1} + w_2x_{i2} + b) \geq 1 \quad \forall i = \{1, 2, \dots, N\}$ The implies condition is not both ways, but still it can be replaced in this problem because ??

Optimization Problem

The implies condition can be replaced in this problem because ??
After some algebra, the optimization problem becomes -

$$\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right)$$

subject to the following **N** constraints -

$$y_i(w_1x_{i1} + w_2x_{i2} + b) \geq 1 \quad \forall i = \{1, 2, \dots, N\}$$

Why the name support vector?

Consider the following two optimization problems -

$$\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right)$$

subject to the following **N** constraints -

$$y_i(w_1x_{i1} + w_2x_{i2} + b) \geq 1 \quad \forall i = \{1, 2, \dots, N\}$$

$$\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right) - \sum_{i=1}^N \alpha_i (y_i(w_1x_{i1} + w_2x_{i2} + b) - 1)$$

subject to no constraints, only the fact that all α_i 's are either zero or positive

Which of these two optimization problems has a lower value?

Why the name support vector?

$$\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right)$$

subject to the following **N** constraints -

$$y_i(w_1x_{i1} + w_2x_{i2} + b) \geq 1 \quad \forall i = \{1, 2, \dots, N\}$$

$$\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right) - \sum_{i=1}^N \alpha_i (y_i(w_1x_{i1} + w_2x_{i2} + b) - 1)$$

subject to no constraints, only the fact that all α_i 's are either zero or positive

Let us say that the optimization problem in blue box is optimal for $w_1 = w_1^*$, $w_2 = w_2^*$ and $b = b^*$. The value of the optimization problem in red box is lower at these values. Thus, the one in the red box may have a further lower optimal value. Now, let us keep playing with putting different values of α_i 's.

Why the name support vector?

Let us keep playing with putting different values of α_i 's and try to solve the following optimization problem.

$$\max_{\alpha_i} \left[\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right) - \sum_{i=1}^N \alpha_i (y_i(w_1 x_{i1} + w_2 x_{i2} + b) - 1) \right]$$

subject to the constraint that all α_i 's are either zero or positive

The following can be shown, with some difficulty (we will not be looking at the proof of this). Refer KKT conditions.

- The value of the optimization problem above will be the same as the value of the optimization problem in the blue box in the previous slide
- The optimal value will be attained for $w_1 = w_1^*$, $w_2 = w_2^*$ and $b = b^*$

Why the name support vector?

$$\max_{\alpha_i} \left[\min_{(w_1, w_2, b)} \left(\frac{w_1^2 + w_2^2}{2} \right) - \sum_{i=1}^N \alpha_i (y_i(w_1 x_{i1} + w_2 x_{i2} + b) - 1) \right]$$

subject to the constraint that all α_i 's are either zero or positive

The inner optimization problem can be solved like a usual minimization problem with no constraints. We take partial derivative with respect to w_1 , w_2 and b to get the following:

- $w_1 = \sum_{i=1}^n \alpha_i y_i x_{i1}$ and $w_2 = \sum_{i=1}^n \alpha_i y_i x_{i2}$
- $\sum_{i=1}^n \alpha_i y_i = 0$

Only the training points for which α_i is non-zero contribute in deciding the value of w_1 and w_2 . These points are called support vectors.

The new optimization problem

On substituting the conditions, w_1 , w_2 and b disappear from the inner optimization problem.

$$\max_{\alpha_i} \left[\sum_{i=1}^n \alpha_i - \frac{1}{2} \left(\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_{i1} x_{j1} + x_{i2} x_{j2}) \right) \right]$$

subject to the following **N+1** constraints that

$$\alpha_i \geq 0$$

$$\sum_{i=1}^n \alpha_i y_i = 0$$

The non-separable cases - Kernel Trick

$$\max_{\alpha_i} \left[\sum_{i=1}^n \alpha_i - \frac{1}{2} \left(\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j) \right) \right]$$

subject to the following **N+1** constraints that

$$\alpha_i \geq 0$$

$$\sum_{i=1}^n \alpha_i y_i = 0$$

Thinking of the points as vectors.

Thank you for your attention

GMM

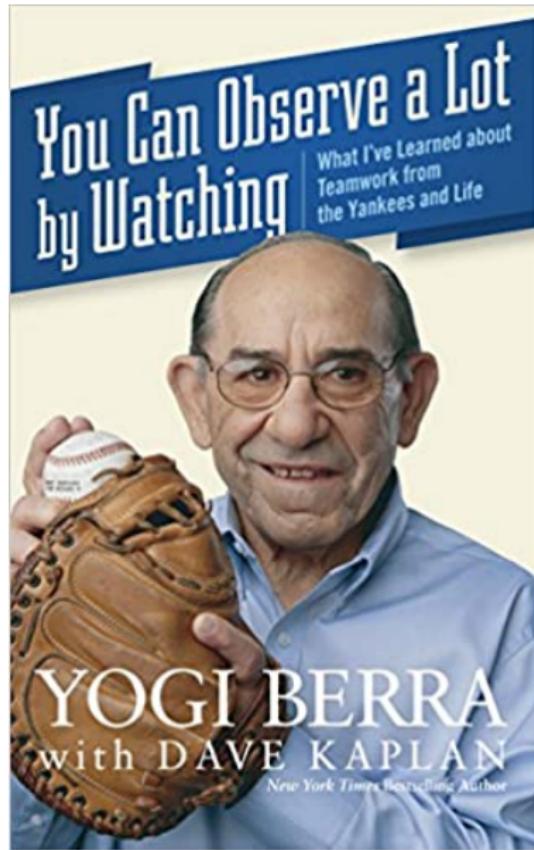
kMeans

28 March 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Clustering



Recap and Today

- kMeans
- Elbow Method
- DB-Indec

Definition Attempt 1 - "subset of points that are closer to each other than to all other data points"

Definition Attempt 1 - "subset of points that are closer to each other than to all other data points

Definition Attempt 2 - Represent a cluster by its center/mean.
Points in a cluster are closer to center/mean of their own cluster
than to the mean of other clusters. (Circular definition because??)

- View points as union of k disjoint clusters - C_1, C_2, \dots, C_k
- Each point lies in exactly one

k-means Clustering problem

- Let the points be x_1, x_2, \dots, x_n
- Mean of the j^{th} cluster =

$$c_j = \frac{1}{m_j} \sum_{i \in C_j} x_i$$

m_j is the number of points in the j^{th} cluster

- Define cost of a cluster as - sum of squared distance from the points to the mean -

$$\sum_{i \in C_j} \|x_i - c_j\|^2$$

- k-means problem : Partition points into k clusters so as to

$$\text{minimize sum of cluster costs} - \sum_{j=1}^k \sum_{i \in C_j} \|x_i - c_j\|^2$$

k-Means algorithm

- Maintain clusters C_1, C_2, \dots, C_k
- Compute the cluster centers for these clusters
- Iteration - For each point, assign it to the c_j that it is closest to. Update C_1, C_2, \dots, C_k and proceed to the next iteration

Finding the value of K

- Elbow Method

Finding the value of K

- Elbow Method
- DB Index

Define cluster dispersion for the j^{th} cluster as -

$$d_j = \sqrt{\frac{1}{m_j} \sum_{i \in C_j} \|x_i - c_j\|^2}$$

- Define cluster similarity between 2 clusters j and l as -

$$S_{jl} = \frac{d_j + d_l}{\|c_j - c_l\|}$$

- $V_{DB} = \frac{1}{K} \sum_{i=1}^K \max_{l \neq i} S_{il}$

Gaussian Mixture Models

Pre-requisites for GMM

- Normal distribution
- Multivariate normal distribution
- Probability Basics
- Maximum Likelihood

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1 . We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1 . We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.

Case - 1 : Assume these parameters to be known,
 $p_1 = 0.8, p_2 = 0.2, p_{h1} = 0.9, p_{h2} = 0.75$

Analogous problem

$$P(\text{coin 1} | 95 \text{ heads}) = \frac{P(C_1 \cap 95H)}{P(95H)}$$

There are 2 coins. We pick coin 1 with probability p_1 . We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.

Case - 1 : Assume these parameters to be known,

$$p_1 = 0.8, p_2 = 0.2, p_{h1} = 0.9, p_{h2} = 0.75$$

We do the experiment once and observe 95 heads. What is the probability it came from coin 1?

→ 0.8

$$\frac{\frac{100}{95} \left(\begin{matrix} 0.9 \\ 0.1 \end{matrix} \right) P(95H | C_1) \cdot P(C_1)}{P(95H \cap C_1) + P(95H \cap C_2)}$$

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1 . We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.

Analogous problem

$$p_1 = 0.4, p_{H_1} = 0.2 \quad || \quad p_2 = 0.6, p_{H_2} = 0.9$$

There are 2 coins. We pick coin 1 with probability p_1 . We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{H1} and p_{H2} respectively.

Case - 2 : The parameters are not known, all we observe is data from several trials of this experiment. Let us say that the observations are -

19, 24, 89, 88, 92, 16, 94, 86, 21, 92

What are the guesses we would like to make for the parameters?

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1 . We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.

Case - 2 : The parameters are not known, all we observe is data from several trials of this experiment. Let us say that the observations are -

19,24,89,88,92,16,94,86,21,92

What are the guesses we would like to make for the parameters? Can you group the data points into 2 and say one group came from coin 1, and other came from coin 2?

Comparison with the analogous case - In the background, we don't have coins generating the data, but we have normal distributions,

Comparison with the analogous case - In the background, we don't have coins generating the data, but we have normal distributions, and we are interested in making the best guess for the parameters of the normal distributions along with the probability that a randomly chosen data point will come from.

Example

Consider the 30 data points -

109, 10079, 8, 106, 9898, 7, 117, 9920, 11, 84, 10034, 11, 116,
9951, 10, 117, 9980, 13, 115, 9970, 11, 94, 9948, 11, 95, 12, 106,
12, 8, 7

Alright, it is too messy, let us organize it better maybe.

Example

$$p_1 = 0.4$$

$$p_2 = 0.5$$

$$p_3 = 0.1$$

S.No	Set-1	Set-2	Set-3
1	109	10079	8
2	106	9898	7
3	117	9920	11
4	84	10034	11
5	116	9951	10
6	117	9980	13
7	115	9970	11
8	94	9948	11
9	95		12
10	106		12
11			8
12			7

$$\mu_1 = 10, \sigma_1^2 = 5$$

$$\mu_2 = 100, \sigma_2^2 = 10$$

$$\mu_3 = 1000, \sigma_3^2 = 50$$

If we have to think of this as data coming from 3 normal distributions, what could be some sensible parameters of the data generation process.

Example - How about this??

S.No	Set-1	Set-2	Set-3
1	109	10079	8
2	106	9898	7
3	117	9920	11
4	84	10034	11
5	116	9951	10
6	117	9980	13
7	115	9970	11
8	94	9948	11
9	95		12
10	106		12
11			8
12			7
mean	10	100	1
sigma	5	50	0.5
probability	1/3	1/3	1/3

Example - How about this one??

S.No	Set-1	Set-2	Set-3
1	109	10079	8
2	106	9898	7
3	117	9920	11
4	84	10034	11
5	116	9951	10
6	117	9980	13
7	115	9970	11
8	94	9948	11
9	95		12
10	106		12
11			8
12			7
mean	100	10000	10
sigma	10	100	2
probability	10/30	8/30	12/30

Example

We will not get into how these parameters are estimated.
We will just keep in mind that it is done with an approach that is similar to what we did in Logistic Regression or SoftMax.

Maximum Likelihood approach
This is usually done using an Iterative algorithm called Expectation Maximization algorithm

In the example, the data was 1 dimensional. It will not always be the case.

(μ_1, μ_2) (μ_1, μ_2, μ_3)

In the example, the data was 1 dimensional. It will not always be the case.

Welcome Multi-variate normal distribution

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_L^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

In the example, the data was 1 dimensional. It will not always be the case.

Welcome Multi-variate normal distribution

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\}$$

The above equation is density of a D-dimensional normal distribution, Σ is the variance-covariance matrix

So, if we want to make 3 clusters from the data, we would think of the data as a simulation of a data generation process going on in the background. The data generation process will come from 3 normal distributions with their respective parameters. Each normal distribution will be picked with some probability.

So, the parameters will be -

$$p_1, \mu_1, \Sigma_1$$

$$p_2, \mu_2, \Sigma_2$$

$$p_3, \mu_3, \Sigma_3$$

with the condition that $p_1 + p_2 + p_3 = 1$

Deciding the value of k

Two ways - AIC and BIC,
pick the one for which this is minimum.
Again, we will not be getting into details of these.

Thank you for your attention

Principal Component Analysis + Singular Value Decomposition

13 April 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Recap and Today

- PCA theory
- PCA some more theory
- PCA Implementation
- SVD Theory + Implementation

PCA Motivation

- To reduce the number of dimensions in the data
- To visualize the data
- To avoid over fitting

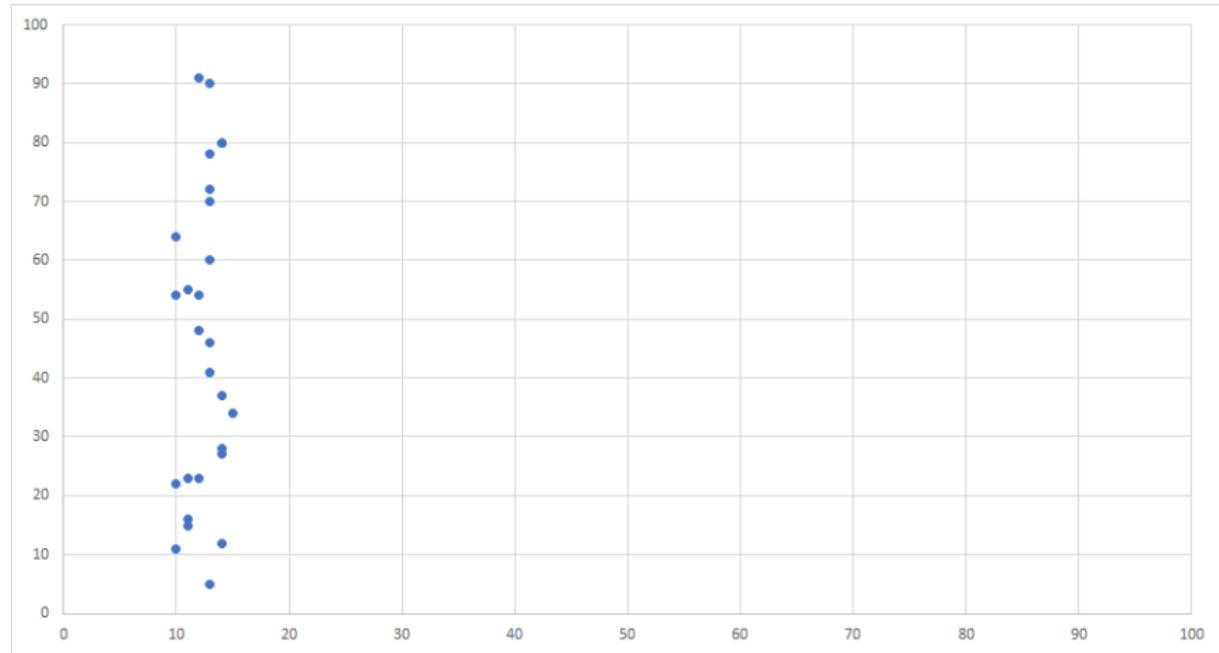
PCA Motivation

- To reduce the number of dimensions in the data
- To visualize the data
- To avoid over fitting

	Location	City	Society	Ambience	Airport
House 1	8	4	9	1	5
House 2	9	6	9	5	5
House 3	10	8	9	7	5
House 4	10	5	9	6	5
House 5	5	4	9	2	5
House 6	2	7	9	9	5
House 7	7	5	9	8	6
House 8	3	4	9	8	5
House 9	4	2	9	7	5
House 10	1	4	9	10	5

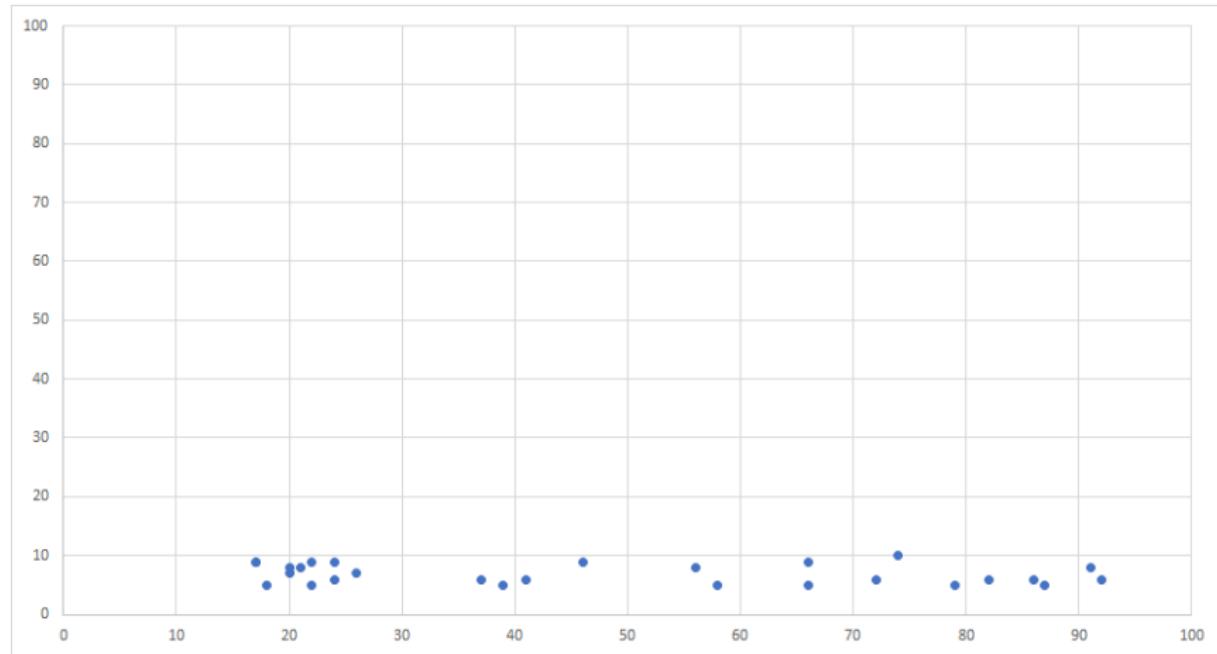
PCA Motivation

Which direction to skip?



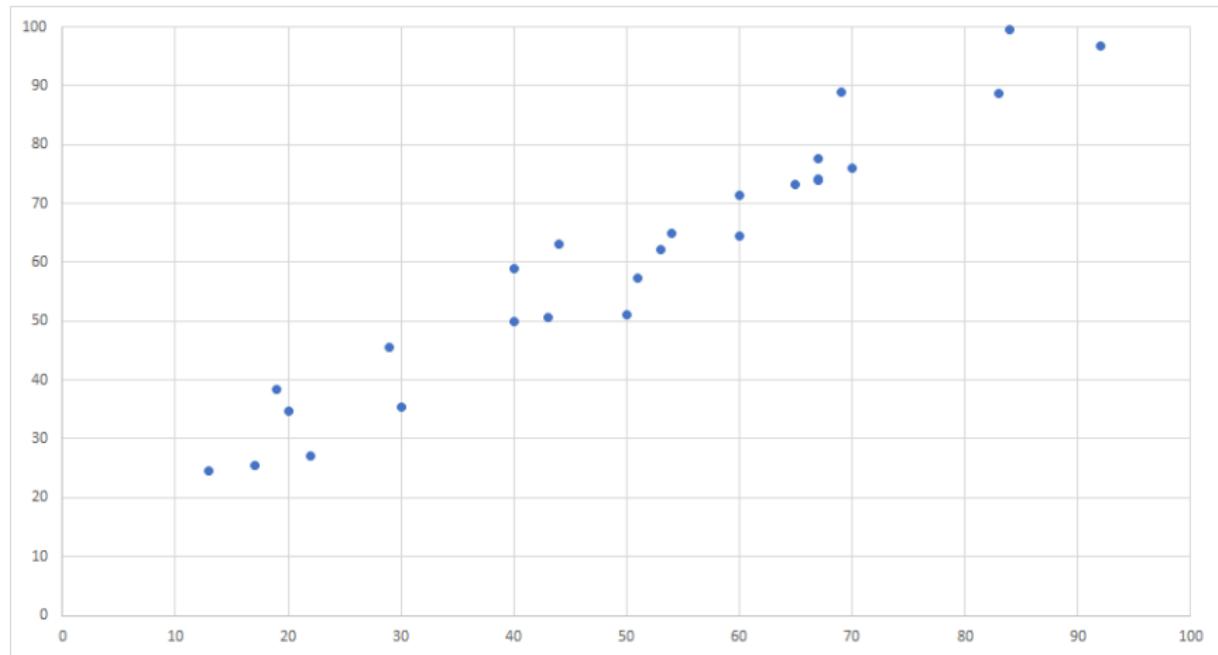
PCA Motivation

Which direction to skip?



PCA Motivation

Which direction to skip?



PCA Motivation

Looks like we are trying to capture as much variance as possible while reducing the number of dimensions.

What is a vector?

- a mathematical object that encodes a length and direction
- A vector is often represented as a 1-dimensional array of numbers, referred to as components and is displayed either in column form or row form
- Represented geometrically, vectors typically represent coordinates within a n-dimensional space

Vectors in R^n /Concept of Vector Space

- Vector Space - Closed under addition and multiplication
- Vectors on line $y = 2 * x + 1$ form a vector space??
- Addition/Subtraction (Graphical Representation)
- Multiplication by a scalar
- Dot Product (Inner Product)
- Length/Magnitude
- Angle between two vectors - $\cos \theta = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \cdot ||\vec{y}||}$
- Dot product of perpendicular vectors = ??

Linearly Independent and Dependent Vectors

- concept of zero vector

Consider m vectors in \mathbb{R}^n ,

If $\alpha_1 \vec{v}_1 + \alpha_2 \vec{v}_2 + \cdots + \alpha_n \vec{v}_n$ implies
 $\alpha_1 = \alpha_2 = \cdots = \alpha_m = \vec{0}$, then

the vectors are said to be linearly independent.

In mathematics, a matrix (plural matrices) is a rectangular array or table of numbers, symbols, or expressions, arranged in rows and columns, which is used to represent a mathematical object or a property of such an object. - Wikipedia

- Matrix Size
- Representing any element of a matrix
- $(1,n)$ and $(m,1)$ matrices
- Square Matrix

Operations on Matrices

- Addition/Subtraction
- Multiplication of Matrix by a scalar
- Transpose of a Matrix
- Multiplication of two Matrices

Matrix as Linear Transformation

Types of Matrices

- Diagonal
- Identity
- Upper Triangular
- Symmetric
- Singular
- Mirror Matrix

- Inverse of a Matrix
- Trace of a Matrix
- Relationship of Eigenvalues with trace and Determinant
- Symmetric Matrix
- Orthogonal Matrix

Eigenvalues and Eigenvectors

Let M be a $n \times n$ matrix. A non-zero vector \vec{X} is said to be an eigenvector of M corresponding to eigenvalue λ if

Spectral Decomposition

A $n \times n$ matrix can be written in terms of its eigenvalues and eigenvectors as follows -

$$M = \lambda_1 \vec{v}_1 \vec{v}_1^T + \lambda_2 \vec{v}_2 \vec{v}_2^T + \dots + \lambda_n \vec{v}_n \vec{v}_n^T$$

Positive Semi-definite Matrix

$$x^\top M x \geq 0$$

All the eigenvalues are greater than zero.

PCA - Objectives

We have a Data Matrix(D), whose dimension is $n * f$.

$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1f} \\ d_{21} & d_{22} & \dots & d_{2f} \\ \vdots & \vdots & \dots & \vdots \\ d_{(n-1)1} & d_{(n-1)2} & \dots & d_{(n-1)f} \\ d_{n1} & d_{n2} & \dots & d_{nf} \end{bmatrix}$$

We want to reduce the number of features to f_s .

PCA - Objectives

We have a Data Matrix(D), whose dimension is $n * f$.

$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1f} \\ d_{21} & d_{22} & \dots & d_{2f} \\ \vdots & \vdots & \dots & \vdots \\ d_{(n-1)1} & d_{(n-1)2} & \dots & d_{(n-1)f} \\ d_{n1} & d_{n2} & \dots & d_{nf} \end{bmatrix}$$

We want to reduce the number of features to f_s .

Let us solve a simpler problem first. Let us say we want to reduce the number of dimensions/features to just 1. Which is the best way to go about it?

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, \dots, v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, \dots, v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

Variance($D\vec{v}$) can be written as $\vec{v}^T \Sigma \vec{v}$, where Σ is the variance covariance matrix of D . (Try to prove this or atleast verify this)

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, \dots, v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

$\text{Variance}(D\vec{v})$ can be written as $\vec{v}^T \Sigma \vec{v}$, where Σ is the variance covariance matrix of D . (Try to prove this or atleast verify this)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1f} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2f} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{(f-1)1} & \sigma_{(f-1)2} & \dots & \sigma_{(f-1)f} \\ \sigma_{1f} & \sigma_{f2} & \dots & \sigma_f^2 \end{bmatrix}$$

PCA Optimization problem

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $\|\vec{v}\| = 1$

PCA Optimization problem

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $\|\vec{v}\| = 1$

Σ can be shown to be positive semi-definite. Thus, all the eigenvalues are greater than zero, and all eigenvectors are orthogonal.

PCA Optimization problem

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $\|\vec{v}\| = 1$

Σ can be shown to be positive semi-definite. Thus, all the eigenvalues are greater than zero, and all eigenvectors are orthogonal.

$$\Sigma = \lambda_1 \vec{v}_1 \vec{v}_1^T + \lambda_2 \vec{v}_2 \vec{v}_2^T + \dots + \lambda_r \vec{v}_r \vec{v}_r^T$$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$$

PCA Optimization problem

Maximize - $\vec{v}^T (\lambda_1 \vec{v}_1 \vec{v}_1^T + \lambda_2 \vec{v}_2 \vec{v}_2^T + \dots + \lambda_n \vec{v}_n \vec{v}_n^T) \vec{v}$

such that $||\vec{v}|| = 1$

PCA Optimization problem

$$\text{Maximize} - \vec{v}^T (\lambda_1 \vec{v}_1 \vec{v}_1^T + \lambda_2 \vec{v}_2 \vec{v}_2^T + \dots + \lambda_n \vec{v}_n \vec{v}_n^T) \vec{v}$$

such that $\|\vec{v}\| = 1$

All \vec{v}_i 's are unit vectors and orthogonal to each other, so the best choice for \vec{v} is ??

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots \\ \sigma_{12} & \sigma_2^2 & \dots \\ \vdots & \vdots & \ddots \\ 1 & \dots & \sigma_f^2 \end{bmatrix} \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_f$$

\vec{v}_1 \vec{v}_2 \vec{v}_f

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

$$D_{n \times f} \quad | \quad D_{\text{trans}} = D \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \vec{v}_3 & \cdots & \vec{v}_f \\ | & | & | & \ddots & | \\ 1 & 1 & 1 & \ddots & 1 \end{bmatrix}$$

+ Ef x fs

$$D_{\text{reduced}} = D_{\text{trans}} E^T$$

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

Fortunately, this happens if we keep going in the same sequence of the principal components.

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

Fortunately, this happens if we keep going in the same sequence of the principal components.

We can now try to maximize the sum of the variance terms in the 2 projected directions.

So, it looks like that the most logical thing to do is to pick up the eigenvector corresponding to the second largest eigenvector. This is what we do in PCA.

Reconstructing the data

One could start with the objective that we want to minimize the reconstruction error, and we would get the same result as what we have described above.

Basic Ideas - Any k-dimensional vector can be represented in terms of k-orthonormal vectors.

Reconstructing the data

One could start with the objective that we want to minimize the reconstruction error, and we would get the same result as what we have described above.

Basic Ideas - Any k-dimensional vector can be represented in terms of k-orthonormal vectors.

Eg - Consider the vector $(2,3)$ in 2-D space and two orthonormal vectors $(1,0)$ and $(0,1)$

$$(2,3) = ((2,3).(1,0))(1,0) + ((2,3).(0,1))(0,1)$$

Reconstructing the data

One could start with the objective that we want to minimize the reconstruction error, and we would get the same result as what we have described above.

Basic Ideas - Any k-dimensional vector can be represented in terms of k-orthonormal vectors.

Eg - Consider the vector $(2,3)$ in 2-D space and two orthonormal vectors $(1,0)$ and $(0,1)$

$$(2,3) = ((2,3).(1,0))(1,0) + ((2,3).(0,1))(0,1)$$

Too trivial, isn't it??

Reconstructing the data

Eg - Consider the vector $(2,3)$ in 2-D space and two orthonormal vectors $(0.6,0.8)$ and $(-0.8,0.6)$

$$(2,3) = ((2,3) \cdot (0.6,0.8))(0.6,0.8) + ((2,3) \cdot (-0.8,0.6))(-0.8,0.6)$$

Reconstructing the data

Eg - Consider the vector (2,3) in 2-D space and two orthonormal vectors (0.6,0.8) and (-0.8,0.6)

$$(2,3) = ((2,3).(0.6,0.8))(0.6,0.8) + ((2,3).(-0.8,0.6))(-0.8,0.6)$$

Thus, in general for a k-dimensional vector \vec{d} , and k-orthonormal vectors $\vec{v}_1, \vec{v}_2, \vec{v}_3, \dots, \vec{v}_k$

$$\vec{d} = (\vec{d} \cdot \vec{v}_1)\vec{v}_1 + (\vec{d} \cdot \vec{v}_2)\vec{v}_2 + (\vec{d} \cdot \vec{v}_3)\vec{v}_3 + \dots + (\vec{d} \cdot \vec{v}_k)\vec{v}_k$$

PCA reconstruction

$$\vec{v}_1, \vec{v}_2, \vec{v}_3, \dots, \dots, \vec{v}_f$$

In PCA, once we have found out the eigenvalues and eigenvectors and projected the data in the lower dimensional space (f_s), the way to reconstruct for a data point back to the original dimensions (f) is -

$$\vec{d}_1 = (\vec{d}_1 \cdot \vec{v}_1) \vec{v}_1 + (\vec{d}_2 \cdot \vec{v}_2) \vec{v}_2 + \dots + (\vec{d}_1 \cdot \vec{v}_f) \vec{v}_f$$

$$\vec{d}_2 = (\vec{d}_2 \cdot \vec{v}_1) \vec{v}_1 + (\vec{d}_2 \cdot \vec{v}_2) \vec{v}_2 + \dots + (\vec{d}_2 \cdot \vec{v}_f) \vec{v}_f$$

.

$$\vec{d}_n = (\vec{d}_n \cdot \vec{v}_1) \vec{v}_1 + (\vec{d}_n \cdot \vec{v}_2) \vec{v}_2 + \dots + (\vec{d}_n \cdot \vec{v}_f) \vec{v}_f$$

PCA reconstruction

In PCA, once we have found out the eigenvalues and eigenvectors and projected the data in the lower dimensional space(f_s), the way to reconstruct for a data point back to the original dimensions (f) is -

$$\vec{d}_{inv} = (\vec{d} \cdot \vec{v}_1) \vec{v}_1 + (\vec{d} \cdot \vec{v}_2) \vec{v}_2 + (\vec{d} \cdot \vec{v}_3) \vec{v}_3 + \dots + (\vec{d} \cdot \vec{v}_{f_s}) \vec{v}_{f_s}$$

PCA reconstruction

In PCA, once we have found out the eigenvalues and eigenvectors and projected the data in the lower dimensional space(f_s), the way to reconstruct for a data point back to the original dimensions (f) is -

$$\vec{d}_{inv} = (\vec{d} \cdot \vec{v}_1) \vec{v}_1 + (\vec{d} \cdot \vec{v}_2) \vec{v}_2 + (\vec{d} \cdot \vec{v}_3) \vec{v}_3 + \dots + (\vec{d} \cdot \vec{v}_{f_s}) \vec{v}_{f_s}$$

This is essentially what we get after the matrix manipulations we saw in the previous session.

Percentage of Variance Captured in PCA

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1f} \\ \sigma_{12} & \sigma_2^2 & & \vdots \\ \vdots & \ddots & & \sigma_f^2 \\ \sigma_{1f} & \sigma_{2f} & \dots & \end{bmatrix}$$
$$\sigma_1^2 + \sigma_2^2 + \dots + \sigma_f^2$$

$$\lambda_1 + \lambda_2 + \dots + \lambda_f = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_f^2$$

Singular Value Decomposition - A way to approximate a general matrix of dimension $m \times n$.

$$M = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \dots - \dots + \lambda_{1000} v_{1000} v_{1000}^T$$

Singular Value Decomposition - A way to approximate a general matrix of dimension $m \times n$.

Can we use PCA to approximate the matrix of size $n \times n$?

$$(1000) \times 5 + 5 = 5005$$

$$1000 \times 1000 = 10 \text{ lakh}$$

Singular Value Decomposition - A way to approximate a general matrix of dimension $m \times n$.

Can we use PCA to approximate the matrix of size $n \times n$? Can we do it always??

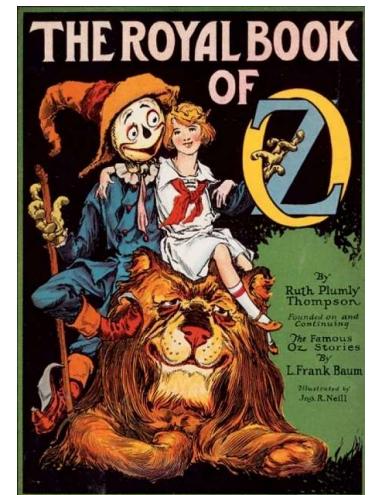
Let A be a general matrix of size $m \times n$.
What can be said about AA^T and A^TA

Let A be a general matrix of size $m \times n$.
What can be said about AA^T and A^TA

Thank you for your attention

Who wrote Royal Book of Oz?

Erica Klarreich “Statistical tests are unraveling knotty literary mysteries”,
Science News Dec 2003.



Books of Oz

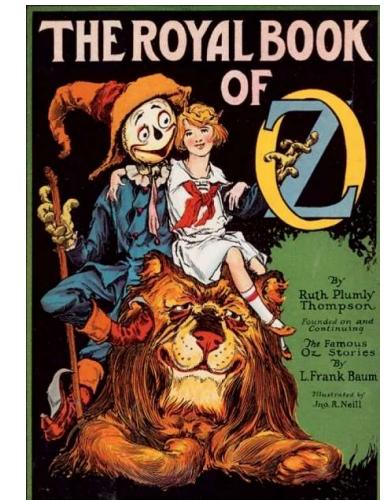
L. Frank Baum (1900-20)



26 Oz books after 1920, most written by Ruth Plumly Thompson

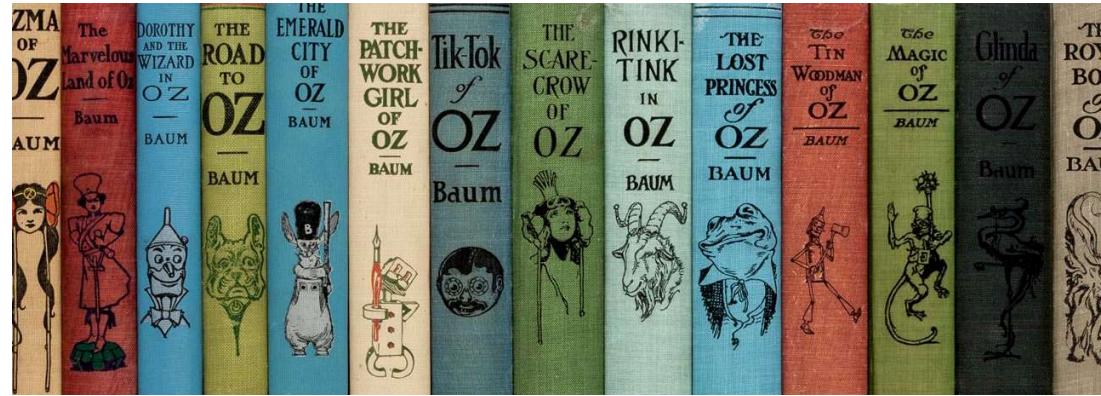
??

(published 1921 under
Baum's name, but
did he truly write it?)





Poll



How many Books of Oz have you read?

- (a) 0 (b) 1-5 (c) 6-20 (d) All of them

Stylometry

Seeks to attribute authorship (e.g. in coauthored texts like *Federalist Papers*, *King James Bible*) via quantitative analyses of author styles.

Surprising finding: Style is easier to spot from usage of *function words* (“to”, “with”, “then”, “however”, etc.) than from less common words.

Example: Alexander Hamilton used “upon” 10 times more frequently than James Madison.



Method

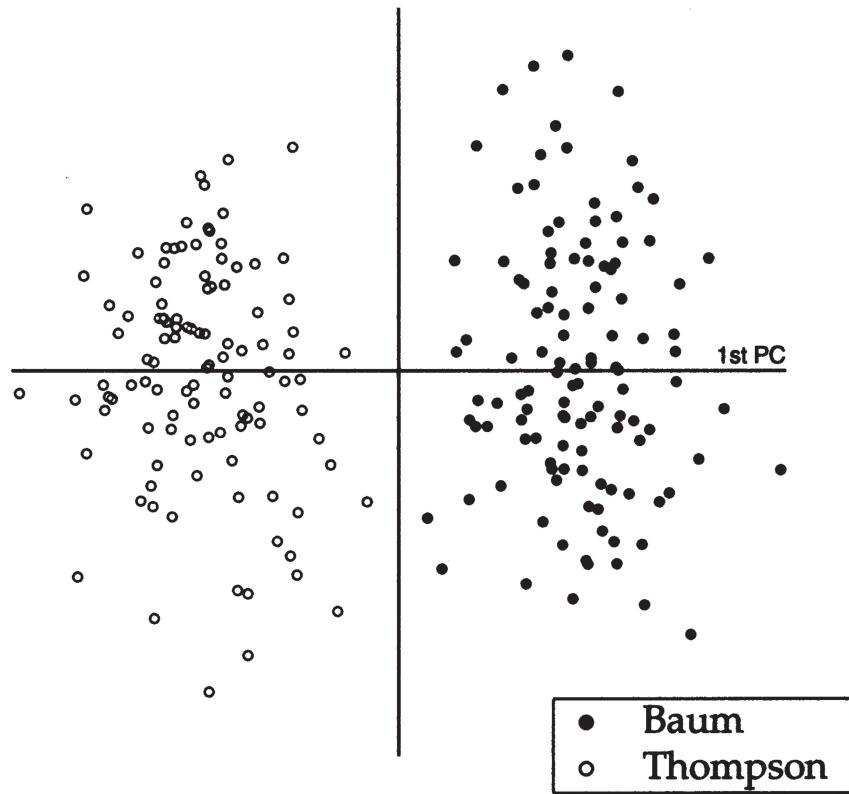
the (6.7%)	with (0.7%)	up (0.3%)	into (0.2%)	just (0.2%)
and (3.7%)	but (0.7%)	no (0.3%)	now (0.2%)	very (0.2%)
to (2.6%)	for (0.7%)	out (0.3%)	down (0.2%)	where (0.2%)
a/an (2.3%)	at (0.6%)	what (0.3%)	over (0.2%)	before (0.2%)
of (2.1%)	this/these (0.5%)	then (0.3%)	back (0.2%)	upon (0.1%)
in (1.3%)	so (0.5%)	if (0.3%)	or (0.2%)	about (0.1%)
that/those (1.0%)	all (0.5%)	there (0.3%)	well (0.2%)	after (0.1%)
it (1.0%)	on (0.5%)	by (0.3%)	which (0.2%)	more (0.1%)
not (0.9%)	from (0.4%)	who (0.3%)	how (0.2%)	why (0.1%)
as (0.7%)	one/ones (0.3%)	when (0.2%)	here (0.2%)	some (0.1%)

Most frequent 50 function words in
Baum/Thompson Oz books

- 1) Divide all Oz books (except “royal book of oz”) into 223 text blocks of 5000 words ea.
- 2) For each block, **measure #occurrences** for each function word.
Obtain 223 vectors in \mathbb{R}^{50}
- 3) Compute their **2D approximation** and visualize in 2D

J.N.G. Binongo, Chance Magazine 2003
<http://dh.obdurodon.org/Binongo-Chance.pdf>

Method

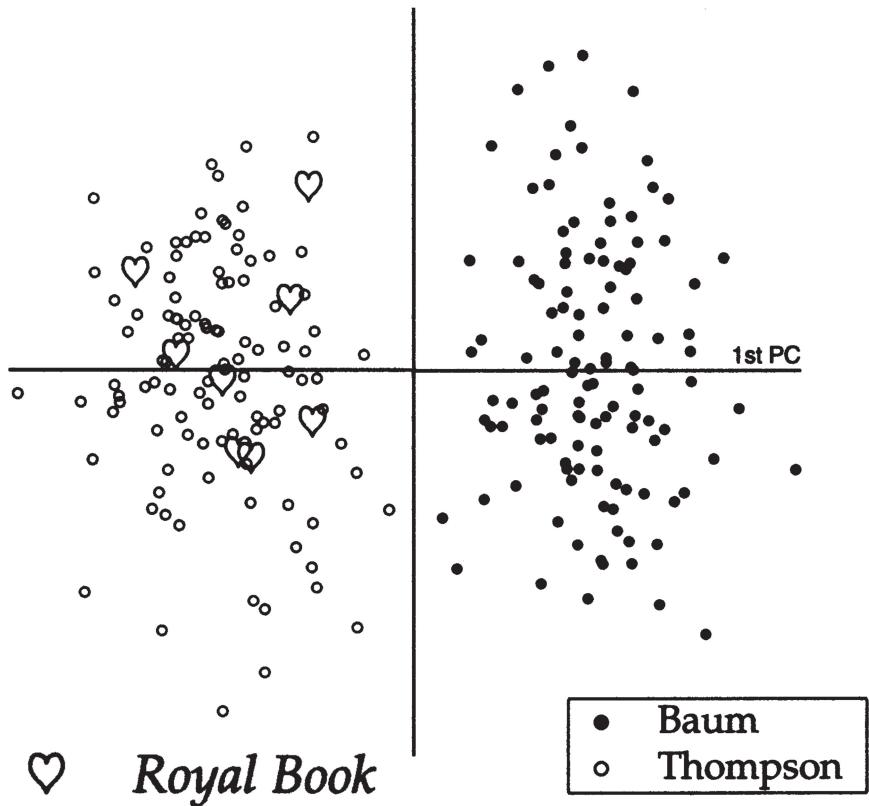


[J.N.G. Binongo, Chance Magazine 2003](#)
<http://dh.obdurodon.org/Binongo-Chance.pdf>

- 1) Divide all Oz books (except “royal book of oz”) into 223 text blocks of 5000 words ea.
- 2) For each block, measure #occurrences for each function word.
Obtain 223 vectors in \mathbb{R}^{50}
- 3) Compute their 2D approximation and visualize in 2D

Heldout data: Non-Oz writings by Baum and Thompson!
Falls neatly onto the correct side;
see Binongo’s paper.

Method



- 1) Divide all Oz books (except “royal book of oz”) into 223 text blocks of 5000 words ea.
- 2) For each block, measure #occurrences for each function word.
Obtain 223 vectors in \mathbb{R}^{50}
- 3) Compute their 2D approximation and visualize in 2D

Ruth Plumly Thompson!

J.N.G. Binongo, Chance Magazine 2003
<http://dh.obdurodon.org/Binongo-Chance.pdf>

Text Analytics

24 April 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Recap and Today(in this session)

- Sentiment Analysis
- Converting text to numbers
- DTM and TF-IDF matrix
- Some more text cleaning exercises and examples
- Cosine Similarity

Consider the following 3 sentences -

- S1 = Text Analytics is boring boring boring
- S2 = Analytics is interesting
- S3 = We want interesting sports analytics

TF-IDF Motivation

Consider the following 3 sentences -

- S1 = Text Analytics is boring boring boring
- S2 = Analytics is interesting
- S3 = We want interesting sports analytics

We can choose to remove the stopwords, convert everything to lowercase and construct the following matrix. We call this DTM or Document Term Matrix.

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1

TF-IDF Motivation

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1

TF-IDF Motivation

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1

We see that analytics and sports is getting the same weightage in S-3, whereas "sports" is exclusive to S-3, "analytics" can be found in all sentences.

In TF-IDF Matrix, we increase the weightage of the words that are exclusive to a document/sentence and decrease the weightage of the words that are common to many sentences.

TF-IDF Motivation

DF = Document Frequency(computed for each term),

IDF = Inverse Document Frequency(computed for each term),

TF = Term Frequency (essentially the DTM matrix),

n = number of documents

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1

TF-IDF Motivation

DF = Document Frequency(computed for each term),

IDF = Inverse Document Frequency(computed for each term),

TF = Term Frequency (essentially the DTM matrix),

n = number of documents

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1

As the name suggests, we would multiply the elements in TF with the corresponding IDF.

Several methods have been proposed in literature for the formula of IDF, one of the common ones is -

$$\text{IDF} = 1 + \ln \left(\frac{1+n}{1+DF} \right)$$

TF-IDF

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1
IDF	$1+\ln(1)$	$1+\ln(2)$	$1+\ln(4/3)$	$1+\ln(2)$	$1+\ln(2)$	$1+\ln(2)$

TF-IDF

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1
IDF	$1+\ln(1)$	$1+\ln(2)$	$1+\ln(4/3)$	$1+\ln(2)$	$1+\ln(2)$	$1+\ln(2)$

We then multiply the TFs with the corresponding IDFs to get-

	analytics	boring	interesting	sports	text	want
S-1	$1*1$	$3*1.693$	$0*1.287$	$0*1.693$	$1*1.693$	$0*1.693$
S-2	$1*1$	$0*1.693$	$1*1.287$	$0*1.693$	$0*1.693$	$0*1.693$
S-3	$1*1$	$0*1.693$	$1*1.287$	$1*1.693$	$0*1.693$	$1*1.693$

Finally, we convert every row vector to a unit vector.

TF-IDF Matrix finally

From the previous slide,

	analytics	boring	interesting	sports	text	want
S-1	1*1	3*1.693	0*1.287	0*1.693	1*1.693	0*1.693
S-2	1*1	0*1.693	1*1.287	0*1.693	0*1.693	0*1.693
S-3	1*1	0*1.693	1*1.287	1*1.693	0*1.693	1*1.693

After normalization of each row,

TF-IDF Matrix						
	analytics	boring	interesting	sports	text	want
S-1	0.1836	0.9326	0	0	0.3109	0
S-2	0.6134	0	0.7898	0	0	0
S-3	0.3452	0	0.4445	0.5845	0	0.5845

Sentiment Analysis - VADER

<https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399>

Cosine Similarity

$$\vec{a}, \vec{b}$$

$$\cos \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

How do we measure similarity of two documents?

S1 = "Phone is good, phone is good"

ph	good	not
2	2	0
1	1	1
1	1	0

S2 = "Phone is not good"

S3 = "It is a good phone"

(Removing the stopwords except "not")

Euclidean distance from DTM would suggest that distance of S3 and S1 is $\sqrt{2}$, and distance of S3 and S2 is 1.

Cosine Similarity

Thank you for your attention

Brief history

- Inspired by 1940s understanding of neurons in animal brains (McCullough-Pitt)
- Basic training algorithm (backpropagation) discovered in 1986 (Rumelhart, Hinton, Williams).
- Convolutional nets + modern training from late 1980s.

[Published: 09 October 1986](#)

Learning representations by back-propagating errors

[David E. Rumelhart](#), [Geoffrey E. Hinton](#) & [Ronald J. Williams](#)

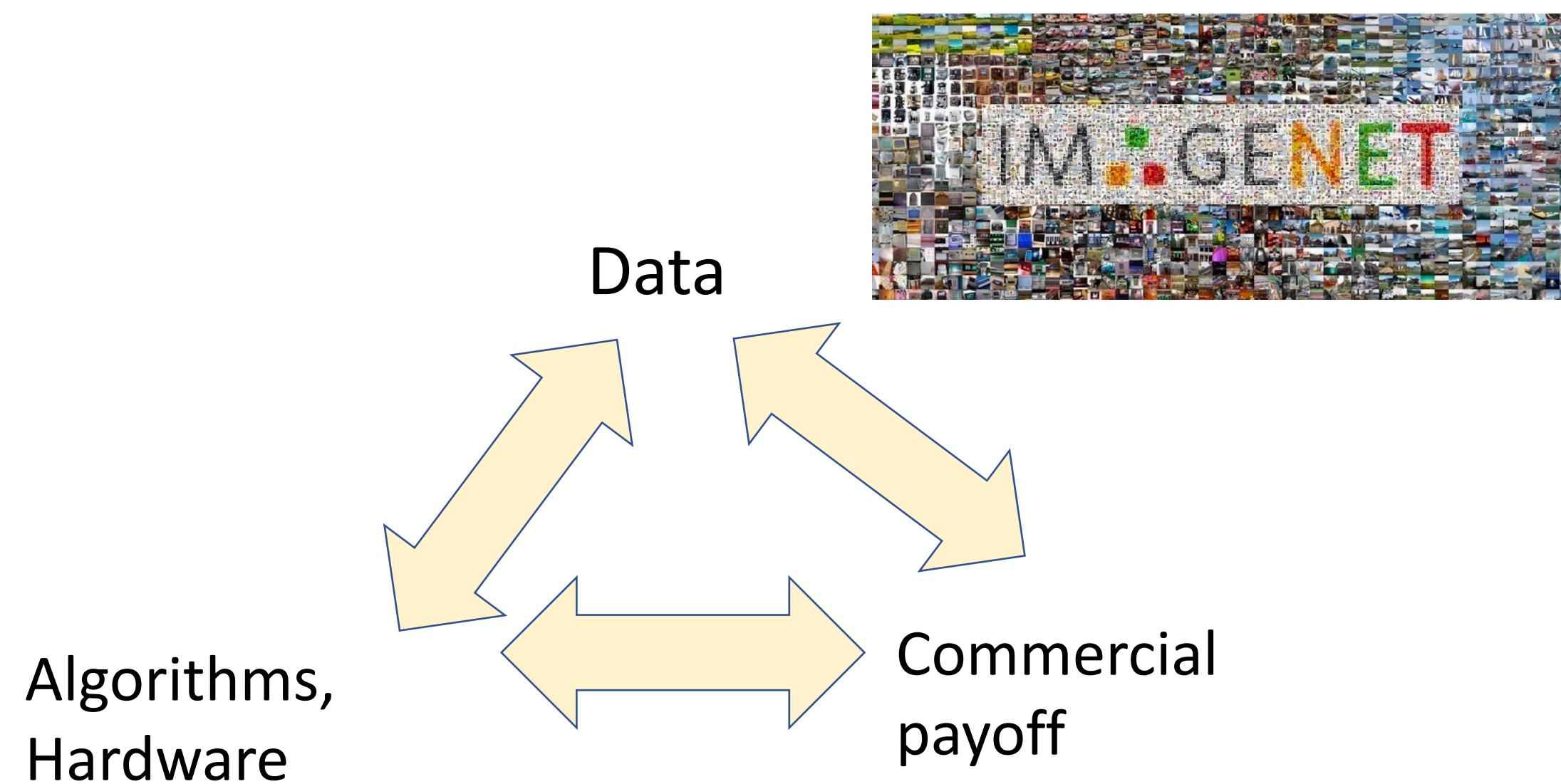
[Nature](#) **323**, 533–536 (1986) | [Cite this article](#)

75k Accesses | **11068** Citations | **238** Altmetric | [Metrics](#)

Brief history

- Out of fashion for a decade; came back strong with around 2012. Now dominant in AI (computer vision, NLP, robotics, etc).

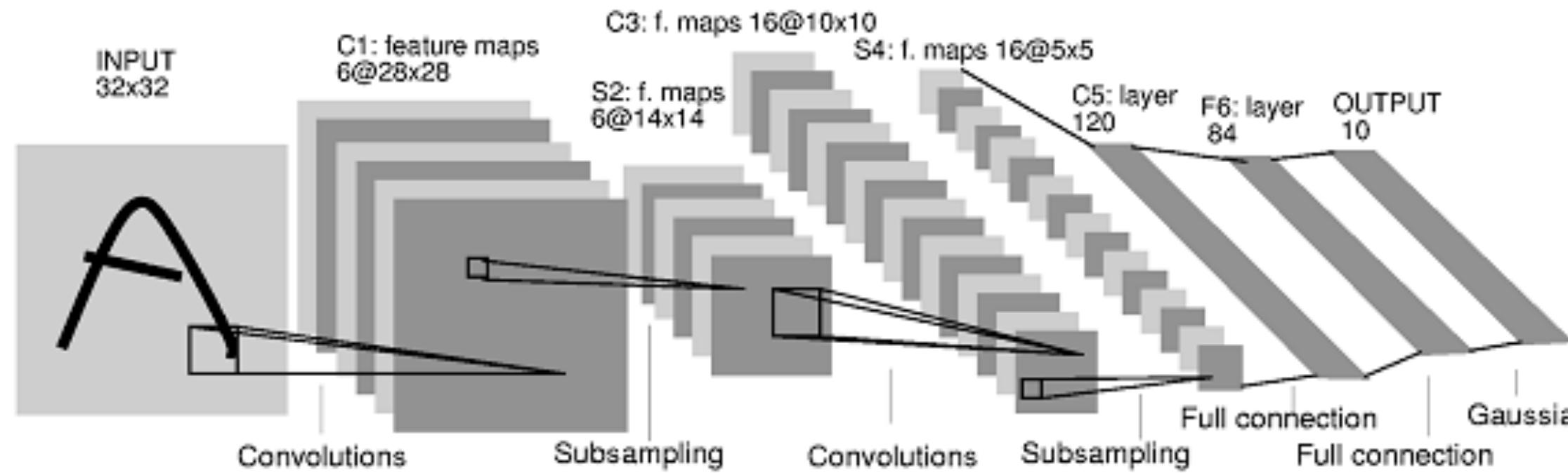
Why popular now?



hardware

1998

LeCun et al.



of transistors



10^6

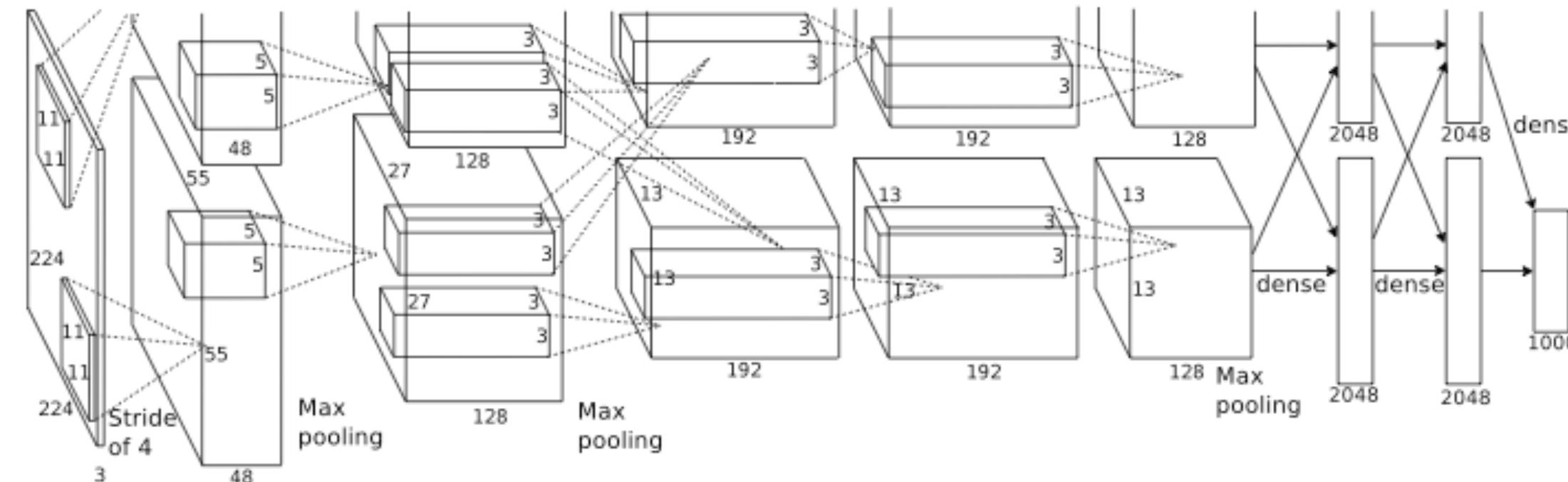
of pixels
used in training

10^7

NIST

2012

Krizhevsky
et al.



of transistors

10^9

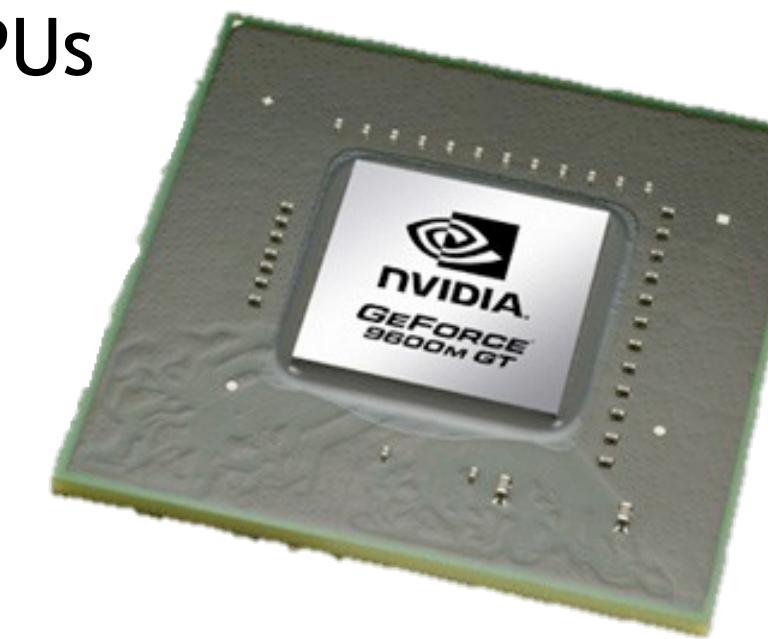


of pixels
used in training

IMAGENET
 10^{14}

2022

GPUs have 8×10^{10} transistors.



Commercial payoff

- Works well enough to monetize, across a range of applications
- Open-source tools enable democratization of access



Tensorflow

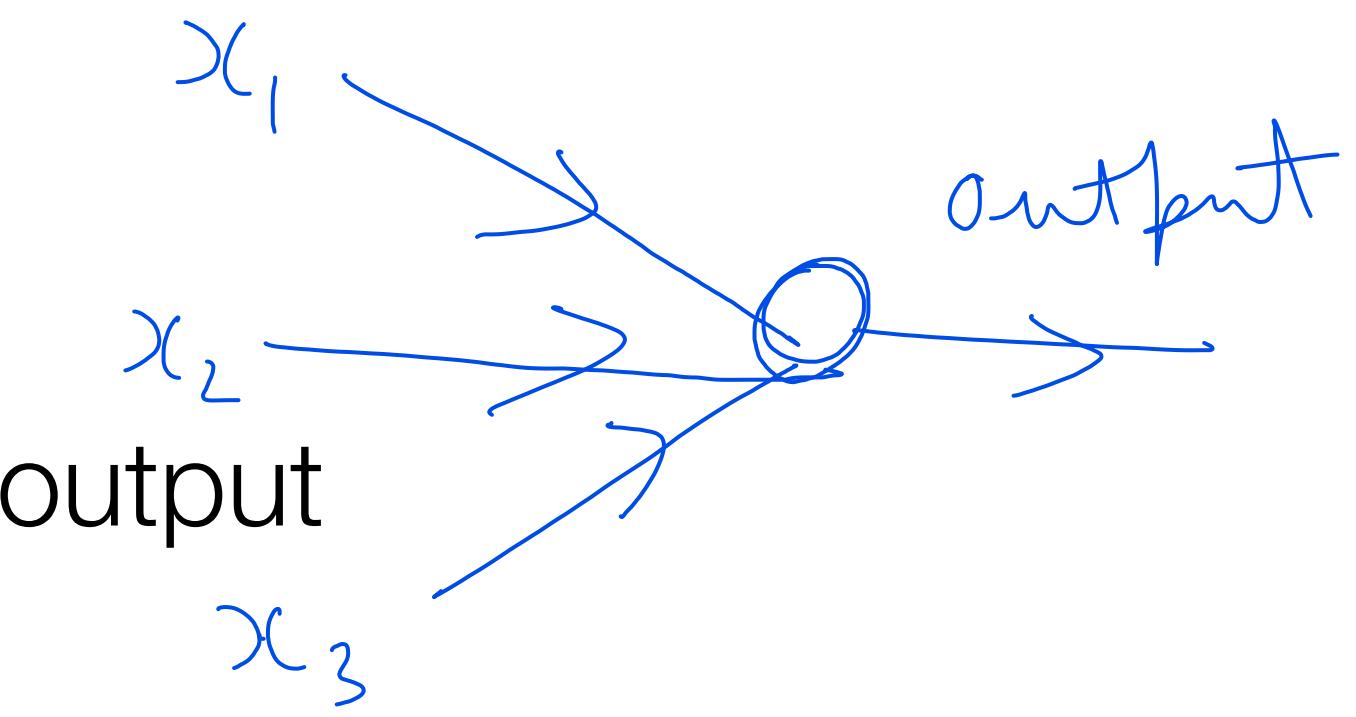


Pytorch

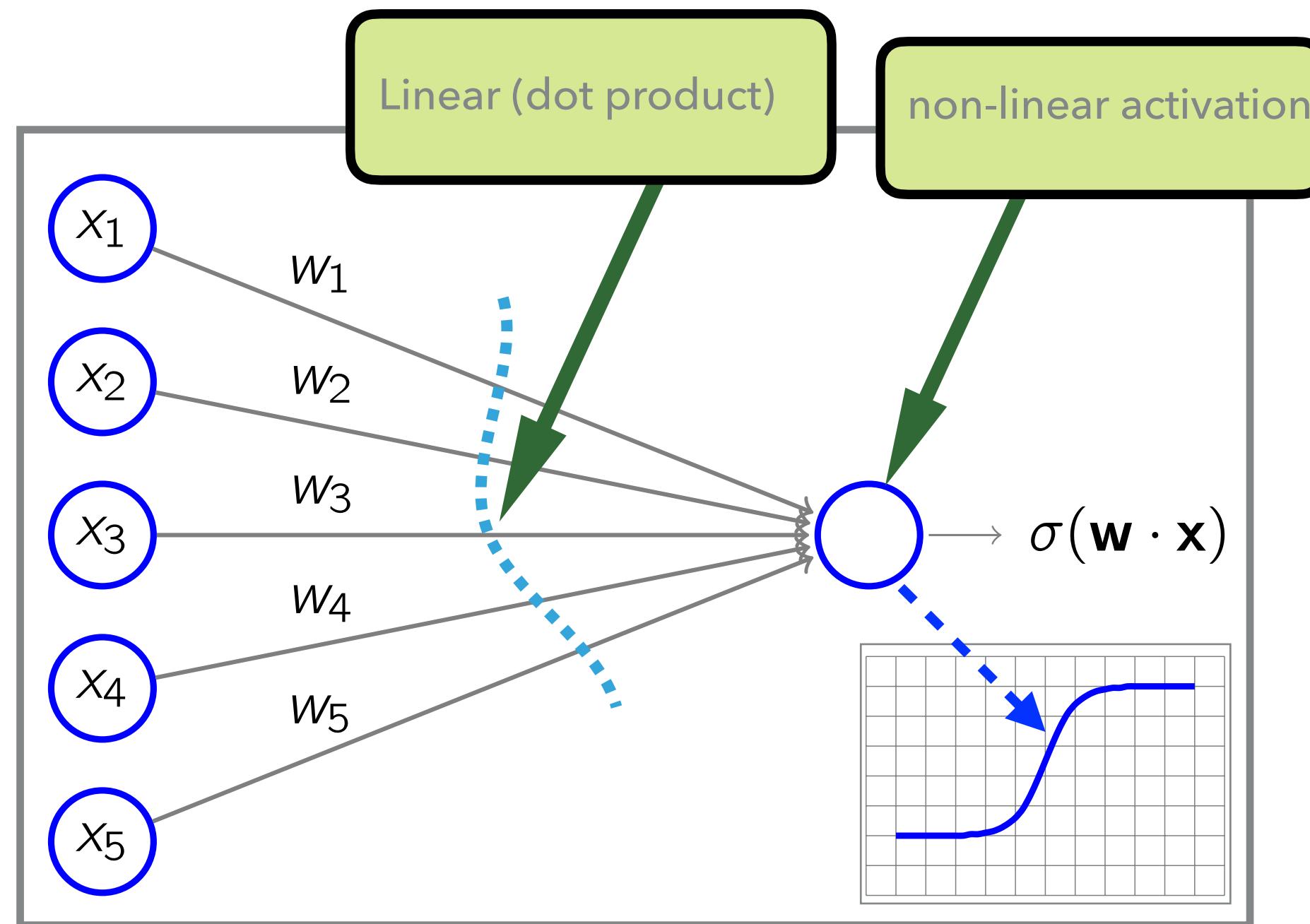
Neural Network Basics

An artificial neuron

- A neuron is a computational unit that has scalar inputs and an output
- Each input has an associated weight.
- The neuron multiples each input by its weight, sums them, applied a **nonlinear activation function** to the result, and passes it to its output.



$$\boxed{w_1x_1 + w_2x_2 + w_3x_3}$$



$$\sigma(z) = 1/(1 + e^{-z})$$

Input: x_1, x_2, x_3, x_4, x_5

weights: w_1, w_2, w_3, w_4, w_5

output:

$$y = \sigma\left(\sum_{j=1}^5 w_j x_j\right) \text{ or } y = \sigma(\mathbf{w} \cdot \mathbf{x})$$

A neuron can be a logistic regression unit!

Popular Activation functions

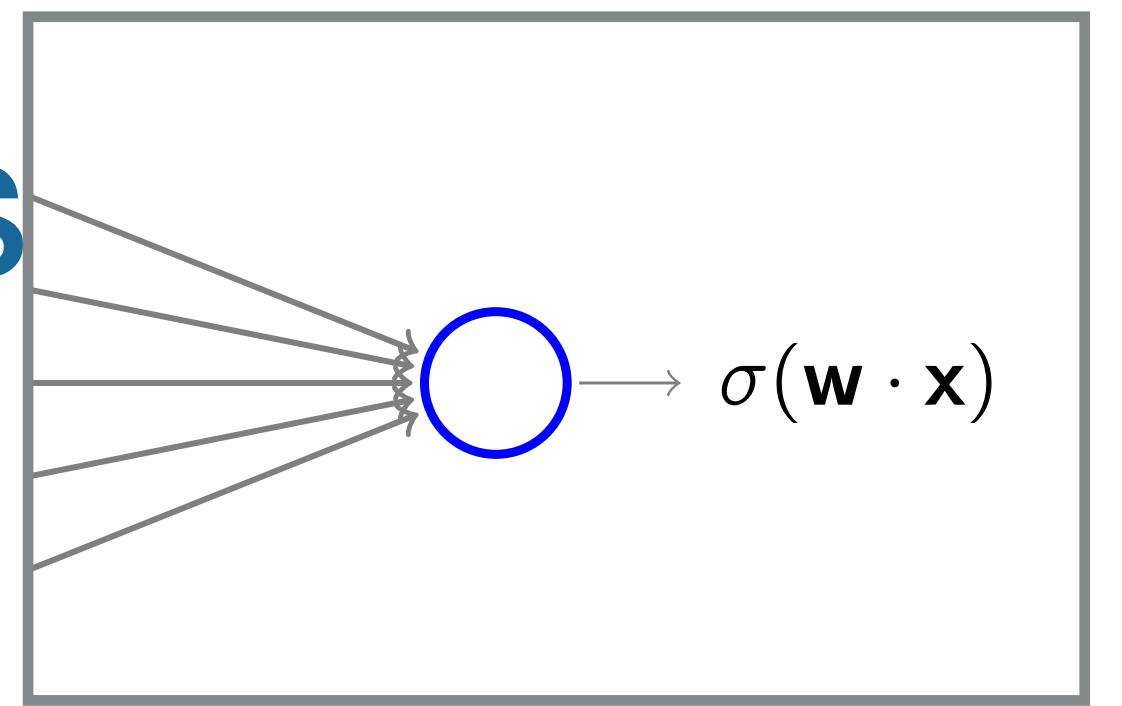
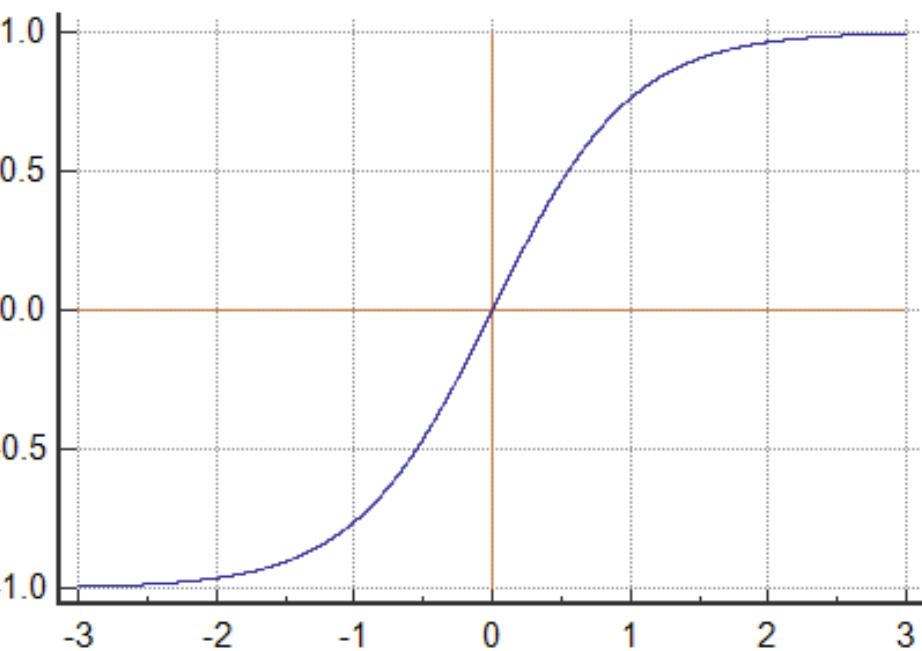
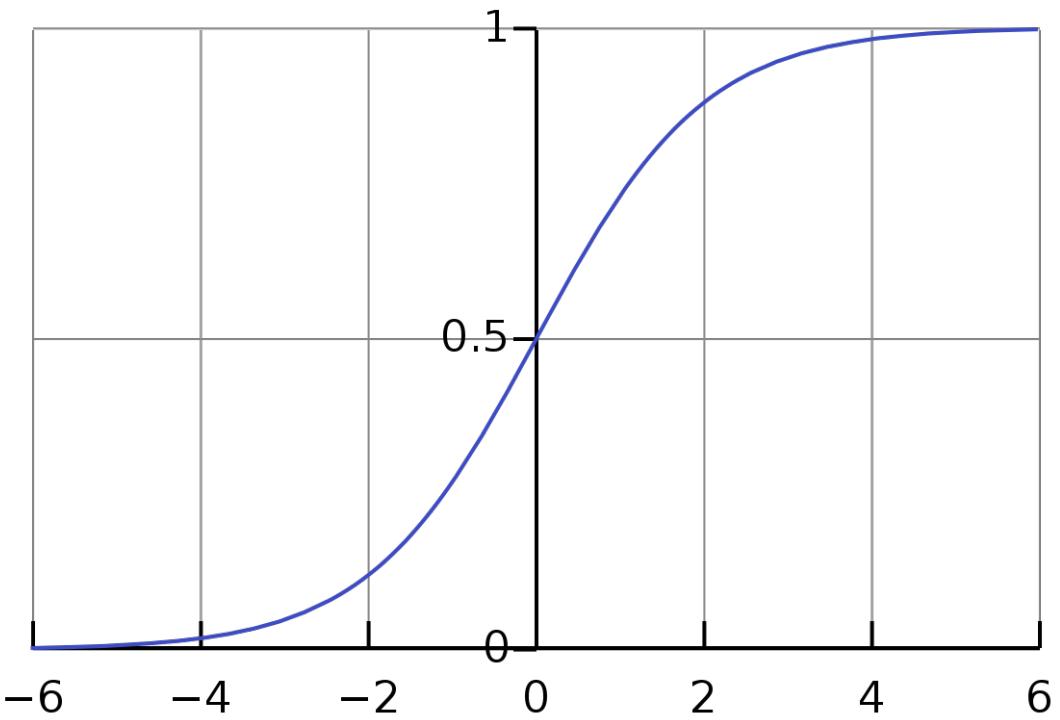
sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

tanh:

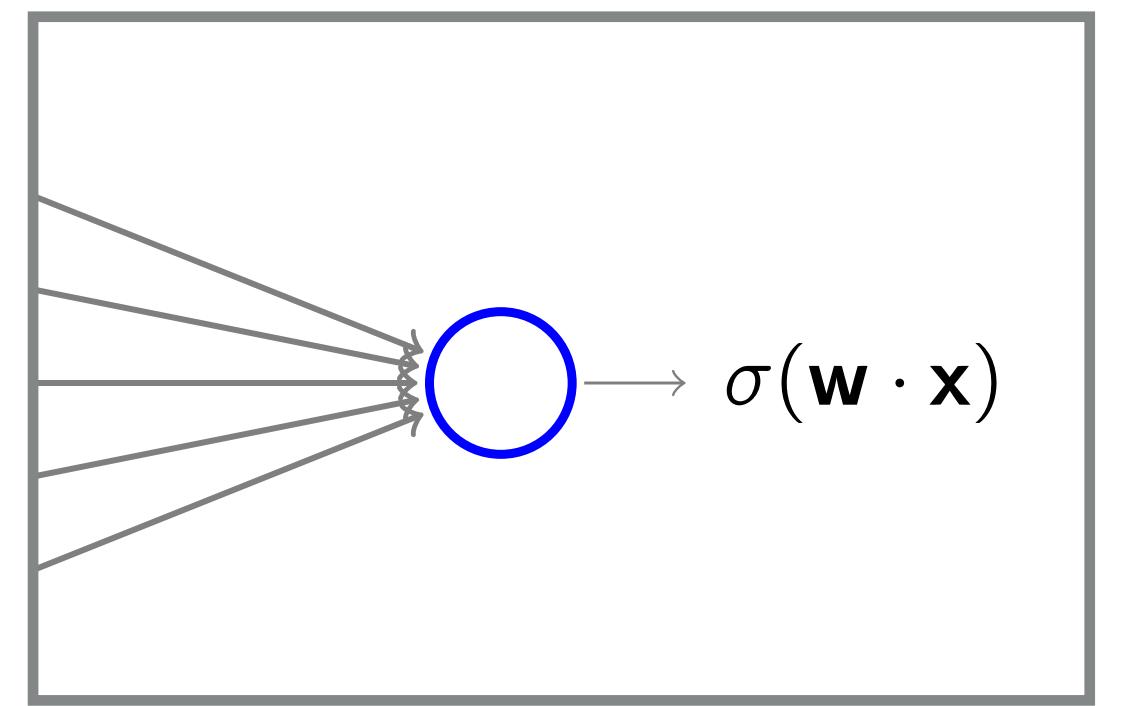
$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$(\tanh(z) = 2\sigma(2z) - 1)$$



Activation function: ReLU

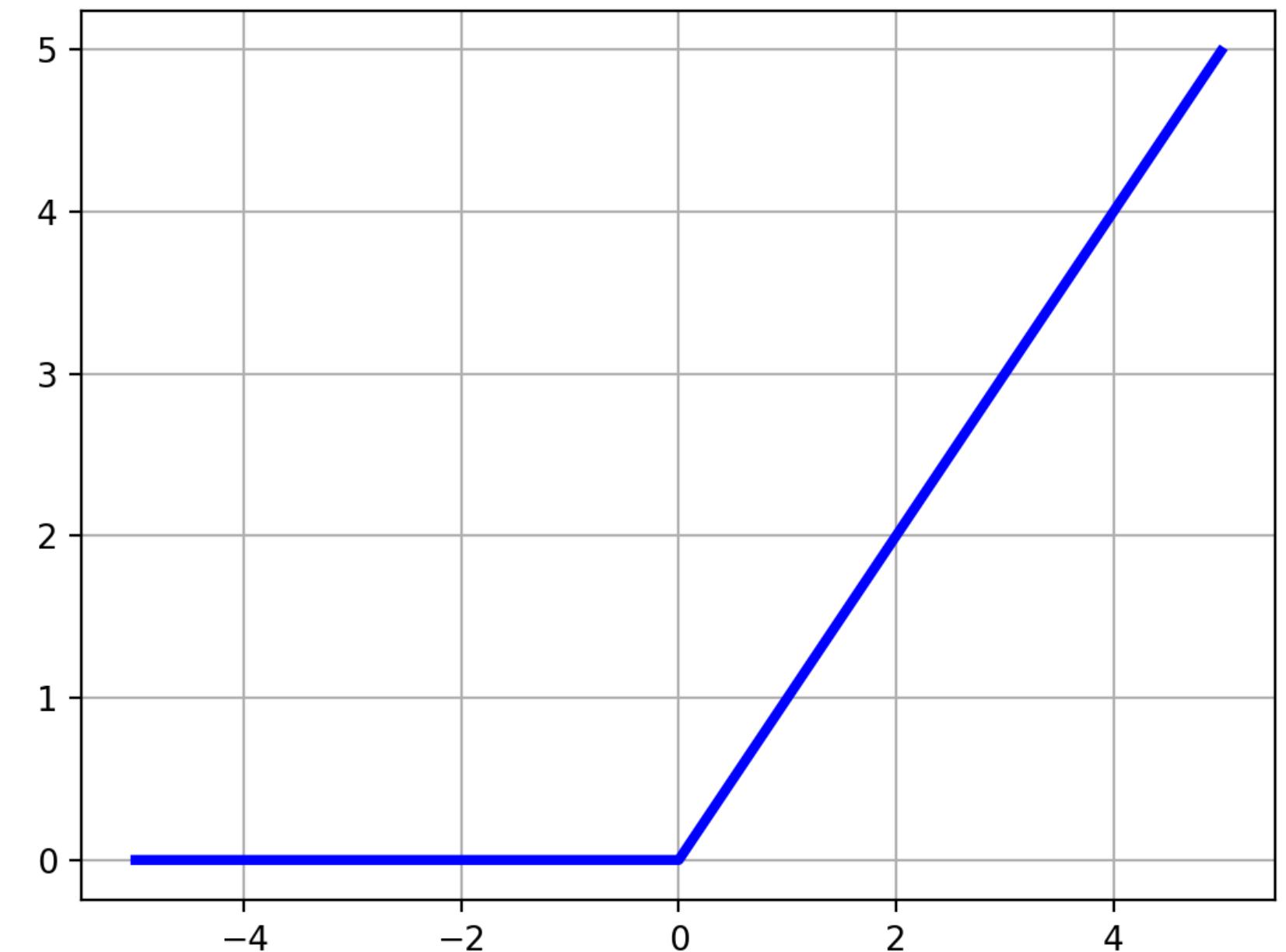
A popular choice: **R**ectified **L**inear **U**nit (ReLU)



$$\text{ReLU}(z) = \max\{z, 0\} = \begin{cases} 0 & z \leq 0 \\ z & z > 0 \end{cases}$$

Shorthand: $\text{ReLU}(z) = [z]_+$

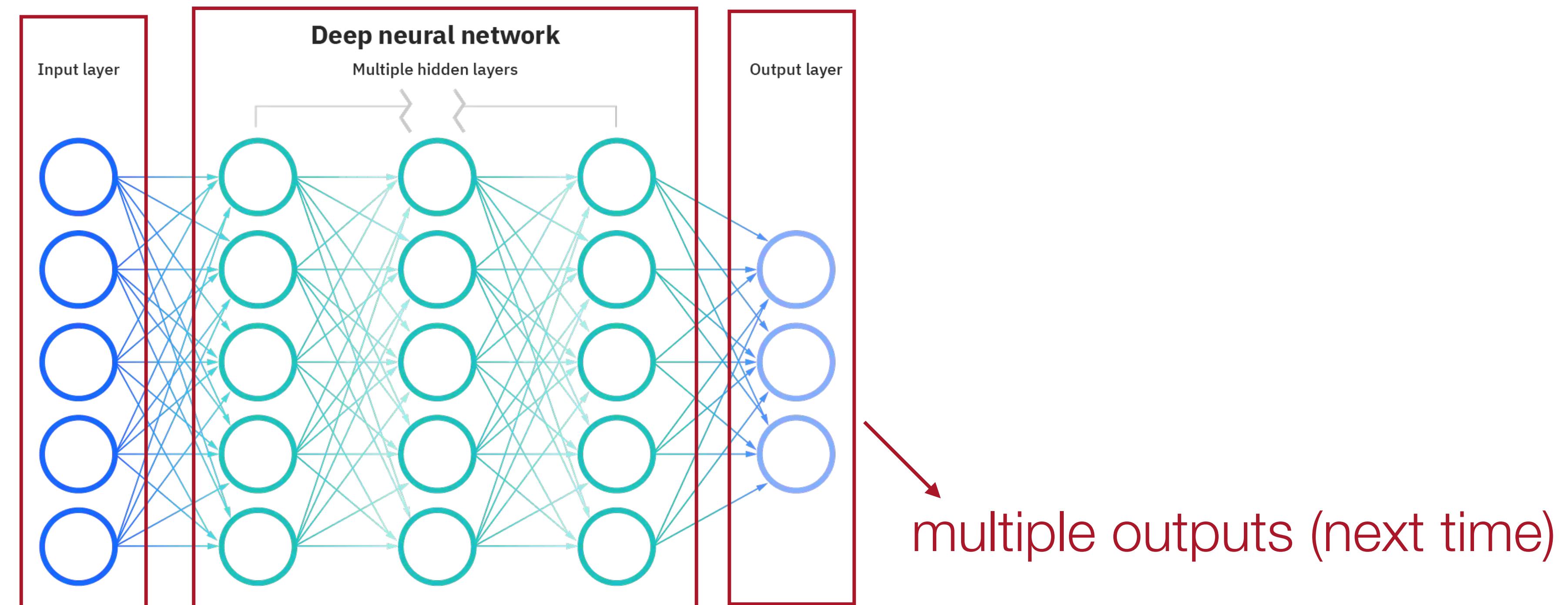
Q: What is the advantage of ReLU?



- 1) avoid gradient vanishing; 2) cheaper to compute

Neural networks (high level)

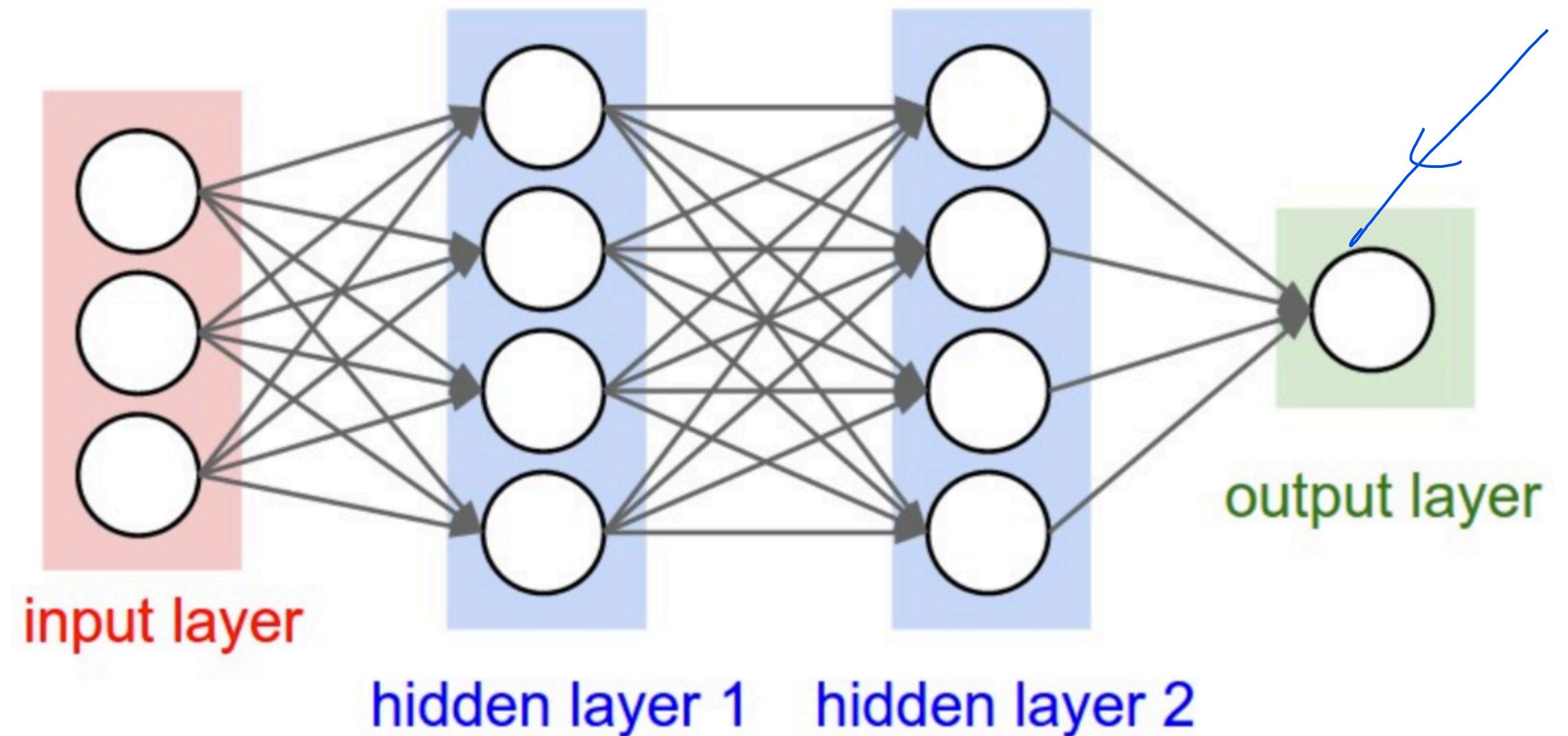
- The artificial neurons are connected to each other, forming a network
- The output of a neuron may feed into the inputs of other neurons



The intermediate layers are called hidden layers

Feedforward neural networks

- A feedforward network is a multilayer feedforward network:
- The units are connected with **no cycles**
- The outputs from units in each layer are passed to units in the next higher layer
- No outputs are passed back to lower layers



Fully-connected layers:

All the units from one layer are fully connected to every unit of the next layer.

Note: each edge has an associated weight