Math 4570 Matrix Methods for DA and ML

Section 10. Probability Review

1. Probability functions
2. Random Variables
3. Probability density functions
4. Expected values and variance
5. Classical distributions

Instructor: He Wang
Department of Mathematics
Northeastern University

➢ Terminologies:

- **Experiment:** A repeatable procedure with a set of possible results.
- **Sample Space S={all possible outcomes of an experiment}** Set
- **Event:** A subset of S. $A \subseteq S$

Example 1. Experiment: Flipping a Coin once.

S={Face, Tail}

Example 2. Experiment: Rolling a 6-sided die once.

S={1, 2, 3, 4, 5, 6}

$$\mathbb{R} \xrightarrow{f} \mathbb{R}$$

$$a \to f(a)$$

➢ The Probability Function

Definition (*1930s, Kolmogorov*)

Let S be a finite sample space. A **probability function** P is a function

{all subsets of S}

$$P: S \to [0,1]$$

satisfying the following two axioms:

1.) $\boxed{P(S) = 1}$

2.) If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

If S is an infinite set, we need one more axiom:

3.) $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \quad \text{if } A_i \cap A_j = \emptyset \text{ for any } i \neq j$$

Classical definition of probability (a special case for probability function)

Suppose the outcomes of an experiment are <u>all equally likely</u>, and the total number of all possible outcomes is <u>finite.</u> #(S)

$$\textit{Probability of an event} := \frac{\textit{Number of ways it can happen}}{\textit{Total number of all possible outcomes}} \qquad \frac{\#(A)}{\#(S)}$$

A

That is, $$P(A) := \frac{|A|}{|S|}$$

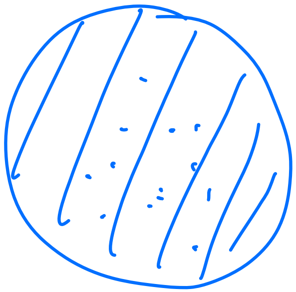
Example 1. Flipping a **fair** coin **2 times**

S={FF, FT, TF, TT} $\longrightarrow$ [0,1]

$\frac{1}{4}$

"fair"

Example. Rolling a 6-sided die once.



Example. flip an **unfair** coin once.

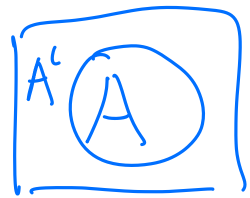
$$S=\{F, T\}$$

$$F \xrightarrow{p} p \in [0,1]$$

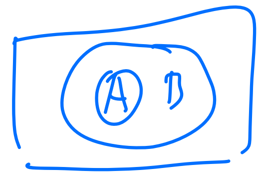
$$T \longrightarrow 1-p \in$$

Some properties

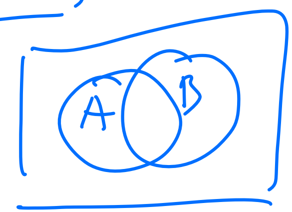
1. $P(A^C) = 1 - P(A)$.



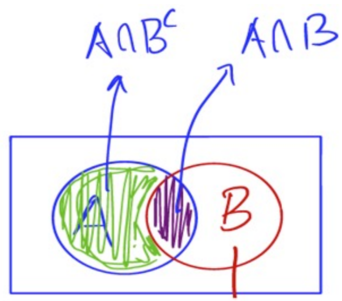
2. If $A \subset B$ then $P(A) \leq P(B)$.



3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.



4. $P(A) = P(A \cap B^C) + P(A \cap B)$.



➢ Conditional Probability

Definition. Probability that event A occurs **given** that event B already occurs, denoted by **P(A|B)** is a conditional probability, defined by

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

$$\boxed{P(A \cap B) = P(A|B)\,P(B)}$$

Example. Rolling a fair 6-sided die once.

$S = \{1, 2, 3, 4, 5, 6\}$

$\{2, 4, 6\} = B$

Given that the result is an even number, what is the probability that the result is 6?

$A = \{6\}$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$

- **Impendence**

$$\boxed{P(A\cap B) = P(A|B)P(B)}$$

Definition: The events A and B are called **independent** if

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

If A and B are not empty set, A and B are **independent** if and only if

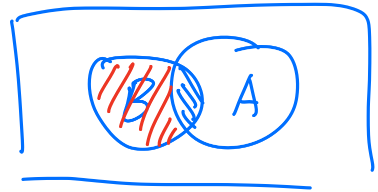
$$P(A|B) = P(A) \text{ if and only if } P(B|A) = P(B)$$

Example. Rolling a 6-sided die twice. $\{(a,b)\}$ $\begin{matrix} a \in \\ b \in \end{matrix} \{1, \ldots 6\}$

A: the first face is even number
B: the second face is 6.

$$P(A\cap B) = P(A)P(B) = \frac{3}{6}\cdot\frac{1}{6}$$

Theorem 1. Law of Total Probability



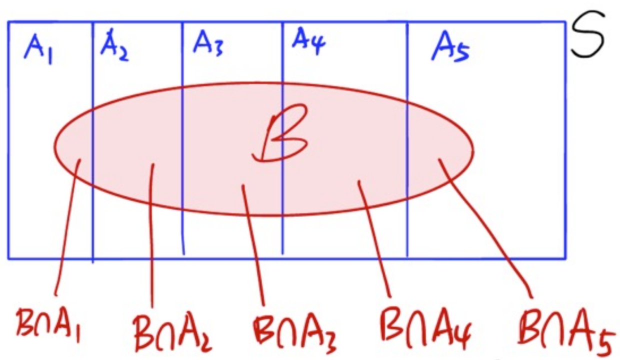
$S = A \cup A^c$

$P(B) = P(B \cap A) + P(B \cap A^c)$

Theorem. Law of Total Probability

Let $A_1, A_2, \ldots, A_n$ be a sequence of events such that $S = \bigcup_{i=1}^{n} A_i$ and $A_i \cap A_j = \emptyset$. Then, for any event B,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$



$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4) + P(B \cap A_5) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4) + P(B|A_5)P(A_5) \end{aligned}$$

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)}$$

Theorem 2. Bayes' Theorem

Theorem. Bayes' Theorem

Let $A_1, A_2, \ldots, A_n$ be a sequence of events such that $S = \bigcup_{i=1}^{n} A_i$ and $A_i \cap A_j = \emptyset$. Then, for any event B,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

for any $j = 1, \cdots, n$.

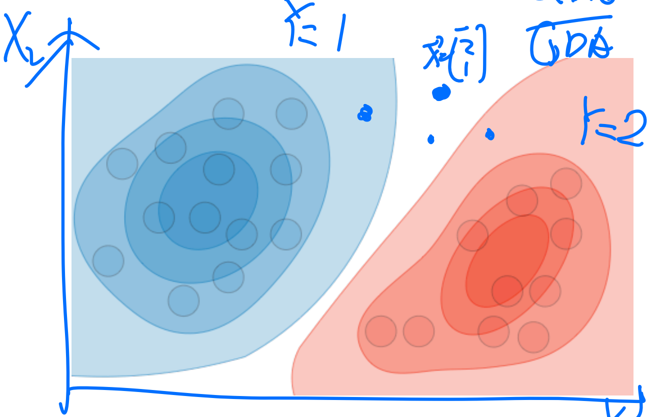
Example: in classification problem.

$$P(Y=1|\vec{X}) = \frac{P(\vec{X}|Y=1)\,P(Y=1)}{P(\vec{X}|Y=1)P(Y=1) + P(\vec{X}|Y=2)P(Y=2)}$$

LDA

QDA

GDA



➢ Random Variables

want (hard)

$P: S \to [0, 1] \subseteq \mathbb{R}$

Let S be a sample space.

A **random variable** is a function

$$X: S \to \mathbb{R}$$

want (easy)

Example 1. Flipping an **unfair** coin once.

S={Face, Tail} $\longrightarrow \mathbb{R}$

Tail $\to 0$, Face $\to 1$

Probability density(mass) function

Definition.

For every discrete random variable X, we define a **probability density function (pdf)** by

$$p_X(k) = P(X = k) := P(\{s \in S | X(s) = k\})$$

(If $k \notin X(S)$, then $p_X(k) = 0.$)

$$\begin{tikzcd}
\{F, T\} \arrow[rr] \arrow[d, phantom, "\parallel"] & & \{1, 0\} \arrow[d, phantom, "\cap" rotate=90] \\
S \arrow[rr, mapsto, "\text{random variable } X"] \arrow[dr, "\text{Probability function } P(\ )"'] & & \mathbb{R}^n \arrow[dl, "\textbf{pdf } p_X(\ )"] \\
& {[0,1]} &
\end{tikzcd}$$

Example 1. Flipping an un**fair** coin once.

X

$S=\{H, T\} \to \{1, 0\} \to p,\ 1-p$

$P(\ \ )$

$$\begin{cases} P(X=1)=p \\ P(X=0)=1-p \end{cases}$$

$$P_X(k)=\begin{cases} p & \text{if } k=1 \\ 1-p & \text{if } k=0 \end{cases} = p^k(1-p)^{1-k} \qquad k\in\{0,1\}$$

Random Variable X ~ **Bernouli** ($\overset{p}{\phi}$)

Pdf function $p_X(k) = \boxed{\phi^k(1-\phi)^{1-k}} = \begin{cases} \phi & if\ k=1 \\ 1-\phi & if\ k=0 \end{cases}$

Example. Rolling an unfair 6-sided die once.

$$S \xrightarrow{X} \mathbb{R} \qquad 1, 2, 3, 4, 5, 6$$

$$P_X(k) = \begin{cases} \phi_1 & \text{if } k=1 \\ \vdots & \vdots \\ \phi_6 & \text{if } k=6 \end{cases}$$

$$\phi_1 + \dots + \phi_6 = 1$$

Assume $Y \sim$ $\boxed{\textbf{Categorical}(\phi_1, \dots, \phi_K)}$ such that $\phi_1 + \cdots + \phi_K = 1$

indicator function

Pdf function $p_Y(y) = \boxed{\phi_1^{\mathbb{I}(y=1)} \phi_2^{\mathbb{I}(y=2)} \dots \phi_K^{\mathbb{I}(y=K)}} = \begin{cases} \phi_1 & if\ y = 1 \\ \phi_2 & if\ y = 2 \\ \vdots & \vdots \\ \phi_K & if\ y = K \end{cases}$

$$\mathbb{1}(\text{true}) = 1$$

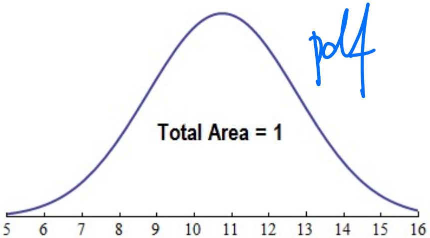
$$\mathbb{1}(\text{false}) = 0$$

Continuous Random variables:

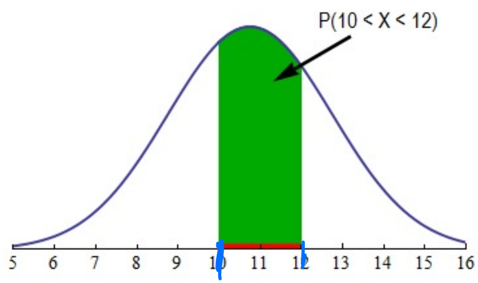
Definition.

The **probability density function (pdf)** of a **continuous** random variable X is a piece-wise continuous function $f_X(x)$ satisfying

1. $f_X(x) \geq 0$
2. $\displaystyle\int_{-\infty}^{\infty} f_X(x)dx = 1.$



pdf



Definition.

The **probability** that X is in an interval $[a, b]$ is

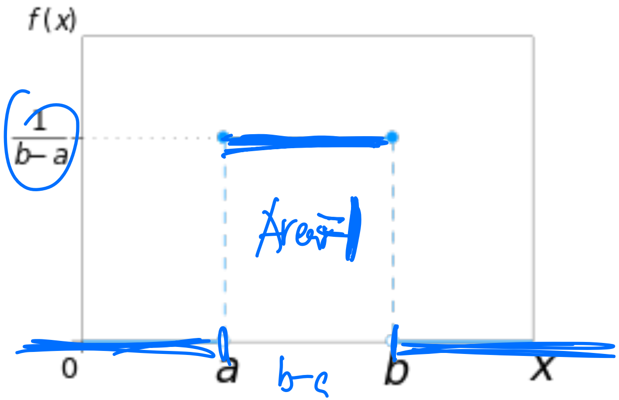
$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Example: **Uniform Distribution.**

The uniform distribution describes an experiment that choose a number randomly from the interval [a, b].

The *probability density function* of the uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$



➢ Expected Value

Expected value is a generalization of the concept "average".

Definition: Expected Value of **discrete** random variable

If X is a discrete random variable with probability function $p_X(k)$, then the **expected value** (or Mean) of X is

$$E(X) = \sum_{\text{all } k} k \cdot p_X(k).$$

Definition: (Expected Value of **continuous** random variable)

If X is a continuous random variable with probability function $p_X(k)$, then the **expected value** (or Mean) of X is

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

Property: $E(aX + b) = aE(X) + b$

Variance and Standard deviation

Definition. (Variance)

The **variance** of a random variable X is

$$\mathrm{Var}(X) := E\big((X-\mu)^2\big)$$

Here $\mu = E(X)$ is the mean of X.
The **standard deviation** is $\sigma := \sqrt{\mathrm{Var}(X)}$

Variance is expected squared distance from the mean.

It measures the spread of the data.

Calculation formula: $\mathrm{Var}(X) = E(X^2) - \big(E(X)\big)^2$

Property: $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$

Standard Normal Distribution

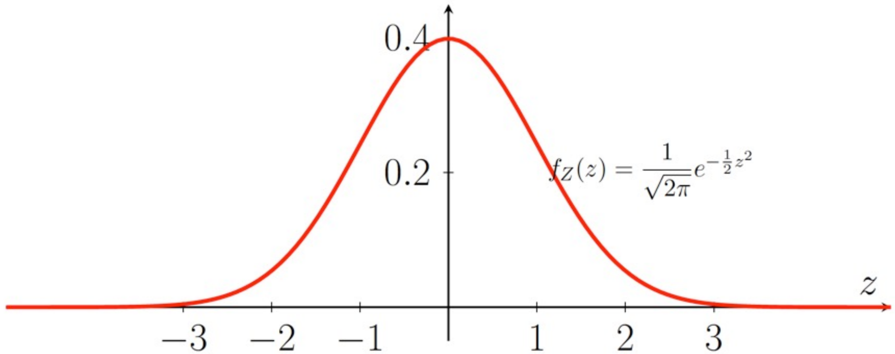
Definition.

The **standard normal distribution** is a continuous **pdf** defined by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

for $-\infty < z < \infty$.

The graph is **Gaussian** curve (bell curve).

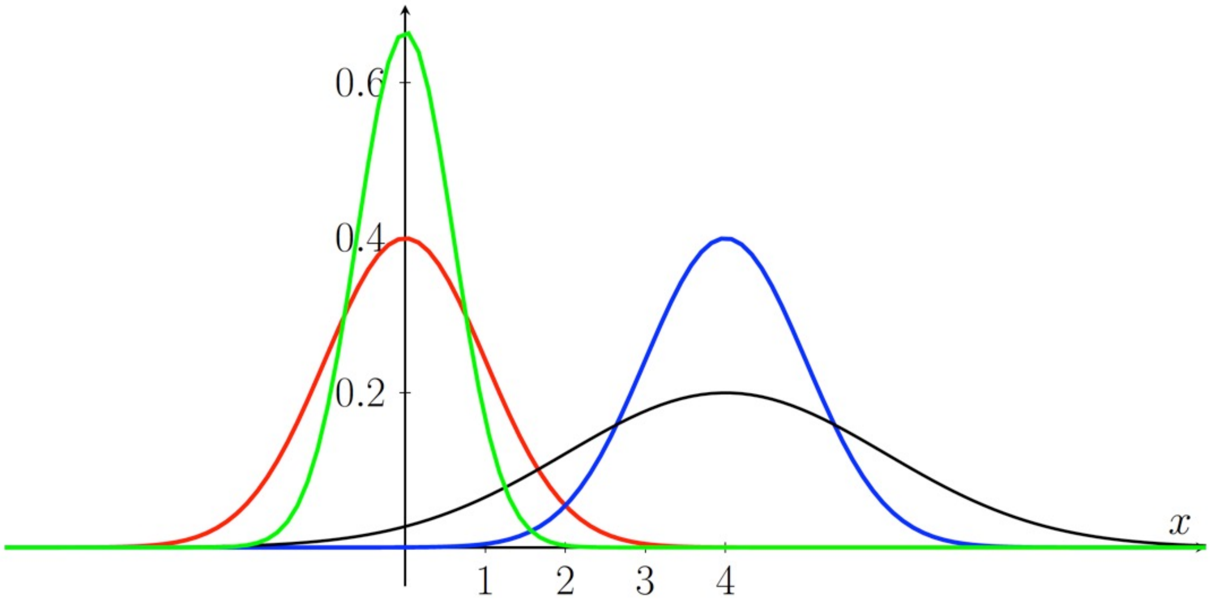


Normal distributions $X = \sigma Z + \mu$:

Definition. (Normal Distribution)

The **Normal Distribution** is a continuous **pdf** function defined as

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty.$$



Red: $\mu = 0, \sigma = 1$. Green: $\mu = 0, \sigma = 0.6$. Blue: $\mu = 4, \sigma = 1$. Black: $\mu = 4, \sigma = 2$.

Covariance and independence

Suppose X and Y are **any** random variables on the same sample space.

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X, Y)$

$Cov(X, Y)$ is the **covariance** of X and Y defined as

$$Cov(X, Y) := E(XY) - E(X)E(Y)$$

If X and Y are independent, then

- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$
- $Cov(X, Y) = 0$
- $E(XY) = E(X)E(Y)$

The converse is not true.

Joint distribution (multi random variables):

Definition. Discrete Joint Density

Let S is a **discrete** sample space. Let X and Y be two random variables on S. The **joint probability density function** (**joint pdf**) of X and Y is denoted by $p_{X,Y}(x,y)$ defined as

$$p_{X,Y}(x,y) := P(X = x, Y = y).$$

Here, $P(X = x, Y = y)$ is the probability when $X = x$ **and** $Y = y$.

Definition.

If X and Y are **continuous** random variables. the **joint pdf** $f_{X,Y}(x,y)$ of X and Y is a piecewise continuous multi-variable function satisfying

(1.) $f_{X,Y}(x,y) \geq 0.$

(2.) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dxdy = 1.$

Multivariate normal distribution.

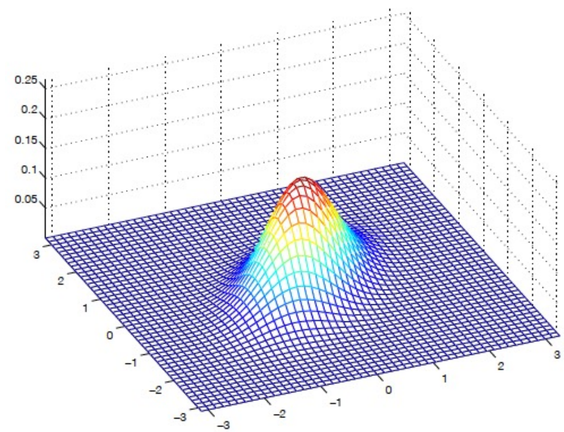
Vector random variable $\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix} \sim \mathrm{Normal}(\vec{\mu}, \Sigma)$

Here $\vec{\mu} \in \mathbb{R}^d$ and Σ is an $d \times d$ symmetric, positive definite matrix.

- The **joint** probability density function (**pdf**) for $\vec{X}$ is

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \right)$$

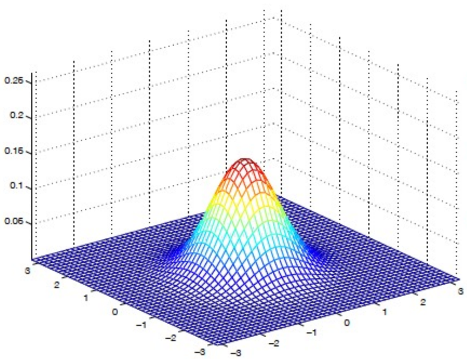
$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\vec{X}}(\vec{x}) \; d\vec{x} = 1$$



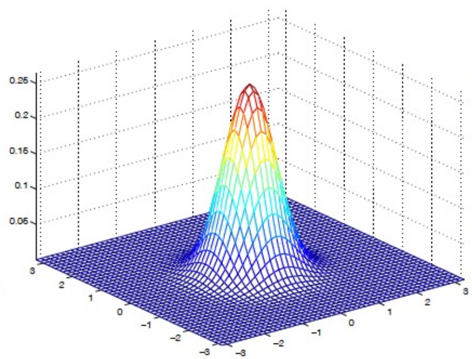
- The **mean vector** of $\vec{X}$ is $E(\vec{X}) = \vec{\mu}$

- The **(co)variance matrix** is Cov($\vec{X}$)= Σ

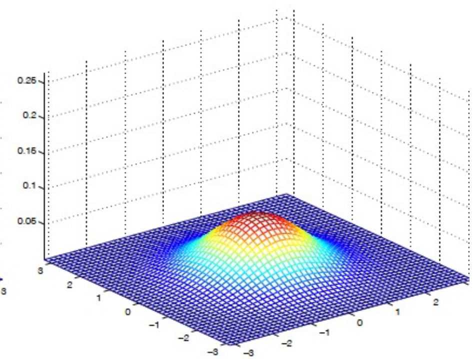
Standard normal

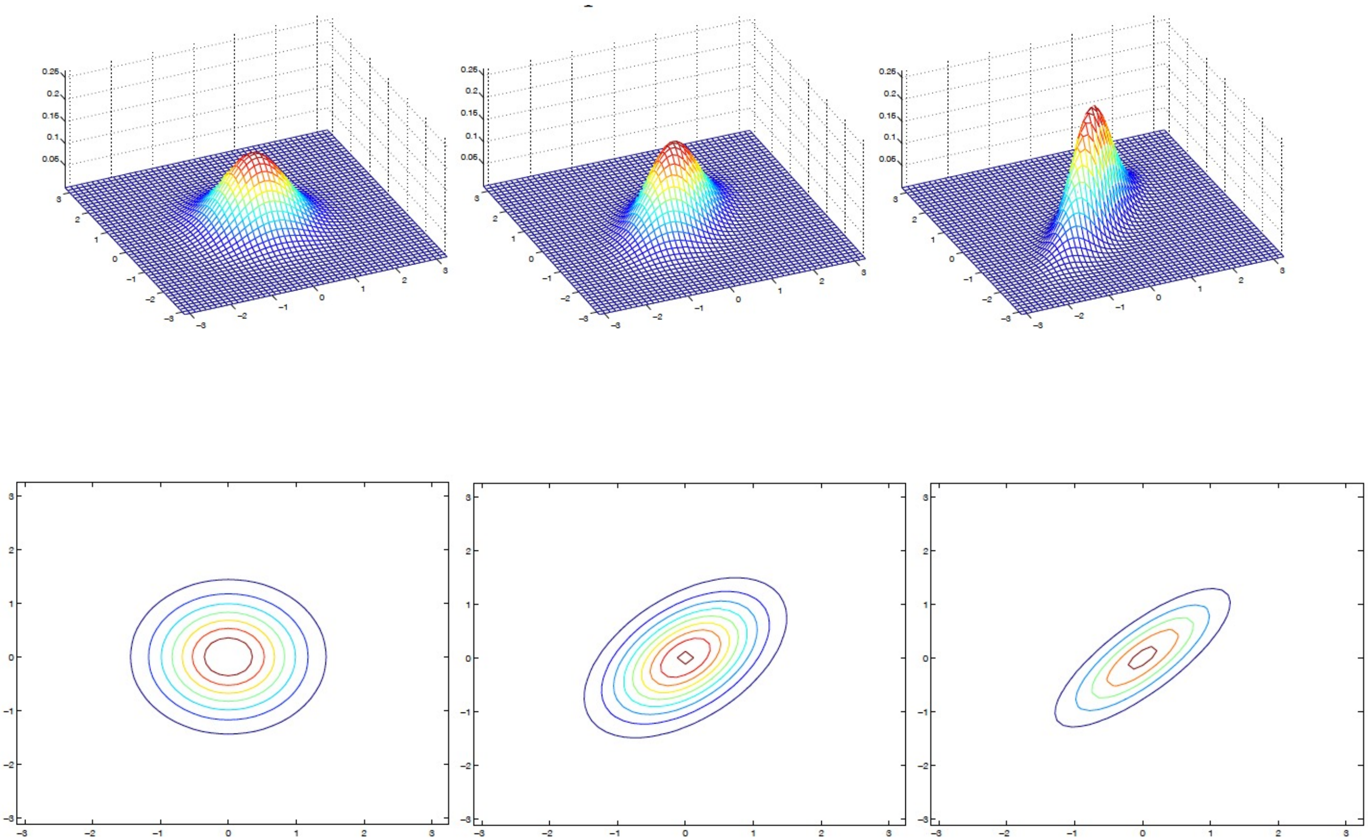


Compressed



Spread-out





0.25
0.2
0.15
0.1
0.05
−3 −2 −1 0 1 2 3

➢ More examples of distributions:

1. Binomial distribution is a generalization of Bernoulli distribution.

Given a series of n independent trials with two outcomes (T or F) with constant probability p and $1-p$.

Let X be the number of T appears in the n trials. Then $\mathrm{X} \backsim Binomial(n, p)$

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

For example, flip a coin n times.

2. Multinomial is a generalization of Categorical distribution.

Given a series of n independent trials with m outcomes $(O_1, \ldots O_m)$ with constant probability $(\phi_1, \ldots ,\ \phi_m)$.

Let $\vec{X}$ be the number of O_i appears in the n trials.

$$\text{Then } \vec{X} \backsim Multinomial(n, \phi_1, \ldots ,\ \phi_m)$$

$$P(X_i = n_i) = \frac{n!}{n_1! \cdots n_m!}\, \phi_1^{n_1} \cdots\ \phi_m^{n_m}$$

for each $i = 1, \ldots, m$, and each $n_1 + \cdots + n_m = n$

For example, Toss a K-side die n times.

3. Exponential random variable is a continuous random variable with pdf given by

$$f_X(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$

where λ is a fixed positive number.

- The **mean** and **variance** are given by

$$E(X) = \frac{1}{\lambda}$$

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}$$

- Exponential distribution model the time between occurrences in a time interval.

4. **Poisson Distribution**

Definition. (Poisson Distribution)

The **Poisson Distribution** Piosson(λ) is a discrete **pdf** function defined as

$$p_X(k) = P(X = k) := \frac{\lambda^k e^{-\lambda}}{k!}$$

for $k = 0, 1, 2, 3, \ldots$. Here, λ is a positive constant.

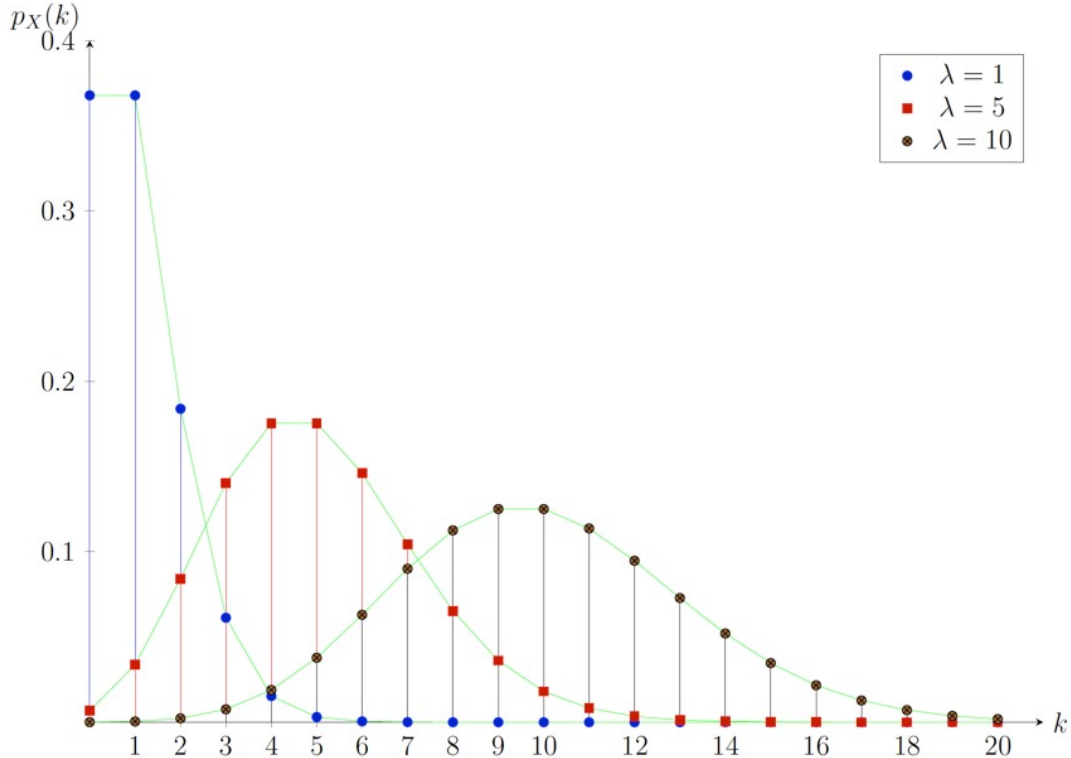
Theorem.

(1) It is a well defined **pdf** , i.e., $\sum_k p_X(k) = 1$
(2) The mean is $E(X) = \lambda$.
(3) The variance is $\mathrm{Var}(X) = \lambda$.

Applications:

1. Poisson approximation for binomial distribution

2. Poisson Model. The number of occurrences in a time interval with a given rate.



$p_X(k)$
0.4
0.3
0.2
0.1
$\lambda = 1$
$\lambda = 5$
$\lambda = 10$
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
k

More references:

1. My lecture notes for Math3081:

   https://web.northeastern.edu/he.wang/Teaching/Teaching3081/Math3081.html

2. Deep Learning by Goodfellow, Bengio, Courville (Chapter 2)

   https://www.deeplearningbook.org/

3. Pattern Recognition and Machine Learning, by Chris Bishop. (Chapters 1.2 and 2.1-2.3)

   https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

4. Review of Probability Theory
   https://cs229.stanford.edu/section/cs229-prob.pdf