

IIT KHARAGPUR

IIT GANDHINAGAR


NPTEL ONLINE
CERTIFICATION COURSES

Scalable Data Science

Lecture 2: Background Probability Theory

Sourangshu Bhattacharya
Computer Science and Engineering
IIT KHARAGPUR

Probability: Definition

- ***Experiment***: toss a coin twice
- ***Sample space***: possible outcomes of an experiment
  - S = {HH, HT, TH, TT}
- ***Event***: a subset of possible outcomes
  - A={HH}, B={HT, TH}
- ***Probability of an event*** : an number assigned to an event Pr(A)
  - Axiom 1: $\Pr(A) \geq 0$
  - Axiom 2: $\Pr(S) = 1$
  - Axiom 3: For every sequence of disjoint events

$$\Pr\left(\bigcup_i A_i\right) = \sum_i \Pr(A_i)$$

  - Example: Pr(A) = n(A)/N: frequentist statistics

Probability

- **Joint Probability**: For events *A* and *B*, joint probability *Pr(AB)* stands for the probability that both events happen.
- **Indepedence**: Two events *A* and *B* are independent in case

$$\Pr(AB) = \Pr(A)\Pr(B)$$

- A set of events $\{A_i\}$ are independent in case

$$\Pr\left(\bigcap_i A_i\right) = \prod_i \Pr(A_i)$$

- **Conditional Probability**: If A and B are events with Pr(A) > 0, the ***conditional probability of B given A*** is

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

Random Variable and Distributions

- A ***random variable X*** is a numerical outcome of a random experiment.
- **Discrete random variable**
  - Takes on one of a finite (or at least countable) number of different values.
  - Examples:
    - X = 1 if heads, 0 if tails
    - Y = 1 if male, 0 if female (phone survey)
    - Z = # of spots on face of thrown die
- **Distribution function or mass function**: For a discrete r.v. *X*, we have $\Pr(X = x)$ or $\Pr(x)$ or $P(x)$, i.e., the probability that r.v. X takes on a given value x.
- **Properties**: $\Pr(X = x) > 0$ and $\sum_x \Pr(x) = 1$.

Random Variable and Distributions

- **Continuous random variable**
  - Takes on one in an infinite range of different values
  - Examples:
    - W = % GDP grows (shrinks?) this year
    - V = hours until light bulb fails
- **Distribution function**:
  - What is the probability that a continuous r.v. takes on a specific value?
    e.g. *Prob(V = 3.14159265 hrs)* = **0**
  - However, ranges of values can have non-zero probability.
    e.g. *Prob(3 hrs <= V <= 4 hrs) = 0.1*
  - For a continuous r.v. *X*, we have $\Pr(x)$ or $P(x) = \Pr(x \leq X \leq x + dx)$.
- **Properties**: $\mathrm{P}(x) > 0$ and $\int_x P(x)dx = 1$.

Expectation

- A discrete random variable $X \sim P(X = x)$. Then, its expectation is:

$$E[X] = \sum_x x\, P(X = x)$$

- For an empirical sample, $x_1, x_2, \ldots, x_N$, expectation can be estimated as:

$$E[X] = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Continuous random variable: $E[X] = \int_x xP(x)dx$
- Expectation of sum of random variables: $E[X_1 + X_2] = E[X_1] + E[X_2]$.
- A measure of central tendency. Other measures: median, mode, etc.

Variance

- The variance of a random variable X is the expectation of $(X - E[x])2$:

$$
\begin{aligned}
Var(X) &= E((X - E[X])^2) \\
&= E(X^2 + E[X]^2 - 2XE[X]) \\
&= E(X^2 - E[X]^2) \\
&= E[X^2] - E[X]^2
\end{aligned}
$$

Bernoulli Distribution

- The outcome of an experiment can either be success (i.e., 1) and failure (i.e., 0).
- $\Pr(X = 1) = p, \Pr(X = 0) = 1 - p$, or

$$p_\theta(x) = p^x(1-p)^{1-x}$$

- $E[X] = p, Var(X) = p(1-p)$

Binomial Distribution

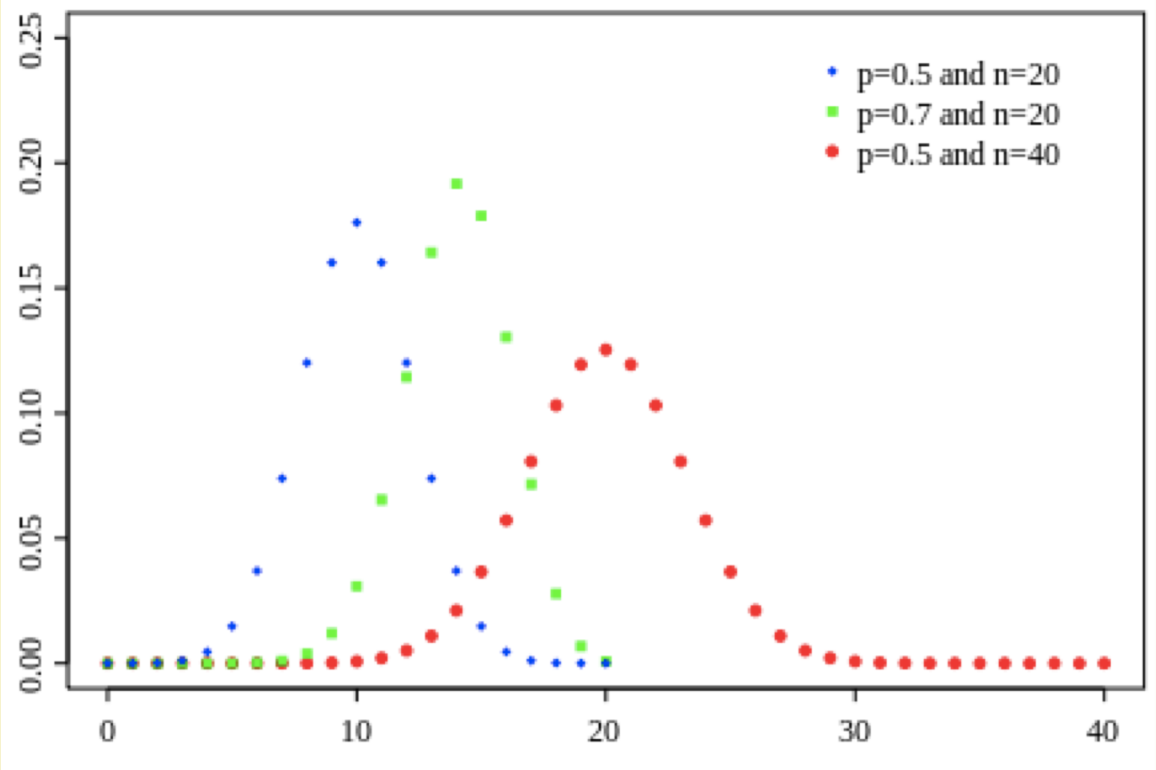
- n draws of a Bernoulli distribution
  $X_i \sim Bernoulli(p), X = \sum_{i=1}^{n} X_i, X \sim Bin(p, n)$

- Random variable X stands for the number of times that experiments are successful.

$$\Pr(X = x) = p_\theta(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

- $E[X] = np, Var(X) = np(1-p)$

Plots


p=0.5 and n=20
p=0.7 and n=20
p=0.5 and n=40
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40

Poisson Distribution

- Distribution of number of arrivals, given the average rate of arrival, λ.

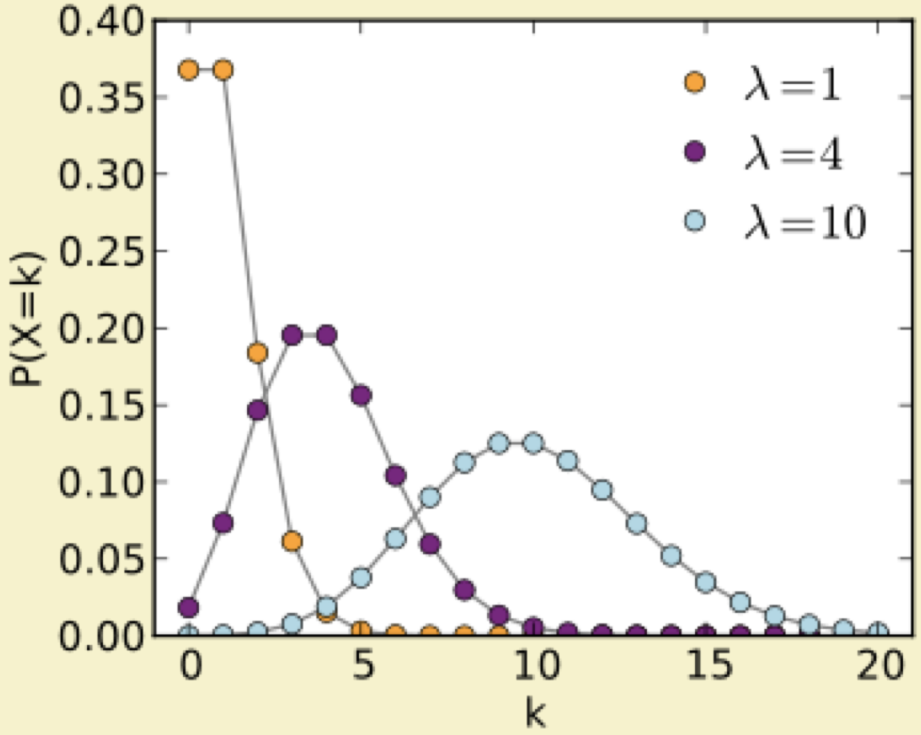
- Coming from Binomial distribution
  - Fix the expectation $\lambda = np$
  - Let the number of trials $n \to \infty$

  A Binomial distribution will become a Poisson distribution

$$\Pr(X = x) = p_\theta(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- $E[X] = \lambda$, $Var(X) = \lambda$

Plots



Normal (Gaussian) Distribution

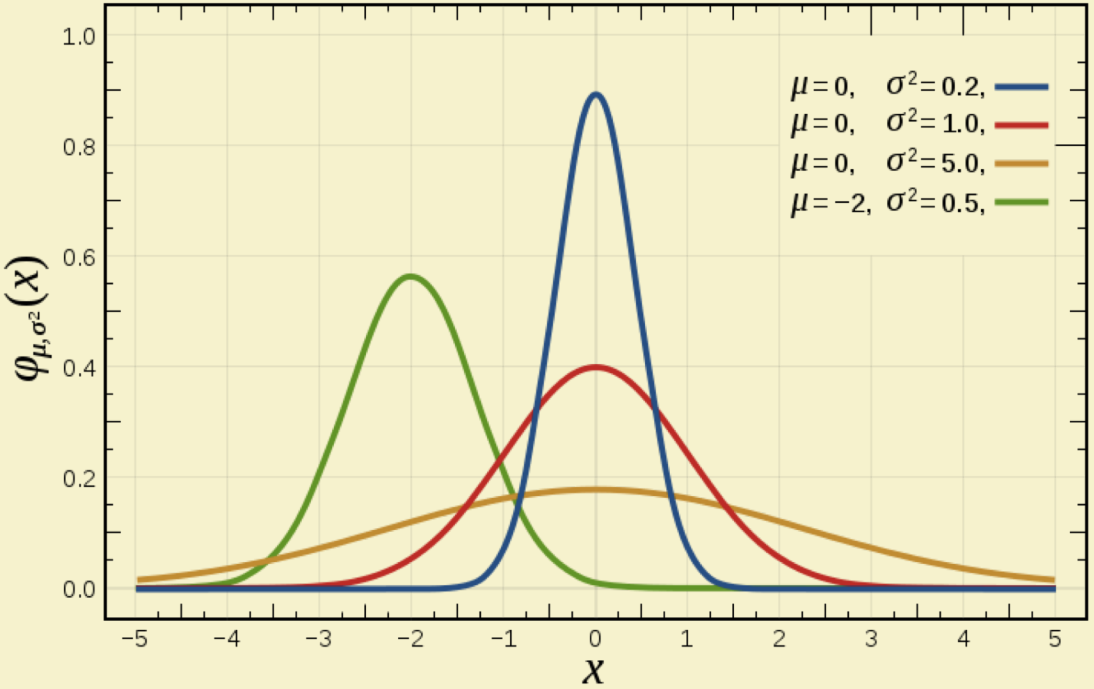
- Continuous valued distribution
- $X \sim N(\mu, \sigma)$

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

$$\Pr(a \le X \le b) = \int_a^b p_\theta(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

- $E[X] = \mu, Var(X) = \sigma^2$
- If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, $X = X_1 + X_2$, then $X \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Plots


1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(x)$
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
$\mu = 0, \quad \sigma^2 = 0.2,$
$\mu = 0, \quad \sigma^2 = 1.0,$
$\mu = 0, \quad \sigma^2 = 5.0,$
$\mu = -2, \quad \sigma^2 = 0.5,$

Concentration Inequalities

Motivation

Many times we do not need to calculate probabilities exactly.

Sometimes it is enough to know that a probability is very small (or very large)

E.g. $P(\text{earthquake tomorrow}) = ?$

This is often a lot easier

I toss a coin 1000 times. The probability that I get 14 consecutive heads is

A	B	C
$< 10\%$	$\approx 50\%$	$> 90\%$

Consecutive heads

Let N be the number of occurrences of 14 consecutive heads in 1000 coin flips.

$$N = I_1 + \ldots + I_{987}$$

where I_i is an indicator r.v. for the event

"14 consecutive heads starting at position i"

$$E[I_i] = P(I_i = 1) = 1/2^{14}$$

$$E[N] = 987 \cdot 1/2^{14} \quad = 987/16384 \approx 0.0602$$

Markov's inequality

For every non-negative random variable X and every value a:

$$P(X \geq a) \leq E[X] \,/\, a.$$

$$E[N] \approx 0.0602$$

$$P[N \geq 1] \leq E[N] \,/\, 1 \leq 6\%.$$

Markov's inequality

For every non-negative random variable X: and every value a:

$$P(X \geq a) \leq E[X] \,/\, a.$$

$$E[X] = E[X \mid X \geq a]\, P(X \geq a) + E[X \mid X < a]\, P(X < a)$$

$\geq a$ ≥ 0 ≥ 0

$$E[X] \geq a\, P(X \geq a) + 0.$$

Patterns

A coin is tossed 1000 times. Give an upper bound on the probability that the pattern HH occurs:

(a) at least 500 times

(b) at most 100 times

Patterns

(a) **Let N be the number of occurrences of HH.**
Last time we calculated $E[N] = 999/4 = 249.75.$

$$P[N \geq 500] \leq E[N] \,/\, 500 = 249.75/500 \approx 49.88\%$$

so 500+ HHs occur with probability $\leq 49.88\%$**.**

(b) $P[N \leq 100] \leq ?$

$$P[N \leq 100] = P[999 - N \geq 899] \leq E[999 - N] \,/\, 899$$
$$= (999 - 249.75)/\, 899$$
$$\leq 83.34\%$$

Chebyshev's inequality

For every random variable X and every t:

$$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$$

where $\mu = E[X]$, $\sigma = \sqrt{Var[X]}$.

Patterns

$$E[N] = 999/4 = 249.75 \qquad \mu = 249.75$$

$$Var[N] = (5 \cdot 999 - 7)/16 = 311.75 \qquad \sigma \approx 17.66$$

(a) $$P(X \geq 500) \leq P(|X - \mu| \geq 14.17\sigma)$$
$$\leq 1/14.17^2 \approx 0.50\%$$

(b) $$P(X \leq 100) \leq P(|X - \mu| \geq 8.47\sigma)$$
$$\leq 1/8.47^2 \approx 1.39\%$$

Chebyshev's inequality

For every random variable X and every a:

$$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$$

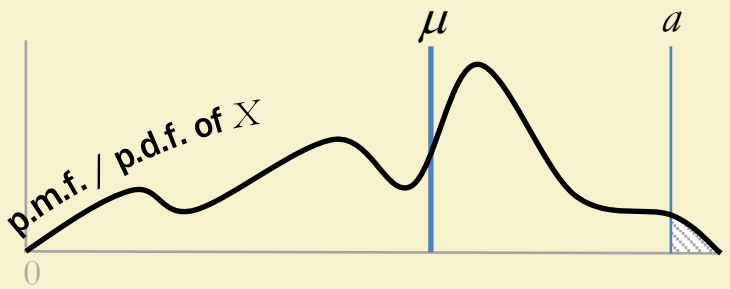
where $\mu = E[X]$, $\sigma = \sqrt{Var[X]}$.

$$P(|X - \mu| \geq t\sigma) = P((X - \mu)^2 \geq t^2\sigma^2) \leq E[(X - \mu)^2] / t^2\sigma^2 = 1 / t^2.$$

An illustration

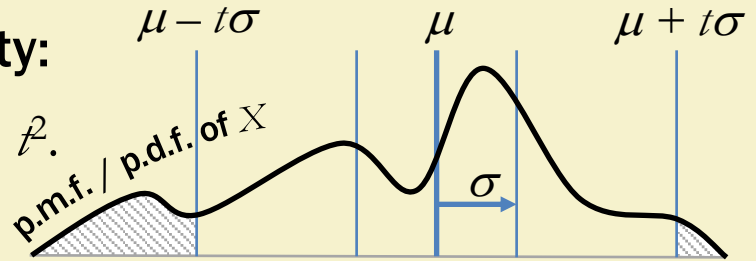
Markov's inequality:

$P(X \geq a) \leq \mu / a.$



Chebyshev's inequality:

$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$



Repertoire of tools

- **Linearity of expectation**: For any random variables $X_1, X_2, \ldots, X_n$, we have
  - $E[\Sigma_i\, Xi] \;=\; \Sigma_i\, E[Xi]$
- **Markov's inequality**: For any random variable X
  - $\Pr[X \;\geq\; c] \;\leq\; E[X]/c$
- **Union bound:** For any sequence of events $E_1, E_2, \ldots, En$, we have
  - $\Pr[U_i\, E_i] \;\leq\; \Sigma_i \Pr[E_i]$

Chernoff bounding

The Chernoff bound for a random variable X is obtained as follows: for any t >0,

$$\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}] \leq E[e^{tX}] / e^{ta}$$

Similarly, for any $t < 0$,

$$\Pr[X \leq a] = \Pr[e^{tX} \geq e^{ta}] \leq E[e^{tX}] / e^{ta}$$

The value of t that **minimizes** $E[e^{tX}] / e^{ta}$ gives the best possible bounds.

When $X = X_1 + \cdots + X_n$:

$$\Pr[X \leq a] \leq \min_{t>0} e^{-ta} \prod_i E[e^{tX_i}]$$

Chernoff bounding

- **Def**: The **moment generating function** of a random variable X is $M_X(t) = E[e^{tX}]$.
- $E[X^n] = M_X^n(0)$, which is the nth derivative of $M_X(t)$ evaluated at $t = 0$.
- Fact: If $M_X(t) = MY(t)$ for all t in (-c, c) for some $c > 0$, then X and Y have the same distribution.
- If X and Y are independent r.v., then $M_{X+Y}(t) = MX(t)\,MY(t)$.

Chernoff bounding

- Let $X_1, X_2, \ldots, X_n$ be n independent random variables in {0,1}, with $X = X_1 + X_2 + \cdots + X_n$.

- For any nonnegative δ

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu}$$

- For any δ in [0,1]

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu\delta^2/3}$$

Chernoff bounding

- Extensions:
  - Chernoff-Hoeffding bounds (bounded r.vs)
  - Azuma's inequality for martingales

- Applications in this course:
  - Sketching: Hashing.
  - Random Projection: Proof of Johnson-Lindenstrauss lemma.
  - Dimensionality reduction: CUR decomposition.

References:

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

- Lecture notes of Andrej Bogdanov. http://www.cse.cuhk.edu.hk/~andrejb/engg2040c/s13/

- Wikipedia.

Thank You!!