ECE253/CSE208 Introduction to Information Theory

Lecture 2: Probability Theory Revisited

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- H. Stark and J. Woods, "Probability, Random Processes, and Estimation Theory for Engineers".
- A. Papoulis, "Probability, Random Variables, and Stochastic Processes".

Probability Space

Probability space is a three-tuple $(\Omega, \mathcal{F}, P)$:

- **Sample space** Ω: the set of all outcomes of a random experiment.
  - Ω may be finite ($\{H, T\}$), countably infinite ($\Omega = \{1,2,3,...\}$), or uncountably infinite ([0,1]).
- **Event space** $\mathcal{F}$: a set whose elements $A \in \mathcal{F}$ are subsets of Ω.
- **Probability function** P: satisfies three axioms:
  - $P(A) \geq 0, \forall\ A \in \mathcal{F}$.
  - $P(\Omega) = 1$.
  - If $A_1, A_2, ...$ are mutually exclusive events ($A_i \cap A_j = \emptyset$, for $i \neq j$), then $P(\bigcup_i A_i) = \sum_i P(A_i)$.

Properties of Probability

Consider events $A, B, C \subseteq \Omega$. Let $P(AB)$ denote $P(A \cap B)$.

- $A \subseteq B \Rightarrow P(A) \leq P(B)$.
- $P(AB) \leq \min\{P(A), P(B)\}$.
- $P(A \cup B) = P(A) + P(B) - P(AB)$.
- $P(\bar{A}) = 1 - P(A)$: $\bar{A} = \Omega \backslash A$.
- Complement rule (De Morgan's Law): $\overline{AB} = \bar{A} \cup \bar{B}$; $\overline{A \cup B} = \bar{A}\bar{B}$
- $P(AB) = P(A|B)P(B) = P(B|A)P(A)$.
- **Conditional probability**: $P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$ (Bayes' theorem).
- If A and B are independent, then $P(AB) = P(A)P(B)$.
- Distributive law: $(A \cup B) \cap C = (AC) \cup (BC)$.

Conditional vis-à-vis Unconditional Probability

Q: $P(A|B) \lessgtr P(A)$?

A: It can be any case in general.

Some special cases:

- If A and B are independent (denoted as $A \perp\!\!\!\perp B$), then $P(A|B) = P(A)$.
- If $AB = \emptyset$ and $0 < P(A), P(B) < 1$, then
  $P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0 < P(A) \Rightarrow P(A|B) < P(A)$.
- If $B \subset A$ and $P(A) < 1$, then
  $P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B)}{P(B)} = 1 > P(A) \Rightarrow P(A|B) > P(A)$.
- If $A \subseteq B$, and $0 < P(B) < 1$, then
  $P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)}{P(B)} > P(A)$.

A Concrete Example

Take a standard deck of cards (52 cards without Jokers). Remove all black queens and kings. When picking a card, define 3 events:

- A: The card picked is a face card.
- B_1: The card picked is a heart.
- B_2: The card picked is a spade.

Question: Find the values of $P(A), P(A|B_1), P(A|B_2)$.

- A: 48 cards left with 8 face cards: $P(A) = \frac{1}{6}$.
- B_1: 13 cards of hearts left with 3 face cards: $P(A|B_1) = \frac{3}{13}$.
- B_2: 11 card of spades left with only 1 face card (the Jack of Spades): $P(A|B_2) = \frac{1}{11}$.

We have $P(A|B_2) < P(A) < P(A|B_1)$.

Chain Rule for Conditional Probability and Partition Theorem

For 3 events, A, B, and C, we have $P(ABC) = P(C|AB)P(B|A)P(A)$.

This can be extended to the case of n events $A_1, A_2, ..., A_n$:

$$P(A_1A_2\cdots A_n) = P(A_n|A_1A_2\cdots A_{n-1}) \times P(A_{n-1}|A_1A_2\cdots A_{n-2}) \times \cdots \times P(A_2|A_1) \times P(A_1)$$

Theorem (Partition Theorem, a.k.a. Law of Total Probability)

Let $B_1, B_2, ..., B_n$ form a **partition** *(i.e., $B_iB_j = \emptyset, \forall i \neq j$ and $\bigcup_i B_i = \Omega$) of the sample space Ω, and assume $P(B_i) \neq 0$ for all i. Then,*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

.

The Birthday Paradox

Given n people named $p_1, p_2, ..., p_n$, what is the probability that at least two of them have the same birthday?

Solution 1:

Let A_2 be the event that p_2 has a different birthday from p_1. p_1 only has their birthday on one of the days of the entire year. Hence, this means that: $P(A_2) = 1 - \frac{1}{365}$.

Let A_3 be the event that p_3 has a different birthday from p_2 and p_1. So, $A_3|A_2$ denotes the event that p_3 has a different birthday from p_2 and p_1 given that p_2 and p_1 have different birthdays, and we have $P(A_3|A_2) = 1 - \frac{2}{365}$.

A_2A_3 is the event that p_1, p_2, and p_3 all have different birthdays.

$$P(A_2A_3) = P(A_3|A_2)P(A_2) = \left(1 - \frac{2}{365}\right)\left(1 - \frac{1}{365}\right)$$

$$\begin{aligned} P(A_3) &= P(A_3|A_2)P(A_2) + P(A_3|\bar{A}_2)P(\bar{A}_2) \\ &= \left(1 - \frac{2}{365}\right)\left(1 - \frac{1}{365}\right) + \left(1 - \frac{1}{365}\right) \times \frac{1}{365} = 0.9945 \end{aligned}$$

The Birthday Paradox (Cont'd)

Now, define a general A_i as the event that the birthday of p_i is not the same day as any of the birthdays of $p_1, p_2, ..., p_{i-1}$. We have $P(A_i|A_1A_2...A_{i-1}) = 1 - \frac{i-1}{365}$.

The probability that all n people have different birthdays is

$$q_n := P(A_1A_2...A_n) = P(A_n|A_1A_2...A_{n-1})P(A_{n-1}|A_1A_2...A_{n-2})...P(A_2|A_1)P(A_1)$$
$$= \left(1 - \frac{n-1}{365}\right) \times \left(1 - \frac{n-2}{365}\right) \times \cdots \times \left(1 - \frac{1}{365}\right)$$

Note that $P(A_1) = 1 - \frac{0}{365} = 1$.

The complement event of $A_1A_2 \cdots A_n$ is at least two people have the same birthday.

Hence, the probability we try to find is $b_n = 1 - q_n$, which is an increasing function in n.

$\implies$ $b_{23} = 0.507$; $b_{30} = 0.706$; $b_{40} = 0.891$; $b_{70} \approx 0.999$.

Solution 2:

To directly calculate q_n. The total number of all possibilities is 365^n since each person has 365 days to choose as his/her birthday. The number of cases that they all have different birthdays is 365 permute n: ${}^{365}P_n = \frac{365!}{(365-n)!}$. Hence, $q_n = \frac{365!}{(365-n)! \times 365^n}$.

Random Variables

Definition (Random Variable (RV))

Let Ω be the sample space of an experiment, and $\mathbb{R}$ denote the set of real numbers. Then, a *random variable* $X : \Omega \mapsto \mathbb{R}$ associated with this experiment is a function that assigns each outcome in Ω to a real number. The range of X is denoted as $\mathrm{val}(X)$.

Example. Flip a coin 5 times. Let X denote the random variable for the number of times the coin came up heads. Then $X(\omega_0) = 3$, for the outcome $\omega_0 = \{\mathrm{HHTHT}\}$. There are two different types of random variables that are often studied: *discrete* and *continuous*.

Random Variables (Cont'd)

If X is a discrete random variable, we use the notation

$$\Pr(X = k) := \Pr(\{\omega : X(\omega) = k\})$$

for the probability of the event $X = k$.

If X is a continuous random variable variable, we use the notation

$$\Pr(a \leq X \leq b) := \Pr(\{\omega : a \leq X(\omega) \leq b\})$$

for the probability of the event $a \leq X \leq b$.

Cumulative Distribution Function (CDF)

Definition

Let X be a random variable associated with an experiment. Then,

$$\boxed{F_X(x) \coloneqq \Pr(X \leq x)}$$

is a *cumulative distribution function*.

Properties of CDFs:

(a) $0 \leq F_X(x) \leq 1$.

(b) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$.

(c) F_X is nondecreasing, namely if $x \leq y$ then $F_X(x) \leq F_X(y)$.

(d) F_X is right-continuous, i.e. $\lim_{x \to a^+} F_X(x) = F_X(a)$.

PMF and PDF

For discrete random variables, we have the *probability mass function* (PMF):

$$p_X(x) \coloneqq \Pr(X = x),$$

where $\sum_{x \in \mathrm{val}(X)} p_X(x) = 1$.

For continuous random variables, we instead consider *probability density function* (PDF),

$$f_X(x) \coloneqq \frac{dF_X(x)}{dx} = F'_X(x),$$

provided that F_X is differentiable at x.[1] Notice that $f_X(x)$ and $\Pr(X = x)$ are two different concepts, which can be related by

$$\Pr(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x$$

and

$$\Pr(X \in A) = \int_{x \in A} f_X(x)dx,$$

where $A \subseteq \mathrm{val}(X)$.

[1] Note that F_X may not be everywhere differentiable even for continuous random variable X.

Expectation

Definition (Expectation)

Let $g : \mathbb{R} \to \mathbb{R}$ be a function and X be a random variable, discrete or continuous, then the *expectation* of the random variable $g(X)$ is

$$\mathrm{E}[g(X)] := \sum_{x \in \mathrm{val}(X)} g(x) p_X(x) \quad \text{or} \quad \mathrm{E}[g(X)] := \int_{x \in \mathrm{val}(X)} g(x) f_X(x) dx,$$

respectively.

Linearity of the expectation operator:

(a) $\mathrm{E}[ag(X) + bh(X)] = a\mathrm{E}[g(X)] + b\mathrm{E}[h(X)]$ for any constants a, b, and arbitrary functions $g(\cdot), h(\cdot)$.

(b) $X \perp\!\!\!\perp Y \Rightarrow \mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

The *indicator function*: $\mathbf{1}_A := \begin{cases} 1, & \text{if } A \text{ is true} \\ 0, & \text{otherwise.} \end{cases} \Rightarrow$

$\mathrm{E}[\mathbf{1}_A] = \Pr(A) \times 1 + \Pr(\bar{A}) \times 0 = \Pr(A), \; F_X(x) = \mathrm{E}[\mathbf{1}_{\{X \leq x\}}]$.

Variance

Given a distribution of a random variable, we use the notion of variance to measure how concentrated that distribution is around the expectation (mean). Formally, we have

Definition (Variance)

$\mathrm{Var}[X] := \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$, where $\mathrm{E}[X^2]$ is the second moment of X.

The following can be derived immediately from the definition:

(a) $\mathrm{Var}[cX] = c^2\mathrm{Var}[X]$.

(b) $\mathrm{Var}[c] = 0$ for any constant c.

(c) $\mathrm{Var}[aX \pm bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] \pm 2ab \times \mathrm{Cov}[X, Y]$.

Characteristic Function and Moment-generating Function

Definition (Characteristic Function)

The characteristic function of a scalar random variable X is defined as $\varphi_X : \mathbb{R} \to \mathbb{C}$:

$$\varphi_X(t) = \mathrm{E}[e^{itX}] = \int_{\mathbb{R}} e^{itx} dF_X(x) = \int_{\mathbb{R}} e^{itx} f_X(x) dx, \quad t \in \mathbb{R}$$

where $i = \sqrt{-1}$.

Definition (Moment-generating Function)

The moment-generating function of a scalar random variable X is defined as $M_X : \mathbb{R} \to \mathbb{R}$:

$$M_X(t) := \mathrm{E}[e^{tX}] = \int_{\mathbb{R}} e^{tx} f_X(x) dx, \quad t \in \mathbb{R}$$

Note that

(a) $\varphi_X(-it) = M_X(t)$.

(b) $\mathrm{E}[X^n] = \frac{d^n M_X(t)}{dt^n}\Big|_{t=0}$.

(c) If $S_n = \sum_{i=1}^n a_i X_i$ for independent RVs $\{X_i\}_{i=1}^n$, then

$$M_{S_n}(t) = M_{X_1}(a_1 t) M_{X_2}(a_2 t) \cdots M_{X_n}(a_n t)$$

Multivariate Random Variables

Definition (Joint CDF)

Let X, Y be two random variables associated with an experiment. Then the *joint cumulative distribution function* of X and Y is

$$F_{XY} \coloneqq \Pr(X \leq x, Y \leq y)$$

and the *marginal cumulative distribution function* of X is

$$F_X(x) \coloneqq \lim_{y \to +\infty} \Pr(X \leq x, Y \leq y).$$

Similarly, the *joint* PMF $p_{XY}(x, y) \coloneqq \Pr(X = x, Y = y)$

the *marginal* PMF of X, $p_X(x) \coloneqq \sum_{y \in \mathrm{val}(Y)} \Pr(X = x, Y = y)$.

The *joint* PDF of continuous X and Y is $f_{XY}(x, y) \coloneqq \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$

the *marginal* PDF of X: $f_X(x) \coloneqq \int_{y \in \mathrm{val}(Y)} f_{XY}(x, y) dy$.

Conditional Distribution

The above summation or integral operation is called *marginalization*. Note that

$$\iint_A f_{XY}(x,y)dxdy = \Pr\Big((x,y) \in A\Big).$$

Definition (Conditional Distribution for Discrete RVs)

Let X and Y be discrete random variables. The *conditional distribution* of Y given $X = x$ is

$$\boxed{p_{Y|X}(y \mid x) := \frac{p_{XY}(x,y)}{p_X(x)}},$$

provided that $p_X(x) \neq 0$.

We say that X and Y are *independent* if $p_{Y|X}(y \mid x) = p_Y(y)$.

Note that $X \perp\!\!\!\perp Y \Leftrightarrow p_{XY}(x,y) = p_X(x)p_Y(y)$.

Q: If $X \perp\!\!\!\perp Y$. For arbitrary functions $g(\cdot)$ and $h(\cdot)$, are $g(X) \perp\!\!\!\perp h(Y)$?

A: Yes.

Covariance & Correlation

Definition (Covariance)

Let X and Y be two random variables. Their *covariance* is defined as

$$\mathrm{Cov}[X, Y] := \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y].$$

Definition (Correlation)

Let X and Y be two random variables. Their *correlation* is defined as

$$\mathrm{Corr}[X, Y] := \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

$\mathrm{Corr}[X, Y] \in [-1, 1]$ is a measure of linear association between X and Y. They are uncorrelated if $\mathrm{Corr}[X, Y] = 0$. $\mathrm{Corr}[X, Y] = \pm 1 \Leftrightarrow Y = aX + b$ for some constants a, b.

Note that if $X \perp\!\!\!\perp Y \implies$ they are uncorrelated (but NOT vice versa).

Example: $X \sim N(0, 1)$ and $Y = X^2$

Law of Total Expectation and Variance

- The law of total expectation: $\mathrm{E}(X) = \mathrm{E}_Y\Big(\mathrm{E}(X|Y)\Big)$.
- The law of total variance: $\mathrm{Var}(X) = \mathrm{E}_Y\Big(\mathrm{Var}(X|Y)\Big) + \mathrm{Var}_Y\Big(\mathrm{E}(X|Y)\Big)$.

Finally, the PDF for multivariate Gaussian distribution is given as

$$\boxed{f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) \coloneqq \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$$

where $\mathbf{x} = [x_1,\ldots,x_n]^\top$, $\boldsymbol{\mu} = [\mu_1,\ldots,\mu_n]^\top \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix, of which the (i,j)-entry is $\mathrm{Cov}[X_i, X_j]$.

LLN and CLT

Theorem (Law of Large Numbers (LLN))

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables so that $\mathrm{E}[X_1] = \mathrm{E}[X_2] = \cdots = \mathrm{E}[X_n] < \infty$. Let

$$\bar{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

denote the sample mean of those n random variables. Then $\bar{X}_n \to \mathrm{E}[X_1]$ as $n \to \infty$ almost surely (a.s., strong law) and in probability (i.p., weak law).

Theorem (Central Limit Theorem (CLT))

Let $X_1, X, \ldots, X_n$ be a sequence of i.i.d. random variables, and assume that $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, for all i. Let $S_n := X_1 + X_2 + \cdots + X_n$. Then, $\mathrm{E}[S_n] = n\mu$, $\mathrm{Var}[S_n] = n\sigma^2$ and we have the standardization of S_n,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\text{i.p.}} N(0,1) \text{ as } n \to \infty$$

where $N(0,1)$ denotes the standard normal random variable.

Convergence of Random Variables

Definition (Convergence of random variables)

Given a sequence of random variables, $X_1, X_2, \ldots$, we say that the sequence $X_1, X_2, \ldots$ converges to a random variable X:

1. In probability ($X_n \xrightarrow{P} X$) if for every $\epsilon > 0, \lim_{n\to\infty} \Pr\{|X_n - X| > \epsilon\} = 0$
2. Almost sure [a.k.a. convergence with probability 1] ($X_n \xrightarrow{a.s.} X$) if $\Pr\{\lim_{n\to\infty} X_n = X\} = 1$
3. In distribution ($X_n \xrightarrow{dist} X$) if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$
4. In r^{th}-order mean ($X_n \xrightarrow{L^r} X$) if $\lim_{n\to\infty} E|X_n - X|^r = 0$
5. In mean square (special case when $r = 2$) if $\lim_{n\to\infty} E(X_n - X)^2 = 0$

Strong & Weak Convergence

Strong convergence: Convergence almost surely and convergence in r^{th}-order mean.

Weak convergence: Convergence in probability and convergence in distribution.

Their relationships are given as follows:

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{dist} X$$

$$X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{dist} X$$

$$X_n \xrightarrow{a.s.} X \nLeftrightarrow X_n \xrightarrow{L^r} X$$

Concentration Inequalities

Concentration inequalities provide bounds on how a random variable deviates from some value (e.g., its expected value):

- Markov's inequality: If X is a nonnegative RV, then $P(X \geq t) \leq \frac{\mathrm{E}(X)}{t}$, for any $t > 0$.

- Chernoff's inequality: If X is a nonnegative RV, then $P(X \geq t) = P(e^{aX} \geq e^{at}) \leq \frac{\mathrm{E}(e^{aX})}{e^{at}}$, for any $a > 0$.

- Chebyshev's inequality: $P(|X - \mathrm{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$, for any $t > 0$.

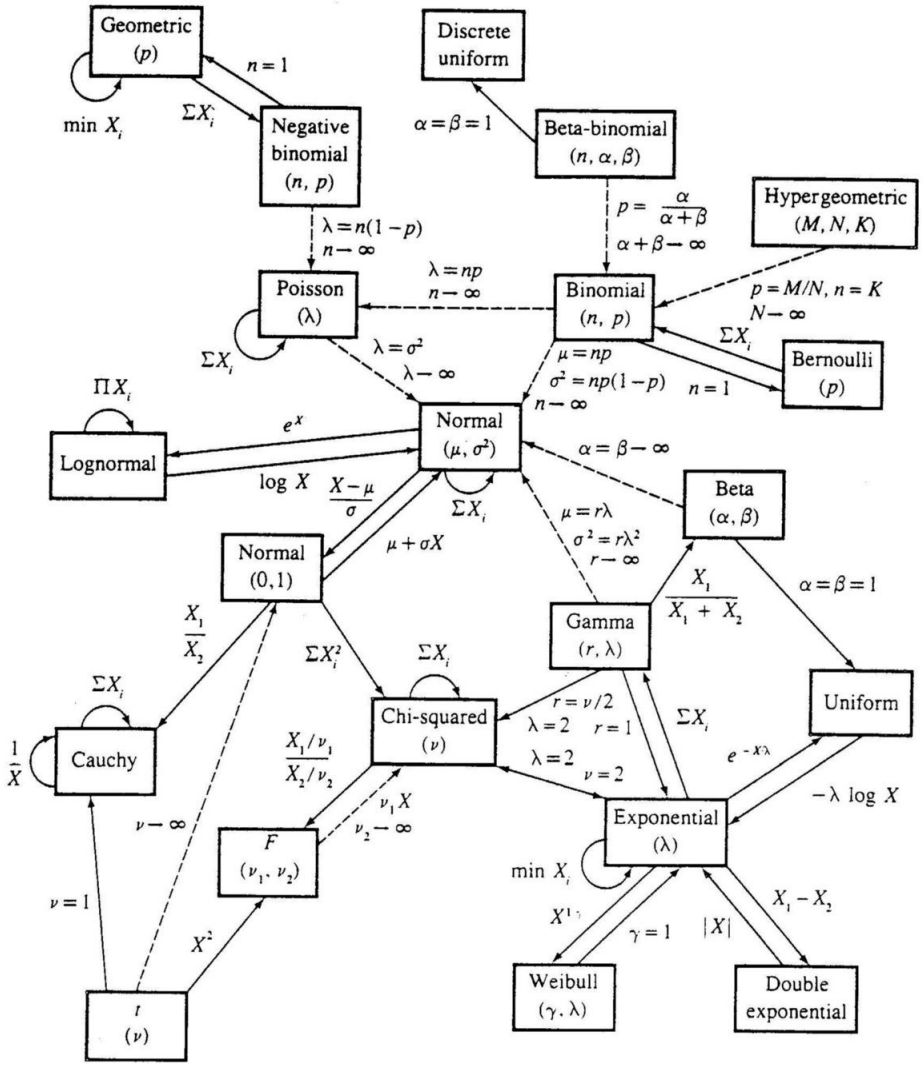
- Hoeffding's inequality: Consider the empirical mean $\bar{X}_n := \frac{1}{n}(X_1 + \cdots + X_n)$ for independent random variables $X_i \in [a_i, b_i]$ for all i. Then, $P(|\bar{X}_n - \mathrm{E}(\bar{X}_n)| \geq t) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$, for any $t > 0$.

Common Distributions

Name of the probability distribution	Probability distribution function	Mean	Variance
Binomial distribution	$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Geometric distribution	$\Pr(X = k) = (1-p)^{k-1} p$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$
Normal distribution	$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Uniform distribution (continuous)	$f(x \mid a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential distribution	$f(x \mid \lambda) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Figure: PDF or PMF of commonly used random variables (Wiki).

Common Distributions (Cont'd)



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Thank You!

Email: <zhangy@ucsc.edu>

Homepage: https://people.ucsc.edu/~yzhan419/