Probability Theory and Stochastic Processes

ITÜ

İstanbul Teknik Üniversitesi

Mustafa Kamasak, PhD

v2018.10.03

Stochastic vs. Deterministic Systems

- Deterministic system
  - no randomness
  - same output for the same input/other conditions
- Stochastic system
  - randomness due to
    - Limited capabilities of production, measurement,
    - various unknown factors (noise, uncertain parameters etc.)
  - different output even for the same input/other conditions
- Only of the theoretical systems are deterministic. Their physical implementations and measurements are stochastic.

Stochastic vs. Deterministic Systems

- This course deals with modeling the output (events / outcomes) of stochastic systems.
- Outcomes are the observation of results from experiments, trials etc.
- Events are the observation of events that happen without a human setup
- Random output instance (at a certain time) can be modelled with probability theory.
- The output can be
  - nominal
  - ordinal
  - interval (continuous-valued or discrete-valued)
- Time series of output can be modelled as a stochastic process.

Sets

- A set is a collection of objects/elements

  $$A = \{\zeta_1, \zeta_2, \cdots, \zeta_N\}$$

  There are N elements in set A
- Notation:
  $\zeta_2 \in A$ means ζ_2 is in set A
  $\zeta_2 \notin A$ means ζ_2 is **not** in set A
- Empty (or null) set Φ contains no elements by definition
- If a set has n elements, it has 2^n subsets including the empty set and itself
- $A \supseteq B$ means B is a subset of A
- $A \supseteq B$ and $B \supseteq A$ then $A = B$

Set Operators

- Complement:

$$A^c = \{x : x \in S \text{ but } x \notin A\}$$

- Union:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

- Intersection:

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

- Symmetric difference:

$$A \triangle B = (A^c \cap B) \cup (A \cap B^c)$$

Properties of Sets

For any subset of S

- Commutative:
  $A \cup B = B \cup A$
  $A \cap B = B \cap A$
- Associative law:
  $(A \cup B) \cup C = A \cup (B \cup C)$
  $(A \cap B) \cap C = A \cap (B \cap C)$
- Distributive law:
  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
  $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Disjoint Sets

- Sets A and B are disjoint (mutually exclusive) if $A \cap B = \Phi$
- Several sets $A_1, A_2, \cdots, A_N$ are mutually exclusive if $A_i \cap A_j = \Phi$ when $i \neq j$.
- A_i are called a partition of S if
  - A_i are mutually exclusive
  - $\cup A_i = S$

De Morgan Law

- De Morgan law is used to find the complement of complicated operations on sets
- De Morgan Law
  $(\cup_i A_i)^c = \cap_i A_i^c$ $(\cap_i A_i)^c = \cup_i A_i^c$
- *General form*
  Replace all sets with its complement
  Replace union with intersection
  Replace intersection with union
  Replace Φ with S
  Replace S with Φ
- For example:

$$[A \cap (B \cup \Phi)]^c = A^c \cup (B^c \cap S)$$

Duality

- If a complicated equality is proven then its dual is also correct.
- General form
  Exchange union with intersection
  Exchange intersection with union
  Exchange Φ with S
  Exchange S with Φ
- For example: if $S \cap A = A$ is proven then its dual $\Phi \cup A = A$ is also correct

Sample space and empty set

- S: sample space / certain event
  It is the set of all possible outcomes/events
- Φ: empty set/ impossible event
- Field:
  Let A_i be a subset of S
  If A_i are **finitely** many $F = \{A_i : A_i \subseteq S, i \leq N\}$ and
  - $\Phi \in F$
  - If $A_i \in F$ then $A_i^c \in F$
  - If $A_i \in F$ for $i = 1, 2...N$ then $\cup_{i=1}^{N} A_i \in F$

  Then F is called a field.
- Borel field:
  If A_i are **infinitely** many then it is called a Borel field.
  A Borel field is closed under complement and countable union operations
- Suppose $\mathcal{B} = \{A_i : A_i \subseteq S \text{ and } i \in \mathbb{N}\}$ is a Borel field. Any subset A of S is called and event iff $A \in \mathcal{B}$

Axioms of Probability

- Probability assigns a unique number in [0,1] range to each event
- Axioms of probability (by Kolmogorov in 1933)
  - $P(S) = 1$
  - $P(A) \geq 0$ for every $A \in \mathcal{B}$
  - $P(\cup_i A_i) = \sum_i P(A_i)$ for all $A \in \mathcal{B}$ is A_i are mutually disjoint
- Axioms are accepted without a proof.

Theorems

Suppose A and B are two events and B_i forms a partition of S then

- $P(\Phi) = 0$ and $P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- $P(B \cap A^c) = P(B) - P(B \cap A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If $A \subseteq B$ then $P(A) \leq P(B)$
- $P(A) = \sum_i P(A \cap B_i)$

These theorems can be proven using axioms and other proven theorems

Conditional Probability

If S is the sample space, $\mathcal{B}$ is Borel field and let $A, B \in \mathcal{B}$ then

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

Independence

- If $P(A|B) = P(A)$ then events A and B are independent.
- Observing event B has no effect (gives us no extra information) on observation of event A.
- Events A and B are independent iff $P(A \cap B) = P(A)P(B)$
- The following statements are equal
  - A and B are independent
  - A^c and B^c are independent
  - A and B^c are independent
  - A^c and B are independent
    Proof:

$$\begin{aligned} P(A^c \cap B) &= P(B) - P(A \cap B) \\ &= P(B)(1 - P(A)) \\ &= P(B)P(A^c) \end{aligned}$$

Mutual Independence

- A collection of events A_i are called mutually independent iff every subcollection consists of independent events
- It is possible to have pairwise independent events, but the whole set may not be mutually independent

Example:

Consider tossing a fair coin ($S = \{H, T\}$)
A_1: H on the first toss
A_2: H on the second toss
A_3: same outcome on both tosses
Are events A_1, A_2, A_3 mutually independent?

Bayes Theorem

Suppose $A_1, ..., A_N$ form a partition of the sample space S For an arbitrary event B

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

- $P(A_i)$ is called prior information
- $P(B|A_i)$ is called *likelihood*
- $P(A_j|B)$ is called *posteriori* probability

Bayes Theorem – Example 1

- Consider a rare disease that is seen 1 in every million.
- Consider a medical test that is 99% accurate. This means if a person has this disease (case positive), the test will detect it correctly with the probability of 0.99
- When someone takes this test and the test result turns out to be positive, what is the probability of this person having this disease?
- Prior information: $P(\text{disease}+) = 0.000001$
  Likelihood: $P(test + |disease+) = 0.99$
  Posterior probability: $P(disease + |test+) =?$

Bayes Theorem – Example 1

- Prior information: $P(\text{disease}+) = 0.000001$
  Likelihood: $P(test + |disease+) = 0.99$
  Posterior probability: $P(disease + |test+) =?$
- Using Bayes theorem:

$$P(disease + |test+) = \frac{P(test + |disease+)P(disease+)}{P(test+)}$$

- What is $P(test+) =?$
- The test can result positive when there is disease (true positive), or when there is no disease (false positive). Hence

$$\begin{aligned} P(test+) = & \, P(test + |disease+)P(disease+) \\ & + P(test + |disease-)P(disease-) \end{aligned}$$

Bayes Theorem – Example 1

- Prior information: $P(\text{disease}+) = 0.000001$
  Likelihood: $P(test + |disease+) = 0.99$

$$\begin{aligned} P(test+) &= P(test + |disease+)P(disease+) \\ &\quad + P(test + |disease-)P(disease-) \\ &= 0.99 \times 0.000001 + 0.01 \times 0.999999 \\ &= 0.01000097999901 \approx 0.01 \end{aligned}$$

- Hence

$$\begin{aligned} P(disease + |test+) &= \frac{P(test + |disease+)P(disease+)}{P(test+)} \\ &= \frac{0.99 \times 0.000001}{0.01} \\ &= 0.000099 < 0.01\% \end{aligned}$$

- Although it is quite an accurate test, it seems meaningless

Bayes Theorem – Example 2

A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:[1]

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue.
- The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?

[1] Example taken from A. Tversky, D. Kahneman, Evidential impact of base rates, in Judgement under uncertainty: Heuristics and biases, D. Kahneman, P. Slovic, A. Tversky (editors), Cambridge University Press, 1982

Bayes Theorem – Example 2

- Apriori probabilities: $P(Green) = 0.85$ and $P(Blue) = 0.15$
- Likelihoods: $P(Witness = Blue|Blue) = 0.8$
- From Bayes theorem

$$P(Blue|Witness = Blue) = \frac{P(Witness = Blue|Blue) \times P(Blue)}{P(Witness = Blue)}$$

$$\begin{aligned}
P(Witness = Blue) &= P(Witness = Blue|Blue) \times P(Blue) \\
&\quad + P(Witness = Blue|Green) \times P(Green) \\
&= 0.8 \times 0.15 + 0.2 \times 0.85 \\
&= 0.29
\end{aligned}$$

$$P(Blue|Witness = Blue) = \frac{0.8 \times 0.15}{0.29} \approx 41\%$$

Permutation and Combination

In a repeated trial, we want to enumerate the number of possible outcomes (without repetition of objects)

- Permutation:
  The number of possible arrangements of k objects from a collection of n objects when the **ordering is important**.

$$\begin{aligned} P_k^n &= n(n-1)(n-2)..(n-k+1) \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

- Combination:
  The number of possible arrangements of k objects from a collection of n objects when the **ordering is NOT important**.

$$\begin{aligned} C_k^n = \left(\begin{array}{c} n \\ k \end{array} \right) &= \frac{n(n-1)...(n-k+1)}{k!} \\ &= \frac{n!}{k!(n-k!)} \end{aligned}$$

Properties of Combination

- $C_0^n = \left(\begin{array}{c} n \\ 0 \end{array} \right) = 1$ if $n > 0$
- $C_k^n = C_{n-k}^n$
- Binomial theorem: The combinations C_k^n gives the binomial coefficients.

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}$$

Permutation and Combination

In a repeated trial, we want to enumerate the number of possible outcomes (with repetition of objects)

- Permutation:
  The number of possible arrangements of k objects from a collection of n objects when the **ordering is important**.

$$\tilde{P}_k^n = n^k$$

- Combination:
  The number of possible arrangements of k objects from a collection of n objects when the **ordering is NOT important**.

$$\tilde{C}_k^n = \left(\begin{array}{c} n+k-1 \\ k \end{array} \right)$$

Permutation Examples

Example: How many different 2 digit numbers can you obtain using digits $\{2, 5, 8\}$ without repeating digits?

- Ordering is important as $25 \neq 52$
- For the first digit there are 3 options from $\{2, 5, 8\}$, for the second digit there are 2 options.
- Hence there are $3 \times 2 = 6$ possibilities
  258, 285, 528, 582, 825, 852

Example: Assuming 20 letters are used to form 3-letter license plates. How many different possibilities if the letters can be repeated?

- License plate ABC $\neq$ ACB, ordering is important
- $20 \times 20 \times 20 = 8000$
- If repeated letters is not permitted, then $20 \times 19 \times 18 = 6840$

Combination Examples

Example: A thesis committee will be formed with 4 professors out of 10 professors in a department. How many different committees can be formed?

- The order of the committee members is not important.
- Hence there are C_4^{10} different committees that can be formed.

Example: There are

Combination Examples

Example: A thesis committee will be formed with 2 professors from mechanical engineering and 3 professors from computer engineering. If mechanical engineering department has 10 and computer engineering department has 8 professors, how many different committees can be formed?

- The order of the committee members is not important.
- There can be C_2^{10} different selection from mechanical eng. and C_3^8 different possibilities from computer eng.
- Hence, there area $C_2^{10} \times C_3^8$ different committees

Random Variables

- A random variable X is a mapping from the sample space S to a subset $\mathcal{X}$ of the real line $\mathbb{R}$

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

- Using a random variable (rv) a real number can be assigned to an event/outcome
- For example, the experiment of coin flipping can generate $S = \{H, T\}$

$$\begin{aligned} X :& H \to 1 \\ & T \to -1 \end{aligned}$$

or

$$\begin{aligned} Y :& H \to 100 \\ & T \to 40 \end{aligned}$$

Both X and Y are random variables.

Discrete Random Variables

- A discrete rv takes a finite or countably infinite number of possible values with specific probabilities assigned to each value.
- If X is a discrete rv it assigns discrete values such as $x_1, x_2, ...$ to the events/outcomes
- Then p_i means probability of X generating the value of x_i: $p_i = P(X = x_i)$
- It is possible for X to assign multiple events/outcomes to a certain value such as x_i. Hence $p_i = P(X = x_i) = \sum_{s \in S, X(s) = x_i} P(s)$
- While assigning probabilities
  - $p_i \geq 0$ for all i
  - $\sum_i p_i = 1$

Probability Mass Function (pmf)

- An assignment $x_i \to p_i$ is called discrete distribution or a discrete probability distribution
- A function $f(x) = P(X = x)$ for $x \in X$ is called a probability mass function (pmf)
- A pmf is discrete in values
- Why do we use "mass" in pmf?

Cumulative Distribution Function (cdf)

- $F(x) = P(X \leq x) = P\{s : s \in SsuchthatX(s) \leq s\}$ is called cumulative distribution function (cdf)
- Properties of cdf
  - $F(x)$ is a nondecreasing function
    $F(x) \leq F(y)$ for all $x \leq y$ where $x, y \in \mathbb{R}$
  - $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
  - F(x) is right continuous
    For all $x \in \mathbb{R}$, $\lim_{h \to 0} F(x + h) = F(x)$

  Examples of valid and invalid cdf comes here!!!
- Probability of x is: $p(x) = F(x) - F(x^{-})$

Probability Density Function (pdf)

- For a continuous rv $p(X = x) = 0$
- Define $f(x)$ associated with $x \in \mathbb{R}$ such that
  - $f(x) \geq 0$ for all $x \in \mathbb{R}$
  - $\int_{\mathbb{R}} f(x)dx = 1$
- For a given pdf f(x)

$$f(x \in A) = \int_A f(x)dx$$

which is the area under pdf for the given region A

Cumulative Distribution Function (cdf)

- cdf is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{\infty} f(t)dt$$

  for all $x \in \mathbb{R}$
- F(x) should have finite or countably infinite number of discontinuities.
- $P(X < a)$ and $P(X \leq a)$ are the same, which is

$$F(a) = \int_{-\infty}^{a} f(t)dt$$

- $P(X > b)$ and $P(X \geq b)$

$$1 - F(b) = \int_{b}^{\infty} f(t)dt$$

- $P(a < X < b)$ or $P(a \leq X < b)$ or $P(a < X \leq b)$ or $P(a \leq X \leq b)$

$$F(b) - F(a) = \int_{a}^{b} f(t)dt$$

Relation of pdf with cdf

For discrete rv

- pdf → cdf

$$F(x) = \sum_{-\infty}^{x} f(x)$$

- cdf → pdf

$$f(x) = F(x) - F(x^{-})$$

For continuous rv

- pdf → cdf

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

- cdf → pdf

$$f(x) = \frac{d}{dt}F(t)\Big|_{t=x}$$

Expected Value of a Distribution

- All possible values of a rv X such that $f(x) > 0$ is called the support of the distribution of X. Support of X is denoted by $\mathcal{X}$
- Expected value of a distribution is denoted by $E(x)$
  - For discrete distributions:

$$E(X) = \sum_{x_i \in \mathcal{X}} x_i f(x_i)$$

  - For continuous distributions:

$$E(X) = \int_{\mathcal{X}} x f(x) dx$$

- The expected value is also called the **mean** of the distribution and typically denoted by μ

Variance of a Distribution

- The variance of rv X is
  - For discrete distributions:

$$\sigma_x^2 = \sum_{x_i \in \mathcal{X}} (x_i - \mu)^2 f(x_i)$$

  - For continuous distributions:

$$\sigma_x^2 = \int_{\mathcal{X}} (x - \mu)^2 f(x) dx$$

- Variance is a measure of deviation of a rv from its mean

Standard Deviation of a Distribution

- The square root of variance is called the standard deviation of the distribution and it is typically denoted by σ
  - For discrete distributions:

$$\sigma = \sqrt{\sum_{x_i \in \mathcal{X}} (x_i - \mu)^2 f(x_i)}$$

  - For continuous distributions:

$$\sigma = \sqrt{\sigma_x^2 = \int_{\mathcal{X}} (x - \mu)^2 f(x) dx}$$

- Both variance and standard deviation of all distributions are nonnegative $\sigma > 0$ $\sigma^2 > 0$

What does SD mean?

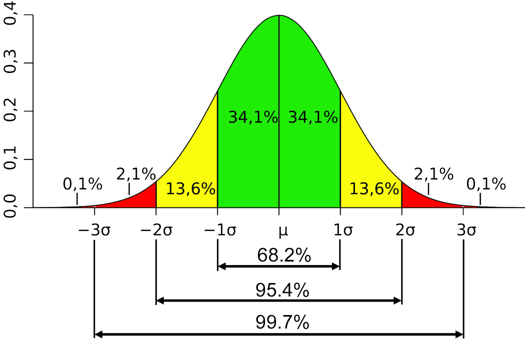
- For a Gaussian (will cover this later) distributed rv X, the range
  - $[\mu_X - \sigma, \mu_X + \sigma]$ contains %68.2
  - $[\mu_X - 2\sigma, \mu_X + 2\sigma]$ contains %95.4
  - $[\mu_X - 3\sigma, \mu_X + 3\sigma]$ contains %99.7

  of the values of this rv.
- Hence for a continuous rv:

$$\int_{\mu-\sigma}^{\mu+\sigma} f(x)dx = 0.682$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} f(x)dx = 0.954$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} f(x)dx = 0.997$$



Expected Value of a Function of rv

- Consider a function of a rv: $X \to g(X)$
- $g(X)$ is also a rv as it maps events/outcomes to another subset of $\mathbb{R}$
- The expected value of $g(X)$ is:
  For discrete rv:

$$E(g(X)) = \sum_{x_i \in \mathcal{X}} g(x_i) f(x_i)$$

  For continuous rv:

$$E(g(X)) = \int_{\mathcal{X}} g(x) f(x) dx$$

Mean an Variance of Translation and Scaling

Consider a rv X and constant values T and S

- $Y = X + T$
- $Z = SX$

What is the mean and variance of rv's Y and Z?

Expectation is a Linear Operator

- Expectation is a **linear operator**
- It can exchange order with other linear operators such as summation, integration etc.
- For example:
  Consider a series of functions $g_i(X)$ and constants a_i. What is the expected value of $Y = \sum_{i=1}^{N} a_i g_i(X)$?
- The variance was given as:

$$\begin{aligned}\sigma^2 &= E((X-\mu_x)^2) \\ &= E(X^2 - 2X\mu_x + \mu_x^2) \\ &= E(X^2) - 2\mu_x E(X) + \mu_x^2 \\ &= E(X^2) - \mu_x^2\end{aligned}$$

  Hence:

$$E(X^2) = \sigma_X^2 + \mu_X^2$$

Median of a distribution

- x_m is the median of $f(x)$ if $F(x_m) = 0.5$
- For a continous distribution:

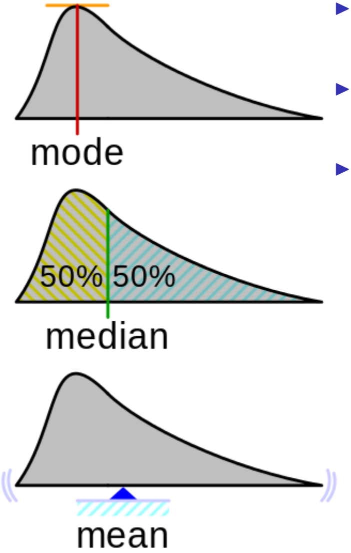
$$P(x \geq x_m) = P(X \leq x_m) = 0.5$$

- How is it defined for discrete rv?

Mode of a distribution

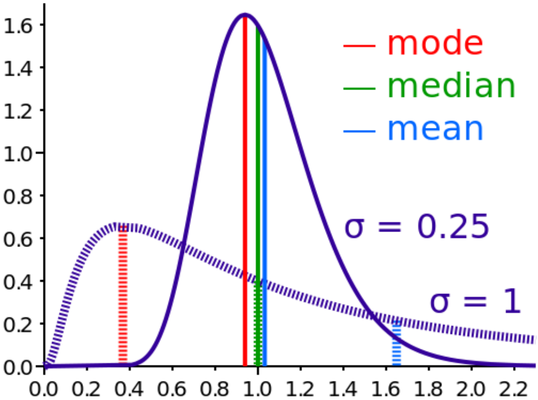
- For discrete rv: The mode is the value x at which its pmf takes its maximum value.
- For a continuous rv: The mode is the value x at which its pdf has its maximum value

Comparison of Mean, Median and Mode



- Mode is the most likely value of an rv that has the highest value of pmf/pdf
- Median is the value of an rv that divides the pmf/pdf in half
- Mean is the value of an rv that is the center of mass of pmf/pdf

Comparison of Mean, Median and Mode



- Mean median and mode have very close values for some distributions
- For other distributions, their values can be quite different.

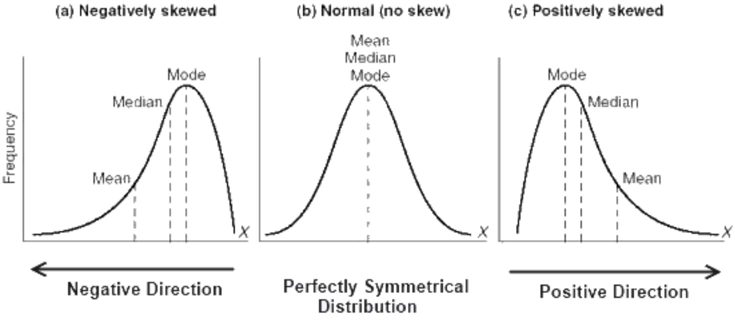
Comparison of Mean, Median and Mode

- If a random variable has symmetric (no skew) distribution, its mean and median are the same.
- However, having the same median and mean does not necessarily imply a symmetric distribution. For example: Consider a discrete rv with a support of $\mathcal{X} = \{-2, 0, 4\}$. The probabilities for these values are $P(X = -2) = 1/3$, $P(X = 0) = 1/2$, $P(X = 4) = 1/6$. Then:

$$\begin{aligned}\mu_X &= \frac{1}{3}(-2) + \frac{1}{2}0 + \frac{1}{6}4 \\ &= 0\end{aligned}$$

and since $C_X(-2) = 1/3$, $C_X(0) = 5/6$, $C(4) = 1$ the median of X is also 0.

Comparison of Mean, Median and Mode



- If a random variable has symmetric and unimodal (single peak) distribution, its mode, mean and median are the same.
- If it is positively skewed then mode<median<mean
- If it is negatively skewed then mode>median>mean

Standard Probability Distributions

For discrete rv

- Bernoulli
- Binomial
- Poisson
- Geometric
- Uniform

For continuous rv

- Uniform
- Normal (Gaussian)
- Standard normal
- Gamma
- Exponential
- Chi-square
- Lognormal
- Student's t
- Cauchy
- F
- Beta
- Negative exponential
- Weibull
- Rayleigh

Bernoulli Distribution

- Discrete distribution
- Single parameter p
- Bernoulli(p)

$$f(x) = P(X = x) = p^x (1 - p)^{1-x}$$

  for $x = \{0, 1\}$ and $0 \leq p \leq 1$

- This means:

$$P(X = 0) = 1 - p$$

  and

$$P(X = 1) = p$$

Bernoulli Distribution – Mean and Variance

$$f(x) = P(X = x) = p^x(1-p)^{1-x}$$

Expected value:

$$\begin{aligned} E(X) = \mu_X &= (1)p + (0)(1-p) \\ &= p \end{aligned}$$

Variance:

$$\begin{aligned} E(X^2) &= (1^2)p + (0^2)(1-p) \\ &= p \end{aligned}$$

Hence:

$$\begin{aligned} \sigma_X^2 &= E(X^2) - \mu_X^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

Binomial Distribution

- Discrete distribution
- Two parameters (n, p)
- Binomial(n,p)

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{(n-x)}$$

- Repeated Bernoulli trials lead to Binomial distribution.
  Example: If a fair coin is tossed 20 times, what is the probability of getting 6 tails?
  Probability of getting 6 tails and 19 heads with a particular order is $0.5^6(1-0.5)^{19}$. There are C_6^{25} different cases (order is not important) with 6 tails.
  Example: If a fair coin is tossed 25 times, what is the probability of getting 6 consecutive tails?

Binomial Distribution – Mean and Variance

Mean:

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{(n-x)} \\ &= \sum_{x=1}^{n} \frac{n}{x} \left(\begin{array}{c} n-1 \\ x-1 \end{array} \right) p^x (1-p)^{(n-x)} \\ &= np \sum_{x=1}^{n} \frac{n}{x} \left(\begin{array}{c} n-1 \\ x-1 \end{array} \right) p^{x-1} (1-p)^{(n-x)} \\ &= np[p + (1-p)]^{(n-1)} \\ &= np \end{aligned}$$

Remember binomial theorem:

$$(a+b)^n = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) a^x b^{n-x}$$

Binomial Distribution – Mean and Variance

Variance:

$$\begin{aligned} E(X(X-1)) &= n(n-1)p^2 \sum_{x=2}^{n} \left(\begin{array}{c} n-2 \\ x-2 \end{array} \right) p^{x-2}(1-p)^{n-x} \\ &= n(n-1)p^2 \end{aligned}$$

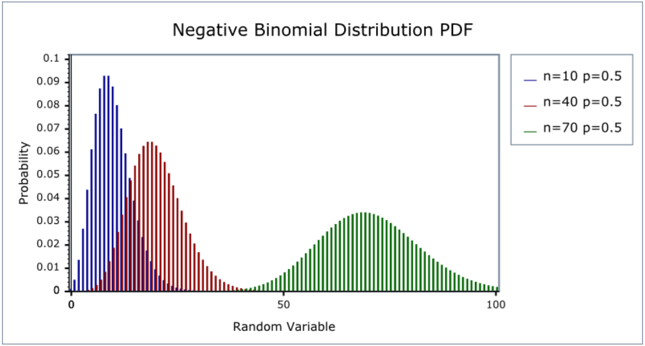
Furthermore

$$\begin{aligned} E(X^2) &= E(X(X-1)) + E(X) \\ &= n(n-1)p^2 + np \end{aligned}$$

Then

$$\begin{aligned} \sigma_X^2 &= E(X^2) - \mu_X^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p) \end{aligned}$$

Binomial Distribution – Example



For the same p value:

- Expected value increases linearly with n (see the shift in pdf)
- Variance increases linearly with n (see the expansion in pdf)
- Hence for a fixed p value, the pdf shifts right and expands as n increases

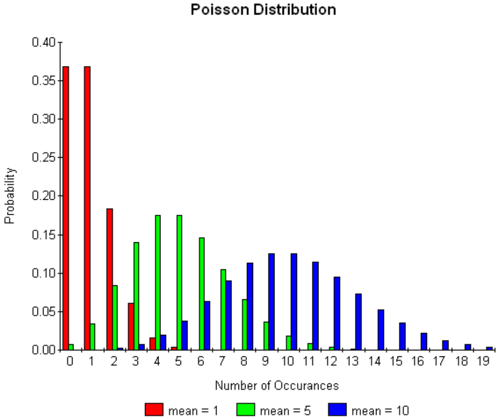
Poisson Distribution

- Discrete distribution
- Single parameter $\lambda > 0$
- Poisson(λ)

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

- Mean: $\mu_X = \lambda$ (Derivation is left as an exercise)
- Variance: $\sigma_X^2 = \lambda$
- Mean is equal to variance.

Poisson Distribution – Example



- As λ increases the mean increases linearly. Observe the shift in the pdf.
- As λ increases the variance increases linearly. Observe the expansion in the pdf.

Geometric Distribution

- Discrete distribution
- Single parameter p
- Geometric(p)

$$f(x) = p(1-p)^{x-1}$$

- x is the number of trials needed for the Bernoulli trials to produce "1" for the first time. This can also be regarded as the number of trials before a success.
- Hence, it should produce $x-1$ times "0" and "1" in the x^{th} trial.
- Mean: $\mu_X = \frac{1}{p}$
- Variance: $\sigma_X^2 = \frac{1-p}{p^2}$

Uniform Distribution

- Discrete distribution
- Each of the possible K outcomes are equally likely

  $$f(x_i) = \frac{1}{K}$$

  for $i \in \{0, 1, ..., K-1\}$
- Assuming $x_i \in [a, b]$ with $b > a$
  - Mean:

    $$\mu_X = \frac{a+b}{2}$$
  - Variance:

    $$\sigma_X^2 = \frac{(b-a+1)^2 - 1}{12}$$

Uniform Distribution

- Continuous distribution $x \in \mathbb{R}$
- Two parameters: (a, b) with $b > a$
- Uniform(a,b)

$$f(x) = \frac{1}{b - a}$$

Normal Distribution

- Continuous distribution $x \in \mathbb{R}$
- Typically referred as Gaussian distribution
- Widely used
- Two parameters (μ, σ)
- $\mathcal{N}(\mu, \sigma^2)$

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

Standard Normal Distribution

- Gaussian Distribution with zero mean and unit variance
- $\mathcal{N}(0,1)$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

- It is possible to normalize X from $\mathcal{N}(\mu,\sigma)$ into $\mathcal{N}(0,1)$ with the following transformation

$$Z = \frac{X-\mu}{\sigma}$$

- It is possible to normalize Z from $\mathcal{N}(0,1)$ into $\mathcal{N}(\mu,\sigma)$ with the following transformation

$$X = \sigma(Z+\mu)$$

Gamma Distribution

- Continuous distribution
- Two parameters (α, β) with $0 < \alpha, \beta < 1$
- $\text{Gamma}(\alpha, \beta)$

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \exp\left\{-\frac{x}{\beta}\right\} x^{\alpha-1}$$

  where $0 < \alpha, \beta < 1$ and $\Gamma(\alpha) = \int_0^{\infty} e^{-x}x^{\alpha-1}dx$
- For Γ function
  - $\Gamma(\alpha) = \alpha\Gamma(\alpha - 1)$ for any positive α. For large values of α $\Gamma(\alpha) \approx \sqrt{2\pi}e^{-\alpha}\alpha^{\alpha-0.5}$ (Stirling's approximation)
  - $\Gamma(n) = (n-1)!$ for any positive integer n. For large values of n $n! \approx \sqrt{2\pi}e^{-n-1}n^{n+0.5}$
  - $\Gamma(1/2) = \sqrt{\pi}$

Exponential Distribution

- Continuous distribution
- Single parameter β
- $\mathrm{Exp}(\beta)$

  $$f(x) = \frac{1}{\beta} \exp\left\{-\frac{x}{\beta}\right\}$$

  where $0 < x, \beta < \infty$
- This is a specific case of Gamma function with $\alpha = 1$
- If $\beta = 1$ this distribution is called standard exponential distribution
- cdf

  $$F(X) = \begin{cases} 0 & \text{if } -\infty < 0 \leq 0 \\ 1 - e^{-x/\beta} & \text{if } x > 0 \end{cases}$$

Chi-square Distribution

- Continuous distribution
- Single parameter υ (this is also called degrees of freedom)
- Chi(υ) is chi-square distribution with υ degrees of freedom

$$f(x) = Gamma(\upsilon/2, 2)$$

Lognormal Distribution

- Continuous distribution
- Two parameters (μ, σ)
- pdf

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

  with $0 < x < \infty, -\infty < \mu < \infty$, and $0 < \sigma < \infty$
- cdf

$$F(X) = \Phi\left(\frac{\log(x) - \mu}{\sigma}\right)$$

  where Φ is the cdf of standard normal distribution
- $$P(\log(X) \leq x) = P(X \leq e^x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$
- Logarithm of the rv X has $\mathcal{N}(\mu, \sigma)$ distribution. Hence, it is called lognormal distribution.

Student's t Distribution

- Continuous distribution
- Single parameter υ (degrees of freedom)
- pdf

$$f(x) = a(\upsilon)\left(1 + \frac{1}{\upsilon}x^2\right)^{-\frac{1}{2}(\upsilon+1)}$$

  where

$$a(\upsilon) = \frac{1}{\sqrt{\upsilon\pi}\frac{\Gamma(0.5(\upsilon+1)}{\Gamma(0.5\upsilon)}}$$

  for integer values of υ
- Distribution is symmetric around $x = 0$
- Discovered by W.S. Goset under pseudonym "Student" in 1908

Cauchy Distribution

- Continuous distribution
- Specific case of Student's t distribution with $\upsilon = 1$
- pdf
$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$
- cdf
$$F(x) = \frac{1}{\pi} arctan(x) + \frac{1}{2}$$

for $x \in \mathbb{R}$

F Distribution

- Continuous distribution
- Two parameters (υ_1, υ_2)
- Order of parameters are important. Hence $f_{\upsilon_1,\upsilon_2}(x) \neq f_{\upsilon_2,\upsilon_1}(x)$
- pdf

$$f(x) = k(\upsilon_1, \upsilon_2)x^{0.5(\upsilon_1 - 2)} \left(1 + \frac{\upsilon_1}{\upsilon_2}x\right)^{-0.5(\upsilon_1+\upsilon_2)}$$

where

$$k(\upsilon_1, \upsilon_2) = \frac{\upsilon_1}{\upsilon_2}^{0.5\upsilon_1} \frac{\Gamma(0.5(\upsilon_1 + \upsilon_2))}{\Gamma(0.5\upsilon_1)\Gamma(0.5\upsilon_2)}$$

Beta Distribution

- Beta function

$$
\begin{aligned}
b(\alpha, \beta) &= \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}
\end{aligned}
$$

  where α, β are positive real numbers.
- Beta distribution
- Continuous distribution
- Two parameters (α, β)
- Beta(α, β)

$$
f(x) = \frac{1}{b(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}
$$

- Beta(1,1) is equal to uniform distribution with range [0,1]

Negative Exponential Distribution

- Continuous distribution
- Two parameters (γ, β)
- pdf
$$f(x) = \frac{1}{\beta} \exp\left\{-\frac{(x-\gamma)}{\beta}\right\}$$
where $0 < \beta < \infty$, $-\infty < \gamma < \infty$ and $\gamma < x < \infty$

Weibull Distribution

- Continuous distribution
- Two parameters (α, β)
- pdf

$$f(x) = \alpha\beta^{-\alpha}x^{-\alpha-1}\exp\left\{(\frac{x}{\beta})^{\alpha}\right\}$$

where $0 < \alpha, \beta$ and $0 < x < \infty$

Rayleigh Distribution

- Continuous distribution
- Single parameter θ
- pdf

$$f(x) = \frac{2}{\theta} x \exp\left\{ -\frac{x^2}{\theta} \right\}$$

where $0 < x < \infty$ and $0 < \theta$

Laplace Distribution

- Continuous distribution
- Two parameters (μ, σ)
- pdf

$$f(x) = \frac{1}{2\sigma} \exp\left\{ -\frac{|x-\mu|}{\sigma} \right\}$$

Moments

- r^{th} moment of a rv X is

$$n_r = E(X^r)$$

- For $r = 1$, the first moment of an rv is its mean

Central Moments

- r^{th} central moment of a rv X is

$$c_r = E((X - \mu)^r)$$

- $c_1 = 0$
- $c_2 = \sigma^2$
- If n_r is finite then
  - c_r is finite
  - n_s is finite for $s \in \{1, 2, ..., r - 1\}$
  - c_s is finite for $s \in \{1, 2, ..., r - 1\}$

Moment Generating Function

$$M_X(t) = E(e^{(tX)})$$

- Why do we need it?
  Remember
  $$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \dots$$
- Hence the moments of an rv X can be computed from its moment generating function
  $$n_r = \left.\frac{d^r}{dt^r} M_x(t)\right|_{t=0}$$

Moment Generating Function of Distributions

The moment generating function of some of the distributions that we have covered

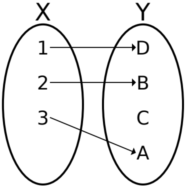
- Bernoulli: $M_X(t) = ((1-p) + pe^t)^n$
- Poisson: $M_X(t) = e^{(-\lambda + \lambda e^t)}$
- Normal: $M_X(t) = e^{(t\mu + 0.5t^2\mu^2)}$
- Gamma: $M_X(t) = (1 - \beta t)^{-\lambda}$ for all $t < 1/\beta$
- Exponential: $M_X(t) = (1 - \beta t)^{-1}$
- Chi-square: $M_X(t) = (1 - 2\upsilon)^{-0.5\upsilon}$ for $t < 0.5$

Functions of Random Variables

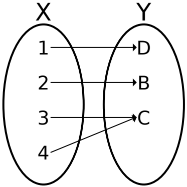
- Functions of an rv are also rv
- Consider a function that maps an rv X into another rv Y ie. $Y = g(X)$
- If the pdf of X is given as $f_X(x)$, how do we find the pdf of Y, $f_Y(y)$? How are they related?

Functions of Discrete Random Variables

- The function $g(.)$ may or may not be a one-to-one function.



One-to-one function (injection)



Not one-to-one (surjection)

- If function $g(.)$ is a one-to-one function there will be a single y_i for each x_i value. Hence: $P(Y = y_i) = P(X = x_i)$ for $y_i = g(x_i)$. Then
$$f(y_i) = f(x_i)$$
for $y_i = g(x_i)$

Functions of Discrete Random Variables

- If function $g(.)$ is not a one-to-one function Hence there may be many x_i that maps to the same y_i such as $y_i = g(x_i)$.
- In this case

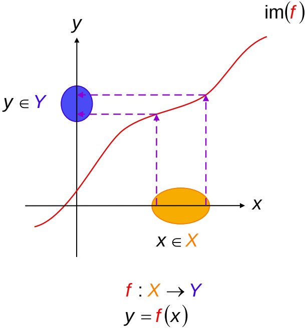
$$P(Y = y_i) = \sum_{j, y_i = g(x_j)} P(X = x_j)$$

and

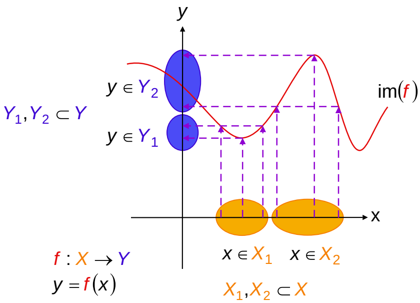
$$f(y_i) = \sum_{j, y_i = g(x_j)} f(x_j)$$

Functions of Continuous Random Variables

- Continuous functions can also be one-to-one or not.



One-to-one



Not one-to-one

- To find $f_Y(y)$, first consider

$$P(y \leq Y \leq y + dy) = f_Y(y)dy = F_Y(y + dy) - F_Y(y)$$

Functions of Continuous Random Variables

- If the function is one-to-one then

$$P(y \leq Y \leq y + dy) = P(x \leq X \leq x + dx)$$

which is

$$f_Y(y)dy = f_X(x)dx$$

- If the function is not one-to-one and there are N different values of x that maps to the same y value: $y = g(x_i)$ $i = 1, 2..N$. In this case

$$\begin{aligned} P(y \leq Y \leq y + dy) = {} & P(x_1 \leq X \leq x_1 + dx) \\ & + P(x_2 \leq X \leq x_2 + dx) + \cdots \\ & + P(x_N \leq X \leq x_N + dx) \end{aligned}$$

which is

$$f_Y(y)dy = f_X(x_1)dx + f_X(x_2)dx + \cdots + f_X(x_N)dx$$

Functions of Continuous Random Variables

From the graph $dy/dx = g'(x)$. Hence $dx = dy/g'(x)$

$$f_Y(y)dy = f_X(x_1)\frac{dy}{|g'(x_1)|} + f_X(x_2)\frac{dy}{|g'(x_2)|} + \cdots + f_X(x_N)\frac{dy}{|g'(x_N)|}$$

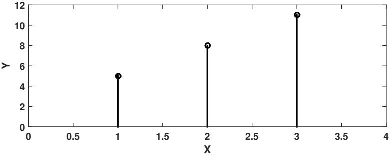
- Absolute value is taken to avoid negativity in pdf

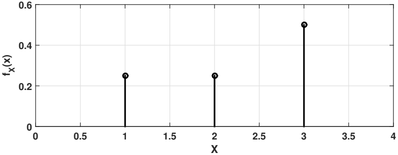
Finally:

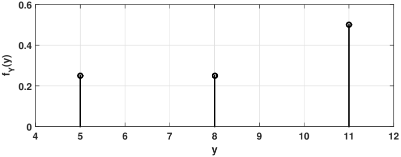
$$f_Y(y) = \sum_{i, y=g(x_i)} \frac{f_X(x_i)}{|g(x_i)|}$$

Example for Function of Discrete RV – 1

- One-to-one discrete rv $Y = 3X + 2$ and $\mathcal{X} = \{1, 2, 3\}$.







Example for Function of Continuous RV – 1

- Consider a linear transformation $Y = g(X) = aX + b$ where a and b are constant real numbers.
- For this function $\frac{d}{dX}(aX + b) = a$ and $x = \frac{y-b}{a}$. Hence

$$f_Y(y) = \frac{f_X(x)}{|a|} = \frac{f_X(\frac{y-b}{a})}{|a|}$$

Linear transformation of rv

Linear transformation of a random variable does not change the type of distribution (ie. uniform, Gaussian etc.). It may change the parameters such as mean, variance etc.

- If X has uniform distribution in $[x_1, x_2]$ range, then Y has also uniform distribution in $[ax_1 + b, ax_2 + b]$ range.

Example for Function of Continuous RV – 2

- Consider $Y = g(X) = 1/X$. Find $f_Y(y)$ in terms of $f_X(x)$.
- This is a one-to-on function with a single value of X such that $X = 1/Y$.

$$
\begin{aligned}
g'(X) &= -\frac{1}{X^2} \\
&= -\frac{1}{(1/Y)^2} = -Y^2
\end{aligned}
$$

- Hence

$$
f_Y(y) = \frac{1}{Y^2} f_X\left(\frac{1}{Y}\right)
$$

Example for Function of Continuous RV – 3

- Consider $Y = g(X) = aX^2$ where $a > 0 \in \mathbb{R}$ is a constant. Find $f_Y(y)$ in terms of $f_X(x)$.
- This is not a one-to-on function
  - For $y < 0$ there is no x
  - For $y > 0$ there are two values of X: $x_1 = \sqrt{y/a}$ and $x_1 = -\sqrt{y/a}$

  that satisfies $Y = g(X) = aX^2$.
- $g'(X) = 2aX$ and $1/|g'(X)| = 1/(2aX) = 1/(2a\sqrt{y/}) = 1/(2\sqrt{ay})$ for both $x_1 = \sqrt{y/a}$ and $x_1 = -\sqrt{y/a}$ when $y > 0$
- Then

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{ay}}(f_X(\sqrt{\frac{y}{a}}) + f_X(-\sqrt{\frac{y}{a}})) & \text{if } y > 0 \\ 0 & \text{if } y < 0 \end{cases}$$

Example for Function of Continuous RV – 4

- Consider $Y = g(X) = e^X$. Find $f_Y(y)$ in terms of $f_X(x)$.
- This is not a one-to-on function with a single solution at $x = \ln y$
-

$$\begin{aligned} \frac{1}{|g'(x)|} &= \frac{1}{e^x} \\ &= \frac{1}{e^{\ln y}} \\ &= \frac{1}{y} \end{aligned}$$

- Note that $y > 0$ for all values of x
- Then

$$f_Y(y) = \begin{cases} \frac{1}{y} f_X(\ln y) & \text{if } y > 0 \\ 0 & \text{if } y \leq 0 \end{cases}$$

PDF Conversion

- We have seen how to find the pdf of a function of a rv.
- Now a different problem: How can we convert a rv X with $f_X(x)$ to another rv Y with $f_Y(y)$ using a function $y = g(x)$?
- We will use 2 steps
  1. Convert X into another temporary rv Z that has uniform distribution in $[0, 1]$
  2. Convert Z into Y