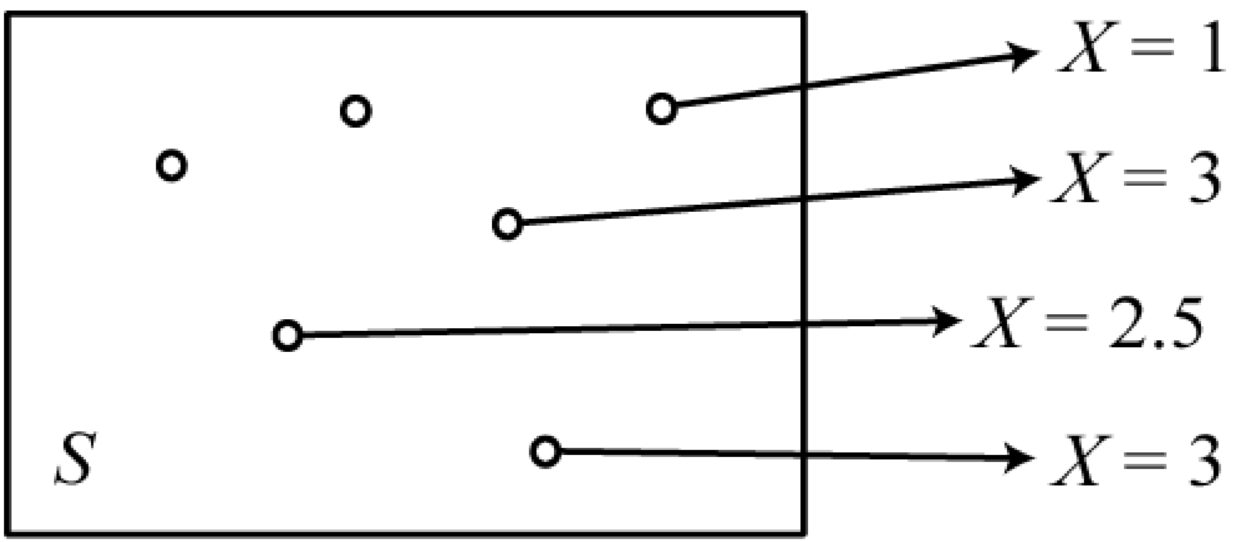
SDSC2102
Statistical Methods and Data Analysis

Topic 1. Basic Probability and Statistics Theory

Random Variables and Probability Distributions

Random Variables

- X is a mapping from each simple event in the sample space S to a real number
  - Discrete: finite or countable set of values
  - Continuous: uncountable set of values



Discrete Random Variables

- **Discrete random variable:** only takes on certain values in a finite or countable set.
  - Flip two coins: X = # of heads

    HH $\leftrightarrow$ $X = 2$
    HT $\leftrightarrow$ $X = 1$
    TH $\leftrightarrow$ $X = 1$
    TT $\leftrightarrow$ $X = 0$
  - Roll a die: X = # of dots on the side facing up
  - Color: X = # of students in a class whose favorite color is red
  - Post office: X = # of people in line

Continuous Random Variables

- **Continuous random variable**: takes on any values within some range (i.e., infinitely many values in an uncountable set).
  - X = the lifetime (hours) of a light bulb
  - X = the weight of the next package that you take to the post office
  - X = the length of time to play 18 holes of golf
  - X = the annual income of Texas residents

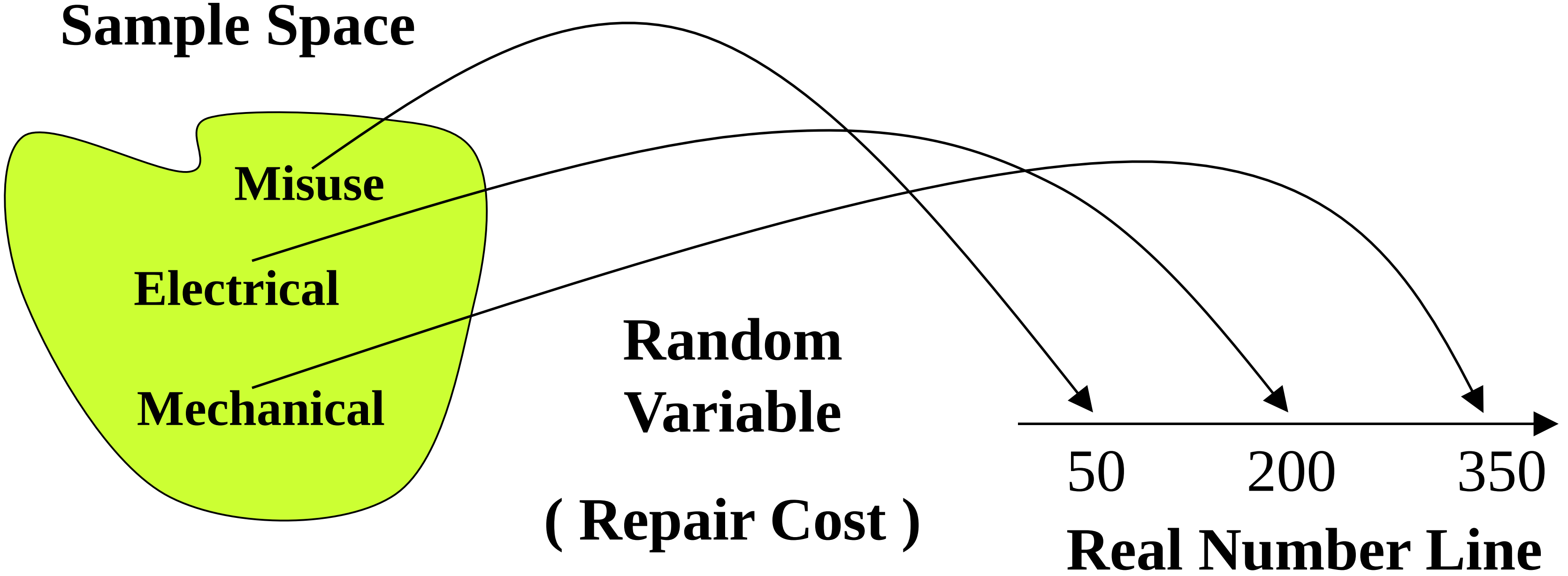
Example

- Machine Breakdown Example

$$S = \{\text{electrical, mechanical, misuse}\}$$

- Let X = the repair cost associated with a failure

Sample Space



Random Variable and Events

- Flip two coins: X = # of heads
  - Exactly one head: $[X = 1]$
  - No more than one head: $[X \leq 1]$
- Post office: X = # of people in line
  - At least 3 and fewer than 10 people: $[3 \leq X < 10]$
  - More than 3 and at most 10 people: $[3 < X \leq 10]$
- Light bulb: X = lifetime of the bulb
  - More than 50 hours: $[X > 50]$
  - Never turns on: $[X = 0]$

Class Problems on Random Variables

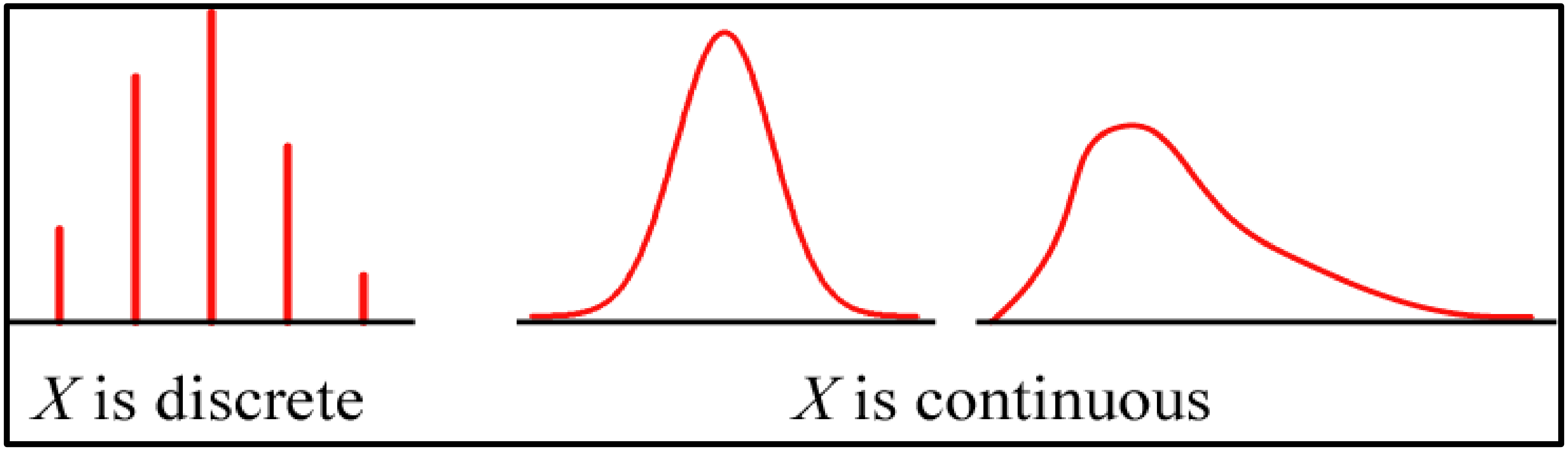
1. A company has 7 machines on its shop-floor of which 4 are lathes. The service personnel pick two machines at random and check to see if they have any maintenance problems. Let X be the number of lathes selected. Find the probability mass function (or probability distribution) of X.

$$P(X=0)=\frac{\binom{3}{2}}{\binom{7}{2}}=\frac{6}{42} \quad P(X=1)=\frac{\binom{4}{1}\binom{3}{1}}{\binom{7}{2}}=\frac{24}{42}$$

$$P(X=2)=\frac{\binom{4}{2}}{\binom{7}{2}}=\frac{12}{42}$$

Probability Distribution

- The probabilities assigned to the values of X
- A discrete r.v. X has **probability mass function (p.m.f.)**: $p(x) = P[X = x]$
- A continuous r.v. X has **probability density function (p.d.f.)**: $f(x)$



Probability Mass Function

- $p(x)$ maps the possible values of the discrete r.v. X to probabilities on the interval [0,1].
- $p(x) = P[X = x]$ = probability that $X = x$, where x is from a finite or countable set.
- p.m.f. values are probabilities.
- Properties of the p.m.f.:

  1) $0 \le p(x) \le 1$ for all x

  2) $\sum_x p(x) = 1$

Probability Mass Function

- Example: Flip two coins
  - $S = \{ \text{HH, HT, TH, TT} \}$
  - $X = \#$ of heads

  $p(0) = P[X = 0] = P(\text{TT}) = 1/4$

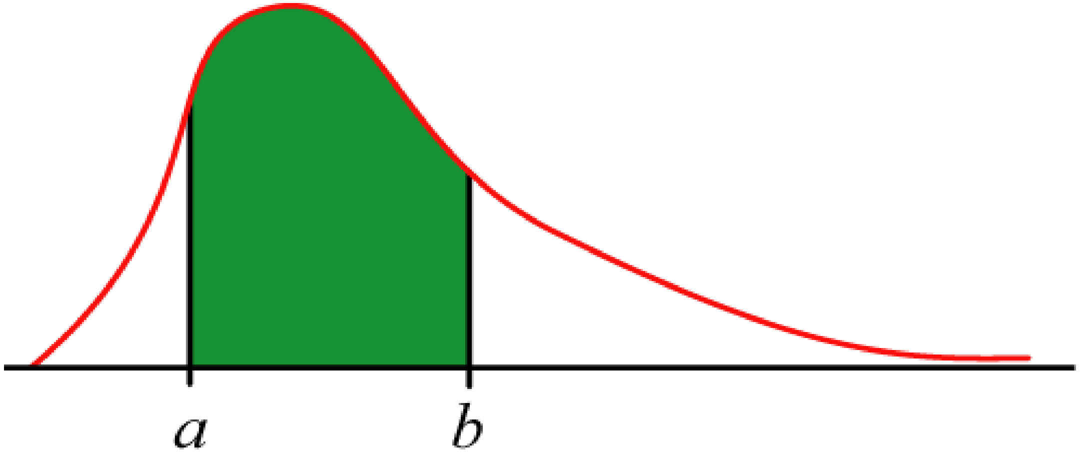
  $p(1) = P[X = 1] = P(\text{HT or TH}) = 1/4 + 1/4 = 1/2$

  $p(2) = P[X = 2] = P(\text{HH}) = 1/4$
- $0 < p(x) < 1$ for $x = 0, 1, 2$; o/w $p(x) = 0$
- $\sum_x p(x) = p(0) + p(1) + p(2) = 1$

Probability Density Function

- $f(x)$ describes the distribution of values of X over a continuous range.
- p.d.f. values are NOT probabilities.
- Probabilities are found by calculating the area under the $f(x)$ curve.

$$P(a < X < b)$$
$$= \int_a^b f(x)dx$$



Probability Density Function

- Properties of the p.d.f.:

  1) $f(x) \geq 0$ for all x

  2) $\int_{-\infty}^{\infty} f(x)dx = 1$

- The probability of a single point is zero

$$P(X = a) = \int_{a}^{a} f(x)dx = 0$$

- If X is a continuous r.v., then for any a and b,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

Probability Density Function

- X = lifetime (hrs) of a certain kind of radio tube

$$f(x) = \begin{cases} \dfrac{100}{x^2}, & x > 100 \\ 0, & x \le 100 \end{cases}$$

$$P(X < 150) = \int_{-\infty}^{150} f(x)dx = 100\int_{100}^{150} x^{-2}dx$$

$$= 100\left(\frac{1}{100} - \frac{1}{150}\right) = \frac{1}{3}$$

- Note:

$$\int_{100}^{\infty} f(x)dx = 1$$

Cumulative Distribution Function

- $F(x) = P[X \leq x]$
  - Discrete r.v. X:

$$F(x) = \sum_{t \leq x} p(t) \qquad \text{for } -\infty < x < \infty$$

  - Continuous r.v. X:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \qquad \text{for } -\infty < x < \infty$$

$$f(x) = \frac{dF(x)}{dx}$$

- c.d.f. values are probabilities

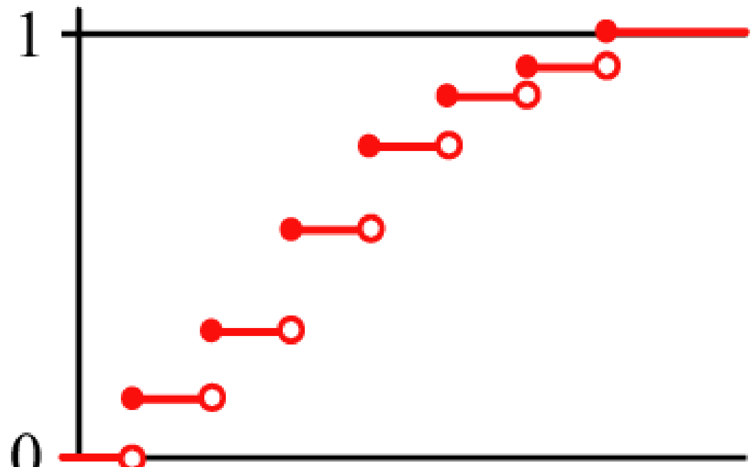
Cumulative Distribution Function

- Properties of the c.d.f.:

  1) $0 \le F(x) \le 1$ for all x

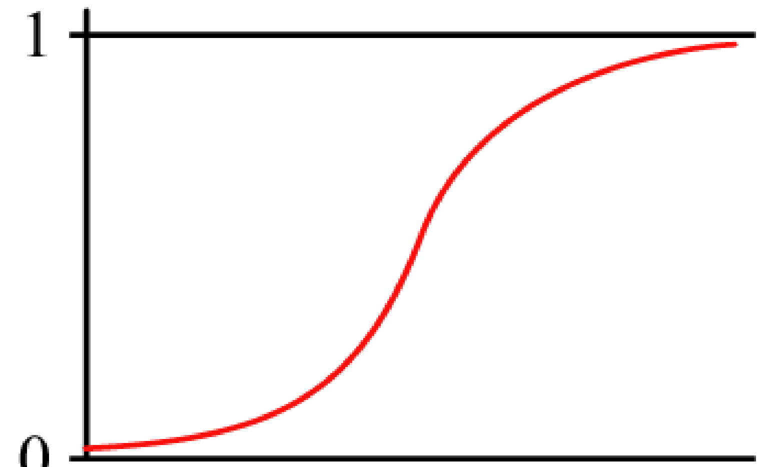
  2) If $x \le y$, then $F(x) \le F(y)$

  3) $F(-\infty) = 0$, $F(\infty) = 1$



X is discrete

Right-continuous



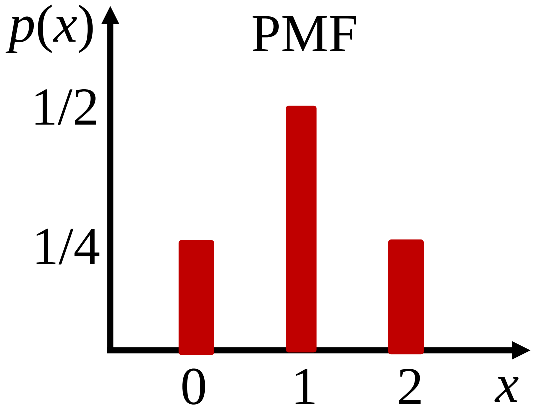
X is continuous

Continuous

Cumulative Distribution Function

- Discrete Example: Flip 2 coins
  - $F(0) = P[X \leq 0] = p(0) = 1/4$
  - $F(1) = P[X \leq 1] = p(0) + p(1) = 1/4 + 1/2 = 3/4$
  - $F(2) = P[X \leq 2] = p(0) + p(1) + p(2) = 1$

$$\Rightarrow \quad F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{1}{4}, & 0 \leq x < 1 \\ \dfrac{3}{4}, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$



Cumulative Distribution Function

➢ Continuous Example: Lifetime of radio tube

- For $x > 100$

$$F(x) = \int_{-\infty}^{x} f(t)dt = 100 \int_{100}^{x} t^{-2}dt = 1 - \frac{100}{x}$$

$$\Rightarrow \quad F(x) = \begin{cases} 0, & x \le 100 \\ 1 - \dfrac{100}{x}, & x > 100 \end{cases}$$

Class Problems on Random Variables

2. The probability mass function, $p(x)$, of a random variable X is

$$p(x) = \begin{cases} \frac{1}{6} & \text{if } x = 0 \\ \frac{1}{3} & \text{if } x = 2 \\ \frac{1}{4} & \text{if } x = 3 \\ \frac{1}{4} & \text{if } x = 4 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the following probabilities

(a) $P\{X \geq 3\}$

(b) $P\{X \leq 2\}$

(c) $P\{X > 4\}$

Class Problems on Random Variables

$$(a)\ P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$$
$$= \frac{1}{4} + \frac{1}{4} + 0 = \frac{1}{2}$$

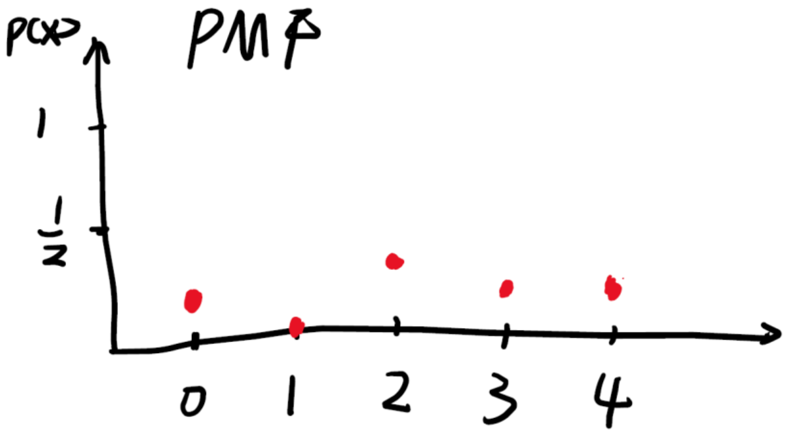
$$(b)\ P(X \leq 2) = P(X = 0) + P(X = 2)$$
$$= \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

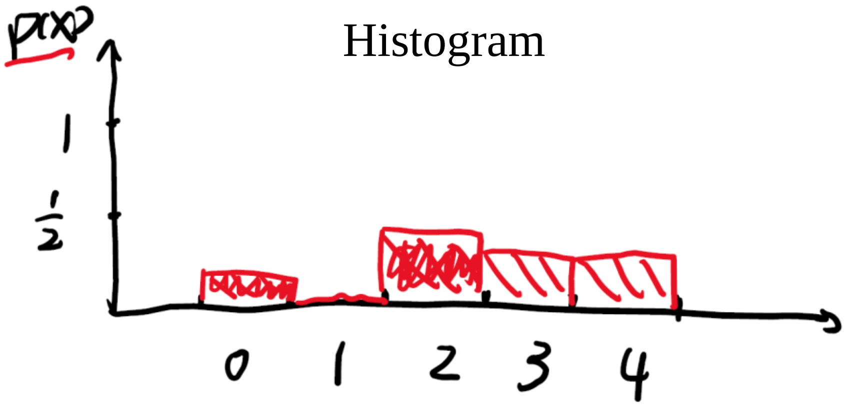
$$(c)\ P(X > 4) = 0$$

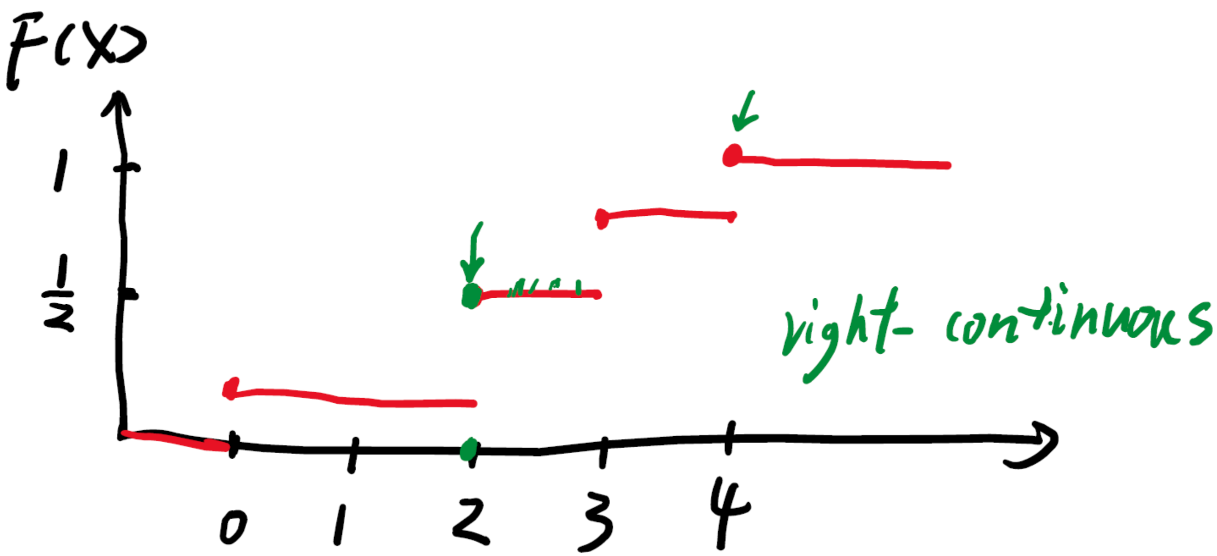
Class Problems on Random Variables

$$p(x) = \begin{cases} \frac{1}{6} & \text{if } x = 0 \\ \frac{1}{3} & \text{if } x = 2 \\ \frac{1}{4} & \text{if } x = 3 \\ \frac{1}{4} & \text{if } x = 4 \\ 0 & \text{otherwise} \end{cases} \Rightarrow F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{1}{6}, & 0 \leq x < 2 \\ \dfrac{1}{2}, & 2 \leq x < 3 \\ \dfrac{3}{4}, & 3 \leq x < 4 \\ 1, & x \geq 4 \end{cases}$$

Class Problems on Random Variables







Class Problems on Random Variables

3. Let the probability density function of a continuous random variable, X be given by

$$f(x) = \begin{cases} c(4x - 2x^2), & 0 < x < 2 \\ 0, & otherwise. \end{cases}$$

(a) What is the value of c?
(b) What is the cumulative distribution function of X?
(c) $P\{\frac{1}{2} < X < \frac{3}{2}\} = ?$

$$(a) \int_{-\infty}^{\infty} f(x)dx = 1$$

$$\Rightarrow \int_0^2 c(4x - 2x^2)dx = c\left(2x^2 - \frac{2}{3}x^3\right)\Big|_0^2 = 1$$

$$\Rightarrow c\left(8 - \frac{16}{3}\right) = 1 \Rightarrow c = \frac{3}{8}$$

Class Problems on Random Variables

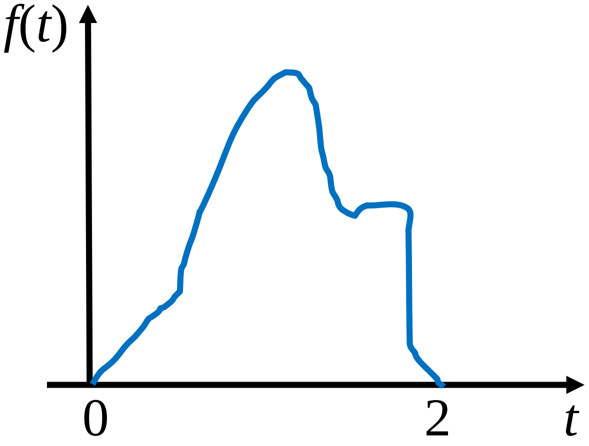
(b) $For\ x \le 0,$ $f(x) = 0, so\ F(x) = 0$

$For\ 0 < x < 2,$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} \frac{3}{8}(4t - 2t^2)dt$$

$$= \frac{3}{8}\left(2x^2 - \frac{2}{3}x^3\right)$$

$For\ x \ge 2,$ $F(x) = 1$



$$\Rightarrow F(x) = \begin{cases} 0, & x \le 0 \\ \frac{3}{8}\left(2x^2 - \frac{2}{3}x^3\right), & 0 < x < 2 \\ 1, & x \ge 2 \end{cases}$$

Class Problems on Random Variables

$$(c)\ P\left(\frac{1}{2} < X < \frac{3}{2}\right) = P(X < 1.5) - P(X \leq 0.5)$$

$$= P(X \leq 1.5) - P(X \leq 0.5)$$

$$= F(1.5) - F(0.5)$$

$$= \frac{3}{8}\left(2 \times 1.5^2 - \frac{2}{3} \times 1.5^3\right) - \frac{3}{8}\left(2 \times 0.5^2 - \frac{2}{3} \times 0.5^3\right)$$

$$= 0.6875$$

Class Problems on Random Variables

4. The cummulative distribution function of a continuous random variable, X, is given by

$$F(x) = \begin{cases} 1 - e^{-5x} & x \geq 0 \\ 0, & otherwise. \end{cases}$$

(a) What is $P\{X = 2\}$
(b) What is the probability density function $f(x)$?
(c) What is $P\{3 \leq X \leq 5\}$?

$$(a)\ P(X = 2) = 0$$

$$(b)\ f(x) = \frac{dF(x)}{dx} = 5e^{-5x}$$

$$(c)\ P(3 \leq X \leq 5) = F(5) - F(3)$$

$$= 1 - e^{-5\times 5} - (1 - e^{-5\times 3})$$

$$= e^{-15} - e^{-25}$$

Special Discrete Distributions

- Binomial distribution
- Geometric distribution
- Negative Binomial distribution
- Hypergeometric distribution
- Poisson distribution

Bernoulli Distribution

- Consider a r.v. X with exactly two possible outcomes
  - 0 = "failure" or 1 = "success"
- Define $p = P[\text{success}] = P[X = 1]$
  - The p.m.f. depends on the parameter p

$$P(X = x) = p^x(1 - p)^{1-x} \qquad \text{for } x = 0, 1$$

- Example: Toss a coin
  - 0 = "tail" or 1 = "head"
  - $p = 0.5$
  - p.m.f: $P(X = x) = 0.5^x(1 - 0.5)^{1-x} = 0.5$

Binomial Distribution

- One iteration of a Bernoulli experiment is called a Bernoulli trial
- Conduct *n* independent Bernoulli trials
- Let r.v. *X* be the # of successes in *n* trials
  - The p.m.f. depends on the parameters *n* and *p*

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0,1,2, \ldots, n$$

Class Problems on Discrete Random Variables

1. Historical data has revealed that a manufacturer's current yield is 85%. That is 15% of the parts they produce are defective.
   a. If 50 parts are produced, what is the probability that exactly 8 parts are defective?
   b. What is the probability that the third part produced is the first defective part?
   c. What is the probability that the fifth defective part is the fifteenth part produced?

1 = "defective", 0 = "not defective"

(a) Let X be the # of defective parts among the 50 parts
$X \sim$ Binomial (n=50, p=0.15)

$$P(X = 8) = \binom{50}{8} 0.15^8 0.85^{42}$$

Geometric Distribution

- Conduct independent Bernoulli trials until the first "success" occurs ($p = P$[success])
- Let r.v. X be the # of trials required
  - The p.m.f. depends on the parameter p

$$P(X = x) = p(1 - p)^{x-1} \qquad \text{for } x = 1,2, \ldots$$

- Example: Roll a die until we see a "6"
  - 1 = "6", 0 = "not a 6"
  - $p = P(6) = 1/6$

$$P(X = 2) = \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{2-1} = \frac{5}{36}$$

Class Problems on Discrete Distributions

1. Historical data has revealed that a manufacturer's current yield is 85%. That is 15% of the parts they produce are defective.
   a. If 50 parts are produced, what is the probability that exactly 8 parts are defective?
   b. What is the probability that the third part produced is the first defective part?
   c. What is the probability that the fifth defective part is the fifteenth part produced?

1 = "defective", 0 = "not defective"

$p = 0.15$

(b) Let X be the # of produced parts to have the first defective

$X \sim$ Geometric (p=0.15)

$$P(X = 3) = 0.15 * 0.85^2$$

Negative Binomial Distribution

- Conduct independent Bernoulli trials until *k* "successes" have occured ($p = P$[success])
- Let r.v. *X* be the # of trials required
  - The p.m.f. depends on the parameters *k* and *p*

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

for $x = k, k+1, k+2, \ldots$

- Ex: Roll a die until "6" has appeared 3 times

$$P(X = 5) = \binom{5-1}{3-1} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{5-3}$$

Class Problems on Discrete Random Variables

1. Historical data has revealed that a manufacturer's current yield is 85%. That is 15% of the parts they produce are defective.
   a. If 50 parts are produced, what is the probability that exactly 8 parts are defective?
   b. What is the probability that the third part produced is the first defective part?
   c. What is the probability that the fifth defective part is the fifteenth part produced?

1 = "defective", 0 = "not defective"

$p = 0.15$

(c) Let X be the # of produced parts to have 5 defective parts

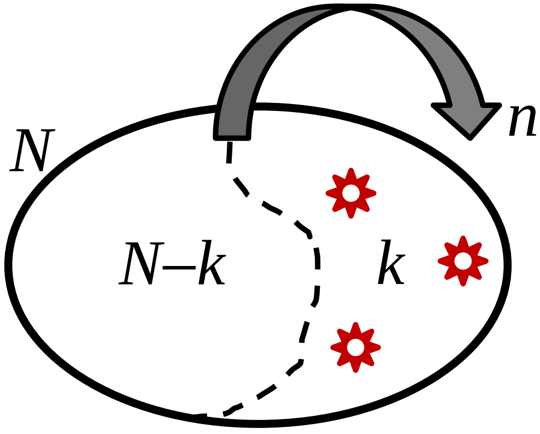
$X \sim$ Negative binomial ($k = 5$, $p = 0.15$)

$$P(X = 15) = \binom{14}{4} 0.15^5 * 0.85^{10}$$

Hypergeometric Distribution

- We have N items, of which k are "successes"
- Sample n items without replacement
- Let r.v. X be the # of successes in the sample
  - The p.m.f. depends on the parameters N, n, and k

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad \text{for } x = 0,1,2,\ldots,n$$



Hypergeometric Distribution

- Example: Batch of 100 parts has 10 defectives. Sample 5 parts without replacement.
  - $N = 100$, $n = 5$, and $k = 10$

$$P(X = 2) = \frac{\binom{10}{2}\binom{90}{3}}{\binom{100}{5}} = 0.0702$$

Class Problems on Discrete Distributions

2. An urn contains 10 balls of which 3 are white and 7 are blue. If 3 balls are chosen at random, what is the probability that exactly 2 are blue? Assume that balls are drawn randomly one after the other.
   a. Without replacement
   b. With replacement

1 = "blue ball", 0 = "not blue ball"

(a) Let X be the # of blue balls among the 3 selected balls

X ~ Hypergeometric (N=10, n= 3, k=7)

$$P(X=2)=\frac{\binom{7}{2}\binom{3}{1}}{\binom{10}{3}}$$

Class Problems on Discrete Distributions

2. An urn contains 10 balls of which 3 are white and 7 are blue. If 3 balls are chosen at random, what is the probability that exactly 2 are blue? Assume that balls are drawn randomly one after the other.
   a. Without replacement
   b. With replacement

1 = "blue ball", 0 = "not blue ball"

(b) Let X be the # of blue balls among the 3 selected balls

$X \sim$ Binomial (n=3, p=7/10=0.7)

$$P(X = 2) = \binom{3}{2} 0.7^2 0.3$$

Poisson Distribution

- Count the occurrences of a specific event over a specific time period or a specific region
  - # of customers entering a post office in one hour
  - # of misprints on a page (or group of pages)
  - # of car accidents in one month
- Define λ = rate of occurrence
- Let r.v. X be the # of outcomes occurring over the time period (or region)
  - The p.m.f depends on the parameter λ

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{for } x = 0,1,2,\ldots$$

Class Problems on Discrete Distributions

3. The number of cars X that arrive at a certain yield sign on a road during an interval t minutes long is according to the following probability mass function (PMF):

$$P(X = x) = e^{-8t}\frac{(8t)^x}{x!}$$

for $x = 0, 1, 2, 3, \ldots$

   a. What is the probability that exactly 10 cars arrive in a 1-minute interval?
   b. What is the probability that two or more cars arrive in 90 seconds?
   c. What is the probability that fewer than 3 cars arrive in 2 minutes?

(a) $$P(X = 10) = e^{-8\times 1}\frac{8^{10}}{10!}$$

Class Problems on Discrete Distributions

(b) $P(X \geq 2) = 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)]$

$$= 1 - \left(e^{-12}\frac{12^0}{0!} + e^{-12}\frac{12^1}{1!}\right) = 1 - 13e^{-12}$$

(c) $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$

$$= e^{-16}\frac{16^0}{0!} + e^{-16}\frac{16^1}{1!} + e^{-16}\frac{16^2}{2!}$$

$$= e^{-16}(1 + 16 + 256/2)$$