Biostatistics
Week V

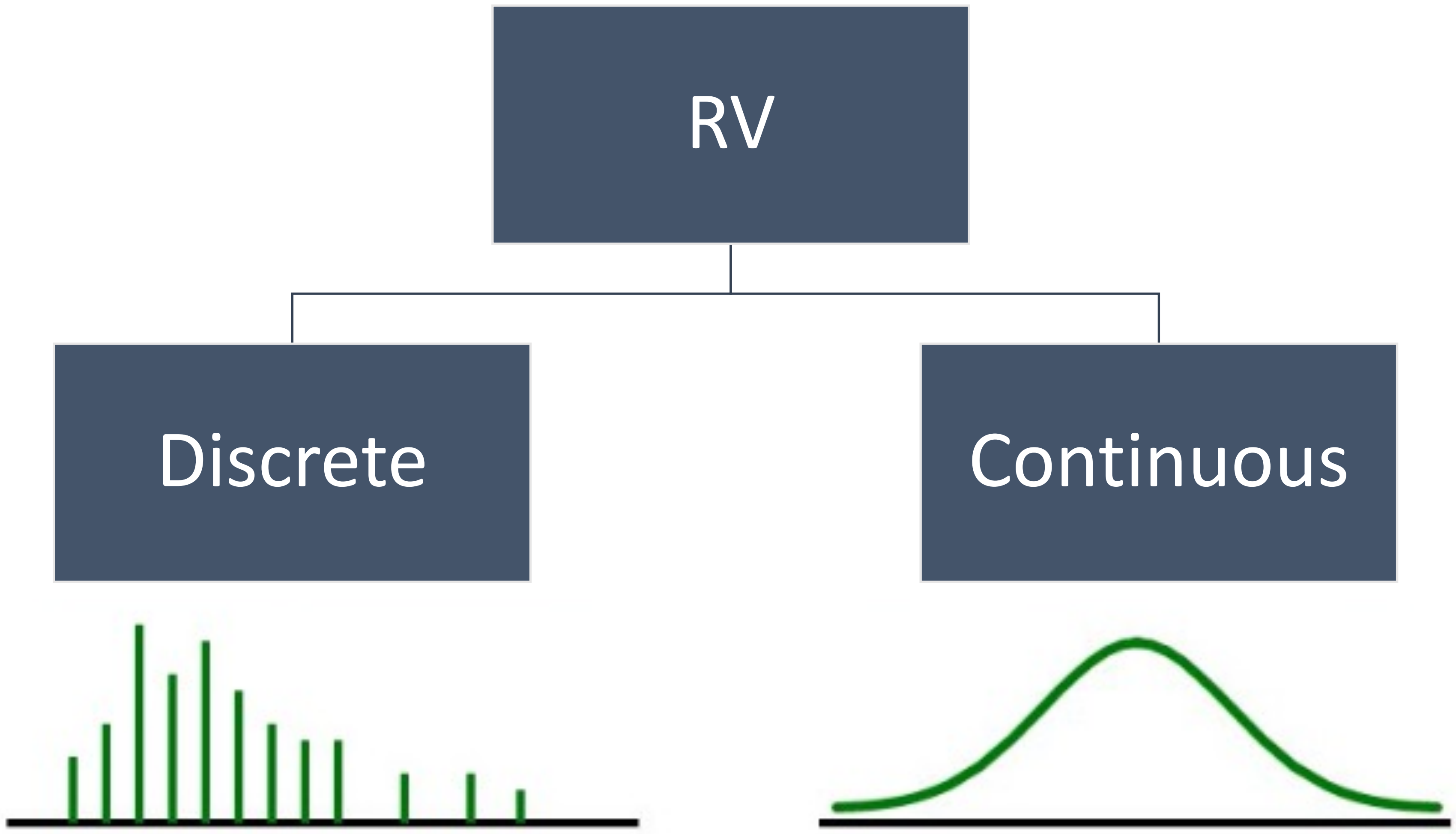
Ege Ülgen, MD, PhD

3 November 2022

Random Variable

- A random variable (RV) is a variable whose possible values are **numerical outcomes of a random phenomenon**

- There are two types of random variables:
  - ***Discrete*** – flipping a coin, rolling a die, number of pancreatic cancer cases in a year ...
  - **Continuous** – systolic blood pressures of hypertensive patients, progression-free survival time of glioblastoma patients, expression level of a certain gene …


RV
Discrete
Continuous

Probability Mass Function (PMF)

- Probability mass function is the probability distribution of a discrete random variable
- It provides the possible values and their associated probabilities

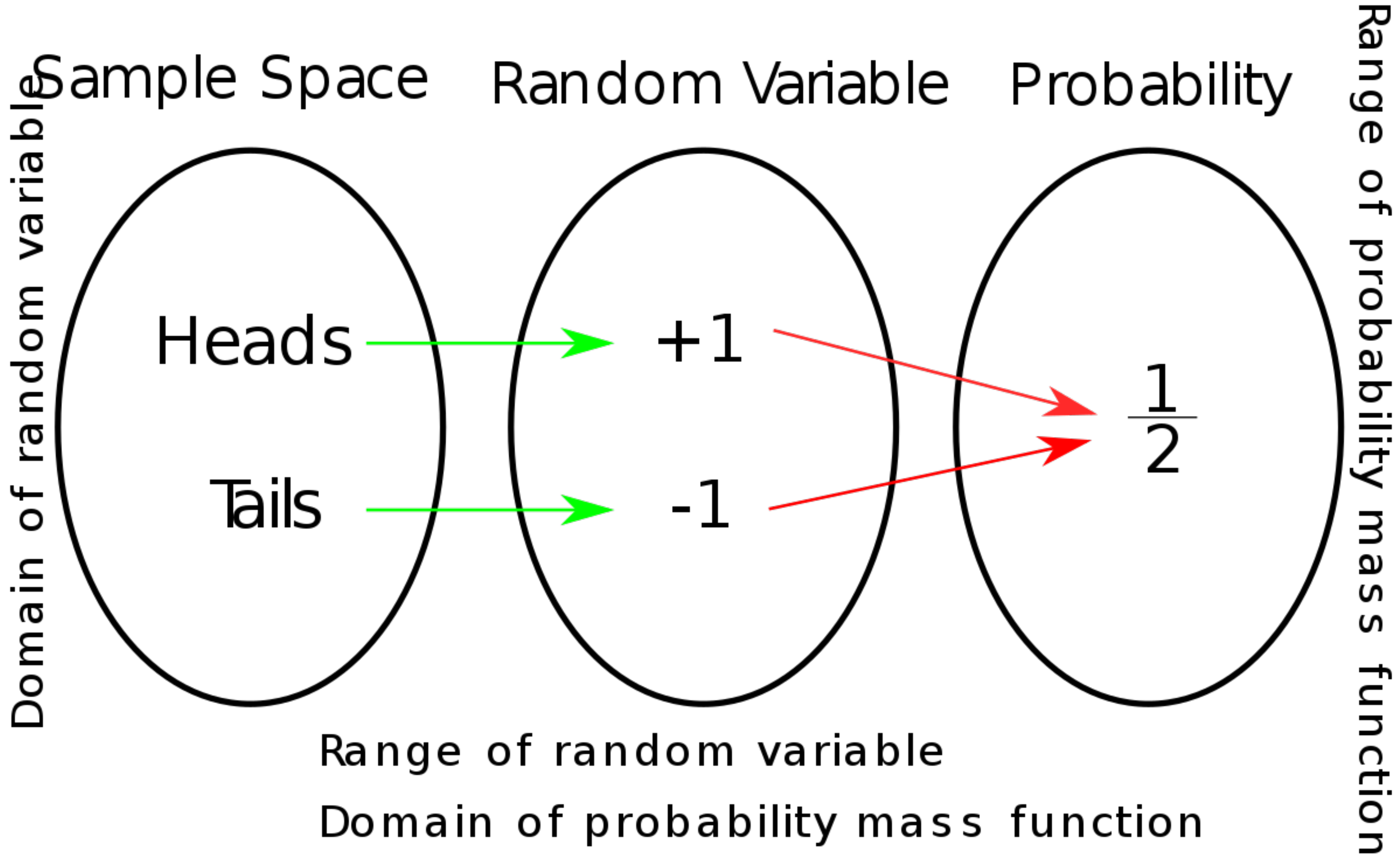
$$p: \mathbb{Z} \to [0,1]$$

$$p_X(x) = P(X = x)$$

Properties of a Proper PMF (p_X)

1. $p_X(x)$ is defined for all *x* over the given domain
2. $0 \leq p_X(x) \leq 1$
3. $\sum_x p_X(x) = 1$

https://en.wikipedia.org/wiki/Random_variable



Probability Density Function (PDF)

- Probability density function is the probability distribution of a continuous random variable
- It provides the possible values and their associated probabilities

$$f: \mathbb{R} \to [0,1]$$

$$f_X(x) = P(X = x)$$

Properties of a Proper PDF (f_X)

1. f_X is continuous over the given range
2. $0 \leq f_X(x) \leq 1$
3. $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Cumulative Density Function (CDF)

$$F_X(x) = P(X \leq x)$$

Survival Function

$$S(x) = P(X > x)$$

Expected Value

- The weighted average of all the possible values of a RV by the associated probabilities
- For discrete RVs:

$$E[X] = \sum_{i=1}^{n} P(X = x_i)x_i$$

- For continuous RVs:

$$E[X] = \int_{-\infty}^{\infty} f(x)x \, dx$$

Expected Value

- Expectation can be interpreted as the average outcome value over a large number of repetitions

- Properties:
  - $E[X + c] = E[X] + c$
  - $E[X * c] = E[X] * c$
  - $E[X + Y] = E[X] + E[Y]$
  - $E[X * Y] = E[X] E[Y]$ if X and Y are independent

Variance

- Expected squared distance of the RV values from the expected value

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Properties:

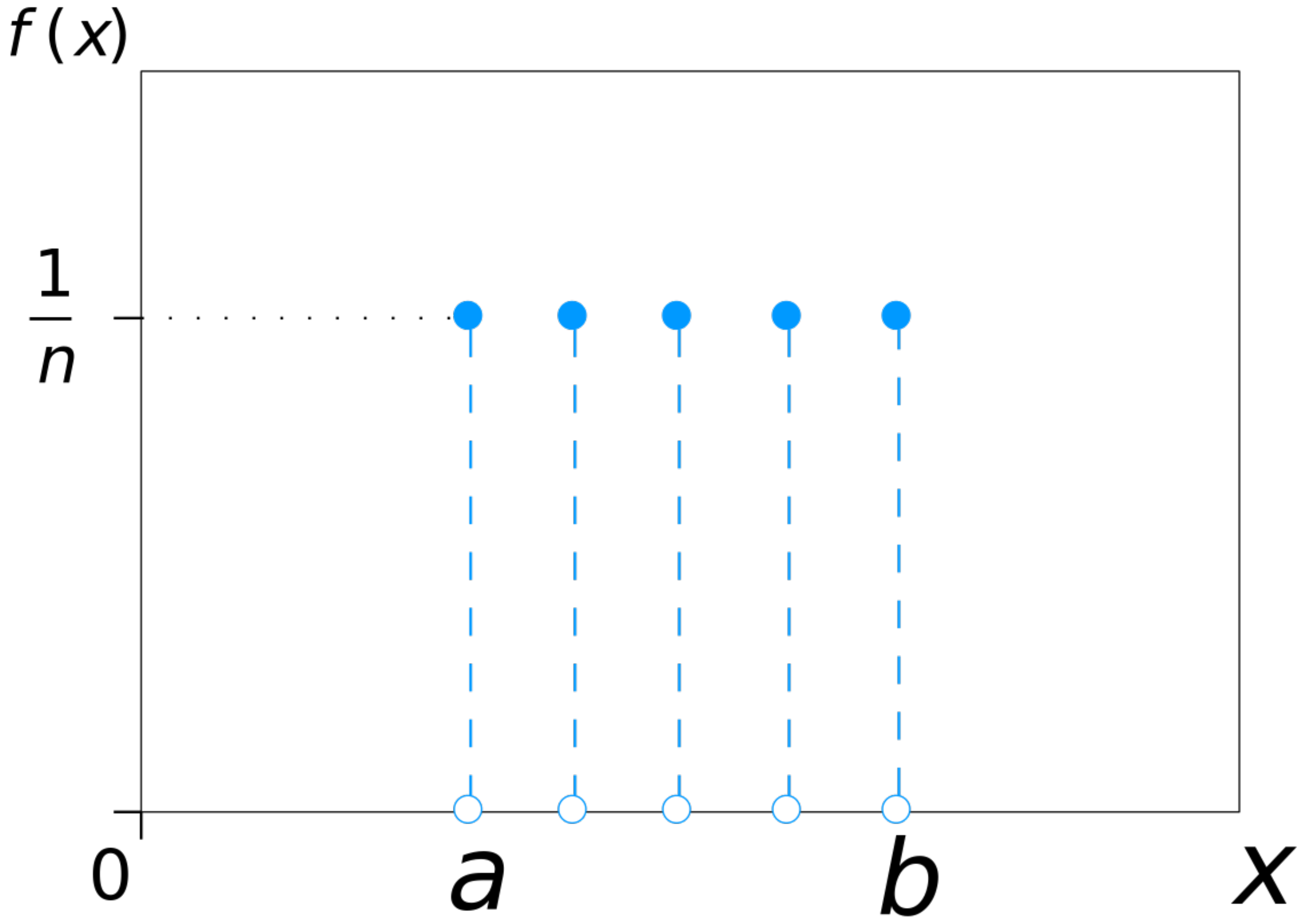
- $Var(X + c) = Var(X)$
- $Var(Xc) = Var(X)c^2$
- $Var(X + Y) = Var(X) + Var(Y)$ if X and Y are independent.

Commonly Used Discrete Distributions

- Discrete Uniform Distribution
- Bernoulli Distribution
- Binomial Distribution
- Geometric Distribution
- Hypergeometric Distribution
- Poisson Distribution

https://en.wikipedia.org/wiki/Discrete_uniform_distribution

Discrete Uniform Distribution



$a \leq k \leq b, \qquad n = b - a + 1$

PMF $\quad P(X = k) = \dfrac{1}{n}$

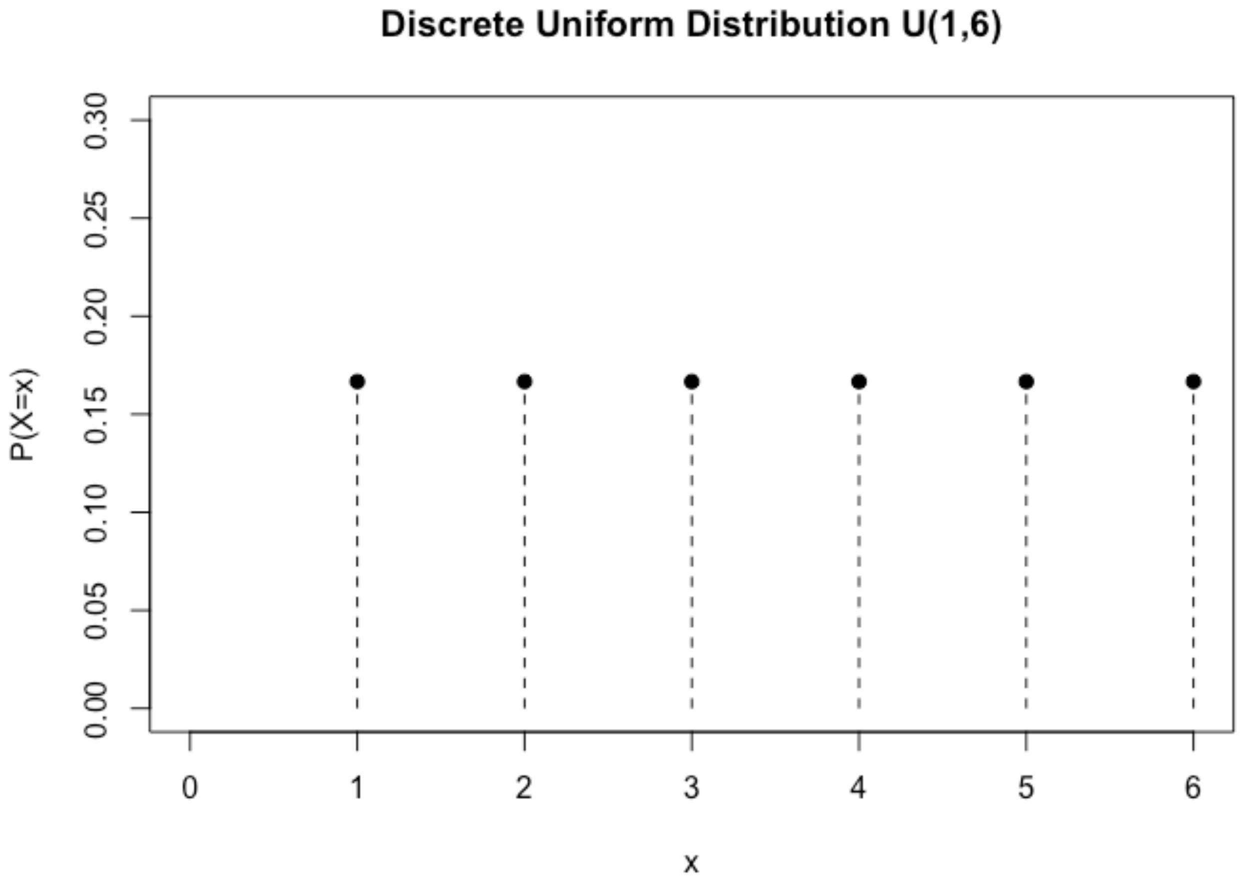
E[X] $\quad \dfrac{a+b}{2}$

Var(X) $\quad \dfrac{n^2 - 1}{12}$

CDF $\quad P(X \leq k) = \dfrac{k - a + 1}{n}$

Discrete Uniform Distribution

- Rolling a die



$1 \le k \le 6, \qquad n = 6$

PMF $\quad P(X = k) = \dfrac{1}{6}$

E[X] $\quad \dfrac{1+6}{2} = 3.5$

Var(X) $\quad \dfrac{6^2 - 1}{12} = \dfrac{35}{12} \approx 2.92$

Bernoulli Distribution

Let X be a random variable with possible values 0 and 1, and let $P(X = 1) = p$.

$$pmf = P(X = x) = \begin{cases} p^x(1-p)^{1-x} & x \in 0, 1 \\ 0 \quad otherwise \end{cases}$$

$$cdf = F_x(x) = P(X \leq x) = p^x(1-p)^{1-x}$$

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

Example: Flipping a fair (p = 0.5) coin

Binomial Distribution

- Used to describe the number of successes in *n* binary trials
- *n:* number of trials
- *p:* probability of success in one trial

$$X \sim B(n, p)$$

PMF $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, *k = 0, 1, 2, …, n*

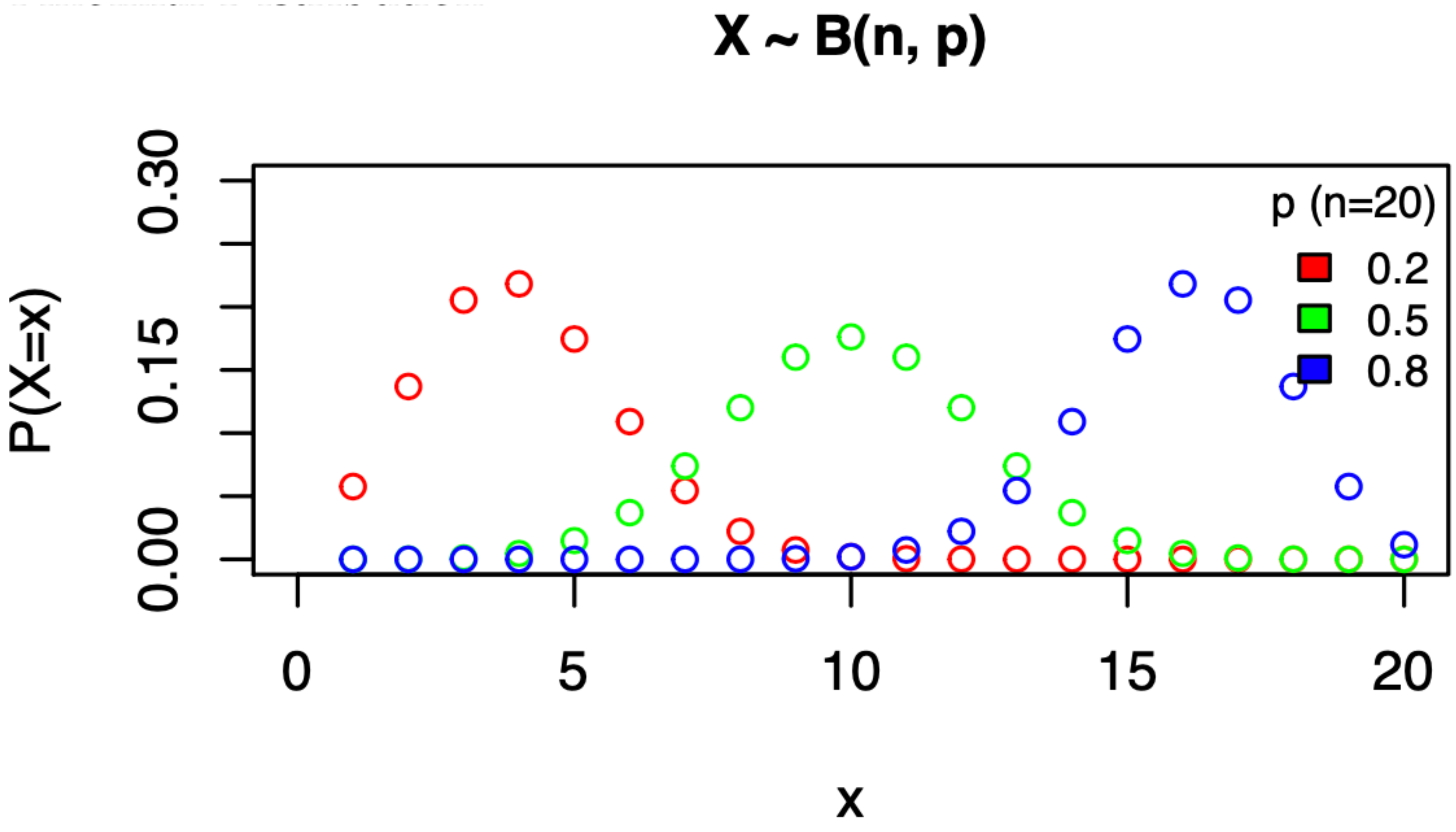
E[X] *np*

As $X = \sum_{i=1}^{n} Y_i \ where \ Y_i \sim Bernoulli(p)(iid)$

Var(X) *np(1-p)*

Binomial Distribution

- e.g., flipping a coin 20 times



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

Binomial Distribution Example

- A novel treatment has a success rate of 80%. Out of 10 patients who underwent the novel treatment:
  a) What is the probability that exactly 6 recovers?
  b) What is the probability that at least 9 recovers?
  c) What is the expected value and variance?

a) $P(X = 6) = \binom{10}{6} 0.8^6 (1 - 0.8)^{10-6} = 0.88$

b) $P(X \geq 9) = P(X = 9) + P(X = 10)$

$$= \binom{10}{9} 0.8^9 (1 - 0.8)^{10-9} + \binom{10}{10} 0.8^{10} (1 - 0.8)^{10-10}$$

$$= 0.2684 + 0.1073 = 0.3758$$

c) $E[X] = np = 10 \times 0.8 = 8$

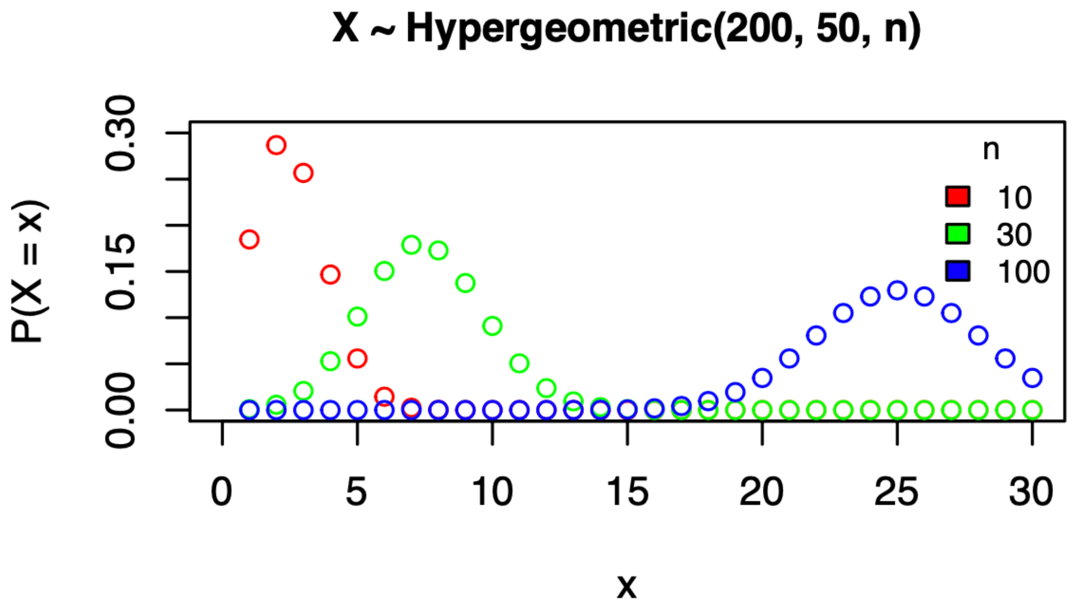
$np(1 - p) = 10 \times 0.8 \times 0.2 = 1.6$

Hypergeometric Distribution

- Describes the probability of *k* successes in *n* draws, **without replacement***, from a finite population of size *N* that contains exactly *K* objects with that feature

*Contrary to the binomial distribution which describes the probability of *k* successes in *n* draws **with replacement**

Hypergeometric Distribution



$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

$E[X] = np$ and $Var(X) = np(1-p)\binom{N-n}{N-1}$ where $p = K/N$

Example: Drawing n balls from an urn that contains 50 white (desired) and 150 red balls (the above plots) and getting x white balls

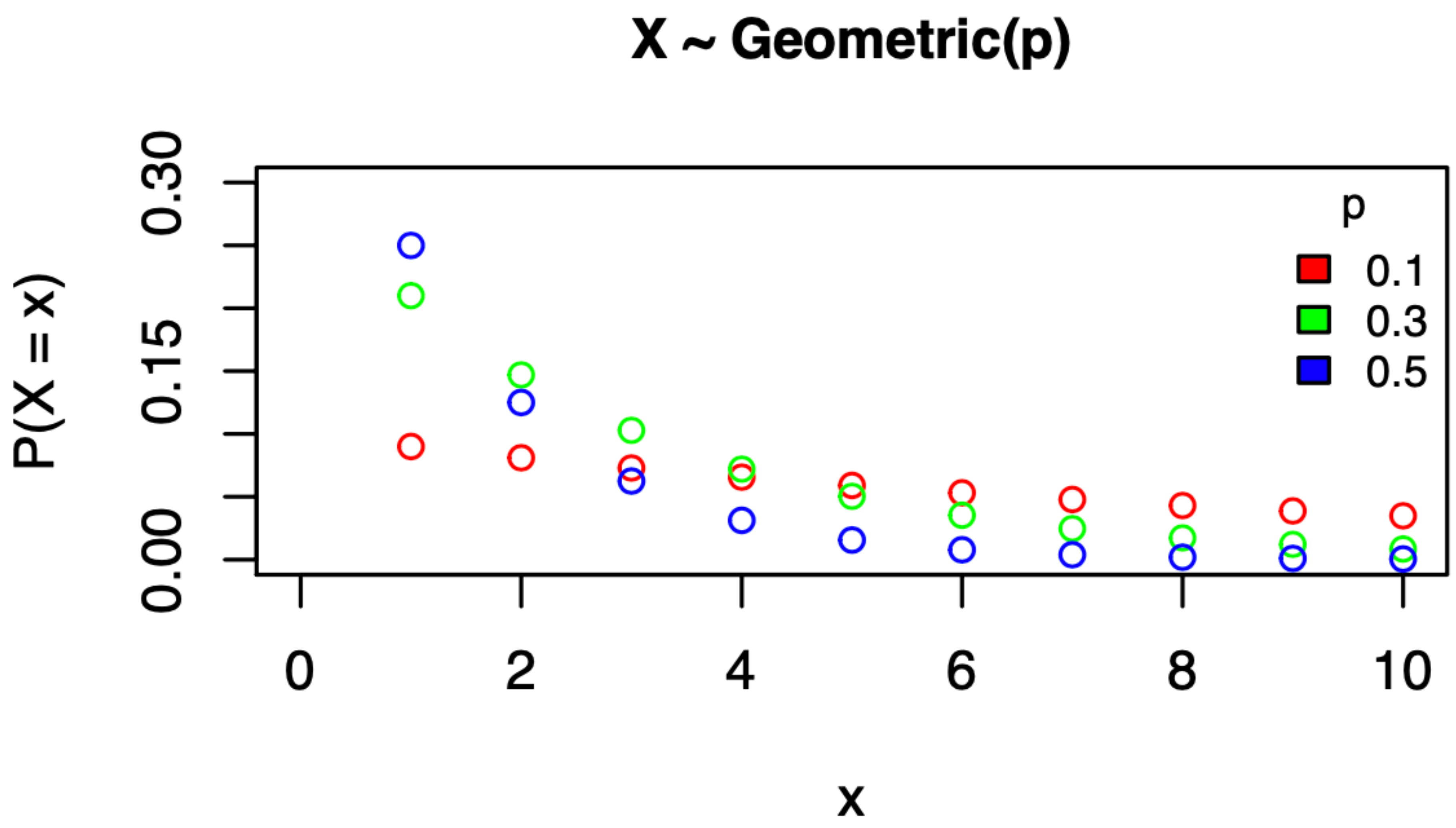
Geometric Distribution

- The probability distribution of the number X of Bernoulli trials needed to get one success

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

$$E[X] = \frac{1}{p}, Var(X) = \frac{1-p}{p^2}$$

Example: Number of times a coin is flipped before getting heads.

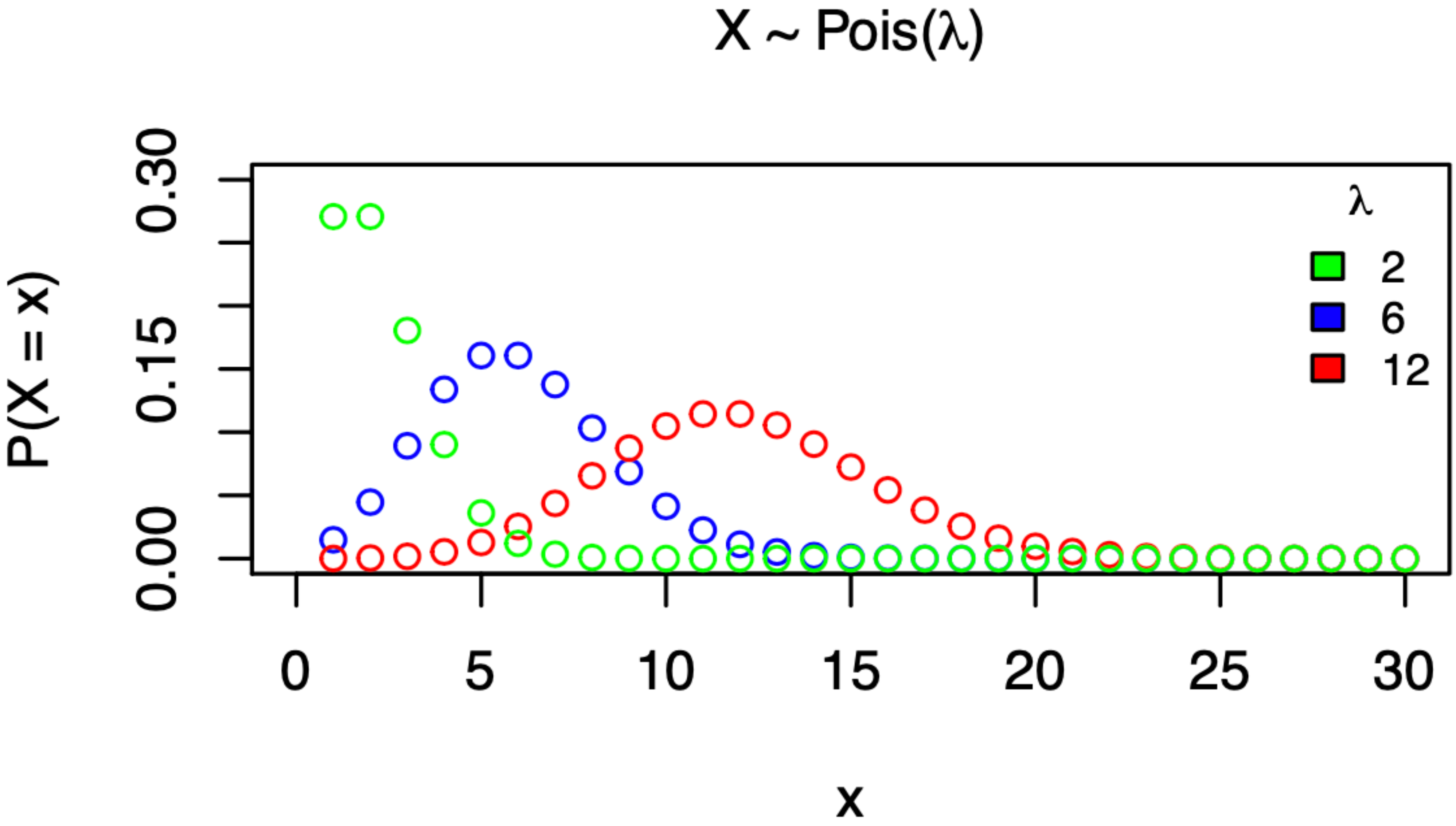


X ~ Geometric(p)
P(X = x)
0.00
0.15
0.30
0
2
4
6
8
10
x
p
0.1
0.3
0.5

Poisson Distribution

- expresses the probability of a given number of events occurring in a **fixed interval of time or space** if these events occur with a known constant rate and independent of time
- useful to model counts. E.g.,
  - number of rare diseases diagnosed in a certain year
  - number of mutations in a certain region within a chromosome
  - number of births per hour in a certain day

PMF	$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0, 1, 2, \ldots$
E[X]	λ
Var(X)	λ


X ~ Pois(λ)
P(X = x)
0.00
0.15
0.30
0
5
10
15
20
25
30
x
λ
2
6
12

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

Poisson Distribution

In a city, the mean number of people dying from a rare disease is 4 in a week. In a certain week,

a) What is the probability that no one dies from the disease?

b) What is the probability that at least 2 people die from the disease?

a) $P(X = 0) = \frac{e^{-4}4^0}{0!} \approx 0.0183$

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

$$= 1 - \left(\frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!}\right)$$

$$= 1 - (0.0183 + 0.0733) = 0.9084$$

Poisson Distribution

- As *n* gets larger, and *p* gets smaller, binomial distribution approximates to Poisson distribution

Brief Summary

- A RV is a variable whose possible values are numerical outcomes of a random phenomenon
- RV can either be discrete or continuous
- Commonly used discrete distributions include:
  - Discrete Uniform Distribution
  - Bernoulli Distribution
  - Binomial Distribution
  - Hypergeometric Distribution
  - Geometric Distribution
  - Poisson Distribution