STAT 408
Applied Regression Analysis

Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

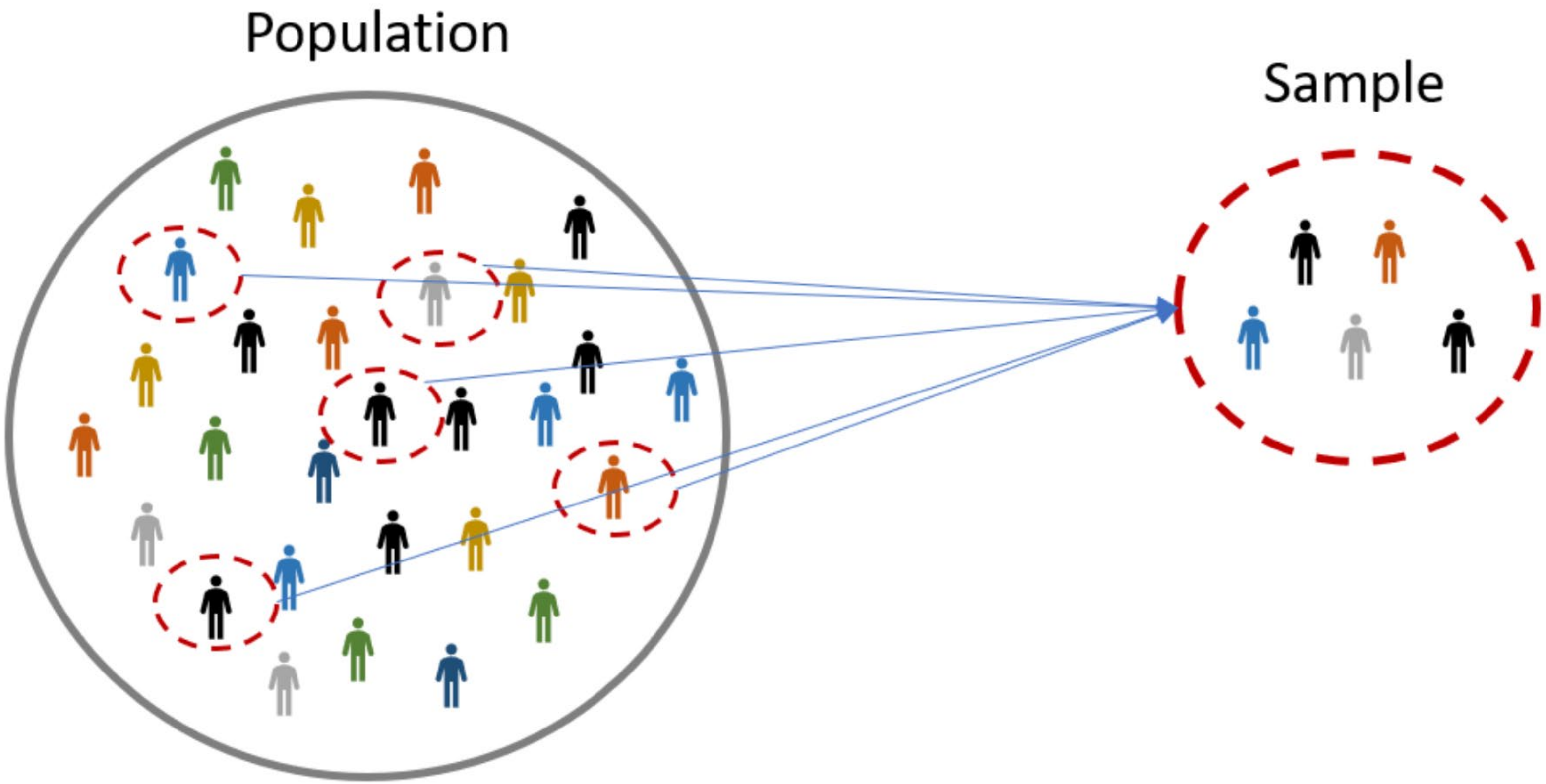
Part 1: Probability and Random Variable

Population and Sample

- Population is a well-defined collection of objects that we are interested in
    - All individuals who received a B.S. in engineering during the most recent academic year

- A sample is a subset of the population selected in some prescribed manner
    - A group of last year's engineering graduates to obtain feedback about the quality of the engineering curricula

Populations and Samples



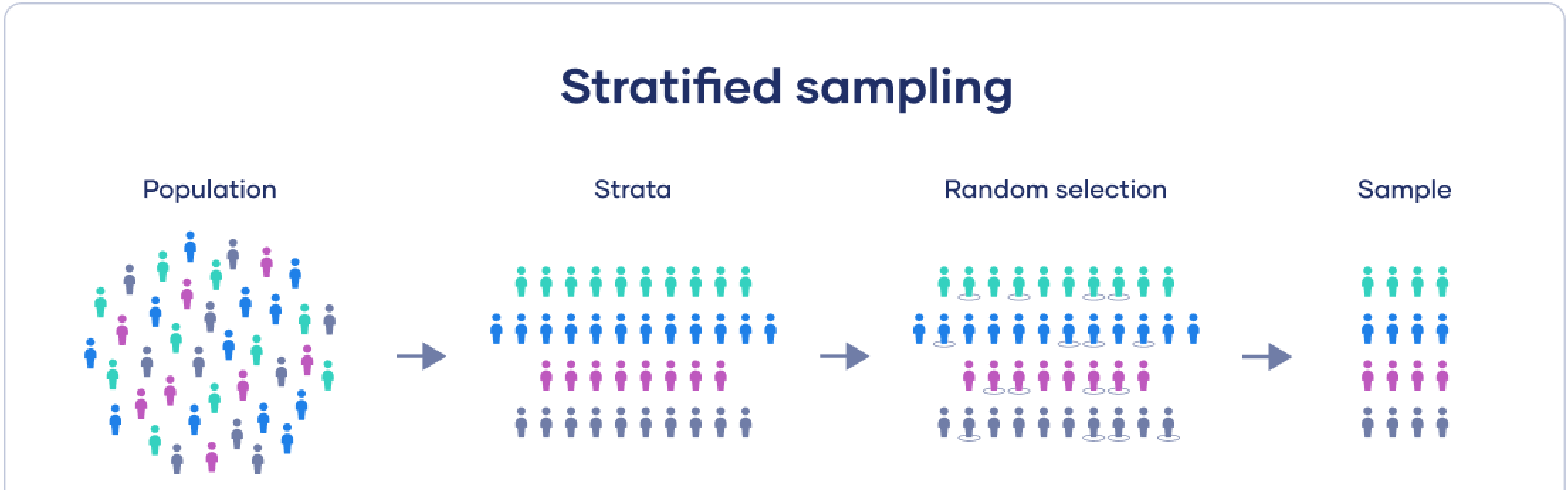
Data Collection

- Data Collection should guarantee that the sample is representative of the target population

- Random sampling is the simplest method for ensuring a representative selection
  - Any individual in the population has the same chance to be selected

- Stratified sampling separates the population into non-overlapping groups and takes a random sample from each one

Data Collection



Probability

- Given an experiment and a sample space S, probability is to assign to each event A a number P(A), which will give a precise measure of the chance that A will occur

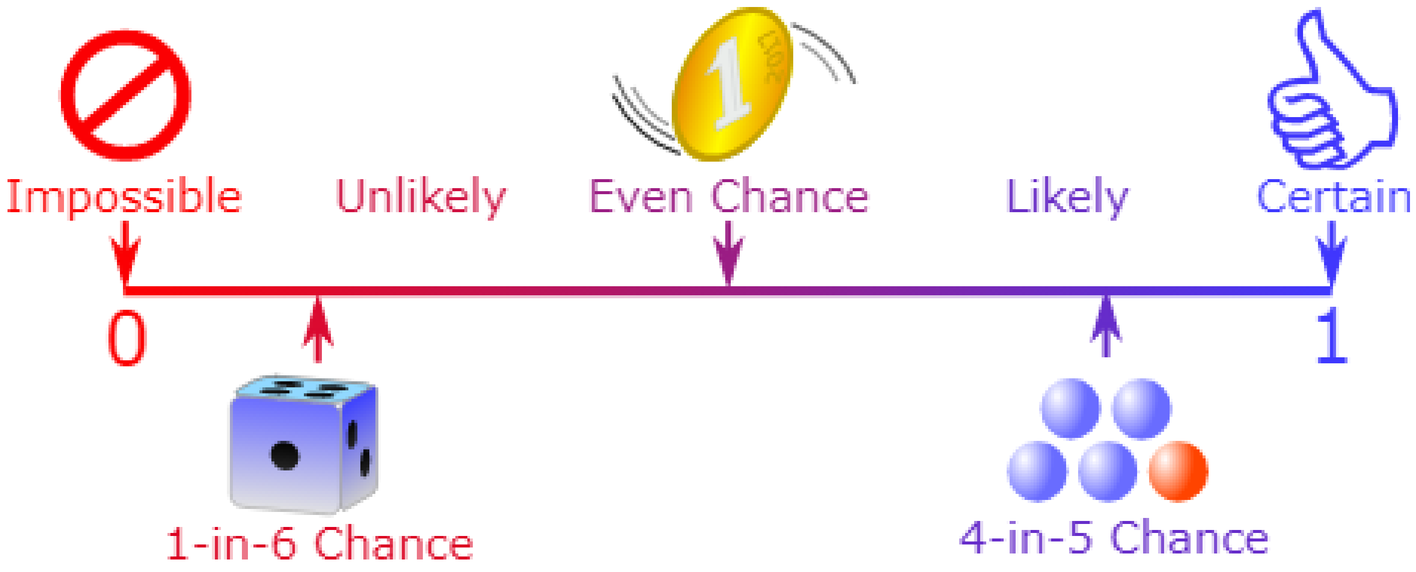
- Sample space S: all possible outcomes

- Event: a set of outcomes

Probability

AXIOM 1	For any event A, $P(A) \geq 0$.
AXIOM 2	$P(\mathcal{S}) = 1$.
AXIOM 3	If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint events, then $$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

Probability



Conditional Probability

- Sometimes, we examine how the "an event B has occurred" affects the probability of event A

- We will use the notation <u>P(A | B) </u>to represent the conditional probability of A given that the event B has occurred

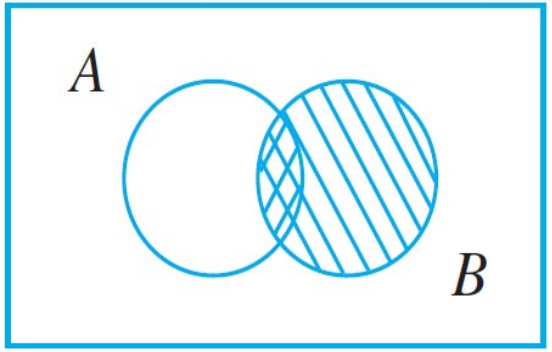
- For example, what is the probability of getting covid 19 if vaccinated

- This will be different from the probability without any conditions

Conditional Probability

For any two events A and B with $P(B) > 0$, the conditional probability of A given that B has occurred is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.3}$$



Independence

- Independence means whether event A happens or not does not affect event B

- Formal definition
  - A and B are independent if and only if $P(B|A) = P(B)$ and $P(A|B) = P(A)$

- Another definition
  - $P(A \cap B) = P(A) * P(B)$

Random Variable

- Random variable is a variable whose values depend on the outcomes of a random experiment
  - The number of responses in one survey
  - The gene expression from different biological samples

- Discrete random variable
  - Take a countable number of distinct values such as 0,1,2,3,4, ...

- Continuous random variable
  - Take an infinite number of possible values

Distribution

- The probability distribution shows the probability for each outcome of one random variable

- Discrete distribution describes the probabilities of the possible values of a discrete random variable

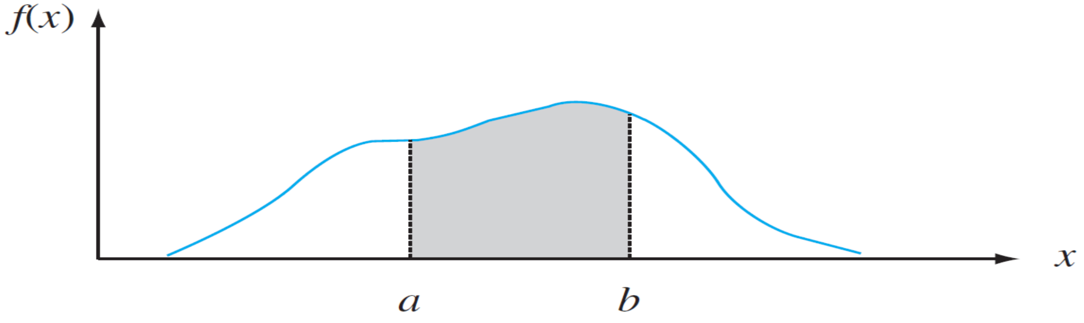
x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

- This table defines the probability mass function (pmf) of a discrete RV

Distribution

- Continuous distribution describes the probabilities of the possible values of a continuous random variable

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



- Function f(x) is the probability density function (pdf) of a continuous RV

Distribution

- Question
  - What is the relationship between discrete and continuous distributions?

Expectation

- We use expectation to measure the average of a random variable

- The expectation of a <u>discrete</u> random variable:

Let X be a discrete rv with set of possible values D and pmf $p(x)$. The **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X or just μ, is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

Expectation

- Question
  - What is the expectation for the following random variable X?

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

Expectation

- The expectation of a continuous random variable:

The **expected** or **mean value** of a continuous rv X with pdf $f(x)$ is

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

Expectation

- Some useful properties

1. $E(aX + b) = aE(X) + b$
2. $E(X + Y) = E(X) + E(Y)$
3. If X and Y are independent, then $E(XY) = E(X)E(Y)$

Variance

- We use variance to measure the average deviance of a random variable

- The variance of a discrete random variable:

Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2, or just σ^2, is

$$V(X) = \sum_D (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The **standard deviation** (SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

Variance

- Question
  - What is the variance for the following random variable X?

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

Variance

- The variance of a continuous random variable:

The **variance** of a continuous random variable X with pdf $f(x)$ and mean value μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2]$$

The **standard deviation** (SD) of X is $\sigma_X = \sqrt{V(X)}$.

Variance

- Some useful properties

- $V(X) = E(X^2) - E^2(X)$

- $V(aX + b) = a^2V(X)$

- If X and Y are independent, then $V(X + Y) = V(X) + V(Y)$

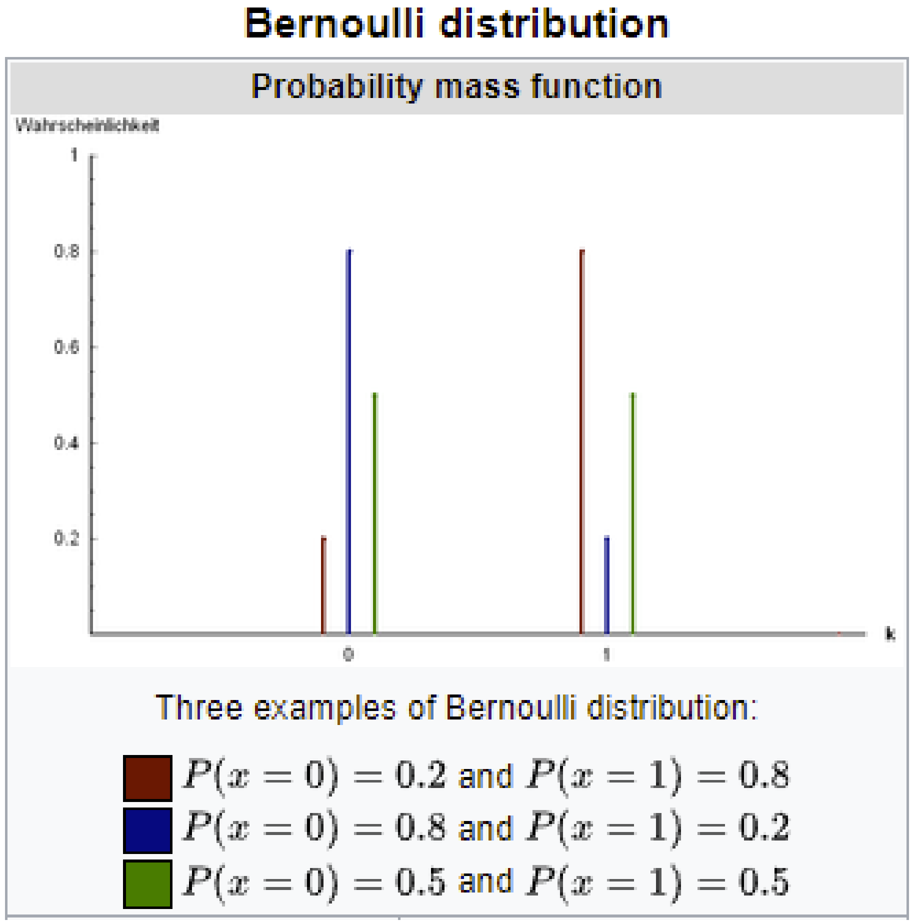
Some Useful Discrete Random Variables

- Bernoulli Random Variable
  - Take the value 1 with probability p and the value 0 with probability q = 1 - p

- Bernoulli distribution
  - $P(X=1) = p$; $P(X=0) = 1 – p = q$

- Expectation = p
- Variance = pq

Some Useful Discrete Random Variables



Some Useful Discrete Random Variables

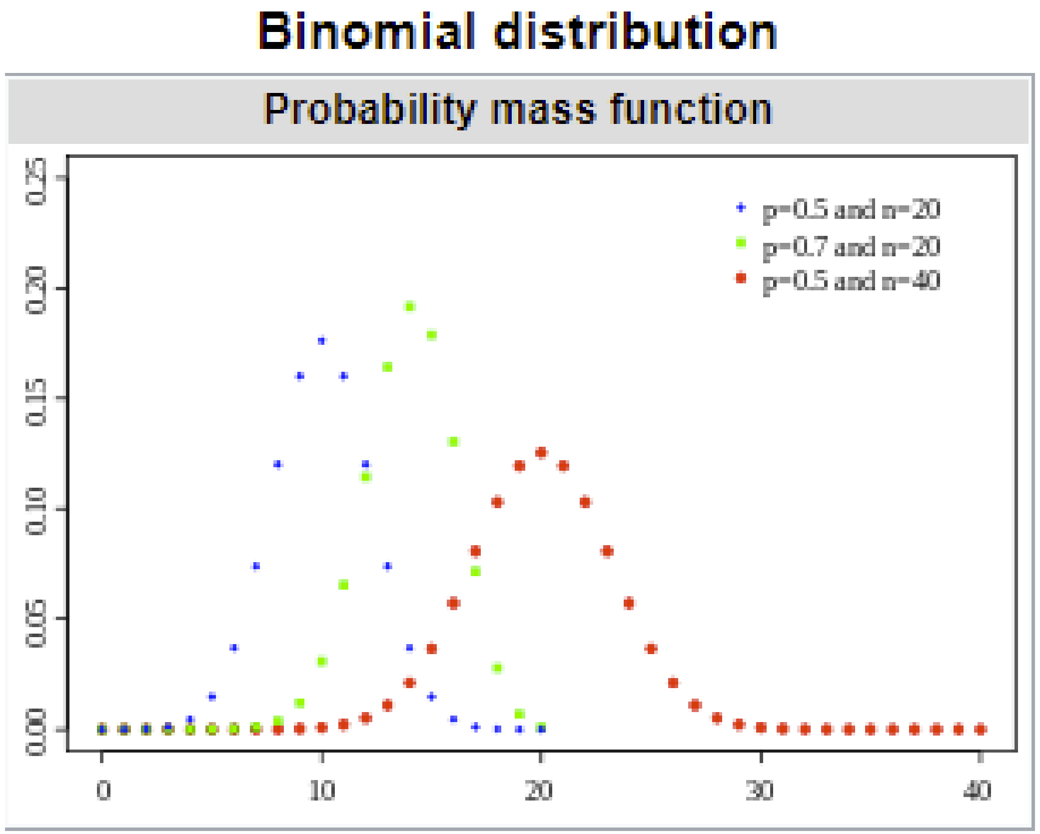
- Binomial Random Variable
  - Sum of n independent Bernoulli random variables
  - Number of successes in n Bernoulli experiments

- Binomial Distribution

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

- Expectation = np
- Variance = npq

Some Useful Discrete Random Variables



Some Useful Discrete Random Variables

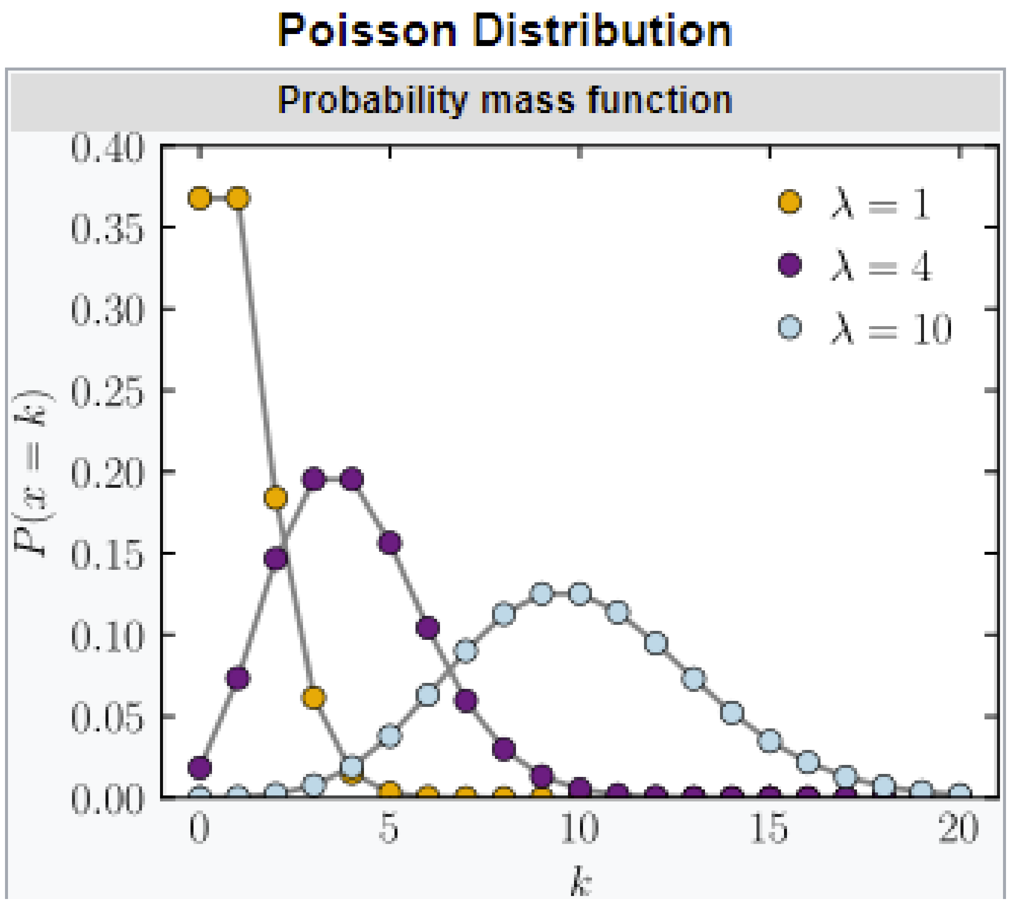
- Poisson Random Variable
  - Number of independent events occurring in a fixed time period
  - The events occur with a constant rate

- Poisson Distribution

$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Expectation = λ
- Variance = λ

Some Useful Discrete Random Variables



Some Useful Continuous Random Variables

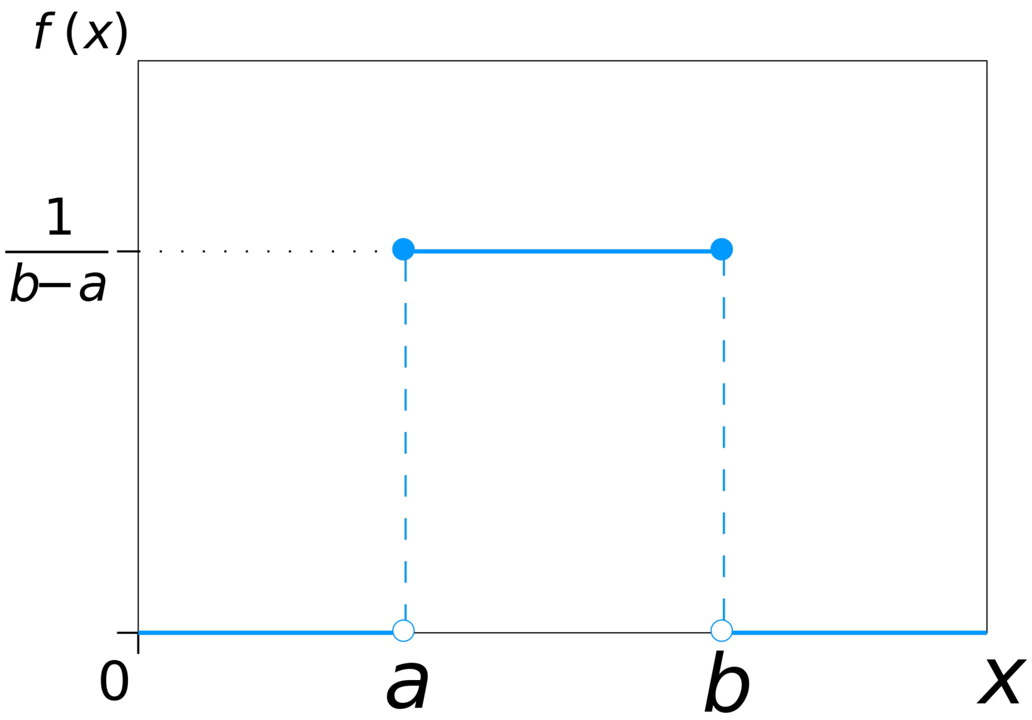
- Unform Random Variable
  - Take the same value in its domain with the same probability

- Unform distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

- Expectation = (a + b)/2
- Variance = $(b - a)^2/12$

Some Useful Continuous Random Variables



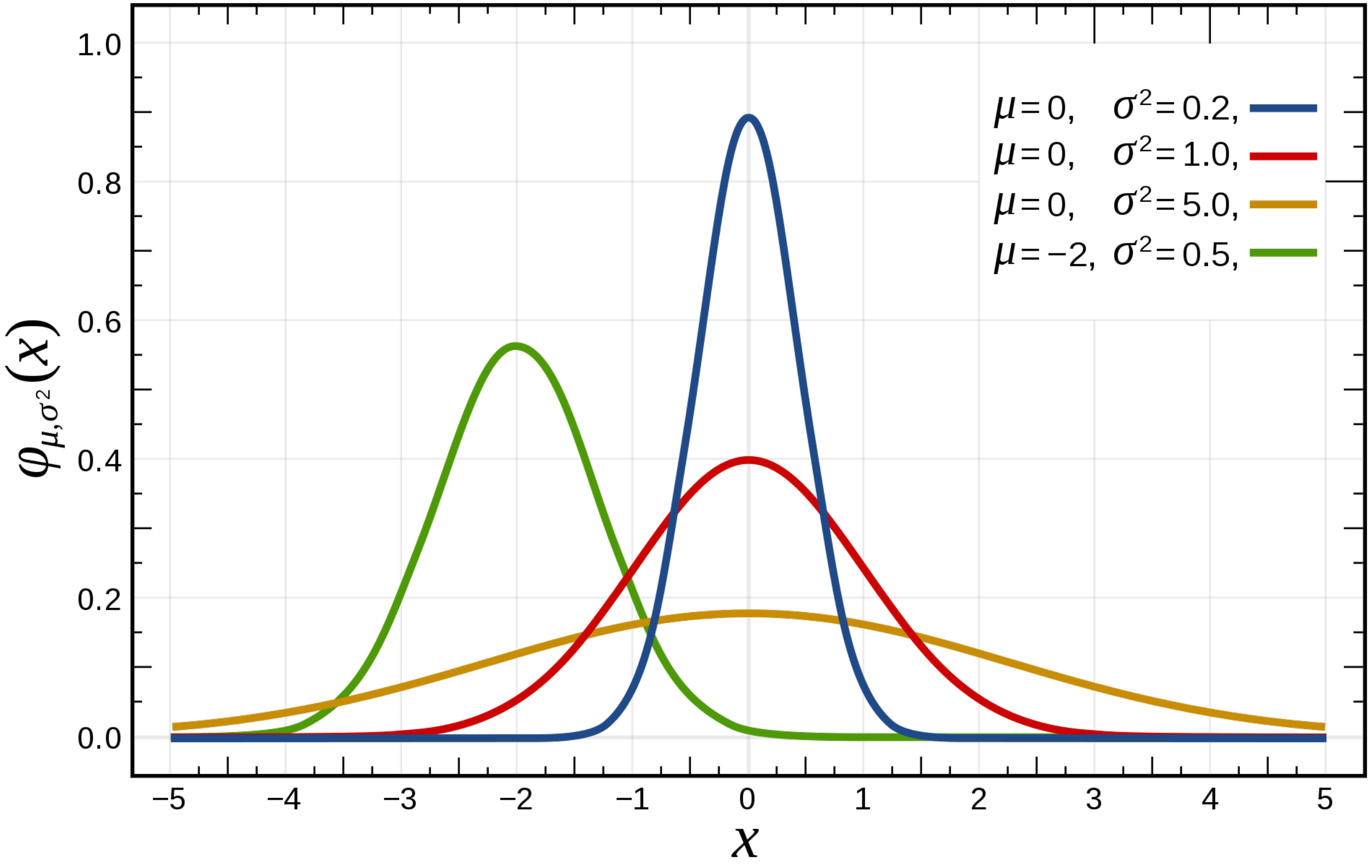
Some Useful Continuous Random Variables

- Normal/Gaussian Random Variable
  - Most important, widely used
- Normal distribution

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

- Expectation = μ
- Variance = σ^2

Some Useful Continuous Random Variables



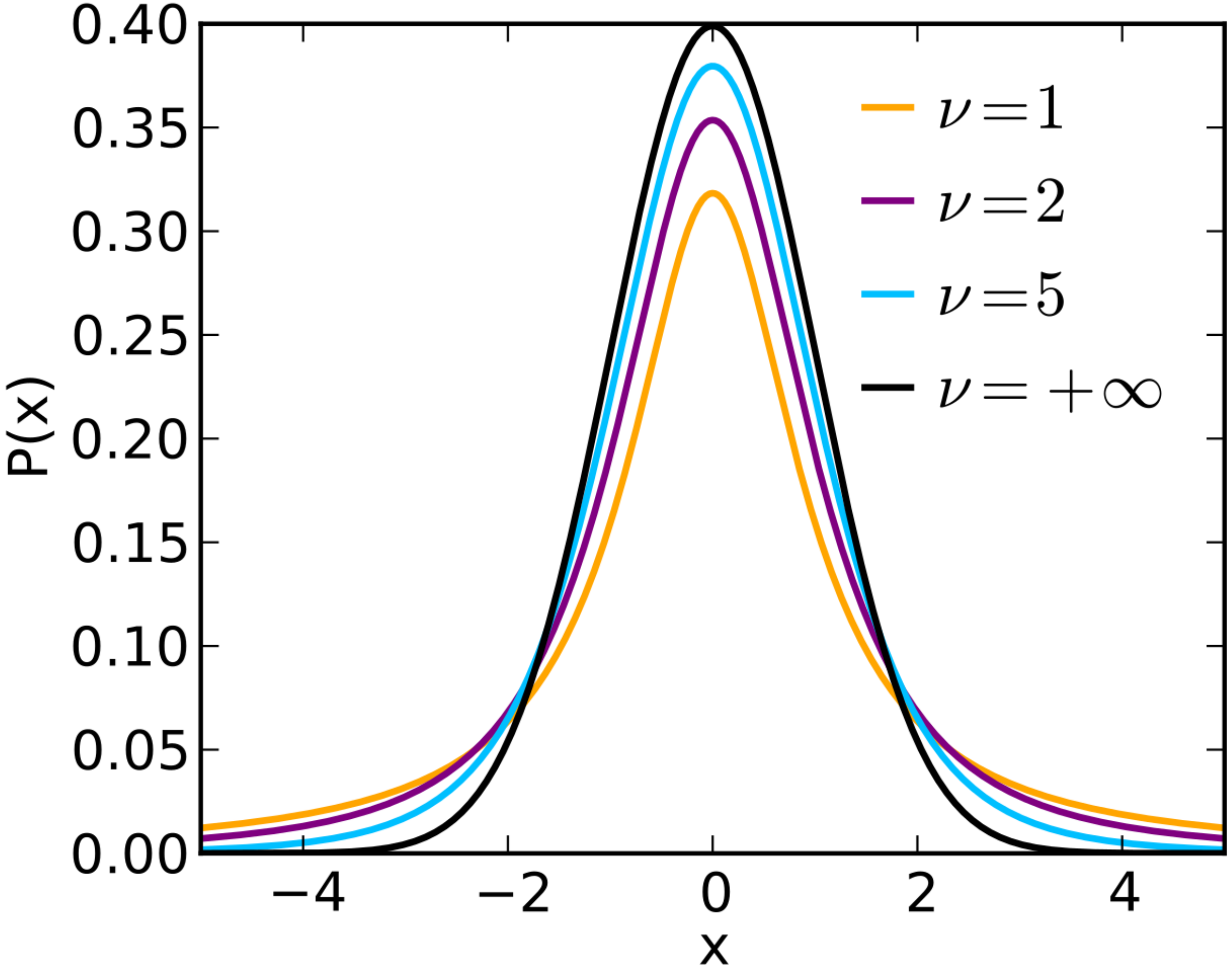
Some Useful Continuous Random Variables

- t Random Variable
  - Similar shape to normal, but always centered at zero

- t distribution
  - Complicated density function, not shown
  - Degree of freedom (DF or v) controls the spread

- Expectation = 0
- Variance = ?

Some Useful Continuous Random Variables



List of Probability Distributions

- https://en.wikipedia.org/wiki/List_of_probability_distributions