



Data Science (COSE471) 2021 Spring

Probability Review (Optional)

Dept. of Computer Science and Engineering
Korea University

* This material is adapted from Berkeley CS 100 (ds100.org) and may be copyrighted by them.

Random Variables

What is a random variable?

A random variable is a **variable whose possible values are numerical outcomes of a random phenomenon.**

- A number is assigned to every outcome of an experiment.

We typically denote random variables with uppercase letters late in the alphabet (e.g. X , Y).

What is a random variable?

A random variable is a **variable whose possible values are numerical samples of outcomes of a random phenomenon.**

- Mapping from an outcome (or sample) to a number

We typically denote random variables with uppercase letters late in the alphabet (e.g. X , Y).

Why **random**? Because the sample or the outcomes was drawn at random.

Why **variable**? Because its value depends on how the sample came out.

Example of random variable



- Let s be an outcome, i.e. a sample of size 3.
- Let X be the number of blue people in our sample.
 - X , then, is a random variable!

In general, the input to a random variable is a sample, and the output is some function of that sample.

- **Domain:** all samples.
- **Range:** real numbers.



$$X(s) = 1$$



$$X(s) = 2$$



$$X(s) = 0$$

Functions of random variables

- **A function of a random variable is also a random variable!**
 - A function of a random variable is a “function of a function of the sample”, which itself is also a function of the sample.
- If you create multiple random variables based on your sample, then functions of them are also random variables.

For instance, if X_1, X_2, \dots, X_n are random variables, then so are all of these:

$$\begin{array}{ccccccc} X_n^2 & & \#\{i : X_i > 10\} & & \frac{1}{n} \sum_{i=1}^n X_i & & \\ & & & & & & \\ \max(X_1, X_2, \dots, X_n) & & & & & & \frac{1}{n} \sum_{i=1}^n (X_i - c)^2 \end{array}$$

Distribution

Probability mass function (PMF)

For now, assume our random variables have a finite number of possible values.

$$P(X = x)$$

This is the **probability that random variable X takes on the value x** .

- For instance, $P(X = 20)$ is the chance that X has the value 20.
- The **distribution** of X is a description of how the total probability of 100% is split over all possible values of X .
- The probabilities of each possible value must each be non-negative, and must sum to 1:

$$\sum_{\text{all } x} P(X = x) = 1$$

Example distribution

Consider a random variable X with the following **distribution table**:

<i>x</i>	$P(X = x)$
3	0.1
4	0.2
6	0.4
8	0.3

values

probabilities of those values

$$P(X = 4) = 0.2$$

$$P(X < 6) = 0.3$$

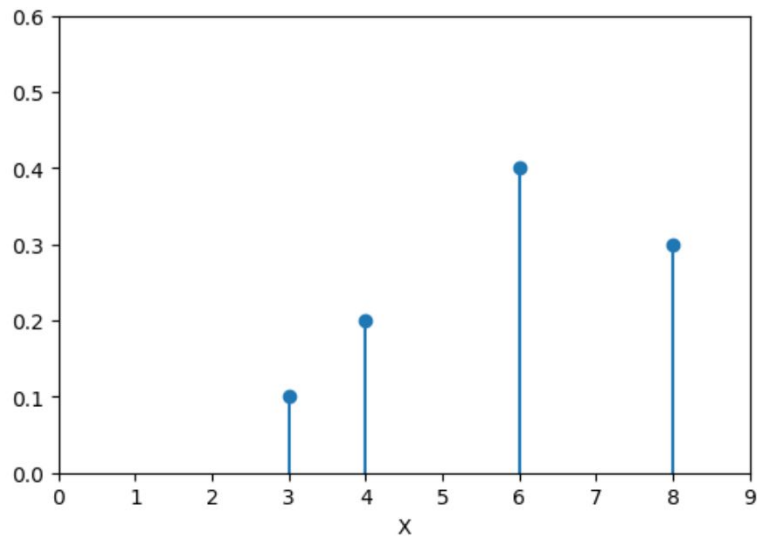
$$P(X \leq 6) = 0.7$$

$$P(X = 7) = 0$$

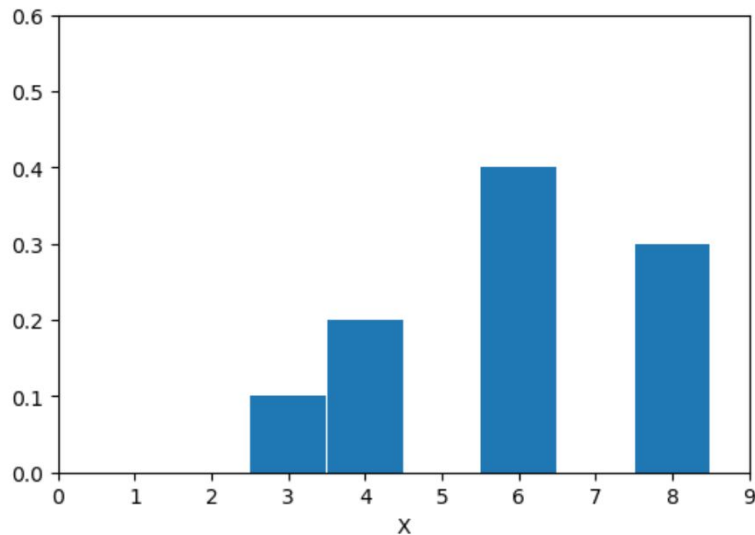
To compute related probabilities, we add up the probabilities belonging to that event.
(For instance, $X < 6$ happens if $X = 3$ or $X = 4$).

Example distribution

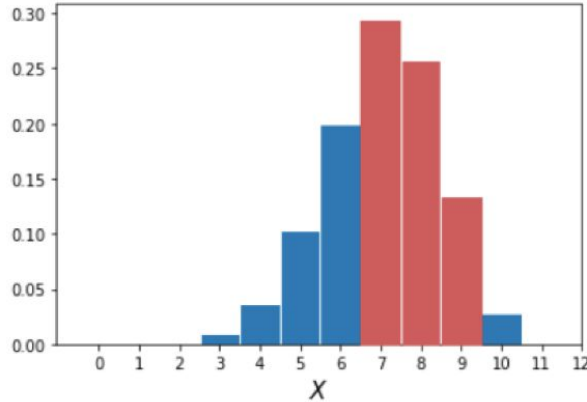
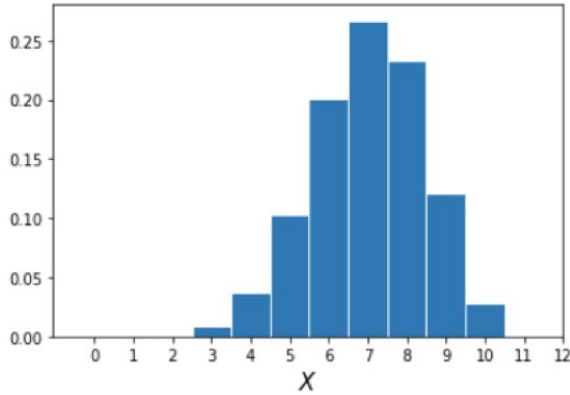
Since all of the probability is on the values in $[3, 4, 6, 8]$, we can visualize the distribution of X like this:



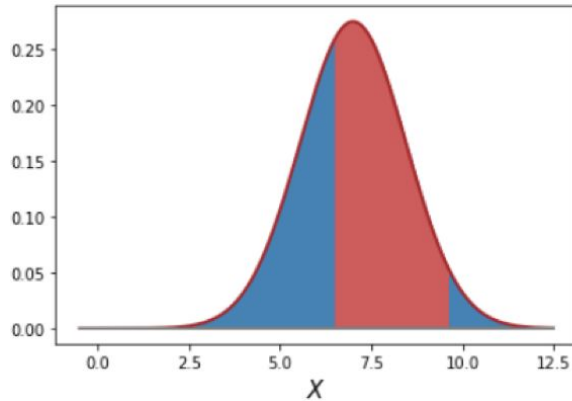
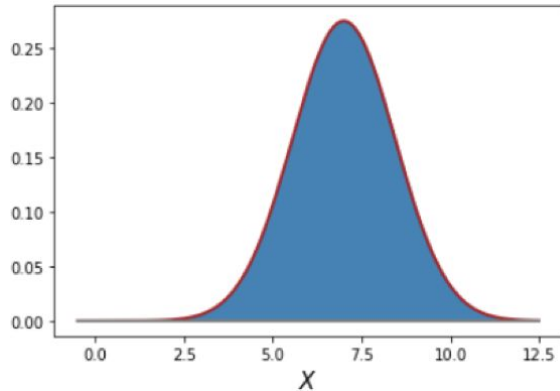
But it can be beneficial to visualize our distribution as being binned. This helps us think of probabilities as areas.



Probabilities are areas, regardless of the distribution type!



The red bars correspond to the region where $7 \leq X \leq 9$. Thus, the area of the red bars is **$P(7 \leq X \leq 9)$** .



The red bars correspond to the region where $6.8 \leq X \leq 9.5$. Thus, the area of the red bars is **$P(6.8 \leq X \leq 9.5)$** .

Types of distributions

Probability distributions largely fall into two main categories.

- **Discrete.**
 - The set of possible values that X can take on is either finite or countably infinite.
 - Values are separated by some fixed amount.
 - For instance, $X = 1, 2, 3, 4, \dots$
- **Continuous.**
 - The set of possible values that X can take on is uncountable.
 - Typically, X can be any real number in some interval (not just our counting numbers).

Here, we will focus almost exclusively on discrete distributions. However, it's important to know that continuous distributions exist. They will reappear later on! (bias-variance tradeoff, KDEs).

Common distributions

Discrete

- **Bernoulli** (p).
 - Takes on the value 1 with probability p , and 0 with probability $1-p$.
- **Binomial** (n, p).
 - Number of 1s in n independent Bernoulli (p) trials.
 - Probabilities given by the binomial formula.
- **Uniform on a finite set.**
 - Probability of each value is $1 / (\text{size of set})$. For example, a standard die.

Continuous

- Uniform on the unit interval.
 - U could be any real number in the range $[0, 1]$.
- Normal (μ, σ^2).

Parameters of a distribution are the constants associated with it. These define its shape and the values it takes on. These are the numbers provided in parentheses.

Bernoulli and Binomial Distributions

Bernoulli (p)

An **indicator** random variable I has the value 1 if a specified event happens, and 0 if it doesn't happen. You can think of it as a numerical code for a yes/no answer. Some examples:

- Flips of a coin (1 if heads, 0 if tails).
- Draws from a sample (1 if person has certain characteristic, 0 if not).
- Rolls of a dice (1 if roll is a 3, 0 if not).

Events with just two possible outcomes are sometimes called **trials**. If such an event occurs with probability p , then the distribution of I is

$$P(I = 1) = p$$

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

This is called the Bernoulli (p) distribution. (Indicators are sometimes called “0-1 RVs”.)

Motivating the binomial distribution

Suppose we have a coin that flips heads with probability 0.3. Let's say we flip it 10 times (where each flip is independent of one another) and want to count the number of heads we see.

To compute the probability of seeing 6 heads (and hence, 4 tails), we

- 6 heads, and each occurs with probability 0.3, so we multiply 0.3^6
- 4 tails, and each occurs with probability 0.7, so we multiply by 0.7^4
- We also need to account for the number of ways you can arrange 6 heads and 4 tails.
 - e.g. HHHHHHTTTT and HHTTHHTTHH are different sequences, but both have the same chance of occurring
 - From last lecture, this is $\binom{10}{6}$ – we have 10 “positions”, choose 6 of them to be heads.
- Putting these pieces together

$$P(6 \text{ heads}) = \binom{10}{6} 0.3^6 0.7^4$$

Binomial distribution

In general, suppose you have n independent, repeated trials, each of which has a probability p of succeeding. The number of successes, X , has the binomial (n, p) distribution.

We compute probabilities using the following function:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad 0 \leq k \leq n$$

This function is called the **probability mass function (PMF)** of the binomial distribution. As an input, it takes in a possible value of the random variable, and as an output, it returns the probability of the random variable having that value.

Equality

Comparing two draws

Suppose X and Y are the ages of two people drawn at random with replacement from a population.

- Is X **equal to** Y ?
 - **No!** The age of the first person drawn could be different from the age of the second person drawn.
- Do X and Y have the **same distribution**?
 - **Yes!** The distribution of X is the distribution of ages in the population. So is the distribution of Y .

Two kinds of equality

Consider two random variables X and Y based on our sample.

X and Y are **equal** if, for every sample s , $X(s) = Y(s)$.

- If this is true, we can write $X = Y$.

X and Y are **identically distributed** if the distribution of X is the same as the distribution of Y .

- If this is true, we can say “ X and Y are equal in distribution.”

If $X = Y$, then X and Y are also identically distributed. But the converse is not true!

Expectation

Definition of expectation

The **expectation** of a random variable X is the weighted average of the values of X , where the weights are the probabilities of the values.

The most common formulation applies the weights one possible value at a time:

$$\mathbb{E}(X) = \sum_{\substack{\text{all possible} \\ x}} x \mathbb{P}(X = x)$$

However, an equivalent formulation applies the weights one sample at a time:

$$\mathbb{E}(X) = \sum_{\substack{\text{all samples} \\ s}} X(s) \mathbb{P}(s)$$

Interpretation of expectation

- Expectation is a **number**, not a random variable!
- It is analogous to the average.
 - It has the same units as the random variable.
 - It doesn't need to be a possible value of the random variable.
 - For instance – what is the expectation of a roll of a die?
 - It is the center of gravity of the probability histogram.
- It is the long run average of the random variable, if you simulate the variable many times.

Examples

Consider the random variable X we defined earlier:


x	$P(X = x)$
3	0.1
4	0.2
6	0.4
8	0.3

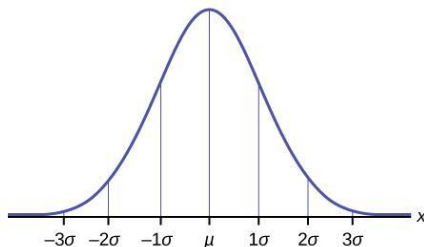
$$\begin{aligned} E(X) &= \sum_x x \cdot P(X = x) \\ &= 3 \cdot 0.1 + 4 \cdot 0.2 + 6 \cdot 0.4 + 8 \cdot 0.3 \\ &= 0.3 + 0.8 + 2.4 + 2.4 \\ &= 5.9 \end{aligned}$$

Note, 5.9 is not one of the possible values that X can take on!

Examples without calculation

Expectation is the “balance point” of the distribution histogram of a random variable. Sometimes, using this property reduces the need for any calculation. Here are some examples.

- X is $\mathbb{E}(X) = (N + 1)/2$, $N\}$.
 -
 - For instance, a standard die takes on values $\{1, 2, 3, 4, 5, 6\}$, all with the same chance. 3.5 is in the “middle”.
- U is $\mathbb{E}(U) = 0.5$ interval $[0, 1]$.
 - (μ, σ^2)
- X is $\mathbb{E}(X) = \mu$. 
-



Transformations

Let X be a random variable, and a and b be constants.

- We call $aX + b$ a **linear transformation** of X .
- The expectation of a linear transform of X is equal to the linear transform applied to the expectation of X .
- Put less cryptically:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Note, this means that $\mathbb{E}[c] = c$, where c is a constant.

Why does this matter?

- We will often manipulate the sample mean of several random variables. This is a linear transformation of the sample sum.
- Many unit conversions are also linear transformations (e.g. $^{\circ}\text{F} = 9/5 * ^{\circ}\text{C} + 32$).

Expectation of functions of random variables

More generally, if X is a random variable and g is any function (not necessarily linear), we have

$$\mathbb{E}(g(X)) = \sum_x g(x)P(X = x)$$

For example, if X is uniform on $\{1, 2, 3, 4, 5, 6\}$, we have

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_x x^2 P(X = x) \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} \\ &= \frac{91}{6}\end{aligned}$$

$$\mathbb{E}(X^2) = 91/6 = 15.166\dots$$

$$\mathbb{E}(X)^2 = 3.5^2 = 12.25$$

$\mathbb{E}(X^2)$ and $\mathbb{E}(X)^2$ are different!

Note: The property that held on the last slide for linear functions g does not hold in general!

$$\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$$

Additivity

For **any** two random variables X and Y (regardless of their relationship):

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

We call this property the **additivity of expectation**. We will not prove this, but you can do so yourself using the second definition of expectation (summing over all samples).

For example: Consider the “sample sum” $S_n = \sum_{i=1}^n X_i$, where $\mathbb{E}(X_i) = \mu$ for each i . Then,

$$\begin{aligned}\mathbb{E}(S_n) &= \sum_{i=1}^n \mathbb{E}(X_i) \\ &= n\mu\end{aligned}$$

Linearity

Two of the properties we just established were

- Linear transformations apply to expectations.
- Expectation is additive.

Combining these gives us a single property, which is sometimes referred to as the **linearity of expectation**. For any random variables X , Y and constants a , b :

$$E(aX + bY) = aE(X) + bE(Y)$$

This more general form won't appear often in this class, but it is good to be aware of.

Expectation of Bernoulli

Recall, if X follows the Bernoulli distribution with parameter p , then

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

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

The expectation of X , then, is

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

Expectation of the binomial distribution

Let X have the binomial (n, p) distribution. We know that X is the number of “successes” in n independent trials of some event, each of which occur with probability p .

- Each trial can be thought of as a single Bernoulli (p) trial.
- We can then write:

$$X = I_1 + I_2 + \cdots + I_n$$

where I_j is the **indicator** of success on trial j . $I_j = 1$ if trial j is a success, and 0 else.

- For each j , $\mathbb{E}(I_j) = p$.
- Then, by the additivity of expectation, $\mathbb{E}(X) = np$.
- Thus, the **expectation of a binomial (n, p) random variable is np** .