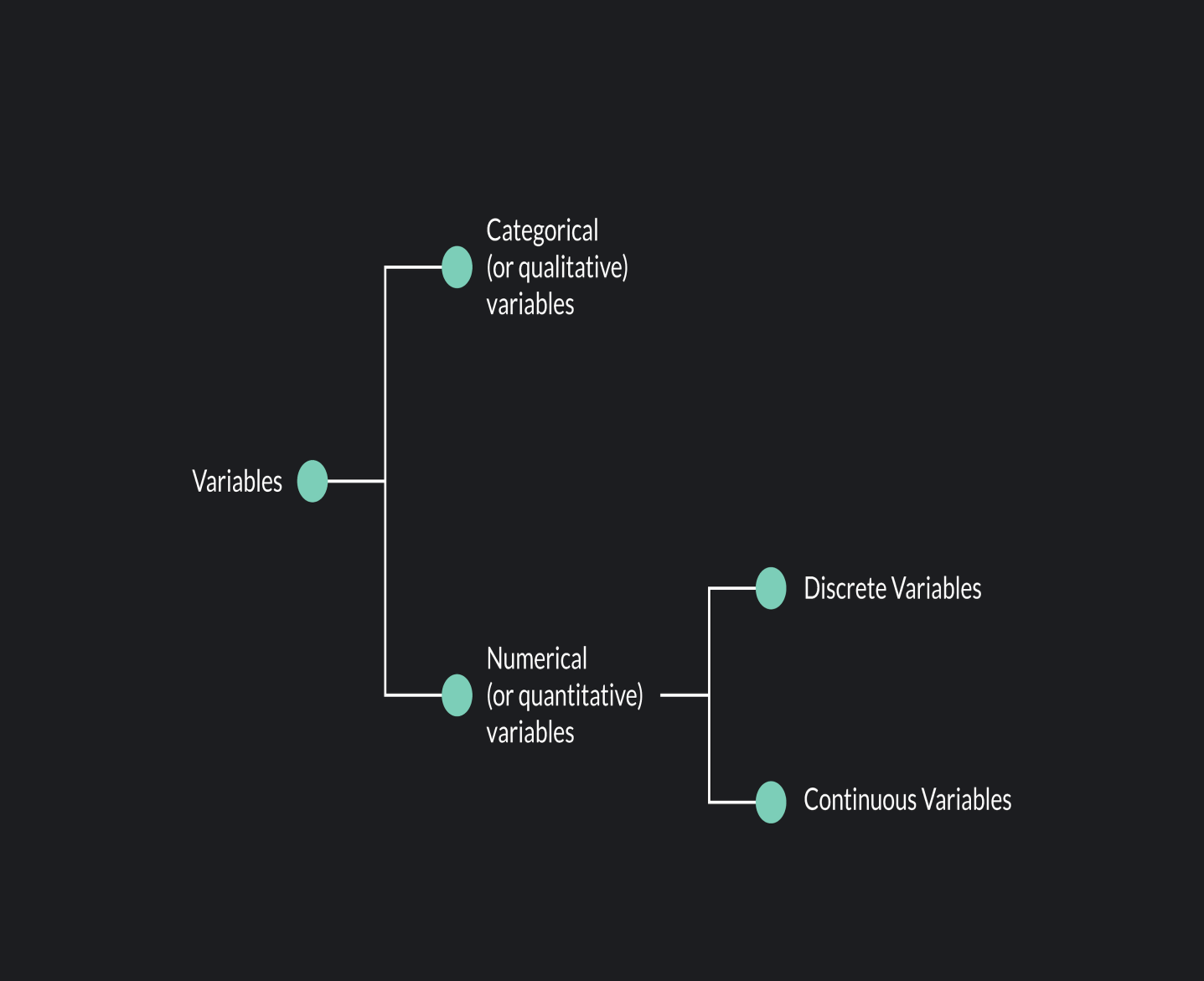
Categorical Variables and Numerical Variables

Variables can be categorical or numerical.

Categorical—also called qualitative—variables consist of names and labels that divide data into specific categories. When you select your nationality or your race on a survey, those responses are categorical.

Numerical—also called quantitative—variables have values that can either be counted or measured. Discrete and continuous variables are specific types of numerical data.



inNow that you know what a discrete random variable is, you may be wondering: Is there a difference between the terms “discrete variable” and “discrete data”?

In [statistics](https://articles.outlier.org/what-is-statistics) and data analysis, a variable is any characteristic or attribute that can be measured. [Data](https://articles.outlier.org/types-of-data-in-statistics) refers to the values or observations that are collected for a particular variable. For example, if you have a discrete random variable representing years of schooling, the data you collect would be discrete data.

Discrete data has the following main characteristics:

A data set that consists only of data collected for discrete variables is called a discrete data set.

WhA continuous variable is a variable that can take on any value within a range. A continuous variable takes on an infinite number of possible values within a given range.

Because the possible values for a continuous variable are infinite, we measure continuous variables (rather than count), often using a measuring device like a ruler or stopwatch. Continuous variables include all the fractional or decimal values within a range.

Examples

Examples of continuous variables include:

* The time it takes sprinters to run 100 meters
* The size of real estate lots in a city
* The weight of baby elephants
* The body temperature of patients with the flu
* The deployment altitude of skydivers

None of these variables are countable. This is the key difference between discrete and continuous variables. A continuous variable can take on an infinite number of values within a range.

What is continuous data? Continuous data are observations or data points collected for a continuous random variable. Let’s say you’re interested in the time it takes 5th graders to run a 50-yard dash.

This is a continuous random variable. Now, if you go out and collect a sample of 100 5th graders and record the time it takes each of them to run the dash, you’ll have a continuous data set consisting of 100 data points.

Continuous data has the following main characteristics:s Variables Sometimes we treat continuous variables as if they were discrete.

Age is an excellent example of this. If you know a person’s time of birth, you could measure their age precisely up to the second or even millisecond if you wanted to. In this sense, age is a continuous variable. However, we don’t usually care about a person’s exact age. Instead, we treat age as a discrete variable and count age in years.a Random Variable?

A random variable is a variable where the values are the outcome of a random process.

An easy example of a random variable is:

|  |
| --- |
| X = the number you get when you roll a die |

When you roll a die, the roll itself is a random event. The possible values of X are 1, 2, 3, 4, 5, or 6, but the specific value you get depends on the randomness of the event. It’s uncertain which number will appear on any given roll. You can learn more about events and the odds of of results when you read our article about [math probability](https://articles.outlier.org/understanding-math-probability).

Random variables can be numerical or categorical, continuous or discrete.

Main Differences Between Discrete and Continuous Variables

The table below summarizes the key differences between [discrete and continuous variables with examples](https://articles.outlier.org/discrete-variable-examples).

|  |  |
| --- | --- |
| **DISCRETE VARIABLES** | **CONTINUOUS VARIABLES** |
| Definition- A discrete variable is a variable that takes on distinct, countable values. | Definition- A continuous variable is a variable that takes on any value within a range, and the number of possible values within that range is infinite. |
| Discrete variables have values that are counted. | The values of a continuous variable are measured. |
| Discrete Variable Examples  - The number of workers in an office  - The number of steps you take in a day  - The number of babies born each day | Continuous Variable Examples  - The time it takes for office employees to commute to work  - The distance you walk in a day  - The weight of newborn babies |

The main difference between discrete data and continuous data is that discrete data is data collected for a discrete random variable, while continuous data is data collected for a continuous random variable.

**Binomial distribution**

There are some sorts of experiments which have only two possible outcomes, either a “**success**” or a “**failure**” – these types of random experiments are called binomial experiments or “**Bernoulli trials**”. For example, the experiment of tossing a coin and getting a head.

Thus, in a [probability distribution](https://byjus.com/maths/probability-distribution/), binomial distribution denotes the success of a random variable X in an n trials binomial experiment. Following are the conditions to find binomial distribution:

* n is finite and defined.
* Each trial has only two possible outcomes: success and failure.
* The result of each trial is independent of other trials.
* The probability of success and failure remains the same in each trial.

**Bernoulli’s Theorem for Binomial Distribution**

Let there be ‘n’ binomial experiment trials and let the random variable X denote the success of these trials. If p is the probability of success and 1 – p = q is the probability of failure in each trial, then,

| **P(X = r) = nCr pr q(n – r)** |
| --- |

As P(X) is the term of the binomial expansion of (p + q)n, it is called the binomial distribution.

**Note :**

* Sum of all probabilities in the distribution sums up to 1
* Probability of success in all n trials is pn
* Probability of failure in all n trials is (1 – p)n = qn
* Probability of success in at least one trial = P(X ≥ 1) = 1 – P(X = 0) = 1 – qn.
* Probability of at least r successes = P(X ≥ r) = ∑knCk pk qn – k (k = r, r + 1,…, n)
* Probability of at most r successes = P(X ≤ r) = ∑knCk pk qn – k (k = 0, 1, …, r)
* If in n trials, the experiment is repeated N times, the expected frequencies are N.P(r) for r = 0, 1, 2, 3, …, n.

**Binomial Distribution Questions with Solutions**

Let us practice some important questions on binomial distribution in probability.

**Question 1:**

Find the binomial distribution of getting a six in three tosses of an unbiased dice.

**Solution:**

Let X be the random variable of getting six. Then X can be 0, 1, 2, 3.

Here, n = 3

p = Probability of getting a six in a toss = ⅙

q = Probability of not getting a six in a toss = 1 – ⅙ = ⅚

P(X = 0) = nCr pr q(n – r) = 3C0 (⅙)0 (⅚)3 – 0

= 1 × 1 × 125/216 = 125/216

P(X = 1) = nCr pr q(n – r) = 3C1 (⅙)1 (⅚)3 – 1

= 3 × ⅙ × 25/36 = 25/72

P(X = 2) = nCr pr q(n – r) = 3C2 (⅙)2 (⅚)3 – 2

= 3 × 1/36 × ⅚ = 5/72

P(X = 3) = nCr pr q(n – r) = 3C3 (⅙)3 (⅚)3 – 3

= 1 × 1/216 × 1 = 1/216

The required binomial distribution of X is:

| **X** | **0** | **1** | **2** | **3** |
| --- | --- | --- | --- | --- |
| **p(X)** | **125/216** | **25/72** | **5/72** | **1/216** |

**Question 2:**

Find the probability distribution of the number of doublets in four throws of a pair of dice.

**Solution:**

There are 36 total possible outcomes for a throw of dice, for which the following outcomes are the success of the experiment: {(1,1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)}

p = probability of getting doublets = 6/36 = ⅙

q = probability of getting not getting doublets = 1 – ⅙ = ⅚

X: numbers of doublets, then X = 0, 1, 2, 3, and 4.

P(X = 0) = nCr pr q(n – r) = 4C0 (⅙)0 (⅚)4 – 0

= 1 × 1 × 625/1296

P(X = 1) = nCr pr q(n – r) = 4C1 (⅙)1 (⅚)4 – 1

= 4 × ⅙ × 125/216 = 125/324

P(X = 2) = nCr pr q(n – r) = 4C2 (⅙)2 (⅚)4 – 2

= 6 × 1/36 × 25/36

= 25/216

P(X = 3) = nCr pr q(n – r) = 4C3 (⅙)3 (⅚)4 – 3

= 4 × 1/216 × ⅚ = 20/1296

P(X = 4) = nCr pr q(n – r) = 4C4 (⅙)4 (⅚)4 – 4

= 1 × 1/1296 × 1 = 1/1296.

∴ The required probability distribution is:

| **X** | **0** | **1** | **2** | **3** | **4** |
| --- | --- | --- | --- | --- | --- |
| **P(X)** | **625/1296** | **125/324** | **25/216** | **20/1296** | **1/1296** |

**Question 3:**

Find the probability of getting at least 5 times head-on tossing an unbiased coin for 6 times by using the binomial distribution.

**Solution:**

p = P(getting an head in a single toss) = ½

q = P(not getting an head in a single toss) = ½

X = successfully getting a head

P(X ≥ 5) = P(getting at least 5 heads) = P(X = 5) + P(X = 6)

= 6C5 (½)5 (½)(6 – 5) + 6C6 (½)6 (½)6 – 6

= 6 × (½)6 + 1 × (½)6 = 7/24.

Hence, the probability of getting at least 5 heads is 7/24.

**Question 4:**

There are four fused bulbs in a lot of 10 good bulbs. If three bulbs are drawn at random with replacement, find the probability of distribution of the number of fused bulbs drawn.

**Solution:**

This is a problem of binomial distribution as the event of drawing a fused bulb is independent.

p = P(drawing a fused bulb) = 4/(10 + 4) = 2/7

q = P(drawing a bulb which is not fused) = 1 – 2/7 = 5/7

X = event of drawing a fused bulb

X can take up the values 0, 1, 2, 3

P(X = 0) = P(getting zero fused bulbs in all draws)

= nCr pr q(n – r)

= 3C0 (2/7)0 (5/7)(3 – 0)

= 1 × 1 × (125/343) = 125/343

P(X = 1) = P (getting one time fused bulb)

= nCr pr q(n – r)

= 3C1 (2/7)1 (5/7)(3 – 1)

= 3 × (2/7) × (25/49) = 150/343

P(X = 2) = P(getting two times fused bulbs)

= nCr pr q(n – r)

= 3C2 (2/7)2 (5/7)(3 – 2)

= 3 × 4/49 × (5/7) = 60/343

P(X = 3) = (P(getting three times fused bulb)

= nCr pr q(n – r)

= 3C3 (2/7)3 (5/7)(3 – 3)

= 1 × 8/343 × 1 = 8/343

The required probability distribution:

| **X** | **0** | **1** | **2** | **3** |
| --- | --- | --- | --- | --- |
| **P(X)** | **125/343** | **150/343** | **60/343** | **8/343** |

**Question 5:**

On average, every one out of 10 telephones is found busy. Six telephone numbers are selected at random. Find the probability that four of them will be busy.

**Solution:**

Let X: event of getting a busy phone number

p = P(probability of getting a phone number busy) = 1/10

q = P(probability of not getting a phone number busy) = 9/10

The required probability = P(X = 4) = 6C4 p4 q(6 – 4)

= 15 × (1/10)4 × (9/10)2

= 15 × 81/106

= 0.001215.

**Question 6:**

An unbiased dice is thrown until three sixes are obtained. Find the probability of obtaining the third six in the sixth throw.

**Solution:**

Since each throw is independent of the previous throws, we can apply the binomial distribution formula to find the probability.

p = P(getting a six in a throw) = ⅙

q = P(not getting a six in a throw) = 1 – ⅙ = ⅚

According to the question, two sixes are already obtained in the previous throws.

∴ Required probability = P(getting exactly two sixes in five throws) × P(getting a six in the sixth throw)

= 5C2 p2q3 × 1C1p1 q1 – 1

= 10 × (⅙)2 × (⅚)3 × 1 × (⅙)

= 10 × (⅙)3 × (⅚)3

= 625/23328.

**Question 7:**

The probability of a boy guessing a correct answer is ¼. How many questions must he answer so that the probability of guessing the correct answer at least once is greater than ⅔?

**Solution:**

p = P(guessing a correct answer) = ¼

q = P(not guessing a correct answer) = ¾

Let him answers n number of questions, then

P(X ≥ 1) = P(guessing at least one correct answer out of n questions) = 1 – P(no success) = 1 – qn

Given, 1 – qn > ⅔ ⇒ 1 – (¾)n > ⅔

⇒ (¾)n < ⅓

Now, let us check the above inequality for different values of n = 1, 2, 3, 4, …

When n = 1

¾ ≮ ⅓

When n = 2

(¾)2 ≮ ⅓

When n = 3

(¾)3 ≮ ⅓

When n = 4

(¾)4 < ⅓.

Thus, he must answer at least 4 questions.

**Question 8:**

When a biased coin is tossed, the probability of getting a head 3 times more than the probability of getting a tail. Find the probability distribution for getting a tail, if the coin is tossed twice.

**Solution:**

Let the probability of getting a tail be p, then the probability of getting a head will be 3p

Now, p + 3p = 1 ⇒ p = ¼

q = P(not getting a tail) = 1 – ¼ = ¾

X = event of getting a tail in a toss

Then, possible values of x will be 0, 1, 2

P(X = 0) = 2C0 p0 q2 – 0

= 1 × 1 × (¾)2

= 9/16

P(X = 1) = 2C1 p1 q2 – 1

= 2 × (¼) × (¾)

= ⅜

P(X = 2) = 2C2 p2 q2 – 2

= 1 × (¼)2 × 1

= 1/16

The probability distribution for getting the tail is:

| **X** | **0** | **1** | **2** |
| --- | --- | --- | --- |
| **P(X)** | **9/16** | **3/8** | **1/16** |

**Question 9:**

A bag contains 5 green balls and 3 red balls. If two balls are drawn from the bag randomly with replacement, find the probability distribution of the number of green balls drawn.

**Solution:**

Let p = P(getting a green ball) = 5/(5 + 3) = 5/8

q = P(not getting a green ball) = 1 – 5/8 = 3/8

X = event of drawing the green ball, then the value of X could be 0, 1, 2

P(X = 0) = Probability of getting no green ball = 2C0 p0 q2 – 0 = 1 × 1 × (3/8)2 = 9/64

P(X = 1) = Probability of getting one green ball = 2C1 p1 q2 – 1 = 2 × (⅝) × (⅜) = 15/32

P(X = 2) = Probability of getting 2 green balls = 2C2 p2 q2 – 2 = 1 × (⅝)2 × (⅜)0 = 25/64

The required probability distribution is:

| **X** | **0** | **1** | **2** |
| --- | --- | --- | --- |
| **P(X)** | **9/64** | **15/32** | **25/64** |

**Question 10:**

Find the probability distribution of getting the number of fours in three throws of a dice. Also, find the mean and variance of the distribution.

**Solution:**

Let, p = P(getting a four in a throw of dice) = ⅙

q = P(not getting a four in a throw of dice) = ⅚

X: number of four obtained, then the value of X could be 0, 1, 2, 3.

P(X = 0) = 3C0 p0q3 – 0 = 1 × (⅚)3 = 125/216

P(X = 1) = 3C1 p1q3 – 1 = 3 × (⅙) × (⅚)2 = 75/216

P(X = 2) = 3C2 p2q3 – 2 = 1 × (⅙)2 × (⅚)3 – 2 = 15/216

P(X = 3) = 3C3 p3q3 – 3 = 1 × (⅙)3 × (⅚)3 – 3 = 1/216

The required probability distribution

| **X** | **0** | **1** | **2** | **3** |
| --- | --- | --- | --- | --- |
| **P(X)** | **125/216** | **75/216** | **15/216** | **1/216** |

Mean = np = 3 × ⅙ = 1.2

Variance = npq = 3 × ⅙ × ⅚ = 5/12.

The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables. The advantage of the CDF is that it can be defined for any kind of random variable (discrete, continuous, and mixed).

**Definition**  
The cumulative distribution function (CDF) of random variable X is defined as

FX(x)=P(X≤x), for all x∈R.

Note that the subscript X indicates that this is the CDF of the random variable X. Also, note that the CDF is defined for all x∈R. Let us look at an example.

**Example**

I toss a coin twice. Let X be the number of observed heads. Find the CDF of X.

* **Solution**
  + Note that here X∼Binomial(2,1/2). The range of  is RX={0,1,2}={0,1,2}

PX(0)=P(X=0)=1/4

PX(1)=P(X=1)=1/2,

PX(2)=P(X=2)=1/4.

To find the CDF, we argue as follows. First, note that if x<0, then

FX(x)=P(X≤x)=0, for x<0.

Next, if x≥2

FX(x)=P(X≤x)=1, for x≥2.

Next, if 0≤x<1

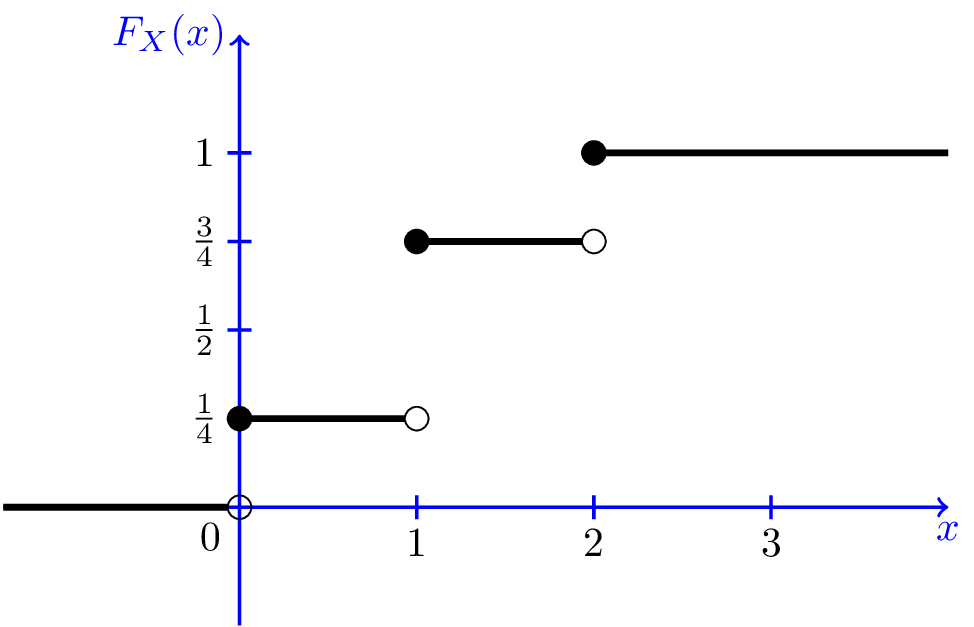
FX(x)=P(X≤x)=P(X=0)=14, for 0≤x<1

Finally, if 1≤x<2

FX(x)=P(X≤x)=P(X=0)+P(X=1)=1/4+1/2=3/4, for 1≤x<2

Thus, to summarize, we have

Note that when you are asked to find the CDF of a random variable, you need to find the function for the entire real line. Also, for discrete random variables, we must be careful when to use "<<" or "≤≤". Figure 3.3 shows the graph of FX(x). Note that the CDF is flat between the points in RX and jumps at each value in the range. The size of the jump at each point is equal to the probability at that point. For, example, at point x=1, the CDF jumps from ¼  to 3/4. The size of the jump here is 3/4−1/4=1/2 which is equal to PX(1). Also, note that the open and closed circles at point x=1 indicate that FX(1)=3/4 and not 1/4.

Fig.3.3 - CDF for Example 3.9.