# Random Variables and Distributions, Bernoulli and binomial distribution

STA 032: Gateway to data science Lecture 14

Jingwei Xiong

May 3, 2023

# Recap

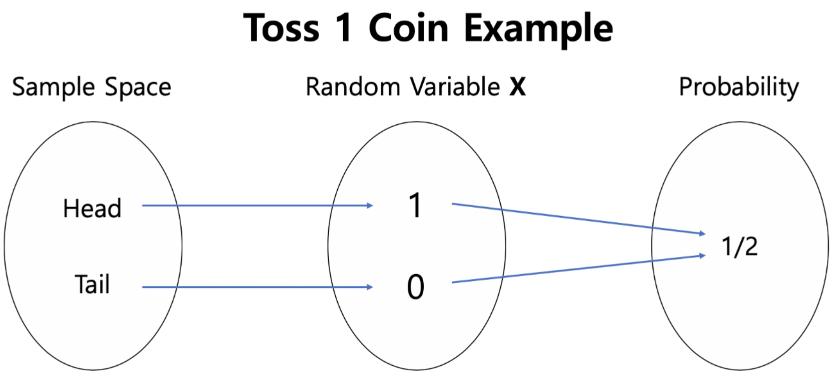
- Conditional probability
    - General multiplication rule: $P(A \cap B) = P(B)P(A|B)$
    - Sum of conditional probabilities: $P(A_1|B) + \ldots + P(A_k|B) = 1$
    - Law of total probability: $P(B) = P(B \cap A_1) + \ldots + P(B \cap A_k) = P(B \mid A_1)P(A_1) + \ldots + P(B \mid A_k)P(A_k)$
- Marginal and joint probability
- Revisiting independence
    - $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
- Bayes' Theorem
    - $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$

# Today

- Random variables
  - Expectation and variance
  - Discrete and continuous random variables
- Common probability distributions

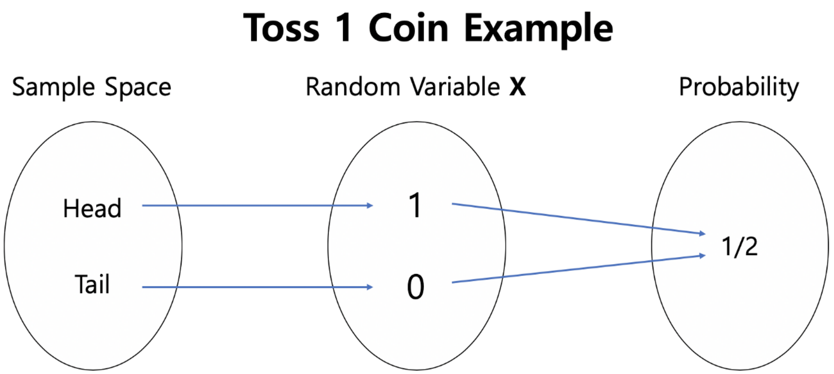
# Random variables

- A random variable is a mapping or a function from possible outcomes in a sample space to a probability space.
- Recall: **sample space** is the set of all possible outcomes from a **random process**




Source: https://medium.com/jun94-devpblog/prob-stats-1-random-variable-483c45242b3c

# Random variables




Source: https://medium.com/jun94-devpblog/prob-stats-1-random-variable-483c45242b3c

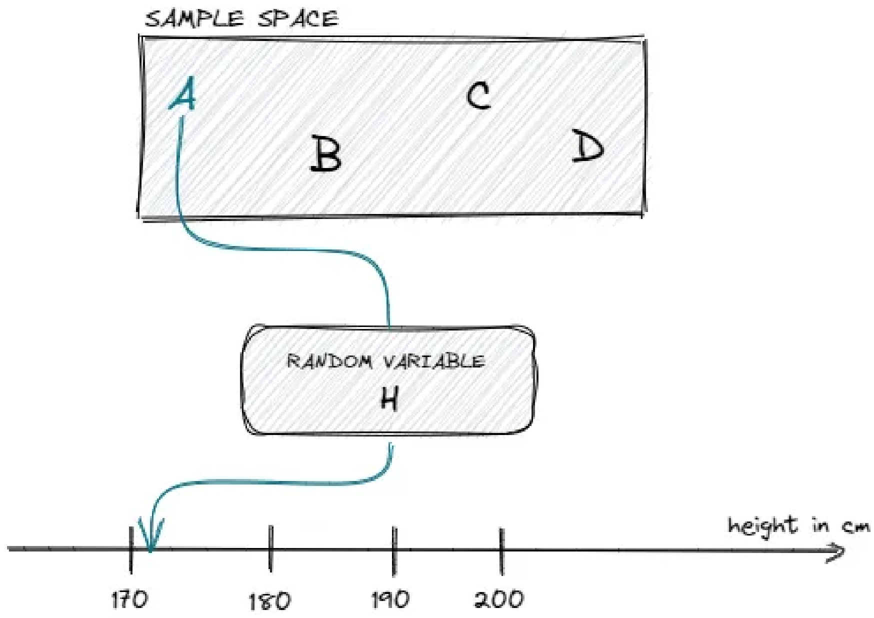
- Let $X$ be the random variable indicating whether a coin flip results in heads.
- Instead of saying $P(\texttt{heads})$, we say $P(X = 1)$
- This representation allows us to apply mathematical frameworks and get a better understanding of real-world phenomenon

# Random variables

- Random variables are usually denoted by capital letters, most commonly $X$, $Y$, $Z$
- A **realization** or draw of the random variable is denoted by a lowercase letter, $x$, $y$, $z$
- Other examples of random variables:
  - Mass of classroom chairs
  - Ages of students at UC Davis
- For discrete random variables:
  - Each outcome has an associated probability $P(X = x_i)$ where $i = 1, \ldots, k$ ( $k$ outcomes are denoted by lower-case, $x_1, \ldots, x_k$)
  - Sometimes also written as $p_1, \ldots, p_k$

Another example:

Let's assume, we have a sample space containing 4 students `{A, B, C, D}`. If we now randomly pick student `A` and measure the height in centimeters, we can think of the random variable (`H`) as the function with the input of student and the output of height as a real number.




Depending on the outcome — which student is randomly picked — our random variable (`H`) can take on different states or different values in terms of height in centimeters.

# Probability Distribution

The description of how likely a random variable takes one of its possible states can be given by a probability distribution.

Thus, the probability distribution is a mathematical function that gives the probabilities of different outcomes for a random variable.

- More generally it can be described as the function

$$P : A \to \mathbb{R}$$

Which maps an input space $A$ — related to the sample space — to a real number, namely the probability.

For the above function to characterize a probability distribution, it must follow all of the axioms:

1. All probabilities must between 0 and 1. $0 \leq p(x) \leq 1$
2. All probabilities must summed up to be 1. $\sum p(x) = 1$

# Discrete RV distribution

Discrete RV: it can only take on a finite number of values within an interval, and the value has the natural gaps

The **probability mass function (PMF)** describes the probability distribution over a **discrete** random variable.

> In other terms, it is a function that returns the probability of a random variable being exactly equal to a specific value.
> $P(X = i) = P(i)$

The **cumulative distribution function (CDF)** describes the probability that a random variable is less than or equal than a given value.

> $F(x) = P(X \leq x) = \sum_{X \leq x} P(X = x)$

# probability mass function (PMF) for Discrete RV

- Example: Two books are assigned for a statistics class: a textbook and its corresponding study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one quarter to another.

- Formally: Let $X =$ number of books sold per student

- The three possible outcomes are $x_1$ = 0 books, $x_2$ = 1 book (1 textbook for each student), $x_3$ = 2 books (1 textbook and 1 study guide for each student)

| **i** | **1** | **2** | **3** |
|---|---|---|---|
| $x_i$ | 0 | 1 | 2 |
| $P(X = x_i)$ | .2 | .55 | .25 |

# CDF for discrete RV

- Example: Two books are assigned for a statistics class: a textbook and its corresponding study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one quarter to another.

- Formally: Let $X =$ number of books sold per student

- The three possible outcomes are $x_1$ = 0 books, $x_2$ = 1 book (1 textbook for each student), $x_3$ = 2 books (1 textbook and 1 study guide for each student)

| **i** | **1** | **2** | **3** |
|---|---|---|---|
| $x_i$ | 0 | 1 | 2 |
| $P(X = x_i)$ | .2 | .55 | .25 |
| $F(X \leq x_i)$ | .2 | .75 | 1 |

# Expectation

Two books are assigned for a statistics class: a textbook and its corresponding study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one quarter to another.

- How many books should the bookstore expect to sell per student?
- Intuitively: $.2 * 0 + .55 * 1 + .25 * 2 = 1.05$
- Another way to think about it: say the class has 100 students. How many books should the bookstore expect to sell to the class?

# Expectation

- Formally: Let $X =$ number of books sold per student
- The three possible outcomes are $x_1$ = 0 books, $x_2$ = 1 book (1 textbook for each student), $x_3$ = 2 books (1 textbook and 1 study guide for each student)

| **i** | **1** | **2** | **3** |
|---|---|---|---|
| $x_i$ | 0 | 1 | 2 |
| $P(X = x_i)$ | .2 | .55 | .25 |

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + x_2 \times P(X = x_2) + \ldots + x_k \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i P(X = x_i) \end{aligned}$$

- Using this definition: $E(X) = 0 * .2 + 1 * .55 + 2 * .25 = 1.05$.

# Expectation

- Say we are interested in the amount of revenue that the bookstore can expect to earn per student. Say the textbook costs \$ 137 and the study guide costs \$ 33.

- What modifications do we need?

  - Formally: Let $X =$ number of books sold per student

  - The three possible outcomes are $x_1$ = 0 books, $x_2$ = 1 books (1 textbook for each student), $x_3$ = 2 books (1 textbook and 1 study guide for each student)

| **i** | **1** | **2** | **3** |
|---|---|---|---|
| $x_i$ | 0 | 1 | 2 |
| $P(X = x_i)$ | .2 | .55 | .25 |

- Using this definition: $E(X) = 0 * .2 + 1 * .55 + 2 * .25 = 1.05$.

# Expectation

- Say we are interested in the amount of revenue that the bookstore can expect to earn per student. Say the textbook costs \$ 137 and the study guide costs \$ 33.

- What modifications do we need?

    - Formally: Let $X =$ revenue from books sold per student

    - The three possible outcomes are $x_1$ = \$ 0, $x_2$ = \$ 137 (1 textbook for each student), $x_3$ = \$ 170 (1 textbook and 1 study guide for each student)

| **i** | **1** | **2** | **3** |
|---|---|---|---|
| $x_i$ | 0 | 137 | 170 |
| $P(X = x_i)$ | .2 | .55 | .25 |

- Using this definition: $E(X) = 0 * .2 + 137 * .55 + 170 * .25 = 117.85$.

# Expectation

- The expectation is denoted by $E(X)$, $\mu$ or $\mu_x$
- This is the expected or average outcome of $X$, where $X$ is a random variable
- Given a probability distribution for a discrete random variable, we can calculate it using

$$
\begin{aligned}
E(X) &= x_1 \times P(X = x_1) + x_2 \times P(X = x_2) + \ldots + x_k \times P(X = x_k) \\
&= \sum_{i=1}^{k} x_i P(X = x_i)
\end{aligned}
$$

- Recall: this is a **population parameter**, a fixed quantity
  - The sample version, the **sample statistic**, is the sample mean $\bar{x}$

# Properties of the expectation

- $E[c] = c$, where c is a constant
- $E[aX] = aE[X]$
- $E[aX + c] = aE[X] + c$
- To calculate $E(X^2)$ (will be useful later), simply replace $x_i$ in the sum by $x_i^2$, i.e.,

$$\begin{aligned} E(X^2) &= x_1^2 \times P(X = x_1) + x_2^2 \times P(X = x_2) + \ldots + x_k^2 \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i^2 P(X = x_i) \end{aligned}$$

- More generally, $E[g(X)] = \sum_{i=1}^{k} g(x_i) P(X = x_i)$ (Law of the unconscious statistician)

# Variance

- Recall: we saw the **sample variance**, calculated for a data set
  - Take the square of deviations and find the mean
  - $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n-1}$
- **Population variance** is often denoted by $\sigma^2$, $\sigma_x^2$, or $Var(X)$
- Given a probability distribution for a discrete random variable, we can calculate it using

$$\begin{aligned} Var(X) &= E[(X-\mu)^2] \\ &= (x_1-\mu)^2 \times P(X=x_1) + (x_2-\mu)^2 \times P(X=x_2)+\ldots+(x_k-\mu)^2 \times P(X=x_k) \\ &= \sum_{i=1}^{k}(x_i-\mu)^2 P(X=x_i) \end{aligned}$$

- Note: rather than summing over observations, these are over possible outcomes, weighted by their probabilities

# Variance

Another common way to write the variance is

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] \\ &= E[(X^2 - 2XE(X) + [E(X)]^2)] \\ &= E(X^2) - 2E(X)E(X) + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

- Recall: $E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$
- To calculate $E(X^2)$, simply replace $x_i$ in the sum above by $x_i^2$, i.e.,

$$\begin{aligned} E(X^2) &= x_1^2 \times P(X = x_1) + x_2^2 \times P(X = x_2) + \ldots + x_k^2 \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i^2 P(X = x_i) \end{aligned}$$

# Properties of the variance

- $Var[c] = 0$, where c is a constant
- $Var[aX] = a^2 Var[X]$
- $Var[aX + c] = a^2 Var[X]$

# Linear combinations of random variables

- Often we care not just about a single random variable, but a combination of them
- E.g.,
  - The total revenue of our bookstore is a combination of books from different classes, not just our one statistics class
  - The total gain or loss in a stock portfolio is the sum of the gains and losses in its components
  - Total weekly commute time is a combination of daily commute
- Let $W$ be the weekly commute time per student at UC Davis
  - $X_1$ = commute time per student on Monday
  - $X_2$ = commute time per student on Tuesday
  - ...
  - $X_5$ = commute time per student on Friday
  - $W = X_1 + X_2 + \ldots + X_5$ is also a random variable

# Linear combinations of random variables

- A **linear combination** of two random variables, $X$ and $Y$, is of the form

$$aX + bY,$$

where $a$ and $b$ are constants.

- $a$ and $b$ are also called coefficients

- In our example, $W = X_1 + X_2 + \ldots + X_5$ is a linear combination with coefficients 1.

# Expectation of linear combinations of random variables

- The expectation for a linear combination of random variables is given by

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

- In our example, say $E(X_1) = \ldots = E(X_5) = 21$ minutes.
- Then, $E(W) = 1 * 21 + 1 * 21 + 1 * 21 + 1 * 21 + 1 * 21 = 105$ minutes.

# Variance of linear combinations of random variables

- The variance for a linear combination of **independent** random variables

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y)$$

- Note: this is only true if $X$ and $Y$ are independent.
- In our example, say $Var(X_1) = \ldots = Var(X_5) = 5$ minutes.
- Commute times on each day of the week are independent.
- Then, $Var(W) = 1^2 * 5 + 1^2 * 5 + 1^2 * 5 + 1^2 * 5 + 1^2 * 5 = 25$ minutes.

# Bernoulli random variable

- Note on terminology: when we say we have a Bernoulli random variable, we mean that the random variable follows a Bernoulli distribution

- Same with normal (or Gaussian) random variable, …

- Consider a binary random variable $Y$. By definition, $Y$ must assume one of two possible values, e.g.

    - failure or success
    - dead or alive
    - UC Davis student or not
    - current smoker or not
    - heads or tails (coin flip)
    - Y chromosome or not

- A random variable of this type is known as a **Bernoulli** random variable, and we describe the probability of response using the parameter $\pi$ or $p$.

# Bernoulli random variable

- $Y = 1$ is often called a "success," $Y = 0$ is called a "failure", and $\pi$ or $p$ is defined as the probability of a success, i.e., $P(Y = 1)$.
- The probability of a "failure," $P(Y = 0)$ is then $1 - p$
- Examples:
    - Coin flip: let $Y = 1$ if heads and $Y = 0$ if tails, then $P(Y = 1) = p = 0.5$
    - Vegetarian in US: $Y = 1$ if vegetarian and $Y = 0$ if not, then $P(Y = 1) = p = 0.05$ and $P(Y = 0) = 1 - p = 1 - 0.05 = 0.95$
    - Vegetarian in India: $Y = 1$ if vegetarian and $Y = 0$ if not, then $P(Y = 1) = p = 0.31$ and $P(Y = 0) = 1 - p = 1 - 0.31 = 0.69$

# Bernoulli random variable

- Probability mass function for a Bernoulli distributed random variable is $P(Y = y) = p^y(1 - p)^{1-y}$
  - $P(Y = 1) = p^1(1 - p)^0 = p$ (remember $x^0 = 1$ for any $x$)
  - $P(Y = 0) = p^0(1 - p)^1 = 1 - p$
- For the Bernoulli random variable, we don't really need this formality
- However, we want to extend this to more complex settings
- $E(Y) = \sum_{i=1}^{k} y_i P(Y = y_i) = p$ and $Var(Y) = p(1 - p)$

# From Bernoulli to binomial...

- $Y$ takes value 1 with probability $p$ and value 0 with probability $1-p$
- $P(Y=y) = p^y(1-p)^{1-y}$
    - $P(Y=1) = p^1(1-p)^0 = p$ (remember $x^0 = 1$ for any $x$)
    - $P(Y=0) = p^0(1-p)^1 = 1-p$
- For the Bernoulli random variable, we don't really need this formality
- However, we want to extend this to more complex settings
- For example, in a randomly-selected group of 3 high school students, how surprising would it be to get 2 who have smoked e-cigarettes in the past month?
- Could consider three draws from a Bernoulli distribution

# Case Study: E-Cigarettes

- The CDC reports that 19.6% of high school students have smoked e-cigarettes in the past 30 days. We'll round this to 20% for simplicity.
- $P(Y = 1) = P(Smoker) = p = 0.2$ and $P(Y = 0) = 0.8$
- Now suppose we randomly select two independent high school students and define a new random variable $X$ representing the number of smokers. X can take the values 0, 1, or 2.
- Let $Y_1$ be the smoking status of the first student and $Y_2$ be the smoking status of the second student, where $Y_j = 1$ if student $j$ smokes and 0 otherwise.

Next we'll talk about how to get the *probability distribution* of $X$.

| $Y_1$ | $Y_2$ | $X$ | $P(X)$ |
|---|---|---|---|
| 0 | 0 | 0 | |
| 1 | 0 | 1 | |
| 0 | 1 | 1 | |
| 1 | 1 | 2 | |

# Case Study: E-Cigarettes

- Recall: If events A and B are independent, then $P(A \cap B) = P(A) \times P(B)$.
- Let $A_1$ be the event that $Y_1 = 1$ and let $A_2$ be the event that $Y_2 = 1$.
- Since the students are independent,

$$\begin{aligned} P(Y_1 = Y_2 = 1) &= P(A_1 \cap A_2) \\ &= P(A_1)P(A_2) \\ &= p \times p \\ &= 0.2(0.2) \\ &= 0.04. \end{aligned}$$

# Case Study: E-Cigarettes

Now we can fill in the bottom row of the probability distribution of $X$.

| $Y_1$ | $Y_2$ | $X$ | $P(X)$ |
|---|---|---|---|
| $0$ | $0$ | $0$ | |
| $1$ | $0$ | $1$ | |
| $0$ | $1$ | $1$ | |
| $1$ | $1$ | $2$ | $0.02 \times 0.02 = 0.04$ |

# Case Study: E-Cigarettes

It's straightforward to fill in the rest of the table in the same way

| $Y_1$ | $Y_2$ | $X$ | $P(X)$ |
|---|---|---|---|
| $0$ | $0$ | $0$ | $0.8 \times 0.8 = 0.64$ |
| $1$ | $0$ | $1$ | $0.2 \times 0.8 = 0.16$ |
| $0$ | $1$ | $1$ | $0.8 \times 0.2 = 0.16$ |
| $1$ | $1$ | $2$ | $0.2 \times 0.2 = 0.04$ |

# Case Study: E-Cigarettes

Recall our table:

| $Y_1$ | $Y_2$ | $X$ | $P(X)$ |
|---|---|---|---|
| 0 | 0 | 0 | $0.8 \times 0.8 = 0.64$ |
| 1 | 0 | 1 | $0.2 \times 0.8 = 0.16$ |
| 0 | 1 | 1 | $0.8 \times 0.2 = 0.16$ |
| 1 | 1 | 2 | $0.2 \times 0.2 = 0.04$ |

We can clean up the table to get the probability distribution of $X$:

| | | | |
|---|---|---|---|
| $X$ | 0 | 1 | 2 |
| $P(X = x)$ | 0.64 | 0.32 | 0.04 |

So if we randomly sample two US high schoolers, the probability that both are recent e-cig smokers is 0.04 (4% chance), the probability only one recently smoked is 0.32 (this can happen two ways -- either only the first smoked or only the second smoked), and the probability neither smoked e-cigs recently is 0.64.

# Case Study: E-Cigarettes

Now suppose we randomly sample 3 independent high school students

| $Y_1$ | $Y_2$ | $Y_3$ | $X$ | $P(X)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | |
| 1 | 0 | 0 | 1 | |
| 0 | 1 | 0 | 1 | |
| 0 | 0 | 1 | 1 | |
| 1 | 1 | 0 | 2 | |
| 1 | 0 | 1 | 2 | |
| 0 | 1 | 1 | 2 | |
| 1 | 1 | 1 | 3 | |

# Case Study: E-Cigarettes

Because these are independent high school students, we can calculate the probabilities in the same manner as before.

| $Y_1$ | $Y_2$ | $Y_3$ | $X$ | $P(X)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0.8(0.8)(0.8)=0.512 |
| 1 | 0 | 0 | 1 | 0.2(0.8)(0.8)=0.128 |
| 0 | 1 | 0 | 1 | 0.8(0.2)(0.8)=0.128 |
| 0 | 0 | 1 | 1 | 0.8(0.8)(0.2)=0.128 |
| 1 | 1 | 0 | 2 | 0.2(0.2)(0.8)=0.032 |
| 1 | 0 | 1 | 2 | 0.2(0.8)(0.2)=0.032 |
| 0 | 1 | 1 | 2 | 0.8(0.2)(0.2)=0.032 |
| 1 | 1 | 1 | 3 | 0.2(0.2)(0.2)=0.008 |

The probability that 2 of 3 are recent e-cig smokers is $0.032 + 0.032 + 0.032 = 0.096$ or $9.6\%$

# Case Study: E-Cigarettes

| $Y_1$ | $Y_2$ | $Y_3$ | $X$ | $P(X)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0.8(0.8)(0.8)=0.512 |
| 1 | 0 | 0 | 1 | 0.2(0.8)(0.8)=0.128 |
| 0 | 1 | 0 | 1 | 0.8(0.2)(0.8)=0.128 |
| 0 | 0 | 1 | 1 | 0.8(0.8)(0.2)=0.128 |
| 1 | 1 | 0 | 2 | 0.2(0.2)(0.8)=0.032 |
| 1 | 0 | 1 | 2 | 0.2(0.8)(0.2)=0.032 |
| 0 | 1 | 1 | 2 | 0.8(0.2)(0.2)=0.032 |
| 1 | 1 | 1 | 3 | 0.2(0.2)(0.2)=0.008 |

The probability distribution of $X$, the number of recent e-cig smokers out of three high school students, is now

| | | | | |
|---|---|---|---|---|
| $X$ | 0 | 1 | 2 | 3 |
| $P(X = x)$ | 0.512 | 0.384 | 0.096 | 0.008 |

# Case Study: E-Cigarettes

- Extending to 4 and more students, we can see why computing the probablities by hand, as we've done, is intractable

- We can use the **binomial distribution** to describe this random variable

# Binomial random variable

- The **binomial distribution** gives us the probability of $X$ "successes" from a sequence of $n$ independent Bernoulli trials (size $n$). This is often denoted binomial(n, p).

- In our example, each student would represent an independent Bernoulli trial (either an e-cig smoker, or not).

    - 1 draw from the binomial distribution is made of 3 independent draws from the Bernoulli distribution

- This distribution involves three assumptions.

    - There is a fixed number $n$ of Bernoulli trials, each of which results in one of two mutually-exclusive outcomes

    - The outcomes of the $n$ trials are independent

    - The probability of success $p$ is the same for each trial

# Binomial distribution

- The probability mass function for the binomial distribution is given by $P(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$

- Compare this with $P(Y = y) = p^y(1-p)^{1-y}$ for the Bernoulli distribution.

- First, look at the second part, $p^x(1-p)^{n-x}$. This is just multiplying the right combination of $p$ and $1-p$ as in the previous tables.

  - There will be a total of $n$ terms being multiplied, one probability for each draw of the distribution (each student in this case)

  - For example, if we want the probability of 3 e-cig smokers, $X = 3$, the second part is $p^x(1-p)^{n-x} = 0.2^3(0.8)^0 = 0.008$, just as in the table.

# Binomial distribution

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

- If we want the probability of 2 e-cig smokers and 1 non-smoker, i.e., $x = 2$, the second part is $p^x(1-p)^{n-x} = 0.2^2(0.8)^{3-2} = 0.032$, which is what we see in any single row in which we have two smokers and one non-smoker.
  - This is the probability of any one specific combination of 2 smokers and 1 nonsmoker. Then we need to figure out how many combinations of 2 smokers and 1 nonsmoker we could get.
- The first part, $\binom{n}{x}$, accounts for all the possible ways in which we can have 2 smokers out of 3 people.

# Summary

- Random variables
  - Expectation and variance
  - Discrete and continuous random variables
- Common probability distributions
  - Bernoulli
  - Binomial distribution