Covariance and Correlation

PHS Launch

Sudipta Saha

Aug 26, 2022

Outline

1. **Expectation**
2. **Variance**
3. **Covariance**
4. **Correlation**

But first, the important stuff!

- You are here for a reason!
- Sometimes learning new things can be scary, but that doesn't mean you can't do it! You can!
- You are capable and we believe in you!

(Thanks Gabe!)

The important stuff!

- Somtimes stats can seem like a topic where you just learn the "correct" answer. And this can make participating hard if you think you don't know what the "correct" answer is

The important stuff!

- Somtimes stats can seem like a topic where you just learn the "correct" answer. And this can make participating hard if you think you don't know what the "correct" answer is
- But at the end of the day we are just trying to understand and describe the world we are in. The quantitative approaches in this course are just one way of doing that.

The important stuff!

- Somtimes stats can seem like a topic where you just learn the "correct" answer. And this can make participating hard if you think you don't know what the "correct" answer is
- But at the end of the day we are just trying to understand and describe the world we are in. The quantitative approaches in this course are just one way of doing that.
- Sometimes the "correct" answer on a slide may not be what you had initially thought - but that's actually great for learning!

The important stuff!

- Somtimes stats can seem like a topic where you just learn the "correct" answer. And this can make participating hard if you think you don't know what the "correct" answer is
- But at the end of the day we are just trying to understand and describe the world we are in. The quantitative approaches in this course are just one way of doing that.
- Sometimes the "correct" answer on a slide may not be what you had initially thought - but that's actually great for learning!
- By contrasting different lines of reasoning we can highlight assumptions and norms in statistics that are often left unsaid
- So your participation makes us better collectively!!

Good Morning

Let's Get Started!

Expected Values

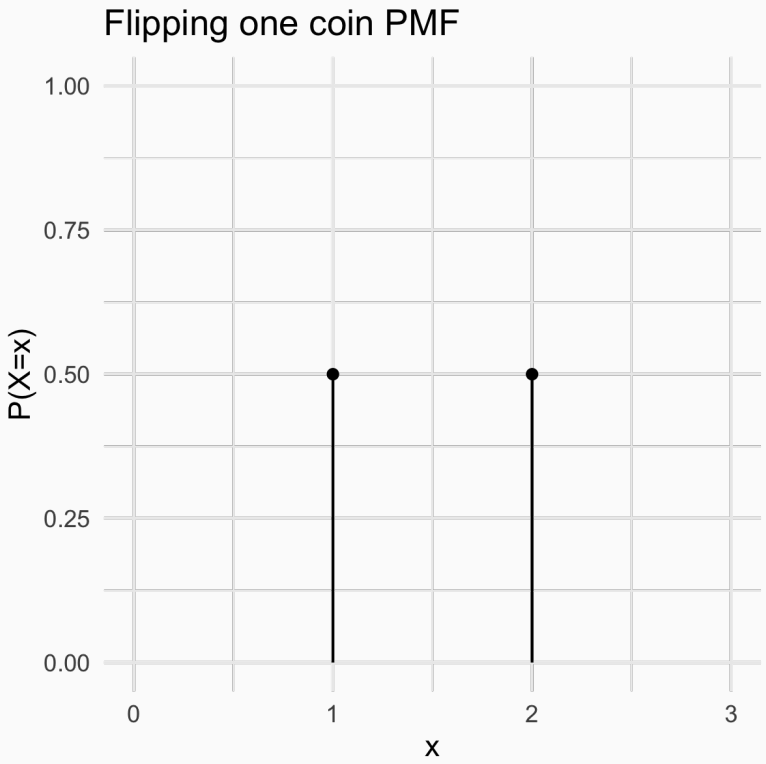
Revisiting Probability Distributions

Recall that the probability distribution of a **random variable** describes the probabilities of all the possible values the random variable can take.

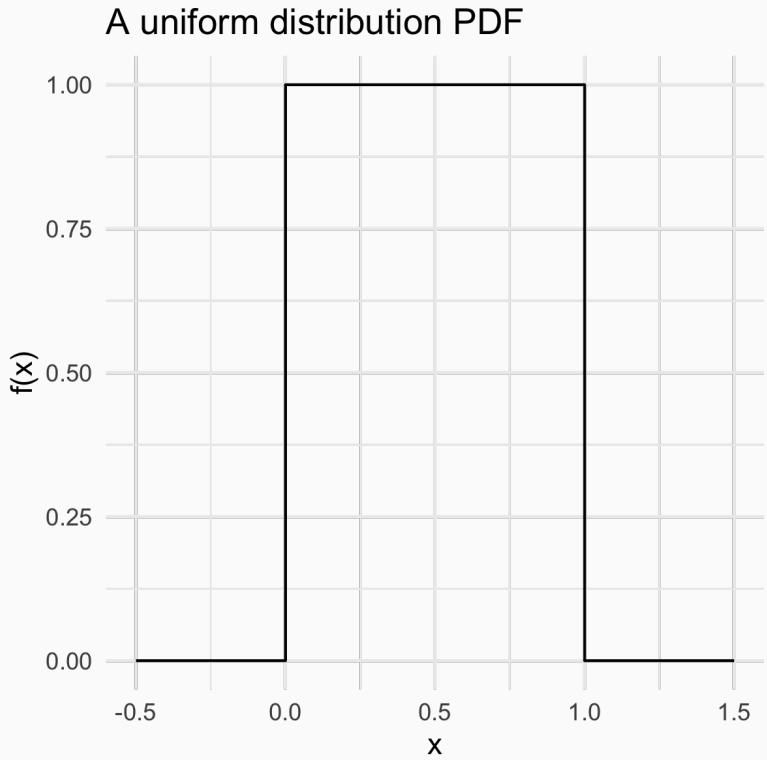
Recall that we talked about **probability mass functions** for discrete random variables, and **probability density functions** for continuous random variables.

Discrete and Continuous RVs

Discrete Case



Continuous Case



For each of these, what is the mean of the distribution? How are you thinking about this intuitively?

Expectation

The expectation or the **expected value** of a random variable X is written as $E[X]$. The $E[X]$ of the population is also the mean, μ

For a discrete random variable X, $E[X] = \sum_x xP(X = x)$

In words: it is a weighted average of all the values that X can take, with their probabilities as the weights

For a continuous random variable X, $E[X] = \int_{-\infty}^{\infty} xf(x)dx$, where the $f(x)$ is the PDF.

An example involving cookies and happiness

An example involving cookies and happiness

Imagine a world where HSPH has been taking on 1000 students per year for 20 years.

All of these 20,000 students are given a a bag of cookies during orientation. The quality control on these cookie-bags is not great, and each bag has a different number of cookies. Students are asked to rate their happiness upon receiving the cookie-bag on a scale of 1 to 10, with 10 being the happiest.

In this world, we, for some reason, recorded the number of cookies in each bag and the happiness of the recipient. We are interested in the relationship between *number of cookies received* and *happiness*.

Getting formal with cookies and happiness

X is the number of cookies in a bag received by a student

Y is the self-rated happiness

The size of the population, N, is 20,000

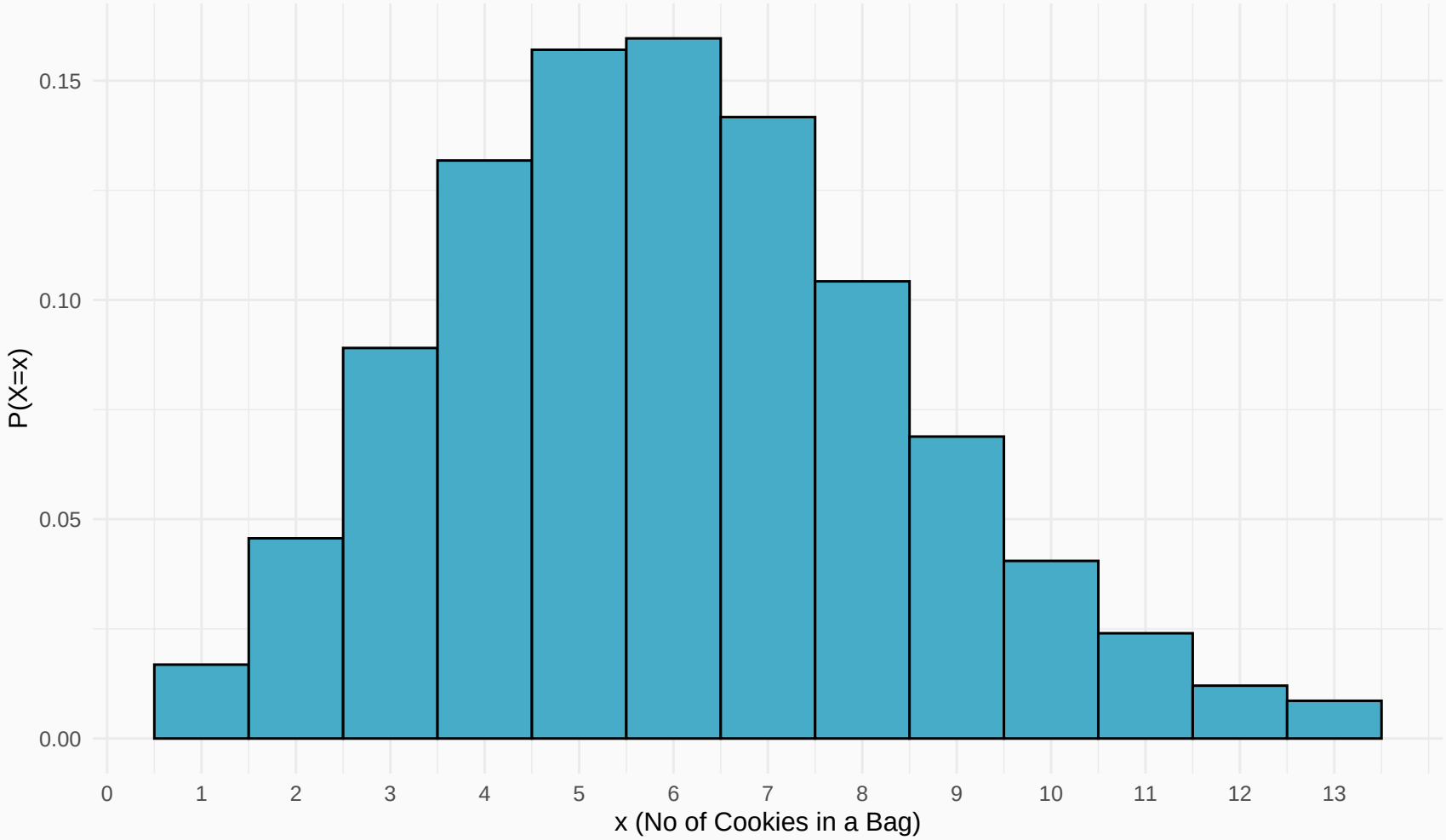
We are interested in the relationship between X and Y in this population.

Note that, in this hypothetical example, we are assuming this is the *total* population, whereas we usually worry about *samples* (more on this later).

Cookie Number Distribution

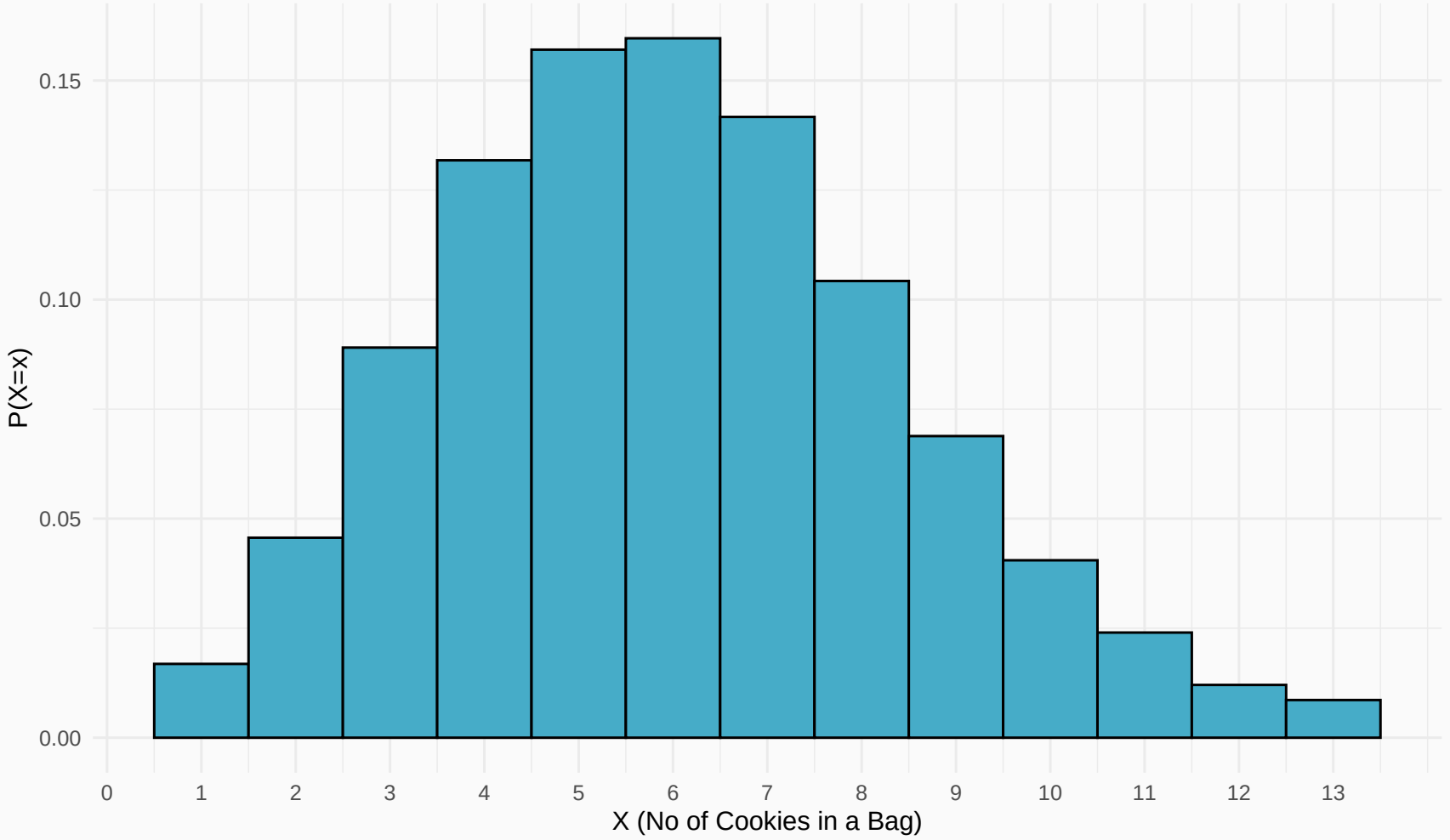
Number of Cookies in a Bag, X, is a discrete random variable. And in this case, we know P(X=x) for every value of x.

Let us plot this PMF.



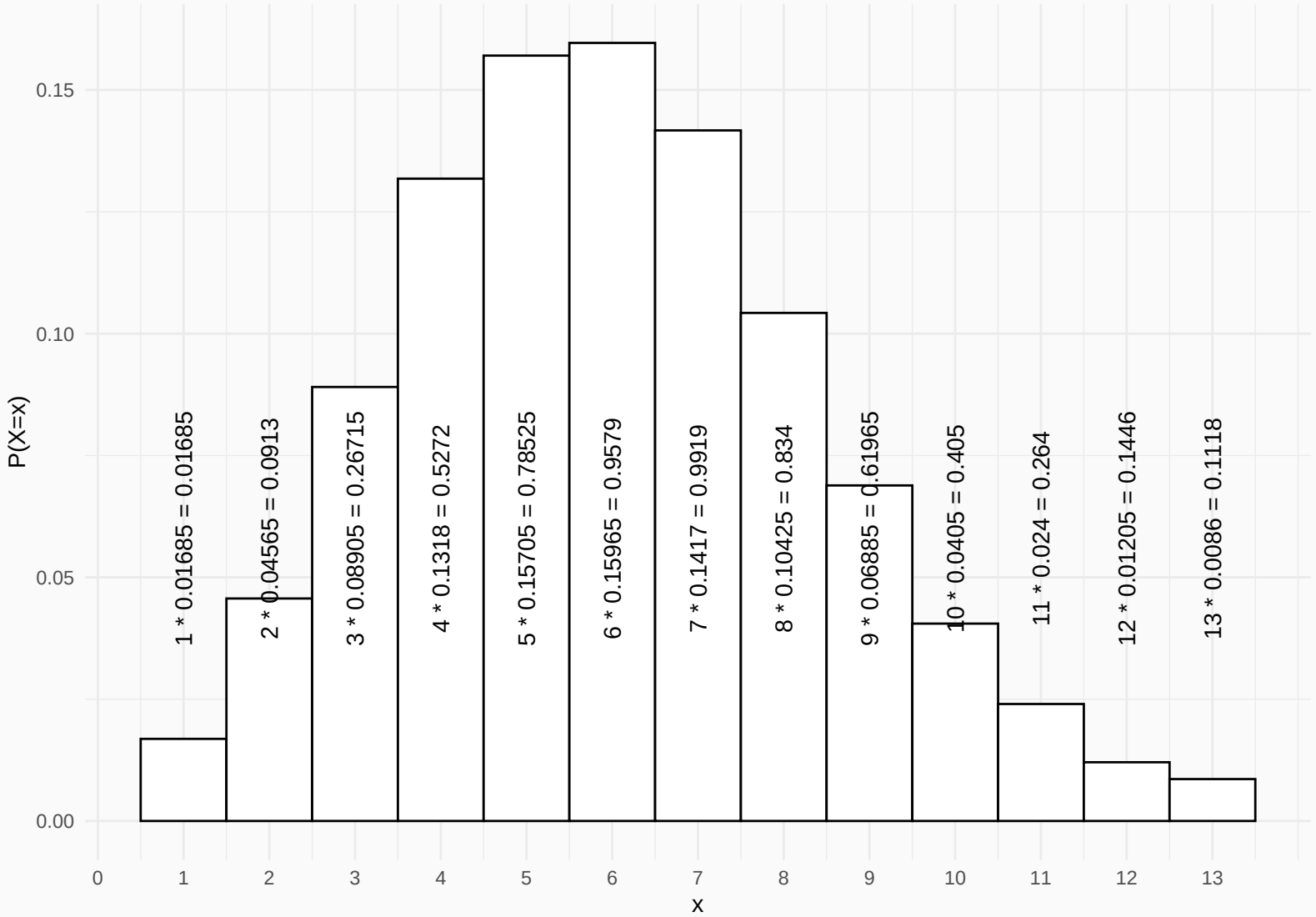
Expected number of cookies in bag

How would you go about calculating $E[X]$, the expected value of cookies in a bag?



Expected number of cookies in bag

We know that: $E[X] = \sum xP(X = x)$. Let's do this "by hand":



Properties of Expectations

It is useful to know some properties of expectations:

- If c is a constant, $E[c] = ?$
- $E[cX] = ?$
- $E[X + Y] = ?$
- $E[aX + bY] = ?$

Properties of Expectations

It is useful to know some properties of expectations:

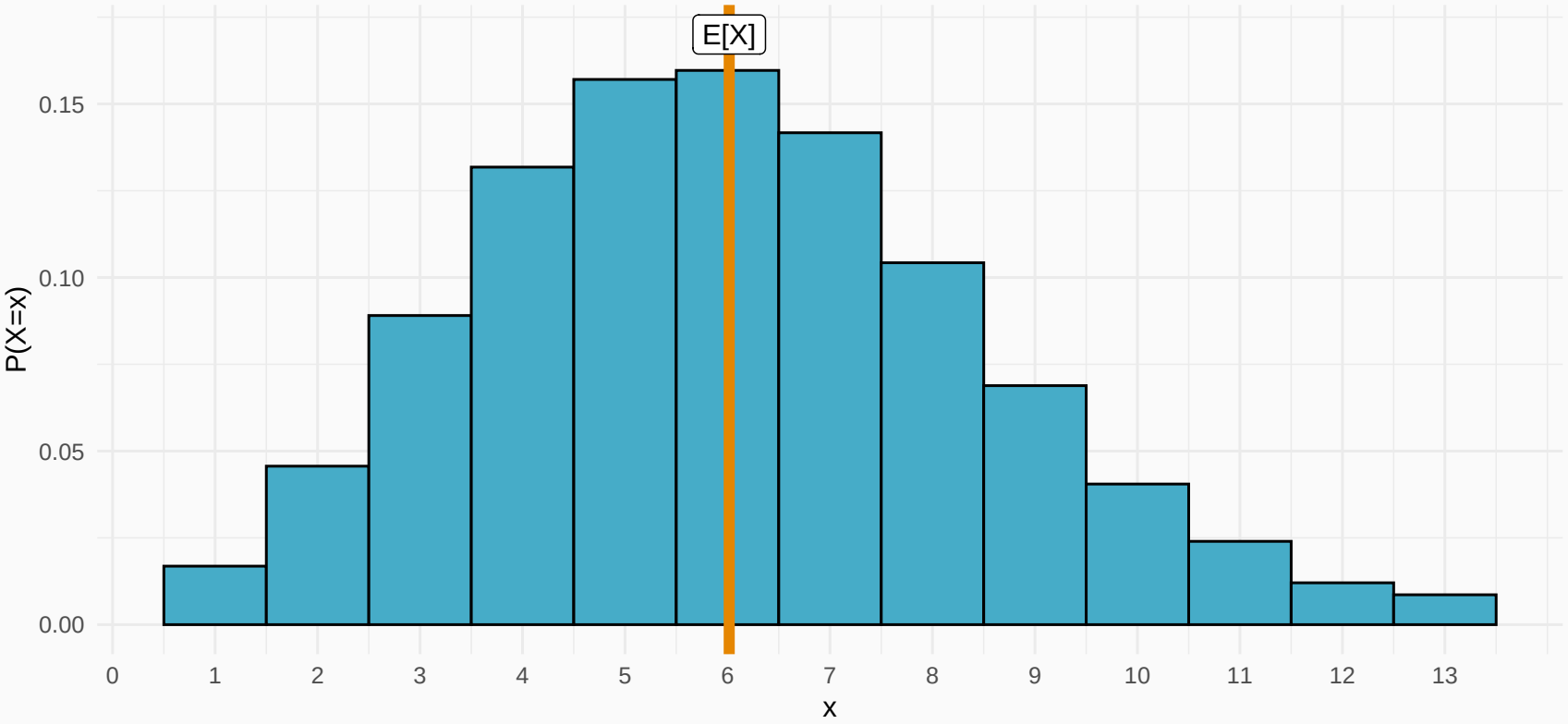
- If c is a constant, $E[c] = c$
- $E[cX] = cE[X]$
- $E[X + Y] = E[X] + E[Y]$
- $E[aX + bY] = aE[X] + bE(Y)$

Variance

The "spread" of the distribution

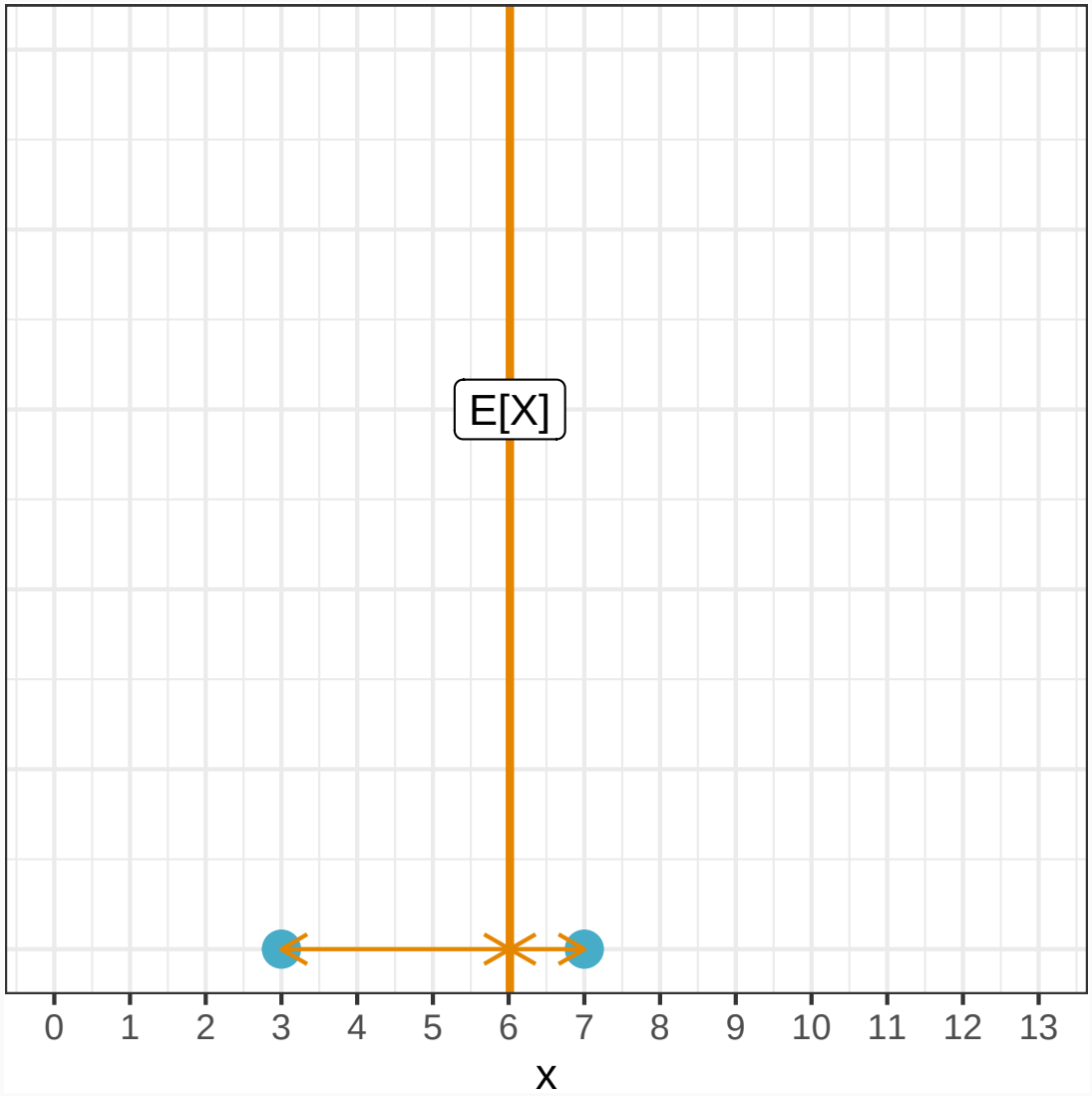
Another important characteristic is how "spread out" the distribution is.

After all, if you know the expected value is 6, you would also want to know how much the # of cookies in bag varies if you want to know how likely you are to snag a 10-cookie bag!



Intuition behind variance

Let us try to reason through how we can measure how the # of cookies vary. Consider just two bags of cookies:



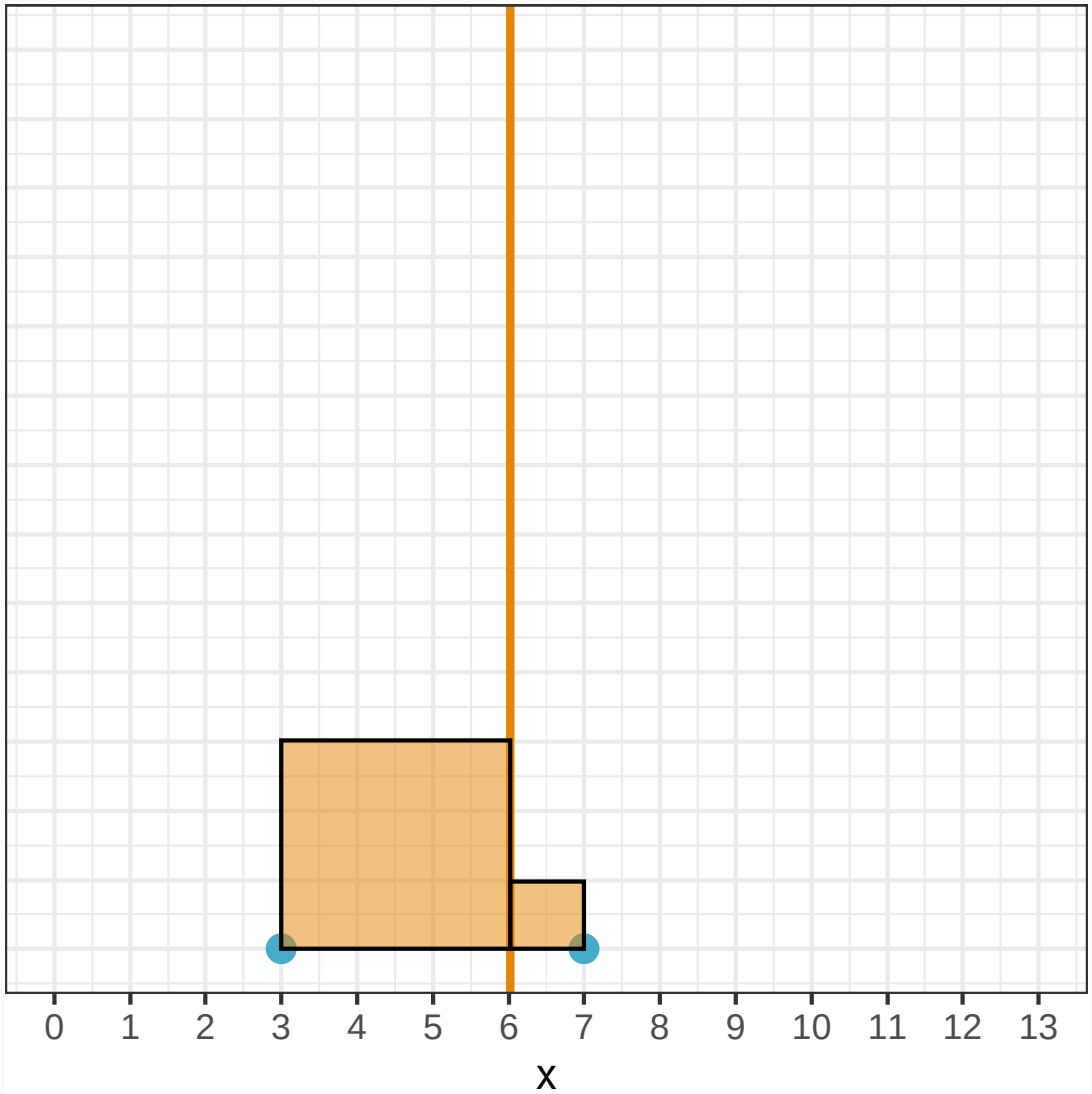
Intuition behind variance

One option is to measure the difference between each observed x, and $E[X]$ and add them all up.

Can you spot an issue with this? How would you solve this issue?

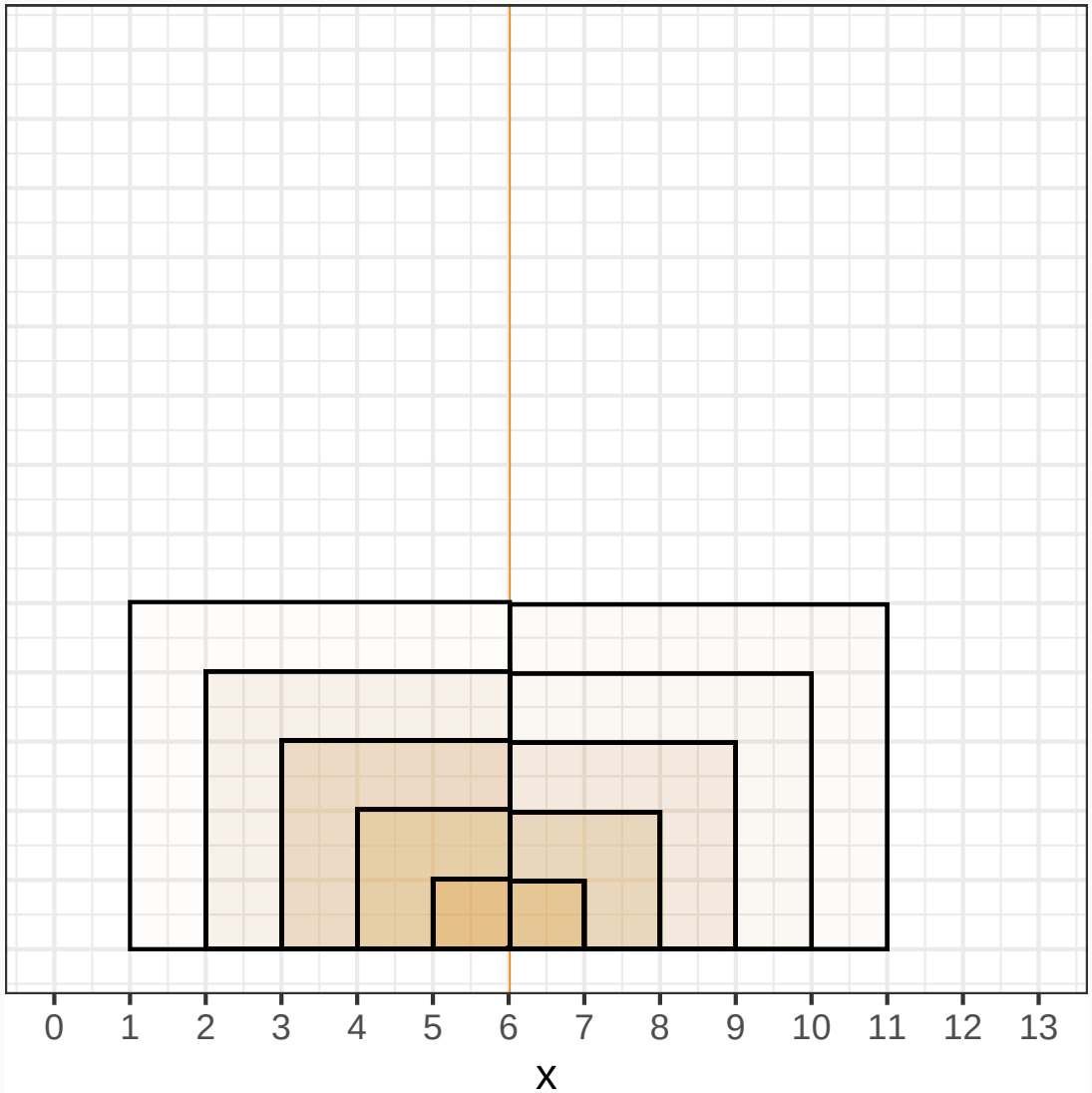
Intuition behind variance

Another option is to square the differences and take their average. Visually, it would look like this for two points:



Intuition behind variance

We can get the variance by taking the mean, or expected value, of the squared differences from $E[X]$.



Variance

The **variance** of a random variable X is written as $Var(X)$

For a random variable X,

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

In the *population*, this is:

$$Var(X) = \sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Properties of Variance

It is also useful to know some properties of expectations:

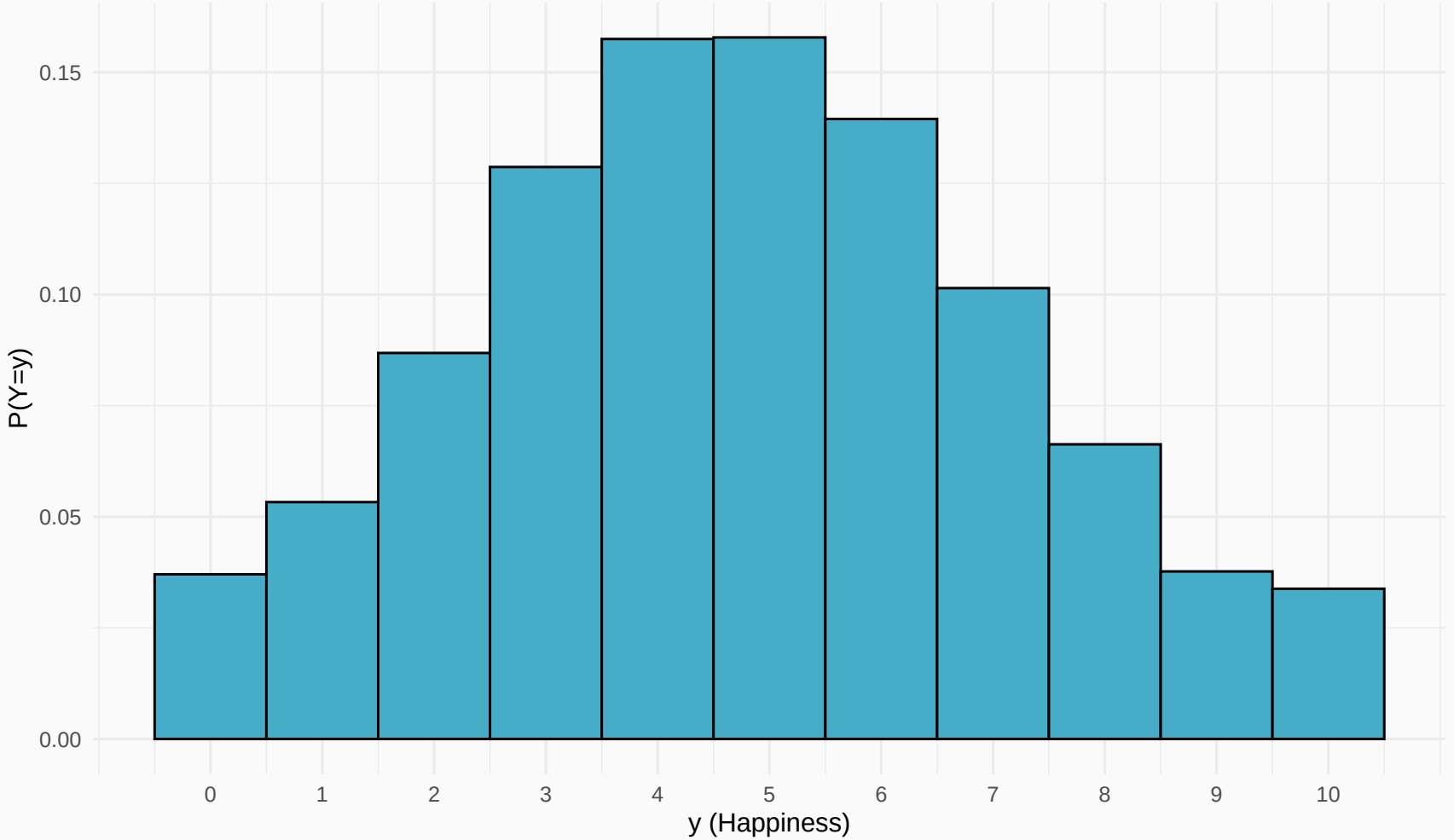
- $Var(aX) = a^2 Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$

Covariance

What about happiness?

We are also interested in students' happiness, Y.

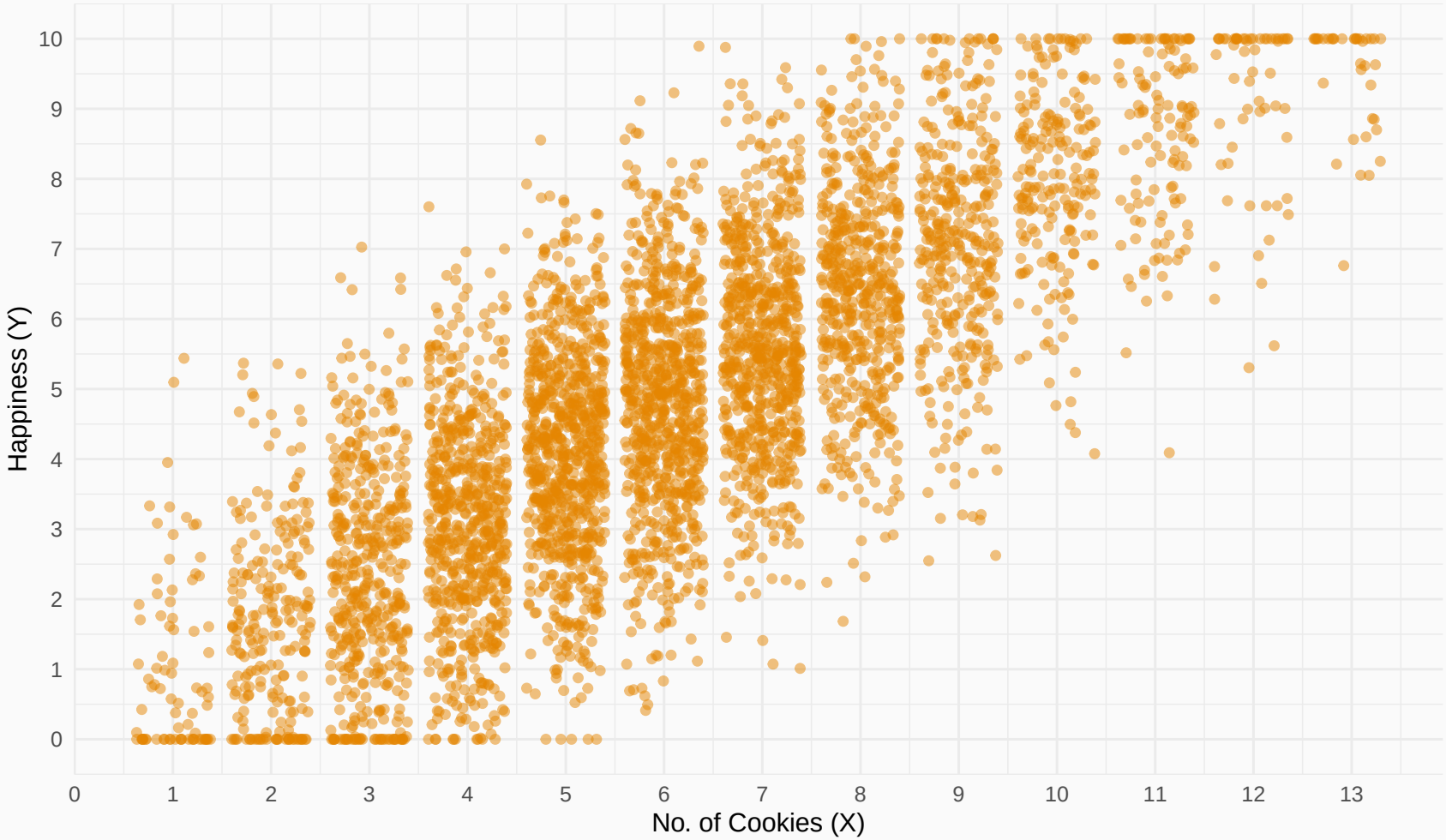
Let us take a look at the distribution and its characteristics



How do no. of cookies and happiness co-vary?

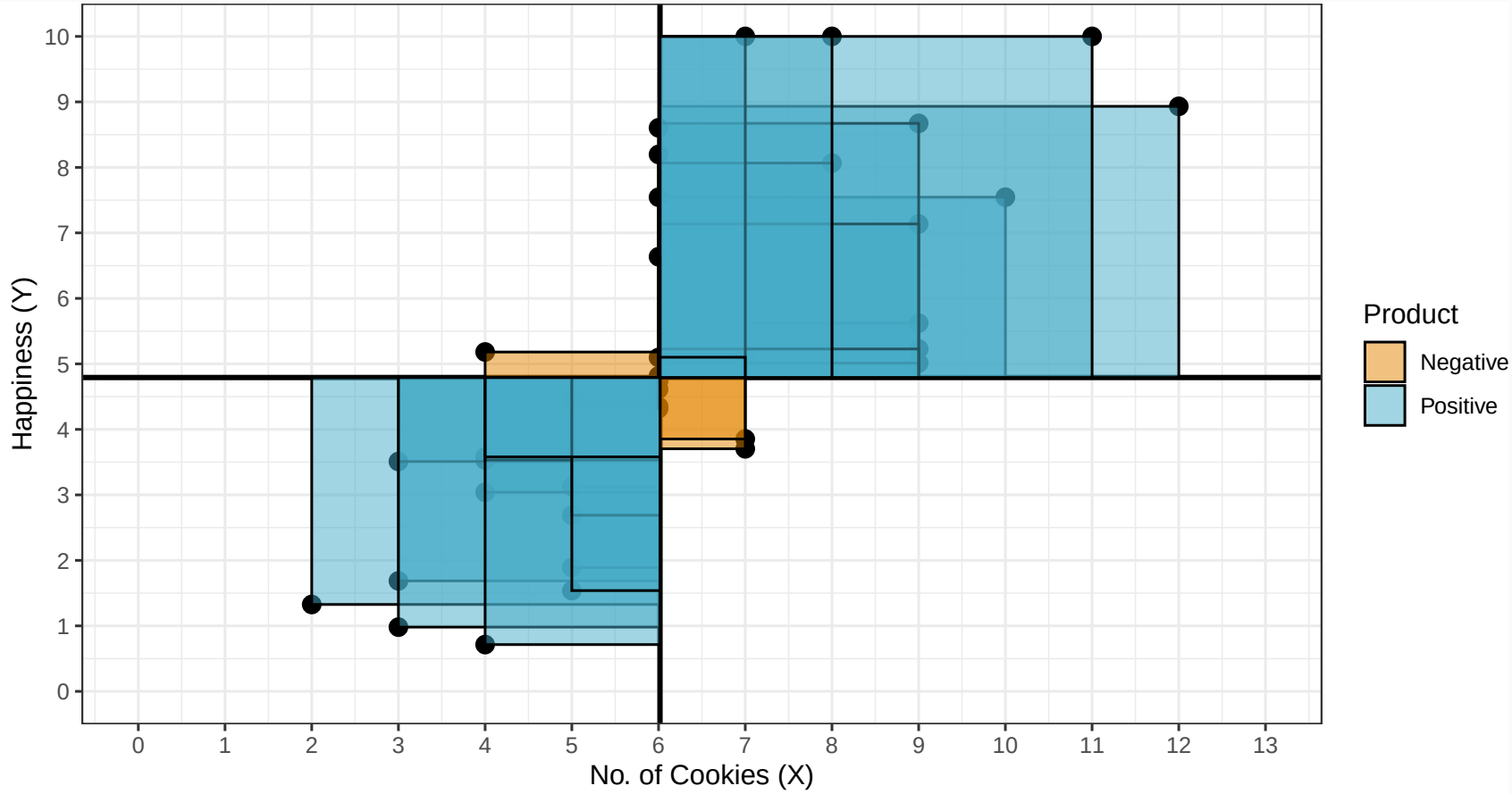
You want to know the relationship between cookies and happiness. What would your knee-jerk attempt be to assess this?

How do no. of cookies and happiness co-vary?



Building Intuition for Covariance

You want to know how the two variables vary together. Perhaps the intuition behind calculating variance can be helpful. Let us again take just a few points and visualize:



Covariance

The **covariance** of two random variables X and Y is written as $Cov(X, Y)$

$$Cov(X, Y) = E[(X - E[X])] * E[(Y - E[Y])]$$

It measures the tendency of two random variables to "move together". If they tend to move in similar directions, the covariance is positive, and if the tend to move in opposite directions, it is negative.

Covariance Properties

It might also be useful to know some properties of covariances

- $Cov(X, a) = 0$
- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(aX, bY) = abCov(X, Y)$
- $Cov(X + a, Y + b) = Cov(X, Y)$

To Cov or Not To Cov

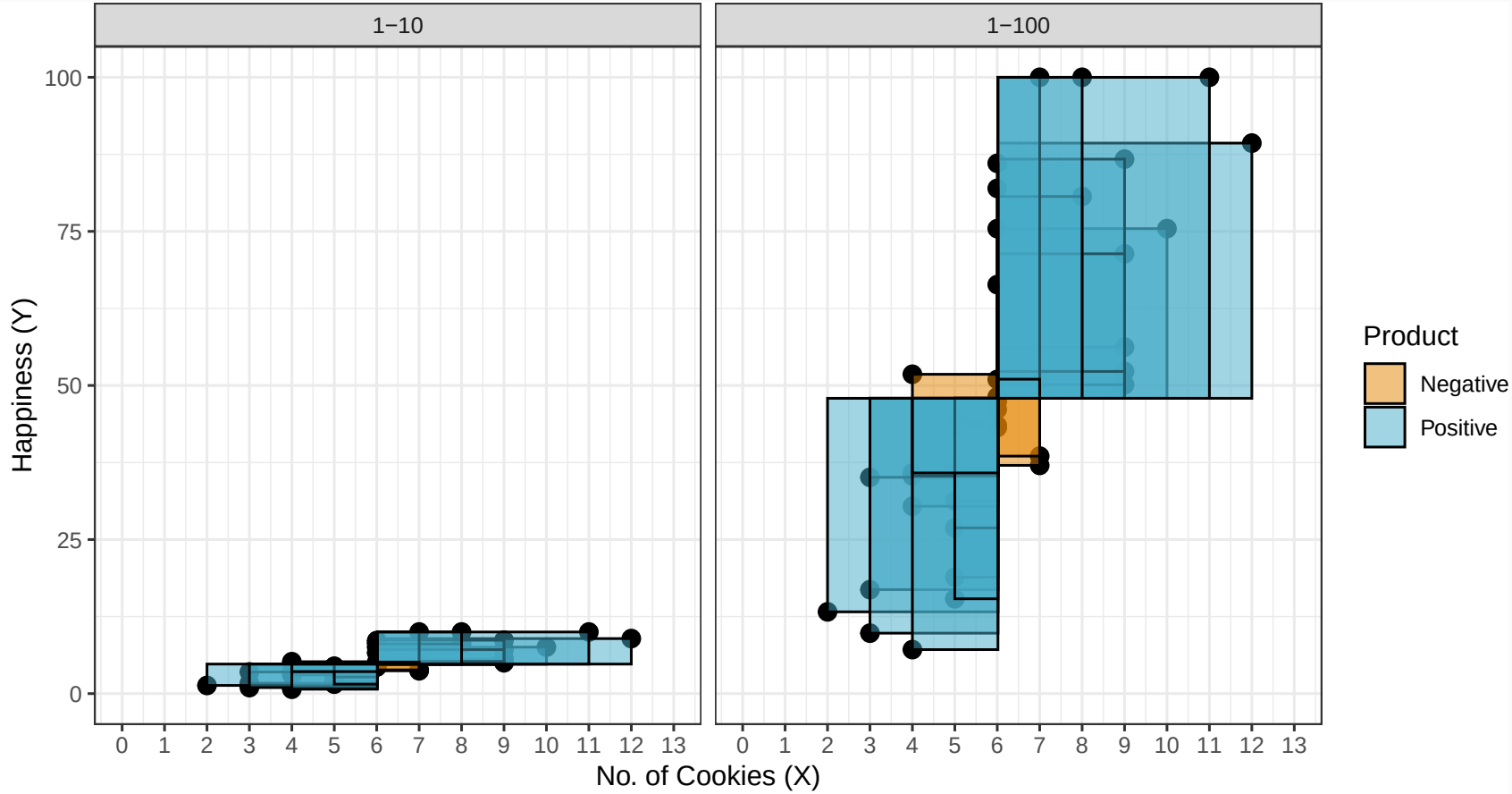
Why might we not want to use Cov(X,Y) for everything?

To Cov or Not To Cov

- Covariance is sensitive to the scale of the random variables
- It does not provide information on the strength of the relationship
- A bit difficult to interpret

Covariance is Scale Sensitive

What if we measured happiness on scale of 1-100?



Correlation

We can fix the scaling issue of covariance by dividing it by the standard deviations of the random variables. The resulting quantity is **correlation**

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Correlation

- Correlation is not sensitive to scale
- Bounded between -1 and 1
- More intuitive (for e.g. $\rho_{X,X} = 1$)

Covariance, Correlation and Independence

If two variables are independent, what would you expect their covariance and correlation to be?

Covariance, Correlation and Independence

If two variables are independent, what would you expect their covariance or correlation to be?

- Zero!

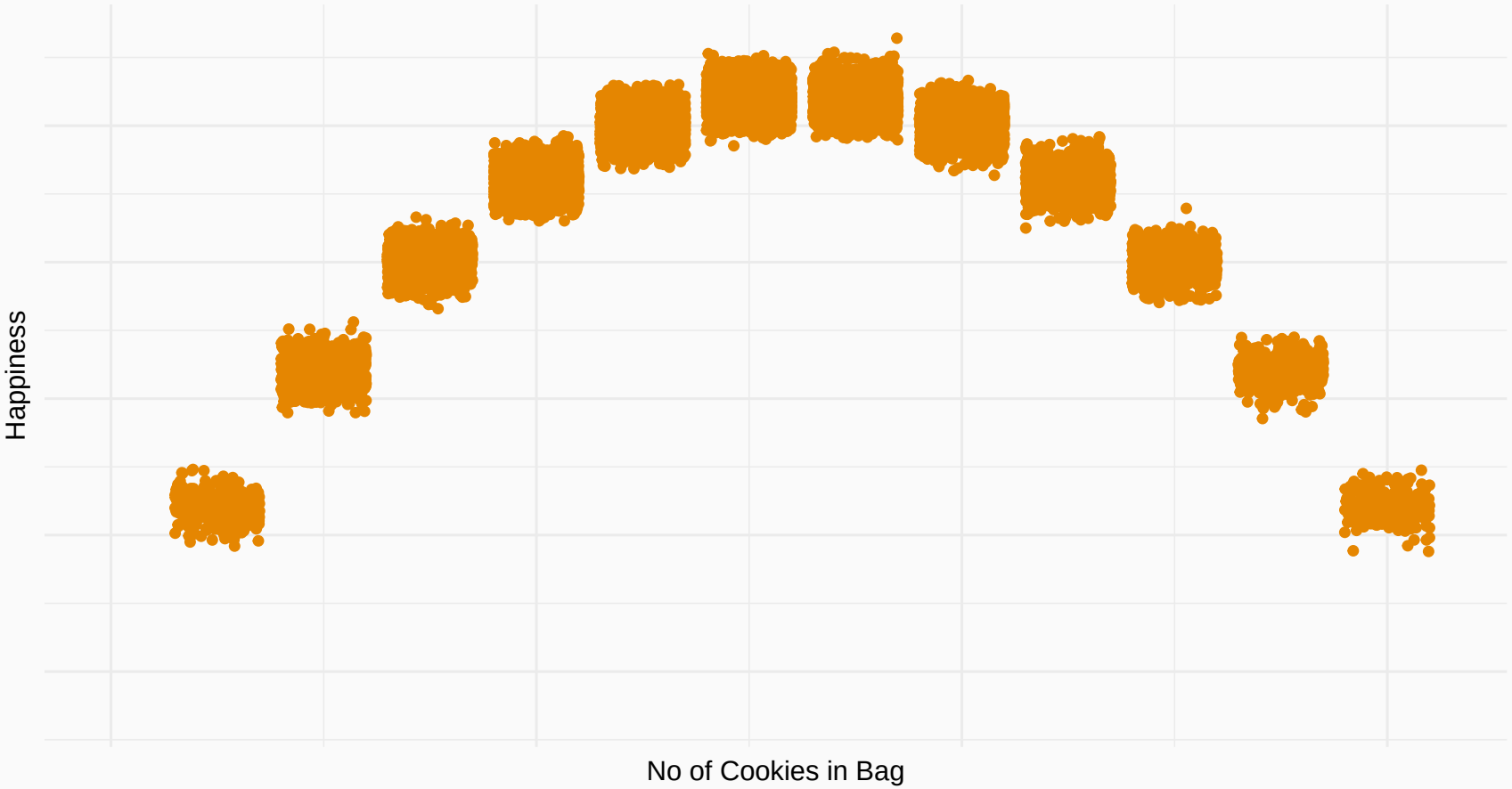
Covariance, Correlation and Independence

If two variables have covariance of 0, are they always independent?

Covariance, Correlation and Independence

If two variables have covariance of 0, are they always independent?

- No!



Joint Distributions

We may be interested in *joint* distribution of X and Y. That is what is the probability $P(X = x, Y = y)$ for each value of X and Y?

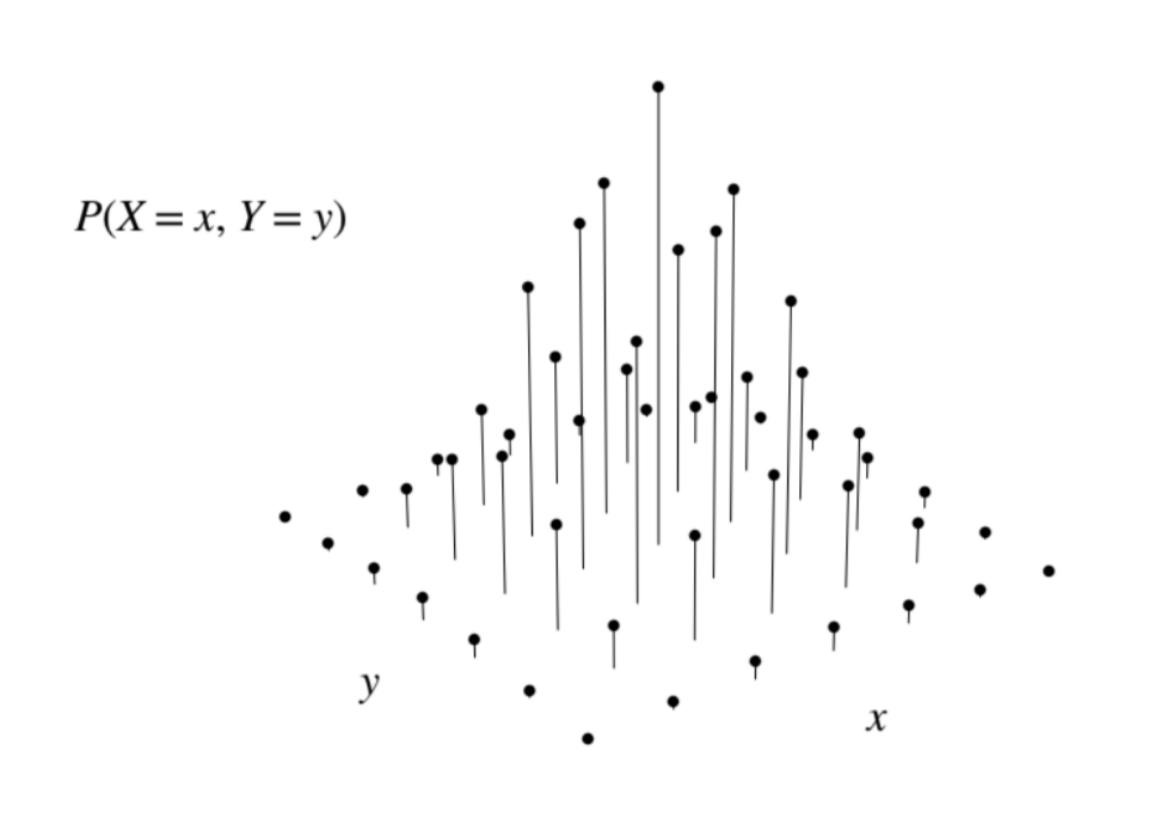
For discrete cases we can write down the $P(X = x, Y = y)$ for all combinations of possible values.

We could also visualize this information

Visual intuition about joint distributions

What would correspond to the conditional probability $P(Y|X = x)$ here?[1]

What about the marginal distribution of $P(Y = y)$?



[1] Figure from Introduction to Probability. Blitzstein and Hwang. CRC Press 2019

Linear Regression

Why regression?

Often, the question that you are interested in maybe something like: "For each additional cookie in a bag, what is the change in self-rated happiness of a HSPH student during orientation week?"

Unfortunately, the correlation and covariance values cannot directly answer this (although they are related!)

A regression approach can help us answer this question, while also allowing us to include multiple variables of interest (e.g. no. of ice cream scoops, no. of swag items)

The question in notation

"For each additional cookie in a bag (X), what is the change in self-rated happiness of a HSPH student (Y) during orientation week?"

$$E[Y|X = (x + 1)] - E[Y|X = x]$$

The Statistical Model

For simplicity, let us assume that the relationship between no. of cookies and happiness is a straight line.

If you just wanted to draw a straight line on two axes, you would write:

$$y = mx + b$$

Similarly, we can write our statistical model assuming that there is a linear relationship with a particular intercept and a slope:

$$E(Y|X = x) = \beta_0 + \beta_1 X$$

Answering the question using regression

We said that we want to know $E(Y|X = (x+1)) - E(Y|X = x)$.

And the regression equation is: $E(Y|X = x) = \beta_0 + \beta_1 X$.

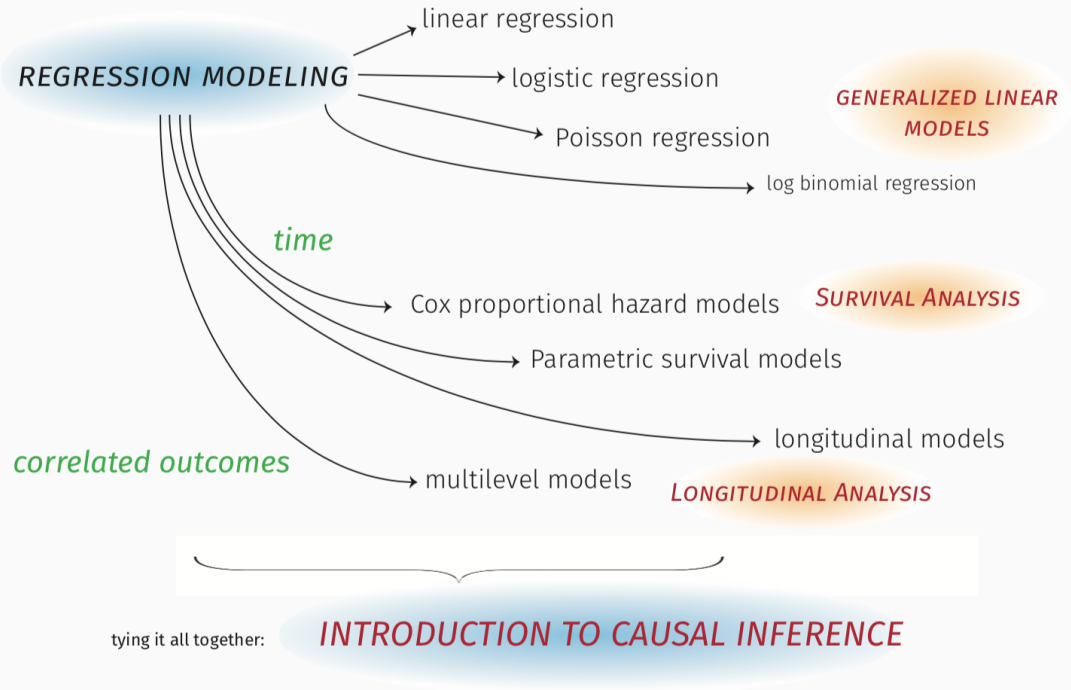
So, we can write our question of interest as:

$$\begin{aligned} E(Y|X = (x+1)) - E(Y|X = x) &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

You might have questions..

- What if we did not have data on all 20,000 students??
- Are cookies the secret of happiness? Do they *cause* happiness??
- What if students in some departments are happier with cookies than others??
- Do orientation cookies affect stress during PQE1??

These and more to be answered in the universe of PHS!!



Estimating beta

Now let us look at the guts of estimation in generalized linear models.

Psych lol

That's what we are going to be during first few weeks of PHS2000A!!

Thanks to past PHS TAs!!

- Gabe Schwartz
- Matt Lee
- Louisa Smith

THE END