3.2 More Probability/Statistics Foundations

PLSC30500, Fall 2021

co-taught by Molly Offer-Westort & Andy Eggers

(This lecture with references to Aronow & Miller (2019) and Wasserman (2004))

Recall our terms from probability. Our random process: flipping a fair coin twice.

- Ω : Sample space. Describes all possible outcomes in our setting.
- ω : Generic notation for the realized outcomes in the sample space.
- Here, $\Omega = \{HH, HT, TH, TT\}$.
- Event: a subset of Ω.
- We will often use terms like A or B to define events.
- In our example, the event that we get a head on the first flip is $A = \{HT, HH\}$.
- S : Event space. Describes all subsets of events, including the null set. [Full event space]
- We use this in addition to the sample space, so we can describe all types of events that we can define the probability for.
- P : Probability measure. A function that assigns probability to all of the events in the event space.
- Here, since our coin is fair, for the event that we get a head on the first flip, $\mathrm{P}(A) = 1/2$.

Random variables

- A random variable is a mapping X from our sample space Ω, to the Real numbers.

$$X : \Omega \rightarrow \mathbb{R}$$

- Random variables are ways to quantify random events described by our sample space.

- We'll mostly work with random variables going forward, but it's important to remember that the random variable is built on the foundations of the sample space -- and often, **you'll be the one deciding how that quantification happens.**

For example, with our two coin flips, let $X(\omega)$ be the number of heads in the sequence ω. Then the random variable, and its probability distribution, can be described as:

ω	$\mathrm{P}(\{\omega\})$	$X(\omega)$
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

and,

x	$\mathrm{P}(X = x)$
0	1/4
1	1/2
2	1/4

We can simulate this in R as well.

```
X <- c(0, 1, 2)
probs <- c(0.25, 0.5, 0.25)

sample(x = X,
       size = 1,
       prob = probs)
```

```
## [1] 0
```

```
n <- 1000
result_n <- sample(x = X,
                   size = n,
                   prob = probs,
                   replace = TRUE)

table(result_n)
```

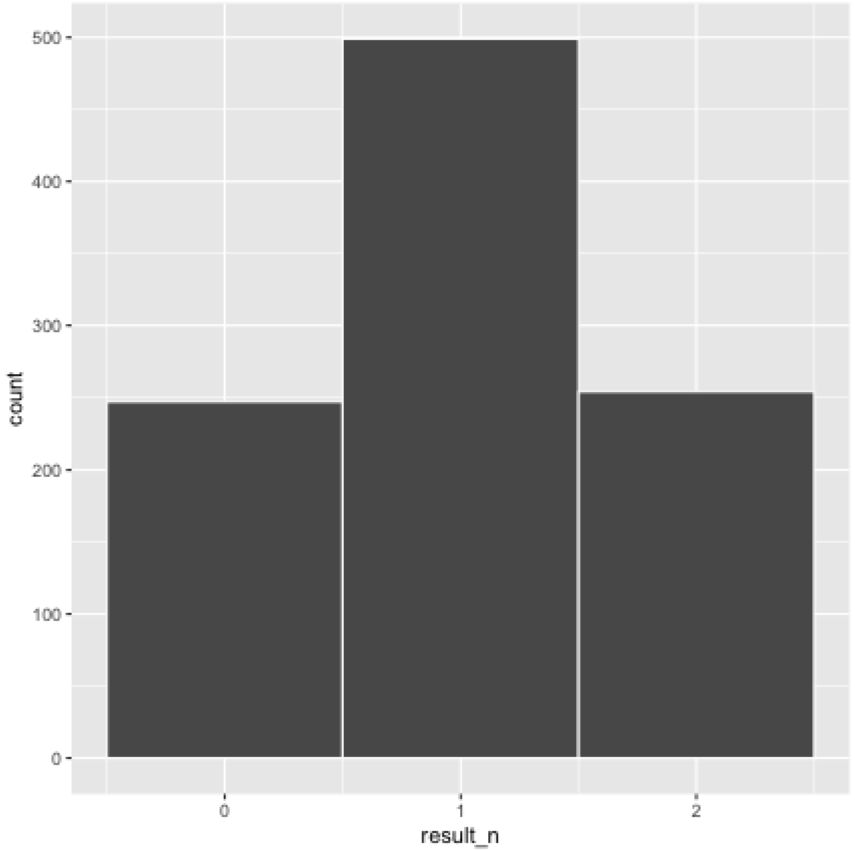
```
## result_n
##   0   1   2 
## 247 499 254
```

```
prop.table(table(result_n))
```

```
## result_n
##     0     1     2 
## 0.247 0.499 0.254
```

We can plot a histogram to look at the distribution of results.

```
ggplot(tibble(result_n), aes(x = result_n)) +
  geom_histogram(bins = 3, position = 'identity', color = 'white')
```



Probability Mass Function of a discrete random variable

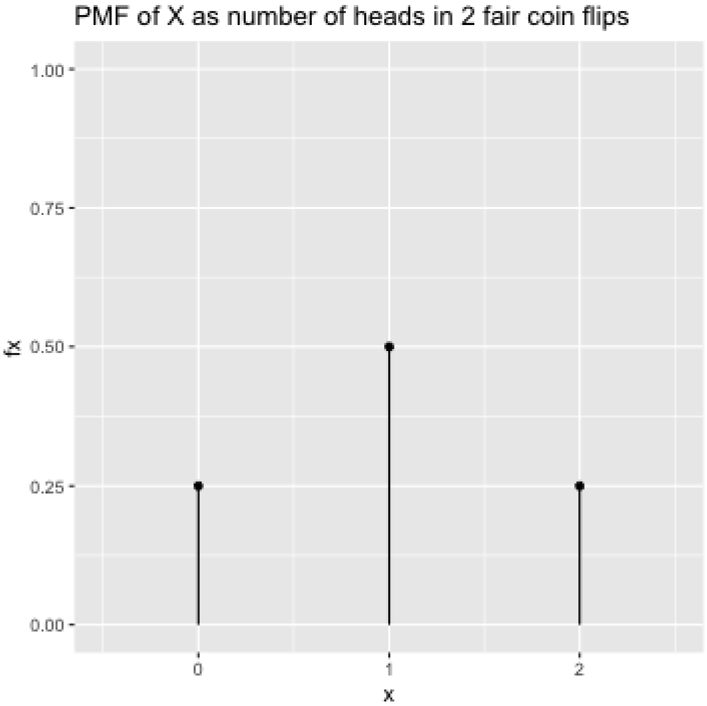
- A random variable is *discrete* if it takes countably many values.
- The probability mass function of a discrete RV X tells us the probability we will see an outcome at some value x.

$$f(x) = \mathrm{P}(X = x)$$

For our coin flip example,

$$f(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Illustrating the PMF of a discrete RV



Note that the probabilities sum to 1. This is one of the foundational axioms of probability.

Cumulative Distribution Functions

- The cumulative distribution function of X tells us the probability we will see an outcome less than or equal to some value x.

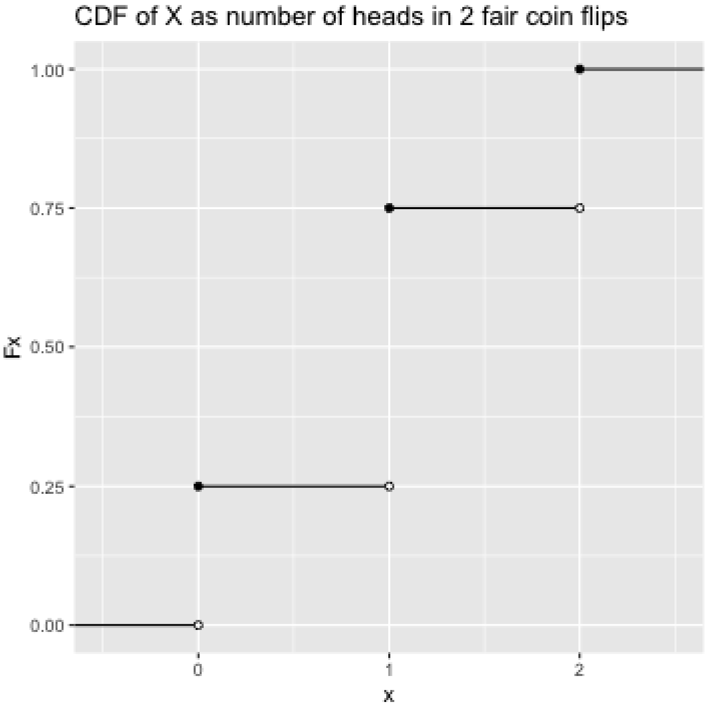
$$F(x) = \mathrm{P}(X \leq x)$$

For our coin flip example,

$$F(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

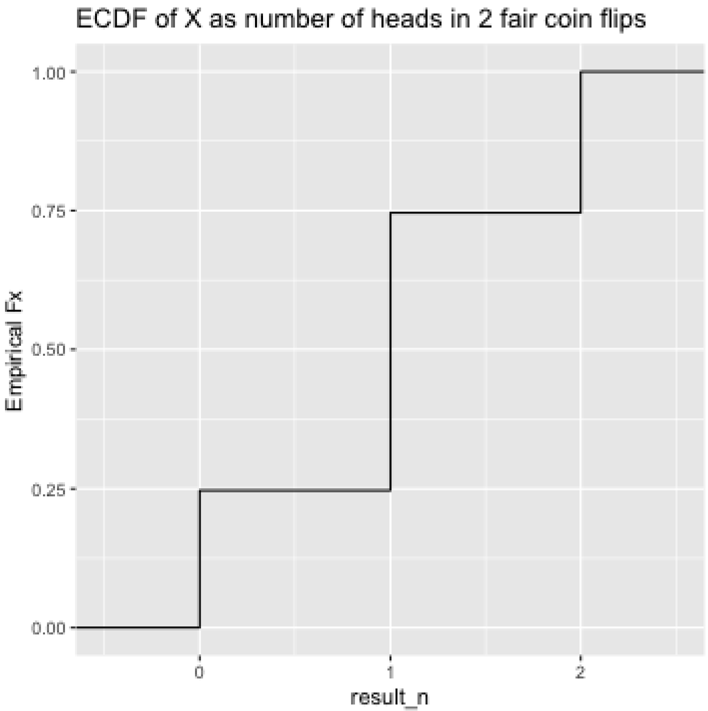
CDFs are really useful, because if we know the CDF, we can fully describe the distribution of *any* random variable.

Illustrating the CDF of a discrete RV



And we can use `ggplot2` to see what the *Empirical* CDF looks like

```
ggplot(tibble(result_n), aes(x = result_n)) +
  stat_ecdf() +
  coord_cartesian(xlim = c(-0.5, 2.5)) +
  ylab('Empirical Fx') +
  ggtitle('ECDF of X as number of heads in 2 fair coin flips')
```



Joint and conditional relationships

Bivariate relationships

We often care about how random variables vary with each other

- age and voter turnout
- sex and income
- education and earnings

Just like with univariate random variables, we can describe these bivariate relationships by their distributions

Joint PMF of discrete random variables

$$f(x, y) = \mathrm{P}(X = x, Y = y)$$

Returning to our example of flipping two fair coins

- Let X be 1 if we get *at least one heads*, and 0 otherwise
- Let Y be 1 if we get *two* heads in our two coin flips, and 0 otherwise

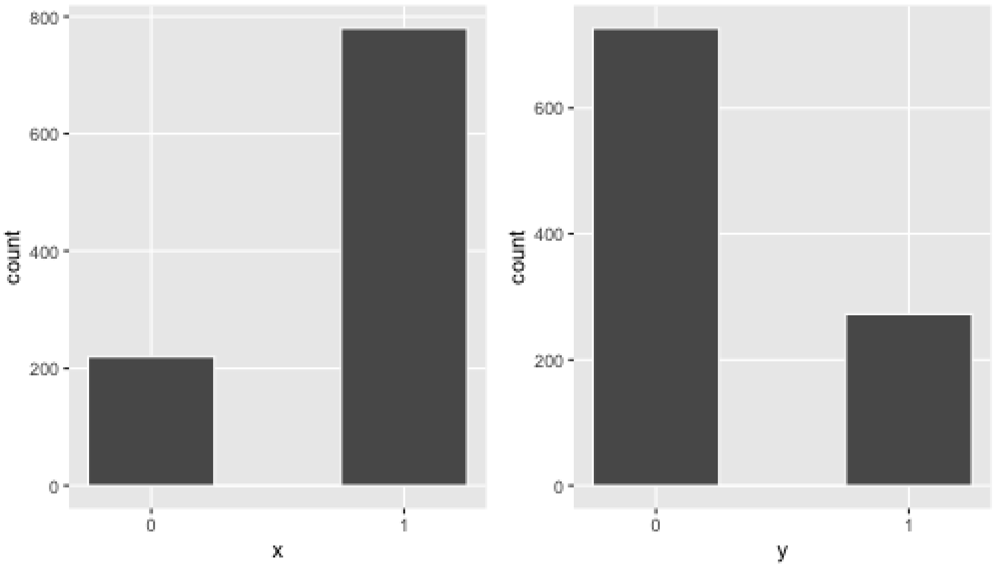
Then the joint probability distribution can be described as:

ω	$\mathrm{P}(\{\omega\})$	$X(\omega)$	$Y(\omega)$
TT	1/4	0	0
TH	1/4	1	0
HT	1/4	1	0
HH	1/4	1	1

and,

$$f(x, y) = \begin{cases} 1/4 & x = 0, y = 0 \\ 1/2 & x = 1, y = 0 \\ 1/4 & x = 1, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

```r
Omega <- c('HH', 'HT', 'TH', 'TT')
probs <- c(0.25, 0.25, 0.25, 0.25)

result_n <- sample(x = Omega,
                   size = n,
                   prob = probs,
                   replace = TRUE)

result_mat <- tibble(omega = result_n,
                     x = case_when(result_n == 'TT' ~ 0, TRUE ~ 1),
                     y = case_when(result_n == 'HH' ~ 1, TRUE ~ 0))

options <- list(theme(panel.grid.minor = element_blank()), scale_x_continuous(breaks = c(0, 1))) # save some style options

p1 <- ggplot(result_mat) + geom_histogram(aes(x = x), bins = 3, position = 'identity', color = 'white') + options

p2 <- ggplot(result_mat) + geom_histogram(aes(x = y), bins = 3, position = 'identity', color = 'white') + options

grid.arrange(p1, p2, ncol = 2)
```



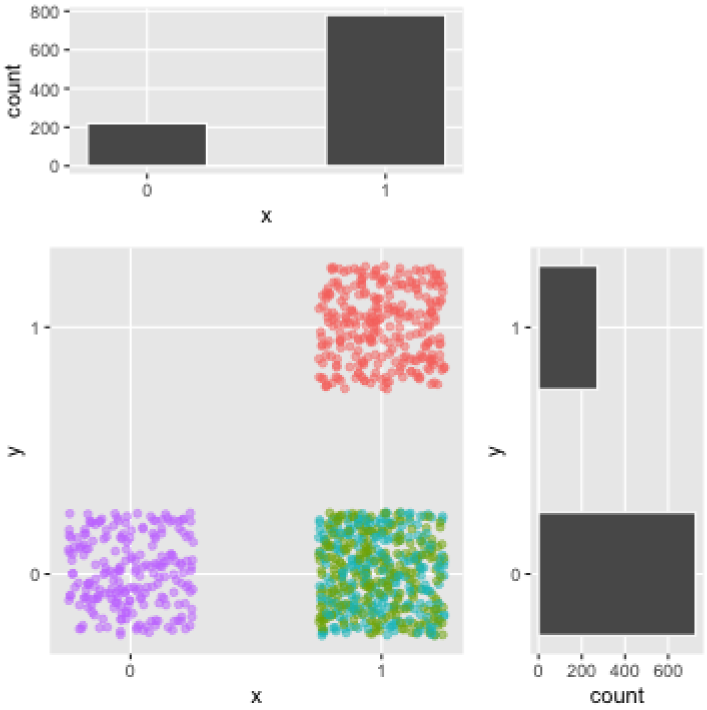
Seeing X and Y plotted side by side doesn't really give us a full picture of their relationship. These are the *marginal* distributions of X and Y, i.e., their distributions where we *marginalize* or sum over the distribution of the other random variable.

Marginal distributions

$$f_X(X) = \mathrm{P}(X = x) \sum_y \mathrm{P}(X = x, Y = y) = \sum_y f(x, y)$$

	$Y = 0$	$Y = 1$	
$X = 0$	1/4	0	**1/4**
$X = 1$	1/2	1/4	**3/4**
	3/4	**1/4**	

Notational aside: we subscript X in f_X to denote that it is the mass function of X specifically, as X and Y have different probability mass functions. If we are only dealing with one random variable at a time, we don't need to specify which distribution we're dealing with.



Plotting X and Y jointly gives us a better understanding of their joint relationship.

Conditional distributions

We are also often interested in conditional relationships.

$$f_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x) = \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(X = x)} = \frac{f(x, y)}{f_X(x)}$$

Here, what is the probability of observing two heads, conditional on having observed at least one heads?

$$f_{Y|X}(y|x) = \begin{cases} 1 & x = 0, y = 0 \\ 1/2 & x = 1, y = 0 \\ 1/2 & x = 1, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Summarizing single variable distributions

Expectation

$$\mathrm{E}[X] = \sum_x x f(x)$$

- Expectation is an *operator* on a random variable; it maps the distribution of X to a specific number.
- Specifically, the expectation operator tells us about the mean, or average value of X across its distribution.

Notational aside: it is common to write the expectation of a distribution as μ.

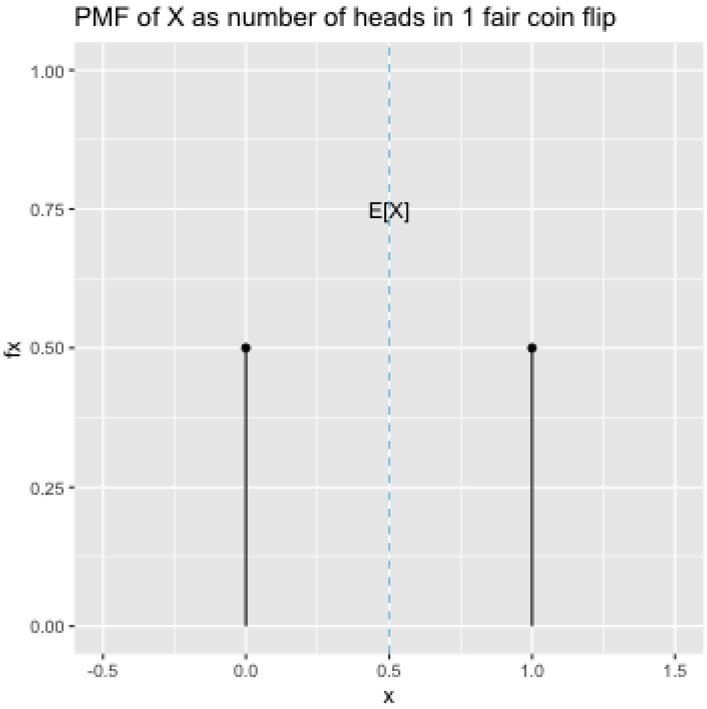
Let's flip a single coin, and let X be 1 if we get a head, and 0 otherwise.

$$f(x) = \begin{cases} 1/2 & x = 0 \\ 1/2 & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Mathematically,

$$\begin{aligned} \mathrm{E}[X] &= \sum_x x f(x) \\ &= 0 \times \frac{1}{2} + 1 \times \frac{1}{2} \\ &= \frac{1}{2} \end{aligned}$$

Visually,


PMF of X as number of heads in 1 fair coin flip
1.00
0.75
0.50
0.25
0.00
fx
E[X]
-0.5
0.0
0.5
1.0
1.5
x

Spread of a distribution

We often describe the spread of a distribution by its variance

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

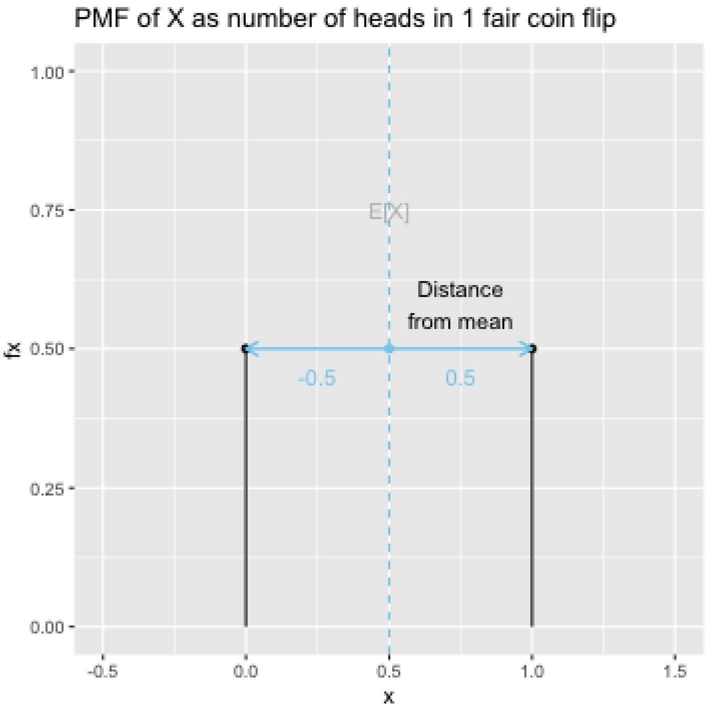
Or equivalently,

$$= \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

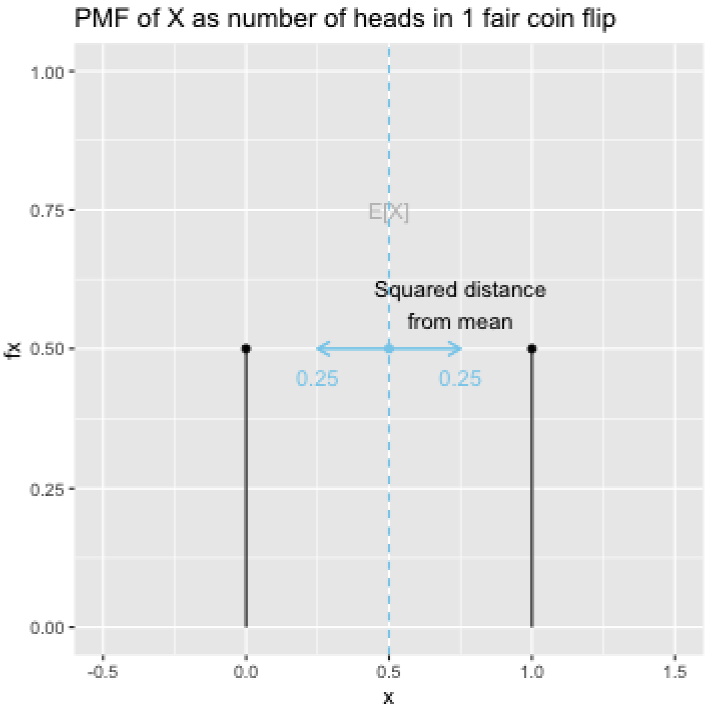
The standard deviation is the square root of the variance.

Notational aside: it is common to write the variance of a distribution as σ^2, or the standard deviation as σ.

The variance is the average squared distance from the mean. The standard deviation is the square root of this.

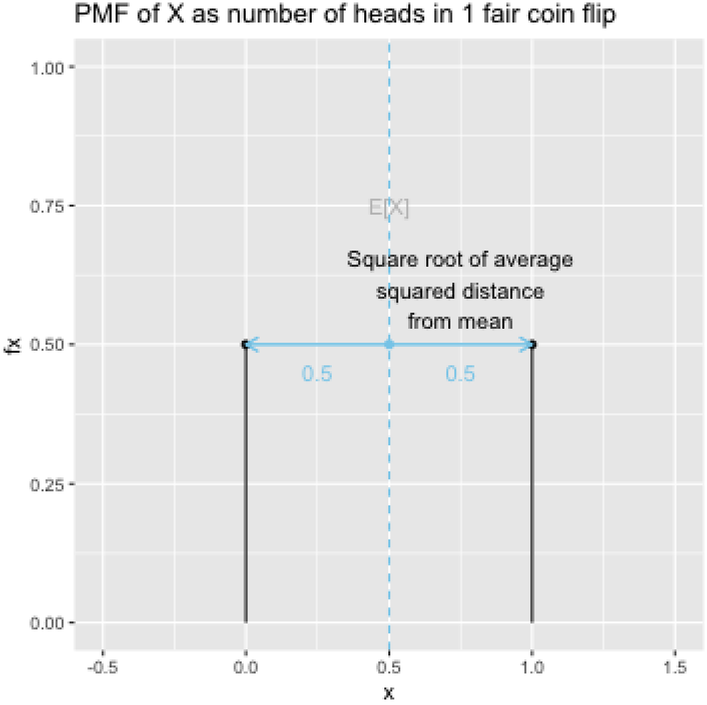


The variance is the average squared distance from the mean. The standard deviation is the square root of this.



Variance = $0.25 \times 0.5 + 0.25 \times 0.5 = 0.25$

The variance is the average squared distance from the mean. The standard deviation is the square root of this.



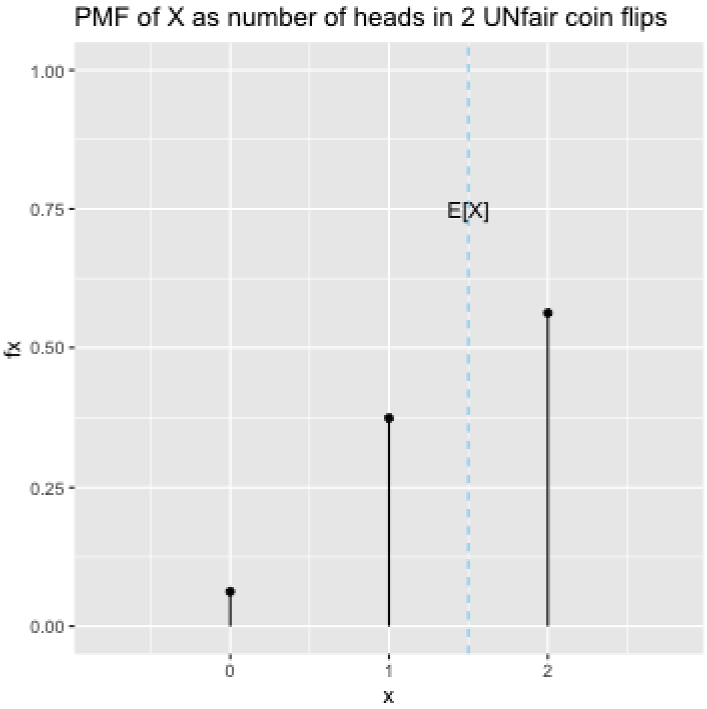
SD = $\sqrt{0.25} = 0.5$

Let's take another example, where we flip a coin twice, and let X be the number of heads. However, let's say our coin is *not* fair, and the probability of getting a heads is 0.8.

The random variable's probability distribution is then:

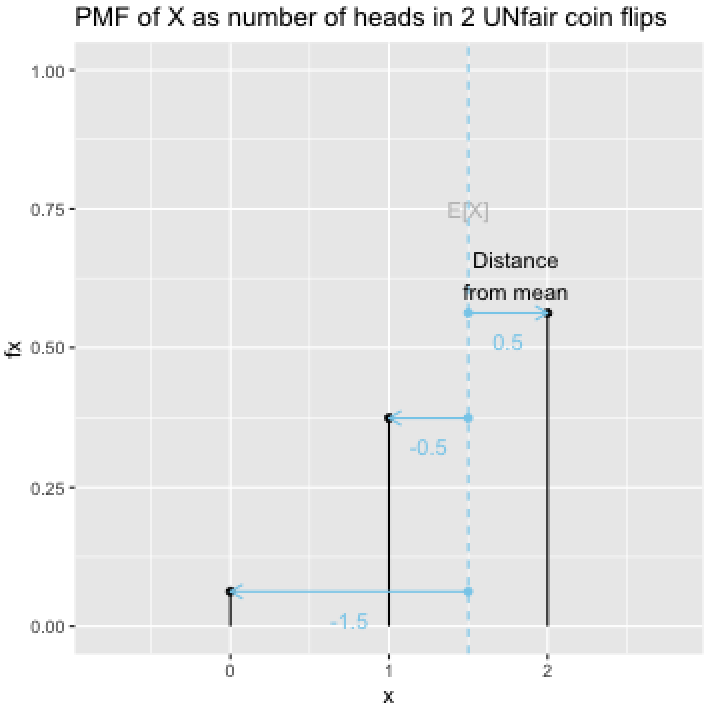
$$f(x) = \begin{cases} 1/16 & x = 0 \\ 3/8 & x = 1 \\ 9/16 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Let's take a look at the mean.



$$
\begin{aligned}
\mathrm{E}[X] &= \sum_x x f x \\
&= 0 \times \frac{1}{16} + 1 \times \frac{3}{8} + 2 \times \frac{9}{16} \\
&= \frac{24}{16} \\
&= 1.5
\end{aligned}
$$

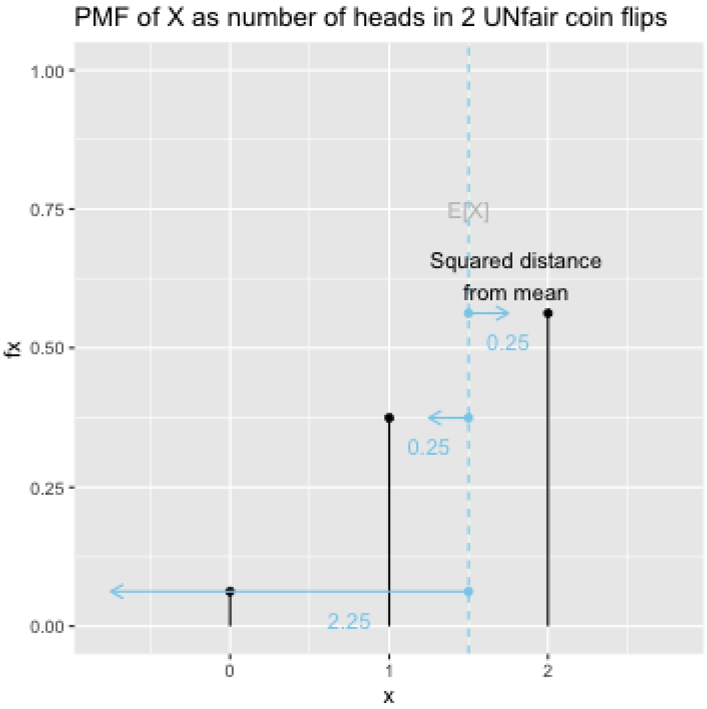
And the spread.



Variance = average squared distance from the mean

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

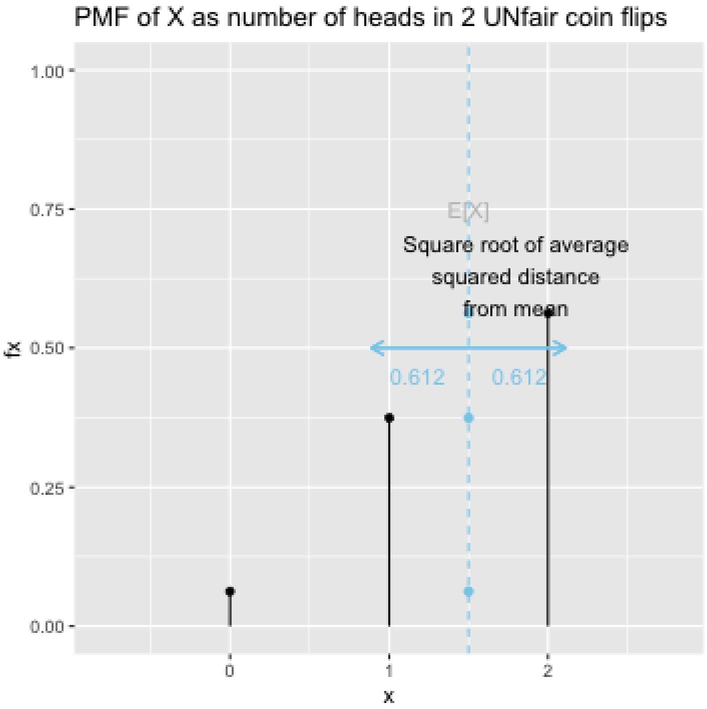
And the spread.



Variance = average squared distance from the mean

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= 2.25 \times \frac{1}{16} + 0.25 \times \frac{3}{8} + 0.25 \times \frac{9}{16} \\ &= 0.375\end{aligned}$$

And the spread.



SD = square root of variance

$$= \sqrt{0.375} = 0.612$$

Summarizing joint distributions

Covariance

$$\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[\mathrm{Y}])]$$

Covariance is how much X and Y vary together.

- If covariance is positive, when the value of X is large (relative to its mean), the value of Y will also tend to be large (relative to its mean)
- If covariance is negative, when the value of X is large (relative to its mean), the value of Y will tend to be small (relative to its mean)

Correlation

$$\rho[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma[X]\sigma[Y]}$$

Rescaled version of covariance

- positive when covariance is positive
- negative when covariance is negative

$$-1 \leq \rho[X, Y] \leq 1$$

Suppose we observe the following data from the joint distribution of X and Y. We are given that $\mathrm{Var}[X] = 3$ and $\mathrm{Var}[Y] = 6$.

```r
df <- tibble(x = rpois(1e3, lambda = 3),
             y = rpois(1e3, lambda = 3) + x)


ggplot(df, aes(x = x, y = y)) +
  geom_jitter(height = .2, width =.2, alpha = 0.6) +
  geom_smooth(method = lm, se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Following content if time, or in handout.

Continuous random variables

- So far, our coin flip example was for a *discrete* random variable.
- A random variable is *continuous* if it has a continuous density function
- Practically, we will treat RVs as discrete if they have countably many outcomes, and RVs as continuous if the number of values they can take on is only constrained by our measurement tool.

Uniform distribution

- If you take a draw from the standard uniform distribution, you are equally likely to draw any number between zero and one.

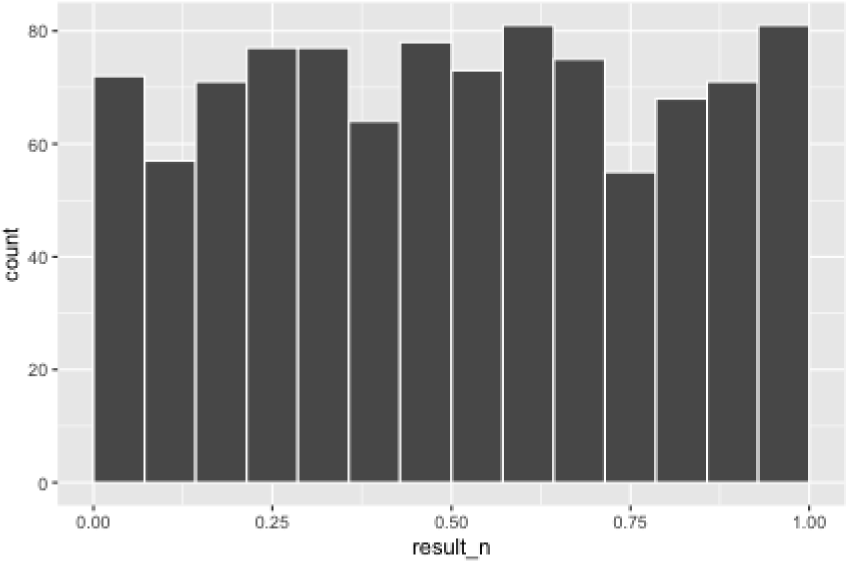
- We can simulate this in R. R allows you to sample from a number of canonical distributions; to see which distributions are available, search `?Distributions`.

```r
runif(n = 1, min = 0, max = 1)
```

```
## [1] 0.7550855
```

We can again sample from the distribution many times, and plot a histogram to look at the distribution of results.

```
result_n <- runif(n, min = 0, max = 1)

ggplot(tibble(result_n), aes(x = result_n)) +
  geom_histogram(breaks = seq(0, 1, length.out = 15),
                 position = 'identity', color = 'white')
```



Probability Density Function of continuous random variables

- Discrete random variables have non-zero mass on specific points, but for continuous random variables, $\mathrm{P}(X = x) = 0$. Instead of mass, we refer to *density* for continuous variables. [2]

- The *probability density function* $f(x)$ for a continuous random variable gives the slope of the CDF at any given point. This means that we can integrate the area under the PDF to get the relative probability of being between two points.

$$\mathrm{P}(a < X < b) = \int_a^b f(x)dx$$

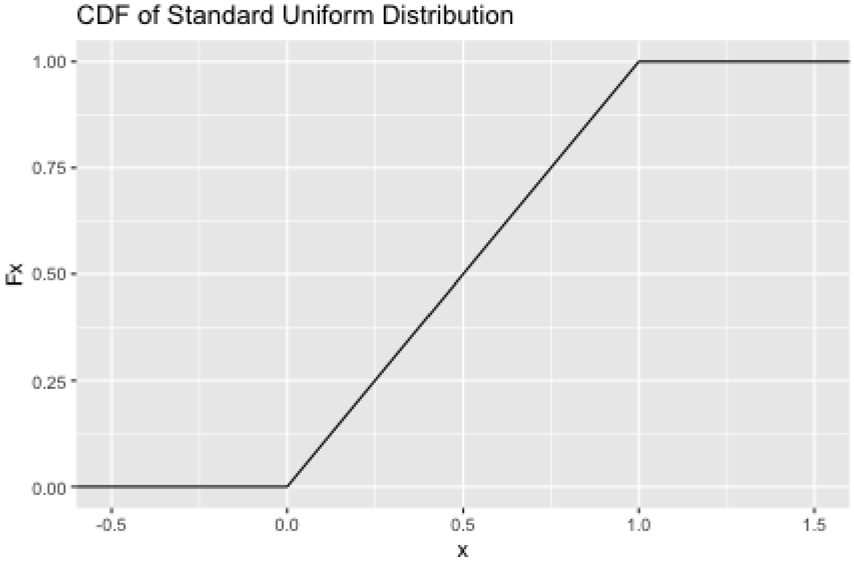
[2] Measure theory give a unified approach to measuring discrete and continuous random variables, but for simplicity, we'll keep the dichotomy of mass vs. density for discrete/continuous here.

Illustrating the CDF of a continuous RV

- We start by showing the CDF of the standard uniform distribution, to illustrate how the PDF relates to the CDF. The CDF for the standard uniform distribution is:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

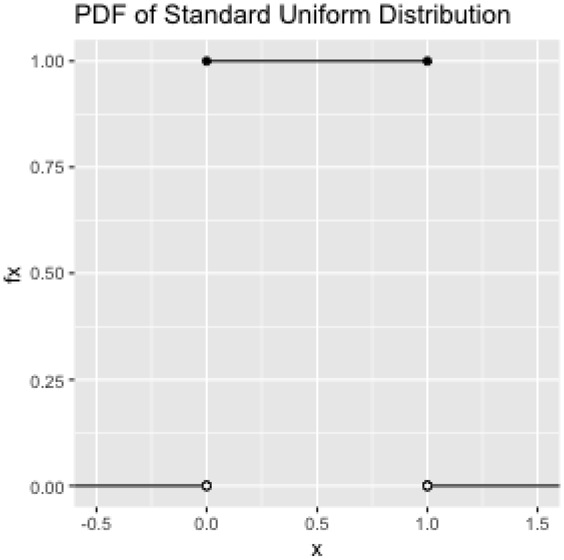
- Notice that the slope is 1 between 0 and 1.



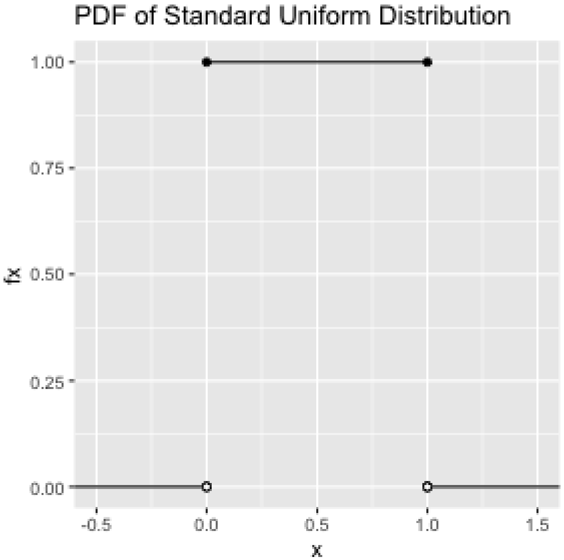
Illustrating the PDF of a continuous random variable

- The PDF for the standard uniform distribution is:

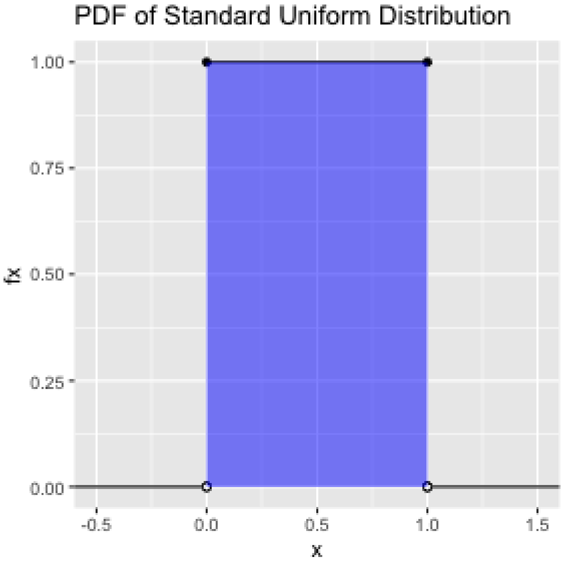
$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$



- Notice that if we take the area under the density curve, the total area will sum to 1.



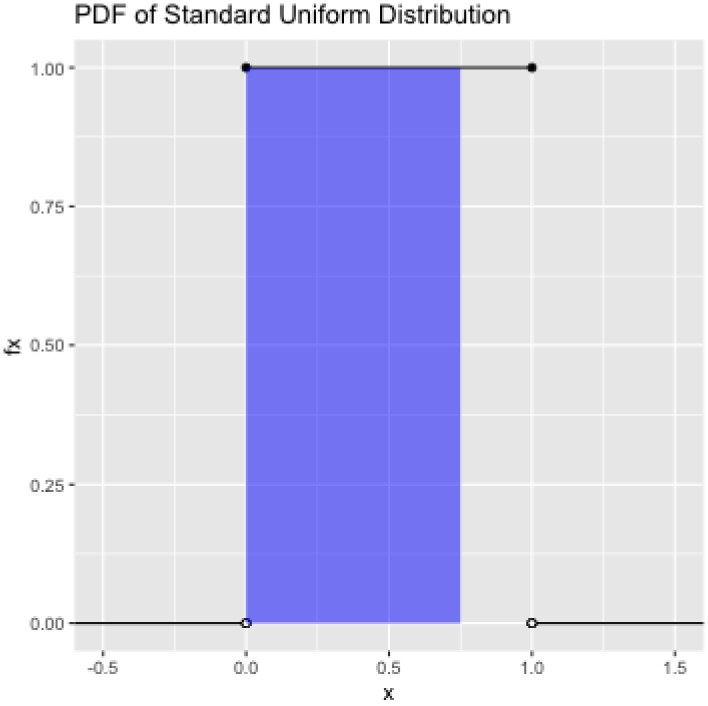
- Notice that if we take the area under the density curve, the total area will sum to 1. Relative density gives us relative probability.



- If we want to get the probability X is between 0 and 0.75,

$$\mathrm{P}(0 \leq x \leq .75) = \int_0^{.75} f(x)dx$$

we take the area under the density curve between 0 and 0.75 -- which is also 0.75. (Notice that we don't need to use calculus here.)



Normal distribution

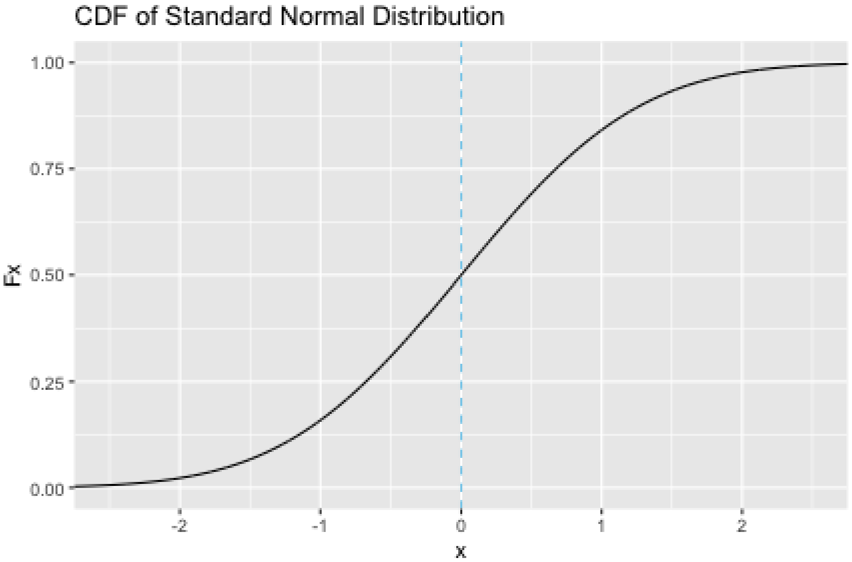
The Normal distribution is frequently used in probability and statistics, because it is a useful approximation to many natural phenomena.

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

It is defined by two parameters, μ, the center of the distribution, and σ, which defines the distribution's standard deviation, or spread.

It has a bell curve shape, with more density around the middle, and less density at more extreme values.

Normal distribution


CDF of Standard Normal Distribution
Fx
x