Social Science Inquiry II

Week 3: A brief introduction to probability

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2023

Housekeeping

- Homework assignments
  - Feedback is very helpful (thank you!)
  - If you spend $>$ 30 minutes without making progress, go to Stack Overflow or solutions.

Loading packages for this class

```
> library(ggplot2)
> library(gridExtra)
> set.seed(60637)
```

What do we do with data

Now that we've gotten started on working with data . . . what do we want to get from that data?

- Describe what's going on in the data
- We can do a pretty good job of this with the data visualization tools we have, along with summary statistics for numerical descriptions

This is already a really useful start, but beyond just describing the data we see in front of us, we may have other goals. We may want to:

- Make generalizations to a larger population
- Make informed guesses about how things would have turned out differently, if something different had happened

These latter two are called *inference*, and we'll need some additional tools and assumptions to make headway on them

Probability

Why do we use probability theory?

- Probability theory allows us to talk about *random* events in structured way.
    - Often, we see only part of the picture we'd like to talk about. We only see some of the data, or we can only see one version of events.
    - Probability theory gives us a way to describe the process that results in the data that we observe.
    - It also allows us to *formalize our uncertainty* when making inference.

Randomness

What does it mean to describe real world events as “random”?

- Probability theory is an abstract construct, but it is useful for empirical research to create a model of the world in which events are probabilistic.
    - If we are conducting measurement on some population but can only observe a sample, we assume that there is some randomness to who we observe and who we don’t observe.
- If we are describing events with counterfactuals, such as turning out to vote or testing positive for COVID, it can be useful to describe those events as probabilistic.
- These are *models* of how the world works, and they help us make sense of the fundamentally squishy nature of social science research. With limited information about the world, we operate with uncertainty. Assigning probabilities conditional on the information we *do* have helps us formalize that uncertainty. Even if we don’t necessarily believe that human behavior is “random.”

Flipping a coin twice

Suppose we are flipping a coin twice, and the coin is fair. This is a random process, and we will describe the probability space associated with this process.

Useful terms in probability

- 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.
  - Here, the event that we get a head on first flip is $A = \{HT, HH\}$.
- S : Event space. Describes all subsets of events, including 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. An operator that assigns probability to all events in the event space.
  - Here, the event that we get a head on the first flip, $\mathrm{P}[A] = 1/2$.

An aside on mathematical notation

Mathematical notation is a tool that gives us a common language to express concepts with precision. There are different conventions in different communities, and no approach is "right" or "wrong"–it's just a question of whether your notation is appropriately communicating to your audience what you want it to.

To use mathematical notation in R Markdown, write LaTeX typesetting commands inside of dollar signs.

For example,

```
$Y = \beta_0 + \beta_1 X$
```

is rendered as

$Y = \beta_0 + \beta_1 X$.

We have just used probability notation as a way to fully describe any random generative process.

Sampling

We can simulate our double coin flip process in R, using the `sample()` function. There are four possible outcomes, and all are equally likely.

```
> Omega <- c('HH', 'HT', 'TH', 'TT')
> probs <- c(0.25, 0.25, 0.25, 0.25)
> sample(x = Omega,
+        size = 1,
+        prob = probs)
```

```
[1] "HH"
```

We can run this simulation many times, and our results should *approximately* follow the probabilities we assigned.

```
> n <- 1000
> result_n <- sample(x = Omega,
+        size = n,
+        prob = probs,
+        replace = TRUE)
> table(result_n)
```

```
result_n
 HH  HT  TH  TT
247 241 258 254
```

Independent events

Two events are *independent* if

$$\mathrm{P}[AB] = \mathrm{P}[A]\mathrm{P}[B]$$

Notational aside: The event AB is that both A and B happen. There are other ways to write this, including $A \cap B$.

Returning to our example of flipping two fair coins. Let's say:

- Event A: we get a head on the first coin flip; $A = \{HT, HH\}$.
- Event B: we get a head on the second coin flip; $B = \{TH, HH\}$.
- We can see the event AB as the overlap in their respective sets, $AB = \{HH\}$

The coin flips are unrelated, so the events should be independent. We can check this mathematically.
First, we know that all of the outcomes $\Omega = \{HH, HT, TH, TT\}$ are equally likely.

$$
\begin{aligned}
\mathrm{P}[A] &= \mathrm{P}[\{HT\}] + \mathrm{P}[\{HH\}] = 0.25 + 0.25 = 0.5 \\
\mathrm{P}[B] &= \mathrm{P}[\{TH\}] + \mathrm{P}[\{HH\}] = 0.25 + 0.25 = 0.5
\end{aligned}
$$

Then, we can calculate the product of the probabilities, and the probability of the joint event.

$$
\begin{aligned}
\mathrm{P}[A]\mathrm{P}[B] &= 0.5 \times 0.5 = 0.25 \\
\mathrm{P}[AB] &= \mathrm{P}[\{HH\}] = 0.25
\end{aligned}
$$

We see that they are the same, so we have independence.

$$\mathrm{P}[A]\mathrm{P}[B] = \mathrm{P}[AB]$$

We can also check if observed proportions in our simulations show the same thing

```
> 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)
> (observed_props <- prop.table(table(result_n)))
```

```
result_n
   HH    HT    TH    TT 
0.273 0.264 0.243 0.220 
```

```
> (PA <- mean(result_n == 'HT' | result_n == 'HH'))
```

```
[1] 0.537
```

```
> (PB <- mean(result_n == 'TH' | result_n == 'HH'))
```

```
[1] 0.516
```

```
> (PAB <- mean(result_n == 'HH'))
```

```
[1] 0.273
```

```
> PA*PB
```

```
[1] 0.277092
```

The proportions look pretty close.

What about for a case where A and B are not independent?

- A is still the event we get a head on the first coin flip; $A = \{HT, HH\}$.
- B is now the event that we get a head on both coin flips; $B = \{HH\}$.
- AB is the intersection of these two sets, which is just $AB = \{HH\}$.

```
> (observed_props <- prop.table(table(result_n)))

result_n
   HH    HT    TH    TT
0.273 0.264 0.243 0.220

> (PA <- mean(result_n == 'HT' | result_n == 'HH'))

[1] 0.537

> (PB <- mean(result_n == 'HH'))

[1] 0.273

> (PAB <- mean(result_n == 'HH'))

[1] 0.273

> PA*PB

[1] 0.146601
```

Conditional probability

If $\mathrm{P}[B] > 0$, the *conditional probability* of an event A occurring, given event B has occurred is:

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[AB]}{\mathrm{P}[B]}$$

This can also be read as, out of all of the times event B occurs, how many times does event A also occur?

For our coin flip example, we'll stick with:

- A is the event that we get a head on the first coin flip; $A = \{HT, HH\}$.
- B is the event that we get a head on the both coin flips; $B = \{HH\}$.
- AB is $\{HH\}$

$$\begin{aligned}
\mathrm{P}[A|B] &= \frac{\mathrm{P}[AB]}{\mathrm{P}[B]} \\
&= \frac{\mathrm{P}[\{HH\}]}{\mathrm{P}[\{HH\}]} \\
&= 1
\end{aligned}$$

What is $\mathrm{P}[B|A]$?

Bayes Rule

A useful theorem to return to is *Bayes Rule*

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[B|A]\mathrm{P}[A]}{\mathrm{P}[B]}$$

Why is Bayes rule so useful? It basically tells us how we update probability based on observed data.

Bayes Rule

Example: Suppose everyone in the University of Chicago community is given a new rapid test for COVID.

- We are concerned with *false negatives*, when we get a negative test for a person who actually *is* infected.
- And *false positives*, when we get a positive test for a person who *is not* infected.

Bayes Rule

Example: Suppose everyone in the University of Chicago community is given a new rapid test for COVID.

A student gets a positive test back. What is the probability that they have COVID?

Bayes Rule

- Event A: person has COVID
- Event B: a positive test

Should we think these events are independent?

Bayes Rule

- Event A: person has COVID
- Event B: a positive test

Doctors know that the probability of having COVID is 3% in this population : $\mathrm{P}[A]$

```
> PA <- 0.03
```

The overall rate of positive tests is 5% : $\mathrm{P}[B]$

```
> PB <- 0.05
```

If you have COVID, your test will turn up positive 95% of the time : $\mathrm{P}[B|A]$

```
> PB_if_A <- 0.95
```

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[B|A]\mathrm{P}[A]}{\mathrm{P}[B]}$$

```
> ( PB_if_A * PA )/PB
[1] 0.57
```

Prosecutor's fallacy

$$\mathrm{P}[A|B] \overset{?}{=} \mathrm{P}[B|A]$$

As a general rule, we cannot assume the conditional probability of A given B is the same as the probability of B given A

- In a court case, the probability that a person is guilty given that we see a DNA match is NOT the same as the probability of a DNA match given that they are guilty.
- This fallacy often occurs when we observe evidence that we are very unlikely to see among innocent people, and very likely to observe among guilty people
- But the overall number of innocent people in the population is very large, so the number of false positives is much higher than the number of true positives

Why thinking about conditional probability is so important in social science research

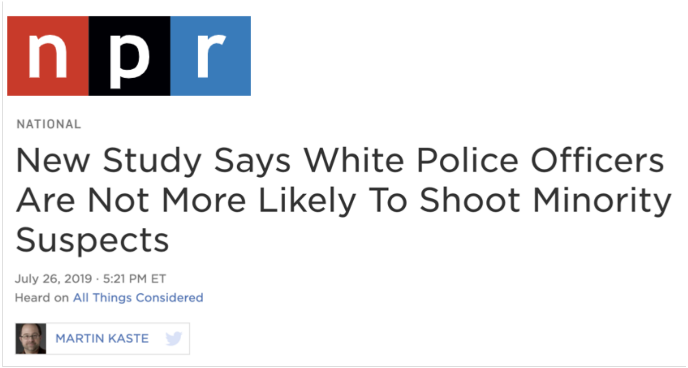
"20% of people hospitalized with COVID-19 are vaccinated"
Does this tell us that vaccines aren't very effective?
Some things we should think about:

- What percentage of the population in areas served by hospitals are vaccinated?
- Are people who are vaccinated at higher risk for breakthrough cases? Or more likely to be hospitalized? Are they older, or have pre-existing conditions?

Why thinking about conditional probability is so important in social science research

What data do you need to answer the question: “Do white officers shoot minority citizens at a higher rate than non-white officers?”

Remember this article?



npr

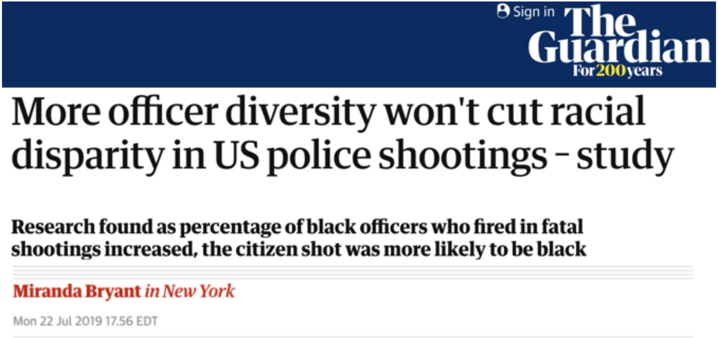
NATIONAL

New Study Says White Police Officers Are Not More Likely To Shoot Minority Suspects

July 26, 2019 · 5:21 PM ET
Heard on All Things Considered

MARTIN KASTE

Remember this article?

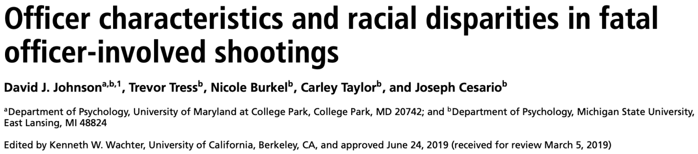


More officer diversity won't cut racial disparity in US police shootings – study

Research found as percentage of black officers who fired in fatal shootings increased, the citizen shot was more likely to be black

Miranda Bryant *in New York*

Mon 22 Jul 2019 17.56 EDT



Officer characteristics and racial disparities in fatal officer-involved shootings

David J. Johnson[a,b,1], Trevor Tress[b], Nicole Burkel[b], Carley Taylor[b], and Joseph Cesario[b]

[a]Department of Psychology, University of Maryland at College Park, College Park, MD 20742; and [b]Department of Psychology, Michigan State University, East Lansing, MI 48824

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved June 24, 2019 (received for review March 5, 2019)

- In a 2020 PNAS paper, Johnson and co-authors evaluated the relationship between race of victims shot by police, and the characteristics of the police shooters
- The authors claim, “White officers are not more likely to shoot minority civilians than non-White officers”–and the paper has been used in congressional testimony to support the claim that diversity in police forces would not be beneficial in reducing bias in officer-involved shootings
- Their database only has data on fatal shootings–not on cases where people were NOT shot.
- Their results report “whether a person fatally shot was more likely to be Black (or Hispanic) than White” conditional on the race of the officer involved
- Does this address original claim?

“White officers are not more likely to shoot minority civilians than non-White officers”

$$\mathrm{P}[\text{shot}|\text{White officer, minority civilian}] -$$
$$\mathrm{P}[\text{shot}|\text{Minority officer, minority civilian}]$$

What data are we missing to address this question?

- Dean Knox & Jonathan Mummolo wrote a letter critiquing the original article, based on the authors’ failure to appropriately apply Bayes Rule;
- The article was eventually retracted
- (It’s a bit more complicated than that, but know your conditional probability when thinking about difficult subjects!)

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,

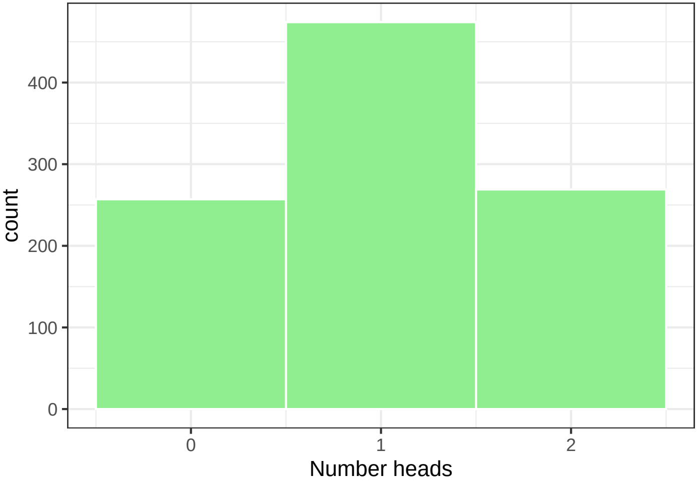
x	$\mathrm{P}[X = x]$
0	1/4
1	1/2
2	1/4

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
257 474 269
> prop.table(table(result_n))
result_n
    0     1     2
0.257 0.474 0.269
>
```

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

```
> ggplot(data.frame(result_n), aes(x = result_n)) +
+   geom_histogram(bins = 3, color = 'white', fill = 'lightgreen') +
+   theme_bw() + xlab('Number heads')
>
```



Probability Mass Function of a discrete random variable

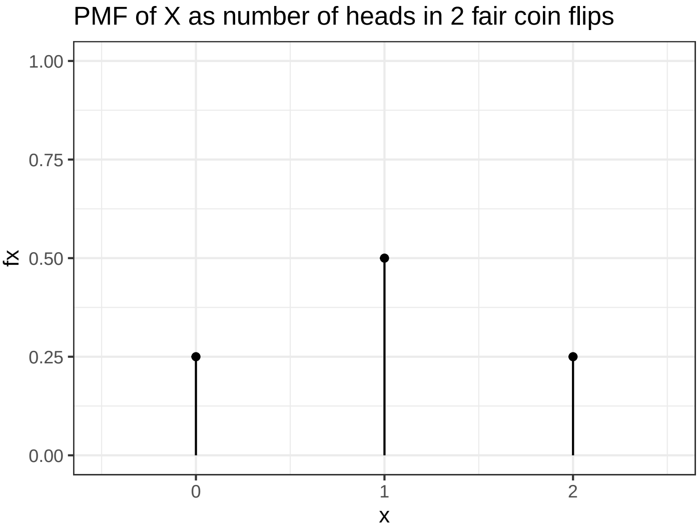
- A random variable is *discrete* if it takes countably many values.
- The probability mass function of a discrete RV X tells us the probability we will see an outcome at some value x.

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

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.

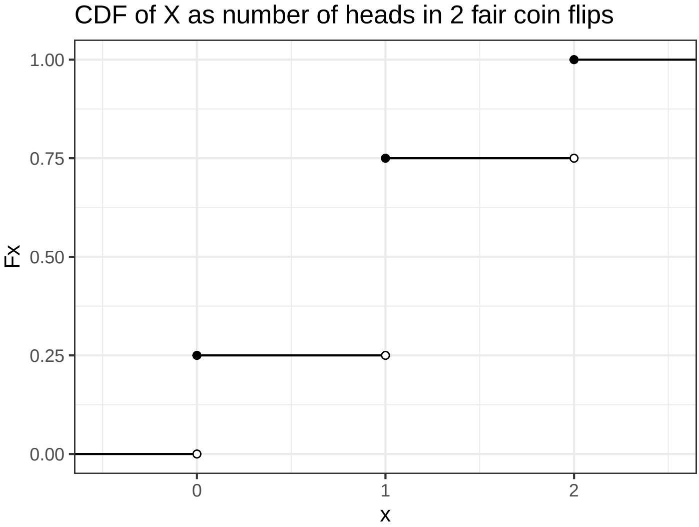
$$F(x) = \mathrm{P}[X \leq x]$$

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.

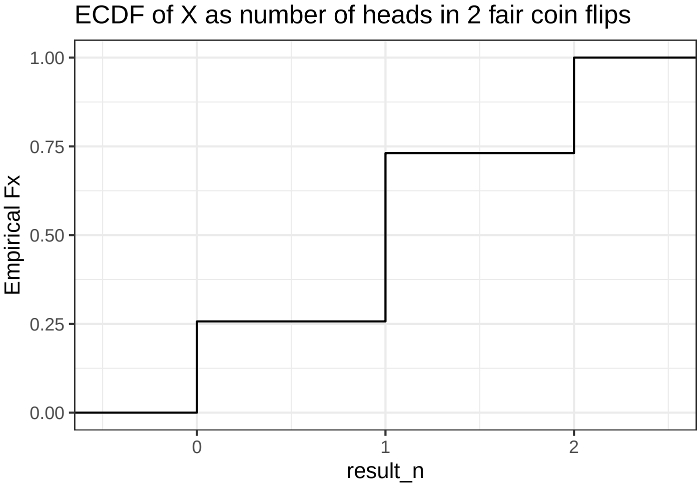
Illustrating the CDF of a RV



Illustrating the CDF of a RV

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

```
> ggplot(data.frame(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') + theme_bw()
>
```



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:

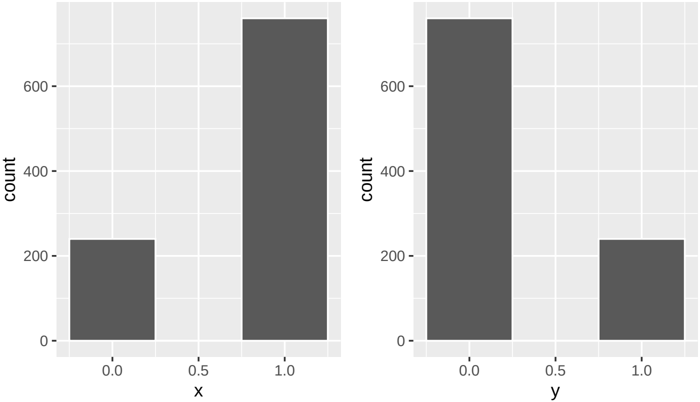
ω	$\text{P}[\{\omega\}]$	$X(\omega)$	$Y(\omega)$
TT	1/4	0	0
TH	1/4	1	0
HT	1/4	1	0
HH	1/4	1	1

or, considering the joint PMF,

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

```
> 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 <- data.frame(omega = result_n,
+                          x = ifelse(result_n == 'TT', 0, 1),
+                          y = ifelse(result_n == 'HH', 1, 0))
> options <- list(theme(panel.grid.minor = element_blank()),
+                 # save some style options
+                 scale_x_continuous(breaks = c(0, 1))) +
+   theme_bw()
> 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.

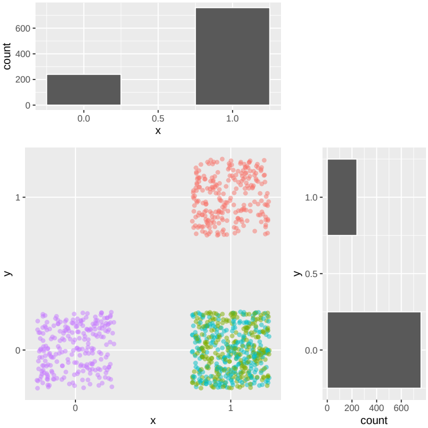
These are the *marginal* distributions of X and Y, i.e., their distributions where we *marginalize* or sum over the distribution of the other random variable.

Marginal distributions

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

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

Notational aside: we can 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. But often we will just omit the subscript for convenience.



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

Conditional distributions

We are also often interested in conditional relationships.

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

For example,

$$f_{Y|X}(y|x) = \begin{cases} 1 & x = 0, y = 0 \\ 2/3 & x = 1, y = 0 \\ 1/3 & x = 1, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

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

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.

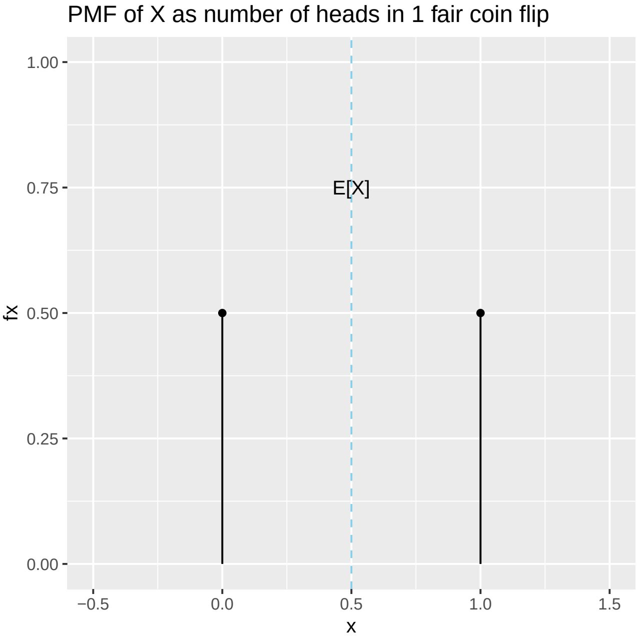
Let's flip a single coin, and let X be 1 if we get a head, and 0 otherwise.

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

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
E[X]
fx
1.00
0.75
0.50
0.25
0.00
−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

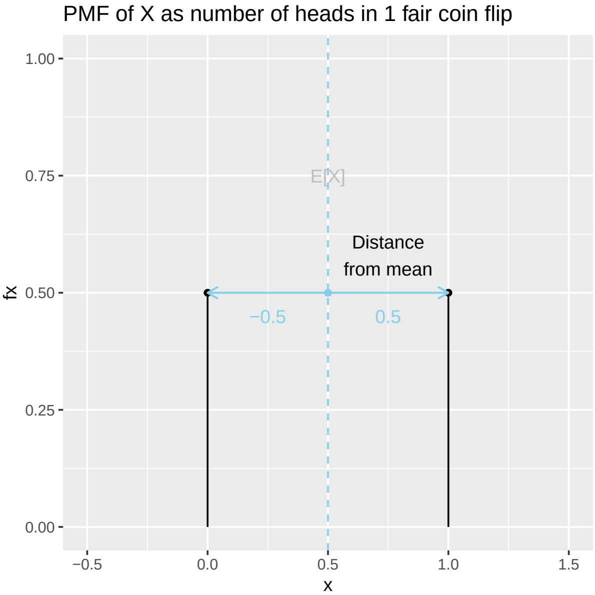
$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

Or equivalently,

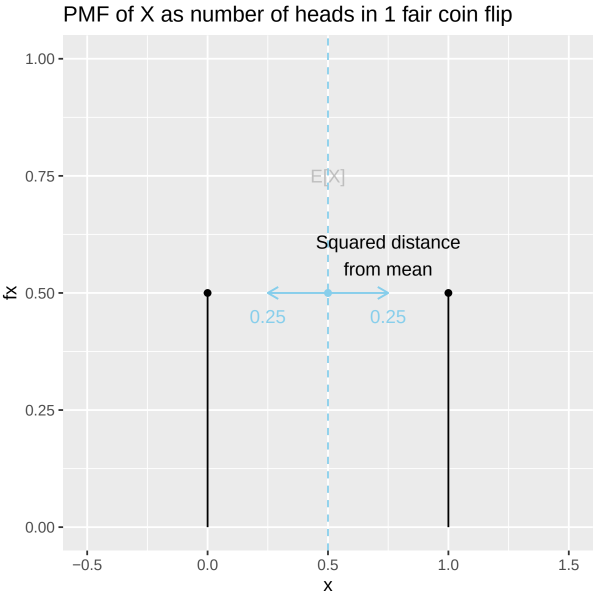
$$= \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

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.

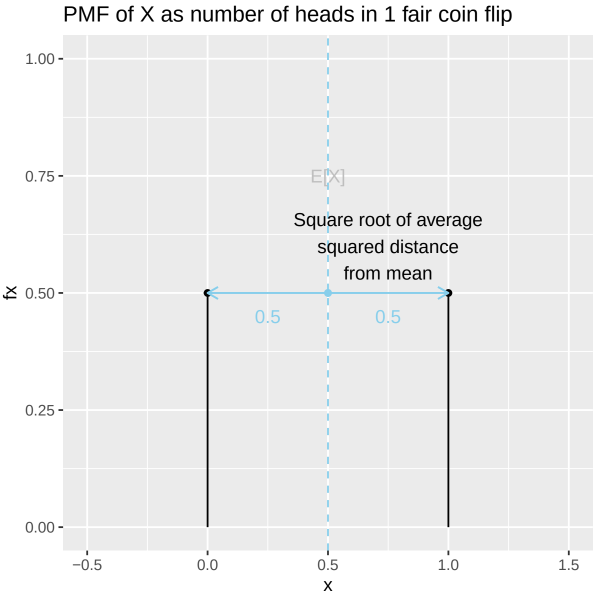
The variance is the average squared distance from the mean. The standard deviation is the square root of this.



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



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

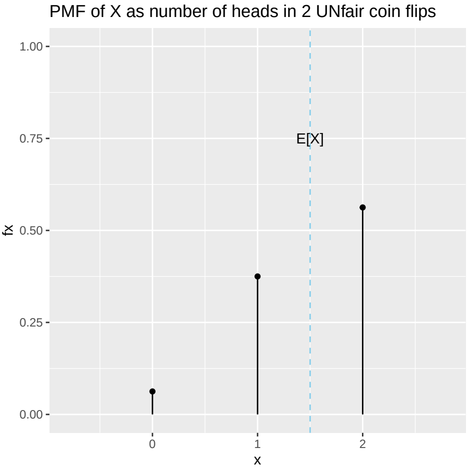


Let's take another example, where we flip a coin twice, and let X be the number of heads. However, let's say our coin is *not* fair, and the probability of getting a heads is 0.75.

The random variable's probability distribution is then:

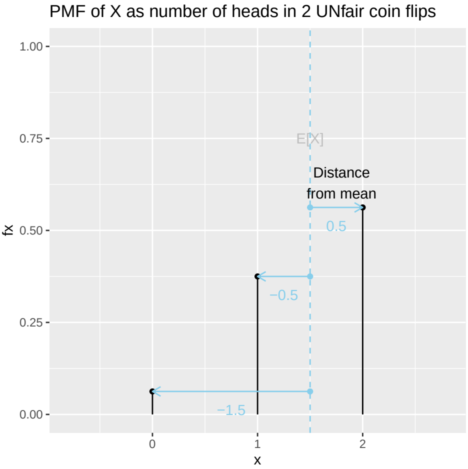
$$f(x) = \begin{cases} 1/16 & x = 0 \\ 3/8 & x = 1 \\ 9/16 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Let's take a look at the mean.



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

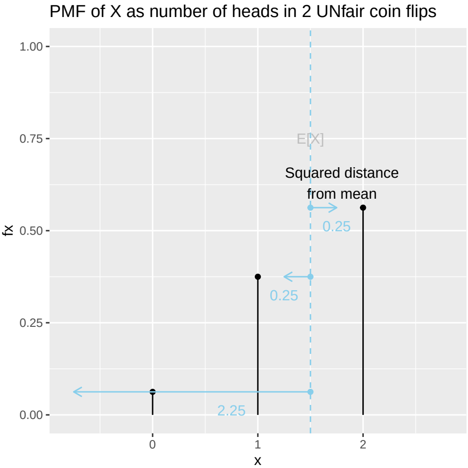
And the spread.



Variance = average squared distance from the mean

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

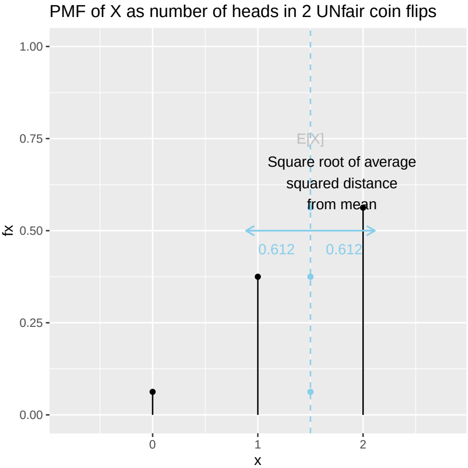
And the spread.



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

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

Applications

- Coin flips are a pretty trivial example of a random event $\rightarrow$ random variable.
- But often, as researchers, our job is to map events that happen in the world to variables in our data sets.

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

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

References I

Aronow, P. M. and Miller, B. T. (2019). Foundations of agnostic statistics. Cambridge University Press.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Wasserman, L. (2004). All of statistics: a concise course in statistical inference, volume 26. Springer.

Flipping two coins event space:

$$S = \{\emptyset,$$
$$\{HH\}, \{HT\}, \{TH\}, \{TT\},$$
$$\{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}$$
$$\{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\},$$
$$\{HH, HT, TH, TT\}\}$$

Back to terms