Probability in Plain Language + (little math)

PHS Launch!

Nicky Rahim (nrahim@g.harvard.edu)

August 29, 2025

Outline

Definitions

Game Time!!

Operating with probability

Conditional probability

Working with functions

DISCLAIMER: whatever KPop Demon Hunters references made in this presentation come from googling "synopsis of kpop demon hunters" and clicking two (2) links. I cannot guarantee 100% accuracy

Why is probability useful?

Probability allows us to make concrete statements about whether or not an event is likely to occur, taking into account both what we know and the randomness that exists in the world.

When applied in a formal context, probability is a key tool that allows us to make **inference** [1] both about relationships between variables and about whether those relationships are due to random variation or something more causal.

[1] see the rest of PHS2000A for more!

Definitions

Probability

Probability

Technical definition: a quantitative (i.e., numerical) measure of how likely a given event is to occur. Probability is quantified as a ratio of occurrences of the event of interest over all possible events, over an infinite number of attempted events. [2] Probabilities can take values between 0 and 1, inclusive.

We sometimes call the "occurrence of the event of interest" a "success."

[2] Adapted from Rosner. (1995). Fundamentals of biostatistics (4th ed.). Duxbury Press.

Question:

Why would we quantify probability over an infinite number of attempted events?

Random variable

Random variable

A **random variable** can be thought of as any event or measurement (e.g, rolling a dice, taking a blood pressure reading) that has probability associated with it. That is, the exact value the random variable will take is not known with certainty before the event or measurement.

We can develop more technical definitions of random variables as those that have a probability density/mass function associated with them. In other words, the possible values of a random variable follow a (possibly knowable) distribution of values, with some values possibly more common than others.

Random variable

Random variables can be either **discrete** or **continuous**.

- Discrete: the variable can only take on integer (0, 1, 2, 3, ...) values, either up to some finite limit or for an infinite number of such values.
- Continuous: the variable can take on any real value (e.g., $\sqrt{2}$, 3.2, $\frac{15}{4}$, but not $\sqrt{-2}$). These values could be over some finite range (e.g., between -1 and 1) or could include all real numbers (i.e., the real line $\mathbb{R}$ from $-\infty$ to ∞).

Random variables: Example

- Let's think of the number of concerts Huntr/x performs. What type of random variable would this be?
- Now, let's think of the length of each song by the Saja Boys? What type of random variable would this be?

Sample space

Sample space

The *sample space* ($\mathcal{S}$) is the set containing all possible outcomes of a trial.

- *Question*: What does the sample space for flipping a coin contain? What about rolling a dice?

Events of interest (e.g., rolling a 6) can be thought of as a **subset** of this sample space. We denote subsets as $A \subset \mathcal{S}$, where A is one possible outcome of the sample space $\mathcal{S}$.

This framework can be helpful if you want to think visually about probabilities.

Game Time !!!$@@!

DnD X Kpop Demon Hunters X Low effort

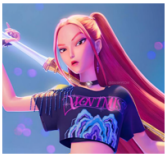
Setting:

You are studying at the Countway Library late at night when suddenly the building is overrun with demons trying to make a comeback tour (oh no!)

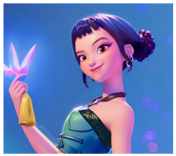
Fortunately, you just so happen to have a a magic D8 die that randomly summons a character from the hit Netflix movie Kpop Demon Hunters (w0w)

DnD X Kpop Demon Hunters X Low effort

The Cast:

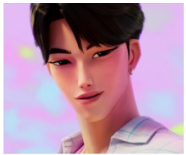


Mira



Zoey

Rumi



Jinu

Mystery

Abby

Romance

Baby

DnD X Kpop Demon Hunters X Low effort

Goal:

The goal is summon a certified demon hunter (Zoey, Mira, or Rumi) to save you. If you summon any other character you. . . lose

DnD X Kpop Demon Hunters X Low effort

Mechanics:

```
hunters <- cbind("Mira", "Zoey", "Rumi")
D8 <- cbind("Mira", "Zoey", "Rumi", "Jinu", "Abby", "Mystery", "Romance", "Baby")

total_wins <- 0

n <- 1 ## Number of game rounds
for (i in 1:n) {
  Roll <- sample(1:length(D8), 1) ## Randomly selects a number 1 through 8
  Summon <- D8[[Roll]] ## Picks the character the number is  associated with
  round_win <- Summon %in% hunters
  ## Checks if the rolled character is a hunter (i.e. are you saved)
  total_wins <- total_wins + round_win
  win_rate <- total_wins/n
}
win_rate
```

DnD X Kpop Demon Hunters X Low effort

Sample Space:

Probability:

Distribution??:

Operating with probability

Ands and *Ors*

When we conceive of probabilities as taking place in a sample space, it's easy to use some basic set operators to think about how we can combine the probabilities of multiple events.

AND

When we are interested in the probability of the co-occurrence of two (or more!) events of interest, we look at their **intersection**.

The intersection of A and B is the subset of the sample space where both A **and** B occur.

We denote intersections with the symbol $\cap$.

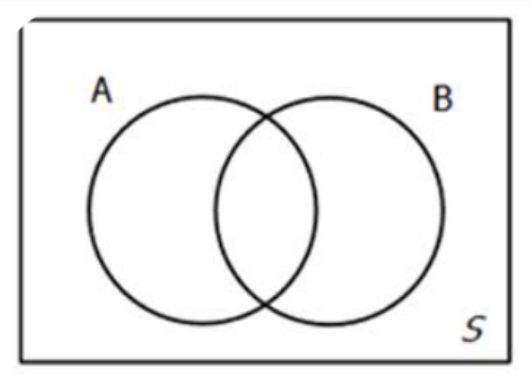
OR

When we are interested in whether at least one of a subset of events of interest occur, we look at their **union**.

The union of A and B is the subset of the sample space where A **or** B **or** both occur.

We denote unions with the symbol $\cup$.

Ands and *Ors*: Venn diagrams

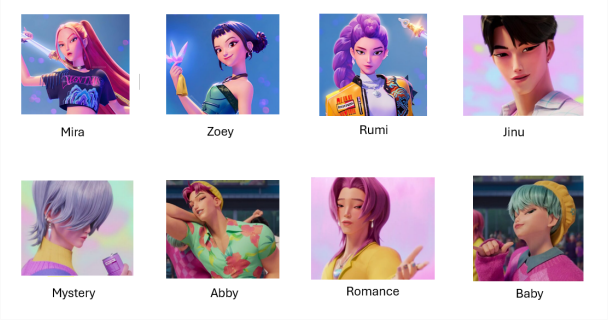


Ands and *Ors*: Kpop

Let's start with considering our kpop sample space, $(\mathcal{S})$.

What is the probability that a character is hunter **AND** is a demon?

What is the probability that a character is a hunter **OR** a demon?



Ands and *Ors*: Public Health Examples

Can you provide some public-health related examples of situations where the probabilities of an intersection of two events may be of interest?

What about situations where the union of two events is of interest?

Mathematical definitions

Intersections

$$P(A \cap B) = P(A|B) \times P(B)$$
$$= P(B|A) \times P(A)$$

Why? We need to consider first the probability of the first happening, **and then** the probability of the second even happening. This "and then" happens mathematically with multiplication. There are two ways we can define "first event" and "second" event, leading to two equivalent formulas.

Unions

$$P(A) \cup P(B) = P(A) + P(B) - P(A \cap B)$$

Why? We want to count everything that is in either the A or B portion of the sample space. The portion of the sample space that is $(A \cap B)$ is part of the subset A and part of the subset B. We need to subtract this value to avoid it being counted twice in our probability.

There are also some special cases where we can define intersections and unions that depend on the precise way that all of the events in our subset of interest relate to each other.

Two important types of events are **mutually exclusive** and **independent events**.

Mutually exclusive events

Events in a set are said to be mutually exclusive if the occurrence of one event means that no other event in the set can occur.

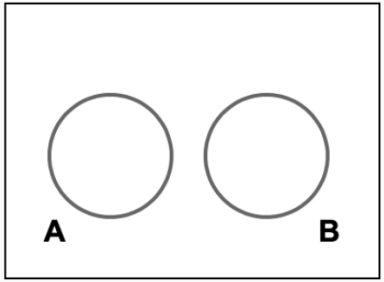
If event A is flipping a coin and landing on heads, while B is landing on tails, these events are mutually exclusive. Similarly, outcomes of a dice roll are mutually exclusive. These are simple examples, but helpful for thinking about how mutually exclusive events operate.

Can folks think of less trivial (and more public health-related) examples?

Mutually exclusive events

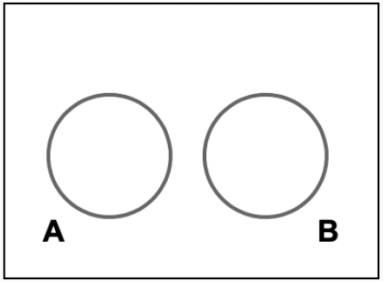
Intersections

$P(A \cap B) =$



Unions

$P(A \cup B) =$



Mutually exclusive events

Intersections

$$P(A \cap B) = 0$$

Why? Since A occurring means B cannot occur, and vice versa, the probability of A **and** B occurring is 0.

This will be true for any pair of mutually exclusive events, even if they don't make up the whole sample space.

Unions

$$P(A \cup B) = P(A) + P(B)$$

Why? We are interested in the probability of A occurring **or** B occurring, and know that only one can occur. We can list each event of interest and add up their probabilities.

Independent events

Two events A and B are independent (denoted by $A \perp\!\!\!\perp B$) if the probability of A occurring does not impact the probability of B occurring.

Two coin flips are independent, since the outcome of flip number 1 doesn't impact what we expect to see for flip number 2. However, probability of a Huntr/x single going Tik Tok viral and hiting the top of the music charts are *not* independent, as both are influenced by the songs underlying popularity and can influence each other.

Question: are mutually exclusive events independent?

Independent events

Intersections

$$P(A) \cap P(B) = P(A) \times P(B)$$

Why? How does this relate to the original expression for the intersection of probabilities?

Unions

$$P(A) \cup P(B) = P(A) + P(B) \times [1 - Pr(A)]$$

How did we end up here?

Independent events: A disclaimer

The rule $P(A) \times P(B) = Pr(A \cap B)$ is sometimes used to define independent events. However, whether you use this to try and identify independent events or claim that two events are independent so that you can use this rule can get a little circuitous.

Independence is a property we need to assume for a lot of the models and methods we use in quantitative research to hold, but it's not something we can always check with this rule!

Conditional probability

Conditional vs Unconditional

A few slides back, we sneakily introduced the concept of **conditional** probability. Now we want to make that piece explicit. A probability is conditional when we are concerned with the probability of an event **among** some group with a particular characteristic, or **among** a group who experiences a second event of interest. In notation, a conditional probability is denoted $P(A|B)$.

Conditional probabilities are common in the quantitative methods learned in this course. Sometimes conditional probabilities are necessary to make the methods "work" and remove bias; sometimes conditional probabilities are of scientific interest in their own right.

Conditional probabilities: K-Pop Demon Hunters

Unconditional

- probability a person is a Rumi stan
- probability of having black hair

Conditional

- probability a person is a Rumi stan **among** Huntr/x fans
- probability of having black hair **among** the Saja boys

Conditional probabilities: Public health examples

Unconditional

- probability of being hospitalized with COVID-19
- probability of death after a heart attack
- probability of developing childhood asthma

Conditional

- probability of being hospitalized with COVID-19 **among** those who are vaccinated
- probability of death after a heart attack **among** those over age 70
- probability of developing childhood asthma **among** those living in historically redlined neighborhoods

Conditional probability notation

We notate conditional probability with a vertical bar. If we want to know the probability of A conditional on event B, we write $Pr(A|B)$.

Mathematically, we define a conditional probability as

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

Question

Is there a difference between and intersection and a conditional probability?

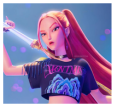
Are $P(A \cap B)$ and $P(A|B)$ the same expression?

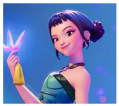
Conditional probabilities vs intersections

$P(A \cap B)$ and $P(A|B)$ are *not* the same expression! Let's explore this using kpop characters first.

$P(\text{demon} \cap \text{hunter}) = ?$

$P(\text{demon}|\text{hunter}) = ?$


Mira


Zoey


Rumi


Jinu


Mystery


Abby

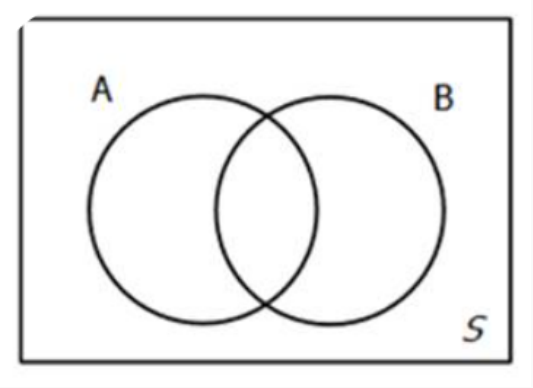

Romance


Baby

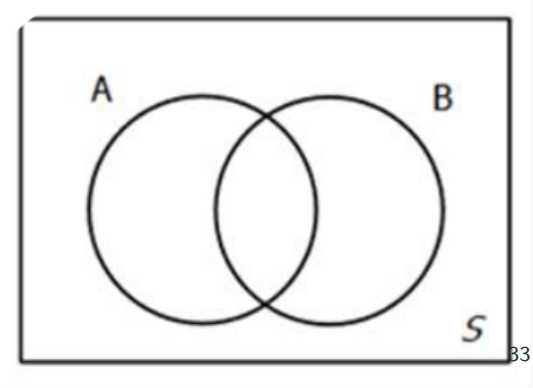
Conditional probabilities vs intersections

$P(A \cap B)$ and $P(A|B)$ are *not* the same expression!

$\mathbf{P(A \cap B)}$



$\mathbf{P(A|B)}$



Conditional probabilities vs intersections

A coin flipping example helps illustrate this. Let A be heads and B be tails, and let the subscripts denote the sequence of flips we're performing.

What values do you expect for $P(A_2 \cap A_1)$ and $P(A_2|A_1)$?

Conditional probabilities vs intersections

$P(A_2 \cap A_1)$ is the probability of flipping heads on flip one **and then** flipping heads on flip two. In our universe, this has a value of $P(A_2 \cap A_1) = 0.25$.

In contrast, $P(A_2|A_1)$ is the probability of flipping heads on flip two, **among** the set of sequences that started with a heads on flip one. Since we know successive flips are independent, in our universe, this has a value of $P(A_2|A_1) = 0.5$.

Notice that in the example above, $P(A_2|A_1) = 0.5$, where A_1 and A_2 are independent ($A_1 \perp\!\!\!\perp A_2$). In fact, this will always be true of independent random variables.

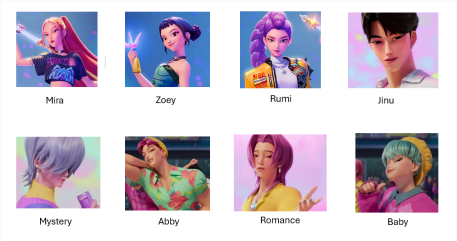
If A and B are two random variables such that $A \perp\!\!\!\perp B$, then $P(A = a) = P(A = a|B = b)$ for all possible values of b. This is a useful identity that you will use a *lot* in this class and others!

Flipping the conditioning statement

In general, it is unfortunately true that $P(A|B) \neq P(B|A)$. Returning to demons and/or hunters.

$P(\text{demon}|\text{hunter}) = ?$

$P(\text{hunter}|\text{demon}) = ?$



But wait! Statistics has a solution

Bayes' Theorem

Bayes' Theorem allows us to go from $P(A|B)$ to $P(B|A)$. Let's start with two things we know and do a small derivation:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Bayes' Theorem

Thus, $P(B|A) = \frac{P(A|B)*P(B)}{P(A)}$!

working with functions

Logarithmic and Exponent rules

Law of Exponents	Law of Logarithms
$\exp(a+b) = \exp(a) \times \exp(b)$ [3]	$\log(a) + \log(b) = \log(a \times b)$
$\exp(a-b) = \exp(a)/\exp(b)$	$\log(a) - \log(b) = \log(a/b)$
$(\exp(a))^b = \exp(a \times b)$	$\log(a^b) = b \times \log(a)$

[3] $\exp(a) = e^a$

Logit and expit functions

The logit function:

$y = \text{logit}(x) = \log(\frac{x}{1-x})$

- Commonly encountered in logistic regressions to convert probabilities back to a wider range of real number
- Domain: [0,1]
- Range: $[-\infty, +\infty]$

The expit function:

$y = \text{expit}(x) = (\frac{1}{1+e^{-x}})$

- Often used to transform data back to a probability scale. **Note: this is the inverse of logit function**
- Domain: $[-\infty, +\infty]$
- Range: [0,1]

4

[4] We often just assume a log-base of e, such that log(x) = ln(x)

Probability and odds

	Probability	**Odds**
Definition	The number of successes, textbfx, out of a total number of trials, **n** (i.e. the fraction of times you expect to see that event in many trials)	The number of successes, **x**, compared to the number of failures, **n-x** (i.e. the probability that the event will occur divided by the probability that the event will not occur)
Measure	Proportion	Ratio
Range	[0,1]	$[0,+\infty]$
Formula	$\frac{x}{n}$	$\frac{x}{n-x}$


You're
GOLDEN
HUNTR/X

Thank you!

Many thanks to Logan Beyer (TF for PHS2000A 2023) for creating the previous version of this presentation, and to the rest of the PHS2000A teaching fellows for comments, edits, and sense-checks on these slides.

And thank you to you all for putting up with me for the past hour-ish