Center for Statistics and the Social Sciences Math Camp 2021

Lecture 5: Introduction to Probability

Peter Gao & Jessica Kunke

Department of Statistics
University of Washington

September 15, 2021

Probability

Motivating Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the probability that you have the disease. That is, given someone has tested positive for the disease what is the chance that they have the disease?

Probability

Motivating Example

What information do we have?

- $P(+\ \mathrm{test}|\ \mathrm{diseased}) = 0.98$
- $P(-\ \mathrm{test}|\ \mathrm{healthy}) = 0.95$

What quantity do we want?

- $P(\ \mathrm{diseased}| +\ \ \mathrm{test})$

So, what is the probability of disease given a positive test? 16.5%

Set Notation

A **set** is a collection of elements from a population.
Examples

- Positive Integers ≤ 5 : $A = \{1, 2, 3, 4, 5\}$
- Primary Colors: B={blue, red, yellow }
- Odd Numbers: $C = \{1, 3, 5, 7, 9...\}$

A set is an empty set if it contains no elements: written $\emptyset$ or $D = \{\emptyset\}$. An example of an empty set would be integers that are greater than 4 and less than 1.

A set is called the universal set if it contains all the elements in the population: written Ω or $E = \{\Omega\}$

Set Notation

Intersection, $\cap$

The **intersection** of two sets A, B is the set of all elements that are in A **AND** B. The intersection is denoted $A \cap B$.

Examples

- $A = \{1, 2, 3, 4, 5\}$, $B = \{2, 4, 6, 8, 10\}$, $A \cap B = \{2, 4\}$
- $A = \{$ Odd numbers $\}$, $B = \{$ Even numbers $\}$, $A \cap B = \{\emptyset\}$
- $A = \{$ Integers less than 5 $\}$, $B = \{$ Integers greater than 2$\}$, $A \cap B = \{3, 4\}$

Set Notation

Union, $\cup$

The **union** of two sets A, B is the set of all elements that are in A **OR** B. The intersection is denoted $A \cup B$.

Examples

- $A = \{1, 2, 3, 4, 5\}$, $B = \{2, 4, 6, 8, 10\}$,
  $A \cup B = \{1, 2, 3, 4, 5, 6, 8, 10\}$
- $A = \{$ Odd numbers $\}$, $B = \{$ Even numbers $\}$, $A \cup B = \{$All integers $\}$
- $A = \{$ Integers less than 5 $\}$, $B = \{$ Integers greater than 2$\}$,
  $A \cup B = \{$ All integers $\}$

Set Notation

Subest, $\subseteq$

One set can be contained inside another. If all elements of A are also in B, then A is a **subset** of B. The subset is denoted $A \subseteq B$.

Examples

- $A = \{1, 5\}$, $B = \{1, 2, 3, 4, 5\}$, $A \subseteq B$
- $A = \{1, 5\}$, $B = \{1, 4, 9, 18\}$, A is not a subset of B, because 5 is not in B.

If $A \subseteq B$ AND $B \subseteq A$ then all of the elements in A are in B and all of the elements of B are in A, so $A = B$.

Set Notation

Mutually Exclusive

If the intersection of A and B is $\emptyset$, i.e. they have no elements in common, A and B are called **mutually exclusive**.

Examples

- $A = \{1, 5\}$, $B = \{2, 3\}$
- $A = \{$ Odd numbers $\}$, $B = \{$ Even numbers $\}$
- $A = \{$ Integers less than 0 $\}$, $B = \{$ Integers greater than 0$\}$

Set Notation

Complement

The **complement** of a set is the set of elements in the population that are not in A. The complement is denoted by A^c.

Examples

- Population: Integers 1-10$\rightarrow A = \{1, 2, 9\}, A^c = \{3, 4, 5, 6, 7, 8, 10\}$
- Population: All integers but zero $\rightarrow A = \{$ Odd numbers $\}$, $A^c = \{$ Even numbers $\}$

Sample Space

An **experiment** is an action or process of observation. It has only one outcome but we do not know what it will be with certainty until the experiment is carried out.

Familiar examples would be rolling a dice or flipping a coin.

The **sample space** is made up of all the possible outcome of the experiment and usually denoted by S. In the experiments above we would have $S = \{1, 2, 3, 4, 5, 6\}$ or $S = \{$ Heads, Tails $\}$.

Determining the sample space can be tricky. Before you start listing the possible outcomes it is wise to think about how many there will be.

Sample Space

Examples

Flipping 3 coins:
There are 2 choices for each coin.
The number of outcomes is then $2 \cdot 2 \cdot 2 = 8$.
$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

Rolling a dice and flipping a coin:
There are 6 choices for the die and 2 choices for the coin.
The number of outcomes is then $6 \cdot 2 = 12$.
$S = \{1H, 2H, 3H, 4H, 5H, 6H, 1T, 2T 3T, 4T, 5T, 6T\}$

What would be the sample space be for rolling 2 dice? How many outcomes? $6 \cdot 6 = 36$.

Sample Space

Events

An **event** is a subset of the sample space.

Examples:

$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

- Getting 2 heads: HHT, HTH, THH
- Getting an odd number of tails: HHT, HTH, THH, TTT
- Getting more than 1 head: HHH, HHT, HTH, THH

$S = \{1H, 2H, 3H, 4H, 5H, 6H, 1T, 2T3T, 4T, 5T, 6T\}$

- Rolling higher than a 4: 5H, 6H, 5T, 6T
- Getting a head: 1H, 2H, 3H, 4H, 5H, 6H
- Rolling a 3 or a 2: 2H, 3H, 2T, 3T

Probability

We can find the probability of an event by adding up the probabilities of the elements in the event.

Rolling a fair die:
$S = \{1, 2, 3, 4, 5, 6\}$
Each element has probability $1/6$.

- $A = \{\text{roll} \leq 4\} = \{1, 2, 3, 4\}$,
  $P(A) = P(1){+}P(2){+}P(3){+}P(4) = 1/6{+}1/6{+}1/6{+}1/6 = 4/6$
- $B = \{\text{roll odd}\} = \{1, 3, 5\}$,
  $P(A) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 3/6$

Probability

Unions

We can find the probability of the union of two events.

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

We are adding up the probability of the elements in the even A and the elements in the event B. However, in doing that we count the elements that are in both A and B ($A \cap B$) twice. So we must subtract the intersection of A and B.

Rolling a fair die:
$S = \{1, 2, 3, 4, 5, 6\}$
Each element has probability $1/6$.
$A = \{\text{roll} \leq 4\}$ and $B = \{1, 3, 5\}$
We know that $P(A) = 4/6$ and $P(B) = 3/6$
The intersection $A \cap B = \{1, 3\}$, thus $P(A \cap B) = 2/6$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/6 + 3/6 - 2/6 = 5/6$$

Probability

Unions

Selecting cards from a deck (52 total cards):
Each element has probability $1/52$.

$A = \{\mathrm{Hearts}\}$ and $B = \{\mathrm{King}\}$

There are 13 Hearts in the deck, so $P(A) = 13/52$.
There are 4 Kings in a deck, so $P(B) = 4/52$.

The intersection $A \cap B = \{\mathrm{King\ of\ Hearts}\}$, thus $P(A \cap B) = 1/52$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 13/52 + 4/52 - 1/52 = 16/52$$

Probability

The sum of the probabilities of all elements in the sample space MUST be 1. Since the complement of a set A is everything in the population that is not A, the probability of of the complement can be found:

$$P(A^c) = 1 - P(A)$$

Note:

$$\begin{aligned} 1 &= P(S) \\ &= P(A \cup A^c) \\ &= P(A) + P(A^c) - P(A \cap A^c) \\ &= P(A) + P(A^c) - P(\emptyset) \\ &= P(A) + P(A^c) - 0 = P(A) + P(A^c) \end{aligned}$$

Rolling a fair die example: $A = \{\text{Roll a } 5\}$, $A^c = \{1, 2, 3, 4, 6\}$.

$$P(A^c) = 1 - P(A) = 1 - 1/6 = 5/6$$

Conditional Probability

Sometimes knowing that one event has occurred changes what you know about the probability of another event. For example if the sidewalk is wet in the morning you might think it is more likely that it rained last night than if you didn't know anything about the sidewalk.

The **conditional probability** of A given B is the probability that A occurs given that B has been observed. It is denoted $P(A|B)$.

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

The denominator is the probability of B occurring. The numerator is the probability of A and B happening.

Conditional Probability

Example

Rolling a fair die:
$A = \{\text{odd number}\}$ and $B = \{\text{roll} < 3\}$

We know that $P(A) = P(1) + P(3) + P(5) = 3/6$ and
$P(B) = P(1) + P(2) = 2/6$
The intersection $A \cap B = \{1\}$, thus $P(A \cap B) = 1/6$

What is the probability that your roll is odd if you know that it is less than 3? That is, $P(A|B) = P(\text{Odd}| < 3)$?

$$\begin{aligned} P(\text{Odd}| < 3) &= \frac{P(\text{Odd } \cap < 3)}{P(< 3)} \\ &= \frac{P(1)}{P(1) + P(2)} = \frac{1/6}{1/6 + 1/6} = \frac{1/6}{2/6} = 1/2 \end{aligned}$$

Conditional Probability

Example

Rolling a fair die:
$A = \{\text{odd number}\}$ and $B = \{\text{roll} < 3\}$

We know that $P(A) = P(1) + P(3) + P(5) = 3/6$ and
$P(B) = P(1) + P(2) = 2/6$
The intersection $A \cap B = \{1\}$, thus $P(A \cap B) = 1/6$

What is the probability that your is less than 3 given that the roll is odd? That is, $P(B|A) = P(< 3|\text{Odd})$?

$$\begin{aligned} P(< 3|\text{Odd}) &= \frac{P(\text{Odd} \cap < 3)}{P(\text{Odd})} \\ &= \frac{P(1)}{P(1) + P(3) + P(5)} = \frac{1/6}{3/6} = 1/3 \end{aligned}$$

Independence

What if knowing B does not give us any information about A? That is, if $P(A|B) = P(A)$, then we say that A and B are **independent**.

Independence also means:

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

Thus, $P(A) \cdot P(B) = P(A \cap B)$ allows us to check for independence.

Independence

Example

Rolling a fair die:
$A = \{1, 2, 3, 4\}$ and $B = \{\text{odd number}\}$

We know that $P(A) = 4/6$ and $P(B) = 3/6$
The intersection $A \cap B = \{1, 3\}$, thus $P(A \cap B) = 2/6 = 1/3$

$P(A) \cdot P(B) = 4/6 \cdot 3/6 = 12/36 = 1/3$

Knowing that you have rolled a 1, 2, 3 or 4 doesn't give you any information about whether or not you rolled an odd number (because there are 2 even and 2 odd) and vice a versa.

Dependent Events

Example

Rolling a fair die:
$A = \{1, 2, 3, 5\}$ and $B = \{\text{odd number}\}$

We know that $P(A) = 4/6$ and $P(B) = 3/6$
The intersection $A \cap B = \{1, 3, 5\}$, thus $P(A \cap B) = 3/6 = 1/2$

$P(A) \cdot P(B) = 4/6 \cdot 3/6 = 12/36 = 1/3$

Knowing that you have rolled a 1, 2, 3 or 5 does give you information about whether or not you rolled an odd number (because there is 1 even and 3 odd).

Bayes Rule

Sometimes you may know one conditional probability, but not the other. How can you use the first conditional probability to find the other one?

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

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

This implies:

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

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

Bayes Rule

Testing Example

Let's return to our motivating example. What do we know? Let's use D^+ =diseased, D^- =healthy, $+$ =positive test, and $-$ =negative test.

- $P(+|D^+) = 0.98$, $P(-|D^+) = 0.02$
- $P(-|D^-) = 0.95$, $P(+|D^-) = 0.05$
- $P(D^+) = 0.01$, $P(D^-) = 0.99$

$$
\begin{aligned}
P(D^+|+) = \frac{P(+|D^+)P(D^+)}{P(+)} &= \frac{P(+|D^+)P(D^+)}{P(+\cap D^+) + P(+\cap D^-)} \\
&= \frac{P(+|D^+)P(D^+)}{P(+|D^+)P(D^+) + P(+|D^-)P(D^-)} \\
&= \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.05 \cdot 0.99} \\
&= 0.165
\end{aligned}
$$

Bayes Rule

Testing Example

As the population prevalence increases the $P(D^+|+)$ (this is called the positive predictive value) increases:

Prevalence	$P(D^+|+)$
0.01	16.5%
0.02	28.6%
0.05	50.8%
0.10	68.5%
0.25	86.7%

Relative Risk

Straightforward application of conditional probability

The Relative Risk (or Risk Ratio) is very easy to calculate when you have a binary outcome and binary exposure.

$$RR = \frac{P[\text{Disease} \mid \text{Exposed}]}{P[\text{Disease} \mid \text{Not Exposed}]}$$

The RR is useful when you are interested in comparing the probability of some event (or disease) based on exposure status.

Calculating the Relative Risk

Given a 2x2 table:

	Disease	Not Disease	
Exposed	a	b	(a+b)
Not Exposed	c	d	(c+d)
	(a+c)	(b+d)	

you can calculate the RR based on the values in the cells.

$$RR = \frac{\text{P[Disease}|\text{Exposed]}}{\text{P[Disease}|\text{Not Exposed]}} = \frac{a/(a+b)}{c/(c+d)}$$

Relative Risk

Example

Let's use lung cancer as an example:

	Lung Cancer	No Lung Cancer	
Smoker	40	160	200
Non-Smoker	5	195	200
	45	355	

$$RR = \frac{\text{P[Disease}|\text{Exposed]}}{\text{P[Disease}|\text{Not Exposed]}} = \frac{a/(a+b)}{c/(c+d)} = \frac{40/200}{5/200} = 8$$

This suggests that smokers are 8 times more likely to get lung cancer than non-smokers.

Case-Control Studies

and the Relative Risk

For rare diseases it is often necessary to select study participants based on their disease status (instead of totally at random or based on exposure). In case-control studies we fix the number of diseased and non-diseased participants and look at exposure.

Thus, we can estimate $P[\text{Exposed}|\text{Disease}]$ and $P[\text{Exposed}|\text{Not Disease}]$.

However, without population-level prevalence of disease we cannot estimate the RR of disease given exposure based on these quantities.

Case-Control Studies

and the Relative Risk

IF you have an estimate of population-level prevalence of disease you could estimate the RR using Bayes Rule:

$$P[\text{Disease}|\text{Exposed}] = \frac{P[\text{Exposed}|\text{Disease}]P[\text{Disease}]}{P[\text{Exposed}]}$$

where,

$P[\text{Exposed}] = P[\text{Exposed}|\text{Disease}]P[\text{Disease}] +$
$P[\text{Exposed}|\text{Not Disease}]P[\text{Not Disease}]$.

You can calculate $P[\text{Disease}|\text{Not Exposed}]$ in the same fashion.

Odds Ratio

For probabilities p_1 and p_2 the Odds Ratio (OR) is calculated:

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

and the RR is calculated:

$$RR = \frac{p_1}{p_2}$$

For rare events (resulting in small probabilities) $1 - p_1 \approx 1$ and $1 - p_2 \approx 1$, thus $OR \approx RR$.

Case-Control

and the Odds Ratio

The odds ratio has a nice property of being equivalent for [Exposure|Disease] and [Disease|Exposure]. Given a 2x2 table:

	Disease	Not Disease	
Exposed	a	b	(a+b)
Not Exposed	c	d	(c+d)
	(a+c)	(b+d)	

$$OR = \frac{\text{odds[Disease|Exposed]}}{\text{odds[Disease|Not Exposed]}} = \frac{\frac{a}{a+b}/\frac{b}{a+b}}{\frac{c}{c+d}/\frac{d}{c+d}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$OR = \frac{\text{odds[Exposed|Disease]}}{\text{odds[Exposed|Not Disease]}} = \frac{\frac{a}{a+c}/\frac{c}{a+c}}{\frac{b}{b+d}/\frac{d}{b+d}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Odds Ratio vs. Relative Risk

Odds ratios are statistically convenient and the only viable option for case-control settings.

However, it is important to note that when events are not rare the odds ratio is not a very good approximation of the relative risk.

For the smoking example we have:

$$RR = \frac{40/200}{5/200} = 8$$

$$OR = \frac{40/160}{5/195} = 9.75$$

The OR is higher than the RR, but not totally unreasonable.

Odds Ratio vs. Relative Risk

	Passed Exam	Failed Exam	
Took Math Camp	80	20	100
No Math Camp	40	60	100
	120	80	

$$RR = \frac{80/100}{40/100} = 2$$

$$OR = \frac{80/20}{40/60} = 6$$

The OR is three times larger than the RR. Without careful interpretation, the results could be quite misleading.