Probability (part 2)

Ryan Miller

Outline

Introduction

- Previously, we counted outcomes and divided by the size of the sample space to determine probabilities
- This approach isn't always viable, so we'll now cover a few *probability laws* to help us make more sophisticated probability calculations

Disjoint Events

- Two *events* are **disjoint** if they have no *outcomes* in common
  - Consider rolling a six-sided die, the event of rolling a six is disjoint from the event rolling an odd number

Disjoint Events

- Two *events* are **disjoint** if they have no *outcomes* in common
  - Consider rolling a six-sided die, the event of rolling a six is disjoint from the event rolling an odd number
- For two disjoint events, we can find the probability of *unions* by addition
  - $P(A \text{ or } B) = P(A) + P(B)$
  - For a six-sided die, $P(\text{Six or Odd Number}) = P(\text{Six}) + P(\text{Odd Number}) = 1/6 + 3/6 = 2/3$

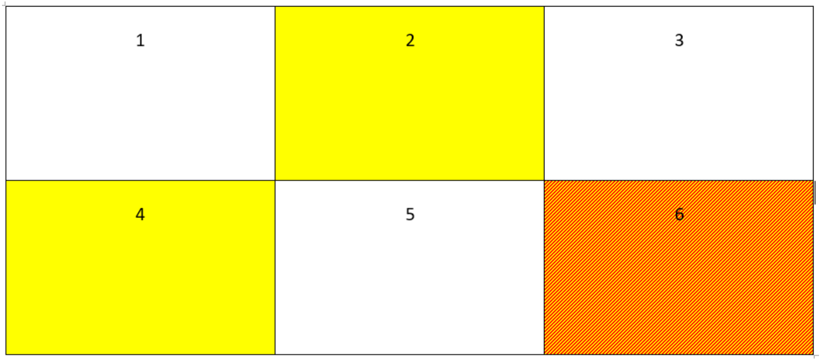
Disjoint Events

It's easy to visually confirm this example by looking at a simple representation of the sample space:

1	2	3
4	5	6

Non-disjoint events

In contrast, consider $P(\text{Six or Even Number})$, clearly these events are *not disjoint*, so adding their probabilities would be a mistake



The Addition Rule

- In general, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  - This is known as the **addition rule**
  - In the special case where events A and B are *disjoint*, $P(A \text{ and } B) = 0$

The Addition Rule

- In general, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  - This is known as the **addition rule**
  - In the special case where events A and B are *disjoint*, $P(A \text{ and } B) = 0$
- In our previous examples:

$P(\text{Six or Odd Number}) = P(\text{Six}) + P(\text{Odd Number}) -$
$P(\text{Six and Odd Number}) = 1/6 + 3/6 - 0 = 2/3$

The Addition Rule

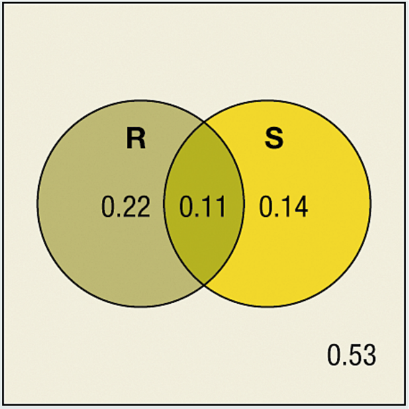
- In general, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  - This is known as the **addition rule**
  - In the special case where events A and B are *disjoint*, $P(A \text{ and } B) = 0$
- In our previous examples:

$P(\text{Six or Odd Number}) = P(\text{Six}) + P(\text{Odd Number}) - P(\text{Six and Odd Number}) = 1/6 + 3/6 - 0 = 2/3$

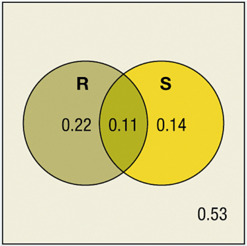
$P(\text{Six or Even Number}) = P(\text{Six}) + P(\text{Even Number}) - P(\text{Six and Even Number}) = 1/6 + 3/6 - 1/6 = 1/2$

Venn Diagrams

- Venn diagrams are frequently used as a visual aid when learning the addition and complement rules
- The diagram below depicts survey results where 33% of college students were in a relationship (R), 25% were involved in sports (S), and 11% were in both

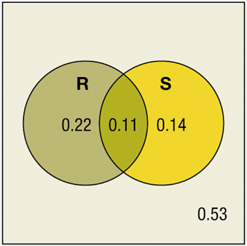


Venn Diagrams - Example



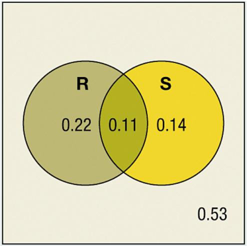
1) $P(R \text{ or } S) = 0.22 + 0.11 + 0.14 = 0.47$ (direct calculation)

Venn Diagrams - Example



1) $P(R \text{ or } S) = 0.22 + 0.11 + 0.14 = 0.47$ (direct calculation)
2) $P(R \text{ or } S) = 0.33 + 0.25 - 0.11 = 0.47$ (addition rule)

Venn Diagrams - Example



1) $P(R \text{ or } S) = 0.22 + 0.11 + 0.14 = 0.47$ (direct calculation)
2) $P(R \text{ or } S) = 0.33 + 0.25 - 0.11 = 0.47$ (addition rule)
3) $P(R \text{ or } S) = 1 - P(\text{Neither}) = 1 - 0.53 = 0.47$ (complement rule)

Practice

Consider a high school graduating class of 100 students. Among them, 50 had taken calculus, 60 had taken physics, and 40 had taken both.

1) Draw a Venn diagram to represent this scenario
2) Using the addition rule, find the probability that a randomly selected student has taken either calculus or physics
3) Find the probability that a randomly selected student has taken neither calculus nor physics

Practice (solution)

1) Sketched
2) $P(C \text{ or } P) = 50/100 + 60/100 - 40/100 = 70/100 = 0.7$
3) $P((C \text{ or } P)^C) = 1 - 0.7 = 0.3$

Conditional Probability

- Analogous to *conditional proportions* (row and column percentages), is the concept of a **conditional probability**
- Conditional probability is used in scenarios involving *dependent events*
  - For example, the probability that Steph Curry makes a free throw could depend on whether he's playing in a home game or an away game

Conditional Probability

- Analogous to *conditional proportions* (row and column percentages), is the concept of a **conditional probability**
- Conditional probability is used in scenarios involving *dependent events*
  - For example, the probability that Steph Curry makes a free throw could depend on whether he's playing in a home game or an away game
- We use a vertical bar to denote conditional probabilities: $P(A|B)$
  - In this example, we might define "A" to be making the free throw and "B" to be playing at home
  - As you'd expect, conditional probabilities can be *estimated* from a contingency table

Example - Conditional Probabilities

ACTN3 is known as the fast twitch gene, everyone has one of three genotypes (XX, RR, or RX). The table below summarizes a sample of 301 elite athletes:

	RR	RX	XX	Total
Sprint/power	53	48	6	107
Endurance	60	88	46	194
Total	113	136	52	301

From this table, let's estimate a few different probabilities:

1) An endurance athlete has genotype XX?

Example - Conditional Probabilities

ACTN3 is known as the fast twitch gene, everyone has one of three genotypes (XX, RR, or RX). The table below summarizes a sample of 301 elite athletes:

	RR	RX	XX	Total
Sprint/power	53	48	6	107
Endurance	60	88	46	194
Total	113	136	52	301

From this table, let's estimate a few different probabilities:

1) An endurance athlete has genotype XX? $46/194 = 0.237$
2) An athlete with the XX genotype is an endurance athlete?

Example - Conditional Probabilities

ACTN3 is known as the fast twitch gene, everyone has one of three genotypes (XX, RR, or RX). The table below summarizes a sample of 301 elite athletes:

	RR	RX	XX	Total
Sprint/power	53	48	6	107
Endurance	60	88	46	194
Total	113	136	52	301

From this table, let's estimate a few different probabilities:

1) An endurance athlete has genotype XX? $46/194 = 0.237$
2) An athlete with the XX genotype is an endurance athlete? $46/52 = 0.885$
3) An athlete has the XX genotype *and* is an endurance athlete?

Example - Conditional Probabilities

ACTN3 is known as the fast twitch gene, everyone has one of three genotypes (XX, RR, or RX). The table below summarizes a sample of 301 elite athletes:

	RR	RX	XX	Total
Sprint/power	53	48	6	107
Endurance	60	88	46	194
Total	113	136	52	301

From this table, let's estimate a few different probabilities:

1) An endurance athlete has genotype XX? $46/194 = 0.237$
2) An athlete with the XX genotype is an endurance athlete? $46/52 = 0.885$
3) An athlete has the XX genotype *and* is an endurance athlete? $46/301 = 0.153$

The Multiplication Rule

The relationship between these probabilities motivates the **multiplication rule**, which states:

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

In our previous example, notice:

1) $P(\text{XX}|\text{End}) = 46/194 = 0.237$
2) $P(\text{End}) = 194/301 = 0.645$
3) $P(\text{XX and End}) = 46/301 = 0.153$

	RR	RX	XX	Total
Sprint/power	53	48	6	107
Endurance	60	88	46	194
Total	113	136	52	301

It's easy verify the multiplication rule: $46/194 = \frac{46/301}{194/301}$

Independence

- Two events are considered *independent* when $P(A \text{ and } B) = P(A) * P(B)$

Independence

- Two events are considered *independent* when $P(A \text{ and } B) = P(A) * P(B)$
  - Notice that independence does not mean the events are *disjoint*
  - $P(A \text{ and } B) = 0$ for disjoint events

Independence

- Two events are considered *independent* when $P(A \text{ and } B) = P(A) * P(B)$
  - Notice that independence does not mean the events are *disjoint*
  - $P(A \text{ and } B) = 0$ for disjoint events
- Independence can greatly simplify a probability calculation
  - Consider 3 coin flips:

$$P(H_1 \text{ and } H_2 \text{ and } H_3) = P(H_1)*P(H_2)*P(H_3) = (1/2)^3 = 1/8$$

- This is a much easier calculation to think about compared to:

$$P(H_1 \text{ and } H_2 \text{ and } H_3) = P(H_3) * P(H_2|H_1) * P(H_3|H_1 \text{ and } H_2)$$

Independence and Sampling from a Population

- Independence is particularly useful when analyzing data that are a *simple random sample* from a population
- In general, if the sample data are less than 5% of the population statisticians will assume each selected case is independent other selections

Example #1 - part 1

A local hospital has 22 patients staying overnight, 15 are adults and 7 are children. Among the adults, this is the first ever hospital stay for 4 of them. Among the children, this is the first ever hospital stay for 5 of them. Use this information to calculate the following probabilities:

1) A randomly selected patient is an adult
2) A randomly selected patient is an adult, given it's their first ever hospital stay
3) A randomly selected patient is in their first ever hospital stay, given they are a child
4) A randomly selected patient is in their first ever hospital stay, *or* they are a child

Example #1 - part 1 (solution)

1) $P(\text{Adult}) = 15/22$, there are 22 patients and 15 are adults

Example #1 - part 1 (solution)

1) $P(\text{Adult}) = 15/22$, there are 22 patients and 15 are adults
2) $P(\text{Adult}|\text{First}) = 4/9$, there are 9 first-time patients and 4 are adults

Example #1 - part 1 (solution)

1) $P(\text{Adult}) = 15/22$, there are 22 patients and 15 are adults
2) $P(\text{Adult}|\text{First}) = 4/9$, there are 9 first-time patients and 4 are adults
3) $P(\text{First}|\text{Child}) = 5/7$, there are 7 children and 5 of them are first-time patients

Example #1 - part 1 (solution)

1) $P(\text{Adult}) = 15/22$, there are 22 patients and 15 are adults
2) $P(\text{Adult}|\text{First}) = 4/9$, there are 9 first-time patients and 4 are adults
3) $P(\text{First}|\text{Child}) = 5/7$, there are 7 children and 5 of them are first-time patients
4) $P(\text{First or Child}) = P(\text{First}) + P(\text{Child}) - P(\text{First and Child}) = 9/22 + 7/22 - 5/22 = 0.5$, notice we could have calculated this *directly* by realizing there are 7 children and 4 first-time adults (totaling 11 of 22 patients)

Example #1 - part 2

A local hospital has 22 patients, 15 are adults and 7 are children. Among the adults, this is the first ever hospital stay for 4 of them. Among the children, this is the first ever hospital stay for 5 of them. Now let's consider randomly selecting two patients sequentially:

1) What is the probability that *both* selections are adults?
2) What is the probability that *at least one* of the selections is an adult?

Example #1 - part 2 (solution)

1) Let A_1 and A_2 denote the selection of adults, then $P(A_1 \text{ and } A_2) = P(A_2|A_1) * P(A_1) = \frac{14}{21} * \frac{15}{22} = 0.45$; notice these events are not independent

Example #1 - part 2 (solution)

1) Let A_1 and A_2 denote the selection of adults, then $P(A_1 \text{ and } A_2) = P(A_2|A_1) * P(A_1) = \frac{14}{21} * \frac{15}{22} = 0.45$; notice these events are not independent
2) Using the additional rule could get complicated here because the events are not independent. Instead, let C_1 and C_2 denote the selection of children and consider $P(A_1 \text{ or } A_2) = 1 - P(\text{Neither}) = 1 - P(C_2|C_1) * P(C_1) = 1 - \frac{6}{21} * \frac{7}{22} = 1 - 0.09 = 0.91$

Example #2

Consider a well-shuffled deck of 52 playing cards and the random selection of two cards, a “top” card and a “bottom” card

1) The following line of reasoning is incorrect: “Because of the addition rule, the probability that the top card is the jack of clubs *and* the bottom card is the jack of hearts is $2/52$.” Point out the flaw in this argument.
2) The following line of reasoning is also incorrect: “Because of the addition rule, the probability that the top card is the jack of clubs *or* the bottom card is the jack of hearts is $2/52$.” Point out the flaw in this argument.
3) The statements in 1 and 2 both contain flaws, but these mistakes are not equally bad. Which approach will result in an answer closer to the truth (for the situation it describes)?

Example #2 (solution)

1) The addition rule pertains to intersections or “or” statements, so it shouldn’t be applied here

Example #2 (solution)

1) The addition rule pertains to intersections or "or" statements, so it shouldn't be applied here
2) The events involved are not disjoint, it is possible for the top card to be the jack of clubs and the bottom card to be the jack of hearts.

Example #2 (solution)

1) The addition rule pertains to intersections or "or" statements, so it shouldn't be applied here
2) The events involved are not disjoint, it is possible for the top card to be the jack of clubs and the bottom card to be the jack of hearts.
3) The second statement is much closer to the truth, because the possibility for both is very small ($\frac{1}{52} * \frac{1}{51}$ by the multiplication rule)

Conclusion

We've now covered three different probability rules:

1) The addition rule, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$, allows us to calculate the probability of *unions* of events
2) The multiplication rule, $P(A \text{ and } B) = P(A|B) * P(B)$, allows us to calculate the probability of *intersections* of events
3) The complement rule, $P(A) + P(A^C) = 1$, allows simpler calculations for large sample spaces