The Center of Applied Data Science

python™

Statistical Data Analysis

Content outline

1. Introduction to Probability Theory

Say we have a random sample that is representative of the population. However, each random sample is not exactly the same. One random sample may represent the population very accurately, while another random sample might not be representative. Therefore, by looking at any one sample, we will never know how much a sample resembles the population. We can use probability to describe the variation in random samples. Probability is a critical tool for **statistical inference** — drawing reliable conclusions about the population considering the uncertainty that is generated by taking a random sample of a population. The following example will illustrate this important point.

Example Suppose that we are interested in estimating the percentage of Malaysian adults aged 16–65 who are digitally literate. To do so, we choose a random sample of 1,200 Malaysian adults to take a digital problem-solving assessment. We find that 970 out of the 1,200, or 81%, are digitally literate.

Our goal here is to do inference — draw conclusions about the percentage of the entire population of Malaysian adults who are digitaly literate, based on data from only 1,200 of them. Can we conclude that 81% of the population are digitally literate?

If we examine another random sample, it may give us a very different result. So we cannot be sure about how well the sample describes the population. However, this uncertainty is due to randomness, not due to problems with how the sample was collected. Therefore, we can use probability to describe the likelihood that our sample is within a desired level of accuracy. For instance, using probability we can answer the question, "How likely is it that our sample estimate is no more than 3% from the true percentage of all digitally literate Malaysian adults?”

The answer to this question (which we find using probability) will have an important impact on how confident we are with our estimates. In particular, if we find it quite unlikely that the sample estimate will be very different from the population estimate, then we have a lot of confidence that we can draw conclusions about the population based on the sample.

2. Probability Definition

What is Probability?

Probability is a mathematical description of randomness and uncertainty. It is a way to measure or quantify uncertainty. Another way to think about probability is that it is the official name for "chance."

Probability deals with calculating the likelihood of a given event's occurrence, which is expressed as a number between 0 and 1. For instance, an event with a probability of 1 can be considered a certainty. A probability of 0 indicates no chance of that event occurring.

Probability is used to answer the following types of questions:

- What is the chance that school will be closed tomorrow?
- What is the chance that a stock will go up in price?
- What is the chance that I will have a heart attack in the next 5 years?
- What is the likelihood that when rolling a pair of dice, it will roll doubles?
- What is the probability that I will win the lottery?

Each of these examples has some uncertainty. The better the chance, the higher the probability.

Exercise Probability is a measure of how likely an event is to occur. Choose the probability that best matches each of the following statements: 1. This event is impossible: **0** 0.01 0.50 0.60 0.99 1 2. This event will occur frequently, but is not extremely likely: 0 0.01 0.50 **0.60** 0.99 1 3. This event is extremely unlikely, but it will occur once in a while in a long sequence of trials: 0 **0.01** 0.50 0.60 0.99 1 4. This event will occur for sure: 0 0.01 0.50 0.60 0.99 **1**

Probability Computation

P(desirable outcomes) = (number of desirable outcomes) / (total number of possible outcomes)

Example

A single 6-sided dice is rolled. What is the probability of rolling a 2?

$$P(2) = 1/6 = 0.17$$

Example

Suppose there are 4 freshmen, 2 sophomore, and 1 junior in a study group. You want to select one person. What is $P(F)$? $P(S)$? $P(J)$?

$$P(F) = 4/(4+2+1) = 4/7 = 0.57$$

$$P(S) = 2/(4+2+1) = 2/7 = 0.28$$

$$P(J) = 1/(4+2+1) = 1/7 = 0.14$$

Exercise

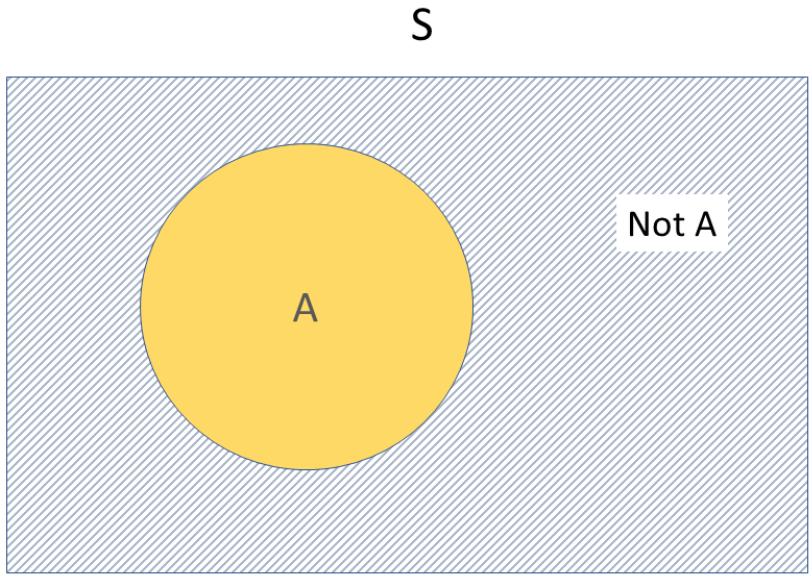
A group of 460 college students was surveyed over several typical weekdays. 253 students reported that they had eaten breakfast that day. Let B be the event of interest — that a college student eats breakfast. Based on this information, what is $P(B)$, the probability that a randomly chosen college student eats breakfast?

In [ ]:

```
```

3. Complement Rule

Complement Rule: $P(not\ A) = 1 - P(A)$



Example

A single 6-sided dice is rolled. What is the probability of not rolling a 2?

$$P(not\ 2) = 1 - P(2) = 1 - 1/6 = 0.83$$

Example

Suppose there are 4 freshmen, 2 sophomore, and 1 junior in a study group. You want to select one person. What is $P(not\ F)$? $P(not\ S)$? $P(not\ J)$?

$$P(not\ F) = 1 - 4/7 = 3/7 = 0.46$$

$$P(not\ S) = 1 - 2/7 = 5/7 = 0.72$$

$$P(not\ J) = 1 - 1/7 = 6/7 = 0.86$$

Exercise

Assume the likelihood a person has blood type O is 0.44, blood type A is 0.42, blood type B is 0.10, and blood type AB is 0.04 in the U.S.

The events O, A, B and AB are defined as follows:

- O: Person has blood type O
- A: Person has blood type A

  <p>

- B: Person has blood type B

  <p>

- AB: Person has blood type AB

Calculate $P(not\ O)$, $P(not\ A)$, $P(not\ B)$, and $P(not\ AB)$.

In [ ]:

4. The Additive and Multiplicative Rules

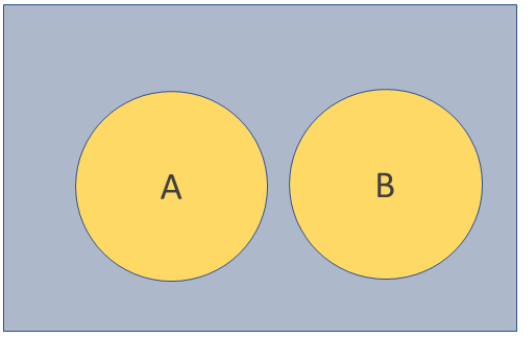
Additive Rules of Probability

Addition Rule 1

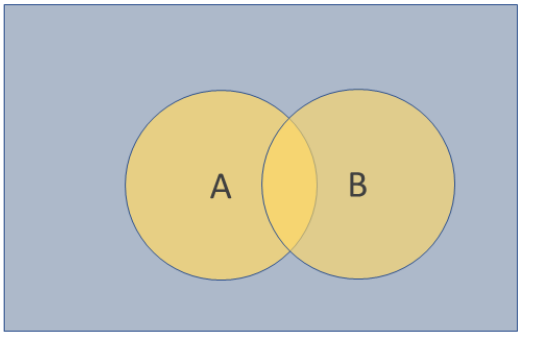
When two events, A and B are **mutually exclusive**, the probability that A or B will occur is the sum of the probability of events A and B.

Additive Rule for mutually exclusive events: $P(A \ or \ B) = P(A) + P(B)$

Mutually Exclusive



Non-Mutually Exclusive



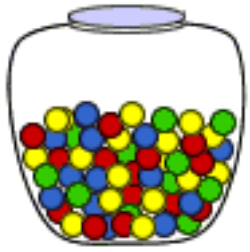
Example

A single 6-sided dice is rolled. What is the probability of rolling a 2 or a 5?

$$P(2) = 1/6$$
$$P(5) = 1/6$$
$$P(2 \ or \ 5) = P(2) + P(5) = 2/6 = 0.3$$

Example

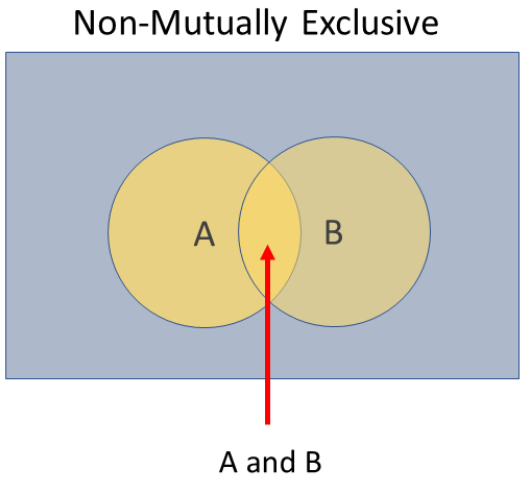
A glass jar contains 1 red, 3 green, 2 blue, and 4 yellow marbles. If a single marble is chosen at random from the jar, what is the probability that it is yellow or green?



$$P(yellow) = 4/10$$
$$P(green) = 3/10$$
$$P(yellow \ or \ green) = P(yellow) + P(green) = 7/10 = 0.7$$

Addition Rule 2

When two events, A and B, are **non-mutually exclusive**, the probability that A or B will occur is the sum of the probability of each event, minus the probability of the overlap.



Additive Rule for non-mutually exclusive events: $P(A \text{ or } B) \quad \text{and } B)$

$= P(A)$

$+ P(B)$

$- P(A$

Example

In a math class of 30 students, 17 are boys and 13 are girls. On a unit test, 4 boys and 5 girls made an A grade. If a student is chosen at random from the class, what is the probability of choosing a girl or an A student?

$$P(girl\ or\ grade_A) = P(girl) + P(grade_A) - P(girl\ and\ grade_A) = 13/30 + 9/30 - 5/30 = 17/30 = 0.57$$

Exercise

In a group of 101 students, 40 are juniors, 50 are female, and 22 are female juniors. Find the probability that a student picked from this group at random is either a junior or female.

In [ ]:

Exercise

Assume the chance a person has blood type O is 0.44, blood type A is 0.42, blood type B is 0.10, and blood type AB is 0.04 in the U.S. You are given the following additional information:

* A person with type A can donate blood to a person with type A or AB.

* A person with type B can donate blood to a person with type B or AB.

* A person with type AB can donate blood to a person with type AB only.

* A person with type O blood can donate to anyone.

Suppose that there are two patients who are each in need of a blood donation. Patient 1 has blood type A and patient 2 has blood type B. Consider the following events:

* D1: a randomly chosen person can be a donor for patient 1.

* D2: a randomly chosen person can be a donor for patient 2.

We are interested in finding the probability that a randomly chosen person can be a donor for patient 1 or patient 2. In other words, we are interested in finding P(D1 or D2).

In [ ]:

In [ ]:

Multiplication Rule for Independent Events

Independent events

Two events A and B are said to be independent if the fact that one event has occurred **does not affect** the probability that the other event will occur.

Multiplication Rule for Independent Events:

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

Example

A woman's pocket contains two quarters and two nickels. She randomly picks one of the coins and, after looking at it, puts it back in her pocket before picking a second coin. Let Q1 be the event that the first coin is a quarter and Q2 be the event that the second coin is a quarter. Are Q1 and Q2 independent events? What is P(Q1 and Q2)?

> Yes. Q1 and Q2 are independent events since she has **replaced** the first coin before picking the second coin. Therefore, the combination of the coins (two quarters and two nickels) is fixed in both cases and Q1 and Q2 are independent.
>
> $$P(Q1 \ and \ Q2) = 2/4 \times 2/4 = 4/16$$
> $$= 0.25$$

Example

A seminar class consists of 5 male students and 5 female students. Two of the 10 students are chosen at random for a role-playing exercise.

* A: the first chosen is male

* B: the second chosen is female

Are the two events independent or dependent?

> A and B are dependent since whether or not event A occurs (the first student chosen is male) has affect on the probability that event B will occur (the second student chosen is female).
> If A occurs, then P(second student chosen is female) = 5/9.
>
> If A does not occur (i.e., if the first student chosen is female), then P(second student chosen is female) = 4/9.
>
> Therefore, the two events are not independent, but rather dependent.

Example

Recall the blood type example from the previous exercise.

$P(O) = 0.44\,;\, P(A) = 0.42\,;\, P(B) = 0.10\,;\, P(AB) = 0.04$

Two people are selected simultaneously and at random from all people in the United States. What is the probability that both have blood type O?

Let O1= "person 1 has blood type O" and O2= "person 2 has blood type O"
We need to find $P(O1 \text{ and } O2)$

Since they were chosen simultaneously and at random, the blood type of one has no effect on the blood type of the other. Therefore, O1 and O2 are independent, and we may apply Rule 4:

$$\begin{aligned} P(O1 \text{ and } O2) &= P(O1). \\ &\times P(O2) \\ &= 0.44 \\ &\times 0.44 \\ &= 0.1936 \end{aligned}$$

Exercise

A 2011 poll by the Pew Research Center for People and the Press estimated that 62% of U.S. adults favor the death penalty for persons convicted of murder, 31% oppose it, with the remaining 7% undecided. What is the probability that two randomly chosen U.S. adults support the death penalty for persons convicted of murder?

* S: person supports death penalty
* O: person opposes death penalty
* U: person undecided
* S1: person 1 supports death penalty
* S2: person 2 supports death penalty

In [ ]:

Exercise

In the **blood type** example, two people are chosen simultaneously and at random. What is the probability that both have the same blood type?

In [ ]:

5. Conditional Probability

The conditional probability of event A given event B is:

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

Example

It is vital that a certain document reaches its destination within one day. To maximize the chance of the delivery arriving on time, two copies of the document are sent using two services, service A and service B, and the following probability table summarizes the chances of on-time delivery:
| |B|not B|Total| |------|-|-----|-----| |A|0.75|0.15|0.90| |not A|0.05|0.05|0.10| |Total|0.80|0.20|1.00| If the document has reached its destination on time through service A, what is the probability that it will also reach its destination through service B? > We are told that the document has arrived on time using service A, and we want to know the probability is that it will also arrive on time using service B. We are therefore looking for $P(B|A)$.
Using the definition of conditional probability and the probability table, we get that:

$$> P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$
$$= 0.75/0.9 = 0.833$$

Example

If service A has failed to deliver the document on time, what is the probability that it arrives on time using service B? As before, first write down the conditional probability that you are asked to find, and then apply the definition of conditional probability to find it.

> We are told that the document was not delivered on time using service A (not A), and we want to know how likely is it that it was delivered on time using service B. We are therefore looking for $P(B|not\ A)$.
Using the definition of conditional probability and the probability table, we get that:

$$P(B|not\ A) = P(not\ A\ and\ B)/P(not\ A) = 0.05/0.1 = 0.50$$

Exercise

In the above delivery example, if service A delivers the document on time, what is the probability that it is not delivered on time using service B?

In [ ]:

6. Multiplicative Rule

By rearranging the definition of conditional probability, we obtain the following:

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

Example

Suppose you pick two cards at random from four cards consisting of one of each suit: club, diamond, heart, and spade, where the first card is replaced before the second card is picked. What is the probability of

picking a club and then a diamond? Given the following events:

* C: the event picking a club

* D: the event picking a diamond > Since the second card is picked after replacement, C and D are independent. Therefore, $P(C \text{ and } D) = P(C) \times P(D)$.

$P(C) = 1/4 = 0.25$
$P(D) = 1/4 = 0.25$
$P(C \text{ and } D) = 0.25 \times 0.25$
$= 0.06$

Example

Suppose you pick two cards at random from four cards consisting of one of each suit: club, diamond, heart, and spade. What is the probability of picking a club and then a diamond without replacement?

> Since we have not replaced the first card before picking the second card, the probability of picking a diamond as the second card has been affected by the type of the first card. For example, if the first card is a diamond, the probability of the second card as a diamond is 0. Therefore, in this problem C and D are dependent. So, $P(C \text{ and } D) = P(D|C) \times P(C)$

$P(C) = 1/4 = 0.25$
$P(D|C) =$ Probability of diamond given the first card is club
$= 1/3 = 0.33$
$P(C \text{ and } D) = 0.25 \times 0.33$
$= 0.08$

Exercise

A woman's pocket contains 2 quarters and 2 nickels; she randomly extracts one of the coins, and without replacing it picks a second coin. What is the probability of getting a quarter both times, $P(Q1 \text{ and } Q2)$? What is the probability of getting a quarter and then a nickel, $P(Q1 \text{ and } N2)$?

In [ ]:

Exercise

In a certain region, one in every thousand people (0.001) of all individuals are infected by the HIV virus that causes AIDS. Tests for presence of the virus are fairly accurate but not perfect. If someone actually has HIV, the probability of testing positive is 0.95. Let H denote the event of having HIV, and T the event of testing positive for HIV. (a) Express the information that is given in the problem in terms of the events H and T.
(b) Use the General Multiplication Rule to find the probability that someone chosen at random from the population has HIV and tests positive.
(c) If someone has HIV, what is the probability of testing negative? Here we need to find $P(not\ T|H)$.

In [ ]:

In [ ]: