Unit 2: Probability

Statistics 102 Teaching Team

February 10, 2020

Basic Concepts of Probability

Intro to Probability

People often colloquially refer to probability...

- "What are the chances the Red Sox will win this weekend?"
- "What's the chance of rain tomorrow?"
- "What is the chance that a patient responds to a new therapy?"

Formalizing concepts and terminology around probability is essential for better understanding probability.

Random Experiments

A *random experiment* is an action or process that leads to one of several possible outcomes.

- For example, flipping a coin leads to two possible outcomes: either heads or tails.

The *probability* of an outcome is the proportion of times the outcome would occur if the random phenomenon could be observed an infinite number of times.

- If a fair coin is flipped an infinite number of times, heads would be obtained 50% of the time.

Outcomes and events

An *outcome* in a study or experiment is the observable result after conducting the experiment.

- The sum of the faces on two dice that have been rolled.
- The response of a patient treated with an experimental therapy.
- The total volume of eggs in a clutch laid by a frog.

An *event* is a collection of outcomes.

- The sum after rolling two dice is 7.
- 22 of 30 patients in a study have a good response to a therapy.
- The total volume of eggs in a clutch is larger than 750 mm^3.

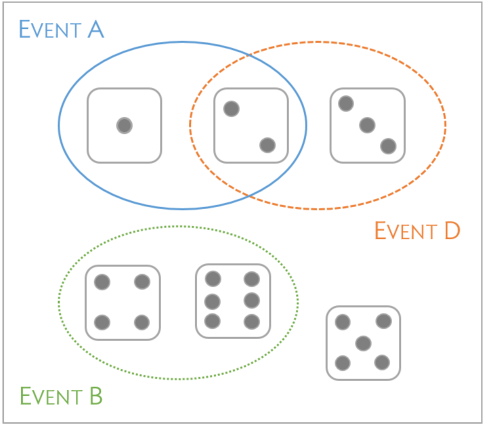
Events can be referred to by letters.

- Suppose event A is the event of rolling a number smaller than 3 on a die.

$$A = \{1, 2\}$$

Disjoint / Mutually Exclusive Events

Two events or outcomes are called *disjoint* or *mutually exclusive* if they cannot both happen at the same time.



Addition Rule for Disjoint Events

If A and B represent two disjoint events, then the probability that either occurs is

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

The $\cup$ symbol denotes the *union* of two events; i.e., $P(A \text{ or } B)$.[1]

As shown on the previous slide, events A and B are disjoint.

- $P(A \cup B) = P(A) + P(B) = 2/6 + 2/6 = 4/6$
- Intuitively, this makes sense; the probability of rolling either a 1, 2, 4, or 6 on a six-sided die is $4/6$.

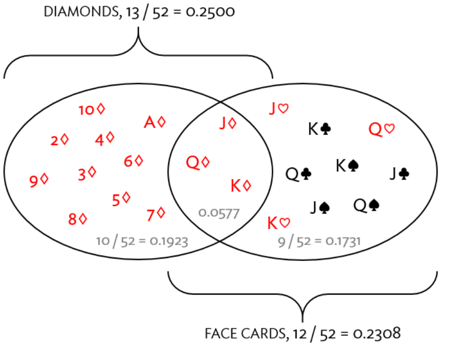
If there are k disjoint events $A_1, \ldots, A_k$, then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k)$$

[1] Statistics uses the inclusive "or", such that A or B occurring means A, B, or both A and B occur.

General Addition Rule

Suppose that we are interested in the probability of drawing a diamond or a face card out of a standard 52-card deck.



Does $P(\text{diamond or face card}) = 13/52 + 12/52$?

General Addition Rule...

To correct the double counting of the three cards that are in both events, subtract the probability that both events occur...

$$\begin{aligned} P(\text{diamond or face card}) =& P(\text{diamond}) + P(\text{face card}) - P(\text{diamond and face card}) \\ =& 13/52 + 12/52 - 3/52 \\ =& 22/52 \end{aligned}$$

Thus, for any two events A and B, the probability that either occurs is

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

The $\cap$ symbol denotes the *intersection* of two events; i.e., $P(A \text{ and } B)$.

Sample Space

A *sample space* is a list of exhaustive and mutually exclusive outcomes.

Suppose the possible k outcomes are denoted $O_1, O_2, \ldots, O_k$. The sample space can be expressed as $S = \{O_1, O_2, \ldots, O_k\}$.

- For the die tossing experiment, $S = \{1, 2, 3, 4, 5, 6\}$

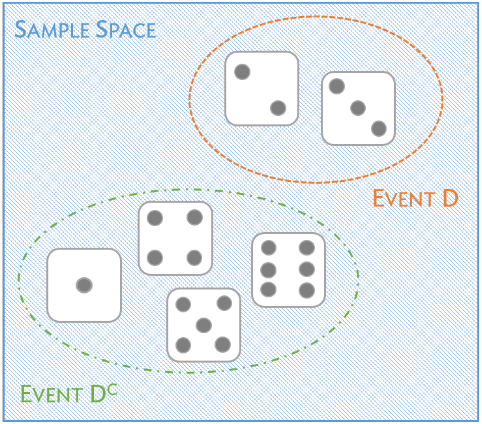
Given a sample space $S = \{O_1, O_2, \ldots, O_k\}$, the sum of the probabilities of each outcome must equal 1.

$$\sum_{i=1}^{k} P(O_i) = 1$$

Complement of an Event

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3.

The *complement* of D represents all outcomes in the sample space that are not in D.



Complement of an Event...

The complement of an event A is denoted by A^C.

An event and its complement are mathematically related:

$$P(A) + P(A^C) = 1 \qquad P(A) = 1 - P(A^C)$$

Independent Events

Two events A and B are *independent* if the probability that both A and B occur equal the product of their separate probabilities.

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

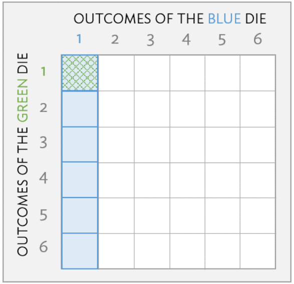
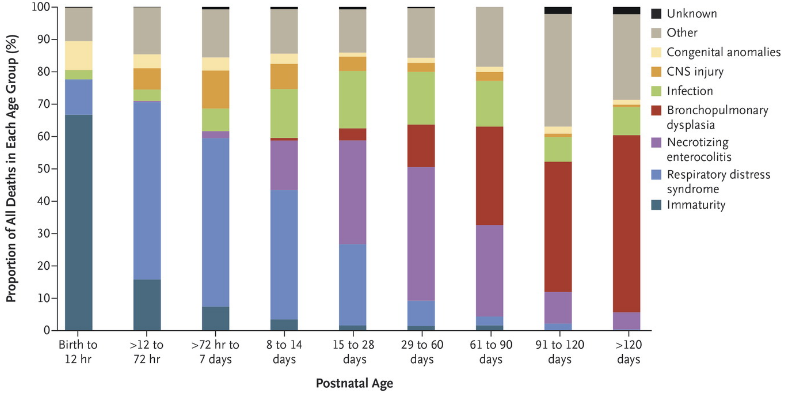


Figure 1: A blue die and a green die are rolled. What is the probability of rolling two 1's?

Conditional Probability

An example from childhood mortality



No. of Patients	2453	854	417	603	696	509	184	92	265
Proportion of All Deaths (%)	40.4	14.1	6.9	9.9	11.5	8.4	3.0	1.5	4.4

Published in Patel, et al., *NEJM* (2015) Vol 372, pp 331 - 340.

Conditional Probability: Intuition

Consider height in the US population.

What is the probability that a randomly selected individual in the population is taller than 6 feet, 4 inches?

- Suppose you learn that the individual is a professional basketball player.
- Does this change the probability that the individual is taller than 6 feet, 4 inches?

Conditional Probability: Concept

The **conditional probability** of an event A, given a second event B, is the probability of A happening, knowing that the event B has happened.

- This conditional probability is denoted $P(A|B)$.

Toss a fair coin three times. Let A be the event that *exactly* two heads occur, and B the event that *at least* two heads occur.

- $P(A|B)$ is the probability of having exactly two heads among the outcomes that have at least two heads.
- Conditioning on B means that the sample space consists of $\{HHH, HHT, HTH, THH\}$, all possible sets of three tosses where at least two heads occurred.
- In this set of outcomes, A, consists of the last three, so $P(A|B) = 3/4$.

Conditional Probability: Formal Definition

As long as $P(B) > 0$,

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

From the definition,

$$\begin{aligned}
P(A|B) =& \frac{P(\text{at least two heads and exactly two heads})}{P(\text{at least two heads})} \\
=& \frac{P(\text{exactly two heads})}{P(\text{at least two heads})} \\
=& \frac{(3/8)}{(4/8)} = 3/4
\end{aligned}$$

Independence, Again...

A consequence of the definition of conditional probability:

- If $P(A|B) = P(A)$, then A and B are independent; knowing B offers no information about whether A occurred.

General Multiplication Rule

If A and B represent two outcomes or events, then

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

This follows from rearranging the definition of conditional probability:

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

Unlike the previously mentioned multiplication rule, this is valid for events that might not be independent.

Positive Predictive Value of a Diagnostic Test (Bayes' Theorem)

Pre-natal testing for trisomy 21, 13, and 18

Some congenital disorders are caused by an additional copy of a chromosome being attached (translocated) to another chromosome during reproduction.

- Trisomy 21: Down syndrome, approximately 1 in 800 births
- Trisomy 13: Patau's syndrome, physical and mental disabilities, approximately 1 in 16,000 newborns
- Trisomy 18: Edward's syndrome, nearly always fatal, either in stillbirth or infant mortality. Occurs in about 1 in 6,000 births

Until recently, testing for these conditions consisted of screening the mother's blood for serum markers, followed by amniocentesis in women who test positive.

Cell-free fetal DNA (cfDNA)

cfDNA consists of copies of embryo DNA present in maternal blood.

Advances in sequencing DNA provided possibility of non-invasive testing for these disorders, using only a blood sample.

Initial testing of the technology was done using archived samples of genetic material from children whose trisomy status was known.

The results are variable, but generally very good:

- Of 1000 unborn children with the one of the disorders, approximately 980 have cfDNA that tests positive. The test has high *sensitivity*.
- Of 1000 unborn children without the disorders, approximately 995 test negative. The test has high *specificity*.

Cell-free fetal DNA (cfDNA)...

The designers of a diagnostic test want the test to be accurate.

- In other words, the test should have high sensitivity and specificity.

A family with an unborn child undergoing testing, however, wants to know the likelihood of the condition being present if the test is positive.

Suppose a child has tested positive for trisomy 21. What is the probability that the child does have the trisomy 21 condition, given the positive test result?

Defining events in diagnostic testing

Events of interest in diagnostic testing:

- $D = \{\text{disease present}\}$
- $D^C = \{\text{disease absent}\}$
- $T^+ = \{\text{positive test result}\}$
- $T^- = \{\text{negative test result}\}$

Could use T and T^C, but T^+ and T^- are consistent with notation in medical and public health literature.

Characteristics of a diagnostic test

The following measures are all characteristics of a diagnostic test.

- Sensitivity $= P(T^+|D)$
- Specificity $= P(T^-|D^C)$
- False negative rate $= P(T^-|D)$
    - Note that $P(T^-|D) = 1 - P(T^+|D)$, i.e., 1 - sensitivity
- False positive rate $= P(T^+|D^C)$
    - Note that $P(T^+|D^C) = 1 - P(T^-|D^C)$, i.e., 1 - specificity

Positive predictive value of a test

Suppose an individual tests positive for a disease.

The **positive predictive value (PPV)** of a diagnostic test is the probability that a person has the disease, given that he/she has tested positive for it.

- PPV $= P(D|T^+)$

The characteristics of a diagnostic test include $P(T^+|D)$, among other probabilities, but not the reverse conditional probability $P(D|T^+)$.

Bayes' Theorem, aka Bayes' Rule

Bayes' Thorem (simplest form):

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

Follows directly from the definition of conditional probability by noting that $P(A)P(B|A)$ equals $P(A \text{ and } B)$:

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

The denominator P(B) in Bayes' Theorem

Bayes' Theorem is seldom stated in its simplest form, because in many problems, $P(B)$ is not given directly, but is calculated using the general multiplication formula for probabilities:

Suppose A and B are events. Then,

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap A^C) \\ &= P(B|A)P(A) + P(B|A^C)P(A^C) \end{aligned}$$

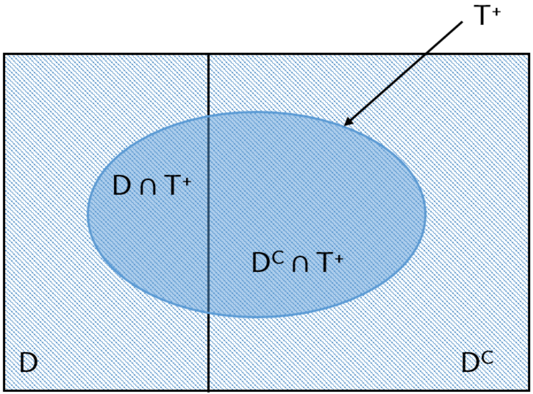
Bayes' Theorem can be written as:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

Bayes' Theorem for diagnostic tests

$$
\begin{aligned}
P(D|T^+) &= \frac{P(D \cap T^+)}{P(T^+)} \\
&= \frac{P(D \cap T^+)}{P(D \cap T^+) + P(D^C \cap T^+)} \\
&= \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|D^C)P(D^C)} \\
&= \frac{\text{sensitivity} \times \text{prevalence}}{[\text{sensitivity} \times \text{prevalence}] + [(1 \text{ - specificity}) \times (1 \text{ - prevalence})]}
\end{aligned}
$$

Bayes' Theorem for diagnostic tests...



$$P(D|T^+) = \frac{P(D \cap T^+)}{P(T^+)} = \frac{P(D \cap T^+)}{P(D \cap T^+) + P(D^C \cap T^+)}$$