Statistics

Probability concepts

Probabilities: models of uncertainty

Random events

Random event: event with an uncertain outcome

Probability model for a random event

- **Sample space** S:
  description of all possible outcomes of the event
- Probability set:
  a value assigned to each simple element in S

Sample space

Discrete sample space a set of distinct outcomes (numbers or descriptions)

- Blood types
  For a random person:
  S = {O+, O-, A+, A-, B+, B-, AB+, AB-}

Continuous sample space a continuum of outcomes on a scaled interval

- Cholesterol level, in mg/dl
  For a random person:
  S = any reasonable positive value (or the interval zero to infinity)

Obtaining a model

Probability models may be based on:

- Properties of the event studied
- Observed frequencies
- Subjective / personal evaluation

MODEL
MUST BE
ADEQUATE!

Mendelian genetic theory predicts 50% males (XY) and 50% females (XX) among newborns. Birth certificates show that, in the U.S., there are actually more male (~51.2%) than female (~48.8%) live births every year.

Probability notation

P(some specified outcome) = a value between 0 and 1

The probability (chance) of the specified outcome is "this much." There is this much probability (chance) that the specified outcome will happen or is true.

Example

Probability model for the sex of a newborn in the United States:

S = {Male, Female}

P(Male) = 0.512; P(Female) = 0.488

"In the United States, the probability that a randomly selected newborn is male is 0.512 (51.2%). There is 0.488 (48.8%) probability that a randomly selected newborn is female."

Probability concepts

Basic rules and definitions

Probability rules

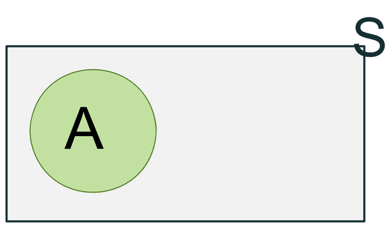
P(sample space) = 1

For any event A:

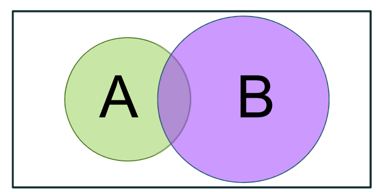
$0 \leq P(A) \leq 1$

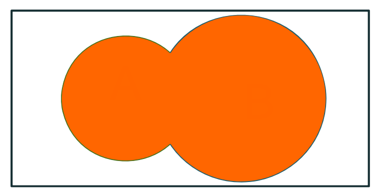
P(not A) = 1 – P(A)

"not A" is the **complement** of "A"

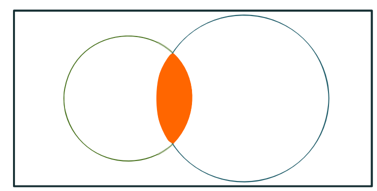


A and B vs A or B





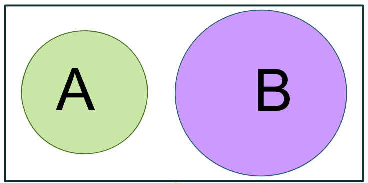
Union: A <u>or</u> B
(at least one occurs)



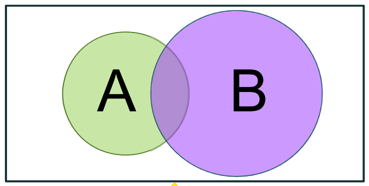
Intersection: A <u>and</u> B
(both occur together)

Disjoint events

$\underbrace{P(\text{A and B})}_{\textit{joint probability}} = 0 \quad \leftrightarrow \quad$ A, B are **disjoint** (mutually exclusive)



These 2 events are disjoint: their joint probability *P*(A and B) is zero.



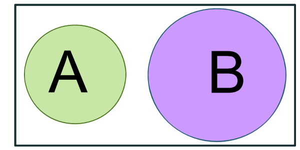
These 2 events are NOT disjoint: they do occur together sometimes.

Addition rule

Addition rule for disjoint events:

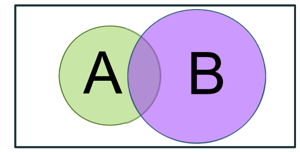
When two events A and B are disjoint:

P(A or B) = *P*(A) + *P*(B)



General addition rule for <u>any</u> two events:

***P*(A or B) = *P*(A) + *P*(B) – *P*(A and B)**



Example

In a large city, data on public swimming pools indicate that 15% of pools have inadequate levels of chlorine, 25% have dirty filters, and 70% have neither.

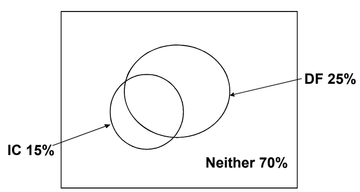
1. Translate into probability notation:

$P(IC) = 0.15$

$P(DF) = 0.25$

$P(\text{not IC and not DF}) = 0.70$

2. Organize the information:



	DF	not DF	Total
IC			15%
not IC		70%	
Total	25%		100%

two-way table

Example

In a large city, data on public swimming pools indicate that 15% of pools have inadequate levels of chlorine, 25% have dirty filters, and 70% have neither.

What is *P*(IC or DF)?

$P(IC) = 0.15$

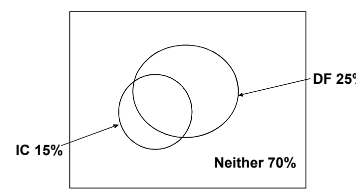
$P(DF) = 0.25$

$P(\text{not IC and not DF}) = 0.70$

$P(\text{IC or DF}) = 1 - P(\text{neither})$
$= 1 - 0.70 = 0.30$

$P(\text{IC or DF}) = P(IC) + P(DF) - P(\text{IC and DF})$
$= 0.15 + 0.25 - 0.10 = 0.30$

	DF	not DF	Total
IC	10%	5%	15%
not IC	15%	70%	85%
Total	25%	75%	100%



Statistics

image: NASA

Probability concepts

The concept of independence

Example

Consider a typical six-sided die.

Your roll the die. What's the chance you get a "5"? 1/6 ≈ 0.17

If you did get a "5" the first time, what is the chance that you will again get a "5" on our second roll?

1/6 (the previous roll doesn't matter)

Consider a box of 4 hazelnut chocolates and 4 cream chocolates.

You blindly pick one chocolate. What's the chance it is a hazelnut one? 4/8 = 0.5

If you did pick and eat a hazelnut chocolate first, what is the chance that your second chocolate is again a hazelnut one?

3/7 ≈ 0.43 (the missing first chocolate makes a difference)

Example

Consider a typical six-sided die.

Your roll the die. What's the chance you get a "5"? 1/6

If you did get a "5" the first time, what is the chance that you will again get a "5" on our second roll?

1/6 (the previous roll doesn't matter)

A die roll does not depend on the outcome of the previous die roll.

The successive rolls are independent.

Example

The probability of a chocolate selection depends on what happened in the previous selection (because, by eating the chocolates picked, we change the composition of the box). The successive picks are dependent.

Consider a box of 4 hazelnut chocolates and 4 cream chocolates.

You blindly pick one chocolate. What's the chance it is a hazelnut one? 4/8 = 0.5

If you did pick and eat a hazelnut chocolate first, what is the chance that your second chocolate is again a hazelnut one?

3/7 ≈ 0.43 (the missing first chocolate makes a difference)

Independence

Two events are independent if the knowledge that one event is true or has happened does not affect the probability of the other event.

"male" and "getting head on a coin flip" → independent (physics law)

"male" and "pregnant" → not independent (biology law)

"male" and "taller than 6 ft" → not independent (we know from experience/data)

"male" and "high cholesterol" → it's not obvious (we would need to collect data to find out)

When two events are independent, no information is gained from the knowledge of the other event.

This is completely different from the idea of events that are disjoint.

Conditional probabilities

Conditional probabilities reflect how the probability of an event can be different if we know that some other event has occurred or is true.

Notation:

***P*("a specific outcome" | "some relevant information")**

***P*(B | A)** is the **conditional probability of B given A**. That is, the probability that event B would happen, when we have the extra knowledge that event A is true or has happened.

Example

Consider a typical six-sided die.

Your roll the die. What's the chance you get a "5"? P(rolling a 5) = 1/6

If you did get a "5" the first time, what is the chance that you will again get a "5" on our second roll?

P(rolling a 5 second | first roll was a 5) = P(rolling a 5) = 1/6

The successive rolls are independent.

Consider a box of 4 hazelnut chocolates and 4 cream chocolates.

You blindly pick one chocolate. What's the chance it is a hazelnut one? P(hazelnut) = 4/8

If you did pick and eat a hazelnut chocolate first, what is the chance that your second chocolate is again a hazelnut one?

P(hazelnut second | first eaten was hazelnut) = 3/7 ≠ P(hazelnut)

The successive picks are not independent.

Probability concepts

Conditional probabilities

Conditional probability

P(B | A) is the **conditional probability of B given A**: the probability that event B would happen, when we have the extra knowledge that event A is true or has happened.

Conditional probability notation:

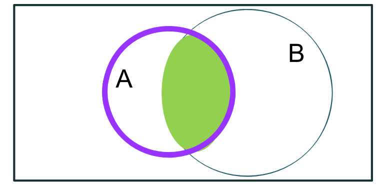
P("a specific outcome" | "some relevant information")

Computation

P(B | A), the conditional probability of event B, given the knowledge of event A, can be computed from other probabilities representing "out of all A outcomes, how often does B also occur":

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

[provided that P(A) ≠ 0]



Example

In a large city, data on public swimming pools indicate that 15% of pools have inadequate levels of chlorine, 25% have dirty filters, and 70% have neither.

$P(\text{IC}) = 0.15$
$P(\text{DF}) = 0.25$
$P(\text{not IC and not DF}) = 0.70$

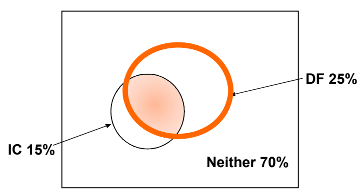
If you find out that a public swimming pool in this city has dirty filters, then what is the chance that it has inadequate levels of chlorine too?

A) 0.10 B) 0.15
C) 0.40 D) 0.50

$P(\text{IC} \mid \text{DF}) = P(\text{IC and DF}) / P(\text{DF})$
$= 0.10 / 0.25 = 0.40$

40% of pools that have dirty filters also have inadequate chlorine levels.

	DF	not DF	Total
IC	10%	5%	15%
not IC	15%	70%	85%
Total	25%	75%	100%



Independence

Two events are independent if the knowledge that one event is true or has happened does not affect the probability of the other event.

P(B | A) = P(B | not A) = P(B) ⇔ A, B are independent

When two events are independent, no information is gained from the knowledge of the other event.

This is completely different from the idea of events that are disjoint.

Example In a large city, data on public swimming pools indicate that 15% of pools have inadequate levels of chlorine, 25% have dirty filters, and 70% have neither.

$P(IC) = 0.15$
$P(DF) = 0.25$
$P(\text{not } IC \text{ and not } DF) = 0.70$

What can we say about events DF and IC?

A) DF, IC are independent.

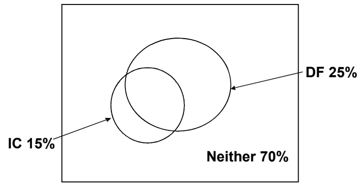
B) DF, IC are not independent.

C) DF, IC may or may not be independent.

$P(IC \mid DF) \neq P(IC)$
$0.40 \neq 0.15$

The probability that a pool has inadequate chlorine levels depends on whether or not we know that it has dirty filters.

	DF	not DF	Total
IC	10%	5%	15%
not IC	15%	70%	85%
Total	25%	75%	100%



Multiplication rule

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

General multiplication rule:

The probability that any two events, A and B, both occur at the same time can be computed as:

***P*(A and B) = *P*(A)*P*(B|A)**

Multiplication rule for independent events:

If A and B are independent, the equation simplifies to:

P(A and B) = *P*(A)*P*(B)

Example In a large city, data on public swimming pools indicate that 15% of pools have inadequate levels of chlorine, 25% have dirty filters, and 70% have neither.

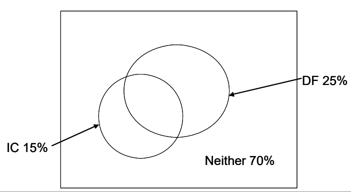
$P(IC) = 0.15$
$P(DF) = 0.25$
$P(\text{not } IC \text{ and not } DF) = 0.70$

Another way to find out if two events are independents:

$P(IC) * P(DF) = 0.15 * 0.25 = 0.0375 = 3.75\%$
$< P(IC \text{ and } DF) = 10\%$

→ IC and DF are not independent

	DF	not DF	Total
IC	10%	5%	15%
not IC	15%	70%	85%
Total	25%	75%	100%



IC and DF are more likely to occur together than we would expect if they occurred independently.

Confusion of the inverse

Confusing *P*(B|A) with *P*(A|B) is a common mistake called the "**confusion of the inverse**." It's a lot less likely to happen if you translate the probability notation into a full sentence.

***P*(>6ft | man)**

***P*(man | >6ft)**