Data Science Math Skills
Paul Bendich and Daniel Egger
Duke University

Basic Probability Definitions: Probability Definitions and Notation

Video companion

1 Introduction

Definition
probability—the degree of belief in the truth or falsity of a statement

Range of uncertainty from 0 to 1
Certain statement is true: probability 1
Certain statement is false: probability 0

Example Statement x: "It is raining."

2 Notation

$P(x)$	probability of x
$\sim x$	negation of statement x

Law of excluded middle

$$P(x) + P(\sim x) = 1$$

Probability of a statement and the probability of the negation of a statement must sum to 1.

If $P(x) = 1$, then $P(\sim x) = 0$, and vice versa.

In general, all outcomes of a probability distribution must sum to 1.

Definitions
probability distribution—collection of statements that are *exclusive* and *exhaustive*
exclusive—given complete information, no more than one of the statements can be true

exhaustive—given complete information, at least one of the statements must be true

A distribution X consisting of n statements would be denoted

$$X = \{x_1, x_2, x_3, ..., x_n\}.$$

The probability of each statement must sum to 1, which is denoted.

$$P(x_1) + P(x_2) + P(x_3) + ... + P(x_n) = 1.$$

3 Principle of indifference

For the i-th outcome x_i in a distribution with n possible outcomes,

$$P(x_i) = \frac{1}{n}.$$

Example: Drawing an ace of spades from a well-shuffled deck of 52 cards. The probability of drawing the ace of spades is $\frac{1}{52}$.

General statement
When there is no basis to choose some outcomes as more likely than others,

$$P(\text{event}) = \frac{\text{number of outcomes as defined in event}}{\text{total number of possible outcomes in universe}}.$$

Example: Event is drawing a queen, which has four outcomes in the event. The total number of outcomes is 52, so the probability of drawing a queen is $\frac{4}{52} = \frac{1}{13}$.

Example: Event is rolling an even number on a six-sided die, which has three outcomes in the event. The total number of outcomes is 6, so the probability of rolling an even is $\frac{3}{6} = \frac{1}{2}$.

Basic Probability Definitions: Joint Probabilities

Video companion

1 Introduction

Definition
joint probability—probability that two separate events with separate probability distributions are both true

$P(A \text{ and } B)$ is written $P(A,B)$, and read "the joint probability of A and B" or "the probability that A is true and B is true."

2 Order of joint probabilities

For probability distributions X and Y:

$$X = \{x_1, x_2, x_3, ..., x_n\}$$
$$Y = \{y_1, y_2, y_3, ..., y_n\}$$

Ordering does not matter in joint probabilities, for either the probability distributions or the individual events.

$$P(X,Y) = P(Y,X)$$
$$P(x_1,y_1) = P(y_1,x_1)$$

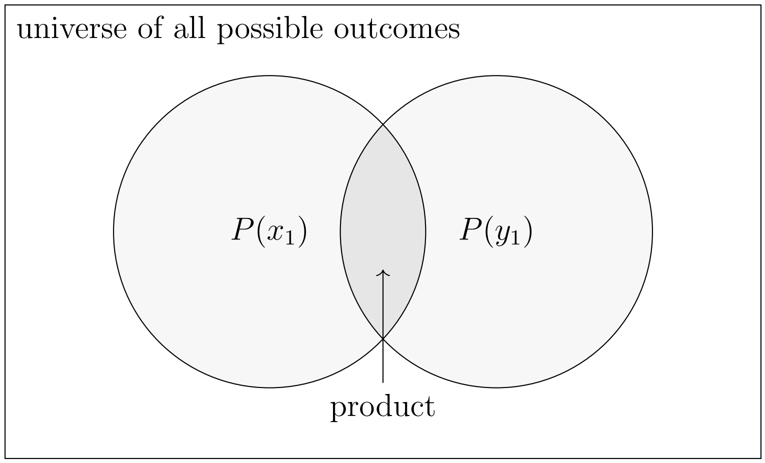
3 Independence

Definition
independence—knowing the outcome of one event does not change the probability of the other

The probability of two independent events:

$$\underbrace{P(x_1,y_1)}_{\text{"joint distribution"}} = \underbrace{P(x_1)P(y_1)}_{\text{"product distribution"}}$$

Venn diagrams—show the intersection and union of sets

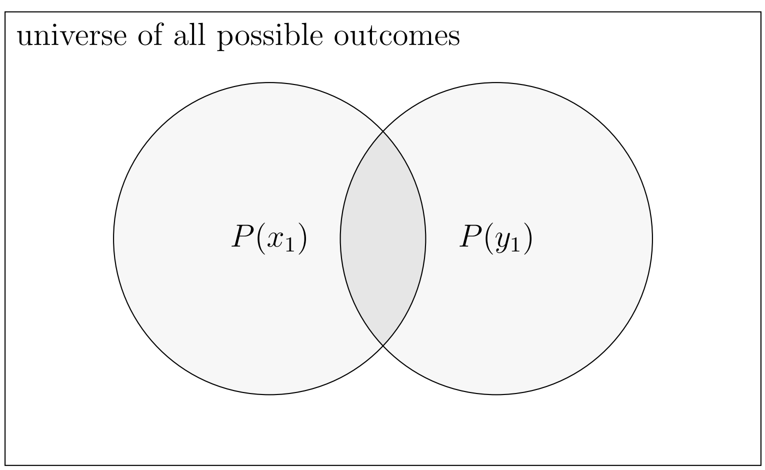


4 OR probability

Probability that either of two events occurs:

$$P(x_1 \text{ or } y_1) = P(x_1) + P(y_1) - P(x_1, y_1)$$

Venn diagram:



Data Science Math Skills
Paul Bendich and Daniel Egger
Duke University

Problem Solving Methods: Permutations and Combinations

Video companion

1 Introduction

Topic: Probability of events occurring in an order or the probability of a group of events occurring

Definitions # of permutation = n! / (n-m)! , n = # of object, m = # of unique attributes

permutation—order matters, e.g. placing five people in five different positions: 120 ways
combination—order does not matter, e.g. forming a five-person team from five people: 1 way

# of combination = n! / (n-m)! m! > n choose m formula

2 Replacement

Sampling *with replacement* (independent), e.g. drawing a card and putting it back in the deck

Sampling *without replacement*, e.g. drawing a card from a deck and not putting it back

With the options permutation, combination, with replacement, and without replacement, we have most of the probability situations that are likely to arise in a basic probability course.

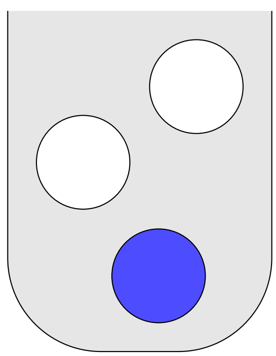
Problem Solving Methods: Using Factorial and "M Choose N"

Video companion

1 Introduction

Urn—a container you cannot see into

Example Drawing a marble from an urn containing two white and one blue marble. Can draw with or without replacement. Drawing with replacement means events are independent.



With replacement:

Draw	Probability
1 white	2/3
1 blue	1/3
2 white (in a row)	(2/3)(2/3) = 4/9

Without replacement:

Draw	Probability
1 white	2/3
1 blue	1/3
2 white (in a row)	(2/3)(1/2) = 1/3

2 Factorial

A factorial is the operation where we take a number and multiply it by each integer that is 1 less until we get down to 1.

Notation: 5! is read "five factorial." The operation means:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

Factorial quotients:

$$\frac{7!}{5!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 7 \cdot 6 = 42$$

Convention:

$$0! = 1$$

3 "m choose n"

Draw n items from a group of m items without replacement.

Example: How many unique committees of five people from a group of ten people? In this example, "10 choose 5," $m = 10$ and $n = 5$. The notation is given by:

$$\begin{aligned}\binom{10}{5} &= \frac{10!}{5! \cdot 5!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 2 \cdot 3 \cdot 2 \cdot 7 \cdot 3 = 252\end{aligned}$$

General formula

$$\binom{m}{n} = \frac{m!}{(m-n)! \cdot n!}$$

Data Science Math Skills
Paul Bendich and Daniel Egger
Duke University

Problem Solving Methods: The Sum Rule, Conditional Probability, and the Product Rule

Video companion

1 Marginal probabilities and the sum rule

Often know the joint probabilities, but don't know individual probabilities.

Table of known joint probabilities:

(X,Y)		X		
		x_1	x_2	x_3
	y_1	$P(x_1,y_1)$ 0.01	$P(x_2,y_1)$ 0.02	$P(x_3,y_1)$ 0.03
Y	y_2	$P(x_1,y_2)$ 0.10	$P(x_2,y_2)$ 0.20	$P(x_3,y_2)$ 0.49
	y_3	$P(x_1,y_3)$ 0.04	$P(x_2,y_3)$ 0.05	$P(x_3,y_3)$ 0.06

Can refer to $P(x_1)$ as the "marginal probability of x_1" because it is in the margins of the matrix.

Sum rule: The marginal probability is equal to the sum of the joint probabilities.

For x_1, this means:

$$\begin{aligned} P(x_1) &= P(x_1,y_1) + P(x_1,y_2) + P(x_1,y_3) \\ &= 0.01 + 0.10 + 0.04 = 0.15 \end{aligned}$$

and for y_2:

$$\begin{aligned} P(y_2) &= P(x_1,y_2) + P(x_2,y_2) + P(x_3,y_2) \\ &= 0.10 + 0.20 + 0.49 = 0.79 \end{aligned}$$

Sum rule for binary probability distribution:

$$P(A) = P(A, B) + P(A, {\sim}B)$$

Sum rule for series of n probabilities:

$$P(A) = P(A, B_1) + P(A, B_2) + ... + P(A, B_n)$$

2 Conditional probability

Definition:
conditional probability—the probability that a statement is true given that some other statement is true with certainty.

Symbol $P(A \mid B)$ means the "probability of A given that B is true with certainty."

Example: What is the probability of rolling a 3 on a 6-sided die, given that the roll is odd? Of the outcomes when the roll is odd (3), one is the relevant outcome, so the probability is 1/3.

Example: What is the probability of rolling an odd, given that the roll is a 3? Of the outcomes when the roll is a 3 (1), one is the relevant outcome, so the probability is 1.

Formula for conditional probability:

$$P(A \mid B) = \frac{\text{(relevant outcomes)}}{\text{(total outcomes remaining in universe, when } B \text{ is true)}}$$

3 Product rule

Want to relate concepts of joint probability, marginal probability, and conditional probability.

Product rule:

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

Conditional probability of A given that B is true is equal to the joint probability that A and B are true, divided by the probability that B is true.

Old definition of independence:

$$P(A, B) = P(A)P(B)$$

Dividing by $P(B)$ gives

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

Using the product rule gives another definition of independence.

New definition of independence:

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

Intuitively, this means knowing that B is true tells us nothing about the probabilities of A. The outcome B has no effect on A; therefore, they are independent.

Conversely, if $P(A \mid B) \neq P(A)$, then the events are *dependent*.

Bayes' Theorem, Part 1

Video companion

1 Derivation

Starting from the product rule,

$$P(A \mid B) = \frac{P(A, B)}{P(B)},$$

and multiplying by the probability of B $P(B)$, gives

$$P(A \mid B)P(B) = P(A, B).$$

Substituting the equivalent $P(B, A)$ for $P(A, B)$,

$$P(A \mid B)P(B) = P(B, A),$$

and using the product rule $P(B \mid A) = P(B, A)/P(A)$, gives

$$P(A \mid B)P(B) = P(B \mid A)P(A),$$

which when rearranged is Bayes' theorem.

Bayes' theorem:

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

2 Inverse probability

An inverse probability problem is one where the answer is in the form of the probability that a certain process with a certain probability parameter is being used to generate the observed data.

The symbol B is used to represent the observed data. The symbol A_i is used to represent a possible process with probability parameter θ_i. i = # of possible process

Example: Urn 1 has 20% white marbles, and urn 2 has 10% white marbles. We observe three white marbles in a row drawn with replacement. What is the probability that we are observing urn 1? Urn 2?

What we know:

Process		P(white marble)
Urn 1	A_1	20%
Urn 2	A_2	10%

where "white marble" is the parameter.

In forward probability we are interested in the probability of an event given a known process.

In this problem, we know the outcome and want to know how probable it is that each process was involved.

Written in terms of conditional probability,

$$P(\text{process parameter} \mid \text{observed data}) = \frac{P(\text{observed data} \mid \text{process parameter}_i)P(\text{process parameter}_i)}{P(\text{data} \mid \text{process}_1)P(\text{process}_1) + P(\text{data} \mid \text{process}_2)P(\text{process}_2) + ... + P(\text{data} \mid \text{process}_n)P(\text{process}_n)}$$

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + ... + P(B \mid A_n)P(A_n)}$$ > sum rule = P(B)

Example arithmetic:

First, solve for likelihoods:

Urn 1: $P(3 \text{ white marbles in a row} \mid 20\% \text{ white}) = (0.2)(0.2)(0.2) = 8/1000$
Urn 2: $P(3 \text{ white marbles in a row} \mid 10\% \text{ white}) = (0.1)(0.1)(0.1) = 1/1000$

the likelihood

Using principle of indifference,

$$P(A_1) = 0.5$$
$$P(A_2) = 0.5$$

because we are neutral before observing any data. These are called the "prior probability."

$$\begin{aligned} P(A_1 \mid B) &= \frac{P(B \mid A_1)P(A_1)}{P(B, A_1) + P(B, A_2)} \\ &= \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)} \\ &= \frac{(8/1000)(1/2)}{(8/1000)(1/2) + (1/1000)(1/2)} = 8/9 \end{aligned}$$

Probability that we observed urn 1: 8/9
Probability that we observed urn 2: 1/9

Bayes' Theorem: Updating with New Data

Video companion

1 Updating probabilities

Bayes' theorem allows for updating probabilities based on new data.

From previous solution:

Process		P(white marble)
Urn 1	A_1	20%
Urn 2	A_2	10%

$P(\text{Urn 1} \mid \text{3 white marbles in a row}) = 8/9$
$P(\text{Urn 2} \mid \text{3 white marbles in a row}) = 1/9$

New information: We draw a fourth marble that is also white.

$P(A_1)$ and $P(A_2)$ become our *new prior probabilities* or *new priors*.

$$\begin{aligned}
&P(\text{urn 1} \mid \text{3 white marbles in a row, } \textit{and} \text{ a 4th}) \\
&= \frac{P(\text{white} \mid \text{urn 1})P(\text{urn 1})}{P(\text{white} \mid \text{urn 1})P(\text{urn 1}) + P(\text{white} \mid \text{urn 2})P(\text{urn 2})} \\
&= \frac{(0.2)(8/9)}{(0.2)(8/9) + (0.1)(1/9)}
\end{aligned}$$

$P(\text{urn 1}) = 94.12\%$ $\uparrow$
$P(\text{urn 2}) = 5.88\%$ $\downarrow$

2 Technical vocabulary

Technical vocabulary of Bayesian inverse probability:

$$\underbrace{P(\theta_i \mid D)}_{\text{posterior probability}} = \frac{\overbrace{P(D \mid \theta_i)}^{\text{likelihood}} \quad \overbrace{P(\theta_i)}^{\text{prior probability}}}{\underbrace{P(D)}_{\text{marginal probability}}}$$

posterior probability—probability after new data is observed
prior probability—probability before any data is observed or before new data is observed
likelihood—standard forward probability of data given parameters
marginal probability—probability of the data

Data Science Math Skills
Paul Bendich and Daniel Egger
Duke University

The Binomial Theorem and Bayes' Theorem

Video companion

1 Introduction

Binomial theorem used when there are two possible outcomes—a success or a non-success, for example, flipping a coin—heads are a success, binary outcome.

Not limited to fair coins, where the probability of success is 0.5. Probability can be any value > 0 and < 1.

2 Binomial theorem

Probability of s successes in n trials, when probability of 1 success is p:

$$\binom{n}{s} p^s (1-p)^{n-s}$$

where n is the number of independent trials (with replacement), s is the number of successes, and p is the probability of one success

Example: 72 heads out of 100 coin tosses of a fair coin

$$n = 100$$
$$s = 72$$
$$p = 0.5$$

$$\binom{100}{72} (0.5)^{72} (1-0.5)^{100-72}$$
$$= \binom{100}{72} (0.5)^{72} (0.5)^{28} = 3.94 \times 10^{-6}$$

3 With Bayes' theorem

Question: Is it more likely a fair coin ($p = 0.5$) heads or a bent coin ($p = 0.55$) heads?

$$
\begin{aligned}
&P(\text{fair coin} \mid 72\text{ heads}/100) \\
&= \frac{P(72\text{ heads}/100 \mid \text{fair coin})P(\text{fair coin})}{P(72\text{ heads}/100 \mid \text{fair coin})P(\text{fair coin}) + P(72\text{ heads}/100 \mid \text{bent coin})P(\text{bent coin})} \\
&= \frac{(3.94 \times 10^{-6})(1/2)}{(3.94 \times 10^{-6})(1/2) + (1.972 \times 10^{-4})(1/2)} \\
&= 1.96\%
\end{aligned}
$$

(assuming it is equally likely that the coin is fair or bent)

Therefore, there is $< 2\%$ probability that the coin is fair and $> 98\%$ probability that the coin is bent.

Bayes' theorem together with binomial theorem can tell us the probability of a process given data that we have observed.