Probability basics

Last modified: 2026-11-13 06:54:40 (PST)

Defining probabilities

Definition 0.1 (Step)**.** A **step** is the carrying out of one instruction.

Example 0.1 (Steps of a die roll)**.** Carrying out the instruction "roll a six-sided die once" is one step. Carrying out the instruction "read the number facing up" is another step.

Definition 0.2 (Procedure)**.** A **procedure** is a list of instructions that can be followed, one step at a time, until they are finished.

Example 0.2 (Rolling a die once)**.** The instructions "roll a six-sided die once, then read the number facing up" comprise a procedure. Following the procedure takes the two steps in Example 0.1, and then it is finished.

Definition 0.3 (Outcome)**.** An **outcome** is a single, fully specified way that a procedure can end, recorded as one value.

Example 0.3 (Outcomes of a die roll)**.** The procedure in Example 0.2 ends with one number facing up. Each of the numbers 1, 2, 3, 4, 5 and 6 is an outcome, and one roll ends in exactly one of them.

Definition 0.4 (Random experiment)**.** A **random experiment** is a procedure that ends in exactly one outcome that we have not determined in advance [1].

Remark

Remark. The possible outcomes are collected in the sample space. Physical procedures, such as rolling a die, and sampling procedures, such as picking a unit from a data set without looking, are both random experiments.

Example 0.4 (Rolling a die as a random experiment)**.** The procedure in Example 0.2 is a random experiment: it ends in exactly one of the outcomes 1, 2, 3, 4, 5 or 6 (Example 0.3), and we do not determine in advance which outcome the procedure will produce.

Example 0.5 (Picking a car model from the Auto data)**.** The Auto data record the number of engine cylinders and the fuel economy (miles per gallon) of 392 car models sold between 1970 and 1982. They come from the StatLib library at Carnegie Mellon University and ship with the book *An Introduction to Statistical Learning* and the R package ISLR2 (`ISLR2::Auto`). This site has a copy at `data/auto.csv`.
Pick one car model from the data without looking, and record its number of cylinders. Which model comes up is not determined in advance, so this procedure is a random experiment. Its possible outcomes are the cylinder counts that occur in the data:

```
# read the local copy of the Auto data (one row per car model):
auto <- read.csv("data/auto.csv")

# the possible outcomes are the distinct cylinder counts in the data:
sort(unique(auto$cylinders))
#> [1] 3 4 5 6 8
```

Definition 0.5 (Sample space)**.** The **sample space** of a random experiment, denoted Ω, is the set of all of its possible outcomes [2].

Example 0.6 (Sample space of a die roll)**.** Rolling a six-sided die once has sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Definition 0.6 (Event)**.** An **event** is a subset of the sample space Ω [3]: a set of outcomes.

Remark

Remark. When Ω is uncountable, such as an interval of real numbers, only the subsets in a designated σ-algebra $\mathcal{S}$ on Ω count as events (some sources call a σ-algebra a **Borel field** [4], Def. 1.2.1); every set that arises in these notes is in $\mathcal{S}$. When Ω is finite or countably infinite, every subset of Ω can be an event, because the collection of all subsets of Ω is a σ-algebra (all subsets form a σ-algebra), as in σ-algebras for die rolls.

Example 0.7 (Rolling an even number)**.** In Example 0.6, the event "the roll is even" is $A = \{2, 4, 6\}$.

Lemma 0.1 (The sample space is an event)**.** *The sample space Ω is an event.*

Proof

Proof. The sample space is a subset of itself, $\Omega \subseteq \Omega$. When every subset of Ω can be an event, that is enough. Otherwise the events are the subsets in a σ-algebra $\mathcal{S}$ on Ω, and by its definition $\mathcal{S}$ contains Ω itself. □

Example 0.8 (The die roll's sample space as an event)**.** In Example 0.6, every subset of $\Omega = \{1, 2, 3, 4, 5, 6\}$ can be an event, including Ω itself: the event "the roll is any of 1 through 6", which always occurs.

Definition 0.7 (Occurrence of an event)**.** An event A **occurs** when the outcome of the experiment is in A, and does not occur when the outcome is not in A.

Example 0.9 (Rolling a 4)**.** In Example 0.7, if the die shows 4, the event "the roll is even", $\{2, 4, 6\}$, occurs, because $4 \in \{2, 4, 6\}$, and the event "the roll is at most 2", $\{1, 2\}$, does not occur.

Definition 0.8 (Complement of an event)**.** The **complement** of an event A, denoted $\neg A$, is the event that A does not occur: the set of elements ω of the sample space Ω that are not in A:

$$\neg A \stackrel{\text{def}}{=} \{\omega \in \Omega : \omega \notin A\}$$

Remark

Remark. Other sources write A^c or $\bar{A}$.

Example 0.10 (Rolling an odd number)**.** In Example 0.7, the complement of the event "the roll is even", $A = \{2, 4, 6\}$, is the event "the roll is odd", $\neg A = \{1, 3, 5\}$.

Proposition 0.1 (The complement of an event is a set difference)**.** *For any event A, its complement $\neg A$ is the set difference $\Omega \setminus A$:*

$$\neg A = \Omega \setminus A$$

Proof

Proof.

$$\begin{aligned} \neg A &= \{\omega \in \Omega : \omega \notin A\} && \text{(definition of the complement)} \\ &= \Omega \setminus A && \text{(definition of the set difference)} \end{aligned}$$

□

Example 0.11 (A complement as a set difference)**.** For the fair die in Example 0.6, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. For the event $A = \{2, 4, 6\}$ in Example 0.7, the set difference is $\Omega \setminus A = \{1, 3, 5\}$, which is the complement $\neg A$ in Example 0.10.

Definition 0.9 (Mutually exclusive events (disjoint, pairwise disjoint, mutually disjoint))**.** Finitely or countably many sets $A_1, A_2, ...$, such as events, are **mutually exclusive** (also called **disjoint**, **pairwise disjoint**, or **mutually disjoint**) when no two of them share an element:

$$A_i \cap A_j = \emptyset \quad \text{for all } i \neq j$$

Remark

Remark. The math notes define the same property for sets in general, as pairwise disjoint.

Example 0.12 (Low and high rolls)**.** In Example 0.6, the events "the roll is at most 2", $\{1, 2\}$, and "the roll is at least 5", $\{5, 6\}$, are mutually exclusive: no roll is in both. The events "the roll is even", $\{2, 4, 6\}$, and "the roll is at most 2", $\{1, 2\}$, are not mutually exclusive: a roll of 2 is in both.

Definition 0.10 (Partition of an event)**.** A **partition** of an event A is a finite or countably infinite collection of events $A_1, A_2, \ldots$ that satisfies:

- The events $A_1, A_2, \ldots$ are mutually exclusive.
- Their union is A: $\bigcup_i A_i = A$.

Remark (A partition of the sample space)

Remark 0.1 (A partition of the sample space). A partition of the sample space is the case $A = \Omega$, which is allowed because Ω is an event (Lemma 0.1).

Remark

Remark. Some sources also require each piece A_i to contain at least one outcome.

Example 0.13 (Partitioning the die roll)**.** In Example 0.6, the events "the roll is even", $\{2, 4, 6\}$, and "the roll is odd", $\{1, 3, 5\}$, are mutually exclusive, and their union is Ω, so they partition the sample space. So do the six single-outcome events $\{1\}, \{2\}, \ldots, \{6\}$. The events $\{1, 2\}$ and $\{5, 6\}$ from Example 0.12 do not partition Ω: their union leaves out 3 and 4.

Proposition 0.2 (An event and its complement partition the sample space)**.** *For any event A, the events A and its complement $\neg A$ are a partition of the sample space Ω.*

Proof

Proof. The sample space Ω is an event (Lemma 0.1), so a partition of it is defined. By Proposition 0.1, the complement $\neg A$ is the set difference $\Omega \setminus A$, which is an event because the events are closed under complements in Ω, by the complement requirement of a σ-algebra. The events A and $\neg A$ are mutually exclusive, because $\neg A$ contains only elements not in A. Their union is Ω, because each element of Ω is either in A or not in A. □

Example 0.14 (Even and odd rolls)**.** In Example 0.7 and Example 0.10, the event $A = \{2, 4, 6\}$ and its complement $\neg A = \{1, 3, 5\}$ are mutually exclusive, and their union is $\Omega = \{1, 2, 3, 4, 5, 6\}$, so they partition Ω.

Definition 0.11 (Probability measure (probability function))**.** A **probability measure** (also called a **probability function** [4], Definition 1.2.4) on a sample space Ω, often

denoted $\Pr()$ or $\mathbb{P}()$, is a measure on the events of Ω that gives the whole sample space probability 1 [5]:

$$\Pr(\Omega) = 1$$

Remark (The probability space)

Remark 0.2 (The probability space). The sample space, its events and the probability measure together form a probability space.

Example 0.15 (Probability measure for a fair die)**.** For the die roll in Example 0.6, define $\Pr(A) \stackrel{\text{def}}{=} |A|/6$, where $|A|$ is the number of outcomes in A. This function is $1/6$ times the counting measure $\mu(A) = |A|$ on the die rolls (counting elements is a measure), so it is a measure on the events of Ω, and it gives

$$\begin{aligned}\Pr(\Omega) &= 6/6 \\ &= 1.\end{aligned}$$

The event "the roll is even" from Example 0.7 has probability

$$\begin{aligned}\Pr(\{2,4,6\}) &= 3/6 \\ &= 1/2.\end{aligned}$$

Corollary 0.1 (Probability measures are finitely additive)**.** *Every probability measure is finitely additive: for any mutually exclusive events $A_1, \ldots, A_n$,*

$$\Pr(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \Pr(A_i)$$

Proof

Proof. A probability measure is a measure (Definition 0.11), so measures are finitely additive applies to it. □

Theorem 0.1 (Continuity of probability)**.** *If $A_1 \supseteq A_2 \supseteq \cdots$ are events with $\bigcap_{i=1}^{\infty} A_i = \emptyset$, then:*

$$\lim_{n \to \infty} \Pr(A_n) = 0$$

Proof

Proof. For each i, let $B_i \stackrel{\text{def}}{=} A_i \setminus A_{i+1}$, the outcomes in A_i but not in A_{i+1}. Each $B_i = A_i \cap (\Omega \setminus A_{i+1})$ is an event, by closure properties of a σ-algebra.
The B_i are pairwise disjoint. For $i < j$, $B_j \subseteq A_j \subseteq A_{i+1}$, while B_i has no outcomes in A_{i+1}, so $B_i \cap B_j = \emptyset$.
$A_n = \bigcup_{i=n}^{\infty} B_i$ *for each* n. For $i \geq n$, $B_i \subseteq A_i \subseteq A_n$, so the union is contained in A_n. Conversely, let $\omega \in A_n$. Since $\bigcap_i A_i = \emptyset$, ω is not in every A_i; since the A_i are nested, the indices i with $\omega \in A_i$ are $1, \ldots, m$ for some $m \geq n$. Then $\omega \in A_m$ and $\omega \notin A_{m+1}$, so $\omega \in B_m$.
$\Pr(A_1)$ *is finite.* The events A_1 and $\Omega \setminus A_1$ are disjoint with union Ω, so Corollary 0.1 applies to them:

$$\begin{aligned} \Pr(A_1) &\leq \Pr(A_1) + \Pr(\Omega \setminus A_1) && (\Pr(\Omega \setminus A_1) \geq 0) \\ &= \Pr(\Omega) && (\text{finite additivity of } \Pr) \\ &= 1 && (\text{definition of a probability measure}) \end{aligned}$$

The limit. For each n:

$$\begin{aligned} \Pr(A_n) &= \Pr\left(\bigcup_{i=n}^{\infty} B_i\right) && \left(A_n = \bigcup_{i=n}^{\infty} B_i\right) \\ &= \sum_{i=n}^{\infty} \Pr(B_i) && (\text{countable additivity; the } B_i \text{ are pairwise disjoint}) \end{aligned}$$

With $n = 1$, this says the series $\sum_{i=1}^{\infty} \Pr(B_i)$ has the finite total $\Pr(A_1)$. So, for $n \geq 2$:

$$\begin{aligned} \Pr(A_n) &= \sum_{i=1}^{\infty} \Pr(B_i) - \sum_{i=1}^{n-1} \Pr(B_i) && (\text{split off the first } n-1 \text{ terms; the total is finite}) \\ &= \Pr(A_1) - \sum_{i=1}^{n-1} \Pr(B_i) && (\text{the case } n = 1) \end{aligned}$$

As $n \to \infty$, the partial sums $\sum_{i=1}^{n-1} \Pr(B_i)$ converge to the total $\Pr(A_1)$, so $\Pr(A_n) \to \Pr(A_1) - \Pr(A_1) = 0$. □

Remark

Remark. Requiring countable additivity, not just finite additivity, enables results such as Theorem 0.1, and it is what makes sums over countably infinite partitions valid.

Theorem 0.2 (Kolmogorov axioms)**.** *A function* Pr *that assigns a real number* $\Pr(A)$ *to each event* A *of a sample space* Ω *is a probability measure if and only if it satisfies:*

1. *For any event* A, $\Pr(A) \geq 0$.
2. *The probability of the whole sample space is 1:*

$$\Pr(\Omega) = 1$$

3. Pr *is countably additive: for any mutually exclusive events* $A_1, A_2, \ldots$,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$$

Proof

Proof. Suppose Pr is a probability measure. It is a measure, so its values lie in the extended non-negative reals $[0, \infty]$, which gives axiom 1, and it is countably additive, which is axiom 3. Axiom 2 is the condition $\Pr(\Omega) = 1$ in Definition 0.11.
Conversely, suppose Pr satisfies the three axioms. By axiom 1, Pr takes values in $[0, \infty]$, so axiom 3 says Pr is countably additive in the sense of the definition of countable additivity. Then, by countable additivity and the empty set, $\Pr(\emptyset)$ is 0 or ∞; $\Pr(\emptyset)$ is a real number, so $\Pr(\emptyset) = 0$. So Pr satisfies both conditions of the definition of a measure ($\Pr(\emptyset) = 0$ and countable additivity) and is a measure on the events of Ω. With axiom 2, Pr is a probability measure (Definition 0.11). □

Remark

Remark. Many sources define a probability measure by these three axioms instead; the theorem shows that the two definitions agree. The axioms are named for Kolmogorov, who introduced them in 1933 [6]. The axiom form does not list $\Pr(\emptyset) = 0$; the proof above derives it.

Theorem 0.3 (Probability of a subset's intersection)**.** *If* A *and* B *are events and* $A \subseteq B$, *then* $\Pr(A \cap B) = \Pr(A)$.

Proof

Proof. Since $A \subseteq B$, every outcome in A is also in B, so $A \cap B = A$, and therefore $\Pr(A \cap B) = \Pr(A)$. □

Theorem 0.4 (An event and its complement sum to 1)**.** *For any event A and its complement $\neg A$:*

$$\Pr(A) + \Pr(\neg A) = 1$$

Proof

Proof. By Proposition 0.2, the events A and $\neg A$ are a partition of Ω, so they are disjoint, and their union is Ω.

$$\begin{aligned} \Pr(A) + \Pr(\neg A) &= \Pr(A \cup \neg A) && \text{(additivity of probability for disjoint events)} \\ &= \Pr(\Omega) && (A \cup \neg A = \Omega) \\ &= 1 && \text{(probability of the sample space is 1)} \end{aligned}$$

□

Corollary 0.2 (Complement rule)**.** *For any event A:*

$$\Pr(\neg A) = 1 - \Pr(A)$$

Proof

Proof. Subtract $\Pr(A)$ from both sides of Theorem 0.4. □

Corollary 0.3 (Complement rule in probability (π) notation)**.** *If the probability of an event A is $\Pr(A) = \pi$, then the probability that A does not occur is:*

$$\Pr(\neg A) = 1 - \pi$$

Proof

Proof.

$$\begin{aligned} \Pr(\neg A) &= 1 - \Pr(A) && \text{(complement rule)} \\ &= 1 - \pi && \text{(substitute } \Pr(A) = \pi) \end{aligned}$$

□

Example 0.16 (Probability of not rolling a six)**.** For a fair die, the event "the roll is a six" has probability $\pi = 1/6$, so by Corollary 0.3 the probability of not rolling a six is $1 - 1/6 = 5/6$.

Definition 0.12 (Choosing at random)**.** Let S be a set with a measure μ such that $\mu(S) \in (0, \infty)$. To **choose an element of S at random** is to run a random experiment whose sample space is S and whose probability measure is the uniform distribution on S.

Example 0.17 (Choosing at random from a finite set and from an interval)**.** For a finite set S, let $\mu(A) = |A|$ be the counting measure. Then $\Pr(A) = |A|/|S|$, so every element of S has the same probability $1/|S|$.
For the interval $S = [0, 1]$, let $\mu(A)$ be the length of A. Then $\Pr(A)$ is the length of A, so the probability that the chosen number lies in $[0.2, 0.5]$ is 0.3. The chosen number has the Uniform$(0, 1)$ distribution.
The integers have infinite counting measure, so they have no choice at random in this sense.

Example 0.18 (Choosing a car at random from the Auto data)**.** The Auto data of Example 0.5 have 392 car models. Let S be the set of car models in the data. Choosing a car at random from S (Definition 0.12), with the counting measure, gives each car probability $1/|S|$, so the probability that the car has four cylinders is the fraction of cars with four cylinders:

```
# read the local copy of the Auto data (one row per car model):
auto <- read.csv("data/auto.csv")

# the event "four cylinders" is the set of rows where cylinders == 4;
# the mean of the logical vector is |A| / |S|, the probability in
# the definition of choosing at random:
mean(auto$cylinders == 4)
#> [1] 0.5076531
```

Odds

Definition 0.1 (Odds)**.** Let A be an event with probability $\mathbb{P}(A) \in (0, 1)$. The **odds** of A are [7]

$$\omega \stackrel{\text{def}}{=} \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)} \in (0, \infty).$$

Example 0.1 (Odds that a car in the Auto data is American)**.** In the Auto data of Example 0.5, let A be the event that a car chosen at random is American, which the data code as `origin == 1`. Its probability is the fraction of American cars, as in Example 0.18, and its odds follow from Definition 0.1:

```
auto <- read.csv("data/auto.csv")
p_american <- mean(auto$origin == 1)
c(probability = p_american, odds = p_american / (1 - p_american))
#> probability        odds
#>    0.625000    1.666667
```

The odds 1.67 say that an American car is about 1.7 times as likely as a car of non-American origin.

Lemma 0.1 (One over one plus the odds is the complement probability)**.** *Let A be an event with probability $\mathbb{P}(A) \in (0, 1)$ and odds ω. One over one plus the odds of A is the probability that A does not happen:*

$$\frac{1}{1+\omega} = 1 - \mathbb{P}(A).$$

Proof

Proof.

$$\begin{aligned}
\frac{1}{1+\omega} &= \frac{1}{1+\frac{\mathbb{P}(A)}{1-\mathbb{P}(A)}} && \text{(definition of the odds)} \\
&= \frac{1}{\frac{1-\mathbb{P}(A)}{1-\mathbb{P}(A)} + \frac{\mathbb{P}(A)}{1-\mathbb{P}(A)}} && \left(1 = \tfrac{1-\mathbb{P}(A)}{1-\mathbb{P}(A)}\right) \\
&= \frac{1}{\frac{1-\mathbb{P}(A)+\mathbb{P}(A)}{1-\mathbb{P}(A)}} && \text{(add fractions with a common denominator)} \\
&= \frac{1}{\frac{1}{1-\mathbb{P}(A)}} && (-\mathbb{P}(A) + \mathbb{P}(A) = 0) \\
&= 1 - \mathbb{P}(A) && \text{(the reciprocal of a reciprocal)}
\end{aligned}$$

□

Example 0.2 (The probability that a car in the Auto data is not American, from the odds)**.**

For the event A of Example 0.1, Lemma 0.1 says that one over one plus the odds is the probability that the car is not American:

```
odds_american <- p_american / (1 - p_american)
c(from_odds = 1 / (1 + odds_american), complement = 1 - p_american)
#>  from_odds complement
#>      0.375      0.375
```

Both ways give 0.375.

Definition 0.2 (Log-odds (also called the logit))**.** The **log-odds**, also called the **logit**, of an event A with probability $\mathbb{P}(A) \in (0, 1)$ are the natural logarithm of its odds [8]:

$$\eta \stackrel{\text{def}}{=} \log\{\omega\} = \log\left\{\frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}\right\} \in \mathbb{R}.$$

Example 0.3 (Log-odds that a car in the Auto data is American)**.** For the event A of Example 0.1, the log-odds of Definition 0.2 are:

```
log(p_american / (1 - p_american))
#> [1] 0.5108256
```

The log-odds are positive because the odds exceed 1: an American car is more likely than not.

Lemma 0.2 (The exponential of minus the log-odds is one over the odds)**.** *Let A be an event with probability $\mathbb{P}(A) \in (0, 1)$, odds ω and log-odds η. The exponential of minus the log-odds of A is one over its odds:*

$$e^{-\eta} = \frac{1}{\omega}.$$

Proof

Proof.

$$\begin{aligned} \mathrm{e}^{-\eta} &= \mathrm{e}^{-\log\{\omega\}} && \text{(definition of the log-odds)} \\ &= \frac{1}{\mathrm{e}^{\log\{\omega\}}} && (\mathrm{e}^{-x} = 1/\mathrm{e}^{x}) \\ &= \frac{1}{\omega} && \text{(the exponential undoes the natural logarithm)} \end{aligned}$$

□

Example 0.4 (The exponential of minus the log-odds that a car in the Auto data is American)**.**
For the event A of Example 0.1, Lemma 0.2 says that the exponential of minus the log-odds is one over the odds:

```
odds_american <- p_american / (1 - p_american)
log_odds_american <- log(odds_american)
exp_neg_log_odds <- exp(-log_odds_american)
c(exp_neg_log_odds = exp_neg_log_odds, one_over_odds = 1 / odds_american)
#> exp_neg_log_odds    one_over_odds
#>              0.6              0.6
```

Both ways give 0.6.

Conditional probability

Definition 0.1 (Conditional probability)**.** For two events A and B with $\Pr(B) > 0$, the **conditional probability** of A given B [9], denoted $\Pr(A \mid B)$, is:

$$\Pr(A \mid B) \stackrel{\text{def}}{=} \frac{\Pr(A \cap B)}{\Pr(B)}$$

Remark

Remark. Geometrically and computationally, conditioning on an event B can be viewed as *slicing* the sample space to retain only the outcomes compatible with B, and then *renormalizing* the remaining probabilities (dividing by $\Pr(B)$) so that they sum to 1 [10, Sec. 2.7.4].

Example 0.1 (Rolling a six, given an even roll)**.** For a fair die (Example 0.15), let $A = \{6\}$ and $B = \{2, 4, 6\}$. Then $A \cap B = \{6\}$, so:

$$\begin{aligned}
\Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} && \text{(definition of conditional probability)} \\
&= \frac{1/6}{3/6} && \text{(substitute the probabilities)} \\
&= \frac{1}{3} && \text{(simplify)}
\end{aligned}$$

Theorem 0.1 (Law of conditional probability (probability chain rule, general product rule))**.** *For any two events A and B with $\Pr(B) > 0$:*

$$\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B)$$

Proof

Proof. Rearranging Definition 0.1:

$$\begin{aligned}
\Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} && \text{(definition of conditional probability)} \\
\Pr(A \cap B) &= \Pr(A \mid B) \cdot \Pr(B) && \text{(multiply both sides by } \Pr(B)\text{)}
\end{aligned}$$

□

Example 0.2 (Applying the law of conditional probability)**.** Suppose 30% of adults exercise regularly ($\Pr(E) = 0.30$), and among adults who exercise regularly, 60% have low blood pressure ($\Pr(L \mid E) = 0.60$).
Then, by Theorem 0.1, the probability that a randomly selected adult both exercises regularly and has low blood pressure is:

$$\begin{aligned}
\Pr(L \cap E) &= \Pr(L \mid E) \cdot \Pr(E) && \text{(law of conditional probability)} \\
&= 0.60 \cdot 0.30 && \text{(substitute the given values)} \\
&= 0.18 && \text{(multiply)}
\end{aligned}$$

Theorem 0.2 (Law of total probability)**.** *If $B_1, B_2, \ldots$ is a partition of the sample space, then for any event A [11]:*

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \cap B_i)$$

Proof

Proof. Since $B_1, B_2, \ldots$ partition the sample space, the events $A \cap B_1, A \cap B_2, \ldots$ are mutually exclusive and their union is A. By countable additivity:

$$\begin{aligned}
\Pr(A) &= \Pr\left(\bigcup_{i=1}^{\infty}(A \cap B_i)\right) && \text{(the union of the events } A \cap B_i \text{ is } A\text{)} \\
&= \sum_{i=1}^{\infty} \Pr(A \cap B_i) && \text{(countable additivity for the mutually exclusive events } A \cap B_i\text{)}
\end{aligned}$$

□

Theorem 0.3 (Law of total probability with conditional probabilities)**.** *If $B_1, B_2, \ldots$ is a partition of the sample space with $\Pr(B_i) > 0$ for every i, then for any event A:*

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \mid B_i) \cdot \Pr(B_i)$$

Proof

Proof. Apply Theorem 0.1 to each summand of Theorem 0.2, which needs $\Pr(B_i) > 0$:

$$\begin{aligned}
\Pr(A) &= \sum_{i=1}^{\infty} \Pr(A \cap B_i) && \text{(law of total probability)} \\
&= \sum_{i=1}^{\infty} \Pr(A \mid B_i) \cdot \Pr(B_i) && \text{(law of conditional probability; } \Pr(B_i) > 0\text{)}
\end{aligned}$$

□

Theorem 0.4 (Bayes' theorem)**.** *For any two events A and B with $\Pr(A) > 0$ and $\Pr(B) > 0$ [12]:*

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}$$

Proof

Proof. By Definition 0.1 and Theorem 0.1:

$$
\begin{aligned}
\Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} && \text{(definition of conditional probability)} \\
&= \frac{\Pr(B \cap A)}{\Pr(B)} && \text{(intersection is commutative: } A \cap B = B \cap A\text{)} \\
&= \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} && \text{(law of conditional probability applied to } \Pr(B \cap A)\text{)}
\end{aligned}
$$

□

Example 0.3 (Positive predictive value of a medical test)**.** Suppose a disease test has 99% sensitivity and 99% specificity, and the prevalence of the disease in the population is 7%.
Let D be the event "person has the disease" and $+$ be the event "test is positive". Then:

- $\Pr(+ \mid D) = 0.99$ (sensitivity)
- $\Pr(- \mid \neg D) = 0.99$ (specificity), so the false positive rate is

$$
\begin{aligned}
\Pr(+ \mid \neg D) &= 1 - 0.99 \\
&= 0.01
\end{aligned}
$$

- $\Pr(D) = 0.07$ (prevalence)

By Theorem 0.4, with the denominator expanded by the law of total probability over the partition $\{D, \neg D\}$:

$$
\begin{aligned}
\Pr(D \mid +) &= \frac{\Pr(+ \mid D) \cdot \Pr(D)}{\Pr(+)} && \text{(Bayes' theorem)} \\
&= \frac{\Pr(+ \mid D) \cdot \Pr(D)}{\Pr(+ \mid D) \cdot \Pr(D) + \Pr(+ \mid \neg D) \cdot \Pr(\neg D)} && \text{(law of total probability)} \\
&= \frac{0.99 \cdot 0.07}{0.99 \cdot 0.07 + 0.01 \cdot 0.93} && \text{(substitute the given values)} \\
&= \frac{0.0693}{0.0693 + 0.0093} && \text{(multiply each term in the numerator and denomina} \\
&= \frac{0.0693}{0.0786} && \text{(add the denominator's two terms)} \\
&\approx 0.88 && \text{(divide)}
\end{aligned}
$$

Even with a highly accurate test (99% sensitive and 99% specific), only about 88% of people who test positive actually have the disease, because the disease prevalence is relatively low (7%).

Exercise 0.1 (Turn a conditional around)**.** Definition 0.1 defines $\Pr(A \mid B)$ in terms of $\Pr(A \cap B)$. Write the same definition with A and B swapped, then eliminate $\Pr(A \cap B)$ between the two.
Your result should express $\Pr(A \mid B)$ using $\Pr(B \mid A)$, $\Pr(A)$ and $\Pr(B)$, and none of them jointly.

Solution

Solution 0.1. Multiplying Definition 0.1 through by its denominator, in each direction,

$$\Pr(A \cap B) = \Pr(A \mid B)\ \Pr(B), \qquad \Pr(A \cap B) = \Pr(B \mid A)\ \Pr(A)$$

The left-hand sides are the same quantity, so the right-hand sides are equal. Dividing by $\Pr(B)$ gives **Bayes' rule** (Theorem 0.4):

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\ \Pr(A)}{\Pr(B)} \tag{1}$$

Nothing was assumed beyond $\Pr(A) > 0$ and $\Pr(B) > 0$, which the two conditionals need in order to be defined at all.
As Hutchinson emphasizes [13], each component carries standard Bayesian terminology:

- $\Pr(A \mid B)$ is the **posterior probability** of A given the observed evidence B.
- $\Pr(B \mid A)$ is the **likelihood** of the evidence B given A.
- $\Pr(A)$ is the **prior probability** of A before seeing B.
- $\Pr(B)$ is the marginal probability of the **evidence**.

From this formula, the direction of effects is transparent: increasing the prior $\Pr(A)$ or the likelihood $\Pr(B \mid A)$ increases the posterior $\Pr(A \mid B)$, while increasing the probability of the evidence $\Pr(B)$ decreases it.

Exercise 0.2 (A test that is right 99% of the time)**.** A screening test for a disease is positive for 99% of people who have it and negative for 99% of people who do not. One person in a thousand has the disease.
A randomly screened person tests positive. What is the probability that they have the disease? Guess first, then compute it with Equation 1.

Solution

Solution 0.2. Write S for having the disease and $+$ for a positive test. The three given numbers are

$$\Pr(+ \mid S) = 0.99, \qquad \Pr(+ \mid \text{not } S) = 0.01, \qquad \Pr(S) = 0.001$$

Equation 1 needs $\Pr(+)$, the overall chance of a positive result. Assemble it with Theorem 0.3, splitting on whether the person is sick:

$$\begin{aligned}\Pr(+) &= \Pr(+ \mid S)\Pr(S) + \Pr(+ \mid \text{not } S)\Pr(\text{not } S) \\ &= (0.99)(0.001) + (0.01)(0.999) \\ &= 0.00099 + 0.00999 \\ &= 0.01098\end{aligned}$$

Then

$$\begin{aligned}\Pr(S \mid +) &= \frac{\Pr(+ \mid S)\ \Pr(S)}{\Pr(+)} \\ &= \frac{0.00099}{0.01098} \\ &\approx 0.0902\end{aligned}$$

About **9%**, not the 99% most people guess.
The reason is the **base rate** [14]. Out of 100,000 people screened, about 100 have the disease and 99 of them test positive, while 99,900 do not have it and about 999 of them test positive anyway. The false positives outnumber the true positives ten to one, because there are a thousand times more people available to produce them.
The base rate is fundamental to evaluating screening tests and binary classifiers, and it illustrates why raw accuracy can be misleading on rare events.

The course website's version of this page has an interactive plot here.

Video lecture

Hutchinson's Probability Refresher (27 min) covers conditional distributions, the law of total probability, the chain rule of probability, and Bayes' rule [15]. The login for the video site is posted on Canvas.

Another explanation

The University of Washington's CSE 312 materials and videos by Alex Tsun teach probability and statistics for computing, and cover counting, conditional probability and Bayes' rule from a computer science viewpoint.

Watch

5MinuteAI's five-minute videos follow the same course [16]: introduction to discrete probability and conditional probability.

Another explanation

Lange's textbook [17] reviews probability and expectation in Section 1.2 and conditional probability in Section 1.3.

Further reading

Another explanation

Evans and Rosenthal's free textbook [18] covers probability models, conditional probability and independence in Chapter 1.

References

[1] Wikipedia contributors, "Experiment (probability theory) — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Experiment_(probability_theory)

[2] Wikipedia contributors, "Sample space — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Sample_space

[3] Wikipedia contributors, "Event (probability theory) — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Event_(probability_theory)

[4] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[5] Wikipedia contributors, "Probability space — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Probability_space

[6] Wikipedia contributors, "Probability axioms — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Probability_axioms

[7] Wikipedia contributors, "Odds — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Odds

[8] Wikipedia contributors, "Logit — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Logit

[9] Wikipedia contributors, "Conditional probability — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Conditional_probability

[10] K. Ross, "An introduction to probability and simulation." Bookdown, 2022. Available: https://bookdown.org/kevin_davisross/probsim-book/

[11] Wikipedia contributors, "Law of total probability — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Law_of_total_probability

[12] Wikipedia contributors, "Bayes' theorem — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Bayes%27_theorem

[13] B. Hutchinson, "DATA 471/571: Machine learning." Western Washington University, 2024.

[14] Wikipedia contributors, "Base rate fallacy — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Base_rate_fallacy

[15] B. Hutchinson, "DATA 471/571 (machine learning) and CSCI 481/581 (deep learning) video lectures." Western Washington University. Accessed: Sep. 28, 2026. [Online]. Available: https://facultyweb.cs.wwu.edu/~hutchib2/video_lectures/data371/

[16] 5MinuteAI, "Probability & statistics." Video playlist on YouTube. Accessed: Oct. 12, 2026. [Online]. Available: https://www.youtube.com/playlist?list=PLeB45KifGiuHesi4PALNZSYZFhViVGQJK

[17] K. Lange, *Applied probability*, 3rd ed. in Springer texts in statistics. Springer, 2024. doi: 10.1007/978-1-0716-4172-9.

[18] M. J. Evans and J. S. Rosenthal, *Probability and statistics: The science of uncertainty*, 2nd ed. Freeman, 2010. Available: https://utstat.utoronto.ca/mikevans/jeffrosenthal/