Classification and Diagnostic Tests

Last modified: 2026-11-12 22:02:32 (PST)

Introduction

Classification is a core problem in statistics and machine learning: we seek to assign individuals or observations to one of several discrete categories based on available data. In medicine and epidemiology, classification problems arise constantly; for example, a clinician decides whether a patient has a disease based on evidence such as:

- test results;
- biomarkers;
- clinical signs.

A test can look highly accurate in isolation, yet its predictive value for an individual patient depends heavily on the prevalence of the condition in the population being tested. Understanding this interplay requires Bayes' theorem and the law of total probability [1].

Definition 0.1 (Classification)**.** A **classification problem** is a statistical problem in which we seek to assign each observation to one of two or more discrete categories based on its observed features or predictors x.
Source: Wikipedia contributors [2].

Definition 0.2 (Class)**.** The discrete categories of a classification problem (Definition 0.1) are its **classes**. Let $\mathcal{C}$ be a set of K classes (for example, $\mathcal{C} = \{1, \dots, K\}$, or a set of text labels).

Definition 0.3 (Classification rule)**.** A **classification rule** c maps features to classes (Definition 0.2):

$$c : x \mapsto c(x) \in \mathcal{C} \tag{1}$$

Definition 0.4 (Predicted probability)**.** The **predicted probability** of class k at features x is the conditional probability of that class given those features, under a model:

$$\hat{\mathbb{p}}_k(x) \stackrel{\text{def}}{=} \Pr(Y = k \mid x) \tag{2}$$

When the model is fitted, the predicted probability is an estimate of the true conditional probability. The two can differ, by the estimation error.

Definition 0.5 (Binary classification)**.** A **binary classification problem** is a classification problem whose set of classes $\mathcal{C}$ consists of $K = 2$ distinct categories, one representing the presence of a condition of interest (k_1) and the other representing its absence (k_0):

$$\mathcal{C} = \{k_1, k_0\}, \quad |\mathcal{C}| = 2 \tag{3}$$

Source: Wikipedia contributors [2].

Definition 0.6 (Positive classification)**.** In a binary classification problem with $\mathcal{C} = \{k_1, k_0\}$, for an observation drawn from a population and assigned class $Y \in \mathcal{C}$ (for example, $Y = c(X)$ when a classification rule c classifies the observation from its feature X), a **positive classification** (also called **positive**, and written $+$) is the event that the observation is assigned to k_1:

$$+ \stackrel{\text{def}}{=} \{Y = k_1\} \tag{4}$$

Source: Wikipedia contributors [2]; Wikipedia contributors [3].

Definition 0.7 (Negative classification)**.** In a binary classification problem with $\mathcal{C} = \{k_1, k_0\}$, for an observation drawn from a population and assigned class $Y \in \mathcal{C}$ (for example, $Y = c(X)$ when a classification rule c classifies the observation from its feature X), a **negative classification** (also called **negative**, and written $-$) is the event that the observation is assigned to k_0:

$$- \stackrel{\text{def}}{=} \{Y = k_0\} \tag{5}$$

Source: Wikipedia contributors [2]; Wikipedia contributors [3].

Exercise 0.1 (Negative classification as a complement)**.** In a binary classification problem with $\mathcal{C} = \{k_1, k_0\}$, show that when each observation is assigned a class $Y \in \mathcal{C}$, the negative classification event $-$ is the complement of the positive classification event $+$: $- = \neg +$.

Solution

Solution 0.1. Because $\mathcal{C} = \{k_1, k_0\}$ with $|\mathcal{C}| = 2$ (Definition 0.5), the classes k_1 and k_0 are distinct, so $Y \in \mathcal{C}$ equals k_0 if and only if $Y \neq k_1$:

$$\begin{aligned} - &= \{Y = k_0\} && \text{(definition of negative classification)} \\ &= \{Y \neq k_1\} && \text{(since } Y \in \{k_1, k_0\} \text{ with } k_1 \neq k_0\text{)} \\ &= \neg\{Y = k_1\} && \text{(definition of complement)} \\ &= \neg + && \text{(definition of positive classification)} \end{aligned} \tag{6}$$

Theorem 0.1 (Negative classification is the complement of positive classification)**.** *In a binary classification problem with $\mathcal{C} = \{k_1, k_0\}$, when each observation is assigned a class $Y \in \mathcal{C}$, the negative classification event $-$ is the complement of the positive classification event $+$:*

$$- = \neg + \tag{7}$$

Proof

Proof. This is the solution to Exercise 0.1, worked in Solution 0.1. □

Example 0.1 (A diagnostic test as a classification rule)**.** In a binary classification problem, a COVID-19 test assigns each person tested to one of $K = 2$ classes, "has COVID-19" (k_1) or "does not have COVID-19" (k_0), based on a single feature x: whether the test's result is "positive" or "negative". The test is the classification rule that assigns a positive result to k_1 and a negative result to k_0:

$$c(x) \stackrel{\text{def}}{=} \begin{cases} \text{has COVID-19} & \text{if } x = \text{positive} \\ \text{does not have COVID-19} & \text{if } x = \text{negative} \end{cases} \tag{8}$$

For a person drawn from a population with test result X, the event that the test is positive is $+ = \{c(X) = \text{has COVID-19}\}$, and the event that the test is negative is $- = \{c(X) = \text{does not have COVID-19}\}$.

Remark (Labels in binary classification)

Remark 0.1 (Labels in binary classification). In a binary classification problem, the two classes k_1 and k_0 are often labeled "positive" and "negative", or, in medical diagnostic testing, "has the disease" and "does not have the disease".

Diagnostic test characteristics

Definition 0.1 (Sensitivity)**.** The **sensitivity** of a diagnostic test is the probability that the test is positive $(+)$, given that the person tested has the disease (D):

$$\text{sensitivity} \stackrel{\text{def}}{=} \Pr(+ \mid D) \tag{9}$$

Source: Wikipedia contributors [3].

Example 0.1 (Sensitivity of a COVID-19 test)**.** Suppose a COVID-19 test is positive for 99% of the people who have COVID-19. Let D be the event "the person has COVID-19" and $+$ the event "the test is positive". Then the test's sensitivity (Definition 0.1) is:

$$\Pr(+ \mid D) = 0.99 \tag{10}$$

Definition 0.2 (Specificity)**.** The **specificity** of a diagnostic test is the probability that the test is negative $(-)$, given that the person tested does not have the disease $(\neg D)$:

$$\text{specificity} \stackrel{\text{def}}{=} \Pr(- \mid \neg D) \tag{11}$$

Source: Wikipedia contributors [3].

Example 0.2 (Specificity of a COVID-19 test)**.** Suppose the COVID-19 test of Example 0.1 is negative for 99% of the people who do not have COVID-19. Then its specificity (Definition 0.2) is:

$$\Pr(- \mid \neg D) = 0.99 \tag{12}$$

By the complement rule and Theorem 0.1, its false positive rate is:

$$\begin{aligned}
\Pr(+ \mid \neg D) &= 1 - \Pr(\neg + \mid \neg D) && \text{(complement rule, conditional on } \neg D) \\
&= 1 - \Pr(- \mid \neg D) && \text{(negative classification is the complement of positive)} \\
&= 1 - 0.99 && \text{(substitute the specificity)} \\
&= 0.01 && \text{(subtract)}
\end{aligned} \tag{13}$$

Definition 0.3 (Prevalence)**.** The **prevalence** of a disease in a population is the probability that a person drawn from that population has the disease (D):

$$\text{prevalence} \stackrel{\text{def}}{=} \Pr(D) \tag{14}$$

Source: Wikipedia contributors [4].

Example 0.3 (Prevalence of COVID-19)**.** Suppose 7% of the population being tested with the COVID-19 test of Example 0.1 has COVID-19. Then the prevalence (Definition 0.3) is:

$$\Pr(D) = 0.07 \tag{15}$$

and, by the complement rule, the probability that a person drawn from that population does not have COVID-19 is:

$$\begin{aligned} \Pr(\neg D) &= 1 - \Pr(D) && \text{(complement rule)} \\ &= 1 - 0.07 && \text{(substitute the prevalence)} \\ &= 0.93 && \text{(subtract)} \end{aligned} \tag{16}$$

Predictive values

Definition 0.1 (Positive predictive value (PPV))**.** The **positive predictive value** of a diagnostic test is the probability that the person tested has the disease (D), given that the test is positive ($+$):

$$\text{PPV} \stackrel{\text{def}}{=} \Pr(D \mid +) \tag{17}$$

Source: Wikipedia contributors [5].

Example 0.1 (PPV of a COVID-19 test)**.** For the COVID-19 test of Example 0.1 and Example 0.2, used in the population of Example 0.3, Bayes' theorem gives the positive predictive value (Definition 0.1):

$$\Pr(D \mid +) \approx 0.88 \tag{18}$$

Even with a highly accurate test (99% sensitive and 99% specific), only about 88% of the people who test positive have COVID-19, because the prevalence (7%) is low enough that false positives make up a meaningful fraction of all positive tests.

Example 0.2 (Predicted probability of COVID-19 after a positive test)**.** Consider the model that gives every person with a positive test ($x = +$) the probability of having COVID-19 that Example 0.1 computes. That probability is the positive predictive value, which is the predicted probability (Definition 0.4) of the class "has COVID-19" ($k = 1$):

$$\hat{\mathbb{p}}_1(+) \approx 0.88 \tag{19}$$

The classification rule of Example 0.1 predicts "has COVID-19" for this person, and the predicted probability says how sure that prediction is: about 88%, not 100%.

Definition 0.2 (Negative predictive value (NPV))**.** The **negative predictive value** of a diagnostic test is the probability that the person tested does not have the disease ($\neg D$), given that the test is negative ($-$):

$$\text{NPV} \stackrel{\text{def}}{=} \Pr(\neg D \mid -) \tag{20}$$

Source: Wikipedia contributors [5].

Example 0.3 (NPV of a COVID-19 test)**.** For the COVID-19 test of Example 0.1 and Example 0.2, used in the population of Example 0.3, the probability of a negative test for a person with COVID-19 is, by the complement rule and Theorem 0.1:

$$\begin{aligned}
\Pr(- \mid D) &= \Pr(\neg + \mid D) && \text{(negative classification is the complement of positive)} \\
&= 1 - \Pr(+ \mid D) && \text{(complement rule, conditional on } D\text{)} \\
&= 1 - 0.99 && \text{(substitute the sensitivity)} \\
&= 0.01 && \text{(subtract)}
\end{aligned} \tag{21}$$

By Bayes' theorem, with the denominator expanded by the law of total probability over the partition $\{D, \neg D\}$, the negative predictive value (Definition 0.2) is:

$$
\begin{aligned}
\Pr(\neg D \mid -) &= \frac{\Pr(- \mid \neg D) \cdot \Pr(\neg D)}{\Pr(-)} && \text{(Bayes' theorem)} \\
&= \frac{\Pr(- \mid \neg D) \cdot \Pr(\neg D)}{\Pr(- \mid \neg D) \cdot \Pr(\neg D) + \Pr(- \mid D) \cdot \Pr(D)} && \text{(law of total probability)} \\
&= \frac{0.99 \cdot 0.93}{0.99 \cdot 0.93 + 0.01 \cdot 0.07} && \text{(substitute the given values)} \\
&= \frac{0.9207}{0.9207 + 0.0007} && \text{(multiply each term in the numerator and denomi} \\
&= \frac{0.9207}{0.9214} && \text{(add the denominator's two terms)} \\
&\approx 0.9992 && \text{(divide)}
\end{aligned}
\tag{22}
$$

A negative result from this test is very reliable: only about 8 in 10,000 people who test negative have COVID-19.

PPV in terms of sensitivity, specificity, and prevalence

Exercise 0.1 (Probability of a positive test)**.** Derive an expression for the probability $\Pr(+)$ that a diagnostic test is positive in terms of:

- the test's sensitivity;
- the test's specificity;
- the prevalence of the disease.

Solution

Solution 0.1. By the law of total probability over the partition $\{D, \neg D\}$, the complement rule, and Theorem 0.1:

$$
\begin{aligned}
\Pr(+) &= \Pr(+ \mid D) \cdot \Pr(D) + \Pr(+ \mid \neg D) \cdot \Pr(\neg D) && \text{(law of total probability)} \\
&= \text{sensitivity} \cdot \Pr(D) + \Pr(+ \mid \neg D) \cdot \Pr(\neg D) && \text{(definition of sensitivity)} \\
&= \text{sensitivity} \cdot \text{prevalence} + \Pr(+ \mid \neg D) \cdot \Pr(\neg D) && \text{(definition of prevalence)} \\
&= \text{sensitivity} \cdot \text{prevalence} + (1 - \Pr(\neg + \mid \neg D)) \cdot \Pr(\neg D) && \text{(complement rule, conditional on } \neg D) \\
&= \text{sensitivity} \cdot \text{prevalence} + (1 - \Pr(- \mid \neg D)) \cdot \Pr(\neg D) && \text{(negative classification is the compleme} \\
&= \text{sensitivity} \cdot \text{prevalence} + (1 - \text{specificity}) \cdot \Pr(\neg D) && \text{(definition of specificity)} \\
&= \text{sensitivity} \cdot \text{prevalence} + (1 - \text{specificity}) \cdot (1 - \Pr(D)) && \text{(complement rule)} \\
&= \text{sensitivity} \cdot \text{prevalence} + (1 - \text{specificity}) \cdot (1 - \text{prevalence}) && \text{(definition of prevalence)}
\end{aligned}
\tag{23}
$$

Theorem 0.1 (Probability of a positive test)**.** *Let sens be the sensitivity of a diagnostic test, spec its specificity, and π the prevalence of the disease. Then the probability* $\Pr(+)$ *that the test is positive is:*

$$
\Pr(+) = sens \cdot \pi + (1 - spec) \cdot (1 - \pi) \tag{24}
$$

Proof

Proof. This is the solution to Exercise 0.1, worked in Solution 0.1. □

Exercise 0.2 (PPV from Bayes' theorem)**.** Derive an expression for the positive predictive value of a diagnostic test in terms of:

- its sensitivity;
- its specificity;
- the prevalence of the disease.

Solution

Solution 0.2. Abbreviate the sensitivity as sens, the specificity as spec, and the prevalence as π. By Bayes' theorem, with the probability of a positive test from Exercise 0.1:

$$
\begin{aligned}
\text{PPV} &\overset{\text{def}}{=} \Pr(D \mid +) && \text{(definition of PPV)} \\
&= \frac{\Pr(+ \mid D) \cdot \Pr(D)}{\Pr(+)} && \text{(Bayes' theorem)} \\
&= \frac{\text{sens} \cdot \Pr(D)}{\Pr(+)} && \text{(definition of sensitivity)} \\
&= \frac{\text{sens} \cdot \pi}{\Pr(+)} && \text{(definition of prevalence)} \\
&= \frac{\text{sens} \cdot \pi}{\text{sens} \cdot \pi + (1 - \text{spec}) \cdot (1 - \pi)} && \text{(probability of a positive test)}
\end{aligned} \tag{25}
$$

Theorem 0.2 (PPV from Bayes' theorem)**.** *Let sens be the sensitivity of a diagnostic test, spec its specificity, and π the prevalence of the disease. If the probability of a positive test is positive, $sens \cdot \pi + (1 - spec) \cdot (1 - \pi) > 0$, then the positive predictive value of the test is:*

$$
PPV = \frac{sens \cdot \pi}{sens \cdot \pi + (1 - spec) \cdot (1 - \pi)} \tag{26}
$$

Source: Wikipedia contributors [1].

Proof

Proof. This is the solution to Exercise 0.2, worked in Solution 0.2, which builds on Theorem 0.1. □

Exercise 0.3 (PPV in terms of two ratios)**.** Rewrite the expression for the positive predictive value from Theorem 0.2 so that the sensitivity, specificity, and prevalence appear only through the ratio $\frac{1-\text{spec}}{\text{sens}}$ of the false positive rate to the sensitivity and the ratio $\frac{1-\pi}{\pi}$ of non-diseased to diseased people in the population.

Solution

Solution 0.3. Write $a \overset{\text{def}}{=} \text{sens} \cdot \pi$ and $b \overset{\text{def}}{=} (1 - \text{spec}) \cdot (1 - \pi)$, so that Theorem 0.2 reads $\text{PPV} = \frac{a}{a+b}$, with $a > 0$ whenever the sensitivity and prevalence are positive. Then:

$$
\begin{aligned}
\text{PPV} &= \frac{a}{a+b} && \text{(PPV from Bayes' theorem)} \\
&= \frac{1}{\left(\frac{a+b}{a}\right)} && \text{(a fraction equals one over its reciprocal, since } a > 0\text{)} \\
&= \frac{1}{\frac{a}{a} + \frac{b}{a}} && \text{(distribute the division over the sum in the numerator)} \\
&= \frac{1}{1 + \frac{b}{a}} && \left(\tfrac{a}{a} = 1\right) \\
&= \frac{1}{1 + \frac{(1-\text{spec})\cdot(1-\pi)}{\text{sens}\cdot\pi}} && \text{(substitute the definitions of } a \text{ and } b\text{)} \\
&= \frac{1}{1 + \frac{1-\text{spec}}{\text{sens}} \cdot \frac{1-\pi}{\pi}} && \text{(a quotient of products is the product of the quotients)}
\end{aligned}
\tag{27}
$$

Theorem 0.3 (PPV in terms of sensitivity, specificity, and prevalence)**.** *The positive predictive value of a diagnostic test with positive sensitivity, used in a population with positive prevalence, depends on its sensitivity, its specificity, and the prevalence of the disease only through two ratios:*

- *the ratio of the false positive rate to the sensitivity;*
- *the ratio of non-diseased to diseased people in the population.*

Specifically:

$$
PPV = \frac{1}{1 + \frac{1-spec}{sens} \cdot \frac{1-\pi}{\pi}} \tag{28}
$$

Proof

Proof. This result is the solution to Exercise 0.3, worked in Solution 0.3, which builds on Theorem 0.2 and Theorem 0.1. □

Example 0.1 (PPV of a COVID-19 test, from the two ratios)**.** For the COVID-19 test of Example 0.1 and Example 0.2, used in the population of Example 0.3, Theorem 0.3 gives:

$$
\begin{aligned}
\mathrm{PPV} &= \frac{1}{1+\frac{1-0.99}{0.99}\cdot\frac{1-0.07}{0.07}} && \text{(PPV in terms of the two ratios, with the given values)}\\
&= \frac{1}{1+\frac{0.01}{0.99}\cdot\frac{0.93}{0.07}} && \text{(subtract in each numerator)}\\
&\approx \frac{1}{1+0.0101\cdot\frac{0.93}{0.07}} && \text{(divide 0.01/0.99)}\\
&\approx \frac{1}{1+0.0101\cdot 13.29} && \text{(divide 0.93/0.07)}\\
&\approx \frac{1}{1+0.134} && \text{(multiply)}\\
&= \frac{1}{1.134} && \text{(add)}\\
&\approx 0.88 && \text{(divide)}
\end{aligned}
\tag{29}
$$

This value matches Example 0.1.

If the same test were used in a population with prevalence 0.1% instead, and its sensitivity and specificity were still 99% in that population, the ratio of non-diseased to diseased people would be $0.999/0.001 = 999$, and the PPV would fall to

$$
\begin{aligned}
\mathrm{PPV} &= \frac{1}{1+\frac{0.01}{0.99}\cdot 999} && \text{(PPV in terms of the two ratios, with the new prevalence)}\\
&\approx \frac{1}{1+0.0101\cdot 999} && \text{(divide)}\\
&\approx \frac{1}{1+10.09} && \text{(multiply)}\\
&= \frac{1}{11.09} && \text{(add)}\\
&\approx 0.09 && \text{(divide)}
\end{aligned}
\tag{30}
$$

so about 9 in 10 positive results would be false positives.

Remark (The PPV depends on the population, not only on the test)

Remark 0.1 (The PPV depends on the population, not only on the test). The sensitivity and specificity describe the test's performance in a specified setting, and may differ across settings (for example, with the case mix of the people tested or the threshold used to assign a positive classification), but the ratio $\frac{1-\pi}{\pi}$ in Theorem 0.3 describes the population being tested. When the disease is rare, that ratio is large, so even a small false positive

rate can make the PPV low, as Example 0.1 shows.

Agreement between two classifiers

So far, we have compared a test with the true disease status. The same tools apply whenever we compare two classifiers that label the same observations: for example, a new diagnostic test against a reference test, or an automated grader (an "agent") against a human grader. We call the classifier being evaluated the **agent** and the one we compare it with the **reference**.

Definition 0.1 (Agreement table)**.** For n observations that an agent and a reference each classify as positive or negative, the **agreement table** counts the four combinations:

	Reference positive	Reference negative	Total
Agent positive	a	b	$a+b$
Agent negative	c	d	$c+d$
Total	$a+c$	$b+d$	n

The cells a and d count agreements, the cells b and c count disagreements, and the four cells together count every observation:

$$n \stackrel{\text{def}}{=} a+b+c+d \tag{31}$$

Source: Wikipedia contributors [6].

Example 0.1 (Agreement table for a COVID-19 test)**.** Apply the COVID-19 test of Example 0.1 and Example 0.2 to $n = 10{,}000$ people from the population of Example 0.3, with the true disease status as the reference. With 7% prevalence, $0.07 \times 10{,}000 = 700$ people have COVID-19 and $10{,}000 - 700 = 9{,}300$ do not. With 99% sensitivity, the test is positive for $0.99 \times 700 = 693$ of the 700; with 99% specificity, it is negative for $0.99 \times 9{,}300 = 9{,}207$ of the 9,300. So the test is positive for $9{,}300 - 9{,}207 = 93$ people without COVID-19 and negative for $700 - 693 = 7$ people with it. The agreement table (Definition 0.1) is:

	COVID-19	No COVID-19	Total
Test positive	$a = 693$	$b = 93$	786
Test negative	$c = 7$	$d = 9{,}207$	9,214
Total	700	9,300	10,000

```r
agree_covid <- c(a = 693, b = 93, c = 7, d = 9207)
agree_covid
#>    a    b    c    d
#>  693   93    7 9207
```

Agreement proportions

Definition 0.1 (Overall agreement)**.** Let:

- A be the event that the agent makes a positive classification for an observation drawn from a population, with complement $\neg A$.
- R be the event that the reference makes a positive classification for the same observation, with complement $\neg R$.

The **overall agreement** of the agent and the reference on that population is the probability that they assign the same class to the observation:

$$p_o \stackrel{\text{def}}{=} \Pr(A \cap R) + \Pr(\neg A \cap \neg R) \tag{32}$$

Source: Cohen [7].

Example 0.1 (Overall agreement for the COVID-19 test)**.** Apply the COVID-19 test of Example 0.1 and Example 0.2 to the population of Example 0.3, with the test as the agent ($A = +$) and the disease status as the reference ($R = D$). By the multiplication rule for conditional probability, with sensitivity $\Pr(+ \mid D) = 0.99$, specificity $\Pr(- \mid \neg D) = 0.99$, prevalence $\Pr(D) = 0.07$, and $\Pr(\neg D) = 0.93$ (Example 0.3), the probability that a person has COVID-19 and tests positive is:

$$\begin{aligned} \Pr(+ \cap D) &= \Pr(+ \mid D) \cdot \Pr(D) && \text{(multiplication rule for conditional probability)} \\ &= 0.99 \cdot 0.07 && \text{(substitute the sensitivity and prevalence)} \\ &= 0.0693 && \text{(multiply)} \end{aligned} \tag{33}$$

and the probability that a person does not have COVID-19 and tests negative is:

$$\begin{aligned} \Pr(- \cap \neg D) &= \Pr(- \mid \neg D) \cdot \Pr(\neg D) && \text{(multiplication rule for conditional probability)} \\ &= 0.99 \cdot 0.93 && \text{(substitute the specificity and } \Pr(\neg D)\text{)} \\ &= 0.9207 && \text{(multiply)} \end{aligned} \tag{34}$$

Then the overall agreement is:

$$\begin{aligned} p_o &= \Pr(+ \cap D) + \Pr(- \cap \neg D) && \text{(definition of } p_o\text{)} \\ &= 0.0693 + 0.9207 && \text{(substitute the two joint probabilities)} \\ &= 0.99 && \text{(add)} \end{aligned} \tag{35}$$

Definition 0.2 (Estimated overall agreement)**.** The **estimated overall agreement** of an agreement table is the sample proportion of observations on which the agent and the reference assign the same class, which estimates the overall agreement p_o:

$$\hat{p}_o \stackrel{\text{def}}{=} \frac{a + d}{n} \tag{36}$$

Source: Cohen [7].

Example 0.2 (Estimated overall agreement for the COVID-19 test)**.** For the agreement table of Example 0.1, the estimated overall agreement is:

$$\begin{aligned} \hat{p}_o &= \frac{693 + 9{,}207}{10{,}000} && \text{(definition of } \hat{p}_o\text{, with the table's counts)} \\ &= \frac{9{,}900}{10{,}000} && \text{(add)} \\ &= 0.99 && \text{(divide)} \end{aligned} \tag{37}$$

This estimate equals the overall agreement $p_o = 0.99$ of Example 0.1, because the table's cell sample proportions $a/n = 0.0693$ and $d/n = 0.9207$ equal the joint probabilities $\Pr(+ \cap D)$ and $\Pr(- \cap \neg D)$.

Definition 0.3 (Positive agreement)**.** The **positive agreement** (also called the *average positive agreement*) of an agreement table (Definition 0.1) with $2a + b + c > 0$ is the number of positive classifications the agent and the reference share, counted once for each classifier, as a proportion of all their positive classifications:

$$p_{\text{pos}} \stackrel{\text{def}}{=} \frac{2a}{2a + b + c} \tag{38}$$

Source: Cicchetti and Feinstein [8].

Example 0.3 (Positive agreement for the COVID-19 test)**.** For the agreement table of Example 0.1, the positive agreement (Definition 0.3) is:

$$
\begin{aligned}
p_{\text{pos}} &= \frac{2 \times 693}{2 \times 693 + 93 + 7} && \text{(definition of } p_{\text{pos}}\text{, with the table's counts)} \\
&= \frac{1{,}386}{1{,}386 + 93 + 7} && \text{(multiply)} \\
&= \frac{1{,}386}{1{,}486} && \text{(add)} \\
&\approx 0.933 && \text{(divide)}
\end{aligned}
\tag{39}
$$

The positive agreement is lower than the overall agreement (Example 0.1), because the 93 false positives are large relative to the 693 true positives.

Definition 0.4 (Negative agreement)**.** The **negative agreement** (also called the *average negative agreement*) of an agreement table (Definition 0.1) with $2d + b + c > 0$ is the number of negative classifications the agent and the reference share, counted once for each classifier, as a proportion of all their negative classifications:

$$
p_{\text{neg}} \stackrel{\text{def}}{=} \frac{2d}{2d + b + c} \tag{40}
$$

Source: Cicchetti and Feinstein [8].

Example 0.4 (Negative agreement for the COVID-19 test)**.** For the agreement table of Example 0.1, the negative agreement (Definition 0.4) is:

$$
\begin{aligned}
p_{\text{neg}} &= \frac{2 \times 9{,}207}{2 \times 9{,}207 + 93 + 7} && \text{(definition of } p_{\text{neg}}\text{, with the table's counts)} \\
&= \frac{18{,}414}{18{,}414 + 93 + 7} && \text{(multiply)} \\
&= \frac{18{,}414}{18{,}514} && \text{(add)} \\
&\approx 0.995 && \text{(divide)}
\end{aligned}
\tag{41}
$$

The negative agreement is even higher than the overall agreement (Example 0.1), because the 100 disagreements are small relative to the 9,207 shared negative calls.

Definition 0.5 (Agent-only rate)**.** The **agent-only rate** of an agreement table (Definition 0.1) is the proportion of observations that the agent assigns a positive classification and the reference assigns a negative classification:

$$
p_{\text{agent}} \stackrel{\text{def}}{=} \frac{b}{n} \tag{42}
$$

Example 0.5 (Agent-only rate for the COVID-19 test)**.** For the agreement table of Example 0.1, the agent-only rate (Definition 0.5) is:

$$\begin{aligned} p_{\text{agent}} &= \frac{93}{10{,}000} && \text{(definition of } p_{\text{agent}}\text{, with the table's counts)} \\ &= 0.0093 && \text{(divide)} \end{aligned} \tag{43}$$

so about 9 of every 1,000 people tested are called positive by the test and negative by the reference.

Definition 0.6 (Test measures estimated from an agreement table)**.** Treating the reference as the truth, an agreement table (Definition 0.1) estimates four measures of the agent:

- its sensitivity;
- its specificity;
- its positive predictive value;
- its negative predictive value.

Each estimate is the matching proportion of the table's cells, defined when its denominator is positive:

$$\begin{aligned} \widehat{\text{sensitivity}} &\stackrel{\text{def}}{=} \frac{a}{a+c}, \\ \widehat{\text{specificity}} &\stackrel{\text{def}}{=} \frac{d}{b+d}, \\ \widehat{\text{PPV}} &\stackrel{\text{def}}{=} \frac{a}{a+b}, \\ \widehat{\text{NPV}} &\stackrel{\text{def}}{=} \frac{d}{c+d}. \end{aligned} \tag{44}$$

Example 0.6 (Test measures for the COVID-19 test)**.** For the agreement table of Example 0.1, Definition 0.6 gives:

```r
test_measures <- with(as.list(agree_covid), c(
  sensitivity = a / (a + c),
  specificity = d / (b + d),
  ppv = a / (a + b),
  npv = d / (c + d)
))
round(test_measures, 4)
#> sensitivity specificity         ppv         npv
#>      0.9900      0.9900      0.8817      0.9992
```

These estimates match the test's sensitivity (0.99) and specificity (0.99), its positive predictive value (about 0.88, Example 0.1), and its negative predictive value (about 0.999, Example 0.3), because the table's counts are the expected counts for those values.

Positive and negative agreement as harmonic means

Definition 0.1 (Harmonic mean)**.** The **harmonic mean** of two positive numbers x and y is the reciprocal of the average of their reciprocals:

$$H(x, y) \stackrel{\text{def}}{=} \frac{2}{\frac{1}{x} + \frac{1}{y}} \tag{45}$$

Source: Wikipedia contributors [9].

Example 0.1 (Harmonic mean of 0.5 and 1)**.** By Definition 0.1:

$$\begin{aligned} H(0.5, 1) &= \frac{2}{\frac{1}{0.5} + \frac{1}{1}} && \text{(definition of } H\text{)} \\ &= \frac{2}{2 + 1} && \text{(take each reciprocal)} \\ &= \frac{2}{3} && \text{(add)} \\ &\approx 0.667 && \text{(divide)} \end{aligned} \tag{46}$$

The ordinary average of 0.5 and 1 is 0.75; the harmonic mean sits closer to the smaller number.

Exercise 0.1 (Positive agreement from PPV and sensitivity)**.** For an agreement table (Definition 0.1) with $a > 0$, write the harmonic mean of the estimated positive predictive

value and sensitivity (Definition 0.6) in terms of the table's cells.

Solution

Solution 0.1. Because $a > 0$, both $a + b$ and $a + c$ are positive, so both estimates are defined and positive:

$$
\begin{aligned}
H\left(\widehat{\text{PPV}}, \widehat{\text{sensitivity}}\right) &= \frac{2}{\frac{1}{\widehat{\text{PPV}}} + \frac{1}{\widehat{\text{sensitivity}}}} && \text{(definition of } H\text{)} \\
&= \frac{2}{\frac{1}{a/(a+b)} + \frac{1}{\widehat{\text{sensitivity}}}} && \text{(definition of } \widehat{\text{PPV}}\text{)} \\
&= \frac{2}{\frac{1}{a/(a+b)} + \frac{1}{a/(a+c)}} && \text{(definition of } \widehat{\text{sensitivity}}\text{)} \\
&= \frac{2}{\frac{a+b}{a} + \frac{1}{a/(a+c)}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{a+b}{a} + \frac{a+c}{a}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{(a+b)+(a+c)}{a}} && \text{(add fractions with a common denominator)} \\
&= \frac{2}{\frac{a+b+a+c}{a}} && \text{(remove parentheses)} \\
&= \frac{2}{\frac{a+a+b+c}{a}} && \text{(reorder terms)} \\
&= \frac{2}{\frac{2a+b+c}{a}} && (a + a = 2a) \\
&= 2 \cdot \frac{a}{2a+b+c} && \text{(dividing by a fraction multiplies by its reciprocal)} \\
&= \frac{2a}{2a+b+c} && \text{(multiply)}
\end{aligned}
\tag{47}
$$

Theorem 0.1 (Positive agreement is a harmonic mean)**.** *For an agreement table with $a > 0$, the positive agreement (Definition 0.3) is the harmonic mean (Definition 0.1) of the estimated positive predictive value and sensitivity (Definition 0.6):*

$$p_{pos} = H\left(\widehat{PPV}, \widehat{sensitivity}\right) \tag{48}$$

Proof

Proof. By Exercise 0.1, the harmonic mean equals $2a/(2a+b+c)$, which is p_{pos} by Definition 0.3. □

Example 0.2 (Positive agreement of the COVID-19 test as a harmonic mean)**.** For the agreement table of Example 0.1, the estimated positive predictive value is 693/786 and the estimated sensitivity is 693/700 (Example 0.6). Theorem 0.1 gives:

$$\begin{aligned}
p_{\text{pos}} &= \frac{2}{\frac{1}{693/786} + \frac{1}{693/700}} && \text{(harmonic mean of the two estimates)} \\
&= \frac{2}{\frac{786}{693} + \frac{1}{693/700}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{786}{693} + \frac{700}{693}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{1{,}486}{693}} && \text{(add fractions with a common denominator)} \\
&= 2 \cdot \frac{693}{1{,}486} && \text{(dividing by a fraction multiplies by its reciprocal)} \\
&= \frac{1{,}386}{1{,}486} && \text{(multiply)} \\
&\approx 0.933 && \text{(divide)}
\end{aligned} \tag{49}$$

which matches Example 0.3. The lower positive predictive value (about 0.88) pulls the positive agreement down from the sensitivity (0.99), as in Example 0.1.

Exercise 0.2 (Negative agreement from NPV and specificity)**.** For an agreement table (Definition 0.1) with $d > 0$, write the harmonic mean of the estimated negative predictive value and specificity (Definition 0.6) in terms of the table's cells.

Solution

Solution 0.2. Because $d > 0$, both $c + d$ and $b + d$ are positive, so both estimates are defined and positive:

$$
\begin{aligned}
H\left(\widehat{\text{NPV}}, \widehat{\text{specificity}}\right) &= \frac{2}{\frac{1}{\widehat{\text{NPV}}} + \frac{1}{\widehat{\text{specificity}}}} && \text{(definition of } H\text{)} \\
&= \frac{2}{\frac{1}{d/(c+d)} + \frac{1}{\widehat{\text{specificity}}}} && \text{(definition of } \widehat{\text{NPV}}\text{)} \\
&= \frac{2}{\frac{1}{d/(c+d)} + \frac{1}{d/(b+d)}} && \text{(definition of } \widehat{\text{specificity}}\text{)} \\
&= \frac{2}{\frac{c+d}{d} + \frac{1}{d/(b+d)}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{c+d}{d} + \frac{b+d}{d}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{(c+d)+(b+d)}{d}} && \text{(add fractions with a common denominator)} \\
&= \frac{2}{\frac{c+d+b+d}{d}} && \text{(remove parentheses)} \\
&= \frac{2}{\frac{d+d+b+c}{d}} && \text{(reorder terms)} \\
&= \frac{2}{\frac{2d+b+c}{d}} && (d+d=2d) \\
&= 2 \cdot \frac{d}{2d+b+c} && \text{(dividing by a fraction multiplies by its reciprocal)} \\
&= \frac{2d}{2d+b+c} && \text{(multiply)}
\end{aligned}
\tag{50}
$$

Theorem 0.2 (Negative agreement is a harmonic mean)**.** *For an agreement table with $d > 0$, the negative agreement (Definition 0.4) is the harmonic mean (Definition 0.1) of the estimated negative predictive value and specificity (Definition 0.6):*

$$
p_{neg} = H\left(\widehat{NPV}, \widehat{specificity}\right) \tag{51}
$$

Proof

Proof. By Exercise 0.2, the harmonic mean equals $2d/(2d + b + c)$, which is p_{neg} by Definition 0.4. □

Example 0.3 (Negative agreement of the COVID-19 test as a harmonic mean)**.** For the agreement table of Example 0.1, the estimated negative predictive value is 9,207/9,214 and the estimated specificity is 9,207/9,300 (Example 0.6). Theorem 0.2 gives:

$$\begin{aligned}
p_{\text{neg}} &= \frac{2}{\frac{1}{9{,}207/9{,}214} + \frac{1}{9{,}207/9{,}300}} && \text{(harmonic mean of the two estimates)} \\
&= \frac{2}{\frac{9{,}214}{9{,}207} + \frac{1}{9{,}207/9{,}300}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{9{,}214}{9{,}207} + \frac{9{,}300}{9{,}207}} && \text{(reciprocal of a fraction)} \\
&= \frac{2}{\frac{18{,}514}{9{,}207}} && \text{(add fractions with a common denominator)} \\
&= 2 \cdot \frac{9{,}207}{18{,}514} && \text{(dividing by a fraction multiplies by its reciprocal)} \\
&= \frac{18{,}414}{18{,}514} && \text{(multiply)} \\
&\approx 0.995 && \text{(divide)}
\end{aligned} \tag{52}$$

which matches Example 0.4.

Cohen's kappa

Definition 0.1 (Chance agreement)**.** Let:

- A be the event that the agent makes a positive classification for an observation drawn from a population, with complement $\neg A$.
- R be the event that the reference makes a positive classification for the same observation, with complement $\neg R$.

The **chance agreement** of the agent and the reference on that population is the overall agreement they would have if A and R were statistically independent with their marginal probabilities $\Pr(A)$ and $\Pr(R)$:

$$p_e \stackrel{\text{def}}{=} \Pr(A) \cdot \Pr(R) + \Pr(\neg A) \cdot \Pr(\neg R) \tag{53}$$

Source: Cohen [7].

Example 0.1 (Chance agreement for the COVID-19 test)**.** Apply the COVID-19 test of Example 0.1 and Example 0.2 to the population of Example 0.3, with the test as the agent ($A = +$) and the disease status as the reference ($R = D$). By Theorem 0.1, with sensitivity sens $= 0.99$, specificity spec $= 0.99$, and prevalence $\pi = \Pr(D) = 0.07$, the probability of a positive test is:

$$\begin{aligned}
\Pr(+) &= \text{sens} \cdot \pi + (1 - \text{spec}) \cdot (1 - \pi) && \text{(probability of a positive test)} \\
&= 0.99 \cdot 0.07 + (1 - 0.99) \cdot (1 - 0.07) && \text{(substitute the given values)} \\
&= 0.99 \cdot 0.07 + 0.01 \cdot 0.93 && \text{(subtract inside each pair of parentheses)} \\
&= 0.0693 + 0.0093 && \text{(multiply each product)} \\
&= 0.0786 && \text{(add)}
\end{aligned} \tag{54}$$

By the complement rule and Theorem 0.1, the probability of a negative test is:

$$\begin{aligned}
\Pr(-) &= \Pr(\neg +) && \text{(negative classification is the complement of positive)} \\
&= 1 - \Pr(+) && \text{(complement rule)} \\
&= 1 - 0.0786 && \text{(substitute } \Pr(+)\text{)} \\
&= 0.9214 && \text{(subtract)}
\end{aligned} \tag{55}$$

With $\Pr(D) = 0.07$ and $\Pr(\neg D) = 0.93$ (Example 0.3), the chance agreement is:

$$\begin{aligned}
p_e &= \Pr(+) \cdot \Pr(D) + \Pr(-) \cdot \Pr(\neg D) && \text{(definition of } p_e\text{)} \\
&= 0.0786 \cdot 0.07 + 0.9214 \cdot 0.93 && \text{(substitute the marginal probabilities)} \\
&= 0.005502 + 0.856902 && \text{(multiply each product)} \\
&= 0.862404 && \text{(add)}
\end{aligned} \tag{56}$$

Because nearly everyone tests negative and does not have COVID-19, two independent classifiers with $\Pr(+) = 0.0786$ and $\Pr(D) = 0.07$ would agree on about 86% of people by chance alone.

Definition 0.2 (Estimated chance agreement)**.** The **estimated chance agreement** of an agreement table estimates the chance agreement p_e by replacing the marginal probabilities $\Pr(A)$, $\Pr(R)$, $\Pr(\neg A)$, and $\Pr(\neg R)$ with the table's marginal sample proportions $(a + b)/n$, $(a + c)/n$, $(c + d)/n$, and $(b + d)/n$:

$$\hat{p}_e \stackrel{\text{def}}{=} \frac{(a + b)(a + c) + (c + d)(b + d)}{n^2} \tag{57}$$

Source: Cohen [7].

Example 0.2 (Estimated chance agreement for the COVID-19 test)**.** For the agreement table of Example 0.1, the estimated chance agreement is:

$$\begin{aligned}
\hat{p}_e &= \frac{786 \times 700 + 9{,}214 \times 9{,}300}{10{,}000^2} && \text{(definition of } \hat{p}_e\text{, with the table's counts)} \\
&= \frac{550{,}200 + 85{,}690{,}200}{10{,}000^2} && \text{(multiply)} \\
&= \frac{86{,}240{,}400}{10{,}000^2} && \text{(add)} \\
&= \frac{86{,}240{,}400}{100{,}000{,}000} && \text{(square)} \\
&= 0.862404 && \text{(divide)}
\end{aligned} \tag{58}$$

This estimate equals the chance agreement $p_e = 0.862404$ of Example 0.1, because the table's marginal sample proportions equal the marginal probabilities $\Pr(+)$, $\Pr(D)$, $\Pr(-)$, and $\Pr(\neg D)$.

Definition 0.3 (Cohen's kappa)**.** For an agent and a reference with chance agreement $p_e < 1$, **Cohen's kappa** is the agreement beyond chance, as a fraction of the largest possible agreement beyond chance, built from the overall agreement p_o and the chance agreement p_e:

$$\kappa \stackrel{\text{def}}{=} \frac{p_o - p_e}{1 - p_e} \tag{59}$$

Source: Cohen [7]; Wikipedia contributors [10].

Example 0.3 (Cohen's kappa for the COVID-19 test)**.** For the COVID-19 test of Example 0.1 and Example 0.2 in the population of Example 0.3, the overall agreement is $p_o = 0.99$ (Example 0.1) and the chance agreement is $p_e = 0.862404$ (Example 0.1), so Cohen's kappa is:

$$\begin{aligned}
\kappa &= \frac{0.99 - 0.862404}{1 - 0.862404} && \text{(definition of } \kappa\text{)} \\
&= \frac{0.127596}{1 - 0.862404} && \text{(subtract in the numerator)} \\
&= \frac{0.127596}{0.137596} && \text{(subtract in the denominator)} \\
&\approx 0.927 && \text{(divide)}
\end{aligned} \tag{60}$$

Definition 0.4 (Estimated Cohen's kappa)**.** For an agreement table with estimated chance agreement $\hat{p}_e < 1$, the **estimated Cohen's kappa** estimates Cohen's kappa κ from the estimated overall agreement $\hat{p}_o$ and the estimated chance agreement $\hat{p}_e$:

$$\hat{\kappa} \stackrel{\text{def}}{=} \frac{\hat{p}_o - \hat{p}_e}{1 - \hat{p}_e} \tag{61}$$

Source: Cohen [7]; Wikipedia contributors [10].

Example 0.4 (Estimated Cohen's kappa for the COVID-19 test)**.** For the agreement table of Example 0.1, the estimated overall agreement is $\hat{p}_o = 0.99$ (Example 0.2) and the estimated chance agreement is $\hat{p}_e = 0.862404$ (Example 0.2), so the estimated Cohen's kappa is:

$$\begin{aligned} \hat{\kappa} &= \frac{0.99 - 0.862404}{1 - 0.862404} && \text{(definition of } \hat{\kappa}\text{)} \\ &= \frac{0.127596}{1 - 0.862404} && \text{(subtract in the numerator)} \\ &= \frac{0.127596}{0.137596} && \text{(subtract in the denominator)} \\ &\approx 0.927 && \text{(divide)} \end{aligned} \tag{62}$$

```
kappa_from_table <- function(a, b, c, d) {
  n <- a + b + c + d
  p_o_hat <- (a + d) / n
  p_e_hat <- ((a + b) * (a + c) + (c + d) * (b + d)) / n^2
  (p_o_hat - p_e_hat) / (1 - p_e_hat)
}
do.call(kappa_from_table, as.list(agree_covid))
#> [1] 0.927323
```

This estimate equals Cohen's kappa $\kappa \approx 0.927$ of Example 0.3, because $\hat{p}_o = p_o$ and $\hat{p}_e = p_e$.

Example 0.5 (High agreement but low kappa)**.** Now apply the same test to $n = 100{,}000$ people in a population where the prevalence is only 0.5%. Then:

- $0.005 \times 100{,}000 = 500$ people have COVID-19 and $100{,}000 - 500 = 99{,}500$ do not;
- with 99% sensitivity, the test is positive for $a = 0.99 \times 500 = 495$ people with COVID-19 and negative for $c = 500 - 495 = 5$;
- with 99% specificity, the false positive rate is $1 - 0.99 = 0.01$, so the test is positive for $b = 0.01 \times 99{,}500 = 995$ people without COVID-19 and negative for $d =$

$$99{,}500 - 995 = 98{,}505.$$

```
agree_rare <- c(a = 495, b = 995, c = 5, d = 98505)
rare_metrics <- with(as.list(agree_rare), c(
  overall = (a + d) / (a + b + c + d),
  positive = 2 * a / (2 * a + b + c),
  negative = 2 * d / (2 * d + b + c),
  kappa = kappa_from_table(a, b, c, d)
))
round(rare_metrics, 4)
#>  overall positive negative    kappa
#>   0.9900   0.4975   0.9949   0.4937
```

The estimated overall agreement is still 0.99, but the estimated Cohen's kappa drops to about 0.49 and the positive agreement to about 0.50, while the negative agreement stays at about 0.995. The positive and negative agreements show where the classifiers disagree: the test and the reference agree on negatives and disagree on about half of the positive calls.
Source: Feinstein and Cicchetti [11] (the first paradox).

Remark (Kappa depends on prevalence)

Remark 0.1 (Kappa depends on prevalence). When both classifiers call one class rarely, the chance agreement p_e is close to 1, so the same overall agreement leaves little room for agreement beyond chance, and Cohen's kappa can be low even when the overall agreement is high, as Example 0.5 shows. Cicchetti and Feinstein [8] recommend reporting the positive agreement and negative agreement alongside Cohen's kappa, so that poor agreement on the rarer class is not hidden.
Source: Feinstein and Cicchetti [11]; Cicchetti and Feinstein [8].

References

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

[2] Wikipedia contributors, "Statistical classification — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Statistical_classification

[3] Wikipedia contributors, "Sensitivity and specificity — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

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

[5] Wikipedia contributors, "Positive and negative predictive values — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values

[6] Wikipedia contributors, "Confusion matrix — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Confusion_matrix

[7] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[8] D. V. Cicchetti and A. R. Feinstein, "High agreement but low kappa: II. Resolving the paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 551–558, 1990, doi: 10.1016/0895-4356(90)90159-M.

[9] Wikipedia contributors, "Harmonic mean — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Harmonic_mean

[10] Wikipedia contributors, "Cohen's kappa — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Cohen%27s_kappa

[11] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa: I. The problems of two paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 543–549, 1990, doi: 10.1016/0895-4356(90)90158-L.