# Section 2.0 Conditional Probability

Conditional probability is a critical concept within probability theory. In essence it means that the probability of an event occurring can be influenced by another piece of information. We have already discussed the fact that many people have poor intuitions when probability, uncertainty, and randomness are involved. This is especially true with regard to the concepts of conditional probability.

The basic notation for conditional probability is shown below.

$$P(B \mid A)$$

This is read as the probability of *B* occurring given that *A* is known or has occurred. Knowing that event *A* has occurred sometimes allows you to calculate a different probability of event *B* occurring, simply because you now have more information.

## 2.0.1 Mortality Tables

One classic example of conditional probability involves the probability distribution of future life for people of different ages. The Social Security Administration (SSA) estimate the average lifetime of a male born in 2011 was a little more than 76 years. Poor intuition leads many people to imagine a male who has reached 80 years of age is living on borrowed time. In fact, the SSA estimates the probability of an 80 year old dying within his next year is 0.06. On average, he will live more than eight additional years. Table 2.0 presents actuarial data maintained by the Social Security Administration of the US. The table starts with a cohort of a 100,000 of each gender. It presents a forecast of the number of people in each cohort who will live to a specific age. This table stops at the age of 20. https://www.ssa.gov/oact/STATS/table4c6.html

| Exact | Number of Lives | | | | Exact | Number of Lives | | |
|---|---|---|---|---|---|---|---|---|
| age | Males | Females | Totals | | age | Males | Females | Totals |
| 0 | 100,000 | 100,000 | 200,000 | | | | | |
| 1 | 99,343 | 99,449 | 198,792 | | 11 | 99,155 | 99,296 | 198,451 |
| 2 | 99,299 | 99,411 | 198,710 | | 12 | 99,146 | 99,288 | 198,434 |
| 3 | 99,270 | 99,389 | 198,659 | | 13 | 99,132 | 99,277 | 198,409 |
| 4 | 99,248 | 99,373 | 198,621 | | 14 | 99,112 | 99,265 | 198,377 |
| 5 | 99,230 | 99,358 | 198,588 | | 15 | 99,080 | 99,248 | 198,328 |
| 6 | 99,215 | 99,346 | 198,561 | | 16 | 99,037 | 99,228 | 198,265 |
| 7 | 99,200 | 99,334 | 198,534 | | 17 | 98,983 | 99,204 | 198,187 |
| 8 | 99,187 | 99,324 | 198,511 | | 18 | 98,917 | 99,175 | 198,092 |
| 9 | 99,175 | 99,314 | 198,489 | | 19 | 98,837 | 99,144 | 197,981 |
| 10 | 99,164 | 99,305 | 198,469 | | 20 | 98,744 | 99,109 | 197,853 |

**Table 2.0.1**: 2011 Social Security Administration actuarial data for the US

Of a total of 100,000 male newborns, the table predicts that 98,744 males will still be alive at age 20. Alternatively, it forecasts that 1,256 of the 100,000 males will have died before the age of 20. In contrast, the number of survivors in a female cohort is forecasted to be 99,109. The corresponding number of deaths would be 901. Now consider a combined population of 200,000 newborns that are equally divided between males and females. The total number of forecasted deaths is 2,147.

This data can be converted into probability statements.

Let: $D$ = the age of individual at death
$M$ = individual is male
$F$ = individual is female

The probability of a randomly selected individual in the total 200,000 will die by age 20 is

$$P(D < 20) = (2147 / 200,000) = 0.0107 \quad \text{or} \quad \text{about 1 chance in 93}$$

However, this probability is different for males and females. The probability of death by age 20 for a female infant is 0.0090.

$$P(D < 20 \mid F) = 0.0090 \quad \text{or} \quad \text{about 1 in 111}$$

The corresponding probability of death for a male infant is 0.0122. There is a 36% higher probability than for a female.

$$P(D < 20 \mid M) = 0.0124 \quad \text{or} \quad \text{about 1 in 80}$$

Knowing the gender of an infant clearly affects the probability of death by age 20. Thus, death by age 20 is not independent of an individual's gender.

The concept of conditional probability can be applied another way to the data in Table 2.1. The probability of dying by age 20 can be conditioned on the individual's age at the time

Let A= the age of an individual.

$$P(D < 20 \mid A)$$

In the total cohort, there are 198,328 individuals of age 15. Of that total 197,853 will survive to age 20. Equivalently, the number of deaths in this group is forecasted to be 475. Thus,

$$P(D<20 \mid A=15) = 475/198,328 = 0.0024$$

This probability is much smaller than the corresponding probability for a newborn which is 0.0124.Thus, a youngster's current age and death by age 20 are not independent events.

All of life's risks can be adjusted based on information we know about the individual. This information could be the gender, race, lifestyle, location, age, etc. The likelihood of dying from lung cancer is very different for a smoker and a non-smoker. Similarly, the likelihood of developing diabetes is not the same for a person of average weight as compared to someone who is obese. The likelihood of being a victim of a violent crime is much lower for someone living in an affluent suburb than in a much less affluent inner city neighborhood.

1. Provide your own example of a risk people of your age might face that will vary based on some other given information.

## Section 2.1 Committee Diversity

The concept of ***independent events*** can be defined using conditional probability. Two events are said to be independent of one another if information about one event does not affect the probability of the other event. This statement is represented as

$P(B \mid A) = P(B) \leftrightarrow A$ and $B$ are independent events and
$P(A \mid B) = P(A) \leftrightarrow A$ and $B$ are independent events.

The double-headed arrows (↔) in the statements above are read "if and only if" and show the equivalence of the statements on either side of the arrows.

Flipping a coin is good example of independent events.

Let $H_1$ first coin flip is heads
$H_2$ second coin flip is heads

The likelihood of a coin flip coming up heads is 0.5 on the first flip. The likelihood of the coin flip coming up heads is 0.5 on the second flip. This probability is not affected by what happened on the first flip.

$$P(H_2 \mid H_1) = P(H_2) = 0.5$$

Therefore, $H_2$ and $H_1$ are independent events.

In contrast, consider a group of four boys and three girls who have all volunteered to serve on a trip planning committee for Ms. Doubtful's class. Ms. Doubtful wondered what might happen if she selected the two without paying any attention to their gender. If two students are selected at random by drawing names out of a hat, what is the likelihood that both are boys? Conditional probability is helpful in answering this question by analyzing the sequence of random events.

The probability that the first person selected is a boy is just the ratio of boys to the total group.

$$P(B_1) = \frac{4}{7}$$

If the first person selected is a boy, there are still three boys among the group of six remaining students.

$$P(B_2 \mid B_1) = \frac{3}{6}$$

If, however, a girl was picked first, there are four boys remaining among the group of six students.

$$P(B_2 \mid G_1) = \frac{4}{6}$$

Clearly, the likelihood that the second selection is a boy changes depending on what happened in the first selection. Now to return to original question, what is the likelihood of selecting two boys? This can be expressed as $P(B_1 \cap B_2)$. If these were independent events, Ms. Doubtful could use the multiplication rule and multiply two probabilities. However, they are not independent. The general form for determining the probability of the intersection of two events is

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

If $A$ and $B$ are independent then $P(B|A) = P(B)$. As a result, the more general conditional probability formula reduces to the multiplication formula for independent events.

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B \mid A) \\ &= P(A) \cdot P(B) \end{aligned}$$

In this example,

$$\begin{aligned} P(B_1 \cap B_2) &= P(B_1) \cdot P(B_2 \mid B_1) \\ &= \frac{4}{7} \cdot \frac{3}{6} \\ &= \frac{12}{42} \\ &\approx 0.29 \end{aligned}$$

This formula suggests that the intersection of two events can be analyzed by considering them as a sequence. First determine the likelihood of the first event. Then determine the likelihood of the second event *conditioned on what happened first*.

Ms. Doubtful applied the same logic to determine the likelihood that both will be girls?

$$\begin{aligned} P(G_1 \cap G_2) &= P(G_1) \cdot P(G_2 \mid G_1) \\ &= \frac{3}{7} \cdot \frac{2}{6} \\ &= \frac{6}{42} \\ &\approx 0.14 \end{aligned}$$

It is twice as likely that the committee will consist of two boys, 12/42, as compared to the probability that the committee will consist of two girls, 6/42. Ms. Doubtful thought the planning committee would better represent the class's interest if there were one boy and one girl on the committee. However, she was not prepared to force this to happen. She wanted to give each person who volunteered the same opportunity to be selected.

To calculate the probability of a balanced committee, she reasoned there were two ways this could occur. It could happen that the first student selected was a boy and the second was a girl and vice versa. These two sequences are mutually exclusive; therefore, their probabilities can be added.

$$\begin{aligned} P(\text{one boy and one girl}) &= P(B_1 \cap G_2) + P(G_1 \cap B_2) \\ &= P(B_1) \cdot P(G_2 \mid B_1) + P(G_1) \cdot P(B_2 \mid G_1) \\ &= \frac{4}{7} \cdot \frac{3}{6} + \frac{3}{7} \cdot \frac{4}{6} \\ &= \frac{12}{42} + \frac{12}{42} \\ &= \frac{24}{42} \\ &\approx 0.57 \end{aligned}$$

The equation above represents the concept of a ***partition***. This involves decomposing an outcome into all of its mutually exclusive ways of happening. In this example $B_1$ and $G_1$ are a partition of the outcome of the first student selected. Ms. Doubtful calculated the probability of selecting one girl and one boy by partitioning on the set of all possible outcomes of the first pick.

Another method for determining this probability involves using the concept of complementary events.

2. Use the probabilities calculated for an all boys' committee or an all girls' committee to determine the likelihood of a mixed committee.

It was more likely that the committee would have one boy and one girl than two of the same gender. However, Ms. Doubtful would like to have better odds of there being at least one boy and one girl on the committee. Danielle Wiseman, the class math whiz, suggested that she expand the committee to three people.

Ms. Doubtful used the same logic as before to calculate the chances of all three members being of the same gender. She imagined a sequence of selections. The first person selected is a boy and then the second is a boy and then the third is a boy. She applied the same logic of conditional probability with one modification. When it came to the third boy, she had to condition on what had happened on the first two selections. She needed to determine $P[B_3|(B_1 \cap B_2)]$. The resultant formula was

$$\begin{aligned} P(B_1 \cap B_2 \cap B_3) &= P(B_1) \cdot P(B_2 \mid B_1) \cdot P\left[B_3 \mid (B_1 \cap B_2)\right] \\ &= \frac{4}{7} \cdot \frac{3}{6} \cdot \frac{2}{5} \\ &= \frac{24}{210} \\ &\approx 0.11 \end{aligned}$$

Ms. Doubtful asked Danielle to complete the analysis.

3. What is the probability that the committee would consist of three girls?

4. What is the probability that the committee would be mixed and include at least one girl and at least one boy? (Use the concept of the complement to calculate this probability.)

In the calculations above, we presented the analysis as a sequence of two events. First we pulled one name out of the hat, and then we picked a second name. Hopefully, it is clear that the probabilities calculated above would also apply if Ms. Doubtful put her hand in the hat and selected two names at once. The manner in which the two names are picked should not affect the probabilities. This illustrates an important point. The logic of conditional probability can be used to determine a joint probability even if the events do not occur in sequence.

In this example, each of the four conditional probabilities could be calculated using simple logic. In other contexts, the conditional probabilities are obtained through data analysis. This is demonstrated in the next example.

# Section 2.6 Testing for Celiac Disease

Tests of all kinds are used as tools for classifying. The tests you take in school aim to categorize your level of proficiency on a topic. Medical tests often seek to confirm or exclude the presence of a disease. Your email provider likely tests each message you receive to determine if it is spam (i.e., an unsolicited message) and filters it accordingly. Astronomers use tests to measure how likely it is that a given astronomical object is an Earthlike planet. Hearing aids use tests to determine which sounds to amplify and which to disregard. Cars process signals from sensors to decide when conditions are right to deploy air bags. Computer security software tests files and programs in an effort to quarantine anything infected with a virus. An emergency room physician needs to analyze a patient's electrocardiogram to determine if the patient should be admitted to the hospital or sent home. Tests of all kinds are everywhere!

## 2.6.1 False Positives and False Negatives

No test is perfect, however. A good test is likely to give a positive result when what it is looking for is actually present. It is also unlikely to give a positive result when what it is looking for is not present. But errors can occur, as outlined in Table 2.6.1.

| | | Object of Interest | |
|---|---|---|---|
| | | Present | Not Present |
| Test Result | Positive | True Positive | False Positive |
| | Negative | False Negative | True Negative |

**Table 2.6.1:** Possible true and false outcomes of a test

If a test result indicates a given condition when it is actually not present, the test result is said to be a ***false positive***. A false positive could occur if a woman performed a home pregnancy test and the test result was positive even though the woman was, in fact, not pregnant. False positive test results can have emotional, safety, and financial ramifications.

1. Describe a situation where you could say your score on a multiple-choice test included a false positive result on a question.

A ***false negative*** occurs if a test result is negative when it should not be. This would be the case if a test for a concussion came back negative when the patient actually had a concussion. There are, of course, harmful consequences for false negative test results as well. In the case of medical tests, it is particularly important to develop tests with a very small rate of false negatives.

2. Describe a situation where you could say your multiple choice test score constituted a false negative result on a specific question.

Consider mammography, which is a noninvasive test that utilizes low-energy X-rays. This test is widely used to screen for breast cancer in women over 40 years of age. Occasionally a mammogram will show an abnormal lump when there is no cancer present. When patients are notified of the positive test result, they usually become anxious about their prospects for having

breast cancer. The mammogram is not a definitive test, though. After the positive mammogram, women are often prescribed a biopsy. Biopsies are more expensive, invasive, and more painful than mammography.

3. Describe some adverse effects of a false negative mammogram.

Knowing how to interpret test results is vitally important because of the consequences of errors. Interpreting test results requires an understanding of conditional probability. For example, consider a patient who has blood drawn at a laboratory to screen for HIV. When the report for that blood test goes back to the patient's physician with a positive result, the physician needs to answer the question, "What is the probability the patient actually has HIV, given the positive test result?" The answer depends on the rates of false positives and false negatives of the test as well as demographic information about the patient (e.g., age, gender, race, history of sexual activity and drug use). Medical tests have well documented rates of giving false positive and false negative results. Experts on diseases are also able to identify risk factors (genetic or environmental) that increase a person's likelihood of having a given disease.

As you have seen, conditional probability is not something we human beings have good intuition about. Even highly educated and well trained professionals have difficulty in these situations. To illustrate this point, consider a study conducted by German cognitive psychologist Gerd Gigerenzer in 2002. He gave the following scenario to 100 American doctors.

> *The probability that one of these women has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does* not *have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?*

4. What is your estimate of the likelihood this woman has breast cancer?

Of the 100 doctors studied, 95 of them estimated the woman's probability of having breast cancer to be around 75%. The correct probability, given the provided information, is about 9%.

The correct answer is so low due to the fact that breast cancer is relatively rare at any particular time in the life of a woman. It was stated in the problem that breast cancer affects only 0.8% of women in this population. Imagine a group of 1,000 women, all of whom are going to be screened for breast cancer. On average, only eight of the 1,000 women will actually have breast cancer. Many more than that will receive false positive tests, simply because the vast majority of these women do not have breast cancer. The 0.8% rate of women having breast cancer is called the ***prior probability*** and plays a major role in interpreting test results, as you will soon see.

## 2.6.2 Diagnosing Celiac Disease

Lily has not been feeling well for the past few months. Her symptoms are generally restricted to her digestive system. They include abdominal swelling, frequent diarrhea, and occasional

vomiting. She is not at all sure of what is causing these symptoms. However, she sees the wide variety of gluten-free foods on the market and wonders if she might be gluten intolerant.

Celiac disease is a chronic (i.e., lifelong) autoimmune disease that causes inflammation in the small intestine. Autoimmune diseases are diseases in which one's immune system attacks one's own tissues and cells. In the case of celiac disease, antibodies attack the villi that line the small intestine. Villi are tiny finger-like structures which increase the surface area within the small intestine in order to absorb more nutrients from food. (See Figure 2.6.1.) When people with celiac disease eat gluten, their villi flatten out as their intestinal lining is damaged by antibodies. This leads to a decrease in the person's ability to absorb nutrients.

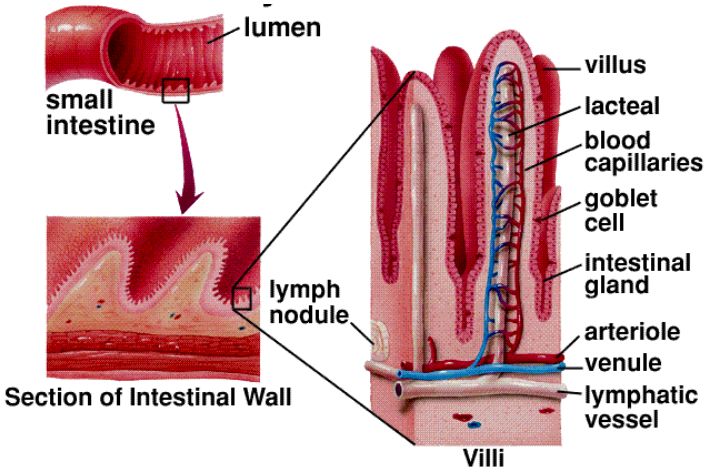


**Figure 2.6.1:** Anatomy of the small intestine

The ***prevalence*** of celiac disease in the United States is approximately 1%. This means that in any random sample of 100 Americans, on average, one person will have celiac disease. The prevalence of a disease shows the proportion of a population who have the disease. It can also be interpreted as the probability that a randomly selected person in the population has the disease.

Lily's physician, Dr. Grano, orders a blood test to check for the presence of a particular antibody, anti-tissue transglutaminase immunoglobulin A (tTG-IgA). This is a biomarker of celiac disease. Medical tests often are unable to directly detect a disease, so they check for biomarkers instead. Biomarkers are other cells, molecules, or genes that are associated with the disease and can be detected and measured more easily. However, the relationship between a biomarker and a disease is not perfect. All medical tests will occasionally give false results. They report either the existence of a disease that is not present (a false positive) or fail to recognize a disease that is present (a false negative). Sensitivity and specificity are statistical measures that describe how well medical tests work. The ***sensitivity*** of a test is the probability of the test showing a positive result when the disease is actually present. Another way of saying that is that the sensitivity of a test is the rate of true positive results from the test. The ***specificity*** of a test is the probability of the test showing a negative result when the disease is not present. Again, this could also be called the true negative rate of the test.

According to the National Institutes of Health, the sensitivity of the tTG-IgA test is 93% and its specificity is 98%. The tTG-IgA blood test has two possible outcomes: positive or negative. The sensitivity and specificity of tests can be used to calculate the rates of false test results, as shown below.

False positive rate = 1 – specificity

False negative rate = 1 – sensitivity

The rate of false positive test results could also be explained as a conditional probability. Specifically, the false positive rate is the probability of getting a positive (+) test result given no disease is present.

False positive rate = $P(+ \text{ test} \mid \text{no celiac})$

5. Write the conditional probability associated with each rate.
    a. True positive rate
    b. True negative rate
    c. False negative rate

6. The sensitivity of a test indicates the rate of true positive results. Explain why the false positive rate does not equal one minus the sensitivity of the test.

Dr. Grano explains the pathology and physiology of celiac disease to Lily. Before ordering the blood test, Dr. Grano also counsels Lily on the strengths and limitations of the tTG-IgA blood test. Lily hears the 93% sensitivity and 98% specificity; she feels very confident in the ability of the test to confirm or disconfirm her self-diagnosis of celiac disease.

Several days later, the results of Lily's blood work comes back positive for tTG-IgA. Dr. Grano interviews Lily to determine if there is a family history of celiac disease. Dr. Grano also asks if Lily has experienced any fatigue or weight loss that coincided with her other symptoms. Lily reported that there is no family history of celiac disease. She also indicated that she did not experience the fatigue or weight loss that are common to celiac disease.

If Dr. Grano endorsed the celiac disease diagnosis, her next step would be to prescribe a treatment plan. The typical treatment for celiac disease is a gluten-free diet. This lifestyle change would mean a radical shift in how Lily currently eats. Before making such a drastic recommendation, Dr. Grano considers ordering a second test more reliable to confirm the diagnosis.

The "gold standard" test for diagnosing celiac disease is an upper gastrointestinal endoscopy. An endoscope is a thin scope with a light and camera at its tip. To check for celiac disease, the endoscope is inserted through the mouth and guided through esophagus and the stomach until entering the beginning of the small intestine. (See Figure 2.6.2.)

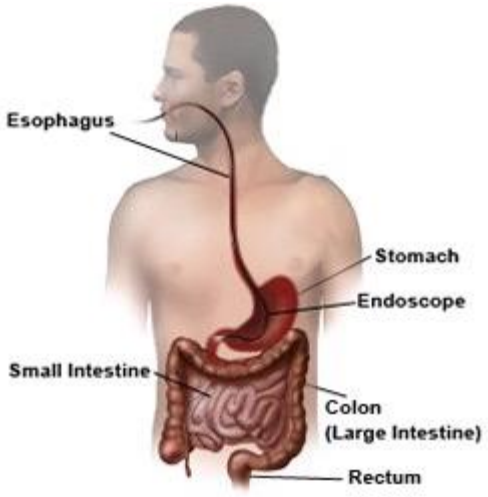


**Figure 2.6.2:** Digestive system with endoscope

If abnormal tissue is observed during the endoscopy, it can be sampled and sent to a lab for biopsy. In the biopsy, a pathologist will examine the abnormal tissue under a microscope to determine if the patient actually has celiac disease. This is the gold standard test because it is the best test available to diagnose celiac disease.

7. Give some reasons why the "gold standard" test would not be the first test used to diagnose a disease.

Dr. Grano advises Lily to undergo the upper gastrointestinal endoscopy. Lily listens patiently as Dr. Grano explains how to prepare for the test and what will occur during the procedure. The preparation includes not eating anything the day of the procedure and not drinking anything for the final four hours before the endoscopy. This ensures that the stomach will be empty. This will allow the endoscope to get a clear picture from within Lily's small intestine. She will also have to be sedated during the procedure.

Lily is not thrilled at the prospect of having to undergo this expensive, invasive, and inconvenient test. She is confident in the blood test and does not see the need for the biopsy.

Dr. Grano knows that the vast majority of people have poor intuition regarding probability in general and conditional probability in particular. The relevant probabilities are shown in Figure 2.6.3. These numbers are abstract and difficult to interpret meaningfully.

Let $T$ = positive test result
$T^c$ = negative test result
$C$ = has celiac disease
$C^c$ = does not have celiac disease

The four possible combinations of test results and disease can be represented with both a probability tree, Figure 2.6.3, and a table, Table 2.6.2. In the tree, a path corresponds to a pair of events. At the origin node, there are two possible branches that represent the random event

whether or not the individual has the disease. This leads to a second random event, the test results. The test results can either be positive or negative. On each branch is placed the probability of that event. For example, the topmost path begins with the probability of having celiac. This probability is 0.01. Next is the conditional probability of obtaining a positive test result given the individual has the disease. This probability is 0.93. The probability at the end of the path represents the joint probability, $P(C \cap T)$. It is calculated by using the formula

$$\begin{aligned} P(C \cap T) &= P(C) \cdot P(T \mid C) \\ &= (0.01)(0.93) \\ &= 0.0093 \end{aligned}$$

or about 1 in 108

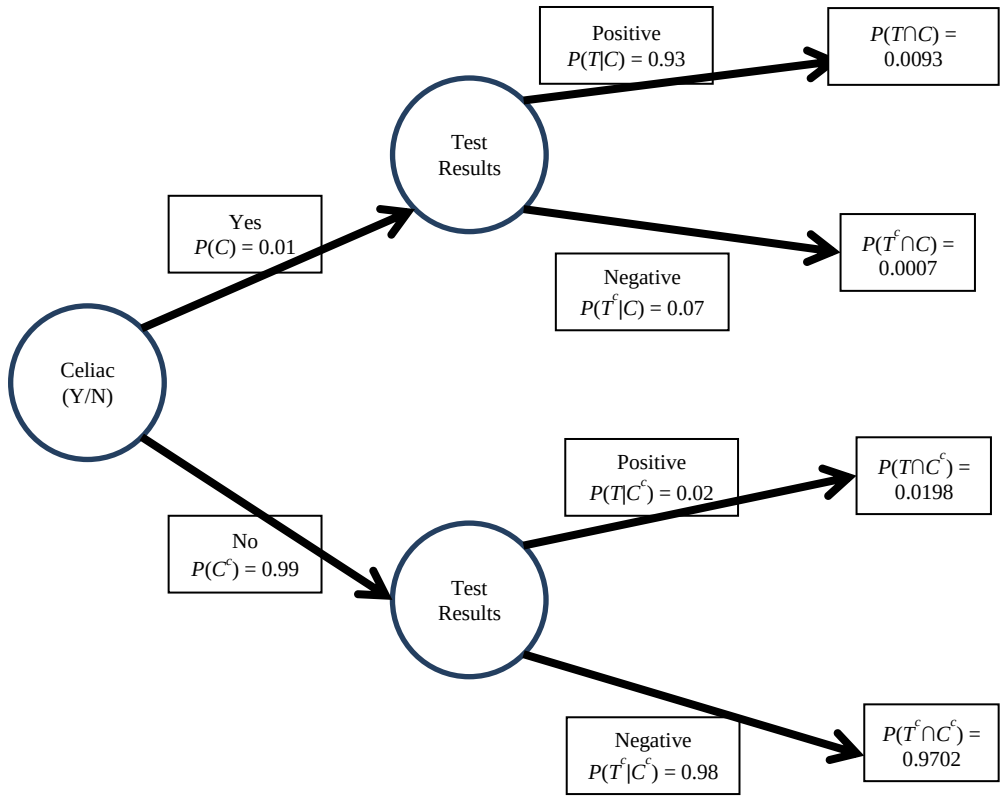


**Figure 2.6.3:** Tree diagram showing probabilities of possible outcomes for celiac disease test

Instead of focusing on all the probabilities, Dr. Grano tries to justify her recommendation by asking Lily to consider a sample of 10,000 people. The prevalence of celiac disease in the US is 1%. Therefore, on average, 100 people in the sample of 10,000 will have celiac disease. The sensitivity of the tTG-IgA test is 93%. This means that on average 93% of the 100 people who have celiac disease will test positive for it. Thus, there will be seven people who have the disease but will get a negative test result.

If only 100 of the 10,000 people actually have celiac, then 9,900 of them do not have the disease. The specificity of the tTG-IgA test is 98%. This means that 98% of the 9,900 people without

celiac will receive negative test results. That leaves 198 people who do not have celiac but receive positive test results. The results of this discussion are summarized in Table 2.6.2. Each of the four center cells represents the intersection of two events. For example, there are 9,702 people who do not have celiac and test negative out of the 10,000. Thus, the probability that a randomly selected individual who is tested will not have the disease and test negative is 0.9702. If we want to calculate the conditional probability within a row, we divide by the row total. There were a total of 291 people with positive test results. Out of this total, 93 have the disease. This discussion leads to the following conclusion. Given that Lily received positive test results, the probability she has the disease is (93/291) or 0.32. For most people this low probability is counterintuitive.

| | | True condition | | Totals |
|---|---|---|---|---|
| | | Celiac | No celiac | |
| tTG-IgA test result | Positive test | 93<br>True positive<br>$T \bigcap C$ | 198<br>False positive<br>$T \bigcap C^c$ | 291<br>Positive test results |
| | Negative test | 7<br>False negative<br>$T^c \bigcap C$ | 9702<br>True negative<br>$T^c \bigcap C^c$ | 9,709<br>Negative test results |
| Totals | | 100<br>Celiac | 9,900<br>No celiac | |

**Table 2.6.2:** Average outcomes for a sample of 1,000 people being tested for celiac disease

8. Given these calculations, do you think it is reasonable for Lily to undergo the upper gastrointestinal endoscopy procedure to confirm the celiac disease diagnosis before changing her diet?

The test is a much better predictor if the test results had been negative.

9. How many negative test results on average occur in the hypothetical sample of 10,000 people?

10. How many of the negative test results were true negatives?

11. Based on the above information, what would be the probability that Lily did not have celiac disease, if she got a negative tTG-IgA test result?

The impact of a false negative is more consequential than the impact of a false positive. A false negative means that the patient's illness was not discovered. In contrast, a false positive often entails additional tests. These tests are usually much more expensive and invasive. In addition, patients worry about their future until receiving the more accurate test results. Recently, the medical community has begun to review their recommendations on annual testing for relatively rare ailments with significant false positive rates.