STA 6384, Report 2.7

Carson Slater *Baylor University*

Problem: Work problem 2.25, p. 65 of Agresti.

For a diagnostic test of a certain disease, let π_1 denote the probability that the diagnosis is positive given that a subject has the disease, and let π_2 denote the probability that the diagnosis is positive given that a subject does not have it. Let ρ denote the probability that a subject has the disease.

(a)

More relevant to a patient who has received a positive diagnosis is the probability that he or she truly has the disease. Given that a diagnosis is positive, show that the probability that a subject has the disease (called the *positive predictive value*) is

$$\frac{\pi_1\rho}{\pi_1\rho + \pi_2(1-\rho)}.$$

Let D be the event that a subject has the disease, and let T^+ be the event that the diagnosis is positive. The given probabilities can be written in standard notation as:

- The prevalence of the disease: $P(D) = \rho$
- The probability of not having the disease: $P(D^c) = 1 - P(D) = 1 - \rho$
- The sensitivity of the test (true positive rate): $P(T^+|D) = \pi_1$
- The false positive rate: $P(T^+|D^c) = \pi_2$

We want to find the probability that a subject truly has the disease given that they received a positive diagnosis. This is the conditional probability $P(D|T^+)$. We can find this using Bayes' Theorem:

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+)}$$

The numerator is the joint probability of having the disease and testing positive, which is given by:

$$P(T^+ \cap D) = P(T^+|D)P(D) = \pi_1\rho$$

The denominator, $P(T^+)$, is the overall probability of getting a positive test result. This can be found using the Law of Total Probability by summing the probabilities of the two ways a positive result can occur: a

true positive or a false positive.

$$\begin{aligned} P(T^+) &= P(T^+ \cap D) + P(T^+ \cap D^c) \\ &= P(T^+|D)P(D) + P(T^+|D^c)P(D^c) \\ &= \pi_1\rho + \pi_2(1-\rho) \end{aligned}$$

By substituting the expressions for the numerator and denominator back into the Bayes' Theorem formula, we arrive at the expression for the positive predictive value:

$$P(D|T^+) = \frac{\pi_1\rho}{\pi_1\rho + \pi_2(1-\rho)}$$

(b)

Suppose that a diagnostic test for HIV+ status has both sensitivity and specificity equal to 0.95, and $\rho = 0.005$. Find the probability that a subject is truly HIV+, given that the diagnostic test is positive.

First, we must define our variables based on the information provided in the problem. The formula for the positive predictive value (PPV) is:

$$\text{PPV} = \frac{\pi_1\rho}{\pi_1\rho + \pi_2(1-\rho)}$$

The problem gives us the following values:

- **Sensitivity** is the true positive rate, so $\pi_1 = 0.95$.
- **Specificity** is the true negative rate ($P(T^-|D^c)$). The variable π_2 represents the false positive rate ($P(T^+|D^c)$), which is calculated as $1 -$ specificity. Therefore, $\pi_2 = 1 - 0.95 = 0.05$.
- **Prevalence** is given as $\rho = 0.005$.

Next, we substitute these values into the PPV formula to find the probability that a subject with a positive test is truly HIV+.

$$\begin{aligned} \text{PPV} &= \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.05)(1-0.005)} \\ &= \frac{0.00475}{0.00475 + (0.05)(0.995)} \\ &= \frac{0.00475}{0.00475 + 0.04975} \\ &= \frac{0.00475}{0.0545} \\ &\approx 0.0871559... \end{aligned}$$

So, the probability that a randomly selected person from this population who tests positive for HIV is actually HIV+ is approximately 0.0872, or 8.72%.

(c)

To better understand the answer in (b), using the probabilities given there either (i) find the joint probabilities relating diagnosis to actual disease status and discuss their relative sizes, or (ii) construct a tree diagram showing what you would expect to happen for a typical sample of 1000 subjects (first branching from the root according to whether a subject is truly HIV+ and then branching according to the test result), showing that of the subjects with a positive diagnosis, the proportion actually HIV+ agrees with the result in (b).

To better understand the answer in part (b), we can model what would happen to a hypothetical cohort of 1000 people.

1. **Step 1: Apply Prevalence**

   First, we divide the 1000 subjects based on the disease prevalence, $\rho = 0.005$.

   - Expected number of subjects who are truly HIV+: $1000 \times 0.005 = 5$ subjects.
   - Expected number of subjects who are not HIV+: $1000 \times (1 - 0.005) = 995$ subjects.

2. **Step 2: Apply Test Sensitivity and Specificity**

   Next, we determine the test outcomes for each group. The test has a sensitivity of 0.95 and a specificity of 0.95 (which means the false positive rate is $1 - 0.95 = 0.05$).

   - For the 5 subjects who are HIV+, we expect:
     - **True Positives** (correctly identified): $5 \times 0.95 = 4.75$
     - **False Negatives** (missed cases): $5 \times 0.05 = 0.25$
   - For the 995 subjects who are not HIV+, we expect:
     - **False Positives** (incorrectly identified): $995 \times 0.05 = 49.75$
     - **True Negatives** (correctly identified): $995 \times 0.95 = 945.25$

   (Note that expected values are statistical averages and do not need to be whole numbers.)

3. **Step 3: Calculate the Positive Predictive Value**

   Now, we can find the proportion of those who tested positive that are actually HIV+.

   - Total subjects with a positive diagnosis = (True Positives) + (False Positives)

   $$\text{Total Positive Tests} = 4.75 + 49.75 = 54.5$$

   - The number of these individuals who are truly HIV+ is the number of True Positives, which is 4.75.

   The proportion is therefore:

   $$\frac{\text{Number of True Positives}}{\text{Total Number of Positive Tests}} = \frac{4.75}{54.5} \approx 0.0871559\ldots$$

This proportion, approximately 0.0872, is identical to the result found using Bayes' theorem in part (b). This approach clearly demonstrates why the positive predictive value is so low: even with a highly accurate test, the vast number of healthy individuals generates a high number of false positives, which greatly outweighs the number of true positives from the small, infected population.

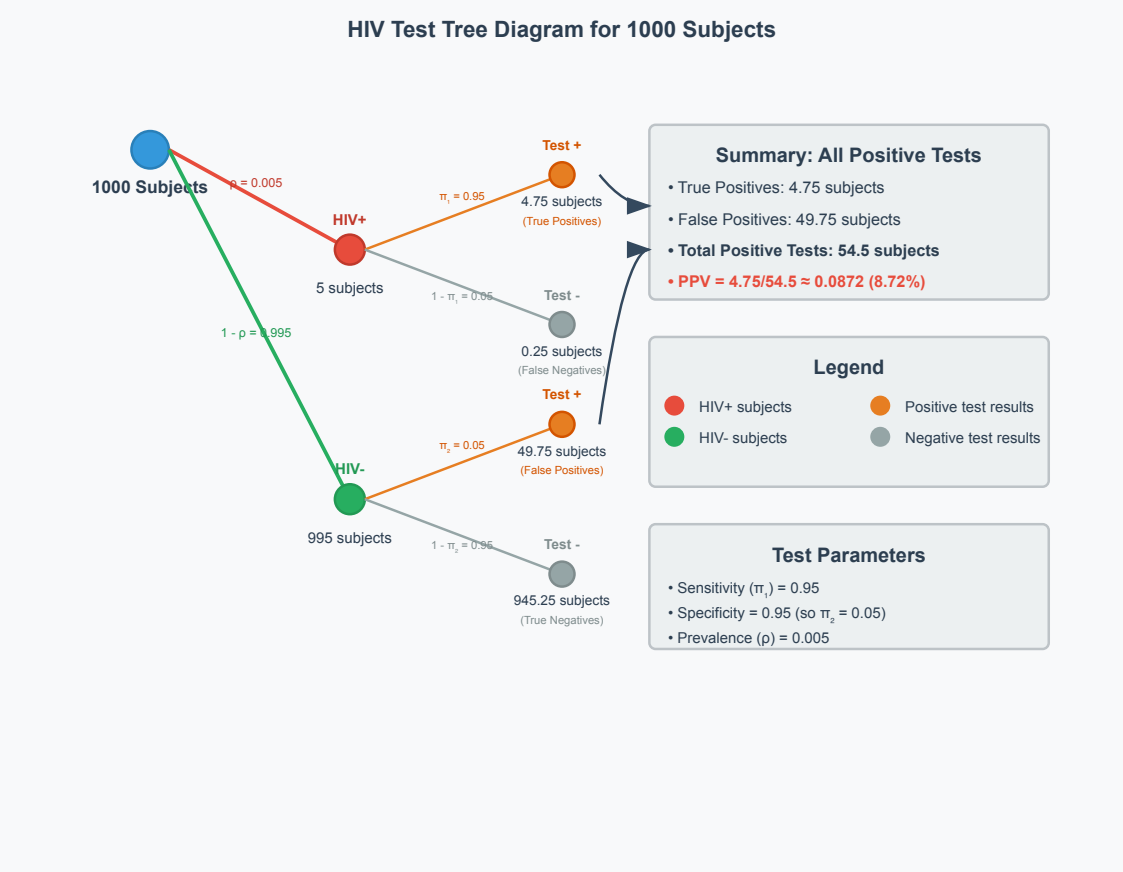

Figure 1: HIV Test Tree Diagram for 1000 Subjects

(d)

Discuss how the answer in (b) depends on the prevalence ρ. Illustrate by finding the answer when $\rho = 0.10$ instead of 0.005.

The positive predictive value (PPV) is highly dependent on the prevalence of the disease, ρ. An examination of the formula from part (a) reveals why:

$$\text{PPV} = \frac{\pi_1 \rho}{\pi_1 \rho + \pi_2 (1 - \rho)}$$

The prevalence, ρ, is a key component of both the numerator (representing true positives) and the denominator (representing all positives).

As disease prevalence ρ **increases**:

- The number of **true positives** ($\pi_1 \rho$) increases proportionally.
- The number of **false positives** ($\pi_2 (1 - \rho)$) decreases.

Both effects work together to significantly **increase the PPV**. In low-prevalence settings, the large number of disease-free individuals generates a substantial number of false positives, which can easily outnumber

the true positives. In high-prevalence settings, the opposite is true, and a positive test is much more likely to be accurate.

Illustration with $\rho = 0.1$

To illustrate this strong dependence, we recalculate the PPV from part (b) using the higher prevalence of $\rho = 0.10$. The sensitivity ($\pi_1 = 0.95$) and specificity ($1 - \pi_2 = 0.95$) remain the same.

$$\begin{aligned}
\text{PPV} &= \frac{\pi_1 \rho}{\pi_1 \rho + \pi_2 (1 - \rho)} \\
&= \frac{(0.95)(0.10)}{(0.95)(0.10) + (0.05)(1 - 0.10)} \\
&= \frac{0.095}{0.095 + (0.05)(0.90)} \\
&= \frac{0.095}{0.095 + 0.045} \\
&= \frac{0.095}{0.140} \\
&\approx 0.67857\ldots
\end{aligned}$$

With a prevalence of 10%, the PPV is approximately 67.9%.

Comparison:

- When $\rho = 0.005$ (0.5% prevalence), PPV $\approx 8.7\%$.
- When $\rho = 0.10$ (10% prevalence), PPV $\approx 67.9\%$.

This dramatic increase confirms that the exact same test can have vastly different predictive meanings depending on the underlying prevalence of the disease in the group being tested.