Positive Predictive Value (Bayes' Theorem)

Chapter 2, Lab 3

OpenIntro Biostatistics

Topics

- Positive Predictive Value
- Bayes' Theorem
- Simulation

The **positive predictive value (PPV)** of a diagnostic test is the probability that a person has a disease, given that they tested positive for it. This lab illustrates three common approaches for calculating PPV: contingency tables, tree diagrams, and simulation. The last two questions are more concept-oriented and probe the relationships between prevalence, sensitivity, specificity, PPV, and NPV.

The material in this lab corresponds to Section 2.2.5 of *OpenIntro Biostatistics*.

Introduction

Suppose a child tests positive for trisomy 21 from a cell-free fetal DNA (cfDNA) test. What is the probability that the child does have trisomy 21, given the positive test result?

This conditional probability, $P(D|T^+)$, where D is the event of having the disease and T^+ is the event of a positive test, has a specific name: **positive predictive value (PPV)** . Similarly, the **negative predictive value (NPV)**, $P(D^c|T^-)$, is the probability that disease is absent when test results are negative. These values are of particular importance to clinicians in assessing patient health.

The characteristics of diagnostic tests are given with the reverse conditional probabilities—the probability that the test correctly returns a positive test in the presence of disease and a negative result in the absence of disease. The probability $P(T^+|D)$ is referred to as the **sensitivity** of the test and $P(T^-|D^c)$ is the **specificity** of the test.[1] Given these two values and the marginal probability of disease $P(D)$, also known as the **prevalence**, it is possible to calculate both the PPV and NPV of a diagnostic test.

- Sensitivity: Of 1,000 children with trisomy 21, approximately 980 test positive.
- Specificity: Of 1,000 children without trisomy 21, approximately 995 test negative.
- Prevalence: Trisomy 21 occurs with a rate of approximately 1 in 800 births.

There are several possible strategies for approaching this type of calculation: 1) creating a contingency table for a large, hypothetical population, 2) running a simulation, 3) using an algebraic approach with Bayes' Theorem.

[1] Note that the sensitivity can also be referred to as the probability of a true positive and the specificity is the probability of a true negative. Thus, the probability of a false negative equals (1 - sensitivity) and the probability of a false positive equals (1 - specificity).

1. **Table-based approach**. Constructing a contingency table for a large, hypothetical population offers an intuitive way to understand the distribution of disease incidence and test outcome. Suppose that the total population is 100,000. The following table can be filled in based on the known information about disease prevalence, test specificity, and test sensitivity.

	Disease Present	Disease Absent	Sum
Test Positive			
Test Negative			
Sum			100,000

   a) Calculate the two column totals—out of 100,000 children, how many are expected to have trisomy 21? How many are expected to not have trisomy 21?

   b) The most efficient way to compute each table value is to use R as a calculator. Run the following code chunks and populate the rest of the table cells. Note that the first chunk simply re-calculates the values from part a).

```
#parameters
prevalence = 1/800
sensitivity = 0.980
specificity = 0.995
population.size = 100000

#expected number with trisomy 21
expected.cases = population.size * prevalence
expected.cases

#expected number without trisomy 21
expected.noncases = population.size - expected.cases
expected.noncases
```

```
#expected number with trisomy 21, tested positive (true pos)
expected.true.positives = expected.cases * sensitivity
expected.true.positives

#expected number without trisomy 21, tested positive (false pos)
expected.false.positives = expected.noncases * (1 - specificity)
expected.false.positives

#total expected positives
total.expected.positives = expected.true.positives + expected.false.positives
total.expected.positives
```

```
#expected number with trisomy 21, tested negative (false neg)
expected.false.negatives = expected.cases * (1 - sensitivity)
expected.false.negatives

#expected number without trisomy 21, tested negative (true neg)
expected.true.negatives = expected.noncases * specificity
expected.true.negatives

#total expected negatives
total.expected.negatives = expected.true.negatives + expected.false.negatives
total.expected.negatives
```

c) Using values from the table, calculate an estimate of the probability that a child who tests positive actually has trisomy 21. *Hint:* Think about how the definition of conditional probability could be applied here... $P(D|T^+) = \frac{P(D \text{ and } T^+)}{P(T^+)}$

2. **Simulation approach**. Rather than building a population based on expected values, a population can be simulated using the provided probabilities. This approach was introduced in the previous lab.

   The following code creates a simulated dataset of 100,000 individuals, each of whom has a disease status and test result. Run the code then answer the following questions. There are comments in the code that correspond to the following questions; i.e., when answering part a), look for the comment `(part a)`.

```
#define parameters
population.size = 100000
prevalence = 1/800
sensitivity = 0.980
specificity = 0.995

#create empty vectors to store results
disease.status = vector("numeric", population.size)
test.result = vector("numeric", population.size)

#set the seed for a pseudo-random sample
set.seed(2018)

#assign disease status (part a)
disease.status = sample(c(0,1), size = population.size,
                    prob = c(1 - prevalence, prevalence),
                    replace = TRUE)

#assign test result (part b)
for(k in 1:population.size){

  if(disease.status[k] == 0){
test.result[k] = sample(c(0,1), size = 1,
                        prob = c(specificity, 1 - specificity))
```

```
  }

  if(disease.status[k] == 1){
test.result[k] = sample(c(0,1), size = 1,
                        prob = c(1 - sensitivity, sensitivity))
  }

}

#create matrix of disease status and test result (part c)
disease.status.and.test.result = cbind(disease.status, test.result)

#create a table of test result by disease status
addmargins(table(test.result, disease.status))

#calculate ppv (part d)
ppv = sum(test.result[disease.status == 1])/sum(test.result)
ppv

#calculate npv (part e)
```

a) Explain how `sample()` is being used to fill in `disease.status`. If an individual is assigned a `0`, what is their disease status?

b) How is test outcome assigned if an individual has disease status `0`? How is test outcome designed if an individual has disease status `1`?

c) Take a look at `disease.status.and.test.result`. What does a single row ith a `0` in both columns represent?

d) Explain the line of code used to calculate PPV based on the results of the simulation.

e) Is the PPV calculated through the simulation different from the result using the table method? If so, explain why the two numbers differ.

f) Estimate the NPV based on the results of the simulation.

3. **Algebraic approach**. The expanded formula for PPV is based upon the definition of conditional probability, and rewriting joint probabilities as the product of a conditional and marginal probability by applying the general multiplication rule. The logic behind the formula can be illustrated via a tree diagram.

$$\begin{aligned} P(D|T^+) &= \frac{P(D \text{ and } T^+)}{P(T^+)} \\ &= \frac{P(D \text{ and } T^+)}{P(D \text{ and } T^+) + P(D^c \text{ and } T^+)} \\ &= \frac{P(T^+|D)P(D)}{[P(T^+|D) \times P(D)] + [P(T^+|D^c) \times P(D^c)]} \end{aligned}$$

a) Draw a tree diagram that organizes the four possible combinations of disease incidence and test outcome: D and T^+, D and T^-, D^C and T^+, D^C and T^-.

b) Use the formula to calculate $P(D|T^+)$.

c) The formula can be generalized in terms of events A and B, where

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}.$$

In this form, it is commonly known as Bayes' Theorem. Apply Bayes' Theorem to calculate the NPV.

Challenge Questions

4. The strongest risk factor for breast cancer is age; as a woman gets older, her risk of developing breast cancer increases. The following table shows the average percentage of American women in each age group who develop breast cancer, according to statistics from the National Cancer Institute. For example, approximately 3.56% of women in their 60's get breast cancer.

Table 1: Prevalence of Breast Cancer by Age Group

Age Group	**Prevalence**
30 - 40	0.0044
40 - 50	0.0147
50 - 60	0.0238
60 - 70	0.0356
70 - 80	0.0382

A mammogram typically identifies a breast cancer about 85% of the time, and is correct 95% of the time when a woman does not have breast cancer.

a) If a woman in her 60's has a positive mammogram, what is the likelihood that she has breast cancer? Solve this problem algebraically.

b) Use an R simulation to simulate the results for administering mammograms to a population of 100,000 women in their 30's. How many women in this hypothetical population are expected to test positive for breast cancer? Estimate the PPV of a mammogram for a woman in her 30's.

c) Using whatever methods you wish, calculate the PPV for each age group; show your work. Describe any trends you see in the PPV values as prevalence changes. Explain the reason for the trends in language that someone who has not taken a statistics course would understand.

d) Suppose that two new mammogram imaging technologies have been developed which can improve the PPV associated with mammograms; one improves sensitivity to 99% (but specificity remains at 95%), while the other improves specificity to 99% (while sensitivity remains at 85%). Which technology offers a higher increase in PPV? Explain your answer.

5. Prostate-specific antigen (PSA) is a protein produced by the cells of the prostate gland. Blood PSA level is often elevated in men with prostate cancer, but a number of benign (not cancerous) conditions can also cause a man's PSA level to rise. The PSA test for prostate cancer is a laboratory test that measures PSA levels from a blood sample. The test measures the amount of PSA in ng/ml (nanograms per milliliter of blood).

   The sensitivity and specificity of the PSA test depend on the cutoff value used to label a PSA level as abnormally high. In the last decade, 4.0 ng/ml has been considered the upper limit of normal, and values 4.1 and higher were used to classify a PSA test as positive. Using this value, the sensitivity of the PSA test is 20% and the specificity is 94%.

   The likelihood that a man has undetected prostate cancer depends on his age. This likelihood is also called the prevalence of undetected cancer in the male population. The following table shows the prevalence of undetected prostate cancer by age group, where age is measured in years. Prevalence is measured as a proportion of the population. For instance, the third row shows that 6.0% of males age 61 - 70 have undetected prostate cancer.[2]

Table 2: Prevalence of Prostate Cancer by Age Group

Age Group	**Prevalence**	**PPV**	**NPV**
< 50	0.001		
50 - 60	0.020		
61 - 70	0.060		
71 - 80	0.100		

   a) Calculate the missing PPV and NPV values, using any method.

   b) Describe any observable trends in the PPV and NPV values.

   c) Explain the reason for the trends in part b), in language that someone who has not taken a statistics course would understand.

   d) The cutoff for a positive test is somewhat controversial. Explain how lowering the cutoff for a positive test from 4.1 ng/ml to 2.5 ng/ml would affect sensitivity and specificity. The answer to this part does not use any of the calculations in the previous parts of the problem.

[2] The numbers are only approximate, and vary by country.