**Lab 5 – A conditional probability application: sensitivity and specificity**

**To submit before your next lab: answers to all numbered questions. Also submit all commands and/or functions you used to generate your output, and submit a single .R file containing all of the scripts you wrote for this lab. If a question asks you to submit the output of one of your functions, submit that as well.**

In our last lab, we simulated various simple experiments: flipping a fair coin, rolling a fair die, and drawing cards from a deck. We are now going to simulate a real-life situation. Simulations are often conducted in order to help policymakers make better decisions.

# Application: testing water for impurities

Many companies must monitor the effluent that is discharged from their plants. It is the law that the concentration of arsenic, which is used in the production of integrated circuits, be below a certain limit before it is released into rivers.

1. Suppose that the concentration of arsenic in an effluent from a particular company is below the limit 95% of the time. In a new script file called **Lab5.R**, write a function, **aboveLimit(n)** that simulates testing **n** samples of effluent and returns the number of samples that are **above** the limit. Assume the samples are independent and that each sample has a 95% chance of being below the limit. Run your function for n=100000 and give your output. (You should be able to tell whether your function returns a reasonable result.)

Our company uses a test to measure whether the concentration of arsenic in a sample of water is below limit. The testing method, of course, is not perfect: it has a **sensitivity** of 90%; that is, 90% of above-limit concentrations are correctly identified as being above limit. The sensitivity is also known rate of **true positives**: that is, the proportion of the time an above-limit sample “tests positive”. We can express the sensitivity in the language of conditional probability: the sensitivity is the probability that a sample tests positive, given that it is above-limit.

1. Write a function called **truePositives(n)** that simulates testing **n** samples that are above limit, and returns the number that test positive. Run your function for n=100000 and give your output.

The **specificity** of our test is 92%; that is, 92% of the time, a sample whose concentration of arsenic is below the limit is correctly identified as having a concentration that is below the limit.

1. In a sentence, express the rate of **false positives** (that is, the proportion of the time that a below-limit sample tests as **above** limit) in terms of conditional probabilities.
2. Write a function called **falsePositives**(n) that simulates testing **n** samples that are **below** limit, and returns the number that test **positive** (ie, your function returns the number of below limit samples out of **n** that the test **incorrectly** identifies as being above limit). Run your function for n=100000 and give your output.

Now, in real life, we don’t know for sure whether the concentration of arsenic in a sample of effluent is in fact above limit – all we know is the result of our test. The sensitivity tells us the probability that a sample tests as above limit, given that it is actually above limit.

We want to know the opposite: what is the probability that a sample is above limit, given that it tests as above limit? In other words…how reliable our test is. We will guess this probability and then run a simulation to get a precise answer.

1. Suppose our test tells us that the concentration of arsenic in a sample is below the limit. What do you think is the probability that it actually is below the limit? (There is no right or wrong answer to this question – we’ll get an answer with a simulation in Question 7.)
2. Now suppose our test tells us that the concentration of arsenic in a sample is above the limit. What do you think is the probability that our test is correct – ie, the sample actually contains an above-limit concentration of arsenic?

Now we will run some simulations to see how accurate our guesses in Questions 5 and 6 were.

1. In this question, we will create a simulation that uses the relative frequency approach to answer Question 5: *Suppose our test is negative – that is, the test tells us that the concentration of arsenic in a sample is below the limit. What is the probability that the sample actually is below limit?*

To do this, you will write a function called **probBelowGivenTestsBelow(n)** that simulates testing **n** samples and returns the following proportion:  
  
 .   
  
Note that the denominator will be a sum, because there are two types of negative results: false negatives, which arise from samples that are above limit; and true negatives, which arise from samples that are below the limit.

Run your function for n=100000 and give your output. (Hint: your output should be above 0.99) Note: By calling the functions you wrote in Question 1, 2, and 3, you can write this function in just a few lines – my **probBelowGivenTestsBelow(n)** function was six lines long. But you will need to be very careful with your function calls. Hint: start by finding the number of those **n** samples that are below limit.

1. Now write a function called **probAboveGivenTestsAbove(n)** to get an answer to the following question: *suppose our test is positive – that is, our test tells us that the concentration of arsenic in a sample is above the limit. What is the probability that our test is correct – that is, the sample actually contains an above-limit concentration of arsenic?*

Run your function for n=100000 and give your output.

1. How do your outputs for questions 7 and 8 compare to your guesses for questions 5 and 6? If you worked at the lab that tested samples, would you consider a negative result to be fairly accurate? How about a positive result?
2. Now let’s see what happens if the incidence of above-limit samples were different. Suppose that at a different plant, only 90% of samples had a below-limit concentration of arsenic. Now what are your results for Questions 7 and 8? What if a plant is very good at treating its samples, and 99% are below limit? Give your results, clearly labeled.