STAT 88: Lecture 4

Contents

Warm up: Let B_i be the event that a black card appears at Position i and R_i be the event that a red card appears at Position i.

(a) If you deal 2 cards, what is the chance the 2nd card is red? i.e. find $P(R_2)$.

(b) Find $P(R_{20} \cap R_{33}), P(R_{20} \cap B_{33}), P(B_{52}|R_{21}R_{40})$.

20th card is red

33rd card is black

Last time

Sec 2.1 The Chance of an Intersection

Multiplication (AND) rule:

$$P(A \cap B) = P(A|B)P(B).$$

Inclustion Exclustion (OR) rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Sec 2.2 Symmetry in Simple Random Sampling

When randomly sampling from a population (with or without replacement), be aware of the difference between unconditional and conditional probability.

Example: Consider a deck of cards.

- $P(B_2) = 26/52$.
- $P(B_2|R_1) = 26/51$.

2.3. Bayes' Rule

So far we have used the multiplication rule $P(A \cap B) = P(B|A)P(A)$ only in the settings where we can calculate $P(B|A)$ directly, e.g. $P(R_2|B_1) = 26/51$. Here we have 2 stages and we condition on what happens in the first stage.

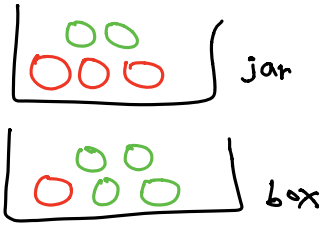
The general definition of conditional probability, regardless of setting, is just a rearrangement of the multiplication rule.

Conditional Probability (Division Rule) Let $A, B \subseteq \Omega$ and $P(A) > 0$. The conditional probability of B given A is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Example: (a random container) I have two containers: a jar and a box. Each container has five balls.

- The jar contains three red balls and two green balls.
- The box contains one red ball and four green balls.



Randomly pick a container and then a ball from that container. Given that the ball is red, what is the chance you pick the box?

Let J = jar
B = box
R = red
G = green

Find P(B|R)

We have updated our opinion about whether you picked the box or jar.

- Before we knew the color of the ball, we said the chance of drawing the box is 0.5. This is called the prior probability of drawing the box.
- After we saw that the ball is red, we said that the chance that the box was drawn is 0.25. This is called the posterior probability of drawing the box.

This way of updating probabilities based on new information is the basis for much inference in data science.

Bayes' Rule For $A, B \subseteq \Omega$,

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

Note that this is just the division rule in a particular scenario.

The rule helps you find a posterior probability: a conditional chance for the first stage, given the result of a second stage. The calculation has two ingredients:

- Probabilities for the first stage; these are called prior probabilities.
- Conditional probabilities for the second stage given the first; these are called likelihoods.

Example: (Excercise 2.6.9) A factory has two widget-producing machines. Machine I produces 80% of the factory's widgets and Machine II produces the rest. Of the widgets produced by Machine I, 95% are of acceptable quality. Machine II is less reliable – only 85% of its widgets are acceptable.

Suppose you pick a widget at random from those produced at the factory.

1. Find the chance that the widget is acceptable, given that it is produced by Machine I.
2. Find the chance that the widget is produced by Machine I, given that it is acceptable.

2.4. Use and Interpretation

Harvard Medical School Survey (60 participants):

"If a test to detect a disease whose prevalence is 1/1,000 has a false positive rate of 5 per cent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?"

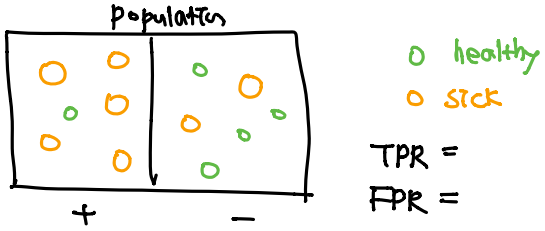
Harvard medical student answers ranged from 2% to 95%, with 27 out of the 60 Medical School members surveyed answering 95%. What do you say?

(Assume true pos rate is 1)

Terminology:

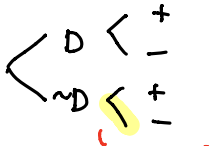
- Prevalence (also called the base rate of the disease) in the population is the percent of people who have the disease.
- True positive rate is the rate of positive results among those who do have the disease.
- False positive rate is the proportion of positive results among people who don't have the disease.

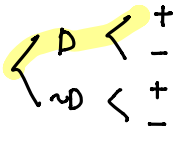
positive result - according to test the person has the disease
negative result - according to test the person doesn't have the disease.



Is this surprising?

Base Rate Fallacy





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

How did so many people get 95%?

$P(-|\sim D) = 95\%$

people confuse $P(D|+)$ for $P(-|\sim D)$

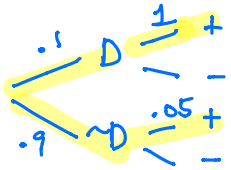
To compute $P(D|+)$ you need to take into account the base rate and people tend not to do this, instead focusing on likelihoods such as $P(-|\sim D)$

Key point Posterior probability is effectively by base rate as well as the likelihoods

Suppose you have a 10% chance of having the disease because you show some symptoms or have a family history.

This changes prevalence to .1 from .001.

Now



$$P(D|+) = \frac{0.1 * 1}{0.1 * 1 + 0.9 * 0.05} = 0.69$$

Example: A True/False test consists of 60 questions. A student knows the answers to 45 of the questions. The remaining 15 answers he guesses at random by tossing a fair coin each time. If it lands heads he answers True and if it lands tails he answers False.

A question is picked at random from the 60 questions on the test. Given that the student got the right answer, what is the chance that he knew the answer?