Inclusion-Exclusion Principle

Math 145, Spring 2019

Suppose we survey the 70 student in the class with the following questions:

- Do you like chocolate icecream? (Yes/No)
- Do you like vanilla icecream? (Yes/No)
- Do you like strawberry icecream? (Yes/No)

We enter each response into a spreadsheet.

	Chocolate	Vanilla	Strawberry
Survey 1	1	1	0
Survey 2	0	1	0
Survey 3	1	0	0
Survey 4	0	1	1
Survey 5	1	1	1
$\vdots$			

Here a 0 means No, I do not like that flavor, and 1 means Yes, I do like that flavor.

Then we use this data to determine four additional columns. If we consider the Chocolate column to be Column C, the Vanilla column to be Column V, and the Strawberry Column to be Column S, then the remaining columns are calculated by the following formulas:

- $CV : i = (C : i) \cdot (V : i)$ – this records if they like both Chocolate and Vanilla
- $CS : i = (C : i) \cdot (S : i)$ – this records if they like both Chocolate and Strawberry
- $VS : i = (V : i) \cdot (S : i)$ – this records if they like both Vanilla and Strawberry
- $CVS : i = (C : i) \cdot (V : i) \cdot (S : i)$ – this records if they like all three flavors

Make sure you understand why these formulas represent what they claim to represent.

When we add these extra columns our spreadsheet looks like:

	Chocolate	Vanilla	Strawberry	CV	CS	VS	CVS
Survey 1	1	1	0	1	0	0	0
Survey 2	0	1	0	0	0	0	0
Survey 3	1	0	0	0	0	0	0
Survey 4	0	1	1	0	0	1	0
Survey 5	1	1	1	1	1	1	1
$\vdots$							

Now, let's put the take the sums in each column and record these sums at the bottom of the spreadsheet (using all 70 surveys not just these first 5). Suppose these sums give us the following data:

	Chocolate	Vanilla	Strawberry	CV	CS	VS	CVS
Total sum	45	30	30	25	20	15	10

1 The basic counting question

By the **Inclusion-Exclusion Principle** the *total number of students who like at least one of the three flavors* is:

$$|C|+|V|+|S|-|C\cap V|-|C\cap S|-|V\cap S|+|C\cap V\cap S| = 45+30+30-25-20-15+10 = 55.$$

Therefore we conclude that of the 70 students in the class, 15 do not like any of the flavors chocolate, vanilla, or strawberry icecream.

2 Related more difficult questions

Now, let's answer some related trickier questions using the same data we are given above.

Question: What is the number of students who like at least two flavors out of the three options?

If we rephrase this question in terms of sets, we are trying to find the size of the set

$$(C\cap V)\cup(C\cap S)\cup(V\cap S)$$

Now, we know the size of each of these sets we are taking the union of from the table:

$$|C\cap V| = 25, \quad |C\cap S| = 20, \quad \text{and} \quad |\mathrm{V}\cap \mathrm{S}| = 15.$$

Now we need to think about if we add these numbers up, will anyone be overcounted and how many times will they be overcounted by?

Well the only way someone can appear in more than one of these three sets is if they like Chocolate, Vanilla, *and* Strawberry icecream. In this case, that person would be counted once in each of the three sets, so they would be counted 3 times instead of 1 time. Therefore we are overcounting those people by a factor of 2. We know how many people like all three flavors from the table:

$$|C\cap V\cap S| = 10.$$

Therefore, **the total number of students who like at least two of the three flavors is:**

$$|C\cap V|+|C\cap S|+|V\cap S|-2|C\cap V\cap S| = 25+20+15-2\cdot 10 = 40.$$

Because we only removed the overcounting, but still counted each student who likes all three flavors exactly once, this is the count of students who like either 2 or 3 of the flavors.

If we want to only count the students who like *exactly* 2 of the three flavors we should subtract out the number of students who like all three flavors. **The number of students who like exactly two flavors is:** $40-10=30$.

$$|C\cap V|+|C\cap S|+|V\cap S|-3|C\cap V\cap S| = 25+20+15-3\cdot 10 = 30.$$

3 When you have less information

In the above example, we had all the information about the size of every set and every possible intersection of sets. Now we will consider situations where we have more limited information and work backwards.

Suppose we don't have the full spreadsheet data, but we are only told the following information

- There are 70 students
- 45 students like chocolate icecream
- 30 students like vanilla icecream
- 30 students like strawberry icecream
- 30 students like *exactly* 2 of the three flavors.
- 10 students like *all* 3 flavors.

Problem: How many of the 70 students do not like any of the three flavors of icecream?

Now we need to work backwards since we are missing the information of the exact numbers for $|C \cap V|$, $|C \cap S|$ and $|V \cap S|$. Let's give variable names to these missing pieces of data:

$$x = |C \cap V|, \quad y = |C \cap S|, \quad \text{and} \quad \mathrm{z} = |\mathrm{V} \cap \mathrm{S}|.$$

Then **the calculation that gives the number of students who do not like any of the three flavors is:**

$$70 - (|C| + |V| + |S| - |C \cap V| - |C \cap S| - |V \cap S| + |C \cap V \cap S|)$$

$$= 70 - (45 + 30 + 30 - x - y - z + 10)$$

We are missing the values of x, y, z to finish this calculation. What we do know is the number of students who like exactly 2 of the flavors. As above, we know that the formula for the number of students who like exactly 2 of the flavors is:

$$|C \cap V| + |C \cap S| + |V \cap S| - 3|C \cap V \cap S| = x + y + z - 3 \cdot 10$$

This tells us that

$$30 = x + y + z - 3 \cdot 10$$

so

$$x + y + z = 60.$$

Then we can plug this into the formula above to find **the number of students who do not like any of the three flavors is:**

$$= 70 - (45 + 30 + 30 - x - y - z + 10) = 70 - (45 + 30 + 30 - (x + y + z) + 10)$$

$$= 70 - (45 + 30 + 30 - 60 + 10) = 70 - 55 = 15.$$

Try out this example where instead of knowing how many students like *exactly* two flavors, you are told that there are 40 who like *at least* two flavors.