The correct statement of Inclusion-Exclusion

Because I keep forgetting lol

Ivan Aidun

Let $\mathcal{S}$ be a finite set, let $\{P_1, \dots, P_n\}$ be properties that elements of $\mathcal{S}$ can have, and for a subset $I \subset \mathcal{I} \coloneqq \{1, \dots, n\}$ let $\mathcal{N}_I$ be the subset of $\mathcal{S}$ consisting of elements having property P_i for all $i \in I$, where we adopt the convention that $\mathcal{N}_\emptyset = \mathcal{S}$. Then the number of elements of $\mathcal{S}$ which have none of the properties P_i is

$$\begin{aligned}\#\left(\mathcal{S} \setminus \bigcup_{i=1}^{n} \mathcal{N}_i\right) &= \sum_{I \subset \mathcal{I}} (-1)^{|I|} \cdot \#\mathcal{N}_I \\ &= \#\mathcal{S} - \sum_{i=1}^{n} \#\mathcal{N}_i + \sum_{\{i,j\} \subset \mathcal{I}} \#\mathcal{N}_{i,j} - \dots.\end{aligned}$$

Furthermore, the successive partial sums are over/underestimates for $\#(\mathcal{S} \setminus \bigcup_{i=1}^{n} \mathcal{N}_i)$.

Example 1. Let $\mathcal{S}$ be the integers $2 \leq n \leq 1000$, and let $\mathcal{N}_p$ be those integers divisible by p for $p = 2, 3, 5, 7$. Then the elements of $\mathcal{S}$ in none of the $\mathcal{N}_p$ are the numbers whose prime factors are all greater than 7. (Which is pretty much just the primes between 7 and 1000, let's be honest.) The above tells us that the number of such numbers is

$$\begin{aligned}&999 - \sum_{p \in \{2,3,5,7\}} \lfloor \tfrac{1000}{p} \rfloor + \sum_{\{p,q\}} \lfloor \tfrac{1000}{pq} \rfloor - \sum_{\{p,q,r\}} \lfloor \tfrac{1000}{pqr} \rfloor + \lfloor \tfrac{1000}{2 \cdot 3 \cdot 5 \cdot 7} \rfloor \\ &= 999 - (500 + 333 + 200 + 142) + (166 + 100 + 71 + 66 + 47 + 28) - (33 + 23 + 14 + 9) + 4 \\ &= 999 - 1175 + 478 - 79 + 4 \\ &= 227.\end{aligned}$$

(In fact, there are only 164 primes between 7 and 1000.)

Example 2. (From Puzzling Stackexchange) There are 1000 people in a conference, 500 of whom speak English, 500 of whom speak Spanish, and 500 of whom speak Hindi. What is the greatest the number of monolinguals in the conference could be?

Based on the wording of the problem, it seems implied that every speaker at the conference speaks at least one of English, Spanish, or Hindi, so we will proceed on this assumption. I often find that with Inclusion-Exclusion puzzles, it's useful to first think of the most extreme scenario you can, then later prove that this is essentially the optimal scenario for the problem. In this case,

there could be as many as 750 monolinguals if 250 people are polyglots who speak all 3 languages, and the remaining 750 are all monolingual, evenly divided among the remaining three languages.

Let $\mathcal{S}$ be the set of all people at the conference, and let $\mathcal{N}_\ell$ be the set of people who do *not* speak language ℓ, with $\ell = 1, 2, 3$ corresponding to English, Spanish, and Hindi. (This is an Inclusion-Exclusion trick: often the properties P_i you want to choose are "elements in the complement of some other set".) In this case, the number of people who are in none of the $\mathcal{N}_\ell$ is the number of polyglots who speak all 3 languages—call this number P. The number of monolingual people M is at most $\sum_{\ell_1,\ell_2} \#\mathcal{N}_{\ell_1,\ell_2}$, and can be less if e.g. every English speaker happens to also speak Pennsylvania Dutch. Since we have assumed every person speaks one of English, Spanish, or Hindi, we have that $\mathcal{N}_{1,2,3} = \emptyset$. Applying Inclusion-Exclusion, we get that

$$P = 1000 - 3 \cdot 500 + \sum_{\ell_1,\ell_2} \#\mathcal{N}_{\ell_1,\ell_2},$$

so we get that $M \leq \sum_{\ell_1,\ell_2} \#\mathcal{N}_{\ell_1,\ell_2} = P + 500$. On the other hand, the set of monolinguals and the set of polyglots are disjoint, so $M + P \leq 1000$. Solving for M gives us that $M \leq 750$, and indeed we can see equality occurs exactly when both $M = P + 500$ and $P + M = 1000$.

Even if we hadn't made the assumption that every attendee speaks at least one of the three languages we are given data about, the answer above would still be optimal. To see this, note that our reasoning above is still valid on the subset of conference attendees who *do* speak at least one of English, Spanish, or Hindi. Letting M still represent the number of monolinguals who speak at least one of English, Spanish, or Hindi, if there are M' monolinguals who speak none of the three then instead we will obtain

$$P = (1000 - M') - 3 \cdot 500 + \sum_{\ell_1,\ell_2} \#\mathcal{N}_{\ell_1,\ell_2}.$$

Now we will get that $M \leq P + 500 - M'$. The total number of monolinguals at the conference is $M + M' \leq P + 500$, and similar reasoning as above gives us $M + M' \leq 750$.