# Have you tested for HIV?

## What this series is about

Following from my previous post on misconceptions about statistics, I have decided (as suggested by my friend Desmond) to continue with the trend of trying to explain how statistics can be misleading in our everyday lives. From HIV testing, to game show problems and more!

## Pregnancy/ HIV Tests

If you are in your mid 20s or older, you would have probably heard of (or joked about) this brand called Clear Blue. Clear Blue, like all the other pregnancy test-kit brands, always claim to be the number 1 brand in the world in terms of reliability and accuracy. Most of the pregnancy test kits would also have a print on the box stating that it has a "99% accuracy rate" or "more than 99% accurate". So, what does 99% accuracy actually mean? To explore this further, we must re-visit the concept of true positives, false positives, true negatives and false negatives.

## Hypothetical Scenario

Imagine we have a population of 10,000 people and we have with us a HIV test kit that is 99% accurate in predicting both HIV positive and negative individuals. That is, if I had HIV, the test would predict that I have HIV 99% of the time and if I did not have HIV, the test would similarly predict that I do not have HIV 99% of the time. That leaves a 1% error in both directions of the results. Let’s also imagine that out of the 10,000 people, only 100 individuals (1%) actually has HIV. Here are some of the outcomes summarized into a table:

|  |  |  |
| --- | --- | --- |
| **Actual / Tested** | **HIV** | **No HIV** |
| **HIV** | True Positive | False Negative |
| **No HIV** | False Positive | True Negative |

## The Issue

Assume we force all the 10,000 people to take the HIV test. We can make this example clearer by splitting our population up into 2 groups:

Individuals with HIV = 100

Individuals with no HIV = 9,900

Since the test is 99% accurate, this means that (according to our parameters) out of the group of individuals with HIV, 99 people tested will be "True positives" and 1 person would be a "False negative". Also, from the No HIV group, 9801 of the people will be "True negatives" (99% of 9,900) and 99 people will fall under "False positives" (1% of 9,900). Summarizing our results, we have:

|  |  |  |
| --- | --- | --- |
| **Actual / Tested** | **HIV** | **No HIV** |
| **HIV** | 99 | 1 |
| **No HIV** | 99 | 9801 |

Okay, now to put things in perspective. If we collate the results from all the individuals, we find that the tests output negative results 9801 + 1 = 9802 times, out of which only 1 was a false prediction. A 0.01% error, amazingly accurate isn't it! However, if we take a look at the other side of the coin, the test output a positive result 99 + 99 = 198 times. This means that the test is only correct 50% in this direction as 99 of the positives results were actually false. (Note that this is an extension of the "Precision", "Recall" and "Sensitivity" metric commonly discussed in classification problems.

## So, what does this mean?

If you happen to be part of that particular population described above with the prescribed parameters and you test negative for HIV, there is almost 0% chance that you have actually contracted it, so don't worry. If, however, the test shows a positive result, don't start quitting your job and dropping all your responsibilities just yet! There is quite a high chance (50%) that the test is incorrect, and you have not actually contracted HIV! This is drastically different from the initial idea of 99% accuracy isn't it!

## Okay, so what gives?

The factor that drives the misleading results is the actual underlying proportion of individuals with HIV. Let's do another quick hypothetical example with 10,000 people, but now with 50 (0.5%) HIV positive individuals and 9,950 non-HIV positive individuals. Going through the same calculations as above, we actually have the *"rate of predicting I have HIV when I actually don't"* increase from 50% to approximately 67%.

|  |  |  |
| --- | --- | --- |
| **Actual / Tested** | **HIV** | **No HIV** |
| **HIV** | 49 | 1 |
| **No HIV** | 150 | 9850 |

This shows that as the underlying proportion of HIV positive individuals decrease below the error rate of the test (1%), the more likely the test is going to predict you have HIV even when you don't!

From the discussion above, it seems that this "underlying HIV proportion" is extremely important in determining the "accuracy" of the results. However, how does one actually know what the underlying proportion is? This is unfortunately, one of the limitations of classical statistics but, also the underlying concept of "Bayesian statistics" where, the underlying proportion is estimated as a "prior belief" of a fact, whether qualitatively or quantitatively derived. More on Bayesian statistics next time. For now, have fun, be responsible and remember, *"Blame the person, don't blame the test!"***.**