Consistent Tests Within Science

Science uses statistical tests to analyse noisy data to find patterns. The collection of tests used creates a pseudo-hypothesis that represents the complete absence of an effect, in other words, we expect the two groups we are comparing to have a difference of zero. To say that a drug works or an invention helped, we have to show this pseudo-hypothesis is unreasonable by falsifying it. There's a little problem, it is not just a difference of around zero, it is a perfect zero, which is simply impossible to observe as there is always uncertainty and inaccuracy within our measurements.

This means this pseudo-hypothesis is vulnerable to the tiny inaccuracies and noise within our imperfect data. Then the pseudo-hypothesis is too easily rejected when the evidence doesn't properly confirm there is an effect. If we found a drug that worked 0.1% of the time, would we say this drug is useful? As 0.1% is not zero, our pseudo-hypothesis could be rejected. If the drug simply doesn't work (meaning that 0.1% is just noise in our data and the pseudo-hypothesis is true), then rejecting the pseudo-hypothesis is called a false positive or false alarm. To ensure we are not rejecting our pseudo-hypothesis too often, we set a maximum error rate. This error rate also tells us when to reject the pseudo-hypothesis, so smaller error rates mean we reject the hypothesis less often. Maybe more precisely, if we make the error rate more strict, we would need clearer evidence of an effect to reject the hypothesis that there is no effect. Whilst setting this maximum error rate does prevent too many false positives, it also sets a minimum error rate when there genuinely is not an effect (i.e. the drug does nothing useful). We could make the data better by collecting more samples, more participants, but that set error rate does not change, even if we collected millions of samples (**Figure 1**). When we do not consistently make the right decision with very large sample sizes, then the system is deemed *inconsistent.* We have a consistent system when there is an effect (i.e. the drug does something useful), as having more data makes patterns easier to find, even the subtle ones that are not obvious.

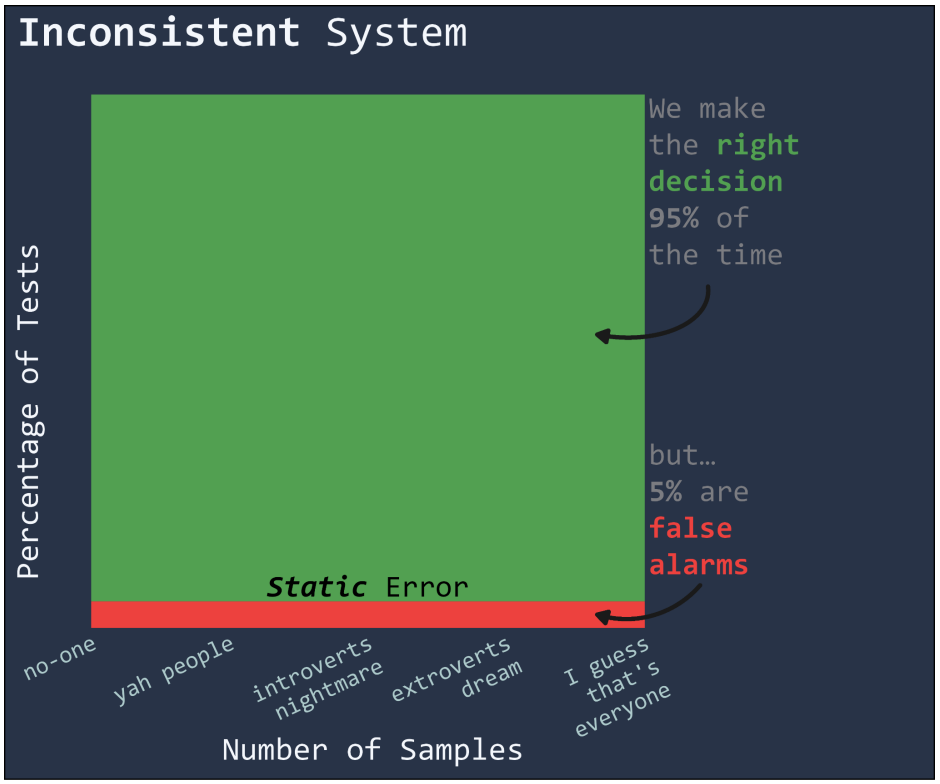

Figure 1: When we use a static error rate, we have an *inconsistent* system when there is no real effect.

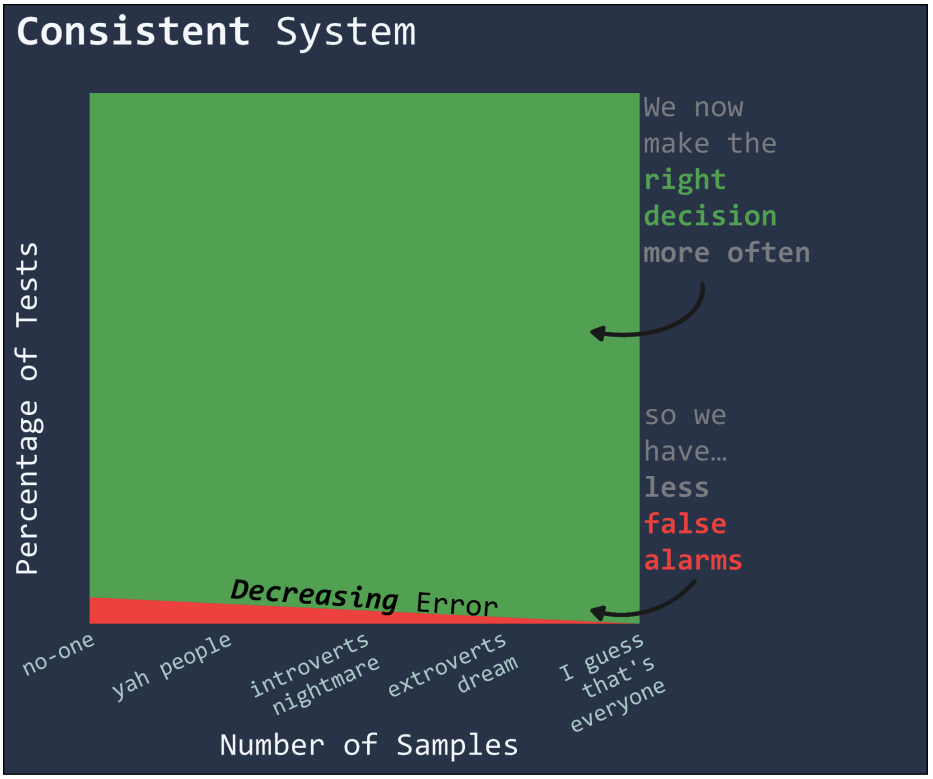

Figure 2: When we don't use a static error rate, we have a *consistent* system when there is no real effect.

To counter this issue of *inconsistency*, we can use an error rate which becomes smaller with larger samples. We could even create a region around zero where the difference is not good enough to be useful. So if a drug needed to work 10% of the time before we considered prescribing it and we found a new drug that worked 5% of the time. We wouldn't reject our pseudo-hypothesis (even though we didn't find the illusive 0%) because we would say the drug isn't good *enough*. Both of these changes (not using a static error rate and defining not *good enough* values) to our statistical tests would achieve consistency, where we continue to make the right decision as we collect more data/samples (**Figure 2**). My project tried to explore ways we can create a consistent test by defending our little pseudo-hypothesis from irrelevant or meaningless differences that are not helpful to the people science is trying to help.

By Mark Gallacher
Supervisor : Dr Guillaume Rousselet