# **Using Bayesian decision theory when evaluating A/B tests instead of Frequentist approach**

Until recently statistical significance has been widely used to evaluate A/B test results. However, in recent times we are moving toward alternative approaches of using statistical significance in online experimentation.

This is due to the fact that although statistical significance successfully reduces false positives it also misses out valuable gains giving rise to more false negatives.

Stanford Encyclopedia of Philosophy entry on Decision Theory considers EU (Expected Utility) theory or Bayesian decision theory as the appropriate account of scientific inference. When it comes to scientific inference it considers Classical or Error Statistics i.e., statistical significance to be the major competitor of Bayesian Decision Theory.

When it comes to online experimentation the major difference between statistical significance and expected utility is with regards to differences in goals between the two approaches. While statistical significance is based on controlling false positive (Type 1) error and false negative (Type II) error over all experiments, expected utility approach has the goal of making the best decision for each experiment given the evidence. Since the goals of the two approaches are different their performance would be dependent on the measurement of the performance.

When the aim is to make the best decision in online experimentation given the evidence, it can be easy to see how statistical significance might not result in the best decision being made in many cases. In this article, we will look at the performance of statistical significance and expected utility approach in terms of the count of correct and incorrect decisions using some stimulated data.

## An A/B Test Setting

Suppose we have a marketing site for an e-commerce business, and we want to see whether the site can be improved to increase sales. A run a design version of the marketing site and run an experiment where we measure revenue. The existing website is regarded as 'Treatment A' and the new design 'Treatment B'. Table 1 shows what the steps we would be taking under the two frameworks- Statistical Significance and Expected Utility approach to obtain results from this experiment.

| | Statistical Significance | Expected Utility |
|---|---|---|
| 1 | Calculate a sample size | Sample size not strictly necessary. Can estimate a sample size |
| 2 | Specify the control treatment arm | Not necessary |
| 3 | Run the experiment | Run the experiment |

|  |  |  |
|---|---|---|
| 4 | Run the experiment until the specified sample size is reached | Run the experiment until<br>a) probability threshold is reached OR<br>b) sample size is reached OR<br>c) a fixed period of time |
| 5 | If difference is statistically significant, make B the default. Otherwise stick to the original treatment. | Choose the treatment with the higher probability. |

In the last step where the decision of whether to make B the default or stick to A is made, in the statistically significant approach, we need significant evidence in favor of treatment B

Now onto Type II errors: the way that stat sig controls the false negative rate, i.e., the percentage of times that you will not see stat sig even when there is a change, is using a sample size calculator. With a large enough sample size, you can see any size effect. But with small samples you can only get statistical significance with very large effects.

Nonetheless, just like with stat sig if we want more confidence we could collect more data. In future posts we'll talk about cases when we might want to do so. But for most experiments, we should just be selecting the treatment that is most likely to be best.

In the code file, simulating many experiments, with and without differences between treatments, and comparing the different decision-making methods. You'll notice that the expected utility method for making decisions performs better when the goal is to pick the best option. Here's an example output from the notebook showing how many times stat sig and expected utility get the same, correct result, how many times only expected utility gets the correct result, and how many times only stat sig gets the correct result.

We see here that expected utility results in the same, correct decision in all the cases that statistical significance does, plus additional cases where stat sig would result in the incorrect decision.

In this particular simulation, the percentage of correct outcomes for stat sig and expected utility are 93.8% and 98.9% respectively. Again, we are taking "correct" to mean that we picked the better option or, in the case of no difference, picked an equally good option.