Challenge

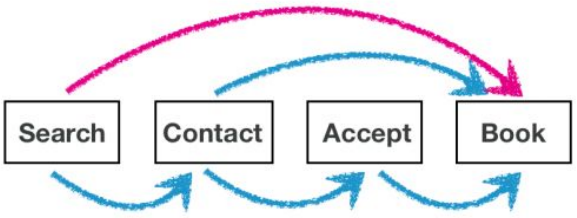
Since the professional photography service consumes so many operational and financial resources, AirBnB management are unsure if they should continue. AirBnB management have asked the Data Science team to analyse the impact of the professional photography service in order to determine whether or not they should continue funding the service.

- Provide full details about how you will run experiments to assess the impact of this service on both hosts and guests. How will you ensure that the experiments are valid and not biased?

A/B Test framework

1. Defining the customer journey and the KPI's

Metric	Δ	p
Search to Book	-0.31%	0.37
Search to Contact	**-1.29%**	0.04
Contact to Book	0.99%	0.06
Contact to Accept	**1.58%**	0.00
Accept to Book	-0.58%	0.11



On the image above we can see on the right the user flow. At the left a table with the main goal metrics for each step of the funnel.

The p values are just used to control significance of the tests, however I wouldn't stop the test when hitting the p-value but as Airbnb business is very dependable on time fluctuations I would leave it running for at least 2 months before taking any decission.

2. Splitting the test

I would also perform the A/B test using the following table:

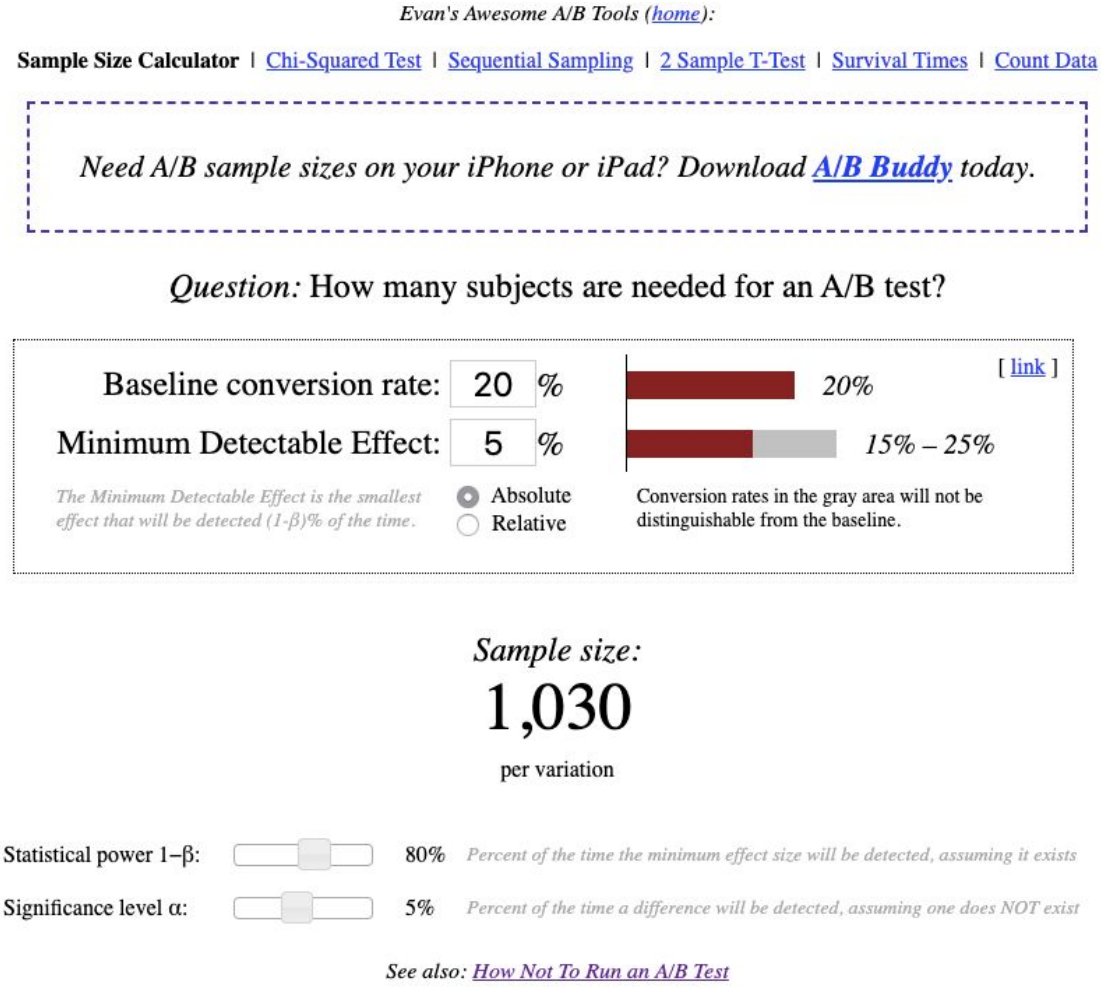
Alternative	Control	Variation	p-value
50%	50%	variation in the KPI	statistic signifcance
75%	25%	variation in the KPI	statistic signifcance
25%	75%	variation in the KPI	statistic signifcance
20%	80%	variation in the KPI	statistic signifcance

Evaluating the different values for different split sizes I would minimize the bias. That's why first row is for equally split and then I start playing with the splits in order to see how they change and their statistical significance (usually provided by third party tools).

Another idea to minimize the bias would be to run an A/A dummy test, where everyone sees the same type of photography and if it is neutral we should see very low variance in the KPI's.

3. Setting sample sizes

I used the Evan Miller Sample Size Calculator , to help me determine we need at least 1030 experiments per variation to make it significative.



I'm assuming a baseline conversion rate of 20% and want to detect at least a 5% variation on it. As Airbnb users decission making processes are long and diverse I want to see at least a 5% change (for better or worse) before taking any decission.