There are two contrasting views of Hypothesis testing. There is a group which is of the view that Hypothesis testing is essential for establishing statistical significance of a metric, and it is essential part of validating models. There is another group which believes that Hypothesis does not really prove or disprove a theory, but is merely a measure to check if there is sufficient data to be able to trust a model.

For now, let us assume that Hypothesis testing is useful, and narrate an example

Marketing Campaigns are done in the form of sending messages from a company to its customers with an objective of increasing sales. It is believed that when messages are sent, some of the customers see them, and few among them will be driven to make a purchase-typically over a period of time. So messages are sent to specific customers, and their purchases are tracked for a period of time. Any sale happening during the period is attributed to the message sent.

(it is assumed that there is only one marketing message during the period, if there are multiple messages then the sale is attributed either to the first or last message)

The attribution process is not always robust or fool-proof, and is driven by a lot of theory than actual facts. It is hard to believe that an SMS can drive crores of sales (RoI is absurdly high in these cases). In reality, the reasons can be different - Offers, Events like Birthday wedding etc., or any unknown needs of customers. While it is difficult to predict why a customer buys, it is possible to trigger a need for purchase in a customer by sending a communication at right time.

Nevertheless, it is important to have a strong measure for lift and attribution in order to establish whether a marketing campaign works or not. One way of measuring effectiveness of campaign is by creating test and control groups. So how do we create test and control groups? To understand this, we need to know the meaning of "Strength of Test". The article below gives good information -

<https://www.markhw.com/blog/control-size>

The strength of test is highest when sample sizes are equally split among test and control, and rapidly diminishes when control group goes below 20%. Depending on the total size of the sample set, any number between 50-20% would be ideal.

How to measure statistical significance of lift?

In this case,

H0: Response rate of Test Group = Response rate of Control Group (There is no significance difference in response rates between Test and Control Groups)

Ha : Response rate of Test Group is not equal to Response rate of Control Group

We use 2 sample z test for proportion and calculate Z value

Z = (p1-p2)/(sqrt(p(1-p)x(1/n1 + 1/n2))

Where p = (n1\*p1+n2\*p2)/(n1+n2)

P1= response proportion in test group

P2 = response proportion in control group

N1 = size of test group

N2 = size of control group

The value of Z for two tailed test at 95% significance level is 1.96. That is - there is a 95% chance that you will retain null hypothesis when it is true, if your Z statistic is 1.96 or more.

P value = chances of observing the test statistic value given that the null hypothesis is true. P value should be as low as possible for us to be confidently reject null hypothesis. It should definitely be < 0.05 base minimum for us to reject null hypothesis

Let us assume that you are unable to reject null hypothesis because Z value < 1.96

You can run an experiment where you keep p1, p2 constant, but increase N1 and N2, you will see that the Z value increases even if you increase the numbers in the original proportions. This clearly indicates that as the size of data grows even if the proportion of test and control group remains same, the significance increases.

Similarly, you can keep n1 and n2 constant and vary values of p1 and p2 such that p1-p2 increases, in this case, as p1-p2 increases, the significance value increases

For those who believe that Hypothesis testing never works -- we could say that hypothesis testing does not really provide conclusive evidence for accepting null or alternate hypothesis, rather it could be used as a measure to check if we have sufficient data to reject either