Context: The design of a user study in which we want to compare two systems for recommending products that people may be interested in buying, one of which uses a recommender system based on past purchase behavior. That recommender system first finds other people with similar shopping interests and then recommends products that those users have purchased (this is a so-called "user-item" recommender system). The other system works by simply suggesting the (same) most popular items to everyone. For example, if I buy a lot of pirate movies, the first system might recommend a pirate movie that many other pirate movie fans have bought (and that I haven't seen), but the second system might recommend the new Star Wars movie to me (and to everyone else who uses that second system) just because it is popular this week. The goal of this user study is to determine which system provides the more useful recommendations when used by fairly new users who have (up until now) purchased only a few (3 to 5) products.

Study Design:

Participants: Customers who have purchased 3 to 5 products. These customers will be randomly selected.

No of participants: More than 30, but 100 random customers should be good for normal distribution.

User Tasks: Construct a survey form by mixing top 10 recommendations from both systems. The products in the form should be coded so that the corresponding recommender can be identified later. Mixing recommended products from both the systems might avoid bias. The users will be asked to check any products that fits their interests. The goal of evaluation will be to compare the average precision score of each system at N (here N=10). In this case, it will be more appropriate to compute the precision score of the system because we will have an idea of what the user is likely to buy through survey forms. The survey results will help to sort out relevant items from the recommended list and hence help to evaluate precision of the system. However, in cases when we have no way to know what the user is likely to buy (relevant products) and only have information of what the user has (already) bought, recall may be a more appropriate measure. However, it is also important to note that we can always achieve a perfect recall score by simply giving all of the items.

Data Processing: Let us say that a recommended product is "Hit" if the customer is interested in the product and "Miss" if not. The number of "hit" divided by n=10 is a precision score. Let us also denote the two recommender systems as R1 and R2

-On the basis of survey results, for each customer, find precision for each recommender system.

-Find the difference in precision score between R1 and R2 for each customer. Let us denote this by P_{diff}.

-Find mean P_{diff}. Lets denote this as X_{diff}. Also find standard deviation of P_{diff}. Lets denote this by S_{diff}.

-Let μ_{diff} denote the mean precision difference in population. Now, in order to compare two recommender systems, let us set hypothesis as following:

Null hypothesis (Ho): $\mu_{diff} = 0$; (there is no difference between two systems)

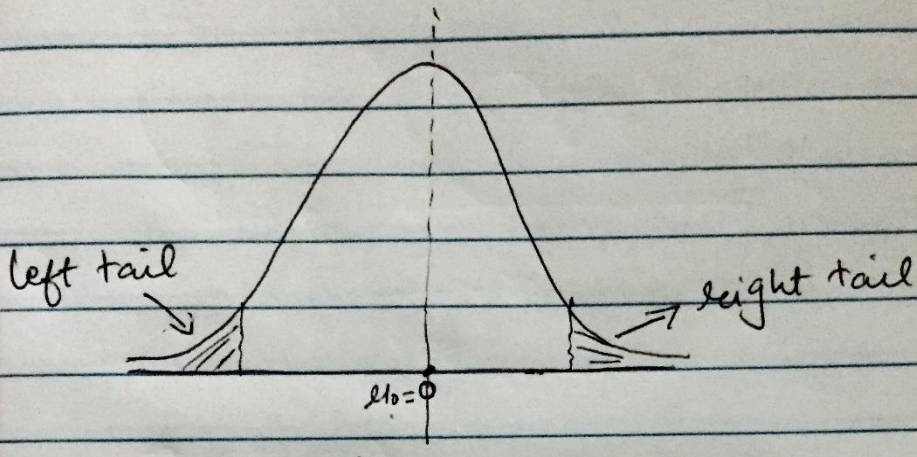
Alternate hypothesis Ha $\mu_{diff} \neq 0$ (there is a difference).

-Compute standard error associated with x_{diff} using the standard deviation of differences (S_{diff}) for n=100 using the formulae: $SEx_{diff} = S_{diff}$ / **(**sq. root(n))

-Compute Z score of x_{diff} under the null condition that the actual mean difference is zero using the formulae:

$Z = (X_{diff} - 0) / (SEx_{diff})$

-Using the Z score table, find the corresponding p-value. Consider the following figure for further computations.



- Compute $p(z)$ using z score table.
- If $2 * p(z) < 0.05$; we reject null hypothesis.
- If $z > 0$, then ~~it falls~~ test statistic falls on Right tail. This suggests $\mu_1 > \mu_2$; which in turn suggests that (statistically) R_1 on average gives better results than R_2.

- If $z < 0$; then test statistic falls on left tail. This suggests $\mu_1 < \mu_2$; which in turn suggests that (statistically) R_2 on average gives better results than R_1.