Statistical Inference

B. Statistical Data Science 2nd Year Indian Statistical Institute

Teacher: Soham Sarkar

Exercise Series 1

In today's exercise session, we will revisit the notions of *population*, *sample*, *statistical models* and *inference*. In the following, we have some scenarios, where we have some data and are faced with certain questions. In each case, determine a suitable statistical model for the scenario and identify the parameter(s) of the model. Specify the parameter space. In each scenario, we encounter some problem. Cast the problem as a statistical inference problem and classify it into (a) estimation or (b) test of hypothesis. Clearly specify the inference problem in terms of the model parameter(s). Identify the relevant and nuisance parameter(s) for the inference problem.

Scenario 1. In a manufacturing company, an item produced by a machine can either be defective or non-defective. The company wants to find the chance of an item being defective. You know the status (defective/non-defective) of 50 items produced by the machine.

The company has the following policy: If the chance of an item being defective is determined to be more than 1%, the production would be halted and the machine would be re-calibrated (which costs time and money). How would you approach if you are asked to determine whether to halt production?

Scenario 2. Again consider the manufacturing company in Scenario 1. In this case, we inspect the items one by one and stop as soon as we see 5 defective items. We record the number of items inspected to get 5 defective items. How would you modify your statistical model for this scenario?

Scenario 3. The traffic department in a city is interested in the number of accidents per day at a particular junction. You have records of the daily number of accidents for the past month.

If there is more than 50% chance of having at least one accident at the junction, then the traffic department will position a traffic control station there. How would you cast this as a statistical inference problem?

Scenario 4. A new vaccine is produced by a pharmaceutical company and they want to study its efficacy (probability of cure) against a disease. The results (cured/not cured) of applying the vaccine on $10,000$ patients are known to you.

A standard vaccine is known to have 75% efficacy against the disease and the government also distributes the same for free to its populace. The government may change to the new vaccine if its efficacy is more than the existing one.

Scenario 5. The government is interested in the average household income (in rupees) of a particular district. A survey is conducted where 500 households from the district are selected and their incomes are noted.

The government wants to know the household income profile: The income threshold with x% households below that income, for $x = 10, 20, \ldots, 90$.

Scenario 6. An LED bulb manufacturing company is interested in the lifetime of its bulbs. The lifetimes of $1,000$ such bulbs (in hours) are recorded.

The company replaces the bulb purchased by a customer if it stops working within 1 year from purchase. The company wants to check whether they need to issue more than 5% replacements.

Scenario 7. Two different brands of tires are available in the market. We have access to the longevities (in kms run before wearing out) of 75 tires of Brand A and 60 tires of Brand B.

A customer wants to know which brand of tires is likely to last longer to decide which one to choose.

Scenario 8. A new diet for weight loss has been introduced. A group of 100 volunteers are chosen and given the diet. The weights of these volunteers at the time of beginning the diet and six months into the diet are recorded.

A dietitian wants to verify whether the new diet is effective before recommending it to others. How would you help the dietitian in determining whether to recommend the new diet?

Scenario 9. A telephone company wants to determine the need of one of its towers, tower T. The number of calls (both incoming and outgoing) passing through tower T at each minute-interval for the last 30 days are available.

If at any point of time, the number of calls passing through the tower is more than a particular threshold (e.g., 1000), then the excess calls are dropped. The company wants to ensure that no more than 0.1% of the calls are dropped.

Scenario 10. Consider the *German Tank* problem. Tanks are made and numbered serially as $1, 2, \ldots$. If we spot a tank, then we are able to note down its serial number. We want to know the total number of tanks.

Scenario 11. We want to find the number of elephants in a jungle. To achieve this, we first catch 10 elephants, tag them as $1, 2, \ldots, 10$, and let them return to the jungle. After one month, we capture 15 elephants and find out how many of them are tagged. This technique is known as *capture-recapture*.

Scenario 12. A project team in a company takes a certain amount of time to complete a project assigned to them. The time taken by the team (in hours) to complete their last 5 projects are recorded.

The CEO of the company wants to know how many new projects can be assigned to the team, all of which should be completed by the current month.