# Unit - 2 : Statistical Analysis :

1) What are statistics?

- Statistic is the process identifying, collecting, organizating, analysis, interpretation and presentation of data.
- The goal of this field is to try to explain & model the world arround us.
  To do that we have to take a look at the population.
- Statistics are the numbers you always see on the news and in the paper.
- Population : We can define a population as the entire pool of subjects of an experiment or a model.
- Sample : We will take a sample of the population.
  - It is sub-set of population.
- Parameter :
  - if your popopulation is all your employees, you want know what percentage of them drinks Alcohol.
  - This question is called a parameter.

Example: Suppose we selected a random sample of 100 students from a school with 1000 students. The average height of the sampled students would be an example of a statistic.

2) How do we obtain & sample data?

- If let's focus on ways of obtaining & sampling data.

* Obtaining data:

- there are 2 main way of collecting data for our analysis:

i) Observational

ii) Experimental

i) Observational:

- We may obtain data through observational.
- which consist of measuring specific characteristics.
- but not attempting to modify the subject being studied.

Example:

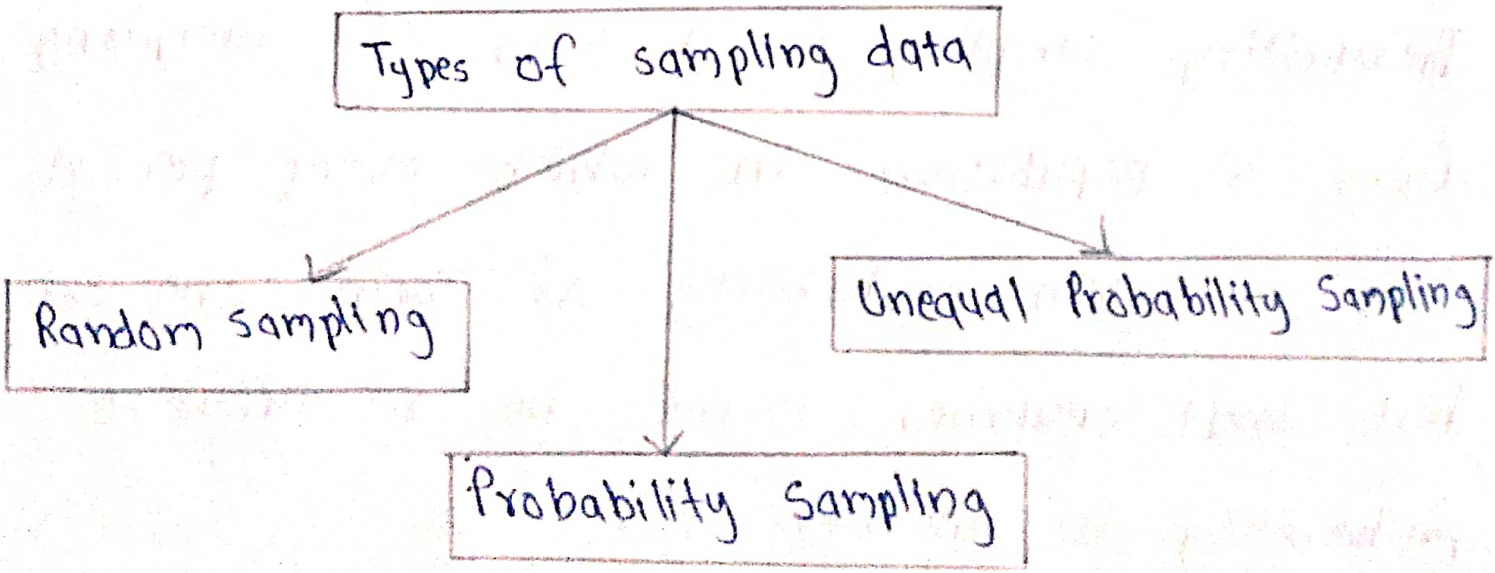
Tracking software on website that observes users behavior on the website such as length of time

spent on certain pages & rate of clicking on ads, this all not affect user's experience. It's is observational study.

ii) Experimental :

- A experiment consist of a treatment and observation it's effect on the subjects.
- Subjects are experimental Unit.
- This is usually how most scientific labs collect data.
- They will put people into 2 or more groups and call them the Control & the experimental group.

3) Sampling data :



ii) Random Sampling :

- Suppose that we are running an A/B test and we need to figure out who will be in group A & who will be in group B.
- There are the following 3 suggestions from your data team :

① Seperate users based on locations.

② Seperate users based on the time of day they visit the site.

③ Make it completely random.

iii) Probability sampling :

Probability sampling is a way of sampling from a population in which every person has a known probability of being choosen but that number might be a different probability to another user.

iii) Unequal Probability Sampling :

- Suppose we are interested in measuring the happiness level of our employees.
- We already know that we can't ask every single person on the staff because that would be silly and exhausting.
- So we need to take a sample.

4) Point estimates :

- A point estimate is an estimate of a population parameter based on sample data.
- We use point estimates to estimate population means, variances & other statistics.
- Example :

Suppose there is a company of 9000 employees & we are interested in ascertaining the average length of breaks taken by the employees in a single day.

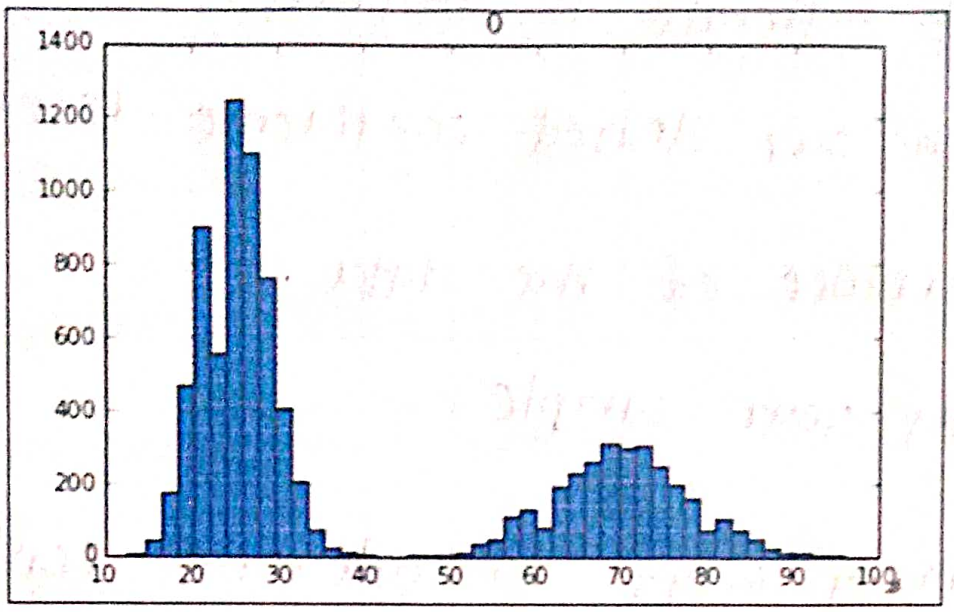
- As we probably cannot ask every single person we will take a sample of the 9000 people and take a mean of the sample.
- This sample mean will be our point estimate.

5) Sampling Distribution.

- Our procedure for creating a sample distribution will be the following:

1) Take 500 different samples of the break time of size 100 each.
2) Take a histogram of these 500 different point estimates.

- The number of elements in the sample (100)
- The number of samples i took (500)




6) Confidence intervals:

- A confidence interval is a range of values based on a point estimate that contains the true population parameter at some confidence level.
- Calculating a confidence interval involves:
  - finding a point estimate
  - incorparating a margin of error to create a range.

-

- The margin of error is a value that represents:
  - Our point estimate is accurate
  - Based on our desired confidence level
  - The variance of the data.
  - How big your sample.
- There many ways to calculate confidence intervals:

  for the purpose of brevity & simplicity, for this confidence intervals we need following

  - A point estimate.
  - A point estimate of the population standard deviation, which represends the variance in the data.
  - The degrees of freedom.

7) Hypothesis test :

- Hypothesis test are one of the most widely used tests in statistics.
- A hypothesis test is a statistical test that is used to ascertain whether we are allowed to assume that a certain condition is true for the entire population, given a data sample.
- Hypothesis test is a test for a certain hypothesis that we have about an entire population.
- Result of the test tells us wheather we should believe the hypothesis or reject it for an alternative one.

* Conducting a hypothesis test.

- There are five basic steps that most hypothesis tests follows:

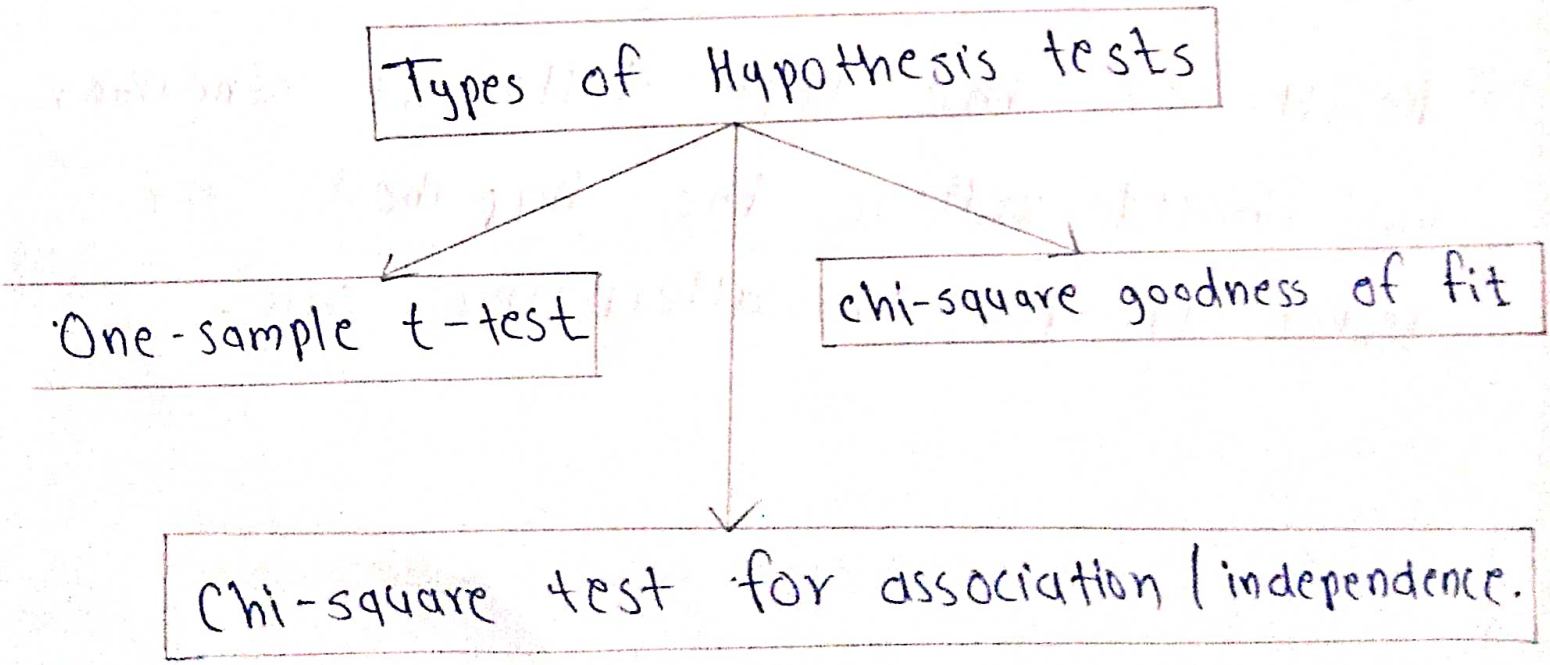
① Specify the hypotheses:

② Determine the sample size of the test sample.

③ choose a significance level

④ collect the data.

⑤ Decide whether to reject or fail to reject the null hypothesis.



## ii) One-Sample t-tests:

- The one-sample t-test is a statistical test used to determine whether a quantitative data sample differs significantly from another dataset.

- Example

  Example of employee break time.

- long_breaks_in_engineering = stats.poisson.rvs (loc=10, mu=55, size=100)

- short_breaks_in_engineering = stats.poisson.rvs (loc=10, mu=15, size=300)

- engineering_breaks = np.concatenate (long_breaks_in_engineering, short_breaks_in_engineering))

- Print breaks.mean()
  # 39.99

- Print engineering_breaks.mean()
  # 34.825.

## 2] Chi-square goodness of fit test.

- The chi-square goodness of fit test is very similar to the one sample t-test in that it tests whether the distribution of the sample data matches an expected distribution, while the big difference is that it is testing for categorical variables.
- For example, a chi-square goodness of fit test would be used to see if the race demographics of your company match that of the entire city of the U.S. population.
- It can also be used to see if users of your website show similar characteristics to average Internet users.
- As we are working with categorical data, we have to be careful because categories like "male", "female," or "other" don't have any mathematical meaning.
- Therefore, we must consider counts of the variables rather than the actual variables themselves.
- In general, we use the chi-square goodness of fit test in the following cases:
  - We want to analyse one categorical variable from one population
  - We want to determine if a variable fits a specified or expected distribution
- In a chi-square test, we compare what is observed to what we expect.

## 3) Chi-square test for association/independence

- Independence as a concept in probability is when knowing the value of one variable tells you nothing about the value of another.
- For example, we might expect that the country and the month you were born in are independent.

- However, knowing which type of phone you use might indicate your creativity levels. Those variables might not be independent.
- The chi-square test for association/independence helps us ascertain whether two categorical variables are independent of one another.
- The test for independence is commonly used to determine whether variables like education levels or tax brackets vary based on demographic factors, such as gender, race, and religion.
- Let's look back at an example posed in the preceding chapter, the A/B split test.
- Recall that we ran a test and exposed half of our users to a certain landing page (Website A), exposed the other half to a different landing page (Website B), and then, measured the sign up rates for both sites.