**Probability tasks**

1. Sampling Methods :

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population.

1. Probability sampling : is a sampling technique in which researchers choose samples from a larger population using a method based on the theory of probability. This sampling method considers every member of the population and forms samples based on a fixed process.

* Simple random sampling : every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.
* Cluster sampling: the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc.
* Systematic sampling: choose the sample members of a population at regular intervals.
* Stratified random sampling: a method in which the researcher divides the population into smaller groups that don’t overlap but represent the entire population.

Ex-🡪  a researcher looking to analyze the characteristics of people belonging to different annual income,  By doing this, the researcher concludes the characteristics of people belonging to different income groups. Marketers can analyze which income groups to target and which ones to eliminate.

1. Non-probability sampling : method that involves a collection of feedback based on a researcher or statistician’s sample selection capabilities and not on a fixed selection process. In most situations, the output of a survey leads to skewed results, which may not represent the desired target population.

* Convenience sampling: startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the mall entrance and giving out pamphlets randomly.
* Judgmental or purposive sampling: when researchers want to understand the thought process of people interested in studying for their master’s degree. The selection criteria will be: “Are you interested in doing your masters in …?” and those who respond with a “No” are excluded from the sample.
* Snowball sampling: researchers apply when the subjects are difficult to trace.
* Quota sampling: sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.

1. Statistical test types :

|  |  |
| --- | --- |
| **Type of Test** | **Use** |
| **Correlational**: these tests look for an association between variables |  |
| Pearson Correlation | Tests for the strength of the association between two continuous variables |
| Spearman Correlation | Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normally distributed data) |
| Chi-Square | Tests for the strength of the association between two categorical variables |
| **Comparison of Means**: these tests look for the difference between the means of variables |  |
| Paired T-Test | Tests for the difference between two variables from the same population (e.g., a pre- and posttest score) |
| Independent T-Test | Tests for the difference between the same variable from different populations (e.g., comparing boys to girls) |
| ANOVA | Tests for the difference between group means after any other variance in the outcome variable is accounted for (e.g., controlling for sex, income, or age) |
| **Regression**: these tests assess if change in one variable predicts change in another variable |  |
| Simple Regression | Tests how change in the predictor variable predicts the level of change in the outcome variable |
| Multiple Regression | Tests how changes in the combination of two or more predictor variables predict the level of change in the outcome variable |
| **Non-Parametric**: these tests are used when the data does not meet the assumptions required for parametric tests |  |
| Wilcoxon Rank-Sum Test | Tests for the difference between two independent variables; takes into account magnitude and direction of difference |
| Wilcoxon Sign-Rank Test | Tests for the difference between two related variables; takes into account the magnitude and direction of difference |
| Sign Test | Tests if two related variables are different; ignores the magnitude of change—only takes into account direction |

1. Hypothesis testing real life examples: the process used to evaluate the strength of evidence from sample.

### **EX1)** Clinical Trials :  determine whether some new treatment, drug, procedure, etc. causes improved outcomes in patients. suppose a doctor believes that a new drug is able to reduce blood pressure in obese patients. To test this, he may measure the blood pressure of 40 patients before and after using the new drug for one month

* **H0:** μafter = μbefore (the mean blood pressure is the same before and after using the drug)
* **HA:** μafter < μbefore (the mean blood pressure is less after using the drug)

If the p-value of the test is less than some significance level (e.g. α = .05), then he can reject the null hypothesis and conclude that the new drug leads to reduced blood pressure.

### **EX2)  Advertising Spend:**  determine whether or not some new advertising campaign, marketing technique, etc. causes increased sales. suppose a company believes that spending more money on digital advertising leads to increased sales. To test this, the company may increase money spent on digital advertising during a two-month period and collect data to see if overall sales have increased.

They may perform a hypothesis test using the following hypotheses:

* **H0:** μafter = μbefore (the mean sales is the same before and after spending more on advertising)
* **HA:** μafter > μbefore (the mean sales increased after spending more on advertising)

If the p-value of the test is less than some significance level (e.g. α = .05), then the company can reject the null hypothesis and conclude that increased digital advertising leads to increased sales.

1. Confidence interval : range of values we are fairly sure our true value lies in.
2. P-value : a probability measure of finding the observed, or more extreme, results, when the null hypothesis of a given statistical test is true.
3. predictive analysis:  the use of data, statistical algorithms and [machine learning](https://www.sas.com/en_us/insights/analytics/machine-learning.html) techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future history.

EX) predicting suicide attempts : Suicide is [the tenth leading cause of death](https://healthitanalytics.com/news/machine-learning-uses-ehr-data-to-predict-suicide-attempt-risk) in the US, taking lives of 14 Americans out of 100,000 annually. To improve this sad situation, a research team at the Vanderbilt University Medical Center (VUMC) developed a predictive analytics model that uses patients’ electronic health records to forecast the likelihood of suicide attempts by particular patients. The tool was tested at VUMC for 11 months. While doctors were receiving patients, the algorithm was running in the background, predicting the risk of patients returning for attempted suicide treatment.

EX) **Predicting and decreasing churn** : Churn, or attrition, rate indicates the percentage of customers who abandon a brand or discontinue a service. To keep it low, companies can implement predictive analytics and identify signals that may indicate that a customer is likely to drop out. With this intelligence, businesses can act in time, turning up the value of their services to retain hesitant customers.

1. Real life Examples of Probability:

**EX1) Weather :** A probability forecast is an assessment of how likely an event can occur in terms of percentage and record the risks associated with weather. Meteorologists around the world use different instruments and tools to predict weather changes. They collect the weather forecast database from around the world to estimate the temperature changes and probable weather conditions for a particular hour, day, week, and month.

**EX2) Insurance :** insurance companies use the theory of probability or theoretical probability for framing a policy or completing at a premium rate. The theory of probability is a statistical method used to predict the possibility of future outcomes.

### 