**INFERENTIAL STATS**

**3 Probability Sampling Techniques**

When each entity of the population has a definite, non-zero probability of being incorporated into the sample, the sample is known as a probability sample.

Probability samples are selected in such a way as to be representative of the population. They provide the most valid or credible results because they reflect the characteristics of the population from which they are selected.

Probability sampling techniques include random sampling, systematic sampling, and stratified sampling.

[](https://humansofdata.atlan.com/wp-content/uploads/2017/07/1-random-image-673-1.jpg)

**When:** There is a very large population and it is difficult to identify every member of the population.

**How:**The entire process of sampling is done in a single step with each subject selected independently of the other members of the population. The term random has a very precise meaning and you can’t just collect responses on the street and have a random sample.

**Pros:**In this technique, each member of the population has an equal chance of being selected as subject.

**Cons:**When there are very large populations, it is often difficult to identify every member of the population and the pool of subjects becomes biased. Dialing numbers from a phone book for instance, may not be entirely random as the numbers, though random, would correspond to a localized region. A sample created by doing so might leave out many sections of the population that are significant to the study.

**Use case:**Want to study and understand the rice consumption pattern across rural India? While it might not be possible to cover every household, you could draw meaningful insights by building your sample from different districts or villages (depending on the scope).

[](https://humansofdata.atlan.com/wp-content/uploads/2017/07/2-systematic-image-673-1.jpg)

**When:** Your given population is logically homogenous.

**How:** In a systematic sample, after you decide the sample size, arrange the elements of the population in some order and select terms at regular intervals from the list.

**Pros:** The main advantage of using systematic sampling over simple random sampling is its simplicity. Another advantage of systematic random sampling over simple random sampling is the assurance that the population will be evenly sampled. There exists a chance in simple random sampling that allows a clustered selection of subjects. This can be avoided through systematic sampling.

**Cons:** The possible weakness of the method that may compromise the randomness of the sample is an inherent periodicity of the list. This can be avoided by randomizing the list of your population entities, as you would randomize a deck of cards for instance, before you proceed with systematic sampling.

**Use Case**: Suppose a supermarket wants to study buying habits of their customers. Using systematic sampling, they can choose every 10th or 15th customer entering the supermarket and conduct the study on this sample.

[](https://humansofdata.atlan.com/wp-content/uploads/2017/07/3-stratified-image-673-1.jpg)

**Stratified Sampling**

**When:** You can divide your population into characteristics of importance for the research.

**How:**A stratified sample, in essence, tries to recreate the statistical features of the population on a smaller scale. Before sampling, the population is divided into characteristics of importance for the research — for example, by gender, social class, education level, religion, etc. Then the population is randomly sampled within each category or stratum. If 38% of the population is college-educated, then 38% of the sample is randomly selected from the college-educated subset of the population.

**Pros:**This method attempts to overcome the shortcomings of random sampling by splitting the population into various distinct segments and selecting entities from each of them. This ensures that every category of the population is represented in the sample. Stratified sampling is often used when one or more of the sections in the population have a low incidence relative to the other sections.

**Cons:**Stratified sampling is the most complex method of sampling. It lays down criteria that may be difficult to fulfill and place a heavy strain on your available resources.

**Use Case:**If 38% of the population is college-educated and 62% of the population have not been to college, then 38% of the sample is randomly selected from the college-educated subset of the population and 62% of the sample is randomly selected from the non-college-going population. Maintaining the ratios while selecting a randomized sample is key to stratified sampling.

**3 Non-Probability Sampling Techniques**

Non-probability sampling techniques include convenience sampling, snowball sampling and quota sampling.

In these techniques, the units that make up the sample are collected with no specific probability structure in mind. The selection is not completely randomized, and hence the resultant sample isn’t truly representative of the population.

[](https://humansofdata.atlan.com/wp-content/uploads/2017/07/4-convenience-image-673-1.jpg)

**When:** During preliminary research efforts.

**How:**As the name suggests, the elements of such a sample are picked only on the basis of convenience in terms of availability, reach and accessibility.

**Pros:** The sample is created quickly without adding any additional burden on the available resources.

**Cons:**The likelihood of this approach leading to a sample that is truly representative of the population is very poor.

**Use Case:** This method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample.

[](https://humansofdata.atlan.com/wp-content/uploads/2017/07/5-snowball-image-673-1.jpg)

**When:** When you can rely on your initial respondents to refer you to the next respondents.

**How:** Just as the snowball rolls and gathers mass, the sample constructed in this way will grow in size as you move through the process of conducting a survey. In this technique, you rely on your initial respondents to refer you to the next respondents whom you may connect with for the purpose of your survey.

**Pros:** The costs associated with this method are significantly lower, and you will end up with a sample that is very relevant to your study.

**Cons:** The clear downside of this approach is that you may restrict yourself to only a small, largely homogenous section of the population.

**Use Case:**Snowball sampling can be useful when you need the sample to reflect certain features that are difficult to find. To conduct a survey of people who go jogging in a certain park every morning, for example, snowball sampling would be a quick, accurate way to create the sample.

[](https://humansofdata.atlan.com/wp-content/uploads/2017/07/6-quota-image-673-1.jpg)

**When:** When you can characterize the population based on certain desired features.

**How:**Quota sampling is the non-probability equivalent of stratified sampling that we discussed earlier. It starts with characterizing the population based on certain desired features and assigns a quota to each subset of the population.

**Pros:** This process can be extended to cover several characteristics and varying degrees of complexity.

**Cons:** Though the method is superior to convenience and snowball sampling, it does not offer the statistical insights of any of the probability methods.

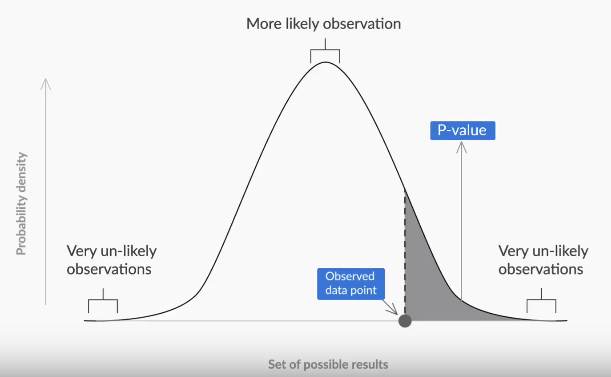
**Use Case:**If a survey requires a sample of fifty men and fifty women, a quota sample will survey respondents until the right number of each type has been surveyed. Unlike stratified sampling, the sample isn’t necessarily randomized.

Probability sampling techniques are superior, but the costs can be prohibitive. For the initial stages of a study, non-probability sampling techniques might be sufficient to give you a sense of what you’re dealing with. For detailed insights and results that you can bank upon, move on to the more sophisticated techniques as the study gathers pace and takes a more concrete structure.

**HYPOTHESIS TESTING**

**p-value** as the **probability that the null hypothesis** will not be rejected. This statement is not the technical (or formal) definition of p-value; it is used for better understanding of the p-value.

The higher the p-value, the higher is the probability of failing to reject a null hypothesis. And the lower the p-value, the higher is the probability of the null hypothesis being rejected.



**Figure 1 -Interpretation of p-value**

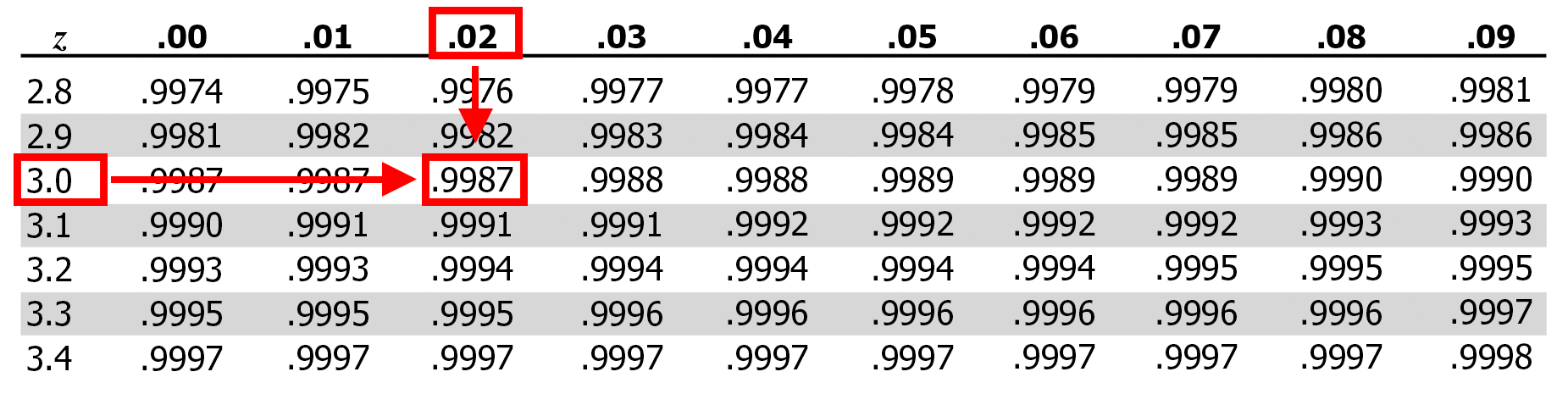
After formulating the null and alternate hypotheses, the steps to follow in order to**make a decision** using the **p-value method** are as follows:

1. Calculate the value of the z-score for the sample mean point on the distribution.
2. Calculate the p-value from the cumulative probability for the given z-score using the z-table.
3. Make a decision on the basis of the p-value (multiply it by 2 for a two-tailed test) with respect to the given value of α (significance value).

To find the correct p-value from the z-score, find the **cumulative probability** first, by simply looking at the z-table, which gives you the area under the curve till that point.

**Situation 1:**The sample mean is on the right side of the distribution mean (the z-score is positive).

**Example:**z-score for sample point = + 3.02



**Figure 2 - z-table for positive z-scores**

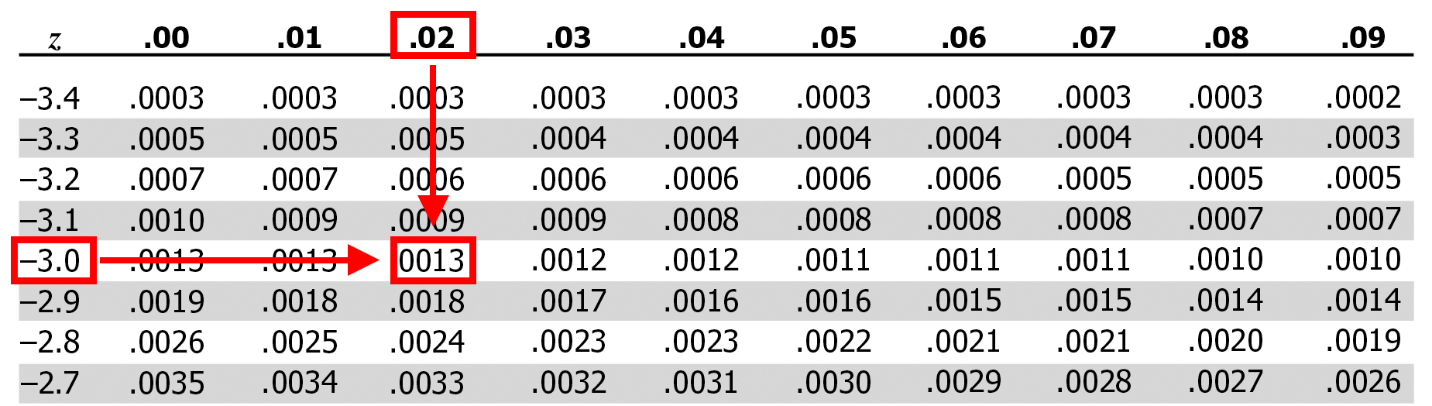
Cumulative probability of the sample point = 0.9987

For a one-tailed test: p = 1 - 0.9987 = 0.0013

For a two-tailed test: p = 2 (1 - 0.9987) = 2 \* 0.0013 = 0.0026

**Situation 2:**The sample mean is on the left side of the distribution mean (the z-score is negative).

**Example:** The z-score for the sample point = -3.02          



**Figure 3 - z-table for negative z-scores**

Cumulative probability of the sample point = 0.0013

For a one-tailed test: p = 0.0013

For a two-tailed test: p = 2 \* 0.0013 = 0.0026

**Let’s solve the following problem stepwise** to consolidate your learning on how to make a decision about any hypothesis using the p-value method.

You are working as a data analyst at an auditing firm. A manufacturer claims that the average life of its product is 36 months. An auditor selects a sample of 49 units of the product and calculates the average life to be 34.5 months. The population standard deviation is 4 months. Test the manufacturer’s claim at a 3% significance level using the p-value method.

First, **formulate the hypotheses** for this two-tailed test, which would be:

                                   H₀: μ = 36 months and H₁: μ ≠ 36 months