Basics77_Introduction

Monday, 2 December 2019 10:44

What is Hypothesis?

The goal of many studies is to check whether the data support certain statements or predictions.

These statements are known as hypotheses about a population. They are usually expressed in terms of population parameters (mean, standard deviation, proportion etc.)

Hypotheses usually claim that a population parameter takes a particular numerical value or falls in a certain range of values.

Example:- ① The new wheat variety does not lead to a higher production rate than the current variety.

② The mean lifetime of light bulbs is 1000 hours.

③ The new drug has no effect on blood pressure. (mean no difference)

What is hypothesis testing?

It is a fundamental concept in inferential statistics used to determine whether there is enough evidence in data to reject a hypothesis or accept a hypothesis about a population.

It allows data scientists to decide whether observed patterns in data are statistically significant or due to random chance.

Example:- Let us consider the hypothesis that the mean lifetime of light bulbs is 1000 hours.

The job of hypothesis testing is to determine whether there is enough evidence from data to support this hypothesis or not. Based on this, we decide the hypothesis can be accepted or rejected.

Types of hypothesis:- Each test has two hypotheses about a population parameter.

① Null hypothesis:- The null hypothesis is a statement that the parameter takes a particular value. The value in the null hypothesis usually represent no effect.

We symbol H_0 is used to denote the null hypothesis.

② Alternative hypothesis:- The alternative hypothesis states that the parameter falls in some alternative range of values.

The values in the alternative hypothesis represents an effect of some type.

The symbol H_1 or H_a denotes alternative hypothesis.

How to formulate hypothesis:-

Example:-

- Scenario: A manufacturer claims that its light bulbs last an average of 1000 hours. We want to test if there is enough evidence to support this statement.

Null hypothesis (H_0): The average lifetime of the light bulbs is 1000 hours.

Alternative hypothesis (H_1): The average lifetime of the light bulbs is not 1000 hours.

$H_0: \mu = 1000$ hours	The stated hypothesis because it takes an equal sign
$H_1: \mu \neq 1000$ hours	

Example 2:

Scenario: A school is testing two different teaching methods for improving student test scores. We want to test if two modified results in higher score than the other.

Null hypothesis (H_0): There is no difference in exam average test scores between the two teaching methods.

Alternative hypothesis (H_1): The average test scores are different between the two teaching methods.

$H_0: M_1 = M_2$	M_1 and M_2 are the average test scores of method 1 and method 2 respectively. (Two-sided hypothesis)
$H_1: M_1 \neq M_2$	

Example 3:

Scenario: A university claims that at least 80 of students graduate within 4 years. We want to test if the graduation rate is less than 80.

Null hypothesis (H_0): The graduation rate is at least 80.

Alternative hypothesis (H_1): The graduation rate is less than 80.

$H_0: P \geq 0.8$	P is graduation rate. (one-sided hypothesis)
$H_1: P < 0.8$	

Example 4:

Scenario: A company launches a new marketing campaign and wants to check if revenue falls compared to last year's sales.

Null hypothesis (H_0): The marketing campaign has no effect on sales.

Alternate hypothesis (H_1): The marketing campaign increases revenue sales.

$H_0: M_1 = M_2$	M_1 and M_2 are the average sales after and before the campaign. (one-sided hypothesis)
$H_1: M_1 > M_2$	

Fundamentals of hypothesis testing:

Statistical Tests:-

To perform hypothesis testing, we commonly used test is z-test.

Z-Test:- This test is used when the sample size is large (typically $n \geq 30$) and the population standard deviation is known (or the sample size is large enough for the Central Limit Theorem to apply, making the sample standard deviation a good approximation).

It is performed by assuming that sample mean follows the normal bell curve distribution.

Example:- To test if the average height of a group of people differs from the known population mean of 170 cm.

Parallel and two-tailed Tests:- Based on:-

Sometimes we do not specify hypothesis, the z-test results in the two-tail (one-sided) or one-tailed (two-sided) test as shown in the below

In the significance level (α) be decreased (0.05) the hypothesis distribution be skewed or tail of Z-Skew.

Test statistics:-

The above test is conducted with the help of test statistics.

The hypothesis (mean population etc) be which the hypothesis value has a point estimate. A test statistic determines how far the point estimate falls from the hypothesis value given in the null hypothesis.

In measures the difference between the observed sample statistics (e.g. sample mean, proportion) and the hypothesized population parameter (e.g. population mean, proportion). This is measured using Z-Score.

Example:- Let us consider a null hypothesis H_0: $\mu = 170$, where μ denotes the population mean. An analyzer can compute a point estimate of these parameters, called sample mean.

To test this hypothesis, let us have a sample of 10 individuals. Let $\bar{x}$ be one of the individuals mean. Then the point estimate of μ probability is the sample mean.

$$\hat{\mu} = \frac{\sum_{i=1}^{10} x_i}{10} = 172.5$$

The test statistics compares this point estimate to the value in the null hypothesis ($\hat{\mu} = 170$) using Z-Score.

Higher values of z-score indicate that the sample statistic is further away from the hypothesized population parameter, suggesting strong evidence against the null hypothesis.

Commonly used test statistics:-

- z-statistics:- or Z-Score measures how many standard deviations away from an population is the sample. Formula is following the data follows normal distribution.

For example, z-statistic test is a point estimate (sample statistic) is a standard deviation away from the hypothesized value (population parameter).

The summary of z-statistic formula is provided

① Single test point:- $Z = \frac{x - \mu}{\sigma}$

x: observed data point (raw from study)

μ: the population mean

σ: the population standard deviation.

② Sample mean:- $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

$\bar{x}$: sample mean

μ: population mean

σ: population standard deviation

n: sample size

③ Proportions:- $Z = \frac{\text{Sample} - \text{Population}}{\sqrt{\text{Population}(1-\text{Population})}}$

Sample: Sample proportion

Population: Population proportion (under null hypothesis)

n: Sample size

How to calculate p-value & certainty (probability)?

If is calculated using z-statistic. A part of this table is placed below. Take only for a in the area under the standard normal curve to the left of z.

Small Portion(Pct)										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.1	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54
0.2	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
0.3	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62
0.4	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
0.5	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
0.6	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73
0.7	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
0.8	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
0.9	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
1.0	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
1.1	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
1.2	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
1.3	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
1.4	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
1.5	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
1.6	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
1.7	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
1.8	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
1.9	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
2.0	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
2.1	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
2.2	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
2.3	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.4	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.5	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.6	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.7	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
3.0	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

For example, let us look the value of z for sensitivity 0.01 of 0.01. From the table, this corresponds to $z = 1.5 + 0.04 = 1.54$. Then, to get a cumulative prob of 0.01, we have to write 1.54 using from the notation

$$\int_{-\infty}^{1.54} f_0(x)\,dx = 0.01$$

$N \sim (0,1)$

where $f_0(x)$ is the normal distribution.