Hypothesis Testing Part-I

Fundamental concepts of hypothesis testing:

Statistical inference may be divided into two major areas: estimation and testing of hypothesis. While the estimation is the procedure of making judgement about a population parameter, the testing procedure is an attempt to arrive at a correct decision about a pre-stated hypothesis concerning a population parameter. In both the cases, knowledge of the sampling distribution of the estimator under consideration is required to establish the degree of accuracy of our estimate.

To distinguish between the two areas, namely estimation and hypothesis testing, let us consider the following example:

An Internet server claims that computer users in the country spend on an average 15 hrs. per week on browsing. To verify if this were overstated, a competitor conducted a survey of 250 users. The average time in the sample could be used as an estimate of the true mean, which is unknown. This problem falls in the area of estimation. If on the other hand, we are interested in arriving at a decision on whether the claim of the internet server is justified, then the problem falls in the area of hypothesis testing. In this case, we donot attempt to estimate a parameter, but instead, we try to arrive at a correct decision based on the sample of 250 users. The server's claim in this instance is referred to as a statistical hypothesis.

Def^n of statistical hypothesis :

A statistical hypothesis is an assertion or statement about a population or equivalently about the probability distribution characterizing a population, which we want to verify on the basis of information contained in a sample.

Examples :

① A physician may hypothesize that the recommended drug is effective in 90% cases.

② A nutritionist claims that at most 75% of the preschool children in a certain country have protein-deficient diets.

③ An administrator of a business farm claims that the average work efficiency of his workers is at least 90%.

(4) A sewing machine company claims that their new machine is superior to the one available in the market.

(5) The court assumes that the indicted person is innocent.

From the foregoing discussion, we might be tempted to know how a hypothesis testing differs from a scientific investigation. We enumerate below a few of these points that attempt to clarify this issue.

In a scientific observation, a scientist

- Observes nature
- Formulates a theory
- Test the theory against the observations

In a hypothesis testing, a researcher

- ✓ Sets up a hypothesis concerning one or more population parameters.
- ✓ Selects samples from the population and th
- ✓ Compares his observations against the stated hypothesis.

And finally the researcher

- Rejects the hypothesis if the observation disagree with the stated hypothesis, if not
- Concludes that either the hypothesis is true or that the sample failed to detect the difference betwe. the true and the hypothesized value.

Here, it should be emphasized that the statistical tests to be discussed in the text are, in most cases, parametric tests, which are primarily based on the assumptions on the forms of population distributions. Such tests require to having hypotheses stated in terms of specified parameter values. In contrast, the discussions on non-parametric tests donot require rigorous assumptions about the population distributions.

Rationale of hypothesis testing:

The reasoning used in a statistical test of hypothesis is similar to the process in a criminal court trial. In trying a person for theft, the court must decide between innocence and guilt. As the trial begins, the indicted person is assumed to be innocent.

The
e

The prosecution collects and presents all available evidences in an attempt to contradict the assumption or hypothesis of innocence and hence obtains a conviction. If there is enough evidence against innocence, the court will reject the hypothesis of innocence and declare the defendant guilty. If the prosecution fails to present firm evidence to prove the defendant guilty, the court will set him free from the indictment of theft. This statement however does not prove that the defendant is innocent, but merely that there was insufficient evidence to conclude that the defendant was guilty.

The truth or falsity of a statistical hypothesis is never known with certainity unless we examine the entire population or we have firm evidence in support of the claim.

Therefore, we make only tentative statements concerning the population parameters or the state of nature. Great uncertanities usually prevail regarding the state of nature, yet we have to make decision relative to them.

The primary objective or function of statistical test of hypothesis is to assist us in reaching sound decision in spite of such uncertainities.

Generally speaking, decision making is the process of making a choice among several alternatives. In our discussion, decision-making refers to the conclusion reached with respect to two different hypotheses on the basis of sample observations. Given the level of confidence, we accept a stated hypothesis if the sample data favor it, or reject it if the sample data show little evidence to support the hypothesis.

This procedure of decision-making through proper statistical tests between two contending hypotheses is what we refer to as hypothesis testing.

Clearly, the purpose or logic behind the test of hypothesis is to help us to draw conclusions about population parameters based on results observed in a random sample.

Statistical significance:

Following the sampling-theory approach, we accept or reject a statistical hypothesis on the basis of sample information only. Since any sample will almost surely vary somewhat from its population, we must judge whether these differences are statistically significant or insignificant.

We call a difference to be statistically significant if there is good reason to believe the difference does not represent chance variation (sampling fluctuation) alone. The term significance is used because the difference between the hypothesized value and the sample outcome is considered to be significant, or too great to be attributable to chance.

For example, we find in a study that 30% of the female children included in the sample suffer from malnutrition, while it is prevalent only among 20% of the male children. How does the result can be interpreted?

(i) The observed difference of 10% might reflect a true difference, which also exists in the population from which the sample was drawn.

(ii) The difference might also be due to chance. That is, in reality there is no difference between the prevalence between sexes of the children on average, but the sample of the girls just happened to differ from the sample of boys. One can also say that the observed difference is just due to sampling fluctuation.

(iii) One can also argue that the observed difference of 10% is due to the faulty study design (also referred to as bias); with an appropriate study design, no such difference would have been found.

If we feel confident that the observed difference between two groups can not be explained by the third factor (i.e., bias), we would like to find out whether this difference can be considered as a true difference (first factor). This leads us to formulate a statistical hypothesis, which we might put to test.

Thus, the hypothesis may be phrased as follows: "There is no difference in the population between the rates of prevalence of malnutrition between the male and female children."

How do we test the truth or falsity of the above hypothesis and arrive at a decision?

We can only conclude that this is the case if we can rule out the chance factor (sampling variation) as an explanation. We accomplish this by applying a significant test.

A significant test is used to find out whether a study result, which is observed in the sample observation can be considered as a result which is exists in the population from which the sample was drawn. This is the test that tells us whether the observed difference can be considered as significant or not in the light of the sample observations.

Definition of significant test:

A significant test is a statistical test that estimates the likelihood that an observed study result (e.g., an observed difference between two or more groups, association between variables) is due to chance or not.

If it is unlikely (a likelihood of less than some pre assigned level) that our study occurred by chance, reject the chance explanation and accept that there is a real difference (or association). We may then say that the difference (association) is statistically significant

If it is likely that our study result occurred by chance, we cannot conclude that a real difference (association) exists. We then can say that the difference (association) is not statistically significant.

Types of hypothesis :

Null hypothesis & Alternative hypothesis :

In classical tests of significance, two mutually exclusive and opposite hypotheses are used; namely ~~nlll~~ null ~~nlly~~ hypothesis and alternative hypothesis. A null hypothesis represents a theory that has been put forward either because it is believed ~~that~~ to be true or because it is to be used as a basis for argument but has not been proved. More specifically, a null hypothesis is a statement which tells us that no difference exists between the parameter and the statistic being compared to it. That "there is no difference in the population between the rates of prevalence of malnutrition between the male and female children" is an example of null hypothesis.

In a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug.

On the orther hand, the alternative hypothesis is the logical opposite of the null hypothesis. In other words, it is a statement of what a hypothesis test is set up to ~~test~~ establish.

The rejection of a null hypothesis leads to the acceptance of the alternative hypothesis. It is thus a statement of what we accept if our sample data cause us to reject the null hypothesis. The alternative hypothesis against the null hypothesis stated above, may be formulated as "there exists significant difference in the population between the rates of prevalence of malnutrition between the male and female children".

It is important to understand that the rejection of a hypothesis is to conclude that it is false, the acceptance of a hypothesis does not necessarily mean that it is true. We are accepting it since we have no evidence to believe otherwise. Because of this terminology, the statistician should always state as his null hypothesis that which he hopes to reject.

For example, if the experimenter is interested in a new drug, he should assume that it is no better than the drug now available on the market and then set out to reject this contention. Although it is possible to consider either of the two as null hypothesis, it is a statistical convention that ~~consider~~ in most cases, the null hypothesis is the one that asserts the absence of any effect claimed for a certain action or treatment.

A null hypothesis is usually denoted by H_0 and an alternative hypothesis by H_1. With reference to the above example, suppose that 70% of all those suffering from a certain disease are cured with drug A or $p = 0.70$. The new drug B is developed which is claimed to be more effective or $p > 0.70$. Then we state $p = 0.70$ as H_0, the null hypothesis, and $p > 0.70$ as H_1, the alternative hypothesis. The former is the null hypothesis because it asserts that the new drug does not have the effect claimed for it.

With a coin tossing experiment. the null and the alternative hypotheses may be formulated as follows:

H_0 : the coin is unbaised (i.e., $p = q = 0.5$)

H_1: the coin is baised (i.e., $p \neq q$)

The null hypothesis is thus a hypothesis which specifies hypothesized values for one or more of the population parameters, while the alternative hypothesis H_1 is the one which asserts that the population parameter is some value other than the one hypothesized.

Few examples of null hypothesis:

(a) There is no difference in the incident of malnutrition between vaccinated and non-vaccinated children.

(b) Males donot smoke more than females

(c) There is no association between level of education and Knowledge of child nutrition among women

(d) Two teaching methods A and B are equally effective.

Few examples of the alternative hypothesis:

The corresponding alternative hypotheses of (a), (b), (c) & (d) can be stated ~~below~~ as follows:

(a) There is a difference in the incidence of malnutrition between vaccinated and non-vaccinated children (or that the incidence malnutrition among the non-vaccinated children is less than among those were vaccinated.)

(b) Males smoke more than females do.

(c) The level of education and knowledge of child nutrition among women are associated.

(d) The two teaching methods A and B are different (or are not equally effective)

It is sometimes convenient to express the hypothesis in symbolic forms. Thus for the hypothesis (a) mentioned above, we can use p_1 to stand for the incidence among the vaccinated children and p_2 to stand for the incidence among the non-vaccinated children and state $P_1 = P_2$ as H_0 and $P_1 \neq P_2$ as H_1. One can also formulate the alternative hypothesis as $P_1 < P_2$ or $P_1 > P_2$.

Simple hypothesis :

If a hypothesis completely specifies the distribution of a population, it is called a simple hypothesis. Suppose, for example, that a coin is tossed 30 times ($n = 30$) to determine whether the coin is an ideal one. Then the hypothesis

$$H : p = 0.50$$

is a simple hypothesis since it completely specifies the population distribution. (As we know that a binomial distribution is completely defined by n and p) Similarly, if a random variable X is normally distributed with standard deviation, $\sigma = 10$, the hypothesis that its mean μ is, say, 50, that is

$$H : \mu = 50$$

is a simple hypothesis since a normal distribution is completely defined by its mean and standard deviation.

Composite hypothesis:

If a hypothesis does not specify the population distrib. completely, it is called a composite hypothesis. In t above-mentioned coin-tossing example, if we donot specify that $n=30$, then the hypothesis

$$H: p = 0.50$$

would be a composite hypothesis since there is a family of distributions all with $p=0.50$. Similarly, if we know that $n=30$, the hypothesis

$$H: p \neq 0.50$$

would be a composite hypothesis since there is a family of distributions all with $n=30$. Likewise, the hypothesis $H: \mu \neq 50$ is a composite hypothesis because we donot know the exact distribution of the population even if we know it is a normal distribution with a known standard deviation.

Exact and Inexact hypotheses:

An exact hypothesis is one that specifies a unique value for the population parameter such as $H: p = 0.50$ or $H: \mu = 50$.

An inexact hypothesis is one that states the population parameter falls within an interval of values, such as $H: p > 0.50$ or $H: \mu \neq 50$.

A simple hypothesis must be an exact one, while it is not necessarily true that an exact hypothesis is a simple hypothesis.

One-tailed and Two-tailed Tests:

The alternative hypothesis may be either unidirectional or non-directional. When H_1 only asserts that the population parameter is different from the one hypothesized, it is referred to as a non-directional or two-tailed hypothesis and the test associated with it is called a two-tailed test.

Defn:

A test of any statistical hypothesis where the alternative is located in both tails of the distribution is called a two-tailed test.

Thus if the null hypothesis is stated as $H_0: \mu = \mu_0$, the alternative hypothesis in a two-tailed test may be specified as $H_1: \mu \neq \mu_0$. It seems obvious that in a two-tailed test, no direction of difference is given. More specifically, the alternative hypothesis states that either $\mu < \mu_0$ or $\mu > \mu_0$.

To test this hypothesis, the region of rejections is divided into two tails of the distribution. A two-tailed test will be used for malnutrition problem stated in (a) above to test the null hypothesis that $P_1 = P_2$ against the two-sided alternative $P_1 \neq P_2$.

Further suppose that there are on average 50 matches in a box. We could then set the null hypothesis and alternative hypotheses as $\mu = 50$ and $\mu \neq 50$, respectively.

Occasionally, H_1 is unidirectional or one-tailed. In this instance, in addition to asserting that the population parameter is different from the one hypothesized, we assert the direction of that difference. The associated test is called one-tailed test

Def^n:

A test of any statistical hypothesis for which we can reject the null hypothesis, is located in one tail of the probability distribution is called a one-tailed test.

In other words, the critical region for a one-sided test is the set of values less than the critical value of the test or the set of values greater than the critical value of the test.

Suppose we want to test a manufacturer's claim that there are, on average, 50 matches in a box. We could set the following hypotheses:

$$H_0 : \mu = 50$$
$$H_1 : \mu > 50$$

or perhaps

$$H_0 : \mu = 50$$
$$H_1 : \mu < 50$$

A one-tailed test places the entire probability of an unlikely outcome into the tail specified by the alternative hypothesis.

Further to conceptualize the one-tailed and two-tailed tests, suppose, for instance, that a manufacturer of ballpoint pen, whose machine produces 1500 pens per hour on the average, is considering the purchase of a new machine. If he does not want to buy the new machine, unless it is definitely proved superior, he would test the null hypothesis that the new machine is no better than similar machin now available on the market and test this against the alternative hypothesis that the new machine is superior to the old one. In other words, he will test the null hypothesis $\mu = 1500$ against the alternative hypothesis $\mu > 1500$, and buy the new machine only if the null hypothesis can be rejected. Such an alternative test will result in a one-tailed test with the critical region in the right tail. If he wants to buy the new machine (which has some other nice features) unless it is actually slower than the old one, he would test the null hypothesis $\mu = 1500$ against the alternative $\mu < 1500$, and buy the new machine unless the null hypothesis can be rejected.

value
coin
P

In general, we use the two-sided alternative $\mu \neq \mu_0$ if we want to reject the null hypothesis regardless of whether μ_0 happens to be too large or too small. The choice of an appropriate one-tailed alternative depends usually on what we hope to be able to show, or better, perhaps, where we want to put the burden of proof. In the above example, the manufacturer in the first case would be putting the burden of proof on the new machine, and in the second case he would be putting the burden of proof on the old machine.

Level of significance:

Careful analysis of the logic of statistical inference reveals that we cannot prove the null hypothesis, nor can we directly prove the alternative hypothesis. However, if we can reject the null hypothesis, we can assert its alternative, namely, that the population parameter is some value other than the one hypothesized. Applied to the coin problem, if we can reject the null hypothesis that $p = q = 1/2$, we can assert the alternative, namely that $p \neq q$. Note that the proof of the alternative hypothesis is always indirect. We have proved it by rejecting the null hypothesis. On the other hand, since the alternative hypothesis can neither be proved nor disproved directly we can never prove the null hypothesis by rejecting the alternative hypothesis.

The strongest statement we are entitled to make in this respect is that we failed to reject the null hypothesis.

How do we decide on the cut-off points, called critical values that separate the acceptance & rejection regions? That is, how do we decide how much statistical evidence we need before we reject the null hypothesis? This depends on the amount of confidence that we want to attach to the test conclusions and the significance level α.

The significance level of a test is a fixed probability of wrongly rejecting the null hypothesis when in fact it is true. This asserts that α is the maximum probability with which we are willing to risk an error of making incorrect decision rejecting the null hypothesis even though it is true. Sometimes level is also known as the size of the test. The choice of α is usually made between 0.01 and 0.0

In any case, we probably want the probability of rejecting a true null hypothesis to be small and thus we select a small value of α in order to run the risk of rejecting a true null hypothesis as small as possible. In a hypothesis test, the level of significance is expressed as

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

The choice of α is somewhat arbitrary. While some researchers are satisfied with 5% (i.e., $\alpha = 0.05$) level of significance, the others are yet stricter and thus use 1% (i.e., $\alpha = 0.01$) level of significance. Thus, when employing a 0.05 level of significance we reject the null hypothesis when a given result occurs, by chance 5% of the time or less. When employing a 0.01 level of significance, we reject the null hypothesis when a given result occurs, by chance 1% of the time or less. Under these circumstances of course, we affirm the alternative hypothesis.

Type-I and type-II errors :

If we reject a null hypothesis (finding a statistically significant difference), then we are accepting the alternative hypothesis. You might now ask aren't we taking a risk that we will be wrong in rejecting the null hypothesis ?

Certainly, in either accepting or rejecting a null hypothesis, we can make incorrect decisions. A null hypothesis can be accepted when it should have been rejected or rejected when it should have been accepted.

The experimenter is thus faced with two types of errors in reaching a decision: Type I error and Type II error. A type I error has been committed if we reject the null hypothesis when it is true and the probability of committing a Type I error is identical with the level of significance that we defined earlier. Thus

$$P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha$$

Type-I error : (Def^n)

A type-I error for a statistical test is the error of rejecting the null hypothesis when the null hypothesis is true.

With a type II error, one fails to reject a false null hypothesis. In other words, Type-II error has been committed if we accept the null hypothesis when it is false and the probability of committing a type-II error is denoted by β.

Hypothesis testing places greater emphasis on Type-I error than on type-II error.

$$P(\text{type-II error}) = P(\text{accept } H_0 \mid H_0 \text{ is false}) = \beta.$$

<u>Def^n: (Type-II error):</u>

A type-II error for a statistical test is the error of a accepting the null hypothesis when the null hypothesis is false.

The complement of the probability of type-II error is called the power of a test.

That is, Power of a test $= P(\text{reject } H_0 \mid H_0 \text{ is false})$

$\therefore$ Power of a test $= 1-\beta$.

Def^n: (Power of a test):

The probability of rejecting a false null hypothesis is referred to as the power of a test.

The power of a test is a measure of our ability to demonstrate that the alternative hypothesis, rather than the null hypothesis is the correct decision. Given the size of α, a test is considered most powerful than another if the value of $1-\beta$ of the test is greater than that of the other.

(#) Both α and β play a vital role in hypothesis testing. They are used in measuring the goodness of fit of a statistical test. A good test is one for which both of these are small.

A graph of $1-\beta$, as a function of the true value of the parameter of interest is called the power curve for the statistical test. Ideally, we want α to be small and $(1-\beta)$ to be large.

We can calculate the power of a test only when H_0 is false and we are given the true value of the population parameter under H_1. A value of $1-\beta = 0.75$ implies that we will correctly reject the false null hypothesis with 75 percent probability.

Example:

Compute the power of the test if $H_0 : \mu = 138$ against the alternative $H_1 : \mu = 142$ at $\alpha = 0.05$. The standard deviation of the sample mean is known to be 2.

Sol^n:

Since $\sigma_{\bar{x}}$ is known, $z = \dfrac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$ is the appropriate test statistic

Since we are employing a one-tailed test, the critical region consists of all values greater than 1.645. The critical value of the sample statistic is therefore,

$$\bar{x} = \mu_0 + z_{0.05}\,\sigma_{\bar{x}}$$

$$= 138 + 1.645 * 2$$

$$\therefore \bar{x} = 141.30$$

If the alternative hypothesis is true (i.e., $H_1 : \mu = 142$) the corresponding z-value is

$$z = \frac{141.30 - 142}{2}$$

$$\Rightarrow z = -0.35$$

Therefore the value of $Z < -0.5$ under the null hypothesis constitute the acceptance region. The type-II error β is therefore

$$\beta = P[H_0 \text{ is accepted} \mid \mu = 142]$$

$$= P[Z < -0.35 \mid \mu = 142]$$

$$= 1 - P[Z < 0.35 \mid \mu = 142]$$

$$= 1 - 0.6368$$

$$= 0.3632$$

Hence, the power of the test for $H_1 : \mu = 142$ is

$$1 - \beta = 1 - 0.3632$$

$$= 0.64$$

Hence, the chances of correcting rejecting the null hypothesis when it is false is 64%.

Example:

The daily production of milk in a diary firm has averaged 880 tons for the last several months. The management would like to know whether this average has changed in recent time. The management sele

50 days randomly from the computer database and computes the average and standard deviation of these 50 days' production volume as 870 tons and 21 tons respectively. If α is set at 0.05, compute the power of the test and hence draw the power curve.

Soln:

The null hypothesis is $H_0: \mu = 880$ against the alternative $H_1: \mu \neq 880$

Since it is a two-tailed test and $\alpha = 0.05$, the decision rule is: reject H_0 if

$$\bar{x} \geq \mu_0 + z_{\alpha/2}\sigma_{\bar{x}} \quad \text{or if} \quad \bar{x} \leq \mu_0 - z_{\alpha/2}\sigma_{\bar{x}}$$

For $\mu_0 = 880$, $z_{\alpha/2} = 1.96$

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{21}{\sqrt{50}} = 2.97$, the critical values are therefore

$\bar{x} \geq 880 + (1.96)\cdot(2.97) = 885.82$ and

$\bar{x} \leq 880 - (1.96)\cdot(2.97) = 874.18$

Thus, the acceptance region will be

$$874.18 < \bar{x} < 885.82$$

Let us compute β for the alternative $\mu = 870$, which is less than the hypothesized value 880.

The probability of accepting the null hypothesis given $\mu = 870$ is equal to the area under the sampling distribution of the mean in the interval between 874.18 and 885.82.

Since $\bar{x}$ is a normal variate with mean 870 and standard deviation 2.97, β is equal to the area under the normal on the left located between 874.18 and 885.82.

The z-values corresponding to these values are

$$z_1 = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{874.18 - 870}{2.97} = 1.41$$

$$\& \; z_2 = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{885.82 - 870}{2.97} = 5.33$$

Hence,

$$\beta = P\left[\text{accept } H_0 \mid \mu = 870\right]$$

$$= P\left[874.18 < x < 885.82 \mid \mu = 870\right]$$

$$= P\left[1.41 < z < 5.33 \mid \mu = 870\right]$$

$$= P\left[z < 5.33 \mid \mu = 870\right] - P\left[z < 1.41 \mid \mu = 870\right]$$

$= \varphi(5.33) - \varphi(1.41)$

$= 1 - 0.9207$

$= 0.0793$

Hence, the power of the test is

$1-\beta = 1 - 0.0793 = 0.9207$

Since the alternative hypothesis is $\mu \neq 880$, we can compute $1-\beta$ for various values of μ under the null hypothesis different from $\mu = 880$.

For example, if under the null hypothesis μ is chosen to be 885, we find that β is 0.6103 and hence $1-\beta = 0.3897$.

The table below shows the values of $1-\beta$ for values of μ under H_1.

When these values are plotted against the values of μ, we obtain a U-shaped curve known as power curve.

μ	865	870	872	875	877	880	883
$1-\beta$	0.999	0.921	0.767	0.390	0.173	0.05	0.173