Statistical Inference I
Introduction to Statistical Hypothesis Testing

Kiranmoy Chatterjee*

1 Introduction

Suppose, we consider a population which is characterized by some parameters such as mean(*for* location), variance(*for* scale), skewness or kurtosis(*for* shape). In statistical analysis, one of the major aim is to make inference about the population that means about some of its unknown parameter(s). In *Estimation* chapter we estimate those parameters with the help of the given sample data. Now in *Statistical Hypothesis Testing* chapter, we will test or validate a statement H (called hypothesis) about some unknown parameter of our interest in the light of the given sample data. Depending on the data, finally that hypothesis will either be rejected or it would not be rejected.

Statistical hypothesis: A statement about a parameter characterising a population.

Let $(x_1, x_2, \ldots, x_n)$ be a random sample of size $n(\geq 1)$ and x_i is an observed value of r.v. X_i such that $X_i \sim f_\theta(x)$, $\forall i = 1(1)n$. Suppose, a statement $H : \theta = \theta_0$ (known) is given about the unknown population parameter θ. Our task is to validate the given statement H against some alternative statement like $\theta > \theta_0$ or $\theta < \theta_0$ or $\theta \neq \theta_0$. Sometimes, a statement H may be given as $\theta > \theta_0$ (or, $\theta < \theta_0$). Hence, it is clear that any hypothesis about a parameter θ is nothing but a proper subset of the parameter space Θ and hence of two types. If the subset is singleton set then it can specify the population completely (e.g. $\theta = \theta_0$) but when it is not a singleton, then it cannot specify the population completely (e.g. $\theta \geq \theta_0$, $\theta < \theta_0$, $\theta \neq \theta_0$, etc.). We define these two types of hypothesis below.

Simple hypothesis: Any statistical hypothesis which specifies the population distribution completely. *Example:* Suppose, observations are coming from $N(\mu, \sigma^2)$ population with σ known (say, 10) then if a hypothesis say that $\mu = 50$, we call it *Simple hypothesis* because specific value of the unknown parameter (under the given hypothesis) completely specify the population.

Composite hypothesis: Any statistical hypothesis which does not specify the population distribution completely. *Example:* Suppose, observations are coming from $N(\mu, \sigma^2)$ population and σ is known (say, 10) then if a hypothesis say that $\mu > 50$, we call it *Composite hypothesis*. Because population is not completely specified as value of the only unknown parameter μ is not exactly known under the given hypothesis.

In general, for any random experiment, there is always a belief (or hypothesis) H about a possible subset of Θ where true θ may belong but that can not be proved. We would like to validate that belief through a statistical hypothesis testing procedure based on sample data only.

*Department of Statistics, Bidhannagar College, Salt Lake, Kolkata-700064; E-mail: *kiranmoy07@gmail.com*
Target Students: B.Sc.(Hons.) in Statistics

So far we have understood the meaning of hypothesis along with its different nature and the ultimate aim of a hypothesis testing procedure. Now, we will see how to frame such a hypothesis testing problem and how to perform it statistically.

Let us start with a simple example. Let p proportion of voters in a population favour a candidate A against another candidate B. Now, a statement is made that $H : p > 1/2$. Our task is to validate or test the given statement. Suppose 20 voters are sampled and X denotes number of voters favour candidate A out of 20 samples. If we observe $x = 2$, we do not agree with given hypothesis. If $x = 5$, still we do not like to accept this. But, if $x = 9$ or 10 or more, then we may agree with the given statement that $p > 1/2$. Hence, it is clear that test of a hypothesis means to construct a subset $R \subseteq \mathcal{X}$, where $\mathcal{X}$ being sample space of X, and if the observed sample point x ($\in \mathcal{X}$) falls in R, we reject the given hypothesis H otherwise accept it. R can be said as rejection region of the hypothesis $H : p > 1/2$.

If you carefully observe that what we actually do for this given problem, then it will be clear that at first we take $p = p_0 = 1/2$ and calculate $np_0 = 15/2 = 7.5$ as $E(X) = np$. Then we reject the given hypothesis $p > 1/2$, if x is too smaller than 7.5. This can be interpret in a equivalently way that if x is too smaller than $E(X|p_0) = np_0 = 7.5$, the expected value of X assuming $p = p_0 = 1/2$, we will reject $p > 1/2$. From the above example we feel a need to frame another hypothesis $p = 1/2$ or $p \leq 1/2$ containing equality which contradicts the given hypothesis $p > 1/2$.

From the example of statistical hypothesis testing problem two things are clear. First is that the statistical testing procedure is a battle of two defined hypotheses (or, statements) H and $H^{'}$ which contradicts the given hypothesis H. Sample data helps to choose the one which is more likely. So, if the hypothesis $H^{'}$ is not given, one should define it before starting the statistical test procedure. This is Fishers approach for hypothesis testing which requires the specification of only one hypothesis, known as null hypothesis (H_0). The null hypothesis H_0 either be same as the existing belief H or it may contradict H. If the sample data are not consistent with H_0, i.e., if improbable outcomes are obtained, H_0 is rejected. In this sense, Fisherian statistical hypothesis testing can be characterized as a validation procedure for the statement in H_0 (equivalently, in H) based on sample data only.

Generally, we mark a given statement $H : \theta = \theta_0$ as *null hypothesis* as $\theta - \theta_0 = 0$ implies there is no (or, null) difference between θ and its specified value θ_0. Null hypothesis is denoted as H_0. Hence, $H \equiv H_0$. We also call any hypothesis (may or may not given) as *alternative hypothesis* which contradict H_0. Alternative hypothesis is denoted as H_A. This does not mean that the given hypothesis H is to be considered always as null hypothesis. When a given statement H about θ consists equality, we will frame it as H_0. When given H does not consist any equality, we will set H as H_A and consider a hypothesis, which contradict H, as H_0. Obviously, then H_0 will consist equality as H does not. For example, if $H : \theta > \theta_0$ then $H \equiv H_A$ and H_0 will be $\theta \leq \theta_0$ or $\theta = \theta_0$.

Null hypothesis: A hypothesis (may be simple or composite) that includes an equality of a population parameter with a specified value. In other words, null hypothesis is that which says there is no difference (i.e. *null* difference) between the true (unknown) parameter and a specified value. Usually, it is denoted as H_0. For example, $H_{01} : \theta = \theta_0$. For two samples problem, often we wish to test $H_{02} : \theta_1 = \theta_2$ (*null* difference) $\Leftrightarrow H_{02} : \theta_1 - \theta_2 = 0$ or $H_{02} : \frac{\theta_1}{\theta_2} = 1$.

Alternate hypothesis: A hypothesis (often composite) which always contradicts the null hypothesis, and associated with a theory one would like to believe. Usually, it is denoted as H_A. For example, $H_{A1} : \theta \neq \theta_0$ or $H_{A2} : \theta > \theta_0$ or $H_{A3} : \theta < \theta_0$.

Thus in connection with the above example, we set $p \leq 1/2$ as null hypothesis here (as it includes equality) and frame $p > 1/2$ as alternative hypothesis. Now, the problem is defined as to test $H_0 : p \leq 1/2$ against $H_A : p > 1/2$.

Generally it is seen that a belief about θ is a composite type statement or a composite hypothesis. The null hypothesis is often the reverse of what the experimenter actually believes; then it is put forward in the hope that the data will contradict the null hypothesis. Though of course, there is also some exceptional real life examples.

The next issue in the above example of testing $H_0 : p \leq 1/2$ against $H_A : p > 1/2$ is to choose a suitable subset W, called critical region or rejection region for H_0, from the sample space $\mathcal{X} = 0, 1, 2, \ldots, 20$. If anyone suggests $W = \{X : X \geq 8\}$, we cannot judge it is good enough or not unless we calculate its probability of occurrence as X is a r.v. Now, to calculate the Prob$(X \in W)$, we have to fix θ at some point. Here we have to look into the matter quite critically.

Critical Region or Rejection Region (Definition 1): The set of all possible sample values $W \subseteq \mathcal{X}$ for which the null hypothesis (H_0) is rejected, is known as critical region or rejection region associated with the test.

Statistical Hypothesis Testing: A statistical hypothesis testing procedure or hypothesis test is a rule that specifies for which sample values the decision is made to reject the null hypothesis (H_0).

By a statistical hypothesis testing procedure either we reject or accept H_0 based on the available data. Many statisticians, however, take issue with the notion of "accepting H_0." Instead, they say: you reject H_0 or you fail to reject H_0. Why the distinction between "acceptance" and "failure to reject?" Acceptance implies that H_0 is true which cannot be concluded. Failure to reject implies that the data are not sufficiently persuasive for us to prefer the H_A over H_0. Hence, we say: either reject or do not reject the H_0 on the basis of the data in hand.

For an instance, suppose we accept H_0 but actually H_0 is not true at all. Then, we will commit an error. Another way, we may commit another type of error if we reject a true H_0. Hence, there are two kind of errors that may occur in any statistical hypothesis testing procedure.

Type-I Error: When we reject H_0 but actually it is true, then Type-I Error occurs.

Type-II Error: When we accept H_0 but actually it is false, then Type-II Error occurs.

Which error is more harmful that cannot be concluded generally. It is completely case specific and depends on which hypothesis (H_0 or H_A) experimenter is giving more importance. We can summarize the all possible situations in a testing procedure in Table 1. Since all the four possibilities in Table 1 are random events due to the randomness of the data, we may measure the extent of two errors separately through their respective probabilities and try to reduce them in order to built efficient test procedure.

Probability of Type-I Error: Probability of rejecting H_0 when actually it is true. So, Probability of Type-I Error=Pr($\underline{X} \in W|H_0$ is true)=Pr($\underline{X} \in W|H_0$), where W denotes *critical region* associated with the test. In practice, it is equated with the level of the test $'\alpha'$ for $H_0 : \theta = \theta_0$, but actually they are not same.

Table:1 Four Possible Results of a Hypothesis Test

State of the World	Decision: H_0 Not Rejected	Decision: H_0 Rejected
If H_0 is true	Correct Decision Probability $= 1 - \alpha$ $=$ Confidence level	Type I error Probability $= \alpha$
If H_0 is false	Type II error Probability $= \beta$	Correct decision Probability $= 1 - \beta$ $=$ Power of test

Probability of Type-II Error: Probability of accepting H_0 when actually it is false. So, Probability of Type-II Error=Pr($\underline{X} \neq W|H_0$ is false)=Pr($\underline{X} \neq W|H_1$), where W denotes *critical region* associated with the test. Usually denoted as $'\beta'$.

Though our aim is to reduce both the error but it is not possible simultaneously. To minimize Prob(Type-I Error), if we choose a smaller critical region W_{α_2} than W_{α_1} (in $\mathcal{X}_n$), for level $\alpha_2 < \alpha_1$, then probability of accepting H_0 will increase and that will inflate the extent of probability of type-II error (*see* Figure 1 for illustration in case of unidimensional r.v. X, i.e. $n = 1$). Hence, the best strategy

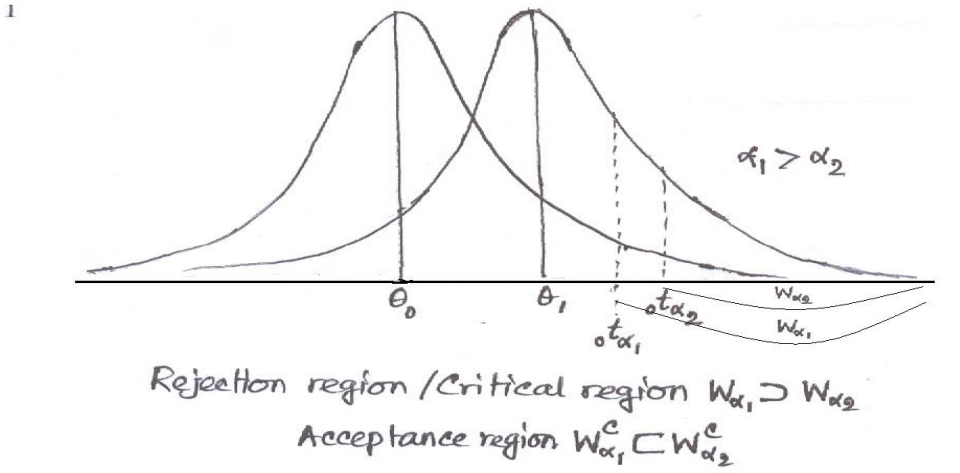

Figure 1: Change of acceptance and rejection region for changing α values

is to keep one error within a specified upper limit and then try to choose a W that minimize the other error. Now, in testing procedure, W is chosen from the relation that Probability of Type-I Error $=$ Prob($X \in W|H_0$) $\leq \alpha$, where α is called the level of the significance or simply *level of the test* which is to be suggested by the experimenter. $'H_0$ is true$'$ can be accounted by considering $\theta = \theta_0$ as $\theta_0 \in \Theta_0$ for any Θ_0 (singleton or not). Hence for practical purpose, Probability of Type-I Error is calculated from Prob($X \in W|\theta = \theta_0$) and this is equated with α to find the critical region W. Thus, if the observed sample $x \in W$, we reject $\theta = \theta_0$ i.e. H_0 otherwise accept H_0. Then, it is said that the test is level α test.

Size: The upper bound of the Probability of Type-I Error considered for a test i.e. $sup_{\theta\in\Theta_0}\Pr(\text{Rejecting } H_0|\theta)=sup_{\theta\in\Theta_0}\Pr(\underline{X}\in W|\theta)$.

Level of significance (α): The probability on the basis of which the differences are being regarded as significant of the falsity of the null hypothesis or not for a test. Conventionally, it is denoted by α. Experimenter has to suggest the value of $'\alpha'$ to conduct a test procedure.

Note 1: Level of significance (α) of a test is taken as the upper bound of its size. So, $sup_{\theta\in\Theta_0}\Pr(\text{Rejecting } H_0|\theta)\leq \alpha$. Hence, for any $\theta\in\Theta_0$, Prob(Type-I Error) $\leq$ size $\leq$ level of significance. In particular, when H_0 is simple hypothesis, i.e., $H_0:\theta=\theta_0$, then Prob(Type-I Error) = size.

Note 2: If experimenter gives more importance to the type-I error than type-II error, then α should be set at a very small value, like 0.01 or 0.005. If he/she gives more importance to the type-II error, then α should be set at a relatively large value (like 0.05 or 0.1) so that there is a possibility to have relatively smaller Prob(Type-II Error) than the former situation.

Size α test: For $0\leq\alpha\leq 1$, a test is called *size α test* if $sup_{\theta\in\Theta_0}\Pr(\text{Rejecting } H_0|\theta)=\alpha$.

Level α test: For $0\leq\alpha\leq 1$, a test is called *level α test* if $sup_{\theta\in\Theta_0}\Pr(\text{Rejecting } H_0|\theta)\leq\alpha$.

Power: Probability of rejecting H_0 when actually it is false i.e. $\Pr(t\in\mathcal{C}|H_0 \text{ is false})$, where $\mathcal{C}$ denotes *critical regions.*

But most often the calculation of Prob(Rejecting $H_0|\theta_0$) or Prob($\underline{X}\in W|\theta_0$) is very cumbersome as it require the joint distribution of r.v.s $\underline{X}=(X_1,X_2,\ldots,X_n)$ on the n dimensional domain $\mathcal{X}_n\subseteq\mathbb{R}^n$. Hence, in many cases it will be possible to find a function $T(\underline{X},\theta)$ of $\underline{X}$ and θ so that univariate T has a well behaved sampling distribution independent of θ under H_0. Under H_0, i.e. if H_0 is true, $T(\underline{X},\theta)$ will be $T(\underline{X},\theta_0)$ or simply, *T($\underline{X}$)* as θ is fixed at known θ_0. Typically, *T($\underline{X}$)* is a function of *sufficient statistic*[1] for the model parameter θ for a sample of size n. Hence one can easily find a subset $\mathcal{C}$ of $\mathbb{T}$ (sample space of T) based on which we can decide whether to reject H_0. This $\mathcal{C}$ is called critical region/ rejection region corresponding to the test statistic $T(\underline{X},\theta_0)$ for a given α and it is found from equation $\text{Prob}(T(\underline{X},\theta_0)\in\mathcal{C})=\alpha$.

Rejection Region/ Critical region (Definition 2): The set of values of the test statistic for which the null hypothesis is rejected. H_0 is rejected if the observed value of T assuming H_0 is true, $T(\underline{x},\theta_0)$, belongs to $\mathcal{C}$. Then $\mathcal{C}$ will be called critical region for the test associated with the test statistic $T(\underline{X},\theta_0)$.

Following steps that are followed in a usual statistical hypothesis testing problem.

Testing Procedure:

Step 1: Formulate the null hypothesis H_0 and the alternative hypothesis H_A

Step 2: Identify a test statistic $T(\underline{X},\theta)\in\mathbb{T}$ and its sampling distribution assuming H_0 is true that means $\theta=\theta_0$ (which contradicts H_A). Also find the observed value of T for given data and assuming $\theta=\theta_0$.

[1] A function of data is called *sufficient statistic* for a parameter θ of underlying statistical model if it is just as informative about θ as the full data. For example, $\sum_{i=1}^n X_i$ or $\overline{X}$ is sufficient statistic for μ in $\text{N}(\mu,\sigma^2)$; n^{th} order statistic $X_{(n)}$ is sufficient for θ in $\text{Uniform}(0,\theta)$.

Step 3: Choose a value of α and identify the critical region $\mathcal{C} \subseteq \mathbb{T}$ by equating α with $\text{Prob}(T(\underline{X}, \theta_0) \in \mathcal{C})$.

Step 4: If observed value of T, $T(\underline{x})$ (from Step 2), falls in the critical region $\mathcal{C}$ (already identified in Step 3), then we reject the H_0, otherwise we do not reject H_0.

Step 5: Finally, we conclude that the statement in H_0 is rejected or not rejected (whichever comes in Step 4) on the basis of the given data with $100\alpha\%$ level of significance.

Note 3: When $T(\underline{x}) \in \mathcal{C}$, we may say that the observed value is significant at level α and hence, we reject H_0. Otherwise we may say that the observed value is insignificant at level α.

Note 4: Since $T(\underline{X}, \theta_0)$ or $T(\underline{X})$ is a one-to-one function defined on the domain $\mathfrak{X}_n$ of $(X_1, X_2, \ldots, X_n)$, then there exists an equivalent Critical Region W as a subset of $\mathfrak{X}_n$, the sample space.

p-value: A different idea of Testing Procedure:

From previous discussion, now it is clear to us that every test statistic $T(\underline{X}, \theta_0)$ or $(T(\underline{X}))$ has a well-behaved explicit form of its theoretical sampling distribution free of θ_0 and observed $T(\underline{x})$ is just a point on its domain $\mathbb{T}$. So, there exists some $\delta \in (0, 1)$ such that $T(\underline{x})$ is $100(1-\delta)$ percentile point of $T(\underline{X})$, i.e. $\text{Prob}(T(\underline{X}, \theta) > T(\underline{x})|\theta = \theta_0) = \delta$. Now, if α is set as level of significance, then $\delta < \alpha$ says the observed value is very unlikely with respect to H_0. This kind of alternative interpretation is the genesis of the idea of *p-value*.

p-value: This is the probability of sampling a test statistic at least as extreme as that which was observed assuming null hypothesis H_0 is true (i.e. $\theta = \theta_0$). Notationally, *p-value* $= \text{Prob}(T(\underline{X}, \theta) \geq T(\underline{x}, \theta_0)|\theta = \theta_0)$.

An alternative process to perform a statistical hypothesis testing is commonly used based on this *p-value*. The *p-value* can be regarded as a measure for the (strength of) evidence against H_0. If the *p-value* is less than or equal to the chosen significance level α (equivalently, if the observed test statistic is in the critical region $\mathcal{C}$), then we will reject H_0. This is like a *guilty* verdict in a criminal trial the evidence is sufficient to reject innocence, thus proving guilt. When *p-value* is greater than the chosen α (equivalently, if the observed test statistic is outside the critical region $\mathcal{C}$), then there is not enough information in the data to reject H_0. When, *p-value* is very close to α, then available evidence is insufficient to reach a conclusion and in that situation, the experimenter typically gives extra consideration from the beneficiary's point of view.

References

[1] Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd Edition. Cengage Learning India Private Ltd., New Delhi, India.

[2] Kale, B. K. (2005). *A First Course on Parametric Inference*, 2nd Edition. Narosa Publishing House, India.

[3] Gun, A. M., Gupta, M. K. and Dasgupta, B. (2008). *Fundamental of Statistics, Volume One*, 8th Edition. The World Press Private Ltd., Kolkata, India.