Hypothesis Testing and Le Cams Method

Information Theory and Coding Techniques
Term project

Joel Antony Thomas
17EC10023

Nomenclature

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

How Hypothesis Testing Works

In hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to test two different hypotheses: the **null hypothesis** and the **alternative hypothesis**.

Null Hypothesis (H_0)

The null hypothesis H_0 represents a theory that has been put forward either because it is believed to be true, or because it is used as a basis for an argument and has not been proven.

Eg : In a clinical trial of a new drug, the null hypothesis may be that the new drug is of the same effect(on average) as the current drug. In this example,

H_0 : there is no difference between the two drugs(on average)

Alternate Hypothesis (H_A)

Alternative hypothesis is a position that states something is happening, in which a new theory is preferred instead of the old one (Null hypothesis). We will take the same case as we did for the Null hypothesis, ie, the new Vs old drug example.

The alternate hypothesis may be that the new drug has a different effect (on average), compared to that of the current drug. Note here that, we can have infinite possibilities for the alternate hypothesis. The thing to look out for is that, they should be mutually exclusive, and only one can be true (and one of the hypothesis will always be true). One such example is:

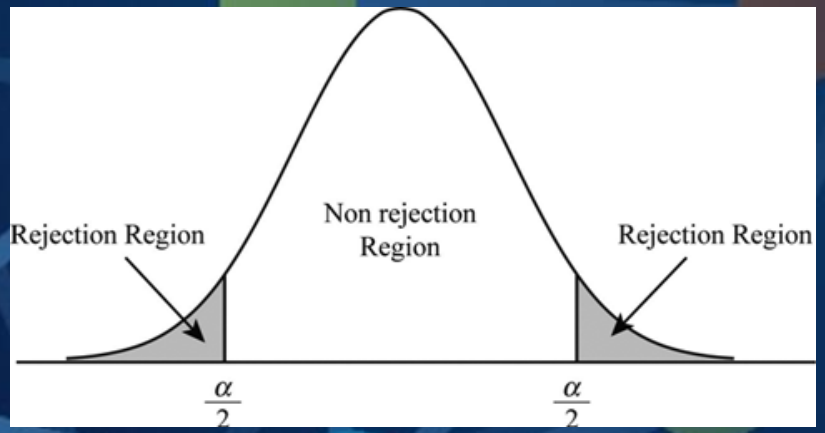
H_A : the new drug is better than the current drug (on average)

Significance Level (α)

We need to decide on a criterion for accepting or rejecting the Null hypothesis. This is where **Significance level** comes into the picture.

Significance Level (α) refers to the percentage of sample means that is outside certain prescribed limits.

If α = 5, the nomenclature used will be : testing a hypothesis at 5% level of significance. This means that we reject the null hypothesis if it falls in the two regions of areas of *0.05/2*.



So, both the regions on either side of the extreme will be for Alternate hypothesis(grey area), while the area in between is for the Null hypothesis(white area).

For our particular example, we took α = 5.

General procedure of Hypothesis Testing

All hypotheses are tested using the following steps:

1. The first step is for the analyst to state the two hypotheses so that only one can be right.
2. The second step is to decide on a value for Significance level, so we can test our hypothesis against the sample.
3. The third step is to carry out the plan and physically analyze the sample data.
4. The fourth and final step is to analyze the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

Real-World Example of Hypothesis Testing

We take a simple real world example to understand how we go about tackling problems, and how we navigate the 4 steps of hypothesis testing.

We take the example of coin flipping.

Test : a coin has exactly a 50 percent chance of landing on heads

H_0 : 50 percent is correct ($P = 0.5$)

H_A : 50 percent is not correct ($P \neq 0.5$)

Let's assume that we did 100 flips for this experiment. We now test the hypothesis based on the results of this experiment. We will look at two different results to better understand.

Let's take our Significance Level (α) = 5%.

Real-World Example of Hypothesis Testing

Result 1 : It is found that the 100 coin flips were distributed as 40 heads and 60 tails.

Inference 1 : Since $\alpha = 0.05$, we could have accepted values from $(50 - \alpha/2)$ heads to $(50 + \alpha/2)$ head, ie, from 47.5 heads to 52.5 heads. However, our results fall outside this region.

Conclusion 1 : We reject the Null hypothesis, and accept the alternate hypothesis, ie, the coin doesn't have a 50% chance of landing on heads.

Result 2 : It is found that the 100 coin flips were distributed as 48 heads and 52 tails.

Inference 2 : Using the same logic and α value as before, we find that our results lie in the region of acceptance (since $48 > 47.5$)

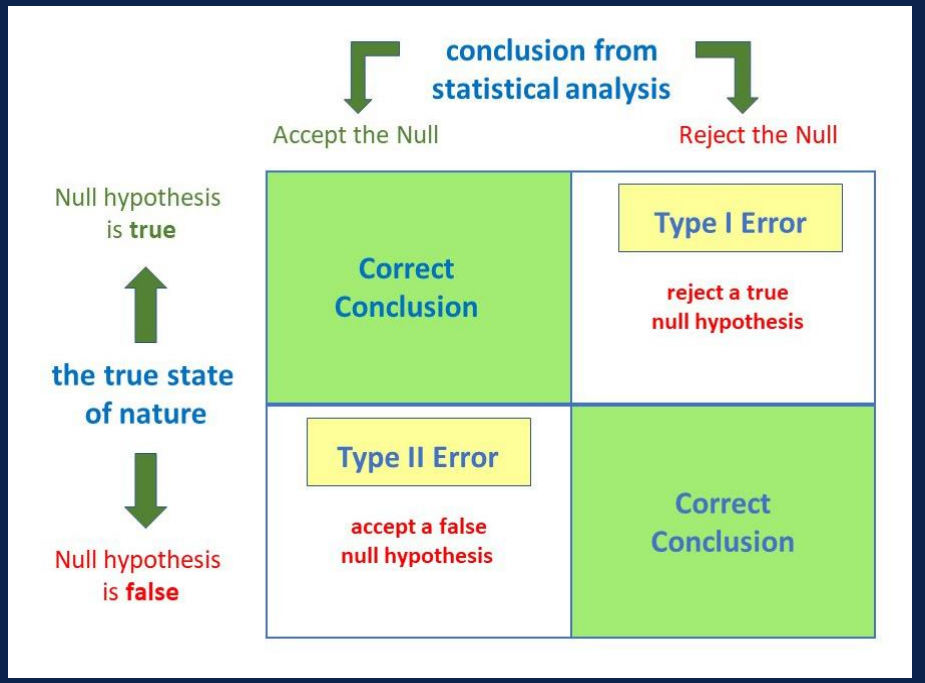
Conclusion 2 : We accept the Null hypothesis, ie, the coin has a 50% chance of landing on heads.

Type I and Type II Error

Type I error is the mistaken rejection of an actually true null hypothesis (also known as a "false positive" finding or conclusion).

Eg : H_0 = There is no difference between the two drugs on average. Type 1 error will occur if we conclude that the two drugs produce different effects when there actually isn't a difference.

The probability of making a Type I Error is the significance level, denoted by alpha(α).



Type II error is the mistaken acceptance of an actually false null hypothesis (also known as a "false negative" finding or conclusion).

Eg : H_0 = There is no difference between the two drugs on average

Type 2 error will occur if we conclude that the two drugs produce the same effect when actually there is a difference.

Hypothesis testing for population mean (critical value approach)

Step 1: State Null Hypothesis.

$H_0 : \mu = \mu_0$ (where μ_0 is a specified value)

Step 2: State Alternative Hypothesis.

1) $H_a : \mu \neq \mu_0$ (two-tailed test)

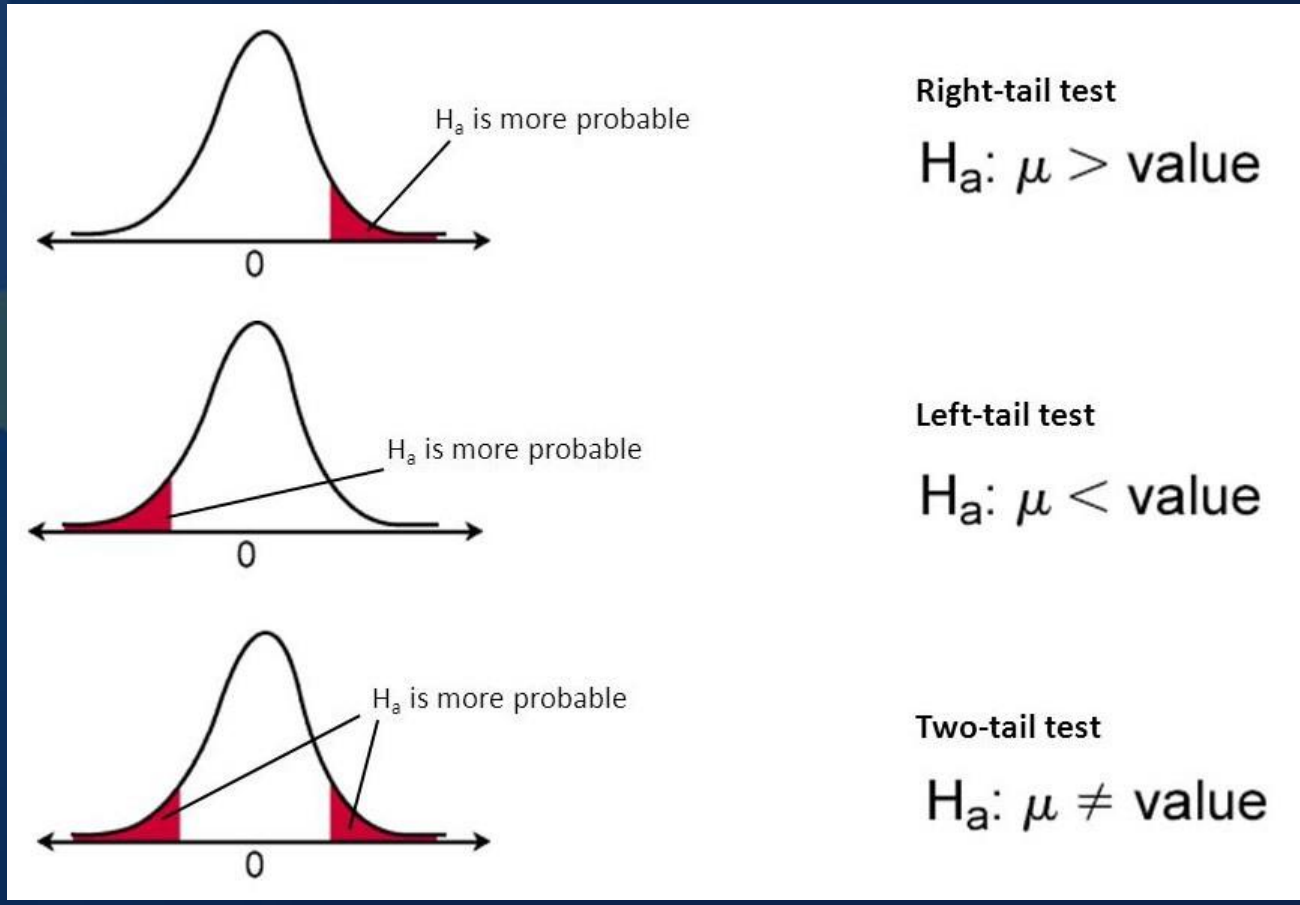
2) $H_a : \mu > \mu_0$ (one-tailed test)

3) $H_a : \mu < \mu_0$ (one-tailed test)

Step 3: State α.

Step 4: Determine Rejection Region.

We have two possibilities here depending on the data given



Hypothesis testing for population mean (critical value approach)

Use when σ is known

Use critical value(z) table

two-tailed (≠) : ***Reject*** H_0 ***if*** $z > z_{\alpha/2}$ ***or*** $z < -z_{\alpha/2}$

one-tailed (>) : ***Reject*** H_0 ***if*** $z > z_{\alpha}$

one-tailed (<) : ***Reject*** H_0 ***if*** $z < z_{\alpha}$

Step 5: Calculate the test statistic:

$$z = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})}$$

Use when σ is unknown

Use critical value(t) table using $df = n - 1$

two-tailed (≠): ***Reject*** H_0 ***if*** $t > t_{\alpha/2}$ ***or*** $t < -t_{\alpha/2}$

one-tailed (>): ***Reject*** H_0 ***if*** $t > t_{\alpha}$

one-tailed (<): ***Reject*** H_0 ***if*** $t < -t_{\alpha}$

Step 5: Calculate the test statistic:

$$t = \frac{(\bar{x} - \mu)}{(s/\sqrt{n})}$$

Step 6: Determine if the calculated test statistic is in the critical region or not. Reject or Accept H_0.

Hypothesis testing for population mean (critical value approach)

Lets look at an example :

Does the evidence support the idea that the average lecture consists of 3000 words if a random sample of the lectures of 16 professors had a mean of 3472 words, given the population standard deviation is 500 words? Use $\alpha = 0.01$. Assume that lecture lengths are approximately normally distributed.

1) $H_0: \mu = 3000$

2) $H_a: \mu \neq 3000$

3) $\alpha = 0.01$

4) Reject H_0 *if* $z < -2.576$ *or* $z > 2.576$ (We get this from the critical value(z) table)

5) $z = (3472-3000) / (500 / \sqrt{16}) = 3.78$

6) Reject H_0, because $3.78 > 2.576$

Hence, the population mean is not equal to 3000 words

Hypothesis testing for population mean (critical value approach)

Lets look at an example when σ isn't given :

The secretary of an association of professional landscape gardeners claims that the average cost of services to customers is $90 per month. Feeling that this figure is too high, we question a random sample of 14 customers. Our sample yields a mean cost of $85 and a standard deviation of $10. Test at the 0.10 significance level. Assume that such costs are normally distributed.

1) $H_0: \mu = 90$

2) $H_a : \mu < 90$

3) $\alpha = 0.10$

4) $(df = 13)$ Reject H_0 $if\ t < -1.350$ (We get this value from the critical value(t) table.)

5) $t = (85-90) / (10 / \sqrt{14}) = -1.87$

6) Reject H_0, because $-1.87 < -1.350$

Hence, the population mean is less than $90

Hypothesis testing for population mean (p value approach)

Step 1: State Null Hypothesis.

$H_0 : \mu = \mu_0$ (where μ_0 is a specified value)

Step 2: State Alternative Hypothesis.

1) $H_a : \mu \neq \mu_0$ (two-tailed test)

2) $H_a : \mu > \mu_0$ (one-tailed test)

3) $H_a : \mu < \mu_0$ (one-tailed test)

Step 3: State α.

Step 4: Determine p-value from Minitab printout

Step 5: Compare the p-value with the α value; If P–value $\leq \alpha$, reject H_0, otherwise accept H_0.

Lets look at an example to understand this approach.

Hypothesis testing for population mean (p value approach)

The mean GPA at a certain university is 2.80 with a population standard deviation of 0.3. A random sample of 16 business students from this university had a mean of 2.91. Test to determine whether the mean GPA for business students is greater than the university mean at the 0.10 level of significance.

1) $H_0: \mu = 2.8$

2) $H_a: \mu > 2.8$

3) $\alpha = 0.10$

4) ***p-value = 0.075*** (We get this from the minitab)

5) $0.075 < 0.1$, hence reject H_0

Hence, the population mean is greater than 2.8.

One-Sample Z - GPA

Descriptive Statistics

N	Mean	SE Mean	95% Lower Bound for μ
16	2.9081	0.0750	2.7847

μ: mean of Sample

Known standard deviation = 0.3

Test

Null hypothesis	$H_0: \mu = 2.8$
Alternative hypothesis	$H_1: \mu > 2.8$

Z-Value	P-Value
1.44	0.075

Binary Hypothesis Testing

In **Binary Hypothesis Testing**, we will be dealing with two probability distributions. Let us call them P_1 and P_2. We will then choose a sample from either P_1 or P_2 (let us call this nature as V). For simplicity, let it be equal (ie 0.5 each)

$$V = \begin{cases} 1 \text{ with probability } 0.5 \\ 2 \text{ with probability } 0.5 \end{cases}$$

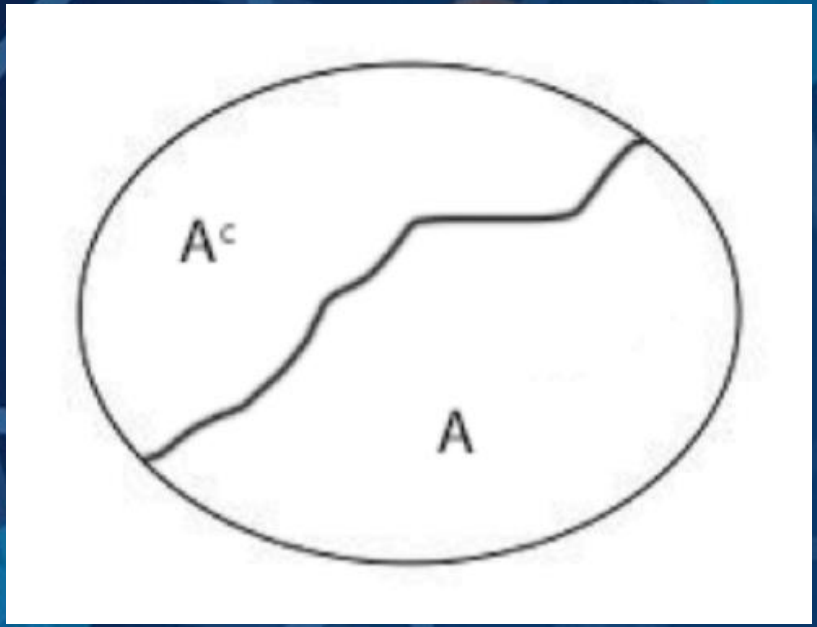
$g(X)$ is an estimator that operates on X and tells about whether X is from distribution P_1 or P_2.

We can then find ther **probability of error** as follows :

$$P_r(error) = P_1(g(x) = 2) * P(V = 1) + P_2(g(x) = 1) * P(V = 2)$$

$$= 0.5 * (P_1(g(x) = 2) + P_2(g(x) = 1))$$

Binary Hypothesis Testing Error



The ***supremum(sup)*** of a subset ***S*** of a partially ordered set ***P*** is the least element in ***P*** that is greater than or equal to all elements of ***S***, if such an element exists.

The estimator ***g()*** will divide the random variable ***X*** into two parts as follows.

A = region where ***g(x) = 1***

A^c = region where ***g(x) = 2***

We now find the best probability of error(P_e^*) :

$$(P_e^*) = 0.5 * inf\,[P_1(A^c) + P_2(A)]$$

Infimum(inf) of a subset ***S*** of a partially ordered set ***P*** is a greatest element in ***P*** that is less than or equal to all elements of ***S***, if such an element exists.

$$= 0.5 * inf\,[\,1 - P_1(A) + P_2(A)\,]$$

Since $P_1(A^c) = 1 - P_1(A)$

$$= 0.5 * [\,1 - sup\,[\,P_1(A) - P_2(A)\,]\,]$$

Le Cam's

We define

$$sup[P_1(A) - P_2(A)] = ||P_1 - P_2||_{TV}$$

Where **TV = Total Variation Distance**.

Hence,

$$P_e^* = 0.5 * [1 - ||P1 - P2||_{TV}]$$

This is the **Le Cam's equation**.

Observe that the worst case error is 0.5 , which occurs when the total variation distance between P_1 and P_2 is 0, at which point the classifier is random guessing. Conversely, when the data is linearly separable, the total variation between the two distributions is 1, the error is 0

We will now look at an example to understand the concept of Total Variation better.

TV Distance

Let our sample space $X = \{1, 2, 3\}$, and

P_1 : $P_1(1) = 0.5$ $P_1(2) = 0.25$ $P_1(3) = 0.25$

P_2 : $P_2(1) = 0.3$ $P_2(2) = 0.3$ $P_2(3) = 0.4$

subset	$P_1(A) - P_2(A^c)$
{1}	0.2
{2}	-0.05
{3}	-0.15
{1, 2}	0.15
{1, 3}	0.05
{2, 3}	-0.2

Total Variation(TV) distance is the maximum distance of all the possible subsets of X

Hence,

$$||P_1 - P_2|| = \sup[P_1(A) - P_2(A)]$$

$$= 0.2$$

Properties of TV Distance

Symmetry of TV Distance :

$$||P_1 - P_2||_{TV} = \sup[P_1(A) - P_2(A)]$$
$$= \sup[1 - P_1(A^c) - 1 + P_2(A^c)]$$
$$= \sup[P_2(A^c) - P_1(A^c)]$$
$$= \sup[P_2(B) - P_1(B)]$$

Where ***B*** also belongs to the set ***X***

$$= ||P_2 - P_1||_{TV}$$

Similarly, it can be shown that :

1. TV distance is a valid distance
2. $||P_1 - P_2||_{TV} = 0.5 * ||P_1 - P_2||_{l1}$ $l1$ = $l1$ norm
3. TV distance satisfies triangularity property
4. $(2 / \ln 2) * ||P_1 - P_2||_{TV}^2 \leq D(P_1||P_2)$

(**Pinsker's Inequality**)

Multi-class Classification

Let's now look at the case of **multi-class classification**. Here, instead of two classes(binary), we are faced with multiple classes.

Let's assume

$$J \sim Uniform(1, 2, 3...M)$$

$$(Z|J = j) \sim P_{\vartheta j}$$

We need to determine index ***j*** from which the samples have been taken. Of course, the difficulty in this depends on the dependence between ***Z*** and index ***j***.

Our goal is to determine the index ***J*** of the probability distribution from which a given sample has been drawn. Intuitively, the difficulty of this depends on the amount of dependence between ***Z*** and index ***J***.

Fano's Method

Let $Q_{Z,J}$ denote the joint distribution of inputs and labels, and $Q_Z Q_J$ denote the product of the marginal distributions. If we have

$$Q_{Z,J} = Q_Z Q_J$$

Z and ***J*** are statistically independent, which means knowing one of them does not let us infer anything about the other. In this case, the best we can do is random guessing. To quantify the amount of dependence of ***Z*** and ***J*** we define mutual information ***I*** as the Kullback-Leibler divergence:

$$I(Z, J) = D_{KL}(Q_{Z,J} \mid Q_Z Q_J) \geq 0$$

Now Fano's inequality gives us a lower bound on the classification error:

$$Q(\psi(P) \neq J) \geq 1 - \frac{(I(Z, J) + \log 2)}{\log M}$$

Ψ = the distribution for the probability

This is the Fano's method used for multiple class classification.

Fano's Method

There are various ways to upper bound the mutual information ***I(Z, J)*** in this setting. On way is to use the convexity of the KL-divergence which leads us the following inequality.

$$I(Z,J) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} \mathcal{D}_{KL}(\mathbb{P}_{\theta_j}^n, \mathbb{P}_{\theta_k}^n)$$

Thank You !