DA-IICT, Gandhinagar
Lecture Notes
Subject: Probability, Statistics and Information Theory (SC222)

Date: 24/03/2021 **Reg. No.: 201901306**

Name: Raj Patel **Lecture No.: 19**

Contents

1 Quick Recap

In the previous lecture, we have started hypothesis testing and studied these topics.

Statistical Hypothesis
Definition: It's a statement about the nature of a population distribution. It's in terms of a set of parameters(e.g., mean, variance) of population distribution. We don't know about its truth that's why it's called hypothesis.

1. **Null Hypothesis -** H_0
   Definition: It's a kind of hypothesis which explains the population parameter whose purpose is to test how valid the given experimental data is. It's denoted by H_0. This hypothesis will be rejected or not rejected is based on the consistency with population or sample.

2. **Alternative Hypothesis -** H_1
   Definition: It's the contrary of the null hypothesis. It's denoted by H_1. It's alternative to null hypothesis.

Null hypothesis will be rejected if it appears to be inconsistent with the given sample data. Otherwise it will not be rejected. What we mean by this statement is that, by accepting a hypothesis we are not guaranteeing that it's indeed true but instead we are interpreting that it's not much differing from the resulting sample data. More formally certain margin is acceptable in the actual correct value for accepting a hypothesis.

Example 1: Apollo tyre manufacturing company makes a statement about the quality of their manufactured tyre is that their tyre on an average gives 50k kilometers' trip. Clearly it's hypothesis, as we don't know the truth. Now suppose a standard checking agency is randomly taking 15 tyres manufactured by Apollo. If it comes out that average trip of those 15 tyres is 47 - 48k kilometers, then clearly this is somewhat bearable static, and so agency is not rejecting the statement. But if that average trip comes out to be 41 - 42k kilometers, then it's too much differing from what manufacturing company is been telling to have, and thus agency is rejecting the Apollo company's statement regarding their tyre quality.

Let's see a further example of hypothesis where we are going to point out which is null hypothesis - H_0 and which is alternative hypothesis - H_1.

Example 2: Suppose a telecom company claims that their broadband connection is having the mean link speed - 5 Mbps or more. So, Telecom Regulatory Authority of India(TRAI) wants to verify this whether the claim made by the telecom company is correct or not.

So, here TRAI wants to test the null hypothesis,

$$H_0 : \mu \geq 5\ Mbps$$

as against the alternative hypothesis,

$$H_1 : \mu < 5\ Mbps$$

Test Statistic - TS
Definition: It's a statistic whose value is determined from the population data. It's denoted by TS. TS determines whether the null hypothesis will be rejected or not rejected. In the previous example, the mean link speed of broadband connection is the test statistic.

Critical Region
Definition: It's the set of values of TS for which the null hypothesis H_0 is rejected. It's denoted by C. It's also called **Rejection region**.

- Reject H_0 if TS belongs to C.
- Not reject H_0 if TS doesn't belong to C.

In this lecture, we studied the topics coming afterwards.

2 Formulating Appropriate Hypothesis

The result that the null hypothesis is rejected is a strong statement, meaning which null hypothesis H_0 can be rejected if we have a strong evidence that proves H_0 is inconsistent with the observed sample data. On the other hand the result that the null statement is not rejected(simply accepted) is a weak statement, meaning which H_0 is consistent with the observed sample data. So, whether to consider the given claim or statement as null hypothesis H_0 or alternative hypothesis H_1 is important. Before we go towards formulation of appropriate hypothesis, we see what kind of errors may occur while testing the null hypothesis.

Errors In Null Hypothesis Testing
In the procedure for testing the null hypothesis H_0, two different types of errors are possible.

- **Type 1 Error:** The test rejects H_0 when H_0 is indeed correct.
- **Type 2 Error:** The test doesn't reject H_0 when H_0 is not correct.

As the result that the null hypothesis is rejected is a strong statement, we are more concerning with the Type 1 error. And we are willing to minimize Type 1 error as small as possible to maintain the correctness of this strong statement. H_0 should be rejected only when the sample data shows much variation when H_0 is true.

Now we can formulate whether a particular claim or statement is null hypothesis H_0 or alternative hypothesis H_1.

- **When To Choose Alternative hypothesis H_1:**
  If one tries to establish a certain hypothesis, then that should be alternative hypothesis H_1. As it's a strong statement and consistent with the sample data.
- **When To Choose Null hypothesis H_0:**
  If one tries to discredit a hypothesis(against some one), then that should be null hypothesis H_0. As we require strong evidence to reject this null hypothesis H_0.

Example 3: Processor manufacturing company, Intel claims that they have achieved the 10 GHz clock speed in their latest processor model, fastest ever in the world. The company wants to prove this to others.

Here company wants to prove their claim, therefor their claim will be alternative hypothesis H_1.

So, the alternative hypothesis,

$$H_1 : Clock\ speed \geq 10\ GHz$$

and the null hypothesis,

$$H_0 : Clock\ speed < 10\ GHz$$

But if one of the competitors of the company wants to discredit the company's claim, then the company's claim will be taken as null hypothesis H_0.

Level of Significance - α
It's the likelihood of making a mistake in testing given hypothesis. It's measure of how strong observed sample data have to be before determining the results. The test has the property that if it is given that H_0 is true, then the probability that it will be rejected is less than equal to α. Formally, it's the probability of type 1 error.

We want this error proportion, α to be as minimum as possible i.e., $\alpha = 0.01$ (1% of error). Suppose if a person being sent to the court for some crime, then we don't want that he or she will be convicted when it is given that he or she is innocent.

Level of Confidence - $1 - \alpha$
It's the probability of failing to reject the null hypothesis. Means if it is given that H_0 is true, then the probability that H_0 will not be rejected is level of confidence. It's just the complement of level of significance.

Example 4: Let's say about some test it is given that, test results in rejection of H_0 at 10% level of significance. How do we interpret it?
Level of significance is 10%, it means that when it's given that H_0 is true, only 10% of the times test rejects H_0.
Here level of confidence, $1 - \alpha = 1 - 0.1 = 0.9$, that is equal to 90%.

3 Test concerning Mean When Variance Is Known

Let $X_1, X_2, ..., X_n$ be the sample from a normal distribution. This distribution have unknown mean - μ and known variance - σ^2.

For some specified value of μ_0, we want to test the null hypothesis,

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis,

$$H_1 : \mu \neq \mu_0$$

Now we know that the point estimator of μ is the sample mean,

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

H_0 is rejected when this estimator of μ, $\overline{X}$ is too large or too small as compared to μ_0. So, the rejection region, C of the test will be,

$$C = \left\{X_1, X_2, ..., X_n \mid |\overline{X} - \mu_0| \geq c\right\} \tag{1}$$

for some suitable value c.

When H_0 is true, here $\mu = \mu_0$, still test rejects H_0 as $\overline{X}$ is sufficiently far from μ. So, here probability of type 1 error, level of significance - α is given by,

$$\boxed{P\left[|\overline{X} - \mu_0| \geq c\right] = \alpha} \tag{2}$$

So, if $\overline{X}$ falls in the interval $[\mu_0 - c,\ \mu_0 + c]$ we are assured that the null hypothesis H_0 is correct with the sufficiently small error probability α, otherwise reject it.

Now the thing is remained to be calculated is the suitable value of c. We know that when $\mu = \mu_0$, $\overline{X}$ is having normal distribution with mean μ_0 and STD $\sigma/\sqrt{n}$.

So standard normal variable will be,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{3}$$

Now from equation (1),

$$\begin{aligned}
P\left[\frac{|\overline{X} - \mu_0|}{\sigma/\sqrt{n}} \geq \frac{c}{\sigma/\sqrt{n}}\right] &= \alpha \\
\therefore P\left[|Z| \geq \frac{c\sqrt{n}}{\sigma}\right] &= \alpha & \text{(From eq (3))} \\
\therefore P\left[\left\{Z \leq -\frac{c\sqrt{n}}{\sigma}\right\} \cup \left\{Z \geq \frac{c\sqrt{n}}{\sigma}\right\}\right] &= \alpha
\end{aligned}$$

As we know Z is having normal distribution which is symmetric about the line $Z = 0$ so,

$$2P\left[Z \geq \frac{c\sqrt{n}}{\sigma}\right] = \alpha$$
$$\therefore P\left[Z \geq \frac{c\sqrt{n}}{\sigma}\right] = \frac{\alpha}{2} \tag{4}$$

And we know that,

$$P\left[Z \geq Z_{\alpha/2}\right] = \frac{\alpha}{2} \tag{5}$$

Thus from eq (4) and eq (5),

$$\frac{c\sqrt{n}}{\sigma} = Z_{\alpha/2}$$
$$So, \boxed{c = \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}} \tag{6}$$

So, reject the null hypothesis H_0 if

$$|\overline{X} - \mu_0| \geq \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}$$

Therefore in terms of Z value,

$$\boxed{\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| \geq Z_{\alpha/2}} \tag{7}$$

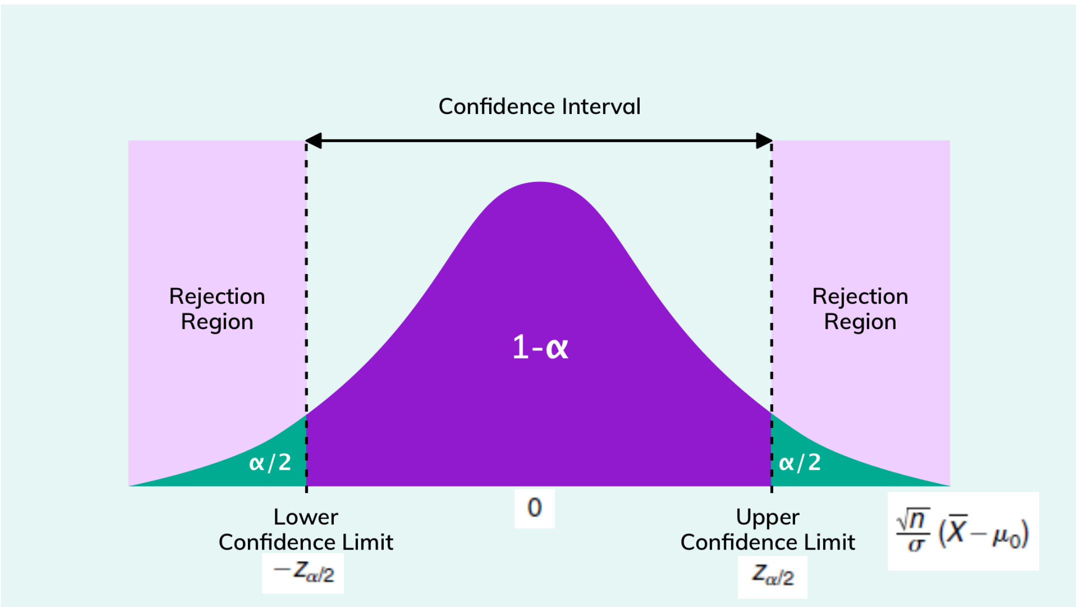

Figure 1: Standard Normal Distribution of Z with Confidence Interval and Rejection Region

Not reject H_0 otherwise.

Here $Z_{\alpha/2}$ and $-Z_{\alpha/2}$ are respectively called **Upper Confidence Limit** and **Lower Confidence Limit**.

From the figure 1, the region(interval) between the $Z_{\alpha/2}$ and $-Z_{\alpha/2}$ is **Confidence Interval** and the two sided regions are **Rejection Regions.** making total error probability α.

Example 5: Suppose in a production company, a technician claims that a machine produces 490 units per day with STD of 16.5. Manager wants to check the efficiency of production. So, he randomly picks up 50 machines and observe the statistics. Here 10% of level of significance is allowed. It comes out that the observed production mean is 486 units per day for a machine. Verify technician's claim.

Here the null hypothesis H_0 will be the technician's claim. Level of significance, $\alpha = 0.1$. So, $Z_{\alpha/2} = Z_{0.05} = 1.645$.
So,

$$\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| = \frac{\sqrt{50}}{16.5}|486 - 490|$$
$$\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| = 1.714$$

Therefore,

$$\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| \geq Z_{\alpha/2}$$

Thus, technician's claim is rejected.

Here If $\alpha = 0.05$ is to be taken instead of $\alpha = 0.1$ then, $Z_{\alpha/2} = Z_{0.025} = 1.96$.

Clearly,

$$\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| < Z_{\alpha/2}$$

And so for $\alpha = 0.05$, technician's claim is not rejected.

I've prepared two different plots in MATLAB for $\alpha = 0.1$ and $\alpha = 0.05$. We can show from the plot when $\alpha = 0.1$ the Z value, $\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| = 1.714$ falls in rejection region, so it's rejected. while for $\alpha = 0.05$ it falls in not rejection region, that is confidence interval so it's not rejected.

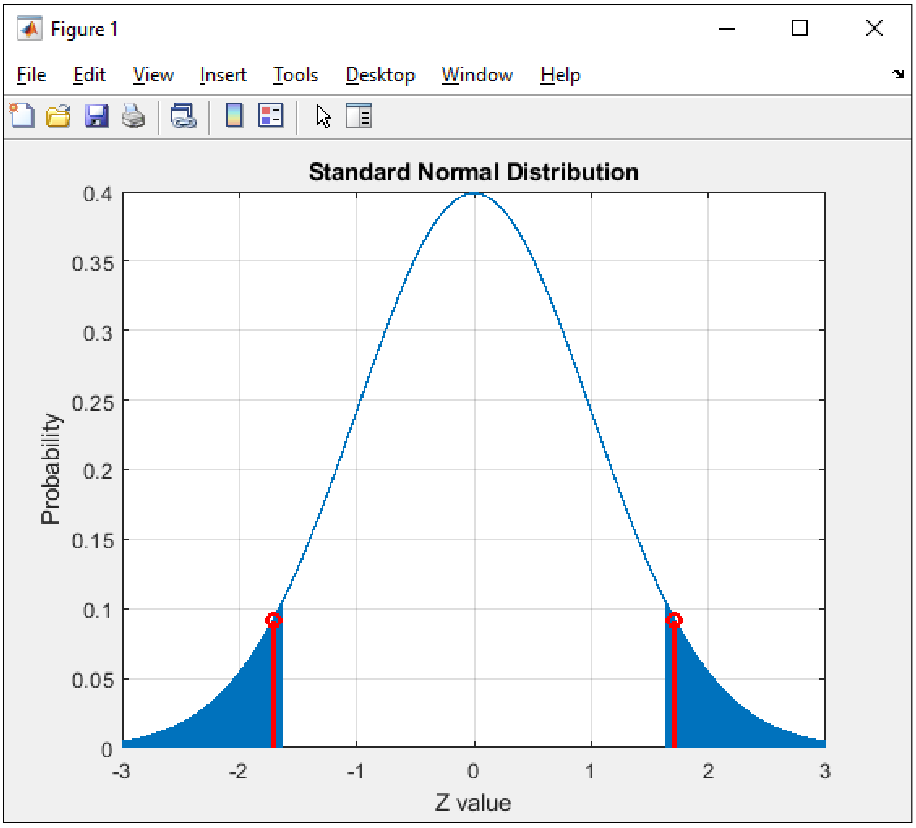

Figure 2: **For Level of Significance $\alpha = 0.1$**

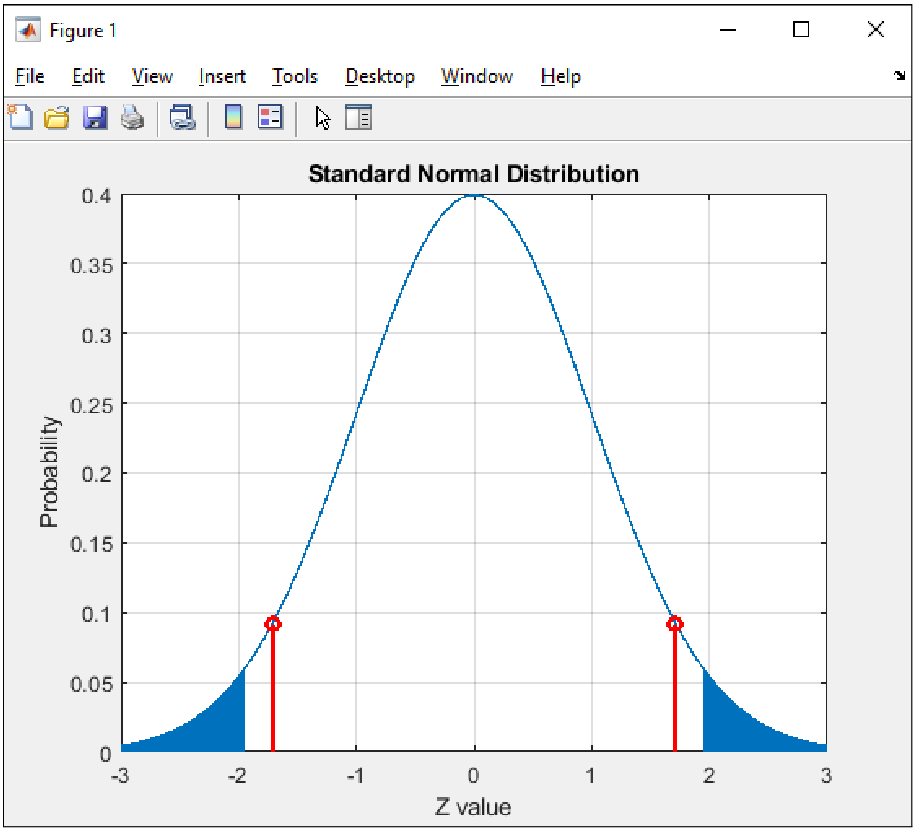

Figure 3: **For Level of Significance $\alpha = 0.05$**

4 p - Value

p - value is the critical significance level, that is the smallest significance value at which the null hypothesis H_0 is rejected. More formally, it gives the probability that the data as unsupportive of H_0 as those observed will occur when H_0 is true.

By using the p - value one can easily determine whether to reject or not to reject the null hypothesis H_0.

- **p - value** $\leq \alpha$**:** Null hypothesis H_0 is rejected.
- **p - value** $> \alpha$**:** Null hypothesis H_0 is not rejected.

If a test has small p - value(0.01 or 0.05) then it's most likely that the null hypothesis H_0 may be rejected. Smaller the p - value means higher chances for the falsity of H_0.

The p - value can't tell you whether your alternative hypothesis is correct, or why. It can only tell you whether the null hypothesis is supported or not.