greatlearning
Learning for Life

# Statistical Learning - Hypothesis Testing -Week 3

# Topics covered in Week 3

- Sampling distribution
- Central Limit Theorem
- Confidence intervals
- Hypothesis Formulation
- Null and Alternative Hypothesis
- Type I and Type II Errors
- Hypothesis Testing
  - One tailed v/s two tailed test
  - Test of mean
  - Test of proportion
  - Test of variance
  - One way ANOVA
- Hands-on exercises

greatlearning
Learning for Life

# Session Agenda

- Understand the need to sample data
- Central Limit Theorem
- “Hypothesis” in the context of statistics
- Type I and Type II Errors
- Confidence intervals
- Hypothesis Testing
  - Test of mean
  - Test of proportion
  - Test of variance
  - One way ANOVA
- Case Study
- Questions

Need for sampling as opposed to using the entire population for analysis:

- Time factor
- Effort factor

*It is usually not feasible to make a complete census of a population because of time and budget constraints. Therefore, a sample of the population is used to make inferences about the whole population. The goal of this type of sampling is to collect data that are representative of the entire population of interest.*

# Central Limit Theorem

- “Sampling Distribution of the mean of any independent random variable will be normal”
- This applies to both discrete and continuous distributions.
- The random variable should have a well defined mean and variance (standard deviation).
- Applicable even when the original variable is not normally distributed.

*Let’s watch central limit theorem in action. (please open the notebook titled ‘Central Limit Theorem’)*

# Hypothesis

An assumption about certain characteristics of a population.

Null hypothesis (H0) -> The hypothesis that does not challenge the status quo

Alternative hypothesis(Ha) -> The hypothesis that challenges the status quo

# Type I and Type II Errors

greatlearning
Learning for Life

Type I Error:

- Rejection of null hypothesis when it should not have been rejected.
- Incorrectly rejecting the null hypothesis.

Type II Error:

- Failure to reject the null hypothesis, when it should have been rejected.
- Incorrectly not rejecting the null hypothesis.

| Decision/ Reality | $H_o$ True (Should not reject) | $H_o$ False (Should reject) |
|---|---|---|
| Reject $H_o$ | Type I Error ($\alpha$) | Correct Rejection (No error) |
| Fail to Reject $H_o$ | Correct Decision (No error) | Type II Error ($\beta$) |

Causes of Type I and Type II Errors:

- By random chance, we may select a sample which is not representative of the population.
- Sampling techniques may be flawed.
- Assumptions in our null hypothesis may be flawed.

# Type of hypothesis tests

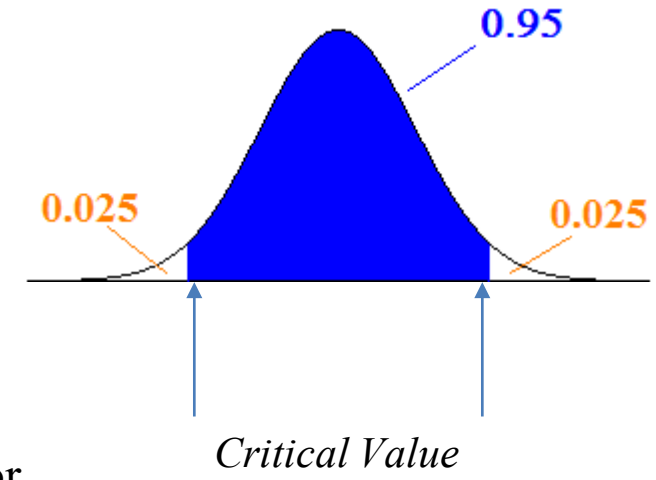
- Single sample or two or multiple samples
- One tailed or two tailed
- Tests of mean, proportion or variance

*Before we discuss hypothesis testing, let's do a quick recap on confidence intervals*

# Confidence Intervals

- 95% of all sample means ($\bar{x}$) are hypothesized to be in this region.

⇒ This is called as 95% confidence interval.




- If sample mean is in the blue region, we fail to reject the null hypothesis
- If sample mean is in the white region, we reject the null hypothesis.

- Here, $\alpha = 0.05$

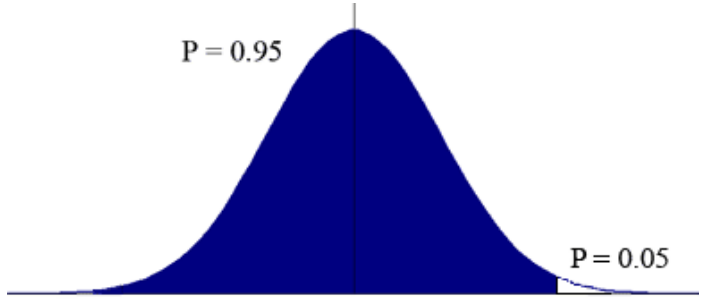
⇒ $\alpha$ is the level of significance or our tolerance level towards making a Type I error.

- If the null hypothesis is correct, ($\alpha$ * 100)% of the sample means should lie in the rejection region.

In case of one-tailed situation:

- All of $\alpha$ is in one tail or the other, depending on the alternative hypothesis.
- $H_a$ points to the tail, where the critical value and the rejection region are.

(Case when observed mean > hypothesized mean)



## Example scenario to perform an hypothesis test

greatlearning
Learning for Life

A study was done to see the effect of presence of dogs as pets on kids (ages 10 to 18). Two groups of teenagers, one group with teenagers who owned a dog for minimum 5 years and another group of kids who never owned a dog, were presented a questionnaire and scores were computed. High score corresponds to higher cheerfulness and low score corresponds to lower cheerfulness.

Do dogs have a significant effect (either positive or negative) on the cheerfulness of kids?

Dog: 6.6, 7.8, 4.6, 7.8, 7, 8, 9, 9, 8.8, 9.9, 8.5, 7.7, 8.6, 8, 7, 5.8, 7.4

No_dog: 9.8, 8.3, 7.1, 7.2, 8.1, 8.9, 6, 7, 7.5, 7.8, 7.6, 7.3, 6.4, 6.8, 7, 6.4, 7.9

*What are the null and alternative hypothesis?*

*Is it a right tailed or a left tailed test?*

*Is it a one sample or a two sample test?*

*Is it a test of mean, proportion or variance?*

*Which statistical test do you think is appropriate?*

**Let's perform a two sample t-test to check if there is a significant difference in means of the two samples**

Avg_score of kids with dogs, m1, s1 = 7.73, 1.24

Avg_score of kids without dogs, m2, s2 = 7.58, 1.23

|A – B| = 0.10

*Is the difference significant at 5% significance level?*

Ho: m1 = m2 ( pets have no effect on the cheerfulness of kids)

Ha: m1 ≠ m2 ( pets have an effect on the cheerfulness of kids)

alpha = 0.05

t_critical = ± 2.11 ( for a dof of 16, and a confidence of 95% in case of a two tailed test)

$$t = \frac{m1 - m2}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$$

t_statistic = **0.35**

*It is well inside the critical level. We could not prove that pets either increase or decrease the cheerfulness of kids. We fail to reject the null hypothesis*

greatlearning
Learning for Life

# What just happened:

- The test we just performed is called as 2-sample t-test or Independent samples t-test or student's t-test.
- We perform this test to see if there is a statistically significant difference between means of two independent groups.
- The null hypothesis will be that there is no difference in means
- The alternative hypothesis will be that there is a significant difference in means
- To perform this test, we will need one independent qualitative variable with two levels and one dependent quantitative variable.

# Example – 2

The following is the income data of blue collar workers who are at the same skill level. There are two groups of workers. Workers of the textile company “Lori’s and Co.” and workers of the general population. We want to check weather the variance in the income of Lori’s is higher than of the general population.

Mean and variance of general population = 100,  16.11

Lori’s : 105, 95, 90,  98, 110, 104, 108, 111, 110, 102, 98, 105, 105, 105, 115

*What are the null and alternative hypothesis?*

*Is it a right tailed or a left tailed test?*

*Is it a test of mean, proportion or variance?*

*Which statistical test do you think is appropriate?*

Pop: m1, v1 = 100, 16.11

Lori's: m2, v2 = 104.06, 40.99

N = 15

Dof = 15 – 1 = 14

Ho: v2 = v1 ( variance in income of Lori's is same as the population)

Ha: v2 > v1 (variance in income of Lori's is higher than the population)

alpha = 0.05

Chi_critical = 23.68 ( any value beyond 23.68 falls in the rejection region)

$$\text{Chi\_statistic} => x^2 = dof\left(\frac{v2}{v1}\right)$$

$$x^2 = 35.62$$

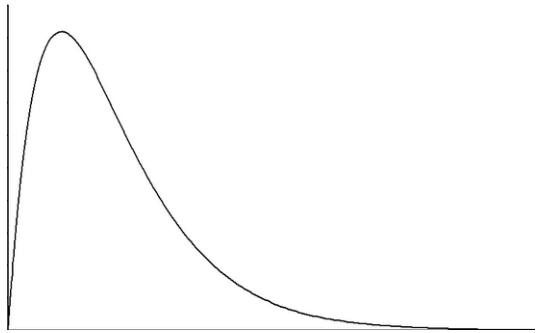
$$x^2 > 23.68$$

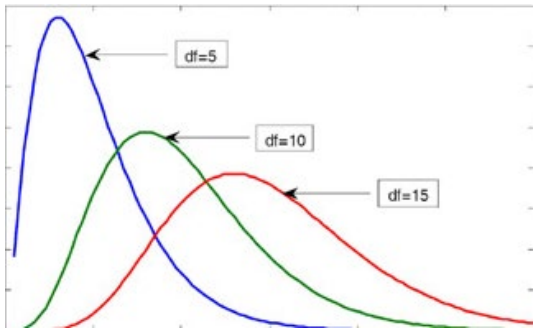
*It is well beyond the critical value. The variation in income of blue collar workers at Lori's is significantly higher than the population variance.*

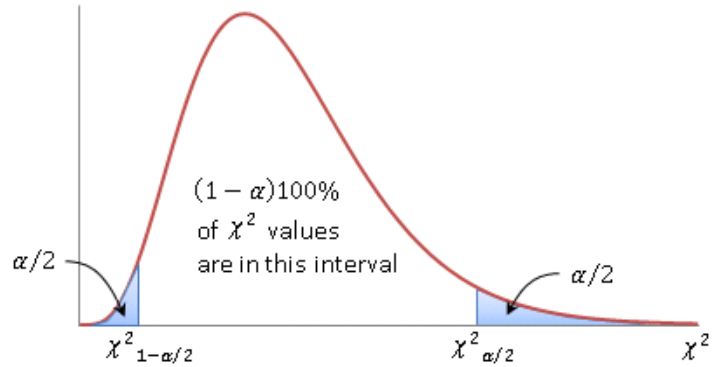
# Chi square test of variance

When we take many samples of the same size from a normal population and find the sample means, they follow a normal distribution.

When we take many samples of the same size from a normal population and find the sample variances, they DO NOT follow a normal distribution; instead they follow a **chi-square ($\chi^2$) distribution** , which is dependent on the degrees of freedom.






- Area under the curve is always 1.
- Cumulative Probability runs from right to left; 1 is towards the left end, while 0 is towards the right.

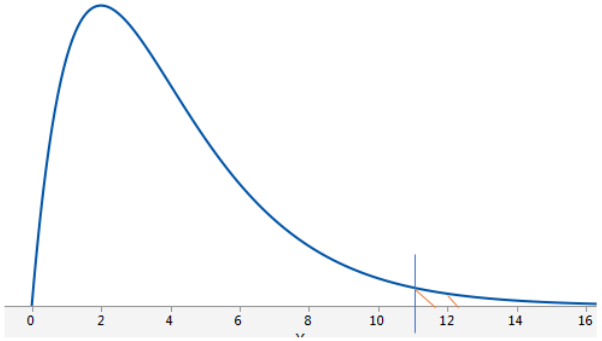
# Chi square test of variance

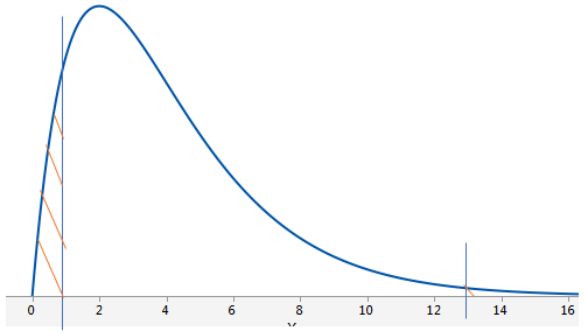
**Chi-square ($\chi^2$)** test compares the population variance, with the hypothesized variance.

$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$ where, n = sample size

$s^2$ = sample variance and $\sigma^2$ = population variance (which we wish to test)

At $\alpha = 0.05$ and $n = 5$ ($df = 4$)




p-value: How much of the area is above the test-statistic? *(Does test statistic fall in the rejection region?)*

If it is less than the specific α, we reject the null hypothesis

greatlearning
Learning for Life

# Example 3

Three groups of samples of factory emissions of different plants of the same company were collected. The score is computed based on the composition of the emissions. We want to find out if there is any inconsistency or difference across the three groups.

A = 57,56,58,58,56,59,56,55,53,54,53,42,44,34,54,54,34,64,84,24

B = 49,47,49,47,49,47,49,46,45,46,41,42,41,42,42,42,14,14,34

C = 49,48,46,46,49,46,45,55,61,45,45,45,49,54,44,74,54,84,39

# Hypothesis of One-Way ANOVA

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$

All population means are equal

$H_a$ : Not all of the population means are equal

For at least one pair, the population means are unequal.

Dof(between) = k – 1 = 3 – 1 = 2

Dof(within) = N – k = 59 – 3 = 56

Dof(total) = 56 + 2 = 58

For the above degrees of freedom,

$$F_{critical} = 3.161$$

Mean(A) = 52.45

Mean(B) = 41.36

Mean(C) = 51.45

Overall Mean = 2864/59 = 48.54

$$SS_{total} = \sum(x_i - overall\_mean)^2 = 8548.64$$

$$SS_{within} = \sum(a_i - mean(A))^2 + \sum(b_i - mean(B))^2 + \sum(c_i - mean(C))^2 = 7096.32$$

$$SS_{between} = SS_{total} - SS_{within} = 1452.32$$

Contd..

greatlearning
Learning for Life

$$MS_{between} = \frac{SS_{between}}{dof_{between}} = 726.16$$

$$MS_{within} = \frac{SS_{within}}{dof_{within}} = 126.72$$

$$F_{statistic} = \frac{MS_{between}}{MS_{within}} = 5.73$$

$$F_{statistic} > F_{critical}$$

*Since our f-statistic is beyond the critical value, we reject the null.*

# One way ANOVA:

- The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups

- For one-way ANOVA, the ratio of the between-group variability to the within-group variability follows an F-distribution when the null hypothesis is true. When you perform a one-way ANOVA for a single study, you obtain a single F-value

Example 4:

A study found that as of 2015, In the world population of 7.7 billion, 36.7 million are diagnosed HIV positive. Of the 1.3 billion Indians, 2.1million were diagnosed positive.

Is the population of aids significantly different in India that it is in the world?

H0: $P_{(aids\ in\ India)} = P_{(aids\ in\ world)}$

Ha: $P_{(aids\ in\ India)} \neq P_{(aids\ in\ world)}$

# Solution:

Proportion of aids in the world = 0.0047 = 0.47%

Proportion of aids in India = 0.0016 = 0.16%

Overall proportion =

$$Z_{statistic} = \frac{(P_{sample} - P_0)}{\sqrt{\frac{(P_0(1-P_0)}{N}}}$$

$$Z_{statistic} = \frac{(0.0016 - 0.0047)}{\sqrt{\frac{(0.0047\ (1-0.0047)}{2100000}}} = -65.68$$

*The proportion of aids population is significantly different from the world population.*

# Case Study:

- Z-test to compare means
- One sample t-test
- Two sample t-test (paired)
- Paired sample t-test

# Dataset information:

- The data used to perform one of the tests is of different wines and their attributes.
- We will be using one continuous column which in a way quantifies how good a particular wine is across different attributes.

Lets go ahead and implement a few hypothesis tests in Python.

Let’s summarize what we have learnt….


# Questions?