08 April 2023 20:05

F-dist

HT

Z-test

t-test

Chi-square

ANOVA

f-test

One way

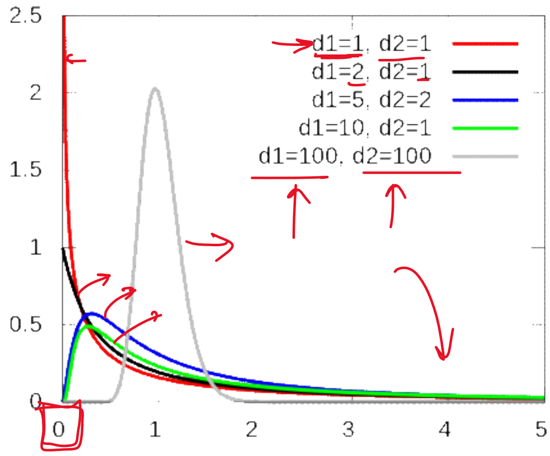
Two way

F - Distribution

08 April 2023 13:04

$\rightarrow$ chi-square $\rightarrow (\text{normal})^2$

1. Continuous probability distribution: The F-distribution is a continuous probability distribution used in statistical hypothesis testing and analysis of variance (ANOVA).
2. Fisher-Snedecor distribution: It is also known as the Fisher-Snedecor distribution, named after Ronald Fisher and George Snedecor, two prominent statisticians.
3. Degrees of freedom: The F-distribution is defined by two parameters - the degrees of freedom for the numerator (df1) and the degrees of freedom for the denominator (df2).
4. Positively skewed and bounded: The shape of the F-distribution is positively skewed, with its left bound at zero. The distribution's shape depends on the values of the degrees of freedom.
5. Testing equality of variances: The F-distribution is commonly used to test hypotheses about the equality of two variances in different samples or populations.
6. Comparing statistical models: The F-distribution is also used to compare the fit of different statistical models, particularly in the context of ANOVA.
7. F-statistic: The F-statistic is calculated by dividing the ratio of two sample variances or mean squares from an ANOVA table. This value is then compared to critical values from the F-distribution to determine statistical significance.
8. Applications: The F-distribution is widely used in various fields of research, including psychology, education, economics, and the natural and social sciences, for hypothesis testing and model comparison.



SSB $\leftarrow \chi_1^2 \rightarrow df_1$

SSW $\leftarrow \chi_2^2 \rightarrow df_2$

$$F = \boxed{\frac{\chi_1^2/df_1}{\chi_2^2/df_2}} \rightarrow \text{f-distribution}$$

$df_1 \; df_2$

One way ANOVA test

08 April 2023 13:12

One-way ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more independent groups to determine if there are any significant differences between them. It is an extension of the t-test, which is used for comparing the means of two independent groups. The term "one-way" refers to the fact that there is only one independent variable (factor) with multiple levels (groups) in this analysis.

t-test → 2 sample t-test → 2 groups → two ways

1 class, Age, gender (2 col) → M, F, O

f / m / o

The primary purpose of one-way ANOVA is to test the null hypothesis that all the group means are equal. The alternative hypothesis is that at least one group mean is significantly different from the others.

Steps

- Define the null and alternative hypotheses.
- Calculate the overall mean (grand mean) of all the groups combined and mean of all the groups individually.
- Calculate the "between-group" and "within-group" sum of squares (SS).
- Find the between group and within group degree of freedoms
- Calculate the "between-group" and "within-group" mean squares (MS) by dividing their respective sum of squares by their degrees of freedom.
- Calculate the F-statistic by dividing the "between-group" mean square by the "within-group" mean square.

$\bar{\bar{X}}$

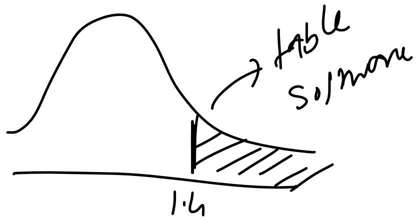
ANOVA table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS) (This is SS Divided by d.f.) and is an Estimation of Variance to be Used in F-ratio	F-ratio
Between samples or categories	SSB $n_1(\bar{X}_1 - \bar{\bar{X}})^2 + \cdots + n_k(\bar{X}_k - \bar{\bar{X}})^2$ 24	2 $(k-1)$	$\frac{SS \text{ between}}{(k-1)}$	$f = \frac{MS \text{ between}}{MS \text{ within}}$
Within samples or categories	SSW $\sum(X_{1i} - \bar{X}_1)^2 + \cdots + \sum(X_{ki} - \bar{X}_k)^2$, $i = 1, 2, 3, \ldots$ 52	4 $(n-k)$ 6	$\frac{SS \text{ within}}{(n-k)}$	
Total	SST $\sum(X_{ij} - \bar{\bar{X}})^2$, $i, j = 1, 2, 3, \ldots$ 76	$(n-1)$ 8		

F-ratio → test statistic

$$F = \frac{\frac{SSB}{df_{ssb}}}{\frac{SSW}{df_{ssw}}}$$

f-dist



- Calculate the p-value associated with the calculated F-statistic using the F-distribution and the appropriate degrees of freedom. The p-value represents the probability of obtaining an F-statistic as extreme or more extreme than the calculated value, assuming the null hypothesis is true.
- Choose a significance level (alpha), typically 0.05.
- Compare the calculated p-value with the chosen significance level (alpha).

  a. If the p-value is less than or equal to alpha, reject the null hypothesis in favour of the alternative hypothesis, concluding that there is a significant difference between at least one pair of group means.
  b. If the p-value is greater than alpha, fail to reject the null hypothesis, concluding that there is not enough evidence to suggest a significant difference between the group means.

It's important to note that one-way ANOVA only determines if there is a significant difference between the group means; it does not identify which specific groups have significant differences. To determine which pairs of groups are significantly different, post-hoc tests, such as Tukey's HSD or Bonferroni, are conducted after a significant ANOVA result.

Example 1

08 April 2023 13:22

number cont

A	B	C
3	1	8
6	8	6
3	9	10

$\bar{X}_A$ $\bar{x}_B$ $\bar{x}_c$

ANOVA

$H_0: \mu_A = \mu_B = \mu_C$

H_1: atleast 1 of them is signifi diff

$n = 9 \quad K = 3$

$\boxed{\bar{\bar{x}} = 6} \quad \bar{X}_A = 4 \quad \bar{X}_B = 6 \quad \bar{X}_c = 8$

$\boxed{SST} \rightarrow$ Sum of square Total $\rightarrow df = n-1 = 9-1 = 8$

$(6-3)^2 + (6-6)^2 + (6-3)^2 + (6-1)^2 + (6-8)^2 + (6-9)^2 +$
$(6-8)^2 + (6-6)^2 + (6-10)^2$

$9 + 0 + 9 + 25 + 4 + 9 + 4 + 0 + 16 = \boxed{76}$

$\boxed{SSW} \rightarrow$ sum of squares within

$\downarrow df = 6$

$n-k = 9-3 = 6$

A	B	C
3	1	8
6	8	6
3	9	10

$\bar{X}_A$ $\bar{X}_B$ $\bar{x}$

$\boxed{4}$ $\boxed{6}$ 8

(2) (2) (2)

$\bar{\bar{x}} = 6$

$(4-3)^2 + (4-6)^2 + (4-3)^2 +$
$(6-1)^2 + (6-8)^2 + (6-9)^2 +$
$(8-8)^2 + (8-6)^2 + (8-10)^2$ $\boxed{SSW = 52}$

$1 + 9 + 1 + 25 + 4 + 9 + 0 + 4 + 4$

$SSB \rightarrow df = 2$

$3 \times (6-4)^2 + 3 \times (6-6)^2 + 3 \times (6-8)^2$

$3 \times 4 + 0 + 3 \times 4 = \boxed{24 = SSB}$

quantity value df

SSB 24 [2] $F = \frac{24}{2}$ 12×6

quantity

→ SSB 24 2

→ SSW 52 6

→ $\boxed{SST}$ 76 $\boxed{8}$

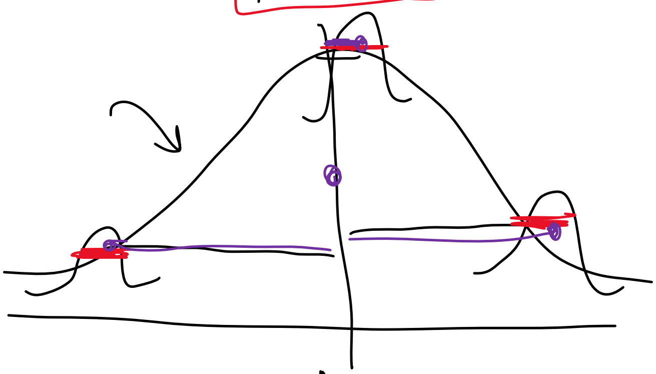
$$F = \frac{\frac{24}{2}}{\frac{52}{6}} = \frac{12 \times 6}{52} = \frac{72}{52}$$

$\boxed{f = 1.4}$

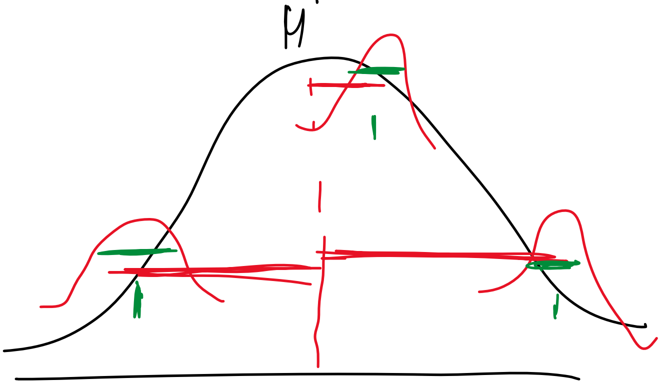
SST = SSW + SSB

3 pop (A), (B), (C)

μ μ μ



μ



μ

SST

interval variance grops (SSW)

SSB

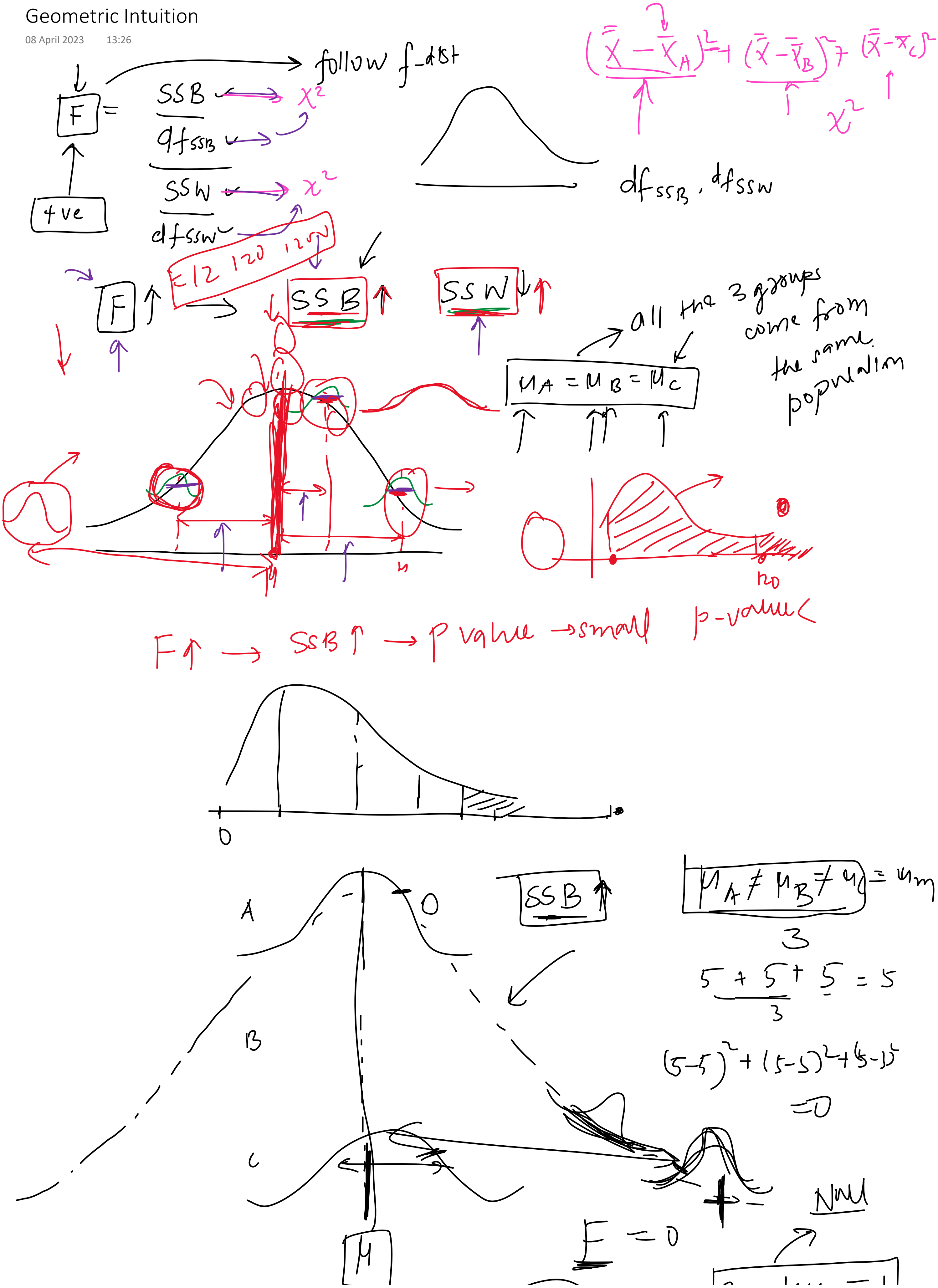
p-value = 0.31 > α (0.05)

Null hy can't reject

$\boxed{\mu_A = \mu_B = \mu_C}$

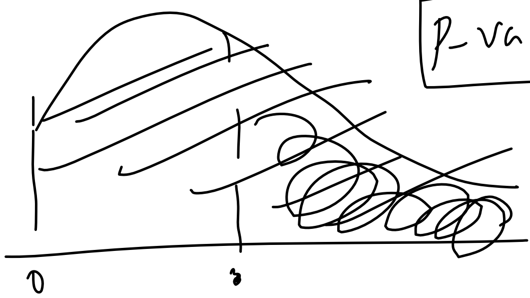
Geometric Intuition

08 April 2023 13:26



$\boxed{\mu}$

$t = 0$



$\boxed{p\text{-value} = 1}$

Assumptions

08 April 2023 16:48

Assumptions

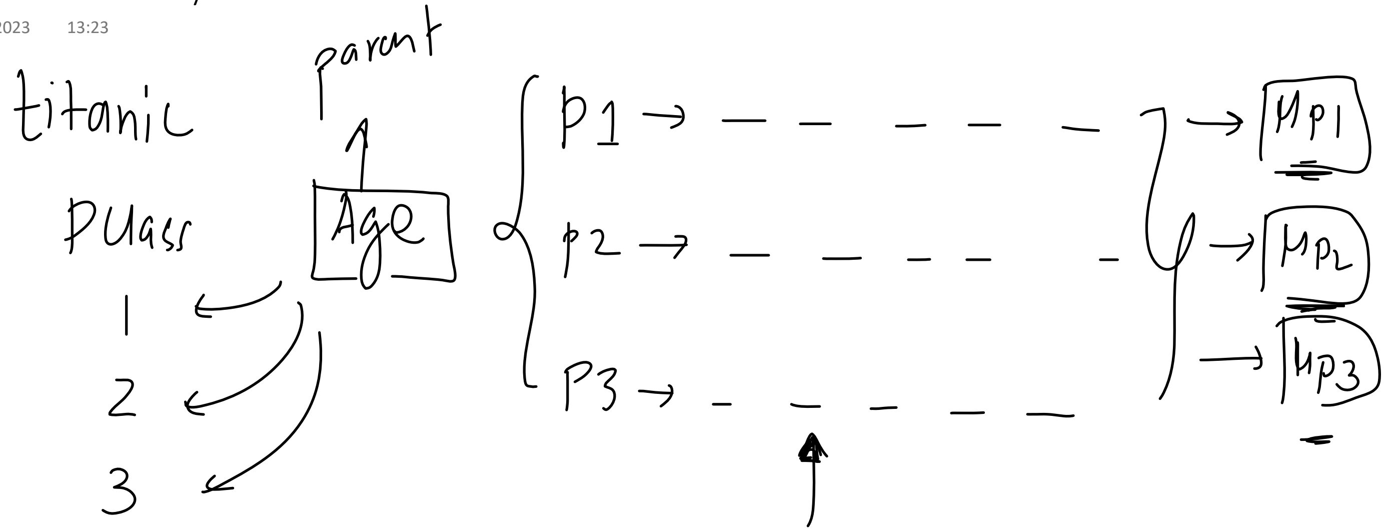
1. Independence: The observations within and between groups should be independent of each other. This means that the outcome of one observation should not influence the outcome of another. Independence is typically achieved through random sampling or random assignment of subjects to groups.

2. Normality: The data within each group should be approximately normally distributed. While one-way ANOVA is considered to be robust to moderate violations of normality, severe deviations may affect the accuracy of the test results. If normality is in doubt, non-parametric alternatives like the Shapiro-wilk test can be considered.

3. Homogeneity of variances: The variances of the populations from which the samples are drawn should be equal, or at least approximately so. This assumption is known as homoscedasticity. If the variances are substantially different, the accuracy of the test results may be compromised. Levene's test or Bartlett's test can be used to assess the homogeneity of variances. If this assumption is violated, alternative tests such as Welch's ANOVA can be used.

Python Case Study

08 April 2023 13:23



Post-hoc Test

08 April 2023 13:23

1 - 39 2 - 29 - 3 → 24

Post hoc tests, also known as post hoc pairwise comparisons or multiple comparison tests, are used in the context of ANOVA when the overall test indicates a significant difference among the group means. These tests are performed after the initial one-way ANOVA to determine which specific groups or pairs of groups have significantly different means.

The main purpose of post hoc tests is to control the family-wise error rate (FWER) and adjust the significance level for multiple comparisons to avoid inflated Type I errors. There are several post hoc tests available, each with different characteristics and assumptions. Some common post hoc tests include:

1. Bonferroni correction: This method adjusts the significance level (α) by dividing it by the number of comparisons being made. It is a conservative method that can be applied when making multiple comparisons, but it may have lower statistical power when a large number of comparisons are involved.
2. Tukey's HSD (Honestly Significant Difference) test: This test controls the FWER and is used when the sample sizes are equal and the variances are assumed to be equal across the groups. It is one of the most commonly used post hoc tests.

When performing post hoc tests, it is essential to choose a test that aligns with the assumptions of your data (e.g., equal variances, equal sample sizes) and provides an appropriate balance between controlling Type I errors and maintaining statistical power.

3 groups

$M_A = M_B = M_C$

less

pvay ← A - B → t-test →

pvy ← B - C → t-test → less

pv ← C - A → t-test →

$\alpha = 0.05$ → type I error

5% ?

0.05

$0.95 \times 0.95 \times 0.95$

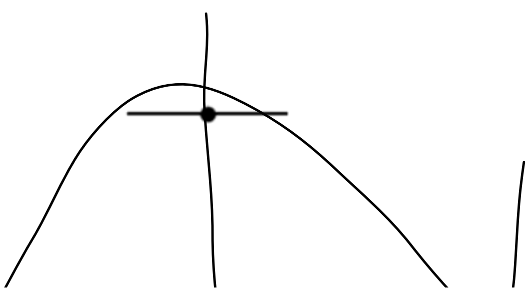
$1 - 0.85 = 0.15$

15%

$\frac{0.05}{3}$ → $0.05 \times 0.05 \times 0.05$

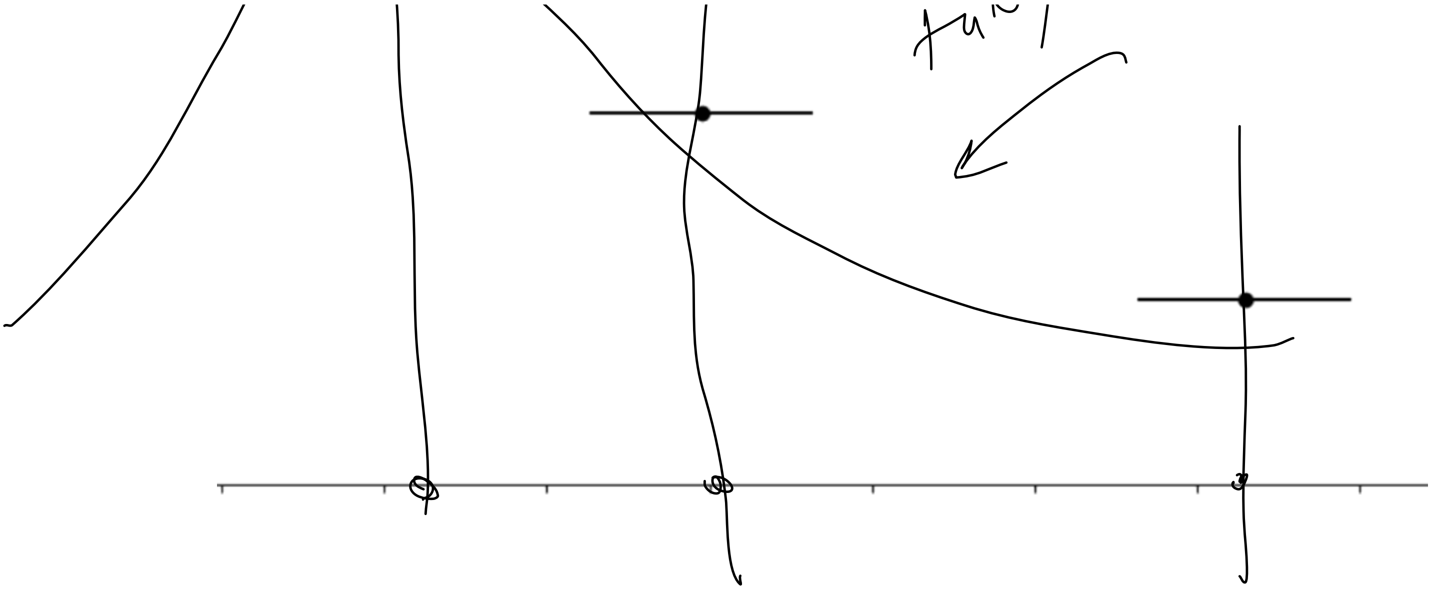
reduce

$\frac{0.15}{3} = 0.05$ → 5%



ANOVA

Tukey



Why t-test is not used for more than 3 categories?

08 April 2023 13:23

1. Increased Type I error: When you perform multiple comparisons using individual t-tests, the probability of making a Type I error (false positive) increases. The more tests you perform, the higher the chance that you will incorrectly reject the null hypothesis in at least one of the tests, even if the null hypothesis is true for all groups.
2. Difficulty in interpreting results: When comparing multiple groups using multiple t-tests, the interpretation of the results can become complicated. For example, if you have 4 groups and you perform 6 pairwise t-tests, it can be challenging to interpret and summarize the overall pattern of differences among the groups.
3. Inefficiency: Using multiple t-tests is less efficient than using a single test that accounts for all groups, such as one-way ANOVA. One-way ANOVA uses the information from all the groups simultaneously to estimate the variability within and between the groups, which can lead to more accurate conclusions.

2 sample indep

AB → 5
BC → 5
CA → 5
→ 15%

Applications in Machine Learning

08 April 2023 13:27

1. Hyperparameter tuning: When selecting the best hyperparameters for a machine learning model, one-way ANOVA can be used to compare the performance of models with different hyperparameter settings. By treating each hyperparameter setting as a group, you can perform one-way ANOVA to determine if there are any significant differences in performance across the various settings.

2. Feature selection: One-way ANOVA can be used as a univariate feature selection method to identify features that are significantly associated with the target variable, especially when the target variable is categorical with more than two levels. In this context, the one-way ANOVA is performed for each feature, and features with low p-values are considered to be more relevant for prediction.

3. Algorithm comparison: When comparing the performance of different machine learning algorithms, one-way ANOVA can be used to determine if there are any significant differences in their performance metrics (e.g., accuracy, F1 score, etc.) across multiple runs or cross-validation folds. This can help you decide which algorithm is the most suitable for a specific problem.

4. Model stability assessment: One-way ANOVA can be used to assess the stability of a machine learning model by comparing its performance across different random seeds or initializations. If the model's performance varies significantly between different initializations, it may indicate that the model is unstable or highly sensitive to the choice of initial conditions.