MATH 5010 –Foundations of Statistical Theory and Probability

ANOVA –Applications

Instructor: He Wang
Department of Mathematics
Northeastern University

1. **Review and summary of One-Way ANOVA**

2. **Examples Computation**

3. **Applications**

Randomized Block Design

➢ Review and summary of One-Way ANOVA

Background: **One-Way ANOVA**

ANOVA is a statistical method used to compare the means of k populations.

This method is appropriate in the following settings:

1. When k independent random samples are drawn from k populations.
2. When k different treatments are applied to a homogeneous group of experimental units.

The group is subdivided into k subgroups, and each treatment is applied to one subgroup.

One-Way Analysis of Variance Assumptions:

The ANOVA model requires:

1. Random samples are independently selected from k populations.
2. The k populations are approximately normally distributed.
3. All k population variances are equal.

Equivalently,

$$X_{ij} = \mu_i + \epsilon_{ij}$$

$i = 1, \ldots, k$: Group index.

$j = 1, \ldots, n_i$: Observation index within group i.

μ_i :(treatment) mean of group i

ϵ_{ij}: IID error term $\sim Normal(0, \sigma^2)$.

Classic ANOVA Hypothesis

Null Hypothesis: All treatment means are exactly equal:

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_k$$

Alternative Hypothesis:

$$H_1: \theta_i \neq \theta_j \text{ for some } i, j$$

Treatments				
1	2	3	…	k
y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	y_{k3}
		y_{3n_3}		$\vdots$
y_{1n_1}				
	y_{2n_2}			y_{kn_k}

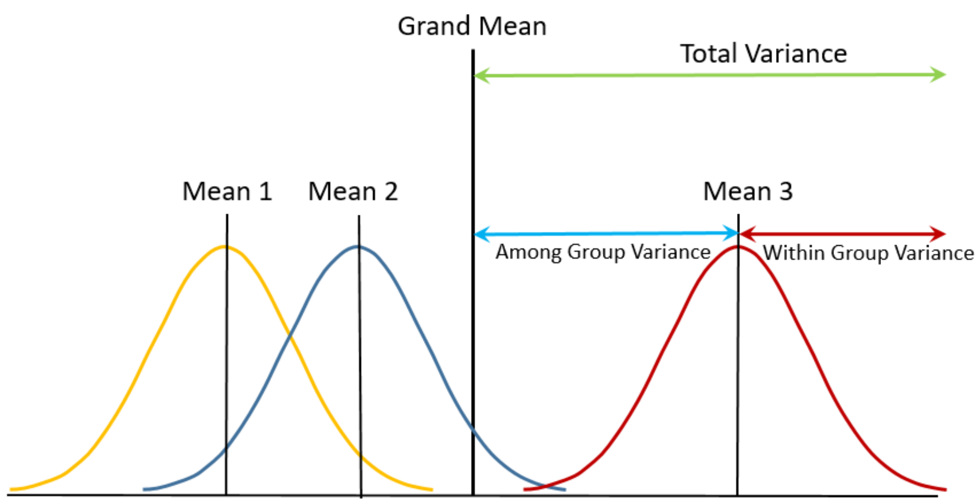
Group Mean and Overall Mean

Group Mean

$$\overline{Y_{i.}} = \frac{T_{i.}}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

Overall Mean

$$\overline{Y_{..}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{k} T_{i.} = \frac{1}{n}\sum_{i=1}^{k} n_i\overline{Y_{i.}}$$



Partitioning Sums of Squares

$$SS_{Total} = SS_{between} + SS_{Within}$$

- **Total** sum of squares(SST or SS_T or SS_{Total})

$$SS_{Total} := \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{..}\right)^2$$

- **Between Treatment** sum of squares (SS_B, SSB, or $SS_{bewteen}$,)

$$SS_{Between} := \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

- **Within(Error)** sum of squares (SS_E , SSW, SS_{within}, SS_{Error}):

$$SS_{Error} := \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

Mean Sum of Squares

- Mean **error** sum of squares

$$MS_E = \frac{SS_E}{N - k}$$

Further more, if $\theta_1 = \theta_2 = \cdots = \theta_k$, then SS_B is an unbiased estimator of σ^2, and

$$\frac{1}{\sigma^2} SS_E \sim \chi^2_{N-k}$$

- Mean **treatment** sum of squares

$$MS_B \coloneqq \frac{SS_B}{k - 1}$$

Under ANOVA assumptions, SS_E is an unbiased estimator of σ^2 and

$$\frac{1}{\sigma^2} SS_B \sim \chi^2_{k-1}$$

The ANOVA F-Test

Under the ANOVA assumptions, suppose $H_0\colon \theta_1 = \theta_2 = \cdots = \theta_k$ is true.

Then the **F-Statistic**

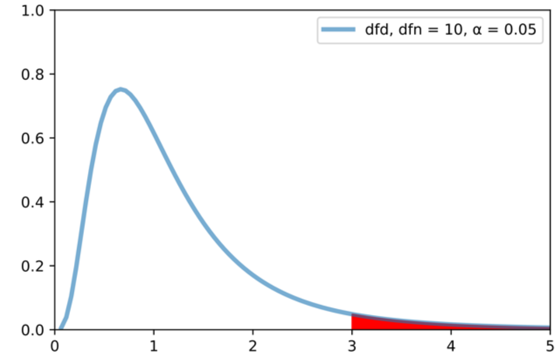
$$F := \frac{MS_B}{MS_W}$$

has a F-distribution with $k-1$ and $N-k$ degrees of freedom.

Hence, the p-value of the ANOVA-Test is

$$p\text{-value} = P(F > F_{obv})$$

Large F values F_{obv} provide evidence against H_0.



One-way ANOVA table

Source of Variation	Sum of Squares	Degrees of Freedom df	Mean Square (MS)	**F-Statistic**
Between Groups	$SS_{between}$	$(k-1)$	$MS_B = \frac{SS_B}{k-1}$	$F_{obv} = \frac{MS_B}{MS_E}$
Within Groups	SS_{Error}	$(N - k)$	$MS_E = \frac{SS_E}{N-k}$	
Total	SS_{Total}	$(N - 1)$		

***p*-value**$= P\left(F_{(k-1,N-k)} > F_{obv}\right)$

Example: Comparing Three Golf Ball Brands

A test was conducted to compare the mean distance (in yards) traveled by three brands of golf balls hit by a robotic golfer.

Brand A	Brand B	Brand C
251.2	263.2	269.7
245.1	262.9	263.2
248.0	265.0	277.5
251.1	254.5	267.4
260.5	264.3	270.5

Step 1: Summary Statistics.

i	$\bar{x}_i$	s_i^2	n_i
1	251.18	33.487	5
2	261.985	18.197	5
3	269.66	27.253	5

$$n = 15, \quad k = 3, \quad \text{Grand mean: } \bar{x}_{..} = \frac{1}{15}(5 \times 251.18 + 5 \times 261.985 + 5 \times 269.66) = 260.94.$$

Step 2: Treatment Sum of Squares.

$$SS_B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{..})^2 = 5[(251.18-260.94)^2+(261.985-260.94)^2+(269.66-260.94)^2] = 861.89.$$

Step 3: Error Sum of Squares.

$$SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2 = 4(33.487 + 18.197 + 27.253) = 315.75.$$

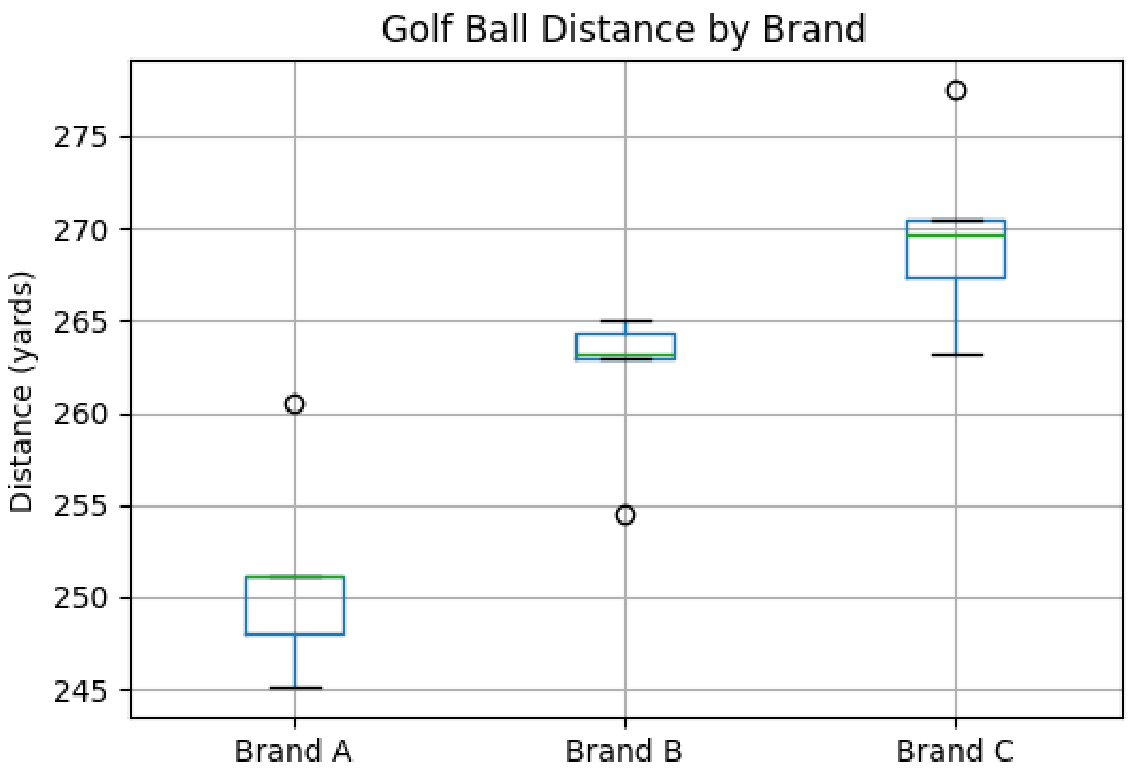
Step 4: ANOVA Table.

Source	**SS**	**df**	**MS**	**F**	**p-value**
Treatment	861.29	2	430.645	16.37	< 0.01
Error	315.75	12	26.312		
Total	1177.64	14			

Step 5: Conclusion.

$$F_{obs} \approx 16.37, \qquad p \approx 0.00037 < 0.01.$$

Since p < 0.01, we reject H_0 at the 1% level and conclude that the mean travel distances for at least one brand differ significantly.



https://drive.google.com/file/d/1Ndta5hdaHUwqRexp1r7Xk97MRiLAP-kf/view?usp=sharing

Example: Smoking and Heart Rate

Subject	Nonsmoker	Light Smoker	Moderate Smoker	Heavy Smoker
1	69	55	66	91
2	52	60	81	72
3	71	78	70	81
4	58	58	77	67
5	59	62	57	95
6	65	66	79	84
T_j	374	379	430	490
$\bar{Y}_{j\cdot}$	62.3	63.2	71.7	81.7

Goal. We test whether the true group means are equal:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \qquad \text{vs.} \qquad H_a : \text{ at least one } \mu_j \text{ differs.}$$

We use a one–way ANOVA F test with significance level $\alpha = 0.05$. There are $k = 4$ groups and $n = 24$ total observations.

Overall (Grand) Mean. The overall mean is

$$\bar{Y}_{..} = \frac{1}{n}\sum_{j=1}^{k} T_j = \frac{374 + 379 + 430 + 490}{24} = 69.7.$$

Between–Group Variability (Treatment Sum of Squares).

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{Y}_{j.} - \bar{Y}_{..})^2.$$

Since $n_j = 6$ for all j,

$$\text{SSTR} = 6\Big[(62.3 - 69.7)^2 + (63.2 - 69.7)^2 + (71.7 - 69.7)^2 + (81.7 - 69.7)^2\Big] = 1464.125.$$

Within–Group Variability (Error Sum of Squares).

$$\text{SSE} = \sum_{j=1}^{k}\sum_{i=1}^{6}(Y_{ij} - \bar{Y}_{j.})^2.$$

Computing group-by-group deviations yields:

$$\text{SSE} = 1594.833.$$

Degrees of Freedom and Mean Squares.

$$df_{\text{treat}} = k - 1 = 3, \qquad df_{\text{error}} = n - k = 20.$$

$$\text{MST} = \frac{\text{SSTR}}{k-1} = \frac{1464.125}{3} = 488.04,$$

$$\text{MSE} = \frac{\text{SSE}}{n-k} = \frac{1594.833}{20} = 79.74.$$

Test Statistic.

$$F = \frac{\text{MST}}{\text{MSE}} = \frac{488.04}{79.74} = 6.12.$$

The critical value for $\alpha = 0.05$ with $(3, 20)$ degrees of freedom is

$$F_{0.95;3,20} = 3.10.$$

Since

$$6.12 > 3.10,$$

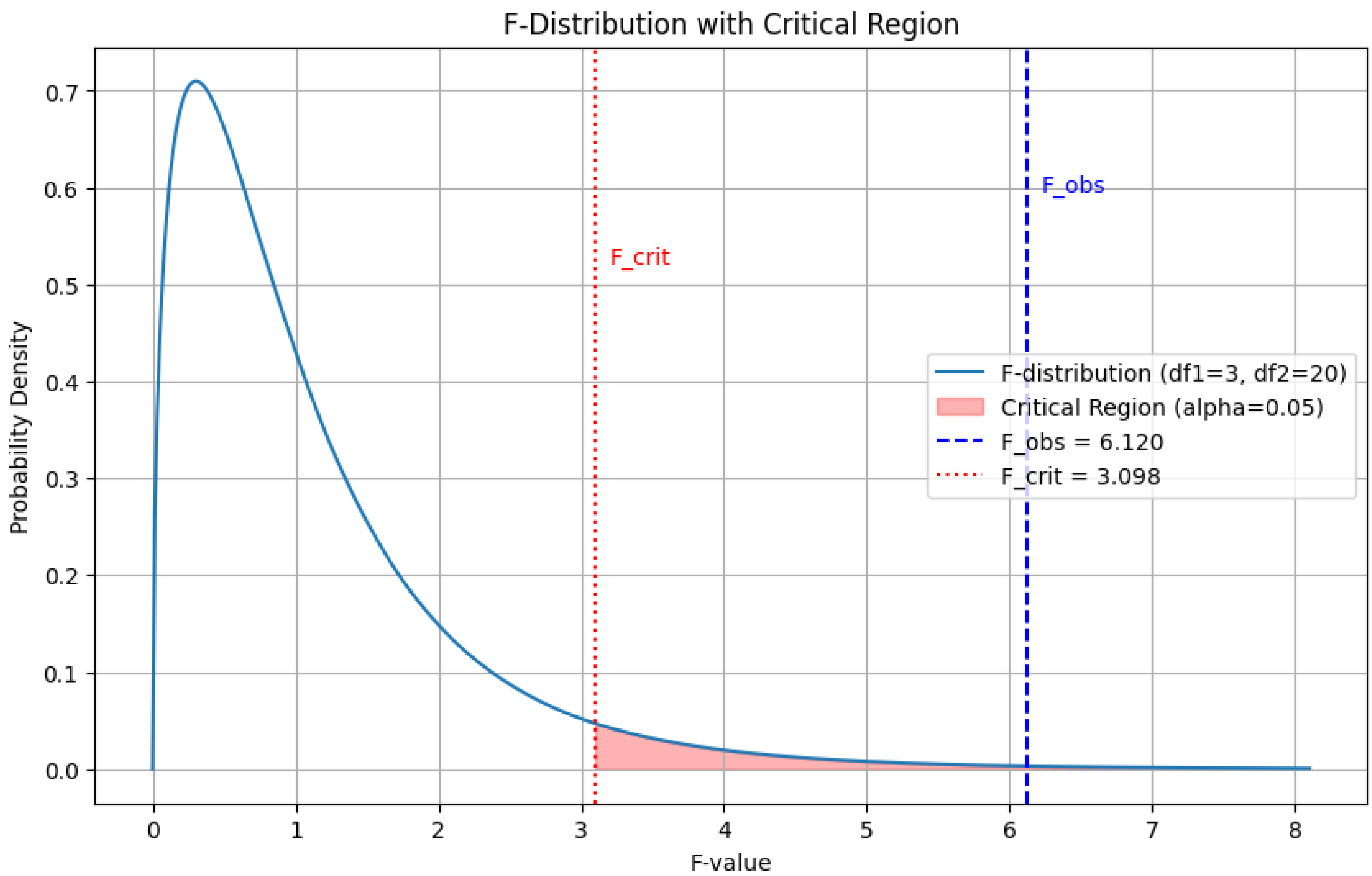
we reject H_0.

The ANOVA software output gives $p = 0.004$, consistent with the same conclusion.

Conclusion. There is statistically significant evidence that smoking level affects mean heart rate after exercise. At least one group's mean heart rate differs from the others.

ANOVA Table.

Source	df	SS	MS	F	p
Treatment	3	1464.125	488.04	6.12	0.004
Error	20	1594.833	79.74		
Total	23	3058.958			



https://drive.google.com/file/d/1WIkzNO7WPlrirmVNYuoxF2bnUKMsmRAT/view?usp=sharing

❖ Randomized Block Designs

The F Test for a Randomized Block Design