MATH 5010 –Foundations of Statistical Theory and Probability

❖ Analysis of Variance (ANOVA) 1

Instructor: He Wang
Department of Mathematics
Northeastern University

1. Review of Pooled t-test
2. One-way ANOVA Model Assumptions
3. Classic ANOVA Hypothesis
4. Linear Combination & Contrast
5.
6. One-way ANOVA table

Review: Two independent samples Pooled t-test

A hypothesis test based on the t −distribution for $\mu_1 - \mu_2$ when the (unknown) population variances σ_X^2 and σ_Y^2 are equal.

Two independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$,both assumed to come from normal distributions with the **same variance** σ^2.

Goal: Test

$$H_0: \mu_X = \mu_Y$$

Define the **pooled variance** estimator:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

The test statistic:

$$T_{n+m-2} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Distribution: $T_{n+m-2} \sim t_{n+m-2}$ Student t distribution with $n+m-2$ degree of freedom.

Under Null assumption $H_0: \mu_X = \mu_Y$, the test statistic:

$$t = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Right-tailed test $H_0 \ v.s. H_1: \mu_X > \mu_Y$

Reject H_0 if: $t \geq t_{\alpha,n+m-2}$

Left-tailed test $H_0 \ v.s. H_1: \mu_X < \mu_Y$

Reject H_0 if: $t \leq -t_{\alpha,n+m-2}$

Two-tailed test $H_0 \ v.s. H_1: \mu_X \neq \mu_Y$

Reject H_0 if: $|t| \geq t_{\alpha,n+m-2}$

One-way Analysis of Variance (ANOVA)

One-way ANOVA is a generalization of two independent samples Pooled t-test to k-**independent** random samples from normal distributions.

(Cell Means) Model assumption:

$$Y_{ij} = \theta_i + \epsilon_{ij}$$

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

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

θ_i :(treatment) mean of group i

ϵ_{ij}: error term (random noise).

Data:

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}

Sample size:

Sample totals:

Sample means:

True means:

Dot notations

Dot in a subscript means sum over that index.

Group (Treatment) Total $$T_{i.} = \sum_{j=1}^{n_i} Y_{ij}$$

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

Overall (Treatment) Total $$T_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^{k} T_{i.}$$

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

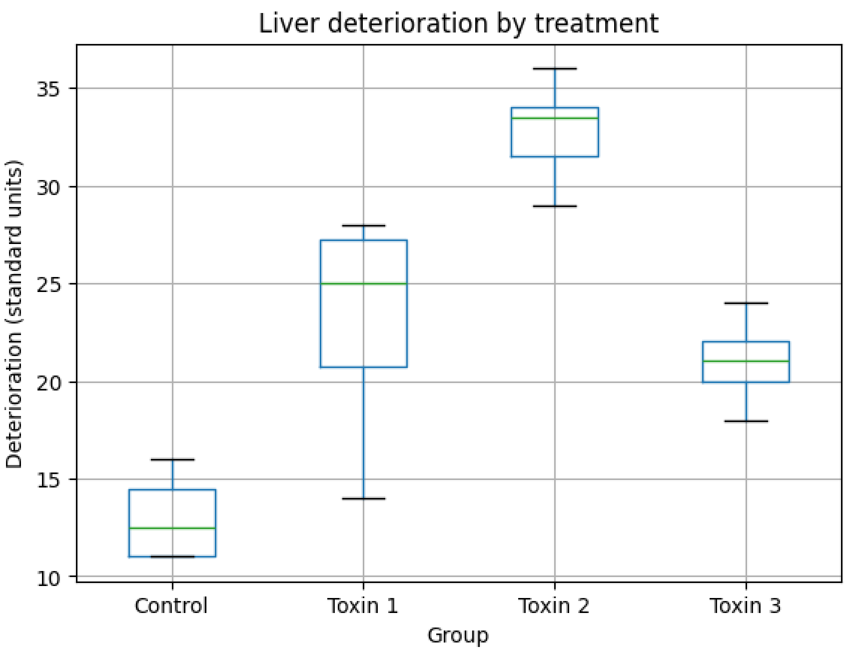
Example:

Toxin 1	Toxin 2	Toxin 3	Control
28.0	33.0	18.0	11.0
23.0	36.0	21.0	14.0
14.0	34.0	20.0	11.0
27.0	29.0	22.0	16.0
nan	31.0	24.0	nan
nan	34.0	nan	nan

Group Descriptives:

	Group	n	mean	var	sd
0	Control	4	13.000	6.000	2.449
1	Toxin 1	4	23.000	40.667	6.377
2	Toxin 2	6	32.833	6.167	2.483
3	Toxin 3	5	21.000	5.000	2.236

Boxplot


Liver deterioration by treatment
Deterioration (standard units)
35
30
25
20
15
10
Control
Toxin 1
Toxin 2
Toxin 3
Group

Overparameterized Model (optional)

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

μ: grand mean (common baseline across all treatments).
τ_i :treatment effect (deviation from the grand mean).

We can't uniquely estimate both μ and all τ_i's:

$\mu' = \mu + c,\ \ and\ \tau_i' = \tau_i - c$ give the same fit. (non-identifiable)

To fix identifiability, we impose a constraint, typically:

$$\sum_{i=1}^{k} \tau_i = 0$$

A parameter is **identifiable** if different values lead to different distributions.

One-way ANOVA Assumptions

Random variables Y_{ij} are observed according to the model

$$Y_{ij} = \theta_i + \epsilon_{ij}$$

Assume:

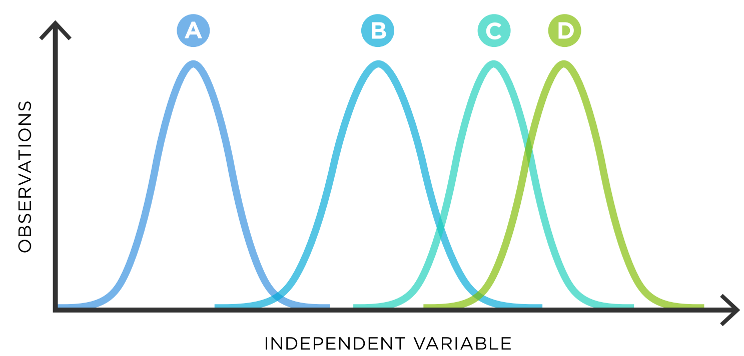
1. Errors have zero mean and finite variance $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma_i^2 < \infty$

2. Errors are independent and normally distributed

$$\epsilon_{ij} \sim Normal(0, \sigma_i^2)$$

3. Equal variances across groups

$$\sigma_i^2 = \sigma^2 \text{ for all } i = 1, \dots, k$$



Classic ANOVA Hypothesis

Null Hypothesis: All treatment means are exactly equal:

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

Alternative hypothesis:

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

Rejecting H_0 implies :

- We are **not** saying *all* means differ, only that *at least two* are different.
- We know **some means differ**, but not **which** ones.

The real interest of ANOVA is **not** in proving equality, but in **estimating and comparing** differences.

Example (ANOVA Hypothesis in Agriculture)

Effect of fertilizers on the zinc content of spinach plants.

Treatments: Mixtures of magnesium, potassium, and zinc (in pounds per acre).

Treatment	Magnesium	Potassium	Zinc
1	0	0	0
2	0	200	0
3	50	200	0
4	200	200	0
5	0	200	15

The **real scientific question**: *how much effect* each mixture has, and how they compare.

Linear Combination & Contrast

Given parameters/statistics $t = (t_1, \dots, t_k)$ and constants $a = (a_1, \dots, a_k)$

A **linear combination** of t_i is

$$\sum_{i=1}^{k} a_i t_i$$

The linear combination is called a **contrast**, if

$$\sum_{i=1}^{k} a_i = 0$$

Example: Compare one treatment v.s. control: $\theta_{Toxin} - \theta_{Control}$

Example : Average of several vs another: $\frac{1}{2}(\theta_{T1} + \theta_{T2}) - \theta_{Control}$

Contrasts allow meaningful **inference** beyond “some difference exists.”

➢ Union–Intersection View

Theorem: The ANOVA null $H_0: \theta_1 = \theta_2 = \cdots = \theta_k$ is **equivalent** to:
Every possible contrast must equal zero.

$$\sum_{i=1}^{k} a_i\theta_i = 0 \qquad \text{with} \quad \sum_{i=1}^{k} a_i = 0$$

Expressing ANOVA in terms of contrasts makes hypotheses:

- Easier to understand (direct comparisons between treatments).
- Easier to interpret (each contrast maps to a scientific question).

Inferences Regarding Linear Combinations of Means

Under the one-way ANOVA model:

$$Y_{ij} \sim Normal(\theta_i, \sigma^2)$$

Each **group sample mean**:

$$\overline{Y_{i.}} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \sim Normal\left(\theta_i, \frac{\sigma^2}{n_i}\right)$$

Consider linear combinations (**Normal distribution**):

$$\sum_{i=1}^{k} a_i \overline{Y_{i.}}$$

Expectation:

$$E\left[\sum_{i=1}^{k} a_i \overline{Y_{i.}}\right] = \sum_{i=1}^{k} a_i \theta_i$$

Variance:

$$Var\left[\sum_{i=1}^{k} a_i \overline{Y_{i.}}\right] = \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}$$

Standardized Test Statistic

$$Z = \frac{\sum_{i=1}^{k} a_i \overline{Y_{i.}} - \sum_{i=1}^{k} a_i \theta_i}{\sqrt{\sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \sim Normal(0,1)$$

In practice, since σ^2 is unknown, we replace it by the **pooled** ANOVA estimate,

$$S_p^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y_{i.}}\right)^2$$

S_p^2 pools the within-group variances.

$$\frac{(N-k)S_p^2}{\sigma^2} \sim \chi_{N-k}^2$$

Test Statistic for Contrasts

For a **contrast** defined by $L = \sum_{i=1}^{k} a_i \theta_i$

We estimate it by $\hat{L} = \sum_{i=1}^{k} a_i \overline{Y_i}$

The variance $Var(\hat{L}) \approx S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}$

Thus, the **t-statistic** is:

$$t = \frac{\sum_{i=1}^{k} a_i \overline{Y_i} \; - \sum_{i=1}^{k} a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \sim t_{N-k}$$

The distribution approximately a **t-distribution** with residual df= $N - k$.

Hypothesis Testing for General Linear Contrasts

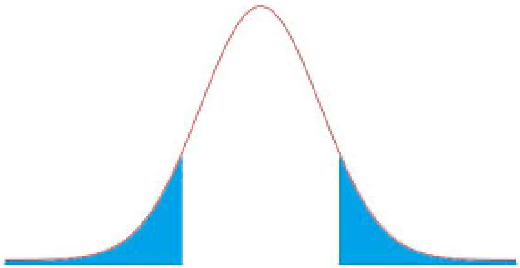
We want to test whether a linear combination of the group means equals zero.

To Test at level α

$$H_0: \sum_{i=1}^{k} a_i\theta_i = 0 \;\; v.s. \;\; H_1: \sum_{i=1}^{k} a_i\theta_i \neq 0$$

We would reject H_0 if

$$\left| \frac{\sum_{i=1}^{k} a_i \overline{Y}_i}{\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \right| > t_{N-k,\alpha/2}$$



Confidence Interval for the Contrast

From the pivot, a $100(1-\alpha)\%$ CI for the Contrast is:

$$\sum_{i=1}^{k} a_i \overline{Y}_i \pm t_{N-k,\alpha/2} \sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}$$

https://colab.research.google.com/drive/1zXX4E64YCg65zMvydgTTyToLgUt1_Rwq?usp=sharing

Example (**ANOVA Contrasts**)

Case 1: Compare two treatments directly (each toxin vs control.)

To compare treatment 1 v.s. 2, choose contrast vector $a = (1, -1, 0, \dots, 0)$, with contrast

$$\bar{Y}_1 - \bar{Y}_2$$

Test statistic:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Reject $H_0 : \theta_1 = \theta_2$ if

$$|t| > t_{N-k,\alpha/2}$$

This looks like a **two-sample t-test**, except that here the pooled variance S_p^2 uses *all groups*, not just the two being compared.

Case 2: Compare one treatment vs average of others (control vs average of all toxins**)**

Suppose treatment 1 is a **control**, and treatments 2 and 3 are experimental.

Contrast vector: $a = \left(1, -\frac{1}{2}, -\frac{1}{2}, 0, \ldots, 0\right)$ with contrast $\overline{Y_1} - \frac{1}{2}(\overline{Y_2} - \overline{Y_3})$

Test statistic:

$$t = \frac{\overline{Y_1} - \frac{1}{2}(\overline{Y_2} - \overline{Y_3})}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{4n_2} + \frac{1}{4n_3}\right)}}$$

Reject $H_0 : \theta_1 = \frac{1}{2}(\theta_2 + \theta_3)$ if

$$|t| > t_{N-k,\alpha/2}$$

Classic ANOVA Hypothesis

Null Hypothesis: All treatment means are exactly equal:

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

Alternative hypothesis:

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

In theory, we can test multiple pairs of means using multiple t-tests.

$$H_0: \theta_i = \theta_j \ v.s. \ H_1: \theta_i \neq \theta_j$$

However, it inflates the Type I error rate (the chance of a false positive).

Each t-test carries a probability of committing a **Type I error (reject|$\boldsymbol{H_0}$)**(e.g., 5%). When you run multiple t-tests on the same dataset, the chance of obtaining at least one false positive increases.

For example, with two independent t-tests, the probability of making **at least one** Type I error is

$$1 - (1 - 0.05)^2 = 1 - 0.952 \approx 9.75\%.$$

For three t-tests (as would happen when comparing three groups pairwise), the probability increases to

$$1 - (0.95)^3 \approx 14.3\%.$$

Next, we will have the single test, controlled error: ANOVA

➢ The ANOVA F-Test

Let $\mathcal{A} = \left\{a = (a_1, \dots, a_k) \colon \sum_{i=1}^{k} a_i = 0\right\}$ be the set of contrast vectors.

The ANOVA hypothesis test

$$H_0\colon \sum_{i=1}^{k} a_i\theta_i = 0 \text{ for } \boldsymbol{all}\ a \in \mathcal{A} \quad \text{v.s.} \quad H_1\colon \sum_{i=1}^{k} a_i\theta_i \neq 0 \text{ for } \boldsymbol{some}\ a \in \mathcal{A}$$

The ANOVA null is the **intersection** of all individual contrast nulls.

Define:

$$\Theta_a = \left\{\theta\colon \sum_{i=1}^{k} a_i\theta_i = 0\right\}$$

The ANOVA null H_0 is equivalent to $H_0\colon \theta \in \bigcap_{a \in \mathcal{A}} \Theta_a$

For each contrast vector a, we test

$$H_{0a}: \theta \in \Theta_a \;\; v.s. \; H_{1a}: \theta \notin \Theta_a$$

The **F-test** arises when combining these contrast tests into a single test.

$$T_a = \frac{\left| \sum_{i=1}^{k} a_i \bar{Y}_i \; - \sum_{i=1}^{k} a_i \theta_i \right|}{\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}}$$

To reject the global ANOVA null H_0 ,it suffices to reject for some contrast a.

Thus, the union–intersection test of the ANOVA null is to reject H_0 if the supremum

$$\sup_{a \in \mathcal{A}} T_a > c$$

where c denotes the critical constant such that $P\left(\sup_{a \in \mathcal{A}} T_a > c\right) = \alpha$.

One-way ANOVA - Partitioning Sums of Squares

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

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

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

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

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

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

Relations

Total sum of squares

$$SS_T = SS_B + SS_W = SS_{between} + SS_{Within}$$

In the **union–intersection contrasts derivation**, we maximized the test statistic

$$F = \frac{SS_B/(k-1)}{SS_W/(N-k)} = \frac{MS_B}{MS_W}$$

$MS_B \coloneqq SS_B/(k-1)$ Mean treatment sum of squares

$MS_W \coloneqq SS_W/(N-k)$ Mean error sum of squares

$MS_W \coloneqq \frac{SS_W}{N-k} = S_p^2$ pooled ANOVA estimate for σ^2.

Under the ANOVA assumptions, in particular if $Y_{ij} \sim N(\theta_i, \sigma^2)$

$$\frac{1}{\sigma^2} SS_W = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y_i}\right)^2 \sim \chi^2_{N-k}$$

If $\theta_i = \theta_j$ for all i, j, then

$$\frac{1}{\sigma^2} SS_B = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i (\overline{Y_{i.}} - \overline{Y_{..}})^2 \sim \chi^2_{k-1}$$

$$\frac{1}{\sigma^2} SS_W = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y_{i.}}\right)^2 \sim \chi^2_{N-1}$$

If $H_0: \theta_1 = \theta_2 = \cdots = \theta_k$ is true, **F-Statistic** $\frac{\frac{SS_B}{k-1}}{\frac{SS_W}{N-K}}$ has a F distribution with $k-1$ and $N-k$ degrees of freedom.

Thus, for an α level test of the ANOVA hypotheses

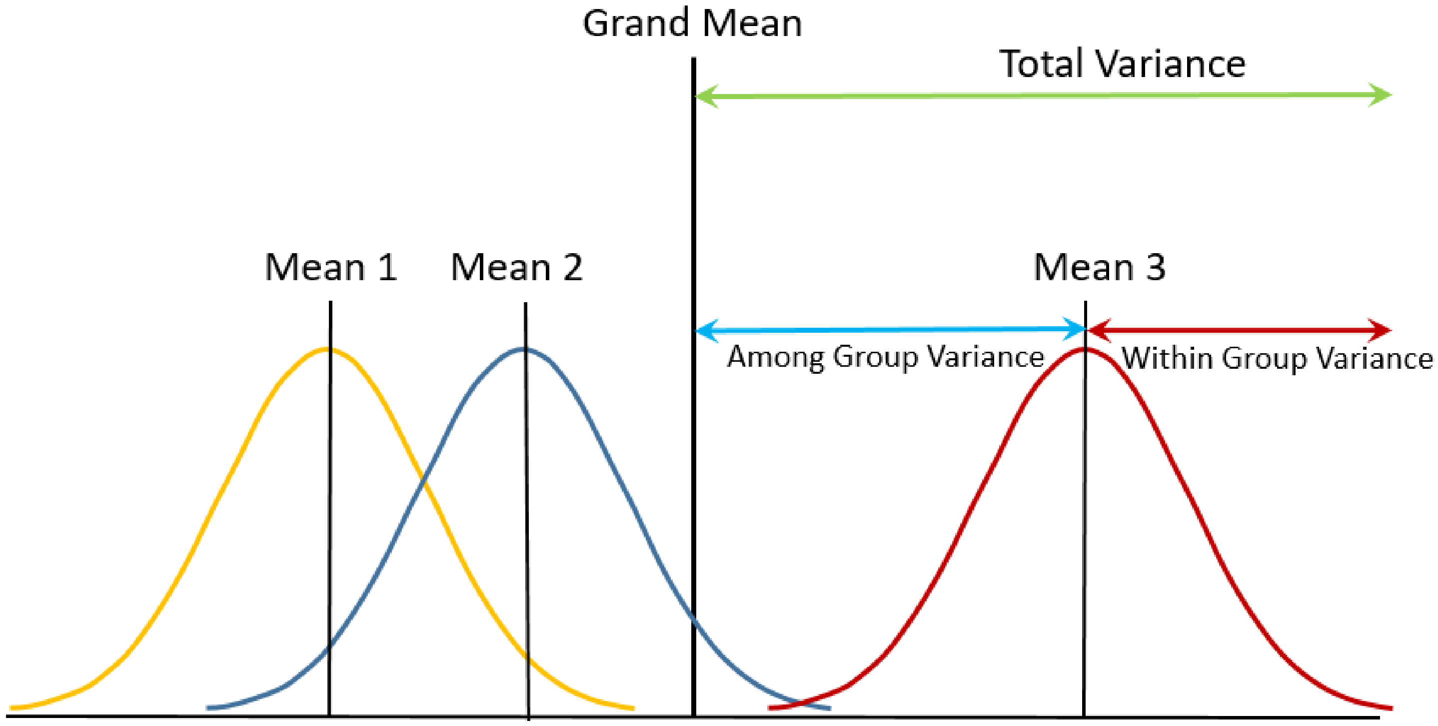
$H_0: \theta_1 = \theta_2 = \cdots = \theta_k$ versus $H_1: \theta_i \neq \theta_j$ for some i, j

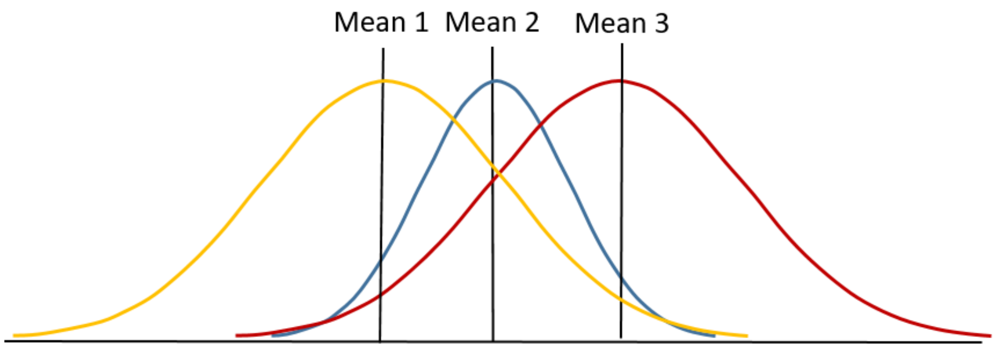
We reject H_0 if

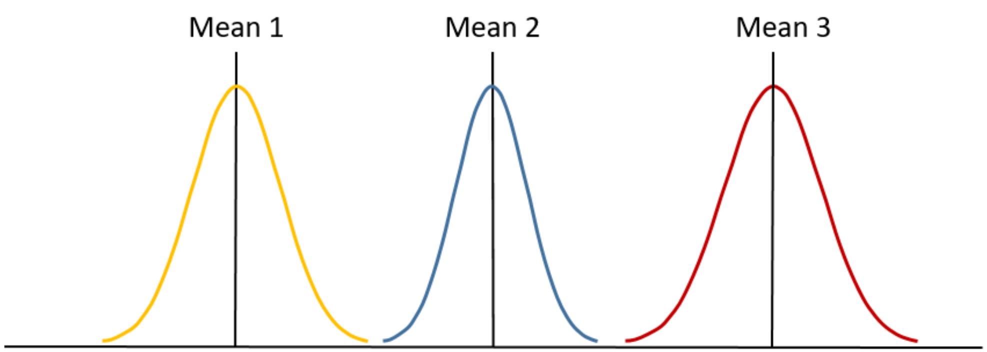
$$\frac{\frac{\sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}{(k-1)}}{\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2}{N-k}} > F_{k-1, N-k, \alpha}$$

One-way ANOVA table

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


Grand Mean
Total Variance
Mean 1
Mean 2
Mean 3
Among Group Variance
Within Group Variance


Mean 1
Mean 2
Mean 3


Mean 1
Mean 2
Mean 3

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**

Online books and courses:

- https://www.probabilitycourse.com/
- https://online.stat.psu.edu/stat415/
- https://stat110.hsites.harvard.edu/
- https://bookdown.org/egarpor/inference/

https://bookdown.org/mcbroom_j/Book/week-7-anova.html