Probability and Mathematical Statistics
503 Stat

Lecture 12

Analysis of Variance

Chapter Goals

After completing this chapter, you should be able to:

- Recognize situations in which to use analysis of variance
- Understand different analysis of variance designs
- Perform a one-way and two-way analysis of variance and interpret the results

One-Way Analysis of Variance

- Evaluate the difference among the means of three or more groups

Examples:	Average production for 1^{st}, 2^{nd}, and 3^{rd} shifts
	Expected mileage for five brands of tires

- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Samples are randomly and independently drawn

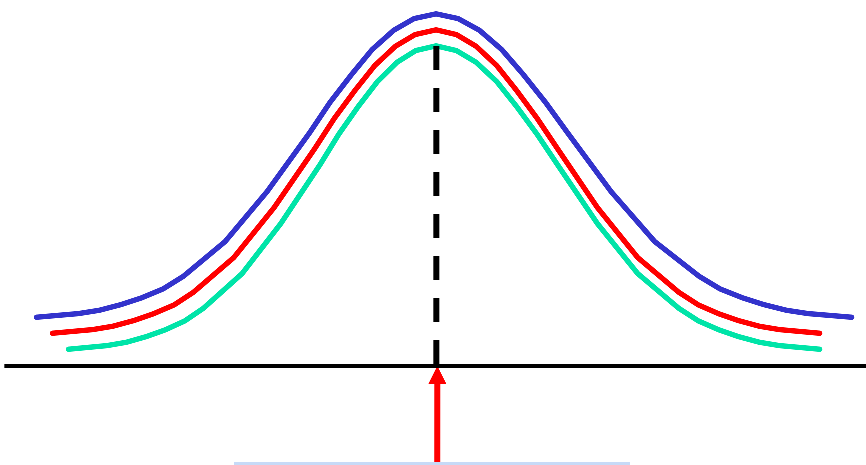
Hypotheses of One-Way ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_K$
  - All population means are equal
  - i.e., no variation in means between groups
- $H_1 : \mu_i \neq \mu_j \quad \text{for at least one } i, j \text{ pair}$
  - At least one population mean is different
  - i.e., there is variation between groups
  - Does not mean that all population means are different (some pairs may be the same)

One-Way ANOVA

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_K$$

$$H_1 : \text{Not all } \mu_i \text{ are the same}$$



$$\mu_1 = \mu_2 = \mu_3$$

All Means are the same:

The Null Hypothesis is True

(No variation between groups)

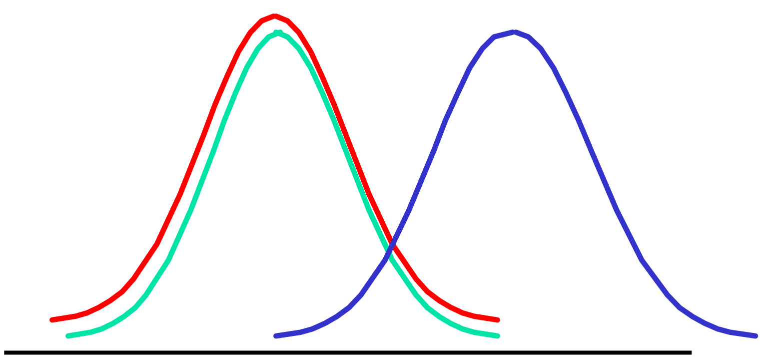
One-Way ANOVA

(continued)

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_K$$

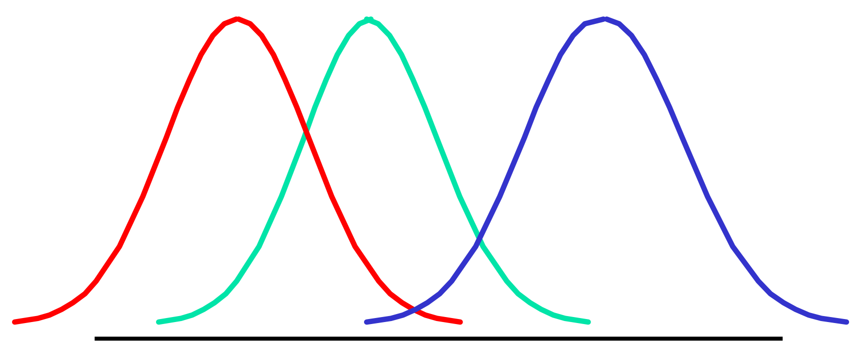
$$H_1: \text{Not all } \mu_i \text{ are the same}$$

At least one mean is different:
The Null Hypothesis is NOT true
(Variation is present between groups)



$$\mu_1 = \mu_2 \neq \mu_3$$

or

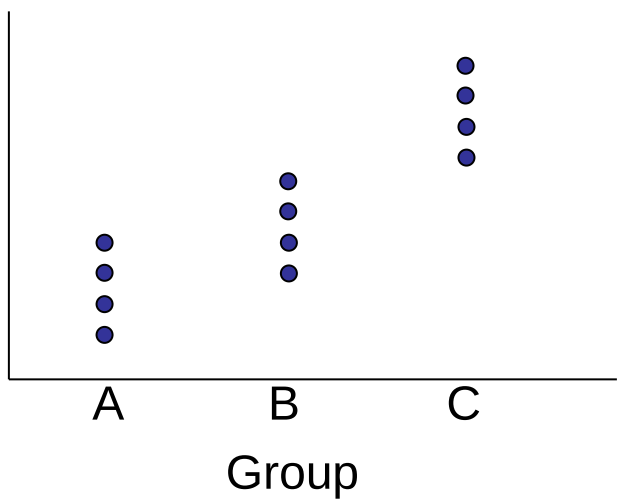


$$\mu_1 \neq \mu_2 \neq \mu_3$$

Variability

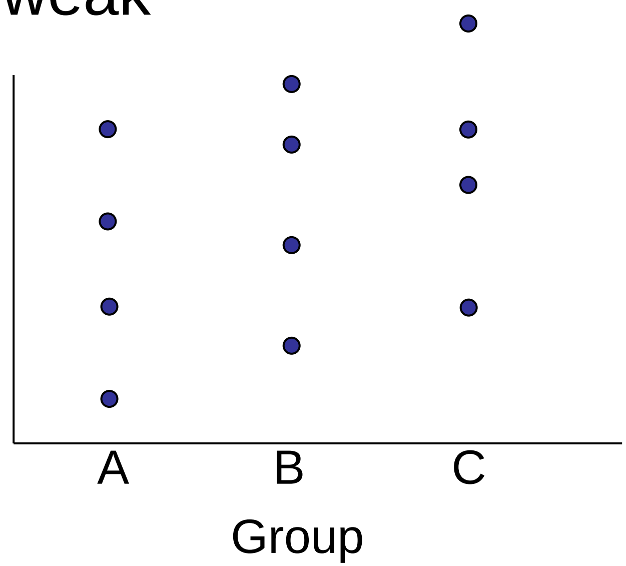
- The variability of the data is key factor to test the equality of means
- In each case below, the means may look different, but a large variation within groups in B makes the evidence that the means are different weak

A



Small variation within groups

B



Large variation within groups

Partitioning the Variation

- Total variation can be split into two parts:

$$\boxed{SST = SSW + SSG}$$

SST = Total Sum of Squares

Total Variation = the aggregate dispersion of the individual data values across the various groups

SSW = Sum of Squares Within Groups

Within-Group Variation = dispersion that exists among the data values within a particular group

SSG = Sum of Squares Between Groups

Between-Group Variation = dispersion between the group sample means

Partition of Total Variation

Total Sum of Squares (SST)

= **Variation due to random sampling (SSW)** + **Variation due to differences between groups (SSG)**

Total Sum of Squares

$$\boxed{SST} = SSW + SSG$$

$$\boxed{SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2}$$

Where:

SST = Total sum of squares

K = number of groups (levels or treatments)

n_i = number of observations in group i

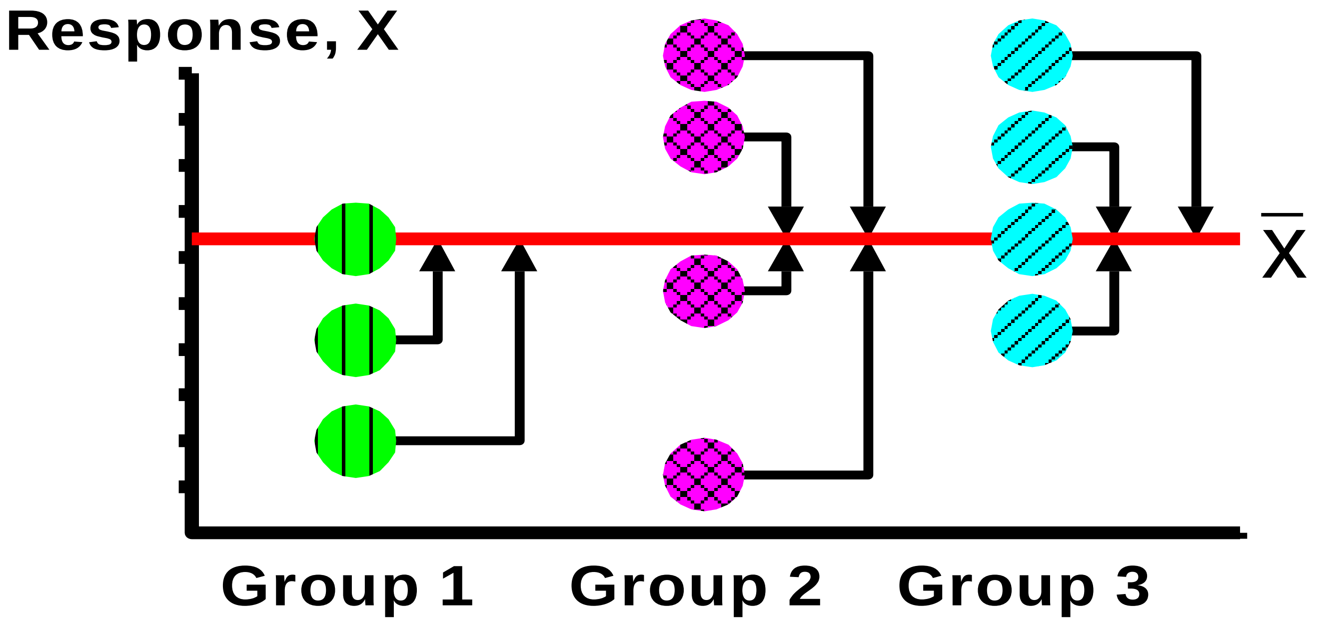
x_{ij} = j^{th} observation from group i

$\bar{x}$ = overall sample mean

Total Variation

(continued)

$$SST = (x_{11} - \bar{x})^2 + (X_{12} - \bar{x})^2 + \ldots + (x_{Kn_K} - \bar{x})^2$$



Within-Group Variation

$$SST = \boxed{SSW} + SSG$$

$$\boxed{SSW = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}$$

Where:

SSW = Sum of squares within groups

K = number of groups

n_i = sample size from group i

$\bar{x}_i$ = sample mean from group i

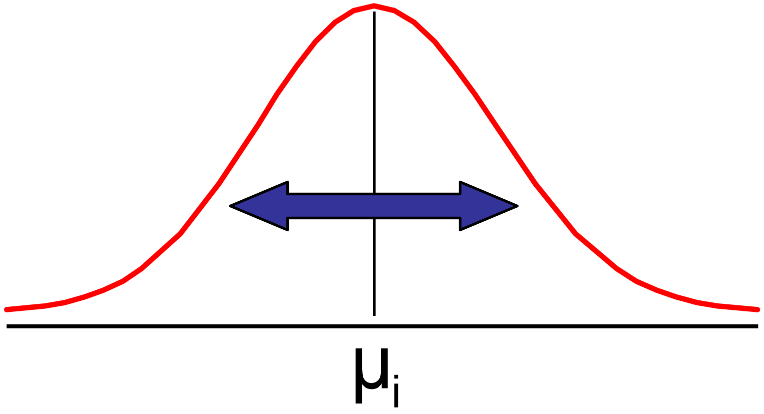
x_{ij} = j^{th} observation in group i

Within-Group Variation

(continued)

$$SSW = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Summing the variation within each group and then adding over all groups



$$MSW = \frac{SSW}{n - K}$$

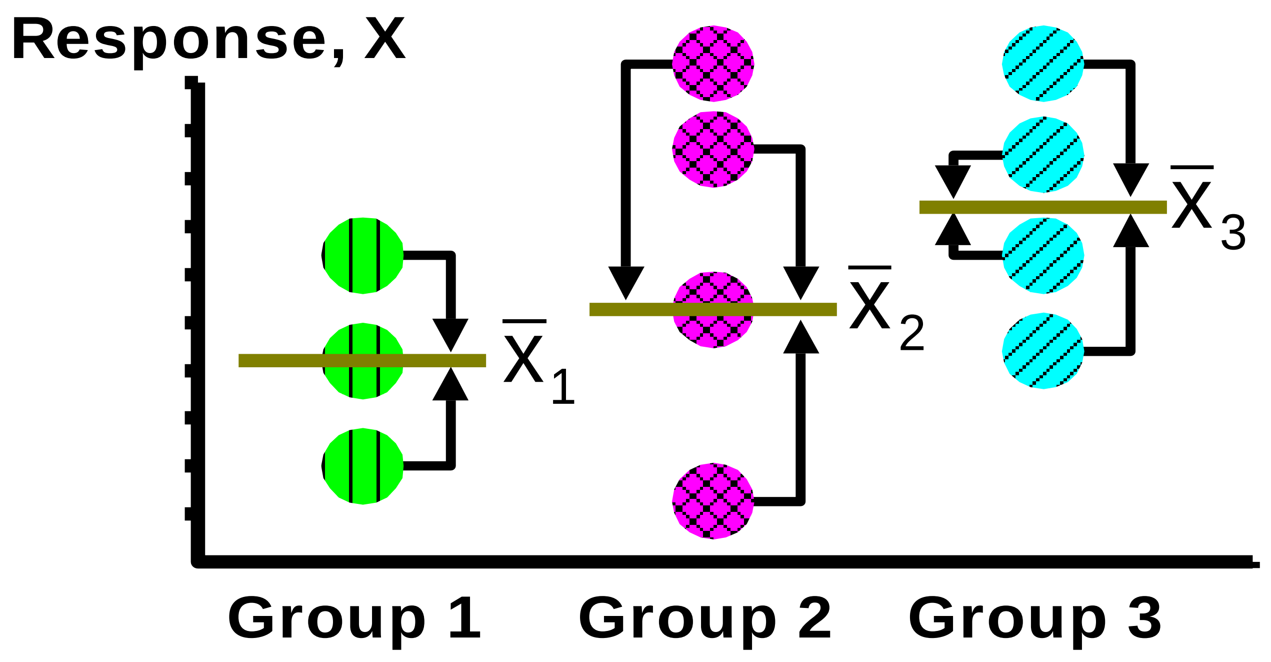
Mean Square Within =

SSW/degrees of freedom

Within-Group Variation

(continued)

$$SSW = (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_1)^2 + \ldots + (x_{Kn_K} - \bar{x}_K)^2$$



Between-Group Variation

$$SST = SSW + \boxed{SSG}$$

$$\boxed{SSG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2}$$

Where:

SSG = Sum of squares between groups

K = number of groups

n_i = sample size from group i

$\bar{x}_i$ = sample mean from group i

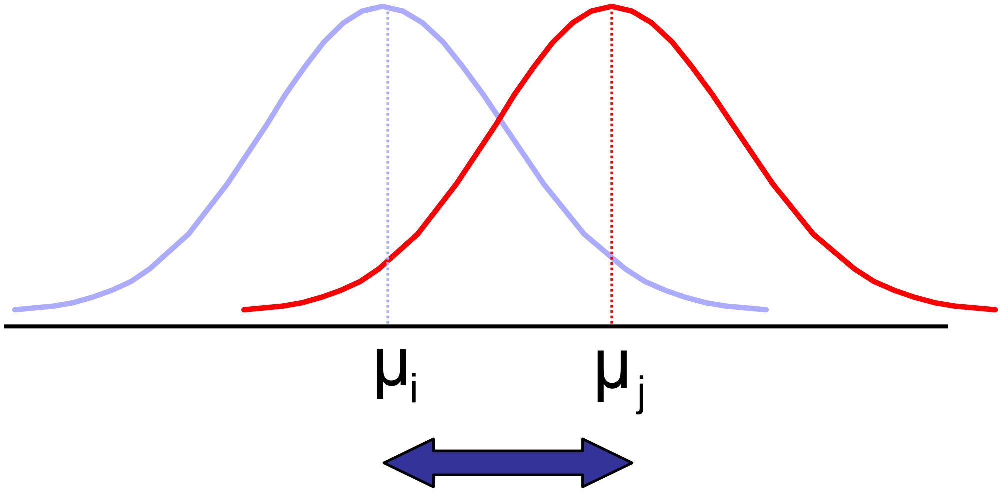
$\bar{x}$ = grand mean (mean of all data values)

Between-Group Variation

(continued)

$$SSG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2$$

Variation Due to
Differences Between Groups



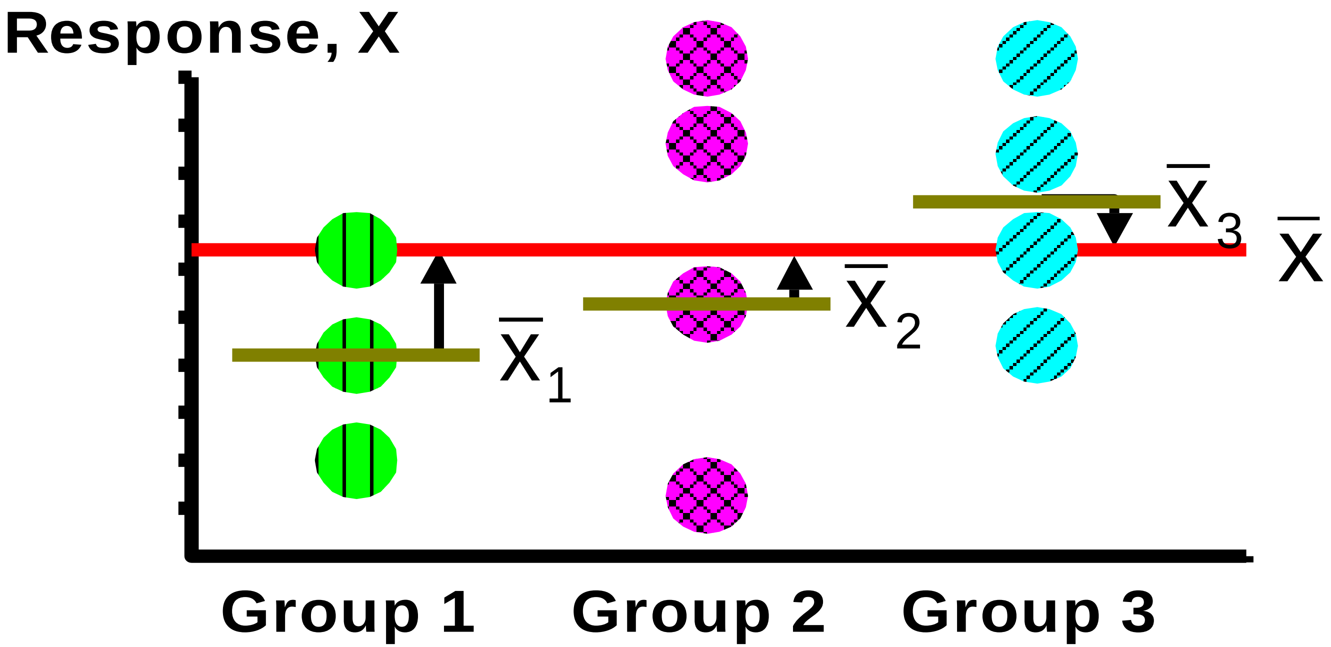
$$MSG = \frac{SSG}{K-1}$$

Mean Square Between Groups
= SSG/degrees of freedom

Between-Group Variation

(continued)

$$SSG = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_K(\bar{x}_K - \bar{x})^2$$



Obtaining the Mean Squares

$$\boxed{MST = \frac{SST}{n-1}}$$

$$\boxed{MSW = \frac{SSW}{n-K}}$$

$$\boxed{MSG = \frac{SSG}{K-1}}$$

One-Way ANOVA Table

Source of Variation	SS	df	MS (Variance)	F ratio
Between Groups	SSG	K - 1	$MSG = \frac{SSG}{K - 1}$	$F = \frac{MSG}{MSW}$
Within Groups	SSW	n - K	$MSW = \frac{SSW}{n - K}$	
Total	SST = SSG+SSW	n - 1		

K = number of groups

n = sum of the sample sizes from all groups

df = degrees of freedom

One-Factor ANOVA F Test Statistic

H_0: $\mu_1 = \mu_2 = \ldots = \mu_K$

H_1: At least two population means are different

- Test statistic

$$F = \frac{MSG}{MSW}$$

MSG is mean squares between variances

MSW is mean squares within variances

- Degrees of freedom
  - $df_1 = K - 1$ (K = number of groups)
  - $df_2 = n - K$ (n = sum of sample sizes from all groups)

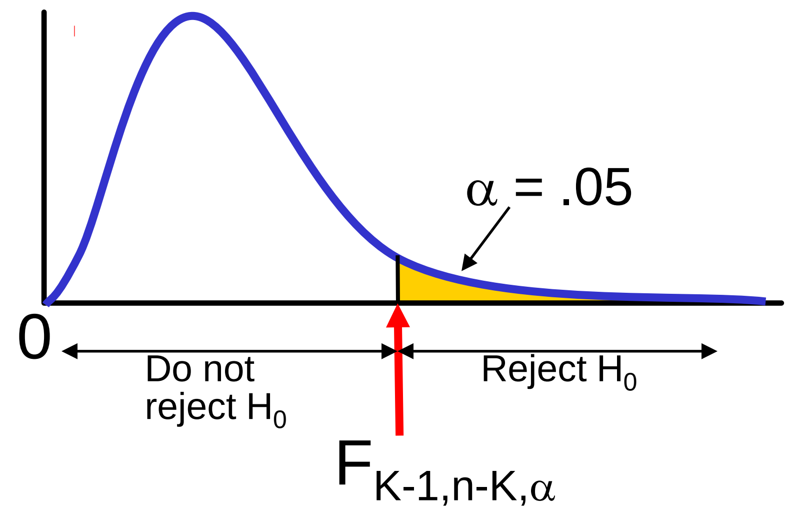
Interpreting the F Statistic

- The F statistic is the ratio of the between estimate of variance and the within estimate of variance
  - The ratio must always be positive
  - $df_1 = K - 1$ will typically be small
  - $df_2 = n - K$ will typically be large

Decision Rule:

- Reject H_0 if

  $F > F_{K-1,n-K,\alpha}$



One-Factor ANOVA F Test Example

You want to see if three different golf clubs yield different distances. You randomly select five measurements from trials on an automated driving machine for each club. At the .05 significance level, is there a difference in mean distance?

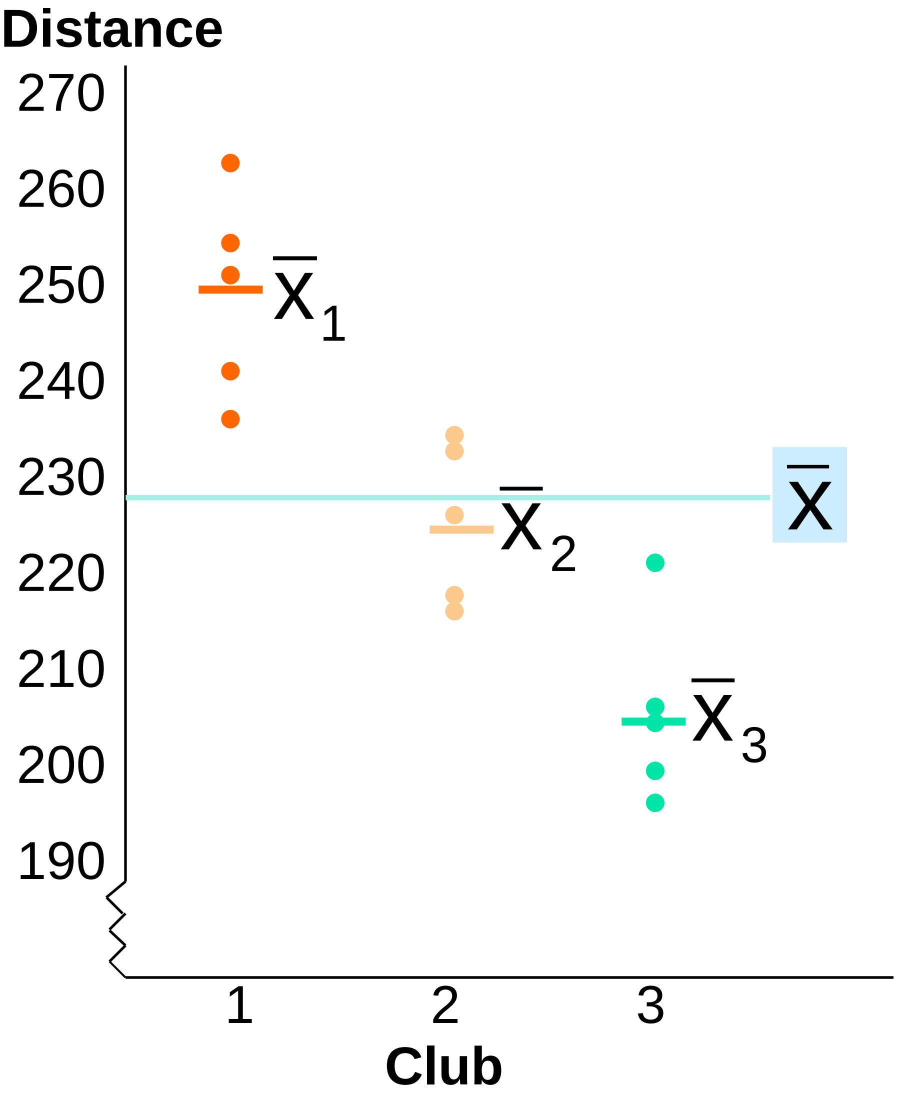
Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

One-Factor ANOVA Example: Scatter Diagram

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$	$\bar{x}_2 = 226.0$	$\bar{x}_3 = 205.8$

$\bar{\bar{x}} = 227.0$



One-Factor ANOVA Example Computations

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$ $\quad n_1 = 5$

$\bar{x}_2 = 226.0$ $\quad n_2 = 5$

$\bar{x}_3 = 205.8$ $\quad n_3 = 5$

$\bar{x} = 227.0$ $\quad n = 15$

$K = 3$

$SSG = 5\,(249.2 - 227)^2 + 5\,(226 - 227)^2 + 5\,(205.8 - 227)^2 = 4716.4$

$SSW = (254 - 249.2)^2 + (263 - 249.2)^2 + \ldots + (204 - 205.8)^2 = 1119.6$

$MSG = 4716.4 \,/\, (3\text{-}1) = 2358.2$

$MSW = 1119.6 \,/\, (15\text{-}3) = 93.3$

$$F = \frac{2358.2}{93.3} = 25.275$$

One-Factor ANOVA Example Solution

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: μ_i not all equal

$\alpha = .05$

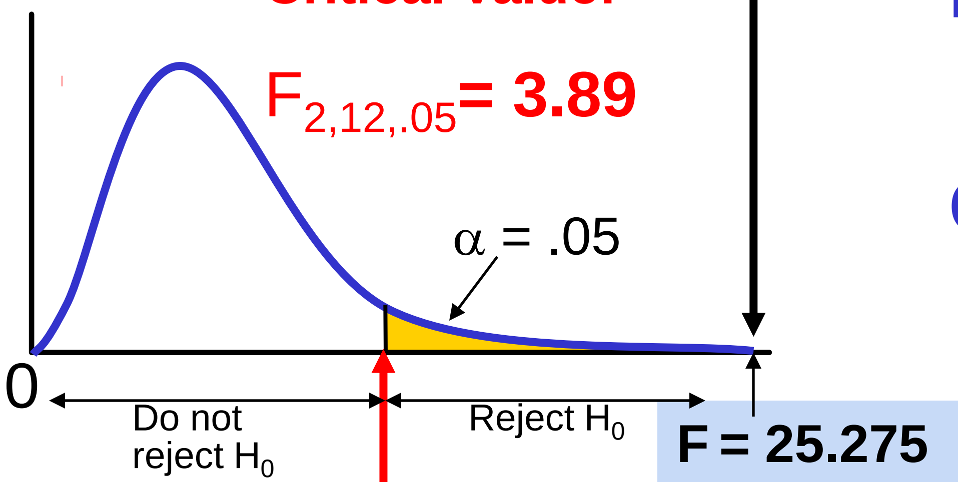
$df_1 = 2$ $df_2 = 12$

Test Statistic:

$$F = \frac{MSA}{MSW} = \frac{2358.2}{93.3} = 25.275$$

Critical Value:

$F_{2,12,.05}$ **= 3.89**



$F_{2,12,.05}$ **= 3.89**

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence that at least one μ_i differs from the rest

ANOVA -- Single Factor: Excel Output

EXCEL: data | data analysis | ANOVA: single factor

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Club 1	5	1246	249.2	108.2		
Club 2	5	1130	226	77.5		
Club 3	5	1029	205.8	94.2		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**4716.4**	**2**	**2358.2**	**25.275**	4.99E-05	**3.89**
Within Groups	**1119.6**	**12**	**93.3**			
Total	5836.0	14				

Two-Way Analysis of Variance

- Examines the effect of
  - Two factors of interest on the dependent variable
    - e.g., Percent carbonation and line speed on soft drink bottling process
  - Interaction between the different levels of these two factors
    - e.g., Does the effect of one particular carbonation level depend on which level the line speed is set?

Two-Way ANOVA

(continued)

- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Independent random samples are drawn

Randomized Block Design

Two Factors of interest: A and B

K = number of groups of factor A

H = number of levels of factor B

(sometimes called a blocking variable)

Block	Group			
	1	2	…	K
1	x_{11}	x_{21}	…	x_{K1}
2	x_{12}	x_{22}	…	x_{K2}
.	.	.	.	.
.	.	.	.	.
H	x_{1H}	x_{2H}	…	x_{KH}

Two-Way Notation

- Let x_{ji} denote the observation in the j^{th} group and i^{th} block
- Suppose that there are K groups and H blocks, for a total of $n = KH$ observations
- Let the overall mean be $\bar{x}$
- Denote the group sample means by

$$\boxed{\bar{x}_{j\bullet} \quad (j = 1, 2, \ldots, K)}$$

- Denote the block sample means by

$$\boxed{\bar{x}_{\bullet i} \quad (i = 1, 2, \ldots, H)}$$

Partition of Total Variation

- SST = SSG + SSB + SSE

Total Sum of Squares (SST)	=	**Variation due to differences between groups (SSG)**
		+
		Variation due to differences between blocks (SSB)
		+
The error terms are assumed to be independent, normally distributed, and have the same variance	→	**Variation due to random sampling (unexplained error) (SSE)**

Two-Way Sums of Squares

- The sums of squares are

Degrees of Freedom:

Total: $SST = \sum_{j=1}^{K}\sum_{i=1}^{H}(x_{ji} - \bar{x})^2$ — $n - 1$

Between - Groups: $SSG = H\sum_{j=1}^{K}(\bar{x}_{j\bullet} - \bar{x})^2$ — $K - 1$

Between - Blocks: $SSB = K\sum_{i=1}^{H}(x_{\bullet i} - \bar{x})^2$ — $H - 1$

Error: $SSE = \sum_{j=1}^{K}\sum_{i=1}^{H}(x_{ji} - \bar{x}_{j\bullet} - \bar{x}_{\bullet i} + \bar{x})^2$ — $(K - 1)(K - 1)$

Two-Way Mean Squares

- The mean squares are

$$\mathrm{MST} = \frac{\mathrm{SST}}{n-1}$$

$$\mathrm{MSG} = \frac{\mathrm{SST}}{K-1}$$

$$\mathrm{MSB} = \frac{\mathrm{SST}}{H-1}$$

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{(K-1)(H-1)}$$

Two-Way ANOVA: The F Test Statistic

F Test for Groups

H_0: The K population group means are all the same

$$F = \frac{MSG}{MSE}$$

Reject H_0 if

$F > F_{K-1,(K-1)(H-1),\alpha}$

F Test for Blocks

H_0: The H population block means are the same

$$F = \frac{MSB}{MSE}$$

Reject H_0 if

$F > F_{H-1,(K-1)(H-1),\alpha}$

General Two-Way Table Format

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Ratio
Between groups	SSG	K – 1	$MSG = \frac{SSG}{K-1}$	$\frac{MSG}{MSE}$
Between blocks	SSB	H – 1	$MSB = \frac{SSB}{H-1}$	$\frac{MSB}{MSE}$
Error	SSE	(K – 1)(H – 1)	$MSE = \frac{SSE}{(K-1)(H-1)}$	
Total	SST	n - 1		

Example

Example:
Four methods of manufacturing penicillin were compare in a randomized block design. The block are blends of the raw material, corn steep liquor, known to be quite variable. The yield of each method for five blends is given below.

	blend (block)				
Method	1	2	3	4	5
A	89	84	81	87	79
B	88	77	87	92	81
C	97	92	87	89	80
D	94	79	85	84	88

At level $\alpha = 0.01$,

1. are there significant differences between the blends (blocks)?
2. are there significant differences between the Methods (treatments)?

Example

Method	blend (block) 1	2	3	4	5	Method mean
A	89	84	81	87	79	$\bar{y}_{1.} = 84$
B	88	77	87	92	81	$\bar{y}_{2.} = 85$
C	97	92	87	89	80	$\bar{y}_{3.} = 89$
D	94	79	85	84	88	$\bar{y}_{4.} = 86$
blend mean	$\bar{y}_{.1} = 92$	$\bar{y}_{.2} = 83$	$\bar{y}_{.3} = 85$	$\bar{y}_{.4} = 88$	$\bar{y}_{.5} = 82$	$\bar{y}_{..} = 86$

We have $k = 4$, $b = 5$, $\bar{y}_{..} = \frac{1}{4 \times 5} \sum_{i=1}^{4} \sum_{j=1}^{5} y_{ij} = \frac{89+84+81+87+79+\ldots+94+79+85+84+88}{20} = \frac{1720}{20}$.

$SSA = b \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2 = 5\left[(84-86)^2+(85-86)^2+(89-86)^2+(86-86)^2\right] = 5\left[4+1+9+0\right] = 70.$

$SSB = k \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 = 4\left[(92-86)^2 + (83-86)^2 + (85-86)^2 + (88-86)^2 + (82-86)^2\right] =$
$4\left[36 + 9 + 1 + 4 + 16\right] = 264.$

$SST = \sum_{i=1}^{k} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{4} \sum_{j=1}^{5} (y_{ij} - \bar{y}_{..})^2 =$
$(89-86)^2 + (84-86)^2 + \ldots + (84-86)^2 + (88-86)^2 = 9 + 4 + \ldots + 4 + 4 = 560$
$SSE = SST - SSA - SSB = 560 - 70 - 264 = 226$

Example

ANOVA of CRD table

(Source)	(SS)	(df)	(MS)	F
Methods (A)	70	3	23.333	1.239
blends (B)	264	4	66.000	3.504
Error	226	12	18.833	
Total	560	19		

1. Test the blend (blocks) effects

$$H_0 : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.5}$$

H_1 :at least one blend mean is different.

$F = 3.504 > F_{0.05,4,12} = 3.26$, then we reject H_0, which means there do appear to be significant differences between blends.

2. Test the method effects

$$H_0 : \mu_{1.} = \mu_{2.} = \ldots = \mu_{.4}$$

H_1 :at least one method mean is different.

$F = 1.239 < F_{0.05,3,12} = 3.49$, then we fail to reject H_0, which means there is no significant differences between methods.

Chapter Summary

- Described one-way analysis of variance
  - The logic of Analysis of Variance
  - Analysis of Variance assumptions
  - F test for difference in K means
- Described two-way analysis of variance
  - Examined effects of multiple factors
  - Examined interaction between factors