Analiza wariancji i metody klasyfikacyjne

Analysis of variance and classification methods

lecture 3

21 October 2019

Ilona Anna Urbaniak (PK)

Marcin Wolter (IFJ PAN)

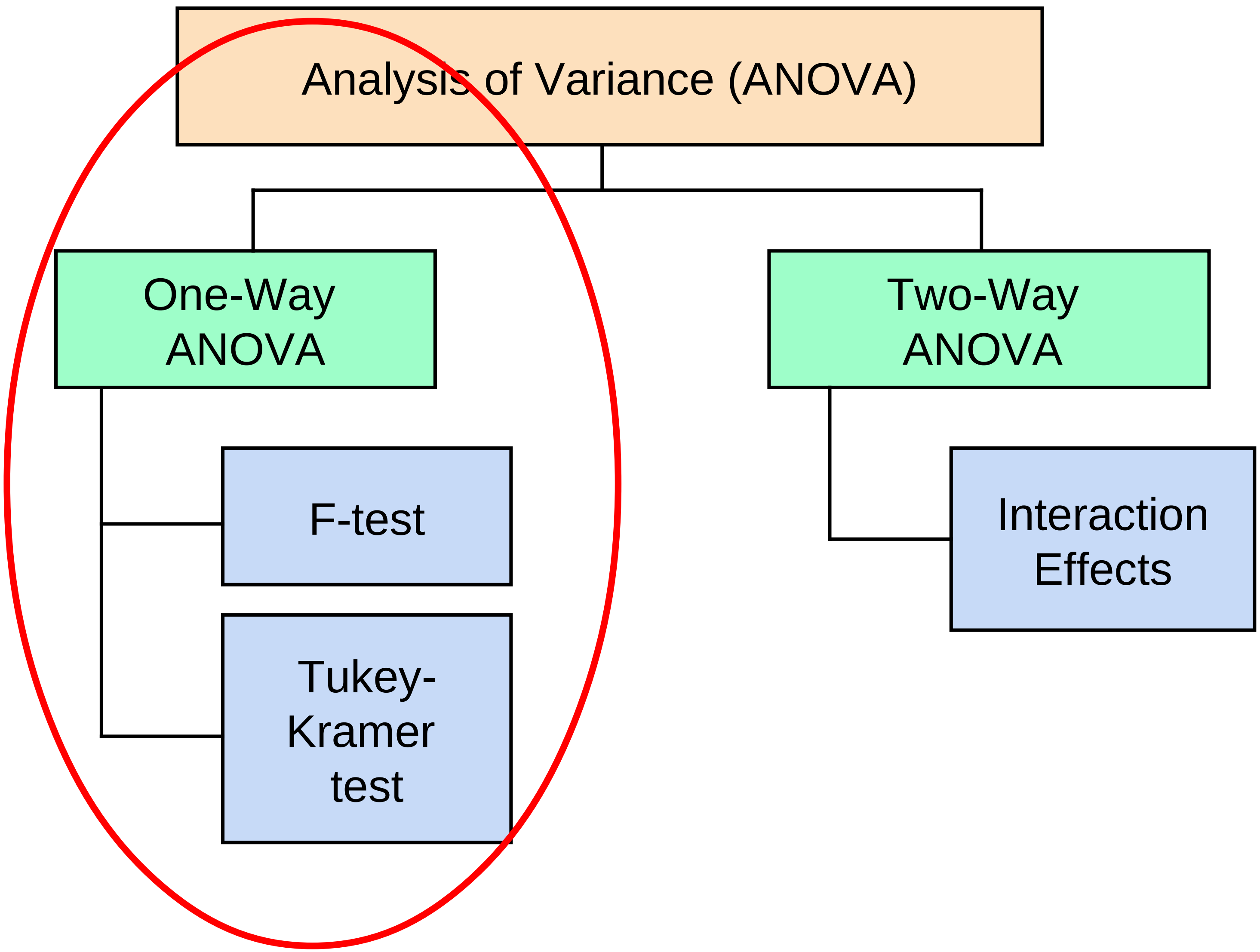
e-mail: marcin.wolter@ifj.edu.pl, phone: 12 662 8024

Slides: https://indico.ifj.edu.pl/event/271/

Analysis of Variance ANOVA

- Ronald Fisher introduced the term variance and proposed its formal analysis in 1918
- Used first in agriculture
- Set of experiments, in which we vary ONE (or more) variables, which take discrete values:
  - Example: we grow the same plant with different types of fertilizers (1 parameter)
  - We may vary the amount of water
- Question: **does the different fertilizer (amount of water) has a significant impact on the plant growth?**

Chapter Overview



General ANOVA Setting

- Investigator controls one or more independent variables
  - Called factors (or treatment variables)
  - Each factor contains two or more levels (or groups or categories/classifications)
- Observe effects on the dependent variable
  - Response to levels of independent variable
- Experimental design: the plan used to collect the data

Completely Randomized Design

- Experimental units (subjects) are assigned randomly to treatments
  - Subjects are assumed homogeneous
- Only one factor or independent variable
  - With two or more treatment levels
- Analyzed by one-factor analysis of variance (**one-way ANOVA**)

One-Way Analysis of Variance

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

Examples: Accident rates for youngsters, middle-aged, seniors
Expected mileage for five brands of tires

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

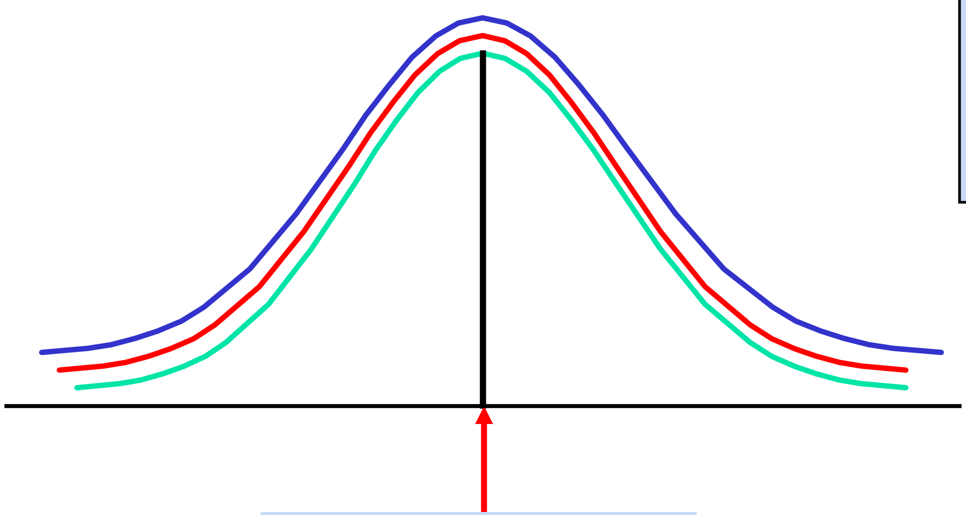
Hypotheses of One-Way ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$
  - All population means are equal
  - i.e., no treatment effect (no variation in means among groups)
- $H_1 :$ Not all of the population means are the same
  - At least one population mean is different
  - i.e., there is a treatment effect
  - Does not mean that all population means are different (some pairs may be the same)

One-Factor ANOVA

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

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



All Means are the same:

The Null Hypothesis is True

(No Treatment Effect)

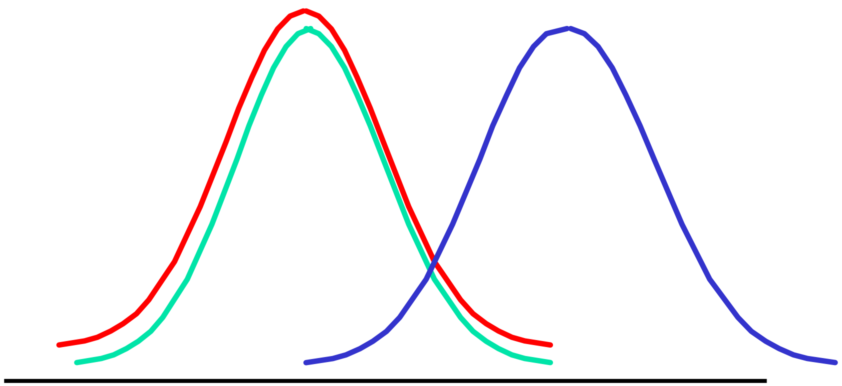
$$\mu_1 = \mu_2 = \mu_3$$

One-Factor ANOVA

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

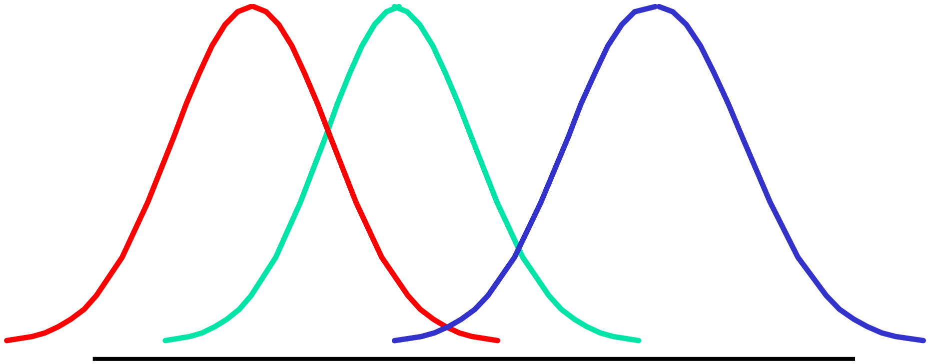
$$H_1: \text{Not all } \mu_i \text{ are the same}$$

At least one mean is different:
The Null Hypothesis is NOT true
(Treatment Effect is present)



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

or



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

Partitioning the Variation

- Total variation can be split into two parts:

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

SST = Total Sum of Squares
(Total variation)
SSA = Sum of Squares Among Groups
(Among-group variation)
SSW = Sum of Squares Within Groups
(Within-group variation)

Partitioning the Variation

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

Total Variation = the aggregate dispersion of the individual data values across the various factor levels (SST)

Among-Group Variation = dispersion between the factor sample means (SSA)

Within-Group Variation = dispersion that exists among the data values within a particular factor level (SSW)

Partition of Total Variation

Total Variation (SST)

= **Variation Due to Factor (SSA)** + **Variation Due to Random Sampling (SSW)**

Commonly referred to as:
- Sum of Squares Between
- Sum of Squares Among
- Sum of Squares Explained
- Among Groups Variation

Commonly referred to as:
- Sum of Squares Within
- Sum of Squares Error
- Sum of Squares Unexplained
- Within Groups Variation

Total Sum of Squares

$$SST = SSA + SSW$$

$$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \bar{\bar{X}})^2$$

Where:

SST = Total sum of squares

c = number of groups (levels or treatments)

n_j = number of observations in group j

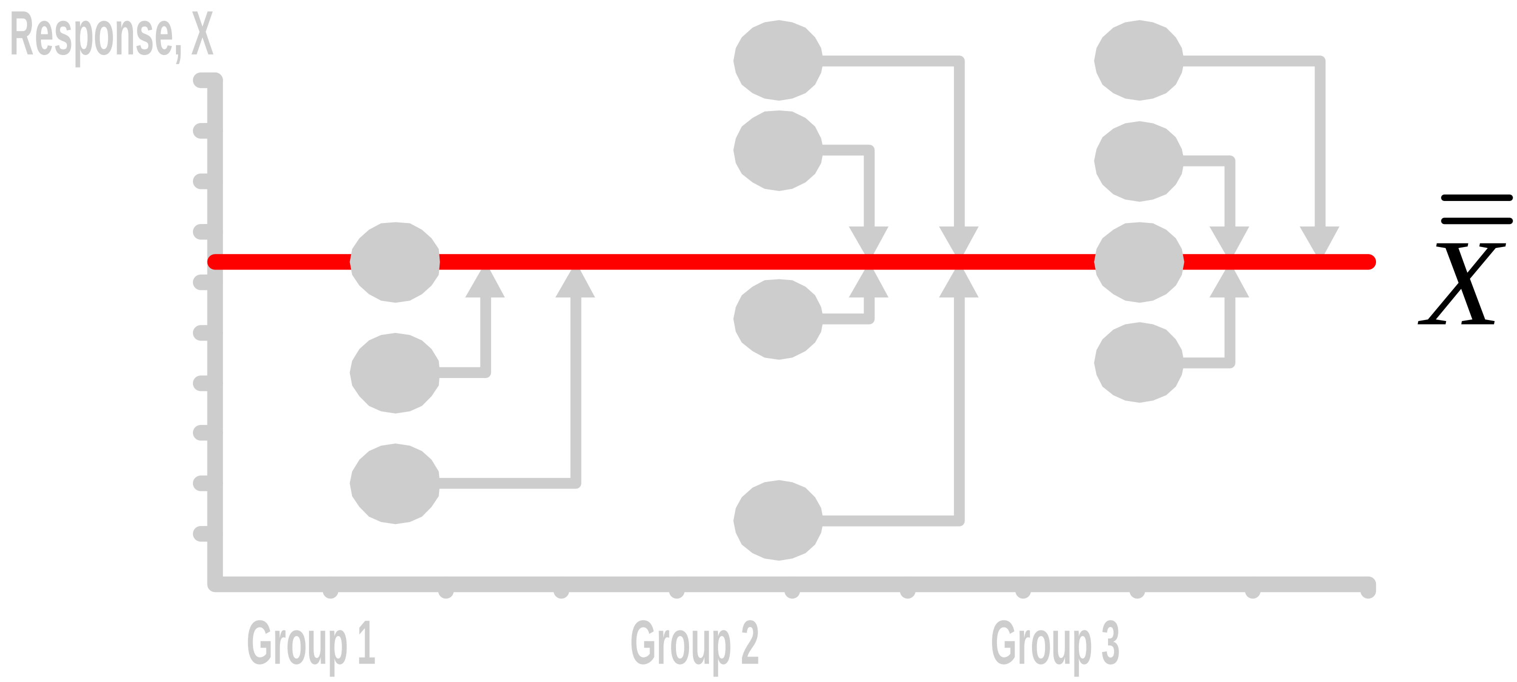
X_{ij} = i^{th} observation from group j

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

Total Variation

(continued)

$$SST = (X_{11} - \bar{\bar{X}})^2 + (X_{12} - \bar{\bar{X}})^2 + \ldots + (X_{cn_c} - \bar{\bar{X}})^2$$



Among-Group Variation

$$SST = SSA + SSW$$

$$SSA = \sum_{j=1}^{c} n_j(\overline{X}_j - \overline{\overline{X}})^2$$

Where:

SSA = Sum of squares among groups

c = number of groups or populations

n_j = sample size from group j

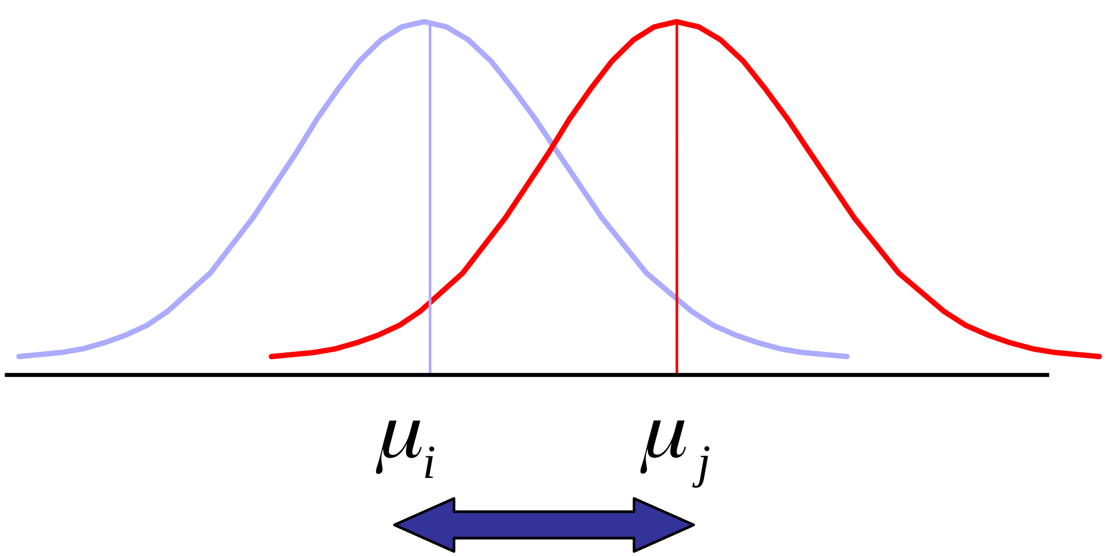
$\overline{X}_j$ = sample mean from group j

$\overline{\overline{X}}$ = grand mean (mean of all data values)

Among-Group Variation

$$SSA = \sum_{j=1}^{c} n_j (\overline{X}_j - \overline{\overline{X}})^2$$

Variation Due to
Differences Among Groups

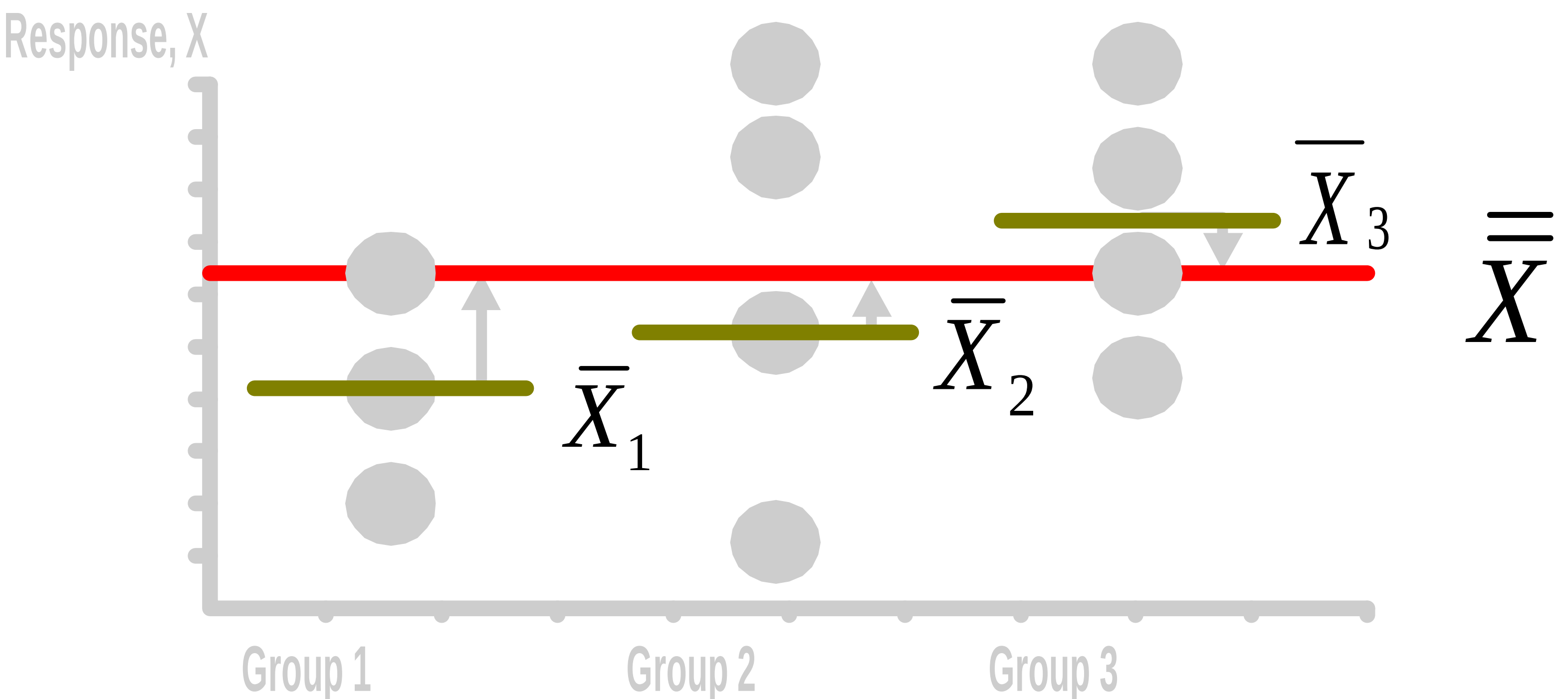


$$MSA = \frac{SSA}{c-1}$$

Mean Square Among =
SSA/degrees of freedom
c – number of groups

Among-Group Variation

$$SSA = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \ldots + n_c(\bar{x}_c - \bar{\bar{x}})^2$$



Within-Group Variation

$$SST = SSA + SSW$$

$$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$$

Where:

SSW = Sum of squares within groups

c = number of groups

n_j = sample size from group j

$\overline{X}_j$ = sample mean from group j

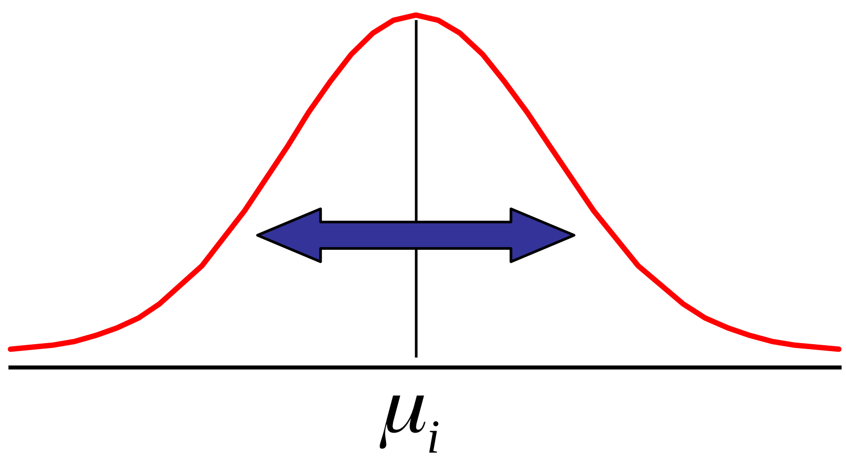
X_{ij} = i^{th} observation in group j

Within-Group Variation

(continued)

$$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$$

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



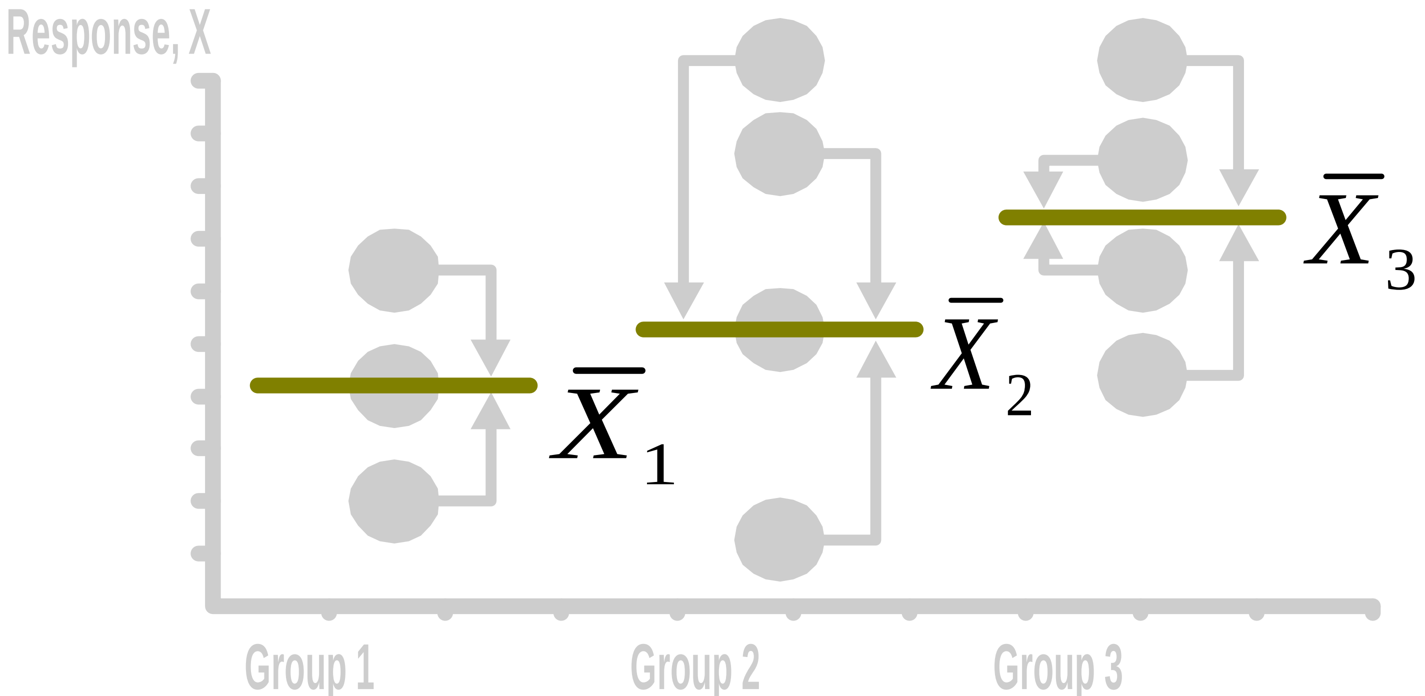
$$MSW = \frac{SSW}{n - c}$$

Mean Square Within = SSW/degrees of freedom

Within-Group Variation

(continued)

$$SSW = (x_{11} - \overline{X}_1)^2 + (X_{12} - \overline{X}_2)^2 + \ldots + (X_{cn_c} - \overline{X}_c)^2$$



Obtaining the Mean Squares

$$MSA = \frac{SSA}{c - 1}$$

Mean Square Among

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

Mean Square Within

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

Mean Square Total

One-Way ANOVA Table

Source of Variation	SS	df	MS (Variance)	F ratio
Among Groups	SSA	c - 1	$MSA = \dfrac{SSA}{c - 1}$	$F = \dfrac{MSA}{MSW}$
Within Groups	SSW	n - c	$MSW = \dfrac{SSW}{n - c}$	
Total	SST = SSA+SSW	n - 1		

c = number of groups
n = sum of the sample sizes from all groups
df = degrees of freedom

One-Factor ANOVA F Test Statistic

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

H_1: At least two population means are different

- Test statistic

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

MSA is mean squares among variances

MSW is mean squares within variances

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

F-test

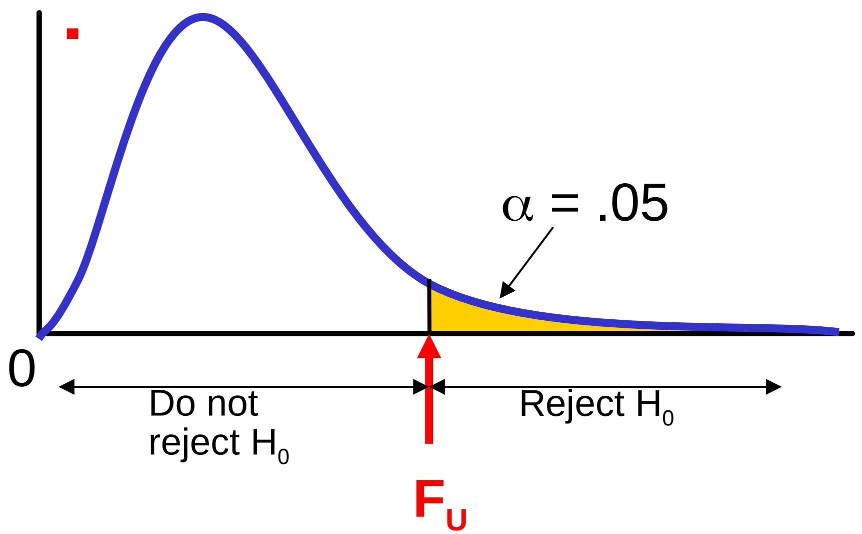
- An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis.
- It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.
- The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher.

Interpreting One-Factor ANOVA F Statistic

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

Decision Rule:

- Reject H_0 if $F > F_U$, otherwise do not reject H_0



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?

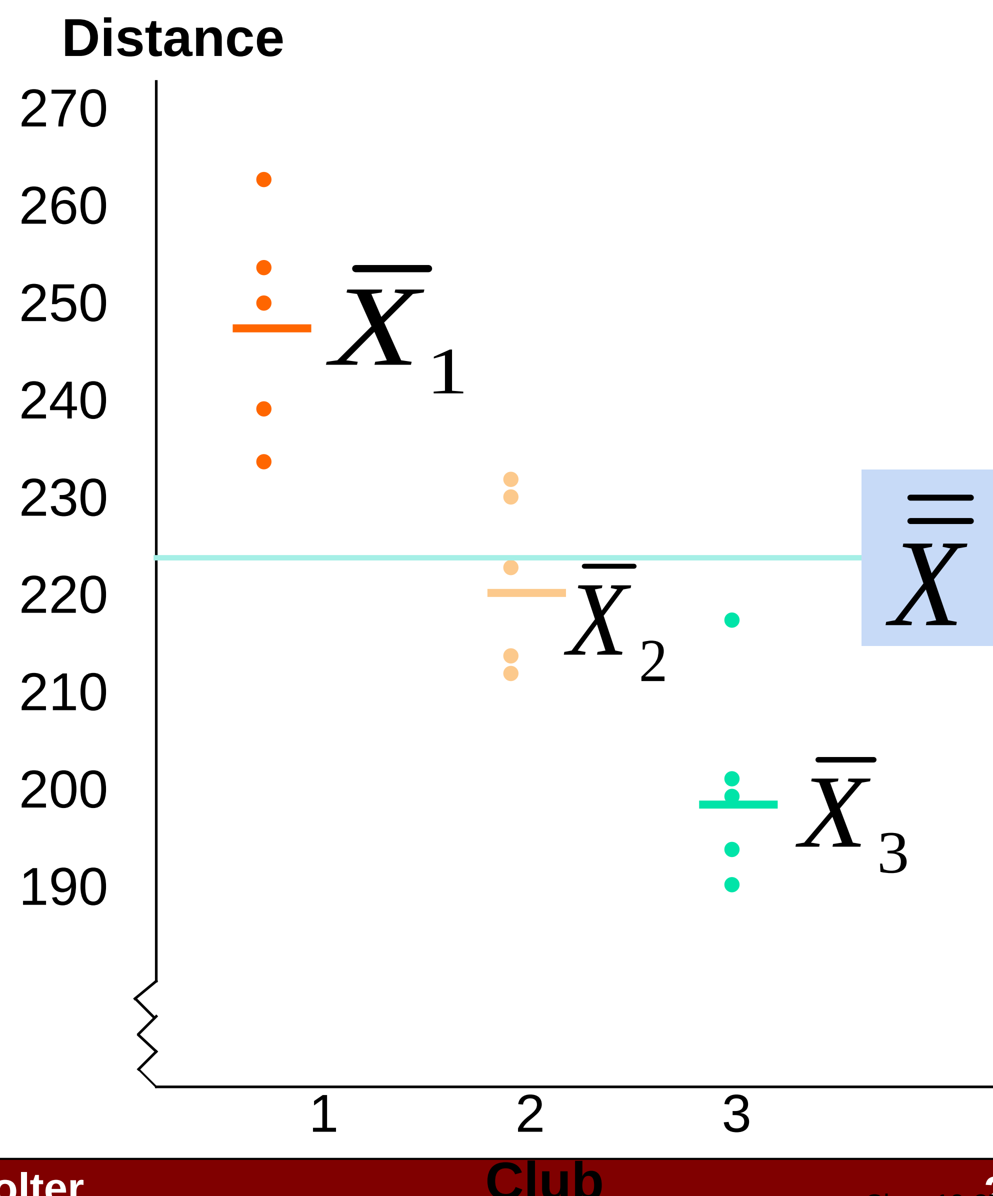
Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

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

$\bar{X}_1 = 249.2$ $n_1 = 5$

$\bar{X}_2 = 226.0$ $n_2 = 5$

$\bar{X}_3 = 205.8$ $n_3 = 5$

$\bar{\bar{X}} = 227.0$ $n = 15$

$c = 3$

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

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

$MSA = 4716.4 / (3-1) = 2358.2$

$MSW = 1119.6 / (15-3) = 93.3$

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

One-Factor ANOVA Example Solution

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

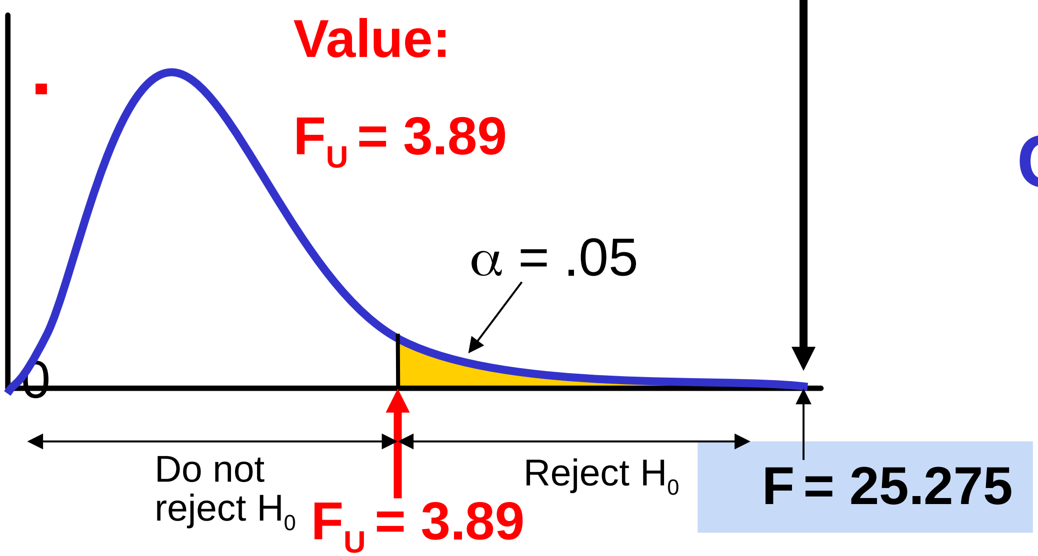
H_1: μ_i not all equal

$\alpha = .05$

$df_1 = 2$ $df_2 = 12$

Critical Value:

$F_U = 3.89$



Test Statistic:

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

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

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

F-distribution (Fisher–Snedecor)

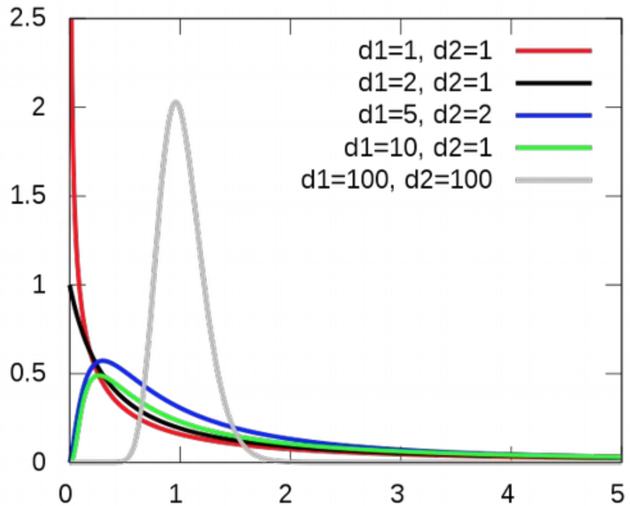
- If a random variable X has an F-distribution with parameters d_1 and d_2, we write $X \sim F(d_1, d_2)$. Then the probability density function for X is given by:

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1\, x)^{d_1}\, d_2^{d_2}}{(d_1\, x + d_2)^{d_1+d_2}}}}{x\, \mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

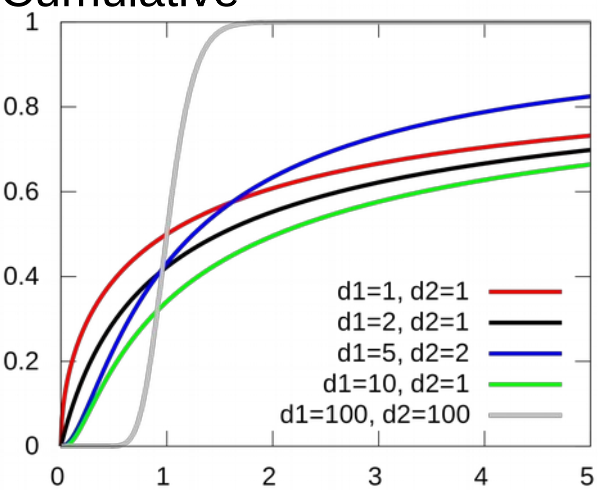
where B is a beta function:

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$$

Probability density function



Cumulative



F-test function:
https://www.stat.purdue.edu/~jtroisi/STAT350Spring2015/tables/FTable.pdf

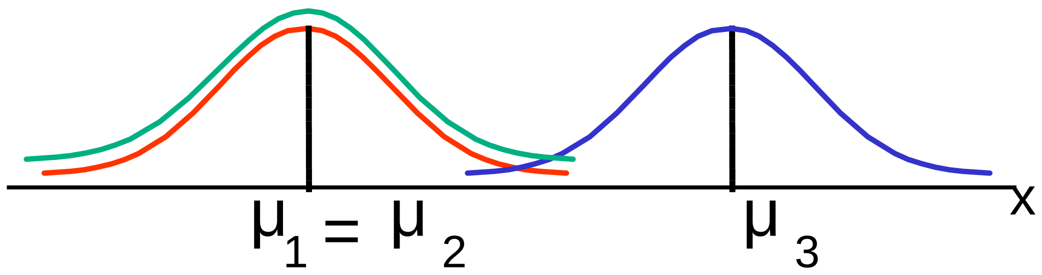
ANOVA -- Single Factor: Excel Output

EXCEL: tools | 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				

The Tukey-Kramer Procedure

- Tells which population means are significantly different
  - e.g.: $\mu_1 = \mu_2 \neq \mu_3$
  - Done after rejection of equal means in ANOVA
- Allows pair-wise comparisons
  - Compare absolute mean differences with critical range



The Tukey-Kramer Procedure

- Compare the difference between means divided by the standard error of the sum of the means SE

$$q_s = \frac{|\overline{X_1} - \overline{X_2}|}{SE}$$

to a q value from the studentized range distribution. If the q_s value is larger than the critical value q obtained from the distribution, the two means are said to be significantly different at level α , 0 ≤ α ≤ 1 .

$$Critical\ Range = Q_U \sqrt{\frac{MSW}{2}\left(\frac{1}{n_{j'}} + \frac{1}{n_j}\right)}$$

Tukey-Kramer Critical Range

$$\text{Critical Range} = Q_U \sqrt{\frac{\text{MSW}}{2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$$

where:

Q_U = Value from Studentized Range Distribution with c and n - c degrees of freedom for the desired level of α

MSW = Mean Square Within

n_j and $n_{j'}$ = Sample sizes from groups j and j'

Studentized Range Distribution

- Tukey's HSD makes use of the studentized range distribution *q*, which describes the expected, normalized difference between the max and min observed means amongst *k* treatments, under the null hypothesis:

$$q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MS_W}{n}}}$$

where

$\bar{X}_i$ = largest mean

$\bar{X}_j$ = smallest mean

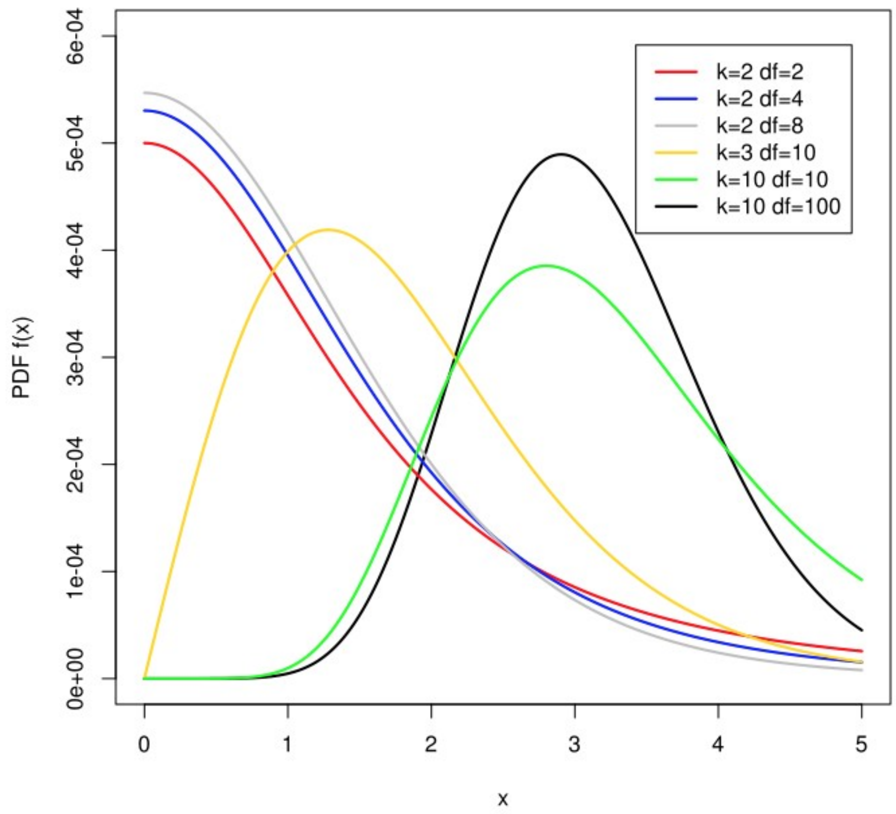
$$f_R(q;k,\nu) = \frac{\sqrt{2\pi}\,k\,(k-1)\,\nu^{\nu/2}}{\Gamma(\nu/2)\,2^{(\nu/2-1)}} \int_0^\infty s^\nu\,\varphi(\sqrt{\nu}\,s) \left[\int_{-\infty}^{\infty} \varphi(z+q\,s)\,\varphi(z)\,[\Phi(z+q\,s)-\Phi(z)]^{k-2}\,\mathrm{d}z\right] \mathrm{d}s$$

Where:

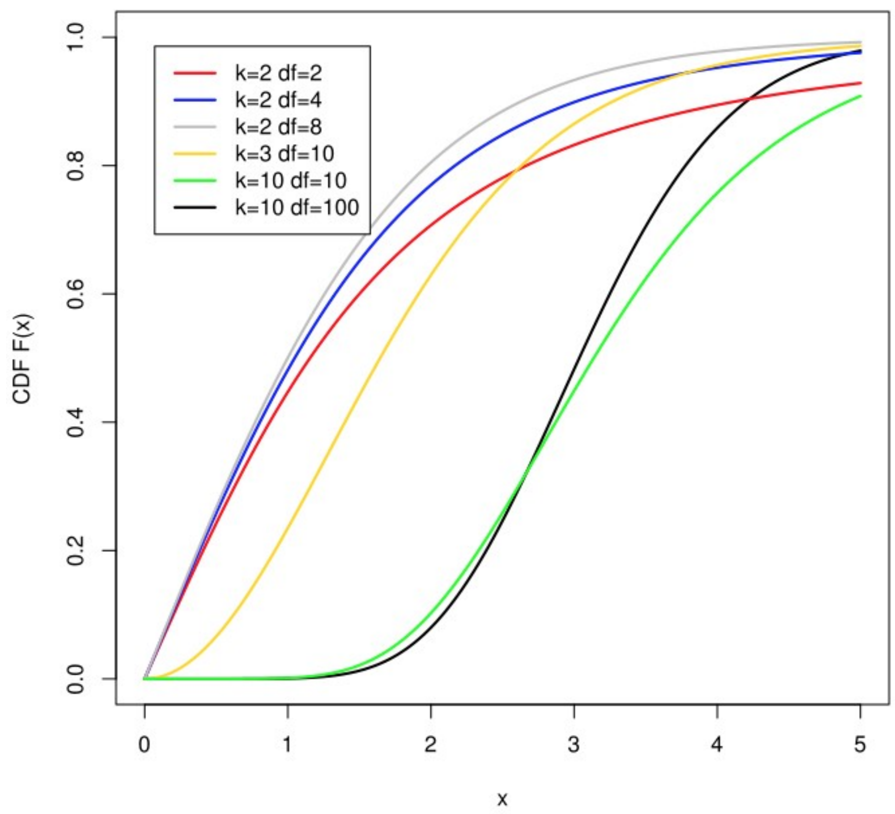
$$\varphi(\sqrt{\nu}\,s)\,\sqrt{2\pi} = e^{-(\nu\,s^2/2)}$$

Studentized Range Distribution

Probability density function



Cumulative distribution



The Tukey-Kramer Procedure: Example

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

1. Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |249.2 - 226.0| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = |249.2 - 205.8| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = |226.0 - 205.8| = 20.2$$

2. Find the Q_U value from the table with $c = 3$ and $(n – c) = (15 – 3) = 12$ degrees of freedom for the desired level of α ($\alpha = .05$ used here):

$$Q_U = 3.773$$

See table:
https://www.stat.purdue.edu/~xbw/courses/stat512/q-table.pdf

The Tukey-Kramer Procedure: Example

(continued)

3. Compute Critical Range:

$$\text{Critical Range} = Q_U \sqrt{\frac{MSW}{2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)} = 3.77\sqrt{\frac{93.3}{2}\left(\frac{1}{5} + \frac{1}{5}\right)} = 16.285$$

4. Compare:

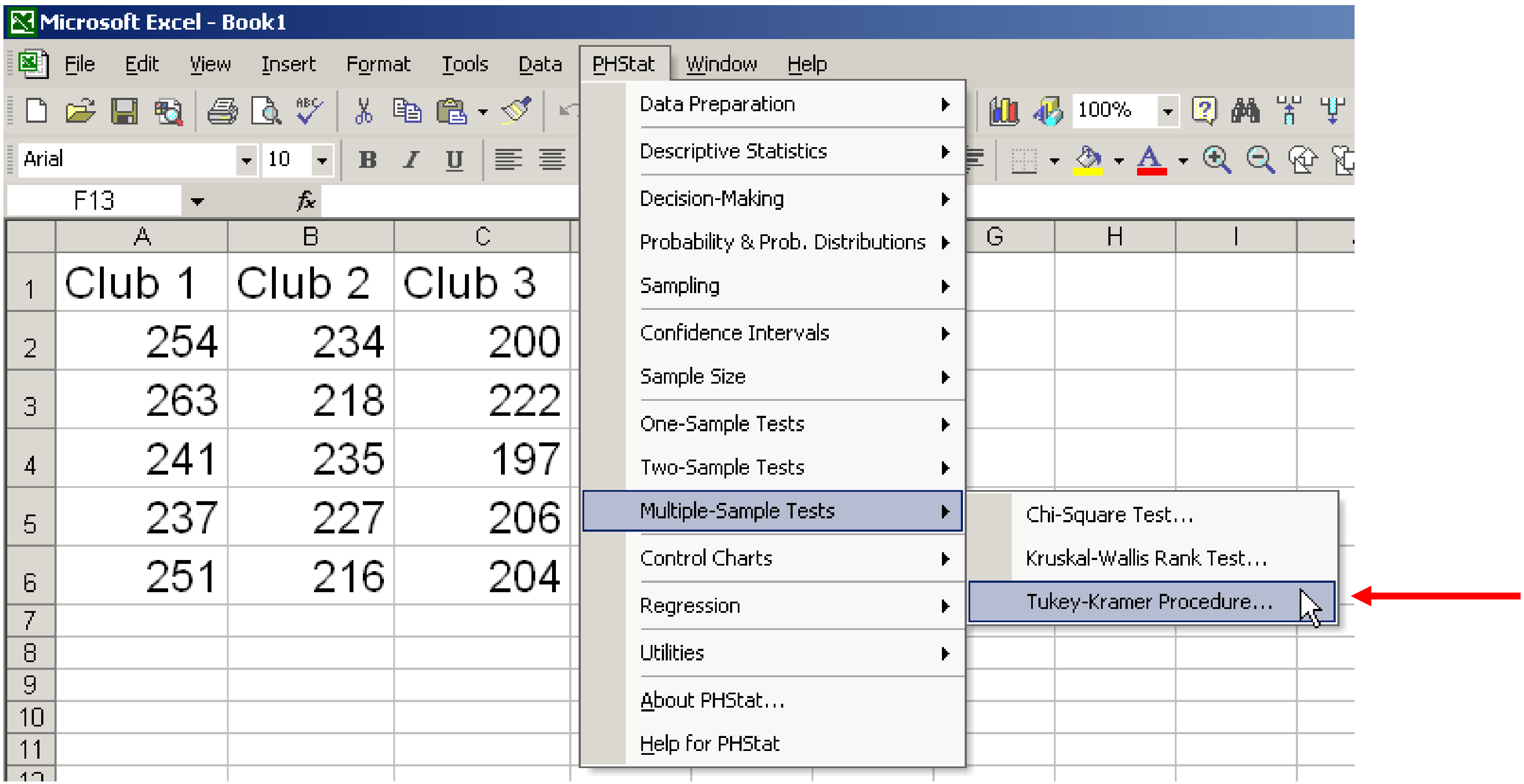
5. All of the absolute mean differences are greater than critical range. Therefore there is a significant difference between each pair of means at 5% level of significance.

$$|\bar{x}_1 - \bar{x}_2| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = 20.2$$

Tukey-Kramer in PHStat



Chapter Summary

- Described one-way analysis of variance
  - The logic of ANOVA
  - ANOVA assumptions
  - F test for difference in c means
  - The Tukey-Kramer procedure for multiple comparisons