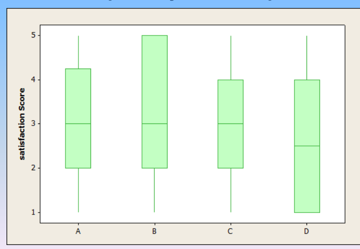
STATS 101
Introductory Statistics

One-way Analysis of Variance
(1-way ANOVA)



Example 1:

Suppose the USGA wants to compare the mean distances obtained when 4 different brands of golf balls are hit with one specific driver. The experiment is carried out with one robot golfer, as follows:

12 golf ball each from the 4 brands (A, B, C, D) are selected. A random sequence is used to hit each of these 48 balls.

This is called a COMPLETELY RANDOMIZED DESIGN.

How is the data collected?

48 golf balls A: 1-12, B: 13 – 24, C: 25 – 36, D: 37 – 48.

Generate a random sequence from integers 1, 2, …, 48. This can easily be done using R – a free programming language for statistical computing.

sample(1:48,replace=FALSE) generated following random sequence: the balls of 4 brands will be hit will be hit in this order.
9 31 16 36 3 7 12 24 19 29 18 44 42 33 37 39 25 1 20 13 47 4 6 15 8 48 46 11 40 26 32 27 21 30 28 41 43 17 35 38 10 23 45 34 2 14 22 5

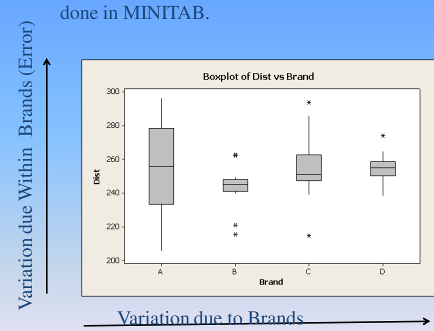
A C B C A A A B B C D D D D C A B B D A B A D D A D C C C B C C D D B C D A B D C A B B A

Suppose the following data set is obtained from the above Completely Randomized Experiment.

A	B	C	D
250.8	241.5	247	254.3
250	246.4	280.8	235.9
235.5	282.8	212.6	222.5
255.4	264.7	245.5	275.1
248.7	269.2	228	254
241.8	263.9	232.3	242.4
253.7	248.5	255.7	286
285.2	219.2	252.9	246.4
282.5	218.3	253.1	234.6
235.9	259.6	231.5	256.2
247.4	234.9	260.2	246.7
246.9	253.5	241.3	241.6

- To test if the 4 brands of golf balls yield same average distance, we test:
- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_1 : H_0 is false (i.e., at least 2 means are different)

EXCEL does not draw BOX PLOT – following was done in MINITAB.



Observe that the data set shown on slides 3 (and the box plot on slide 8) has 2 kinds of variability:

1) Variability WITHIN each brand
2) Variability ACROSS the 4 brands

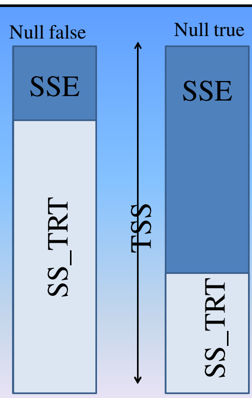
We can in fact show that total variability can be split into two parts:

$$\sum_{i=1}^{4}\sum_{j=1}^{15}(x_{ij}-\bar{x}_{..})^2=\sum_{i=1}^{4}\sum_{j=1}^{15}(x_{ij}-\bar{x}_{i.})^2+15\sum_{i=1}^{4}(\bar{x}_{i.}-\bar{x}_{..})^2$$

where

$$\bar{x}_{..}=\frac{\sum_{i=1}^{4}\sum_{j=1}^{15}x_{ij}}{N}=\text{grand mean},\quad (N=4\times 15=16)$$

$$\bar{x}_{i.}=\frac{\sum_{j=1}^{15}x_{ij}}{15}=\text{ mean distance for Brand } i,\ i=1,\dots,4$$



If average SS_TRT is large compared to average SSE, then there is a treatment effect, i.e., null will be rejected.

If average SS_TRT is small compared to average SSE, then there is NO treatment effect, i.e., null will NOT be rejected.

$$\sum_{i=1}^{4}\sum_{j=1}^{15}(x_{ij}-\bar{x}_{..})^2=\sum_{i=1}^{4}\sum_{j=1}^{15}(x_{ij}-\bar{x}_{i.})^2+15\sum_{i=1}^{4}(\bar{x}_{i.}-\bar{x}_{..})^2$$

Total sum of squares = Error sum of squares + Treatment (Brand) sum of squares

$TSS = SSE + SS\_Trt$

Each SS has a degrees of freedom (df)

df(Total) = N - 1 = 60 - 1 = 59 (for this example)

df(SS_Trt) = g - 1 = 4 - 1 = 3

df(Error) = (N-1) - (g-1) = N-g by substraction

$TSS = SSE + SS\_Trt$

df(Total) = N - 1 = 60 - 1 = 59 (for this example)

df(SS_Trt) = g - 1 = 4 - 1 = 3

df(Error) = (N-1) - (g-1) = N-g by substraction

The mean sum of squares are calculated by dividing by df:

$$\text{MS(Treatment)}=\frac{SS\_Trt}{g-1}$$

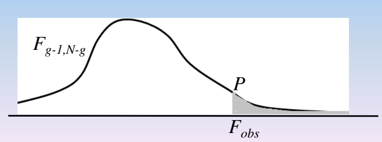
$$\text{MS(Error)}=\frac{SSE}{N-g}$$

Reject null hypothesis if

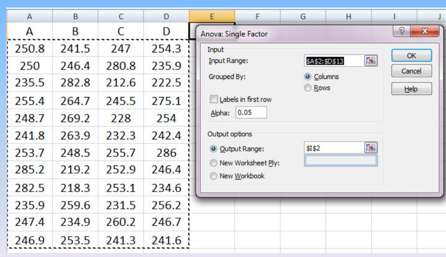
$$\frac{MS\_Trt}{MSE} > \text{cut-off C}$$

The cut-off point C comes from the F-table with degrees of freedom
df(numerator) = *g-1*
df(denominator) = *N-g*

Alternatively, H_0 is rejected if $P < \alpha = .05$
Where P-value is shown in following figure.



To run 1-way ANOVA on Example 1 data in EXCEL:
Data/Data Analysis/Anova:Single Factor
Anova:Single Factor Window will pop open
Select INPUT RANGE
Select a blank cell for output, click OK



A	B	C	D
250.8	241.5	247	254.3
250	246.4	280.8	235.9
235.5	282.8	212.6	222.5
255.4	264.7	245.5	275.1
248.7	269.2	228	254
241.8	263.9	232.3	242.4
253.7	248.5	255.7	286
285.2	219.2	252.9	246.4
282.5	218.3	253.1	234.6
235.9	259.6	231.5	256.2
247.4	234.9	260.2	246.7
246.9	253.5	241.3	241.6

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	12	3033.8	252.8167	248.9033
Column 2	12	3002.5	250.2083	384.2427
Column 3	12	2940.9	245.075	314.1239
Column 4	12	2995.7	249.6417	303.799

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	373.0323	3	124.3441	0.397561	0.755398	2.816466
Within Groups	13761.76	44	312.7672			
Total	14134.79	47				

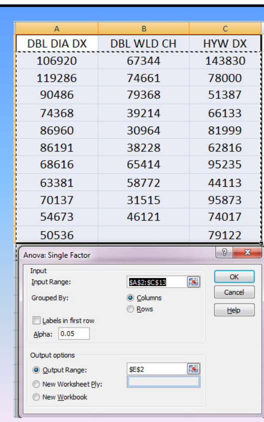
P-value = .755 > .05, H_0 is not rejected.
Conclusion: Each brand of golf ball gives same average distance.

Example 2: Coin-in for 3 different types of slot games on a casino floor are given below. Test if the mean coin-in values are same for the 3 types of slot games.

DBL DIA DX	DBL WLD CH	HYW DX
106920	67344	143830
119286	74661	78000
90486	79368	51387
74368	39214	66133
86960	30964	81999
86191	38228	62816
68616	65414	95235
63381	58772	44113
70137	31515	95873
54673	46121	74017
50536		79122

(Example 2 – continued)

In EXCEL, type data as shown in the table on the right.
Data/Data Analysis/Anova: Single Factor
This will pop open the window on the bottom.
Select Input Range, Click on "Grouped By Columns",
Provide a blank cell for output range.



DBL DIA DX	DBL WLD CH	HYW DX
106920	67344	143830
119286	74661	78000
90486	79368	51387
74368	39214	66133
86960	30964	81999
86191	38228	62816
68616	65414	95235
63381	58772	44113
70137	31515	95873
54673	46121	74017
50536		79122

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	11	871554	79232.18	4.5E+08
Column 2	10	531601	53160.1	3.29E+08
Column 3	11	872525	79320.45	7.19E+08

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	4.69E+09	2	2.34E+09	4.639643	0.017856	3.327654
Within Groups	1.47E+10	29	5.05E+08		Reject H_0 as P < .05	
Total	1.93E+10	31				

Box Plot of Coin-in by SLOT GAME TYPE for Example 2 also shows that the mean coin-in values are not equal. The boxplot was drawn in MINITAB as EXCEL does not draw boxplots.