One-Way Analysis of Variance (ANOVA)

One-Way Analysis of Variance (ANOVA) is a method for comparing the means of a populations. This kind of problem arises in two different settings

1. When a independent random samples are drawn from a populations.
2. When the effects of a different treatments on a homogeneous group of experimental units is studied, the group of experimental units is subdivided into a subgroups and one treatment is applied to each subgroup. The a subgroups are then viewed as independent random samples from a populations.
3. **Assumptions required for One-Way ANOVA**
   (a) Random samples are independently selected from a (treatments) populations.
   (b) The a populations are approximately normally distributed.
   (c) All a population variances are equal.
4. The assumptions are conveniently summarized in the following **statistical model:**

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

   where e_{ij} are independent $N(0, \sigma^2)$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, n_i$

5. Example: Tests were conducted to compare three top brands of golf balls for mean distance traveled when struck by a driver. A robotic golfer was employed with a driver to hit a random sample of 5 golf balls of each brand in a random sequence. Distance traveled, in yards, for each hit is shown in the table below.

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

   Suppose we want to compare the mean distance traveled by the three brands of golfballs based on the three samples. One-Way ANOVA provides a method to accomplish this.

6. The hypotheses of interest in One-Way ANOVA are:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \ldots = \mu_a \\ H_A &: \quad \mu_i \neq \mu_j \text{ for some i, j} \end{aligned}$$

   (a) In the above example, $a = 3$. So the mean distance traveled by the three brands of golfballs are equal according to H_0.
   (b) According to H_A, at least one mean is not equal to the others.

7. The total variability in the response, X_{ij} is partitioned into between treatment and within treatment (error) components. When these component values are squared and summed over all the observations, terms called **sums of squares** are produced. There is an additive relation which states that the total sum of squares equals the sum of the treatment and error sum of squares.

$$SST = SS_{Tr} + SSE$$

The notations SS_{Tr}, SSTr, $SS_{treatment}$, and $SS(Between)$ are synonymous for "treatment sum of squares". The abbreviations SSE, SS_{error}, SS_{Error}, SS_E and $SS(Within)$ are synonymous for "error sum of squares".

Associated with each sum of squares is its degrees of freedom. The **total degrees of freedom** is $n-1$. The **treatment degrees of freedom** is $a-1$ and the **error degrees of freedom** is $n-a$. The degrees of freedom satisfy an an additive relationship, as did the sums of squares.

$$n - 1 = (a - 1) + (n - a)$$

8. Scaled versions of the treatment and error sums of squares (the sums of squares divided by their associated degrees of freedom) are known as mean squares: $MS_{Tr} = SS_{Tr}/(a-1)$ and $MSE = SSE/(n-a)$.

9. MS_{Tr} and MS_E are both estimates of the error variance, σ^2. MSE is always unbiased (its mean equals σ^2), while MS_{Tr} is unbiased only when the null hypothesis is true. When the alternative H_A is true, MS_{Tr} will tend to be larger than MSE. The ratio of the mean squares is $F = MSTr/MSE$. This should be close to 1 when H_0 is true, while large values of F provide evidence against H_0. The null hypothesis H_0 is rejected for large values of the observed test statistic F_{obs}.

10. ANOVA calculations are conveniently displayed in the tabular form shown below, which is known as an ANOVA table. We will be making frequent use of such tables for the remainder of the course.

Source	df	SS	MS	F_{obs}	p-value
Treatments	$a-1$	SS_{Tr}	MS_{Tr}	$\frac{MS_{Tr}}{MSE}$	$P[F \geq F_{obs}]$
Error	$n-a$	SSE	MSE		
Total	$n-1$	SST			

Notation:

a is the number of factor levels (treatments) or populations

x_{ij} is te jth observation in the ith sample, $j = 1, \ldots, n_i$

n_i is sample size for the ith sample

$\bar{x}_{i.} = \sum_{j=1}^{n_i} x_{ij}/n_i$ is the ith sample mean

$s_i^2 = \frac{1}{(n_i-1)} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$ is the ith sample variance

$\bar{x}_{..} = \frac{1}{n} \sum_{i=1}^{a} n_i \bar{x}_{i.}$ is the grand mean of all observations

$n = \sum_{i=1}^{a} n_i$ is the total number of observations

Here are the formulas for sums of squares. We will see that there are simpler formulas when we know the sample means and sample variances for each of the a groups.

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

$$SS_{Tr} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_{i=1}^{a} n_i (\bar{x}_{i.} - \bar{x}_{..})^2$$

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 = \sum_{i=1}^{a} (n_i - 1) s_i^2$$

The test statistic is $F_{obs} = \frac{SS_{Tr}/(a-1)}{SSE/(n-a)}$ and the p-value is $P(F \geq F_{obs})$.

Notes:

- F_{obs} is the observed value of the test statistic
- Under the null hypothesis F has an F distribution with $a-1$ numerator and $n-a$ denominator degrees of freedom
- the p-value is $P(F_{(a-1),(n-a)} \geq F_{obs})$
- reject H_0 at level α when the p-value $< \alpha$
- equivalently, when $F_{obs} \geq F_{\alpha,(a-1),(n-a)}$, where $F_{\alpha,(a-1),(N-a)}$ is the upper α'th percentile of the F distribution $(a-1)$ numerator and $(n-a)$ denominator degrees of freedom.

F Distribution - using the F table

(a) What is the probability that an F variable with 7 numerator and 23 denominator degrees of freedom is less than 3? Ans: the probability is between .01 and .05

(b) What is the probability that an F variable with 3 numerator and 5 denominator degrees of freedom is greater than 12.5? Ans: less than .01.

11. Example: Consider again the tests conducted to compare three brands of golf balls for mean distance traveled when struck by a driver. Again, distances traveled, in yards, for each hit and are shown below.

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

(a) Here we'll compare the mean distance traveled by the different brands of golfballs using ANOVA.

i	$\bar{x}_{i.}$	S_i^2	n_i
1	251.18	33.487	5
2	261.98	18.197	5
3	269.66	27.253	5

- Total sample size $n = \sum_{i=1}^{a} n_i = 5 + 5 + 5 = 15$
- $a = 3$ groups
- The treatment degrees of freedom is $a - 1 = 2$. The error degrees of freedom is $n - a = 15 - 3 = 12$.
- The grand mean is $\bar{x}_{..} = \frac{1}{n} \sum_{i=1}^{a} n_i \bar{x}_{i.} = (5 \times 251.18 + 5 \times 261.985 \times 269.66)/15 = 260.94$
- The treatment sum of squares is $SS_{Tr} = \sum_{i=1}^{a} n_i (\bar{x}_{i.} - \bar{x}_{..})^2$

  $$SS_{Tr} = 5[(251.18 - 260.94)^2 + (261.98 - 260.94)^2 + (269.66 - 260.94)^2] = 861.89$$

- The error sum of squares is

  $$SSE = \sum_{i=1}^{a} (n_i - 1) s_i^2$$

  $$SSE = 4[33.487 + 18.197 + 27.253] = 315.748$$

- The quantities can be summarized in an ANOVA table

Source	SS	df	MS	F
Treatment	861.29	2	430.645	16.37
Error	315.75	12	26.312	
Total	$SS_T = 1177.64$	14		

- The observed test statistic is $F_{obs} = 16.37$ with 2 numerator and 12 denominator degrees of freedom.
- The p-value is $P(F_{2,12} > 16.37) < .01$
- Since the p-value $< .01$, reject H_0 at $\alpha = .01$ and conclude that the mean travel distances for all three brands of golfballs are not the same.

Here are the calculations done three ways.

```
MTB > read c1-c3   read data into columns 1-3
DATA> 251.2 263.2 269.7
DATA> 245.1 262.9 263.2
DATA> 248 265 277.5
DATA> 251.1 254.5 267.4
DATA> 260.5 264.3 270.5
DATA> end
5 rows read.

MTB > aovoneway c1-c3  do the one way ANOVA

One-way ANOVA: C1, C2, C3

Source  DF      SS     MS      F      P
Factor   2   861.9  430.9  16.38  0.000
Error   12   315.7   26.3
Total   14  1177.6
Pooled StDev = 5.13   (The estimate of sigma is 5.13. )


(Now put the data into another format.  Stack the observations
 for brands A,B and C.)

MTB > stack c1-c3 c10

MTB > set c11
DATA> 5(1) 5(2) 5(3)

(C11 is 1 for brand A observations, 2 for brand B, and 3 for brand C.)


DATA> oneway c10 c11    (when the data is stacked, use the oneway command)

One-way ANOVA: C10 versus C11

Source  DF      SS     MS      F      P
C11      2   861.9  430.9  16.38  0.000
Error   12   315.7   26.3
Total   14  1177.6
Pooled StDev = 5.13

(Note that the ANOVA tables are identical.)

(Now calculate sums of squares using the basic formulae.)

MTB > let k1=stdev(c10)   (k1 is the standard deviation of the 15 data points)
MTB > let k2=(k1**2)*14   ( the sample variance * (n-1) = total sum of squares)
MTB > print k2
```

```
Data Display

K2    1177.64              (this agrees with SST from the Anova tables)


MTB > descr c1-c3;         (get the individual sample means)
SUBC> mean.

Descriptive Statistics: C1, C2, C3

Variable    Mean
C1        251.18
C2        261.98
C3        269.66

MTB > set c5               (enter sample means to C5)
DATA> 251.18 261.98 269.66
DATA> end
MTB > set c4               (enter sample sizes to C4)
DATA> 5 5 5
DATA> end

MTB > let k4=sum(c4*c5)/sum(c4)  (the overall mean = sum (ni*xbari)/sum(ni))
MTB > print k4

Data Display

K4    260.940

MTB > let k5=sum(c4*(c5-k4)**2)  (SStreatment = sum(ni*(xbari-xbar)**2) )
MTB > print k5

Data Display

K5    861.888                    (agrees with SStreatment from Anova tables)

MTB > descr c10;                 (here's how to get standard deviations from
SUBC> by c11;                                          stacked data)
SUBC> stdev.

Descriptive Statistics: C10

Variable  C11  StDev
C10       1     5.79
          2     4.27
          3     5.22

MTB > set c6                    (enter standard deviations to C6)
DATA> 5.79 4.27 5.22
DATA> end
MTB > let c7=c6**2              (square to get variances, in C7)
```

```
MTB > let k5=sum((c4-1)*c7)  (SSE = sum (sample size -1)*(sample variance)
MTB > print k5

Data Display

K5    316.022   (SSE - note rounding error from using only 3 digits for
                  standard deviations. Otherwise this agrees with SSE from
                  ANOVA tables.)
```

12. Example: A group of 32 rats were randomly assigned to each of 4 diets labelled (A,B,C,and D). The response is the liver weight as a percentage of body weight. Two rats escaped and another died, resulting in the following data

A	B	C	D
3.42	3.17	3.34	3.65
3.96	3.63	3.72	3.93
3.87	3.38	3.81	3.77
4.19	3.47	3.66	4.18
3.58	3.39	3.55	4.21
3.76	3.41	3.51	3.88
3.84	3.55		3.96
	3.44		3.91

Here is how to carry out the ANOVA in minitab.

```
MTB > set c1
DATA> 3.42 3.96 3.87 4.19 3.58 3.76 3.84 3.17 3.63 3.38 3.47 3.39
DATA> 3.41 3.55 3.44 3.34 3.72 3.81 3.66 3.55 3.51 3.65 3.93
DATA> 3.77 4.18 4.21 3.88 3.96 3.91
DATA> end
MTB > set c2
DATA> 7(1) 8(2) 6(3) 8(4)
DATA> end

MTB > oneway c1 c2

One-way ANOVA: C1 versus C2

Source  DF      SS      MS      F      P
C2       3  1.1649  0.3883  10.84  0.000
Error   25  0.8954  0.0358
Total   28  2.0603

S = 0.1893   R-Sq = 56.54\%   R-Sq(adj) = 51.32\%

Pooled StDev = 0.1893
```

Let's verify the calculations based on the following summary statistics:

```
Level  N    Mean   StDev
1      7  3.8029  0.2512
2      8  3.4300  0.1353
3      6  3.5983  0.1675
4      8  3.9363  0.1884


MTB > set c7
DATA> 7 8 6 8
DATA> set c8
DATA> 3.8028 3.43 3.5983 3.9363
DATA> set c9
DATA> .2512 .1353 .1675 .1884
```

```
DATA> end

MTB > let k1=sum(c7*c8)/sum(c7)
MTB > let k3=sum(c7*(c8-k1)**2)
MTB > print k3
   K3    1.16505       (this is the treatment SS)

MTB > let k2=sum((c7-1)*c9**2)
MTB > print k2
   K2    0.895494      (this is the SSE)
```

You need to be comfortable with the order of calculations in an ANOVA table. Fill in the blanks in the following table.

Source	SS	df	MS	F	p-value
Treatment		4			
Error	900				
Total	1200	12			