5.1 One-way Analysis of Variance

Example. Suppose subjects are recruited for a cohort study from three different communities. The purpose of the study is to analyze health complications caused by obesity. For validity of results, the three cohorts should be as uniform as possible with respect to many characteristics at the baseline. Suppose the measurements of body mass indices (BMI) for five people in each sample are given in the table below.

Community	BMI				
A	29.3	31.4	38.7	33.2	30.3
B	42.0	39.9	44.5	40.7	38.9
C	28.8	36.1	37.5	31.0	33.2

We need to test that the average BMI are the same in all three communities. Assuming that BMI has a normal distribution, we can conduct a **one-way analysis of variance** (ANOVA). In a one-way ANOVA, there is a single **factor** (sometimes referred to as **treatment**) with several **levels**. In our example, the factor is the communities with 3 levels. We have 5 observations per level.

In a one-way ANOVA, the data are typically presented in a table like this:

Factor	Observations				**Totals**	**Averages**
1	y_{11}	y_{12}	$\dots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\dots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
.	.	.	$\dots$	.	.	.
.	.	.	$\dots$	.	.	.
.	.	.	$\dots$	.	.	.
a	y_{a1}	y_{a2}	$\dots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

Definition. In a one-way ANOVA, the statistical model for the observations is $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, where y_{ij} is the j-th observation for the i-th factor (the ij-th observation), $i = 1, \dots, a$, $j = 1, \dots, n$. The parameter μ is the **overall mean**, τ_i is the i-th factor effect (or the **i-th treatment effect**), and ε_{ij} is a **random error**. It is assumed that ε_{ij}'s are independent and identically distributed (iid) with the distribution $N(0, \sigma^2)$, where σ is constant for all observations.

Note that in this model, the measurements y_{ij} are iid $N(\mu + \tau_i, \sigma^2), i = 1, \dots, a,\ j = 1, \dots, n$. Note also that the overall mean μ is the mean of all

the averages and so it must be that $\mu = \frac{1}{na}\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu+\tau_i) = \mu + \frac{1}{a}\sum_{i=1}^{a}\tau_i$. From here, we conclude that $\sum_{i=1}^{a}\tau_i = 0$, that is, the treatment effects sum up to 0. The goal of the analysis is to check if the expected values $\mu+\tau_i$'s are equal for all $i = 1,\ldots,a$. Thus, the statistical hypotheses can be written as $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$ versus $H_1 : \tau_i \neq 0$ for some $i = 1,\ldots,a$.

Next, we derive the test statistic by first proving **decomposition of the total sum of squares**, the key formula in the analysis of designed experiments. To this end, we introduce the following notation.

The **sum** of all observations for the i-th level of the factor is denoted by $y_{i.} = \sum_{j=1}^{n} y_{ij}$. The **mean** is denoted by $\bar{y}_{i.} = \frac{y_{i.}}{n}, \quad i = 1,\ldots,a$. The **overall total** is denoted by $y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$, and the **overall mean** is $\bar{y}_{..} = \frac{y_{..}}{N}$, where $N = na$.

The **total corrected sum of squares** is written as

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2.$$

It represents the total variability in the data. It can be partitioned into component parts (whence, the name "Analysis of Variance"). We derive

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}[(\bar{y}_{i.}-\bar{y}_{..})+(y_{ij}-\bar{y}_{i.})]^2$$

$$= n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.}) + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2.$$

The cross product is equal to zero (!):

$$\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.}) = y_{i.} - n\bar{y}_{i.} = y_{i.} - n(y_{i.}/n) = 0.$$

Thus, $SS_T = SS_{tr} + SS_E$ where $SS_{tr} = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2$ is the **sum of squares due to treatments** (or **sum of squares between treatments**), and $SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$ is the **sum of squares due to error** (or

the **sum of squares within treatments**).

The quantities $MS_{tr} = \frac{SS_{tr}}{a-1}$ and $MS_E = \frac{SS_E}{N-a}$ are called **mean squares.** It can be shown that

$$\mathbb{E}(MS_{tr}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1} \text{ and } \mathbb{E}(MS_E) = \sigma^2.$$

Under H_0, both MS_E and MS_{tr} are estimates of σ^2. Moreover, it can be shown that, under H_0, SS_E/σ^2 and SS_{tr}/σ^2 are independent chi-square random variables with $N-a$ and $a-1$ degrees of freedom. Therefore, if the null hypothesis is true, the ratio

$$F = \frac{SS_{tr}/(a-1)}{SS_E/(N-a)} = \frac{MS_{tr}}{MS_E}$$

has an F-distribution with $a-1$ and $N-a$ degrees of freedom. Thus, one would reject the null, if $F > F_{\alpha, a-1, N-a}$.

Example (continued). To conduct the analysis of variance in our example, we run the following SAS and R codes.

In SAS:

```
data cohorts;
 input community $ BMI @@;
cards;
A 29.3 A 31.4 A 38.7 A 33.2 A 30.3
B 42.0 B 39.9 B 44.5 B 40.7 B 38.9
C 28.8 C 36.1 C 37.5 C 31.0 C 33.2
;

proc anova;
class community;
 model BMI=community;
 run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	228.2440000	114.1220000	10.96	0.0020
Error	12	124.9760000	10.4146667		
Corrected Total	14	353.2200000			

Since the p-value < 0.01, we reject the null hypothesis and conclude at the 1% significance level that the average BMIs in the three communities are not all the same.

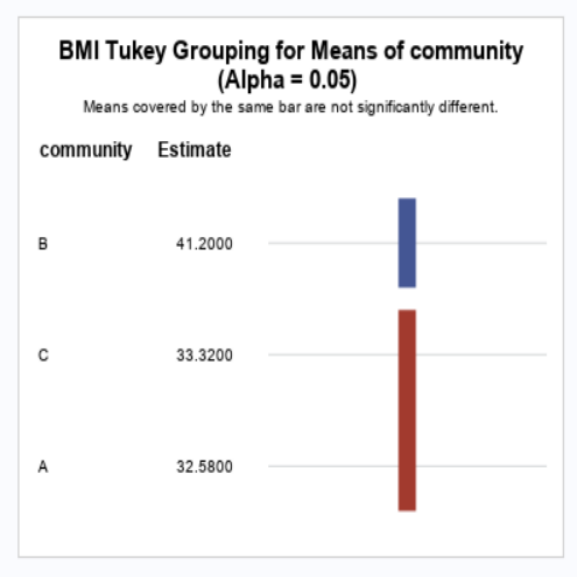
In R:

```
community<- c("A", "A", "A", "A", "A", "B", "B", "B", "B", "B", "C",
              "C", "C", "C", "C")
BMI<- c(29.3, 31.4, 38.7, 33.2, 30.3, 42.0, 39.9, 44.5, 40.7, 38.9,
        28.8, 36.1, 37.5, 31.0, 33.2)

summary(model<- aov(BMI~community))

            Df Sum Sq Mean Sq F value  Pr(>F)
community    2  228.2  114.12   10.96 0.00196
Residuals   12  125.0   10.41
```

Further, since our conclusion was that not all average BMIs are the same, the post hoc procedure would be conduct simultaneous pair-wise testings to see which communities are different from which. Put rigorously, we test simultaneously $\binom{a}{2}$ pairs of hypotheses $H_0 : \tau_i = \tau_j$ for all $i \neq j$ against two-sided alternatives $H_1 : \tau_i \neq \tau_j$ for some $i \neq j$. There are many types of tests that can be conducted. The most common one is called the **Tukey's Honestly Significant Difference (HSD) Test** (or just **Tukey test**). The test declares two means significantly different if the absolute value of the difference of the respective sample treatment means exceeds $q_\alpha(a, N-a)\sqrt{MS_E/n}$ where the critical values $q_\alpha(a, N-a)$ are tabulated quantities (see, for example, Table V on pages 707-708 in the Montgomery's book, 9th edition).

In SAS:

```
proc glm;
  class community;
   model BMI=community;
     means community/tukey;
run;
```



From the output, we can conclude that in Community B, the average BMI is higher than those in the other two communities. But in Communities A and C, the average BMIs are statistically indistinguishable.

In R:

```
TukeyHSD(model)

$community
     diff        lwr       upr     p adj
B-A  8.62   3.174769 14.065231 0.0031308
C-A  0.74  -4.705231  6.185231 0.9305172
C-B -7.88 -13.325231 -2.434769 0.0059301
```

□