12: Analysis of Variance

Introduction | EDA | Hypothesis Test

Introduction

In Chapter 8 and again in Chapter 11 we compared means from two independent groups. In this chapter we extend the procedure to consider means from k independent groups, where k is 2 or greater. The technique is called analysis of variance, or *ANOVA* for short.

Illustrative data **(`anova.sav`).** We consider `AGE` of study participants from three different clinical centers. Ages (years) are:

Center 1 (n_1 = 22): 60, 66, 65, 55, 62, 70, 51, 72, 58, 61, 71, 41, 70, 57, 55, 63, 64, 76, 74, 54, 58, 73
Center 2 (n_2 = 18): 56, 65, 65, 63, ,57, 47, 72, 56, 52, 75, 66, 62, 68, 75, 60, 73, 63, 64
Center 3 (n_1 = 23): 67, 56, 65, 61, 63, 59, 42, 53, 63, 65, 60, 57, 62, 70, 73, 63, 55, 52, 58, 68, 70, 72, 45

Data are entered as two separate variables, one for the **dependent variable** and one for the **group ("independent") variable**. Data are stored in `anova.sav` are stored as variables `AGE` and `CENTER`. The *first five and last five observations* are:

OBS	AGE	CENTER
1	60	1
2	66	1
3	65	1
4	55	1
5	62	1
etc.	etc.	etc.
59	58	3
60	68	3
61	70	3
62	72	3
63	45	3

EDA

Summary statistics. Let N denote the total sample size (e.g., N = 63). Let $\bar{x}$ (no subscript) represent the mean of all N subjects, combined. This is called the **grand mean.** For the illustrative `AGE` data, $\bar{x}$ = 62.127.

Group means, standard deviations, and **sample sizes** are denoted with subscripts. For the illustrative data:

$\bar{x}_1$ = 62.546, s_1 = 8.673, n_1 = 22

$\bar{x}_2$ = 63.278 , s_2 = 7.789, n_2 = 18

$\bar{x}_3$ = 60.826, s_3 = 8.004, n_3 = 23

We note that group means and standard deviations are all within a couple of years of each other.

Graphical Analysis. A **side-by-side boxplot** is one of the best way to compare group locations, spreads, and shapes. (Detailed instruction on how to draw and interpret boxplots was presented in Chapter 4). A side-by-side boxplot for the illustrative data, shown at the bottom of the page, shows distributions with similar shapes, locations and spreads. It may be worth noting that group 3 has a low outside value.

> ***SPSS.*** Click `Analyze > Descriptive Statistics > Explore` and place the dependent variable in the `Dependent list` and the group variable in the `Factor list`.

Other graphical technique such as side-by-side dot plots, stem-and-leaf plots, mean±SD, mean±SE, and confidence interval plots may also be employed to compare the groups.

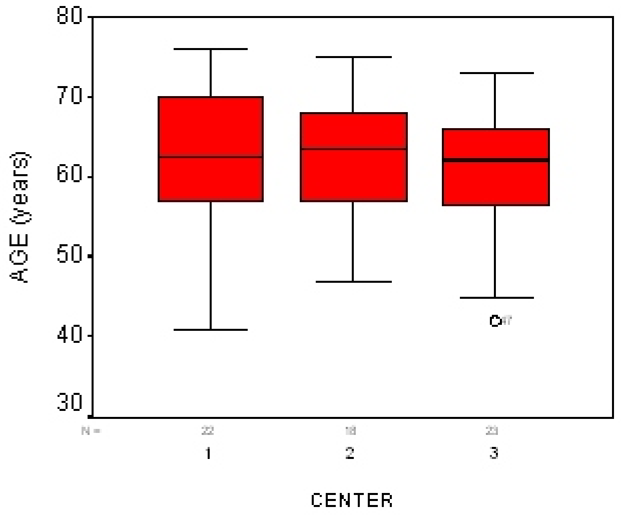


Figure 1

Hypothesis Test (ANOVA)

Null and Alternative Hypotheses

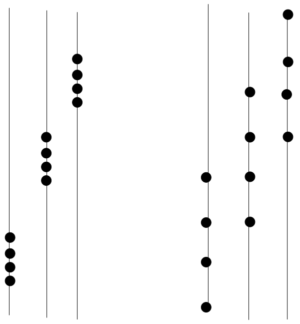
The name analysis of *variance* may mislead some students to think the technique is used to compare group variances. In fact, analysis of variance uses variance to cast inference on group *means*. The null and alternative hypotheses are:

$$H_0\text{: } \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_1\text{: } H_0 \text{ is false (“at least one population mean differs”)}$$

where μ_i represents the population mean of group *i*.

Whether an observed difference between group mean is “surprising” will depends on the spread (variance) of the observations within groups. Widely different averages can more likely arise by chance if individual observations within groups vary greatly. We must therefore take into account the variance within group when assessing differences between groups. Consider the dot-plots below. Surely the difference demonstrated between the first three groups is more likely to be significantly that the difference demonstrated by the second three groups.



Thus, if the variance between groups exceeds what is expected in terms of the variance within groups, we will reject the null hypothesis.

Let σ^2_W represent the variance within groups in the population. Let σ^2_B denote the variance between groups within the population. The null and alternative hypotheses may now be restated as:

$$H_0\text{: } {}^2{}_B \le {}^2{}_W$$

$$H_1\text{: } {}^2{}_B > {}^2{}_W$$

The variance between groups may be thought of as a *signal* of group differences. The variance within groups may be thought of as *background noise*. When the signal exceeds the noise, we will reject the null hypothesis.

Variance Between Groups

Let s^2_B represent the **sample variance between groups**:

$$s_B^2 = \frac{SS_B}{df_B} \tag{1}$$

This statistic, also called the **Means Square Between (MSB)**, is a measure of the variability of group means around the grand mean (Fig. 2).

The SS_B (sum of squares between) is

$$SS_B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \tag{2}$$

where n_i represents the size of group i, $\bar{x}_i$ represents the mean of group i, and $\bar{x}$ represents the grand mean. For the illustrative data, $SS_B = [(22)(62.5-62.127)^2 + (18)(63.3-62.127)^2 + (23)(60.8-62.127)^2] = 66.6$.

This statistic has degrees of freedom:

$$df_B = k - 1 \tag{3}$$

where k represents the number of groups. For the illustrative data, $df_B = 3-1$.

For the illustrative data, $s^2_B = 66.614 / 2 = 33.3$.

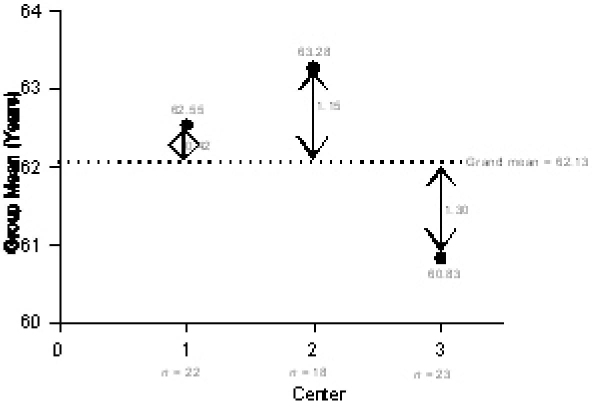


Figure 3. Variability between.

Variance Within Groups

The **variance within (s^2_W)** quantifies the spread of values within groups (Fig. 3). In the jargon of ANOVA, the variance within is also called the **Mean Square Within (MSW)** and is calculated:

$$s^2_W = \frac{SS_W}{df_W} \tag{4}$$

where the **sum of squares within (SS_W)** is:

$$SS_W = \sum_{i=1}^{k} (n_i - 1)s_i^2 \tag{5}$$

and the **degrees of freedom within** is:

$$df_W = N - k \tag{6}$$

For the illustrative data, $SS_W = [(22-1)(8.673^2) + (18-1)(7.789^2) + (23-1)(8.004^2)] = 4020.36$ and $df_w = 63 - 3 = 60$. Thus, $s^2_w = 4020.36 / 60 = 67.006$.

An **alternative formula** for the variance within is:

$$s^2_w = \frac{(df_i)(s_1^2) + (df_2)(s_2^2) + \cdots + (df_k)(s_k^2)}{df_1 + df_2 + \cdots + df_k} \tag{7}$$

where s^2_i represent the variance in group i and df_i represent the degrees in the group ($df_i = n_i - 1$). This alternative formula shows the variance within as a weighted average of group variances with weights determined by group degrees of freedom.

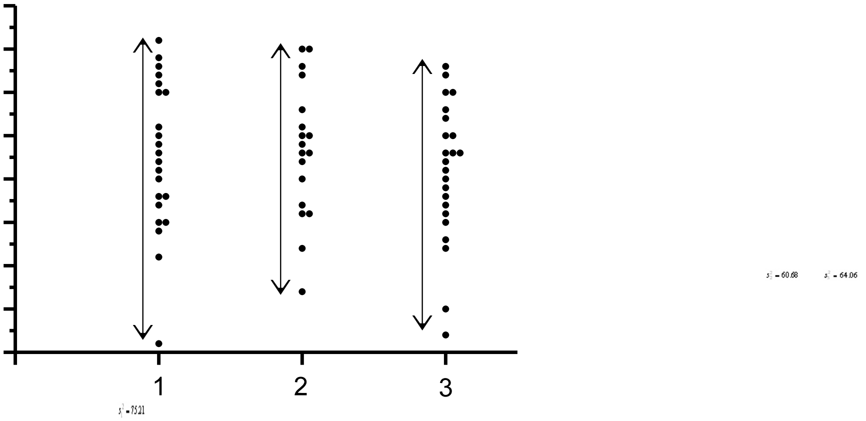


Figure 4. Variability within groups

ANOVA Table and *F* Statistic

The statistics describe thus far are arranged to form an ANOVA table as follows:

Source	**Sum of Squares**	**Degrees of freedom**	**Mean Squares**
Between	SS_B	$\text{df}_B = k \quad 1$	$s_B^2 = SS_B / df_B$
Within	SS_W	$\text{df}_w = N \quad k$	$s_W^2 = SS_W / df_W$
Total	$SS_T = SS_B + SS_W$	$\text{df} = \text{df}_B + \text{df}_w$	

The ANOVA table for the illustrative data is:

Source	**Sum of Squares**	**Degrees of freedom**	**Mean Squares**
Between	66.614	2	33.307
Within	4020.370	60	67.006
Total	4086.984	62	

The ratio of the variance between (s^2_B) and the variance within (s^2_W) is the ANOVA *F* statistic:

$$F_{\text{stat}} = \frac{s_B^2}{s_W^2} \tag{8}$$

Under the null hypothesis, this test statistic has an *F* sampling distribution with df_1 and df_2 degrees of freedom. The *p* value for the test is represented as the area under $F_{\text{df1,df2}}$ to the right tail of the F_{stat}. For the illustrative example, F_{stat} = 33.307 / 67.006 = 0.50 with 2 and 60 degrees of freedom. Using the *F* table we note $F_{2,60,.95} = 3.15$. Therefore, $p > .05$ (Fig. 4). The more precise *p* value can be computed with `WinPepi > WhatIs.EXE` and other probability calculator ($p = .60$).

SPSS: Click `Analyze > Compare Means > One-way ANOVA`. Then select the dependent variable and independent variable, and click `OK`.

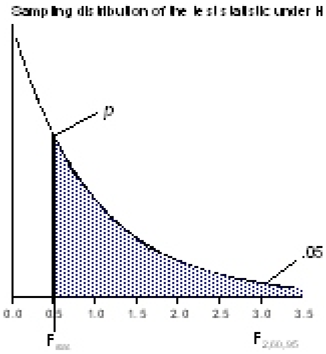


Figure 5.

OPTIONAL: WHEN k = 2, ANOVA = EQUAL VARIANCE *T* TEST

ANOVA is an extension of the equal variance independent *t* test. Whereas the independent *t* tests compares the two groups (k = 2) by testing H_0: $\mu_1 = \mu_2$, ANOVA tests k groups, where k represents any integer greater than 1. Moreover,

1. The variance within groups calculated by ANOVA is equal to the pooled estimate of variance used in the independent t test ($s^2_{\mathrm{w}} = s^2_{\mathrm{p}}$)

2. df_{B} in the ANOVA in testing two groups = 2−1 = 1

3. df_{w} in the ANOVA = $N - 2 = n_1 + n_2 - 2$ = the degrees of freedom in the independent *t* test

4. the F_{stat} from the ANOVA = the $(t_{\mathrm{stat}})^2$ from the independent *t* test

5. when = .05, $F_{1,N\text{-}2,.95} = t^2_{N\text{-}2,.975}$

OPTIONAL: SAMPLE SIZE REQUIREMENTS FOR ANOVA

Sample size requirements for an ANOVA can be determined by asking how big a sample is needed to detect a difference of Δ at a type I error rate of α with power 1−β. It is also necessary to assume a value of variance within groups (σ_{W}). Computational solutions are available in Sokal & Rohlf (1996, pp. 263-264). Calculations have been scripted on the website http://department.obg.cuhk.edu.hk/ and can be accessed by clicking `Statistical Tools > Statistical Tests > Sample Size > Comparing Means`.

Illustrative example. Suppose we test H_0: $\mu_1 = \mu_2 = \mu_3$. Prior studies estimate the measurement has standard deviation (within groups) of 8. To find a mean difference of 5, the above website derives the following results:

Sample size Per Group:

	Type I error=0.05	Type I error=0.01	Type I error=0.001
Power=80%	41	60	87
Power=90%	54	76	107
Power=95%	67	91	125

The output provides samples sizes per group (n_i) at various power and α levels. For example, under the stated assumptions, we need n = 54 (per group) for 90% power at α = .05.