Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Analysis of Variance (ANOVA)

ANOVA

- The process of evaluating hypotheses regarding the group means of multiple populations is called the Analysis of Variance (ANOVA).
- ANOVA models generalize the *t*-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories.
- Since we are only considering one factor only, this method is specifically called one- way ANOVA.
- An ANOVA with two factors is called a two-way ANOVA.
- In general, the between-groups variation is denoted as SS_B and calculated by

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

where k is the number of groups

ANOVA

- The within-groups variation is denoted as SS_W and calculated by

$$SS_w = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{ij} - \bar{y}_i)^2$$

- We measure the total variation in Y by

$$SS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

- The total variation SS is equal to the sum of the between-groups variation SS_B and the within-groups variation SS_W,

$$SS = SS_B + SS_W.$$

- The total variation can be attributed partly to the variation within groups and partly to the variation between groups.
- SS_B is interpreted as the part of total variation SS that is associated with (and can be explained by) the factor variable X (e.g., syndrome type).
- In contrast, SS_W is regarded as the unexplained part of total variation and is regarded as random.

ANOVA

- Let us denote the overall population mean of Y as μ and group-specific population means as $\mu_1, \ldots, \mu_4$.
- Then we can express the null hypothesis of no difference in means between the groups as

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

- The alternative hypothesis H_A is that at least one of the group means μ_i is different from the mean μ.
- The test statistic for examining the null hypothesis is called F-statistic (more specifically, ANOVA F -statistic) and is defined as

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

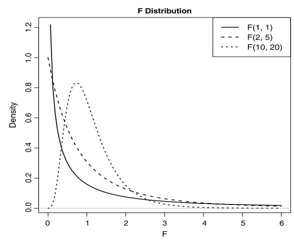
where n is the total sample size, and k is the number of groups.
- The numerator is called the mean square for groups, and the denominator is called the mean square error (MSE).

ANOVA

- For the one-way ANOVA, the F-statistic has $F(df_1 = k - 1, df_2 = n - k)$ distribution under the null hypothesis (i.e., assuming that the null hypothesis is true).
- The F-distribution, which is a continuous probability distribution, is very important for hypothesis testing.
- It is specified by two parameters, df_1 and df_2, and is denoted as $F(df_1, df_2)$.
- We refer to df_1 and df_2 as the numerator degrees of freedom and denominator degrees of freedom, respectively.
- Both parameters must be positive.

ANOVA

- The following figure shows the pdf of F-distribution for different values of df_1 and df_2.



Example

- As an example, we analyze the Cushings data set, which is available from the MASS package.
  - Cushing's syndrome is a hormone disorder associated with high level of cortisol secreted by the adrenal gland.
- The *Type* variable in the data set shows the underlying type of syndrome, which can be one of four categories:
  - adenoma (a),
  - bilateral hyperplasia (b),
  - carcinoma (c),
  - unknown (u).

Example

- Objective is to find whether the four groups are different with respect to urinary excretion rate of Tetrahydrocortisone.
- We denote by Y the urinary excretion rate of Tetrahydrocortisone and by X the *Type* variable,
  - where $X = 1$ for Type = a, $X = 2$ for Type = b, $X = 3$ for Type = c, and $X = 4$ for Type = u.
- Then, our objective could be defined as investigating whether the *mean* of the response variable Y differs for different values (levels) of the factor X.

Example

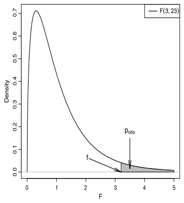
- Denote the individual observations as y_{ij} : the urinary excretion rate of Tetrahydrocortisone of the jth individual in group i.
- Total number of observations is $n = 27$,
- The number of observations in each group is
  $n_1 = 6$, $n_2 = 10$, $n_3 = 5$, and $n_4 = 6$.
- The overall (for all groups) observed sample mean for the response variable is $\bar{y} = 10.46$.
- We also find the group specific means, by clicking (in R-Commander) *Statistics→Summaries→Numerical summaries*
  $\bar{y}_1 = 3.0$, $\bar{y}_2 = 8.2$, $\bar{y}_3 = 19.7$, and $\bar{y}_4 = 14.0$.
- The degrees of freedom parameters are
  $df_1 = 4-1 = 3$ and $df_2 = 27 - 4 = 23$.

Example

- $SS_B = 893.5$ and $SS_W = 2123.6$.
- The observed value of F-statistic is $f = 3.2$ given under the column labeled F value.
- The resulting p-value is then 0.04.
- Therefore, we can reject H_0 at 0.05 significance level (but not at 0.01) and conclude that the differences among group means for urinary excretion rate of Tetrahydrocortisone are statistically significant (at 0.05 level).

Example

- For plotting the $F(3, 23)$ distribution using R-Commander, click *Distribution → Continuous distributions→ F distribution Plot F distribution*.
- Set the *Numerator degrees of freedom* to 3 and the *Denominator degrees of freedom* to 23.



- The density plot of $F(3, 23)$-distribution.
- This is the distribution of F-statistic for the Cushings data assuming that the null hypothesis is true.
- The observed value of the test statistic is $f = 3.2$, and the corresponding p-value is shown as the *shaded area* above 3.2