# Second course in Statistics. Lecture 20.

One-way analysis of variance (ANOVA)

Two-way analysis of variance (ANOVA)

# ANOVA table

| Source of Variance | df | SS | MS | F statistic |
|---|---|---|---|---|
| Groups | $k-1$ | $SSG=\sum_{j=1}^{k} n_j\left(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot}\right)^2$ | $MSG=\frac{SSG}{k-1}$ | $F=\frac{MSG}{MSE}$ |
| Error | $N-k$ | $SSE=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_{\cdot j}\right)^2$ | $MSE=\frac{SSE}{N-k}$ | |
| Total | $N-1$ | $SST=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_{\cdot\cdot}\right)^2$ | | |

$SST=SSG+SSE$

F test is always one-sided because any differences among group means tend to make F large.

## ANOVA table

For the effect of ceiling insulation on energy consumption in note 19

| Source of Variance | df | SS | MS | F statistic |
|---|---|---|---|---|
| Groups | 4 | $SSG = 9.836$ | $MSG = 2.459$ | $F = 36.458$ |
| Error | 18 | $SSE = 1.214$ | $MSE = 0.067$ | |
| Total | 22 | $SST = 11.050$ | | |

Conclusion: Since $F = 36.458 > F_{0.01}(4,18) = 4.579$, we can reject $H_0$ at 1% level of significance and conclude that the thickness of ceiling insulation has effect on energy consumption.

# Coefficient of determination

Total variation can be decomposed as $SSG$ and $SSE$. $SSG$ is the variance between the groups and $SSE$ is the variance between the groups.

Definition of coefficient of determination:
It is the percentage of variation explained by variance between the groups among total variance.

$$R^2 = \frac{SSG}{SST}$$

$R^2$ is the value between 0 and 1. The bigger the $R^2$, the larger variance explained by differences between the groups and smaller variance left within the groups.

# One-way ANOVA

SPSS output of example 3 in note 19

**ANOVA OF SCI FOR UNSKILLE SURPERVISORS**

n15_SCI

| | Sum of Square | df | Mean Sq |
|---|---|---|---|
| Between Groups | 46 | | |
| Within Groups | | | 32 |
| Total | 19 | 58 | |

$$R^2 = \frac{4662.233}{196391.4} = 0.024$$

About 2% of the variation in SCI scores is explained by membership in the groups of workers, unskilled, skilled workers, and supervisors. The other 98% of the variation is due to worker-to-worker variation within each of the three groups.

## Comparing the means

The ANOVA F test gives a general answer to a general question: Are the differences among observed group means significant? Rejection of null hypothesis tells us that the groups mean are not all the same. Unfortunately it does not tell us specifically which means differ from each other.

In SCI example, having the evidence that the three population mean SCI scores are not the same, we would really like to know which alternative is true. Is $\mu_{UN} \neq \mu_{SK}$ or $\mu_{UN} \neq \mu_{SU}$ or $\mu_{SK} \neq \mu_{SU}$ or is any linear combination of these statements true?

Contrast is a useful tool to obtain the results.

# Contrasts

A contrast is a combination of population means of the form

$$\psi = \sum a_i \mu_i$$

where the coefficient $a_i$ have sum 0. The corresponding sample contrasts is

$$c = \sum a_i \bar{x}_i$$

The standard error of c is

$$SE_c = \sqrt{MSE}\sqrt{\sum \frac{a_i^2}{n_i}}$$

To test null hypothesis

$$H_0 : \psi = 0$$

Use the t statistic

$$t = c / SE_c \quad \text{with} \quad \nu = DFE$$

# Contrasts

Example 1:

To compare the supervisors with the other two groups of workers we are interested in whether the mean SCI score for supervisors is higher than the average of the means for the unskilled workers and the skilled workers. We construct the following null hypothesis $H_0 : \mu_{SU} - \frac{1}{2}(\mu_{UN} + \mu_{SK}) = 0$ . The coefficients in the contrasts are $a_{SU} = 1, a_{UN} = -0.5$ and $a_{SK} = -0.5$. Calculate the contrast statistic and draw a conclusion.

Example 2:

Use contrast statistic and compare the mean SCI scores for the unskilled workers and the skilled workers.

# Two-way ANOVA

Two-way ANOVA compares the means of populations that are classified in two ways or the mean response in two factor experiments.

Malaria is a serious health problem causing an estimated 2.7 million deaths per year, mostly in Africa. Some research suggests that vitamin A can reduce episodes of malaria in young children. Red palm oil is a good source of vitamin A and is readily available in Nigeria, a country where malaria accounts for about 30% of the deaths of young children. To see the effect, a group of children who are 2 to 5 years of age will be divided to take a placebo, a low dose of red palm oil, or a high dose of red palm oil. Because boys and girls may differ in exposure to malaria and the response to the red palm oil supplement, our analysis should also take gender into consideration.

# Advantages of Two-way ANOVA

1. It is more efficient to study two factors simultaneously rather than separately.
2. We can reduce the residual variation in a model by including a second factor thought to influence the response.
3. We can investigate interactions between factors.

# Assumptions for two-way ANOVA

We have independent SRSs of size $n_{ij}$ from each of $I \times J$ normal populations. The population means $\mu_{ij}$ may differ, but all populations have the same standard deviation $\sigma$. The $\mu_{ij}$ and $\sigma$ are unknown parameters.

Let $x_{ijk}$ represent the $k$th observation from the population having factor A at level $i$ and factor $B$ at level $j$. The statistical model is

$$x_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

for $i = 1, \cdots I$ and $j = 1, \cdots J$ and $k = 1, \cdots n_{ij}$, where the observations $\varepsilon_{ijk}$ are from an $N(0, \sigma^2)$ distribution.

# Inference for two-way ANOVA

$$SST = SSA + SSB + SSAB + SSE$$

$$DFT = DFA + DFB + DFAB + DFE$$

| Source | Degrees of freedom | Sum of squares | Mean square | F |
|---|---|---|---|---|
| A | $DFA = I - 1$ | $SSA$ | $MSA = SSA / DFA$ | $MSA / MSE$ |
| B | $DFB = J - 1$ | $SSB$ | $MSB = SSB / DFB$ | $MSB / MSE$ |
| AB | $DFAB = (I-1)(J-1)$ | $SSAB$ | $MSAB = SSAB / DFAB$ | $MSAB / MSE$ |
| Error | $DFE = N - IJ$ | $SSE$ | $MSE = SSE / DFE$ | |
| Total | $DFT = N - 1$ | $SST$ | $SST / DFT$ | |

# Inference for two-way ANOVA

F statistics:

If the effect being test is zero, the calculated F statistic has an F distribution with numerator degrees of freedom corresponding to the effect and denominator degrees of freedom equal to DFE. Large values of the F statistic lead to the rejection of null hypothesis. The p-value is the probability that a random variable having the corresponding F distribution is greater than or equal to the calculated values.

Hypotheses:

There are three null hypotheses in two-way ANOVA, with an F test for each. We can test for significance of the main effect of A, the main effect of B and the AB interaction.