**Lab Two: One-Way and Two-Way Analysis of Variance (ANOVA)**

**INTRODUCTION**

Analysis of Variance (ANOVA) is a parametric technique that compares k categories of data, in order to determine whether or not these sub-sets all derive from the same population. As its name implies, ANOVA is concerned with determining whether the variance between groups is attributable to simple random differences; that is, a comparison of between-groups variance to within-groups variance will suggest whether or not the data derive from a single population.

The stipulations of this test are fairly rigid. The data being examined must be measured on either an interval or a ratio scale. In addition, it assumes that the populations from which the samples are drawn must be normally distributed.

Formulae:

Within sample variance:

$$\sigma_w^2 = \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n}(x - \overline{x})^2}{N - k}$$

k - is the number of sample
N - total number of individuals in all samples
X bar - is the mean

Between sample variance:

$$\sigma_b^2 = \frac{\sum\limits_{i=1}^{k} n(\overline{x} - \overline{x}_G)^2}{k - 1}$$

$\overline{x}_G$    Is the grand mean

The short hand formulae are:

$$SST = (\sum_{i=1}^{r} \sum_{j=1}^{k} x_{ij}^2) - \frac{1}{n} T^2$$

$$T = \sum_{i=1}^{r} \sum_{j=1}^{k} x_{ij}$$

$$SSR = (\sum_{i=1}^{r} \frac{T_i^2}{n_i}) - \frac{1}{n} T^2$$

$T_i$ is the sum of all the observations in a row

SSE=SST-SSR

$$MSR = \frac{SSR}{k-1}$$

$$MSE = \frac{SSE}{N-k}$$

$$F = \frac{MSR}{MSE}$$

$$df_1 = k-1 \quad df_2 = N-k$$

**INSTRUCTIONS**

In a survey of intertidal communities, the following data were collected to describe the distribution of three littoral snails of the genus *Littorina*, which are also known as periwinkles. It is suspected that these snails have distinct habitat preferences, which are related to their ability to withstand varying degrees of exposure to air and water. The intertidal zone was divided into four areas, according to height from the water level at low tide.

Using **SPSS for Windows**, perform a one-way analysis of variance with these data. With the aid of the output that you have generated, answer the following questions.

1a.    State the null and the research hypotheses.
1b.    State the within-groups and between-groups degrees of freedom.
1c.    For each of the 95%, and 99% confidence intervals, state the critical values for the F-ratio.
1d.    Show how you would have derived the following quantities manually: between-groups variance, within-groups variance, and the F-ratio.
1e.    Do we accept or reject the null hypothesis at each level of confidence? If there are differences, explain them.

| Distribution of Littorina spp. In the littoral zone (individuals per square decimetre) | Upper | MidUpper | MidLower | Lower |
|---|---|---|---|---|
| | 43 | 59 | 42 | 52 |
| | 46 | 47 | 42 | 51 |
| | 51 | 44 | 43 | 56 |
| | 49 | 50 | 48 | 48 |
| | 47 | 47 | 44 | 43 |
| Green's Point, NB | | 48 | 46 | 47 |
| | | | 49 | 44 |
| | | | 47 | 43 |
| | | | 47 | |

**Two-Way Analysis of Variance (ANOVA)**

INTRODUCTION

The two-way analysis of variance is applied in situations where there is reason to believe that *two* factors may be affecting the values of observations in a contingency table. The distinct advantage associated with two-way ANOVA is that besides comparing variance between two main effect variables (J and K), it also considers the interaction effect. Thus, three significance tests are required: for Factor J, Factor K, and the interaction between Factors J and K. Although the computations can become confusing, the process is very similar to the one-way ANOVA. Remember: you are doing nothing more than comparing the values of each observation, cell, row, and column to its expected mean value.

Fortunately, SPSS for Windows alleviates much of the tedium that is associated with calculating each step. However, performing a manually calculated two-way ANOVA provides you with a better sense of how the statistic is developed and how the significance tests work.

Formula

$$SSR = \frac{1}{k}\left(\sum_{i=1}^{r} T_i^2\right) - \frac{1}{n}T^2 \qquad\qquad SSC = \frac{1}{r}\left(\sum_{j=1}^{k} T_j^2\right) - \frac{1}{n}T^2$$

$$T_i = \sum_{j=1}^{k} x_{ij} \qquad\qquad T_j = \sum_{i=1}^{r} x_{ij} \qquad\qquad SST = \left(\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}^2\right) - \frac{1}{n}T^2$$

SSE=SST-SSR-SSC

Degrees of Freedom

Total df = N - 1
Column $df_k$ = k - 1
Row $df_r$ = r - 1
Residual $df_{ij}$ = (r - 1) (k - 1)

Mean Sums of Squares

Column MSC = SSC / $df_k$
Row MSR = SSR / $df_r$
Residual MSE = SSE / $df_{ij}$

F - ratio

Test column effect $F_c$ = MSC / MSE
Test row effect $F_r$ = MSR / MSE

**INSTRUCTIONS**
For much of the late 1980s and into the early 1990s, Canada has been occasioned with relatively high interest rates. According to the Bank of Canada, the high interest rate policy is meant to keep inflation under control. However, an interurban comparison of inflation rates across Canada's regions suggests that the spatial manifestation of inflationary pressure restricted to Southern Ontario. This observation has prompted many economists, small business organizations, and 'think tanks' to describe inflationary pressure as a 'Toronto-based phenomena'.

Recently, a representative survey solicited opinion from across the nation about the prevailing interest rate policy. People from five different regions were asked the following question: "Should the Federal Government direct the Bank of Canada to pursue a dual interest rate policy, whereby regions outside of Central Canada had their interest rates determined separately from Central Canada's? The regions responded in the following manner:

| Region | High Income | Medium Income | Low Income |
|---|---|---|---|
| British Columbia | 74 | 90 | 81 |
| Northern Canada | 87 | 84 | 77 |
| Prairie Region | 86 | 91 | 82 |
| Central Canada | 67 | 72 | 64 |
| Atlantic Canada | 78 | 86 | 79 |

<u>TO BE CALCULATED BY HAND</u>
1. State the null and research hypotheses.
2. Using the formulae provided, compute the degrees of freedom, sums of squares, mean squares, and the F-statistic for the column and row effect. Be sure to show all of your work.
3. Do we accept or reject the null hypotheses at the 95% significance level? Briefly explain whether or not there is a significant difference between the five regions of Canada with respect to the interest rate question.
4. Rank the different regions according to the degree to which they depart from the grand mean. What does this exercise suggest?

PART TWO: COMPLETE USING **SPSS for Windows**

The following data chronicle the generation of waste paper products for a sample of three types of households. Compare these, with a two-way ANOVA, and determine whether or not there is a significant relationship between the volume of waste paper generation and either the type of paper product or the household category. Remember to pay particular attention to the significance of the interaction between these two factors. Note that the data are meant to represent the weight in kilograms of three types of paper garbage generated by three different types of households.

|  | Urban Family | Apartment Unit | Rural Family |
|---|---|---|---|
| Newsprint | 25, 15, 24, 29, 26 | 36, 43, 38, 35, 33 | 31, 39, 29, 31, 20 |
| Mixed paper | 15, 10, 16, 13, 18 | 42, 32, 36, 40, 41 | 26, 32, 21, 21, 17 |
| Fine Paper | 39, 37, 37, 37, 37 | 28, 35, 25, 26, 18 | 20, 12, 20, 21, 22 |

The interaction assumes the following division of variance

SST = SSC +SSR +SSI +SSE

a. State the null hypotheses. State the research hypotheses.
b. At the 95% confidence interval, do we accept or reject the null hypotheses? Why or why not?
c. What do you conclude about the relationship between volume of waste paper, waste paper type and household type.
d. What do you conclude from the interaction effect?