Analysis of Variance

EDP 613

Week 13

A Note About The Slides

Currently the equations do not show up properly in Firefox. Other browsers such as Chrome and Safari do work.

Bivariate Tables: Reducing Errors

A **proportional reduction of error** (PRE)

- is a general term for statistical measures that tells us how much we can improve predictive the vale of a dependent variable based on available information on the independent variable.

> - includes variables

- E_1: errors of predictions made when the independent variable is *ignored*
- E_2: errors of predictions made when the independent variable is *included*

> - is found by
>
> $$PRE = \frac{E_1 - E_2}{E_1}$$
>
> - takes on values between (0,1) with
>   - 0: no reduction in error
>   - 1: perfect prediction—the error is completely eliminated.

Measures for Strength of Association

Very General Rule of Thumb

Range	Strength
0 - 0.1	Weak
> 0.1 - 0.4	Moderate
> 0.4 - 1	Strong

PRE: Lambda λ

- λ is used to determine the strength of the relationship between two nominal variables.

- Process

> - Find E_1
>
> - Find E_2
>
> - Calculate PRE which is given by λ in these cases

Gamma γ and Kendall's Tau τ-B

- γ and τ-B are used to determine the strength of the relationship between two ordinal variables.

ANOVAS

- An *analysis of variance* (ANOVA) is a test to find a significant relationship between two variables in two or more groups or samples.
    - *one-way ANOVA*:Used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups.
    - *two-way ANOVA*: Used to compare the mean differences between groups that have been split on two independent variables (called *factors*)

Assumptions 1/2: one-way ANOVA

> - Assumption 1: Dependent variable should be measured at the interval or ratio level (i.e., they are continuous)
> - Examples: time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg)

> - Assumption 2: Independent variable should consist of two or more categorical, independent groups
> - Examples: ethnicity (e.g., 3 groups: Caucasian, African American and Hispanic), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist)

> - Assumption 3: *Independence of observations*, which means that there is no relationship between the observations in each group or between the groups themselves

Assumptions 2/2: one-way ANOVA

> - Assumption 4: No significant outliers.
> - Example: in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which places that individual in top 1% of IQ scores globally

> - Assumption 5: Dependent variable should be approximately normally distributed for each category of the independent variable.

> - Assumption 6: **Homogeneity of variances** where all groups have the same or similar variance.

Formulas to know 1/3

between-group sum of squares (SSB) given by

$$\sum n_k \cdot (\overline{Y}_k - \overline{Y})^2$$

where

- k: number of different samples
- n_k: the number of cases in a sample k
- $\overline{Y}_k$: the mean of a sample k
- $\overline{Y}$: the overall mean

Formulas to know 2/3

within-group sum of squares (SSW) given by

$$\sum (Y_i - \overline{Y}_k)^2$$

where

- k: number of different samples
- $\overline{Y}_k$: the mean of a sample k
- Y_i: each individual score in a sample

Formulas to know 3/3

F ratio obtained or **F statistic** (F) given by

$$\frac{SSB/df_b}{SSW/df_w}$$

where

- the numerator is called the **mean square between**
- the denominator is called the **mean square within**
- $df_b = k - 1$
- $df_w = N - k$
- N: total number of cases

Example

> - A clinical trial is run to compare three (3) weight loss programs and participants are randomly assigned to one of the comparison programs and are counseled on the details of the assigned program.

- low-calorie diet
- low fat diet
- low carbohydrate diet
- control

> - Participants follow the assigned program for 8 weeks.
>
> - The outcome of interest is weight loss, defined as the difference in weight measured at the start of the study (baseline) and weight measured at the end of the study measured in pounds.

Data

Low Calorie	Low Fat	Low Carbohydrate	Control
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

Process 1/5

Step 1: Set up hypotheses and determine level of significance

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4; \alpha = 0.05$$

Step 2: Select the appropriate test statistic

Using the *F*-statistic

Step 3: Set up decision rule

- The appropriate critical value can be found in a table of probabilities for the F distribution (Appendix F)
- In order to determine the critical value of F, we need to calculate the degrees of freedom
  - $df_b = 4 - 1 = 3$
  - $df_w = 20 - 4 = 16$

Process 2/5

Step 4: Compute the test statistic

- $N = 20$
- $\overline{Y} = 3.6$

So

$$\begin{aligned} SSB &= 5(6.6-3.6)^2 + 5(3.0-3.6)^2 + 5(3.4-3.6)^2 + 5(1.2-3.6)^2 \\ &= 75.8 \end{aligned}$$

Process 3/5

And

- low calorie diet:

$$\sum (Y - 6.6)^2 = 21.4$$

- low fat diet:

$$\sum (Y - 3.0)^2 = 10.0$$

- low carbohydrate diet:

$$\sum (Y - 3.4)^2 = 5.4$$

- control:

$$\sum (Y - 1.2)^2 = 10.6$$

So $SSE = 21.4 + 10.0 + 5.4 + 10.6 = 47.4$

Process 4/5

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Means Squared	F
Between treatment	75.8	4-1=3	75.8/3=25.3	25.3/3 = 8.43
Error	47.4	20-4=16	47.4/16	
Total	123.2	20-1=19		

Process 5/5

Step 5: Conclusion

Since $8.43 \geq 3.24$, we reject H_0 and say that *there is a statistically significant difference in the mean weight loss among the four diets*.

That's it. Take a break before our R session!