Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Outline

- Make-up Classes
- Problem Set 3 and 4
- GLMs
- Two-way ANOVA
- Checking Model Assumptions
- Robust Tests
- R Lab
  - Loops
  - Book Correction: Fisher's exact example
  - Chapters 4 and 5

One-way analysis of variance

- The one-way analysis of variance (ANOVA) tests the null hypothesis that samples in two or more groups are drawn from populations with the same mean values. Typically, a sample is partitioned into 2 or more samples by a categorical variable.
- Tests the null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$ by comparing the variability between groups to the variability within groups
- A generalization of the pooled-variance two-sample t-Test, and has the same assumptions!
- independent simple random samples
- the populations are normally distributed
- the pooled population variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_n^2$ are equal
- Produces an F-statistic, the ratio of the sample variance calculated among the means to the sample variance within the samples. If the group means are drawn from populations with the same mean values, then the sample variance between the group means should be lower than the sample variance within the samples. A higher ratio therefore implies that the samples were drawn from populations with different mean values.
- When there are only two means to compare the t-test and the one-way ANOVA are equivalent, and the relation between the F-test and the t –test is given by $F = t^2$.

Two-way analysis of variance

- One-way ANOVA test examines the influence of an independent categorical variable on the mean of one dependent variable
- The two-way analysis of variance is an extension to the one-way ANOVA. There are two independent variables (hence the name two-way).
- Assumptions
1. The errors of populations from which the samples are obtained must be normally distributed
2. Observations for within and between groups must be independent.
3. The variances among populations (for each level or patient group) must be equal (homoscedastic).

- Two-way ANOVA determines the main effect of contributions of each independent variable and also identifies if there is a significant interaction effect between the independent variables
- Hypotheses - there are three sets of hypothesis with the two-way ANOVA.
1. The population means of the first factor are equal. This is like the one-way ANOVA for the row factor.
2. The population means of the second factor are equal. This is like the one-way ANOVA for the column factor.
3. There is no interaction between the two factors. This is similar to performing a test for independence with contingency tables.

Two-way analysis of variance

- Factors - the two independent variables in a two-way ANOVA are called factors. The idea is that there are two variables, factors, which affect the dependent variable. Each factor will have two or more levels within it, and the degrees of freedom for each factor is one less than the number of levels.
- Treatment Groups are formed by making all possible combinations of the two factors. For example, if the first factor has 3 levels and the second factor has 2 levels, then there will be 3x2=6 different treatment groups.
- The main effect involves the independent variables one at a time. The interaction is ignored for this part. Just the rows or just the columns are used, not mixed. This is the part which is similar to the one-way analysis of variance. Each of the variances calculated to analyze the main effects are like the between variances
- The interaction effect is the effect that one factor has on the other factor. The degrees of freedom here is the product of the two degrees of freedom for each factor.
- The Within variation is the sum of squares within each treatment group. You have one less than the sample size (remember all treatment groups must have the same sample size for a two-way ANOVA) for each treatment group.
- The total number of treatment groups is the product of the number of levels for each factor. The within variance is the within variation divided by its degrees of freedom.

Two-way analysis of variance

- There is an F-test for each of the hypotheses, and the F-test is the mean square for each main effect and the interaction effect divided by the within variance. The numerator degrees of freedom come from each effect, and the denominator degrees of freedom is the degrees of freedom for the within variance in each case.

Linear Models in R

One-way ANOVA $Y_{i\ k} = \alpha_i + \varepsilon_{i\ k}$

Two-way ANOVA $Y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

α_i is the mean of Group i indicated by the first factor
β_j is the mean of Group j indicated by the second factor
$(\alpha\beta)_{ij}$ the interaction effect
ϵ_{ijk} the error which is distributed according to $N(0,\sigma^2)$

Checking Model Assumptions

- When the linear model is applied for analysis of variance there are two assumptions made.

1. The errors are assumed to be independent and normally distributed
2. The error variances are assumed to be equal for each level (or patient group). (known as the homoscedasticity assumption.)

- The normality assumption can be tested as a null hypothesis by applying the Shapiro-Wilk test on the residuals.
- The homoscedasticity assumption can be tested as a hypothesis by the Breusch-Pagan test on the residuals.
- The Breusch-Pagan test is a generalization of the F-test for equal variances.

Robust Tests

- In the case when homoscedasticity is violated, we are in a situation quite similar to that of t-testing with unequal variances. The null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$ of equal means can still be tested without assuming equal variances (homoscedasticity) by using the oneway.test function.

- In case the normality is violated, a rank type of test is more appropriate. In particular, to test the null-hypothesis of equal distributions of groups of gene expression values, the Kruskal-Wallis rank sum test is recommended.
- The Kruskal-Wallis test is a generalization of the Wilcoxon test for testing the equality of two distributions. Since it is based on ranking the data, it is highly robust against non-normality