Analysis of Variance (ANOVA)

Chapter 5, Lab 2

OpenIntro Biostatistics

Topics

- ANOVA F-test
- Adjustment for multiple comparisons

The previous lab introduced the two-group independent t-test as a method for comparing the means of two groups. In some settings, it is useful to compare the means across more than two groups. The methodology behind a two-group independent t-test can be generalized to a procedure called **analysis of variance (ANOVA)**. Assessing whether the means across several groups are equal by conducting a single hypothesis test rather than multiple two-sample tests is important for controlling the overall Type I error rate.

The material in this lab corresponds to Section 5.5 of *OpenIntro Biostatistics*.

FAMuSS: comparing change in non-dominant arm strength by *ACTN3* genotype

Is change in non-dominant arm strength after resistance training associated with genotype?

In the Functional polymorphisms Associated with Human Muscle Size and Strength study (FAMuSS), researchers examined the relationship between muscle strength and genotype at a particular location on the *ACTN3* gene. The `famuss` dataset in the `oibiostat` package contains a subset of data from the study.

The percent change in non-dominant arm strength, comparing strength after resistance training to before training, is stored as `ndrm.ch`. There are three possible genotypes (CC, CT, TT) at the *r577x* position on the *ACTN3* gene; genotype is stored as `actn3.r557x`.

1. Load the data. Create a plot that shows the association between change in non-dominant arm strength and *ACTN3* genotype. Describe what you see.
2. Check whether the assumptions for conducting an ANOVA are reasonably satisfied: 1) observations are independent within and across groups, 2) the data within each group are nearly normal, and 3) the variability across the groups is about equal.
3. Conduct a hypothesis test to address the question of interest. Let $\alpha = 0.05$.
   a) Let the parameters μ_{CC}, μ_{CT}, and μ_{TT} represent the population mean change in non-dominant arm strength for individuals of the corresponding genotype. State the null and alternative hypotheses.
   b) Use `summary(aov())` to compute the F-statistic and p-value. Interpret the p-value.
   c) Complete the analysis using pairwise comparisons.
      i. What is the appropriate significance level $\alpha^\star$ for the individual comparisons, as per the Bonferroni correction?

ii. Use `pairwise.t.test()` to conduct the pairwise two-sample t-tests.

iii. Summarize the results.

NHANES: factors associated with poverty

This section uses data from the National Health and Nutrition Examination Survey (NHANES), a survey conducted annually by the US Centers for Disease Control (CDC).[1] The dataset `nhanes.samp.adult.500` contains data for 500 participants ages 21 years or older that were randomly sampled from the complete NHANES dataset that contains 10,000 observations.

The variable `Poverty` is a ratio of family income to poverty guidelines. Smaller numbers indicate more poverty, where a value below 1 indicates that the participant is in a family with income below the poverty threshold. Ratios above 5 were recorded as 5.

4. Is there evidence of an association between poverty and BMI category? People with lower incomes tend to have less access to healthy food, and an increased risk of obesity.

   The variable `BMI_WHO` contains BMI data, grouped into the categories defined by the World Health Organization: 12 - 18.5 (underweight), 18.5 - 24.9 (normal), 25 - 29.9 (overweight), and 30+ (obese).

   a) Create a plot that shows the association between poverty ratio and BMI category. Check the assumptions for conducting an ANOVA.

   b) Among individuals with BMI of at least 18.5, is there evidence of an association between poverty and BMI category?

5. Is there evidence of an association between poverty and educational level? It is reasonable to expect that on average, individuals with a higher educational level will also have a higher income.

   The variable `Education` records the highest level of education obtained: 8^{th} grade, 9^{th} - 11^{th} grade, high school, some college, or college degree.

   a)

[1] The dataset was first introduced in CHapter 1, Lab 1 (Introduction to Data).