STATS 8: Introduction to Biostatistics

Analysis of Variance

Babak Shahbaba
UCI, Spring of 2012

Introduction

- We discuss Analysis of Variance (ANOVA) models that generalize the t-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories.

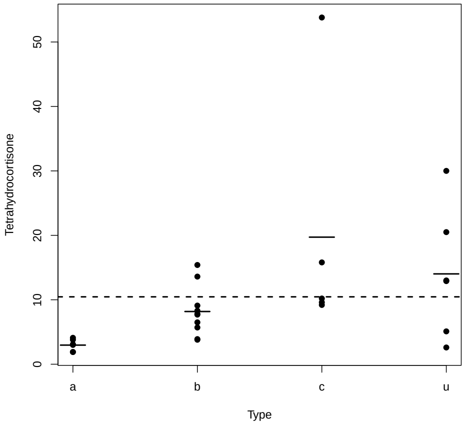
- The categorical variable is called the **factor** and is typically considered as the explanatory variable.

- In contrast, the numerical variable, whose means across different groups are compared, is regarded as the response variable.

- e mainly focus on ANOVA models with only one factor; These models are known as **one-way ANOVA**.

Example

- As an example, we analyze the `Cushings` data set, which is available from the `MASS` package.



Between-groups vs. within-groups variations

- Across the four groups, there appears to be considerable variation in the group means (i.e., deviations of the small solid lines from the dashed line), SS_B

- Likewise, within groups, there are different degrees of variation of the observations from their specific mean (i.e., variation of points around the corresponding small horizontal line), SS_W

- Both sources of variation contribute to the total variation of the observations around the overall mean (dashed line).

$$SS = SS_B + SS_W.$$

Hypothesis testing

- Let us denote the overall population mean of Y as μ and group-specific population means as $\mu_1, \ldots, \mu_4$.

- We want to evaluate the null hypothesis,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu,$$

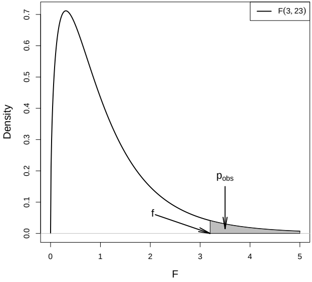
- For this, we use the following test statistic

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)},$$

where n is the total sample size, and k is the number of groups.

Hypothesis testing

- The F-statistic has $F(df_1 = k-1,\ df_2 = n-k)$ distribution under the null hypothesis.
- For the above example, the degrees of freedom parameters are $df_1 = 4-1 = 3$ and $df_2 = 27-4 = 23$.
- The observed value of F is $f = 3.2$.



The assumptions of ANOVA

- To use ANOVA models, we assume that the samples are selected randomly from the population and independently from each other (e.g., by using simple random sampling).

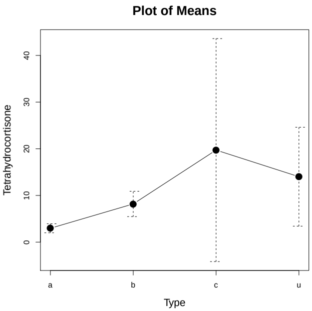
- Further, we assume that the response variable in each group has a normal distribution.

- While the means of these normal distributions can change from one group to another, we assume that they all have the same variance.

The assumptions of ANOVA

- Violation of these assumption could lead to wrong inference.

- For the example discussed above, the constant variance assumption does not seem reasonable.



The assumptions of ANOVA

- Sometimes, we can stabilize the variance (i.e., making it approximately constant) by using simple data transformations such as log or square root.