UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Statistics Workshop: Part I

Bei Jiang

Department of Mathematical and Statistical Sciences
University of Alberta

May 12, 2016

Outline

Overview of ANOVA

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- The analysis of variance (ANOVA) is a method to compare average (mean) responses to experimental manipulations in controlled environments.
- A one-way layout consists of a single factor with several levels and multiple observations at each level. For example, subjects are randomly assigned to either drug or placebo group (two levels of the treatment).

Statistical Hypothesis Testing for one-way ANOVA

- Step 1: State the Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ for k levels of an experimental treatment.
- Step 2: State the Alternative Hypothesis: H_1 : treatment level means not all equal.
  - Many people make the mistake of stating the Alternative Hypothesis as: $H_1 : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k$ which says that every mean differs from every other mean. This is only one of many possibilities.
  - A simpler way of thinking about the alternative of H_0 is that at least one mean is different from all others.

Statistical Hypothesis Testing for one-way ANOVA

- Step 3: Set significance level α: If we look at what can happen, we can construct the following contingency table:

	In Reality	
Decision	H_0 is TRUE	H_0 is FALSE
Accept H_0	OK	Type II Error β = probability of Type II Error
Reject H_0	Type I Error α = probability of Type I Error	OK

The typical value of α is 0.05, establishing a 95% confidence level.

Statistical Hypothesis Testing for one-way ANOVA

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Step 4: Collect Data and Determine the Sample size (R example later!)
- Step 5: Calculate a test statistic: for three or more categorical treatment level means, we use an F statistic. For two treatment levels, we use student's T test.
  - A test statistic is considered as a numerical summary of a data set, reducing the data to one single value.
  - A null hypothesis is typically specified in terms of a test statistic.
  - The sampling distribution under H_0 must be calculable, which allows p-values (a probability) to be calculated.

Statistical Hypothesis Testing for one-way ANOVA

- Step 6: Construct Acceptance / Rejection regions: a critical value (threshold), denoted by F_α, is established, which is the minimum value for the test statistic for us to be able to reject the null. For F test, the location of Acceptance / Rejection regions are shown in the graph below:

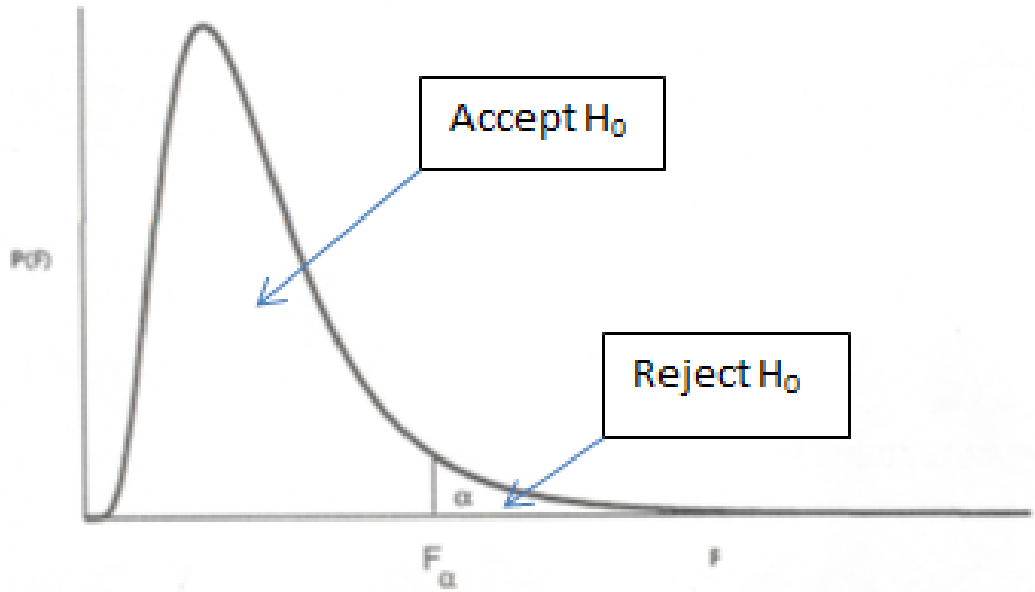


Figure K.1: The F distribution

Note: rejection regions for T-test will depend on whether H_1 is one sided or two sided: $H_1 : \mu_1 \neq \mu_2$ or $H_1 : \mu_1 \geq (\text{or} <)\mu_2$

Statistical Hypothesis Testing for one-way ANOVA

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Step 7: Based on steps 5 and 6, draw a conclusion about H_0.
  - If the calculated F-test statistic from the data is larger than the F_α, then you are in the Rejection region and you can reject the Null Hypothesis with $(1-\alpha)$ level of confidence.
  - Note that modern statistical software condenses step 6 and 7 by providing a p-value. The p-value here is the probability of getting a calculated F test statistic even greater than what you observe.
  - If the p-value obtained from the ANOVA is less than α, then we have evidence to reject H_0 and accept H_1.

The ANOVA foundations

Each observation can be denoted by Y_{ij}, referenced by two subscripts: i refers to the i^{th} level of the treatment (we have k levels), within each level i, j refers to the j^{th} observations (for balanced design $j = 1, \cdots, n$).

- Grand Mean: $\bar{Y}_{..}$, an average over all j observations in all i treatment levels.
- Treatment Mean: $\bar{Y}_{i.}$, an average over all j observations in each of the i treatment levels.

ANOVA is testing the effect of the treatment effect (equal means = no effect); it is also an analysis of variance (why?).

The one-way ANOVA foundations

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- In Statistics, we consider ANOVA testing as partitioning of the total variability into two components 1) variability between groups due to treatment effect and 2) variability within the group (hope this random variation is small?).
- Obviously we are very interested in the variability between groups because this means the treatment makes a difference!

$Y_{ij} - \bar{Y}_{..}$	=	$\bar{Y}_{i.} - \bar{Y}_{..}$	+	$Y_{ij} - \bar{Y}_{i.}$
Total deviation		Deviation of estimated factor level mean around overall mean		Deviation around estimated factor level mean

one-way ANOVA table and F-test

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- ANOVA Table

Source	Degrees of Freedom (df)	Sum of Squares	Mean Squares
treatment	$k-1$	$SST = \sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$MST = SST/df$
residual	$N-k$	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$	$MSE = SSE/df$
total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	

- The F statistic for $H_0 : \mu_1 = \cdots = \mu_k$ is

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2/(N-k)} = \frac{MST}{MSE},$$

which has an F distribution with dfs $k-1$ and $N-k$.

The one-way ANOVA model

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- The model: $Y_{ij} = \mu_i + \epsilon_{ij}$, where ϵ_{ij} i.i.d. $\mathrm{N}(0, \sigma^2)$.
  - Model assumption: Y_{ij}'s are assumed to be normally distributed; independent; and have equal variance among treatment levels (homogeneity)
  - Quick check: within each treatment group, QQ Plot of Y and residual plots of $\hat{\epsilon}$ after model fitting; the spread of residuals is roughly equal per treatment level (rule of thumb $s^2_{max}/s^2_{min} < 3$)
  - Formally, Shapiro Test of Normality and Bartlett Test of Homogeneity of Variances.
  - Box-cox Transformation of Y may help: $\log(Y)$, $1/Y$, Y^2, etc.
  - Outlier detection: fit the ANOVA model with or without potential outliers and report both sets of results if they are dramatically different. Another approach is to use a robust version of ANOVA procedure (assuming T- instead of Normal distribution for Y). In general regression settings, methods include leverage, Cook's distance, etc (more detail later).

Questions should be answered!

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- How often do I need to do experiment X before I can start to analyze the results?
- How do I treat my results if I have different numbers of measurements for different groups?
- How do I recognize an outlier, and when is it OK to remove one?
- When is it OK to use student's T test, and what are the alternatives?

Questions should be answered!

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

One-way ANOVA, T-test, Power Analysis, Diagnostic: R examples

Two-way ANOVA foundations

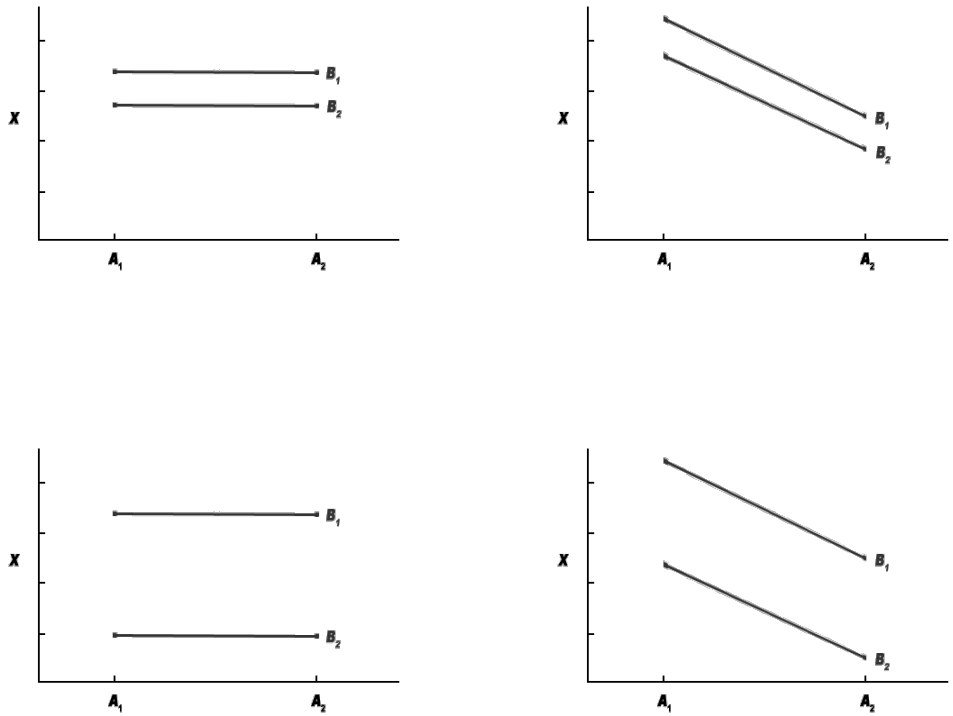
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Two-way ANOVA consists of assigning subjects into combinations of experiment conditions under more than one factors, say, two levels of Factor A and Factor B.

$\bar{Y}_{ij.} - \bar{Y}_{...}$	=	$\bar{Y}_{i..} - \bar{Y}_{..}$	+	$\bar{Y}_{.j.} - \bar{Y}_{...}$	+	$\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$
Deviation estimated treatment mean around overall mean		*A* main effect		*B* main effect		*AB* interaction effect

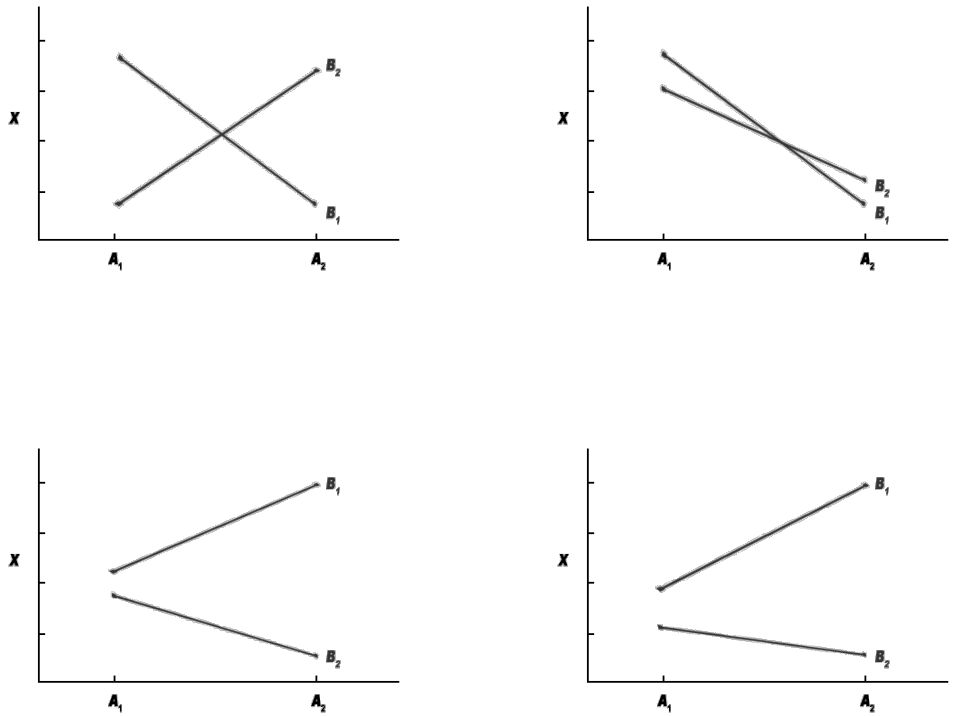
- Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \epsilon_{ijk}$, where Y_{ijk} represents the k^{th} observation under i^{th} level of Factor A and j^{th} level of Factor B, α_i is the main effect of Factor A, β_j is the main effect of Factor B, and τ_{ij} is the interaction effect.
- What is a main effect? No main effect means the effects at different levels of a Factor are the same.
- What is an interaction? It can be interpreted as: the effect of one factor is not the same at different levels of another factor (Examples next slide).

Two-way ANOVA: interaction effect



- top left: no effect of Factor A, a small effect of Factor B, and no interaction between Factor A and Factor B.
- bottom left: no effect of Factor A, a large effect of Factor B, and no interaction.
- top right: a large effect of Factor A small effect of Factor B, and no interaction.
- bottom right: a large effect of Factor A, a large effect of Factor B and no interaction.

Two-way ANOVA: interaction effect



- top left: no effect of Factor A, no effect of Factor B but an interaction between A and B.
- bottom left: No effect of Factor A, a large effect of Factor B, with a very large interaction.
- top right: Large effect of Factor A, no effect of Factor B with a slight interaction.
- bottom right: An effect of Factor A, a large effect of Factor B with a large interaction.

Two-way ANOVE and Repeated Measurements R examples

Survival Analysis: Introduction

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Typically focuses on **time to event** data, e.g.,

- time to death
- time to onset (or relapse) of a disease
- length of stay in a hospital
- time to finish a doctoral dissertation!

Some notations:

- Let T denote the failure/survival time, then $T \geq 0$. T can either be discrete, e.g., age or continuous.
- T may not be always observed due to **ensoring**. Let C denote the censoring time, then we only observe either C or T.
  - does not experience the event before the study ends (uninformative censoring)
  - lost to follow-up during the study period not due to health related problem (uninformative censoring)
  - drops out of the study because he/she got much sicker and/or had to discontinue taking the study treatment (informative censoring)
- Most analysis methods to be valid, the censoring mechanism must be uninformative, i.e., independent of the survival outcome

Survival Analysis: Introduction

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Next, we show R examples:

- describe survival data
- compare survival of several groups
- explain survival with covariates
- design studies with survival endpoints

References

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

some materials from:

- http://www.ats.ucla.edu/stat/r
- https://onlinecourses.science.psu.edu/statprogram/programs