Methods of Data Analysis – STA302 Course Notes

Yuchen Wang

May 14, 2019

Contents

1 May 7th - Lecture 1

Definition 1.1 - Statistical Analysis Data Analysis that relies on Probability theory to account for the variability of the data.

Permutation Test 1.2 Insert random premise, observe two samples Group A and Group B.
If the groups have no effect, all of the permutations are equally likely.
We can plot the Permutation Distribution with respect to difference between sample means.

Characteristics of Permutation Test 1.3

1. Involves simple probability theory
2. distribution-free
3. listing all the permutation for large dataset is almost impossible

Definition 1.4 - Statistical Significance We say a difference is **statistically significant** if it's less probable than our pre-determined significance level. (when p-value $p <$ significance level α)

Definition 1.5 - Significant Effect We say the groups have a **significant effect** if it causes the variable of interest to be significantly different.

2 May 9th - Lecture 2

2.1 The basics

Fact 2.1.1 If H_0 is true, the p-value $\sim U(0,1)$

remarks: Hard to prove, just take it.

Tradeoff Between Type I and Type II Error It's common to fix α (significance level or type-I error) and minimize type-II error.

2.2 One-way Analysis of variance (ANOVA)

One-way ANOVA is an extension of the t-test to 3 or more samples focus analysis on group differences.
If groups are different, we expect there is a bigger difference between groups (reflecting the group effect) than within groups (natural variability of the data).

Basic Definitions Suppose we have T groups and n_t observations for the t-th group, and we denote each observation as y.

1. <u>SST</u>: This is the sum of the squared deviations between each observation and the overall mean:

$$SST = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y})^2$$

2. <u>SSE</u>: This is the sum of the squared deviations between each observation and the mean of the group to which it belongs:

$$SSE = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y}_t)^2$$

3. <u>SSG</u>: This is the sum of the squared deviations between each group mean and the overall mean:

$$SSG = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (\bar{y}_t - \bar{y})^2$$

Sum of Squares Decomposition Total sum of squares = Within group sum of squares + Between group sum of squares

$$\sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y})^2 = \sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y}_t)^2 + \sum_{t=1}^{T} \sum_{i=1}^{n_t} (\bar{y}_t - \bar{y})^2$$

In shorthand:

$$SST = SSE + SSG$$

<u>*proof:*</u>
add $-\bar{y}_t + \bar{y}_t$ inside the squared error term and everything is just like a short proof in STA261, nothing interesting. ■

ANOVA We want to assess how large is SSG relative to SSE, but it would be hard to establish a distribution for SSG/SSE. Knowing a sum of squares divided by its degrees of freedom has a chi-square distribution, we can conclude that

$$SSG/(T-1) \sim \chi^2_{T-1}, \quad SSE/(n-T) \sim \chi^2_{n-T}$$

Theorem 2.2.1 $\frac{SSG/(T-1)}{SSE/(n-T)} \sim F_{T-1,n-T}$ if $SSG/(T-1)$ and $SSE/(n-T)$ are equal.

<u>*proof:*</u>
In STA261, we've proven that if $\sigma_x^2 = \sigma_y^2$, then $\frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \sim F_{n-1,n-1}$.
Since SSG/(T-1) is an estimation for the variation between groups (σ_T) and $SSE/(n-T)$ is an estimation for the variation within groups (σ_ε), then the result follows. ■

Remarks 2.2.2 Thus a small p-value indicates theses variances are different, which is evidence for the existence of some group effect.

Theorem 2.2.3 One-way ANOVA Table if p-value $< \alpha$, we reject H_0: groups have no effect.

Source	Sum of Squares	df	Mean Squares	Test Statistic
Between	SSG	$T-1$	$MSG = \frac{SSG}{T-1}$	$F = \frac{MSG}{MSE}$
Within	$n-T$	$MSE = \frac{SSE}{n-T}$		
Total	SST	$n-1$		

the Effect Model

$$y_{i,t} = \mu + \tau_t + \varepsilon_{i,t}$$

where $\varepsilon \sim N(0, \sigma^2)$.

1. μ: global mean
2. τ_t: the effect of the tth treatment with $\sum_{t=1}^{T} \tau_t = 0$
3. ε: errors representing the natural variability in real-life data