AID-521 Mathematics for Data Science

Module: Statistics | Lecture: 6

HYP. TESTING FOR MULTIPLE SAMPLES

testing two samples, analysis of variance (ANOVA)

Hypothesis Test for Two Samples

- Comparing the means and variances of two populations

- Let $X_{11}, \ldots, X_{1n_1}$ be a random sample from population 1 with mean μ_1 and variance σ_1^2, and $X_{21}, \ldots, X_{2n_2}$ be a random sample from population 2 with mean μ_2 and variance σ_2^2.

- Here, we study **for the case when** samples are independent and $n_1, n_2 \geq 30$.

Hypothesis Test for Two Samples

SUMMARY OF HYPOTHESIS TEST FOR $\mu_1 - \mu_2$ FOR LARGE SAMPLES (n_1 & $n_2 \geq 30$)

To test

$$H_0 : \mu_1 - \mu_2 = D_0$$

versus

$$H_a : \begin{cases} \mu_1 - \mu_2 > D_0, & \text{upper tailed test} \\ \mu_1 - \mu_2 < D_0, & \text{lower tailed test} \\ \mu_1 - \mu_2 \neq D_0, & \text{two-tailed test.} \end{cases}$$

The test statistic is

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Replace σ_i by S_i, if $\sigma_i, i = 1,2$ are not known.
Rejection region is

$$RR : \begin{cases} z > z_\alpha, & \text{upper tail RR} \\ z < -z_\alpha, & \text{lower tail RR} \\ |z| > z_{\alpha/2}, & \text{two tail RR,} \end{cases}$$

where z is the observed test statistic given by

$$z = \frac{\overline{x}_1 - \overline{x}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Introduction to ANOVA

- → Tests to analyze data from more than two populations.

- → The hypothesis that the population means are equal is considered equivalent to the hypothesis that there is no difference in treatment effects (as in experiments).

- → Can be considered as an extension of the test of hypothesis for the equality of two means.

Introduction to ANOVA

- Assume 4 populations. **Why do we need a new method** to test for differences among these 4 population means?
- Can't we use z- or t-tests for all possible pairs and test for differences in each pair?
  - If any one of these tests leads to the rejection of the hypothesis of equal means, then we might conclude that at least two of the four population means differ.
- Actual Type I error becomes amplified than what we might think!
  - For $\binom{4}{2} = 6$ tests, let $\alpha = 0.10$ be the significance level.
  - Probability that at least one of the six tests leads to the conclusion that there is a difference leads to an error $1-(0.9)^6 = 0.46856$.
  - Hence, one is likely to declare significance when there is none.

Introduction to ANOVA -- Common Terms Used

$\rightarrow$ $Total\ SS$ $=$ total sums of squares of values

$\rightarrow$ SST $=$ sum of squares for treatment

$\rightarrow$ SSE $=$ $Total\ SS - SST$ $=$ sum of squares of errors

$\rightarrow$ MSE $=$ $\frac{SSE}{N-k}$ $=$ mean square error

$\rightarrow$ MST $=$ $\frac{SSE}{N-k}$ $=$ mean square treatment

Int. to ANOVA -- Important Result in Use

If χ_1^2 and χ_2^2 are independent, and have ν_1 and ν_2 d.o.f.s respectively, then

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

has a F-distribution with ν_1 numerator d.o.f. and ν_2 denominator d.o.f.

ANOVA for Two Treatments

→ The simplest form of the analysis of variance procedure, the case of studying the means of two populations I and II.

→ For comparing only two means, the ANOVA will result in the same conclusions as the t-test for independent random samples.

→ This section will help to introduce the concept of ANOVA in simpler terms.

ANOVA for Two Treatments

ANALYSIS OF VARIANCE PROCEDURE FOR TWO TREATMENTS

For equal sample sizes $n = n_1 = n_2$, assume $\sigma_1^2 = \sigma_2^2$.

We test

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_a : \mu_1 \neq \mu_2.$$

1. Calculate: $\overline{y_1}, \overline{y_2}, \sum_{ij} y_{ij}^2, \sum_{ij} y_{ij}$, and find

$$SST = \sum_{i=1}^{2} n_i \left(\overline{y}_i - \overline{y}\right)^2.$$

Also calculate

$$Total\ SS = \sum_i \sum_j y_{ij}^2 - \frac{\left(\sum_i \sum_j y_{ij}\right)^2}{n_1 + n_2}.$$

Then

$$SSE = Total\ SS - SST.$$

ANOVA for Two Treatments

2. Compute

$$MST = \frac{SST}{1}$$

$$MSE = \frac{SSE}{n_1 + n_2 - 2}.$$

3. Compute the test statistic,

$$F = \frac{MST}{MSE}.$$

4. For a given α, find the rejection region as

$$RR : F > F_\alpha,$$

based on 1 numerator and $(n_1 + n_2 - 2)$ denominator degrees of freedom.

5. **Conclusion:** If the test statistic F falls in the rejection region, conclude that the sample evidence supports the alternative hypothesis that the means are indeed different for the two treatments.

Assumptions: Populations are normal with equal but unknown variances.

ANOVA for More Than Two Treatments

- Hypothesis testing problem of comparing population means of more than two independent populations
- Data are about several independent groups
- Let $\mu_1, ..., \mu_k$ be the means of *k* normal populations with unknown but equal variance σ^2.
- Are the means of these groups are different, or are all equal?
- Overall variability: (1) between-groups, (2) within-groups
- If between groups is much larger than that within groups, this will indicate that differences between the groups are real, not merely due to the random nature of sampling.

ANOVA for More Than Two Treatments

- Let independent samples be drawn of sizes $n_i,\ i = 1, 2, ..., k$, and let $N = n_1 + \cdots + n_k$.
- Let y_{ij} be the measured response on the jth experimental unit in the ith sample. That is, Y_{ij} is the jth observation from population i, $i = 1, 2, ..., k$, and $j = 1, 2, ..., n_i$.
- Let $\bar{y}$ be the overall mean of all observations.
- The problem can be formulated as a hypothesis testing problem, where we need to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{vs.} \quad H_1 : \text{ Not all } \mu_i\text{s are equal.}$$

ANOVA for More Than Two Treatments

1. Compute

$$T_i = \sum_{j=1}^{n_i} y_{ij}, T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}, \text{ and } \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2,$$

$$CM = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N} = \frac{T^2}{N}, \text{ where } N = \sum_{i=1}^{k} n_i,$$

$$\overline{T}_i = \frac{T_i}{n_i},$$

and

$$\text{Total } SS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - CM.$$

2. Compute the sum of squares between samples (treatments),

$$SST = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - CM$$

$$= \sum_{i=1}^{k} \overline{T}_i - CM.$$

and the sum of squares within samples,

$$SSE = \text{Total } SS - SST$$

ANOVA for More Than Two Treatments

Let

$$MST = \frac{SST}{k-1},$$

and

$$MSE = \frac{SSE}{n-k}.$$

3. Compute the test statistic:

$$F = \frac{MST}{MSE}.$$

4. For a given α, find the rejection region as

$$RR: \; F > F_\alpha$$

with $\upsilon_1 = (k-1)$ numerator degrees of freedom and $\upsilon_2 = \left(\sum_{i=1}^{k} n_i\right) - k = N - k$ denominator degrees of freedom, where $N = \sum_{i=1}^{k} n_i$.

5. **Conclusion:** If the test statistic F falls in the rejection region, conclude that the sample evidence supports the alternative hypothesis that the means are indeed different for the k treatments and are not all equal.

Assumptions: The samples are randomly selected from the k populations in an independent manner. The populations are assumed to be normally distributed with equal variances σ^2 and means $\mu_1, \ldots, \mu_k$.

ANOVA Table

Table 13.3 ANOVA table for a one-way layout

Source	df	SS	MS	F
Treatments	$k-1$	SST	$\text{MST} = \dfrac{\text{SST}}{k-1}$	$\dfrac{\text{MST}}{\text{MSE}}$
Error	$n-k$	SSE	$\text{MSE} = \dfrac{\text{SSE}}{n-k}$	
Total	$n-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2$		