Contents

Day 21

ANOVA → Analysis of VAriance

Why ANOVA?

We want to compare the means of more than two populations and our previous methods of analysis have been limiting.

Example: we want to compare three samples to see if a difference exists somewhere between those groups. Each of them have their own sampling distributions and other characteristics.

Question to ask yourself: Do all three of these means come from a common population.

This type of test takes in n amount of distributions and then distributes them along a unifying curve. We can then see where those distributions lie on the bigger distribution.

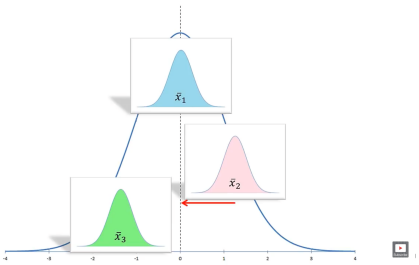

Figure 1: ANOVA Distribution Close to Mean

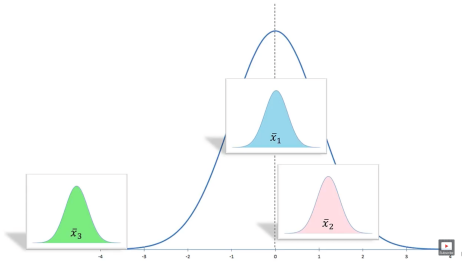

Figure 2: ANOVA Distribution Far from Mean

Mean is in a different location relative to the overall mean of the populations. This might conclude that it is not apart of the common population.

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$

$\implies$ All the samples come from the same population. We are not asking if they are EXACTLY equal. We are asking if each mean likely came from the larger overall population. Variability AMONG/BETWEEN the sample means ($\bar{x}_\text{i}$)

Comparing Analysis Methodology

The problem with running the individual tests on each respective sampling distribution, we run into some issues:

- All three t-tests have $\alpha = 0.05 \therefore$ the all compound
  - This makes the confidence of decrease to $0.95 \times 0.95 \times 0.95 = 0.857$
  - Recall that this is the confidence level or β
- Our new α value need to reflect as such: $\alpha = 1 - 0.857 = 0.143$

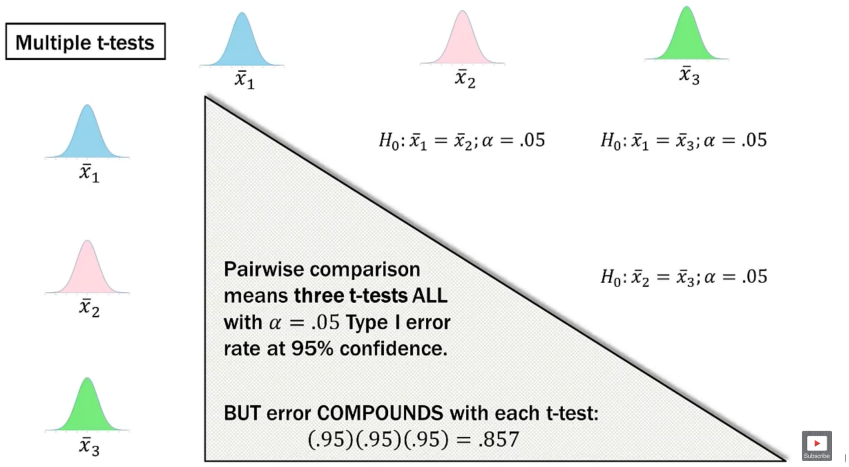

Figure 3: Multiple t-Test Chart

This is why we do not do this, the error rate goes ↑.

Variability Ratio

ANOVA: Analysis of Variance is *variability* ratio.

This can be seen in the following figure:

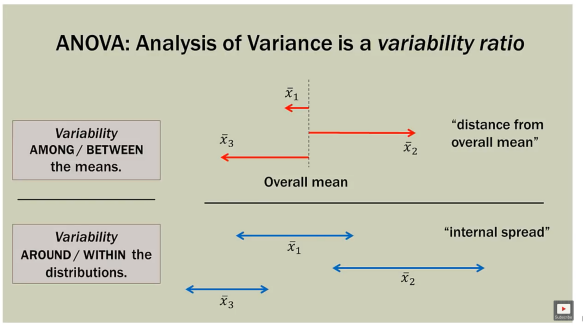

Figure 4: Variability Ratio Visual Example

$$\frac{\text{Variance Between}}{\text{Variance Within}}$$

THE classic Fisher test.

Model: $y \sim N(\mu, \sigma)$

- μ and σ are unknown.
- μ may not be the same for all data points
- σ is assumed same for all data points

In one-way ANOVA:

- Data: one numerical response variable `y` and one categorical explanatory variable whose values are the "groups".
- We need ≥ 2 groups

Example situations where we use it:

- Compare control group to > 1 treatment group
- Observational study comparing 3 or more groups/populations

Use cases

- Between-group effects: variation due to changes in μ
- Within-group effects: variation due to individual differences

Notation

- $\bar{y}$ = grand mean or the mean of all data in the whole sample.
- N = total sample size
- I = number of groups
- $\bar{y}_{\mathtt{i}}$ = sample mean in group `i`
- $s_{\mathtt{i}}$ = sample standard deviation in group `i`
- $n_{\mathtt{i}}$ = sample size in group `i`
- $y_{\mathtt{ij}}$ = value of y for the $\mathtt{j}^{\text{th}}$ case in group `i`.

Hypothesis Testing

H_0: $\mu_1 = \mu_2 = ... = \mu_I$

- All the population means are equal $\implies$ no effect of group on response.
- Under H_0, $y_{ij} \sim N(\mu, \sigma)$
- Also: $\bar{y}_i \sim N(\mu, \frac{\sigma}{\sqrt{n_i}})$
- μ, σ are fixed but unknown

H_a: not H_0 [not really necessary because this is Fisher framework]

- $\implies$ effect of group on response

Implicit Assumptions of Model

- Normal population distribution
- σ is the same for all groups [not as critical]
  - robust to violations of this assumption as long as the largest $s_i < 2 \times$ smallest s_i

Under H_0:

- $F\ simF(DFG, DFE)$
- Think of `F` like χ^2
- t and F are linked
- $N(0,1)$ and $\chi^2 arelinked$
- If p-value $\leq$ significance level, reject H_0 and claim the population means are not all equal
- Do posthoc procedures to figure out which measure different & how different they are.
- If p-value $>$ significance level, we fail to reject the hypothesis we did not prove a difference /an effect.

ANOVA tests **CANNOT** determine/make conclusions about all populations means ($\forall$), only at least one element in the set ($\mu \in \forall$)

Suppose a horticulturist measures the aboveground height growth rate of four different ornamental shrub species grown in a greenhouse. The shrubs were grown from a random sample of seeds, and they were all grown in the same soil mixture and in the same size pot. To ensure that any slight differences in the environmental conditions throughout the greenhouse are not confounded with species, she randomizes the location of the pots throughout the greenhouse. The table contains a summary of her data.

Population	Population description	Sample size	Sample mean	Sample standard deviation
1	Species 1	$n_1 = 20$	$\bar{x}_1 = 13.749$ cm/year	$s_1 = 2.160$ cm/year
2	Species 2	$n_2 = 20$	$\bar{x}_2 = 16.619$ cm/year	$s_2 = 4.284$ cm/year
3	Species 3	$n_3 = 20$	$\bar{x}_3 = 13.608$ cm/year	$s_3 = 3.396$ cm/year
4	Species 4	$n_4 = 20$	$\bar{x}_4 = 12.769$ cm/year	$s_4 = 2.683$ cm/year

The growth rate distributions of each sample are approximately normal, and the data do not contain outliers. The horticulturist uses a one-way analysis of variance (ANOVA) at a significance level of $\alpha = 0.01$ to test if the mean growth rates of all four species are equal. Her results are shown in the table.

Source of variation	SS	df	MS	f	P-value	f-critical
Between groups	169.057	3	56.352	5.398	0.002	4.050
Within groups	793.330	76	10.439			
Total	962.386	79				

Figure 5: ANOVA Example Problem

Homework Example

- $k = 4$
- $n = 20$
- $\alpha = 0.01$
- P-Value $= 0.002$
- H_0: the shrubs are in the same species ($\mu_1 = \mu_2 = \mu_3 = \mu_4$)
- H_1: the shrubs are not in the same species ($\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$)

The decision to reject the null hypothesis at a significance level of $\alpha = 0.01$. There is sufficient evidence to conclude that at least one of the population means different from at least one other population mean.

Computing an F-Statistic

$df_1 = \text{Between} \implies k - 1$

$df_2 = \text{Within} \implies n - k$

- n: total number of elements in the sample
- k: number of populations

$$F = \frac{\text{Between}}{\text{Within}}$$

Example [Made Up]

Four groups:

1. $n_1 = 15$
2. $n_2 = 20$
3. $n_3 = 15$
4. $n_4 = 25$

↑ F-Stat, ↓ p-value

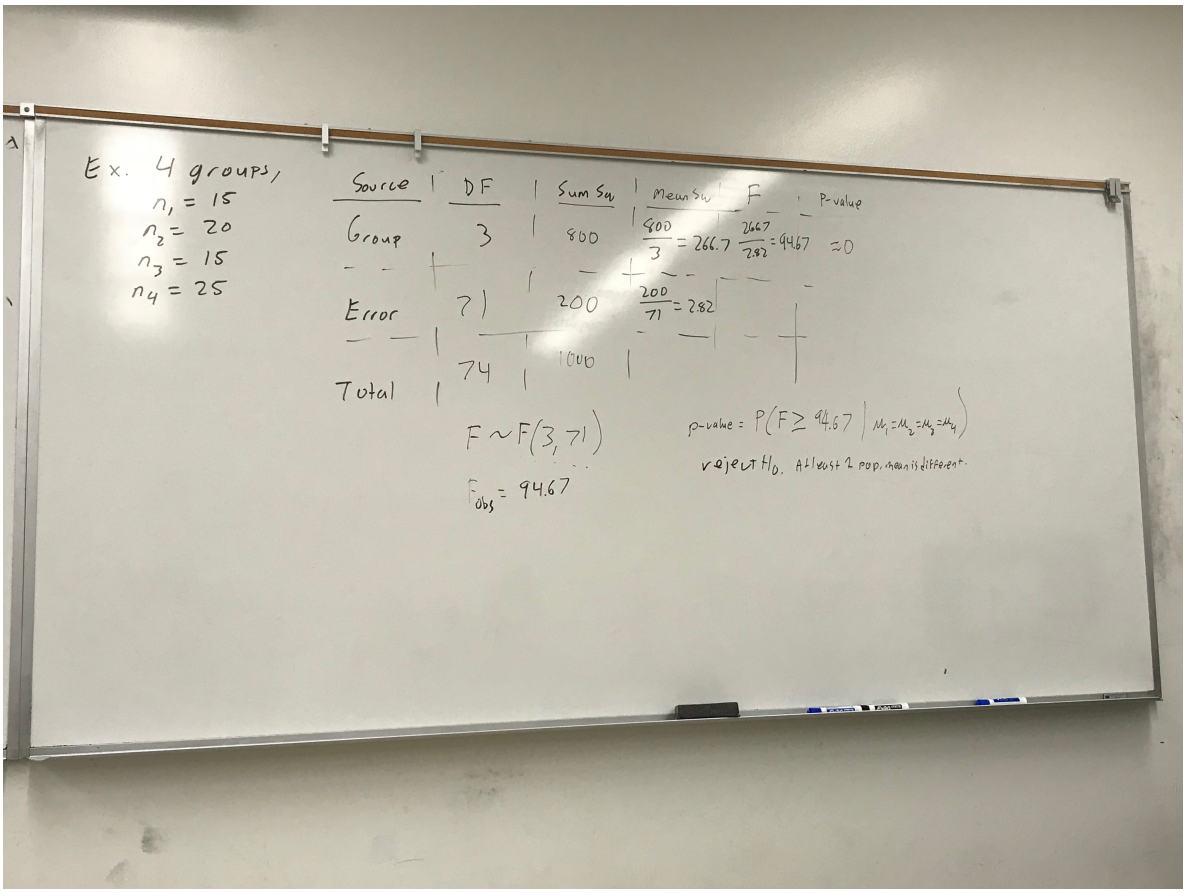

Figure 6: Example Worked in Class

External Links

- ANOVA YouTube Lecture