4 Analysis of Variance (ANOVA)

5 ANOVA

5.1 Introduction

ANOVA

ANOVA is a way to estimate and test the means of multiple populations. We will start with one-way ANOVA.

If the populations included in the study are selected by the experimenter and inferences are to be made *only* about those populations, then the model is called a **fixed effects model**. If instead the populations are *representative* of other populations that are not sampled, and the experimenter wishes to infer properties of *all* populations, then the model is called a **random effects model**.

5.2 Fixed Effects ANOVA

5.2.1 The Model

Cell Means Model

$$Y_{ij} = \theta_i + \epsilon_{ij} \qquad\qquad i = 1, 2, \ldots, k, \qquad j = 1, 2, \ldots, n_i$$

where θ_i are unknown population parameters, ϵ_{ij} are random errors, k is the number of distinct populations, and n_i is the sample size in the ith population. Note, the sample sizes may not be equal.

Note, if we assume $E[\epsilon_{ij}] = 0$, then the expected value of data is

$$E[Y_{ij}] = \theta_i, j = 1, 2, \ldots, n_i.$$

We conclude that the θ_i are the *population means*, so θ_i is the population mean of population i.

If only these k populations are of interest, then the θ_i are viewed as unknown constants in the fixed effects models. If these k populations are only representative of a larger collection of populations, then the θ_i are viewed as k randomly sampled means, i.e. θ_i are random variables in the random effects model. For example, if treatment $1, 2$, and 3 are applied to three random groups of patients, then θ_1, θ_2, and θ_3 are the unknown mean treatment responses and are fixed effects. In contrast, if you sample college students and study their grade point average as a function of the number of school days missed in grades 1-12, then you can separate them into populations based on the number of days missed, but the population means, θ_i for i days missed, are random variables representative of a whole collection $\{\theta_0, \theta_1, \ldots, \theta_M\}$, where M is the maximum number of school days that can be missed. Clearly, in any reasonable sample size, we don't expect to observe students from all possible populations. Instead, we observe students with several values of i and use statistical inference to extend to the whole population.

Henceforth we focus on the fixed effects model.

Alternative Parameterization

Often, you will see another parameterization of the one-way ANOVA model.

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

Then,

$$E[Y_{ij}] = \mu + \tau_i$$

where μ is the *grand mean* and τ_i is the unique effect of treatment i. Note, there are $k + 1$ parameters in this model formulation and this leads to identifiability problems.

Data

The data associated with ANOVA might be summarized in a table of the following form:

Treatment				
1	2	3	$\cdots$	k
y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	y_{2n_2}	y_{3n_3}	$\cdots$	$\vdots$
y_{1n_1}			$\cdots$	y_{kn_k}

where *Treatment* is a label to describe the populations and is borrowed from the common ANOVA application of determining whether different treatments lead to better prognosis in medical applications.

Example: heights of singers in a choir

Suppose, for example, that you are studying the heights of singers in a choir. Your data table is below.

Soprano	Alto	Tenor	Bass
64	65	69	72
62	62	72	70
66	68	71	72
$\vdots$	$\vdots$	$\vdots$	$\vdots$
63	66	66	68
65	66	68	70
62	66	67	75
65	62	64	68
66	70		71
62	65		70
$\vdots$	$\vdots$	$\vdots$	$\vdots$
65	67		70
66	66		75
65	68		72
62			66
			72
			70
			69

[Find the original data at the Data & Story Library.]

The "treatments" are singer types (soprano, alto, tenor, bass). In working with this data, you might be interested in determining whether the mean heights of all treatments are the same. Specifically, you might expect a significant difference in the mean heights of basses and sopranos, because most, if not all, of the former are male, and the latter are female.

5.2.2 Identifiability

Identifiable

Recall our overall framework. We have a population and associated with it are unknown population parameter(s) θ. We assume there is some probability model that describes data X sampled from this population. The probability model defines the pdf $f_\theta(x)$ (or pmf for discrete outcomes) for the data.

Definition: identifiable

A population parameter θ is *identifiable* if distinct θ correspond to distinct pdfs (or pmfs for discrete random variables). That is, if $\theta \neq \theta'$, then the pdf of the data $f_\theta(x) \neq f_{\theta'}(x)$ are distinct functions.

For example, if $\mu_1 \neq \mu_2$, then the corresponding normal pdfs are not the same:

$$f_{\mu_1}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu_1)}{2\sigma}\right] \neq \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu_1)}{2\sigma}\right] = f_{\mu_2}(x)$$

indicating that population mean of normally distributed random variables is identifiable.

1. identifiability is a property of the model (not the estimates of the population parameter), so solving identifability problems involves changing the model
2. if a model is not identifiable, then estimation of or inference on its population parameters is not possible

Alternative Parameterization is Overparameterized

In the alternative formulation, there are $k+1$ parameters and k sample means available from the data. The extra degree of freedom in the data indicates that the model is *unidentifiable*. More than one choice of $(\mu, \tau_1, \ldots, \tau_k)$ can lead to the same data. One restriction on the parameters must be added to make the model identifiable. There are multiple choices for that restriction that change the way the parameters are interpretted.

- $\sum \tau_i = 0$ means that we can interpret the τ_i as deviations from the overall mean attributable to each population i.
- $\tau_1 = 0$ might be useful if population 1 is the control group and we want to interpret the $\tau_i, i > 1$ as deviations from no treatment.

5.2.3 ANOVA Framework

Assumptions

1. $E[\epsilon_{ij}] = 0$, $\text{Var}(\epsilon_{ij}) = \sigma_i^2 < \infty$ for all i, j, $\text{Cov}(\epsilon_{ij}, \epsilon_{kl}) = 0$ for all i, j, k, l with $i \neq j$ or $k \neq l$.
2. $\epsilon_{ij} \sim N(0, \sigma_i^2)$ independent.
3. *Homoscedasticity*: $\sigma_i^2 = \sigma^2$

Comments:

- Assumption 2 is required for hypothesis testing and confidence intervals.
- Without assumption 2, we are limited to do estimation. With assumption 1 about variance, we can find the estimate with minimum variance.
- Non-normality can lead to difficulties, but there are solutions for other kinds of distributions. We will not discuss much here.
- We can use CLT to get normality on population means if n_i is large enough and the real distribution is fairly symmetric.
- Robustness of ANOVA to violations of 2 depends on the extent to which 3 is true. For this reason, people will often transform the Y random variables to achieve 3 so that they do not need to worry so much about normality of their data.

Classic ANOVA Hypothesis

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$$
$$H_A : \theta_i \neq \theta_j \qquad \text{for some } i \neq j$$

This hypothesis is not often so interesting. Take the example of comparing several treatments. One may often include a control as a treatment to make sure that the experiment runs as planned. One knows before even collecting data that the control should have a different outcome compared to the rest, which means this classic H_0 will always be rejected. We might still lilke to know if $\theta_2 \neq \theta_3$.

Contrast

Definition:

Let $t = (t_1, \ldots, t_k)$ be a vector of random variables, their realizations, parameters, or statistics. Let $a = (a_1, \ldots, a_k)$ be constants, then

$$\sum_{i=1}^{k} a_i t_i$$

is a linear combination of t_i's. If $\sum_i a_i = 0$, then the linear combination is called a *contrast*.

Classical ANOVA Hypothesis in Terms of Contrasts

Theorem:

$\theta_1 = \cdots = \theta_k$ if and only if $\sum_i a_i \theta_i = 0$ for all $a \in \mathcal{A}$, where $\mathcal{A} = \{a = (a_1, \ldots, a_k) : \sum_i a_i = 0\}$.

Proof:

The forward implication is obvious

$$\sum_i a_i \theta_i = \theta \sum_i a_i = 0$$

The reverse implication is also quite easy. Consider $a^{(1)} = (1, -1, 0, \ldots, 0) \in \mathcal{A}$. This one shows $\theta_1 = \theta_2$. Similarly, $a^{(2)} = (0, 1, -1, 0, \ldots, 0)$ shows $\theta_2 = \theta_3$. In general, the set $a^{(1)}, a^{(2)}, \ldots, a^{(k-1)}$ spans the space $\mathcal{A}$. Therefore, all possible equalities encoded in $\theta_1 = \cdots = \theta_k$ are implied by combining these vectors appropriately.

5.2.4 Inference on Contrasts

Inference on Contrasts

Under the ANOVA assumptions, we have

$$Y_{ij} \sim N(\theta_i, \sigma^2)$$

Define the population sample means

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{i=1}^{n_i} Y_{ij}$$

and note

$$\bar{Y}_{i\cdot} \sim N(\theta_i, \sigma^2 / n_i)$$

by the CLT.

Also, for any a,

$$\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} \sim N(\cdot, \cdot)$$

with mean and variance

$$E\left[\sum_i a_i \bar{Y}_{i\cdot}\right] = \sum_i a_i \theta_i \qquad \text{Var}\left(\sum_i a_i \bar{Y}_{i\cdot}\right) = \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}$$

Z test for Generic Contrast

Given the above, the statistic

$$Z = \frac{\sum_i a_i \bar{Y}_{i\cdot} - \sum_i a_i \theta_i}{\sigma \sqrt{\sum_i \frac{a_i^2}{n_i}}} \sim N(0,1)$$

t-test for Generic Contrast

But of course, we don't usually know σ^2. Instead, we use

$$S_i = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$

which is unbiased for σ^2 (σ_i^2 with heteroscedasticity) and also has distribution

$$\frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}$$

If assumption 3 of homoscedasticity applies, then we can pool sample variances to get a better estimate of σ^2. Namely, with $N = \sum_i n_i$, we use the pooled sample variance

$$S_p^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) S_i^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$

Because the S_i^2 are independent, we also have

$$\frac{(N-k)S_p^2}{\sigma^2} \sim \chi^2_{N-k}$$

Also, because S_p^2 is independent of $\bar{Y}_{i\cdot}$, we have that statistic

$$\frac{\sum_i a_i \bar{Y}_{i\cdot} - \sum_i a_i \theta_i}{S_p \sqrt{\frac{a_i^2}{n_i}}} \sim t_{N-k}$$

which allows confidence intervals of the usual form

$$\sum_i a_i \bar{Y}_{i\cdot} - t_{N-k,\alpha/2} S_p \sqrt{\sum_i \frac{a_i^2}{n_i}} \leq \sum_i a_i \theta_i \leq \sum_i a_i \bar{Y}_{i\cdot} + t_{N-k,\alpha/2} S_p \sqrt{\sum_i \frac{a_i^2}{n_i}}$$

5.2.5 Classical ANOVA

Partitioning Variance

Often, ANOVA is presented as a way of partitioning the variance. The total variability can be summarized as the total sum of squares

$$SS_{\text{tot}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

Note, this is just ($N-1$ times the combined sample variance).

By adding and subtracting the sample means $\bar{Y}_{i\cdot}$, we can partition the total variance into parts

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + (Y_{ij} - \bar{Y}_{i\cdot})^2\right]$$

Expand the quadratic and recognize the cross-term becomes 0 because

$$\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot}) = n_i\bar{Y}_{i\cdot} - n_i\bar{Y}_{i\cdot} = 0$$

to find

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

Interpretting each part of this sum, we have

$$SS_{\text{tot}} = SS_{\text{treatment}} + SS_E$$

where $SS_{\text{treatment}}$ is the sum-of-squares due to treatments (i.e. between treatments) and SS_E is the sum-of-squares due to error (i.e. within treatments). There are N observations, so there are $N-1$ d.f. for SS_{tot}. There are k treatments, so there are $k-1$ d.f. for $SS_{\text{treatment}}$. Within the ith treatment, there are $n_i - 1$ d.f. for a total of $\sum_i(n_i - 1) = N - k$ d.f. within treatments for the SS_E.

Estimates of σ^2

Under the ANOVA assumptions, $\frac{SS_E}{N-k}$ uses all the data to estimate population variance σ^2 (it is the pooled sample variance).

$$\begin{aligned}
E\left[\tfrac{SS_E}{N-k}\right] &= \tfrac{1}{N-k}\textstyle\sum_{i=1}^{k} E[(Y_{ij} - \bar{Y}_{i\cdot})^2] \\
&= \tfrac{1}{N-k}\textstyle\sum_{i=1}^{k} E[(n_i - 1)S_i^2] \\
&= \tfrac{1}{N-k}\textstyle\sum_{i=1}^{k}(n_i - 1)\sigma^2 \qquad \text{constant variance assumption} \\
&= \sigma^2.
\end{aligned}$$

Under the classic ANOVA null hypothesis $H_0 : \theta_1 = \cdots = \theta_k = \theta$

$$\bar{Y}_{i\cdot} \sim N(\theta, \sigma^2/n_i) \qquad \text{or} \qquad \sqrt{n_i}\bar{Y}_{i\cdot} \sim N\left(\sqrt{n_i}\theta, \sigma^2\right)$$

Intuitively, since $\bar{Y}_{\cdot\cdot}$ provides a sample estimate of θ, we have

$$E\left[\frac{SS_{\text{treatment}}}{k-1}\right] = \frac{1}{k-1}E\left[\sum_{i=1}^{k}(\sqrt{n_i}\bar{Y}_{i\cdot} - \sqrt{n_i}\bar{Y}_{\cdot\cdot})^2\right] = \frac{1}{k-1}E\left[\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2\right] = \sigma^2$$

[If not convinced, you can work out the details by moving the expectation into the sum and using model assumptions.]

Finally, also under H_0,

$$Y_{ij} \sim N(\theta, \sigma^2)$$

so

$$E\left[\frac{SS_{\text{tot}}}{N-1}\right] = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sigma^2$$

is the traditional sample variance, which estimates the population variance.

F Test for Testing Classical ANOVA Hypothesis

These estimates of σ^2 provide the basis of the F test for the classic hypothesis. Under H_0, recall that $\frac{SS_{\text{tot}}}{\sigma^2} \sim \chi^2_{N-1}$.

Theorem: Cochran's Theorem

Let $Z_i \sim N(0,1)$ for $i = 1, \ldots, \nu$ and

$$\sum_{i=1}^{\nu} Z_i^2 = Q_1 + Q_2 + \cdots + Q_s$$

with $s \leq \nu$. Then, $Q_1, Q_2, \ldots, Q_s$ are independent χ^2 random variables with $\nu_1, \nu_2, \ldots, \nu_s$ d.f., respectively if and only if

$$\nu = \nu_1 + \nu_2 + \cdots + \nu_s.$$

Proof: Omitted.

Since $(N-k) + (k-1) = N-1$, Cochran's theorem implies

$$\begin{aligned} \frac{SS_{\text{treatment}}}{k-1} &= \chi^2_{k-1} \\ \frac{SS_E}{N-k} &= \chi^2_{N-k} \end{aligned}$$

Therefore,

$$F = \frac{SS_{\text{treatment}}/(k-1)}{SS_E/(N-k)} \sim F_{k-1,N-k}$$

If the alternative hypothesis is correct, then we expect $SS_{\text{treatment}}/(k-1)$ to *overestimate* the population variance, so large values of this statistic will indicate problems with H_0, thus rejection is according to a one-tailed test when

$$F > F_{k-1,N-k,\alpha}$$

The ANOVA Table

The one-way ANOVA analysis is summarized in the *ANOVA table*.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between treatments	$SS_{\text{treatment}}$	$k-1$	$MS_{\text{treatment}} = \frac{SS_{\text{treatment}}}{k-1}$	$F = \frac{MS_{\text{treatment}}}{MS_E}$
Within treatments	SS_E	$N-k$	$MS_E = \frac{SS_E}{N-k}$	
Total	SS_{tot}	$N-1$		