10 Analysis of Variance (ANOVA)

Chapter Preview

In chapter 9, we considered comparing the population means or population proportions of two populations. In this chapter, we will extend this framework to consider comparing the population means or proportions of three or more populations.

10.1: Single-Factor ANOVA

<u>Illustration of ANOVA</u>

Motivation

Q: Are the means the same in each population? Does $\mu_1 = \mu_2 = \ldots = \mu_k$?
A: Collect a sample from each population and use ANOVA to make the determination.

Our hypotheses will be the following:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
$H_a :$ at least one μ_i differs

- Do the average GPAs differ by class level at this university?

Is there enough evidence to suggest that the average GPA is not the same for all four levels? Our hypotheses for this example will be:

Recall that a two-sample t-test will compare the means of two populations (or treatments). ANOVA extends this to 3 or more populations (or treatments).

Definitions and Notation

Def: A factor is the characteristic that distinguishes the populations/treatments.

Def: A level of a factor is one of the specific populations or treatments.

Example: We want to test the average effectiveness of four drugs.

Example: We want to compare the average growth of sunflowers for six growing conditions.

We will measure a quantitative response (Y) for a sample from each level of the factor.

Often, an ANOVA analysis has data that comes from a controlled experiment.

k = the number of levels/treatments/populations

Notation:

	True Population Parameters		Sample Statistics		
Group	Mean	Variance	Mean	Variance	Sample Size
Group #1	μ_1	σ_1^2	$\bar{x}_1$	s_1^2	n_1
Group #2	μ_2	σ_2^2	$\bar{x}_2$	s_2^2	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group #k	μ_k	σ_k^2	$\bar{x}_k$	s_k^2	n_k

$$\boxed{N = n_1 + n_2 + \ldots + n_k}$$

$$\boxed{T = n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_2}$$

$$\boxed{\bar{\bar{x}} = T/N}$$

Sum of Squares

Def: Sum of Squares (SS) quantify two types of variation.

(1) Treatment Sum of Squares (SSTR) - measures the variability between groups.
How far apart are the sample means of the k populations?

$$\boxed{SSTR = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \ldots + n_k(\bar{x}_k - \bar{\bar{x}})^2}$$

$$\boxed{df_{TR} = k - 1}$$

(2) Error Sum of Squares (SSE) - measures the variability within the k populations.
How spread out are the individual populations?

$$\boxed{SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots (n_k - 1)s_k^2}$$

$$\boxed{df_E = N - k}$$

(3) Total Sum of Squares (SST) - measures the total variability in the data, ignoring populations and treatments.

$$\boxed{SST = SSTR + SSE = \sum (x - \bar{\bar{x}})^2}$$

$$\boxed{df_T = N - 1}$$

Example: Calculate SSTR, SSE, and SST (and the associated degrees of freedom) for GPA data set.

Mean Squares

Def: A mean square is the sum of squares divided by the degrees of freedom.

$$\boxed{MSTR = \frac{SSTR}{k-1}}$$

$$\boxed{MSE = \frac{SSE}{N-k}}$$

Note: Under the null hypothesis H_0 $(\mu_1 = \mu_2 = \dots \mu_k)$, $MSTR \approx MSE$.

Note: Under the ANOVA framework, we are assuming the distributions of the k populations are approximately normally distributed and have equal variances.

- Calculate MSTR, MSE, F_{test}, and the p-value for the GPA data set.

ANOVA Assumptions

ANOVA requires that the population or treatment distributions are normal and have the same variance. These assumptions can be checked in R although we assume these conditions hold in the problems we analyze.

χ^2 distribution

1. χ^2_k is a continuous random variable, where k is the degrees of freedom.
2. A χ^2_k random variable is the sum of k independent squared standard normal random variables.
3. Suppose $Z_1, Z_2, \ldots, Z_k$ are independent standard normal random variables. Then $\sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$.
4. $\chi^2_k \geq 0$
5. pdf of χ^2 distribution:

F distribution

1. F_{df_1,df_2} is a continuous random variable with two parameters, df_1 and df_2, that correspond to the numerator and denominator degrees of freedom.
2. An F random variable is the ratio of two χ^2 distributions, so it will always be non-negative.
3. $F_{df_1,df_2} \sim \frac{\chi^2_{df_1}/df_1}{\chi^2_{df_2}/df_2}$.
4. For our purposes, any hypothesis test will be right-tailed.
5. pdf of F_{df_1,df_2} distribution:

One Way ANOVA F-test

Step 1: State the hypotheses.

Step 2: State the level of significance.

Step 3: Calculate the test statistic.

Step 4: Calculate the p-value and plot.

Step 5: Make a statistical decision.

Step 6: Interpret your decision in the context of the problem.

Using R to complete ANOVA

Since many of these calculations can be tedious to calculate, a statistical program like R can be used to complete the analysis.

ANOVA table:

Source of Variation	df	SS	MS	F	p-value
Treatment					
Error (Residual)					
Total					

R output:

```
> summary(fit)
            Df Sum Sq Mean Sq F value  Pr(>F)
year         3  3.338  1.1125   6.781 0.00367 **
Residuals   16  2.625  0.1641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R code:

```
# GPA data set
fresh.gpa = c(2.05, 2.20, 2.00, 2.50, 2.75)
soph.gpa = c(3.00, 2.80, 3.80, 3.60, 3.30)
junior.gpa = c(3.30, 3.20, 2.30, 2.90, 3.30)
senior.gpa = c(3.50, 3.10, 2.90, 3.00, 4.00)

# Make one list of all GPAs
gpa = c(fresh.gpa,soph.gpa,junior.gpa,senior.gpa)

# Create labels
year = c(rep("fresh",5),rep("soph",5),rep("junior",5),rep("senior",5))

# Make a boxplot
boxplot(gpa~year)

# calculate ANOVA
fit = aov(gpa~year)
summary(fit)
```

- Using the R output from the previous page, complete a six-step hypothesis test to whether the mean GPAs at the four college levels are the same at $\alpha = 0.05$.

10.2: Multiple Comparison in ANOVA

Suppose that we conclude that the means of the k populations are not all equal. Which populations means do differ?

Motivation

Q: Can we just do two-sample t-tests for all pairs of populations?
A: Yes and no. We need to make an adjustment for simultaneously completing a group of tests.

- In the GPA example, we had four populations. Suppose the null hypothesis is true ($\mu_1 = \mu_2 = \mu_3 = \mu_4$) and $\alpha = 0.05$.

Tukey's Procedure

Tukey's Procedure (Tukey's Honest Significant Difference or Tukey HSD) is a multiple comparisons procedure to account for completing a number of simultaneous tests.

Confidence intervals for the difference in means ($\mu_i - \mu_j$) can be computed using Tukey's Procedure. These calculations utilize another continuous distribution, the studentized range distribution (Q).

While these calculations can be completed by hand, we'll focus on calculating them in R.

Example: Returning to the GPA data set, determine which population means differ at $\alpha = 0.05$ using a Tukey adjustment. Some R code and output is provided below.

R code:

```
TukeyHSD(fit)
plot(TukeyHSD(fit))
```

R output:

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = gpa ~ year)

$year
              diff         lwr       upr     p adj
junior-fresh   0.7 -0.03291823 1.4329182 0.0638299
senior-fresh   1.0  0.26708177 1.7329182 0.0062342
soph-fresh     1.0  0.26708177 1.7329182 0.0062342
senior-junior  0.3 -0.43291823 1.0329182 0.6527735
soph-junior    0.3 -0.43291823 1.0329182 0.6527735
soph-senior    0.0 -0.73291823 0.7329182 1.0000000
```

R plot: