Elementary Statistics for the Biological and Life Sciences

STAT 205

University of South Carolina
Columbia, SC

Chapter 11: An Introduction to Analysis of Variance (ANOVA)

Example 11.2.1

- Return to the (indep.) 2-sample model from Sec. 7.2 (t-test). What if there are more than 2 groups?

- <u>Ex. 11.2.1</u>: Lamb response to diff't diets.

  Y_1 = weight gain of lambs after diet type 1

  Y_2 = weight gain of lambs after diet type 2

  Y_3 = weight gain of lambs after diet type 3

So now we're interested in μ_1 vs. μ_2 vs. μ_3!

Multiple Group Model

Observations from:	# obsv'ns	sample mean	sample variance
$Y_1 \sim N(\mu_1,\sigma^2)$	n_1	$\bar{Y}_{1\bullet}$	S_1^2
$Y_2 \sim N(\mu_2,\sigma^2)$	n_2	$\bar{Y}_{2\bullet}$	S_2^2
$Y_3 \sim N(\mu_3,\sigma^2)$	n_3	$\bar{Y}_{3\bullet}$	S_3^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y_I \sim N(\mu_I,\sigma^2)$	n_I	$\bar{Y}_{I\bullet}$	S_I^2

Notice: "dot" notation $\bar{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$

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

Multiple Comparisons

- **It's natural to ask: why not now just compare *all possible pairs* $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_2 - \mu_3$, etc., each with a t-test (or conf. interval)?**

- **DEF'N: The PROBLEM of MULTIPLE COMPARISONS occurs when the same data set is used to make multiple inferences on $I > 1$ associated parameters.**

Error Inflation

- Suppose $I = 3$.
- Perform a t-test of $H_o:\mu_1 = \mu_2$ with false-positive error rate set to $\alpha = .05$.
- Then, perform another test, now on $H_o:\mu_1 = \mu_3$. *Jointly*, the probability of making a false positive error is now *larger* than 5%!!
- In fact, it gets worse as I increases. Table 11.1.2 gives an illustration →

Table 11.1.2

Table 11.1.2 Overall risk of Type I error in using repeated t tests at $\alpha = 0.05$

I	Overall risk
2	0.05
3	0.12
4	0.20
6	0.37
8	0.51
10	0.63

Combined Analysis

- Error considerations make it more efficient to study an overall null hypothesis:

$$H_o: \mu_1 = \mu_2 = \cdots = \mu_I$$

(vs. a non-directional alternative

H_A: some difference among μ_i's).

- Indeed, a more complete analysis seeks to incorporate information among <u>all</u> groups:
  - e.g., we can estimate the common σ^2 using variation from all I groups.

ANOVA

- **DEF'N:** The **ANALYSIS OF VARIANCE (ANOVA)** among $\mathrm{I} > 1$ populations is used to make inferences about the I population means.

- The various components for an ANOVA are assembled from the data. We start with:

$$n^* = \sum_{i=1}^{\mathrm{I}} n_i \quad = \text{ total sample size}$$

$$\bar{y}_{\bullet\bullet} = \frac{1}{n^*} \sum_{i=1}^{\mathrm{I}} \sum_{j=1}^{n_i} y_{ij} = \text{ grand mean}$$

Sums of Squares

DEF'N: **SUMS OF SQUARES** are sums of squared deviations from a central value.

(a) the **WITHIN GROUPS S.S.** is

$$SS(\text{Within}) = \sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2$$

(b) the **BETWEEN GROUPS S.S.** is

$$SS(\text{Between}) = \sum_i \sum_j (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = \sum_i n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$

(c) the **TOTAL S.S.** is

$$SS(\text{Total}) = \sum_i \sum_j (y_{ij} - \bar{y}_{\bullet\bullet})^2$$

SS(Resid.)

- Notice that in

$$SS(\text{Within}) = \sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2$$

the per- group means $\bar{y}_{i\bullet}$ can be viewed as "predictors" of each y_{ij} under our multi-group model, so these deviations are a form of "residual" or "error" from y_{ij}

- So, we often write SS(Within) = SS(Resid.) or sometimes SS(Within) = SS(Error).

Mean Squares

DEF'N: A **MEAN SQUARE** (MS) is the avg. of the squared deviations from a central value. It is a Sum of Squares (SS) divided by the number of informative values in the SS.

called "degrees of freedom", or df

MS(Resid.)

- For $SS(Within) = \sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2$

  we have $df_{within} = n^* - I$. So, MS(Within) is

$$MS(Within) = \frac{SS(Within)}{df_{within}} = \frac{SS(Within)}{n^* - I}$$

- Some other notations:

  MS(Resid.) = SS(Within)/$(n^* - I)$

  MS(Error) = SS(Within)/$(n^* - I)$

MS(Between)

- For $SS(\text{Between}) = \sum_i \sum_j (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$

$df_{betw'n} = I - 1$. So, MS(Between) is

$$MS(\text{Between}) = \frac{SS(\text{Between})}{df_{betw'n}} = \frac{SS(\text{Between})}{I - 1}$$

- Alternate notation:

MS(Groups) = SS(Between)/(I – 1)

Computing Formulae

For computing purposes, we use:

$$SS(\text{Total}) = \sum_i \sum_j y_{ij}^2 - \frac{1}{n^*}\left(\sum_i \sum_j y_{ij}\right)^2$$

and $SS(\text{Resid.}) = \sum_i (n_i - 1)S_i^2 \quad (= SS\{\text{Within}\})$

And it turns out (11.2.1) that

$$SS(\text{Tot.}) = SS(\text{Between}) + SS(\text{Resid.})$$

Pooled Average

Also, since $df_{resid} = n^* - I = \sum_{i=1}^{I} (n_i - 1)$

we can write

$$MS(Resid.) = \frac{\sum (n_i - 1)S_i^2}{\sum (n_i - 1)}$$

as a df-weighted ("pooled") avg. of the per-group variances. If all the popl'n variances are equal to σ^2, then MS(Resid.) is an unbiased estimator of this common σ^2.

Estimating σ

- To estimate the common σ we use

$$\sqrt{\text{MS(Resid.)}}$$

- To emphasize this we use the notation

$$S_{pool} = \sqrt{\text{MS(Resid.)}}$$

which is at times called the "root mean squared error" or RMSE.

ANOVA

- We collect all the
  - Sums of Squares,
  - Mean Squares,
  - df, etc.,

  together into a simple table of values, called an **ANOVA Table**.
- Think of it simply as an accounting device, or perhaps as a "spreadsheet" for arranging all the various terms.

ANOVA Table

Source of Variation	df	SS	MS
Between Groups	$I - 1$	$\sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$\frac{SS(Between)}{I - 1}$
Residual	$n^* - I$	$\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$	$\frac{SS(Resid.)}{n^* - I}$
Total	$n^* - 1$	$\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2$	

NOTE: Recall that SS(Resid.) is also called SS(Within)

Example 11.2.1

Ex. 11.2.1: For the Lamb Weight example, we have the following data:

Table 11.2.1 Weight gains of lambs (lb)*

	Diet 1	Diet 2	Diet 3
	8	9	15
	16	16	10
	9	21	17
		11	6
		18	
n_i	3	5	4
Sum $= \sum_{j=1}^{n_i} y_{ij}$	33	75	48
Mean $= \bar{y}_i$	11.000	15.000	12.000
SD $= s_i$	4.359	4.950	4.967

*Extra digits are reported for accuracy of subsequent calculations.

Example 11.2.6 – ANOVA table

- The ANOVA calculations for the Lamb Weight data done in R give (coming up…) Table 11.2.3:

Table 11.2.3 ANOVA table for lamb weight gains

Source	df	SS	MS
Between diets	2	36	18.00
Within diets	9	210	23.33
Total	11	246	

- So, e.g., $S_{pool} = \sqrt{23.333} = 4.83$ lbs.

F-testing

- We use the ANOVA calculations to assess H_o: $\mu_1 = \cdots = \mu_I$. To do so, we need the following:

- **DEF'N**: The F-DISTRIBUTION with ν_1 and ν_2 degrees of freedom is the dist'n of the ratio of two (indep.) mean squares.

  NOTATION: $F \sim F(\nu_1,\nu_2)$

- In $F(\nu_1,\nu_2)$, ν_1 = df from the numerator MS, and ν_2 = df from the denominator MS.

F-ratio

We use F for testing H_o: $\mu_1 = \cdots = \mu_I$ as follows:

- calculate F_s = MS(Between)/MS(Resid.)
- under H_o, $F_s \sim F(df_{betw'n}, df_{resid})$
- Obtain P-value from R.
- If P-value < α then reject!

Review

- We now have I groups to compare, each with their own mean $\mu_1, \mu_2, \ldots, \mu_I$.
- We assume the variance σ^2 is the same across the groups.
- The ANOVA table places all the SS, MS, F_s , and the P-value into a convenient table, called the ANOVA table.
- The P-value tests $H_o: \mu_1 = \mu_2 = \cdots = \mu_I$ vs. a non-directional alternative. If we reject, then we know there's a group effect!

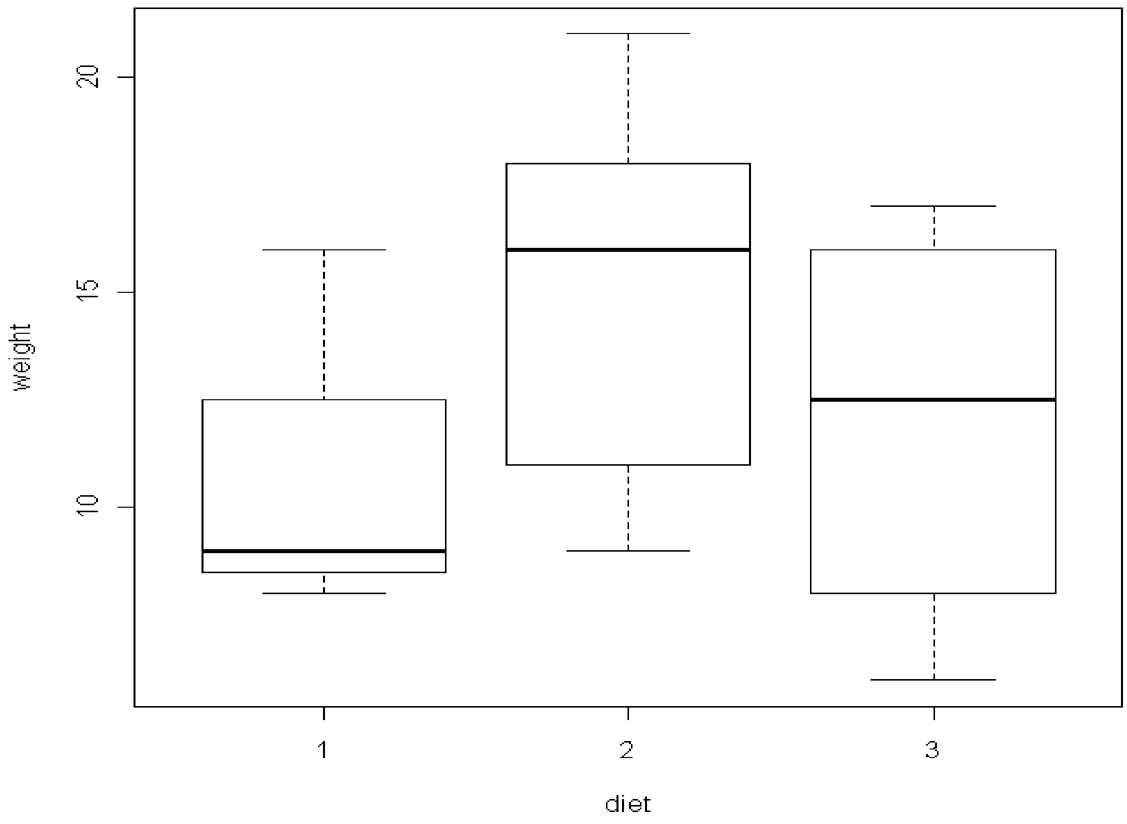
ANOVA in R

- We need to define two lists, a list of the response variable, and list indicating which group the response came from.
- The group list needs to be a "factor" in R.
- Fitting the ANOVA model is carried out through `fit=aov(response~group)` then typing `summary(fit)` to get the ANOVA table and P-value.
- Let's look at the lamb diet data…

R code

```
> weight=c(8,16,9,9,16,21,11,18,15,10,17,6)
> diet =c(1,1,1,2,2,2,2,2,3,3,3,3)
> diet=factor(diet)
> plot(weight~diet)
> fit=aov(weight~diet)
> summary(fit)
            Df Sum Sq Mean Sq F value Pr(>F)
diet         2     36  18.000  0.7714 0.4907
Residuals    9    210  23.333
```

Side-by-side boxplots



Interpretation

- P-value = 0.49 > 0.05, we accept that there is no difference in weight gain due to diet at the 5% level.
- An estimate of σ^2 23.33.
- We will now analyze the "MOA & schizophrenia" data from Chapter 1.
- You will analyze the "radish growth" data from Chapter 2 in your homework.