One-Way Analysis of Variance

BIOS 6611

CU Anschutz

Week 11

1. One-Way Analysis of Variance (ANOVA)

2. ANOVA as a Linear Model

3. ANOVA with Unequal Variances

One-Way Analysis of Variance (ANOVA)

Motivation

One-way ANOVA can be used to compare the means of J groups ($J \geq 2$).

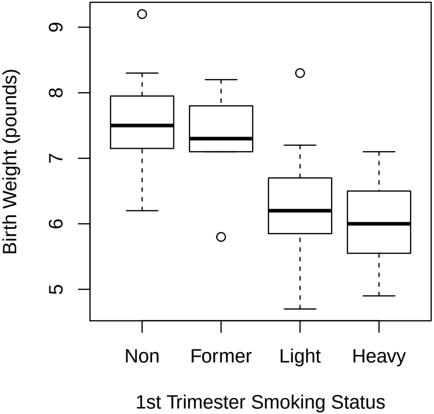
It can be thought of as a generalization of the independent samples t-test (with equal variances).

We will introduce some methods specific to ANOVA, as well as draw connections with how a general linear (regression) model can be used as well.

Motivating Example

Our motivating example will be infant birthweight (pounds) and smoking status of mother during the first trimester.

	Smoking Status			
i	**Non**	**Former**	**Light**	**Heavy**
1	7.5	5.8	5.9	6.2
2	6.2	7.3	6.2	6.8
3	6.9	8.2	5.8	5.7
4	7.4	7.1	4.7	4.9
5	9.2	7.8	8.3	6.2
6	8.3		7.2	7.1
7	7.6		6.2	5.8
8				5.4
$\bar{Y}_j$	7.59	7.24	6.33	6.01
s_j^2	0.93	0.83	1.3	0.52



ANOVA Model Formulation

One formulation for the one-way ANOVA is the *effects model*:

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

- Y_{ij} denotes the outcome for the i^{th} observation of the j^{th} group
- μ is a constant that represents the **grand mean** of all groups taken together
- α_j is a constant that represents the difference between the mean of the j^{th} group and the grand mean (*between group differences*) where $\sum_j \alpha_j = 0$
- ϵ_{ij} is the **error term** that represents the random errors about the group mean, $\mu + \alpha_j$, for an individual observation from the j^{th} group (*within group differences*)
- Each group has n_j observations, with $N = \sum_j n_j$ total observations

One-way ANOVA Assumptions

The assumptions for the one-way ANOVA are:

Independence: The samples are randomly and independently drawn from the respective populations.

Homoscedasticity: The variances of the j populations are the same.

Normal Distribution: Each population is normally distributed, and thus the errors follow a normal distribution:

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ and } Y_{ij} \sim N(\mu + \alpha_j, \sigma^2)$$

Partitioning the Variance

Like regression, we can partition the variance into the *between* and *within* group variability. We can present this as the *sum of squares* that we have seen in regression.

Let $\overline{\overline{Y}}$ denote the grand mean and $\bar{Y}_j$ denote the mean for the j^{th} group, then the deviation of an individual observation from the grand mean can be represented as:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\overline{\overline{Y}})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{Y}_j-\overline{\overline{Y}})^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_j)^2$$

In other words, Total SS $=$ Between SS $+$ Within SS.

This corresponds to our regression presentation for partitioning of the variability: $SS_{\text{Total}} = SS_{\text{Model}} + SS_{\text{Error}}$.

F-test for the One-Way ANOVA

The sums of squares allow us to derive the terms in the eponymous *ANOVA table* and conduct an **overall F-test** if all groups have the same mean.

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Variance Ratio (F)**	**p-value**
Between (Model)	$\text{SS}_{\text{Between}}$	$J-1$	$\text{MS}_{\text{Between}}$	$\text{F}=\frac{MS_{Between}}{MS_{Within}}$	$\Pr(F_{J-1,N-J} > \text{F})$
Within (Error)	$\text{SS}_{\text{Within}}$	$N-J$	$\text{MS}_{\text{Within}}$		
Total	SS_{Total}	$N-1$			

Formally stated,

$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_J = 0$

H_1 : at least one $\alpha_j \neq 0$

Equivalently, for $\mu_j = \mu + \alpha_j$,

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

H_1 : at least one of the means is different

Birthweight Example Code

In SAS we can use PROC ANOVA:

```
PROC ANOVA data=BWT;
    class momsmoke;
    model birthwt = momsmoke;
    means momsmoke;
run;
```

In R we can conduct our one-way ANOVA using oneway.test:

```
BWT <- read.csv('birthweight_smoking_dataset.csv', header=T)
oneway.test(birthwt ~ momsmoke, data=BWT, var.equal = T)
```

Birthweight Example Interpretation

```
##
##  One-way analysis of means
##
## data:  birthwt and momsmoke
## F = 4.4076, num df = 3, denom df = 23, p-value = 0.01371
```

Our hypotheses are:

Our conclusion is:

ANOVA as a Linear Model

ANOVA via Linear Regression

The linear regression model is *extremely* flexible. In fact, we can recreate the one-way ANOVA model in our regression framework with indicator variables. For example, we can represent our single `momsmoke` variable as 3 indicator variables:

Group	X_F	X_L	X_H
Non	0	0	0
Former	1	0	0
Light	0	1	0
Heavy	0	0	1

In terms of our true regression model, it could be represented as

$$Y_i = \beta_0 + \beta_F X_F + \beta_L X_L + \beta_H X_H + \epsilon_i$$

Connection Between Hypotheses

In the context of our birthweight example, the one-way ANOVA hypothesis is

$$H_0 : \mu_{non} = \mu_{former} = \mu_{light} = \mu_{heavy}$$

We can draw a direct connection with our regression model's β coefficients, where the overall F-test evaluates the null hypothesis:

$$\begin{aligned}
H_0 :& \beta_F = \beta_L = \beta_H = 0 \\
& \quad \text{(Step 1: Add } \beta_0 \text{ to the } H_0\text{)} \\
H_0 :& \beta_F + \beta_0 = \beta_L + \beta_0 = \beta_H + \beta_0 = 0 + \beta_0 \\
& \quad \text{(Step 2: Substitute in definition for } \mu_j\text{)} \\
H_0 :& \mu_{former} = \mu_{light} = \mu_{heavy} = \mu_{non}
\end{aligned}$$

Birthweight Example as Regression Model

Using `lm` we can see we arrive at the same results (on the next slide):

```r
# Create indicator variables
BWT$Xf <- BWT$momsmoke=='Former'
BWT$Xl <- BWT$momsmoke=='Light'
BWT$Xh <- BWT$momsmoke=='Heavy'

# Fit our regression model
lm1 <- lm(birthwt ~ Xf + Xl + Xh, data=BWT)

# Note, we can equivalently fit:
lm2 <- lm(birthwt ~ momsmoke, data=BWT)
## This works since momsmoke is a character/factor variable
## R will assign a reference category for us
```

Birthweight Example as Regression Model

Recall, in our one-way ANOVA we previously observed $F = 4.4076$ with (3,23) DF and $p = 0.01371$. If we look at summary output for `lm`:

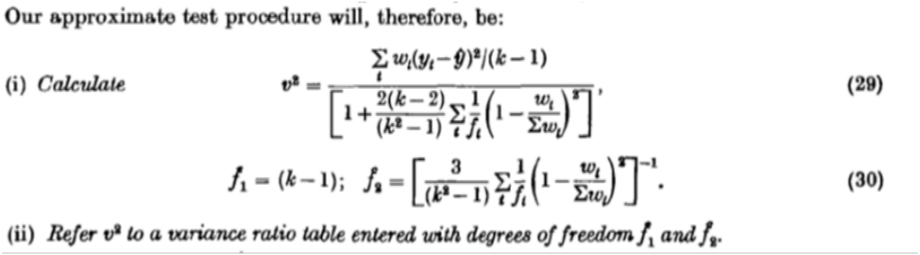
```
##
## Call:
## lm(formula = birthwt ~ Xf + Xl + Xh, data = BWT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6286 -0.4786 -0.1286  0.6371  1.9714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5857     0.3551  21.361  < 2e-16 ***
## XfTRUE       -0.3457     0.5501  -0.628  0.53593
## XlTRUE       -1.2571     0.5022  -2.503  0.01985 *
## XhTRUE       -1.5732     0.4863  -3.235  0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9396 on 23 degrees of freedom
## Multiple R-squared:  0.365,  Adjusted R-squared:  0.2822
## F-statistic: 4.408 on 3 and 23 DF,  p-value: 0.01371
```

ANOVA with Unequal Variances

Unequal Group Variances

We can test if the variances between our groups are equal using **Levene's test** or **Bartlett's test** for homogeneity of variance. For both tests, H_0 *is that the variances are equal across groups.*

If our variances are *not* equal across groups, we can use the approximate F-test proposed by B.L. Welch[1] to implement **Welch's ANOVA** that does not assume equal variances. It can be thought of as an extension of the *two-sample t-test assuming unequal variances* to more than two groups:



Our approximate test procedure will, therefore, be:

(i) *Calculate*

$$v^2 = \frac{\sum_t w_t(y_t-\hat{y})^2/(k-1)}{\left[1+\frac{2(k-2)}{(k^2-1)}\sum_t \frac{1}{f_t}\left(1-\frac{w_t}{\Sigma w_t}\right)^2\right]}, \quad (29)$$

$$f_1 = (k-1); \quad f_2 = \left[\frac{3}{(k^2-1)}\sum_t \frac{1}{f_t}\left(1-\frac{w_t}{\Sigma w_t}\right)^2\right]^{-1}. \quad (30)$$

(ii) *Refer v^2 to a variance ratio table entered with degrees of freedom f_1 and f_2.*

[1] On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330-336, 1951.

Code for Testing Homogeneity of Variances

For SAS we can request hovtest in PROC GLM:

```
PROC GLM DATA = BWT;
  CLASS momsmoke;
  MODEL birthwt = momsmoke;
  MEANS momsmoke/ hovtest=levene(type=abs) hovtest=bartlett WELCH;
RUN;
```

In R we can use leveneTest from the car package and bartlett.test from the default stats package:

```
## Test Equality of Variances
library(car)
leveneTest( birthwt ~ momsmoke, data=BWT)
# add center=mean to match SAS output for leveneTest
bartlett.test( birthwt ~ momsmoke, data=BWT)

## Welch's one-way ANOVA
oneway.test( birthwt ~ momsmoke, data=BWT, var.equal = FALSE)
```

Birthweight Example: Testing Homogeneity of Variances

```
library(car)
leveneTest( birthwt ~ momsmoke, data=BWT) # add center=mean to match SAS
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.1247 0.9446
##       23
```

```
bartlett.test( birthwt ~ momsmoke, data=BWT)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  birthwt by momsmoke
## Bartlett's K-squared = 1.2656, df = 3, p-value = 0.7373
```

Birthweight Example: Welch's One-Way ANOVA

```r
oneway.test( birthwt ~ momsmoke, data=BWT, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  birthwt and momsmoke
## F = 4.6369, num df = 3.000, denom df = 11.529, p-value = 0.02354
```

Closing Comments

In general, always applying Welch's method for a one-way ANOVA when the equality of variances is unknown is a good strategy. If they are equal, the standard and Welch's one-way ANOVA are essentially the same.

Some other alternatives include:

- Using variable transformations (e.g., the $\log(Y)$)
- Fitting unequal variance t-tests on all pairwise comparisons with corrections for multiple comparisons

Many times we also want to adjust for continuous predictors as well, making our regression framework very flexible and more practical than ANOVA which only incorporates categorical predictors.

In the next lecture we will discuss methods to address a typical follow-up question to a significant ANOVA test: *what group means are significantly different?*