Sanchari Chowdhuri
sanchari@umd.edu

Do Songs of different genres have different duration?

Using R to analyze data the million song database. Determining whether songs from different genre's have significantly different durations.

A subset is created from million song database containing genre of music and duration. The subset is created such that it has 2 variables namely genre and duration. The duration of the songs are grouped according to 10 genre divisions. Since the research question is to find if songs from different genre have significantly different durations, the one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups (in this case genre).

The one-way ANOVA compares the means of duration between the groups (genre of music) and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the **null hypothesis**:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

Where μ = group mean (genre) and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the **alternative hypothesis (H_A),** which is that there are at least one group means that are statistically significantly different from each other.

It is important to realize that the one-way ANOVA is an **omnibus** test statistic and cannot tell which specific groups (i.e genre) were statistically significantly different from each other, only that at least one group is different. In order to understand which groups vary we need to do pairwise t testing

Three main assumptions are:

1. The dependent variable is **normally distributed** in each group that is being compared in the one-way ANOVA .So, for example, in this case when we are comparing ten genres (*Classic-Pop-Rock, Classical, Punk, hip-hop, dance electronica, jazz-blue, metal, Pop and Soul-reggae*) based on duration of songs from these genres, ***the duration of songs (Dependent variables)*** would have to be normally distributed for all the ***genres groups*** (***Independent Variables***).

2. **There is homogeneity of variances**. This means that the population variances in each group (genre of music) are equal.
3. **Independence of observations**. This is mostly a study design issue and, as such, you will need to determine whether you believe it is possible that your observations are not independent based on your study design.

Testing of Assumptions

- Firstly we check the **normality of the Independent variables** (genres wise duration). Using the following comments in R

  > qqnorm(a$)

  > qqline(a$)

  > hist(a$)

  Looking at the qq plot and histogram we conclude that although majority of the data is compliant to normality; some of them are not however ANOVA is robust against some deviations from normality.

- **Homogeneity of Variance:**

  To test the homogeneity of variance we conduct Levene Test

  - H0: variances are equal
  - Ha: at least one variance is not equal
    - p < significance level (o.05)
    - Reject H0

  > library(car)

  > levene.test(a$duration~ a$genre, data=a)

  Levene's Test for Homogeneity of Variance (center = median)

Df	F value	Pr(>F)	group
9	5.578	2.246e-07	*
536			

  Signif. codes: 0 ‘**’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since p value is less than 0.05 we reject the null hypothesis. Thus there is at least one variance which is not equal.

- **Independence of Observations**

  Every song should uniquely belongs to one genre and hence have unique duration .Comparing Song title and genre in excel we have realized that there are 3 values which are repeated. Hence Independence test is violated.

The research question is whether songs from different genres have significantly different durations.

Null Hypothesis is Ho= *the mean of song durations from different genres are same*.

Alternate Hypothesis: $\boldsymbol{H_a} = Atleast\ one\ group's\ mean\ is\ not\ equal.$

Performing the ANOVA test in R

```
av=aov(a$duration~a$genre ,data=a)
> summary(av)
             Df  Sum Sq Mean Sq F value   Pr(>F)
a$genre       9  581999   64667   6.988 1.43e-09 ***
Residuals   536 4960031    9254
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Source Variation	Df	SS	MS	F	p
Between	9	581999	64667	6.988	1.43e-09
Within	536	4960031	9254		
Total	df_{total} **(545)**	SS_{total} **(5542030)**			

$df_{total} = df_{between} + df_{within}$ =9+536

$\boldsymbol{df_{total} = 545}$

$df_{between} = No.of\ levels - 1$ = 10-1

$\boldsymbol{df_{between} = 9}$

$df_{total} = No.of\ observations - 1 = 546 - 1$

$$\boldsymbol{df_{total} = 545}$$

$$SS_{total} = SS_{between} + SS_{within} = 581999 + 4960031$$

$$\boldsymbol{SS_{total} = 5542030}$$

$$MS_{between} = SS_{between} / df_{between} = 581999/9$$

$$\boldsymbol{MS_{between} = 64667}$$

$$MS_{within} = SS_{within} / df_{within} = 4960031/536$$

$$\boldsymbol{MS_{within} = 9254}$$

$$F_{between} = \frac{MS_{between}}{MS_{within}}$$

$$\boldsymbol{F_{between} = 6.698}$$

pf(6.698, 9, 536,lower.tail=FALSE)= 1.43e-09

Since p value is less than 0.05 we reject the Null Hypothesis. That means at least one mean is significantly different.

Effect Size: R^2= Percent of variance explained by Independent Variables

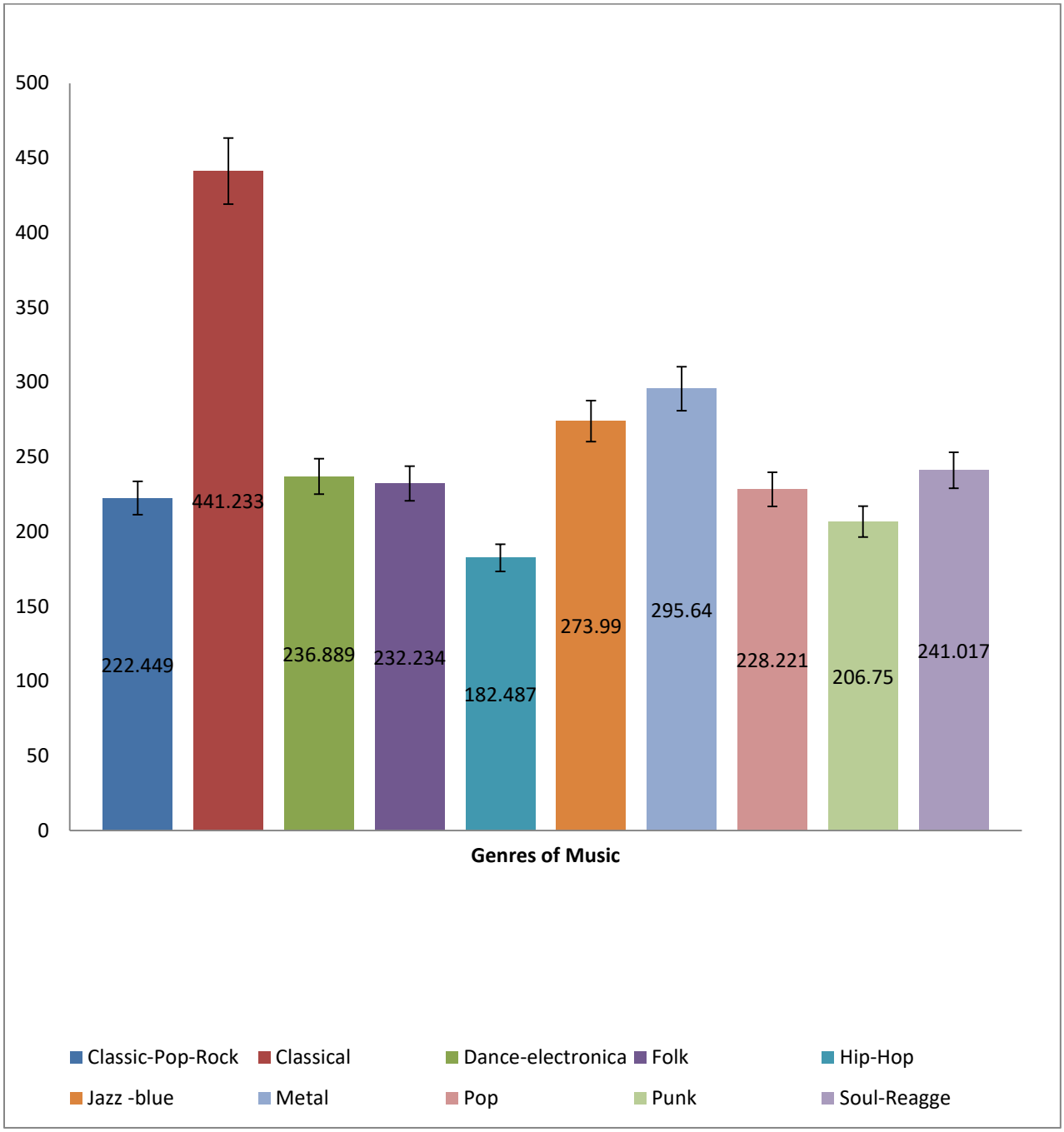
$$R^2 = SS_{between}/SS_{total}$$

$$R^2 = \frac{581999}{5542030} = 0.105$$

This means only 10.5% variance is explained.in DV due IV

Sanchari Chowdhuri
sanchari@umd.edu

Bar graph along with error bars of standard errors of the mean for each genre.



One way ANOVA is an omnibus test doesn't tell us which means are significantly different from each other .Hence a pairwise two sample t-tests is required to compare population means.

```
pairwise.t.test(a$duration,a$genre,p.adj="bonferroni")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  a$duration and a$genre
```

data: a$duration and a$genre

	classic pop and rock	classical	dance and electronica	folk
classical	2.5e-08	-	-	-
dance and electronica	1.00000	6.2e-06	-	-
folk	1.00000	2.1e-07	1.00000	-
hip-hop	1.00000	1.3e-05	1.00000	1.00000
jazz and blues	0.06805	0.00036	1.00000	0.70095
metal	0.00289	0.00648	0.74653	0.04346
pop	1.00000	5.0e-05	1.00000	1.00000
punk	1.00000	3.2e-05	1.00000	1.00000
soul and reggae	1.00000	2.5e-06	1.00000	1.00000

	Hip-hop	jazz and blues	metal	pop	punk
classical	-	-	-	-	-
dance and electronica	-	-	-	-	-
folk	-	-	-	-	-
hip-hop	-	-	-	-	-
jazz and blues	0.90784	-	-	-	-
metal	0.22478	1.00000	-	-	-
pop	1.00000	1.00000	1.00000	-	-
punk	1.00000	1.00000	0.65933	1.00000	-
soul and reggae	1.00000	1.00000	0.46895	1.00000	1.00000

```
P value adjustment method: bonferroni
```

The above mentioned output of pairwise t testing gives us the p value of every pair wise testing of mean duration of songs genre wise and ***if p value is less than 0.05 we reject the null hypothesis*** .Thereby concluding that those genres don't have similar mean duration of song. Looking at the above results we derive the following interpretation:

- The p value of ***classical music and classic pop and rock is 2.5e-08*** which is less than 0.05 hence we reject the null. Thus there is difference in duration of means among songs in these two groups. Classical music has longer duration.
- The p value of ***classical music is less than 0.05*** when compared to other music genres. Hence we reject the null hypothesis. This basically means that the duration of classical music is longer than any other music genre songs.
- The ***p value of metal and classic pop is 0.002 which is less than 0.05*** . Hence we reject the null hypothesis. Thus metal has longer song durations.
- The p value of ***folk and metal is 0.04*** which is less than 0.05. Hence we reject the null hypothesis of having no difference in duration. Going through the average of both the genres we realize that metal has higher average than folk. Hence metal songs have larger duration.

Thus the pairwise t test tells us which two genres in a group have different mean as compared to others.