

One-Way ANOVA to Compare Population Means

In this class, you will work with your group to explore one-way ANOVA models by completing **activities 19-26 from the textbook**. While you work through this activity, make sure that all group members are engaged and contribute ideas, and also follow the code. The [R Manual](#) has useful R code for today's activities.

The ANOVA Model

As you saw before class, the ANOVA model is given by

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \text{for } i = 1, \dots, I \quad \text{and } j = 1, 2, \dots, n_i, \quad \text{where } \varepsilon_{ij} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Once we have data in hand, we can use it to estimate the model parameters of interest. When interest centers on discerning a difference between group means the **effects**, α_i , are of primary interest. Today's activity will guide you through an initial one-way ANOVA model.

To begin, load the **Games1** data set:

```
games1 <- read.csv("https://aloy.rbind.io/kuiper_data/Games1.csv")
```

You should also load the **ggformula** and **mosaic** libraries.

```
library(ggformula)
library(mosaic)
```

Activity 19.

Use the `mean()` function with appropriate formulas to calculate $\bar{y}_{..}$, $\bar{y}_{1.}$, and $\bar{y}_{2.}$.

Activity 20.

Using the sample means you just calculated, estimate the effect sizes, α_i , for the color distractor game and the standard game.

Activity 21.

A **main effects** plot simply plots the average for each factor level.

To create the main effects plot of game type, need to plot the average response for each factor level. You calculated these group means in Activity 19 above. To create this plot, we need to first create a data frame with two columns: the group labels and the group means.

```
group_means <- mean(Time ~ Type, data = games1)
main_effects <- data.frame(
  group = names(group_means), # use group names from the group means
  avg = group_means           # the group means
)
```

Next, we can use `gf_point()` and `gf_line()` to draw the plot:

```
main_effects |>                               # take the effects data
  gf_point(avg ~ group) |>                     # and draw the points
  gf_line(avg ~ group, group = 1) # and then connect the points
```

In the `gf_line()` command above we needed to add `group = 1` because we used the original group variable, not the indicator (dummy) variable. If instead we had used 0s and 1s, this would not have been necessary.

Activity 22.

Calculate the residual for the 20th observation from the standard group, $\hat{\epsilon}_{2,20}$.

In R you can extract the 20th observation for the standard group by first extracting all of the observations from that group, and then extracting the 20th observation.

```
standard <- filter(games1, Type == "Standard") # extract standard group
standard$Time[20] # extract 20th value from Time column
```

Activity 23.

Before class, you saw the formulas to calculate the F-statistic.

A one-way ANOVA model is fit using the `aov()` command in R. The ANOVA table (without the Total row) is produced using the `summary()` command on the fitted model:

```
games_anova <- aov(Time ~ Type, data = games1)
summary(games_anova)
```

Report the value of the F-statistic, the degrees of freedom, and the p-value. Based on your results, draw a conclusion from this study.

Activity 24.

Below is output from a two-sample t-test and the regression model for this same problem. How does the test statistic and p-value compare from the one-sample ANOVA model compare to those you found for the two-sample t-test and the slope of the regression model?

Two-sample model

```
data: Time by Type
t = 2.2862, df = 38, p-value = 0.02791
alt. hypothesis: true difference != 0
95 percent confidence interval:
 0.2920254 4.8079746
```

Regression model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.1000    0.7887   48.308  <2e-16
TypeStandard  -2.5500    1.1154  -2.286   0.0279
```

Activity 25.

Take the square root of the F-statistic you found via one-way ANOVA. Does this look familiar? Explain.

The `sqrt()` function takes the square root of a number in R.

Activity 26.

Recall that the necessary conditions for the one-way ANOVA model are:

1. the error terms are i.i.d.
2. the error terms follow a normal distribution
3. the error terms have mean 0
4. the population variance is the same for each group (i.e., within each factor level)

To check these assumptions, let's explore the residuals. You can calculate the residuals from an ANOVA object using the `resid()` function:

```
games_resids <- resid(games_anova)
```

Now, check the model assumptions by

- creating a histogram of the residuals,
- creating a plots of the residuals vs. the type of game, and
- creating a plot of the residuals against the order in which the data were collected.

You can create these plots with the following code:

```
# Histogram of the residuals with 15 bins
gf_histogram(~games_resids, bins = 15, xlab = "Residuals")

# Plot of the residuals vs. type of game
gf_jitter(games_resids ~ Type, data = games1, width = 0.1, ylab = "Residual") %>%
  gf_hline(yintercept = ~0, color = "blue")

# Plot of the residuals vs. order collected
gf_point(games_resids ~ studentID, data = games1, ylab = "Residuals", xlab = "Order") %>%
```

```
gf_line() %>%  
gf_hline(yintercept = ~0, color = "blue")
```

Are the residuals approximately normal? Are the residual variances (spreads) similar for the two factor levels? Are there patterns in the residual plots that might indicate the residuals are not iid?