

One-Way Analysis of Variance (ANOVA)

1. Hypotheses

One-Way Analysis of Variance or **One-Way ANOVA** is a statistical method to determine if there is a difference in means between two or more *independent* groups, where the groups are defined by the outcomes for a single categorical variable.

In its purpose, it is essentially an extension of the independent samples test for a difference in means, extended to more than two groups.

Suppose there are $K \geq 2$ number of groups. The null hypothesis says that the means of all the groups are equal to each other. The alternative hypothesis says that *at least one* of the means differs from another mean.

Null: $\mu_1 = \mu_2 = \dots = \mu_K$

Alternative: $\mu_i \neq \mu_j$ for some $i \neq j$

2. Assumptions

The following assumptions are necessary for the one-way ANOVA test:

- **Random assignment:** Observations are randomly assigned to the K groups independently of the outcome variable.
- **Independent groups:** Observations in one group are independent of observations in all other groups. There must therefore be different sampling units in each group.
- **Homogeneity of variance:** The population variances of the outcome variable for the different groups are the same, even though the population means may be different.
- **Central Limit Theorem assumptions:** There is a sufficiently large sample size and/or the population distribution for the outcome variable is normal

3. Example: Current Population Survey

The data set below includes data from more than 52,000 individuals over the age of 25 years that participated in the 2016 Current Population Survey. The variables include the following:

- **HOURS:** Usual weekly hours worked over all jobs (ratio)
- **SEX:** Male or female (categorical)
- **RACE:** Racial identity (categorical)
- **EDU:** Education (categorical)
- **MARST:** Marital status (categorical)

The following code downloads the data, opens it, and saves it in a dataframe that we call `wdat`.

```
wdat <- read.csv("http://murraylax.org/datasets/cpshours.csv")
```

We can use this data to determine if the number of usual weekly hours is different for people in different categories for one or more of the categorical variables.

3. Estimating a One-Way Linear Model

The `lm()` function can be used to estimate several types of linear models including a one-way or multi-way ANOVA. The code below estimates a one-way ANOVA that predicts usual weekly earnings by race.

```
lmhours <- lm(HOURS~0+RACE , data=wdat)
```

The first parameter is a *formula* that tells `lm()` to estimate a linear relationship that predicts the outcome variable `HOURS` with the explanatory variable `RACE`. Because `RACE` is a categorical variable, `lm()` will treat each category as an explanatory variable.

Those who are familiar with using the `lm()` function to estimate a regression will find this code familiar. A possible difference is the inclusion of `0+` in the formula preceding the explanatory variable `RACE`. This is entered in the formula so that `lm()` does not estimate an intercept coefficient, but rather estimates the unique intercept (equal the mean in a one-way ANOVA) for each category in `RACE`.

We can see the estimates for the mean number of categories for each race by investigating the coefficients. The code below reports the coefficients and their 95% confidence intervals.

```
lmhours$coefficients
```

```
## RACEAmerican Indian/Aleut/Eskimo      RACEAsian/Pacific Islander
##                                39.44427                        40.40267
##                                RACEBlack                          RACEOther
##                                40.08817                        40.73537
##                                RACEWhite
##                                40.64012
```

```
confint(lmhours, level=0.95)
```

```
##                                2.5 %   97.5 %
## RACEAmerican Indian/Aleut/Eskimo 38.58930 40.29925
## RACEAsian/Pacific Islander      40.05854 40.74680
## RACEBlack                       39.79435 40.38199
## RACEOther                       39.97124 41.49949
## RACEWhite                       40.53238 40.74787
```

4. Conducting One-Way ANOVA Hypothesis Test

Variance Decomposition

The ANOVA procedure uses a *decomposition of variances* to determine how much variability in the outcome variable is explained by different assignments to the categorical variable, and how much variability is unexplained.

The **Between Group Variation** is a measure of explained variation, the measure of variability in the outcome variable that is explained by assignment to different categories of the explanatory variable. In our example, it is the measure of variability in usual weekly hours that is explained by a person being in a given racial category.

The **Within Groups Variation** is a measure of unexplained variation, also called residual variation. It is the measure of variability in the outcome variable that is seen *within* each category of the explanatory variable. In our example, it is the variability in usual weekly hours between individuals within any single racial group.

Finding Statistical Evidence

When *between groups variation* is sufficiently large compared to *within groups variation*, there is evidence to reject the null hypothesis and conclude there is sufficient statistical evidence that the outcome variable has a

different mean for the different categories of the explanatory variable.

Conducting the Test

We can call the `anova()` function to conduct the ANOVA test. This will estimate each of these measures of variability and determine if average usual weekly hours is different for people of different races.

```
anova(lmhours)

## Analysis of Variance Table
##
## Response: HOURS
##           Df    Sum Sq Mean Sq F value    Pr(>F)
## RACE       5 86008324 17201665  138020 < 2.2e-16 ***
## Residuals 52301  6518380     125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We pass in a single variable, the return object from the `lm()` function. We can see here that the p-value is very low (less than 2.2×10^{-16}). We reject the null hypothesis and conclude that the mean usual weekly hours is different depending on the racial group.

4. Post-Hoc Tests for Differences in Means

Having established that mean usual weekly hours is different for people of different races, we can take the analysis a step further and make pairwise comparisons of the means between different races. We can do this by running a series of independent samples t-tests for each pair of races. Since there are 5 racial categories, there are 10 different pairs, and therefore 10 different independent samples t-tests to run.

Remember that any hypothesis test can lead to incorrectly rejecting the null hypothesis by statistical chance. This is called a Type 1 error. The probability of making a type 1 error is equal to the significance level you use for the hypothesis test, usually 5%. When we run 10 independent samples t-tests, the probability of making a type 1 error in *at least one of the 10 tests* is larger. For example, if the chances of making a type 1 error are independent across the 10 statistical tests, the probability of making a type 1 error in at least one of the tests is more than 40%!

The **Bonferonni post-hoc** test is a series of independent samples tests for differences in means that makes a correction for p-values so that the probability of making one or more type 1 errors over the is not more than 5%.

The function call `pairwise.t.test()` that follows estimates independent samples difference in means tests for each pair of race categories, making a Bonferonni corrections to the p-values.

```
pairwise.t.test(wdat$HOURS, wdat$RACE, p.adjust.method="bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  wdat$HOURS and wdat$RACE
##
##           American Indian/Aleut/Eskimo Asian/Pacific Islander
## Asian/Pacific Islander 0.4154 -
## Black                 1.0000 1.0000
## Other                 0.2733 1.0000
## White                 0.0653 1.0000
##
##           Black Other
## Asian/Pacific Islander - -
```

```
## Black          -          -
## Other          1.0000 -
## White          0.0055 1.0000
##
## P value adjustment method: bonferroni
```

The only racial pair where the corrected p-value is less than 5% is white people vs. black / African American people, where the p-value is equal to 0.0055. We find sufficient statistical evidence that there is a difference in usual weekly hours between white people and African American people.