

Multi-Way Analysis of Variance (ANOVA)

*James M. Murray, Ph.D.
University of Wisconsin - La Crosse*

Updated: November 01, 2017

PDF file location: <http://www.murraylax.org/rtutorials/multiway-anova.pdf>

HTML file location: <http://www.murraylax.org/rtutorials/multiway-anova.html>

Multi-Way Analysis of Variance or **Multi-Way ANOVA** is a statistical method to estimate how the mean for an outcome variable depends on **two or more** categorical variables. It is an extension of the the one-way ANOVA that estimates how the mean for an outcome variable depends on a single categorical variable.

1. Assumptions

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

- **Random assignment:** The values that the categorical explanatory variables take are determined independently of the outcome variable.
- **Independent groups:** For every given categorical variable, the observations in one category are independent of observations in all other categories. There must therefore be different sampling units in each category for a given variable.
- **Homogeneity of variance:** The population variances of the outcome variable is the same for all possible realizations of the categorical variables.
- **Central Limit Theorem assumptions:** There is a sufficiently large sample size and/or the population distribution for the outcome variable is normal.

2. 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)
- **EDUC:** Education (categorical)
- **MARST:** Marital status (categorical)

The following code downloads the data, opens it, and saves it in a data frame called `wdat`.

```
load(url("http://murraylax.org/datasets/cpshours.RData"))
```

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 Multi-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 multi-way ANOVA that predicts usual weekly earnings using categorical variables, race, sex, and marital status.

```
lmhours <- lm(HOURS ~ RACE + SEX + MARST , 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 variables `RACE`, `SEX`, and `MARRIED`.

We can see the estimates for the coefficients almost every category of all the explanatory variables. The code below reports the coefficients and their 95% confidence intervals.

```
cbind(lmhours$coefficients)
```

```
##                [,1]
## (Intercept)    37.4211481
## RACEAsian/Pacific Islander  0.5924694
## RACEBlack      0.8454845
## RACEOther      1.2449087
## RACEWhite      0.8861092
## SEXMale        4.6993114
## MARSTSeparated/Divorced    0.3312965
## MARSTSingle    -0.3715599
## MARSTWidowed   -2.7509081
```

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

```
##                2.5 %    97.5 %
## (Intercept)    36.57529518  38.2670010
## RACEAsian/Pacific Islander -0.30888623  1.4938250
## RACEBlack      -0.03786126  1.7288303
## RACEOther      0.12482437  2.3649930
## RACEWhite      0.04370698  1.7285114
## SEXMale        4.51051566  4.8881072
## MARSTSeparated/Divorced    0.04240476  0.6201883
## MARSTSingle    -0.61492048 -0.1281992
## MARSTWidowed   -3.44527794 -2.0565382
```

Notice that each categorical variable is missing one category. We are missing the racial category for Native Americans, the sex category for females, and the marital status category for married people. These excluded categories form the basis for comparing means. The intercept equal to 37.42 is the mean usual weekly hours for a married, female, Native American.

The remaining coefficients are the difference in the usual weekly hours compared to this subgroup when one categorical variable is changed. For example, Asian people on average work 0.59 more hours per week than Native Americans. Single people on average work 0.37 *less* hours per week than married people.

4. Hypothesis Tests

4.1. 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 one of our categorical variables. In our example, we will have a measure of between group variation for each **RACE**, **SEX**, and **MARST**.

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 sub-category of the explanatory variables. In our example, it is the variability in usual weekly hours between individuals that is not explained by race, sex, or marital status.

4.2. Finding Statistical Evidence

For a given categorical variable, when the *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 based on the categorical variable.

4.3. Conducting the Tests

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       4   2435     609    5.1187 0.0004031 ***
## SEX        1 290616 290616 2444.0239 < 2.2e-16 ***
## MARST      3   9185     3062   25.7474 < 2.2e-16 ***
## Residuals 52297 6218579     119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values for every categorical variable are less than 5%. We can make the following three conclusions:

- We have sufficient statistical evidence that mean usual weekly hours is different for different races, accounting for other variables in the model (sex and marital status).
- We have sufficient statistical evidence that mean usual weekly hours is different for males versus females, accounting for other variables in the model (race and marital status).
- We have sufficient statistical evidence that mean usual weekly hours is different for people of different marital statuses, accounting for other variables in the model (race and sex).

5. Post-Hoc Tests for Differences in Means

Having established that mean usual weekly hours is different based on the outcomes of the categorical variables, we can take the analysis a step further and make pairwise comparisons of the means between categories. We can do this by running a series of independent-samples t-tests for differences in means. There are 17 such tests to run:

- Since there are 5 racial categories, there are 10 different pairs, and therefore 10 different independent-samples t-tests to run to identify differences by race.
- Since there are 2 sex categories, there is 1 independent-samples t-test to run to identify differences by sex.

- Since there are 4 marital status categories, are are 6 independent-samples t-tests to run to identify differences by marital status.

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 to identify differences by race, 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 is equal to 5% and these are independent across the 10 statistical tests, the probability of making a type 1 error in at least one of the tests is equal to $1 - (1 - 0.05)^{10} = 0.4013$, or more than 40%!

The **Bonferroni 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 calls `pairwise.t.test()` that follow estimate independent samples difference in means tests for each pair of race categories, each pair of marital status categories, and the only pair of sex categories, making a Bonferroni corrections to each for the p-values.

Post-hoc tests for differences by race

```
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
```

Post-hoc tests for differences by sex

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

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  wdat$HOURS and wdat$SEX
##
##      Female
## Male <2e-16
##
## P value adjustment method: bonferroni
```

Post-hoc tests for differences by marital status

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

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  wdat$HOURS and wdat$MARST
##
##           Married Separated/Divorced Single
## Separated/Divorced 0.007      -           -
## Single              8.9e-05 1.000      -
## Widowed             < 2e-16 < 2e-16    < 2e-16
##
## P value adjustment method: bonferroni
```

6. Interaction Effects

We may choose to allow two or more categorical variables to have **interaction effects**. Suppose for example that we suspect that marital status and sex have an interaction effect in determining usual weekly hours. This means that *in addition* to the difference in usual weekly hours explained by sex and marital status, the impact these two categorical variables make may depend on one another.

For example, while single people have fewer usual weekly hours on average than married people, the difference that getting married has on usual weekly hours may depend on whether a person is male or female.

The following code reruns the `lm()` procedure, allowing for an interaction effect between marital status and sex.

```
lmhours <- lm(HOURS ~ RACE + SEX + MARST + SEX:MARST, data=wdat)
```

The `SEX:MARST` part of the formula refers to the interaction effect.

We can now run the ANOVA test which will include hypothesis tests for each categorical variable like above, but also include a hypothesis tests for the interaction effect between sex and marital status.

```
anova(lmhours)

## Analysis of Variance Table
##
## Response: HOURS
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## RACE       4   2435     609    5.1436 0.0003852 ***
## SEX        1 290616 290616 2455.9304 < 2.2e-16 ***
## MARST      3   9185     3062   25.8728 < 2.2e-16 ***
## SEX:MARST  3   30503   10168   85.9253 < 2.2e-16 ***
## Residuals 52294 6188076     118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a very low p-value on the interaction effect, we do find sufficient statistical evidence that there is an interaction effect between marital status and sex. That is, the impact that marital status has on usual weekly hours does depend on sex. An equivalent way of stating this is the impact that sex has on usual weekly hours does depend on marital status.