QS101: Introduction to Quantitative Methods in Social Science

Week 17: Comparing Groups - Analysis of Variance (ANOVA)

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression

February 20, 2015

Assessment 2

One-Way ANOVA

Two-Way ANOVA

Assessment 2

Assessment: Task

- Which socio-demographic factors are you looking at?
- Which variables are you choosing for this?

One-Way ANOVA

- Stata's command for a one-way ANOVA is pretty straightforward

- Stata's command for a one-way ANOVA is pretty straightforward
- **`oneway`** *`depvar indepvar`*`,` **`bonferroni tabulate`**

- Stata's command for a one-way ANOVA is pretty straightforward
- **`oneway`** *`depvar indepvar`*`,` **`bonferroni tabulate`**
- Who is Bonferroni?

Who is Bonferroni?

- Bonferroni is one of many multiple-comparison tests

Who is Bonferroni?

- Bonferroni is one of many multiple-comparison tests
- So far, we know, that if the p-value is small, we know that the means are different

Who is Bonferroni?

- Bonferroni is one of many multiple-comparison tests
- So far, we know, that if the p-value is small, we know that the means are different
- What we do not know, is how different they are

Constructing Confidence Intervals

- To answer this question, we can estimate the population means and construct confidence intervals around them (see Section 12.2. in Agresti and Finlay for this)

Constructing Confidence Intervals

- To answer this question, we can estimate the population means and construct confidence intervals around them (see Section 12.2. in Agresti and Finlay for this)
- If we know these means and their confidence intervals, we can also say within which range the difference between the means lies

Constructing Confidence Intervals

- To answer this question, we can estimate the population means and construct confidence intervals around them (see Section 12.2. in Agresti and Finlay for this)
- If we know these means and their confidence intervals, we can also say within which range the difference between the means lies
- If we hypothesise that the means should be different, this range must not contain zero

Problems

- If we have many groups (for example, blacks, whites, Hispanis, etc.), we also have many comparisons

Problems

- If we have many groups (for example, blacks, whites, Hispanis, etc.), we also have many comparisons
- To be precise: $g(g-1)/2$ comparisons

Problems

- If we have many groups (for example, blacks, whites, Hispanis, etc.), we also have many comparisons
- To be precise: $g(g-1)/2$ comparisons
- For $g = 10$ we have 45 comparisons

Problems

- If we have many groups (for example, blacks, whites, Hispanis, etc.), we also have many comparisons
- To be precise: $g(g-1)/2$ comparisons
- For $g = 10$ we have 45 comparisons
- If we apply a 95% confidence interval, we would expect that $45 \times 0.05 = 2.25$ of the intervals would not contain the true differences of the means

Conclusion

- The larger the number of groups to compare, the greater is the chance of at least one incorrect inference

Conclusion

- The larger the number of groups to compare, the greater is the chance of at least one incorrect inference
- There are methods to correct this, they are called Multiple Comparisons of Means

Conclusion

- The larger the number of groups to compare, the greater is the chance of at least one incorrect inference
- There are methods to correct this, they are called Multiple Comparisons of Means
- Bonferroni is one such method

Conclusion

- The larger the number of groups to compare, the greater is the chance of at least one incorrect inference
- There are methods to correct this, they are called Multiple Comparisons of Means
- Bonferroni is one such method
- It adjusts the confidence intervals of each comparison of means upwards, so as to arrive at the overall desired level of confidence

Example

- We want 95% confidence level overall

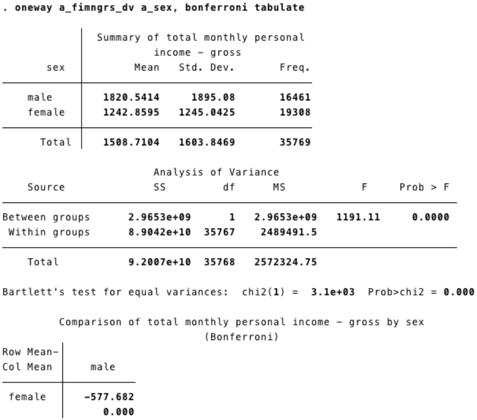
Example

- We want 95% confidence level overall
- We have 3 groups ($g = 3$), and hence $3(3-1)/2 = 3$ comparisons

Example

- We want 95% confidence level overall
- We have 3 groups ($g = 3$), and hence $3(3-1)/2 = 3$ comparisons
- Bonferroni would use error probability $0.05/3 = 0.0167$ for each interval

Output



```
. oneway a_fimngrs_dv a_sex, bonferroni tabulate

            |  Summary of total monthly personal
            |           income - gross
        sex |        Mean   Std. Dev.       Freq.
------------+------------------------------------
       male |   1820.5414     1895.08       16461
     female |   1242.8595   1245.0425       19308
------------+------------------------------------
      Total |   1508.7104   1603.8469       35769

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups     2.9653e+09      1   2.9653e+09   1191.11     0.0000
 Within groups     8.9042e+10  35767    2489491.5
------------------------------------------------------------------------
    Total          9.2007e+10  35768   2572324.75

Bartlett's test for equal variances:  chi2(1) =  3.1e+03  Prob>chi2 = 0.000

         Comparison of total monthly personal income - gross by sex
                                (Bonferroni)
Row Mean-|
Col Mean |       male
---------+-----------
  female |   -577.682
         |      0.000
```

Explanation

- First tabulation shows the mean income, standard deviation and frequency (here equal to n) for each sex

Explanation

- First tabulation shows the mean income, standard deviation and frequency (here equal to n) for each sex
- What can we learn from this?

Explanation

- First tabulation shows the mean income, standard deviation and frequency (here equal to n) for each sex
- What can we learn from this?
- What problem with regards to the standard deviation might occur?

ANOVA Table

- Source: within-groups variance should be small, between-groups variance should be large

ANOVA Table

- Source: within-groups variance should be small, between-groups variance should be large
- Why?

ANOVA Table

- Source: within-groups variance should be small, between-groups variance should be large
- Why?
- Between Group Mean Square is the estimated population variance based on differences between groups

ANOVA Table

- Source: within-groups variance should be small, between-groups variance should be large
- Why?
- Between Group Mean Square is the estimated population variance based on differences between groups
- Again, this should be large

ANOVA Table

- Source: within-groups variance should be small, between-groups variance should be large
- Why?
- Between Group Mean Square is the estimated population variance based on differences between groups
- Again, this should be large
- What does the p-value tell us?

Final table

- Bartlett's test for equal variance tests if the variances of the dependent variable are equal in both groups

Final table

- Bartlett's test for equal variance tests if the variances of the dependent variable are equal in both groups
- The data do not meet this assumption here

Final table

- Bartlett's test for equal variance tests if the variances of the dependent variable are equal in both groups
- The data do not meet this assumption here
- The test is less important with large samples, like this one, however, and is therefore often ignored

Your Turn!

- Look at your independent variables and perform an ANOVA with income for each of them

Your Turn!

- Look at your independent variables and perform an ANOVA with income for each of them
- Interpret the results of each table, and consider the implications for your research project

Your Turn!

- Look at your independent variables and perform an ANOVA with income for each of them
- Interpret the results of each table, and consider the implications for your research project
- Be ready to present some results for the seminar group.

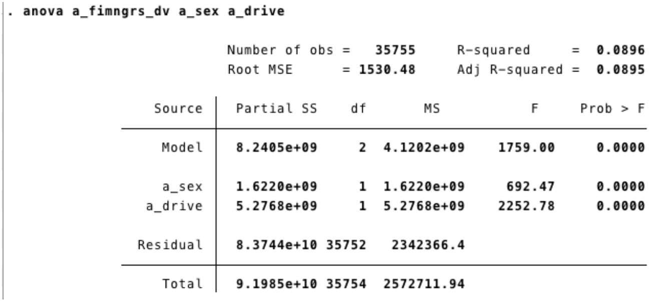
Two-Way ANOVA

- Again, the Stata command is straightforward

- Again, the Stata command is straightforward
- `anova` *`depvar indepvar1 indepvar2`*

- Again, the Stata command is straightforward
- **`anova`** *`depvar indepvar1 indepvar2`*
- This time, we only get one table

Example



```
. anova a_fimngrs_dv a_sex a_drive

                         Number of obs =     35755     R-squared     =  0.0896
                         Root MSE      =   1530.48     Adj R-squared =  0.0895

                  Source | Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model | 8.2405e+09     2  4.1202e+09    1759.00     0.0000
                         |
                   a_sex | 1.6220e+09     1  1.6220e+09     692.47     0.0000
                 a_drive | 5.2768e+09     1  5.2768e+09    2252.78     0.0000
                         |
                Residual | 8.3744e+10 35752   2342366.4
              -----------+----------------------------------------------------
                   Total | 9.1985e+10 35754  2572711.94
```

Queries

- Are our independent variables significant?

Queries

- Are our independent variables significant?
- What does this mean?

Queries

- Are our independent variables significant?
- What does this mean?
    - There is evidence that income varies by sex, within driving categories

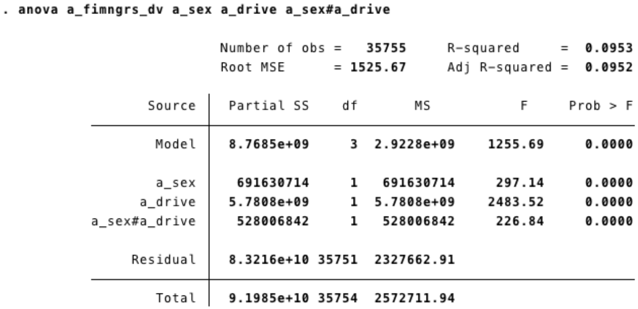
Queries

- Are our independent variables significant?
- What does this mean?
    - There is evidence that income varies by sex, within driving categories
- What's missing here?

Queries

- Are our independent variables significant?
- What does this mean?
  - There is evidence that income varies by sex, within driving categories
- What's missing here?
  - The interaction

Example



```
. anova a_fimngrs_dv a_sex a_drive a_sex#a_drive

                         Number of obs =     35755    R-squared     =  0.0953
                         Root MSE      =   1525.67    Adj R-squared =  0.0952

                  Source | Partial SS    df        MS          F     Prob > F
            -------------+----------------------------------------------------
                   Model | 8.7685e+09     3  2.9228e+09    1255.69     0.0000
                         |
                   a_sex |  691630714     1   691630714     297.14     0.0000
                 a_drive | 5.7808e+09     1  5.7808e+09    2483.52     0.0000
           a_sex#a_drive |  528006842     1   528006842     226.84     0.0000
                         |
                Residual | 8.3216e+10 35751  2327662.91
            -------------+----------------------------------------------------
                   Total | 9.1985e+10 35754  2572711.94
```

Interpreting the Interaction

- Is the Interaction significant?

Interpreting the Interaction

- Is the Interaction significant?
- What does this mean?

Interpreting the Interaction

- Is the Interaction significant?
- What does this mean?
  - We reject H_0: no interaction

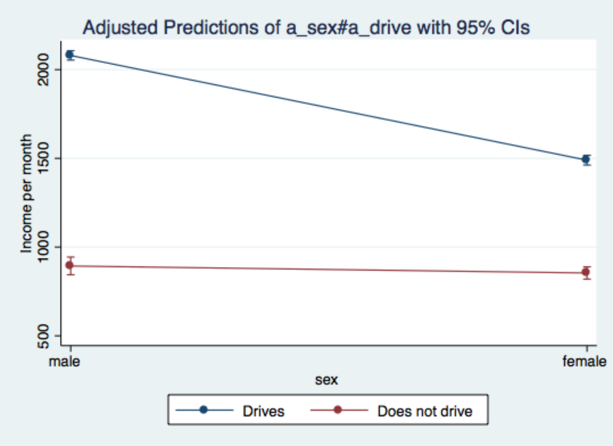
Interpreting the Interaction

- Is the Interaction significant?
- What does this mean?
  - We reject H_0: no interaction
  - Conclusion: each variable has an effect, but the nature of that effect changes according to the category of the other variable

Interpreting the Interaction

- Is the Interaction significant?
- What does this mean?
  - We reject H_0: no interaction
  - Conclusion: each variable has an effect, but the nature of that effect changes according to the category of the other variable
  - A comparison of means would be sensible here

Example


Adjusted Predictions of a_sex#a_drive with 95% CIs
Income per month
2000
1500
1000
500
male
female
sex
Drives
Does not drive

Your Turn!

- Look at your independent variables and perform an ANOVA with income for different pairs of them

Your Turn!

- Look at your independent variables and perform an ANOVA with income for different pairs of them
- Interpret the results of each table, and consider the implications for your research project

Your Turn!

- Look at your independent variables and perform an ANOVA with income for different pairs of them
- Interpret the results of each table, and consider the implications for your research project
- Be ready to present some results for the seminar group.