Module 3: ANOVA and Frequency Analysis

# Reading

*PROJECT: Work on creating a preliminary analysis in SAS of your proposed data. This could be a basic test, visualization, or tackle some challenges with loading, cleaning, and filtering your data. Be prepared to share (via Slack) and talk briefly about your project plan during our meeting on 10/31).*

*Read*

* *(Cody) Chapter 9: Comparing More Than Two Means (ANOVA)*
* *(Cody) Chapter 10: N-Way ANOVA*
* *(Cody) Chapter 14: Analyzing Categorical Data*
* *(OpenIntro) Chapter 6: Inference for categorical data*

*Watch (~1.5 hour total)*

* [*https://youtu.be/\_VFLX7xJuqk*](https://youtu.be/_VFLX7xJuqk) *-- Introduction to One-Way ANOVA*
* [*https://youtu.be/-AeU4y2vkIs*](https://youtu.be/-AeU4y2vkIs) *-- ANOVA and Sums of Squares*
* [*https://youtu.be/k-xZzEYL8oc*](https://youtu.be/k-xZzEYL8oc) *-- F-Statistic and p-value*
* [*https://youtu.be/pscJPuCwUG0*](https://youtu.be/pscJPuCwUG0) *-- ANOVA and Multiple Comparisons*
* [*https://youtu.be/pfc9MUz03XA*](https://youtu.be/pfc9MUz03XA) *-- Chi Squared Test*

*Exercises:*

* *(Cody) Chapter 9 -- Exercise 1 (pg. 109)*
* *(Cody) Chapter 14 -- Exercise 1 and 2 (pg. 196)*
* [*DUE Saturday October 31, 2020 @ 9:00AM (LINK)*](https://elearn.stonehill.edu/webapps/assignment/uploadAssignment?content_id=_634065_1&course_id=_16488_1&group_id=&mode=cpview)

# Discussion for this week:

*Prompt:*

1. *Post at least one (1) comment on someone else’s project post. Offer helpful feedback, suggestions, or (constructive) criticism that will help to improve their work or clarify their reporting.*
2. *Post with one thing that has challenged you about learning SAS (and the various SAS interfaces) in this program and one thing that you find really useful. This is a good time to reflect on what is working and what still needs work.*

POST RESPONSE TO [#dan602\_discuss](https://stonehillmpsda21.slack.com/archives/C01A1MZUGDB) in Slack by November 8 @11:59PM

# SAS Workshop: One Way ANOVA & Multiple Comparison Hypothesis Testing

*Goal: Use ANOVA to test whether people in different body mass index (BMI) categories have different average heights (i.e., obese people are taller).*

H0: There is no difference in average height across 6 categories of body mass index (Extremely Weak, Weak, Normal, Overweight, Obese, and Very Obese).

HA: **(Write a valid alternative hypothesis here in your own words)**

*Additional Stats Guidance:*

One-way ANOVA -- <http://sites.utexas.edu/sos/guided/inferential/numeric/onecat/more-than-2/more-than-two-groups/anova/>

Multi-way ANOVA and GLM -- <http://sites.utexas.edu/sos/guided/inferential/numeric/glm/>

*Data:*

*Information:* [*https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex*](https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex)

*File:* [*https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex?select=500\_Person\_Gender\_Height\_Weight\_Index.csv*](https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex?select=500_Person_Gender_Height_Weight_Index.csv)

*Methods:*

One-Way ANOVA

1. Download and import the 500\_Person\_Gender\_Height\_Weight\_Index.csv data
2. Set up a One-Way ANOVA
   1. Tasks and Utilities 🡪 One-Way ANOVA
   2. Set the dependent Variable as “Height”
   3. Set the categorical variable as “Index” \*these are categorical values for Body Mass Index
   4. Run
3. Review Results tables **(Record Answers)**
   1. What is the F Value for this test? (Bonus: How is this calculated?)
   2. What is the p-value?
   3. Copy the “Distribution of Height” plot here. With what you know now from the ANOVA how do you interpret the boxplots here?
   4. Look at the Adjustment for Multiple Comparisons: Tukey-Kramer tables. (These show the ranked group means and the Tukey-Kramer test result p-values).  
        
      Which BMI categories have average heights? Interpret these results in plain English.

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

N-Way ANOVA with Gender(Index)

1. Tasks and Utilities 🡪 Statistics 🡪 N-Way ANOVA

What is the hypothesis being tested now?

(Hint: This model expands the groups (M/F) for each Index so we end up with ngender \*nindex)

1. Setup
   1. Set dependent variable as “Height”
   2. Set categorical variable as “Index”
   3. In the Model tab Edit the model, “Add” Index to the model effects and “Nest” Gender with Index. Set Index as the Outer term and Gender as Nested within Outer. The final model should look like: “Index(Gender)”.
   4. In the Options tab, choose “Default and additional statistics”. Check “Perform multiple comparisons” and for “Which effects” choose “All effects”. The rest of the defaults are OK.
2. Run
3. Review Results Tables **(Record answers here)**
   1. What is the F-value and p-value for this test? What does this mean for your hypotheses?
   2. Copy the “Interaction Plot for Height” here and write an interpretation that is consistent with your hypothesis interpretation for the ANOVA.
   3. Look at the Adjustment for Multiple Comparisons: Tukey-Kramer tables. Now these show ALL pairwise mean comparisons with adjustments made to the p-values.   
        
      Which BMI category/Gender pairs have the highest average heights? Don’t try to interpret this entire table. Which comparisons stand out within these results?

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

# SAS Workshop: Chi-Squared Testing

*Goal: Perform frequency analysis and chi-squared testing. Interpret your results.*

*Data:* [*https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data*](https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data)

*Additional Stats Guidance:*

[*http://sites.utexas.edu/sos/guided/inferential/categorical/univariate/chi2/*](http://sites.utexas.edu/sos/guided/inferential/categorical/univariate/chi2/)

[*http://sites.utexas.edu/sos/guided/inferential/categorical/chi2/*](http://sites.utexas.edu/sos/guided/inferential/categorical/chi2/)

*Methods:*

Chi-Squared Goodness of Fit Test (1-way categorical analysis)

1. To test the following hypothesis:  
     
   H0: Women and men commit crime at the same rate in NYC.  
   HA: Women and men to not commit crime at the same rate in NYC.  
     
   We are going to perform a basic Chi-squared goodness of fit test using the SUSP\_SEX variable.
2. Import the data (\*\*This one may yield an error, but the data should read with defaults\*\*)
3. Tasks and Utilities 🡪 One-Way Frequencies
   1. Analysis variables as “SUSP\_SEX”
   2. Under Options 🡪 Statistics; choose Chi-squared Goodness-of-Fit, Asymptotic test  
      \*Consider Exact test only for smaller sample sizes (<1000), but results usually do not differ
4. Run
5. Interpret results  
   1. Look at the results table “Chi-Square Test for Equal Proportions”. How do you interpret these results?
   2. Compare the plots “Deviations of SUSP\_SEX” and “Distribution of SUSP\_SEX”. What are these plots showing you given the results of the Chi-squared test?

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

Chi-Squared Test of Independence (2-way categorical analysis)

**Design new hypothesis statements for a Chi-squared test of independence for male suspects comparing SUSP\_AGE\_GROUP and SUSP\_RACE.**H0:

HA:

1. Filter the NYC crime table for only those rows with:

SUSP\_SEX equal to M

SUSP\_RACE equal to WHITE or BLACK

SUSP\_AGE\_GROUP equal to ’18-24’ OR ’25-44’ OR ’45-64’

proc sql noprint;

create table WORK.filter as select \* from WORK.IMPORT where(

SUSP\_SEX EQ "M"

AND (SUSP\_RACE EQ "WHITE" OR SUSP\_RACE EQ "BLACK")

AND (SUSP\_AGE\_GROUP EQ "18-24"

OR SUSP\_AGE\_GROUP EQ "25-44"

OR SUSP\_AGE\_GROUP EQ "45-64"

)

);

quit;

1. Tasks and Utilities 🡪 Table Analysis
2. Enter
   1. Row variables: SUSP\_AGE\_GROUP
   2. Column variables: SUSP\_RACE
3. In the Options tab select
   1. Percentages 🡪 Row, Column
   2. Do NOT Suppress plots
4. Run
5. Interpret Results  
   1. Look at the Statistics for Table of SUSP\_AGE\_GROUP by SUSP\_RACE. What can you conclude about your hypotheses?
   2. Compare the two plots titled “Distribution of SUSP\_AGE\_GROUP by SUSP\_RACE”. Which one is more informative to you? What do you think is going on given your hypotheses and test results?

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

# Post-Class:

\*\*Complete any remaining analyses from today’s SAS Lab\*\*

**Project Draft Report**

Submit to eLearn a ~2-4 page report of your project including background info (with references), data collection methods, preliminary stats, visualizations, and conclusions/results.

*Draft Format:*

*(NOTE: These sections may be very short, the entire report could be 2-4 pages, double spaced, size 11 font).*

*Abstract – 1 paragraph summary of the study*

*Introduction – Present relevant background information about your study system including any previous data analysis you have found that relates to your question.*

*Methods – Describe data collection methods (as much as possible. \*If you are using publicly available data you can cite the dataset and describe what the authors did to put it together\*). Write a paragraph describing the statistics you are going to present.*

*Results & Conclusions – What did you find? Write your conclusions supported by the results (and visualizations/plots) of your analyses.*

*References – Formal citations using APA style (*[*https://owl.purdue.edu/owl/research\_and\_citation/apa\_style/apa\_formatting\_and\_style\_guide/general\_format.html*](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/general_format.html)*) or another standard format appropriate to your field.*

**DUE: November 14, 2020 @ 11:59**