

Directions: Complete the exercises below. When you are finished, turn in any required files online in Canvas, then check-in with the Lab TA for dismissal.

ANOVA in SAS

Consider the donut example we introduced in lecture. The purpose of this experiment is to compare the amount of oil absorbed (`absorbed`) when cooking donuts in four different types of frying oil (`oil`). The dataset is contained in the `donuts.txt` file located in our course's shared folder in SAS Studio.

- First, read in the data set:

```
data donuts;
    infile '~/my_shared_file_links/u63538023/STAT5000_Fall2024_ISU/donuts.txt'
        dlm=', ' firstobs=2;
    input oil $ absorbed;
run;
```

- Then, make a scatterplot of the amount of oil absorbed for each type using the `sgplot` procedure with the `scatter` option with the group variable on the x-axis and the response variable on the y-axis:

```
title 'Donut Oil Absorption';
proc sgplot data=donut;
    scatter x=oil y=absorbed / markerattrs=(size=15 symbol=diamond color=black);
    yaxis label="Oil Absorbed - 150 grams" values=(-5 to 50 by 5)
        labelattrs=(size=15) valueattrs=(size=15);
    xaxis label="Type of Oil" labelattrs=(size=20) values=(1 to 4 by 1)
        valueattrs=(size=15) offsetmin=0.10 offsetmax=0.10;
run;
```

- Next, compute some summary statistics of the amount of oil absorbed for each type using the `means` procedure:

```
proc sort data=donut;
    by oil;

proc means data=donut n mean std var stderr min max;
    by oil;
    var absorbed;
run;
```

- Finally, find the ANOVA table and conduct the F -test using the `glm` procedure where the `class` option specifies the grouping variable and the `model` option is of the form `response = group`:

```
proc glm data=donut;
    class oil;
    model absorbed = oil;
run;
```

Checking ANOVA Assumptions in SAS

- First, conduct ANOVA using the `glm` procedure with some additional options¹:

```
proc glm data=donut plots=(diagnostics residuals) plots=boxplot;  
  /* To obtain individual residual plots, replace  
  plots=(diagnostics residuals) with plots(unpack)=residuals */  
  class oil;  
  model absorbed = oil / p solution xpx inverse ;  
  means oil /hovtest=bf;  
  lsmeans oil / stderr cl pdiff plots=none;  
  output out=set2 residual=r predicted=yhat;  
run;
```

The additional options will show more details of the model, including:

- The first 2 tables are standard output for the `glm` procedure.
- If you look at the 4×4 submatrix of the third *The $X'X$ Matrix* table (a result of the `xpx` option), you will see the $\mathbf{X}^T\mathbf{X}$ matrix shown on slide 14 of the Week 5 notes with the sample sizes along the diagonal and zeros on the off-diagonals.
- The fourth *$X'X$ Generalized Inverse ($g2$)* table (a result of the `inverse` option) shows $(\mathbf{X}^T\mathbf{X})^{-1}$.
- The fifth through eighth tables are the same as the usual ANOVA output.
- The ninth table (a result of the `solution` option) provides estimates of all model parameters (the first row for *Intercept* corresponds to μ and the following rows provide estimates of the effects $\alpha_1, \dots, \alpha_r$).
- The tenth table (a result of the `p` option) shows the observed, predicted, and residual values associate with every single data point.
- The final 4 tables in the overall output are a result of the `lsmeans` options show information for comparing the group means to each other (we will talk more about this next week).

You will also see information relevant to diagnosing the ANOVA assumptions, including:

- The first set of plots (a result of the `plots=(diagnostics residuals)` option), show different plots of the residuals which can be used to assess the equal variance and normality assumptions.
- The second plot (standard output for the `glm` procedure), shows the standard side-by-side boxplot of the responses within each group, which can be used to assess the equal variance and normality assumptions.
- The twelfth table (a result of the `hovtest=bf` option) provides the output of the Brown-Forsythe test for equal variances.
- The third plot and thirteenth table (a result of the `plots=boxplot` option), is a repeat of the standard side-by-side boxplot with accompanying summary statistics which can be used to assess the equal variance assumption.

¹Additional details of the options and their corresponding output can be found in the SAS User Guide:
https://documentation.sas.com/doc/en/statug/15.2/statug_glm_syntax01.htm

- Next, you can assess the equal variance assumption for the residuals using the scatterplot of the residuals from the `sgplot` procedure (which is made possible by saving the output of the ANOVA in the previous procedure):

```

title "Plot Residuals vs. Oils";
proc sgplot data=set2;
    scatter x=oil y=r / markerattrs=(size=15 symbol=diamond color=black);
    yaxis label="Residuals" values=(-20 to 25 by 5)
        labelattrs=(size=17) valueattrs=(size=15);
    xaxis label="Type of Oil" labelattrs=(size=17)
        values=(1 to 4 by 1) valueattrs=(size=15)
        offsetmin=0.10 offsetmax=0.10;
run;

```

If the equal variance assumption holds, then you should see no patterns or trends in the residuals (they should look like random noise).

- Finally, you can further assess the normality assumption using the `univariate` procedure on the residuals:

```

proc univariate data=set2 normal;
    var r;
    qqplot r / normal(mu=est sigma=est) square;
run;

```

The output will show plots, summaries, and tests, including:

- The first table contains the measures of skew (*Skewness*) and excess kurtosis (*Kurtosis*), which should both be near 0 if normality holds.
- The second table contains a comparison of the mean and median, which should also both be near 0 if normality holds.
- The fourth table shows the various tests for normality of the residuals.
- Lastly, the normal Q-Q plot of the residuals should follow the diagonal reference line if normality holds.

Kruskal-Wallis Test in SAS

- Because the Kruskal-Wallis test is just the multi-group extension of the Wilcoxon rank-sum test, the procedure in SAS is the same. You can conduct the test using the `npar1way` procedure with the `wilcoxon` option, where `class` specifies the grouping variable and `var` the response:

```
title 'Kruskal-Wallis test';
proc npar1way data=donut wilcoxon;
    class oil;
    var absorbed;
    exact wilcoxon / mc n=50000;
run;
```

The output will include:

- A table showing the summary statistics for each group, including the sums of the ranks in the *Sum of Scores* column.
- The next table will provide the Chi-squared test-statistic and associated p -value for the approximate test.
- If you used the `exact wilcoxon` option, the last table will provide an estimate of the exact p -value estimated from a random selection of all possible assignments of ranks to groups.
- Lastly, the output will show a side-by-side boxplot of the ranks (or *Scores*) within each group to help visualize whether or not there is a difference in the distributions.

Assignment

1. [+11] Refer to the ANOVA results for the donut example to complete the following exercises:
 - (a) Construct a scatter plot of the absorption to visualize potential differences between oil treatment groups. Include a screenshot of the plot.
 - (b) Compute the relevant summary statistics to compare the amount of oil absorbed for each type. Include a screenshot or record the means and standard deviations.
 - (c) Conduct the F -test for the equality of the four means. Provide a screenshot of the full ANOVA table for the different types of cooking oil and absorption measurements.
 - (d) Given the results in part (c), state the null and alternative hypotheses, test statistic, p -value, and interpretation of the result in the context of the study.
2. [+6] Refer to the process of diagnosing the ANOVA assumptions for the donut example to complete the following exercises:
 - (a) Discuss the independence assumption in the context of the study. Is it satisfied for the donut example?
 - (b) Check the assumption of equal variances and provide a screenshot of any relevant tables and/or figures that you used. Is it satisfied for the donut example?
 - (c) Check the assumptions of normality and provide a screenshot of any relevant tables and/or figures that you used. Is it satisfied for the donut example?
3. [+8] Refer to the results of the Kruskal-Wallis test for the donut example to complete the following exercises:
 - (a) Conduct the Kruskal-Wallis test to analyze the different types of cooking oil on the oil absorbed. Provide a screenshot of the results obtained.
 - (b) Given the results in part (a), state the null and alternative hypotheses, test statistic, and p -value.
 - (c) Interpret the result from part (b) in the context of the study.

Total: 25 points **# correct:** _____ **%:** _____