Module 1

Assignment 1: Applied Statistics and Design of Experiments

Here we become reintroduced to statistical analyses to test for differences between treatments and then proceed to DOE, Design of Experiments. Assignment 1 is the ungraded, practice assessment and Assignment 2 is your graded assessment. See Grading Exercises rubrics for how your report will be graded. Please format the report as described in the 'Tips' document (with the headings: Introduction, Methods, Results, Discussion, Conclusion). Keep it simple for Module1, since it is mainly review (You may role-play to invent a goal, hypothesis etc).

Univariate Statistics

MVA for chemometrics usually involves a specific type of data set- a few independent, categorical variables (things you set prior to the start of the experiment, like 'Product' and 'Replication') and *many* dependent, continuous variables (things you measured, like NMR data). However, 'traditional' statistical tools are often designed to deal with datasets that have some or many independent categorical variables and *one* continuous dependent variable- this is univariate statistics. We will learn in the course that doing MVA *first* allows you to remove the many unimportant dependent variables in your study, so you can then focus on the key dependent variables that matter the most. However, for now we will review univariate statistics, so you know what to do once we start getting results from MVA.

Hypothesis testing

Often the main goal in statistics is to determine if there are differences between treatments when measuring just one continuous dependent variable. In practice this involves hypothesis testing and ANOVA or a non-parametric test, depending on if the data is normally distributed or not. These tests compare variation within treatments to variation between treatments to test for differences. Post-hoc tests are then done to determine what those differences are.

- Load 'impurities_24.csv'. Note the four independent variables: day, filtration, surfactant and sonication.
- Assume the data is normally distributed and carry out an ANOVA followed by a barchart comparing means as the posthoc test. Follow up with t-tests on the levels.
- Assume the data is non-parametric and carry out Kruskal-Wallis followed by a boxplot comparing medians as the posthoc test. Follow up with Mann-Whitney tests on the levels.
- Now see which test you should have used by testing for normality using a histogram.

Briefly introduce the two analyses and insert your figures. Briefly discuss the outcomes. After testing for normality, should you use ANOVA or Kruskal-Wallis? If the experiment was about using filtration to purify a beneficial vitamin from fruit pulp, where the response is a measure of purity (higher the better),

what might a false positive and false negative mean in this context? Extra analyses could involve other normality tests (QQ plots and the Shapiro-Wilks/Anderson-Darling tests), testing for equal variances (homoscedasticity), making cloud plots and exploring interactions between the independent variables using ANOVA-N.

Univariate regression

We have thus far treated filtration as a categorical variable, but it is also numerical (not just a category such as 'low' or 'high'). Since its numeric, you also carry out univariate regression analysis.

- Plot filtration and the response:
    - Add a linear regression line using *scipy stats.linregress*
    - Identify p_value of the regression to determine significance
    - Identify r2 of the regression and plot it on the figure
    - Create CI's for the slope and intercept

Insert your figure and CI table. Was the regression significant? Was it significant based on confidence intervals? What does this regression analysis provide that the ANOVA does not? An extra test would be to calculate the CI for a specific treatment, say Filtration75 and perhaps elaborate on the linear model (ie What is the actual model? What might a significant intercept mean?).

Replication and significance

To see if there are differences between treatments, we need to know the variation within the treatments, thus replication becomes very important. In the impurities data, Day can be considered as replication (rep1 and rep2).

- Erase 'Day 2' from impurities.csv and rerun the linear regression. What happens? What does it say about the variation between sampling days? An extra analysis might be to repeat the entire study for Day1 and Day2 separately and then compare the two studies.

Univariate stats summary

We now have reviewed the basic analyses done on a *single* response variable, testing for normality and then choosing the correct test to see if there are differences between treatments. If there are differences, one proceeds to post-hoc tests to find what those differences are. If the independent variable is numeric, one can also test for significance using univariate regression. Replication is important if you want to find significant differences between treatments.

.

Design of Experiments (DOE)

DOE is started before any data is collected. Imagine wanting to identify the Pressure, Temperature and Flowrate for the synthesis of a product (highest amount). The literature gives you some ranges (40-70 Pa, 290-350 C and 2-4 m3s), but you do not know the best combination to use. Thus, you want to do a preliminary study to identify the best parameters to use, before the larger, main experiment. You think about three types of design- one that samples all possible combinations of your levels of each factor (full factorial), one that is smaller, using less samples and is limited in scope (fractional factorial), and one that is smaller, but to some extent compensates for reduced sampling (CC design). You decide to use the following:

Pressure	Temperature	Flowrate
40	290	2
55	320	4
70	350	

- Use *build* from *doepy* to create design figures (3d plot with each axis one of your factors) and write the results to a csv file for the FullFac, FracFac and CC designs. Briefly describe the three designs in terms of pros and cons. What is special about the CC design (examine the output file)?

An example of a screening experiment

Imagine an experiment to understand the relations between pH and Speed in determining the viscosity of a product. You want to find the variable that best shows a linear relationship with viscosity, while holding the other constant. You know the approximate ranges to study (in real-life you would base this on the literature):

pH: 3 to 9

Speed: 20-90 rpm

You will do this in two steps- first do small screening experiments to see the overall response of *both* variables on viscosity. Consider some of the screening designs you did above. Repeat screens until you not only approximate the conditions for highest viscosity, but also identify one variable that gives a linear response, while holding the other constant. Then do a main experiment that allows you to do theoretical downstream statistical analyses (such as ANOVA and linear regression). Therefore, the main experiment should have replication and at least 3 levels of the variable you chose. You have a total of 30 experimental units to divide between the screening and the main experiments. See the video for details.

Present the results. What variable did you choose and why? Do the parameters chosen show a good linear response and include maximal viscosity? Extra studies would use what we learned in the first part to do down-stream analyses on your final results.

Leverage

We often have a dilemma deciding between assigning experimental units to different factors, to more levels of a given factor, or to more reps of a given factor-level combination. Where should we put them for the best effect? This brings up the concept of leverage.

- Use 'leverage.csv' to determine the leverage each level has on the study.
- Rearrange/change the original levels to improve on the leverage.
- Plot the two leverage analyses on the same graph. What seems to be the best strategy to use and why?
- Extra studies might include multiple tests to optimize leverage and a short, concise description of the outcome that speculate on the pros and cons of each design.

Assignment 2

Assignment2 is graded (see the rubric). Include the headings: Introduction, Methods, Results, Discussion, Conclusion, and possibly References and Appendix too. Follow the 'Tips' guidelines found in the Grading section of Canvas.

You are asked to describe the relationship between viscosity of a product, the pH of your substrate and the processing speed used. Explore the data, compare the treatments and develop a linear model for *either* pH or speed, depending on which best predicts viscosity in a linear model. The ranges to study are:

pH: 3 to 9

Speed: 20-100 rpm

You can use up to 30 experimental units (ie 30 individual treatments, each with a defined pH and speed). You may choose to do either one big experiment or *any number* of smaller screening experiments first. Run the program and collect the results- evaluate the response and continue to run until you do not have any experimental units left, or you feel you have completed data collection. Use your results for down-stream analyses presented in Assignment 1. Consider the aforementioned extra analyses to improve your grade. Grades do not depend on getting the highest viscosity, but rather the path you take in your analyses.

Consider:

Pre-experiment goals (before running the app):

- Present a short, rational argument for your DOE strategy. Feel free to role-play to help with goals, etc.
- Show your initial experimental design.

During experiment goals: Here you are trying to find the pH-speed combination that provides the highest viscosity *and* identifies *the one factor* that best can linearly model the data. So, a few small screening tests might be a good way to find the best factor, then use the remaining experimental units in the final main experiment to fully describe that best factor, while keeping the other factor constant.

- Based on each screening experiment results, present a *short*, rational argument for how and why you designed subsequent screens.
- Evaluate the main experiment design before running.

Post-experiment goals: At this point you should have results describing *one* of the factors well (ie pH *or* speed with the other held constant) in terms of viscosity. Carry out the following:

- Describe the experiment from the perspective of goals and hypothesis testing.
- Briefly discuss the levels you chose and why.
- For the factor of interest, are there significant differences between the levels chosen?
- Describe the treatment (pH-speed combination) that gives the highest viscosity.
- Can you develop a linear model for predicting viscosity for your chosen factor?
- Describe what you might do to improve your results in a future experiment.