

# Improving Student Statistical Skills Via An Interactive Web-based Portal



Tod D. Romo and Alan Grossfield  
Pls: Lynne E. Maquat and Jeffrey J. Hayes  
Department of Biochemistry and Biophysics  
University of Rochester Medical Center  
Rochester, NY, 14642



http://membrane.ums.rochester.edu



http://membrane.ums.rochester.edu/jupyter/

## Abstract

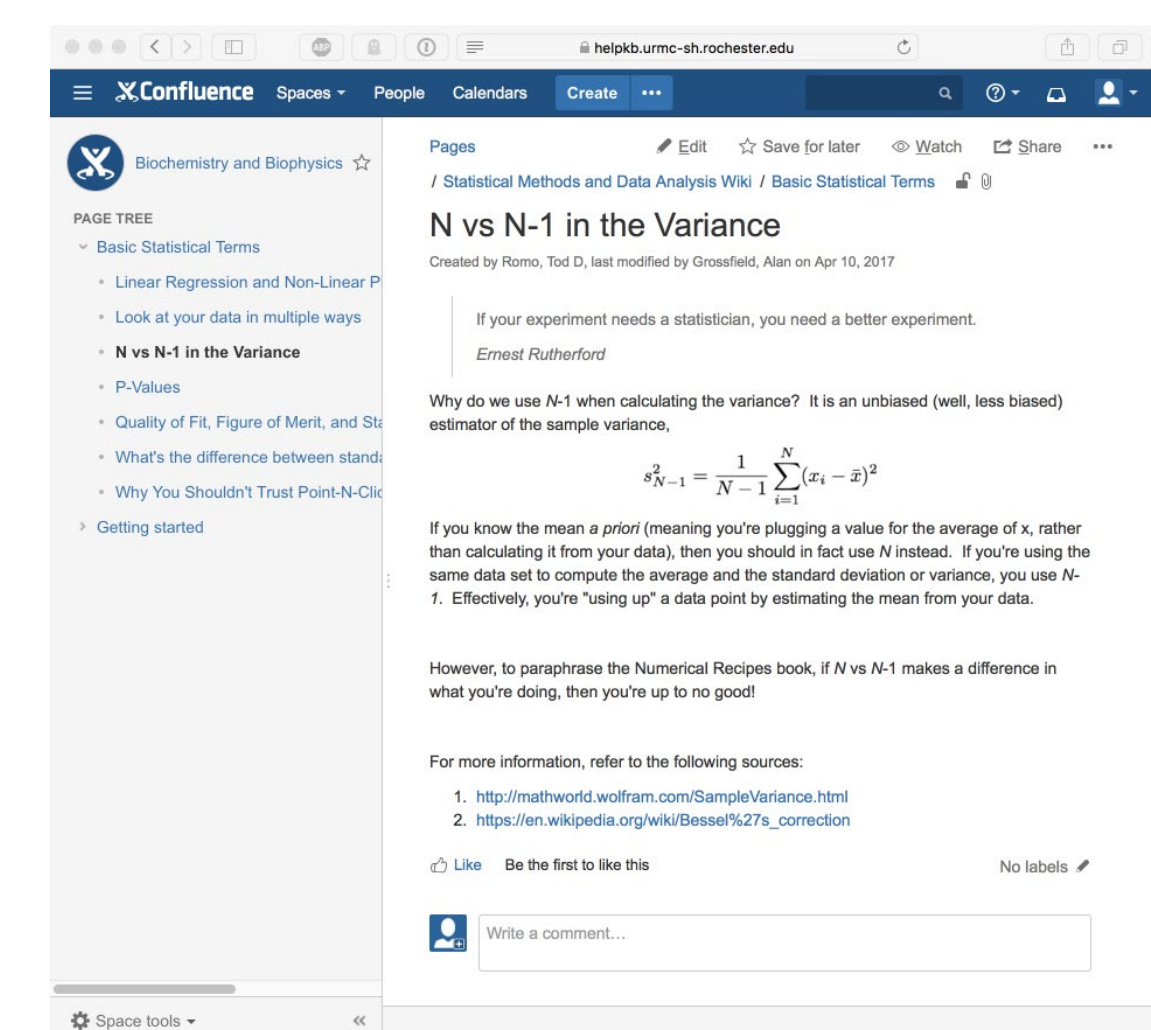
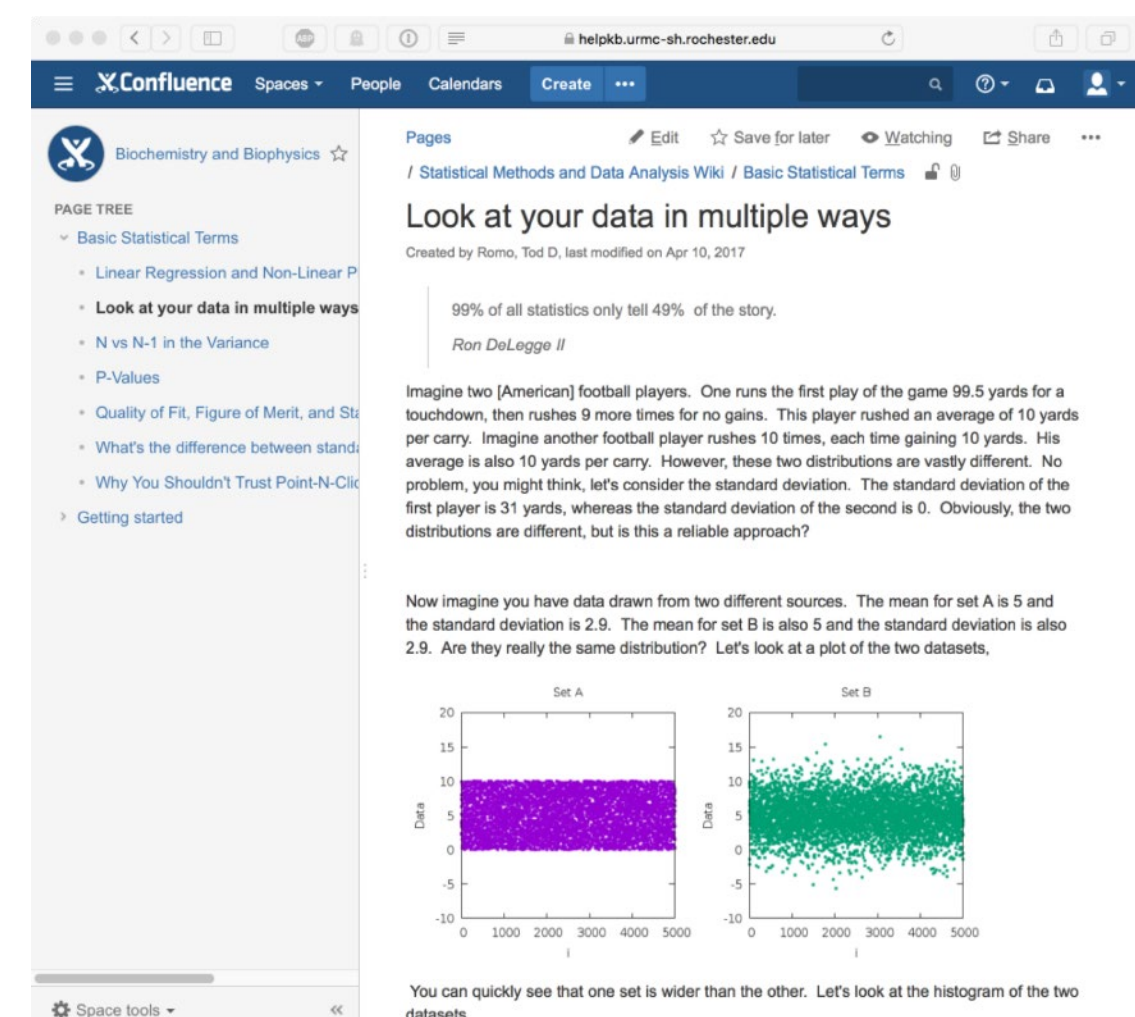
A general lack of statistical awareness is one of the major challenges facing the scientific workforce. Consequently, developing better ways to train graduate students in the sciences is crucial to helping them reach their potential in their future careers. Moreover, it is clear that this training must reach beyond the classroom in order to help students apply the lessons directly to their research. Accordingly, we have created a statistics and data analysis web portal to allow students to learn the best ways to handle their data. Our approach is three-pronged:

- Answer common questions in clear, non-specialist language, while providing links to more rigorous treatments for those with specialized interests
- Provide interactive demonstrations of key concepts to help students build intuition by “playing” with data
- Encourage users to take ownership by facilitating student comments and authorship on the site.

Finally, we recognize that certain disciplines (e.g. proteomics) have very specific data-handling requirements; for these cases, we will recruit local experts to create articles and link to authoritative external sites.

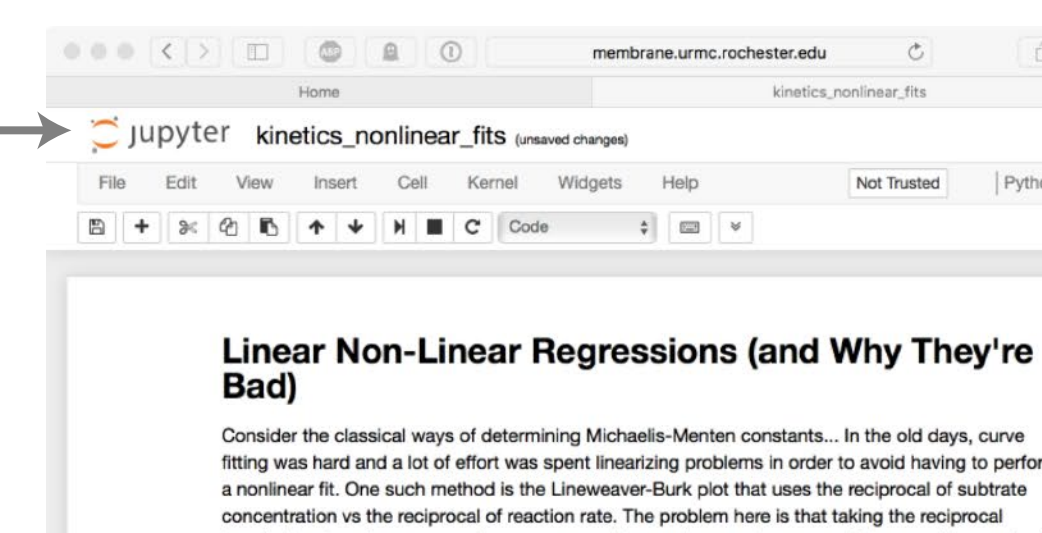
## Statistics Wiki

- Make it easy to find relevant information
- Wiki format, editable by anyone within University
- Proctored by faculty from the T32
- Students encouraged to create new posts:
  - Describe successful applications to their own research
- Include links to outside sources
- Leverages students as a creative workforce



## Interactive Lessons

Runs in Web Browser  
Uses Jupyter (iPython)  
"Temporary Notebook  
Server" Provides  
Resources



Python Code in  
Grey Boxes can  
be Ignored or Edited  
by Student

Can be Replaced with  
Student Data

Linear Non-Linear Regressions (and Why They're Bad)

Consider the classical ways of determining Michaelis-Menten constants. In the old days, curve fitting was used and a lot of effort was spent choosing procedures in order to avoid having to perform a nonlinear fit. One such method is the Lineweaver-Burk plot that uses the reciprocal of substrate concentration vs the reciprocal of reaction rate. The problem here is that taking the reciprocal inverts the size of measurement errors, so small errors become large ones. Moreover, the constants are determined by the axis intercepts, which may be some distance from the data and extrapolation over long distances is always risky. Another method is the Hanes-Woolf plot which plots the substrate concentration divided by the reaction rate against the substrate concentration. The problem here is that both axes depend on the substrate concentration, so typical methods of error estimation are no longer valid.

These methods are still taught today because they provide an intuition into problem and because it's necessary to understand how the kinetics constants were found in older papers. However, we now have access to fast and accurate computing. A much better solution is to use nonlinear fitting. For example, setup the Michaelis-Menten equation in graphpad and use fit function, and you'll get a very clean, very accurate constant estimates with error estimates that actually mean something. For a more detailed comparison of the different methods and their errors, see the following paper: Current statistical methods for estimating the Michaelis-Menten constants (10.1016/S0026-8857(00)00089-1)

Here, we'll work out an example using Python... First, let's import everything we're going to need...

Now, let's load in some kinetics data. You can download the example data with this link (your link here, or you can use your own data if you want). You may need to adjust the code below to suit your data.

Then we define the function we want to fit, which is just the Michaelis-Menten equation.

Now, we use the nonlinear fitting function from SciPy. We're going to treat these answers as the "correct" ones and preserve them for later use.

The classic method of determining Km and Vmax is to use a double-reciprocal plot, also known as the Lineweaver-Burk plot. The data fit using a linear regression model, and we can derive Km and Vmax. Now let's try it using Python...

Handling Measurement Errors

Let's see how these two methods respond to different levels of measurement error. First, let's set up some functions that will let us compare the two methods.

Interactive Session... Compare Nonlinear Fit with Lineweaver-Burk in Real Time

Direct Link  
to Notebook



Introduction, Background  
and Relevant Literature

Import Required Libraries

Load Data for Examples

Define Michaelis-Menten Equation

Example Nonlinear Fit

Compare with Lineweaver-Burk

Interactive Session...

Compare Nonlinear Fit with  
Lineweaver-Burk in Real Time

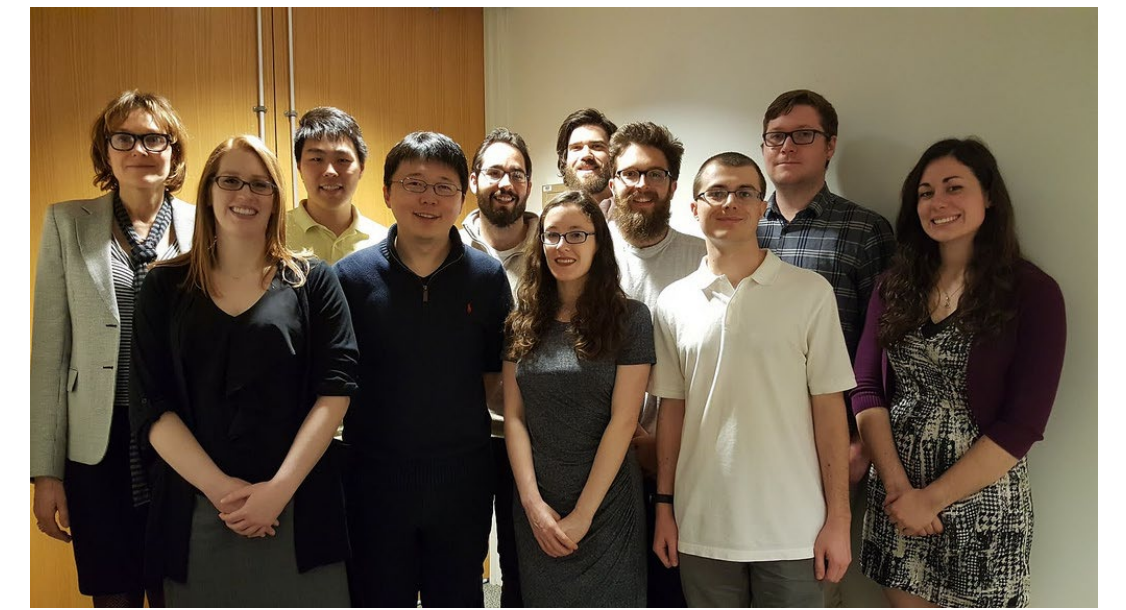
Student Can Adjust Error in  
Measurements

Create Interface

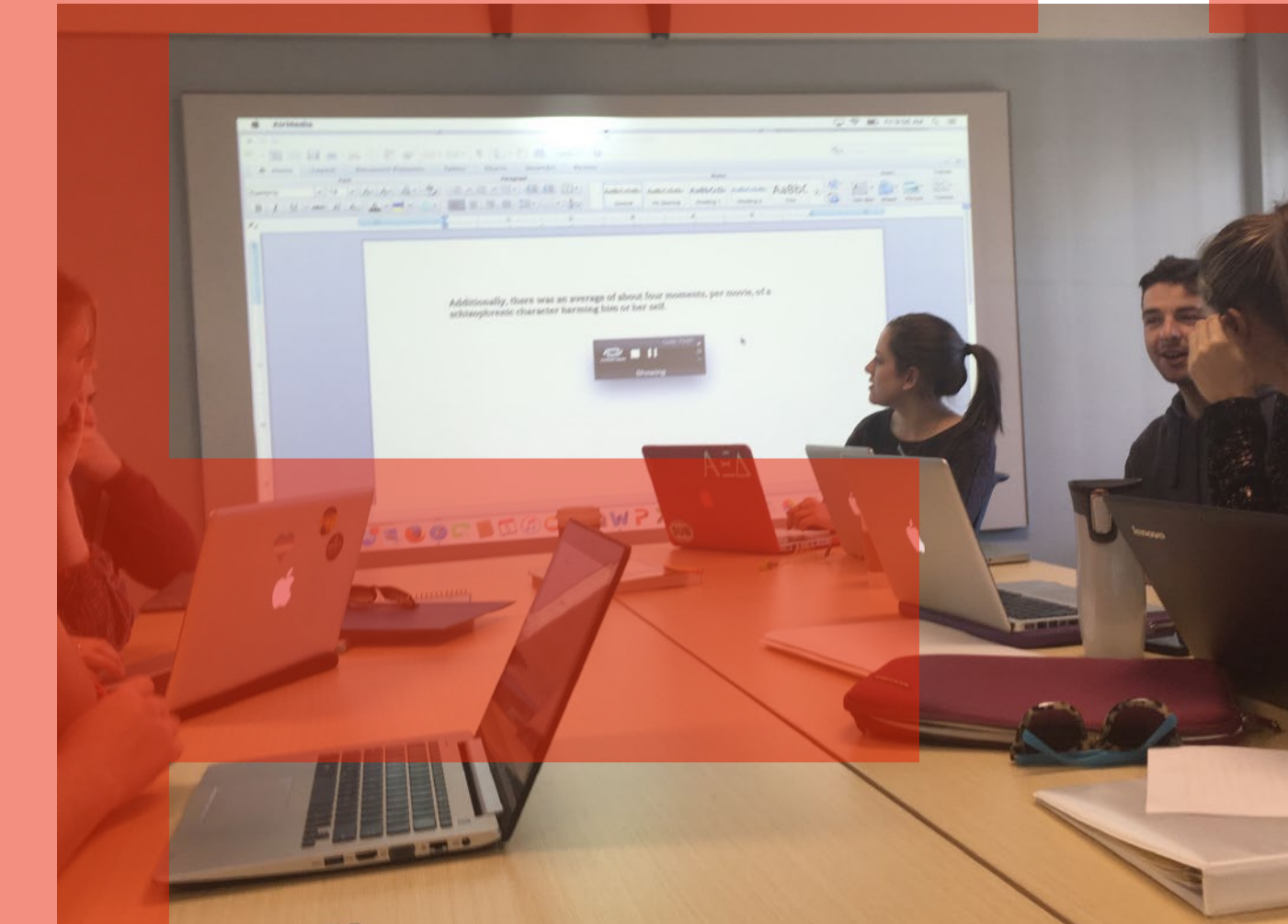
Student Can Move Sliders  
To Adjust Measurement Error

## Summer Mini-Courses

Students will take this course the summer before their qualifying exam, so that they can incorporate the lessons into their thesis proposal and qualifying exam in the fall.



## Statistical Thinking



- Focus on statistics applied to biochemistry & molecular biology
- Emphasize meaning and application, not formal math
- Lecture Component
  - Small sample sizes
  - Measures of error
  - Tests of statistical power
- Small-Group Case Studies
  - Provided by faculty
  - Determine appropriate analyses
  - Implement and present findings

## Presentation Skills



- How to frame a talk for an audience
- How to design clear, effective slides
- Students submit slides to instructor
- Slides will be incorporated into lecture along with critique.
- Small-group workshop where students present and critique each other's work.

## Computer Resources



- Train on practicalities in working with big data sets
- Introduce the University linux cluster
- Discuss range of software available
- Break into small groups:
  - Supervised by CIRC Staff & Faculty
  - Work with real data
- Perform analysis using CIRC resources