Applied Statistics & Data Analytics

1. UMass statistics module
2. Web-based resources
3. MATLAB tools

Applied Statistics & Data Analytics

UMass Statistics Module

Background

- Sophomore level course
- Description: development and analysis of mathematical models for chemical engineering systems
- Topics: statistics for data analysis, linear and nonlinear algebraic equation models, ordinary differential equation models and numerical methods for model solution
- Statistics objective: be able to perform statistical analysis of experimental data and to use computer-based tools for data analysis
- Textbook: E. Kreyzig, Advanced Engineering Mathematics, J. Wiley and Sons, 10th edition (2011).

Statistics Lecture Topics: 4 Weeks

- Introduction
- MATLAB: Introduction
- Probability
- Probability distributions
- MATLAB: Manipulating data
- Binomial & normal distributions
- Confidence intervals
- MATLAB: Functions
- Hypothesis testing
- Correlation & regression analysis
- MATLAB: Statistics toolbox
- Experimental design
- MATLAB: Statistical analysis

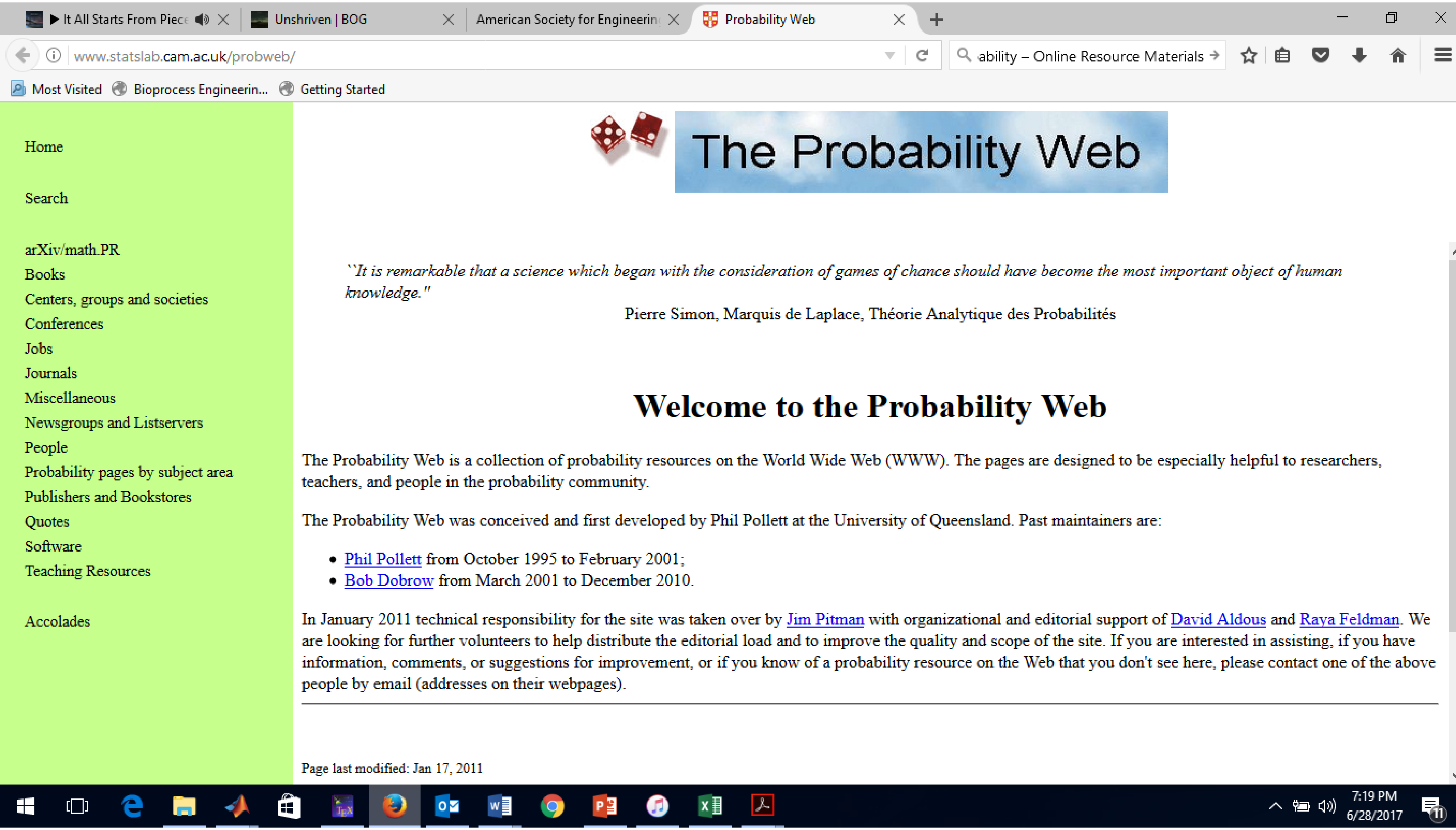
Statistics Homeworks and Tests

- 3 written homeworks
  - » Probability
  - » Probability distributions; binomial, Poisson & normal distributions
  - » Confidence intervals; hypothesis testing
- 1 MATLAB homework
  - » Experimental design; correlation & regression analysis; response surface modeling
- 1 midterm exam
- 1 of 3 options for final project
- All materials available upon request

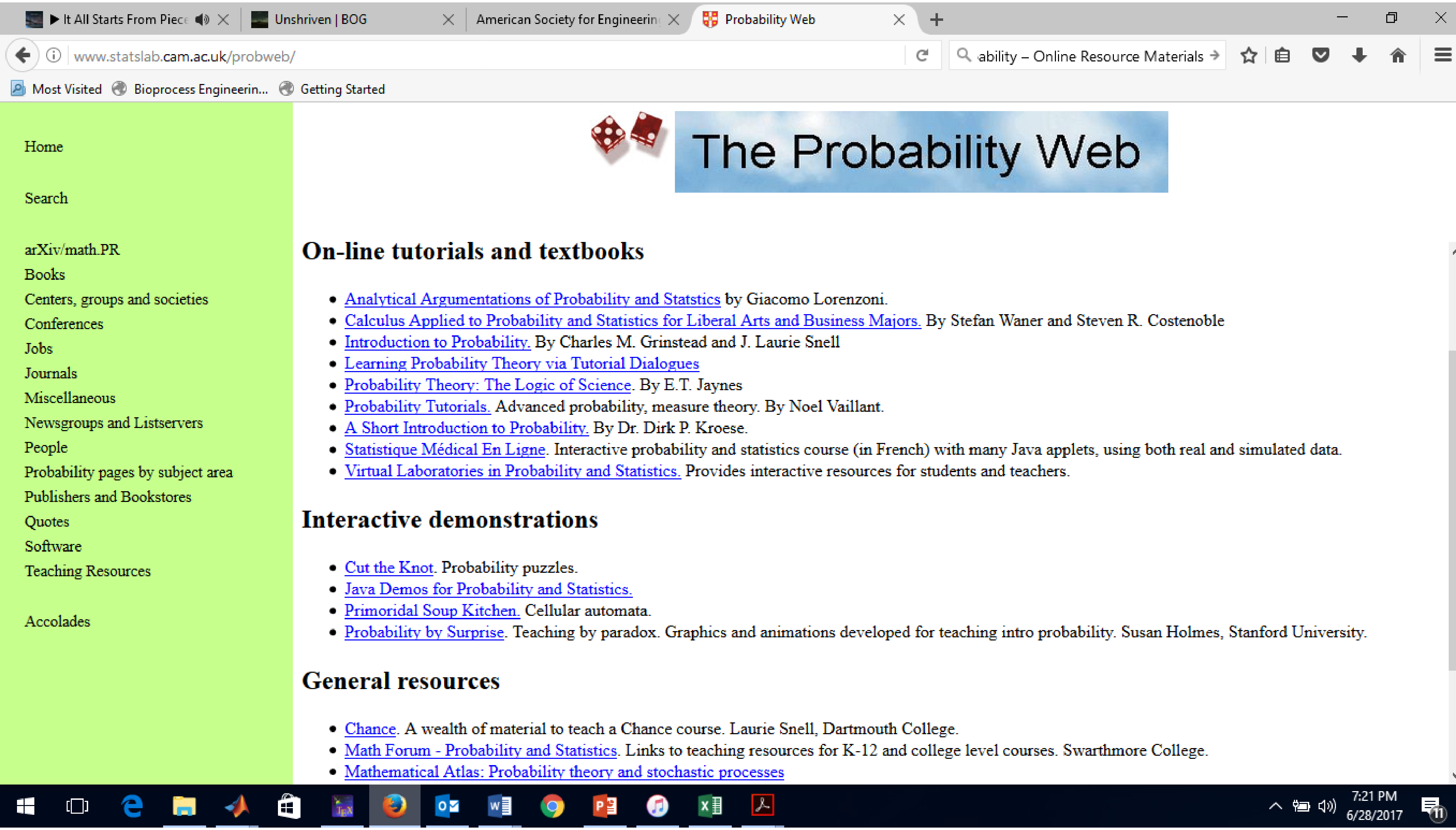
Applied Statistics & Data Analytics

Web-based Resources

The Probability Web



Teaching Resources



The Probability Web

Home

Search

arXiv/math.PR
Books
Centers, groups and societies
Conferences
Jobs
Journals
Miscellaneous
Newsgroups and Listservers
People
Probability pages by subject area
Publishers and Bookstores
Quotes
Software
Teaching Resources

Accolades

On-line tutorials and textbooks

- Analytical Argumentations of Probability and Statstics by Giacomo Lorenzoni.
- Calculus Applied to Probability and Statistics for Liberal Arts and Business Majors. By Stefan Waner and Steven R. Costenoble
- Introduction to Probability. By Charles M. Grinstead and J. Laurie Snell
- Learning Probability Theory via Tutorial Dialogues
- Probability Theory: The Logic of Science. By E.T. Jaynes
- Probability Tutorials. Advanced probability, measure theory. By Noel Vaillant.
- A Short Introduction to Probability. By Dr. Dirk P. Kroese.
- Statistique Médical En Ligne. Interactive probability and statistics course (in French) with many Java applets, using both real and simulated data.
- Virtual Laboratories in Probability and Statistics. Provides interactive resources for students and teachers.

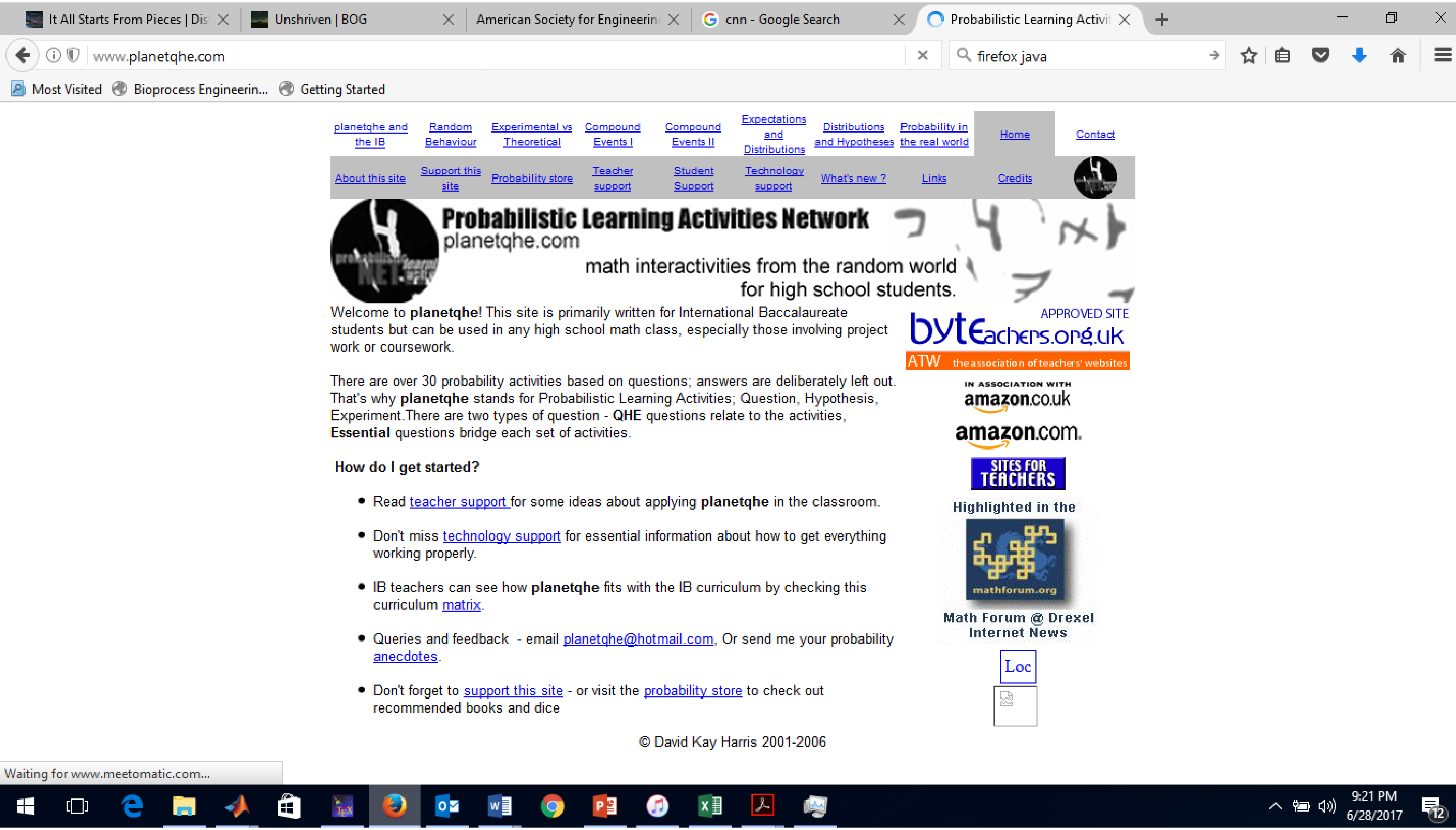
Interactive demonstrations

- Cut the Knot. Probability puzzles.
- Java Demos for Probability and Statistics.
- Primordal Soup Kitchen. Cellular automata.
- Probability by Surprise. Teaching by paradox. Graphics and animations developed for teaching intro probability. Susan Holmes, Stanford University.

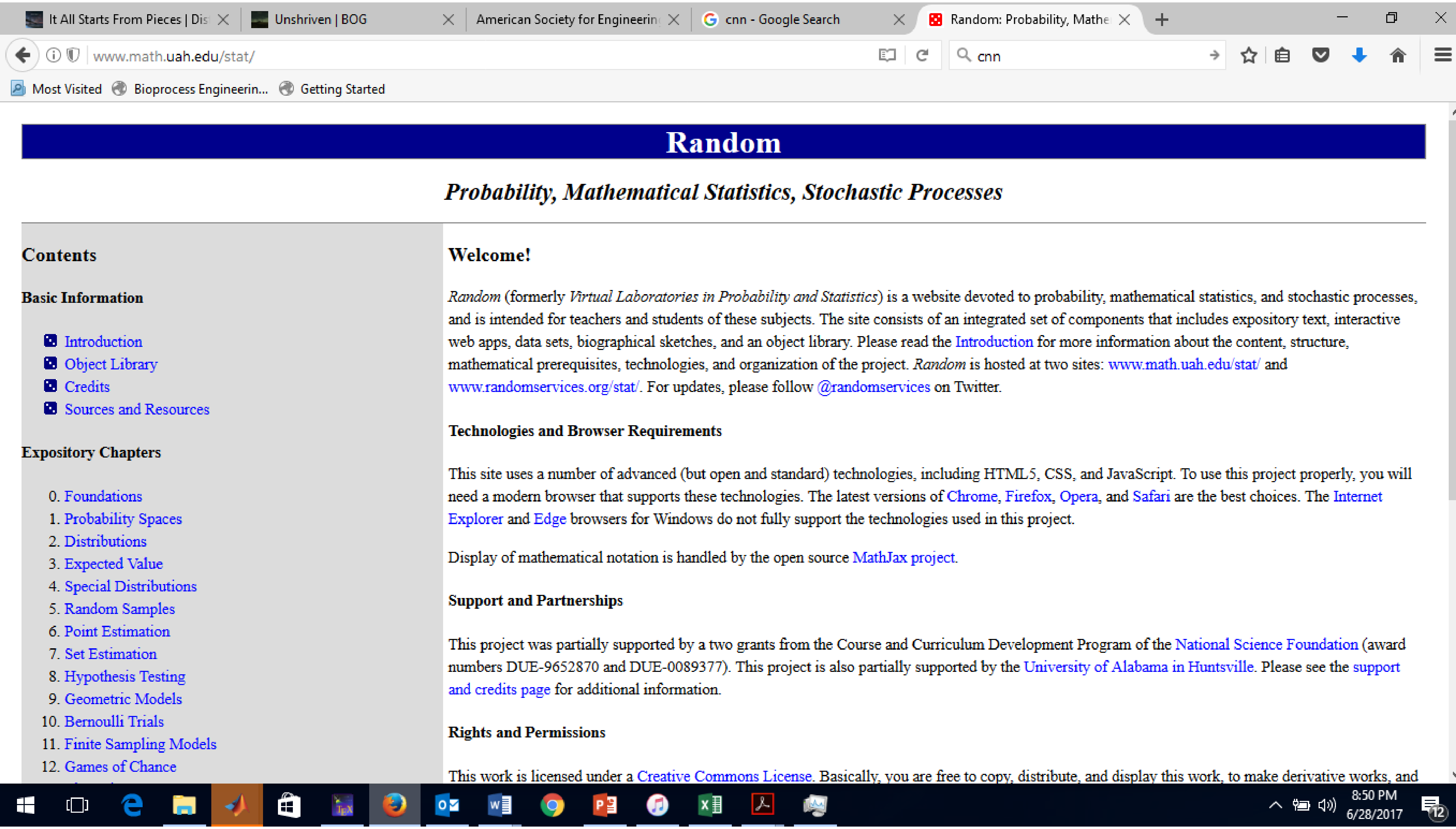
General resources

- Chance. A wealth of material to teach a Chance course. Laurie Snell, Dartmouth College.
- Math Forum - Probability and Statistics. Links to teaching resources for K-12 and college level courses. Swarthmore College.
- Mathematical Atlas: Probability theory and stochastic processes

Probabilistic Learning Activities Network



Virtual Laboratories in Probability & Statistics



Random

Probability, Mathematical Statistics, Stochastic Processes

Contents

Basic Information

Expository Chapters

Welcome!

Random (formerly *Virtual Laboratories in Probability and Statistics*) is a website devoted to probability, mathematical statistics, and stochastic processes, and is intended for teachers and students of these subjects. The site consists of an integrated set of components that includes expository text, interactive web apps, data sets, biographical sketches, and an object library. Please read the Introduction for more information about the content, structure, mathematical prerequisites, technologies, and organization of the project. *Random* is hosted at two sites: www.math.uah.edu/stat/ and www.randomservices.org/stat/. For updates, please follow @randomservices on Twitter.

Technologies and Browser Requirements

This site uses a number of advanced (but open and standard) technologies, including HTML5, CSS, and JavaScript. To use this project properly, you will need a modern browser that supports these technologies. The latest versions of Chrome, Firefox, Opera, and Safari are the best choices. The Internet Explorer and Edge browsers for Windows do not fully support the technologies used in this project.

Display of mathematical notation is handled by the open source MathJax project.

Support and Partnerships

This project was partially supported by a two grants from the Course and Curriculum Development Program of the National Science Foundation (award numbers DUE-9652870 and DUE-0089377). This project is also partially supported by the University of Alabama in Huntsville. Please see the support and credits page for additional information.

Rights and Permissions

Applied Statistics & Data Analytics

MATLAB Tools

MATLAB: Statistics Toolbox

Overview

Statistics Toolbox Capabilities

- Descriptive statistics
- Statistical visualization
- Probability distributions
- Hypothesis tests
- Linear models
- Nonlinear models
- Multivariate statistics
- Statistical process control
- Design of experiments
- Hidden Markov models

Histograms

>> y = [1 3 5 8 2 4 6 7 8 3 2 9 4 3 6 7 4 1
5 3 5 8 9 6 2 4 6 1 5 6 9 8 7 5 3 4 5 2 9
6 5 9 4 1 6 7 8 5 4 2 9 6 7 9 2 5 3 1 9 6
8 4 3 6 7 9 1 3 4 7 5 2 9 8 5 7 4 5 4 3 6
7 9 3 1 6 9 5 6 7 3 2 1 5 7 8 5 3 1 9 7 5
3 4 7 9 1]';

>> mean(y) → ans = 5.1589

>> var(y) → ans = 6.1726

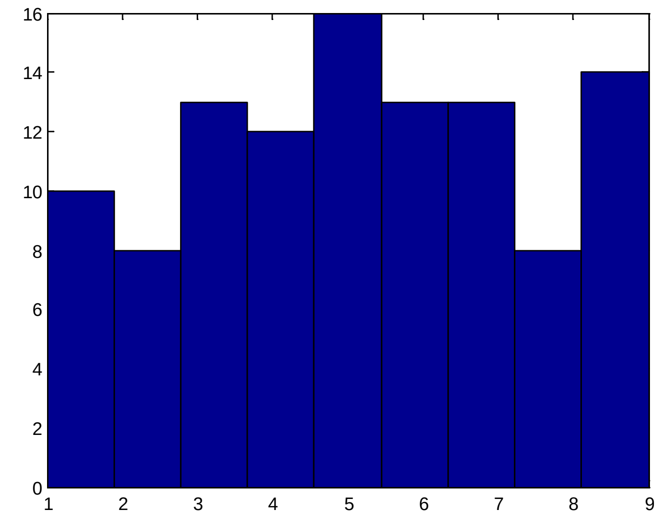
>> std(y) → ans = 2.4845

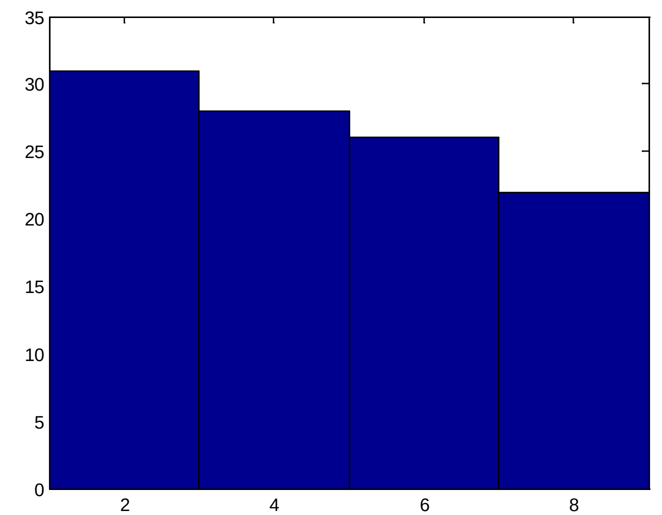
>> hist(y,9) → histogram plot with 9 bins

>> n = hist(y,9) → store result in vector n

>> x = [2 4 6 8]'

>> n = hist(y,x) → create histogram with bin centers specified by vector x





Permutations and Combinations

>> perms([2 4 6]) → all possible permutations of 2, 4, 6

```
6 4 2
6 2 4
4 6 2
4 2 6
2 4 6
2 6 4
```

>> randperm(6) → returns one possible permutation of 1-6

```
5 1 2 3 4 6
```

>> nchoosek(5,4) → number of combinations of 5 things taken 4 at a time without repetitions

```
ans = 5
```

>> nchoosek(2:2:10,4) → all possible combinations of 2, 4, 6, 8, 10 taken 4 at a time without repetitions

```
2 4 6  8
2 4 6 10
2 4 8 10
2 6 8 10
4 6 8 10
```

Probability Distributions

- 21 continuous distributions for data analysis
  - » Includes normal distribution
- 6 continuous distributions for statistics
  - » Includes chi-square and t distributions
- 8 discrete distributions
  - » Includes binomial and Poisson distributions
- Each distribution has functions for:
  - » pdf — Probability density function
  - » cdf — Cumulative distribution function
  - » inv — Inverse cumulative distribution
  - » functionsstat — Distribution statistics function
  - » fit — Distribution fitting function
  - » like — Negative log-likelihood function
  - » rnd — Random number generator

Normal Distribution Functions

- normpdf – probability distribution function
- normcdf – cumulative distribution function
- norminv – inverse cumulative distribution function
- normstat – mean and variance
- normfit – parameter estimates and confidence intervals for normally distributed data
- normlike – negative log-likelihood for maximum likelihood estimation
- normrnd – random numbers from normal distribution

Normal Distribution Examples

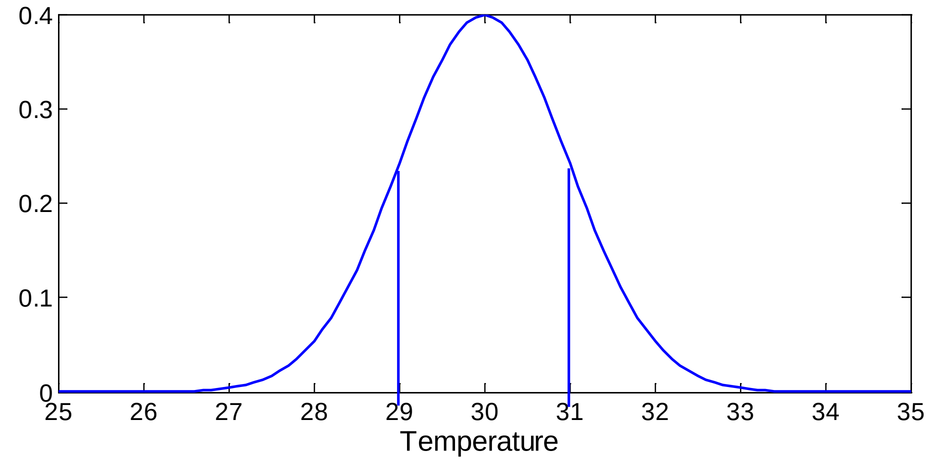
- Normal distribution: normpdf(x,mu,sigma)
  - » normpdf(8,10,2) → ans = 0.1210
  - » normpdf(9,10,2) → ans = 0.1760
  - » normpdf(8,10,4) → ans = 0.0880
- Normal cumulative distribution: normcdf(x,mu,sigma)
  - » normcdf(8,10,2) → ans = 0.1587
  - » normcdf(12,10,2) → ans = 0.8413
- Inverse normal cumulative distribution: norminv(p,mu,sigma)
  - » norminv([0.025 0.975],10,2) → ans = 6.0801 13.9199
- Random number from normal distribution: normrnd(mu,sigma,v)
  - » normrnd(10,2,[1 5]) → ans = 9.1349 6.6688 10.2507 10.5754 7.7071

Normal Distribution Example

- The temperature of a bioreactor follows a normal distribution with an average temperature of 30°C and a standard deviation of 1°C. What percentage of the reactor operating time will the temperature be within +/-0.5°C of the average?
- Calculate probability at 29.5°C and 30.5°C, then calculate the difference:
  - » p=normcdf([29.5 30.5],30,1)

    p = [0.3085 0.6915]
  - » p(2) – p(1)

    0.3829



- The reactor temperature will be within +/- 0.5°C of the average ~38% of the operating time

Confidence Intervals

>> [muhat,sigmahat,muci,sigmaci] = normfit(data,alpha)

- data: vector or matrix of data
- alpha: confidence level = 1-alpha
- muhat: estimated mean
- sigmahat: estimated standard deviation
- muci: confidence interval on the mean
- sigmaci: confidence interval on the standard deviation

>> [muhat,sigmahat,muci,sigmaci] = normfit([1.25 1.36 1.22 1.19 1.33 1.12 1.27 1.27 1.31 1.26],0.05)

```
muhat = 1.2580
sigmahat = 0.0697
muci = 1.2081
       1.3079
sigmaci = 0.0480
          0.1273
```

MATLAB: Statistics Toolbox

In-class Exercise

Membrane Quality

- A membrane manufacturer sells membranes with three different pore sizes
- The average pore size of the membranes supposedly meet the following specifications (in microns):
  - » Small pore membranes: $\mu = 50$, $s = 2.5$
  - » Medium pore membranes: $\mu = 75$, $s = 5$
  - » Large pore membranes: $\mu = 125$, $s = 10$
- The Excel spreadsheet membranes.xls contains the average pore size measured for 25 membranes of each pore size
- Determine if the specifications are satisfied to a 95% confidence level

Exercise

```
>> x=xlsread('membranes')
x =
     1    55    73   125
     2    54    80   123
     3    55    74   118
     4    57    77    99
     5    58    79   123
     6    52    81   112
     7    51    80   121
     8    48    84   132
     9    61    77   121
    10    60    72   127
     ⋮     ⋮     ⋮     ⋮
```

```
>>  x1=x(:,2);
>> x2=x(:,3);
>> x3=x(:,4);
```

Results:

- Small pore membranes
  - » Mean: too high
  - » STD: too high
- Medium pore membranes
  - » Mean: too high
  - » STD: too low (OK)
- Large pore membranes
  - » Mean: too low
  - » STD: in range

Exercise

```
>> [muhat,sigmahat,muci,sigmaci] =
normfit(x1,0.05)

muhat =
  53.9200

sigmahat =
   4.3004

muci =
  52.1449
  55.6951

sigmaci =
   3.3579
   5.9825
```

```
>> [muhat,sigmahat,muci,sigmaci]
= normfit(x2,0.05)

muhat =
  76.7600

sigmahat =
   3.3823

muci =
  75.3639
  78.1561

sigmaci =
   2.6410
   4.7053
```

Exercise

```
>> [muhat,sigmahat,muci,sigmaci] = normfit(x3,0.05)

muhat =
  119.5200

sigmahat =
    9.0789

muci =
  115.7724
  123.2676

sigmaci =
    7.0891
   12.6301
```

MATLAB: Statistical Analysis

1. Overview
2. In-class exercise

MATLAB: Statistical Analysis

Overview

Hypothesis Tests

- 17 hypothesis tests available
- ttest – one-sample or paired-sample t-test. Tests if a sample comes from a normal distribution with unknown variance and a specified mean, against the alternative that it does not have that mean.
- vartest – one-sample chi-square variance test. Tests if a sample comes from a normal distribution with specified variance, against the alternative that it comes from a normal distribution with a different variance.
- chi2gof – chi-square goodness-of-fit test. Tests if a sample comes from a specified distribution, against the alternative that it does not come from that distribution.

Mean Hypothesis Test

```
>> h = ttest(data,m,alpha,tail)
```

- data: vector or matrix of data
- m: expected mean
- alpha: significance level
- Tail = ‘left’ (left handed alternative), ‘right’ (right handed alternative) or ‘both’ (two-sided alternative)
- h = 1 (reject hypothesis) or 0 (accept hypothesis)
- Measurements of polymer molecular weight

$$\{1.25 \quad 1.36 \quad 1.22 \quad 1.19 \quad 1.33 \quad 1.12 \quad 1.27 \quad 1.27 \quad 1.31 \quad 1.26\}$$

$$\bar{x} = 1.258 \quad s^2 = 0.0049$$

- Hypothesis: $\mu_0 = 1.3$ instead of $\mu_1 < \mu_0$

```
>> h = ttest(x,1.3,0.1,'left')
      h = 1
```

Variance Hypothesis Test

\>> h = vartest(data,v,alpha,tail)

- data: vector or matrix of data
- v: expected variance
- alpha: significance level
- Tail = ‘left’ (left handed alternative), ‘right’ (right handed alternative) or ‘both’ (two-sided alternative)
- h = 1 (reject hypothesis) or 0 (accept hypothesis)
- Measurements of polymer molecular weight:

$$\bar{x} = 1.258 \quad s^2 = 0.0049$$

- Hypothesis: $\sigma^2 = 0.0075$ and not a different variance

```
>> h = vartest(x,0.0075,0.1,'both')
h =
    0
```

Linear Regression

```
>> [k, kint] = regress(y, X, alpha)
```

- y is a vector containing the dependent variable data
- X is the independent variable data; must contain a vector of ones or the regression will be calculated to pass through the origin
- Confidence level = 1-alpha
- Returns vector of coefficient estimates k with confidence intervals kint

Linear Regression Example

Experiment	1	2	3	4	5	6	7	8
Reactant Concentration	0.1	0.3	0.5	0.7	0.9	1.2	1.5	2.0
Rate	2.3	5.7	10.7	13.1	18.5	25.4	32.1	45.2

```
>> c = [0.1 0.3 0.5 0.7 0.9 1.2 1.5 2];
>> r  = [2.3 5.6 10.7 13.1 18.5 25.4 32.1 45.2];
>> caug = [ones(length(c), 1), c']
caug =
    1.0000    0.1000
    1.0000    0.3000
    1.0000    0.5000
    1.0000    0.7000
    1.0000    0.9000
    1.0000    1.2000
    1.0000    1.5000
    1.0000    2.0000
```

Linear Regression Example

```
>> [k, kint] = regress(r', caug, 0.05);
>> k
   k =
      -1.1847
      22.5524
>> kint
   kint =
      -2.8338    0.4644
      21.0262   24.0787
```

Correlation Analysis

```
>> [R,P]=corrcoef(x,y)
```

- R is a matrix of correlation coefficients calculated from vectors x and y
- The correlation coefficient of interest is located in the off-diagonal entries of the R matrix
- P a matrix of p-values for testing the hypothesis of no correlation. Each p-value is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero.
- If P(i,j) is small, say less than 0.05, then the correlation R(i,j) is significant

Correlation Analysis Example

Experiment	1	2	3	4	5	6	7	8
Hydrogen Concentration	0	0.1	0.3	0.5	1.0	1.5	2.0	3.0
Polymerization rate	9.7	9.2	10.7	10.1	10.5	11.2	10.4	10.8

```
>> h = [0 0.1 0.3 0.5 1 1.5 2 3];
>> p = [9.7 9.2 10.7 10.1 10.5 11.2 10.4 10.8];
>>  [R,P] = corrcoef(h,p)
R =
    1.0000    0.6238
    0.6238    1.0000
P =
    1.0000    0.0984
    0.0984    1.0000
```

- Accept hypothesis that x and y are uncorrelated at 5% significance level

Response Surface Models

- Example – three inputs (x_1, x_2, x_3) and one output (y)
- Linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- Linear model with interactions

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

- Quadratic model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2$$

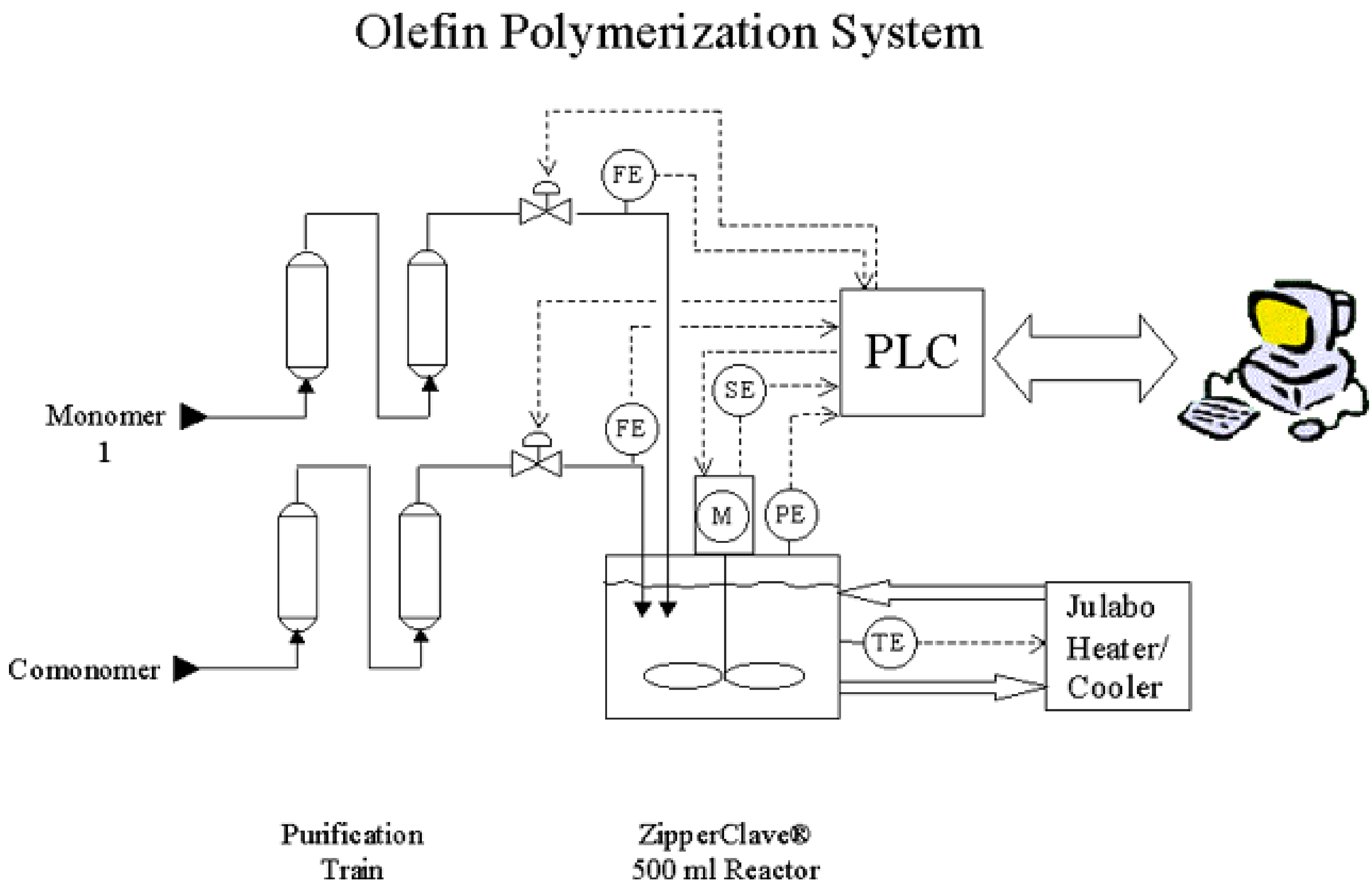
Response Surface Modeling

- RSTOOL(X,Y,MODEL) opens a GUI for fitting a polynomial response surface for a response variable Y as a function of the multiple predictor variables in X.
- Distinct predictor variables should appear in different columns of X. Y can be a single vector or a matrix, with columns corresponding to multiple responses.
- RSTOOL displays a family of plots, one for each combination of columns in X and Y. 95% global confidence intervals are shown as two red curves.
- MODEL controls the regression model.
  - » 'linear ' – Constant and linear terms (the default)
  - » 'interaction' – Constant, linear, and interaction terms
  - » 'quadratic' – Constant, linear, interaction, and squared terms
  - » 'purequadratic' – Constant, linear, and squared terms

MATLAB: Statistical Analysis

In-class Exercise

Response Surface Modeling Example



Polymer Reactor Data Regression

- Input variables
  - » Catalyst and co-catalyst concentrations
  - » Monomer and co-monomer concentrations
  - » Reactor temperature
- Output variables
  - » Polymer production rate
  - » Copolymer composition
  - » 2 molecular weight measures
- Dataset available in Excel spreadsheet reactordata.xls
  - » 27 experiments with different input combinations
  - » Data collected for all 4 outputs
  - » Perform analysis only for polymer production rate

Polymer Reactor Data Regression

```
>> data=xlsread('reactordata');
>> size(data)
ans =
   10    27
>> x=data(2:6,:);
>> y=data(7,:);
>> size(x)
ans =
    5    27
>> size(y)
ans =
    1    27
>> rstool(x',y')
```