MATLAB: Statistics

1. Overview
2. In-class exercise

MATLAB: Statistics

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

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

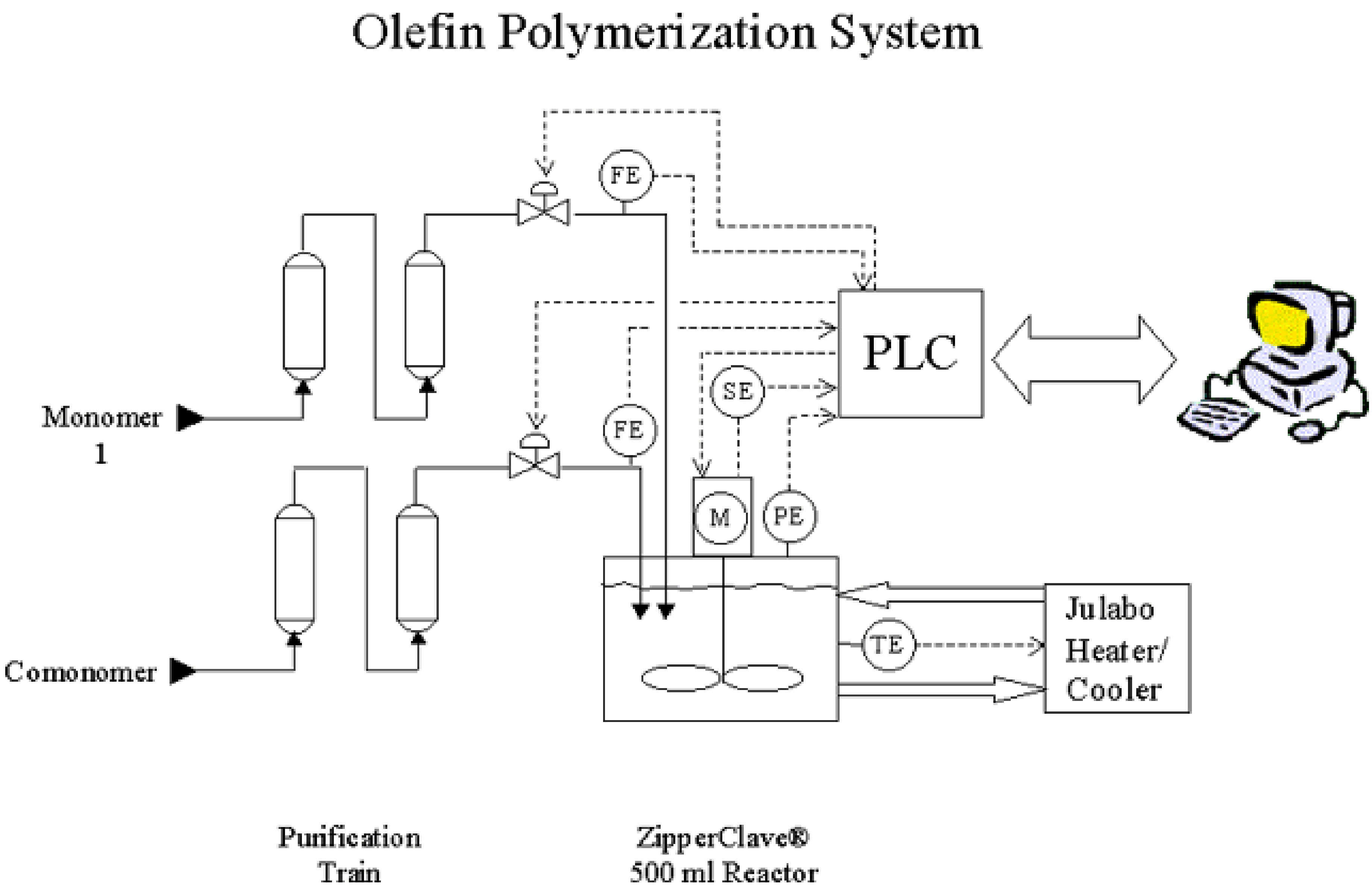
Response Surface Modeling

- 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: Statistics

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