# CE3502 EMMA Final Exam Review

Last updated: 14 April 2010

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wk 12</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/6</td>
<td>4/7 NO CLASS</td>
<td>4/8 No scheduled lab PROJECTS</td>
<td>4/9 Presentations on lab 9 (Groups 6,7,8)</td>
</tr>
<tr>
<td><strong>Wk 13</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/13</td>
<td>4/14 Exam Review</td>
<td>4/15 No scheduled lab PROJECTS</td>
<td>4/16 PROJECT Presentations Groups 3,5,10 Course Eval.</td>
</tr>
<tr>
<td><strong>Wk 14</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/20</td>
<td>4/21 Presentations Groups 1, 4,7,12</td>
<td>4/22 NO CLASS</td>
<td>4/23 Presentations Groups 2,6,8,9,11</td>
</tr>
<tr>
<td><strong>EXAM WEEK</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Midterm – 10%
Project – 15%
Final – 15%

20% of grade

40% of grade

CE3502
Project Evaluation Sheet
Writing (20 pts)

Abstract
(Objective/Purpose, Methods, Key results, Significance)

Introduction
(Concern with this topic (background), Objectives)

Methods

Results
(proper display of results, Text along with results)

Discussion
(Augments results, interprets results, draws logical conclusions)

Conclusions and Recommendations
(Answer the objectives, given results)

Composition
(Spelling, grammar, sentence structure, paragraph structure, flow of ideas, choice of words)

Statistics (50 pts)

Testable hypotheses (10)

Experimental design (replication, factorial or other design) (10)

Correct choice of statistical tools (10)

Correctness of statistical procedures (20)

New knowledge or understanding (10 pts)

Did students learn material above and beyond what was covered in class

Presentation (20 pts)
Topics

Descriptive statistics:
- Mean, Standard Deviation, COV
- Bias (accuracy), precision, Random vs. systematic errors
- Populations vs. samples
- Distributions – descriptions of populations
  - Normal, lognormal
  2. Smoothing Techniques for visualization
    - Running average
    - Weighted running average
  3. Correlation and Regression analysis
  4. Comparisons of numbers:
    - Percentiles, tolerance limits, outliers, DL, MDL
    - Confidence intervals
  - t-test: independent and paired t-tests, 1- vs 2-tailed
  - ANOVA, F-test,
  - Experimental design: Factorial design, single factor

5. Propagation of Error
6. Significant Digits

Types of material to know

- Vocabulary – definitions;
- Equations, calculations (incl. spreadsheets);
- Graphical representations;
- Interpretation of spreadsheet output;
- Able to read statistics tables;
Vocabulary

Mean | Median | Mode
---|---|---
Standard Deviation | Error | Uncertainty
Standard error | Variance | Coefficient of variation
Random | Systematic | Accuracy
Bias | Precision | Population
Distribution | Normal | Lognormal
Gaussian | Smoothing | Running average
Exponentially-weighted running average | | Correlation Coefficient
Intercept | Slope | Least squares
Percentile | Confidence Interval | Upper Tolerance Limit
Outlier | Independent | Degrees of Freedom
Histogram | Error propagation | Null Hypothesis
Residual | ANOVA | Factorial design
Balanced design

Equations

\[ CI = t_{v,n} \cdot \frac{s}{\sqrt{n}} \]
\[ e_{\text{total}}^2 = e_{\text{random}}^2 + e_{\text{systematic}}^2 \]
\[ \bar{x}_i = \sum_{j=0}^{n} (1-\phi^j) \cdot x_{i-j} \]
\[ SE = \frac{\sigma}{\sqrt{n}} \]
\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \]
\[ DL = \frac{S_{\text{null}} - \bar{B}L}{m} \]
\[ S_{\text{null}} = \bar{B}L + K \cdot s_{g_i} \]
\[ MDL = t_{v,n} \cdot s \]
\[ t^* = \frac{x_i - \eta_j}{s^*_{\eta}} \]
\[ \bar{d} \approx \frac{\sum d_i}{n} = \frac{1}{n} \sum (x_i - y_i) \]
\[ s^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} \]
\[ f(x) = \frac{m}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
\[ s_{\text{pool}}^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2} \]
\[ S_{\eta}^2 = \frac{s_d}{\sqrt{n}} \]
\[ \sigma = \frac{\sqrt{\sum (\bar{x} - x_i)^2}}{n-1} \]

4/14/2010
Computer output

- Descriptive statistics
- Regressions;
- Correlation coefficients (table);
- T-tests (F-test not required);
- ANOVA

Statistical tables

- Critical correlation coefficients;
- t tables;
- F tables;
- z table
Graphical representations

- Mean and error bars;
- Box & whisker plots;
- Histograms;
- Regression lines (trend lines);
- Smoothing;
Jeopardy:
Statistically speaking
Vocabulary

True or False (circle one): For a population the mean is defined as
\[ \mu = \frac{\sum_{i=1}^{n} x_i}{n} \]
but for a small sample of the population, the mean is defined as:
\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n - 1} \]

Vocabulary

True or False (circle one): The median is defined as the highest point on a histogram showing the frequency distribution.
True or False (circle one): For a lognormal distribution, the median equals the mean.

True or False (circle one): If the median equals the mean, the data must be normally distributed.
Correlation, Regression

It appears as if the relative contribution of organic matter to the garbage collected in the Dow Building has been increasing over the years. Answer the following questions based on the graph below.

\[ y = 0.009x - 17.897 \]
\[ R^2 = 0.77 \]

a. Does this graph tell us that the amount of organic matter in the garbage has increased over the time period shown? Explain.

b. To determine if the trend line is significant, you need to know how many degrees of freedom are available. For the correlation coefficient is the degrees of freedom equal to \((n - 1)\) or \((n - 2)\)?
Correlation, Regression

Based on the table below and your answer above, would you say that the time trend shown in the graph is statistically significant at the 95% confidence level?

$r^2 = 0.77, r = 0.877, n = 4$

<table>
<thead>
<tr>
<th>Year</th>
<th>Organic Matter (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>4.0%</td>
</tr>
<tr>
<td>1999</td>
<td>8.0%</td>
</tr>
<tr>
<td>2000</td>
<td>12.0%</td>
</tr>
<tr>
<td>2001</td>
<td>16.0%</td>
</tr>
<tr>
<td>2002</td>
<td>20.0%</td>
</tr>
<tr>
<td>2003</td>
<td>24.0%</td>
</tr>
<tr>
<td>2004</td>
<td>28.0%</td>
</tr>
<tr>
<td>2005</td>
<td>32.0%</td>
</tr>
<tr>
<td>2006</td>
<td>36.0%</td>
</tr>
<tr>
<td>2007</td>
<td>40.0%</td>
</tr>
<tr>
<td>2008</td>
<td>44.0%</td>
</tr>
</tbody>
</table>

**Correlation, regression**

The graph below shows the historical relationship between chlorophyll and total phosphorus in Lake Superior. The $r^2$ value is so low that this correlation would not be statistically significant under any circumstances. True or False?

$y = 0.262x - 0.093$

$R^2 = 0.383$
The table below summarizes the results of multiple t-tests following an ANOVA. The last column in the table indicates which years are similar and dissimilar. Answer the questions below based on the table.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9</td>
<td>41.63</td>
<td>4.6</td>
<td>0.22</td>
<td>0.31 b</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>42.40</td>
<td>4.2</td>
<td>0.39 a,b</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>44.86</td>
<td>4.1</td>
<td>0.25 a,b</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>42.20</td>
<td>3.8</td>
<td>0.19 a</td>
<td>0.26 a</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>55.85</td>
<td>5.1</td>
<td>1.04 b,c</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>61.75</td>
<td>6.2</td>
<td>0.76 c</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>55.85</td>
<td>5.1</td>
<td>1.04 b,c</td>
<td></td>
</tr>
</tbody>
</table>

a. Which Group is unlike any other groups?

b. Is Group 4 significantly different from Group 9?
### T-tests

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9</td>
<td>41.63</td>
<td>4.6</td>
<td>0.22</td>
<td>0.31</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>42.40</td>
<td>4.2</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>44.86</td>
<td>4.1</td>
<td>0.25</td>
<td>0.29</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>42.20</td>
<td>3.8</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>45.97</td>
<td>4.2</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>55.85</td>
<td>5.1</td>
<td>1.04</td>
<td>0.60</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>61.75</td>
<td>6.2</td>
<td>0.76</td>
<td>0.54</td>
</tr>
</tbody>
</table>

List three groups that are significantly different from each other.

### Below are descriptive statistics (Excel) for a set of data. Are the data normally distributed?

<table>
<thead>
<tr>
<th>CO₂ (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean: 5300</td>
</tr>
<tr>
<td>Median: 6000</td>
</tr>
<tr>
<td>Mode: 6000</td>
</tr>
<tr>
<td>Std. Dev.: 1445</td>
</tr>
<tr>
<td>Skew: -2.2</td>
</tr>
<tr>
<td>Kurtosis: 4.1</td>
</tr>
</tbody>
</table>

- **a.** Definitely
- **b.** Definitely not
- **c.** Possibly
- **d.** Probably not
- **e.** Cannot tell
Below are descriptive statistics (Excel) for a set of data. Are the data normally distributed?

<table>
<thead>
<tr>
<th>CO₂ (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean: 438</td>
</tr>
<tr>
<td>Median: 437</td>
</tr>
<tr>
<td>Mode: 424</td>
</tr>
<tr>
<td>Std. Dev.: 12</td>
</tr>
<tr>
<td>Skew: 0.52</td>
</tr>
<tr>
<td>Kurtosis: -0.55</td>
</tr>
</tbody>
</table>

a. Definitely  
b. Definitely not  
c. Possibly  
d. Probably not  
e. Cannot tell

The mean ± standard deviation for a set of measurements of CO₂ concentrations was found to be 438.2 ± 12.1. How many digits in the mean are significant?

a. 1 (400 ± 10)  
b. 2 (440 ± 12)  
c. 3 (438 ± 12.1)  
d. 4 (438.2 ± 12.1)
You have taken 5 samples of well water and measured the arsenic concentration. You would like to estimate the 99% upper tolerance limit where
\[ UTL = x + k \cdot s. \]
The definition of \( k \) is shown below. What is the correct value of \( t \) from the table below that you should use to calculate \( k \) and the UTL?

\[ k = t_{\nu, \alpha} \cdot \sqrt{1 + \frac{1}{n}} \]

**TABLE A.3 Upper percentage points for the Student's t distribution**

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>0.040</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nu )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.325</td>
<td>1.000</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
</tr>
<tr>
<td>2</td>
<td>0.289</td>
<td>0.816</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
</tr>
<tr>
<td>3</td>
<td>0.277</td>
<td>0.766</td>
<td>1.638</td>
<td>2.333</td>
<td>3.182</td>
<td>4.541</td>
</tr>
<tr>
<td>4</td>
<td>0.270</td>
<td>0.742</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
</tr>
<tr>
<td>5</td>
<td>0.267</td>
<td>0.727</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
</tr>
</tbody>
</table>

You have collected 5 samples of well water and measured the arsenic concentration in each. Now you want to calculate the 95% confidence interval about the mean. The standard deviation is 1.4 ppb. Based on the \( t \) table below, what would be the confidence interval?

\[ CI = t_{\nu, \alpha} \cdot \frac{s}{\sqrt{n}} \]

- a. \( \mu \pm 1.3 \text{ ppb} \) (\( t = 2.132 \))
- b. \( \mu \pm 2.4 \text{ ppb} \) (\( t = 3.747 \))
- c. \( \mu \pm 1.7 \text{ ppb} \) (\( t = 2.776 \))
- d. \( \mu \pm 1.0 \text{ ppb} \) (\( t = 1.533 \))

**TABLE A.3 Upper percentage points for the Student's t distribution**

<table>
<thead>
<tr>
<th>( \nu )</th>
<th>0.040</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.325</td>
<td>1.000</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
</tr>
<tr>
<td>2</td>
<td>0.289</td>
<td>0.816</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
</tr>
<tr>
<td>3</td>
<td>0.277</td>
<td>0.766</td>
<td>1.638</td>
<td>2.333</td>
<td>3.182</td>
<td>4.541</td>
</tr>
<tr>
<td>4</td>
<td>0.270</td>
<td>0.742</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
</tr>
<tr>
<td>5</td>
<td>0.267</td>
<td>0.727</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
</tr>
</tbody>
</table>
Your consulting firm is working on the remediation of an old gas station that had a leaking underground storage tank. The firm has drilled well holes to determine the extent of the ground water contamination. A total of 10 wells have been drilled, and 3 replicate water samples have been collected and analyzed for volatile hydrocarbons. What statistical test(s) would you employ to determine which wells were contaminated?

a. Correlation and regression analysis;
b. A nonparametric test;
c. Weighted average smoothing;
d. ANOVA
e. ANOVA and t-tests

You would like to determine the relationship between mercury deposition in rain and distance from the nearest coal-fired power plant. You place rain collectors at different distances from several power plants. What statistical test would be the best to analyze your data?

a. Correlation analysis;
b. Regression Analysis
c. T-test;
d. ANOVA;
e. Smoothing with moving averages;