## UMBC UGC New Course Request: BTEC 350: Statistics for Translational Life Science

Date Submitted: 04/04/2014 (revised 10-6-15) Proposed Effective Date: Fall 2015

| | Name | Email | Phone | Dept |
|---|---|---|---|---|
| Dept Chair or | William R. LaCourse | lacourse@umbc.edu | X55827 | CNMS |
| Other Contact | Kathy Lee Sutphin | sutphin@umbc.edu | X52271 | CNMS |

### COURSE INFORMATION:

| | |
|---|---|
| Course Number(s) | BTEC 350 |
| Formal Title | **Statistics for Translational Life Science** |
| Transcript Title (≤30c) | Trans Life Sci Statistics |
| Recommended Course Preparation | |
| Prerequisite | See below |
| Credits | 4 |
| Repeatable? | ☐ Yes ☒ No |
| Max. Total Credits | 4 |
| Grading Method(s) | ☒ Reg (A-F) ☐ Audit ☐ Pass-Fail |

### PROPOSED CATALOG DESCRIPTION (no longer than 75 words):

This course will introduce statistical methods used commonly by Translational Life Science Technology researchers. Some topics covered include organizing and presenting data with charts, histograms and plots, calculating means, standard deviations, and other statistical measures, statistical computing in biology using R, data collection and sampling, Bayesian statistics, developing statistical estimators hypothesis testing and likelihood ratio analysis. Application examples in epidemiology, genome wide association studies, and biological product quality control will illustrate basic and applied concepts.

### RATIONALE FOR NEW COURSE:

This course will be part of the upper level (years 3 and 4) of a joint program in Translational Life Science Technology (2+2 TLST) created and run by UMBC. The Translational Life Science Technology (2+2 TLST) is being managed through the College of Natural and Mathematical Sciences at UMBC due to the interdepartmental, intercollegiate, and inter-institutional nature of the joint program, which is being developed by a faculty committee involving Dr. Mauricio Bustos, Dr. Charles Bieberich, and Dr. Mariejose Castellanos.

Prerequisites for UMBC students (all courses must be passed with a grade of "C" or higher): MATH 150 (or MATH 151/155), BIOL 302.

**ATTACH COURSE OUTLINE (mandatory):**

## **OUTLINE -** BTEC 350: Statistics for Translational Life Science

## Summary

Contemporary life science research and technology depend heavily on collecting and analyzing scientific data from which new concepts and ideas may be inferred. The range of applications of statistics in translational life science is very broad. In order to integrate the study and practice of statistics within the TLST curriculum, the BTEC 350 syllabus will be tailored to a level of competency in calculus corresponding to MA 151, and a broad knowledge of biology. Specific examples will be drawn directly from TLST courses at both lower and upper levels.

## Syllabus

### *Learning objectives*

By the end of the semester, students should have:

1. Acquired the basic notions and vocabulary necessary to understand to design experiments, collect and organize data
2. Gained an understanding of the range of applications of statistics in biology.
3. Grasped the power of quantitative statistical analysis in Translational Life Science Technology research and development.
4. Attained competency on how to apply various statistical tests, and how to estimate point values, means and standard deviations

**Textbook and other didactic materials**
Myra L. Samuels and Jeffrey A. Witmer, "Statistics for the Life Sciences", 5th Edition, (Brooks/Cole, 2012)

### *Calendar (weekly)*

| Week | |
|---|---|
| 1 | Organizing data with charts, histograms, and plots |
| 2 | Calculating the mean, median, trimmed mean, standard deviation, quartiles, weighted mean, and regression lines of a data set |
| 3 | Using the statistical software package R to perform basic statistical procedures |
| 4 | Summarizing empirical probability distributions (cumulative) and survival functions of a dataset |
| 5 | Application example: Estimation of the reproduction number $r_0$ during a human epidemics of influenza A (H1N1) |
| 6 | Collecting data under a well-crafted experimental design |
| 7 | Data collection and experiments |
| 8 | Basic probability theory |
| 9 | Bayes formula and Bayesian statistics |
| 10 | Using Monte Carlo techniques to simulate data generation and sampling |
| 11 | Application example: Genome wide association study simulation using the R package GenABEL |
| 12 | Statistical estimation procedures: developing estimators |
| 13 | Hypothesis testing, likelihood ratio analysis |
| 14 | Application example: "qcc" an R package for product quality and statistical process control |

### *Application examples*

Application examples will provide students with opportunities for putting into practice the knowledge of statistics learned in this course, while at the same time revealing the depth and power of statistical analysis in the life sciences. Each application will be based on one (or more) published references, from which students will obtain data and context for each exercise. Descriptions of the three types of application exercises are provided below:

### *Estimation of the reproduction number $r_0$ during a human epidemics of influenza A (H1N1)*

Reference:

Fraser, C. *et al*. (2009). Pandemic potential of a strain of influenza A (H1N1): Early findings. Science 324: 1557-1561.

Excerpt from Fraser et al. (2009)

"A novel influenza A (H1N1) virus has spread rapidly across the globe. Judging its pandemic potential is difficult with limited data, but nevertheless essential to inform appropriate health responses. By analyzing the outbreak in Mexico, early data on international spread, and viral genetic diversity, we make an early assessment of transmissibility and severity"

In this application exercise students will download and analyze epidemiological data from Fraser et al (2009) to estimate the reproduction number $r_0$ of influenza A (H1N1) virus using the methodology described in the Supplementary information (Methods).

### *Genome wide association study (GWAS ) simulation using GenABEL*

References:

Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12): e1002822. doi:10.1371/journal.pcbi.1002822

Aulchenko Y.S., Ripke S., Isaacs A., van Duijn C.M. GenABEL: an R package for genome-wide association analysis. Bioinformatics. 2007, 23(10):1294-6.

Excerpt from Bush WS, Moore JH (2012)

"A central goal of human genetics is to identify genetic risk factors for common, complex diseases such as schizophrenia and type II diabetes, and for rare Mendelian diseases such as cystic fibrosis and sickle cell anemia. There are many different technologies, study designs and analytical tools for identifying genetic risk factors. We will focus here on the genome-wide association study or GWAS that measures and analyzes DNA sequence variations from across the human genome in an effort to identify genetic risk factors for diseases that are common in the population. The ultimate goal of GWAS is to use genetic risk factors to make predictions about who is at risk and to identify the biological underpinnings of disease susceptibility for developing new prevention and treatment strategies."

In this application exercise students will use the R package GenABEL (Aulchenko et al., 2007, http://www.genabel.org/) to simulate a Genome Wide Association Study. The students will download a dataset from the course web site containing phenotypic and genotypic (SNP frequencies) data, perform data classification, normalization, rank-normalization and non-parametric adjustment procedures, call rate and heterozygosity, and SNP association test using GLM, case-control association analysis, and calculate genome-wide significance, and score tests.

### *"qcc" an R package for product quality and statistical process control*

References:

Scrucca, L. (2004). Qcc: An R package for quality control charting and statistical process control. R News, 4: 11-17.

Charaniya, S. et al. (2008) Mining bioprocess data: opportunities and challenges. Trends in Biotechnology, 26(12): 690-699.

Excerpt from Scrucca (2004)

"In this paper we briefly describe the qcc package. This provides functions to draw basic Shewhart quality control charts for continuous, attribute and count data; corresponding operating characteristic curves are also implemented. Other statistical quality tools available are Cusum and EWMA charts for continuous data, process capability analyses, Pareto charts and cause-and-effect diagram."

In this application exercise students will employ the qcc R package to analyze quality control data referenced in Charaniya et al (2008). Specific goals of this exercise will be to produce Cusum and EWMA charts for continuous biological culture data, dimensionality reduction and bioprocess data mining.

### *Grading*

The final grade will be based on homework (10 best scores out of 12 assignments), two midterms and a final exam. Final grades will be calculated as follows.

| | | Subtotals |
|---|---|---|
| Homework | 25 % | 25 % |
| Midterms (2) | 20 % | 40 % |
| Final exam | 35 % | 35 % |
| Total | | 100 % |