Module 1: Statistical Process

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 1.a - **Define** the term *sample* and *population*. (Supports Course Objective A)
- 1.b - **Describe** the two types of *variables.* (C)
- 1.c - **Describe** the process of *statistical inference*. (H)
- 1.d - **Compare** and **contrast** the terms *statistic* and *parameter*. (A, H)
- 1.e - Given a description of a research goal or question of interest, **identify** the population and the parameter of interest and **state** the *null* and *alternative* hypotheses which capture the question of interest. (A, E)
- 1.f - **Describe** the following sampling techniques: *simple random sample, stratified random sample*. Given a description of how a dataset was collected, **identify** whether the sampling scheme was a simple random sample, a stratified random sample, a random sample not defined in class, or a non-random sample. (B)
- 1.g - **Compare** and **contrast** *observational studies* and *controlled experiments*. Given a description of how a study was conducted, **identify** whether the study was an observational study or controlled experiment. (B, E, H)
- 1.h - **State** the primary benefit of *random sampling* and *random assignment*. (B, E, H)
- 1.i - Given a description of a study design, **identify** whether the conclusion is reliable. (E, H)
- 1.j - Given a dataset, **construct** graphical (*bar charts, histograms, box plots, individual value plots, density plots*) and numerical summaries (*proportions, frequencies, mean, median, quartiles, standard deviation, IQR*) which address the question of interest. (C)
- 1.k - Given a graphical summary, **interpret** the graphic in the context of the question of interest. That is, answer the question: "based on the graphic, do you believe the data supports the research hypothesis? Explain." (C)

Module 2: Sampling Distributions and Confidence Intervals

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 2.a - **Define** the term *sampling distribution* and **describe** how it is constructed conceptually. (Supports Course Objective D)
- 2.b - **Compare** the sampling distribution to the distribution of the population. (D)
- 2.c - **State** two properties of the sampling distribution of most statistics. (D)
- 2.d - **Describe** the idea behind the common intuition that "more data is better." (D)
- 2.e - **State** three general approaches for modeling a sampling distribution. (D)
- 2.f - **Describe** the process of *bootstrapping* for empirically modeling the sampling distribution of a statistic. (D, E)
- 2.g - Given a model for the sampling distribution, **state** whether the sample on which the model is based is consistent with a specified value of a parameter and **justify** your response. (D, E)
- 2.h - Given a research objective for the mean of a single response variable, **state** the model for the data generating process. (A, E, F)
- 2.i - **Define**, in the context of a specific research objective, the terms *independence* and *identically distributed*. (D, H)
- 2.j - **Compare** and **contrast** *point estimation* and *interval estimation*. (E)
- 2.k - Given a set of data, **compute** the confidence interval for the mean response via bootstrapping. (E, G)
- 2.l - **Interpret** a given confidence interval in the context of a problem description. (E, G)
- 2.m - **Identify** common misconceptions about the interpretation of a confidence interval. (E)
- 2.n - Using a confidence interval **discern** the plausibility of a specified model (hypothesis). (E)
- 2.o - **Describe** the effects of changing the sample size, *confidence level*, and variability in the data on the size of the resulting confidence interval. (E)
- 2.p - **Describe** the role of a sampling distribution in conducting statistical inference (D)

Module 3: Null Distributions and P-Values

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 3.a - **Define** the term *null distribution*. (D)
- 3.b - **Compare** and **contrast** the sampling distribution and the null distribution of a statistic. (D, E)
- 3.c - **Describe** the impact of the null hypothesis on a model for the data generating process. (E)
- 3.d - Given a model for the null distribution, **state** whether the sample on which the model is based provides *evidence* against the null hypothesis. (D, E)
- 3.e - **Define** the term *standardized statistic* and **describe** how it quantifies evidence against the null hypothesis. (E)
- 3.f - Using statistical software, **compute** the standardized statistic for a hypothesis about the mean of a single population. (E)
- 3.g - **Define** the term *p-value*. (E)
- 3.h - Given the p-value, **discern** the plausibility of a specified hypothesis. (E)
- 3.i - Given the results of a hypothesis test (confidence interval or p-value), **interpret** the decision in context of the problem statement. (E)
- 3.j - **Describe** the computation of a p-value given a model for the null distribution. (E)
- 3.k - Using statistical software, **compute** the p-value for a hypothesis about the mean of a single population. (E)

Module 4: Simple Linear Regression

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 4.a - **Construct** and **interpret** a *scatterplot* which summarizes the relationship between a response and a quantitative predictor. (C)
- 4.b - Given a graphic, **estimate** the *correlation coefficient* summarizing the relationship between two quantitative variables and **interpret** the value. (C)
- 4.c - **State** the properties of the correlation coefficient and **describe** its limitations - namely, that it only describes linear association. (C)
- 4.d - **Distinguish** between association and causation, and given a description of a data collection scheme, **identify** whether causal inference is appropriate. (C, E, H)
- 4.e - Given a research question, **identify** the *response* and *predictor* of interest and write in mathematical notation the equation for the data generating process. (A, B, E)
- 4.f - **Describe** the motivation for *least squares estimation*. (E)
- 4.g - Using statistical software, **compute** the least squares estimates. (E)
- 4.h - **State** the four conditions which might be placed on the stochastic portion of the data generating process. (F)
- 4.i - Using statistical software, **compute** confidence intervals for the parameters of the data generating process under various conditions. (E)
- 4.j - **State** what is meant by the *classical regression model*. (E, F)
- 4.k - **Describe** two modeling approaches for the sampling distribution of the least squares estimates. (D)
- 4.l - **Interpret** the *population slope* and *population intercept* of a linear model under the condition that the deterministic portion of the model is correctly specified. (A)
- 4.m - **Define** and **identify** *extrapolation* and **describe** the danger of extrapolation. (E)
- 4.n - Given a research question regarding the *marginal relationship* between two quantitative variables, **state** the hypotheses which capture the question of interest in terms of the parameters of the data generating process. (A)
- 4.o - Given the least squares fit and a specified set of values for the predictor, **predict** the value of the response and **interpret** this prediction under the condition that the deterministic portion of the model for the data generating process is correctly specified. (C, E)
- 4.p - Given confidence intervals for the parameters of the data generating process, **describe** the conclusions that can be drawn regarding the research objective. (E)
- 4.q - **State** the properties of a *density function*. (F)
- 4.r - **State** the properties of the *Normal distribution*. (F)

Module 5: Partitioning Variability

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 5.a - **Describe** what the *Total Sum of Squares*, the *Regression Sum of Squares* and the *Error Sum of Squares* quantify. (E)
- 5.b - **State** the equation which partitions the variability in the response. (E)
- 5.c - **Describe** how partitioning the variability in the response is helpful in assessing if a predictor belongs in the data generating process. (E)
- 5.d - **State** the *degrees of freedom* associated with each sum of squares. (E)
- 5.e - **State** the formula for computing the *mean squares*. (E)
- 5.f - **State** the formula for a standardized statistic for testing a hypothesis about the parameter in the model for the data generating process, which results from partitioning the variability. (E)
- 5.g - Using statistical software, **compute** the standardized statistic and corresponding p-value associated with testing a hypothesis about a parameter in the model for the data generating process. (E)
- 5.h - **Define** the term *R-squared*; given the sums of squares, **compute** the R-squared for a regression fit. (C, E)
- 5.i - **Define** the terms *indicator variable* and *reference group*. (E)
- 5.j - Using indicator variables, **state** a model for the data generating process which relates a quantitative response with a two-level categorical predictor. (E)

Module 6: Assessing Conditions in Regression

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 6.a - **Define** the term *residual*. (F)
- 6.b - Using statistical software, **compute** residuals for a set of data given the least squares estimates for the parameters. (F)
- 6.c - **State** the appropriate graphic for assessing whether a dataset is consistent with the condition of the errors having mean 0 for all values of the predictor. (F)
- 6.d - **State** the appropriate graphic for assessing whether a dataset is consistent with the condition that the errors are independent and **identify** when it is appropriate to rely on this graphic instead of solely the context of the data collection scheme. (F)
- 6.e - **State** the appropriate graphic for assessing whether a dataset is consistent with the condition of the errors are identically distributed or that the variability in the error does not depend upon the value of the predictor. (F)
- 6.f - **State** the appropriate graphic for assessing whether a dataset is consistent with the condition that the errors follow a Normal distribution. (F)
- 6.g - Given graphics for assessing the conditions of a regression model as well as a description of how the data was collected, **comment** on whether it is reasonable to assume the data is consistent with each condition. (F)
- 6.h - Given graphics for assessing the conditions on a regression model, **choose** an appropriate approach for modeling the sampling distribution of the parameter estimates or null distribution of the standardized statistic and **justify** your choice. (E, F, J)

Module 7: Analysis of Variance

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 7.a - Given a question of interest comparing the mean of a quantitative response across levels of a factor, **state** the null and alternative hypotheses that capture the question of interest. (A)
- 7.b - **State** and **define** the three aspects of a well-designed study (*replication, randomization,* and *comparative groups*). **Describe** the benefit of each in designing a study. (B, F, G)
- 7.c - **Define** the term *power* used to describe a statistical analysis. (B, H)
- 7.d - **Construct** and **interpret** side-by-side box plots (overlayed density plots, side-by-side violin plots, and side-by-side individual value plots) for comparing a quantitative variable across the levels of a factor. (C)
- 7.e - Given a research objective, **state** the model for the data generating process which explains a quantitative response as a function of a categorical predictor. (A)
- 7.f - **State** three conditions we can impose on the stochastic portion of the data generating process. (F)
- 7.g - **State** the *classical ANOVA model*. (F)
- 7.h - **Describe** the two sources of variability (*between group* and *within group*) and the rationale for partitioning the variability in order to quantify the signal. (E)
- 7.i - **State** the standardized statistic for comparing the mean of a quantitative response across levels of a categorical predictor. (E)
- 7.j - Given a partially complete ANOVA table, **complete** the table using known relations between the key components. (E)
- 7.k - Using statistical software, **compute** the ANOVA table which summarizes the partition of the variability in the response such that the p-value adheres to any stated conditions. (E)
- 7.l - Given an ANOVA table, **state** the conclusions which can be drawn regarding the research objective in the context of the problem. (E)

Module 8: Assessing Conditions in ANOVA

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 8.a - **State** the appropriate graphic for assessing whether a dataset is consistent with the condition that the errors are independent and **identify** when it is appropriate to rely on this graphic instead of solely the context of the data collection scheme. (F)
- 8.b - **State** the appropriate graphic for assessing whether a dataset is consistent with the condition of the errors are identically distributed or that the variability in the error does not depend upon the value of the predictor. (F)
- 8.c - **State** the appropriate graphic for assessing whether a dataset is consistent with the condition that the errors follow a Normal distribution. (F)
- 8.d - Given graphics for assessing the conditions of a regression model as well as a description of how the data was collected, **comment** on whether it is reasonable to assume the data is consistent with each condition. (F)
- 8.e - Given graphics for assessing the conditions on a regression model, **choose** an appropriate approach for modeling the sampling distribution of the parameter estimates or null distribution of the standardized statistic and **justify** your choice. (E, F, J)

Module 9: Block Designs

After completing this module, you should be able to perform the tasks listed below. Mastery of these skills will enable you to successfully complete course assignments and the course objectives discussed in the **Course Syllabus**.

- 9.a - **Define** the term *blocking* and **describe** the use of blocking in a controlled experiment. (B)
- 9.b - **Define** the term *randomized complete block design* and **identify** when it is being employed based on a description of the data collection process. (B)
- 9.c - **Describe** the impact of blocking on the responses - namely that it induces correlation among responses from the same block. (B)
- 9.d - **Construct** graphics which highlight the correlation structure in the responses when it is present. (C)
- 9.e - **State** the form of the data generating process which describes a quantitative response as a function of a categorical predictor while accounting for a fixed number of blocks. (A)
- 9.f - **Interpret** the parameters in the *repeated measures ANOVA model* for the data generating process and **contrast** them with the parameters from the ANOVA model. (A)
- 9.g - **Contrast** *fixed effects* and *random effects* and **identify** each given a study description. (A)
- 9.h - **State** the conditions on the stochastic portion of the model as well as the random effects in the model associated with the *classical repeated measures ANOVA model*. (F)
- 9.i - Given a problem description and graphic of the data, **identify** whether it is safe to assume that the random effects are identical for all observations with a given block. (F)
- 9.j - **Describe** the difference in the partitioning of the variability that occurs for ANOVA models and repeated measures ANOVA models. (E)
- 9.k - Using statistical software, **compute** the test statistic and corresponding p-value for comparing the levels of a categorical fixed effect while accounting for blocks. (E)