* + 1. **Data Analysis**

# Goals:

Students develop skills in drawing conclusions and critically evaluating results based on data.

# Expected Learning Outcomes:

* + - 1. Students understand basic concepts of statistics and probability.
      2. Students comprehend methods needed to analyze and critically evaluate statistical arguments.
      3. Students recognize the importance of statistical ideas.

*Courses proposed for this component of the General Education (GE) should be designed with these goals and expected learning outcomes (ELOs) in mind and considered in terms of their contribution to the requirement as a whole. Courses will be reviewed by the Arts and Sciences Curriculum Committee (ASCC) in light of these goals and expected learning outcomes. All GE courses should be made available to undergraduates with a minimum of prerequisites and not be restricted to majors.*

# Proposals must include the following:

1. The appropriate Course Request Form via [curriculum.osu.edu](http://eca.osu.edu/)
2. A course syllabus that follows the ASC syllabus template guidelines (see pp. 13-15).
3. A concurrence should be solicited from the Department of Statistics specifically addressing the requested GE Data Analysis status.
4. A GE rationale that addresses how the course will meet the required coursework outlined below (pp. 59-60).
5. A GE assessment plan which explains how the faculty teaching the course will assess the effectiveness of the course in achieving the GE expected learning outcomes over time, rather than how individual student grades will be assessed. As you develop your GE assessment plan, please bear in mind that the faculty will need to implement it from the very first offering of the course so keep it simple (a GE assessment plan should not be so complex that it cannot be implemented).

Complete the following table to show how the faculty will assess the three expected learning outcomes. Then, in an appendix, provide one or more specific example(s) for each assessment method you will use.

|  |  |  |  |
| --- | --- | --- | --- |
| **GE Expected Learning Outcomes** | **Methods of Assessment**  *\*Direct methods are required. Additional indirect methods are encouraged.* | **Level of student achievement expected for the GE ELO.**  *(for example, define percentage of students achieving a specified level on a scoring rubric)* | **What is the process that will be used to review the data and potentially change the course to improve student learning of GE ELOs?** |
| **ELO 1** | 1. **Portfolio sample student work.**  * Six copies of problem sets and excel exercises will be made, with one copy representative of work that received a grade of A, B, C, etc. * **SAMPLE – Problem Sets:** See Appendix 1 below. * **SAMPLE –** **Excel Exercises:** See Appendix 2.  1. **Review of course materials.**  * A copy of the syllabus, both exams, and instructions for the problem sets and excel exercises will be kept on file to review prior to the next offering and as a resource for the peer evaluator. | 1. 85% of students to achieve 80% or higher based on grading rubric. | 1. Portfolio of student work of problems and excel exercises 2. Review of course materials 3. Embedded test questions 4. Portfolio of student discussion prompts and responses 5. Peer evaluation 6. Student SEI evaluations   Items 1-5 will be maintained on file in the department (with the Academic Affairs Committee Chair) and with the instructor so that the progress of the course can be monitored and evaluated across time as the course evolves and to enable the department to address any major concerns or drift from the established goals and standards. |
| Students understand basic concepts of statistics and probability. |
| **ELO 2** | **3. Embedded test questions**   * The exams from each time the course is offered will contain several similar questions that address GE Data Analysis, departmental, and course learning objectives. Comparisons will be made between course offerings in terms of content and cumulative scores on the embedded questions. * **SAMPLE – Lecture Exam Question:** See Appendix 1 | 1. 85% of students to achieve 70% or higher based on grading rubric |
| Students comprehend methods needed to analyze and critically evaluate statistical arguments. |
| **ELO 3** | **4. Portfolio of student discussion prompts and responses**   * Students are required to post once per week on that week’s topic. Discussion forum posts count for 8% of the overall grade. * **SAMPLE – Discussion Prompt: See Appendix 1**   **5.** **Peer Evaluation**   * The course will be evaluated by a faculty colleague once a year in a manner consistent with current HCS policy. The evaluation will contain at least one lecture visit. A copy of the written evaluation will be kept with the other evaluation material from the class by the instructor and the Chair of the Department. | 1. 100% of students will participate in the discussion and achieve 90% or higher on grading rubric. |
| Students recognize the importance of statistical ideas. |

**\*Direct Methods** *assess student performance related to the expected learning outcomes. Examples of direct assessments are course-embedded questions; pre/post test; standardized exams; portfolio evaluation; videotape/audiotape of performance; rubric-based evaluation of student work.*

**\*Indirect Methods** *assess opinions or thoughts about student knowledge, skills, attitudes, learning experiences, and perceptions. Examples of indirect measures are student surveys about instruction; focus groups; student self-evaluations.*

After the second offering of the course, please submit an initial report summarizing the GE assessment results following the format of the “Assessment Report Requirements” in Appendix 11.

1. *For ASC units only*: If the GE request applies to a new course and the new course can also count toward the major of the submitting unit (whether as a required course or as an elective), please include the curriculum map of that program to which you have added the newly proposed course, indicating the program goal(s) and levels it is designed to meet. If the course is not new but the request involves moving the course to a new level or place on the major’s curriculum map, the updated map will need to be provided as well.

# Required Coursework for Students:

The intent of this General Education category is to enable students to deal with the gathering, presentation, and interpretation of data. Students should develop an understanding of problems of measurement, be able to deal critically with numerical and graphical arguments, and recognize the uses and misuses of statistics and related quantitative arguments. Courses should include exposure to fundamental ideas of probability, involve the use of computational technology in problems of data analysis, and include opportunities to present data using summary measures and graphical techniques. Specialized courses within the B. S. major may also be proposed to satisfy this requirement.

The ASCC Natural and Mathematical Sciences Panel and the full ASCC will use these guidelines (approved by the ASCC on April 11, 2014) as the basis for evaluation of data analysis courses. The fulfillment of the following criteria would make the course eligible to be considered for GE Data Analysis status with the final decision based on the overall rigor and sophistication of the course. Prerequisite courses can count in the requirement (for example Statistics 4202 meets the requirement because Statistics 4201 is a prerequisite).

Core requirements (at least 4 instructional hours spent on each bullet):

* + Notions of probability. The axioms of probability, and basic probability calculations. Random variables, and probability calculations using random variables. Expected values.
  + Basics of statistical inference. Moving from a sample to a population. Bias and variance. Understanding the margin of error and confidence. The logic of statistical testing. The misuse of statistics.

Additional requirements (At least two out of four, with at least 3 instructional hours spent on each numbered item):

1. Understanding where data come from. Data sources. Discriminating between observational and experimental studies. (Random) sampling.
2. Summarizing data graphically and numerically. Discriminating between good and bad summaries. Understanding the advantages and disadvantages of a given summary.
3. Methods of statistical inference. Statistical testing. Constructing confidence intervals. Making quantitative statistical arguments using data. Understanding and verifying assumptions underlying a given inference.
4. Statistical modeling (e.g., regression models, analysis of variance). Interpreting the parameters underlying statistical models. Model assessment.

Thus, in a three-semester-hour course, for example, at least one third of the class should be spent teaching topics in probability and statistics.

Possible software: The R Project for Statistical Computing ([www.r-project.org](http://www.r-project.org/)) is an open source statistical software package. Commercial packages for which the university has a license include MATLAB, Mathematica, Minitab, JMP, SAS, SPSS, and Stata.

Useful reference: "Statistics: Concepts and Controversies, 8th ed." by Moore and Notz (The Statistics 1350 text).

**APPENDIX 1**

**Sample problem from problem set 1:**

**1)**  A poll was conducted to determine the topics that voters are most concerned with. The findings show that 40% are concerned with the influence of wealthy people and corporations, 35% are concerned about negative attacks and 20% are concerned that nothing will change regardless of who wins. The survey was based on 1005 registered voters selected at random from all US voters.

**a.**  Describe the sample.

**b.** Describe the population from which the sample was selected.

**c.** Is the sample representative of the population?

**d.**  What is the variable of interest?

**e.** How is the inference expressed?

**f.**  Newspaper accounts of most polls usually give a margin of error (i.e. plus or minus 3%) for the survey

result. What is the purpose of the margin of error and what is its interpretation?

**Sample embedded test question:**

**1)** (15 Points) Suppose your objective is to find the mean and range of prices for a gallon of gas at Franklin County gas stations on a particular day. You drive around town, noting the price at different stations, until you’ve visited 20 stations.

**a**. What is the population in this example?

**b.** What is the sample?

**c.** What is the variable of interest?

**d.** Suggest a better way to do this experiment.

**Sample student discussion prompt and responses:**

**Prompt:** At the end of chapter 2, the topic is "Distorting the truth with deceptive statistics" and I spend some timing talking about some examples that are mostly hypothetical, in which you could use statistics or data presentation to make things seem different than what they really are.  This is most often done for the purpose of convincing you to spend money or to agree with a particular interest or point of view.  After viewing my examples, I'd be interested in knowing what examples of this you have seen or have heard of in your day-to-day lives.  A quick internet search of the topic may bring up some articles, many quite humorous, that you can read.

**Sample responses:** During my childhood sugary cereals were strictly banned from the household. If mama caught you sneaking in the savory taste of Lucky Charms or Coco-Puffs into the cart, you'd be leaving the store with a sore bum. However, like in most laws, there were loop holes to this rule. I was allowed to have Mini-Wheats, the smaller, slightly sweeter version of the hay bale in the bowl Shredded Wheat.

Kellogg's at the time was promoting Mini-Wheats being the bee's knees for young students, claiming that if you eat their cereal you'll do better in school. In fact, after finding the commercials again, the quote was  "A clinical study showed kids who had a filling breakfast of Frosted Mini-Wheats cereal improved their attentiveness by nearly 20 percent."

I'm not so sure if mother ever watched the commercials and actually bought into it, but it seemed to work for Kellogg's.  51% of surveyed adults said they bought that cereal because of that claim and were disappointed when they found out there was no improvement with their kids. Surprisingly, this ended with a law suit of $4 million because of the ad.

Kellogg' still say they did nothing wrong, and I'm hungry.

**Appendix 2**

**Computer Exercise 4**

HCS 2260

**Demonstration of the Central Limit Theorem, Probabilities of Normal Random Variables, Random Number Generation and Descriptive Statistics.**

**Objective: The purpose of this lab is to examine the Central Limit Theorem, to see if what it says about the distribution of x is true.**

**Step 1.**Generate random samples of specified size *n* from digits 0 through 9.

Let’s begin with samples of size *n* = 5. *Excel* has a random number function, RAND(), that will generate random numbers from 0 to 1, including 0 but not including 1. We can generate random numbers from 0 to 10, including 0 but not including 10 by multiplying RAND() by 10. *Excel* also has an integer function, INT(), which will find the greatest integer value smaller than the argument

value between the parentheses. For example INT(4.98) will return a value of 4. Thus, the

command INT(10\*RAND()) will return the integer values from 0 to 10, including 0 but

not including 10. That is, INT(10\*RAND()) will return a random integer from 0 to 9,

including both 0 and 9.

In cell A1, type in =INT(10\*Rand()), and press **Enter**. Highlight cells A1 through C1.

Click **Edit,** and choose the **Fill** and **Right** options. When this is done, cells A1 through

C1 will be filled with integer values from 0 to 9, inclusive. This represents our sample of

size 3.

**Note:** As further entries are made in this spreadsheet, the random values previously

determined will be regenerated. That is, the values currently in cells A1 to C1 will change

as further entries are added.

**Step 2.** Compute the sample mean x – of this sample.

We will use *Excel’s* AVERAGE() function to compute x–. In cell E1, type in

=AVERAGE(A1:C1) and press **Enter**. The average of the values in cells A1 to C1 now

appears in cell E1.

**Step 3.** Generate an x – distribution.

Highlight cells A1 through E100. Click on **Edit** and choose the **Fill** and **Down** options.

The 100 entries down the E column constitute a sample x – distribution.

**Step 4.** Find the mean and standard deviation of this x – distribution. That is, find x –

x and Sx.

Click on cell I1; type in =AVERAGE(E1:E100), and press **Enter**. Click on cell G3; type

in =STDEV(E1:E100), and press **Enter**. The mean of the x – distribution now appears in

cell G1, and the standard deviation of x – distribution now appears in cell G3.

**Step 5.** Tally the number of x – values that fall from 0 to 1, 1 to 2, 2 to 3, . . . , 9 to 10.

This requires some slick maneuvering. *Excel* does not have a function that will allow us

to directly make such tallies. Thus, we will find these counts with a two-step process.

First, we will use the COUNTIF() function.

Click on cell G5, type in =COUNTIF(E1:E100,"<1"), and press **Enter.**

Click on cell G6, type in =COUNTIF(E1:E100,"<2"), and press **Enter.**

Click on cell G7, type in =COUNTIF(E1:E100,"<3"), and press **Enter.**

Click on cell G8, type in =COUNTIF(E1:E100,"<4"), and press **Enter.**

Click on cell G9, type in =COUNTIF(E1:E100,"<5"), and press **Enter.**

Click on cell G10, type in =COUNTIF(E1:E100,"<6"), and press **Enter.**

Click on cell G11, type in =COUNTIF(E1:E100,"<7"), and press **Enter.**

Click on cell G12, type in =COUNTIF(E1:E100,"<8"), and press **Enter.**

Click on cell G13, type in =COUNTIF(E1:E100,"<9"), and press **Enter.**

Click on cell G14, type in =COUNTIF(E1:E100,"<10"), and pre**ss Enter.**

Click on cell H5, type in =G5, and press **Enter.**

Click on cell H6, type in =G6-G5, and press **Enter.**

Highlight cells H6 through H14; click on **Edit,** and choose options **Fill** and **Down.**

The values in cells H6 through H14 now contain the desired tallies.

**Step 6.** Construct a chart of the tallies.

Highlight cells H5 through H14. Click the **Insert** menu and the **Chart** option. From the

Chart type: menu, choose **XY (Scatter),** and from the Chart sub-type, click on the icon in

the second row second column. (When you click on this icon, the text below the images

will say, “Scatter with points connected by smoothed Line without markers.”) Click

**Finish,** and the chart will be created.

The graph should appear relatively mound shaped and symmetrical (that is, approximately normal).

Recall that when further entries are made in unused cells, the random numbers are regenerated. Click on an unused cell (such as K1). Press **Delete**. The random numbers are regenerated, and the chart is adjusted accordingly. Repeat this process as often as you like to see that the x – distributions are approximately normal.

**Step 7.** Repeat this process with randomly generated samples of size *n* = 10

We do not need to start from scratch in order to repeat this process with randomly

generated samples of size *n =* 10.

First copy the WHOLE Sheet1, and paste on Sheet 2. Then, we simply need to insert seven additional columns of randomly generated numbers within the five we already have. Highlight columns C by clicking at the top of the column. With the right button of your mouse, click **Insert** option. Repeat this 7 times to insert 7 blank columns to the left.

Highlight columns A through J, by clicking at the top of the columns. Click on the **Home** menu, and choose options **Fill** and **Left**.

The blank columns should be filled in with random numbers from 0 to 9.

Scroll over to column L. The entries in column L are now the sample means of the

samples of size 10. Likewise, the entries in columns N and O are updated using the

information from the samples of size 10.

**Step 8.** However, the chart in sheet 2 is actually a graph of the distribution in sheet 1 (where the sample size was *n*=3). To fix this, click on the graphed line in the chart.

In the formula bar it will say this:

=SERIES(,,Sheet1!$H$5:$H$14,1)

Edit this to read

=SERIES(,,Sheet2!$O$5:$O$14,1)

Cells O5-O14 will be highlighted. Now the chart of the tallies is updated to reflect the distribution when the sample size *n*=10.

**Step 9.** Click on cell N35 in sheet 2. Briefly compare the two distributions you generated and tell me which distribution seems more normal, the one of a sample size *n*=3 in sheet 1 or the one of sample size *n*=10 in sheet 2.