Plenty of Fish in the Sea: Counting Fish Using Probability

Subject (Focus/Topic):	Normal and hypergeometric probability distributions
Grade Level:	Sophomore university level or above (14+)
Average Learning Time:	Two 65-minute class periods
Author:	Dr Emily Cilli Turner TAS-2018 University of La Verne 1950 Third Street, La Verne, CA 91750 ecilli-turner@laverne.edu
Creation Date:	March 31, 2019

LESSON PLAN DESCRIPTION

Lesson Summary (Overview/Purpose)

Students will apply the knowledge they already have on hypergeometric and normal probability distributions to determine properties of the samples of pollock collected onboard the NOAA Ship Oscar Dyson, such as type of distribution, amount of pollock expected in a certain length or weight range, how many pollock should be sampled to ensure certain properties and the basics of how the fish samples are used to estimate numbers of pollock in the Eastern Bering Sea.

Overall Concept (Big Idea/Essential Question)

Students will understand how to apply probability theory techniques learned in class to real-world situations and understand how the NOAA scientists use mathematics in their daily work. This will show students a possible career path for when they finish their degrees.

Specific Concepts (Key Concepts)

Applications of the normal probability distribution, theory behind the Central Limit Theorem, using the Central Limit Theorem to determine sample sizes, applications of the hypergeometric probability distribution as applied to counting fish

Focus Questions (Specific Questions)

List no more than 10 specific questions about the subject that students will explore in order to understand both the listed specific concepts and overall concept.

1. How can we apply the normal probability distribution to analyze real-world data?
2. What is the Central Limit Theorem and how can we use it to determine sample sizes for specified confidence intervals?
3. How can we use the hypergeometric distribution to understand the capture-recapture method of counting fish?

Objectives/Learning Goals

- Students will be able to identify probability distributions by looking at histograms based on their prior knowledge of distribution types.
- Using an online applet, students will investigate underlying distributions and changing sample size to understand the Central Limit Theorem.
- Students will apply the Central Limit Theorem to determine ideal sample sizes for data collection.
- Students will work through an application of the hypergeometric distribution and how it is used to count fish in a big population. This will familiarize them with the work of the NOAA scientists and allow them to see how mathematics is used in this job.

Background Information

Students will have already learned the basics of discrete and continuous random variables. Additionally, students need in depth knowledge of the normal and hypergeometric random variable, including working with specific examples and looking up values in the normal distribution table. Ideally, this lesson will be taught just after learning about Markov's and Chebyshev's Inequalities as these are needed to prove and understand the Central Limit Theorem.

Common Misconceptions/Preconceptions

Students need to remember the mechanics of working with normal and hypergeometric random variables to successfully complete the worksheet. Students also need to carefully follow directions on the lab. I expect the most difficulty on worksheet question 3b as this requires a lot of algebraic manipulation and simplification.

Teaching Materials

Appended Resources (AR's)

Non-interactive, printed materials <u>inserted</u> <u>into</u> <u>this document</u> after the Lesson Plan Description.

- Central Limit Theorem Worksheet (Pages 5-6)
- Applications of Probability Distributions (Pages 7-8)

Digital Resources (DR's)

Interactive, electronic files created by the Teacher-At-Sea to support this lesson.

- **Downloadable data set** = Excel Data Sheet: MACE 2016 EBS LW
  Filename: TAS-LP-2018-cilliTurner-emily-science-lesson-data-508-DR.xlsx
  Link: https://noaateacheratsea.blog/wp-content/uploads/2021/09/TAS-LP-2018-cilliTurner-emily-science-lesson-data-508-DR.xlsx

Physical Resources

Items used in the classroom to facilitate learning, especially during activities:

- see Technical Requirements

Technical Requirements

Computer(s) with internet access to download data spreadsheet, CLT applet, and complete worksheet

Teacher Preparation

The teacher needs to review the random variable types and make sure that students have ready access to the properties of each type. This will most likely be done earlier in the form of an exam note sheet.

Keywords

- Central Limit Theorem
- Capture-recapture method
- Walleye pollock

Lesson Procedure

Day 1:

1. Teacher will pass out lab worksheet (Central Limit Theorem Worksheet) to students and ensure each student has access to a computer to complete the lab. (5 min)
2. An introduction will be given to the data set, where it came from, and the specific data variables it contains. The teacher will discuss how the data was collected and how the NOAA scientists sample the ocean. (10 min)
3. Students will work individually on completing the lab worksheet. (40 min)
4. Wrap-up: Teacher will make sure that students understand the Central Limit Theorem and connect it to the data that they saw on the computer. (10 min)

Day 2:

1. Teacher will group students and pass out worksheet. (5 min)
2. Teacher will review what was done last class and rewrite the central limit theorem. Teacher will also get input from students about what they remember about the normal probability distribution and the hypergeometric probability distribution. (15 min)
3. Students will work in groups on completing the worksheet. Then teacher will call for volunteers to present select problems on the board. (35 min)
4. Wrap-up: Teacher will recap the student presentations and highlight that these are nice applications of the mathematics in the real world. (10 min)

Assessment and Evaluation

Students will complete in-class worksheets in groups. Formative assessment will be done in class while students are working, but later summative assessment would be in the form of an exam question about these topics.

Standards

Next Generation Science Standards (NGSS) or State Science Standards Addressed

There are no state standards for university education. However, the Mathematical Association of America's Curriculum Guide for math majors recommends that:

- "Because statistics is first and foremost about using data to inform thinking about real-world situations, it is critical that prospective teachers have realistic problem-solving experiences with statistics."
- In addition, the MET recommends that, "secondary mathematics teachers should have experience with exploring data, planning a study, anticipating patterns, using statistical inference, and applying appropriate technologies."

Ocean Literacy Principles Addressed

- Students will be "able to make informed and responsible decisions regarding the ocean and its resources."

Topic: Central Limit Theorem

In this lab, we will explore the Central Limit Theorem (CLT), which is one of the most remarkable and useful results in probability theory. We have seen an instance of the CLT previously, but here we will explore the theory and see why empirical data so often exhibits the properties of a normal curve.

1. Download the data set from the digital resources. These data are from the National Oceanic and Atmospheric Administration's (NOAA) 2016 Eastern Bering Sea Pollock Survey. The variables displayed in the file are haul number, common name (species), length (cm), weight (kg), and age. Age is not given for every specimen because only a sample of the pollock are aged in each haul and non-pollock species are not aged.

   (a) Note that the first column of data shows the haul number. Make a separate histogram in Excel of all the data for Haul numbers 5, 108, 119, 136, and 200. Sketch each histogram below and identify the type of each probability distribution from the shape of the histograms.

   (b) Make a histogram in Excel of all the data for common name "walleye pollock". Sketch the histogram and identify the type of the probability distribution from the shape of the histogram. Also, find the mean and variance.

   (c) If X_i represents the length of the ith walleye pollock in the sample, most likely what type of random variable is X_i? Explain.

2. Open the applet found here.

   (a) Choose the Exponential distribution and sample size 1. Click "Generate samples". Describe the shape of the distribution.

   (b) Again, choose the Exponential distribution, but now change the sample size to 2. Click "Generate samples". Describe the shape of the distribution. Repeat for sample sizes of 5, 10, 20 and 100. What do you notice?

   (c) Repeat parts (a) and (b) of this problem with the Uniform distribution. What do you notice?

What you probably noticed above is that regardless of the shape of the initial distribution, when you start taking samples of size n from the data set, the distribution of the sample means tends towards a normal distribution. This idea is known as the Central Limit Theorem.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then the distribution of $\dfrac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$ tends to a standard normal distribution as $n \to \infty$.

That is, for $-\infty < a < \infty$, $P\left\{\dfrac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}\right\} \to \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{a} e^{-x^2/2}\,dx$ as $n \to \infty$.

1. Suppose that we have a haul of walleye pollock that we believe have lengths uniformly distributed between 25 and 42 cm and let X_i be the length of the ith pollock in the sample.

   (a) Use the Markov inequality to obtain a bound on $P\left\{\left(\sum_{i=1}^{20} X_i\right)/20 > 30\right\}$.

   (b) Use the Central Limit Theorem to approximate $P\left\{\left(\sum_{i=1}^{20} X_i\right)/20 > 30\right\}$.

2. Recall that the mean length of the walleye pollock from the 2016 NOAA Eastern Bering Sea Pollock Survey was 38.5047 and the variance was 55.7473. How many walleye pollock must be sampled to ensure that the average length would be within ± 0.5 cm of the mean for this sample with probability 0.95?

3. *(Capture-Recapture Method)* Scientists on the NOAA Eastern Bering Sea Pollock Survey are tasked each year with estimating the number and biomass of the walleye pollock that live in the Eastern Bering Sea. As the pollock cannot simply be counted individually, advanced probability and statistical techniques are used in these estimations. This problem will walk you through a simplified version of how this estimation takes place.

   (a) Suppose we have a bay (not the whole ocean!) where m pollock have been captured, tagged and released. So we have a total population of pollock and this smaller number of tagged pollock. Then we catch a sample of n pollock from the bay and find that i of them are tagged. Write an expression for $P\{X = i\}$, where X represents the number of tagged pollock in the sample. [Hint: What type of random variable is X?]

   (b) We want to determine N, which is the total number of pollock in the bay. The value of N is the value that makes the observed value of 20 tagged pollock the most likely, which is referred to as a *maximum likelihood estimate*. That is, it is the value where $P(N) > P(N-1)$ or $1 < \frac{P(N)}{P(N-1)}$. Find and simplify an expression in terms of N, n, m and i for the right side of the second inequality.

   (c) Solve your simplified inequality for N. What does this imply?

   (d) Suppose 50 pollock are initially tagged in the bay and a sample of 100 has 10 tagged pollock. What does (c) say is the most likely number of total pollock in the bay?