Code

Project 3 Instructions

Last Updated April 22, 2024

Project Outline

This project outline and background information have been provided to assist you as you complete your project. You should assume the reader of your work has no knowledge or access to this information.

How much explosive material remains in the soil? The military is concerned about amount of explosives that remain in the soils of training ranges because soldiers are out on the training range and explosive residues are carcinogens (they cause cancer). Collecting a representative sample from the training ranges is difficult to do and researchers are continuing to work on developing, refining, and validating sampling methods. Source (https://serdp-estcp.org/content/download/5167/73264/file/ER-0628-FR.pdf)

Our work in this project relies on assumptions and use of probability distributions. We will (1) fit probability distributions to data, (2) create simulated samples from those fitted probability distributions, and (3) use the fitted probability distributions to provide probability information about explosive residue in the surface soil of a training range.

The Environment Protection Agency (EPA) specifies regulatory levels of Nitroglycerin concentrations for human heath. For this project, we will use 10 mg/kg[1] as the toxicity threshold for surface soil of this training range. Source (https://cfpub.epa.gov/ncea/pprtv/documents/Nitroglycerin.pdf)

The Four Probability Models

There are 4 probability models we'll be exploring in this project. They are the uniform distribution, the normal distribution, the gamma distribution, and the exponential distribution.

The Uniform Distribution

The probability density function for the uniform distribution is

- $f_0(x; a, b) = \frac{1}{b-a}$ for $a < x < b$ (and 0 otherwise).

We have seen this probability model before in our Example Project. The probability is the same (uniform) for all measurements between a and b. The command `?dunif` or `?runif` in your R console will access the R documentation for the uniform distribution.

The Normal Distribution

The probability density function for the normal distribution is

- $f_1(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ with $-\infty < \mu < \infty$, $\sigma > 0$, and $-\infty < x < \infty$.

The probability that measurements are near μ is larger than the probability that measurements will be far from μ. This distribution is symmetric and defined for all values of x (both positive and negative values are possible). We have seen this probability model before in our Example Project and in Project 2. The command `?dnorm` or `?rnorm` in your R console will access the R documentation for the normal distribution distribution.

The Gamma Distribution

The probability density function for the gamma distribution is

- $f_2(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ with $\alpha > 0$, $\beta > 0$, and $x \geq 0$ (and 0 otherwise).

The function $\Gamma(x)$ is call the **gamma function** (NOT the gamma distribution). **The gamma *distribution*** and **the gamma *function*** are **different functions.** One way to think about the gamma function is a generalization of the factorial to noninteger values. We have seen this probability model before in our Example Project. The command `?dgamma` or `?rgamma` in your R console will access the R documentation for the **gamma distribution**. The command `?gamma` in your R console will access the R documentation pages about the **gamma function**.

The Exponential Distribution

The probability density function for the exponential distribution is

- $f_3(x; \lambda) = \lambda e^{-\lambda x}$ with $\lambda > 0$ and $x \geq 0$ (and 0 otherwise)

The exponential distribution is often thought of as a “waiting time” model. The command `?dexp` or `?rexp` in your R console will access the R documentation for the exponential distribution.

Task 1: Data and General Models

- Create an R Markdown file.
- Use the code below to read in the soil data.

Hide

```
# run this line once in the console to get the data4soils package
#devtools::install_github("byuidatascience/data4soils")

library(data4soils)
```

Hide

```
Ng <- cfbp_fpjuliet$ng
```

This code creates a vector called “Ng”. The vector contains the measurements Nitroglycerin (Ng) measured as mg/kg found in 100 soil samples.

- Familiarize your self with this Nitroglycerin data.
  - Create a **density histogram** of the 100 Nitroglycerin measurements.
  - Calculate the sample mean for these 100 Nitroglycerin measurements. Use the command `?mean` in your Console to access the R documentation for the `mean()` command.
  - Calculate the sample variance for these 100 Nitroglycerin measurements. Use the command `?var` in your Console to access the R documentation for the `var()` command.
  - Include your histogram, as well as your mean and variance computations, in your final report.
- Familiarize yourself with the 4 probability models given before Task 1.
  - We previously explored models f_0, f_1 (now $h = \mu$ and $a = \sigma^2$), and f_2 (now $h = 0$, $a = \alpha$, and $b = \beta$). Read the narrative for Example Project Task 2 to remind yourself what we noticed about how changing the parameters in these models changes the behavior of the functions f_0, f_1, and f_2. In this project, we are exploring explosives in soil rather than brightness of light bulbs. Use the Desmos files below for f_0, f_1, and f_2 in any way they are helpful as you complete this project.
    - Desmos file for function 0 (https://www.desmos.com/calculator/6snrwycie4)
    - Desmos file for function 1 (https://www.desmos.com/calculator/o2lb19ivnk)
    - Desmos file for function 2 (https://www.desmos.com/calculator/h03fdeh2ao)
  - The function f_3 is new. Use the Desmos file below to dynamically explore how changing λ changes the behavior of f_3.
    - Desmos file for function 3 (https://www.desmos.com/calculator/j8peg4jsy4)
  - As part of your analysis, create (in R) plots of at least 2 representative curves illustrating what you learned in your parameter exploration of f_3. (You do not need to include plots of your parameter exploration for the other 3 models). In your narrative, summarize your observations about the parameter λ in terms of transformations of functions (shifts, reflections, stretch) and the mathematical behavior of the functions (increasing, decreasing, constant, positive, negative, nonnegative).
- Visually fit f_2 and f_3 to the **density histogram** of the 100 Nitroglycerin measurements. Your plot should include a histogram (the data) and a curve (the model).
  - We use density histograms to visualize the list of measurements (the data) graphically.
  - A probability density function contains probability information about all possible measurements. When we graph a probability density function it will be a curve.
- Use the parameter values of your visually fitted f_2 model and the `rgamma()` command to simulate a sample of 25000 random measurements. Then use that sample to approximate how many measurements out of 25000 will have more than 10 mg/kg of explosive.
  - Use the `set.seed()` command in R to set the seed so your simulated samples and probability calculations are reproducible. Use your assigned seed.
  - Use the simulated sample to calculate the approximate probability that the amount of explosive in this sample will be more than 10 mg/kg?
    - Adjust the code below to approximate how many measurements out of 25000 will have more than 10 mg/kg of explosive.

Hide

```
#set.seed(2021)
#tmp2 <- rgamma(25000, shape = alpha, rate = beta)
#length(which(tmp2 > 10))
```

- Use the parameter values of your visually fitted f_3 model and the `rexp()` command to simulate a sample of 25000 random measurements. Then use that sample to approximate how many measurements out of 25000 will have more than 10 mg/kg of explosive.
  - Use the `set.seed()` command in R to set the seed so your simulated samples and probability calculations are reproducible. Use your assigned seed.

- Use the simulated sample to calculate the approximate probability that the amount of explosive in this sample will be more than 10 mg/kg?
  - Adjust the code below to approximate how many measurements out of 25000 will have more than 10 mg/kg of explosive.

Hide

```
#set.seed(2021)
#tmp3 <- rexp(25000, rate = lambda)
#length(which(tmp3 > 10))
```

- **CHECK YOUR WORK:**
  - For both f_2 and f_3, do the probability models that you plotted over the histogram appear to be visual fit?
  - Once you've chosen values for the parameters of each model, use the Shiny App (https://shiny.byui.edu/connect/#/apps/3c058e83-8470-40d5-87b5-ce6cb8f16d83/access) to check your probability calculations.
    - In your analysis include an image from the Shiny App to show the probability calculations from your simulations are correct.
- Organize your work into a **cohesive analysis** (https://byuistats.github.io/M119/specs_detail.html) and submit the html file on Canvas.

Task 2: Find the Expected Value and Variance of each Stochastic Variable to Fit the Model ("Modified Method of Moments" Method)

- Create a new R Markdown file.
- For each model (uniform distribution, normal distribution, gamma distribution, and exponential distribution) using mathematical notation write down the definite integral required to calculate the expected value of the distribution (or model).
  - Use Mathematica to compute the definite integrals (find the expected value of f_0, f_1, f_2 and f_3) as a function the model's parameters. Include your work from Mathematica as an image.
- For each model (uniform distribution, normal distribution, gamma distribution, and exponential distribution) using mathematical notation write down the definite integral required to calculate the variance of the model (or distribution).
  - Use Mathematica to compute the definite integrals (find the variance of f_0, f_1, f_2 and f_3) as a function each model's parameters. Include your work from Mathematica as an image.
- Set the sample mean of the Nitroglycerin data equal to the expected value of f_0, set the sample variance of the Nitroglycerin data equal to the variance of f_0, and solve this system of equations for a and b. Then state the fitted model, $f_0(x)$, with the parameters values rounded to 3 significant figures as needed.
- Set the sample mean of the Nitroglycerin data equal to the expected value of f_3 and solve this equation for λ. Then state the fitted model, $f_3(x)$, with the parameters values rounded to 3 decimal places as needed.
- The fitted models f_1 and f_2 are $f_1(x) = \frac{1}{\sqrt{2\pi(10.979)}} e^{-\frac{1}{2}\left(\frac{x-2.792}{3.313}\right)^2}$ where $-\infty < x < \infty$ and $f_2(x) = \frac{0.254^{0.710}}{\Gamma(0.710)} x^{0.710-1} e^{-0.254x}$ where $x \geq 0$.
  - Note when using fitted models it is best practice not to round any preliminary calculations, so when you use the fitted models to answer questions, make sure you use as many decimal places as possible for the parameter values, NOT the rounded values. For f_1, μ is the sample mean and σ^2 is the sample variance. For f_2, $\alpha = \frac{(\text{sample mean})^2}{\text{sample variance}}$ and $\beta = \frac{\text{sample mean}}{\text{sample variance}}$.
- **CHECK YOUR WORK:**
  - Use the Shiny App (https://shiny.byui.edu/connect/#/apps/85c8a9ef-45bf-440b-98ea-d360cb853d91/access) to check the parameters for the fitted functions f_0 and f_3.
    - Include an image from the Shiny App showing you found the correct parameter values.
  - Plot a density histogram of the the data and each the fitted functions. Does it look like you have found a function that fits the data?
- Organize your work into a **cohesive analysis** (https://byuistats.github.io/M119/specs_detail.html) and submit the html file on Canvas.

Project 3: Bringing it All Together (and answer a question)

- Create a new R Markdown file.
- Answer the question, "What is the amount of explosives in the soil?"
  - Begin with background and an introduction to the question(s) you will be answering with the explosives data.
  - Introduce the given data.
  - Introduce the four probability models.
  - Describe how you will fit the models (maybe what it means to fit those models).
  - Provide the fitted models.
    - The work to fit $f_1(x)$ and $f_2(x)$ was completed in class but results from class should be included.
      - Remember when using fitted models it is best practice not to round any preliminary calculations, so when you use the fitted models to answer questions, make sure you use as many decimal places as possible for the parameter values, NOT the rounded values. For f_1, μ is the sample

mean and σ^2 is the sample variance. For f_2, $\alpha = \frac{(\text{sample mean})^2}{\text{sample variance}}$ and $\beta = \frac{\text{sample mean}}{\text{sample variance}}$.

- For the fitted models f_0 and f_3, use an integral to calculate the probability that the amount of explosive in a sample will be between 0 mg/kg and 5 mg/kg? Write down the integral you need to calculate and then use Mathematica to compute the integral. Include your work from Mathematica as an image. What does this mean in the context of the concentration of explosive in a soil sample?
- For the fitted models f_1 and f_3 determine the 99$^{\text{th}}$ percentile of the distribution. Remember to write down the equation you need to solve, it will include an integral, and then solve the equation using Mathematica. Include your work from Mathematica as an image. What does this mean in the context of the concentration of explosive in a soil sample?
- Use each of the four fitted models to calculate the probability that the amount of explosive in a sample will be more than 10 mg/kg? Again, remember to write down the integral you need to calculate and then use Mathematica to compute the integral. Include your work from Mathematica as an image. What does this mean in the context of the concentration of explosive in a soil sample?

- **CHECK YOUR WORK:**
  - Use the Shiny App (https://shiny.byui.edu/connect/#/apps/398b2c19-0d8a-4a9c-8af6-6d8c4b9f6fe0/access) to check your answers.
    - In your analysis include an image(s) from the Shiny App to show the answers from the calculations are correct.
- While it is possible to fit all four probability models, to this data, explain why f_0 and f_1 should not be used as models for the amount of Nitroglycerin in the soil in this situation. How are these models inconsistent with the information we see in the the density histogram of the Nitroglycerin data?
  - Consider calculating $P(X < 0)$ with each of these fitted models. What do the results of these calculations tell you about these models? In what specific way are both f_0 and f_1 stories that are inconsistent with what you know about amount of Nitroglycerin in soil?
- Describe in 4-6 sentences how the information (or answer) you get from the data depends on the general model you assume. Use results from your calculations above to illustrate this idea. Why is this an important concept to understand when working with models and data?
- Organize your work into a **cohesive analysis** (https://byuistats.github.io/M119/specs_detail.html) and submit the html file on Canvas. Your narrative should stand alone apart from the "project instructions" (meaning your reader should not need the instructions for the project to understand what you are doing or explaining) and separate from the individual Tasks (meaning you should not assume your reader has read any of your previous narratives). It is your job in the narrative to lead your reader from the background and question to given data and 4 general models, fitting those models, and answering a question about the data using those fitted models.
- Reflect on your work for this project. At the bottom of your report include the following in a brief (1-2 paragraph) reflection.
  - Identify/explain 2-3 key mathematical ideas you learned (and would like to remember).
  - Identify/explain 1-3 soft skills you needed/improved/learned while working on the project.
    - List of some Soft Skills
      - Dedication
      - Following Directions
      - Motivation
      - Self-directed
      - Organization
      - Planning
      - Time Management
      - Willing to Accept Feedback
      - Perseverance
      - Good attitude
      - Meets deadlines
      - Willingness to learn

---

1. This number is a simplified story for illustrative purposes only. ↩