CS146 Assignment 2

Call center data modeling & other exercises

You should submit your work as a Python notebook, or a Python notebook and PDF both if you want to separate your code and your report. (Please do not submit your Python code as a PDF only. This usually results in lines being truncated.)

Typeset your PDF using Google Docs, LaTeX, Jupyter notebooks, CoCalc, or any other software that allows you to type text and math. Make sure your code is readable and commented.

Show your work for all exercises! Do not simply turn in final answers.

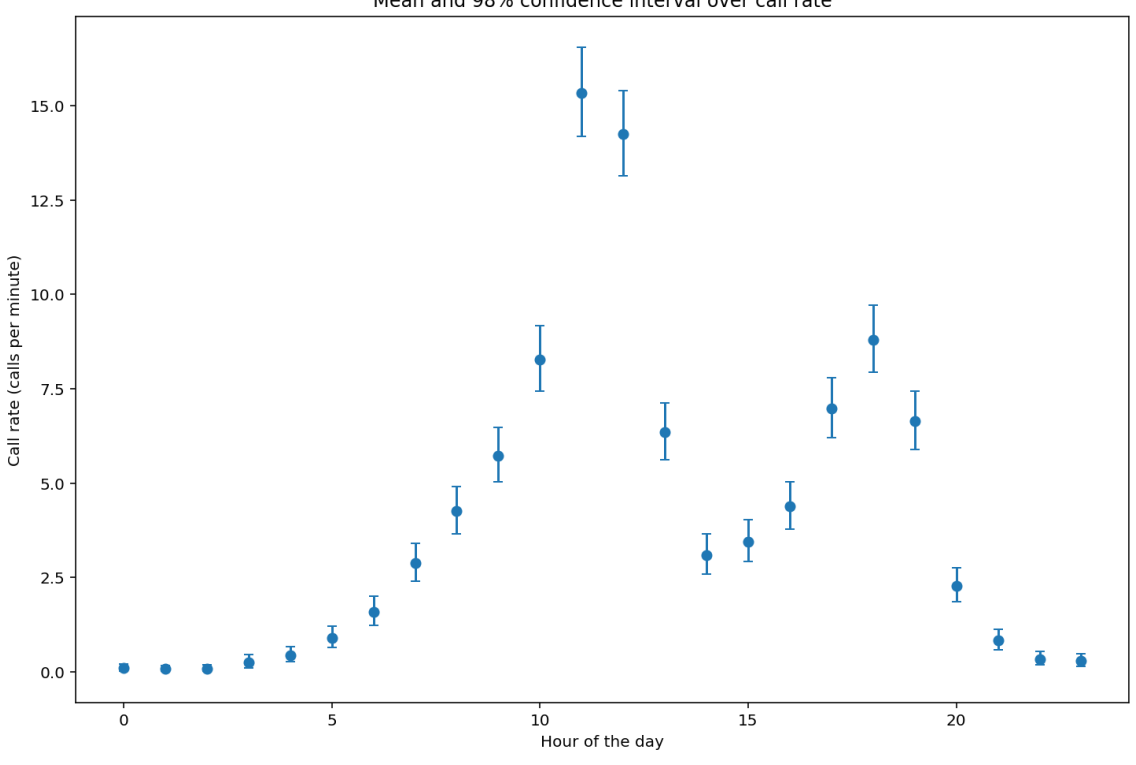
1. Call center data modeling

Complete the call center data modeling assignment we start in the Pre-class work and Activity 2 breakouts of Session 2.2. You may re-use and build on all code or any other work from the class session.

In class, we completed the Bayesian data modeling problem for 1 hour of the day. In this assignment, you need to do the same analysis for all 24 hours of the day.

1. Compute a 98% posterior confidence interval over the number of calls per minute (the call rate λ) for each hour of the day — so you will have 24 confidence intervals. Also, compute the posterior mean of λ for each hour of the day.

2. Present your results graphically using Matplotlib. Make a plot that looks like the one below. Each dot is at the posterior mean and each line shows a 98% confidence interval for a λ. You can use the `errorbar()` function in the plotting library to do this.



3. Write a paragraph (100–200 words) to accompany your plot and present your findings to the client. Carefully summarize how many calls you expect during different parts of the day, and how much uncertainty there is in your estimates. Remember that the client is not an expert in statistics, so make it easy for them to understand. You may also make additional plots to help communicate your results.

2. Stretch goal (optional)

Explain how the `compute_posterior` function (reproduced below) from Lesson 2.2 works. We discussed the function briefly in class. For a complete answer you need to address all the points below. You can also add any other information you think is relevant to the techniques used in the function.

```
 1. def compute_posterior(parameter_values, prior, likelihood, data):
 2.     log_prior = np.log(prior(parameter_values))
 3.     log_likelihood = np.array([
 4.         np.sum(np.log(likelihood(param, data)))
 5.         for param in parameter_values])
 6.     unnormalized_log_posterior = log_prior + log_likelihood
 7.     unnormalized_log_posterior -= max(unnormalized_log_posterior)
 8.     unnormalized_posterior = np.exp(unnormalized_log_posterior)
 9.     area = sp.integrate.trapz(unnormalized_posterior, parameter_values)
10.     posterior = unnormalized_posterior / area
11.     return posterior
```

1. The purpose of the function is to multiply the prior and likelihood passed as input arguments and to return the posterior as output. Explain how the function achieves this purpose using logarithms.
2. What is the purpose of `np.sum()` in line 4?
3. Explain why the maximum of the unnormalized log posterior is subtracted in line 7.
4. Why do we still have to divided by the area in line 10 even after having subtracted the maximum of the unnormalized log posterior in line 7?
5. Create an example where <u>not</u> taking logarithms would cause a problem. Create a prior, likelihood, and data set that fails to produce the correct posterior when we don't take logs. Show all your code and visualize your results on one or more plots.

More practice exercises (optional)

Below are additional practice exercises for you to attempt. These are optional and you can choose to do as many or as few as you want. These exercises will not be graded.

If you get stuck on any of them, contact your instructor with specific questions via email and during office hours. Just saying "I'm stuck" is not enough — explain what you tried and where you got stuck so your instructor can understand your thinking and where you might have missed something or made a mistake.

1. Answer the following questions using Python.

    a. Generate 1000 samples from a normal distribution with mean 100 and standard deviation 10. How many of the numbers are at least 2 standard deviations away from the mean? How many to you expect to be at least 2 standard deviations away from the mean?

    b. Toss a fair coin 50 times. How many heads do you have? How many heads to you expect to have?

    c. Roll a 6-sided die 1000 times. How many 6s did you get? How many 6s do you expect to get?

    d. How much area (probability) is to the right of 1.5 for a normal distribution with mean 0 and standard deviation 2?

2. Let y be the number of 6s in 1000 rolls of a fair die.

    a. Draw a sketch of the approximate distribution of y, based on the normal approximation.

    b. Using the normal distribution function in SciPy, give approximate 5%, 25%, 50%, 75%, and 95% points for the distribution of y.

3. A random sample of n students is drawn from a large population, and their weights are measured. The average weight of the sampled students is $\bar{y} = 75$ kg. Assume the weights in the population are normally distributed with unknown mean μ and known standard deviation 10 kg. Suppose your prior distribution for μ is normal with mean 180 and standard deviation 40.

    a. Give your posterior distribution for μ. (Your answer will be a function of n.)

    b. A new student is sampled at random from the same population and has a weight of y' pounds. Give a posterior predictive distribution for y'. (Your answer will still be a function of n.)

    c. For $n = 10$, give a 95% posterior interval for theta and a 95% posterior predictive interval for y'.

    d. Do the same for $n = 100$.

4. Perfectly and partially observed data in the exponential model.

    a. Suppose $y \mid \lambda$ is exponentially distributed with rate λ, and the prior distribution of λ is Gamma(α, β). Suppose we observe that $y \geq 100$, but do not observe the exact value of y. What is the posterior distribution, $p(\lambda \mid y \geq 100)$, as a function of α and β? Write down the posterior mean and variance of λ.

    b. In the above problem, suppose that we are now told that y is exactly 100. Now, what are the posterior mean and variance of λ?

    c. Explain why the posterior variance of λ is higher in part (b) even though more specific information has been observed.