DATA 5690: Midterm

Tyler J. Brough

2024-03-04

Table of contents

Introduction

This midterm exam covers basic statistical inference from both the frequentist (objectivist) and Bayesian (subjectivist) perspectives. Emphasis is on fundamental understanding and comparative interpretation.

> **Note**
>
> If you get stuck make appropriate assumptions, document them, and proceed.

Frequentist Analysis

1. In this question your task is to carry out statistical inference for a binomial proportion from the frequentist perspective.

The frequentist agent confronts a tootsie roll candy machine with a fixed but unknown probability of dispensing a cherry tootsie roll denoted by θ.

a. Generate artificial data for this scenario with the following code:

```
import numpy as np

np.random.seed(42)

theta = 0.30
tootsie_rolls = np.random.binomial(n=1, p=theta, size=50)
```

The agent is given these data and told they represent draws from the candy machine where an observation of 1 represents a cherry tootsie roll and an observation of 0 represents a vanilla tootsie roll.

b. Compute the maximum likelihood estimator $\hat{\theta}_{MLE}$ as though you were the agent. What is the agent's numerical point estimate of this maximum likelihood estimator?

c. What is the sampling distribution of $\hat{\theta}_{MLE}$ according to the *Central Limit Theorem*? Make a plot of it using `matplotlib.pyplot`.

d. Compute a 95% confidence interval for $\hat{\theta}_{MLE}$. What are the upper and lower bounds? Give a formal interpretation of this confidence interval.

e. Conduct a hypothesis test that the tootsie roll machine is biased towards dispensing vanilla tootsie rolls with a level of significance of 5%.

- State the null hypothesis.
- State the alternative hypothesis.
- Compute the test statistic and report its numerical value.
- Compute the rejection region and report its numerical value.
- Is this a one-tailed or two-tailed test?
- What does the agent conclude? State it formally.

f. Please redo parts a-e for $\theta = 0.45$.

2. In this question your task is to carry out statistical inference for count data from the frequentist perspective. Assume that these data represent visitors that arrive per hour to take a turn at the tootsie roll machine. Let λ be the hourly arrival rate of the visitors.

a. Generate artificial data for this problem with the following code:

```
import numpy as np

np.random.seed(42)

lam = 20
visits = np.random.poisson(lam=lam, size=50)
```

b. The agent doesn't see the data-generating process but assumes that they come from a Poisson distribution with a fixed but unknown λ parameter. Compute the maximum likelihood estimator $\hat{\lambda}_{MLE}$.

c. What is the sampling distribution of $\hat{\lambda}_{MLE}$ according to the *Central Limit Theorem*? Make a plot of it using `matplotlib.pyplot`.

d. Compute a 95% confidence interval for $\hat{\lambda}_{MLE}$. What are the upper and lower bounds? Give a formal interpretation of this confidence interval.

e. Conduct a hypothesis test that the true arrival rate is 18 visitors per hour.

- State the null hypothesis.
- State the alternative hypothesis.
- Compute the test statistic and report its numerical value.
- Compute the rejection region and report its numerical value.
- Is this a one-tailed or two-tailed test?
- What does the agent conclude? State it formally.

3. Use the IID bootstrap procedure to generate an approximate sampling distribution for $\hat{\lambda}_{MLE}$ in the previous problem using the same data that were given to the agent.

a. You can produce a single bootstrap sample with the following code:

```
np.random.seed(42)

x_b = np.random.choice(a=x, size=50, replace=True)
```

Given this bootstrap sample you would then compute a bootstrap replication of the MLE: $\hat{\lambda}^b_{MLE}$.

b. Repeat the above for $b = 1, \ldots, B$ with $B = 10,000$.

c. Reproduce the confidence interval and hypothesis test from question 2 above but using the bootstrap sampling distribution rather than appealing to the CLT.

d. Compare this computational procedure to the classical approach using the CLT.

Bayesian Analysis

4. Reproduce the statistical inference for the data from problem 1 above but from the subjective Bayesian perspective.

- Assume the agent has a prior of $\theta \sim Beta(a = 1, b = 1)$.
- Compute the posterior distribution.
- Make plots of the prior, likelihood and posterior using `matplotlib.pyplot`.
- Calculate the posterior probability that $\theta = 0.5$.
- Compute a 95% equal-tailed credibility interval.
- Using Bayes' factors conduct a hypothesis test for $H_1 : \theta = 0.5$ (i.e. a fair coin) against $H_2 : \theta \neq 0.5$ (i.e. a biased coin). See Clyde et al (2022) Chapter 3 for details on implementing Bayes' factors.
- Interpret the results. Compare the results to the frequentist procedure.

5. Reproduce the statistical inference for the data from problem 2 above but from the subjective Bayesian perspective.

- Assume the agent has the prior: $\lambda \sim Gamma(\alpha, \beta)$, which is the conjugate prior for the Poisson likelihood function.
- Compute the posterior distribution.
- Make plots of the prior, likelihood, and posterior using `matplotlib.pyplot`.
- Compute a 95% equal-tailed credibility interval.
- Using Bayes' factors conduct a hypothesis test for $H_1 : \lambda = 18$ against $H_2 : \lambda \neq 18$. Use a diffuse prior for H_2.

Additional Topics

6. The 18th century mathematician Compte de Buffon conducted one of the first stochastic simulation exercises by paying a child to play the St. Petersburg game $N = 2048$ times. He argued that after 29 rounds of the game that there would not be enough money in all of France to cover the bet. (Verify his calculation). The finite nature of the bookie's cash reserves is a simple and straight forward resolution of the paradox for most people.

But we're not most people and we have cheap computing power at our hands!

- Conduct a Monte Carlo study of the Buffon experiment ($N = 2048$ trials). Set $M = 1,000,000$.
- Describe the pattern you see. What kind of sampling distribution is this?
- Calculate the typical univariate sample statistics.
- Use the *Central Limit Theorem* to describe the results. Does the CLT hold? Why or why not?

NB: You'll want to choose your plotting parameters wisely!

7. Return to problem 4 above in the *Bayesian Analysis* portion of the exam. In this question you will be asked to run a Monte Carlo study to analyze the frequentist properties of the Bayesian "estimator" from problem 4. Take the mean of the posterior distribution for θ (*hint:* the $Beta(a^*, b^*)$ distribution) as your point estimate. Simulate $10,000$ samples each of size $n = 1,000$ from the data-generating process. For each sample apply the Bayesian point estimator (i.e. the posterior mean). Save these and make a histogram plot of them. Provide univariate descriptive statistics (mean, media, standard deviation, range, min, max). Do the same thing for the maximum likelihood estimator and compare. What do you find? What does that mean for frequentist vs Bayesian inference?

NB: Make sure to use an uninformative prior (i.e. $Beta(a = 1, b = 1)$).

8. Whereas your task in problem 7 above was to assess the frequentist properties of the Bayesian point estimator for θ, in this question you are asked to derive an approximate Bayesian predictive density from a maximum likelihood estimator (frequentist). Return to problem 5 above in the *Bayesian analysis* portion of the exam. Compute the maximum likelihood estimator (see also problem 2). Using the MLE point estimate for the original simulated data (problem 2), run a Monte Carlo simulation to generate predictive data from the likelihood function (i.e. Poisson distribution). Re-run the original data-generating simulation but

this time set $n = 1,000$ observations and obtain a new MLE point estimate from these data. Now generate $M = 10,000$ repititions from the likelihood function. This is your approximate (frequentist) predictive distribution for $\tilde{y}$ from $p(\tilde{y}|\widehat{\lambda}_{MLE})$.

Now run a posterior predictive simulation from the fully Bayesian model in problem 5. Make sure to use a very diffuse prior. The posterior predictive distribution is a negative binomial. You can either use that directly or simply take a draw from the posterior (i.e. $\lambda_{draw} \sim Gamma(\alpha^*, \beta^*)$) and then use that to take a draw from the likelihood $p(\tilde{y}|\lambda_{draw})$. Repeat this process $M = 10,000$ times.

Make plots of both predictive distributions and provide univariate summaries of each. Compare thoroughly. What do you find? Is it possible to interpret the frequentist simulation as an *approximate Bayesian* procedure? How so?

References

Clyde et al. 2022. *An Introduction to Bayesian Thinking: A Companion to the Statistics with R Course*. https://statswithr.github.io/book/.