DATA 5690: Midterm Exam

2024-02-28

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:

```python
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 {.unnumbered}

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, Merlise and Çetinkaya-Rundel, Mine and Rundel, Colin and Banks, David and Chai, Christine and Huang, Lizzy (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.