Assignment IV

October 26, 2022

Problem I

Let X be the time (in hours) taken by a person to complete a certain 5K race. Assume that the pdf of X is given by

$$f(x) = \begin{cases} (\theta + 1)x^{\theta} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Note that in the above equations $\theta > -1$. Here are the race times of an individual from 10 different events.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.77	0.47	0.9	0.86	0.65	0.79	0.73	0.97	0.94	0.92

Use the maximum likelihood estimation technique to find (i) an estimator for θ (ii) and an estimate for θ based on the observed data.

Problem II

Complete the Jupyter notebook (`conf_intervals.ipynb`) posted on Canvas with this assignment. The objective of this notebook is to demonstrate how two-sided confidence interval estimates are obtained.

- Complete the function `get_sample_mean_variance`. This function takes a numpy array as an argument. The values stored in the array are values of i.i.d random variables $\{X_i\}$. The function should return the estimates of sample mean $\overline{X}_n$ and sample variance s^2.

  I have provided a function `get_sample(n=20)`, which takes an integer `n` as input and returns a numpy array of size n. This function mimics the process of obtaining a sample of size n. The array returned by the function contains the values of n different i.i.d random variables (sample data $\{X_1, X_2, X_3, \ldots X_n\}$).

Here is a short code snippet which calls the function `get_sample(n=20)` to obtain sample data of size 20. The data obtained can be passed to the function `get_sample_mean_variance`, to get estimates for sample mean $\overline{X}_n$ and sample variance s^2.

```
def get_sample(n=20):
    # Return a sample of size n from population
    rv = expon(scale=5)
    return rv.rvs(size=n)

def get_sample_mean_variance ( data_X ):
    # Complete this function

    return sample_mean, sample_vars

rv_s = get_sample(500)
print(get_sample_mean_variance(rv_s))
```

Here is the output of the code snippet.

```
(5.67740945953341, 28.024301787875093)
```

Note that the result you get will be different since random numbers are used to generate the sample data.

- Complete the function `get_z_limits ( conf_level )`. The function accepts one argument (`conf_level`) and returns a float value z such that $F(z) - F(-z) =$`conf_level`. Here $F(z)$ is the cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$. The value z that is returned is such that area under the standard normal between $-z$ and z equals `conf_level`. You may need to use a root finding method that is available in the `scipy.optimize` library.

  Here is an example which uses the `fsolve` root finding method from `scipy.optimize` to solve a similar problem. It computes the pth quantile of a normal distribution. You can modify this code to complete the function `get_z_limits`.

```
from scipy.optimize import fsolve
from scipy.stats import norm

def norm_cdf_quant_eqn(x, c):
    rv   = norm (0, 1)
    return rv.cdf(x) - c

def get_pth_quantile ( p = 0.5 ):

    p_quant = fsolve(norm_cdf_quant_eqn, x0=0.0, args = (p))
    return p_quant

print(get_pth_quantile(0.9))
```

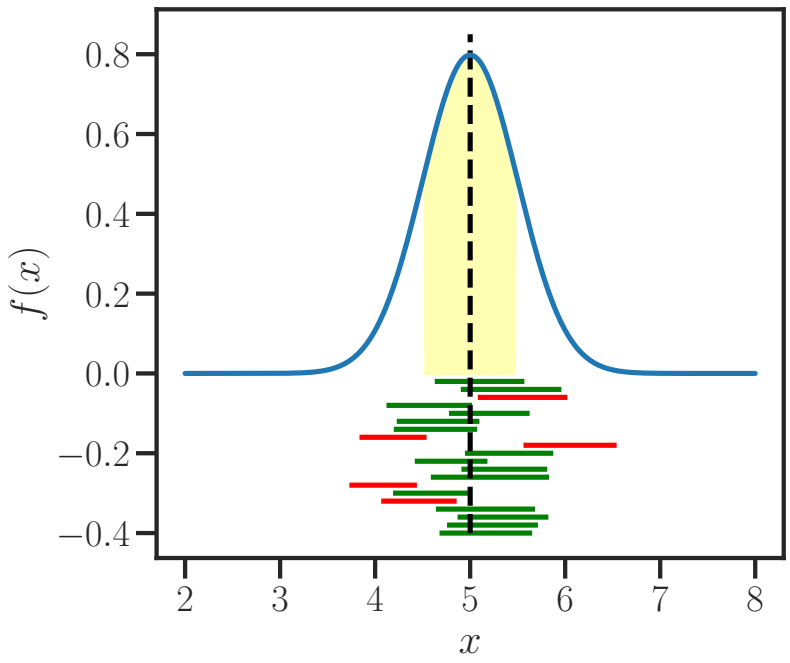

Figure 1: PDF of the $\overline{X}_n$ shown together with confidence intervals of 20 samples of size 100. The area within the yellow shaded region equals 0.65. The green and red bars are confidence intervals. Red bars do not contain the true population mean and the green ones do.

[1.28155157]

- Alternative way to complete `get_z_limits` You may use the `interval` method of the `norm` class which can be imported from `scipy.stats` to obtain this result. The documentation for the `norm` class is available in this link.

- Complete the function `get_confidence_interval`. This function takes a numpy array `data_X` and a float value `conf_level` as arguments. The values stored in the array `data_X` are values of i.i.d. random variables. The `float` corresponds to the confidence level (e.g. 0.95). The function returns the lower and upper bound of the confidence interval estimate estimated from the sample data contained in `data_X`. The code snippet below contains a sample function call.

```
1 import numpy as np
2 from scipy.stats import norm, expon
3 import math
4
```

```
def get_sample(n=20):
    # Return a sample of size n from population
    rv = expon(scale=5)
    return rv.rvs(size=n)

def get_confidence_interval ( data_X, conf_level = 0.95 ):

    # Complete this function
    return l_bound, u_bound

rv_s = get_sample(500)
a,b  = get_confidence_intervals(rv_s, 0.95)
print(a,b)
```

Here is the output of the code snippet.

```
4.526967059211625 5.470951850469932
```

Use `get_sample_mean_variance` and `get_z_limits` as helper functions to complete `get_confidence_interval`.

- The mean and standard deviation of the population, from which you obtained samples in this notebook (`get_sample(n)`) are $\mu = 5.0$ and $\sigma = 5.0$. Given this information and your knowledge of the central limit theorem can you predict (approximately) the sampling distribution of $\overline{X}_n$. Note that in real-world situations this information (μ, σ) is unavailable and you obtain their estimates from sampled data.

- Plot the the approximate pdf of $\overline{X}_n$ (prediction from central limit theorem). Obtain the confidence intervals of 25 different samples at confidence level 65%. Set the sample size of each to be 100. (See code snippet below). Plot the confidence intervals as horizontal lines below the pdf as shown in figure I. Count how many of these intervals do not contain the population mean $\mu = 5$. What do you expect this number to be ? In figure I confidence intervals which do not contain the population mean (dashed black line) are shown in red. Change the confidence level to 80 % and observe how the plot changes. Briefly explain the observations you made from the plots.

```
# Here is code to obtain the confidence intervals of 100 samples
for i in range(25):

    rv_s = get_sample(100)
    a,b  = get_confidence_intervals(rv_s, 0.65)
```

Problem III

A scientist has purchased a micro-balance for his lab. He measures the weight of 100 standardized samples using his new micro-balance. The weight of these samples is accurately known from other independent measurements. He documents the error of each sample measurement. The sample mean error is 12 μ g and standard deviation of the error is 60 μ g. A perfectly calibrated micro-balance should have a mean error of zero. Based on the sample data, the scientist concludes that the device needs re-calibration. Assess the validity of his claim using NHST.

- State the null hypothesis and the alternative hypothesis.
- Compute the p-value.
- Pick $\alpha = 0.05$ as the significance level, and use null hypothesis significance testing to verify the claim of the scientist.
- Does your conclusion change if $\alpha = 0.01$?

Problem IV

A company that sells benzene claims that the purity of its product is greater than 90%. Five independent samples were collected from barrels sold by the company. The purity of the samples were 90.2, 88.7, 89.1, 91.3 and 92.2.

- What assumptions should you make regarding the population in order to obtain a suitable statistic that will enable you to perform hypothesis testing ?
- Use null hypothesis significance testing to check if the claim made by the company is statistically significant. You can pick a significance level of your choice.

Problem V

The tire manufacturer Michelin has developed a new brand of tires that can last at least 50,000 miles. You are designing a significance test to verify the claims of the company. Assume that the sample standard deviation is 10000 miles for the first 4 questions listed below.

- You have decided to test a sample of 100 tires and accept the claim of the company (reject the null hypothesis) if the sample mean (evidence) is at least 52,000 miles. What should the significance level of the test be ?
- Assume that the true lifespan of the tires is 53,000 miles. What is the power of the test at significance level 0.05 ?

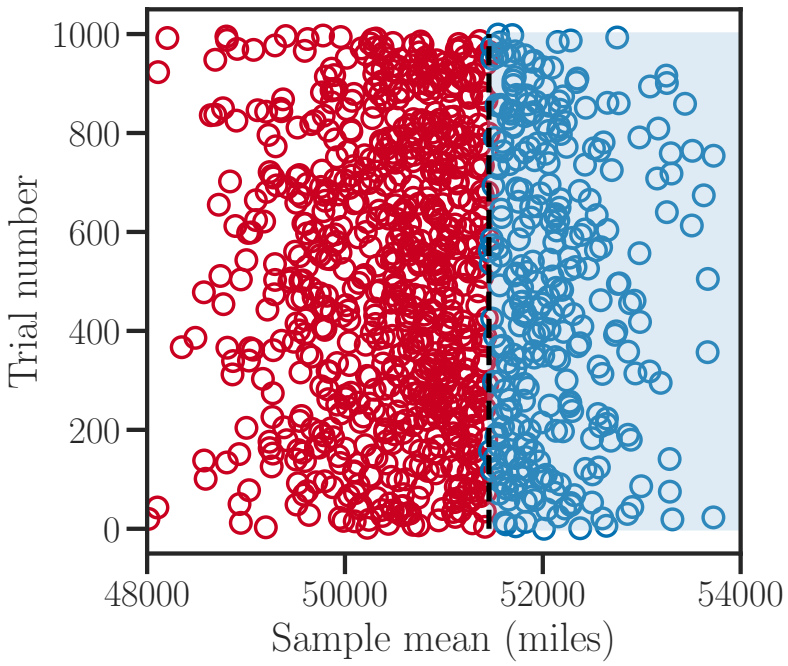

Figure 2: Plot of 1000 different sample means vs trial number. The right of the black dashed line is the rejection region.

- Assume that the true lifespan of the tires is 51,000 miles. What is the power of the test at significance level 0.05 ?
- Assume that the true lifespan of the tires is 51,000 miles. What should the sample size be such that at significance level 0.01 the power of the test is 0.90 ?
- Let's run some simulations of the scenario described in this problem. Here is a code-snippet that simulates the act of testing n different Michelin tires.

```
from scipy.stats import gamma

def measure_tire_lifespan(n = 100, true_pop_mean = 50000):

    sd    = 10000
    beta  = true_pop_mean / sd**2
    alpha = true_pop_mean * beta

    rv = gamma(a = alpha, scale=1/beta)

    # Returns lifetime of n different tire samples
    return rv.rvs(size=n)
```

- The function call `measure_tire_lifespan(n = 100, true_pop_mean = 53000)` returns a `numpy` array of $n = 100$ lifespan measurements assuming that the ground truth or the true population mean lifespan is 53000 miles. From this data compute the sample mean and check if it falls in the rejection region for the null hypothesis that $\mu = 50000$ miles at significance level $\alpha = 0.05$. Repeat this for $m = 1000$ trials and make a plot that looks like figure II. Each circle in this plot is a sample mean measurement and it is shown in blue if it falls in the rejection region and red otherwise. The region to the right of black dashed line is the rejection region. What fraction of sample means fall in the rejection region ? What do you expect this number to be as $m \to \infty$

- Assume that the true lifespan of the tires is 51,000 miles. Use the function call `measure_tire_lifespan(n = 100, true_pop_mean = 51000)` to get a `numpy` array of $n = 100$ lifespan measurements assuming that the ground truth or the true population mean lifetime is 51000 miles. From this data compute the sample mean and check if it falls in the rejection region for the null hypothesis that $\mu = 50000$ miles at significance level $\alpha = 0.05$. Repeat this for $m = 1000$ trials and make a plot that looks like figure II. What fraction of sample means fall in the rejection region ? What do you expect this number to be as $m \to \infty$. Briefly comment on the two plots you have made.

Instructions for submission

- Please submit your answers to the problems as a single PDF file. If your work is hand-written, please make sure to scan them in high resolution.

- Please submit you answer to problem II and V as a Jupyter notebook. Add captions to the plots and write comments to explain your code.

- If you worked as a group to solve these problems please list their names in the PDF file.

- Assignment due by 23:59 on 11/6/22.