PHYS 139 Assignment 2

Yuntong Zhou

October 2022

Problem 1

The null hypothesis gives a probability=0.5. Since this is a binomial distribution, we can approximate the result in a normal distribution.

$$\mu = Np_0 = 0.5 \times 21 = 10.5$$
$$\sigma = \sqrt{Np(1-p)} = \sqrt{0.5 \times 0.5 \times 21} = 2.291$$
$$t = \frac{n-\mu}{\sigma} = 2.400$$
$$p = 1 - P_{normal}(t) = stats.norm.sf(2.400) = 0.0082$$

This p value is smaller than 0.05. Therefore, the experiment is statistically significant to eliminate the null hypothesis that the coin is biased.

Problem 2

(a)

Approximate the binomial distribution to a normal distribution,

$$\mu = Np = 1,000,000,000 \times 0.5 = 500,000,000$$
$$\sigma = \sqrt{Np(1-p)} = \sqrt{1,000,000,000 \times 0.5 \times 0.5} = 15811.4$$

Calculate two t values,

$$t_1 = \frac{500,100,000 - 500,000,000}{\sigma} = 6.32$$
$$t_2 = \frac{500,200,000 - 500,000,000}{\sigma} = 12.65$$
$$p(500,100,000 \leq N \leq 500,200,000) = P_{normal}(t_2) - P_{normal}(t_1)$$
$$= 1.31 \times 10^{-10}$$

The following code is used to perform these,

```
from scipy.stats import norm
import math

N=1e9
p=0.5
sigma=math.sqrt(N*p*p)
t1=100000/sigma
t2=200000/sigma
print (norm.cdf(t2)-norm.cdf(t1))
```

Listing 1: probability calculation by scipy.stats.norm

Using the binom module in scipy.stats, we obtain,

$$binom.cdf(500200000, N, p) - binom.cdf(500100000, N, p) = 1.27 \times 10^{-10}$$

Just in case problem a is asking for code, there's a problem2.py file in Assignment2 folder. Don't try to run it. It takes ages to finish.

(b)

I accidentally did the comparison in part a.
The estimates are pretty close to the exact binomial distribution. The difference is around 3 percent.

Problem 3

(a)

To demonstrate the convergence to classical central limit theorem, simulations are performed with increasing number of dices. For each trial, the average of the dices is computed. The distribution of the averages is plotted. See problem3.py for detailed explanations.
The following graph is plotted,

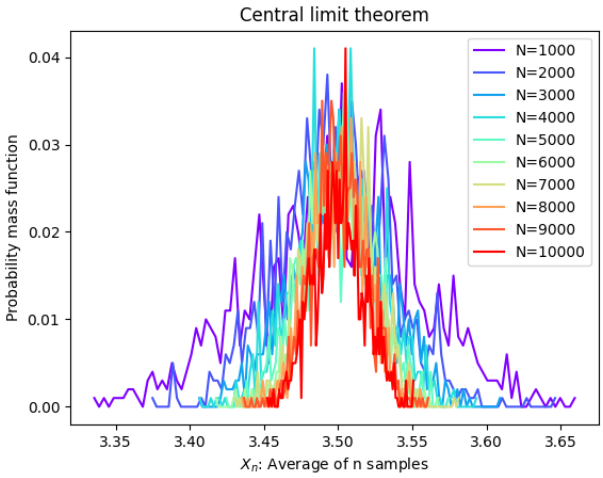

Figure 1: The approximated probability mass distribution of X_N as N (number of dices) increases

By uncommenting several lines in the file and commenting the lines for probability mass function approximation, a probabibility density function can be obtained below,

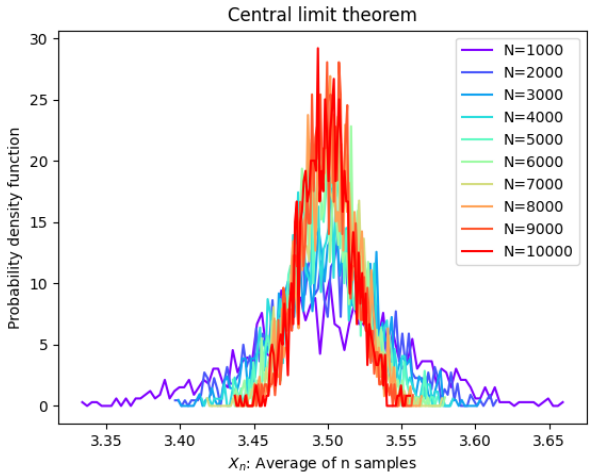

Figure 2: Probability density function

(b)

The expected value and variance of the dice is,

$$E(X) = \frac{1}{6} \times (1+2+3+4+5+6) = 3.5$$
$$Var(X) = E(X^2) - E(X)^2 = \frac{1}{6} \times (1+4+9+16+25+36) - 3.5^2 = \frac{35}{12}$$

Based on the classical central limit theorem, the average of n dices, $\frac{\sum_i^n X_i}{n}$ converges to $Normal(3.5, \sqrt{\frac{\sigma}{n}})$.

Problem 4

(a)

$$P(x|data) \propto x^{N_B}(1-x)^{N-N_B} \times P(x|I)$$

Since $P(x|I)$ is a constant,

$$P(x|data) = A \cdot x^{N_B}(1-x)^{N-N_B} = Ax^{13}(1-x)^{24}$$

Normalization gives the probability density function,

$$\int_X Ax^{13}(1-x)^{24}\ dx = 1$$
$$A = \frac{1}{\int_0^1 x^{13}(1-x)^{24}\ dx}$$

We know that,

$$\int_0^1 p^k(1-p)^{n-k}\mathrm{d}p = \frac{k!(n-k)!}{(n+1)!}$$
$$A = \frac{38!}{13!24!}$$

Therefore,

$$P(x|data) = \frac{38!}{13!24!}x^{13}(1-x)^{24}$$

(b)

problem4.py plots the normalized probability density function. The result is shown below,

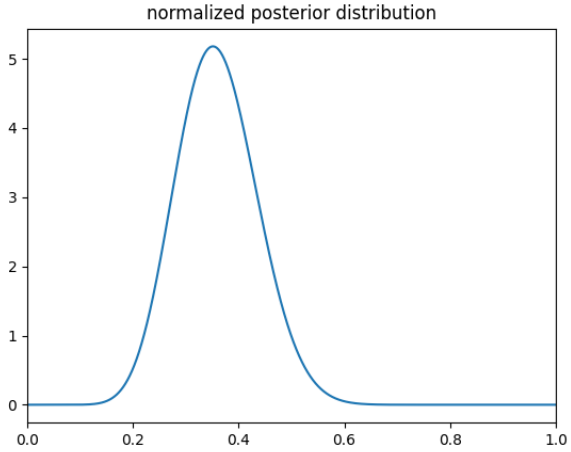

Figure 3: The normalized probability density function for the posterior distribution

Note that in this case, a probability mass function can also be plotted with the approximation of the posterior distribution as a discrete dataset (which it is in the case of the simulation).

c

Disregarding the normalization constant, suppose $\frac{N_B}{N} = c < 1$,

$$P(x|data) = x^{cN}(1-x)^{(1-c)N} = [x^c(1-x)^{1-c}]^N$$

At the limit of $N \to \infty$, we would expect this function to become a delta function. We can check by taking the first derivative of the term inside the power function,

$$\frac{dP}{dx} = cx^{c-1}(1-x)^{1-c} + x^c(1-c)(1-x)^{-c} = 0$$

The power function does not change the local maximum, therefore, multiplying both sides with $x^{-c} \cdot (1-x)^c$

$$\frac{c}{x}(1-x) + 1 - c = 0$$

$$\frac{c}{x} = 2c - 1$$

$$x = \frac{c}{2c-1} = \frac{N_B}{2N_B - N} = \frac{N_B}{N_B - N_C}$$

Problem 5

(a). Analysis

$$I_4 = M_4 - 3M_2^2 =< (x_i - \bar{x})^4 > -3 \cdot [< (x_i - \bar{x})^2 >]^2$$
$$= \int (x - \bar{x})^4 \cdot p(x)\ dx - 3 \cdot [\int (x - \bar{x})^2 \cdot p(x)\ dx]^2$$

Suppose S=X+Y,

$$M_2(S) = M_2(X) + M_2(Y)$$
$$I_4(S) = M_4(S) - 3M_2(S)^2 = M_4(S) - 3(M_2(X) + M_2(Y))^2$$
$$= M_4(S) - 3(M_2^2(X) + M_2^2(Y) + 2M_2(X)M_2(Y))$$

The goal now is to prove,

$$M_4(X) + M_4(Y) = M_4(S) - 6M_2(X)M_2(Y)$$

Expand the $M_4(S)$ term in integral form,

$$M_4(S) = \int_X \int_Y (x + y - \overline{x + y})^4 p_{XY}(x, y)\ dx\ dy$$

Expand the term $(x + y - \overline{x + y})^4$ in terms of $[(x - \bar{x}) + (y - \bar{y})]^4$,

$$M_4(S) = \int \int [(x - \bar{x})^4 + (y - \bar{y})^4 + 6(x - \bar{x})^2(y - \bar{y})^2 \quad (1)$$
$$+4(x - \bar{x})^3(y - \bar{y}) + 4(x - \bar{x})(y - \bar{y})^3]p(x, y)\ dx\ dy \quad (2)$$

Since the X and Y are independent variables, we can say,

$$\int \int f(x)f(y)\ dx\ dy = \int f(x)\ dx \cdot \int f(y)\ dy$$
$$p_{XY}(x, y) = p(x) \cdot p(y)$$

Using this principle, we find that the first three terms in the integral correspond to $M_4(X), M_4(Y), 6M_2(X)M_2(Y)$, respectively. Moreover, since,

$$\int_X (x - \bar{x})p(x)\ dx = 0$$
$$\int_Y (y - \bar{y})p(y)\ dx = 0$$

The two terms drop out.
Therefore,

$$I_4(S) = I_4(X + Y) = I_4(X) + I_4(Y)$$

Thus the additive nature.

(b). Simulation

problem5.py shows how terrible the numpy distribution random generator is. The function to calculate I4 is included in problem5.py, but the result is proved in problem5b.py. A sample output is showed below.

```
yuntongzhou@Yuntongs-MacBook-Air Assignment2 % python3 problem5b.py
Semi-variant I4 for X is  -19.984089951950295
Semi-variant I4 for Y is  -54.88686323587011
The sum of semi-variants I4x and I4y is  -74.8709531878204
The semi-variant I4 for (X+Y) is  -76.7297844736953
```

Figure 4: Sample output from problem5b.py that demonstrates the additive nature of I4.

Problem 6

(a)

Use problem6.py to count the occurrences, we get,
There are 474786 A bases.
There are 289915 G bases.
There are 472303 T bases.
There are 288180 C bases.
There are in total 1525184 bases.

(b)

Model 1: All bases occur with an equal probability of 0.25.

$$p_0 = 0.25$$
$$\mu = 0.25 \times N = 381296$$
$$\sigma = \sqrt{N(1-p)p} = 534.76$$
$$t_A = \frac{N_A - \mu}{\sigma} = 174.8$$
$$t_T = \frac{N_T - \mu}{\sigma} = 170.18$$
$$t_C = \frac{N_C - \mu}{\sigma} = 174.13$$
$$t_G = \frac{N_G - \mu}{\sigma} = 170.88$$

The p-values are so small that they are approximately 0.
Model 2: A, T occur with equal probability and C,G occur with equal proba-

bility.

$$p_{AT} = \frac{N_A + N_T}{2N} = 0.31$$
$$p_{CG} = \frac{N_C + N_G}{2N} = 0.1895$$
$$t_{AT} = \frac{n_A - n_T}{\sqrt{2Np_{AT}(1-p_{AT})}} = \frac{474786 - 472303}{\sqrt{2 \times 1525184 \times 0.31 \times 0.69}} = 3.074$$
$$t_{CG} = \frac{n_C - n_G}{\sqrt{2Np_{CG}(1-p_{CG})}} = \frac{289915 - 288180}{\sqrt{2 \times 1525184 \times 0.31 \times 0.69}} = 2.148$$

The p-values are determined as,

$$t_{CG} = 0.032$$
$$t_{AT} = 0.002$$

Taking the alpha value as 0.05, both p-values are smaller than alpha value. Therefore, the null hypothesis is failing.