Suvo_Ganguli_Assignment_3.1

January 23, 2024

1 Assignment 3.1

Name: Suvo Ganguli

Date: Jan 23, 2024

For this assignment, you will refer to the textbook to solve the practice exercises. **Use Python to answer any coding problems (not R, even if indicated in your textbook).** Use Jupyter Notebook, Google Colab, or a similar software program to complete your assignment. Submit your answers as a **PDF or HTML** file. As a best practice, always label your axes and provide titles for any graphs generated on this assignment. Round all quantitative answers to 2 decimal places.

1.1 Problem 3.2.

In an exit poll of 1648 voters in the 2020 Senatorial election in Arizona, 51.5% said they voted for Mark Kelly and 48.5% said they voted for Martha McSally

a) Suppose that actually 50% of the population voted for Kelly. If this exit poll had the properties of a simple random sample, find the standard error of the sample proportion voting for him.

b) Under the 50% presumption, are the results of the exit poll surprising? Why? Would you be willing to predict the election outcome? Explain by (i) conducting a simulation; (ii) using the value found in (a) for the standard error.

a) The standard error of a sample proportion is given by:

SE = sqrt(p(1-p)/n)

where p is the proportion of interest, and n is the sample size. In this case, p = 0.5 and n = 1648.

SE = sqrt(0.5(1-0.5)/1648) = 0.013 or 1.3%

(b)

(c) Based on the simulation of 1648 voters (see code below), mean = 49.4 and SE = 0.72. Thus 99.7% of the data will be within [mean +/- 3*SE] = [46.5, 50.8]

The actual vote is 50% and is within the bounds found by simulation.

(ii) Prediction using standard error: 99.7% of the data will be within: [p +/- 3*SE] = [0.5 +/- 3*0.013*50] = [48.85, 51.95]

We observe that both Mark Kelley's and Martha McSally's exit poll estimates fall within the 99.7% bounds.

[152]:
```
# Include library
import numpy as np
from numpy import random

# Calculate random samples
n = 1648 # number of samples

np.random.seed(10)
x = random.choice(range(0,100), size=n)

mean_sample = np.mean(x)
std_sample = np.std(x)/np.sqrt(n)

print('Mean of sample = ' + str(mean))
print('StdDev of sample = ' + str(std_sample))
print('Mean - 3 * StdDev of sample = ' + str((mean_sample - 3*std_sample).
 ↪round(2)))
print('Mean + 3 * StdDev of sample = ' + str((mean_sample + 3*std_sample).
 ↪round(2)))
```

```
Mean of sample = 49.406
StdDev of sample = 0.7150331707228658
Mean - 3 * StdDev of sample = 46.51
Mean + 3 * StdDev of sample = 50.8
```

1.2 Problem 3.3.

The 49 students in a class at the University of Florida made blinded evaluations of pairs of cola drinks. For the 49 comparisons of Coke and Pepsi, Coke was preferred 29 times. In the population that this sample represents, is this strong evidence that a majority prefers Coke? Use a simulation of a sampling distribution to answer.

Let E be the event that the students like Coke

Total number of students, n = 49 Success number for code, np = 29 Hence, success probability, p = 0.592 and, failure probability, q = 0.408

Estimate of mean of true probability of success, p0 = 0.5 Estimate of mean of true probability of failure, q0 = 0.5 Estimate of standard deviation of true probability of success, s0 = sqrt(p0*q0/n) = 0.071

Z score = (0.592 - 0.071) = 0.521

Let us consider 95% confidence. Hence alpha = 0.05, and alpha/2 = 0.025 We can create the following null hypothesis:

H0: students like Coke if z < 1.96 or > 1.96

From the z-score table, z = p-po/s0 z = (0.592-0.5)/0.071 = 1.296

Since, z does not fall outside the bound of 1.96, we cannot reject the null hypothesis.

Thus, there is not a strong evidence that the majority of the students like Coke.

1.3 Problem 3.5.

The example in Section 3.1.4 simulated sampling distributions of the sample mean to determine how precise $\bar{Y}$ for $n = 25$ may estimate a population mean μ.

a) Find the theoretical standard error of $\bar{Y}$ for the scenario values of $\sigma = 5$ and 8. How do they compare to the standard deviations of the 100,000 sample means in the simulations?

For the first scenario with = 5 and n = 25:

SE = 5 / √25 = 5 / 5 = 1.

For the second scenario with = 8 and n = 25:

SE = 8 / √25 = 8 / 5 = 1.6

In the book, the mean and standard deviation are given as 19.9981 and 1.003612 for (mu, sigma) of (20,5), and the mean and standard deviation are given as 23.9949 and 1.603475 for (mu, sigma) of (24,8).

Thus, the values of mean and standard deviation from the 100,000 simulations are very close to the actual mean and standard deviations.

b) In the first scenario, we chose $\sigma = 5$ under the belief that if $\mu = 20$, about 2/3 of the sample values would fall between \$15 and \$25. For the gamma distribution with $(\mu, \sigma) = (20, 5)$, show that the actual probability between 15 and 25 is 0.688.

For the gamma distribution in Python, with parameters alpha and lambda

np.random.gamma(shape = alpha, scale = lambda, size = 1000)

Let beta = 1/lambda

Therefore, mu = alpha * beta, sigma^2 = alpha*beta^2

Hence, beta = sigma^2/mu = 5^2/25 = 1 alpha = mu/beta = 20/1 = 20

[150]:
```
import numpy as np
import scipy.stats as stats
import math
from scipy.special import gamma
import matplotlib.pyplot as plt

mu = 20.0
sigma = 5.0

beta = sigma**2/mu
alpha = mu/beta
lmbda = 1/beta

dx = 0.01
x = np.arange(0,60,dx)
```

```
# Calculate the gamma distribution PDF values
pdf = lmbda**alpha * x**(alpha-1) * np.exp(-lmbda*x) / gamma(alpha)

# Calculate and plot the gamma distribution CDF values
cdf = np.cumsum(pdf)*dx
# plt.plot(x,pdf)
# plt.plot(x,cdf)

for i in np.arange(len(x)):
    if x[i] <= 15:
        x1 = x[i]
        p1 = cdf[i]
    if x[i] <= 25:
        p2 = cdf[i]
        x2 = x[i]

print("Area between 15 and 25 = " + str(np.round(p2-p1,3)))
```

```
Area between 15 and 25 = 0.688
```

1.4 Problem 3.8.

Construct the sampling distribution of the sample proportion of heads, for flipping a balanced coin (a) once; (b) twice; (c) three times; (d) four times. Describe how the shape changes as the number of flips n increases. What would happen if n kept growing? Why?

(a) For one flip, there are 2 options: H and T.
(b) For two flips, there are 4 options: HH, HT, TH, TT.
(c) For three flips, there are 8 options: HHH, HHT, HTH, THH, THT, TTH, HTT, TTT.
(d) For four flips, there are 16 options: HHHH, HHHT, HHTH, HTHH, HTTH, HHTT, HTTT, THHH, THHHT, THTH, TTHH, TTTH, THTT, TTTT.

Thus the probability of heads are 1-1/2^0, 1-1/2^1, 1-1/2^2, 1/2^3, … 1-1/2^(inf).

As n increases, the probability of a single head tends to 1.

Thus, the probability exponentially decays.

1.5 Problem 3.13.

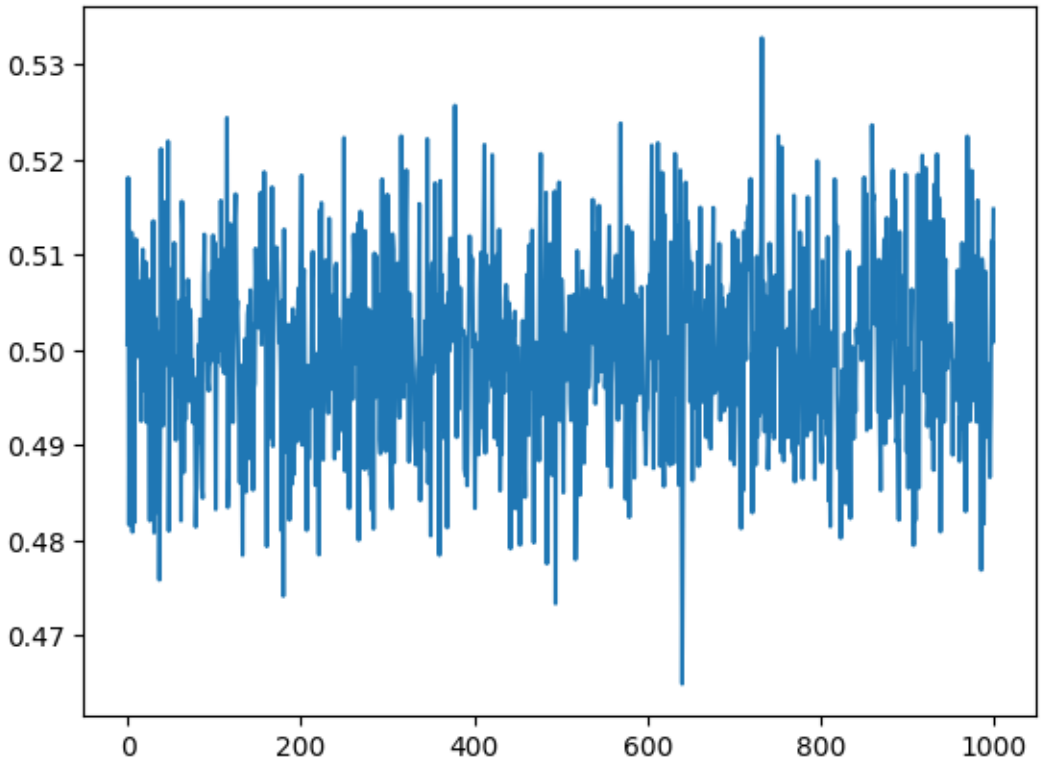
Simulate random sampling from a uniform population distribution with several n values to illustrate the Central Limit Theorem.

We construct a vector of uniform random numbers between 0 and 1. The mean is 0.49995, which is close to 0.5. This shows the Central Limit Theorem.

[149]:
```
n = 1000
np.random.seed(1)

mu_uni = np.zeros(n)
```

```
for i in np.arange(0,n):
    vec = np.random.uniform(low=0.0, high=1.0, size=n)
    mu_uni[i] = np.mean(vec)

plt.plot(np.arange(0,n),mu_uni)

mu = np.mean(mu_uni)
print(mu)
```

```
0.49994825412866567
```



1.6 Problem 3.14.

On each bet in a sequence of bets, you win 1 dollar with probability 0.50 and lose 1 dollar (i.e., win negative 1 dollar) with probability 0.50. Let Y denote the total of your winnings and losings after 100 bets. Giving your reasoning, state the approximate distribution of Y.

This is a case of binomial distribution. This is because the number of observations or trials is fixed, and each observation or trial is independent.

If we want to calculate x number of winnings, them the probsbility of winning is given by 100Cx * 0.5^x * (1-0.5)^(100-x) = 100Cx * 0.5^x * 0.5^(100-x).

The mean is 100*0.5 = 50, and the variance is 100*0.5*(1-0.5) = 100*0.5^2 = 100*0.25 = 25.

1.7 Problem 3.15.

According to a General Social Survey, in the United States the population distribution of Y = number of good friends (not including family members) has a mean of about 5.5 and a standard deviation of about 3.9.

a) Is it plausible that this population distribution is normal? Explain.

Since we have no information about the population distribution so we cannot assume that population distribution is normally distributed.

b) If a new survey takes a simple random sample of 1000 people, describe the sampling distribution of $\bar{Y}$ by giving its shape and approximate mean and standard error.

The sampling distribution of sample mean will be normall distributed with mean

mu = 5.5

and standard error

SE = 3.9 / sqrt(1000) = 0.123

c) Suppose that actually the mean of 5.5 and standard deviation of 3.9 are not population values but are based on a sample of 1000 people. Treating results as a simple random sample, give an interval of values within which you can be very sure that the population mean falls. Explain your reasoning.

The required interval is interval corresponding to 3 standard deviations about the mean for 99.7% confidence. So required interval is mu +/- 3*SE = (5.131,5.869)

1.8 Problem 3.18.

Sunshine City, which attracts primarily retired people, has 90,000 residents with a mean age of 72 years and a standard deviation of 12 years. The age distribution is skewed to the left. A random sample of 100 residents of Sunshine City has $\bar{y} = 70$ and $s = 11$.

a) Describe the center and spread of the (i) population distribution, (ii) sample data distribution. What shape does the sample data distribution probably have? Why?

(i) Population distribution:

- Center: The mean age of the population is 72 years.
- Spread: The standard deviation of the population is 12 years.
- Shape: The population distribution is skewed to the left.

(ii) Sample data distribution:

- Center: The sample mean is 70 years.
- Spread: The sample standard deviation (s) is 11 years.
- Shape: The sample data distribution probably also has a left-skewed shape.

b) Find the center and spread of the sampling distribution of $\bar{Y}$ for $n = 100$. What shape does it have and what does it describe?

For the sampling distribution of with n = 100: - Center: The mean of the sampling distribution of is still 72 years. This is because the sampling distribution of the sample mean is centered around

the population mean. - Spread: The standard deviation of the sampling distribution of is the population standard deviation divided by the square root of the sample size. In this case, sigma = 100/sqrt(12) = 1.2 years. - Shape: The sampling distribution of Y is approximately normally distributed. This is due to the Central Limit Theorem, which states that for large enough sample sizes, the sampling distribution of the sample mean will be approximately normal, regardless of the shape of the population distribution.

c) Explain why it would not be unusual to sample a person of age 60 in Sunshine City, but it would be highly unusual for the sample mean to be 60, for a random sample of 100 residents.

This is because of sampling variability. The sample mean can fluctuate around the population mean due to random sampling. The standard deviation of the sampling distribution is 1.2. Hence with 99.7% confidence, the interval is (64.8,75.6)

d) Describe the sampling distribution of $\bar{Y}$: (i) for a random sample of size $n = 1$; (ii) if you sample all 90,000 residents.

(i) The mean and standard deviation will be different from the population mean and standard deviation of 72 for n = 1. This is shown by 70 and 11 being different from 72 and 12. Although 11 is close to 12, for another sample, the standard deviation of the sample can be quite different from the population standard deviation of 12.

(ii) If we sample 100 times from 90,000 residents, according to the Central Limit Theorem, the mean of the sampling distribution will be close to 72. The standard deviation will be 12/sqrt(100) = 1.2

1.9 Problem 3.21.

- In your school, suppose that GPA has an approximate normal distribution with $\mu = 3.0, \sigma = 0.40$. Not knowing μ, you randomly sample $n = 25$ students to estimate it. Using simulation for this application, illustrate the difference between a sample data distribution and the sampling distribution of Y .

A sample distribution is an observed distribution of the values that a statistic is observed to have for a sample of individuals. For example, the mean can be mu_sample and the standard deviation can be sigma_sample. These will most probably be different from the values for normal distribution (mu = 3.0, sigma = 0.40)

Samping distribution is for a number of simulations. We will get the sample distribution of the sample means and standard deviation. According to the Central Limit Theorem, the mean will be close to 3.0 and the standard distribution will be 0.40/sqrt(25) = 0.08

- When sample data were used to rank states by brain cancer rates, Ellenberg (2014) noted that the highest ranking state (South Dakota) and the nearly lowest ranking state (North Dakota) had relatively small sample sizes. Also, when schools in North Carolina were ranked by their average improvement in test scores, the best and the worst schools were very small schools. Explain how these results could merely reflect how the variability of sample means and proportions depends on the sample size.

The variability for sample size of n is given by sigma_population / sqrt(n). Since n is small for North and South Dakota compared to the other state (n_North_South_Dakota < n_Other_States),

sigma_population_North/South Dakota = sigma_population / sqrt(n_North/South_Dakota) < sigma_population / sqrt(n_Other_States))