Dakota Wilson

DSC-510 T1DQ1

1/27/2024

I think an excellent start is to first give a definition of the basic statistical concepts mentioned: population, sample, parameter, and statistic.

population - defined as "the entire collection of things or subjects about which information is obtained for our analysis" (Rogel-Salazar, 2023, p. 142).

sample - a subset intended to represent the population (Rogel-Salazar, 2023).

parameter - a number that describes the population (OpenAI, 2023).

statistic - a number that describes the sample (OpenAI, 2023).

As long as our sample is a good representation of the population, there is a lot to learn about a population given a good size and statistical analysis.

An excellent example of the use of these statistics can be seen in epidemiology. Python can be used to help determine infection rates in a given city.

In [20]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [21]:

```
# generate synthetic data of 5000 samples with sex, age, and infection status
total_samples = 5000 # represents sample size
sex = np.random.choice(['Male', 'Female'], size = total_samples) # randomize sex
ages = np.random.randint(0, 100, size = total_samples) # randomize age
infected_status = np.random.choice([0, 1], size = total_samples) # randomize infectior
```

A DataFrame is created using the synthetic data.

In [22]:

```
sample_data = pd.DataFrame({'Sex': sex, 'Age': ages, 'Infected': infected_status})
```

The synthetic data can be seen in the table below.

In [23]:

```
# print synthetic data

print(sample_data)
```

```
         Sex  Age  Infected
0     Female   74         0
1       Male   39         0
2     Female   31         1
3     Female   62         0
4     Female    6         0
...      ...  ...       ...
4995  Female   16         0
4996    Male   82         1
4997  Female   32         0
4998  Female   33         0
4999    Male   74         0

[5000 rows x 3 columns]
```

Next we can determine the infection rate in the city.

In [24]:

```python
# calculate total number of infected individuals
total_infected = sample_data['Infected'].sum()
print("Total infected: ", total_infected)

# calculate infection rate
infection_rate_sample_data = (total_infected / total_samples) * 100
print("The COVID infection rate for the synthetic sample data is: {:.2f}".format(infec
```

```
Total infected:  2435
The COVID infection rate for the synthetic sample data is: 48.70 %
```

We can futher analyze the data and determine the infection rate for each gender using the .groupby() method.

In [25]:

```python
# create series with sex and mean infection rate
gender_infection_rate = sample_data.groupby('Sex')["Infected"].mean() * 100
```

Finally, the results of some of our infection rate by gender analysis can be printed.

In [26]:

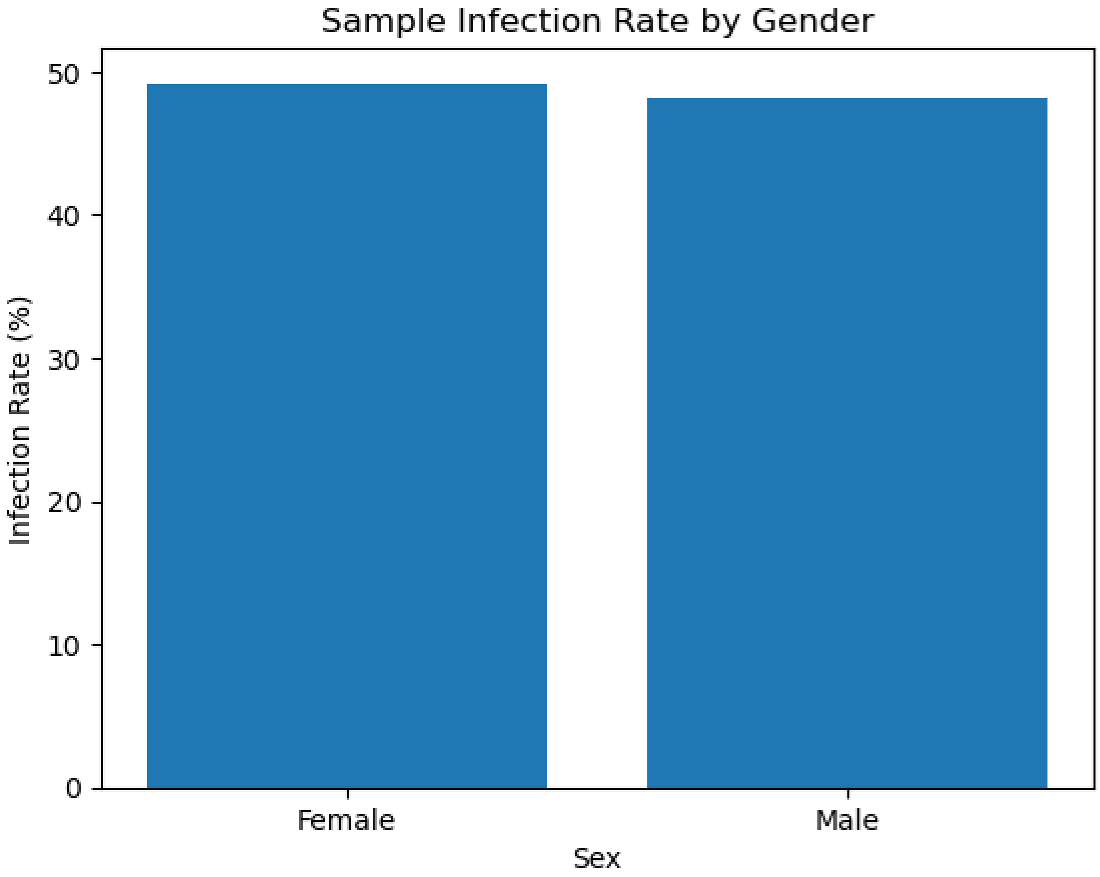
```python
print("Infection Rate by Sex: \n")

# print sex and infection rate
for sex, rate in gender_infection_rate.items():
    print("{}: {:.2f}%".format(sex, rate))
```

```
Infection Rate by Sex:

Female: 49.19%
Male: 48.19%
```

matplotlib.pyplot can be used to visualize this data as a bar graph.

In [27]:

```python
# extract genders and infection rates
genders = gender_infection_rate.index.tolist()
infection_rates = gender_infection_rate.values

# plot bar graph using genders and infection rates
plt.bar(genders, infection_rates)
plt.xlabel('Sex')
plt.ylabel('Infection Rate (%)')
plt.title('Sample Infection Rate by Gender')
plt.show()
```



References:

OpenAI. (2023). ChatGPT. [Large language model]. https://chat.openai.com/chat

Rogel-Salazar, J. (2023). Statistics and data visualisation with python (1st ed.). CRC Press.