Homework 7

1. Textbook 6.10

   Here is my code for this section:

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

url =
'https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/a
balone.data'
column_names = ['Sex', 'Length', 'Diameter', 'Height', 'Whole
weight', 'Shucked weight', 'Viscera weight', 'Shell weight',
'Rings']
abalone_data = pd.read_csv(url, names=column_names)

length_data = abalone_data['Length']
population_median = np.median(length_data)
print(f"Population Median: {population_median}")

def bootstrap_confidence_interval(data, n_samples, sample_size,
confidence_level=0.90):
   boot_medians = []
   for _ in range(n_samples):
       sample = np.random.choice(data, size=sample_size,
replace=True)
       boot_medians.append(np.median(sample))

   lower_percentile = (1 - confidence_level) / 2 * 100
   upper_percentile = (1 + confidence_level) / 2 * 100
   lower_bound = np.percentile(boot_medians, lower_percentile)
   upper_bound = np.percentile(boot_medians, upper_percentile)

   return boot_medians, (lower_bound, upper_bound)

n_samples_a = 2000
sample_size_a = 100
```

```python
boot_medians_a, ci_a = bootstrap_confidence_interval(length_data,
n_samples_a, sample_size_a)
print(f"90% CI for part (a): {ci_a}")

inside_ci_a = np.sum((ci_a[0] <= population_median) &
(population_median <= ci_a[1])) / n_samples_a
print(f"Fraction for part (a): {inside_ci_a}")

plt.hist(boot_medians_a, bins=30, edgecolor='k')
plt.title('Bootstrap Sample Medians - Part (a) (100 Records)')
plt.xlabel('Median')
plt.ylabel('Frequency')
plt.show()

n_samples_c = 2000
sample_size_c = 10
boot_medians_c, ci_c = bootstrap_confidence_interval(length_data,
n_samples_c, sample_size_c)
print(f"90% CI for part (c): {ci_c}")

inside_ci_c = np.sum((ci_c[0] <= population_median) &
(population_median <= ci_c[1])) / n_samples_c
print(f"Fraction for part (c): {inside_ci_c}")

plt.hist(boot_medians_c, bins=30, edgecolor='k')
plt.title('Bootstrap Sample Medians - Part (c) (10 Records)')
plt.xlabel('Median')
plt.ylabel('Frequency')
plt.show()
```

The output of that code is:

Population Median: 0.545

90% CI for part (a): (0.515, 0.5674999999999999)

Fraction for part (a): 0.0005

90% CI for part (c): (0.457375, 0.61)

Fraction for part (c): 0.0005c

2. Textbook problem 7.1

   First, let's calculate the standard error of the Mean.

   $$SEM = \frac{10}{\sqrt{50}} \approx 1.41$$

   Then lets calculate the Z score:

   $$Z = \frac{200-169}{1.41} \approx 21.99$$

   With a Z score so high, the chances of the average being greater than 200cm is basically 0.

3. Textbook problem 7.2
    a. A

First, let’s calculate the standard error of the Mean.

$$SEM = \frac{0.7}{\sqrt{30}} \approx 0.128$$

Then let’s calculate the Z score:

$$Z = \frac{4-5}{0.128} \approx -\ 7.81$$

The chances of the mean being less than 4k are very close to 0.

$$2.55 \times 10^{-15}$$

    b. b

First, let’s calculate the standard error of the Mean.

$$SEM = \frac{0.7}{\sqrt{300}} \approx 0.04$$

Then let’s calculate the Z score:

$$Z = \frac{4-5}{0.04} \approx -\ 25$$

$$1.82 \times 10^{-135}$$

    c. c

    The reason the chances are smaller for larger sample sizes is that in a smaller sample size, it is more realistic to find a couple of outliers from the usual average and make an unusually high mean. When the sample size is larger the chances of that happening diminish.

4. Textbook problem 7.4

The p value in terms of the integral is equal to

$$2 \times \int_{30}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Evaluating this using wolfram I get 2.06×10^{-198}

Since the p value is extremely close to 0 it is very unlikely that the mean of the Parktown Prawns is 10cm. The true mean is probably closer to 7cm.

5. Textbook problem 7.8

   The expected number of boys (E[X]) for this problem is:

   $E[X] = 2009 \times 0.5 = 1004.5$

   The standard error is:

   $SE = \sqrt{2009 \times 0.5 \times 0.5} \approx 22.42$

   The Z score is:

   $\frac{983-1004.3}{22.42} \approx -0.96$

   Now lets look at his hypothesis. The P value integral is:

   $\int_{0.96}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$

   Using Wolfram Alpha we find the p-value to be 0.3372. Since the p value is fairly large, his hypothesis could be true that boys are born with probability 0.5.