# Homework # 1

Bhumsitt "Bose" Pramuanpornsatid

August 31, 2023

# Question 1

(10 points) A teacher gives 5 students a multiple choice test, in which each problem is worth 1 point and there is no penalty with negative points. The median and mean scores turn out to be 9 and 10 points, respectively.

(a) What is the minimum of the possible top scores?

(b) What is the maximum of the possible top score?

(c) What is the minimum of the possible standard deviations?

(d) What is the maximum of the possible standard deviations?

## Answer

As the mean is given to be 10 points and there are 5 students. The total score of all students is $5 * 10 = 50$ points.

(a) The minimum top scores is 12. This case is possible when we arrange the set of score to be $\{9, 9, 9, 11, 12\}$. The median is 9, and the mean is 10 which satisfy the question. By setting the score of the first 2 students to equal the median. The sum of the last 2 students must be 23. The combination of 11 and 12 will result in the minimum of the top scores and sum up to 23. Therefore, 12 is the minimum of the possible top scores.

(b) The maximum top scores is 32. This case is possible when we arrange the set of score to be $\{0, 0, 9, 9, 32\}$. The median is 9, and the mean is 10 which satisfy the question. By setting the score of the first 2 students to equal 0. The sum of the last 2 students must be 41. The combination of 9 and 32 will result in the maximum of the top scores and sum up to 41. Therefore, 32 is the maximum of the possible top scores.

(c) The minimum possible standard deviation is 1.265. This case is possible when we arrange the set of score to have the least variance $\{9, 9, 9, 9, 14\}$.
Calculation of standard deviation: $\sqrt{\frac{1+1+1+1+4}{5}} = \sqrt{1.6} \approx 1.265$

(d) The maximum possible standard deviation is 11.713. This case is possible when we arrange the set of score to have the most variance $\{0, 0, 9, 9, 32\}$.
Calculation of standard deviation: $\sqrt{\frac{100+100+1+1+484}{5}} = \sqrt{137.2} \approx 11.713$

# Question 2

(10 points) Let $\{x_i\}$ be a dataset consisting of $N$ real numbers, $x_1, ..., x_N$.

(a) Prove from definitions or proved properties in the textboook that the standardized data set $\{\hat{x_i}\}$ that is derived from $\{\hat{x_i}\}$ has mean = 0 and standard deviation = 1

(b) If the median of data set $\{\hat{x_i}\}$ is -0.5, is the data symmetric, left-skewed or right-skewed?

## Answer

(a) Mean:

$$\{\hat{x}\} = \{\frac{x_i - mean\{x_i\}}{std\{x_i\}}\}$$

$$mean\{\hat{x}\} = mean\{\frac{x_i - mean\{x_i\}}{std\{x_i\}}\}$$

$$mean\{\hat{x}\} = 0$$

Standard Deviation:

$$std\{\hat{x}\} = std\{\frac{x_i - mean\{x_i\}}{std\{x_i\}}\}$$

$$std\{\hat{x}\} = \{\frac{std\{x_i\}}{std\{x_i\}}\}$$

$$std\{\hat{x}\} = 1$$

(b) Data is right skewed because we know that a standardized data has a mean of 0 and the median is less than the mean.

# Question 3

```
import pandas as pd
import matplotlib.pyplot as plt
```
MagicPython

## Q3 1.1

You can find a dataset giving the cost (in 1976 US dollars), number of megawatts, and year of construction of a set of nuclear power plants at http://lib.stat.cmu.edu/DASL/Datafiles/NuclearPlants.html.

(a) Are there outliers in this data?

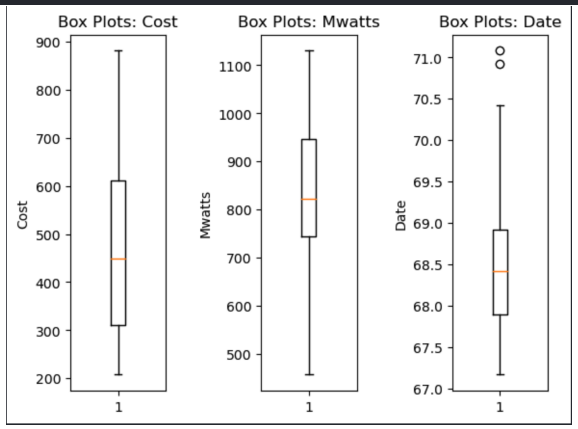
(b) What is the mean cost of a power plant? What is the standard deviation?

(c) What is the mean cost per megawatt? What is the standard deviation?

(d) Plot a histogram of the cost per megawatt. Is it skewed? Why?

```
power = pd.read_csv('./q3.csv', names=['Cost', 'Mwatts', 'Date'], sep=',', header=0, index_col=False)
```
MagicPython

```
# a
plt.subplot(1, 3, 1)
plt.boxplot(power['Cost'])
plt.title('Box Plots: Cost')
plt.ylabel('Cost')
plt.subplots_adjust(wspace = 1)
plt.subplot(1, 3, 2)
plt.boxplot(power['Mwatts'])
plt.title('Box Plots: Mwatts')
plt.ylabel('Mwatts')
plt.subplot(1, 3, 3)
plt.boxplot(power['Date'])
plt.title('Box Plots: Date')
plt.ylabel('Date')
plt.show()
```
MagicPython




(a) According the 3 box plots above, we can see that the "Blot Plots: Date" have two dots outside the range of the box and bracket. This means that there are 2 outliers in the data set. The other graphs doesn't contain any dots outside thus there are no outliers.

(b)
```
print(power.mean(axis=0))
print(power.std(axis=0))
```

Using the code above we can get the mean and standard deviation of each column of the data set. The result for cost is:

```
---mean---
Cost      $ 461.560313
---standard deviation---
Cost      $ 170.120670
```

(c) By dividing the cost by the power we can get the cost per watt. Using this code:

```
print("---mean cost per megawatts---")
power['costPerW'] = (power['Cost']/power['Mwatts'])
print(power.mean(axis=0))
print("---std cost per megawatts---")
print(power.std(axis=0))
```

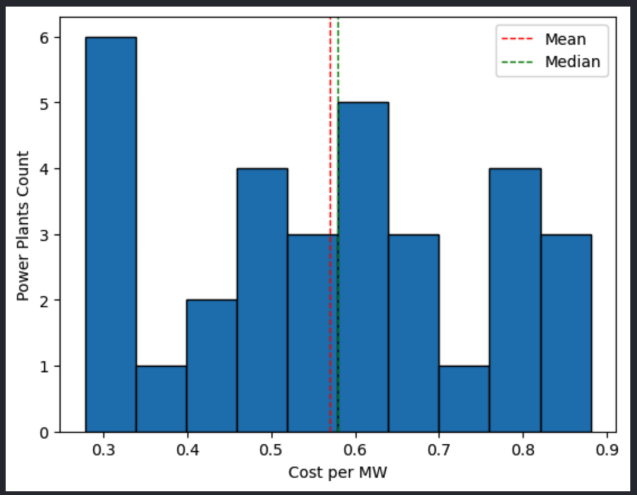
The result is:

```
---mean cost per megawatts---
costPerW      $ 0.569735 per megawatts
---std cost per megawatts---
costPerW      $ 0.187124 per megawatts
```

(d) The cost per megawatts is left skewed. This is because the mean is less than the median. The mean is 0.569735 and the median is 0.57957. Notice the mean line (red dashed line) is on the left of the median line (green dashed line) in the graph below.

```
# d
plt.hist(power['costPerW'], bins=10, edgecolor='black')
plt.ylabel('Power Plants Count')
plt.xlabel('Cost per MW')
plt.axvline(0.569735, color='red', linestyle='dashed', linewidth=1)
plt.axvline(0.57957, color='green', linestyle='dashed', linewidth=1)
plt.legend(['Mean', 'Median'])
plt.show()
```

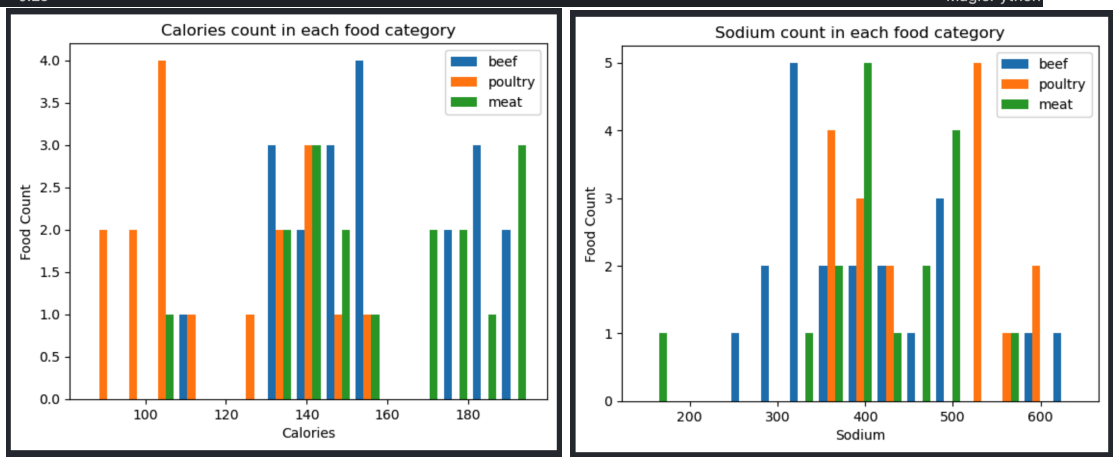
0.1s MagicPython



# Question 4

```python
import pandas as pd
import matplotlib.pyplot as plt

food = pd.read_csv('./q4.csv', names=['Type', 'Calories', 'Sodium'], sep=',', header=0, index_col=False)

beef = food[food['Type'] == 'Beef']
poultry = food[food['Type'] == 'Poultry']
meat = food[food['Type'] == 'Meat']

# Calories
print("---Calories count in each food category---")
plt.hist([beef['Calories'], poultry['Calories'], meat['Calories']], bins=15, label=['beef', 'poultry', 'meat'])
plt.legend(loc='upper right')
plt.title('Calories count in each food category')
plt.xlabel('Calories')
plt.ylabel('Food Count')
plt.show()

# Sodium
print("---Sodium count in each food category---")
plt.hist([beef['Sodium'], poultry['Sodium'], meat['Sodium']], bins=15, label=['beef', 'poultry', 'meat'])
plt.legend(loc='upper right')
plt.title('Sodium count in each food category')
plt.xlabel('Sodium')
plt.ylabel('Food Count')
plt.show()
```

✓ 0.2s MagicPython




The above code generates 2 of the histogram graphs. The one of the left show calories count in each food category and the one on the right show the sodium count in each food category.

By looking at the left graph, we can clearly see that many of the poultry food (orange) have the lowest calories count than the other food categories. The beef(blue) and meat(green) have similarly higher calories.

By looking at the right graph, we can see that the sodium count in each food category is very similar. The poultry(orange) have the highest sodium count spike, the beef(blue) have the lowest sodium count spike. The meet(green) have the second highest sodium count.

# Question 5

The code below plot a box plot showing the statistic of three or more syllable for each magazine group.

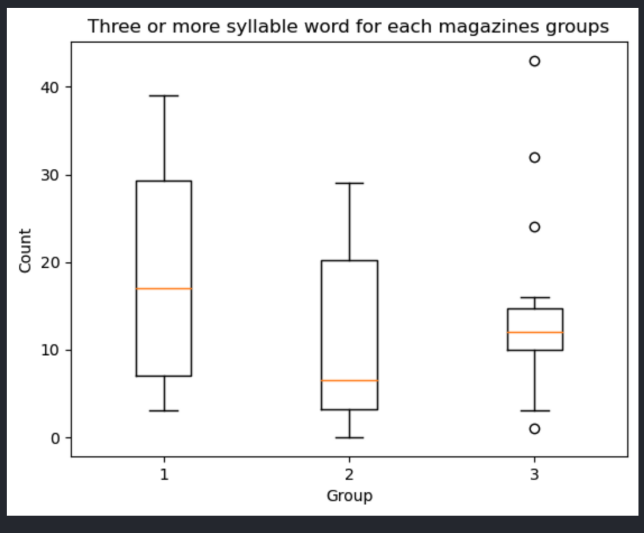
You will find a dataset giving (among other things) the number of 3 or more syllable words in advertising copy appearing in magazines at http://lib.stat.cmu.edu/DASL/Datafiles/magadsdat.html. The magazines are grouped by theeducation level of their readers; the groups are 1, 2, and 3 (the variable is called GRP in the data).

(a) Use a box plot to compare the number of three or more syllable words for the ads in magazines in these three groups. What do you see?

(b) Use a box plot to compare the number of sentences appearing in the ads in magazines in these three groups. What do you see?

```python
import pandas as pd
import matplotlib.pyplot as plt

magazine = pd.read_csv('./q5.csv', names=['WDS', 'SEN', '3SYL', 'MAG', 'GROUP'], sep=',', header=0, index_col=False)

# a
group1 = magazine[magazine['GROUP'] == 1]
group2 = magazine[magazine['GROUP'] == 2]
group3 = magazine[magazine['GROUP'] == 3]

plt.boxplot([group1['3SYL'], group2['3SYL'], group3['3SYL']])
plt.xlabel('Group')
plt.ylabel('Count')
plt.title('Three or more syllable word for each magazines groups')
plt.show()
```

✓ 0.0s MagicPython

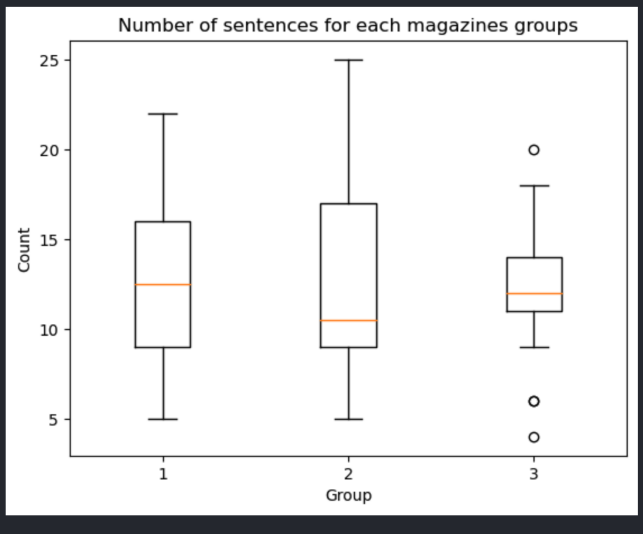



(a) Looking at the box plot we can notice that the median for the amount of three or more syllable word for each group are arrange from least to greatest as follows: 2, 3, 1. One major difference in the groups can be seen in magazine group 3. It has the significantly lower interquartile range compared to other and there are 4 outliers data in this group.

The code below plot a box plot showing the statistic of number of sentences for each magazine group.

```
# b
plt.boxplot([group1['SEN'], group2['SEN'], group3['SEN']])
plt.xlabel('Group')
plt.ylabel('Count')
plt.title('Number of sentences for each magazines groups')
plt.show()
```

0.0s MagicPython




(b) Looking at the box plot we can notice that the median for the amount of sentences for each group are arrange from least to greatest as follows: 2, 3, 1. One major difference in the groups can be seen in magazine group 3. It has the significantly lower interquartile range compared to other and there are 3 outliers data in this group.

# Question 6

(Extra credit 5 points) Let $\{x_i\}$ be a dataset consisting of $N$ real numbers, $x_1, ..., x_N$ . Prove the function $g(m) = \sum_{i=1}^{N} |(x_i - m)|$ is minimized when $m = median(\{x_i\})$ . Hint: try to prove $(\sum_{i=1}^{N} |(x_i - d)|) - (\sum_{i=1}^{N} |(x_i - median)|) \geq 0$ for $d \geq median$ , then for $d \leq median$ , d is any real number.

Let $\{x_i\}$ be a dataset consisting of $N$ real numbers, $x_1, ..., x_N$. Prove the function $g(m) = \sum_{i=1}^{N} |(x_i - m)|$ is minimized when $m = median(\{x_i\})$. Hint: try to prove $(\sum_{i=1}^{N} |(x_i - d)|) - (\sum_{i=1}^{N} |(x_i - median)|) \geq 0$ for $d \geq median$, then for $d \leq median$, d is any real number.

## Answer

First, consider the inequality $(\sum_{i=1}^{N} |x_i - d| - \sum_{i=1}^{N} |x_i - median| \geq 0)$ for $(d \geq median)$.

We break it into cases based on whether $(x_i)$ is less than or greater than the median, and note that the sum on the left side is larger than or equal to the sum on the right side.

Extending this argument to $(d \leq median)$, we find that the same inequality holds.

In both cases, the function $(g(m))$ can't be minimized with $(m)$ away from the median, as moving $(m)$ towards the median reduces the left-hand side of the inequality. Therefore, $(g(m))$ is minimized when $(m = median(\{x_i\}))$.