MATH 254: Introduction to Statistics **Mintaek Lee**

Lesson 1: Describing Data

Corresponding Workbook Module: 1

Example 1.1 *Mintaek is interested in how many geese are in Julia Davis Park. Mintaek sits on a bench from 9 am to 9 pm and counts the number of geese he observes. He repeated it for 10 consecutive days. Mintaek's geese counts of the day are listed below.*

20 24 24 24 24 26 28 28 28 29

Part A. *Find mean, median, and standard deviation of the given data set.*

Solution: We can find the sample mean as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{20+24+24+24+24+26+28+28+28+29}{10} = \frac{255}{10} = 25.5$$

There are 10 observations. Since $n = 10$ is an even number, the median is the average of two middle observations, 24 and 26. Therefore, median is $\frac{24+26}{2} = 25$. If n were an odd number, the median would have been the middle observation.

We can find the standard deviation using the following procedure (proceed from left to right)

x	$-\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$	$\sum(x-\bar{x})^2$
20	-25.5	-5.5	30.25	70.5
24	-25.5	-1.5	2.25	
24	-25.5	-1.5	2.25	
24	-25.5	-1.5	2.25	
24	-25.5	-1.5	2.25	
26	-25.5	$+0.5$	0.25	
28	-25.5	$+2.5$	6.25	
28	-25.5	$+2.5$	6.25	
28	-25.5	$+2.5$	6.25	
29	-25.5	$+3.5$	12.25	

So we get $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{10-1}\times 70.5} = \sqrt{\frac{70.5}{9}} = 2.7988.$

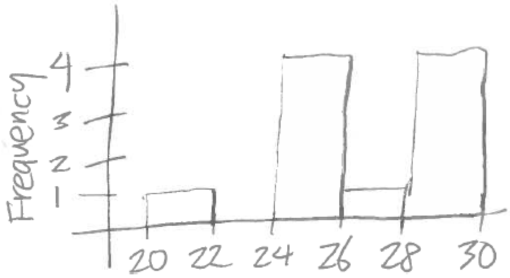
Part B. *Make stem-and-leaf plot and histogram of the data.*

Solution: Stem-and-leaf plot is shown below.

2	0	Stem: tens digit
2		Leaf: ones digit
2	4444	
2	6	Key: 2|4 = 24
2	8889	

Frequency table and the histogram are shown below.

Intervals	Frequency
$[20, 22)$	1
$[22, 24)$	0
$[24, 26)$	4
$[26, 28)$	1
$[28, 30)$	4



For more information on how to make stem-and-leaf plots and histograms, review the videos for pre-class reading or Section 1.2 of your textbook.

Part C. *Find the five number summary of the given data set. Are there any outliers?*

Solution: We found median as 25 from Part A. We also find minimum as 20 and maximum as 29. To find the five number summary, we need first quartile (Q_1) and third quartile (Q_3).

First quartile (Q_1) is the median of the values below (lower than) the median of the entire data set. Median was 25. Values below the median were: 20, 24, 24, 24, 24. Since there are five of them, the median of those five values would just be the middle value, which is 24.

Third quartile (Q_3) is the median of the values above (higher than) the median of the entire data set. Median was 25. Values above the median were: 26, 28, 28, 28, 29. Since there are five of them, the median of those five values would just be the middle value, which is 28.

The five number summary is then:

Min	Q_1	Median	Q_3	Max
20	24	25	28	29

To determine if there are any outliers, we use the following procedure

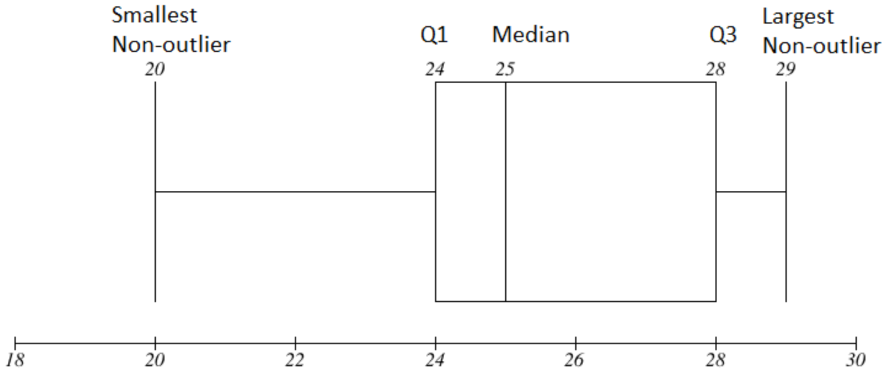
$$Q_1 - 1.5 \times IQR = 24 - 1.5 \times 4 = 24 - 6 = 18$$
$$Q_3 + 1.5 \times IQR = 28 + 1.5 \times 4 = 28 + 6 = 34$$

since $IQR = Q_3 - Q_1 = 28 - 24 = 4$.

We check if any observations fall outside the interval $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR) = (18, 34)$. Since no observation fall outside $(18, 34)$, we determine that there are no outliers.

Part D. *Make a boxplot of the given data set.*

Solution:



For more information on how to make boxplots, see workbook page 16.

Part E. *Fully describe the distribution of Mintaek's geese counts.*

Solution: When describing the distribution of a given data set, researchers generally focus on the four elements: center, unusual features, spread, and shape. The acronym "**CUSS**" may help you remember those four features. For more information, see workbook page 11.

What we found so far about this data set are as follows:

- **C**enter: mean was 25.5 and median was 25
- **U**nusual features: There are no outliers, although there is a gap between 20 and 24.
- **S**pread: sample standard deviation was 2.7988 and IQR was 4. Or the five number summary was 20, 24, 25, 28, 29.
- **S**hape: The distribution appears to be bimodal with slight left-skew.

In most cases, you need not list all findings. General rule of thumb is that mean and standard deviation are preferred when the data appear symmetric, and median and IQR (or five number summary) are preferred when the data appear skewed.

For this data set, you could say

> Median is 25.
>
> There are no outliers, although there is a gap between 20 and 24.
>
> IQR is 4 (OR five number summary is 20, 24, 25, 28, 29).
>
> The distribution appears to be bimodal with slight left-skew.

Note that it is unnecessary to write out CUSS.

Example 1.2 *Suppose that weights of adult emperor penguins are approximately normally distributed with mean 75 lbs and standard deviation 15 cm.*

Part A. *Mintaek is interested in adopting a pet emperor penguin. Mintaek found an emperor penguin weighing 65 lbs. Find a proportion of all penguins weighing less than 65 lbs.*

Solution: Let X be weight of penguins, then $X \sim N(\mu = 75, \sigma = 15)$. We first find the z-score of 65 lbs which is

$$z = \frac{x - \mu}{\sigma} = \frac{65 - 75}{15} = -0.67$$

Since we want to find a proportion of all penguins weighing **less than** 65 lbs, we want the area under the standard normal curve to the **left** of $z = -0.67$. Using the standard normal distribution (z) table, we find 0.2514 as the desired area.

Therefore, 25.14% of all emperor penguins would weigh less than 65 lbs. In other words, a weight of 65 lbs is about the 25-th percentile of all emperor penguin weights.

Part B. *Mintaek found another emperor penguin. This time, weighing 70 lbs. Find a proportion of all penguins weighing more than 70 lbs.*

Solution: We first find the z-score of 70 lbs which is

$$z = \frac{x - \mu}{\sigma} = \frac{70 - 75}{15} = -0.33$$

Since we want to find a proportion of all penguins weighing **more than** 70 lbs, we want the area under the standard normal curve to the **right** of $z = -0.33$. Remember that the standard normal distribution (z) table gives you the area under the standard normal curve to the **left** of $z = -0.33$ which was found as 0.3707. To find the area under the standard normal curve to the **right** of $z = -0.33$, we subtract 0.3707 from 1. We find $1 - 0.3707 = 0.6293$ as the desired area.

Therefore, 62.93% of all emperor penguins would weigh more than 70 lbs.

Part C. *Find a proportion of all penguins weighing between 65 lbs and 70 lbs.*

Solution: From Part A, we found the area under the standard normal curve to the **left** of $z = -0.67$ as 0.2504. It means 25.04% of all emperor penguins would weigh less than 65 lbs.

From Part B, we found the area under the standard normal curve to the **left** of $z = -0.33$ as 0.3707. It means 37.07% of all emperor penguins would weigh less than 70 lbs.

To find a proportion of all penguins weighing between 65 lbs and 70 lbs, we subtract 25.04% from 37.07%, giving us $37.07\% - 25.14\% = 11.93\%$.

Therefore, 11.93% of all emperor penguins would weigh between 65 lbs and 70 lbs.

Part D. *Suppose there is a regulation prohibiting Mintaek from adopting a pet emperor penguin whose weight is more than the 70-th percentile. What is the maximum weight of a penguin that Mintaek can legally adopt?*

Solution: The heaviest penguin that Mintaek can legally adopt would have a weight in the 70-th percentile of all emperor weights. It means we will need to find a z-score such that the area under the standard normal curve to the **left** of it is 0.7.

Looking at the standard normal distribution (z) table, we find that $z = 0.52$ gives 0.6985 and $z = 0.53$ gives 0.7019. We can make a guess that $z = \frac{0.52+0.53}{2} = 0.525$ would give about 0.7 as the area under the standard normal curve to the **left** of it.

To find the weight x corresponding to the z-score $z = 0.525$, we set up our equation as

$$z = \frac{x - \mu}{\sigma} \implies 0.525 = \frac{x - 75}{15}$$

Solving for x gives us $x = 82.875$.

The heaviest penguin that Mintaek can legally adopt would have a weight of 82.875 lbs.

Note that you can still earn full credits even if you used $z = 0.52$ or $z = 0.53$.

Additional Comments: When answering these types of questions, draw the standard normal distribution curve and shade the area you need and areas you find from the table. This practice should help you significantly reduce mistakes.