Organizing Raw Data

What is a **Raw Data**?

Raw Data (sometimes called source data) is data that has not been processed for meaningful use.

What is a **Frequency Distribution Table**?

A **Frequency Distribution Table** is one way that we can organize raw data so that it makes more sense and will give us a better picture of our obtained data than simply a list.

What is in a **Frequency Distribution Table**?

Any **Frequency Distribution Table** consists of

- **Rows** that divides the raw data into classes.
- **Columns** that holds specific information for each class.

How do we determine the number of **Classes**?

The number of **Classes** k for a raw data with **Sample Size** n must be a whole number

- ranging from 5 to 20, and it is usually given, or
- by finding the smallest k such that $2^k > n$.

We label **Columns** of a frequency distribution table as follows and we use combination of these columns to construct different statistical drawing:

- **Class Limits**
- **Class Boundaries**
- **Class Midpoints**
- **Class Frequencies**
- **Cumulative Frequencies**
- **Relative Frequencies**
- **Percentage Frequencies**

Here is an example of **Frequency Distribution Table**:

Class Limits	Class Boundaries	Class Midpoint	Class Frequency	Cumulative Frequency	Relative Frequency	Percentage Frequency

Always sort the raw data in ascending order before you start building the frequency distribution table.

Raw Data:

A sample of 40 exam scores in a math class is given below:

58	72	100	62	74	53	99	66	75	70
61	55	98	61	57	98	69	69	81	61
78	63	87	67	87	70	77	57	57	90
64	55	56	56	70	57	69	71	80	77

Sorted Data:

Here is the same but sorted data in ascending order that will be very useful as we build our frequency distribution table:

53	55	55	56	56	57	57	57	57	58
61	61	61	62	63	64	66	67	69	69
69	70	70	70	71	72	74	75	77	77
78	80	81	87	87	90	98	98	99	100

To make a **Frequency Distribution Table** , we first need to find its **Class Width** .

How do we find the **Class Width**?

Here are some steps that we need to take:

1. **Sort the raw data**
2. **Find the minimum and maximum value of the raw data**
3. **Find the range,** $Range = Max - Min$
4. **Find class width,** $CW = \dfrac{Range}{number\ of\ classes}$
   - **when** CW **is a whole number, add** 1.
   - **When** CW **is a decimal number, round up.**

Objective:

1. Find the minimum and the maximum value.
2. Find the value of the range.
3. Find the class width if we wish to have a frequency distribution table with 5 classes.

Solution:

Since our data is already sorted in ascending order, then we can easily discover that

Min.$=53$, Max.$=100$, Range $=$ Max. $-$ Min. $=100-53=47$,

$$\text{Class Width} = \frac{\text{Range}}{\text{Number of Classes}} = \frac{47}{5} = 9.4 \Rightarrow \boxed{\text{CW}=10}$$

How do we find the **Lower Class Limits**?

Here are some steps that we need to take:

1. **Choose the minimum data value or a convenient value below the minimum data value as your first lower class limit.**
2. **Second Lower Class Limit**
   = First Lower Class Limit + Class Width
3. **Third Lower Class Limit**
   = Second Lower Class Limit + Class Width
4. **Fourth Lower Class Limit**
   = Third Lower Class Limit + Class Width
5. **And so on.**

Objective:

Find all lower class limits.

Solution:

1. We begin by choosing the minimum value 53 as our first lower class limit
2. To find the rest of lower class limits, we just add the class width to each lower class limit in order to get the next lower class limit.

$$\begin{aligned}
\text{Second Lower Class Limit} &= 53 + 10 = 63 \\
\text{Third Lower Class Limit} &= 63 + 10 = 73 \\
\text{Fourth Lower Class Limit} &= 73 + 10 = 83 \\
\text{Fifth Lower Class Limit} &= 83 + 10 = 93
\end{aligned}$$

3. So our lower class limits are 53, 63, 73, 83, and 93.

How do we find the **Upper Class Limits**?

Here are some steps that we need to take for the raw data with D decimal places:

1. **First Upper Class Limit**
   = First Lower Class Limit + Class Width $-.1^D$
2. **Second Upper Class Limit**
   = First Upper Class Limit + Class Width
3. **Third Upper Class Limit**
   = Second Upper Class Limit + Class Width
4. **Fourth Upper Class Limit**
   = Third Upper Class Limit + Class Width
5. **And so on**.

Objective:

Find all upper class limits.

Solution:

To find the upper class limits for each class, we must examine our raw data to see if there are any numbers with decimals, and then we use this formula

First Upper Class Limit
$=$ **First Lower Class Limit** $+$ **Class Width** $-.1^D$ to find the first upper class limit.

1. In our example, we do not have any number with decimal, therefore $D = 0$, so we can find our first upper class limit:

$$\text{First Upper Class Limit} \quad = \quad 53 + 10 - .1^0 = 62$$

Solution Continued:

2. To find the rest of upper class limits, we just add a multiple of the class width to the first upper class limit in order to get the next upper class limit.

$$\begin{aligned}
\text{Second Upper Class Limit} &= 62 + 10 = 72 \\
\text{Third Upper Class Limit} &= 72 + 10 = 82 \\
\text{Fourth Upper Class Limit} &= 82 + 10 = 92 \\
\text{Fifth Upper Class Limit} &= 92 + 10 = 102
\end{aligned}$$

3. So our upper class limits are 62, 72, 82, 92, and 102.

Warning: It is extremely important to make sure that class width, first lower class limit, and first upper class limit have been calculated correctly.

We are now ready with our class limits to update our **Frequency Distribution Table**:

Class Limits	Class Boundaries	Class Midpoint	Class Frequency	Cumulative Frequency	Relative Frequency	Percentage Frequency
53–62						
63–72						
73–82						
83–92						
93–102						

How do we find the **Class Boundaries**?

Here are some steps that we need to take after we have found class limits:

1. **First Upper Class Boundary**

$$= \frac{\text{First Upper Class Limit} + \text{Second Lower Class Limit}}{2}$$

2. **First Lower Class Boundary**
   = First Upper Class Boundary - Class Width

3. **Add class width to each class boundary to get the successive class boundary.**

Objective:

Find all class boundaries.

Solution:

1. **First Upper Class Boundary**
   $= \frac{\text{First Upper Class Limit} + \text{Second Lower Class Limit}}{2}$
   $= \frac{62+63}{2} = 62.5$
2. **First Lower Class Boundary**
   $=$ First Upper Class Boundary $-$ Class Width
   $= 62.5 - 10 = 52.5$
3. **Now add class width to each class boundary to get the successive class boundary.**

Here is a complete list of class boundaries for our sample:

$$\boxed{52.5,\ 62.5,\ 72.5,\ 82.5,\ 92.5,\ 102.5}$$

Let's update our **Frequency Distribution Table** with class boundaries:

Class Limits	Class Boundaries	Class Midpoint	Class Frequency	Cumulative Frequency	Relative Frequency	Percentage Frequency
53–62	52.5–62.5					
63–72	62.5–72.5					
73–82	72.5–82.5					
83–92	82.5–92.5					
93–102	92.5–102.5					

How do we find the **Class Midpoint**?

Here are some steps that we need to take after we have found class limits:

1. **First Class Midpoint**
   $$= \frac{\text{First Lower Class Limit} + \text{First Upper Class Limit}}{2}$$
2. **Second Class Midpoint**
   $=$ **First Class Midpoint** $+$ **Class Width**
3. **Third Class Midpoint**
   $=$ **Second Class Midpoint** $+$ **Class Width**
4. **Fourth Class Midpoint**
   $=$ **Third Class Midpoint** $+$ **Class Width**
5. **And so on**.

Objective:

Find all class midpoints.

Solution:

1. **First Class Midpoint** $= \frac{53+62}{2} = \frac{115}{2} = 57.5$
2. **Second Class Midpoint** $= 57.5 + 10 = 67.5$
3. **Third Class Midpoint** $= 67.5 + 10 = 77.5$
4. **Fourth Class Midpoint** $= 77.5 + 10 = 87.5$
5. **Fifth Class Midpoint** $= 87.5 + 10 = 97.5$

Let's update our **Frequency Distribution Table** with class midpoints:

Class Limits	Class Boundaries	Class Midpoint	Class Frequency	Cumulative Frequency	Relative Frequency	Percentage Frequency
53–62	52.5–62.5	57.5				
63–72	62.5–72.5	67.5				
73–82	72.5–82.5	77.5				
83–92	82.5–92.5	87.5				
93–102	92.5–102.5	97.5				

How do we find the **Class Frequencies**?

Let's begin this by introducing few notations for this topic:

First Class Frequency	$\Rightarrow$	f_1
Second Class Frequency	$\Rightarrow$	f_2
Third Class Frequency	$\Rightarrow$	f_3

and so on.

Class frequencies are simply **tallies** to indicate how many numbers from our raw data falls within each class using class boundaries, this is where we can use the sorted raw data.

Objective:

Find class frequency for each class.

Solution:

1. **First Class Frequency** $f_1 = 14$
2. **Second Class Frequency** $f_2 = 12$
3. **Third Class Frequency** $f_3 = 7$
4. **Fourth Class Frequency** $f_4 = 3$
5. **Fifth Class Frequency** $f_5 = 4$

Here is a updated **Frequency Distribution Table** that includes the **Class Frequencies** .

Class Limits	Class Boundaries	Class Midpoint	Class Frequency	Cumulative Frequency	Relative Frequency	Percentage Frequency
53–62	52.5–62.5	57.5	14			
63–72	62.5–72.5	67.5	12			
73–82	72.5–82.5	77.5	7			
83–92	82.5–92.5	87.5	3			
93–102	92.5–102.5	97.5	4			

What are **Cumulative Frequencies**?

Cumulative frequencies are defined as the sum of all previous frequencies up to and including the frequency of the current class.

How do we find **Cumulative Frequency**?

Let's introduce few notations and formulas that we can use to find **Cumulative Frequencies**

First Class Cumulative Frequency	$\Rightarrow$	$cf_1 = f_1$
Second Class Cumulative Frequency	$\Rightarrow$	$cf_2 = cf_1 + f_2$
Third Class Cumulative Frequency	$\Rightarrow$	$cf_3 = cf_2 + f_3$

and so on.

Objective:

Find cumulative frequency for each class.

Solution:

We use class frequencies and the pattern we just discussed to complete this task.

$$
\begin{array}{lcl}
\text{First Class Cumulative Frequency} & \Rightarrow & cf_1 = 14 \\
\text{Second Class Cumulative Frequency} & \Rightarrow & cf_2 = 14 + 12 = 26 \\
\text{Third Class Cumulative Frequency} & \Rightarrow & cf_3 = 25 + 7 = 33 \\
\text{Fourth Class Cumulative Frequency} & \Rightarrow & cf_4 = 33 + 3 = 36 \\
\text{Fifth Class Cumulative Frequency} & \Rightarrow & cf_5 = 36 + 4 = 40
\end{array}
$$

STOP: Make sure that the last class cumulative frequency is equal to the sample size.

What are **Relative Frequencies**?

The **relative frequency** a class is defined as the frequency of the class divided by the sample size and is generally written as a decimal number rounded to 3 decimal places.

How do we find **Relative Frequency**?

Let's introduce few notations and formulas that we can use to find **Relative Frequencies**

$$\begin{aligned} &\text{First Class Relative Frequency} && \Rightarrow && rf_1 = f_1/n \\ &\text{Second Class Relative Frequency} && \Rightarrow && rf_2 = f_2/n \\ &\text{Third Class Relative Frequency} && \Rightarrow && rf_3 = f_3/n \end{aligned}$$

and so on where n is the size of our sample.

Objective:

Find relative frequency for each class.

Solution:

We use class frequencies and divide that by the sample size of 40 to complete this task. We generally round our answers to 3 decimal places.

$$\begin{aligned}
&\text{First Class Relative Frequency} && \Rightarrow && rf_1 = 14/40 = 0.350\\
&\text{Second Class Relative Frequency} && \Rightarrow && rf_2 = 12/40 = 0.300\\
&\text{Third Class Relative Frequency} && \Rightarrow && rf_3 = 7/40 = 0.175\\
&\text{Fourth Class Relative Frequency} && \Rightarrow && rf_4 = 3/40 = 0.075\\
&\text{Fifth Class Relative Frequency} && \Rightarrow && rf_5 = 4/40 = 0.100
\end{aligned}$$

STOP: Make sure that the sum of all relative frequencies is equal to 1.

What are **Percentage Frequencies**?

The **percentage frequency** a class is the relative frequency of the class written as percentage with 1 decimal place.

How do we find **Percentage Frequency**?

We simply multiply the relative frequency of each class by 100 to find **Percentage Frequency**.

$$\begin{aligned}
&\text{First Class Percentage Frequency} && \Rightarrow && pf_1 = rf_1 \times 100\\
&\text{Second Class Percentage Frequency} && \Rightarrow && pf_2 = rf_2 \times 100\\
&\text{Third Class Percentage Frequency} && \Rightarrow && pf_3 = rf_3 \times 100
\end{aligned}$$

and so on.

Objective:

Find percentage frequency for each class.

Solution:

We use relative frequency of each class and multiply that by 100 to complete this task. We generally round our answers to 1 decimal place, and write as with % symbol.

$$\begin{array}{lcl}
\text{First Class Percentage Frequency} & \Rightarrow & pf_1 = 0.350 \times 100 = 35.0\% \\
\text{Second Class Percentage Frequency} & \Rightarrow & pf_2 = 0.300 \times 100 = 30.0\% \\
\text{Third Class Percentage Frequency} & \Rightarrow & pf_3 = .175 \times 100 = 17.5\% \\
\text{Fourth Class Percentage Frequency} & \Rightarrow & pf_4 = 0.075 \times 100 = 7.5\% \\
\text{Fifth Class Percentage Frequency} & \Rightarrow & pf_5 = 0.100 \times 100 = 10.0\%
\end{array}$$

STOP: Make sure that the sum of all percentage frequencies is equal to 100%.

Let's update our **Frequency Distribution Table**.

Class Limits	Class Boundaries	Class Midpoint	Class Frequency	Cumulative Frequency	Relative Frequency	Percentage Frequency
53–62	52.5–62.5	57.5	14	14	.350	35.0
63–72	62.5–72.5	67.5	12	26	.300	30.0
73–82	72.5–82.5	77.5	7	33	.175	17.5
83–92	82.5–92.5	87.5	3	36	.075	7.5
93–102	92.5–102.5	97.5	4	40	.100	10.0


If I'd known they wanted me to use all this info- I would never have asked for it!