For wherever experimental work in physics, or in science generally is undertaken the degree of accuracy of the measurements and of the results of the experiments, must be of the first importance. The mathematical treatment throughout has been kept as simple as possible. It has seemed advisable, however, to explain the statistical concepts at the basis of the main considerations, and it is hoped that this lectures contains as elementary an account of the leading statistical ideas involved as is possible in such a small compass.

Frequency distribution:

Numerical data, including scientific measurements as well as industrial and social statistics, are often presented graphically to aid their appreciation.

The first step in dealing with such data, if they are sufficiently numerous, is to arrange them in some convenient order. This is often done by grouping them into classes according to their magnitude or according to suitable intervals of a variable on which they depend. For instance, the percentage marks obtained in an examination by a number of students could be grouped by counting the number of students who had marks between 0 and 9, 10 and
19, 29, 39 and 49. Thus dividing them into 10 classes. The data could then be tabulated as shown in table (1), in which the marks of a sample of 120 students have been used.

The number of data in each class is usually called the frequency for that class. Table (1) shows what is called the frequency distribution.
The pairs of numbers written in the columns headed "class", for example, 0 and 9, 10 and 19 and so on, are usually called the lower and upper class limits. The width of any class is the difference between the first number specifying that class and the first number specifying the next, that is, 10 for each of the classes shown in Table 1. For some groupings, however, the widths of the classes may be unequal.

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Class</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 9</td>
<td>2</td>
<td>50 - 59</td>
<td>32</td>
</tr>
<tr>
<td>10 - 19</td>
<td>5</td>
<td>60 - 69</td>
<td>25</td>
</tr>
<tr>
<td>20 - 29</td>
<td>6</td>
<td>70 - 79</td>
<td>10</td>
</tr>
<tr>
<td>30 - 39</td>
<td>14</td>
<td>80 - 89</td>
<td>2</td>
</tr>
<tr>
<td>40 - 49</td>
<td>22</td>
<td>90 - 99</td>
<td>2</td>
</tr>
</tbody>
</table>

The classification shown in Table 1 obviously helps us to appreciate the distribution of marks amongst the students; we can see at a glance, for instance, how many students have fewer than 40 marks and how many have 70 or more. But a graphical representation can make it possibly even clearer. The data are plotted in Fig. 1, where the marks are represented along the horizontal axis and the frequency...
along the vertical axis.

Fig (1). Frequency polygon of examination marks.

The points obtained by plotting the frequency against the mid-value of the corresponding class, namely, 4.5, 14.5, ..., 94.5 are joined by straight lines and the resulting figure is known as a frequency polygon.

A different method is used in Fig (2) where a series of rectangles are constructed of width equal to the class width and of area equal to the frequency of the corresponding class. That is, the rectangles in Fig (2) have areas equal to 2, 5, 6, 14, 22, 32, 25, 10, 2, 2 units respectively.

Fig (2). Histogram of examination marks.
The figure obtained is called a histogram, and the total area of the histogram in this case is 120 units equal to the total number of students.

Since the area of each rectangle in a histogram represents the frequency in the corresponding class, the heights of the rectangles are proportional to the frequencies when the classes have equal widths. In this case, the mean height of the rectangles is proportional to the mean frequency.
1. Relative frequency distribution: normalized.

2. Student were asked to time the oscillations of the same pendulum. Jamal said the period was 1.4 sec. Rami said it was 1.53 sec. Which was to be believed? What matters is how much information comes out from the experimental results period and reliability of the measurements made. A single numerical result from an experiment cannot give both pieces of information. At least two numbers are required to give a result and measure its reliability. After repeating the measurement five times:

- Samir gave: 1.4, 1.7, 1.7, 1.4, 1.6 sec
- Rami gave: 1.53, 1.50, 1.51, 1.52, 1.52 sec

We can compare these results by plotting them as frequency distribution.

![Frequency distribution graphs](attachment:image.png)
This gives a very different view of the matter. The arithmetic average or mean of results are:

Samir: \[
\frac{1.4 + 1.7 + 1.7 + 1.4 + 1.6}{5} = 1.56 \text{ sec} \text{ instead of } 1.4 \text{ sec}
\]

Sara: \[
\frac{1.53 + 1.5 + 1.51 + 1.51 + 1.52}{5} = 1.514 \text{ sec} \text{ instead of } 1.53 \text{ sec}
\]

Comparing these two results:

- Sara's measurements are closely bunched together—none less than 1.50 sec or greater than 1.53 sec, while
- Samir's measurements spread from 1.4 to 1.7 sec

Sara looks more accurate than Samir.

Now instead, Samir came back with 500 measurements and the average value of 1.5326 sec as shown in Figure below.

Fig (4)
This is still much broader on than Sarq's measurement. It is important we would attach greater value to Samir's 500 measurements than to his 5. But we should find some numerical way of expressing this that enables us to compare his 500 with the 5 of Sarq.

It is difficult to compare the results of 5 measurement with those of 500 if we use the same scale. It is better to present the results as relative frequency distributions, where the ordinates are not the actual number of times a measurement is recorded but the ratio of this no. to the total no. of measurements made at that stage. The results of Fig (3) & (4) in relative frequency distribution form are represented in Fig (5) & (6).
From now on we shall use frequency to mean relative frequency rather than absolute frequency.

It is essential, when presenting results as relative frequency distribution, to record with it the total no. of measurements made.

Suppose that when a total of \( (n) \) measurements of a quantity \( (X) \) have been made, an experiment shows that \( X_1 \) was recorded \( (n_1) \) times, \( X_2 \) was recorded \( (n_2) \) times, \( \ldots \), \( X_m \) was recorded \( (n_m) \) times. Now:

\[
N = n_1 + n_2 + n_3 + \ldots + n_m
\]  \( \text{(1)} \)

So that the ordinates of relative frequency distribution at the values \( X_1, X_2, \ldots, X_m \) will be relative:

\[
\begin{align*}
    f_n (X_1) &= \frac{n_1}{N} \\
    f_n (X_2) &= \frac{n_2}{N} \\
    f_n (X_m) &= \frac{n_m}{N}
\end{align*}
\]  \( \text{(2)} \)

Then

\[
f_n (X_1) + f_n (X_2) + \ldots + f_n (X_m) = \frac{(n_1 + n_2 + \ldots + n_m)}{N} = 1
\]
A distribution with this property is said to be normalized so that changing a frequency distribution from absolute to relative form is equivalent to normalizing the distribution.

Limiting frequency distribution

Figs. (5) & (6) show that even when they are normalize frequency distribution fluctuate as the total no. of measurements increase. The distribution for (5) measurements look rather different from that for (10).

However, as the no. increase, the fluctuations decrease. Small difference is marked betw. the distributions for (50) and (100) measurements but when comparing (500) and (1000) measurements the difference between the normalized frequency distribution is negligible. This means that the distribution settles down to a more and more definite shape as the no. of measurements increases, i.e., there is limiting frequency distribution for the infinite experiment. We shall denote this limiting frequency distribution without a subscript, so that
\( f(x_1), f(x_2), \ldots, f(x_n) \) are the relative frequency of recording the measurements \( x_1, x_2, \ldots, x_n \) in the infinite experiment (see Fig. (7))

- Fig. (7) -

The existence of the limiting distribution is always an assumption and cannot be tested satisfactorily. It on experiments is kept on and on making measurements he may become tired and careless or the apparatus will begin to wear, both of which will tend to break the frequency distribution or the quantity to be measured is varying with time and room temperature. So that freq. distribution will drift up or down as the measurements are repeated. So all scientific theories start from some assumptions.
Discret value distribution:

If \( X \) is the no. of days in which rain falls during a year, \( X \) could have any integer value from (0) to (360).

For such an experiment, a discrete value distribution as shown in Fig (8) is appropriate.

![Histogram](image)

- Fig (8) -

Histograms:

If we measuring a group of children each year to see how they were growing, there would be no set of exact weights that could describe every child. Growth is gradual or continuous process.

If \( X \) is a continuous variable (variable which could have any value), we decide on intervals with end points \( X_0, X_1, X_2, \ldots, X_m \) and determine from an experiment the relative frequencies of measurements lying within them:

\[
\frac{n_1}{n} \quad \text{with} \quad X_0 \leq X \leq X_1, \\
\frac{n_2}{n} \quad \text{with} \quad X_1 \leq X \leq X_2, \\
\vdots \\
\frac{n_m}{n} \quad \text{with} \quad X_{m-1} \leq X \leq X_m
\]
where \( n = n_1 + n_2 + \ldots + n_m \) is the total no. of measurements. These values are shown as rectangles. They constitute anormalized histogram of the results (see Fig.(9)).

\[ \begin{align*}
\text{Relative Frequency} \\
\text{Per unit interval in } x \\
X_0 & \quad X_1 & \quad X_2 & \quad X_{m-1} & \quad X_m \\
\frac{n_1}{n} & \quad \frac{n_2}{n} & \quad \frac{n_3}{n} & \quad \frac{n_{m-1}}{n} & \quad \frac{n_m}{n}
\end{align*} \]

- **Fig.(9)** -

The heights of the rectangles, \( f_n(x_i), f_n(x_2), \ldots, f_n(x_m) \), are the ordinates of the histogram and are such that:

\[(X_i - X_0) f_n(x_i) = \frac{n_1}{n} \quad \text{(is the relative frequency of obtaining measurement } X_0 \leq X \leq X_1)\]

\[(X_2 - X_1) f_n(x_2) = \frac{n_2}{n} \quad \text{(is the relative frequency of obtaining measurement } X_1 \leq X \leq X_2)\]

\[\vdots\]

\[(X_m - X_{m-1}) f_n(x_m) = \frac{n_m}{n} \quad \text{(is the relative frequency of obtaining measurement } X_{m-1} \leq X \leq X_m)\]
Area under histogram = \( (x_1 - x_0) f_n(x_1) + (x_2 - x_1) f_n(x_2) + \cdots + (x_m - x_{m-1}) f_n(x_m) \)
\[ = (n_1 + n_2 + \cdots + n_m)/n \]
\[ = 1 \quad (4) \]

Intervals \((x_1 - x_0), (x_2 - x_1), \cdots, (x_m - x_{m-1})\) do not need to be equal.

Continuous distribution:

we can deal with both discrete value distribution and histogram together if we imagine a smooth curve to be drawn through the outline of the chart (see Fig. 10 & 11)

If we label the curve \( F_n(x) \), then when it is used to represent a discrete value distribution it has the property:

\[ F_n(x_1) = f_n(x_1), \quad F_n(x_2) = f_n(x_2), \quad \cdots \quad F_n(x_m) = f_n(x_m) \]
when it is used to represent a histogram, \( F_n(x) \) is approximately equal to the height of each rectangle at the midpoint of the interval.

More precisely, area under each section is equal to the corresponding area under the histogram:

\[
\int_{x_{r-1}}^{x_r} F_n(x) \, dx = (x_r - x_{r-1}) \cdot F_n(x_r)
\]

Hence,

\[
\int_{x_0}^{x_0} F_n(x) \, dx = (x_1 - x_0) F_n(x_1) + (x_2 - x_1) F_n(x_2) + \ldots + (x_m - x_{m-1}) F_n(x_m) = 1
\]

---

Fig (11)
True value: Mean, Mode and Median

The limiting frequency distribution curve represents in a compact from all the information that an experiment can yield. Both, the physical quantity (being measured) and the apparatus are involved in finding position and shape of the curve.

Sometimes the curve is determined entirely by the quantity that is being measured.

Example:

we examined a large no. of ladders to see how many steps they had. Nearly always we count (10). Sometimes we count (9) or (8) and some other times we Count (11) or (12)

If we counted the steps of more and more ladders the frequency distribution would settle down and approach its limiting form as shown.
The form of the distribution tells us that:

1. Ladder can have any no. of steps from 6 - 10.
2. Majority have 8 steps - we call this the normal or natural no.
3. Accidents give fewer or more than 8 steps.

The true value is the value at the central axis of symmetry of the curve, X. The most powerful argument for this is that it could not very well be any other value.

![Diagram](image)

For example: If **X₁** were chosen in above figure, it is difficult to see how any argument for this choice, could not equally well be used for **X₂**.

Only X has unique properties. Among them are the following:

1. it is the mean of all measurements.
2. it is the mode - the value with the greatest frequency.
3. it is the median - measurements above and below X occur equally frequently.
Any of these properties could be used to define the true value and in this case they all give the same value.

If the curve is not symmetrical, we obtain different values for the mean value \( \bar{x} \) and the median \( x' \).

The mean is calculated as follows:

1. **Discrete value distribution:**
   Multiply each measurement value by the relative frequency with which it occurs and sum over all possible values.
   \[
   \bar{x} = \bar{x} = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_m f(x_m)
   \]

2. **Histogram**
   Multiply the mid-value of measurement in each interval by the relative frequency for that interval and sum over all possible values.
   \[
   \bar{x} = \bar{x} = \frac{1}{2} (x_1 + x_2) (x_2 - x_1) f(x_1) + \frac{1}{2} (x_2 + x_3) (x_3 - x_2) f(x_2) + \ldots
   + \frac{1}{2} (x_m + x_{m-1}) (x_{m-1} - x_m) f(x_m)
   \]

3. **Distribution Curve**
   Multiply the measurement value by the corresponding distribution curve function and integrate over all values; that is, the smooth curve approximation to the preceding
expression,

\[ X = \bar{X} = \int_{-\infty}^{\infty} x f(x) \, dx \]
1- Best estimate of the true values:

When using the limiting frequency curve we are considering the results of an experimental measurement repeated an infinite number of lines.

In fact we can make only a finite number of independent measurements (n), and can make only an estimate of the true value.

Samin first five measurements had a mean value \( X_5 \)

\[ X_5 = \frac{1.4 + 1.7 + 1.4 + 1.6 + 1.7}{5} \]

= 1.56 sec.

This is not necessarily the limiting value however, at any stage:

\[ X_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \]

is the mean

In a similar way we could talk of the best estimate of the mode or the median value.

2- Best estimate of precision

From (n) measurement we first calculate the mean (or best estimate of the true value):