Measures of central tendency and frequency distribution

Deependra Dhakal

Assistant Professor

Agriculture and Forestry University

https://rookie.rbind.io

- **Descriptive statistics** are statistics that describe a set of data.
- For quantitative data
  - They show a tendency to concentrate at certain values, usually somewhere in centre of the distribution. Measures of this tendency are called measures of central tendency or averages.
  - The data vary about a measure of central tendency and these measures of deviation are called measures of *variation* or *dispersion*.
  - The data in a frequency distribution may fall into symmetrical or asymmetrical patterns. The measures of the direction and degree of asymmetry are called measures of *skewness*.
  - Polygons of frequency distributions exhibit flatness or peakedness of frequency curves. The measures of peakedness or flatness of the frequency curves are called measures of *kurtosis*.
- For numerical data, the frequency distribution can usefully be supplemented by a few numerical measures.
- For categorical data, the frequency distribution provides a concise and complete summary of a sample.
- Measures of the "center" or "typical value" of the data can be defined in several ways.

- A book-keeper sold following units of the same book during 20 consecutive days. How can we describe the sales of the book ?

$$12, 0, 5, 13, 0, 0, 5, 10, 5, 1, 5, 6, 7, 5, 7, 8, 10, 5, 11, 14$$

- arrange in order

$$0, 0, 0, 1, 5, 5, 5, 5, 5, 5, 6, 7, 7, 8, 10, 10, 11, 12, 13, 14$$

- Mode is most frequently occuring number
- Median is value that divides data in half
- Mean
  - sum of all observations divided by number of observations

$$\frac{12 + 0 + 5 + 13 + 0 + 0 + 5 + 10 + 5 + 1 + 5 + 6 + 7 + 5 + 7 + 8 + 10 + 5 + 11 + 14}{20}$$

Arithmetic mean

- The mean (sample mean) of a sample is the sum of the observations divided by the number of observations.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- where:
  - y_i's are the observations in the sample and n is the sample size (that is, the number of y_i's)

Mean of grouped data

$$\bar{y} = \frac{\sum_{k=1}^{r} f_k y_k}{\sum_{k=1}^{r} f_k}$$

Where:
f_k and y_k are the frequency and numerical value of the k^{th} item in the collection.
y_k also refers to the midpoint value of the k^{th} class in a grouped numeric data having fixed class intervals.

(Statistics for the Life Sciences, 2016) An agronomist counted the number of leaves on each of 150 tobacco plants of the same strain (Havana). The results are shown in Table 1.

Table 1: Number of leaves on tobacco plants

Number of leaves	Frequency (number of plants)
17	3
18	22
19	44
20	42
21	22
22	10
23	6
24	1

For the data given above,
$\bar{y} = 19.78$.

Median

- The sample median is the value that most nearly lies in the middle of the sample – the data value that splits the ordered data into two equal halves.
- Requires construction of ordered array
- Median is denoted by $\tilde{y}$

Median of grouped data

Steps

- Construct the cumulative frequency distribution
- Decide the class that contain the median. Class median is the first class with the value of cumulative frequency equal at least $\frac{n}{2}$
- Median is given by following formula:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right)$$

Where:

- n = total frequency
- F = cumulative frequency before class median
- f_m = frequency of the class median
- i = class width
- L_m = lower boundary of the class median

Based on the grouped data below, find the medium.

Table 2: Frequency distribution of time to travel to work also showing cumulative class interval frequencies for successive classes

Time to travel to work	Frequency	Cumulative frequency
1 - 10	8	8
11 - 20	14	22
21 - 30	12	34
31 - 40	9	43
41 - 50	7	50

$\frac{n}{2} = \frac{50}{2} = 25 \longrightarrow$ class median in the 3$^{\text{rd}}$ class,
So, $F = 22$, $f_m = 12$, $L_m = 20.5$ and $i = 10$.
$\therefore$ Median $= 21.5 + \left(\frac{25-22}{12}\right) \times 10 = 24$.
Thus, 25 persons take less than 24 minutes to travel to work and another 25 persons take more than 24 minutes to travel to work.

Interquartile range of grouped data

- The above expression for Median can be modified so that we can get the Q_1 and Q_3 as follows:

$$\mathrm{Q}_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) \qquad \mathrm{Q}_3 = L_{Q_3} + \left(\frac{\frac{3n}{4} - F}{f_{Q_3}} \right)$$

Now,

Interquartile range (IQR)

$$IQR = \mathrm{Q}_3 - \mathrm{Q}_1$$

Mode

- Mode is the value that has the highest frequency in a data set.
- For grouped data, class mode (or modal class) is the class with the highest frequency.
- Following formula gives the mode of the given set of grouped data:

$$\text{Mode} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times i$$

- Where:
  - i is the class width
  - Δ_1 is the difference between the frequency of class mode and the frequency of the class after the class mode
  - Δ_2 is the difference between the frequency of class mode and the frequency of the class before the class mode
  - L_{mo} is the lower boundary of class mode

Numerical example

The number of days that children were missing from school due to sickness in on year was recorded. Estimate the mean, medium and the modal class.

Number of days off sick	Frequency (f_k)	Mid-point (y_k)	Cumulative frequency	$f_k \times y_k$
1-5	12	3	12	36
6-10	11	8	23	88
11-15	10	13	33	130
16-20	4	18	37	72
21-25	3	23	40	69

Mean $= \frac{395}{40} = 9.925$ days.

As there are 40 pupils, we need to consider the mean of the 20th and 21st values. These both lie in the 6-10 class interval (which is really the 5.5-10.5 class).

As there are 12 values in the first class interval, the medium is found by considering the 8th and 9th values of the second interval.

As there are 11 values in the second interval, the median is estimated as being $\frac{8.5}{11}$ of the way along the second interval.

Use case: library books

- Students borrow and return books from a university library regularly

$$14, 13, 12, 11, 17, 20, 14, 16, 12, 12, 11, 9, 18, 21$$

- One student borrows a book and forgets to return it
- They re-discover the book 1 year later when moving out of the student accommodation and decide to return the book

$$14, 13, 12, 11, 17, 20, 14, 16, 12, 12, 11, 9, 18, 21, 365$$

- What do you think will happen to the mean and median of a data set on **borrowing periods**?

"A statistician drowned while crossing a river that was on average six inches deep"

- The mean is the preferred measure of central tendency when describing a data that do not have outliers.
- A major disadvantage is that it is affected by outliers (i.e. single observations which are very extreme compared with most observations and whose inclusion or exclusion changes results noticeably).
- In the presence of outliers, the median is the preferred measure of central tendency
- Calculation of the median does not involve the use of all available data and is therefore has less power than the mean.

Partition values

Quartiles

A quartile is one of 4 values (lower quartile, median and upper quartile) which divides data into 4 equal groups.

Percentiles

A percentile is one of 99 values which divides data into 100 equal groups. The lower quartile corresponds to the 25th percentile. The median corresponds to the 50th percentile and the upper quartile corresponds to the 75th percentile.
Formally, the percentile is:

$$\text{Percentile (P)} = (1 - w)x_{(j)} + wx_{(j+1)}$$

For some weight w between 0 and 1. Different approaches to choosing w are described – In R, there are nine different alternatives to compute the quantile.

Geometric mean

In mathematics, geometric mean is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the n^{th} root of the product of n numbers, i.e., for a set of numbers $x_1, x_2, x_3, ..., x_n$, the geometric mean is defined as:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln \alpha_i\right)$$

For the example book sales data mentioned above ($0, 0, 0, 1, 5, 5, 5, 5, 5, 5, 6, 7, 7, 8, 10, 10, 11, 12, 13, 14$), the geometric mean is (note the data contain 0 as values, logarithmic transformation of which is undefined so adding a positive value before the transformation and subtraction after exponentiation should give proper geometric mean): 4.74.

Relationship between GM, AM and HM

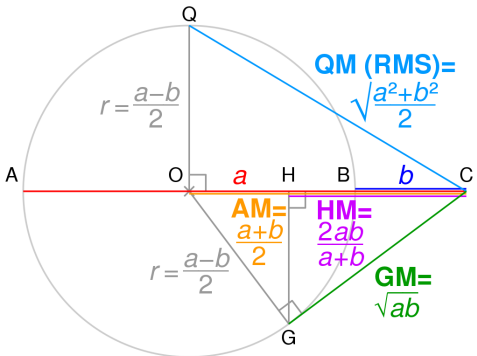


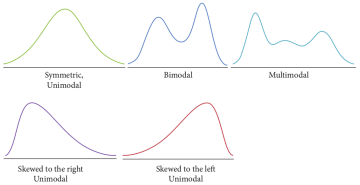
Figure 1: Geometric proof that max(a,b) > root mean square (RMS) or quadratic mean (QM) > arithmetic mean (AM) > geometric mean (GM) > harmonic mean (HM) > min(a,b) of two distinct positive numbers a and b. *Source:https://en.wikipedia.org/wiki/Geometric_mean*

Diagrammatic presentation of data

- Bar
  - A graph of a frequency distribution for a categorical data set. Each category is represented by a segment of the bar, and the area of the segment is proportional to the corresponding frequency or relative frequency.
- Pie
  - A graph of a frequency distribution for a categorical data set. Each category is represented by a slice of the pie, and the area of the slice is proportional to the corresponding frequency or relative frequency.
- Histogram
  - A picture of the information in a frequency distribution for a numerical data set. A rectangle is drawn above each possible value (discretized data) or class interval. The rectangle's area is proportional to the corresponding frequency or relative frequency.
- Frequency polygon
- Frequency curve
- Ogives (cumulative frequency curves)

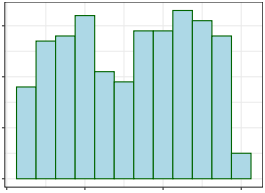
Histogram

- Histogram shows the shape of a distribution.
- Distribution can be called unimodal (one peak), bimodal (two peaks) or multimodal (multiple peaks).
- Distribution can be symmetric or skewed. A skewed distribution is asymmetrical with a longer tail on one side.



- Let us consider the marks in a subject obtained by 300 candidates selected at random among those appearing a certain examination.

Score class	Frequency
(45.1,55.9]	65
(55.9,66.6]	58
(66.6,77.4]	51
(77.4,88.1]	66
(88.1,98.9]	60



Bar diagram

- A bar diagram is a pictorial representation of a frequency distribution of categorical data.
- It is Useful to represent data with multiple occurrence or counts.
- Pie diagram accommodates lesser classes and better shows relative proportion of occurrences or count data.
- For the table 3, we can generate a bar diagram as shown in 2.

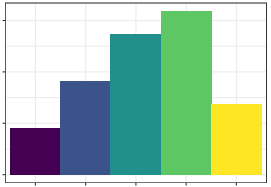


Figure 2: Bar graph showing frequency distribution of mid-term scores of 22 pupils

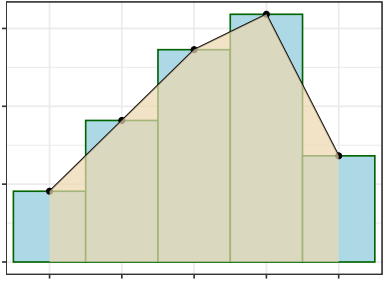
Table 3: Frequency distribution of mid-term scores of 22 pupils in a statistics exam

pupil	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
scores	69	84	52	93	81	74	89	85	88	63	87	90	67	72	74	55	82	91	68	77	70	77

Frequency polygon and cumulative frequency curve (Ogive)

Frequency polygon

- Frequency curve, relatively smooth frequency curves, unlike polygons, is for continuous data. For example assume the mid term score data in its original form as being continuous.



Ogives (cumulative frequency curves)

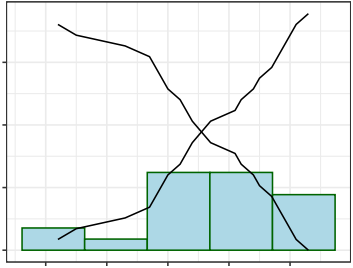


Figure 3: Cumulative frequency curve fitted to the mid-term score distribution

- In the discrete frequency distribution, above, height of the bar shows relative frequency of each class.
- When we use the height as the relative frequency if the interval lengths are unequal, histogram may not appropriately display the frequency distribution.

Frequency distribution

When observations of either discrete or continuous nature are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information. A **frequency distribution** is a table that displays the frequency of observation in each interval in a sample. To build a frequency distribution, we need the following steps.

1. Find the minimum and the maximum values in the dataset
2. Determine class intervals: intervals or cells of equal length that cover the range between the minimum and the maximum without overlapping. e.g., minimum 0, maximum 100: [0, 10), [10, 20), . . . , [90, 100).
3. Find frequency: the number of observations in the data that belog to each class interval. Let's denote the frequencies as $f_1, f_2, ...$
4. Find relative frequency:

$$\frac{\text{Class frequency}}{\text{Total number of observations}}$$

The relative frequency are denoted as $f_1/n, f_2/n, ...$ if the total sample size is n.

- For example, for the Mid-term scores data (Shown in Table 3), we could construct class intervals to show relative frequencies as follows:

Table 4: Relative frequencies of mid-term scores expressed in score classes of interval 10

Class interval	Frequency	Relative frequency
50-59	2	9.1%
60-69	4	18.2%
70-79	6	27.3%
80-89	7	31.8%
90-99	3	13.6%

- There is no gold standard in selecting class intervals, but a rule of thumb is an integer near $\sqrt{n}$ for the number of classes.

- Tabular data facilitates the presentation of large information into concise way under different titles and subtitles so that can further be subjected to statistical analysis.
  - Simple tabulation (for single variable; e.g., Table 4)
  - Double tabulation (tabulation of different crops under condition of irrigation and without irrigation)
  - Triple tabulation
  - Manifold tabulation (Tabulation based on more than 3 characteristics; data of students in a college according to native place, class residence and sex.)

	Residence	Class	n
Rural			
Female			
	Hostellers	Graduate	13
		Intermediate	7
		Post Graduate	7
	Scholars	Graduate	20
		Intermediate	3
		Post Graduate	13
Male			
	Hostellers	Graduate	16
		Intermediate	3
		Post Graduate	6
	Scholars	Graduate	13
		Intermediate	7
		Post Graduate	22

	Residence	Class	n
Urban			
Female			
	Hostellers	Graduate	2
		Intermediate	2
		Post Graduate	2
	Scholars	Graduate	3
		Intermediate	4
		Post Graduate	2
Male			
	Hostellers	Graduate	4
		Intermediate	1
		Post Graduate	1
	Scholars	Graduate	6
		Intermediate	2

Question 1

A group of 50 biomedical students recorded their pulse rates by counting the number of beats for 30 seconds and multiplying by 2.

a. Why are all of the measurements even number ?
b. Draw a stem and leaf plot to describe the data, splitting each stem into two lines.
c. Construct a relative frequency histogram for the data.
d. Write a short paragraph describing the distribution of the student pulse rates.

Solution 1

a. Because the resulting data is constructed from multiplication of original values by an even integer.

b. Stem and leaf plot is:

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##     4 | 2
##     5 | 24688
##     6 | 002266668
##     7 | 0002222248
##     8 | 002244444444468888
##     9 | 0066
##    10 | 04
##    11 | 0
```