Mathematics IV
(Probability and Statistics)

MATH 403

Lecture 1

Dr. Phoebe Edward Nashed

Course Information

- Course Instructor: Dr. Phoebe Edward Nashed Kirelloss
- Email: phoebe.edward@giu-uni.de
- Office Hours: Monday 5th slot, you need to send an email to reserve your appointment.
- TAs:
  - Eng. Mohamed Ayman Karakish: Mohamed.ayman@giu-uni.de
  - Eng. Seif Tarek: seif.tarek@giu-uni.de
- Textbooks:
  - Probability and Statistics for Computer Scientists (Chapman & Hall)

Course Assessment

- 10% Assignments
- 20% Quizzes (best 2 out of 3)
- 30% Midterm
- 40% Final exam

Course Outline

- Descriptive Statistics
- Probability Theory and Random Variables
- Discrete and Continuous Distributions
- Inductive Statistics: Interval Estimates
- Markov Chains and Queuing

Lecture 1 - Outline

- Descriptive Statistics
  - Raw Data
    - Measures of Central Tendency: Mean, Median, Mode
    - Measures of dispersion: Range, Variance, Standard Deviation
  - Grouped Data:
    - Frequency Distribution
    - Histogram
    - Measures of Central Tendency for Grouped Data: Mean, Median, Mode
    - Measures of dispersion: Variance, Standard Deviation

Introduction

- Statistics is a collection of methods which help us to describe, summarize, interpret, and analyze data. Drawing conclusions from data is vital in research, administration, and business.
- Researchers are interested in understanding whether a medical intervention helps in reducing the burden of a disease, how personality relates to decision-making, and many more questions.
- Governments may be interested in the life expectation of a population, the risk factors for infant mortality, migration patterns, or reasons for unemployment.
- In business, identifying people who may be interested in a certain product, optimizing prices, and evaluating the satisfaction of customers are possible areas of interest.

The Need for Data

- No matter what the question of interest is, it is important to collect data in a way which allows its analysis. The representation of collected data in a **data set** or **data matrix** allows the application of a variety of statistical methods. In this lecture, we are going to introduce the framework of statistics which is needed to properly collect, administer, evaluate, and analyze data.

Central Tendency of Data

- A data set may contain many observations. However, we are not always interested in each of the measured values but rather in a summary which interprets the data. Statistical functions fulfil the purpose of summarizing the data in
  a meaningful yet concise way.

Example 1

- Suppose someone from Munich (Germany) plans a holiday in Bangkok (Thailand) during the month of December and would like to get information about the weather when preparing for the trip. Suppose last year's maximum temperatures during the day (in degrees Celsius) for December are as follows:

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,
25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

Example 1

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,

25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

- How do we draw conclusions from this data? Looking at the individual values gives us a feeling about the temperatures one can experience in Bangkok, but it does not provide us with a clear summary.
- It is evident that the **average** of these 31 values as "Sum of all values / Total number of observations"

$$(22 + 24 + \cdots + 28 + 29)/31 = 26.48$$

- is meaningful in the sense that we know what temperature to expect "on average".

Example 1

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,
25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

- To choose the right clothing for the holidays, we may also be interested in knowing the temperature **range** to understand the variability in temperature, which is between 21 and 31℃.
- Summarizing 31 individual values with only three numbers (26.48, 21, and 31) will provide sufficient information to plan the holidays.

Central Tendency and Dispersion

- In this lecture notes, we focus on the most important statistical concepts to summarize data: these are measures of central tendency and dispersion.
  - Measures of Central Tendency:
    - Arithmetic Mean
    - Median
    - Mode
  - Measures of dispersion:
    - Range
    - Variance
    - Standard Deviation

Measures of Central Tendency

- Usually we look for a unique number to represent the raw data, this number is generally called the **average**. We are going to discuss the most common three different types of averages which are frequently used in applications.
  - Arithmetic Mean ($\bar{x}$)
  - Median ($\hat{x}$)
  - Mode ($\tilde{x}$)

Arithmetic Mean ($\bar{x}$)

❖ Arithmetic mean of $\boldsymbol{n}$ observations is the sum of the values divided by its number. If we have the observations $x_1, x_2, x_3, \cdots\cdots, x_n$, then

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots\cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median ($\hat{x}$)

❖ Median of $\boldsymbol{n}$ observations is the midpoint of the observations if $\boldsymbol{n}$ is odd and the average of the two midpoints if $\boldsymbol{n}$ is even. Of course, this must be done after rearrange the observations in ascending order or in descending order.

$$\hat{x} = x_{\left(\frac{n+1}{2}\right)} \qquad \text{if } \boldsymbol{n} \text{ is odd}$$

$$\hat{x} = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] \qquad \text{if } \boldsymbol{n} \text{ is even}$$

Mode ($\tilde{x}$)

The word "Mode" in French means: "Fashion". So, the mode is the value that is mostly wide spread and highly repeated.

- Mode of n observations is the value of the reading that appears most often.
- The mode is undefined for sequences in which no observation is repeated.

Important Remark

- A data set that has only one value that occurs with the greatest frequency is said to be **unimodal**.
- If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal**.
- If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal**.
- When no data value occurs more than once, the data set is said to have **no mode**.

Example 2

❖ For the data sets below, find mean, mode, and median.

(A) 20, 22, 25, 25, 21, 26, 25, 8, 19, 31.

(B) 20, 22, 25, 21, 26, 25, 8, 19, 22, 30, 35.

(C) 20, 22, 24, 21, 26, 25, 8, 18, 23, 30, 10, 31.

Example 2 – Sol.

(A) 20, 22, 25, 25, 21, 26, 25, 8, 19, 31.

- The number of observations is $\boxed{n = 10}$
- Mean $= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(222) = 22.2$
- Mode $= 25$ (The most repeated value)
- The sorted data is:
  - 8, 19, 20, 21, 22, 25, 25, 25, 26, 31.
- Median $= \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] = \frac{x_5+x_6}{2} = \frac{22+25}{2} = 23.5$
- The median is the value which divides the observations into two equal parts.

Example 2 – Sol.

B. 20, 22, 25, 21, 26, 25, 8, 19, 22, 30, 35.

- The number of observations is $\boxed{n = 11}$
- Mean $= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{11}(253) = 23$
- Mode $=$ 22 and 25
- There are two modes 22 and 25 in the data set, so it is called **bimodal** data.
- The sorted data is:
  - 8, 19, 20, 21, 22, 22, 25, 25, 26, 30, 35.
- Median $= x_6 = 22$

Example 2 – Sol.

C. 20, 22, 24, 21, 26, 25, 8, 18, 23, 30, 10, 31.

- The number of observations is $\boxed{n = 12}$
- Mean $= \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{12}(258) = 21.5$
- There is no mode.
- The sorted data is:
    - 8, 10, 18, 20, 21, 22, 23, 24, 25, 26, 30, 31.
- Median $= \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] = \frac{x_6+x_7}{2} = \frac{22+23}{2} = 22.5$

Linearly Transformed Data

- If the data is **linearly** transformed as

$$y_i = a\, x_i + b,$$

- where a and b are known constants, it holds that

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a\, x_i + b) = \frac{1}{n}\left[\sum_{i=1}^{n} a\, x_i + \sum_{i=1}^{n} b\right]$$

$$= \frac{1}{n}\left[a\sum_{i=1}^{n} x_i + \ n\, b\right] = \frac{a}{n}\sum_{i=1}^{n} x_i + b = a\, \bar{x} + b$$

Example 3

- Recall Example 1 where we considered the temperatures in December in Bangkok. We measured them in degrees Celsius, but someone might prefer to know them in degrees Fahrenheit. With a linear transformation, we can create a new temperature variable as

$$\boxed{\text{Temperature in °F} = 32 + (1.8 * \text{Temperature in °C})}$$

$$\text{i.e.} \quad y_i = 1.8\, x_i + 32$$

Remember that $\bar{x} = 26.48$

- Using $\bar{y} = a\,\bar{x} + b$, we get

$$\bar{y} = (1.8 * 26.48) + 32 = 79.664 \text{ °F}$$

- where $\bar{y}$ is the average temperature in degrees Fahrenheit.

Measures of dispersion

- The mean is not enough to represent the data. Why? The answer appears in another question. What about the two different groups

  (22, 36, 23, 35, 20, 34) and (17, 35, 20, 42, 16, 40)?

- Both have the same mean of 28.3 but the **variation** about this mean is different. So, we are in need to find a measure to the dispersion of the data about this mean.
  - Measures of dispersion:
    - Range $(\boldsymbol{R})$
    - Variance $(\boldsymbol{S}^2)$
    - Standard Deviation $(\boldsymbol{S})$

Range (R)

- Range of n observations is the difference between the maximum and minimum value of the data as

$$R = x_{max} - x_{min}$$

Variance (S^2)

- Variance of $\boldsymbol{n}$ observations is the arithmetic mean of the squared deviations from the mean.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- where $\bar{x}$ is the arithmetic mean.
- The mean value $\bar{x}$ of a finite set of measurements differs from the true mean μ of the theoretical infinite population of measurements. Because of this, a biased estimate of the variance and standard deviation are obtained accordingly.
- A better prediction of the variance of the infinite population can be obtained by applying the Bessel correction factor $(\frac{n}{n-1})$ to the variance formula.
- NB: In practice, in case $n > 25$, the Bessel correction factor is quite small and can be neglected.

Standard Deviation (S)

- Standard Deviation of $\boldsymbol{n}$ observations is the square root of the variance.

$$S = \sqrt{S^2}$$

- Note that both variance S^2 and standard deviation S always have positive values.

Note

- The standard deviation has the same unit of measurement as the data whereas the unit of the variance is the square of the units of the observations.
- The standard deviation measures how much the observations vary or how they are dispersed around the arithmetic mean. A low value of the standard deviation indicates that the values are highly concentrated around the mean. A high value of the standard deviation indicates lower concentration of the observations around the mean, and some of the observed values may even be far away from the mean.

Outliers

- If there are extreme values or outliers in the data, then the arithmetic mean is more sensitive to outliers than the median.
- Moreover, extreme values affect strongly on the value of the range.

Example 4

- The following data represents the grades of 20 computer science students on a math test. Find the mean, mode, median, range, variance, and standard deviation of these grades.

87, 86, 85, 87, 86, 87, 85, 81, 76, 85, 84, 85, 83, 82, 80, 79, 80, 74, 78, 90.

Example 4 – Sol.

87, 86, 85, 87, 86, 87, 85, 81, 76, 85, 84, 85, 83, 82, 80, 79, 80, 74, 78, 90

- $\boxed{n = 20}$
- Mean $= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20}(1660) = 83$
- Mode $= 85$ (appeared 4 times)
- <u>The sorted data is</u>:
- 74, 76, 78, 79, 80, 80, 81, 82, 83, 84, 85, 85, 85, 85, 86, 86, 87, 87, 87, 90.
- Median $= \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] = \frac{x_{10}+x_{11}}{2} = \frac{84+85}{2} = 84.5$

Example 4 – Sol.

87, 86, 85, 87, 86, 87, 85, 81, 76, 85, 84, 85, 83, 82, 80, 79, 80, 74, 78, 90

- $\boxed{n = 20}$
- <u>The sorted data is</u>:
- 74, 76, 78, 79, 80, 80, 81, 82, 83, 84, 85, 85, 85, 85, 86, 86, 87, 87, 87, 90.
- Range $= 90 - 74 = 16$
- Variance $= S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- $= \frac{1}{19} [(74 - 83)^2 + (76 - 83)^2 + \cdots \cdots + (90 - 83)^2] = 17.16$
- $\therefore$ Standard deviation $= S = \sqrt{17.16} = 4.14$

Example 4 – Sol. (Another Solution)

87, 86, 85, 87, 86, 87, 85, 81, 76, 85, 84, 85, 83, 82, 80, 79, 80, 74, 78, 90

- $\boxed{n = 20}$, $\bar{x} = 83$

Data	87	86	85	87	86	87	85	81	76	85	84	And so on …
Deviation from mean $x_i - \bar{x}$	4	3	2	4	3	4	2	-2	-7	2	1	…
Squared Deviations $(x_i - \bar{x})^2$	16	9	4	16	9	16	4	4	49	4	1	…

- Variance $= S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- $= \frac{1}{19} [16 + 9 + 4 + 16 + 9 + 16 + 4 + 4 + 49 + 4 + 1 \cdots\cdots] = 17.16$
- $\therefore$ Standard deviation $= S = \sqrt{17.16} = 4.14$
- Note that squared deviations are always positive values.

Important Remarks - 1

❖ The sum of the deviations of each variable around the arithmetic mean is zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\,\bar{x} - n\,\bar{x} = 0$$

Important Remarks - 2

- Let us consider a linear transformation $y_i = a\,x_i + b$ of the original data x_i , $(i = 1,2,\cdots,n)$. We get the arithmetic mean of the transformed data as

$$\bar{y} = a\,\bar{x} + b$$

- and for the variance:

$$(S^2)_y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}([a\,x_i + b] - [a\,\bar{x} + b])^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(a\,x_i - a\,\bar{x})^2 = \frac{a^2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = a^2\,(S^2)_x$$

Important Remarks - 3

- If you're finding the variance by hand, the "usual" formula can be a bit difficult. An alternative version is the **computational** formula, which can be a little easier to work:

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - n(\bar{x})^2\right]$$

- i.e.

$$\text{Variance} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - n(\bar{x})^2\right]$$

Proof of Last Remark

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i)^2 - 2(x_i)(\bar{x}) + (\bar{x})^2]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - 2\,\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}(\bar{x})^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - 2\,\bar{x}\,(n\,\bar{x}) + n\,(\bar{x})^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - n(\bar{x})^2\right].$$

Example 5

- Let x_i , $i = 1,2,\cdots,n,$ denote measurements on time. These data could have been recorded and analyzed in hours, but we may be interested in a summary in minutes. We can make a linear transformation $y_i = 60\, x_i$. Then,

$$\boxed{\bar{y} = 60\, \bar{x}} \quad \text{and} \quad \boxed{(S^2)_y = 60^2\, (S^2)_x}$$

- If the mean and variance of the $\boldsymbol{x_i}$'s have already been obtained, then the mean and variance of the $\boldsymbol{y_i}$'s can be obtained directly using these transformations.

Example 6

- Recall Example 4 where we calculated the variance using the usual formula. Now, we will calculate it using the computational formula.

87, 86, 85, 87, 86, 87, 85, 81, 76, 85, 84, 85, 83, 82, 80, 79, 80, 74, 78, 90.

Example 6 – Sol.

87, 86, 85, 87, 86, 87, 85, 81, 76, 85, 84, 85, 83, 82, 80, 79, 80, 74, 78, 90.

- $\sum_{i=1}^{n} x_i = 87 + 86 + 85 + \cdots\cdots + 90 = 1660$
- $\sum_{i=1}^{n} (x_i)^2 = (87)^2 + (86)^2 + (85)^2 + \cdots\cdots + (90)^2 = 138106$
- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{20}(1660) = 83$
- $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} (x_i)^2 - n(\bar{x})^2\right] = \frac{1}{19}\left[138106 - (20)(83)^2\right] = 17.16$

Grouped Data

- In a survey, people usually collect a lot of data, known as **raw data**. Sometimes, the collected data can be too numerous to be meaningful. We need to organize raw data in some logical manner in order to make sense out of them. Wherever you have a large amount of data, the grouped frequency distribution makes it easy to analyze the data. Therefore, the raw data can be arranged in a frequency distribution.

Grouped Data – Frequency Distribution

- We could group data into subgroups (**classes**). Each class is known as a **class interval**. The intervals in grouped frequency distribution are called class limits.

Raw Data

- For example, suppose that the following raw data comes from a certain experiment.

15.8	26.4	17.3	11.2	23.9	24.8	18.7	13.9	9.0	13.2
22.7	9.8	**6.2**	14.7	17.5	26.1	12.8	28.6	17.6	23.7
26.8	22.8	18.0	20.5	11.0	20.9	15.5	19.4	16.7	10.7
19.1	15.2	22.9	26.6	20.4	21.4	19.2	21.6	16.9	19.0
18.5	23.0	24.6	20.1	16.2	18.0	7.7	13.5	23.5	14.5
14.4	29.6	19.4	17.0	20.8	24.3	22.5	24.6	18.4	18.1
8.3	21.9	12.3	22.3	13.3	11.8	19.3	20.0	25.7	**31.8**
25.9	10.0	15.9	27.5	18.1	17.9	9.4	24.1	20.1	28.5

Grouped Data – Frequency Distribution

- These 80 readings are too hard to deal with. So, we rearrange them into classes, for simplicity with **equal** class width. In order to deal more easily with it, the next table with the so called class midpoint $(\boldsymbol{x_i})$ is introduced.

Class (Interval) Limits	Midpoint (x_i)	Frequency (f_i)
$5 \le x < 9$	7	3
$9 \le x < 13$	11	10
$13 \le x < 17$	15	14
$17 \le x < 21$	19	**25**
$21 \le x < 25$	23	17
$25 \le x < 29$	27	9
$29 \le x \le 33$	31	2
		$n = \sum f_i = 80$

Frequency Distribution

- How to construct **frequency distribution** for the data using an appropriate scale?
- **Step 1:** *Find the range*
  In this example, the greatest value is 31.8 and the smallest value is 6.2. So,

$$\boxed{\text{Range}} = 31.8 - 6.2 = 25.6$$

- The scale of the frequency distribution must contain the range of values.

Frequency Distribution

- **Step 2:** *Find the classes*
- The classes separate the scale into **equal** parts. We can choose let's say 7 classes.

$$\boxed{\text{Class width} = \frac{\text{Range}}{\text{Number of classes}}} = \frac{25.6}{7} = 3.7 \cong 4$$

- Note that it is easier to deal with whole number class width, so we round up.
- We should round up, not approximate in order to incorporate all data values in all classes.

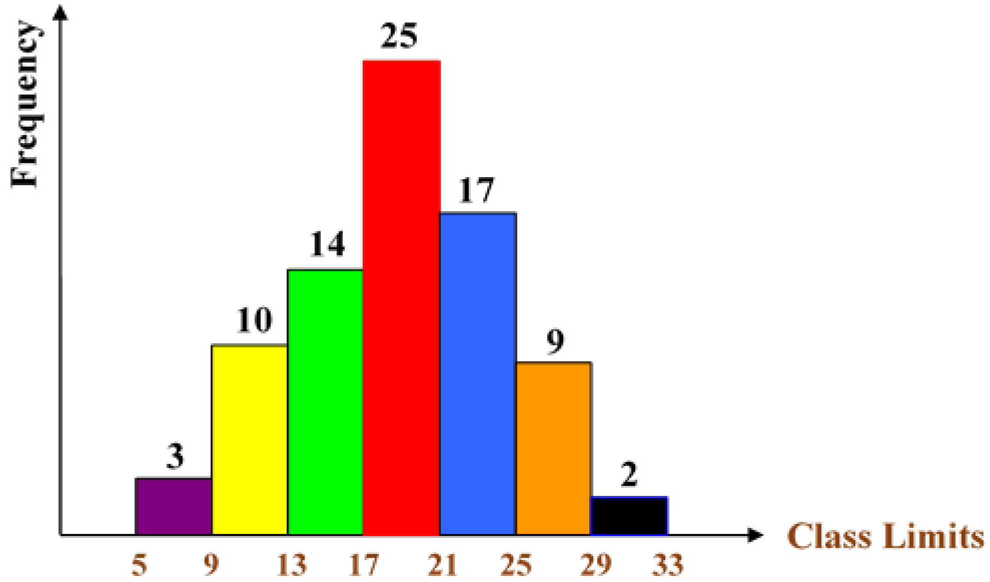
Frequency Distribution

- **Step 3:** *Select a suitable value to start the scale with*
- For our example, we can begin with 5 (an appropriate value ≤ 6.2).
- N.B: You should not choose a very small value to start with, in order to make sure you will include all data values within the designed classes.

Histogram

❖ The histogram consists of a set of rectangles whose heights represent the frequencies. If the class widths are all the same (**always**, we chose the class width to be the same), then the height of the rectangles may be taken to represent the frequencies as in the following figure.

Class (Interval) Limits	Midpoint (x_i)	Frequency (f_i)
$5 \le x < 9$	7	3
$9 \le x < 13$	11	10
$13 \le x < 17$	15	14
$17 \le x < 21$	19	**25**
$21 \le x < 25$	23	17
$25 \le x < 29$	27	9
$29 \le x \le 33$	31	2
		$n = \sum f_i = 80$



Measures of Central Tendency and Dispersion for Grouped Data

- If we have a frequency distribution (as in the previous table), we can calculate the mean, variance, mode, and median as follows

$$\text{Mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{\boldsymbol{m}} x_i\, f_i$$

$$\text{Variance} = S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{\boldsymbol{m}} (x_i - \bar{x})^2 f_i\right] = \frac{1}{n-1}\left[\sum_{i=1}^{\boldsymbol{m}} (x_i)^2 f_i - n(\bar{x})^2\right]$$

- where $\boldsymbol{m}$ is the number of classes, $\boldsymbol{x_i}$ is the class midpoint and $\boldsymbol{f_i}$ is the class frequency.

$$n = \sum_{i=1}^{\boldsymbol{m}} f_i$$

- NB: In practice, in case $n > 25$, the Bessel correction factor is quite small and can be neglected.

Example

- So, for the data given in previous table where $\boxed{m = 7}$

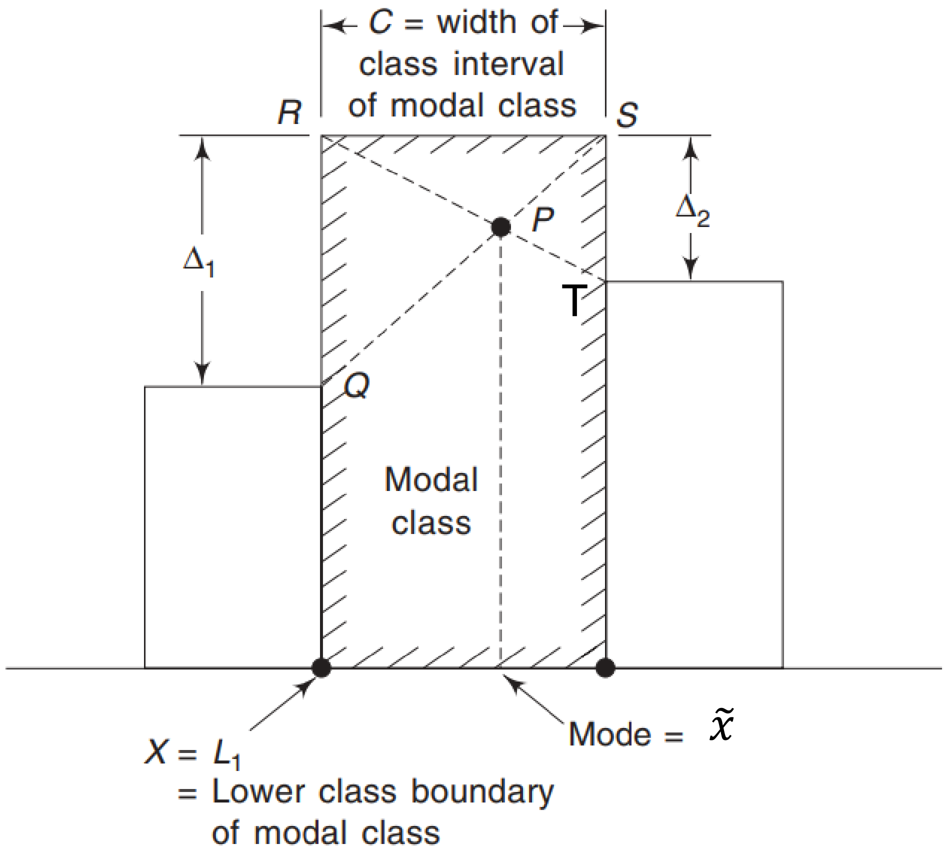
Class (Interval) Limits	Midpoint (x_i)	Frequency (f_i)
$5 \le x < 9$	7	3
$9 \le x < 13$	11	10
$13 \le x < 17$	15	14
$17 \le x < 21$	19	**25**
$21 \le x < 25$	23	17
$25 \le x < 29$	27	9
$29 \le x \le 33$	31	2
		$n = \sum f_i = 80$

- $\bar{x} = \frac{1}{80}[(7)(3) + (11)(10) + \cdots\cdots + (31)(2)] = \frac{1}{80}(1512) = 18.9$
- $S^2 = \frac{1}{n}[\sum_{i=1}^{m}(x_i - \bar{x})^2 f_i] = \frac{1}{80}(2431.2) = 30.39$
- **<u>OR</u>**
- $S^2 = \frac{1}{n}[\sum_{i=1}^{m}(x_i)^2 f_i - n(\bar{x})^2] = \frac{1}{80}[31008 - (80)(18.9)^2] = 30.39$
- NB: In practice, in case $n > 25$, the Bessel correction factor is quite small and can be neglected.

Mode- Grouped Data

- To determine the modal value, consider a part of the histogram showing the highest frequency rectangle along with the adjacent lower class and higher class rectangles. We define the mode as the x-axis value $\tilde{x}$ of the point of intersection P of the lines QS and RT. It can be shown that the value of mode $\tilde{x}$ is given by the formula:

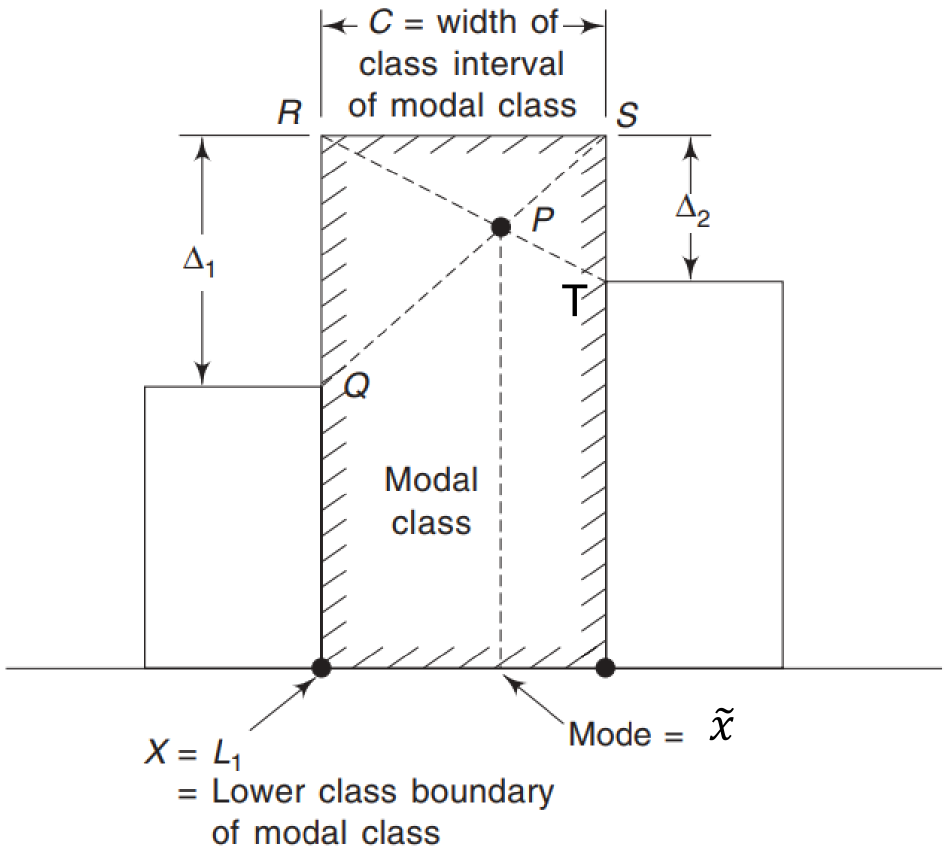
$$\tilde{x} = L_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] C$$



Mode- Grouped Data

$$\tilde{x} = L_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] C$$

- where L_1 = lower class boundary of the modal class (i.e. class containing mode).
- Δ_1 = excess of the modal frequency over frequency of next lower class.
- Δ_2 = excess of the modal frequency over frequency of next higher class
- C = modal class interval size



Example

$$Mode = \tilde{\boldsymbol{x}} = L_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] C$$

- For the data in the given table, the **modal class** is (17 – 21). So,
- Mode $= 17 + \left[\frac{11}{11+8}\right](4) = 19.32$

Class (Interval) Limits	Midpoint (x_i)	Frequency (f_i)
$5 \le x < 9$	7	3
$9 \le x < 13$	11	10
$13 \le x < 17$	15	14
$17 \le x < 21$	19	**25**
$21 \le x < 25$	23	17
$25 \le x < 29$	27	9
$29 \le x \le 33$	31	2
		$n = \sum f_i = 80$

Median – Grouped Data

- To evaluate the median, we first define the **median class** which is the class with cumulative frequencies greater than $(n/2)$ [i.e. the sum of frequencies till this class] and then

$$\text{Median} = \hat{x} = L_1 + \left[\frac{\left(\frac{n}{2}\right) - n_b}{f_m}\right] C$$

Class (Interval) Limits	Midpoint (x_i)	Frequency (f_i)
$5 \le x < 9$	7	3
$9 \le x < 13$	11	10
$13 \le x < 17$	15	14
$17 \le x < 21$	19	**25**
$21 \le x < 25$	23	17
$25 \le x < 29$	27	9
$29 \le x \le 33$	31	2
		$n = \sum f_i = 80$

Median – Grouped Data

$$\text{Median}=\hat{x} = L_1 + \left[\frac{\left(\frac{n}{2}\right) - n_b}{f_m}\right] C$$

- where
- L_1: lower limit of the median class
- n: number of observations
- n_b: cumulative frequency before the median class
- f_m: frequency of the median class
- C: width of the median class

Example

$$\text{Median} = \hat{x} = L_1 + \left[\frac{\left(\frac{n}{2}\right) - n_b}{f_m}\right] C$$

- For the data in the given table, the **median class** is (17 – 21). So,
- Median $= 17 + \left[\frac{\left(\frac{80}{2}\right) - 27}{25}\right](4) = 19.08$
- N.B: Not necessarily the median class is the same as the modal class!

Class (Interval) Limits	Midpoint (x_i)	Frequency (f_i)
$5 \leq x < 9$	7	3
$9 \leq x < 13$	11	10
$13 \leq x < 17$	15	14
$17 \leq x < 21$	19	**25**
$21 \leq x < 25$	23	17
$25 \leq x < 29$	27	9
$29 \leq x \leq 33$	31	2
		$n = \sum f_i = 80$

Example 7

- The data below shows the mass of 40 students in a class. Each measurement is to the nearest kg.

55	70	57	73	55	59	64	72
60	**48**	58	54	69	51	63	78
75	64	65	57	71	**78**	76	62
49	66	62	76	61	63	63	76
52	63	71	61	53	56	67	71

a) Construct a frequency distribution for the masses with 4 classes.

b) Generate a histogram.

c) Find mean, mode, median and variance.

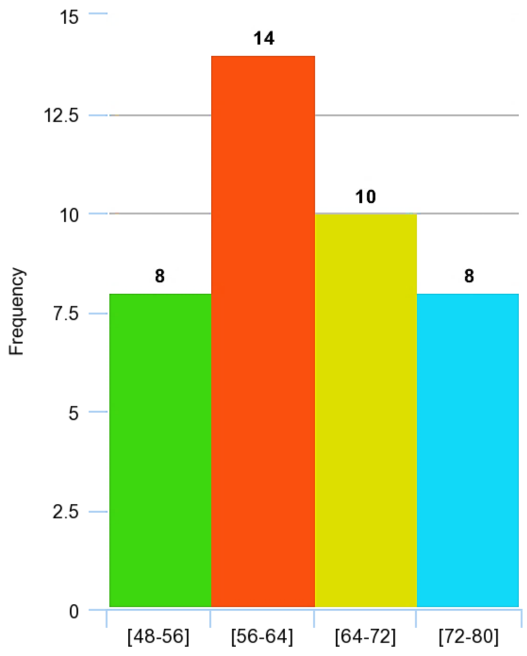
Example 7 – Sol.

- The greatest mass is 78 and the smallest mass is 48. The range of the masses is then
- Range $= 78 - 48 = 30$
- Class width $= 30/4 = 7.5 \cong 8$

Class (Interval) Limits		Midpoint (x_i)	Frequency (f_i)
$48 \le x < 56$	$48 - 56$	52	8
$56 \le x < 64$	$56 - 64$	60	14
$64 \le x < 72$	$64 - 72$	68	10
$72 \le x \le 80$	$72 - 80$	76	8
			$n = \sum f_i = 40$

Example 7 – Sol.

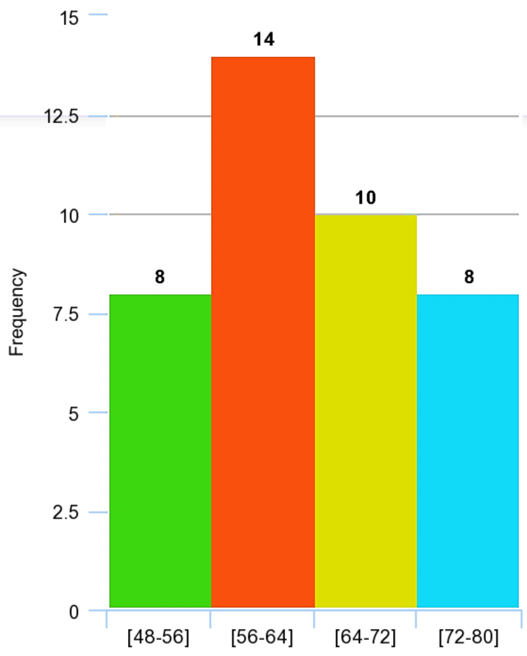
Class (Interval) Limits		Midpoint (x_i)	Frequency (f_i)
$48 \le x < 56$	$48 - 56$	52	8
$56 \le x < 64$	$56 - 64$	60	14
$64 \le x < 72$	$64 - 72$	68	10
$72 \le x \le 80$	$72 - 80$	76	8
			$n = \sum f_i = 40$



- The Mean:
- $\bar{x} = \frac{1}{40}[(52)(8) + (60)(14) + (68)(10) + (76)(8)] = \frac{1}{40}(2544) = 63.6$
- The Variance:
- $S^2 = \frac{1}{n}\left[\sum_{i=1}^{m}(x_i)^2 f_i - n(\bar{x})^2\right] = \frac{1}{40}\left[164480 - (40)(63.6)^2\right] = 67.04$

Example 7 – Sol.

Class (Interval) Limits		Midpoint (x_i)	Frequency (f_i)
$48 \le x < 56$	$48 - 56$	52	8
$56 \le x < 64$	$56 - 64$	60	14
$64 \le x < 72$	$64 - 72$	68	10
$72 \le x \le 80$	$72 - 80$	76	8
			$n = \sum f_i = 40$



- The Median: For the data in the given table, the **median class** is (56 – 64). So,
- Median $= L_1 + \left[\frac{\left(\frac{n}{2}\right) - n_b}{f_m}\right] C = 56 + \left[\frac{\left(\frac{40}{2}\right) - 8}{14}\right](8) = 62.86$
- The Mode: The **modal class** is (56 – 64). So,
- $Mode = \tilde{\boldsymbol{x}} = L_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] C = 56 + \left[\frac{6}{6+4}\right](8) = 60.8$

Thank You ☺