MBA 651: Quantitative Methods for Decision Making

Raghu Nandan Sengupta
Industrial & Management Engineering Department
Indian Institute of Technology Kanpur



MBA 651: Quantitative Methods for Decision Making

Instructor:	Raghu Nandan Sengupta
Department:	Industrial & Management Engineering Department
Office:	206 (New IME Building)
Contact:	6607 (O)
Email:	raghus@iitk.ac.in



Few rules for class

- Class/group assignments: 30%, Quizzes: 20%, Mid-semester examination: 20%, Final Examination: 25%, Attendance: 05%.
- All quizzes, examinations are closed notes/books, while class/group assignments are open book.
- In case a student is absent during class/group assignment (when given in class), quiz/mid-semester/final examination (when conducted) then it will re-conducted for him/her as per PG norms (with proper reasons/documents/paper being submitted by student).
- No mobiles in class.
- Be exactly on time or do not be in class.
- All notes, assignments, cases, etc., can be found at <http://home.iitk.ac.in/~raghus/>

Syllabus (Applied, to do with Statistics)

- Elementary probability theory
- Conditional probability
- Bayesian concepts
- Discrete and continuous random variables
- Generating functions
- Central Limit Theorem and uses
- Functions of random variables
- Basic Jointly distributed random variables
- Sampling theory and sampling distributions
- Method for statistical inference

Syllabus (Applied, to do with Statistics)

- Theory of point estimation and estimation of parameters
- Theory of interval estimation
- Theory of hypotheses testing
- Descriptive and deductive statistics
- Linear and multiple linear regression
- Analysis of variance
- Introduction to statistical Packages, e.g., MATLAB

Syllabus (Applied, to do with OR)

- Introduction to decision analysis and process.
- Aspects of Modelling and general concepts
- Linear Programming
- Network Flows
- Assignment Problems
- Transhipment Problems
- Sequencing, Scheduling basic concepts
- Non-Linear Programming basic concepts

Text Books

1) Quantitative Analysis for Management: B. Render, R. M. Stair, Jr. , M. E. Hanna and T. N. Badri; Pearson Publication, ISBN: 9788131723739.
2) Decision Analysis & Decision making: S. C. Albright, W. Winston and C. J. Zappe, Thomson Press, ISBN: 9812435433.
3) Mathematical Statistics: J.E. Freund; Prentice Hall of India, ISBN: 8120307844.

Software/Language

1) MATLAB <http://www.mathworks.com/> . One can find sever based MATLAB at https://www.iitk.ac.in/ccnew/
2) R <https://www.r-project.org/>
3) SPSS https://www.spss.co.in/
4) SAS <https://www.sas.com/en_in/home.html>

Examples

1) News paper vendor: wants to maximize profit.
2) Production manager: wants to minimize waiting time of jobs on machines and thus reduce inventory and costs.
3) Marketing manager: wants to recommend the best changes for a product (which is being sold in the market) so that it will do well.

Statistics

The word STATISTICS is derived from the Italian word *stato*, which means "state" and *statista* refers to a person involved with the affairs of the state.

Statistics

- Now a days, STATISTICS (in a plural sense) is the study of qualitative and quantitative data from our surrounding, be it environment or any system so as to draw meaningful conclusions about the environment or system.
- It also means (in the singular sense) the body of methods that are meant for treatment of such data

Main steps in the study of Statistics

- Method of collection of data (primary or secondary)
- Scrutiny of data
- Presentation of data (non frequency data, frequency data)

Main steps in the study of Statistics

- Analysis of data through statistical models/methods
- Conclusions from results thus obtained
- Modification of statistical models/methods depending on results obtained

Descriptive Statistics

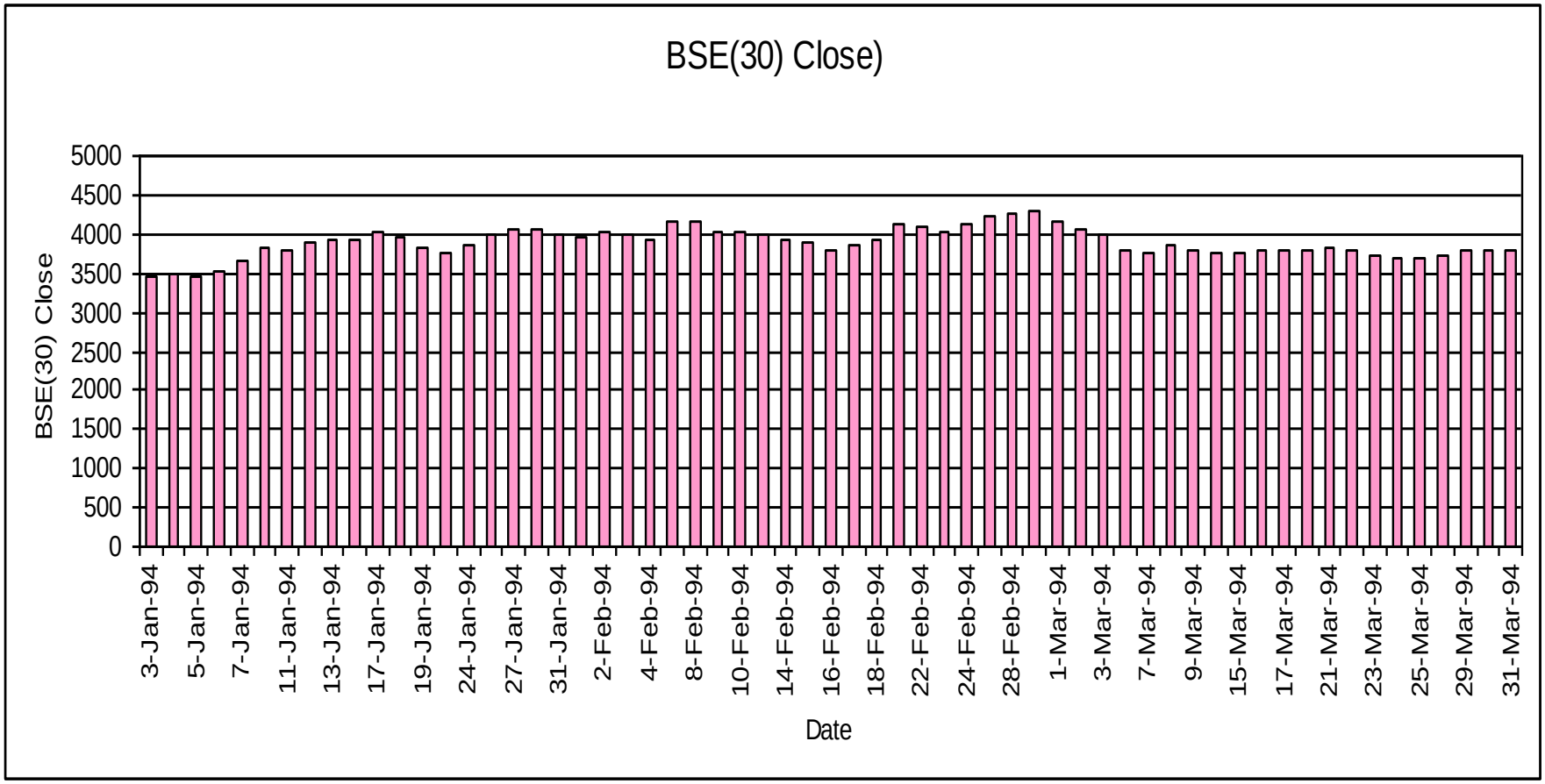
Presentation of data

- Non frequency data
- Frequency data

Non frequency data
Time series or Historical data

Consider the case where the representation of the values of one or more variables like population of India, price of petroleum etc., may be given for different periods of time. For instance we may be interested in knowing the population change over time or the change of production of petroleum over time.

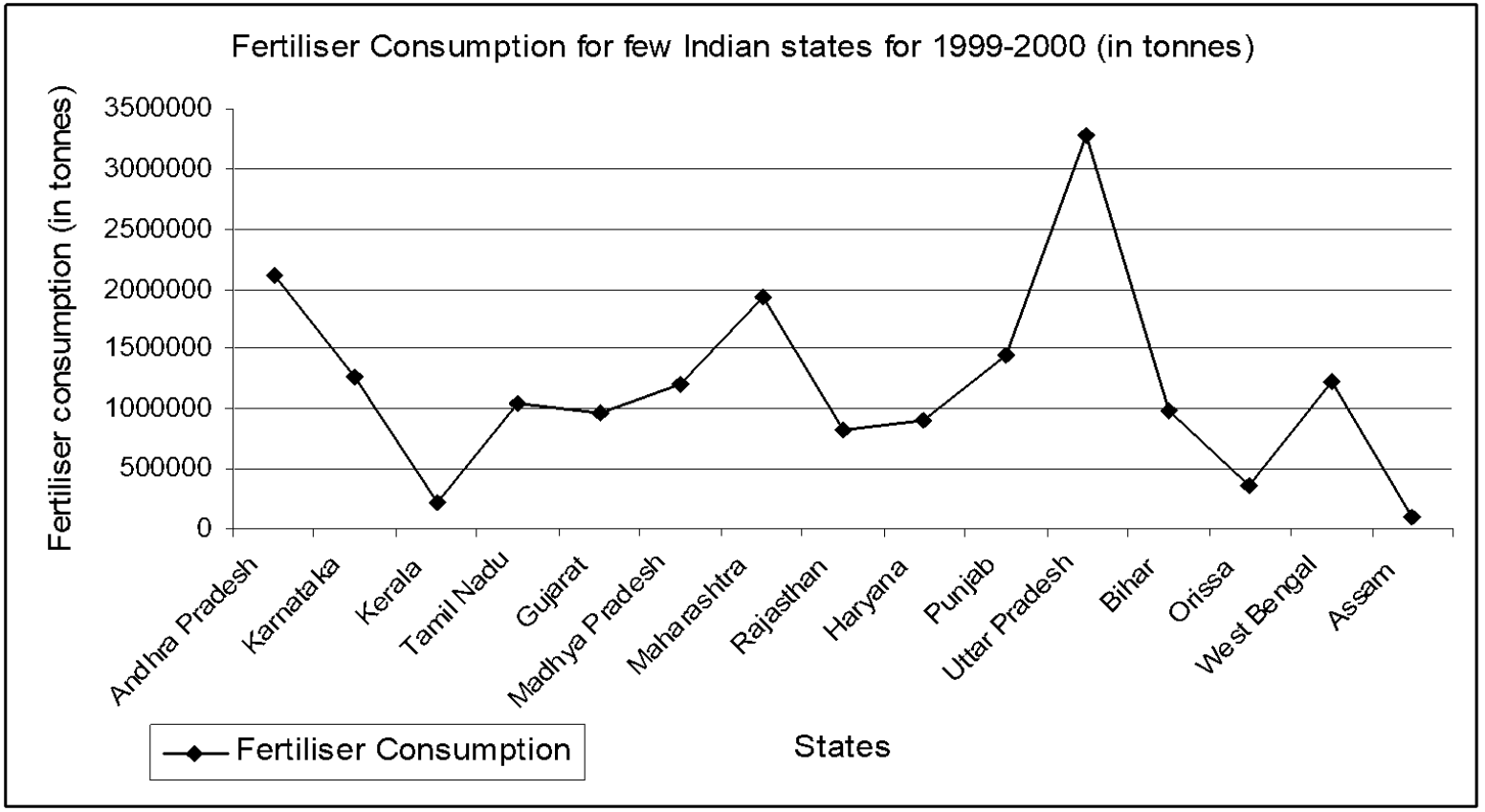
Non frequency data
Time series representation of data



Non frequency data
Spatial series data

It may be that the values of one or more variables are given for different individuals in a group for the same period of time. But instead of considering the group as such we may be more interested in studying the way the values of the variable(s) change from individual to individual in that group.

Non frequency data
Spatial series representation of data



Frequency data

We have the data on one or more variables for different individuals for different periods of time or for different points. But now we are more interested in the characteristic(s) of the group rather than the individuals in that group. In studying the IQ level of students in a school we may be interested in such group characteristics as the percentage of students with IQ higher than 130 or the percentage of students with average IQ less than 90, etc.

Frequency data: Tabular representation

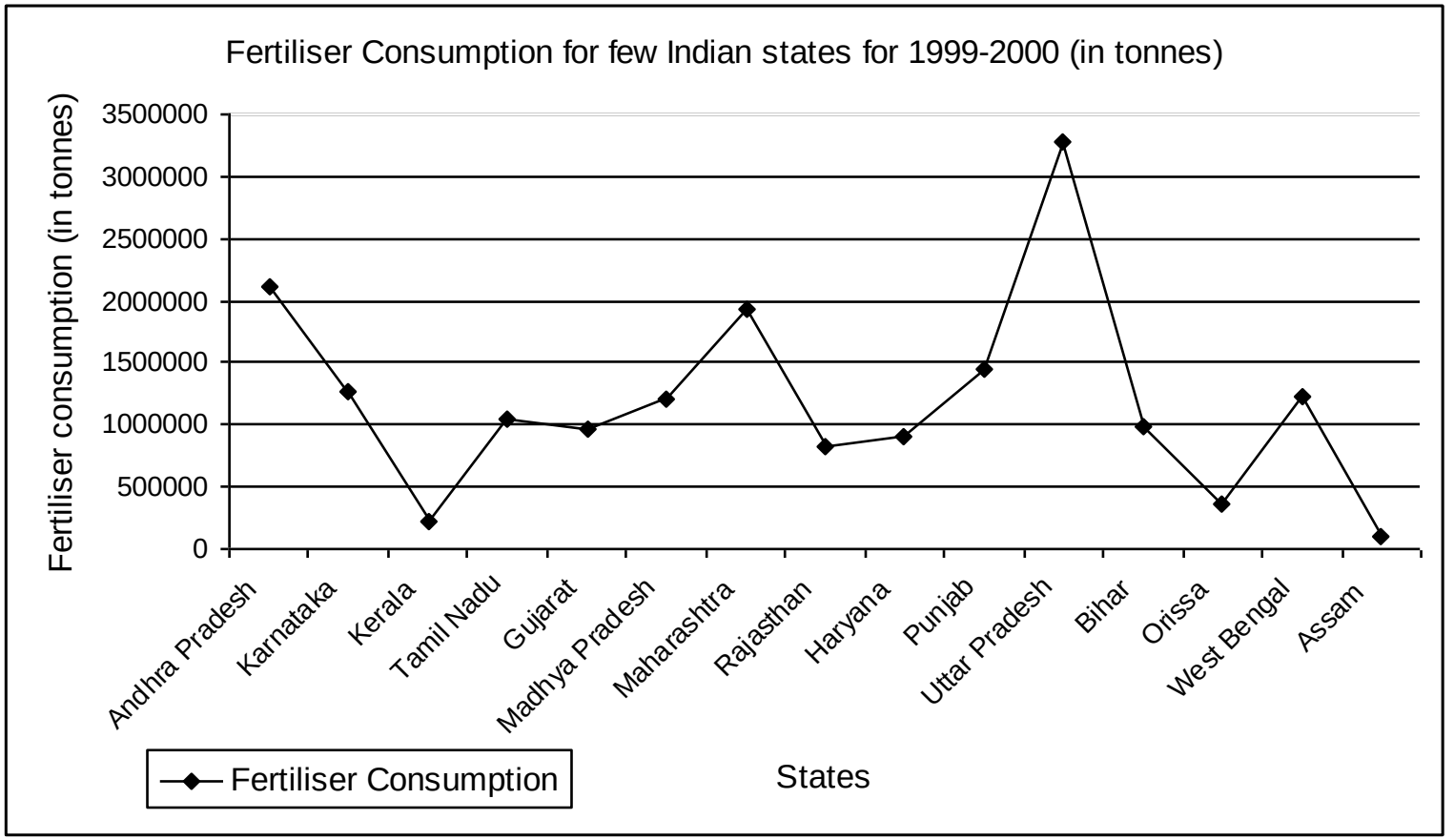
India at a glance	Year			
(% of GDP)	1983	1993	2002	2003
Agriculture	36.6	31.0	22.7	22.2
Industry	25.8	26.3	26.6	26.6
Mfg	16.3	16.1	15.6	15.8
Services	37.6	42.8	50.7	51.2
Pvt Consump	71.8	37.4	65.0	64.9
GOI consump	10.6	11.4	12.5	12.8
Import	8.1	10.0	15.6	16.0
Domes save	17.6	22.5	24.2	22.2
Interests paid	0.4	1.3	0.7	18.3

Note: 2003 refers to 2003-2004; data are preliminary. Gross domestic savings figures are taken directly from India's central statistical organization.

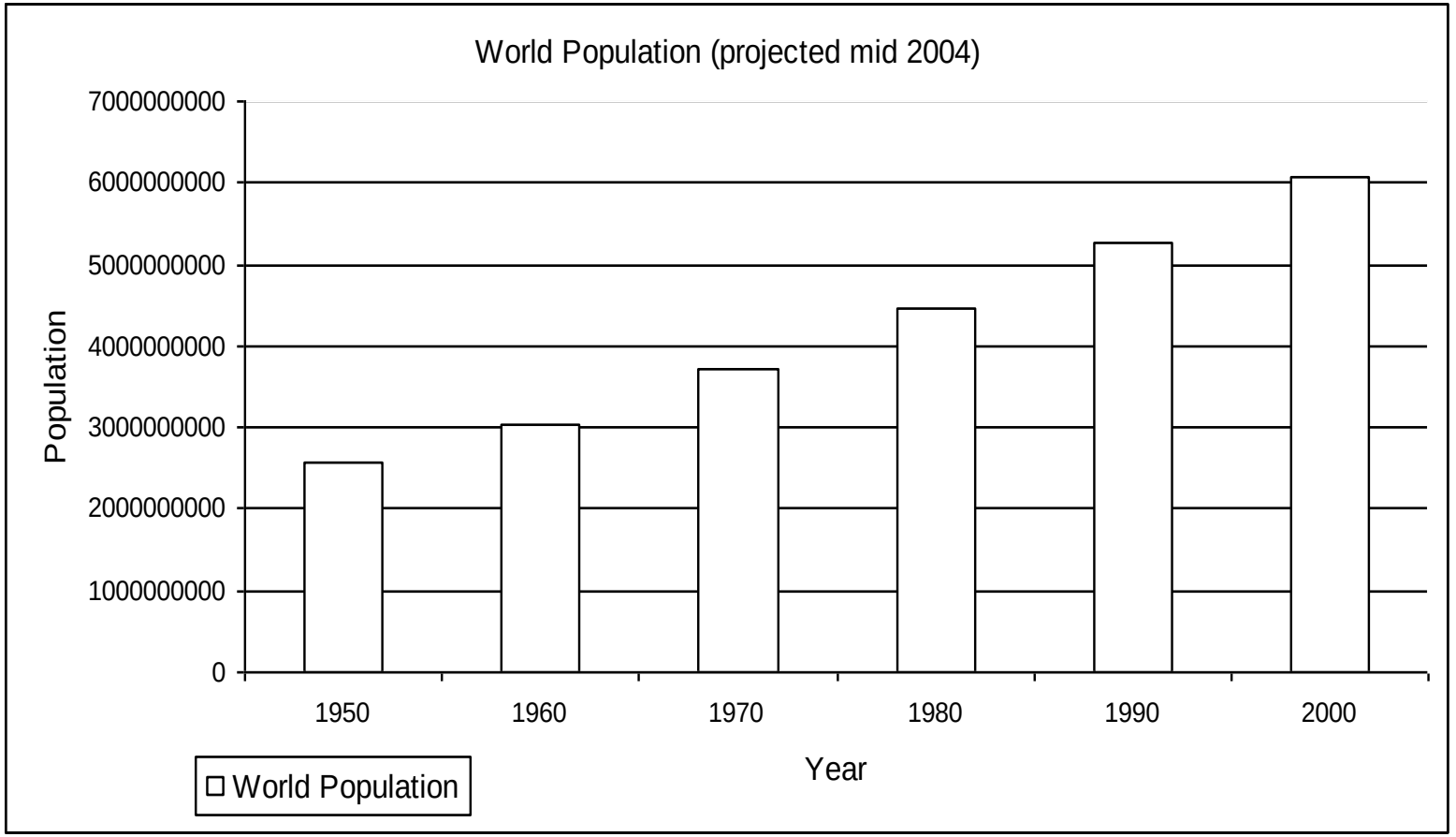
Frequency data
Diagrammatic representation

This is the most commonly used for representing time series. The line diagram (also called histogram) is a graph showing the relationship of the given variable with time. There may be three types of line diagram, for the scales used for both the axes of co-ordinates may be arithmetic (or natural) scales, or one of them may be arithmetic and the other logarithmic, or both may be logarithmic. A line diagram where the vertical scale is logarithmic but the horizontal scale is of the ordinary arithmetic type is called a ratio chart or semi-logarithmic chart. When both the vertical as well as the horizontal axes are logarithmic then the chart is called the doubly-logarithmic chart.

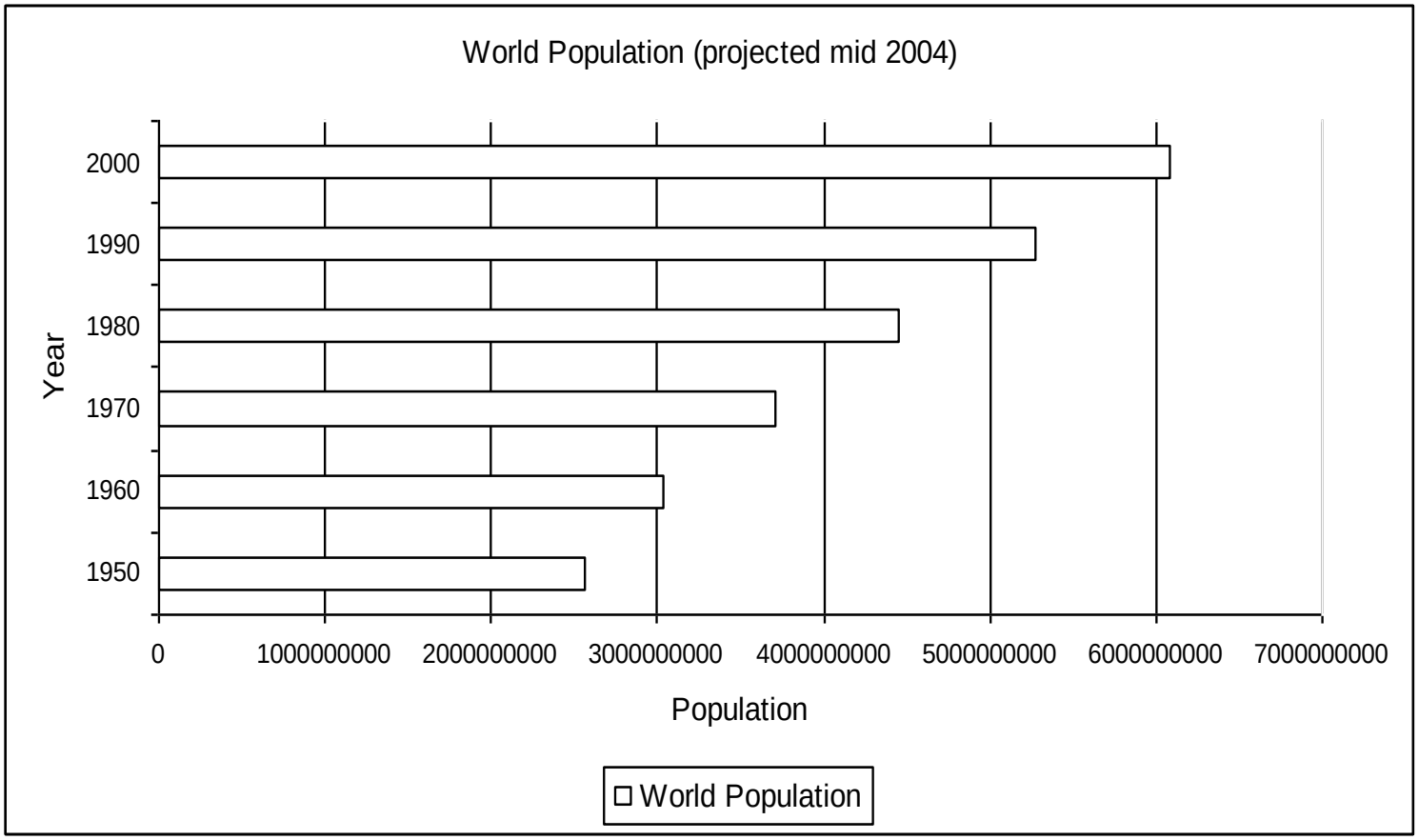
Frequency data
Line diagram representation



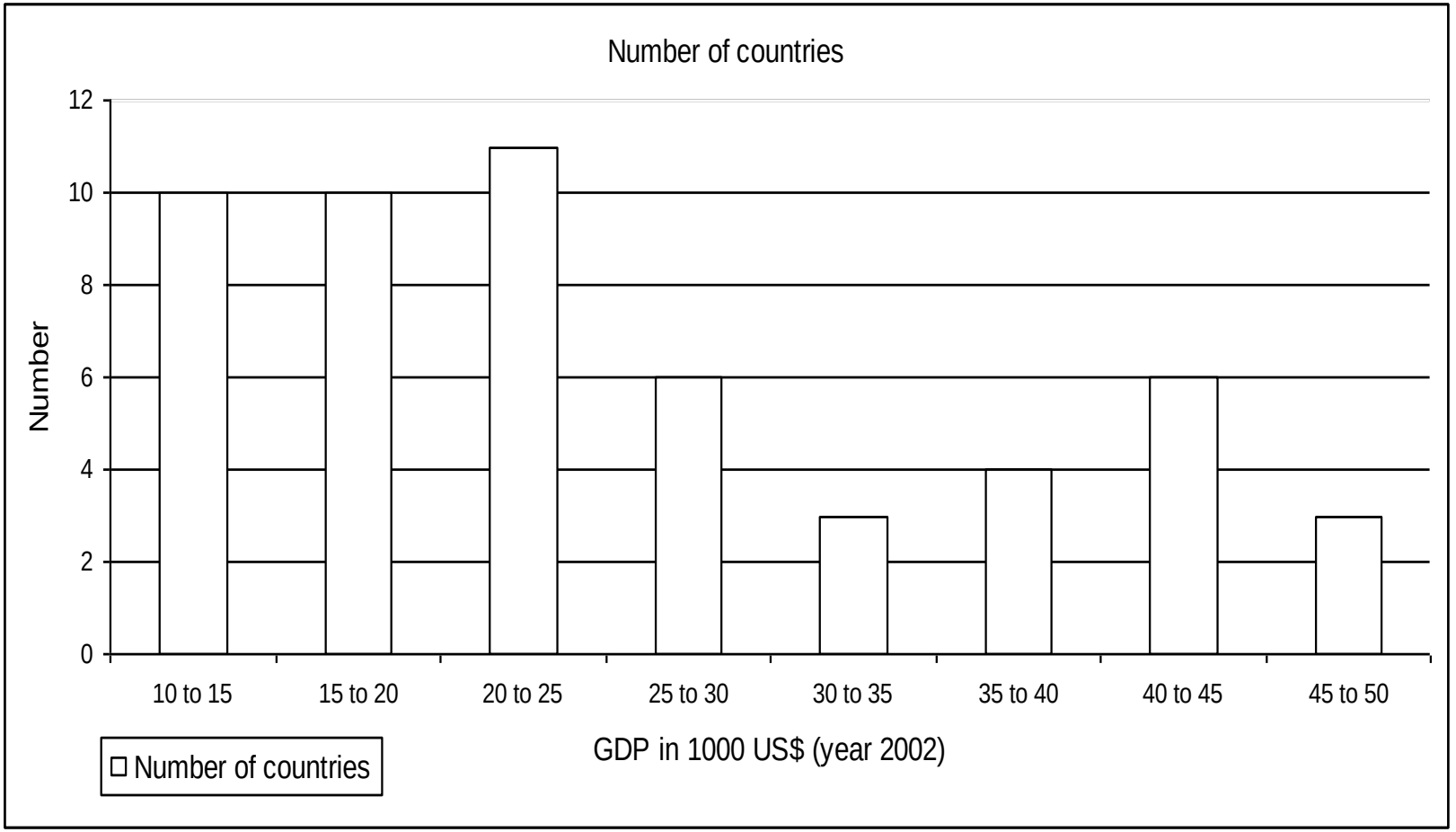
Frequency data
Bar diagram (histogram) representation



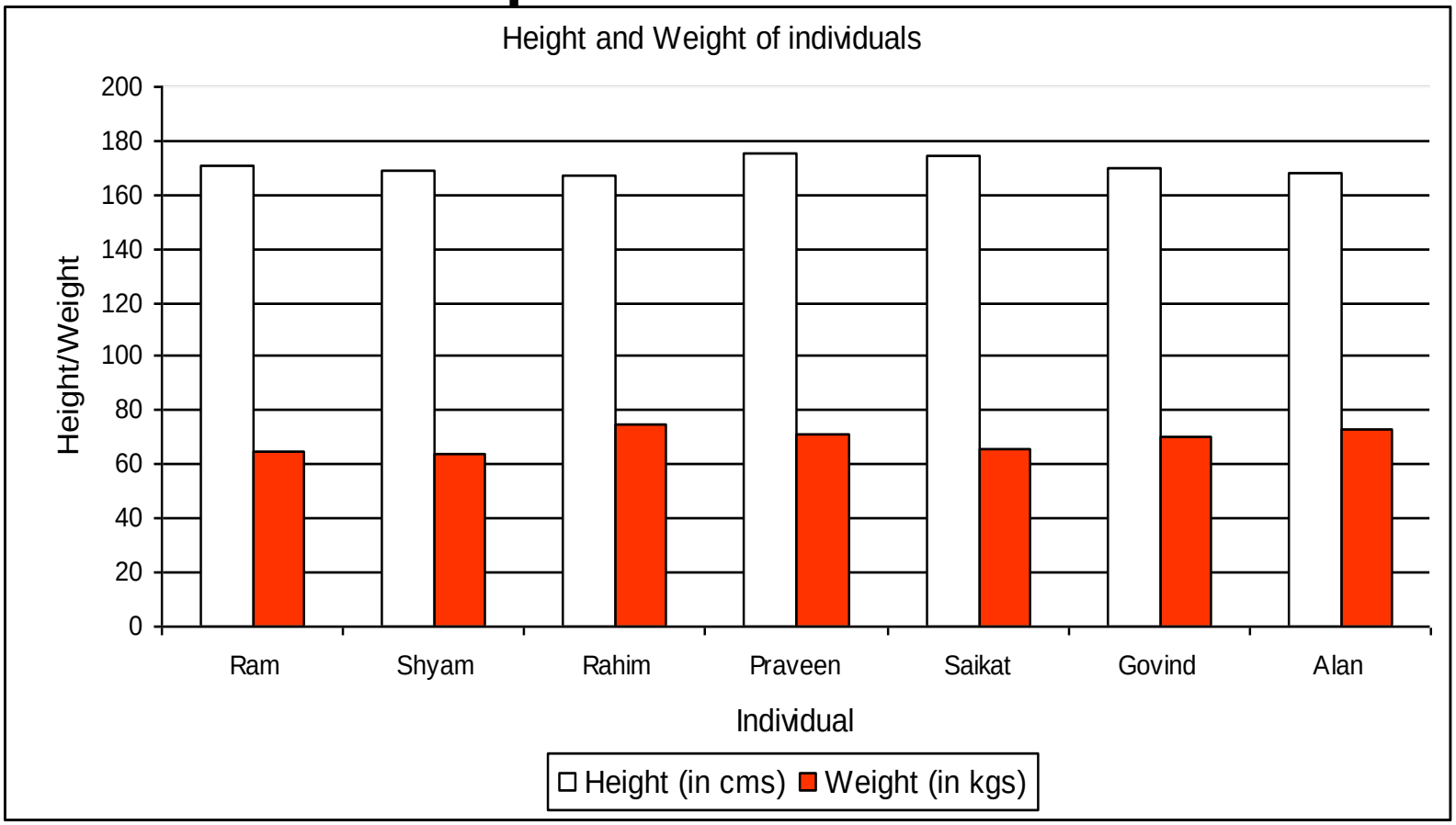
Frequency data
Bar diagram (histogram) representation



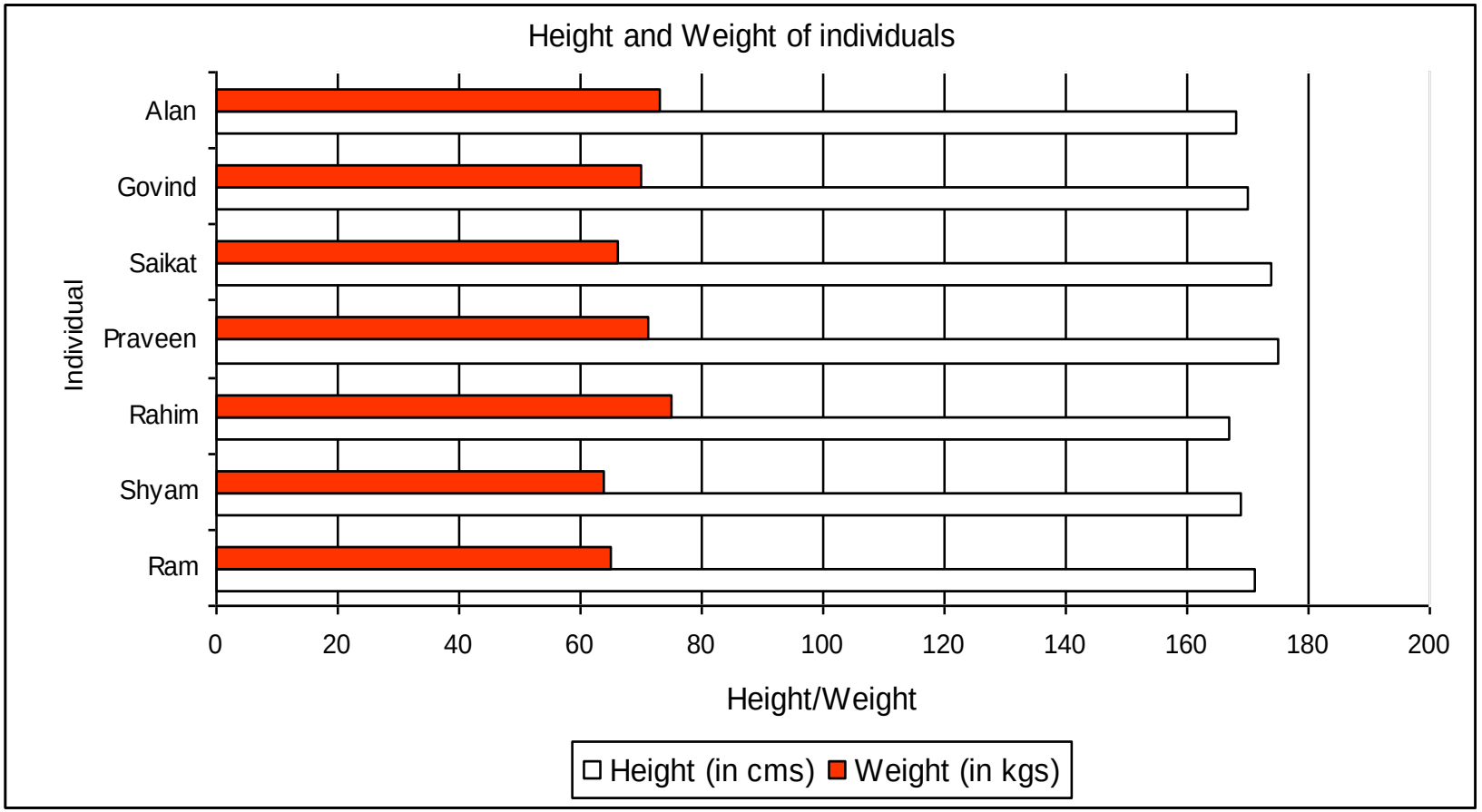
Frequency data
Bar diagram (histogram) representation



Frequency data
Bar diagram (histogram) representation



Frequency data
Bar diagram (histogram) representation



Frequency data
Pictorial diagram representation

In this type of presentation we can represent the data more vividly and in many a cases it is the popular method of representing the data. Here a suitable symbol is first chosen to represent a definite number/quantities of units of the variable. Against each data or observation the symbol is represented proportionally so that we can get the idea of the variable quantities against that data point.

Frequency data
Pictorial diagram representation

Number of universities in the following states in USA

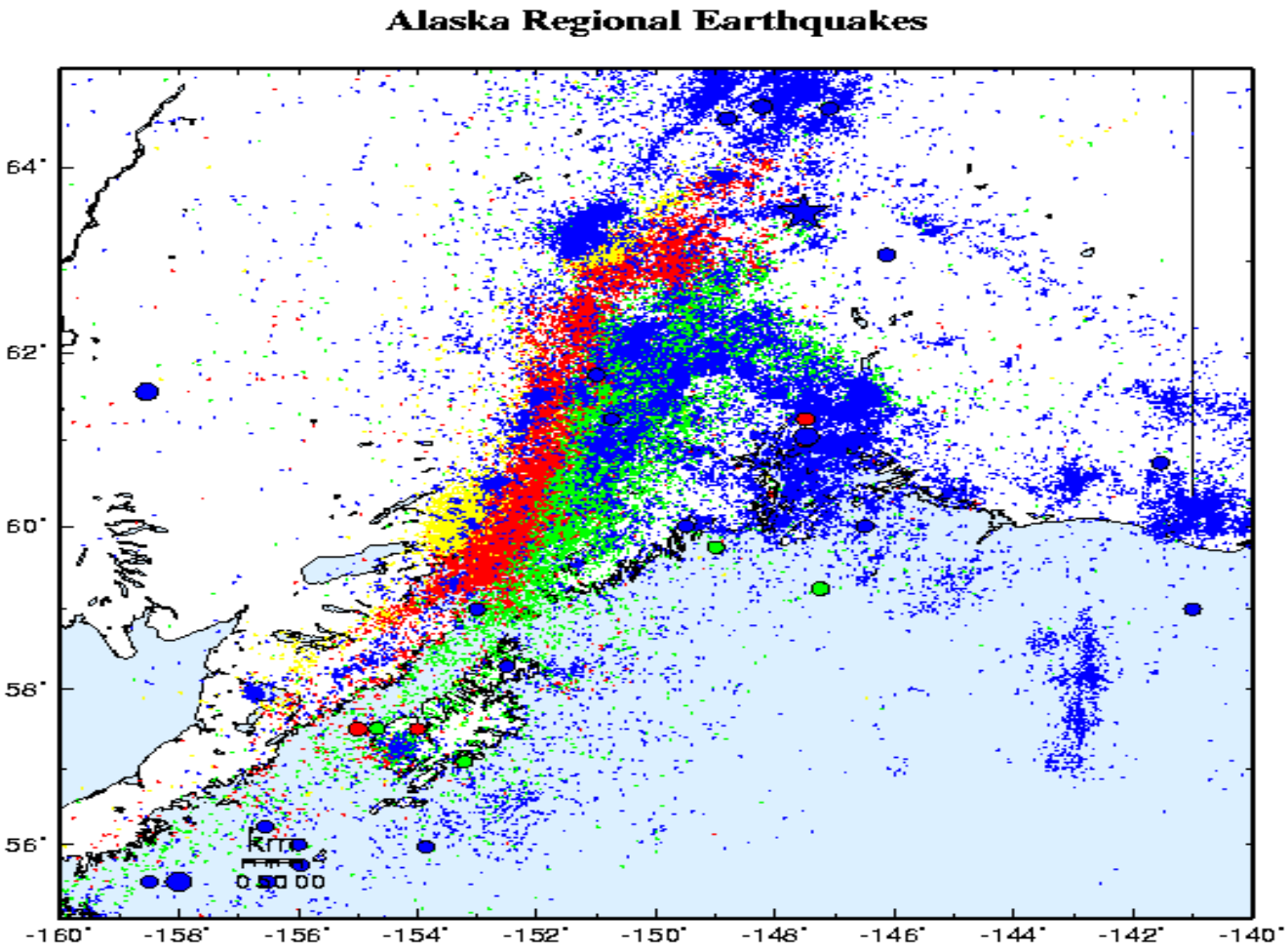
Note: Each of the ∏ symbol represents 5 universities

Alabama	∏ ∏ ∏ ∏ ∏ ∏
Alaska	∏
Arizona	∏ ∏ ∏
Arkansas	∏ ∏ ∏ ∏
Colorado	∏ ∏ ∏ ∏ ∏
Connecticut	∏ ∏ ∏ ∏ ∏
Delaware	∏
Kansas	∏ ∏ ∏ ∏ ∏ ∏

Frequency data
Statistical map representation

If for example we are interested to show diagrammatically regional seismicity in Alaska of earthquakes of all magnitudes reported between 01/01/1960 to 11/09/2002. The colour code as shown indicates the depth of the event. Thus blue: $0 < h \leq 33$ km, green: $33 < h \leq 75$ km, red: $75 < h <= 125$ km and yellow : $h > 125$ km. The larger circles are earthquakes of M 7.0 and higher from 1900-11/09/2002. The colour of circle indicates the depth of the event, as above. The star indicates the location of the 03/11/2002

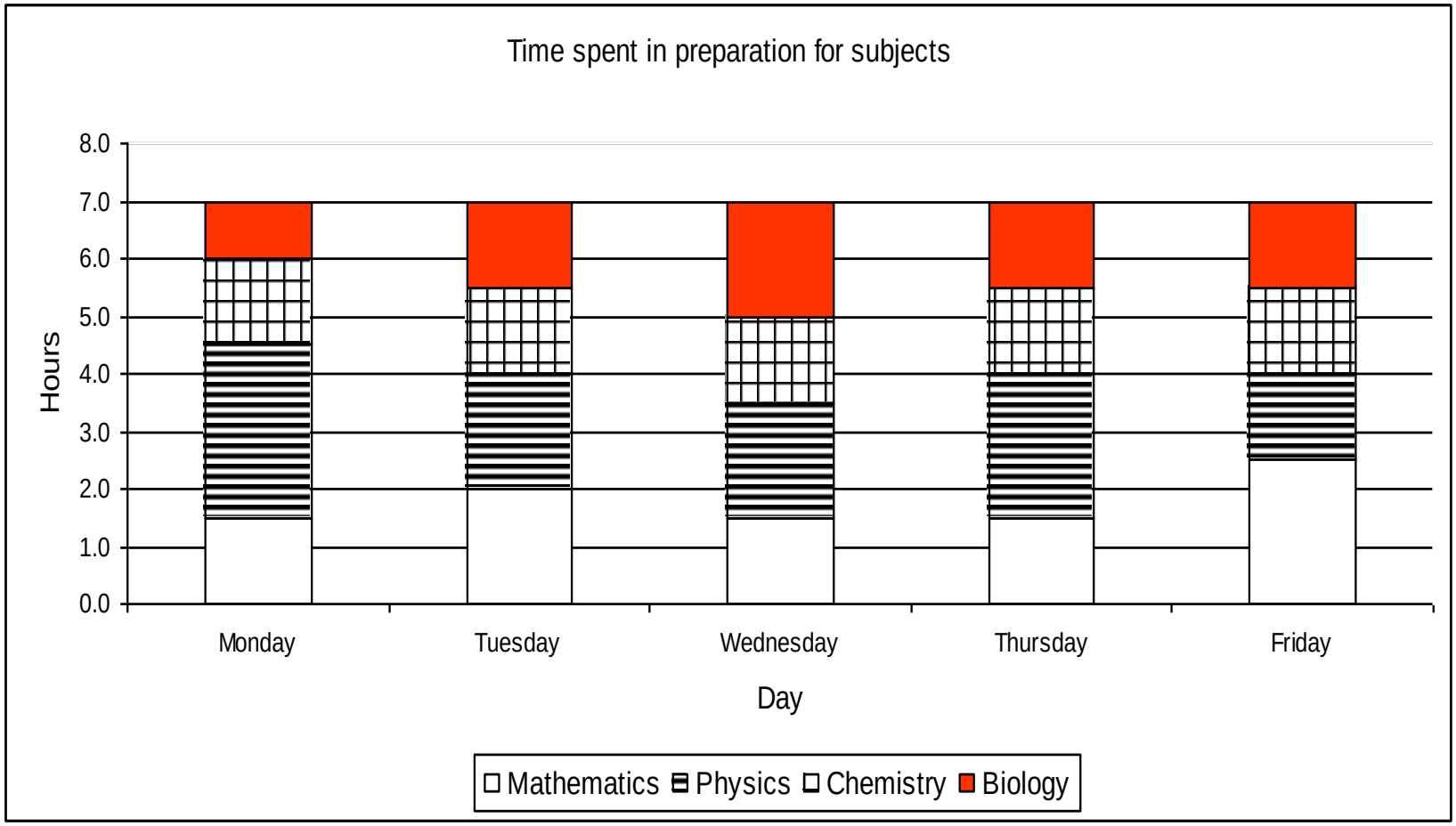
Frequency data
Statistical map representation



Frequency data
Divided bar diagram representation

Consider we have the time spent in hours by student appearing for the CBSE examination for the preparation of Mathematics, Physics, Chemistry and Biology. We collect the student's preparation pattern for a five day period and want to represent the data thus obtained. In that case we would use the divided bar diagram as illustrated below.

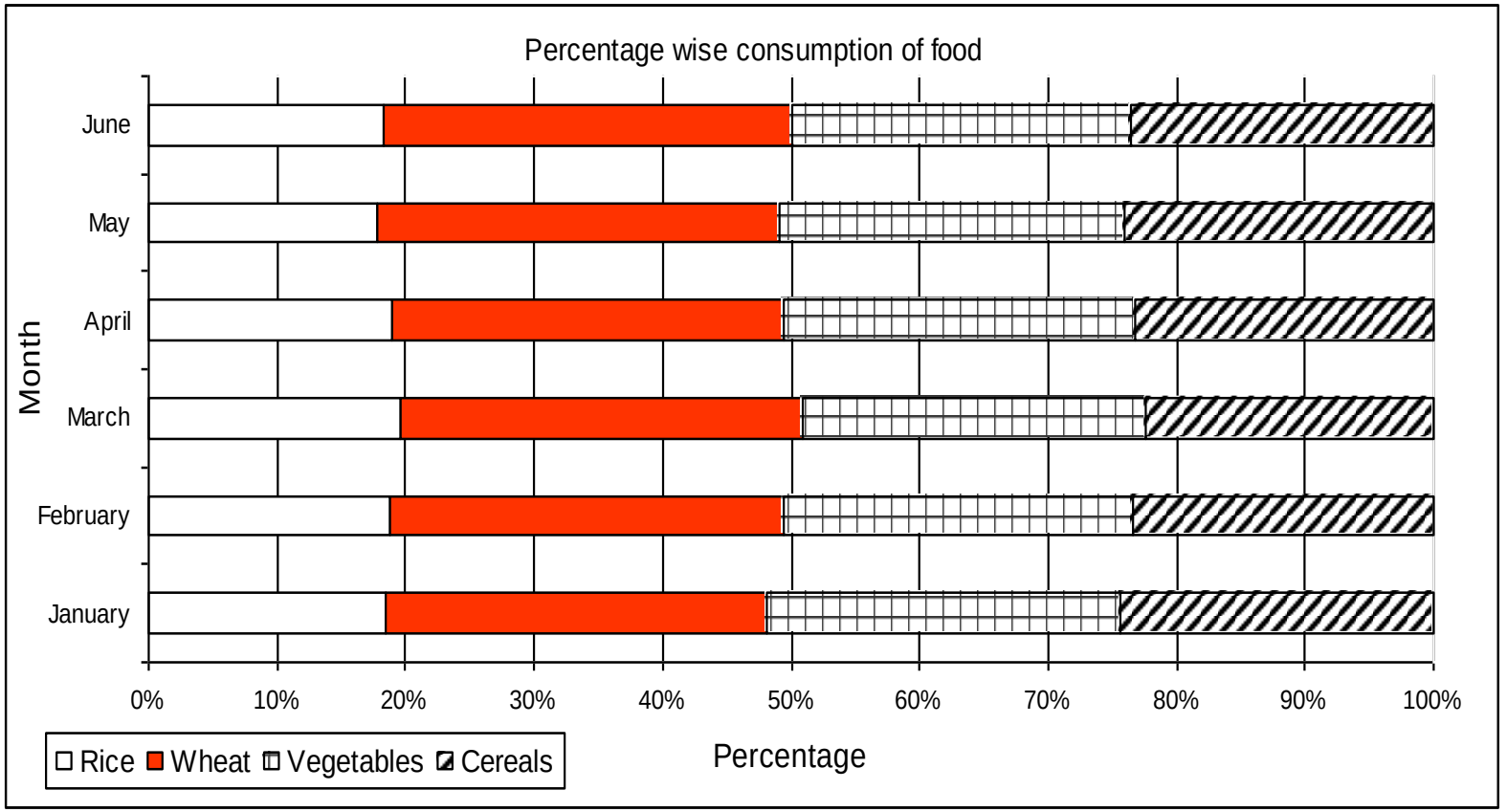
Frequency data
Divided bar diagram representation



Frequency data Stacked column diagram representation

The method of depicting the data is almost similar to the divided bar diagram representation, but here we represent the percentage wise figures for the variables for each data point. Consider we are finding the consumption in rupees for the four main categories of food of a family in the months of January to June. Remembering that the total amount spent for each month can be different, we depict the percentage wise consumption in food for the four categories.

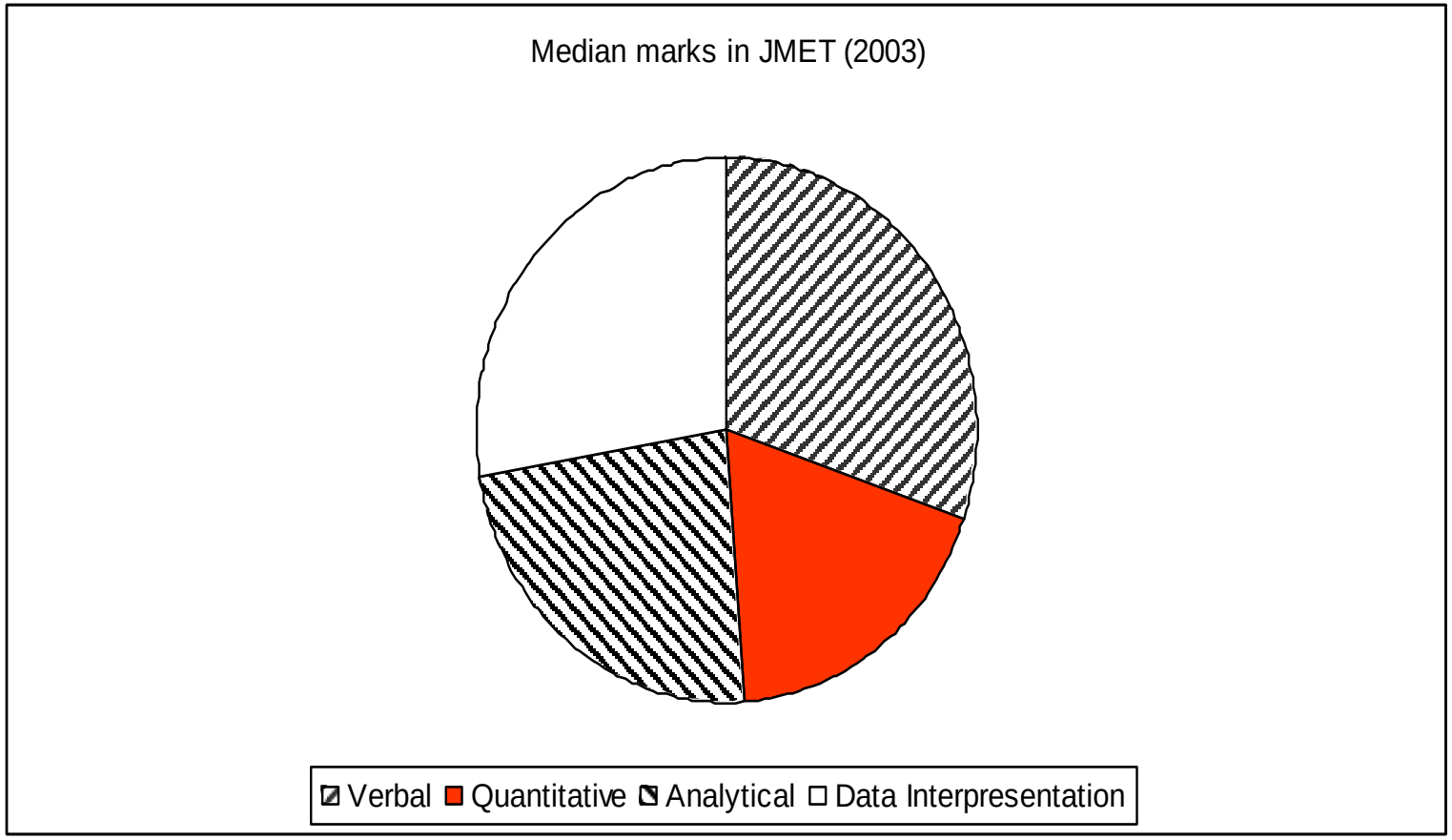
Frequency data Stacked column diagram representation



Frequency data
Pie diagram/chart representation

When the values of a variable are given for a number of categories, as in spatial series, we may be interested in a comparison of the categories or series or the contribution of each category to the total. Here the proportions or percentages of various categories, rather than the absolute values for the categories, will be the principal subject of study

Frequency data
Pie diagram/chart representation



Frequency data
Textual representation

In textual representation of data we depict the information through text.

Consider for the year 2004-2005 we know the number of post graduate students who have registered in different engineering course at IIT Kanpur. The figures are 83 in Aerospace, 88 in Chemical, 139 in Civil, 222 in Electrical, 176 in Mechanical, 115 in Computer Science. Given this data we may be required to utilize this information to answer some queries.

Frequency data
Stem leaf representation

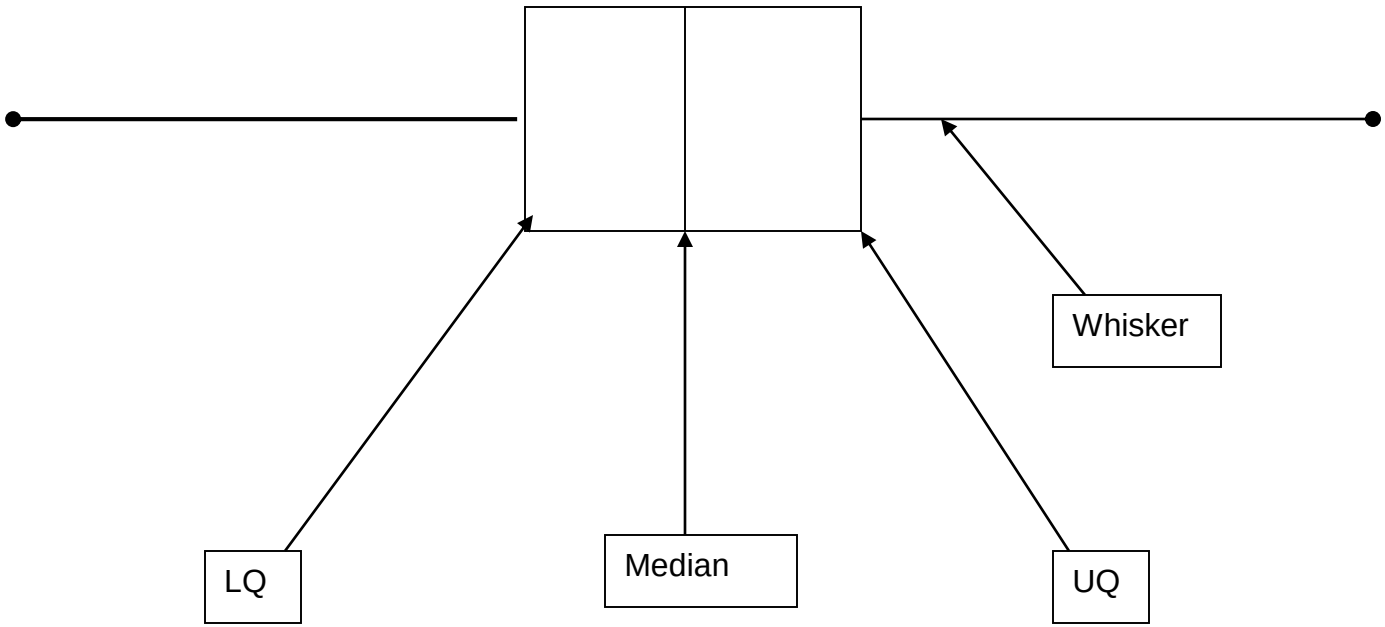
The stem and leaf representation is a quick way of looking at the data set. It contains the information of a histogram but avoids the loss of information in a histogram that results from aggregating the data into intervals. The stem and leaf display is based on the tallying principle but also uses the decimal base of our number system. In the steam and leaf representation, the stem is the number without its rightmost digit (the leaf). The steam is written to the left of a vertical line separating the steam from the leaf. Suppose we have the numbers 105, 106, 107, 109, 100, 108. Then if we use the steam and leaf representation we would depict the numbers as 10 | 567908

Frequency data
Box plot representation

The box plot is also called the box whisker plot. A box plot is a set of five summary measures of distribution of the data which are

- median
- lower quartile
- upper quartile
- smallest observation
- largest observation.

Frequency data
Box plot representation



Frequency data
Box plot representation

Here:

- UQ – LQ = Inter quartile range (IQR)
- X = Smallest observation within 1.5(IQR) of LQ
- Y = Largest observation within 1.5(IQR) of UQ

Definitions

Quantitative variable: It can be described by a number for which arithmetic operations such as averaging make sense.

Qualitative (or categorical) variable: It simply records a qualitative, e.g., good, bad, right, wrong, etc.

As already discusses statistics deals with measurements, some being qualitative others being quantitative. The measurements are the actual numerical values of a variable. Qualitative variables could be described by numbers, although such a description might be arbitrary, e.g., good = 1, bad = 0, right = 1, wrong = 0, etc.

Scales of measurement

Nominal scale: In this scale numbers are used simply as labels for groups or classes. If we are dealing with a data set which consists of colours blue, red, green and yellow, then we can designate blue = 3, red = 4, green = 5 and yellow = 6. We can state that the numbers stand for the category to which a data point belongs. It must be remembered that nothing is sacrosanct regarding the numbering against each category. This scale is used for qualitative data rather than quantitative data.

Scales of measurement

Ordinal Scale: In this scale of measurement, data elements may be ordered according to relative size or quality. For example a customer or a buyer can rank a particular characteristics of a car as good, average, bad and while doing so he/she can assign some numeric value which may be as follows, characteristic good = 10, average = 5 and bad = 0.

Scales of measurement

Interval Scale: For the interval scale we specify intervals in a way so as to note a particular characteristic, which we are measuring and assign that item or data point under a particular interval depending on the data point. Consider we are measuring the age of school going students between classes 5 to 12 in the city of Kanpur. We may form intervals 10-12 years, 12-14 years,....., 18-20 years. Now when we have one data point, i.e., the age of a student we put that data under any one particular interval, e.g. if the student's age is 11 years, we immediately put that under the interval 10-12 years.

Scales of measurement

Ratio Scale: If two measurements are in ratio scale, then we can take ratios of measurements. The ratio scale represents the reading for each recorded data in a way which enables us to take a ratio of the readings in order to depict it either pictorially or in figures. Examples of ratio scale are measurements of weight, height, area, length etc.

Definitions

Population: Consists of the set of all measurements in which the investigator is interested. Example can be all the students in the city of Kanpur. The population is also called the universe. A population is denoted by N

Sample: Is a subset of measurements selected from the population. Sampling from the population is often done randomly, such that every possible sample of n elements will have an equal chance of being selected. A sample selected in this way is called a simple random sample or just a random sample. Example can be the students in Kendriya Vidalaya inside IIT Kanpur campus. A sample is denoted by n

Definitions

- Tally number: By tally number we mean the tally we give depending upon the number of times that particular value of the variable occurs in the total universe or sample.

Definitions

- Frequency (absolute frequency): By frequency (absolute frequency) we mean the number of data points which fall within a given class or for a given value in a frequency distribution. It means that it denotes the number of occurrence or happening of a particular outcome. We denote frequency (absolute frequency) with f_i.

Definitions

Cumulative frequency: The cumulative frequency corresponding to the upper boundary of any class interval or value in a frequency distribution is the total absolute frequency of all values less (greater) than that boundary for the class or value. We denote cumulative frequency less (greater) than type by $F_{\leq n} = \sum_{i \leq n} f_i \; (F_{\geq n} = \sum_{i \geq n} f_i)$

Example

Consider we have the following data related to the size in numbers of thirty families in the city of Jaipur.

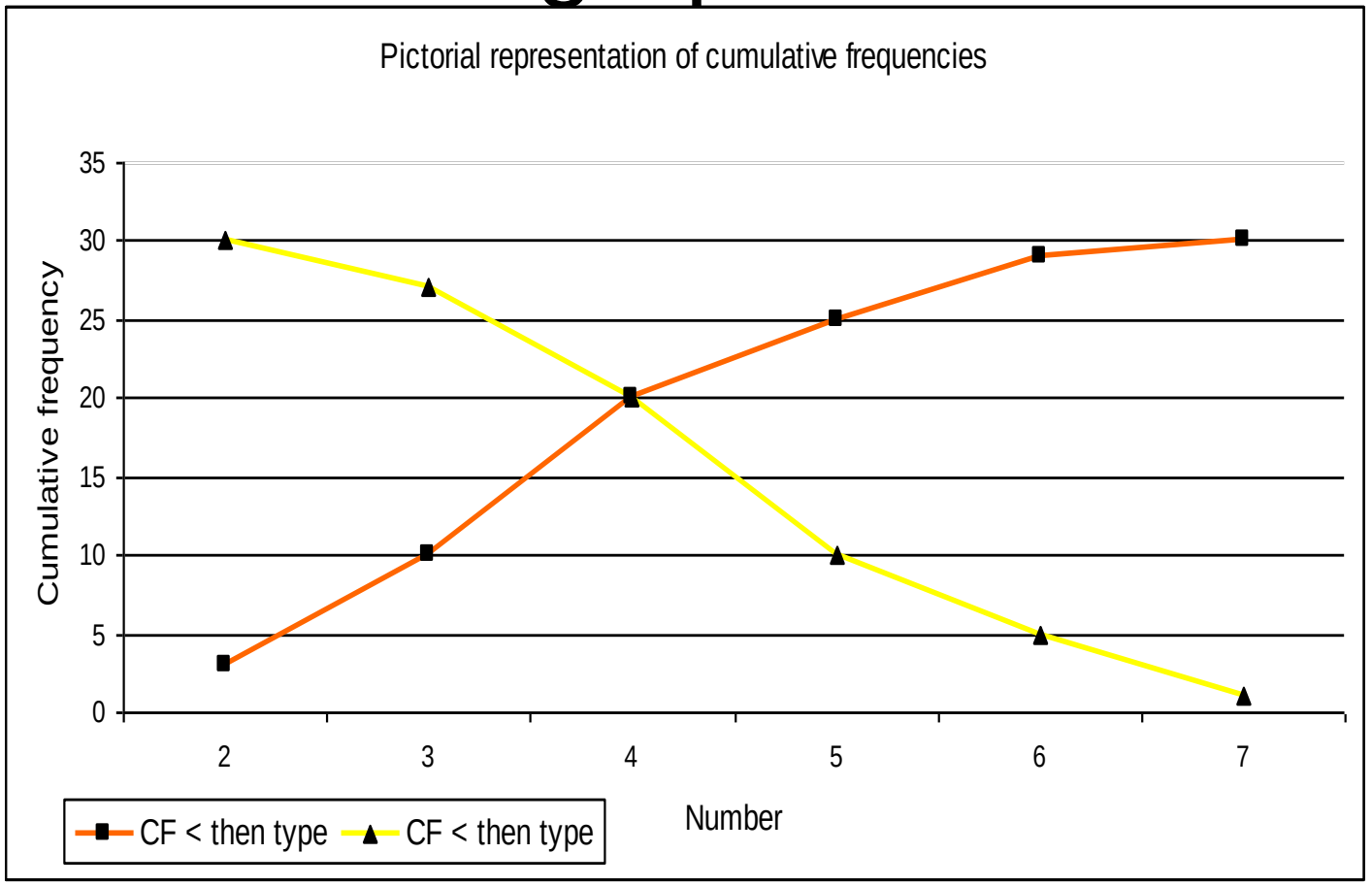
2, 6, 3, 4, 4, 5, 3, 6, 4, 4, 5, 3, 2, 3, 6, 5, 4, 4, 4, 3, 2, 4, 5, 6, 7, 4, 4, 5, 3, 3.

Now the question is how do we represent the data using tally numbers, frequencies, cumulative frequencies?

Example

# of members	Tally #	fi	$F_{\leq n} = \sum_{i \leq n} f_i$	$F_{\geq n} = \sum_{i \geq n} f_i$
2	𝍸	3	3	30
3	𝍸 ||	7	10	27
4	𝍸 𝍸	10	20	20
5	𝍸	5	25	10
6	||||	4	29	5
7	|	1	30	1

Example: cumulative frequency graphs



Important note

A cumulative frequency diagram is called the ogive. The abscissa of the point of intersection represents the median of the data.

Guidelines for depicting frequency tables

1) The classes should be mutually exclusive and exhaustive
2) The number of classes should be neither too small not too large
3) As far as practicable, the classes should be of equal lengths

Example

The following data relates to the height in cms of forty individuals.

160.1, 167.2, 181.3, 154.7, 172.3, 161.3, 182.4, 158.2, 167.3, 159.4, 150.1, 157.3, 152.8, 155.8, 146.0, 162.0, 147.9, 149.9, 173.4, 166.4, 182.3, 151.2, 168.3, 170.1, 187.6, 163.4, 183.3, 171.9, 179.4, 166.8, 179.2, 168.3, 165.2, 166.7, 165.1, 166.3, 166.3, 173.4, 164.2, 164.9

1) We are required to prepare a frequency distribution table showing the frequencies and the cumulative frequencies.
2) We are required to draw a histogram to exhibit the frequency distribution graphically.
3) We are required to draw the two ogives.

Example

Frequency table for the heights

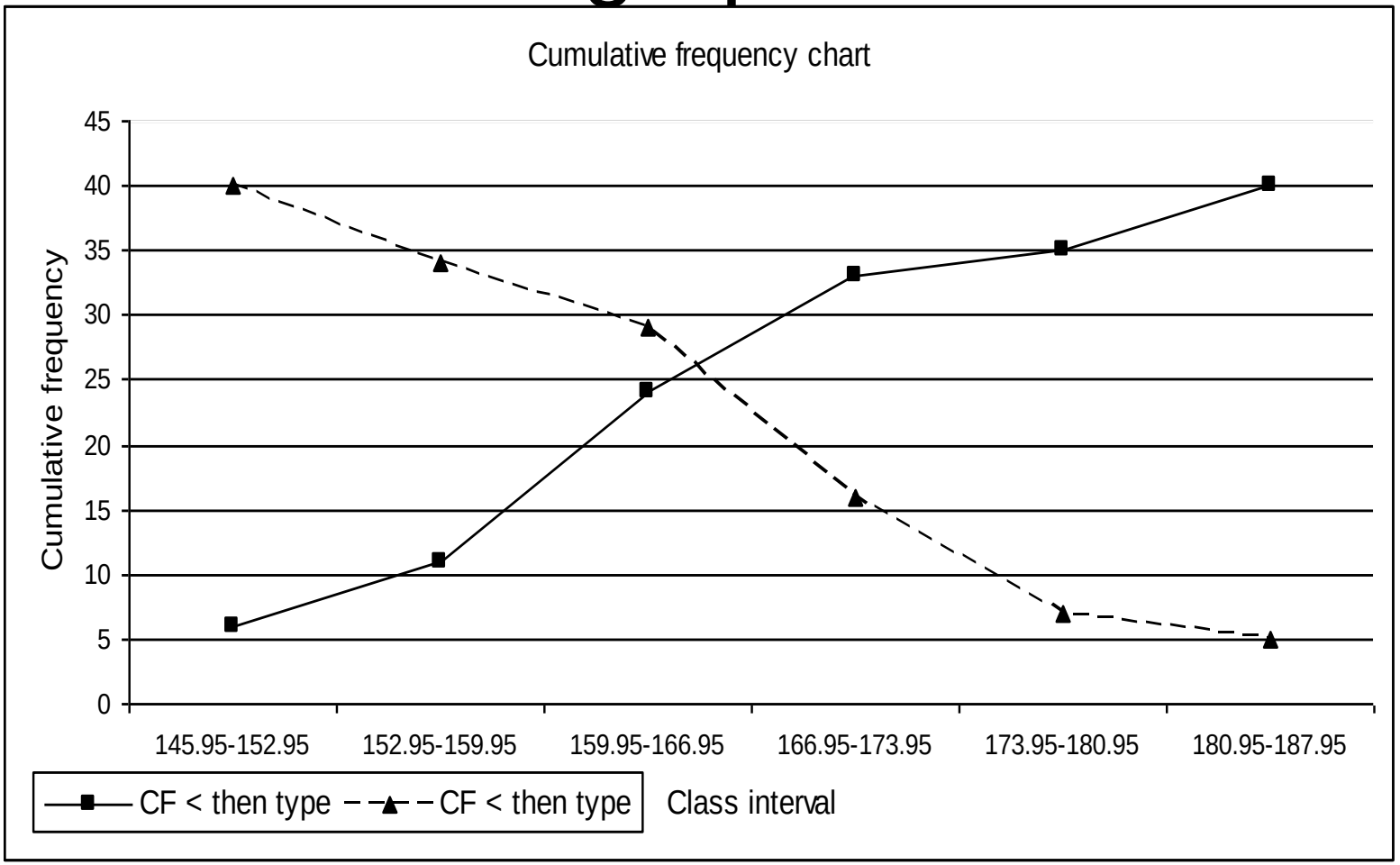
Class Interval	f_i	CF(< then)	CF(> then)
145.95-152.95	6	6	40
152.95-159.95	5	11	34
159.95-166.95	13	24	29
166.95-173.95	9	33	16
173.95-180.95	2	35	7
180.95-187.95	5	40	5

Example

To draw the cumulative frequency less (greater) than type, first specify the class intervals. Then depict the class intervals along the horizontal axis and the cumulative frequency less (greater) than type along the vertical axis.

Against each class interval mark the point by the corresponding cumulative frequency less (greater) than type value. Join the points (the values) depicting the cumulative frequencies less (greater) than type with straight lines, which will give us the respective ogives

Example: cumulative frequency graphs



Definitions

1) Class limit: The end points of any class
2) Class boundaries: The end points of any class interval
3) Relative frequency: It is the ratio of $\frac{f_i}{F}$

Note: Using this concept of relative frequency and cumulative relative frequency less (greater) than type we can also draw the cumulative relative frequencies curves

Definitions: different measures

1) Measure of central tendency
   - Mean (Arithmetic mean (AM), Geometric mean (GM), Harmonic mean (HM))
   - Median
   - Mode
2) Measure of dispersion
   - Variance or Standard deviation
   - Skewness
   - Kurtosis

Definition: Mean

Given N number of observations, x_1, x_2,……, x_n we define the following

$$AM = \frac{1}{n}[X_1 + X_2 + ...+ X_n]$$

$$GM = [X_1 * X_2 * ..... * X_n]^{1/n}$$

$$HM = [\frac{1}{n}(\frac{1}{X_1} + \frac{1}{X_2} + ...+ \frac{1}{X_n})]^{-1}$$

Arithmetic Mean (μ)

When estimating the long-term expectation of a random variable, the arithmetic mean is a natural choice, e.g. finding the average age of a group of persons, average income of a group of people etc.

Use of Harmonic Mean

Consider a car travels a distance x with a velocity of v_1 and returns back the same distance with a velocity of v_2. What is the average velocity?

$$v = \frac{2x}{\frac{x}{v_1} + \frac{x}{v_2}}$$

Hence we see that in this case we use the HM and not AM.

Uses of Harmonic Mean

Application areas are: (1) design stream flow estimation for waste load allocation (2) estimation of effective petrochemical and geophysical properties of a heterogeneous system of porous media.

Use of Geometric Mean

Suppose you have an investment which earns 10% the first year, 50% the second year, and 30% the third year and we are interested in finding an equivalent average rate of return, say r:

Now we have

$P(1+0.1)(1+0.5)(1+0.3) = P(1+r)^3$

Hence r = 28.97%

Definition: Median and Mode

- Median(μ_e) : The median of a data set is the value below which lies half of the data points. To find the median we use $F(\mu_e) = 0.5$.
- Mode (μ_o): The mode of a data set is the value that occurs most frequently. Hence $f(\mu_o) \geq f(x); \forall\ x$.

Definition: Variance, Standard deviation, Skewness, Kurtosis

- Variance: V[X] = $\sigma^2 = E\left[X - E(X)\right]^2$
- Standard deviation (SD) = σ
- Skewness = $\gamma_1 = \sqrt{\beta} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3}$
- Kurtosis = $\gamma_2 = \beta_2 - 3 = \left[\frac{\mu_4}{\sigma^4} - 3\right]$

Example

Consider we have the following data points:

5, 7, 10, 7, 10, 11, 3, 5, 5

For these data points we have

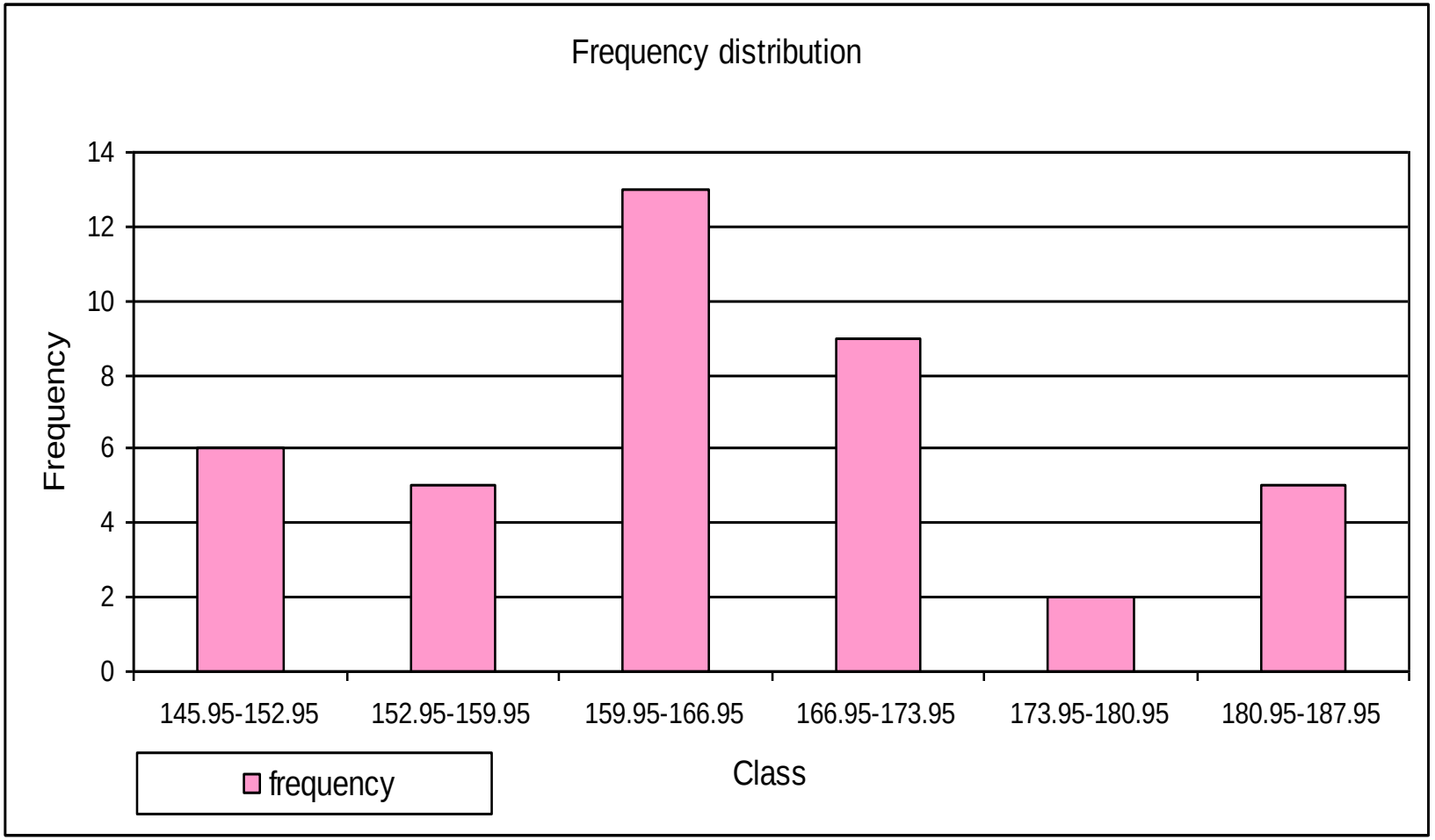
$\mu = 7$; $\mu_e = 10$; $\mu_o = 5$; $\sigma^2 = 6.89$

Descriptive statistics

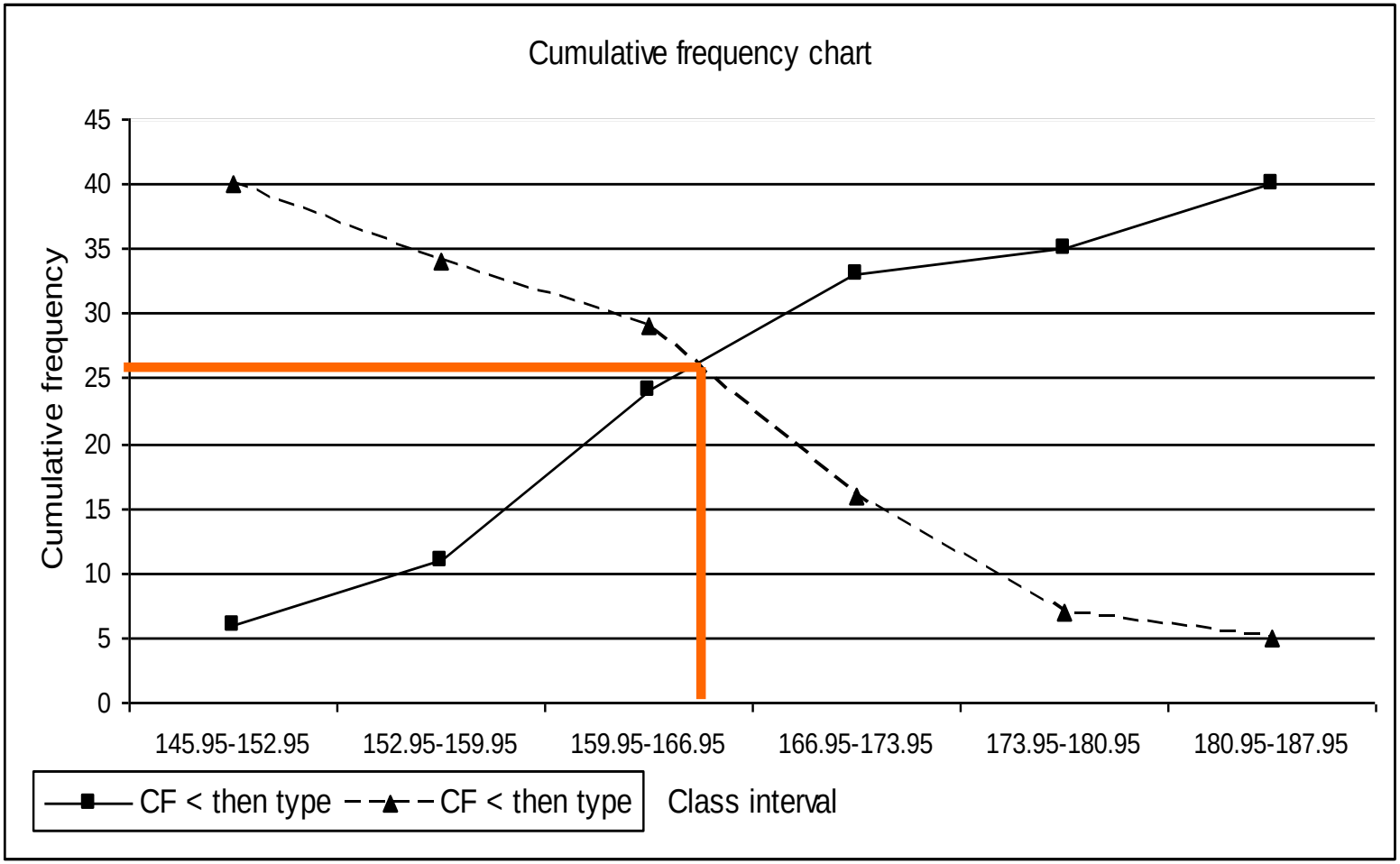
Suppose the data are available in the form of a frequency distribution. Assume there are k classes and the mid-points of the corresponding class intervals being x_1, x_2,…., x_k. While the corresponding frequencies are f_1, f_2,….., f_k, such that $n = f_1+f_2+\ldots..+f_k$

Then: $\mu = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$ $\sigma = \{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i\}^{1/2}$

Height in cms of forty individuals.



Height in cms of forty individuals



Descriptive statistics

Consider m groups of observations with respective means μ_1, μ_2,….., μ_m and standard deviations σ_1, σ_2,….., σ_m. Let the group sizes be n_1, n_2,….., n_m such that $n = n_1 + n_2 + \ldots + n_m$.

Then:

$$\mu_{OVERALL} = \frac{1}{n}\sum_{i=1}^{m} \mu_i f_i$$

$$\sigma_{OVERALL} = [\frac{1}{n}\{\sum_{i=1}^{m} n_i \sigma_i^2 + \sum_{i=1}^{m} n_i (\mu_i - \mu_{OVERALL})^2\}]^{1/2}$$

Example

In a batch of 10 children the IQ of the dull boy is 36 below the average IQ of the other children. Shown that the standard deviation of IQ for all the children cannot be less than 10.8. If the standard deviation is actually 11.4, determine what is the standard deviation when the dull boy is left out.

Example

Take k =2, such that $n_1 = 1$ and $n_2 = 9$.

It is given that $\mu_1 - \mu_2 = 36$ and we also know that $\sigma_1 = 0$.

Hence: $\sigma_{OVERALL} = [0.9\sigma_2^2 + 116.64]^{1/2}$

- From above we have $\sigma_{OVERALL} > 10.8$
- If $\sigma_{OVERALL} = 11.4$, we have $\sigma_2 = 3.9$.

Random event

Random experiment: Is an experiment whose outcome cannot be predicted with certainty.

- Sample space (Ω) : The set of all possible outcomes of a random experiment
- Sample point (ω_i): The elements of the sample space
- Event (A): Is a subset of the sample space such that it is a collection of sample point(s).

Probability

Probability (P(A)): Of an event is defined as a quantitative measure of uncertainty of the occurrence of the event

- Objective probability: Based on game of chance and which can be mathematically proved or verified . If the experiment is the same for two different persons, then the value of objective probability would remain the same. It is the limiting definition of relative frequency. Example: be probability of getting the number 5 when we roll a fair die.
- Subjective probability: Based on personal judgment, intuition and subjective criteria. Its value will change from person to person. Example one person sees the chance of India winning the test series with Australia high while the other person sees it to be low.

Random event

For a random experiment, we denote

$P(\omega_i) = p_i$

$$P(A) = \sum_{\varpi_i \in A} p_i$$

Where:

- $P(\omega_i) = p_i$ = Probability of occurrence of the sample point ω_i
- P(A) = Probability of occurrence of the event
- $$P(\Omega) = \sum_{\varpi_i \in \Omega} p_i = 1$$

Example 1

Suppose there are two dice each with faces 1, 2,....., 6 and they are rolled simultaneously. This rolling of the two dice would constitute our random experiment

Then we have:

- Ω = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1),……, (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}.
- ω_i = (1,1), (1,2),…., (6,5), (6,6)
- We define the event is such that the outcomes for each die are equal in one simultaneous throw, then A = {(1, 1), (2, 2),….., (6, 6)}
- P(ω_i): $p_1 = p_2 = \ldots\ldots = p_{36} = 1/36$
- $P(A) = p_1 + p_8 + p_{15} + p_{22} + p_{29} + p_{36} = 6/36$

Example 2

Suppose a coin is tossed repeatedly till the first head is obtained.

Then we have:

- Ω = {(H), (T,H), (T,T,H),……}
- ω_i = (H), (T,H), (T,T,H),…..
- We define the event such that at most 3 tosses are needed to obtain the first head, then A = {(H), (T,H), (T,T,H)}
- P(ω_i): $p_1 = ½$, $p_2 = (½)^2$, $p_3 = (½)^3$, $p_4 = (½)^4$,…..
- P(A) = $p_1 + p_2 + p_3 = 7/8$

Classical definition of probability

Under this definition we consider the following:

- Sample space is finite.
- All the sample points are equally likely, i.e., they have equal probability of occurrence or equal relative frequency of occurrence.

Example 3

In a club there 10 members of whom5 are Asians and the rest are Americans. A committee of 3 members has to be formed and these members are to be chosen randomly. Find the probability that there will be at least 1 Asian and at least 1 American in the committee

Total number of cases = $^{10}C_2$ and the number of cases favouring the formation of the committee is $^5C_2 * ^5C_1 + ^5C_1 * ^5C_2$

Hence P(A) = 100/120

Exercise

There is box containing 10 coloured balls out of which 4 are red, 4 are blue and 2 are white.

1) If you draw a ball at random what is the probability that it is a white ball?
2) If you draw two balls consecutively and with replacement then what is the probability that both the balls are red?
3) If you draw two balls consecutively and without replacement, then what is the probability that the first is blue and the second is white?

Axiomatic definition of probability

Under this definition we consider the following:

- Sample space is infinite.
- Sample points are not equally likely.

Example 4

Suppose we continue with example 2 which we have just discussed and we define the event B , that al least 5 tosses are needed to produce the first head

- Ω = {(H), (T,H), (T,T,H),…….}
- ω_i = (H), (T,H), (T,T,H),…..
- P(ω_i): $p_1 = ½$, $p_2 = (½)^2$, $p_3 = (½)^3$, $p_4 = (½)^4$,…..
- P(B) = $p_5+p_6+p_7+$ ….. = $1 – (p_1+p_2+p_3+p_4)$

Exercise

Two fair coins are tossed simultaneously and such simultaneously tossing is repeated till you get two tails together.

What is the probability that you need at most two such simultaneous tossing to achieve our objective?

Theorem in probability

For any event A, B $\in \Omega$

- $0 \leq P(A) \leq 1$
- If $A \subset B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A^c) = 1 - P(A)$
- $P(\Omega) = 1$
- $P(\phi) = 0$

Definitions

- Mutually exclusive: Consider n events A_1, A_2,….., A_n. They are mutually exclusive if no two of them can occur together, i.e., $P(A_i \cap A_j) = 0.\ \forall\ i, j\ (i \neq j) \in n$
- Mutually exhaustive: Consider n events A_1, A_2,….., A_n. They are mutually exhaustive if at least one of them must occur and $P(A_1 U A_2 U \ldots . U A_n) = 1$

Example 5

Suppose a fair die with faces 1, 2,….., 6 is rolled. Then Ω = {1, 2, 3, 4, 5, 6}. Let us define the events A_1 = {1, 2}, A_2 = {3, 4, 5, 6} and A_3 = {3, 5}

- The events A_2 and A_3 are neither mutually exclusive nor exhaustive
- A_1 and A_3 are mutually exclusive but not exhaustive
- A_1, A_2 and A_3 are not mutually exclusive but are exhaustive
- A_1 and A_2 are mutually exclusive and exhaustive

Conditional probability

Let A and B be two events such that P(B) > 0. Then the conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Assume Ω = {1, 2, 3, 4, 5, 6}, A = {2}, B = {2, 4, 6}. Then A $\cap$ B = {2} and

$$P(A \mid B) = \frac{1/6}{3/6} = \frac{1}{3} \qquad P(B \mid A) = \frac{1/6}{1/6} = 1$$

Baye's Theorem

Let $B_1, B_2, \ldots, B_n$ be mutually exclusive and exhaustive events such that $P(B_i) > 0$, for every i =1, 2,…., n and A be any event such that $P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$ then we have

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}$$

Example 6

In an examination each question has four alternatives, answer of which only one is correct. If a student knows the correct alternative then he/she is definitely able to identify it. Otherwise he/she picks one of the alternatives at random. Given that a student has identified the correct alternative what is the conditional probability that he/she knew it, assuming 70% of the student know the correct alternative to the question under consideration.

Example 6 (continued)

Let us define the following events

- A_1 = The student identifies the correct alternative
- B_1 = The student knows the correct alternative
- B_2 = The student does not know the correct alternative

Then we know that $P(B_1) = 0.7$, $P(B_2) = 0.3$, $P(A|B_1) = 1$ and $P(A|B_2) = 0.25$ from which we have $P(B_1|A) = 0.9032$

Exercise

The marketing manager of a toy manufacturing company is considering the marketing of a new toy. In the past 40% of the toys introduced by the company have been successful and 60% have been unsuccessful. Before the toy is marketed, a market research is conducted and a report, either favourable or unfavourable, is compiled. In the past, 80% of the successful toys received a favourable market research report, and 30% of the unsuccessful received a favourable market research report. The marketing manager wants to know the probability that the toy will be successful if it receives a favourable report

Independence of events

Two events A and B are called independent if $P(A \cap B) = P(A)*P(B)$

Distribution

Depending what are the outcomes of an experiment a random variable (r.v) is used to denote the outcome of the experiment and we usually denote the r.v using X, Y or Z and the corresponding probability distribution is denoted by f(x), f(y) or f(z)

- Discrete: probability mass function (pmf)
- Continuous: probability density function (pdf)

Discrete distribution

1) Uniform discrete distribution
2) Binomial distribution
3) Negative binomial distribution
4) Geometric distribution
5) Hypergeometric distribution
6) Poisson distribution
7) Log distribution

Bernoulli Trails

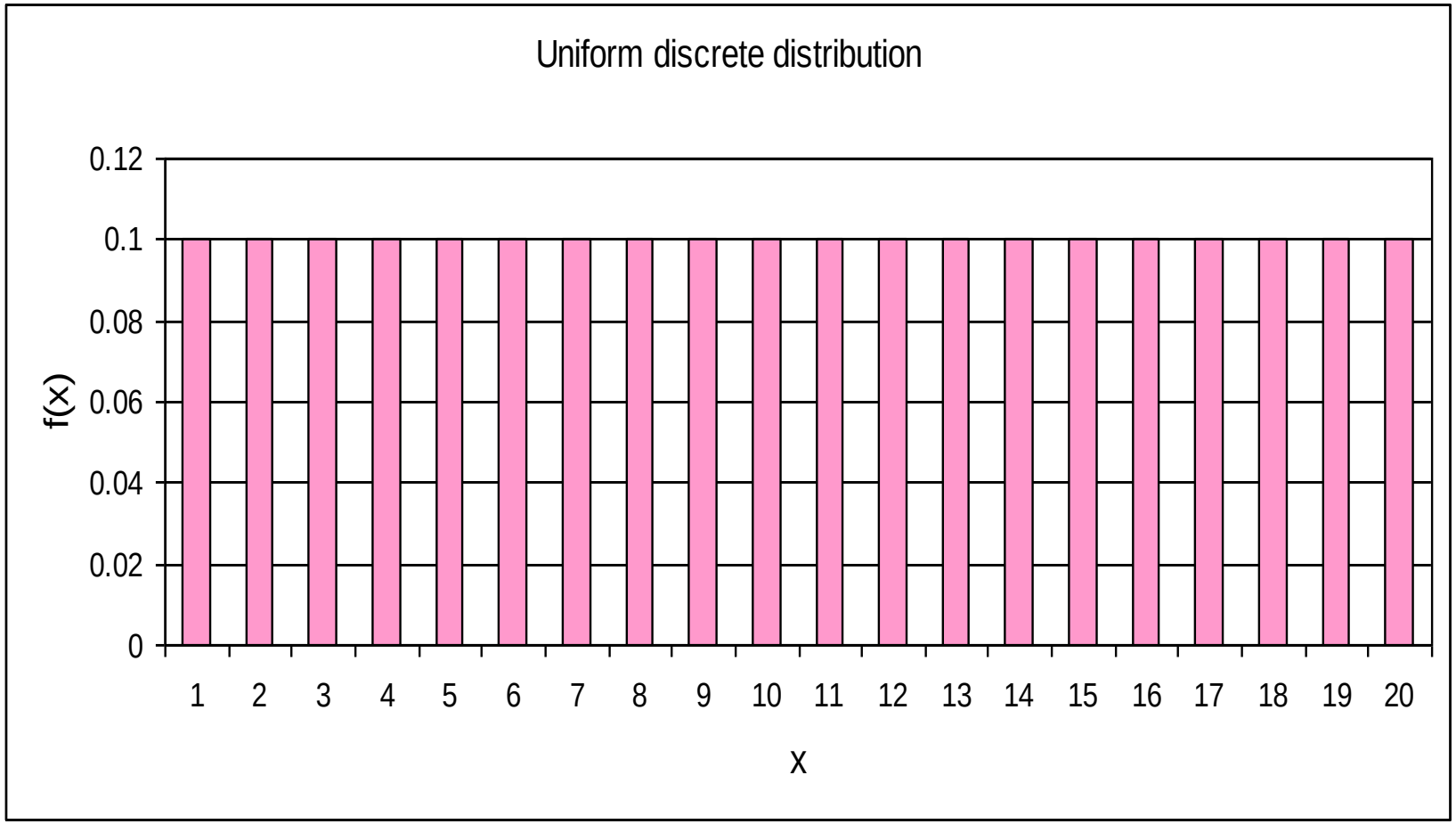
1) Each trial has two possible outcomes, say a success and a failure.
2) The trials are independent
3) The probability of success remains the same and so does the probability of failure from one trial to another

Uniform discrete distribution [X ~ UD (a , b)]

f(x) = 1/n x = a, a+k, a+2k,….., b

- a and b are the parameters where a, b $\in$ R
- E[X] = $a + \frac{k}{2}(n-1)$
- V[X] = $k^2(\frac{n^2-1}{12})$
- Example: Generating the random numbers 1, 2, 3,…, 10. Hence X~UD(1,10) where a=1, k=1, b=10. Hence n=10.

Uniform discrete distribution



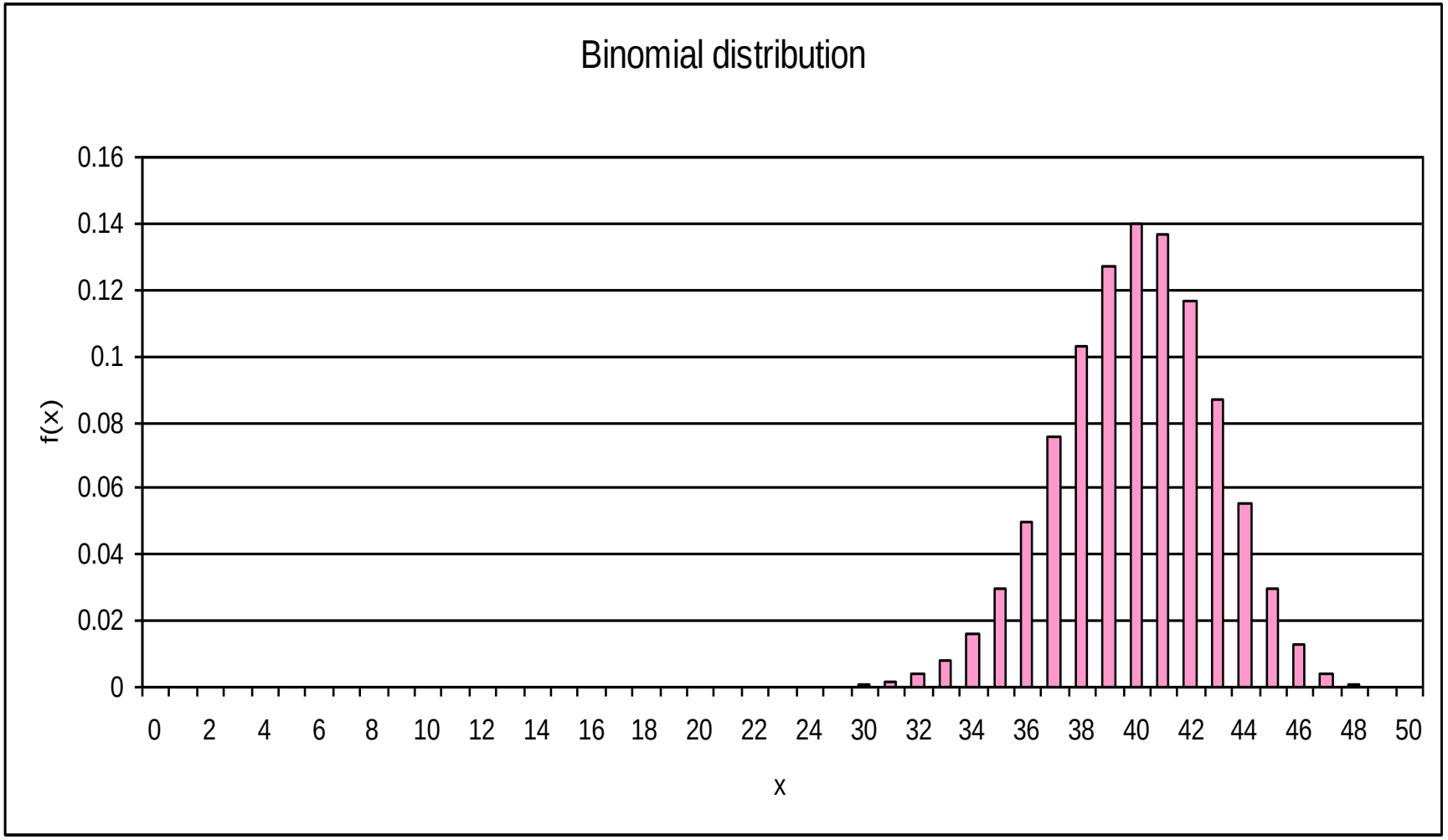
Binomial distribution [X ~ B (p , n)]

$$f(x) = {}^{n}C_{x}p^{x}q^{n-x} \qquad x = 0, 1, 2,\ldots..,n$$

- n and p are the parameters where $p \in [0, 1]$ and $n \in Z^{+}$
- E[X] = np
- V[X] = npq
- Example: Consider you are checking the quality of the product coming out of the shop floor. A product can either pass (with probability p = 0.8) or fail (with probability q = 0.2) and for checking you take such 50 products (n = 50). Then if X is the random variable denoting the number of success in these 50 inspections, we have

$$X \sim {}^{50}C_{x}(0.8)^{x}(0.2)^{50-x}$$

Binomial distribution



Solved example (Binomial distribution)

Question: 20% of the BSNL shares are owned by Mr.Murali Lal. A random sample of 5 shares is chosen. What is the probability that at most 2 of them will be found to be owned by Mr.Murali Lal?

Solved example (Binomial distribution)

Answer: Here p=0.2, q=0.8. Hence the required probability is:

$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

$P(X \leq 2) = {}^5C_0 0.2^0 0.8^5 + {}^5C_1 0.2^1 0.8^4 + {}^5C_2 (0.2)^2 (0.8)^3$

Assignment (Binomial distribution)

Question: Mr. Abhishek Bhatia finds that 70% of the students have answered question # 1 correctly in his chemistry examination. A random sample of 10 answer scripts is chosen. What is the probability that at least 3 of the students have correctly answered the question # 1?

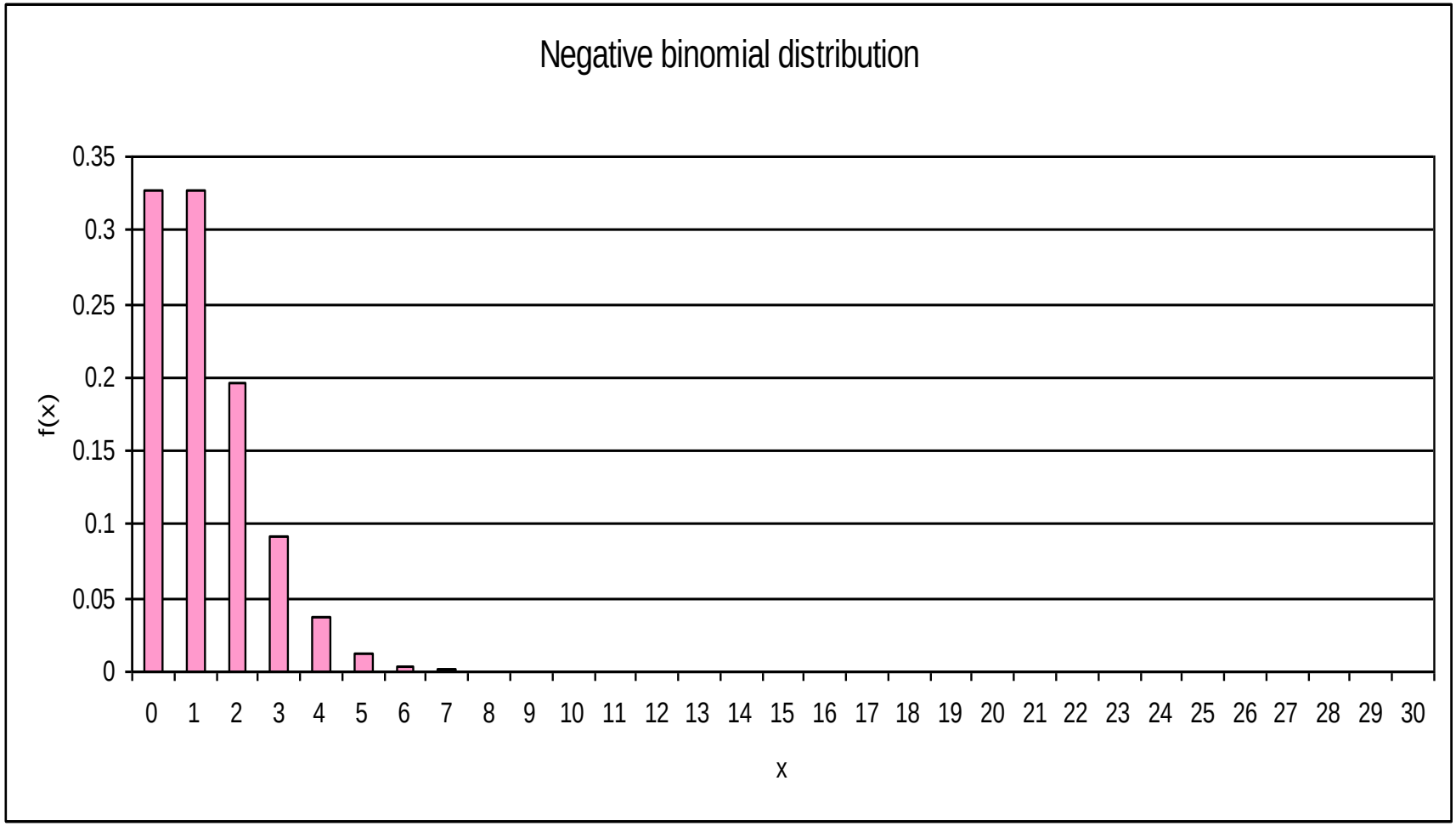
Negative binomial distribution [X ~ NB (p , r)]

$$f(x) = {}^{r+x-1}C_{r-1}p^r q^x \qquad x = r, r+1, \ldots\ldots$$

- p and r are the parameters where $p \in [0, 1]$ and $r \in Z^+$
- $E[X] = rq/p$
- $V[X] = rq/p^2$
- Example: Consider the example above where you are still inspecting items from the production line. But now you are interested in finding the probability distribution of the number of failures preceding the 5^{th} success of getting the right product. Then, we have, considering p=0.8, q=0.2

$$X \sim {}^{5+x-1}C_{5-1}(0.8)^5(0.2)^x$$

Negative binomial distribution



Solved example (Negative binomial distribution)

Question: Suppose that the probability of a manufacturing process producing a defective item is 0.05. Suppose further that the quality of any one item is independent of the quality of any other item produced. If a Mr. Rao the quality control officer selects items at random from the production line and stops his inspection when he gets 10 good items. Then what is the probability of getting 2 defective items before he stops inspection?

Solved example (Negative binomial distribution)

Answer: Here p=0.95, q=0.05. Hence the required probability is

$P(X=2) = {}^{10+2-1}C_{10-1}(0.95)^{10}(0.05)^{2}$

Assignment (Negative binomial distribution)

Question: Mr. Agrawal is the immigration official at IGIA and he knows that the probability of an Indian origin person coming in any flight from abroad is 10%. He notes down the country of origin of persons and stop when he knows he has noted down exactly 3 people of Indian origin. Then what is the probability of Mr.Agrawal noting down 2 non Indian origin people before he stops his checking?

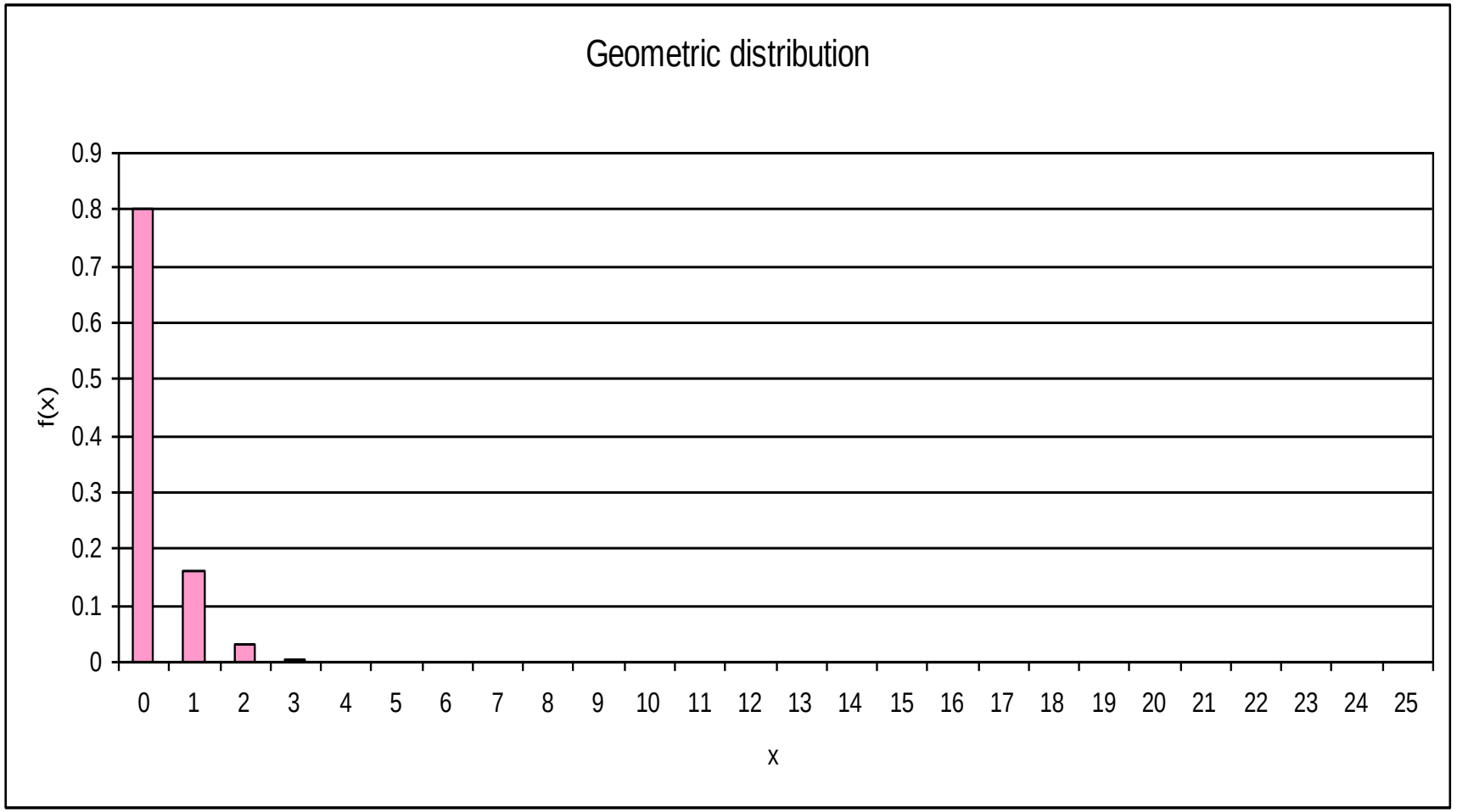
Geometric distribution [X ~ G (p)]

$$f(x) = pq^x \qquad x = 0,1,2,\ldots..$$

- p is the parameter where $p \in [0, 1]$
- E[X] = q/p (r = 1 in the Negative Binomial distribution case)
- $V[X] = q/p^2$ (r = 1 in the Negative Binomial distribution case)
- Example: Consider the example above. But now you are interested in finding the probability distribution of the number of failures preceding the 1st success of getting the right product. Then, we have considering p=0.8, q=0.2

$$X \sim (0.8)(0.2)^x$$

Geometric distribution



Solved example (Geometric distribution)

Question: A recent study indicates that Colgate toothpaste has a market share of 45% (versus 55% of Pepsodent). Miss. Dabawallah the marketing research executive firm wants to conduct a new taste test for which she wants users of Colgate. Potential participants for the test are selected by random screening of users of toothpaste to find Colgate users. What is the probability that 5 users will have to be interviewed by Miss. Dabawallah to find the 1st Colgate toothpaste user?

Solved example (Geometric distribution)

Answer: Here p=0.45, q=0.55. Hence the required probability is

$P(X=4) = (0.45)^1(0.55)^4$

Assignment (Geometric distribution)

Question: Mr.Moti Singh is a lottery ticket sales person and he knows that the probability of a person coming to him to check whether he/she has won the lottery is 1%. What is the probability that 10 persons arrive before the winner arrives?

Hypergeometric distribution [X ~ HG (N, n, p)]

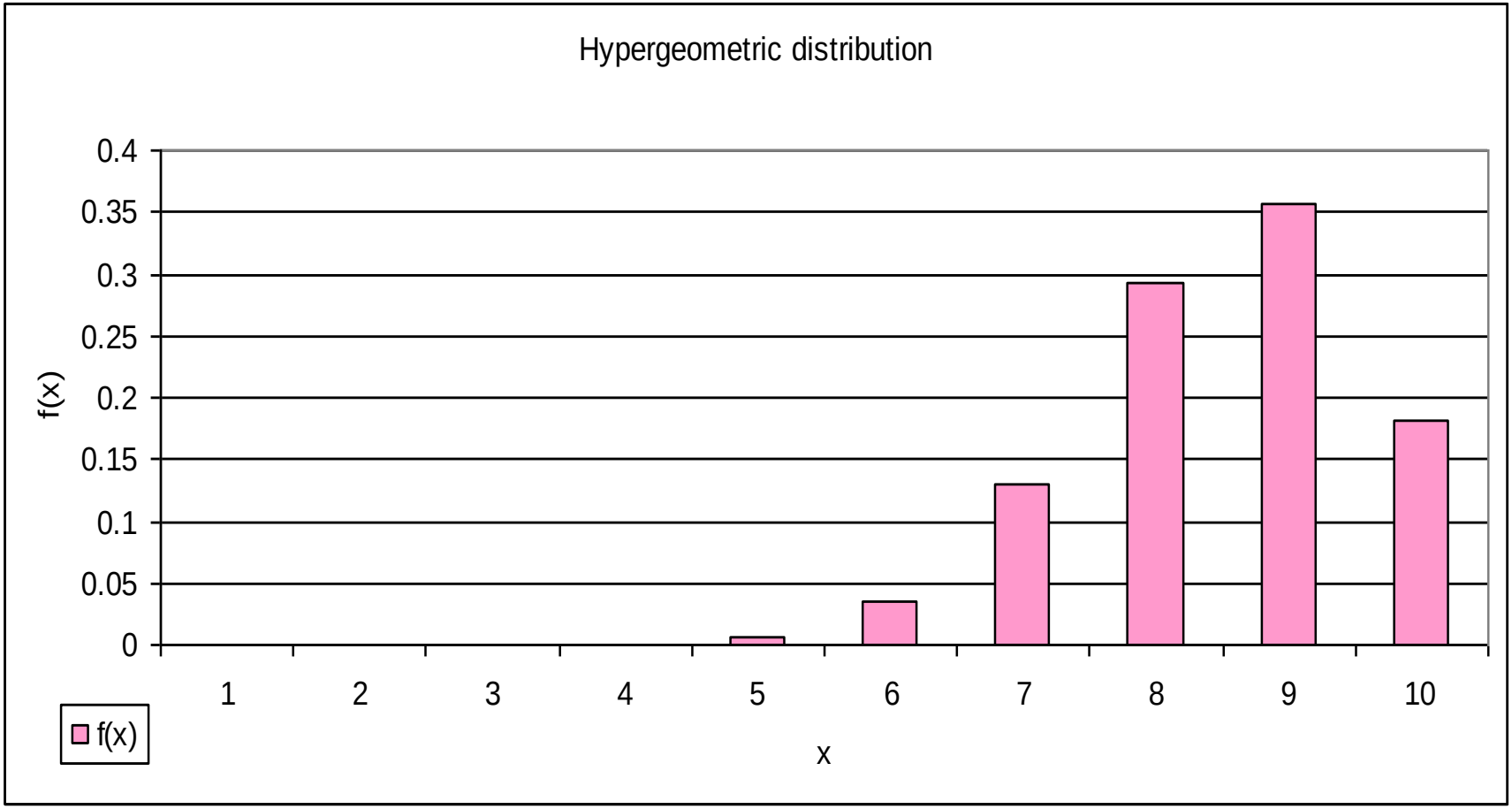
$f(x) = {}^{Np}C_x \, {}^{Nq}C_{n-x} / {}^{N}C_n$ $\quad 0 \le x \le Np$ and $0 \le (n - x) \le Nq$

- N, n and p are the parameters
- E[X] = np
- V[X] = npq{(N – n)/(N – 1)}
- Example: Consider the example above. But now you are interested in finding the probability distribution of the number of failures(success) of getting the wrong(right) product when we choose n number of products for inspection out of the total population N. If the population is 100 and we choose 10 out of those, then the probability distribution of getting the right product, denoted by X is given by

$$X \sim {}^{85}C_x \, {}^{15}C_{10-x} / {}^{100}C_{10}$$

- Remember
  - p (0.85) and q (0.15) are the proportions of getting a good item and bad item respectively.
  - In this distribution we are considering the choosing is done without replacement

Hypergeometric distribution



Solved example (Hypergeometric distribution)

Question: Suppose that automobiles arrive at a Mr. Ghosh′s garage in lots of 10 and that for time and resource considerations, he can inspect only 5 out of each 10 for safety. The 5 cars are randomly chosen from the 10 on the lot. If 2 out of the 10 cars on the lot are below standards for safety, what is the probability that at most 1 out of the 5 cars to be inspected by Mr.Ghosh will be found not meeting the safety standards?

Solved example (Hypergeometric distribution)

Answer: Here N=10, n=5, Np=2, Nq=8

Hence the required probability is

$P(X \leq 1) = P(X=0) + P(X=1)$

$P(X \leq 1) = {}^{2}C_{0}\,{}^{8}C_{5} / {}^{10}C_{5} + {}^{2}C_{1}\,{}^{8}C_{4} / {}^{10}C_{5} =$

Assignment (Hypergeometric distribution)

Question: Mrs. Patanaik is the teacher of class III consisting of 30 students. Daily during morning prayer session she would like to check whether all her students have brought their respective lunches. But due to paucity of time each day she has to randomly select 10 students and finds that on an average out of this 10, 4 do not bring their lunches. If on any particular day she selects 10 students then what is the probability that exactly 5 students would not have brought their respective lunches.

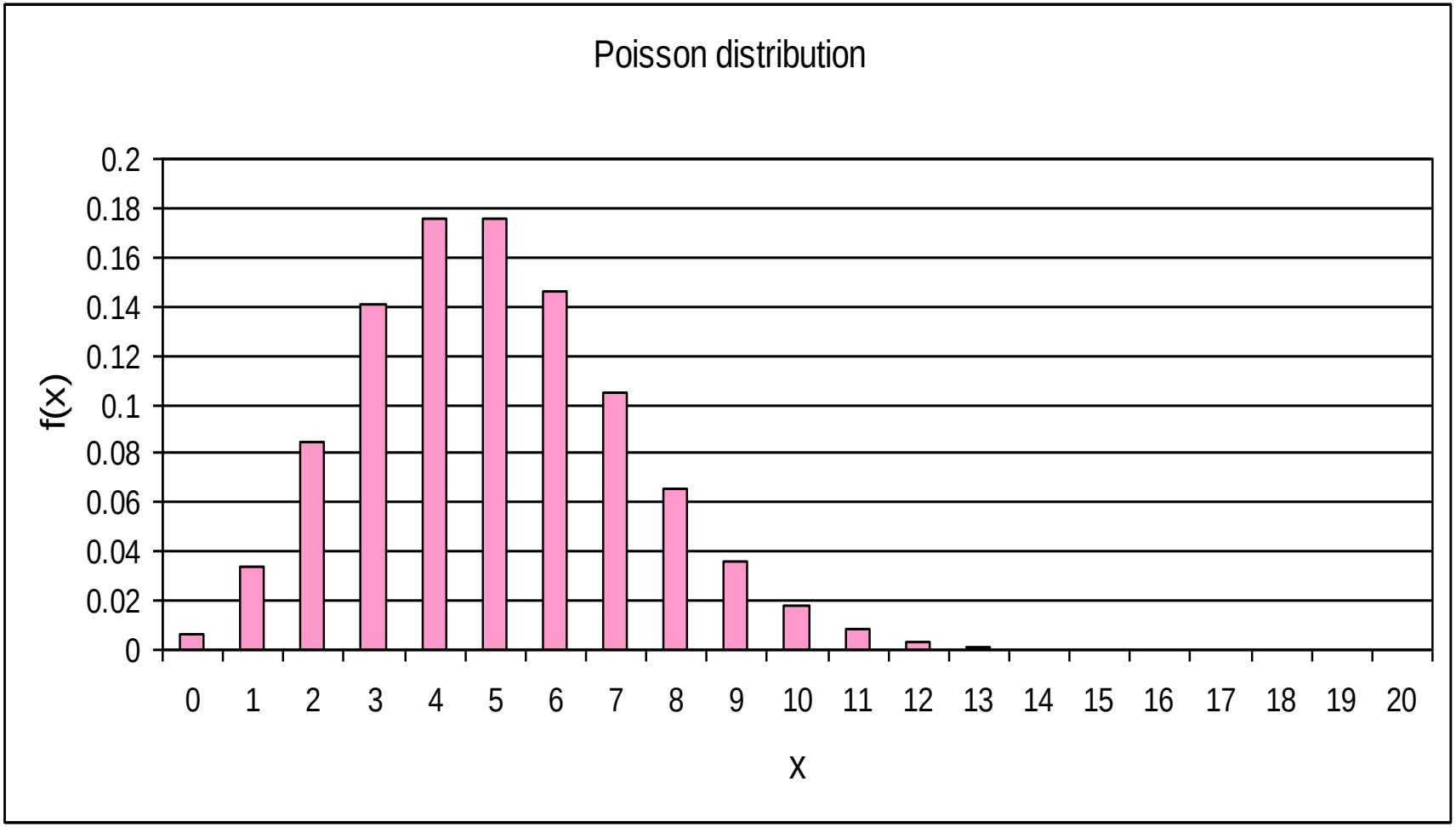
Poisson distribution
[X ~ P (λ)]

$$f(x) = e^{-\lambda}\lambda^{x}/x! \qquad x = 0,1,2,\ldots..$$

- λ is the parameter where $\lambda > 0$
- $E[X] = \lambda$
- $V[X] = \lambda$
- Example: Consider the arrival of the number of customers at the bank teller counter. If we are interested in finding the probability distribution of the number of customers arriving at the counter in specific intervals of time and we know that the average number of customers arriving is 5, then, we have

$$X \sim e^{-5}5^{x}/x!$$

Poisson distribution



Solved example (Poisson distribution)

Question: Mr. Gurneek Singh who is the cashier at the departmental store cash counter notices that the average number of customer arriving at the cash counter per 5 minutes is 10. Then what is the probability that more than 5 customers arrive at the cash counter with the interval of 5 minutes?

Solved example (Poisson distribution)

Answer: Here $\lambda=10$. Hence the required probability is

$$P(X\geq 6) = 1 - \{P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)\}$$

$$P(X\geq 6) = 1 - \{\exp(-10)10^0/0! + e^{-10}10^1/1! + \exp(-10)10^2/2! + \exp(-10)10^3/3! + \exp(-10)10^4/4! + \exp(-10)10^5/5!\}$$

Assignment (Poisson distribution)

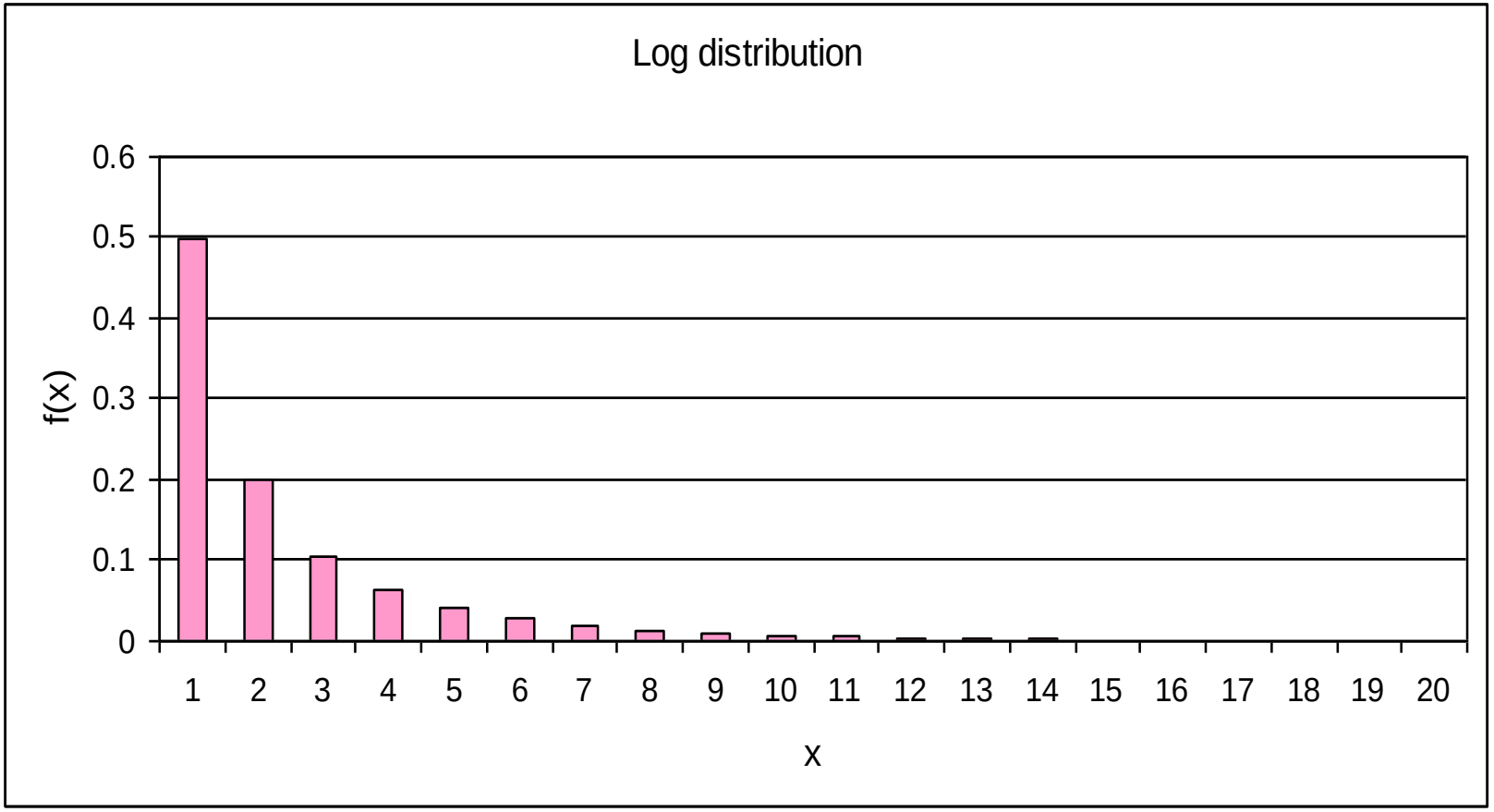
Question: Mr. Pankaj Yadav is the shop floor manager and he notices that the average number of jobs arriving at the lathe machine per hour is 25. He wants to find out what is the probability of exactly 10 jobs arrive in an hour at the lathe machine?

Log distribution
[X ~ L (p)]

$$f(x) = -(\log_e p)^{-1} x^{-1} (1 - p)^x \qquad x = 1,2,3,\ldots..$$

- p is the parameter where $p \in (0, 1)$
- $E[X] = -(1-p)/(p\log_e p)$
- $V[X] = -(1-p)[1 + (1 - p)/\log_e p]/(p^2 \log_e p)$
- Example
  1) Emission of gases from engines against fuel type
  2) Used to represent the distribution of the number of items of a product purchased by a buyer in a specified period of time

Log distribution



Continuous distribution

1) Uniform distribution
2) Normal distribution
3) Exponential distribution
4) Chi-Square distribution
5) Gamma distribution
6) Beta distribution
7) Cauchy distribution

Continuous distribution

8) t-distribution
9) F-distribution
10) Log-normal distribution
11) Weibull distribution
12) Double exponential distribution
13) Pareto distribution
14) Logistic distribution

Uniform distribution [X ~ U (a, b)]

$$f(x) = 1/(b - a) \qquad a \leq x \leq b$$

- a and b are the parameters where a, b $\in$ R and a < b
- E[X] = (a+b)/2
- V[X] = $(b-a)^2/12$
- Example: Choosing any number between 1 and 10, both inclusive, from the real line

Uniform distribution

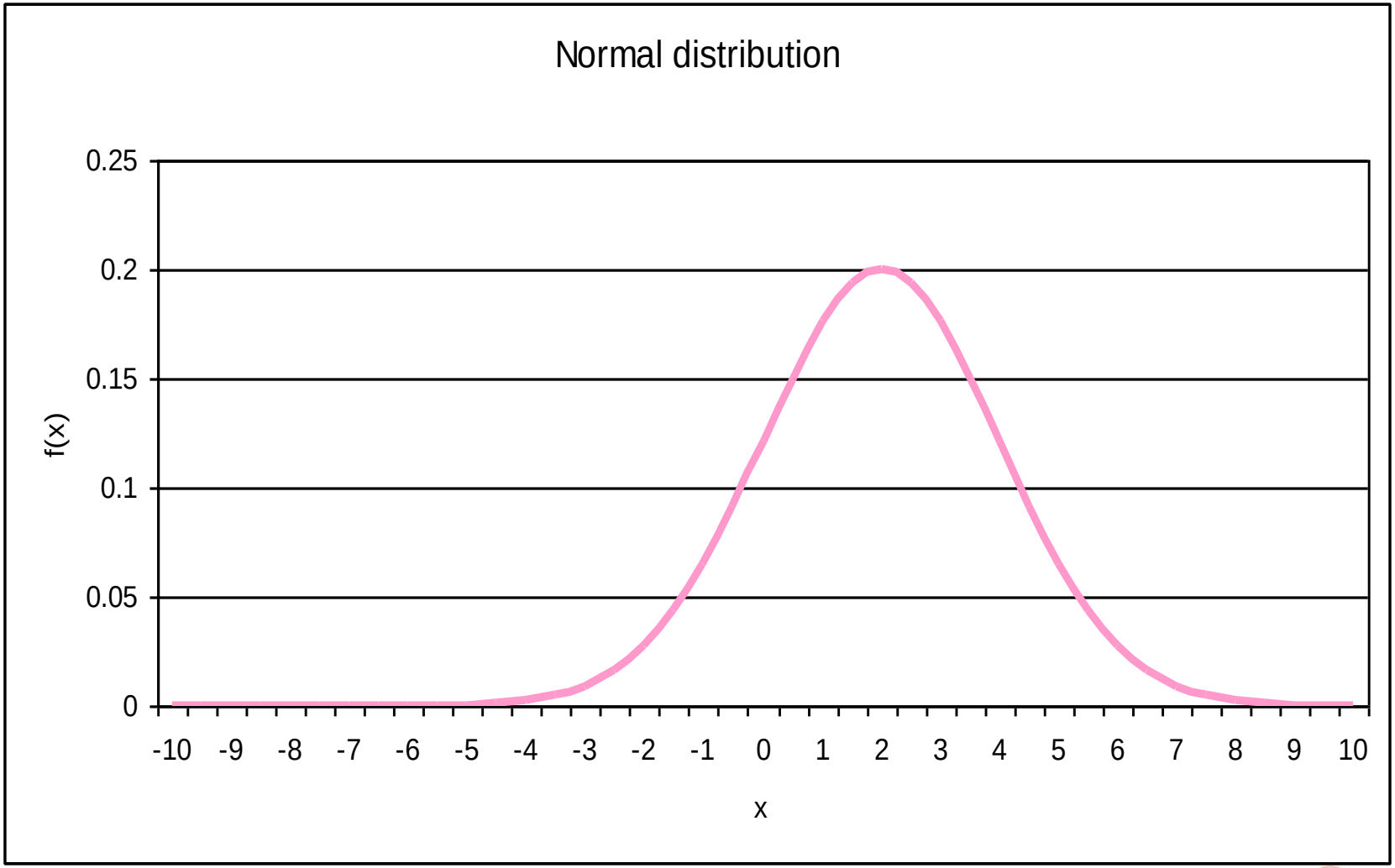


Normal distribution
[X ~ N (μ, σ^2)]

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \qquad -\infty < x < \infty$$

- μ_X, σ^2_X are the parameters where $\mu_X \in R$ and $\sigma^2_X > 0$
- $E[X] = \mu_X$
- $V[X] = \sigma^2_X$
- Example: Consider the average age of a student between class VII and VIII selected at random from all the schools in the city of Kanpur

Normal distribution

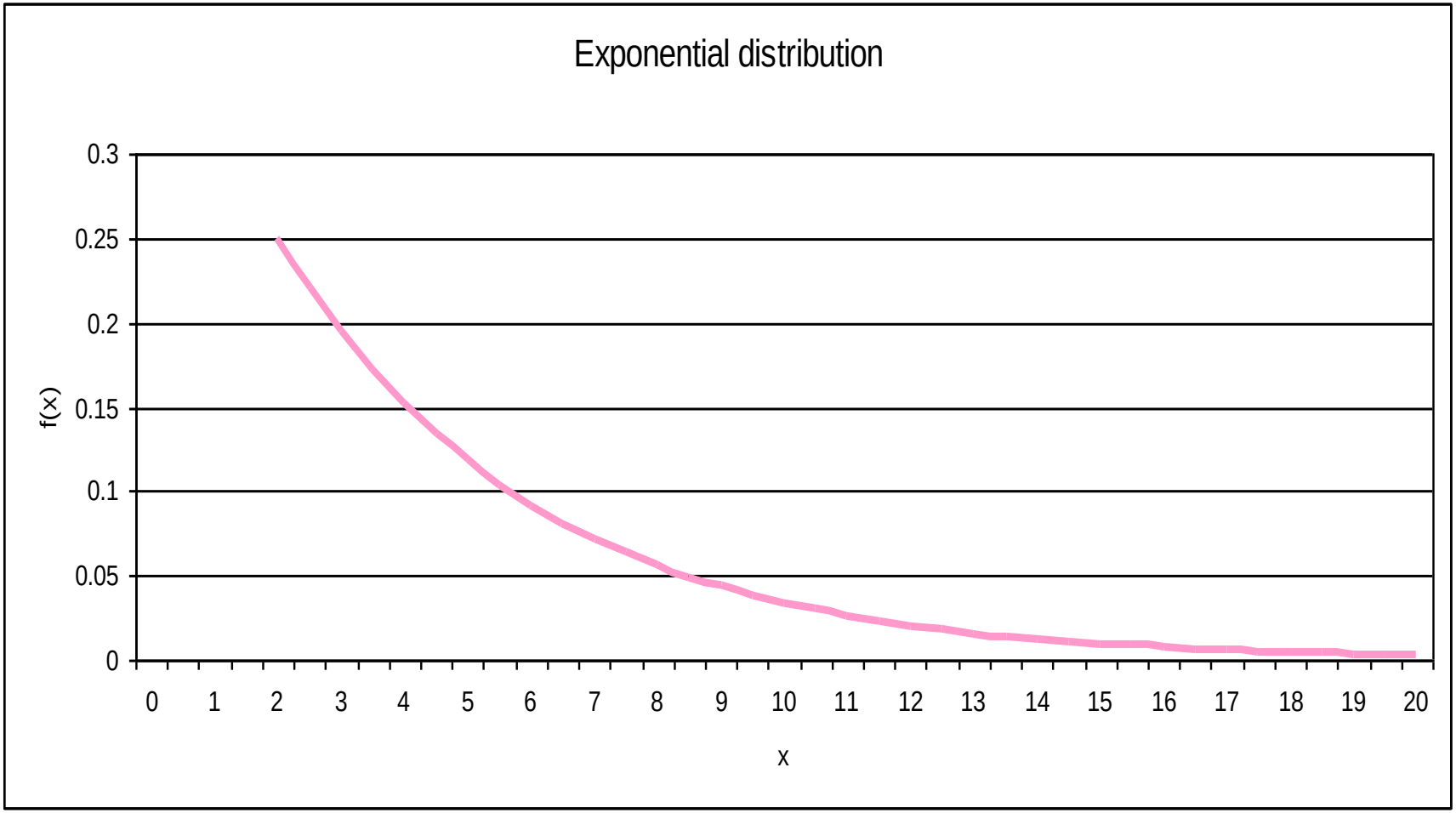


Exponential distribution [X ~ E (a, θ)]

$$f(x) = \frac{1}{\theta} e^{-\frac{(x-a)}{\theta}} \qquad a < x < \infty$$

- a and θ are the parameters where a ∈ R and θ > 0
- E[X] = a + θ
- V[X] = θ^2
- `Example: The life distribution of the number of hours a electric bulb survives.

Exponential distribution



Solved example (Exponential distribution)

Question: Mr. K.Bharadwaj an electrician knows that the time a bulb operates before it gets fused is exponentially distributed with a mean life of 100 hours. What is the probability that any bulb will work continuously for at least 200 hours?

Solved example (Exponential distribution)

Answer: Here $\theta = 100$. hence the required probability is

$P(X \geq 200) = 1 - P(X \leq 200)$

$$P(X \geq 200) = 1 - \int_{x=0}^{200} \frac{1}{100} \exp(-\frac{x}{100}) dx$$

Assignment (Exponential distribution)

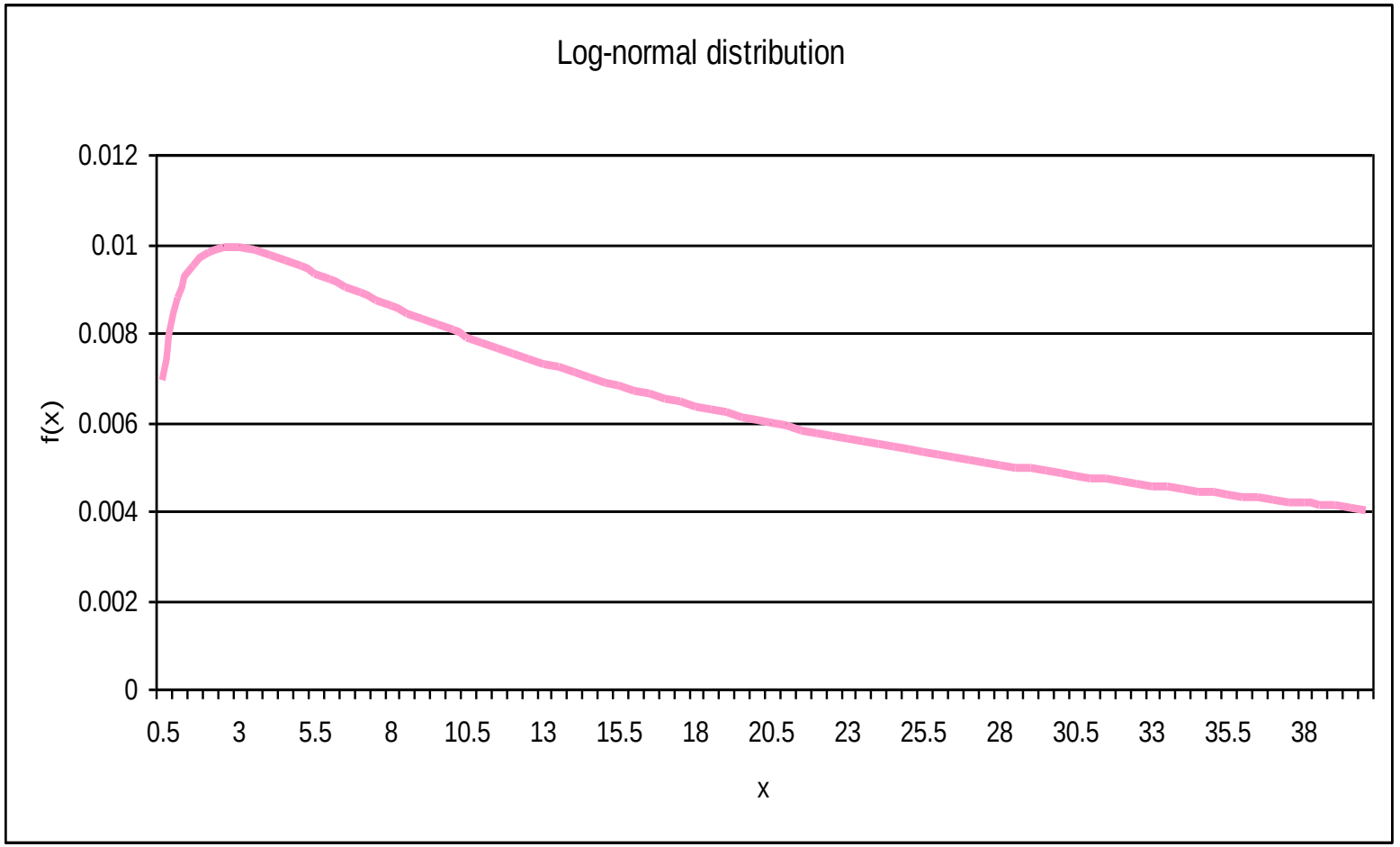
Question: Mr. Lyndoh the owner of an electronics shop knows that the average life of a tape recorder he sells is 1.5 years. What is the probability that any such tape recorder would function for at most 3 years?

Log-normal distribution
[X ~ LN (μ, σ^2)]

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_X}\frac{1}{x}e^{-\frac{(\log_e x-\mu_X)^2}{2\sigma_X^2}} \qquad 0 < x < \infty$$

- μ_X, σ^2_X are the parameters where $\mu_X \in R$ and $\sigma^2_X > 0$
- $E[X] = \exp(\mu_X+\sigma^2_X/2)$
- $V[X] = \exp(2\mu_X+\sigma^2_X)\{\exp(\sigma^2_X)-1\}$
- Example: Stock prices return distribution

Log-normal distribution



Relationship between Poisson and Exponential distribution

If a process has the intervals between successive events as independent and identical and it is exponentially distributed then the number of events in a specified time interval will be a Poisson distribution

Cumulative distribution function (cdf) or the distribution function

We denote the distribution function by F(x)

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} dF(x)$$

Properties of distribution function

1) F(x) is non-decreasing in x, i.e., if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$
2) Lt F(x) = 0 as $x \rightarrow -\infty$
3) Lt F(x) = 1 as $x \rightarrow +\infty$
4) F(x) is right continuous

Standard normal distribution

Putting $Z=(X-\mu_X)/\sigma_X$ in the normal distribution we have the standard normal distribution

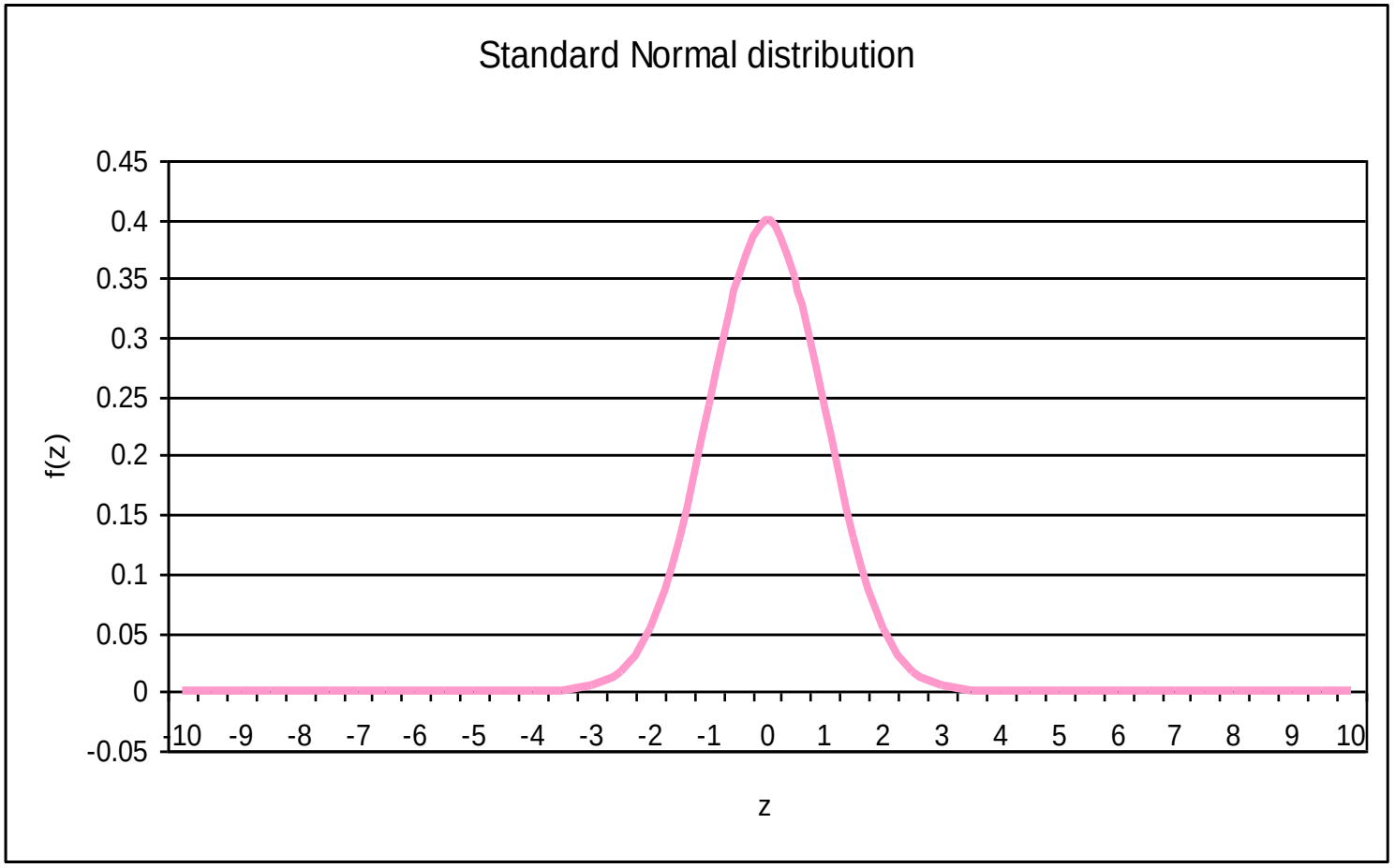
$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

Where: $\mu_Z = 0$ and $\sigma_Z = 1$

<u>Remember</u>

- $F(x) = P(X \le x) = F(z) = F(Z \le z)$
- $f(z) = \phi(z)$
- $F(z)=P(Z\le z)=\int_{-\infty}^{z} f(z)dx=\int_{-\infty}^{z} dF(x)=\Phi(z)$

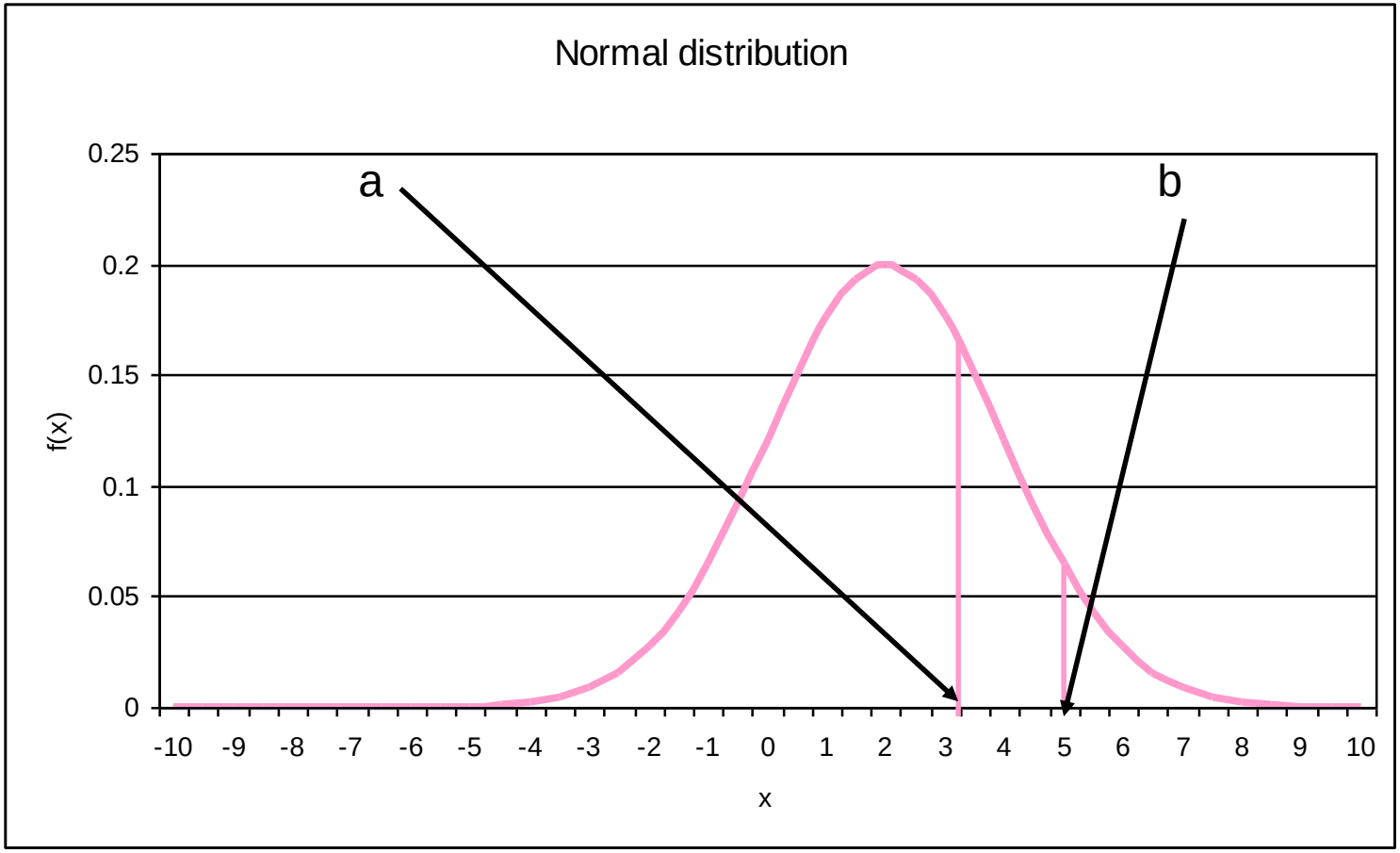
Standard normal distribution



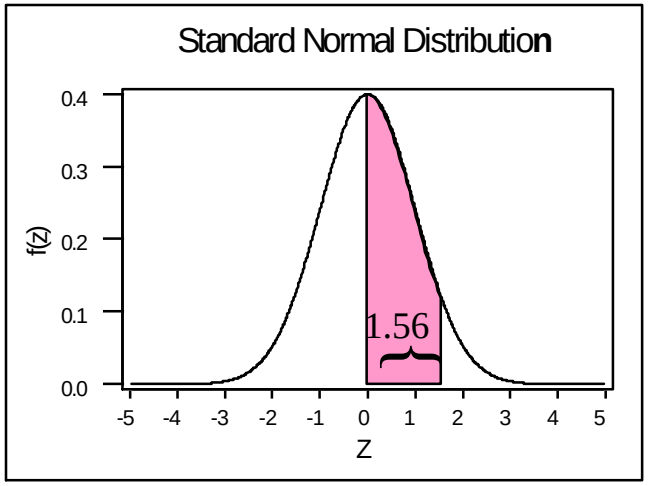
Normal distribution results

- $\phi(-z) = \phi(z)$, i.e., $f(-z) = f(z)$
- $\Phi(-z) = 1 - \Phi(z)$, i.e., $P(Z \leq z) = 1 - P(Z \geq z)$
- $P(a \leq X \leq b) = \Phi[(b - \mu_X)/\sigma_X] - \Phi[(a - \mu_X)/\sigma_X]$
- $P(X \leq b) = \Phi[(b - \mu_X)/\sigma_X]$
- $P(a \leq X) = 1 - \Phi[(a - \mu_X)/\sigma_X]$

Normal distribution results



Finding Probabilities using Standard Normal Distribution: P(0 < Z < 1.56)



Look in row labeled 1.5 and column labeled .06 to find $P(0 \le z \le 1.56) = .4406$

Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Standard Normal distribution

Z ~ N(0,1) given by the equation $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

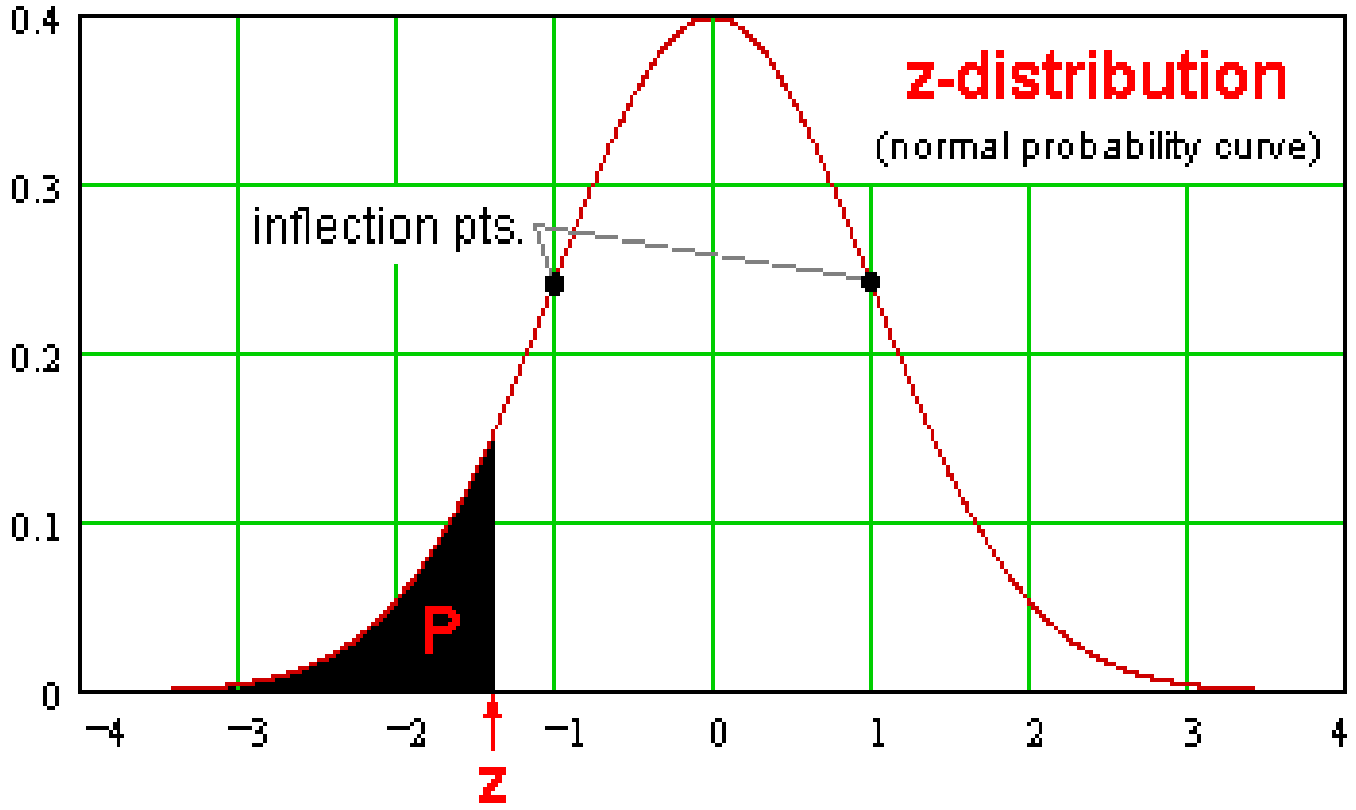
The area within an interval (a,b) is given by

$$F(a \le Z \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$ which is not integrable

algebraically. The Taylor's expansion of the above assists in speeding up the calculation, which is

$$F(Z \le z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k+1}}{(2k+1)2^k k!}$$

Standard normal distribution



Solved example (Normal distribution)

In an examination 20% of the students failed (i.e., obtained a score which is less than or equal to 40 marks out of 100) and 10% of the students obtained a grade A (score of 70 marks of above out of 100). Assuming normal distribution of marks find the mean and the standard deviation of the distribution

Solved example (Normal distribution)

Steps

1) $P(X \leq 40) = 0.2 = P[(X-\mu_X)/\sigma_X \leq (40- \mu_X)/\sigma_X] = P(Z \leq z_1) = \Phi(z_1) = -0.84$
2) $P(X \geq 70) = 0.1 = P[(X-\mu_X)/\sigma_X \geq (70- \mu_X)/\sigma_X] = P(Z \geq z_2) = 1 - P(Z \leq z_2) = 1 - \Phi(z_2)$. Hence $\Phi(z_2) = 0.9$

Hence we have from the above two equations:

- $z_1 = (40 - \mu_X)/\sigma_X = - 0.84$
- $z_2 = (70 - \mu_X)/\sigma_X = + 0.90$
- $\mu_X = 54.12$; $\sigma_X = 17.64$

Solved example (Normal distribution)

Question: In Prof. Ram Pal's mathematics examination 20% of the students failed (i.e., obtained a score which is less than or equal to 40 marks out of 100) and 10% of the students obtained a grade A (score of 70 marks of above out of 100). Assuming normal distribution of marks find the mean and the standard deviation of the distribution of marks in mathematics?

Solved example (Normal distribution)

Answer:

1) $P(X \leq 40) = 0.2 = P[(X-\mu_X)/\sigma_X \leq (40- \mu_X)/\sigma_X] = P(Z \leq z_1) = \Phi(z_1) = -0.84$

2) $P(X \geq 70) = 0.1 = P[(X-\mu_X)/\sigma_X \geq (70- \mu_X)/\sigma_X] = P(Z \geq z_2) = 1 - P(Z \leq z_2) = 1 - \Phi(z_2)$. Hence $\Phi(z_2) = 0.9$

   Hence we have from the above two equations:

- $z_1 = (40 - \mu_X)/\sigma_X = - 0.84$
- $z_2 = (70 - \mu_X)/\sigma_X = + 1.28$
- $\mu_X = 51.9$; $\sigma_X = 14.2$

Solved example (Normal distribution)
Finding Probabilities using Standard Normal

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Assignment (Normal distribution)

Question: Dr. Satish Kashyap finds that his patient's height are normally distributed with mean 165 cms and standard deviation 20 cms. What is the probability of a patient height being between 160 to 170 cms?

Few approximations

Normal approximation to Binomial distribution:

Let X ~ B(p, n) where n is large and p is small. Then the distribution can be approximated by the Normal distribution X ~ N(np, npq)

De Moivre Laplace Limit Theorems

1) $$P(a \le X \le b) \cong \Phi[\frac{b-np}{\sqrt{npq}}] - \Phi[\frac{a-np}{\sqrt{npq}}]$$

2) $$P(a \le X) \cong 1 - \Phi[\frac{a-np}{\sqrt{npq}}]$$

3) $$P(X \le b) \cong \Phi[\frac{b-np}{\sqrt{npq}}]$$

Markov Inequality

Let Z be a non-negative r.v such that E[Z] exists. Then for every positive t we have

$$P(Z \geq t) \leq E[Z]/t$$

Tchebychev's Inequality

Let X be a r.v such that $E[X] = \mu_X$ and $V[X] = \sigma^2_X$. Then for every positive t we have

$$P(|X - \mu_X| \geq t\sigma_X) \leq 1/t^2$$

Bernoulli′s Theorem

Let X_n be the number of success in ′n′ number of Bernoulli trials, each with success probability ′p′. Then for arbitrary positive ε we have

$$\text{Lt } P[|X_n/n - p|) \leq \varepsilon] = 1 \text{ as } n \to \infty$$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d) r.v each with $E[X_i] = \mu_X$ and $V[X_i] = \sigma^2_X$. Then if we define $S = X_1 + X_2 + \ldots + X_n$ and $\frac{X_1 + \ldots + X_n}{n} = \overline{X}_n$

we have $E[S] = n\mu_X$ and $V[S] = n\sigma^2_X$, and for large values of 'n'

$$\frac{\overline{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0,1)$$

Sampling

1) Point estimation
2) Interval estimation
3) Hypothesis testing

Sampling

- Population: N
  - Population distribution
  - Parameter (θ)
- Sample: n
  - Sampling distribution
  - Statistic (t_n)

  Example: Consider a population having the following elements {1, 3, 6, 7}, hence N = 4 and μ = 17/4. If we take n = 2 and chose samples, then the possible values of the sample average are 1, 2,…..,6.5 ,7

Sampling

- Simple random sampling with replacement (SRSWR)
- Simple random sampling without replacement (SRSWOR)
- Note if X has a distribution such that E[X] = μ_X and V[X] = σ^2_X. Then E[X_i] = μ_X and V[X_i] = σ^2_X.

Types of sampling

- Probability Sampling
  - Simple Random Sampling
  - Stratified Random Sampling
  - Cluster Sampling
  - Multistage Sampling
  - Systematic Sampling
- Judgement Sampling
  - Quota Sampling
  - Purposive Sampling

Chi-square distribution

Suppose $Z_1, Z_2, \ldots, Z_n$ are $'n'$ independent observations from N(0, 1), then

$$Z^2_1 + Z^2_2 + Z^2_3 + \ldots + Z^2_n \sim \chi^2_n$$

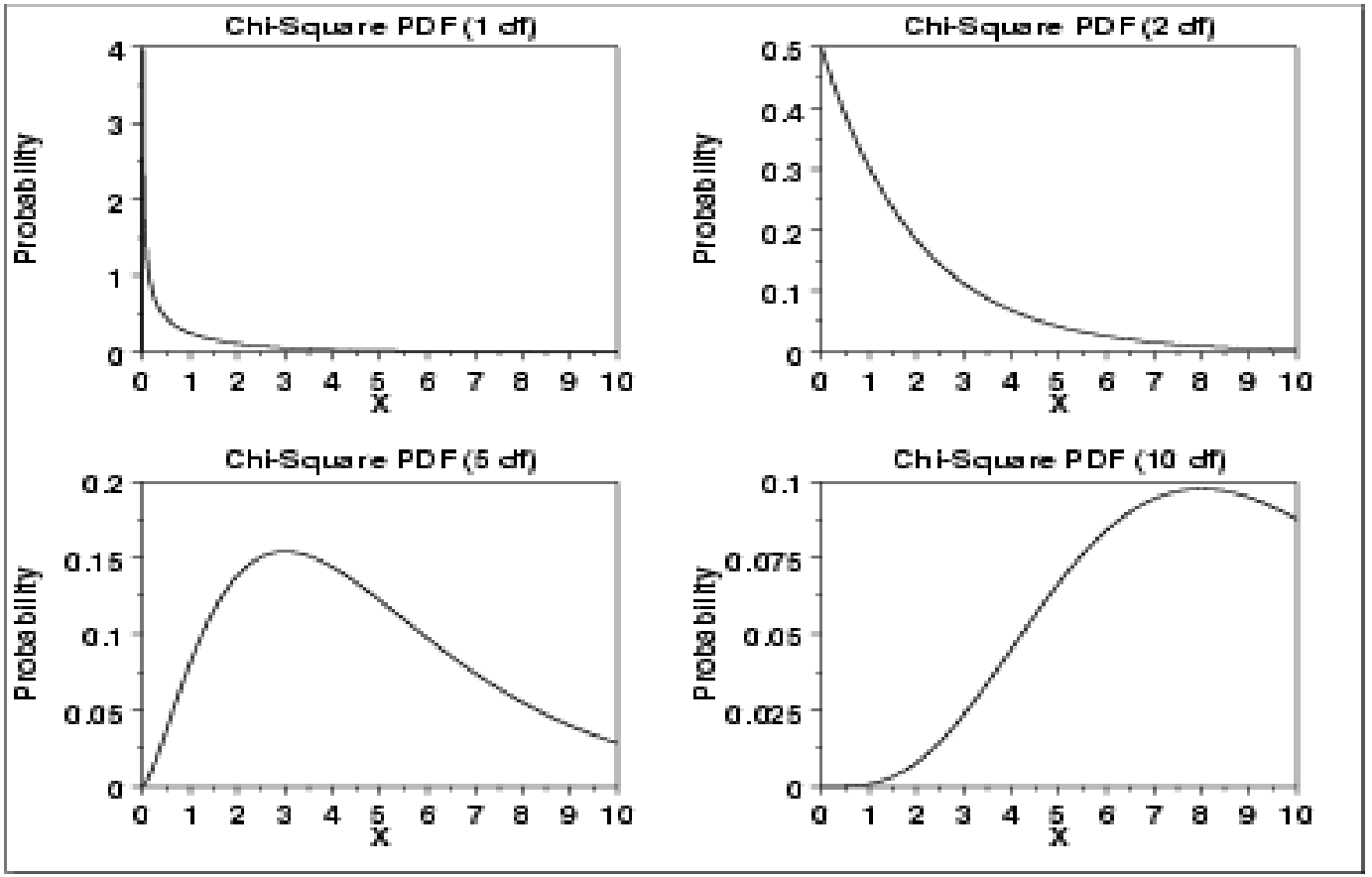
Chi-square distribution

$$[X \sim \chi^2_n]$$

$$f(x) = \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{(\frac{n}{2})-1} e^{-(\frac{x}{2})} \qquad 0 \leq x < \infty$$

- n is the parameter (degree of freedom) where n $\in Z^+$
- E[X] = n
- V[X] = 2n

Chi-square distribution



t-distribution

Suppose Z ~ N(0, 1), Y ~ χ^2_n and they are independent, then

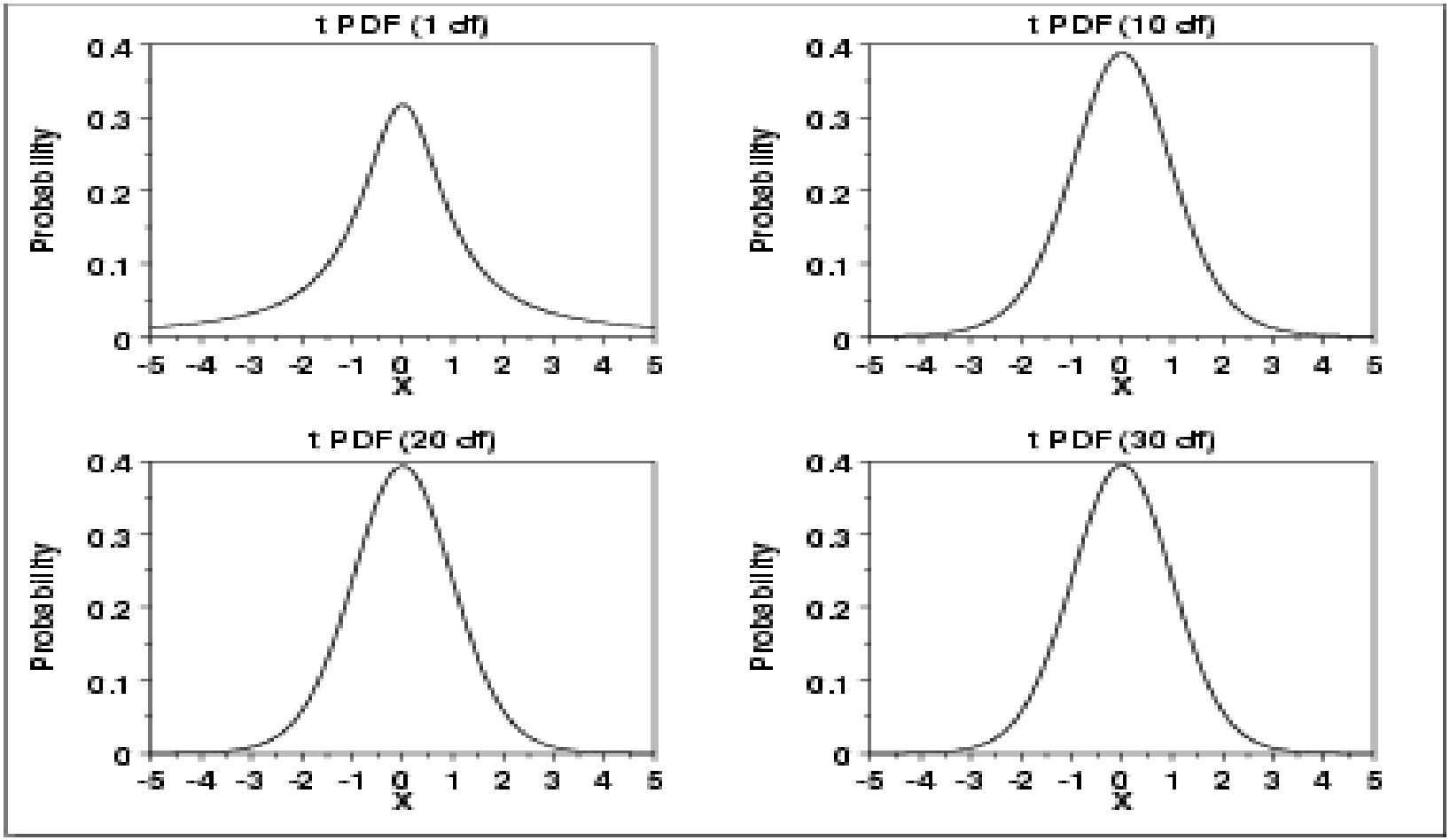
$$\frac{Z}{\sqrt{\frac{Y}{n}}} \sim t_n$$

t-distribution
[$X \sim t_n$]

$$f(x) = \frac{\Gamma[\frac{n+1}{2}]}{\sqrt{n\pi}\Gamma(\frac{n}{2})}[1+\frac{x^2}{n}]$$

- n is the parameter where $k \in Z^+$
- E[X] = 0 (n > 1)
- V[X] = n/(n – 2), (n > 2)

t-distribution



F-distribution

Suppose $X \sim \chi^2_n$, $Y \sim \chi^2_m$ and they are independent, then

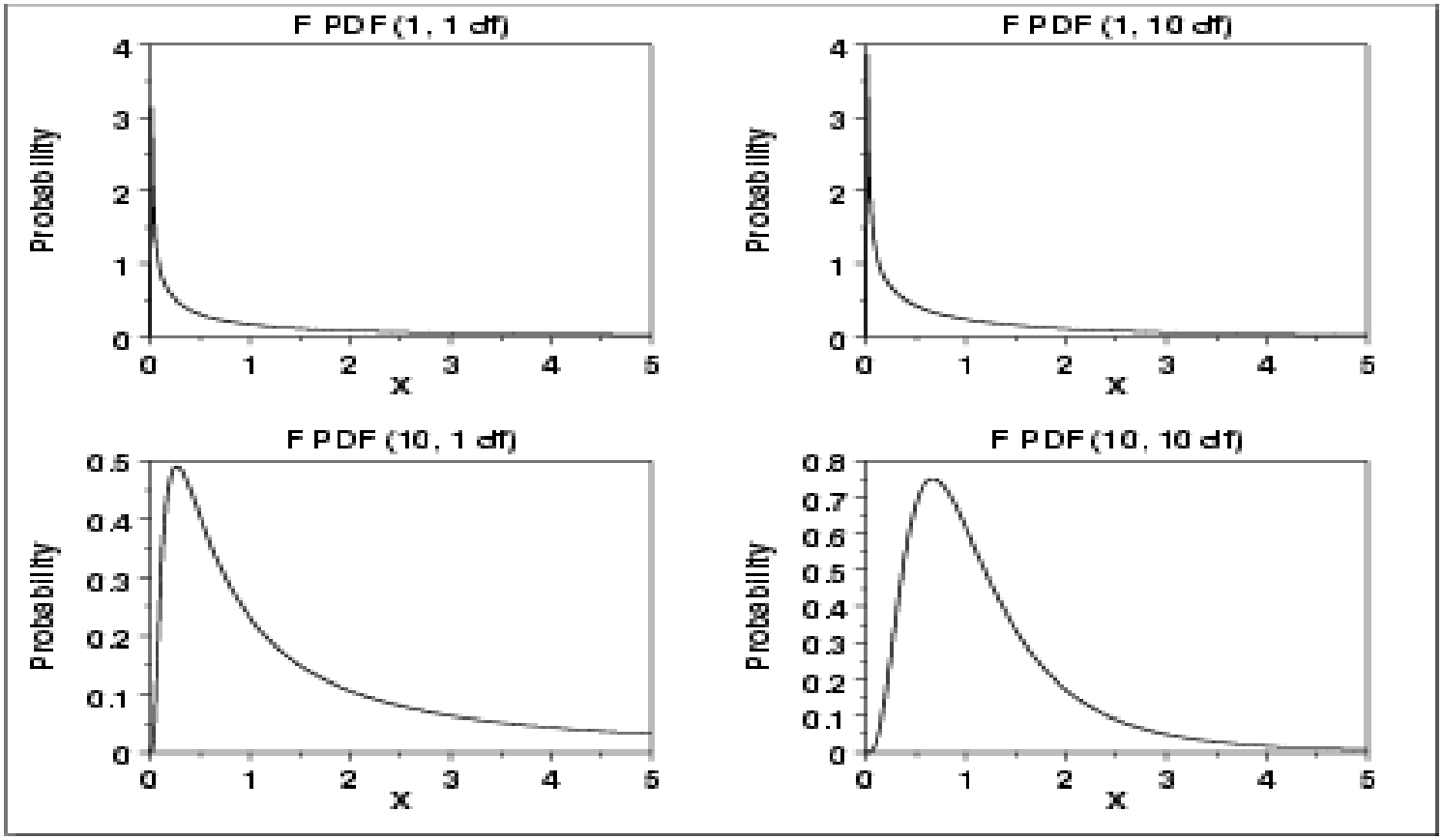
$$\frac{\frac{X}{n}}{\frac{Y}{m}} \sim F_{n,m}$$

F-distribution
[X ~ $F_{n,m}$]

$$f(x) = \frac{n^{\frac{n}{2}} m^{\frac{m}{2}} \Gamma(\frac{n+m}{2}) x^{(\frac{n}{2})-1}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})(nx+m)^{\frac{n+m}{2}}} \qquad 0 < x < \infty$$

- n, m are the parameter (degrees of freedom) where n, m $\in Z^+$
- E[X] = m/(m - 2), (m > 2)
- V[X] = $2m^2(n + m -2)/[n(m – 2)^2(m – 4)]$, (m > 4)

F-distribution



Some results

If $X_1, X_2, \ldots\ldots, X_n$ are ′n′ observations from $X \sim N(\mu_X, \sigma^2_X)$ and $\dfrac{X_1 + X_2 + \ldots\ldots X_n}{n} = \bar{X}_n$

then $\dfrac{\bar{X}_n - \mu_X}{\dfrac{\sigma_X}{\sqrt{n}}} \sim N(0,1)$

Some results

If $S'^{2}_{n,X} = \frac{1}{n}\sum_{j=1}^{n}\{X_j - \mu_X\}^2$ and $S^2_{n,X} = \frac{1}{(n-1)}\sum_{j=1}^{n}\{X_j - \bar{X}_n\}^2$ then

$$\frac{nS'^{2}_{n,X}}{\sigma^2_X} \sim \chi^2_n$$

$$\frac{(n-1)S^2_{n,X}}{\sigma^2_X} \sim \chi^2_{n-1}$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu_X)}{S_{n,X}} \sim t_{n-1}$$

Some results

If $X_1, X_2, \ldots, X_n$ are $'m_X'$ observations from $X \sim N(\mu_X, \sigma^2_X)$ and $Y_1, Y_2, \ldots, Y_m$ are $'m_Y'$ observations from $Y \sim N(\mu_Y, \sigma^2_Y)$ and more over these samples are independent then

$$\frac{\frac{(m_X - 1)S_X^2}{\sigma_X^2} / (m_X - 1)}{\frac{(m_Y - 1)S_Y^2}{\sigma_Y^2} / (m_Y - 1)} = \left(\frac{\sigma_Y^2}{\sigma_X^2}\right)\left(\frac{S_X^2}{S_Y^2}\right) \sim F_{m_X - 1, m_Y - 1}$$

Estimators and their properties

Estimator: Any statistic (a random function) which is used to estimate the population parameter

- Unbiasedness

$$E_\theta(t_n) = \theta$$

- Consistency

$$P[|t_n - \theta| < \varepsilon] = 1 \text{ as } n \to \infty$$

Estimators (Discrete distribution)

1) X ~ UD (a , b) then $\hat{a} = \min(X_1,...,X_n)$ and

   $\hat{b} = \max(X_1,...,X_n)$

2) X ~ B (n, p), then $\hat{p} = \dfrac{\#\ favouring}{n}$

3) X ~ P (λ), then $\hat{\lambda} = \overline{X}_n$

Estimators (Discrete distribution)

1) X ~ UD (a , b) then $\hat{a} = \min(X_1,\ldots,X_n)$ and

   $\hat{b} = \max(X_1,\ldots,X_n)$

2) X ~ P (λ), then $\hat{\lambda} = \overline{X}_n$

3) X ~ B (n, p), then $\hat{p} = \dfrac{\#\ favouring}{n}$

Estimators (Continuous distribution)

1) X ~ N(μ, σ^2), then $\hat{\mu} = \overline{X}_n$
2) X ~ N(μ, σ^2) and if μ is known then

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{X_i - \mu\}^2$$

3) X ~ N(μ, σ^2) and if μ is unknown then

$$\hat{\sigma}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\{X_i - \overline{X}_n\}^2$$

4) X ~ E(θ), then $\hat{\theta} = \overline{X}_n$

Examples (Estimation)

It was found that the respective number of members in 10 different families are 4, 5, 6, 7, 8, 3, 4, 5, 6 and 6. If we consider that the number of members in a family to be uniformly distributed, then the estimated value of a = 3 and that of b = 8.

Examples (Estimation)

You are testing the components coming out of the shop floor and find that 9 out of 30 components fail. Then the estimated value of p (proportion of bad items in the population) = 9/30.

Examples (Estimation)

At a particular teller machine in one of the bank branches it was found that the number of customers arriving in an unit time span were 4, 6, 7, 4, 3, 5, 6, and 5. Then for this Poisson process the estimated value of λ is 5.

Examples (Estimation)

Suppose it is known that the survival time of a particular type of bulb has the exponential distribution. You test 10 such bulbs and find their respective survival times as 150, 225, 275, 300, 95, 155, 325, 75, 20 and 400 hours respectively. Then the estimated value of θ = 202.

Multiple Linear Regression

Given ′k′ independent variables X_1, X_2,….., X_k and one dependent variable Y we predict the value of Y given by $\hat{Y}$ or y using the values of X_i ′s. We need ′n′ ($n \geq k+1$) data points and the multiple linear regression (MLR) equation is as follows:

$$Y_j = \beta_1 x_{1,j} + \beta_2 x_{2,j} + \ldots\ldots + \beta_k x_{k,j} + \varepsilon_j$$

$$\forall\ j = 1, 2, \ldots\ldots, n$$

Multiple Linear Regression

Note

- There is no randomness in measuring X_i
- The relationship is linear and not non-linear. By non-linear we mean that at least one derivative of Y wrt β_i′s is a function of at least one of the parameters. By parameters we mean the β_i′s.

Multiple Linear Regression

Assumptions for the MLR

- X_i, Y are normally distributed
- X_i are all non-stochastic
- $\varepsilon_j \sim N(0,\sigma^2 I)$
- rank(**X**) = k
- $n \geq k$
- No dependence between the X_j's, i.e., the rank of the matrix **X** is
- $E(\varepsilon_j \varepsilon_l) = 0 \qquad \forall\ i, j = 1, 2, \ldots, n$
- $Cov(X_{i,} \varepsilon_j) = 0 \qquad \forall\ i \neq j,\ i, j = 1, 2, \ldots, n$

Multiple Linear Regression

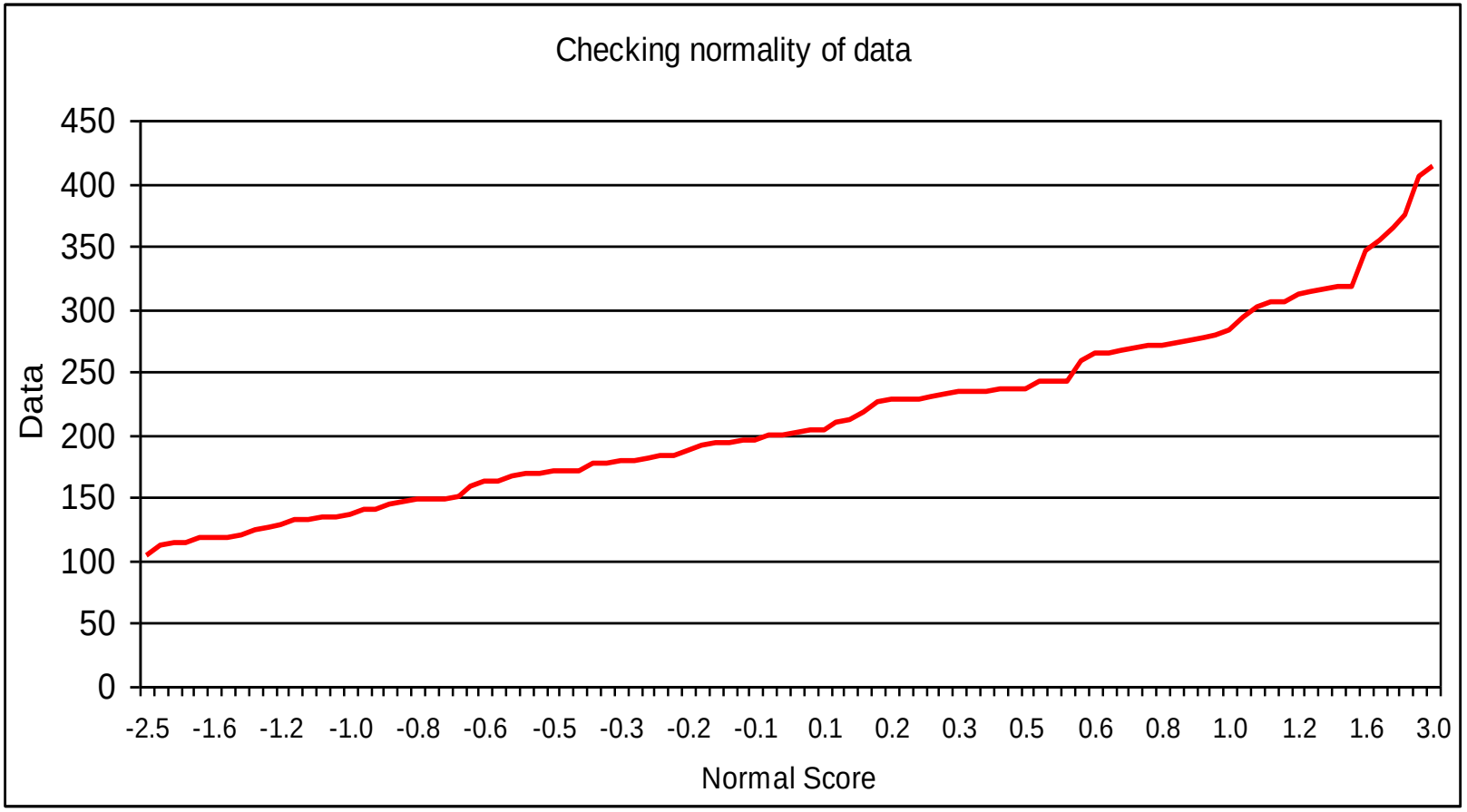
- Find $\beta_1, \beta_2, \ldots\ldots, \beta_k$ using the concept of minimizing the sum of square of errors. This is also known as least square method or method of ordinary least square. The estimates found $\hat{\beta}_1, \hat{\beta}_2, \ldots\ldots, \hat{\beta}_k$ are the estimates of $\beta_1, \beta_2, \ldots, \beta_k$ respectively.
- Utilize these estimates to find the forecasted value of Y (i.e., $\hat{Y}$ or y) and compare those with actual values of Y obtained in future.

To check for normality of data

We need to check for the normality of X_i's and Y

1) List the observation number in the column # 1,call it i.
2) List the data in column # 2.
3) Sort the data from the smallest to the largest and place in column # 3.
4) For each i^{th} of the n observations, calculate the corresponding tail area of the standard normal distribution (Z) as follows, A = (i – 0.375)/(n + 0.25). Put the values in column # 4.
5) Use NORMSINV(A) function in MS-EXCEL to produce a column of normal scores. Put these values in column # 5.
6) Make a copy of the sorted data (be sure to use paste special and paste only the values) in column # 6.
7) Make a scatter plot of the data in columns # 5 and # 6.

To check for normality of data



Basic transformation for MLR

Some transformation to convert to MLR

- $X \rightarrow X^{(p)} = \{X^p - 1\}/p$, as $p \rightarrow 0$ then $X^{(p)} = \log_e X$
- If the variability in Y increases with increasing values of Y, then we use

$$\log_e Y = \beta_1 \log_e X_1 + \beta_2 \log_e X_2 + \ldots + \beta_k \log_e X_k + \varepsilon$$

Simple linear regression

In the simple linear regression we have

$Y_j = \alpha + \beta X_j + \varepsilon_j \qquad \forall\ j = 1,2,\ldots..,n$

The question is how do we find α and β, provided we have 'n' number of observations which constitutes the sample.

We minimize the sum of square of the error wrt to α and β

$$\Delta = \sum_{j=1}^{n} \{Y_j - (\hat{\alpha} + \hat{\beta} X_j)\}^2$$

Finally:

$$E(Y) = \hat{\alpha} + \hat{\beta} E(X)$$

$$\mathrm{cov}(X,Y) + E(X)E(Y) = \hat{\alpha} E(X) + \hat{\beta}\{V(X) + E(X)^2\}$$

Simple linear regression

After we have found out the estimators of α and β, we use these values to predict/forecast the subsequent future values of Y, i.e., we find out y and compare those y′s with the corresponding values of Y. Thus we find $y_k = \hat{\alpha} + \hat{\beta} X_k$ and compare them with corresponding values of Y_k , for k = n+1, n+2,…….

Simple linear regression

