భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Biostatistics BT2023

Lecture 7
Dispersion and shape of distributions

29 August 2023

Mid range

Arithmetical mean of the smallest and largest value of the data set

7 8 9 3 4 7 7 7 9 5

3 4 5 7 7 7 7 8 9 9

Mid range = 6

Geometrical mean

nth root of the data set

$$G = (x_1 \times x_2 .... \times x_n)^{\frac{1}{n}}$$

$$antilog \ \frac{\sum log \ x_i}{n}$$

Harmonic Mean

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{\sum f_i}{\sum f_i \times \frac{1}{x_i}}$$

Example:

first 10 km 30 km/hour
next 10 km 60 km/hour

Average car speed

Useful when data is rate with respect to time

Continuous data

$$mean\ (\bar{x}) = \frac{\sum m \times f}{N}$$

$$M = L + \frac{\frac{n}{2} - C}{f} \times i$$

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f2}$$

Blood Cholesterol and Triglycerides

- A normal fasting level is 150 milligrams per deciliter (mg/dL).
- A borderline high level is 150 to 199 mg/dL.
- A high level is 200 to 499 mg/dL.

Glycerides mg/dL	200-300	300-400	400-500	500-600	600-700
Patient	5	10	20	5	3

Plot the following function

$$y = x$$

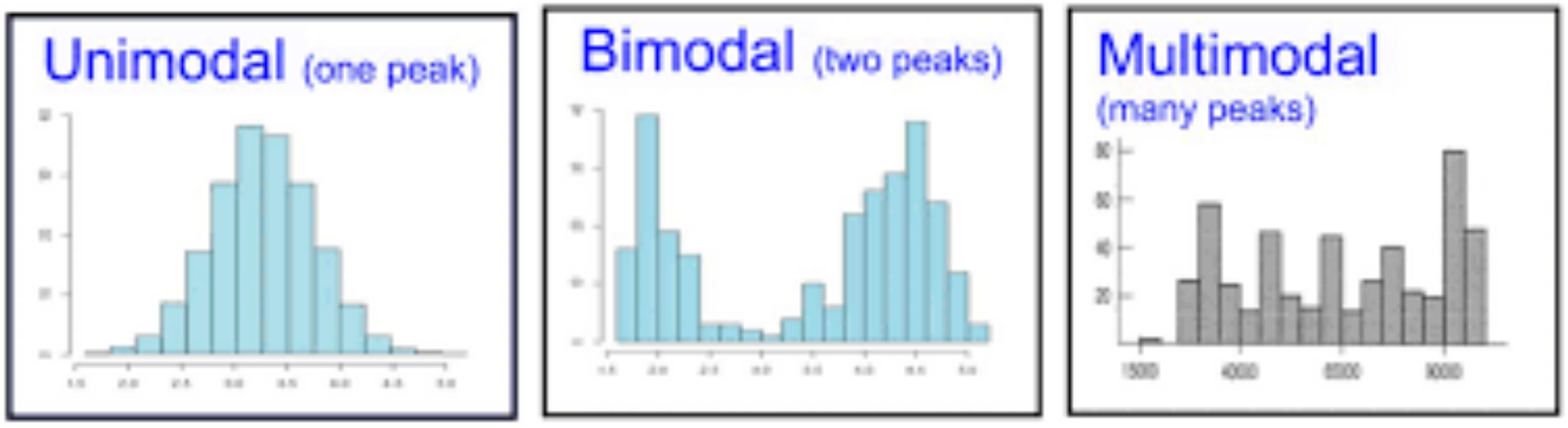
$$y = log(x)$$

$$y = e^{x}$$

$$y = e^{-x^2}$$

Sometime $\overline{x}$ is used for the the mean of the sample and μ for mean of the population

Unimodel, bimodel and multi model data



भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Mode : The grouping method

Size	Frequency
10	3
20	5
30	3
40	1
50	2
60	5
70	13
80	9
90	2

Series

Size	(I) Frequency	(II) (1+2)	(III) (2+3)	(IV) (1+2+3)	(V) (2+3+4)	(VI) (3+4+5)
10	3	8	-	11		
20	5		8		9	
30	3	4				6
40	1		3	8		
50	2	7			20	
60	5		18			27
70	**13**	**22**		**24**		
80	9		11			
90	2	-				

Mode = 70

Column	Size of items containing maximum frequency								
	10	20	30	40	50	60	70	80	90
I							✓		
II							✓	✓	
III						✓	✓		
IV							✓	✓	✓
V					✓	✓	✓		
VI						✓	✓	✓	
Total	-	-	-	-	1	3	6	3	1

Median: Merits and demerits

Demerits

1) For computing median data needs to be arranged in ascending or descending order.

2) It is not based on all the observations of the data.

3) It can not be given further algebraic treatment.

4) It is affected by fluctuation of sampling.

5) It is not accurate when the data is not large.

6) In some cases median is determined approximately as the mid-point of two observations whereas for mean this does not happen.

Useful when the data is skewed or asymmetric

Merits

1) It is easy to compute and understand.

2) It is well defined as an ideal average should be.

3) It can also be computed in case of frequency distribution with open ended classes.

4) It is not affected by extreme values and also interdependent of range or dispersion of the data.

5) It can be determined graphically.

6) It is proper average for qualitative data where items are not measured but are scored.

7) It is only suitable average when the data are qualitative & it is possible to rank various items according to qualitative characteristics.

8) It can be calculated easily by watching the data.

9) In some cases median gives better result than mean.

Mode of group data : Example

Class	100-110	110-120	120–130	130-140	140–150	150–160	160–170
Frequency	4	6	20	32	33	8	2

Class	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
100 – 110	4					
110 – 120	6	4 + 6 = 10				
120 – 130	20		20 + 6 = 26	4 + 6 + 20 = 30		
130 – 140	32	20 + 32 = (52)			6 + 20 + 32 = (58)	
140 – 150	(33)		32 + 33 = (65)			20 + 32 + 33 = (85)
150 – 160	8	33 + 8 = 41		32 + 33 + 8 = (73)		
160 – 170	2		8 + 2 = 10		33 + 8 + 2 = 43	

Class	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Total
100-110							0
110-120					1		1
120-130		1			1	1	3
130-140		1	1	1	1	1	5
140-150	1		1	1		1	4
150-160				1			1
160-170							0

$$Mode = 130 + \frac{1}{12} \times 10 = 130.83$$

Mode: Merits and demerits

Merits

- The mode is easy to understand and calculate.
- The mode is not affected by extreme values.
- The mode is easy to identify in a data set and in a discrete frequency distribution.
- The mode is useful for qualitative data.
- The mode can be computed in an open-ended frequency table.
- The mode can be located graphically.

Demerits

- The mode is not defined when there are no repeats in a data set.
- The mode is not based on all values.
- The mode is unstable when the data consist of a small number of values.
- Sometimes the data has one mode, more than one mode, or no mode at all.

Useful when data is nominal, like in business

Relationship b/w mean, median and mode

If a frequency distribution graph has a symmetrical frequency curve, then mean, median and mode will be equal.

Mode = 3 Median - 2 Mean

or

Median = (2 mean + Mode)/3

Although sometimes MEAN can be effected by the extreme values but still it is the widely used and scientific statistical treatment and interpretations

भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Measure of dispersion

15, 20, 17, 19, 21, 13, 12, 10, 17, 9, 12

Range: subtract two farthest data points

Interquartile range : Divide your data into four parts, the range between first and third quartile is you IQR.

Mean Deviation or average deviation: The arithmetical mean of the mode of all the deviations from the central values,

$$\frac{1}{n}\sum |x_i - \bar{x}|$$

From mean

$$\frac{1}{n}\sum |x_i - M_d$$

From median

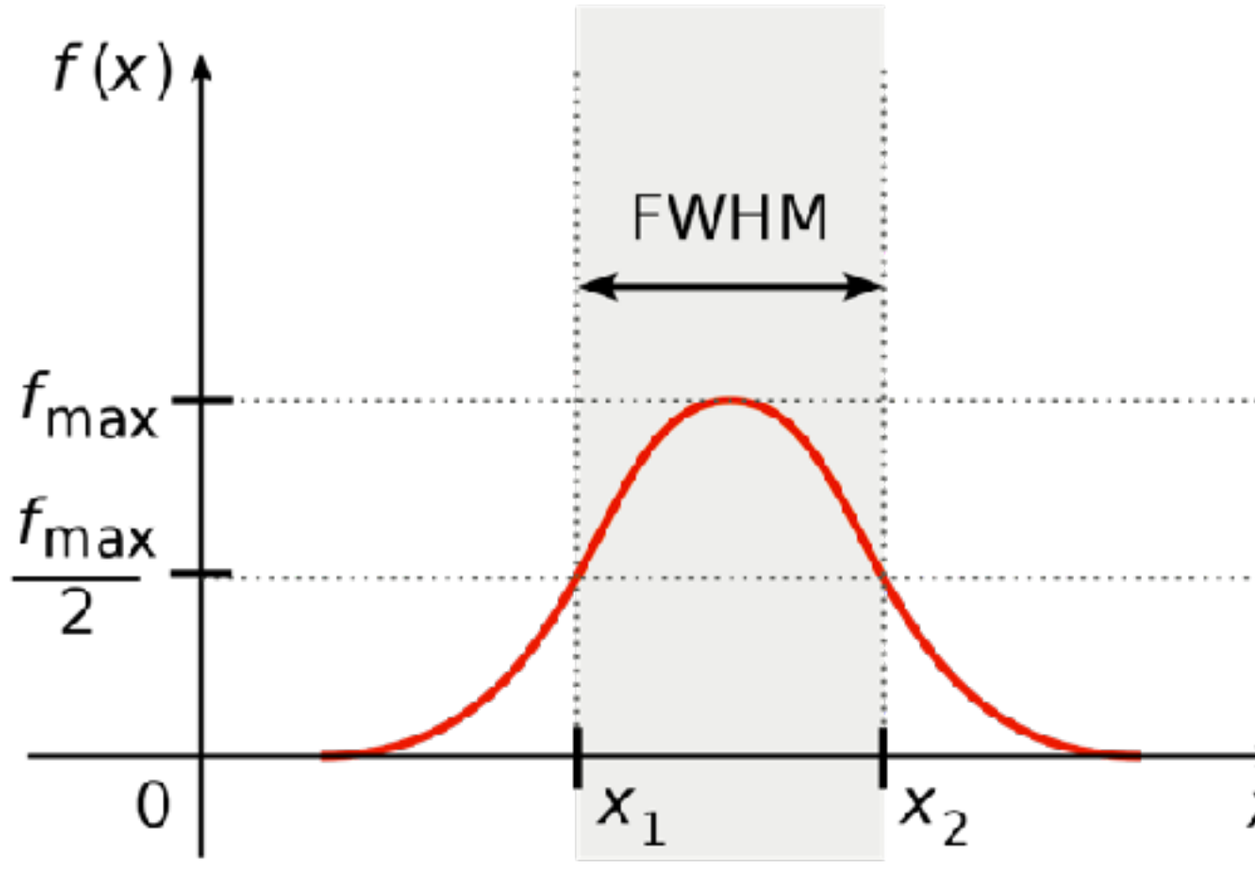
$$\frac{1}{n}\sum |x_i - Z|$$

From mod

Coefficient of mean deviation $\frac{MD_x}{\bar{x}}$

The relative measure of dispersion corresponding to the mean deviation is called coefficient of mean deviation, be it the Mean, Mode or Median

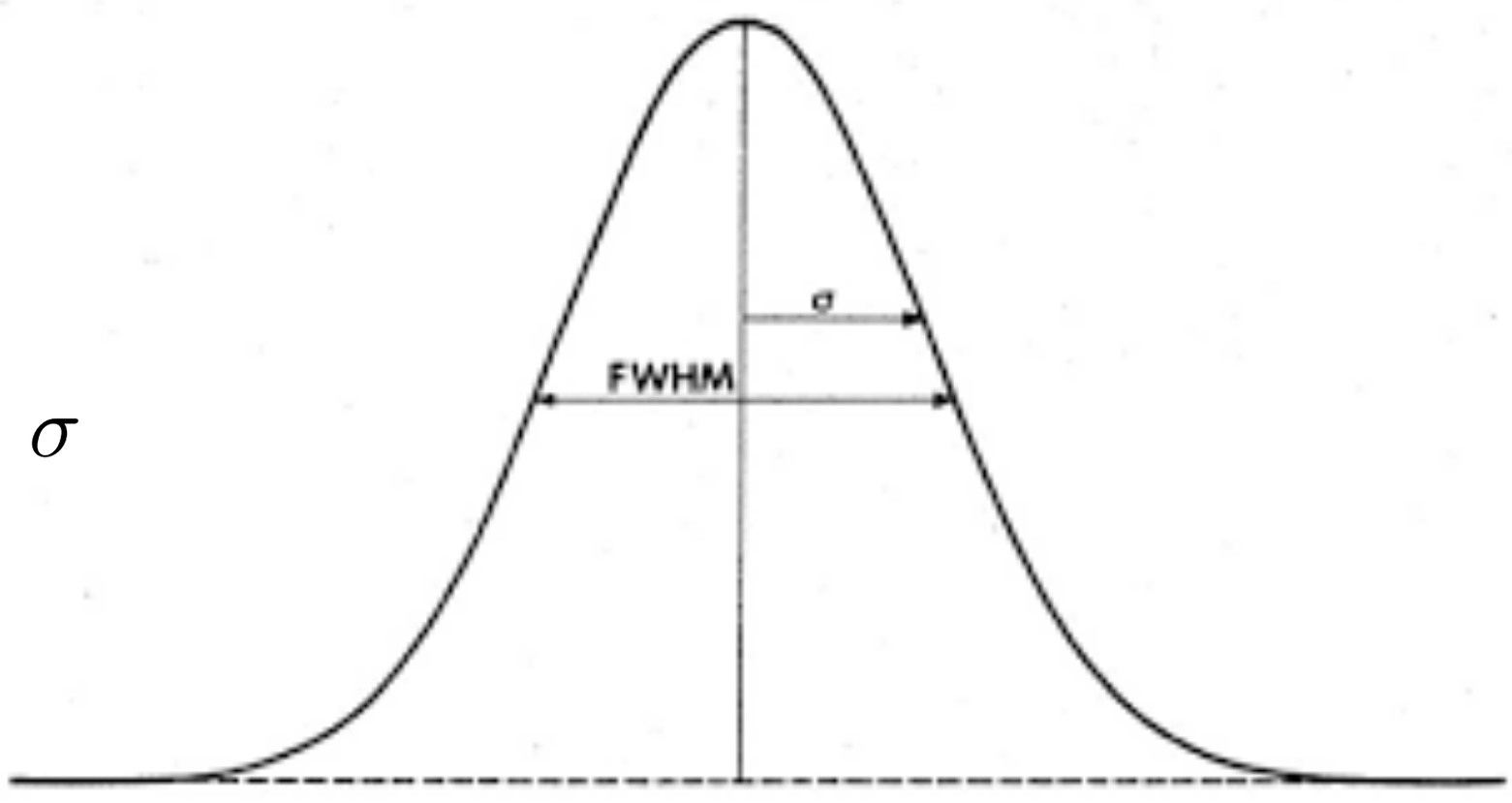
Full width half maximum



Normal or symmetrical distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp - \left[\frac{(x - x_0)^2}{2\sigma^2}\right]$$

$$FWHM = 2\sqrt{2ln2}\ \sigma$$



Plotting this function

Measurement of dispersion

Standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Coefficient of standard deviation $\frac{\sigma}{\bar{x}}$

Variance σ^2

15, 20, 17, 19, 21, 13, 12, 10, 17, 9, 12

Partition values

When a data set is required to be divided into two or three or n equal parts the dividing places are called partition values

Halves points 1

Quartiles points 3

Deciles points 9

Percentiles points 100

Quartiles

Q1 = (N+1)/4

Q2 =(N+1)/2 N is odd
or [N/2 +(N/2 +1)]/2 N is even

Q3 = 3 x (N+1)/4

Quartile values

- Step 1: **Put the numbers in order.**
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.

- Step 2: **Find the median**.
  1, 2, 5, 6, 7**, 9**, 12, 15, 18, 19, 27.

- Step 3: **Place parentheses around the numbers above and below the median.**
  Not necessary **statistically**, but it makes Q1 and Q3 easier to spot.
  (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).

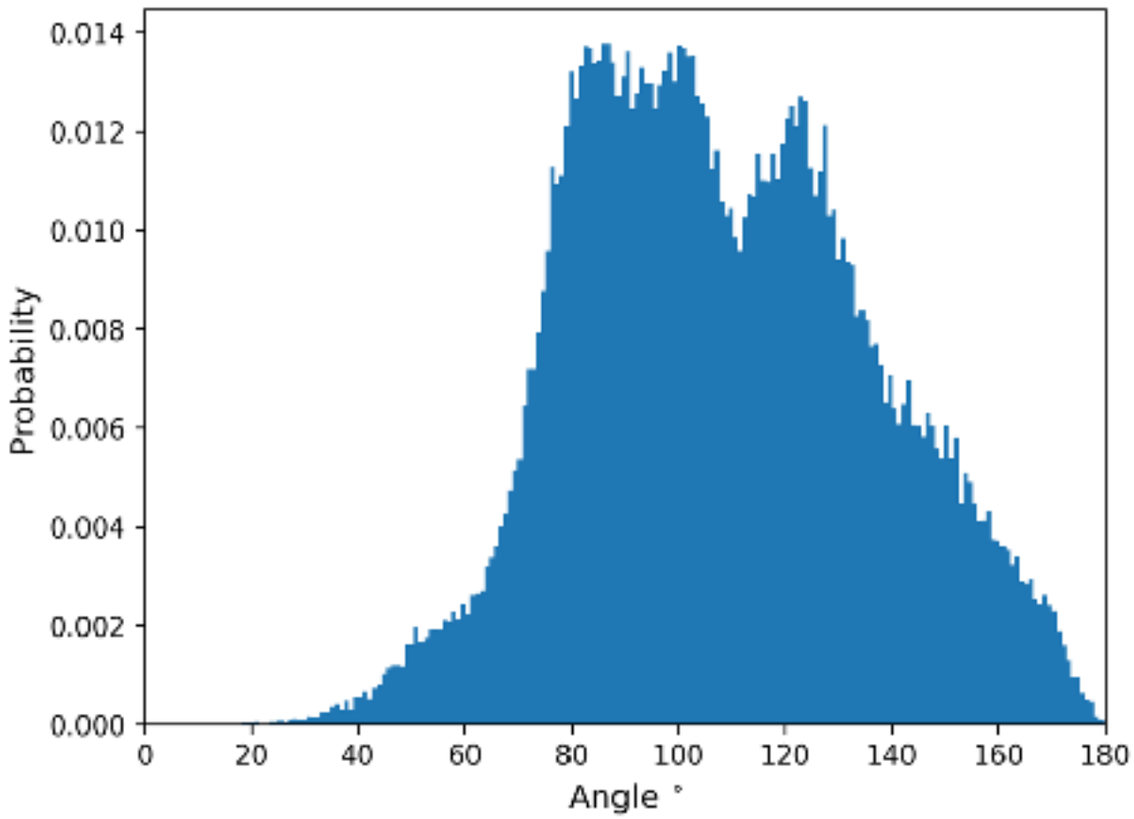
- Step 4: **Find Q1 and Q3**
  Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.
  (1, 2, **5**, 6, 7)**,** **9**, (12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.

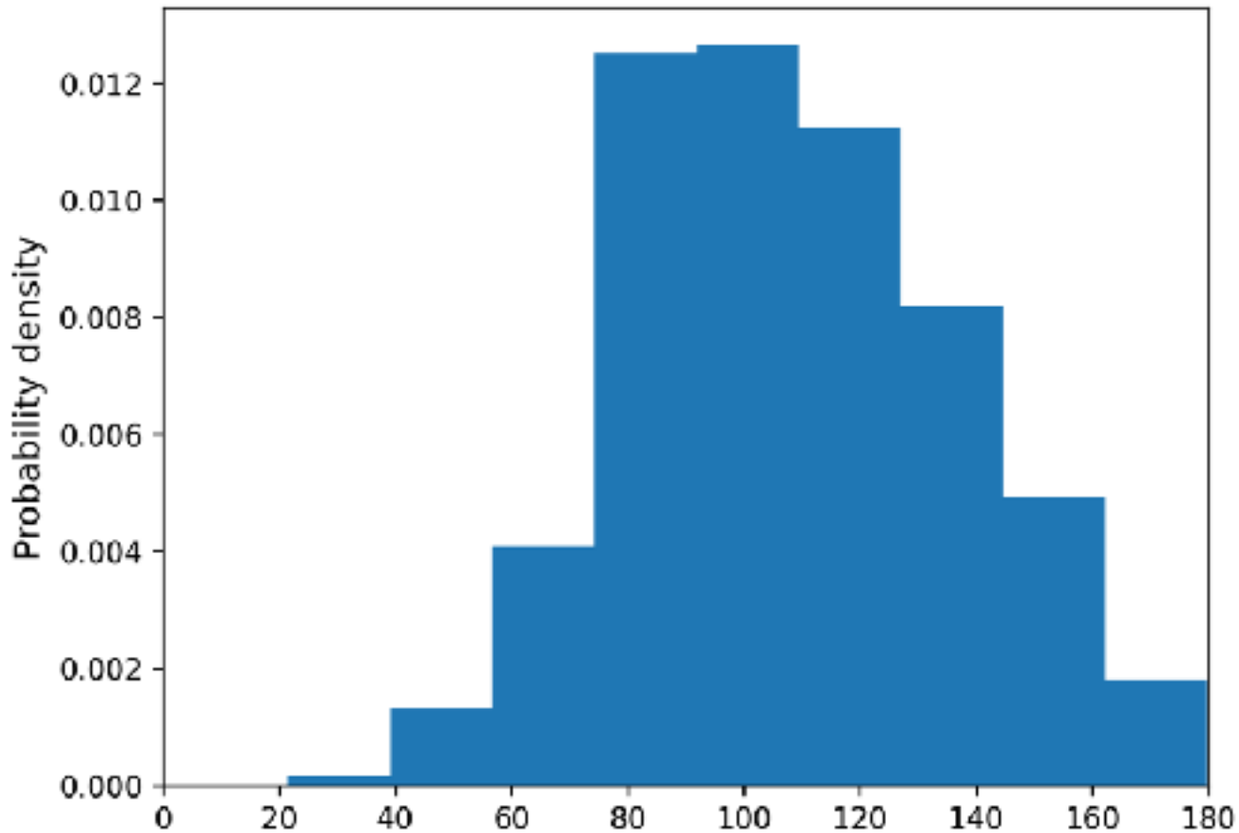
- Step 5: **Subtract Q1 from Q3 to find the interquartile range**.
  18 – 5 = 13.

भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Histogram: Choice of bin size

The most important parameter of a histogram is the bin width





$$h = 2 \times IQR \times n^{-1/3}$$

interquartile range (IQR)

Bin width

$$w = \frac{max - min}{h}$$

Biostatistics : The journal

Among the important scientific developments of the 20th century is the explosive growth in statistical reasoning and methods for application to studies of human health.

Examples include developments in likelihood methods for inference, epidemiologic statistics, clinical trials, survival analysis, and statistical genetics.

Substantive problems in public health and biomedical research have fueled the development of statistical methods, which in turn have improved our ability to draw valid inferences from data.

The objective of Biostatistics is to advance statistical science and its application to problems of human health and disease, with the ultimate goal of advancing the public's health.

Biostatistics is an online only journal publishing papers that develop innovative statistical methods with applications to the understanding of human health and disease, including basic biomedical sciences.

Papers should focus on methods and applications.

Introduction of original methodology should be grounded in substantive problems; there is the opportunity to present extensive analyses of data on the journal's website as supplementary material.

Authors are strongly encouraged to submit code supporting their publications. Authors should submit a link to a Github repository and to a specific example of the code on a code archiving service such as Figshare or Zenodo.

Reference: Oxford university press

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Next Class

2:30 PM Friday, 1 September 2023

Measure of dispersions