



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

# Biostatistics BT2023

## Lecture 5: Measure of central tendency

Himanshu Joshi  
16 August 2022

# Computer virus

How powerful computer and the program that runs these computers !

Just a few lines code can effect our lives in so many ways.

## Examples

Story of Stuxnet

Power grid failure in Mumbai

Organization of the class

# 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.

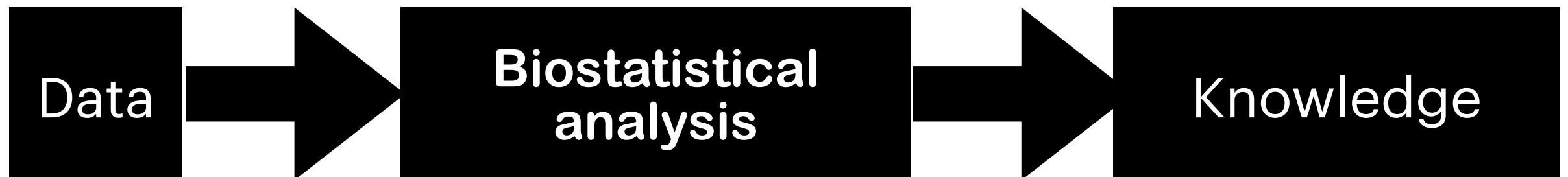
Introduction of original methodology should be grounded in substantive problems; there is the opportunity to present extensive analyses of data on the journal's website as [supplementary material](#).

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](#).



## Goal

Advancing health science research, education, and practice by turning data into knowledge and addressing the greatest public health issues of the 21st century.



Biostatistics department of Harvard university



## Measure of central tendency

**Example:** the average height of the students in class

### Arithmetical mean

$$\bar{x} = \sum_{i=1}^n x_i$$

### Arithmetical mean of a grouped data

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

## Arithmetical mean by transformation

$$y = ax \quad ; \quad \langle y \rangle = a \langle x \rangle$$

$$y = a + x \quad ; \quad \langle y \rangle = a + \langle x \rangle$$

$$y = a + cx \quad ; \quad \langle y \rangle = a + c \langle x \rangle$$

We can use above transformation to calculate mean by what is called “step deviation method”

$$\bar{x} = a + \frac{\sum_{i=1}^n f_i d_i}{\sum_{i=1}^n f_i}$$

$$d = x - a$$

$$a = \text{provisional mean} \quad \bar{x} = 70 + \frac{1}{8} = 70.125$$

$l$  = common size of class interval

$x$	$d$
45	-25
50	-20
60	-10
67	-3
74	4
80	10
90	20
95	25



## Mean for continuous data

Class Interval	0 - 8	8 - 16	16 - 24	24 - 32	32 - 40	40 - 48
Frequency	10	20	14	16	18	22

Class Intervals	Class Mark ( $m_i$ )	Frequency ( $f_i$ )	$d_i = m_i - a = m_i - 28$	$d_i' = \frac{d_i}{l} = \frac{d_i}{8}$	$d_i' f_i$
0 - 8	4	10	-24	-3	-30
8 - 16	12	20	-16	-2	-40
16 - 24	20	14	-8	-1	-14
24 - 32	28	16	0	0	0
32 - 40	36	18	8	1	18
40 - 48	44	22	16	2	44

$$\sum f_i = 100$$

$$\sum d_i' f_i = -22$$

$$\bar{x} = a + l \times \frac{\sum_{i=1}^n f_i d_i'}{\sum_{i=1}^n f_i}$$

$$d_i' = d_i / l$$

$$28 + 8 \times \frac{-22}{100} = 26.24$$

**Its approximation since we don't know the raw data**



## Median

Arrange the data (n) in the ascending order

$$Median = \frac{N^{th}}{2} term$$

N= odd

$$Median = \frac{\frac{N}{2}^{th} + (\frac{N}{2} + 1)^{th} terms}{2}$$

N= even

45. 50. 56 57 74 80 90 95



## Mode

The highest frequency data point in the data set

Mark obtained      7 8 9 3 4 7 7 7 9 5

Mode = 4

Mode of Grouped Data =  $L + W[(F_m - F_1) / ((F_m - F_1) + (F_m - F_2))]$

Range	Frequency
1-10	2
11-20	7
21-30	10
31-40	3
41-50	1

$L = 21$

$W = 9$

$F = 10$

$F_1 = 7$

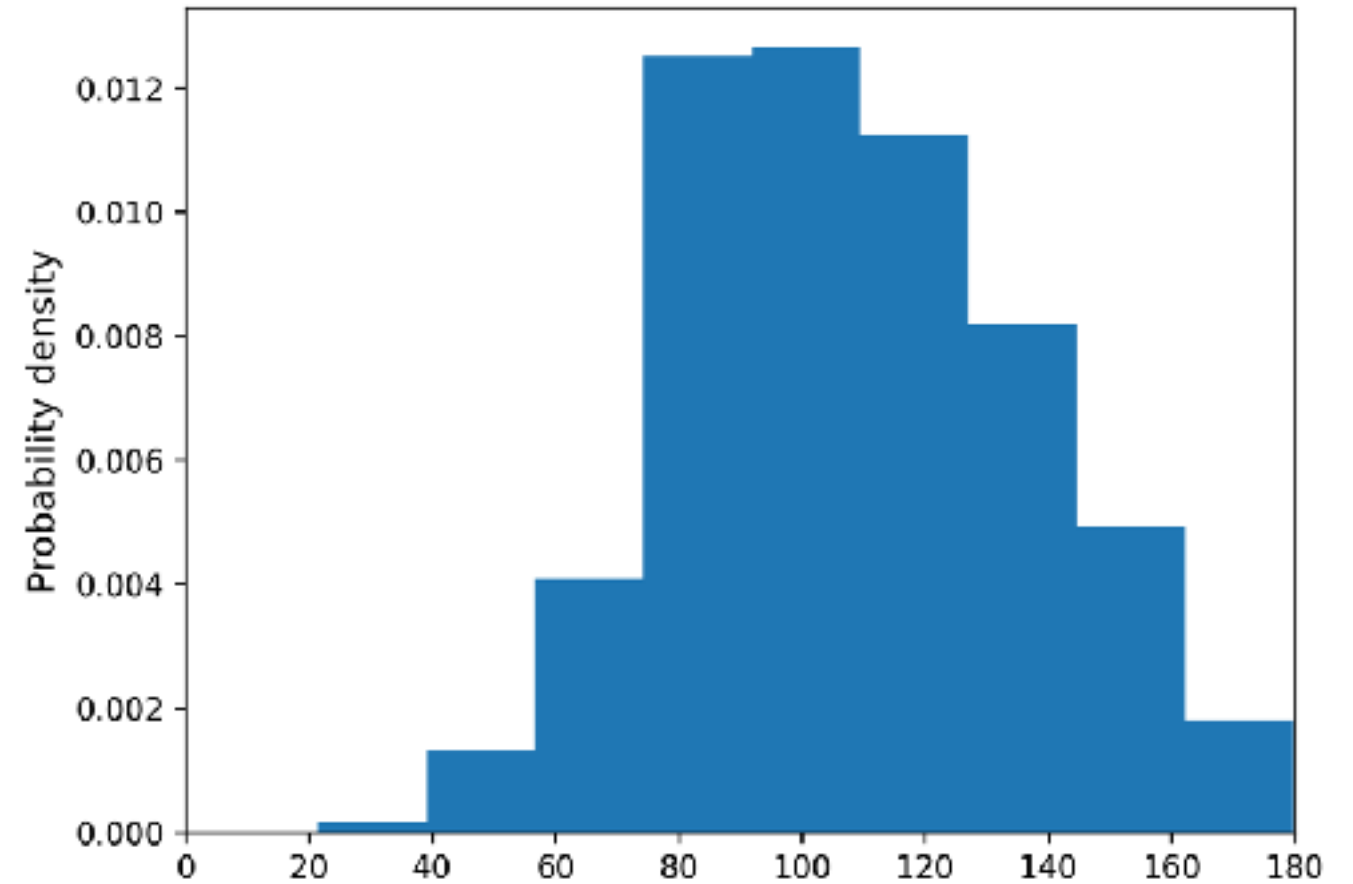
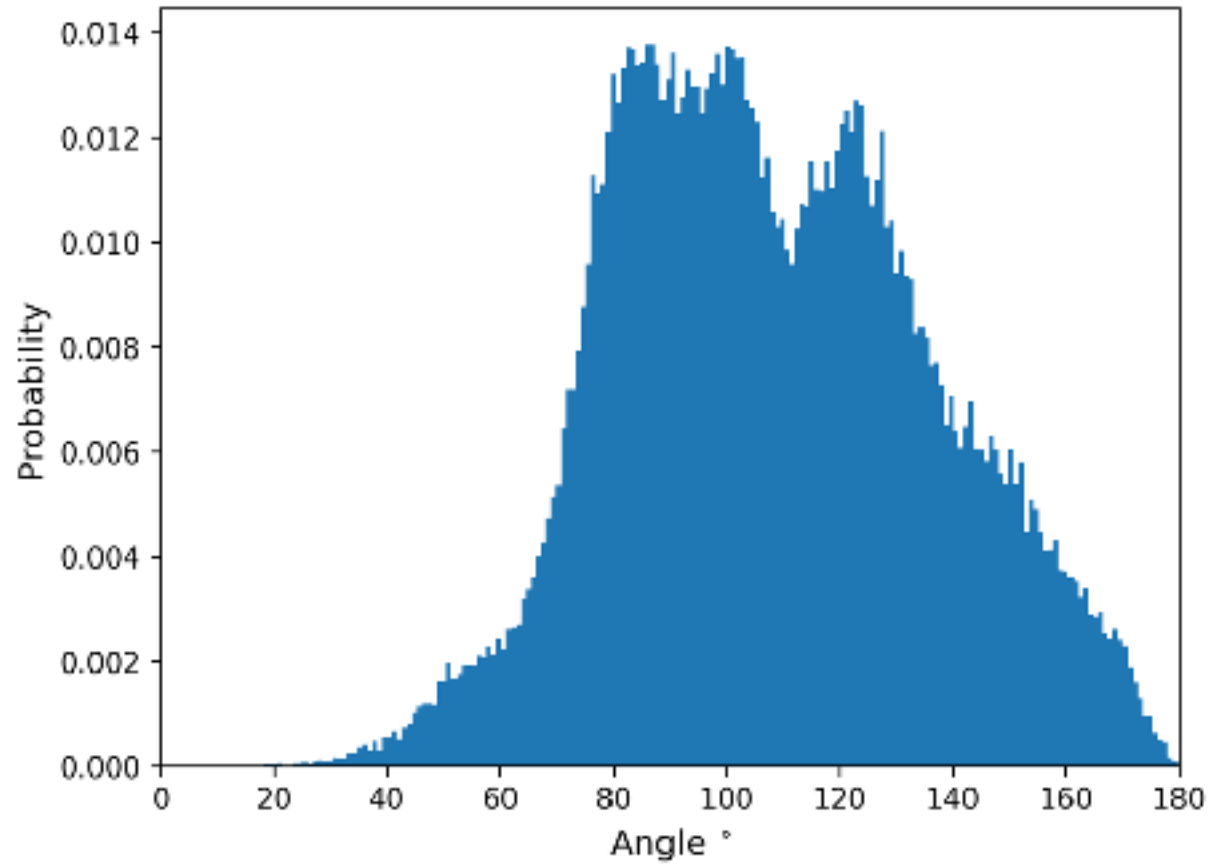
$F_2 = 3$

Mode = 24



# Histogram

## Plot of the probability distribution



How to chose the bin



# Useful commands in data analysis bioinformatics

## SED

```
sed -e 's/\(.*\) /\L\1/'
```

## pr mts

```
pr -mts ' ' col1.dat col2.dat >combine.dat
```

## grep

```
grep -A1 'blah' logfile
```

## Awk

### column to row

```
awk 'BEGIN { ORS = " " } { print }' id2.dat
```

### row to column

```
awk '{for (i=1;i<=NF;i++) print $i}' 2.dat >22.dat
```

# Introduction to “github”

GitHub is a website for developers and programmers to collaboratively work on code



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

## Next Class

**2:30 PM Friday, 23 August 2022**

**Measure of dispersions**