Statistical Methods & Data Analysis (MA 231)

by

Chanchal Kundu

Department of Mathematical Sciences

(गणितीय विज्ञान विभाग)

Rajiv Gandhi Institute of Petroleum Technology, Jais

(राजीव गांधी पेट्रोलियम प्रौद्योगिकी संस्थान, जायस)

Lecture # 11-16

Descriptive Statistics

and

Data Analysis

with

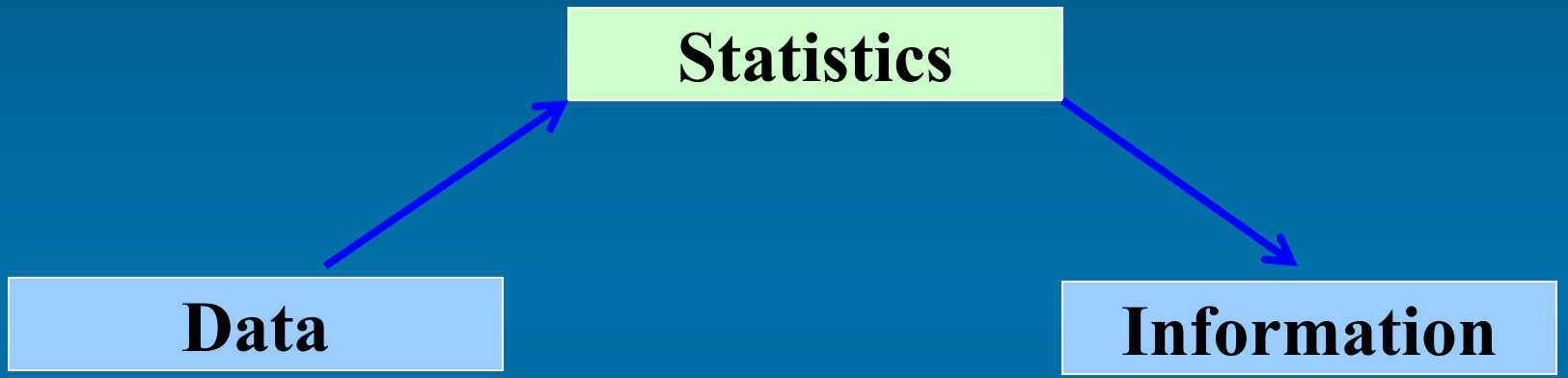
R:

Dr. Chanchal Kundu

RGIPT, Jais Campus

What is Statistics ?

"Statistics is a way to get information from data"



Statistics is a *tool* for creating *new understanding* from a set of numbers.

Where is Wisdom?

We have lost in Knowledge.

Where is Knowledge?

We have lost in Information.

Where is Information?

We have lost in Data.

Where is **Data**?.....

Source of Data

- Population
- Sample

Different Sampling Techniques

- Simple Random Sampling
- Stratified Random Sampling
- Cluster Sampling
- Systematic Sampling
- Convenience Sampling
- Judgment Sampling

Simple Random Sampling: Finite Population

- Finite populations are often defined by lists such as:
  - Organization membership roster
  - Credit card account numbers
  - Inventory product numbers

- A simple random sample of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

Simple Random Sampling: Finite Population

- Replacing each sampled element before selecting subsequent elements is called sampling with replacement.
- Sampling without replacement is the procedure used most often.
- In large sampling projects, computer-generated random numbers are often used to automate the sample selection process.

Simple Random Sampling: Infinite Population

- Infinite populations are often defined by an ongoing process whereby the elements of the population consist of items generated as though the process would operate indefinitely.

- A simple random sample from an infinite population is a sample selected such that the following conditions are satisfied.
  - Each element selected comes from the same population.
  - Each element is selected independently.

Simple Random Sampling: Infinite Population

- In the case of infinite populations, it is impossible to obtain a list of all elements in the population.
- The random number selection procedure cannot be used for infinite populations.

Stratified Random Sampling

- The population is first divided into groups of elements called strata.
- Each element in the population belongs to one and only one stratum.
- Best results are obtained when the elements within each stratum are as much alike as possible (i.e. a homogeneous group).

Stratified Random Sampling

- A simple random sample is taken from each stratum.
- Formulas are available for combining the stratum sample results into one population parameter estimate.
- Advantage: If strata are homogeneous, this method is as "precise" as simple random sampling but with a smaller total sample size.
- Example: The basis for forming the strata might be department, location, age, industry type, and so on.

Cluster Sampling

- The population is first divided into separate groups of elements called clusters.
- Ideally, each cluster is a representative small-scale version of the population (i.e. heterogeneous group).
- A simple random sample of the clusters is then taken.
- All elements within each sampled (chosen) cluster form the sample.

Cluster Sampling

- Example: A primary application is area sampling, where clusters are city blocks or other well-defined areas.
- Advantage: The close proximity of elements can be cost effective (i.e. many sample observations can be obtained in a short time).
- Disadvantage: This method generally requires a larger total sample size than simple or stratified random sampling.

Systematic Sampling

- If a sample size of n is desired from a population containing N elements, we might sample one element for every n/N elements in the population.
- We randomly select one of the first n/N elements from the population list.
- We then select every n/Nth element that follows in the population list.

Systematic Sampling

- This method has the properties of a simple random sample, especially if the list of the population elements is a random ordering.
- Advantage: The sample usually will be easier to identify than it would be if simple random sampling were used.
- Example: Selecting every 100^{th} listing in a telephone book after the first randomly selected listing

Convenience Sampling

- It is a nonprobability sampling technique. Items are included in the sample without known probabilities of being selected.
- The sample is identified primarily by convenience.
- Example: A professor conducting research might use student volunteers to constitute a sample.

Convenience Sampling

- Advantage: Sample selection and data collection are relatively easy.
- Disadvantage: It is impossible to determine how representative of the population the sample is.

Judgment Sampling

- The person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population.
- It is a nonprobability sampling technique.
- Example: A reporter might sample three or four senators, judging them as reflecting the general opinion of the senate.

Judgment Sampling

- Advantage: It is a relatively easy way of selecting a sample.
- Disadvantage: The quality of the sample results depends on the judgment of the person selecting the sample.

Types of Data & Information

Data (at least for purposes of Statistics) fall into three main groups:

Interval Data

Nominal Data

Ordinal Data

Interval Data…..

Interval data

- Real numbers, i.e. heights, weights, prices, etc.
- Also referred to as **quantitative** or **numerical**.

Arithmetic operations can be performed on Interval Data.

Nominal Data…

Nominal Data

- The values of **nominal** data are ***categories.***

E.g. responses to questions about marital status, coded as:

Single = 1, Married = 2, Divorced = 3, Widowed = 4

These data are **categorical** in nature; arithmetic operations don't make any sense (e.g. does Widowed ÷ 2 = Married?!)

Nominal data are also called **qualitative** or **categorical**.

Ordinal Data…

Ordinal Data appear to be categorical in nature, but their values have an ***order***; a ranking to them:

E.g. College course rating system:

poor = 1, fair = 2, good = 3, very good = 4, excellent = 5

While its still not meaningful to do arithmetic on this data (e.g. does 2*fair = very good?!), we can say things like:

`excellent > poor` or `fair < very good`

That is, order is maintained no matter what numeric values are assigned to each category.

Hierarchy of Data…

Interval

Values are real numbers.

All calculations are valid.

Data may be treated as ordinal or nominal.

Nominal

Values are the arbitrary numbers that represent categories.

Only calculations based on the frequencies of occurrence are valid.

Data may not be treated as ordinal or interval.

Ordinal

Values must represent the ranked order of the data.

Calculations based on an ordering process are valid.

Data may be treated as nominal but not as interval.

Entering & Visualization of data in R

- Source of R:- https://cran.r-project.org

Manual:- *http://www.cran.r-project.org/doc/manuals*

- Entering Raw Data

x<-c(raw data separated by comma)

x= c(raw data separated by comma)

print(x)

- Visualization of Data

Histogram & Bar Chart

Curve Plotting

Pie Chart

Some Characteristics of Data

- **Measures of Central Tendency/ Location**
- **Measures of Dispersion/ Variability**
- **Skewness**
- **Kurtosis**

Measures of Central Tendency/ Location

- Mean
- Median
- Mode
- Quartiles

If the measures are computed for data from a sample, they are called sample statistics.

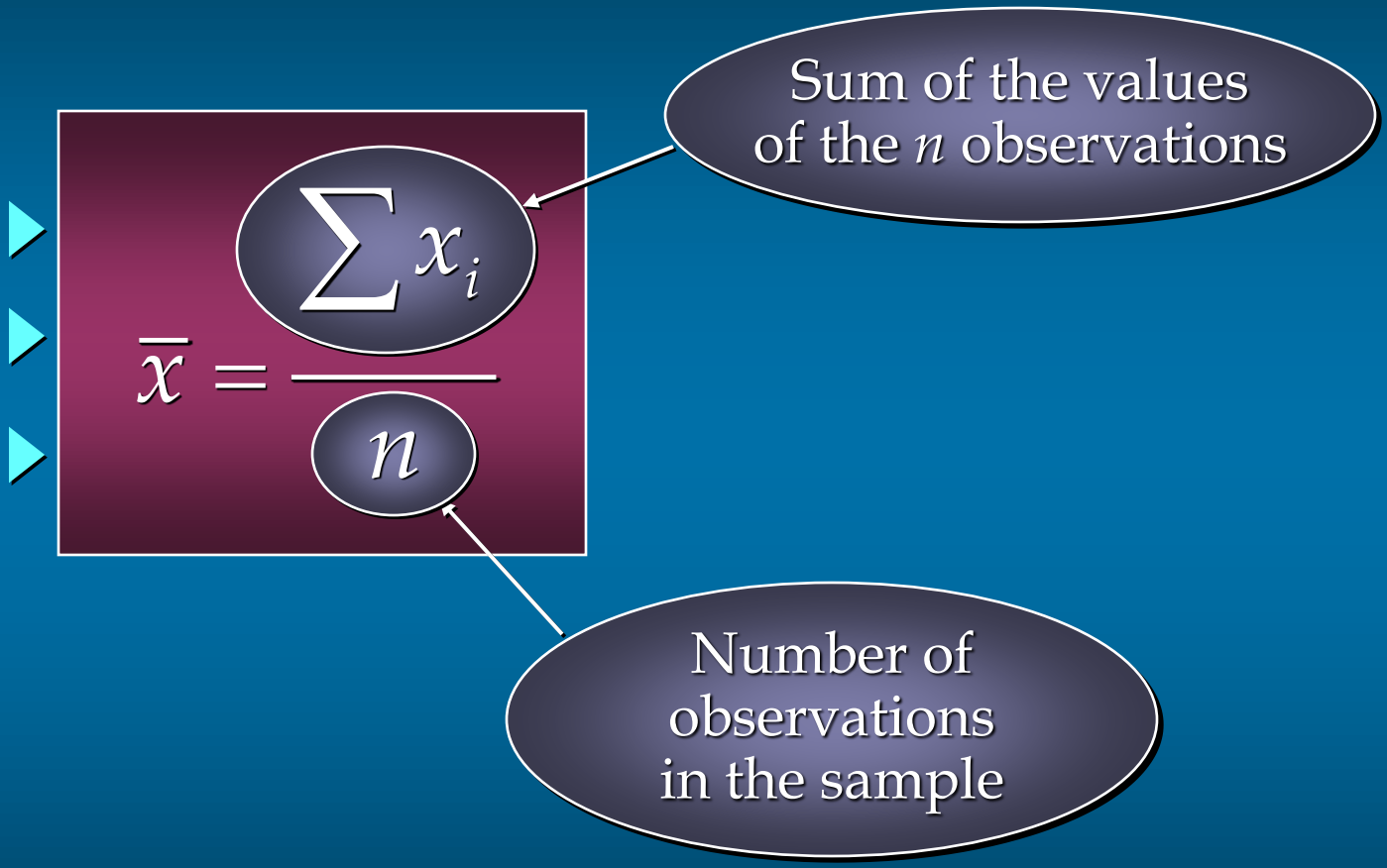
If the measures are computed for data from a population, they are called population parameters.

A sample statistic is referred to as the point estimator of the corresponding population parameter.

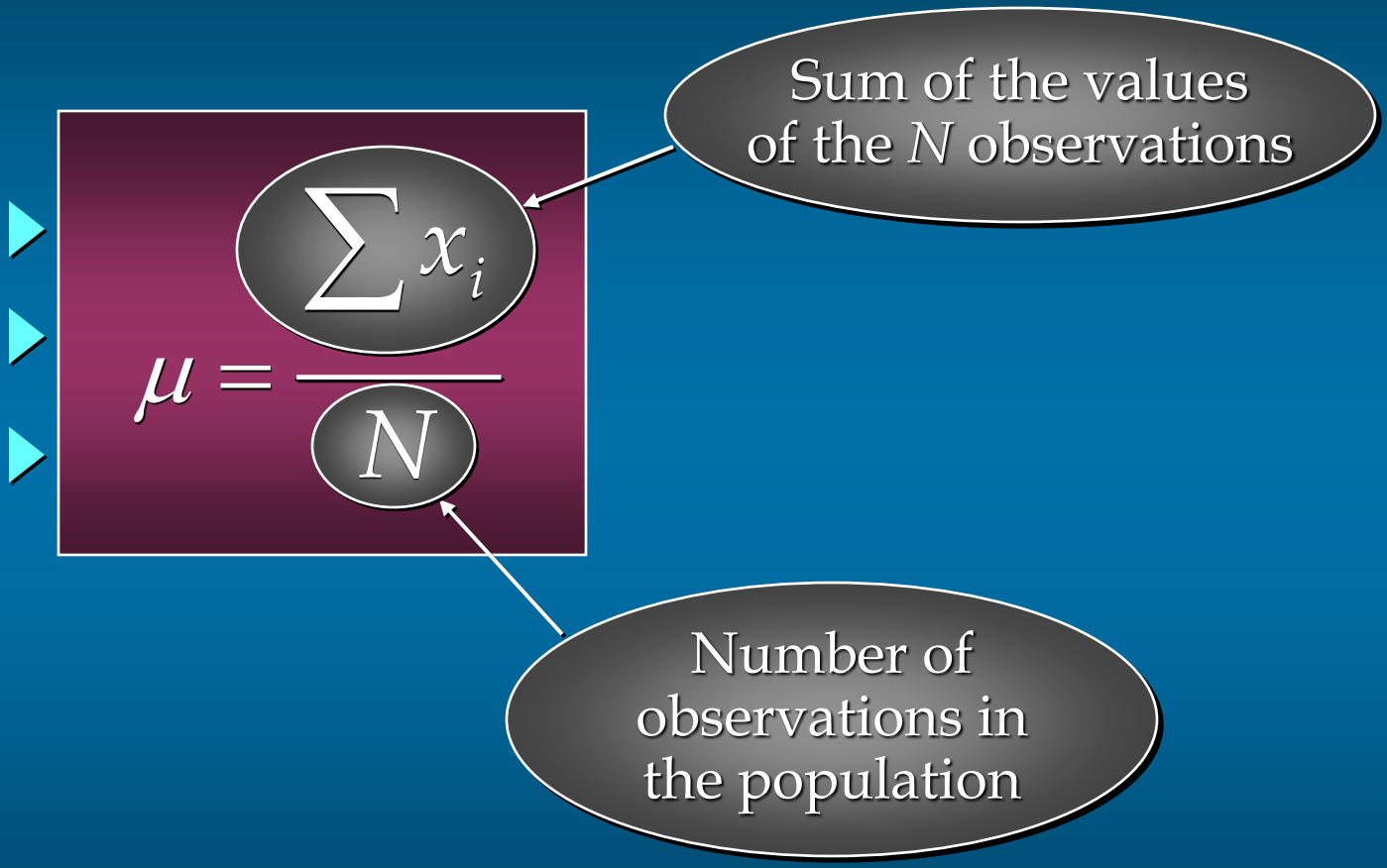
Mean

- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean μ.

Sample Mean $\bar{x}$



Population Mean μ



Sample Mean

- Example: Apartment Rents

  Seventy efficiency apartments were randomly sampled in a small college town. The monthly rent prices for these apartments are listed in ascending order on the next slide.

Sample Mean

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Sample Mean

$$\bar{x} = \frac{\sum x_i}{n} = \frac{34,356}{70} = 490.80$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Mean in R

- Find the mean of the following miles per gallon (mpg) obtained in 20 test runs performed on urban roads with an intermediate-size car:

19.7, 21.5, 22.5, 22.2, 22.6, 21.9, 20.5, 19.3, 19.9, 21.7, 22.8, 23.2, 21.4, 20.8, 19.4, 22.0, 23.0, 21.1, 20.9, 21.3

- Solution

xmpg = c(19.7, 21.5, 22.5, 22.2, 22.6, 21.9, 20.5, 19.3, 19.9, 21.7, 22.8, 23.2, 21.4, 20.8, 19.4, 22.0, 23.0, 21.1, 20.9, 21.3)

```
M1=mean(xmpg)
print(M1)
OR
print(mean(xmpg))
```

Median

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, the median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Median

- For an odd number of observations:

26	18	27	12	14	27	19

7 observations

12	14	18	19	26	27	27

in ascending order

the median is the middle value.

Median = 19

Median

- For an even number of observations:

26	18	27	12	14	27	30	19

8 observations

12	14	18	19	26	27	27	30

in ascending order

the median is the average of the middle two values.

Median = (19 + 26)/2 = 22.5

Median

- Averaging the 35th and 36th data values:
- Median = (475 + 475)/2 = 475

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Median in R

- Find the mean & Median of the Apartment rent data.

Solution:

```
rent=c(70 data separated by comma)
x<-rent
M2=median(x)
print(M2)
OR
print(median(rent))
```

Mode

- The mode of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are bimodal.
- If the data have more than two modes, the data are multimodal.

Mode

- 450 occurred most frequently (7 times)

Mode = 450

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Mean, Median, Mode: Which is Best?

With three measures from which to choose, which one should we use?

The mean is generally our first selection. However, there are several circumstances when the median is better.

The mode is seldom the best measure of central location.

One advantage the median holds is that it not as sensitive to extreme values as is the mean.

Mean, Median, & Modes for Ordinal & Nominal Data

For ordinal and nominal data the calculation of the mean is not valid.

Median is appropriate for ordinal data.

For nominal data, a mode calculation is useful for determining highest frequency but not “central location”.

Measures of Dispersion/ Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.

Measures of Variability

- Range
- Interquartile Range
- **Variance**
- **Standard Deviation**
- **Coefficient of Variation**

Variance

- The variance is a measure of variability that utilizes all the data.
- It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, μ for a population).

Variance

The variance is the average of the squared differences between each data value and the mean.

The variance is computed as follows:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

for a sample

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

for a population

Standard Deviation

- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily interpreted than the variance.

Standard Deviation

The standard deviation is computed as follows:

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

for a sample

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

for a population

Interpreting Standard Deviation…

The standard deviation can be used to compare the variability of several distributions and make a statement about the general shape of a distribution. If the histogram is **bell shaped**, we can use the ***Empirical Rule,*** which states:

1) Approximately 68% of all observations fall within one standard deviation of the mean.
2) Approximately 95% of all observations fall within two standard deviations of the mean.
3) Approximately 99.7% of all observations fall within three standard deviations of the mean.

Coefficient of Variation

The coefficient of variation indicates how large the standard deviation is in relation to the mean.

The coefficient of variation is computed as follows:

$$\left(\frac{s}{\bar{x}} \times 100\right)\%$$

for a sample

$$\left(\frac{\sigma}{\mu} \times 100\right)\%$$

for a population

Measures of Location & Dispersion in R

x=c(70 data separated by comma)

Function	Result
mean(x)	Sample mean
median(x)	Sample median
range(x)	Range
IQR(x)	Interquartile range
summary(x)	Min. Q1 Median Q3 Max.
var(x)	Sample variance
sd(x)	Sample standard deviation

Variance, Standard Deviation, and Coefficient of Variation

- Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 2,996.16$$

- Standard Deviation

$$s = \sqrt{s^2} = \sqrt{2996.47} = 54.74$$

the standard deviation is about 11% of of the mean

- Coefficient of Variation

$$\left(\frac{s}{\bar{x}} \times 100\right)\% = \left(\frac{54.74}{490.80} \times 100\right)\% = 11.15\%$$

Example

The scores of Sachin and Dhoni in ten innings during a certain season are as under:

Sachin	32	28	47	63	71	39	10	60	96	14
Dhoni	19	31	48	53	67	90	10	62	40	80

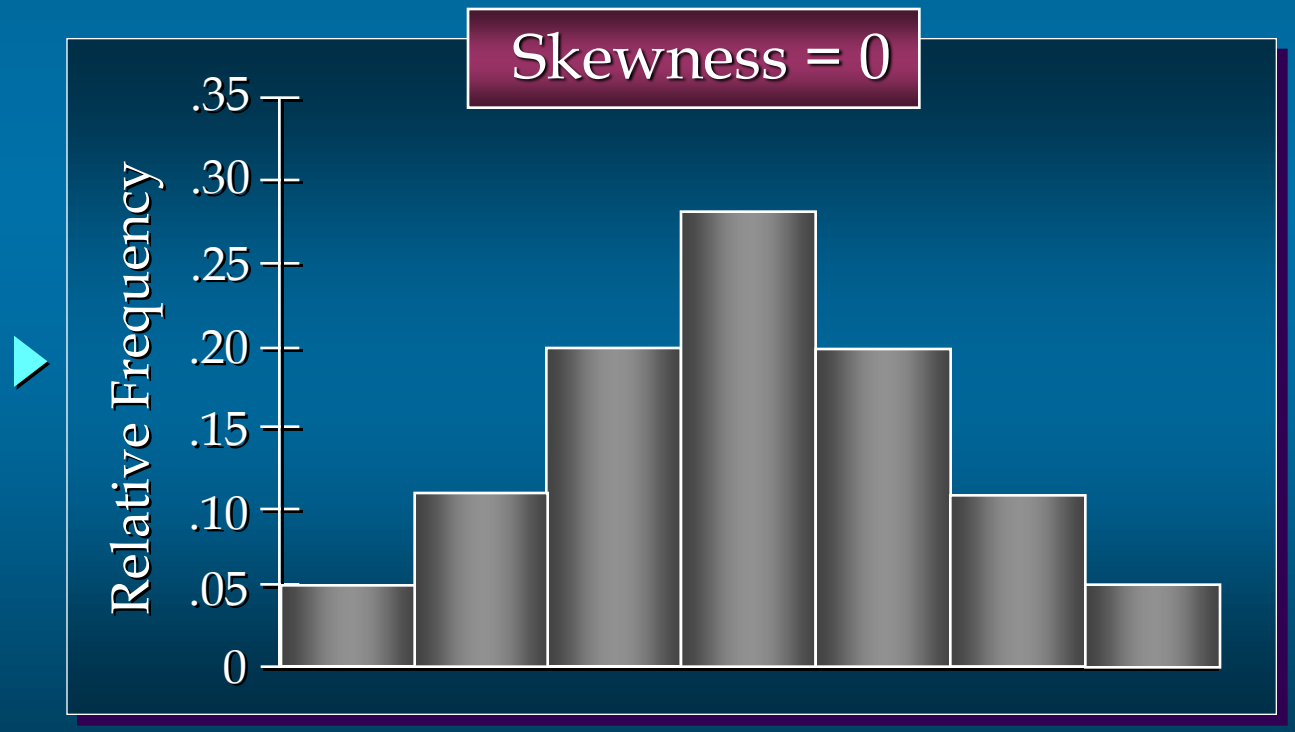
Find who is more consistent in scoring.

Distribution Shape: Skewness

- An important measure of the shape of a distribution is called skewness.
- The formula for computing skewness for a data set is somewhat complex.
- Skewness can be easily computed using statistical software.

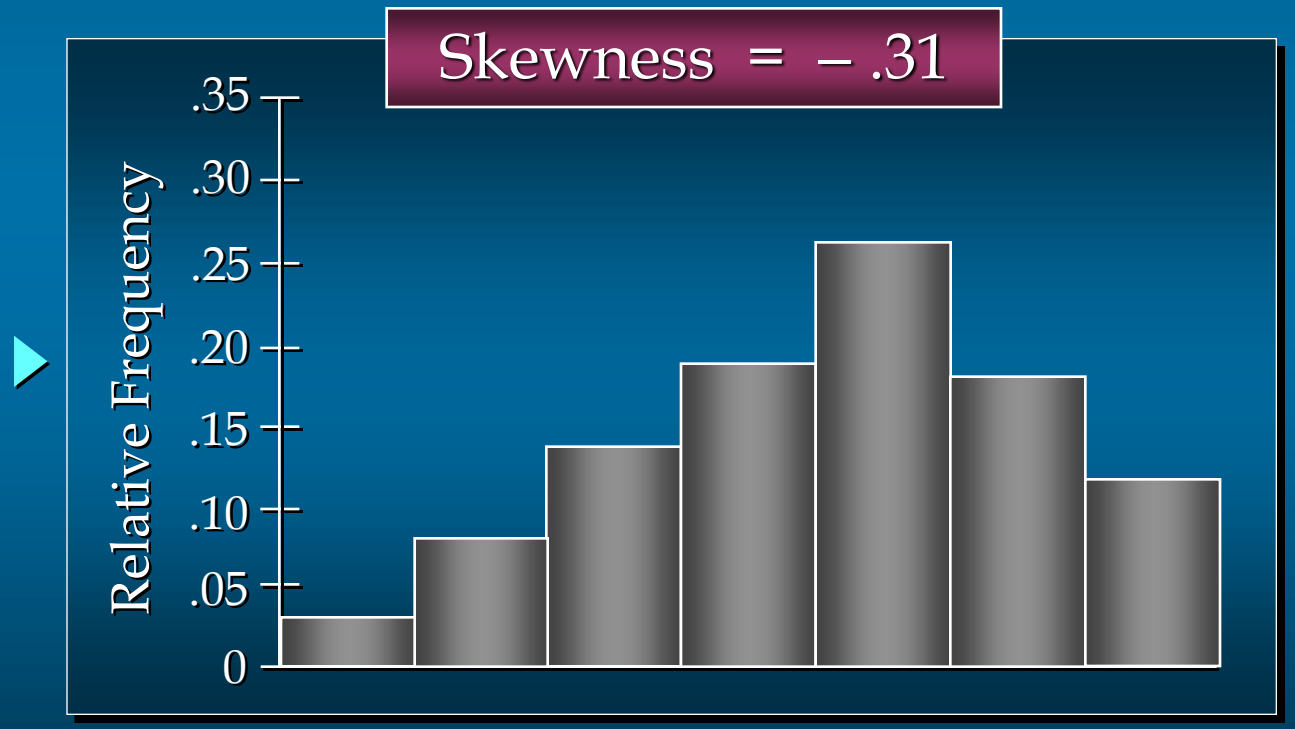
Distribution Shape: Skewness

- Symmetric (not skewed)
  - Skewness is zero.
  - Mean and median are equal.



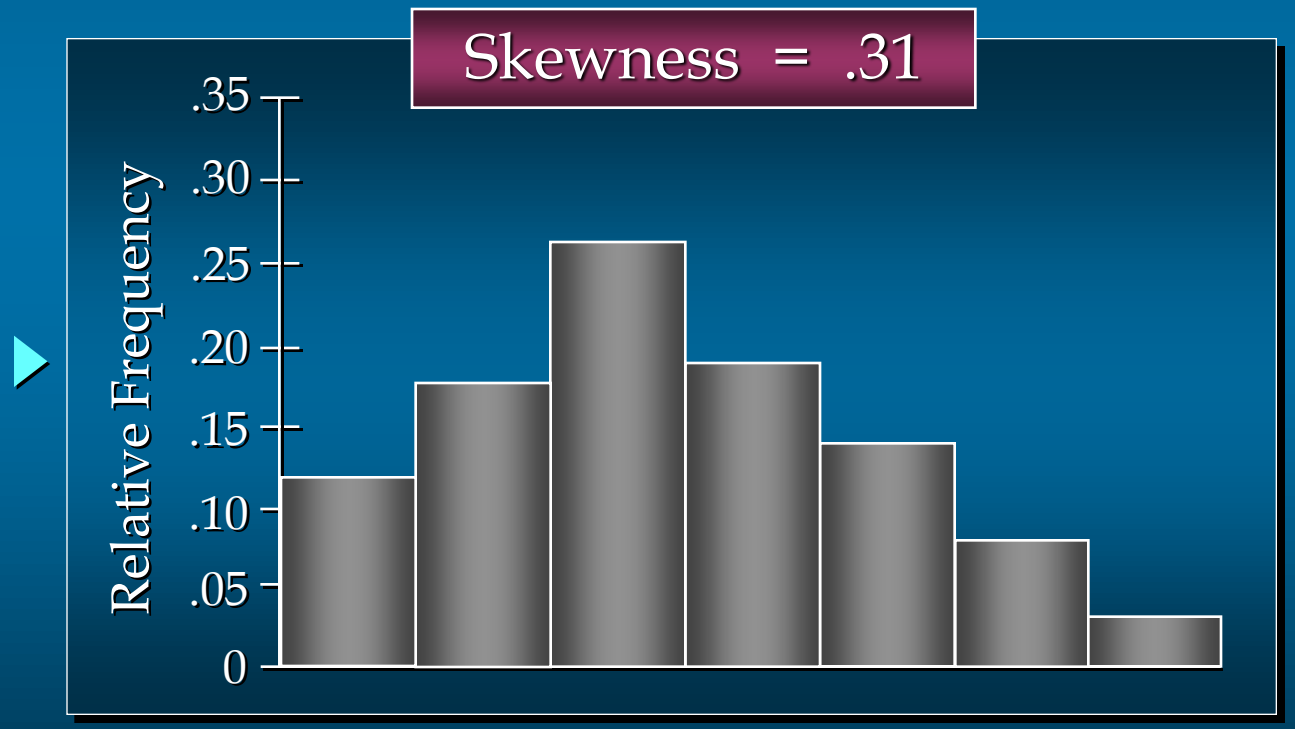
Distribution Shape: Skewness

- Moderately Skewed Left
  - Skewness is negative.
  - Mean will usually be less than the median.



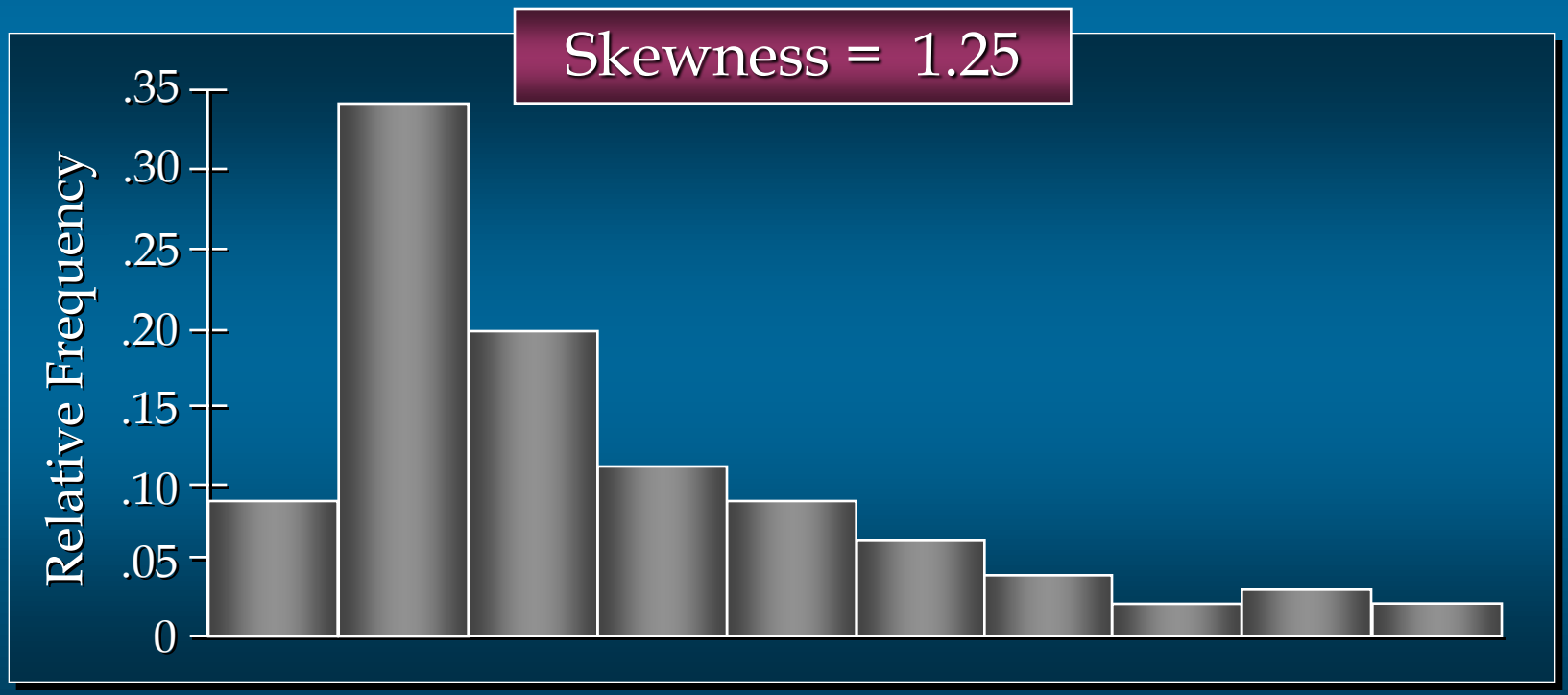
Distribution Shape: Skewness

- Moderately Skewed Right
  - Skewness is positive.
  - Mean will usually be more than the median.



Distribution Shape: Skewness

- Highly Skewed Right
  - Skewness is positive (often above 1.0).
  - Mean will usually be more than the median.



Measure of Skewness

- Bowley's Measure of Skewness

$$\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

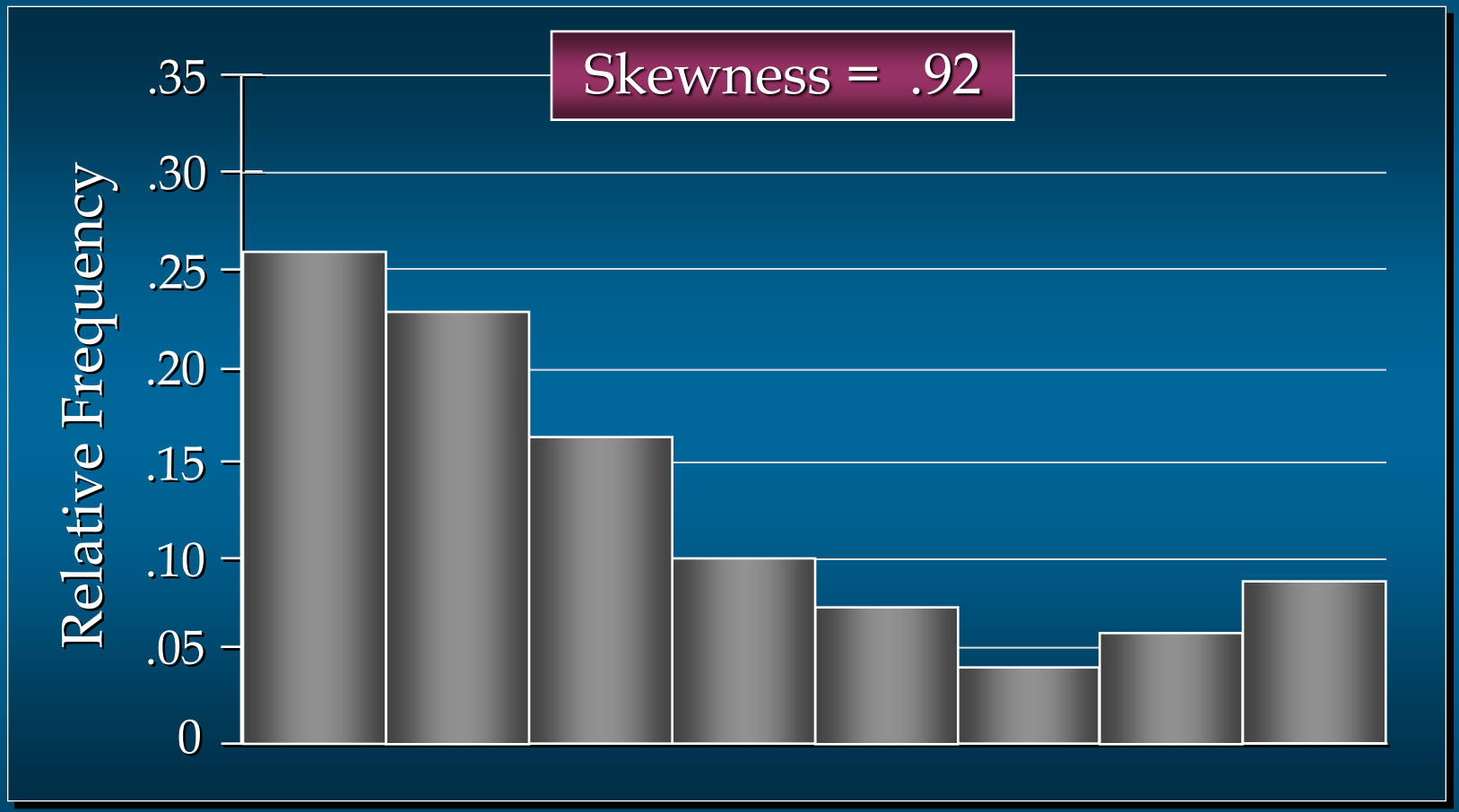
- Pearson's Coefficient of Skewness $\gamma_1 = \sqrt{\beta_1}$

Where,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, and\ \mu_r = \frac{1}{N}\sum f_i (x_i - \bar{x})^r$$

Example: Apartment Rents

x=c(70 data separated by comma)
print(skewness(x))



Kurtosis

x=c(70 data separated by comma)

print(kurtosis(x))

Pearson's Coeff. of Kurtosis

$$\gamma_2 = \beta_2 - 3$$

Where

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, and\ \mu_r = \frac{1}{N}\sum f_i (x_i - \bar{x})^r$$

$\beta_2 > 3$: Leptokurtic

$\beta_2 < 3$: Platykurtic

$\beta_2 = 3$: Mesokurtic

Grouped Data

- Example: Apartment Rents

  Seventy efficiency apartments were randomly sampled in a small college town. The monthly rent prices for these apartments are listed in ascending order on the next slide.

Grouped Data

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Grouped Data

- The weighted mean computation can be used to obtain approximations of the mean, variance, and standard deviation for the grouped data.
- To compute the weighted mean, we treat the midpoint of each class as though it were the mean of all items in the class.
- We compute a weighted mean of the class midpoints using the class frequencies as weights.
- Similarly, in computing the variance and standard deviation, the class frequencies are used as weights.

Grouped Data

Given below is the previous sample of monthly rents for 70 efficiency apartments, presented here as grouped data in the form of a frequency distribution.

Rent ($)	Frequency
420-439	8
440-459	17
460-479	12
480-499	8
500-519	7
520-539	4
540-559	2
560-579	4
580-599	2
600-619	6

Mean for Grouped Data

- Sample Data

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

- Population Data

$$\mu = \frac{\sum f_i M_i}{N(=\sum f_i)}$$

where:

f_i = frequency of class i

M_i = midpoint of class i

Sample Mean for Grouped Data

Rent ($)	f_i	M_i	f_iM_i
420-439	8	429.5	3436.0
440-459	17	449.5	7641.5
460-479	12	469.5	5634.0
480-499	8	489.5	3916.0
500-519	7	509.5	3566.5
520-539	4	529.5	2118.0
540-559	2	549.5	1099.0
560-579	4	569.5	2278.0
580-599	2	589.5	1179.0
600-619	6	609.5	3657.0
Total	70		34525.0

$$\bar{x} = \frac{34,525}{70} = 493.21$$

This approximation differs by $2.41 from the actual sample mean of $490.80.

Median for Grouped Data

Median is given by:

$$l_1 + \frac{\frac{N}{2} - C}{f_{median}} \times i$$

Where l_1 : lower class boundary of the median class.

C : Sum of frequencies of all classes lower than the median class.

f_{med} : Frequency of the median class.

i : Width of the median class.

N.B.: Make the class intervals **Continuous**, if required.

Mode for Grouped Data

Mode is given by:

$$l_1 + \frac{f_{\text{mod}} - f_1}{2f_{\text{mod}} - f_1 - f_2} \times i$$

Where f_1 : frequency of the class immediately preceding the modal class.

f_2 : frequency of the class immediately following the modal class.

l_1, f_{mod} : Lower class boundary & frequency of modal class.

N.B.: Make the class intervals **Continuous**, if required.

Grouped Data

Given below is the previous sample of monthly rents for 70 efficiency apartments, presented here as grouped data in the form of a frequency distribution.

Rent ($)	Frequency
420-439	8
440-459	17
460-479	12
480-499	8
500-519	7
520-539	4
540-559	2
560-579	4
580-599	2
600-619	6

Quartile for Grouped Data

Quartiles are given by:

$$Q_K = l_1 + \frac{K \times \frac{N}{4} - C}{f_{quartile}} \times i$$

for $K = 1, 2, 3$

Note that for $K = 2$ we obtain the median.

N.B.: Make the class intervals **Continuous**, if required.

Variance for Grouped Data

- For sample data

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n-1}$$

- For population data

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N}$$

Sample Variance for Grouped Data

Rent ($)	f_i	M_i	$M_i - \bar{x}$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
420-439	8	429.5	-63.7	4058.96	32471.71
440-459	17	449.5	-43.7	1910.56	32479.59
460-479	12	469.5	-23.7	562.16	6745.97
480-499	8	489.5	-3.7	13.76	110.11
500-519	7	509.5	16.3	265.36	1857.55
520-539	4	529.5	36.3	1316.96	5267.86
540-559	2	549.5	56.3	3168.56	6337.13
560-579	4	569.5	76.3	5820.16	23280.66
580-599	2	589.5	96.3	9271.76	18543.53
600-619	6	609.5	116.3	13523.36	81140.18
Total	70				208234.29

continued →

Sample Variance for Grouped Data

- Sample Variance

$$s^2 = 208{,}234.29/(70 - 1) = 3{,}017.89$$

- Sample Standard Deviation

$$s = \sqrt{3{,}017.89} = 54.94$$

This approximation differs by only \$.20
from the actual standard deviation of \$54.74.

Measure of Skewness for Grouped Data

- Bowley's Measure of Skewness

$$\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

- Pearson's Coefficient of Skewness $\gamma_1 = \sqrt{\beta_1}$

Where,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, and\ \mu_r = \frac{1}{N}\sum f_i (M_i - \bar{x})^r$$

Kurtosis for Grouped Data

Pearson's Coeff. of Kurtosis

$$\gamma_2 = \beta_2 - 3$$

Where

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, and\, \mu_r = \frac{1}{N}\sum f_i (M_i - \bar{x})^r$$

$\beta_2 > 3$: Leptokurtic

$\beta_2 < 3$: Platykurtic

$\beta_2 = 3$: Mesokurtic

Question ?

Thank You