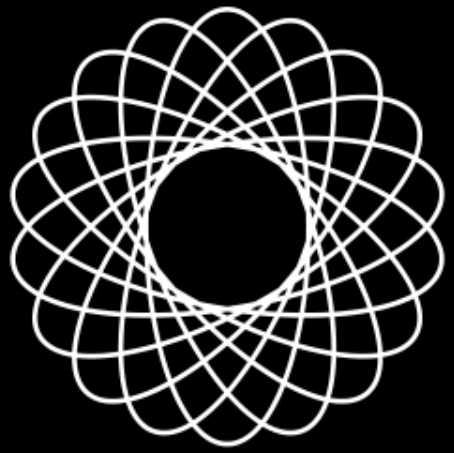
DATA SCIENCE



DATA SCIENCE

STATISTICS

★ Statistical Concepts ★

DATA SCIENCE

STATISTICAL CONCEPTS

What does Statistics cover?

Sample v/s Population

Probability Theory

Probability Distribution Concepts

Types of Distributions

DATA SCIENCE

STATISTICAL CONCEPTS

What does Statistics cover?

Sample v/s Population

Probability Theory

Probability Distribution Concepts

Types of Distributions

DATA SCIENCE

STATISTICAL CONCEPTS

What does Statistics cover?

Sample v/s Population

Probability Theory

Probability Distribution Concepts

Types of Distributions

DATA SCIENCE

STATISTICAL CONCEPTS

What does Statistics cover?

Sample v/s Population

Probability Theory

Probability Distribution Concepts

Types of Distributions

DATA SCIENCE

STATISTICAL CONCEPTS

What does Statistics cover?

Sample v/s Population

Probability Theory

Probability Distribution Concepts

Types of Distributions

DATA SCIENCE

STATISTICAL CONCEPTS

What does Statistics cover?

1. **Summary Statistics**
2. Inferential Statistics

Sample v/s Population

Probability Theory

Probability Distribution Concepts

Types of Distributions

What is Statistics?

Dictionary definition of Statistics:

"the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements."

What is Statistics?

Dictionary definition of Statistics:

"the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements."

Two parts to the definition:

What is Statistics?

Dictionary definition of Statistics:

"the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements."

Two parts to the definition:

1. the collection, classification, analysis and interpretation of numeric data

What is Statistics?

Dictionary definition of Statistics:

"the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements."

Two parts to the definition:

1. the collection, classification, analysis and interpretation of numeric data

2. the use of probability theory to impose order on aggregates of data

Introduction

In general, statistics summarizes information about data in a meaningful, relevant way.

Introduction

In general, statistics summarizes information about data in a meaningful, relevant way.

Describe the population of Bangalore?

http://commons.wikimedia.org/

Introduction

In general, statistics summarizes information about data in a meaningful, relevant way.

Describe the population of Bangalore?

- population in 2010 is 5.4 million
  - That is a statistic – the total sum of all full-time residents of Bangalore

- What other statistics can you think of?

  "Population density" "Median Age" "Distribution by Religion" "Literacy Rate"

http://commons.wikimedia.org/

Introduction

In general, statistics summarizes information about data in a meaningful, relevant way.

Describe the population of Bangalore?

- population in 2010 is 5.4 million
  - That is a statistic – the total sum of all full-time residents of Bangalore

- What other statistics can you think of?

  "Population density" "Median Age" "Distribution by Religion" "Literacy Rate"

All these statistics summarize information because talking about each data point is impossible.

http://commons.wikimedia.org/

Introduction

Some commonly used statistics to summarize information

Data Series: 17,4,33,2,51,23,3,41,18,2,4,2

Mean = 16.67

(17+4+33+2+51+23+3+41+18+2+4+2)/12

Median = 10.5 (4+17)/2 – Why?

Mode = 2 – Why?

Minimum = 2

Maximum = 51

- **Sum:** Total of all values in dataset
- **Mean:** The average of all values in the dataset
- **Median:** Mid value of sorted data
  - If even series?
- **Mode:** Most commonly occurring value in a series
- **Minimum:** Lowest value in series
- **Maximum:** Highest value in series

Introduction

- We can describe the series we looked at in the previous example as:

  "Minimum of 2, Maximum of 51, Average of 16.6."

Introduction

- We can describe the series we looked at in the previous example as:

  "Minimum of 2, Maximum of 51, Average of 16.6."

- Given this description of the data series, what picture do we form of the data?

Introduction

- We can describe the series we looked at in the previous example as:

  "Minimum of 2, Maximum of 51, Average of 16.6."

- Given this description of the data series, what picture do we form of the data?

- The easiest way to visualize data is to look at its **"distribution"**

Introduction

A distribution is a visualization of a frequency distribution table

Introduction

A distribution is a visualization of a frequency distribution table

Taking the data series -

17,4,33,2,51,23,3,41,18,2,4,2

Data Point	Frequency
2	3
3	1
4	2
17	1
18	1
23	1
33	1
41	1
51	1

Introduction

A distribution is a visualization of a frequency distribution table

Taking the data series -

17,4,33,2,51,23,3,41,18,2,4,2

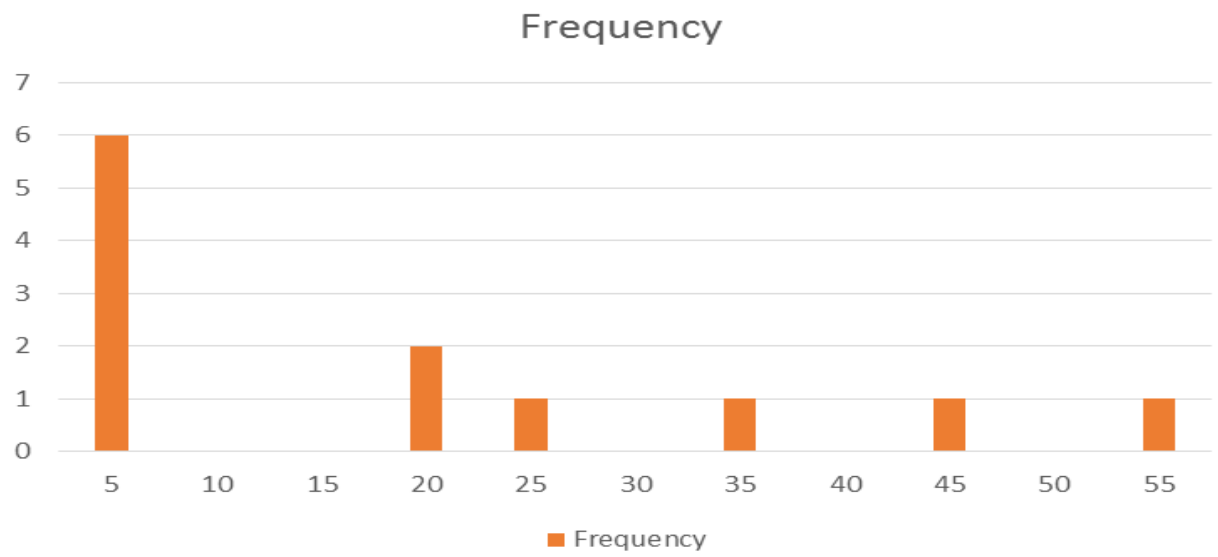
- We create a frequency table which is just counting the number of times each value appears in the data series
- The table shows the frequency count of each data value
- A better way to create this table would be to use ranges of values, rather than individual data points

Data Point	Frequency
2	3
3	1
4	2
17	1
18	1
23	1
33	1
41	1
51	1

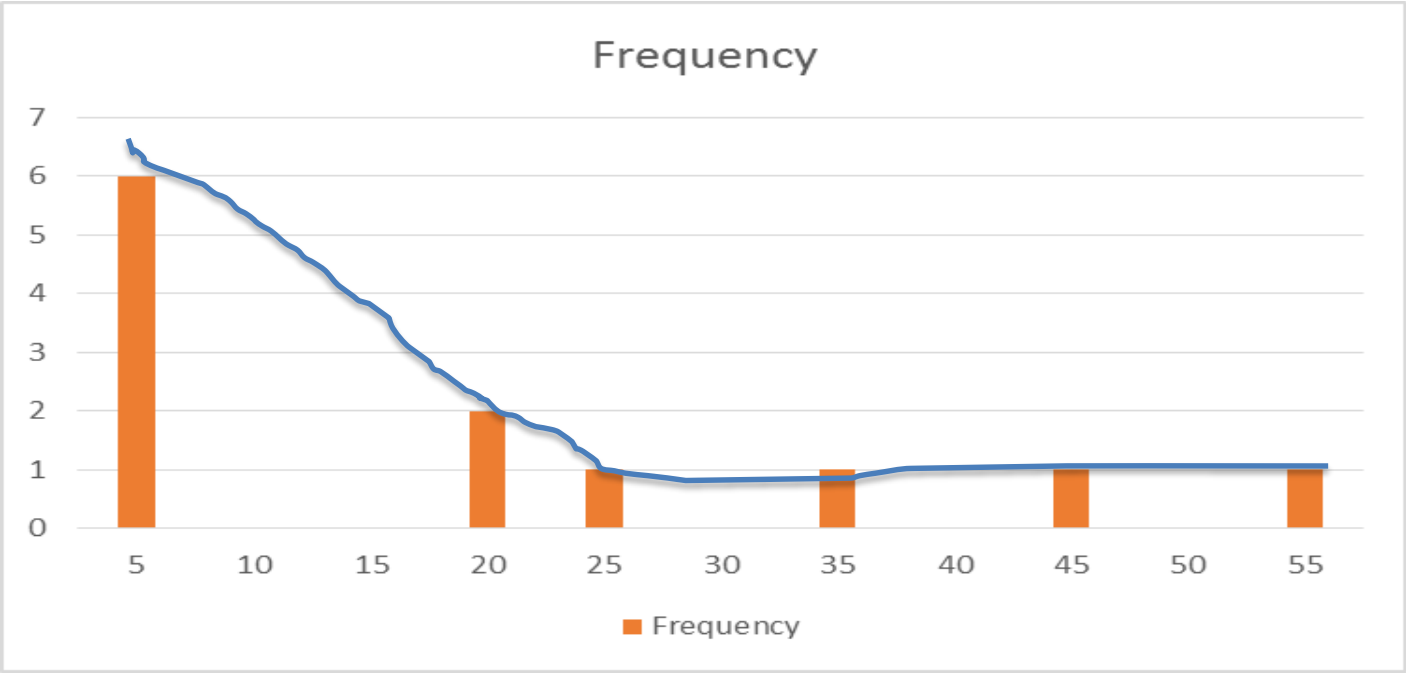
Distributions

- **Bin** refers to the value range, so 5 refers to "0 – 5"
- We have 6 observations that have values between 0 – 5, 2 observations with values between 15 – 20 and so on
- The quickest way make sense of this data is to turn it into a **visualization**

Bin	*Frequency*
5	6
10	0
15	0
20	2
25	1
30	0
35	1
40	0
45	1
50	0
55	1



Distributions



This visualization gives us a “picture” of the data - this is a data distribution

We can also draw a line joining the bars

Introduction

Summary statistics can also be used to understand variation or dispersion in the data

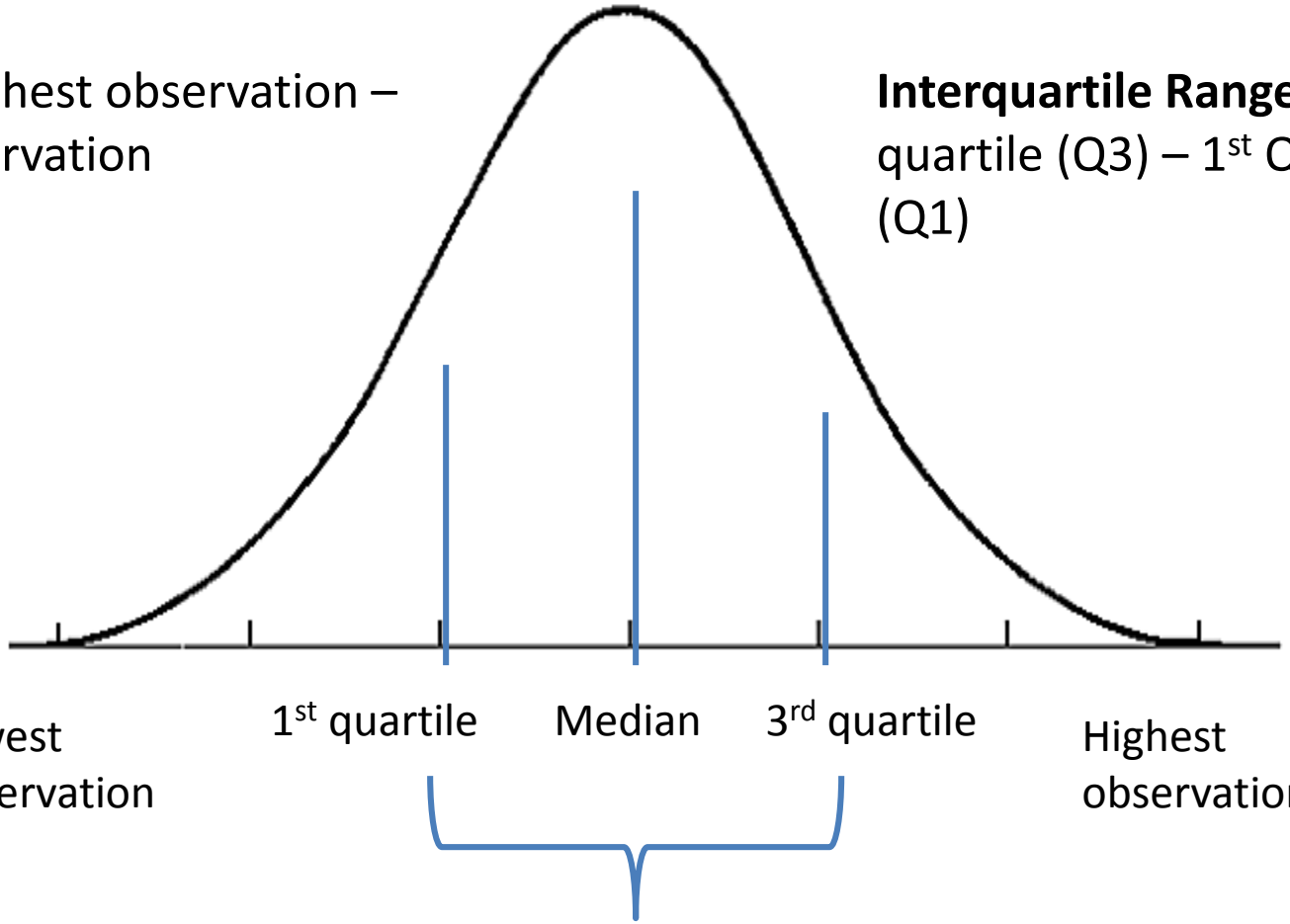
For Example:

- Range
- Variance
- Standard Deviation

Range

Range = Highest observation – lowest observation

Interquartile Range = 3rd quartile (Q3) – 1st Quartile (Q1)



Variance

Customer Number	Average minute usage (monthly)	X-μ	$(x-\mu)^2$
1	228	-6.8	46.24
2	260	25.2	635.04
3	252	17.2	295.84
4	298	63.2	3994.24
5	234	-0.8	0.64
6	50	-184.8	34151.04
7	264	29.2	852.64
8	230	-4.8	23.04
9	304	69.2	4788.64
10	228	-6.8	46.24

Variance = $\sigma^2 = \sum(x-\mu)^2/N$

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

***x**= observation*

***μ**= population mean*

***N**= number of observations in the population*

Variance = (44833.6/10) = **4483**

Standard deviation = √4483 = **67**

Application & uses of Std Deviation

CHEBYSHEV'S INEQUALITY

For any data set, it can be proved mathematically that –

Application & uses of Std Deviation

CHEBYSHEV'S INEQUALITY

For any data set, it can be proved mathematically that –

- atleast 75% of all data points will lie within 2 std deviations of the mean,
- atleast 89% of all data points will lie within 3 std deviations of the mean.

Application & uses of Std Deviation

CHEBYSHEV'S INEQUALITY

For any data set, it can be proved mathematically that –

- atleast 75% of all data points will lie within 2 std deviations of the mean,
- atleast 89% of all data points will lie within 3 std deviations of the mean.

For a data series with a min of 200, max of 1500, a mean of 600, and a std deviation of 80,

Application & uses of Std Deviation

CHEBYSHEV'S INEQUALITY

For any data set, it can be proved mathematically that –

- atleast 75% of all data points will lie within 2 std deviations of the mean,
- atleast 89% of all data points will lie within 3 std deviations of the mean.

For a data series with a min of 200, max of 1500, a mean of 600, and a std deviation of 80,

- atleast 75% of all the data points in the series will be within the range:
  600 – 2*80, 600 + 2*80 : (440, 760)

Application & uses of Std Deviation

CHEBYSHEV'S INEQUALITY

For any data set, it can be proved mathematically that –

- atleast 75% of all data points will lie within 2 std deviations of the mean,
- atleast 89% of all data points will lie within 3 std deviations of the mean.

For a data series with a min of 200, max of 1500, a mean of 600, and a std deviation of 80,

- atleast 75% of all the data points in the series will be within the range:
  600 – 2*80, 600 + 2*80 : (440, 760)
- atleast 89% of all data points in the series will be within the range:
  600 – 3*80, 600 + 3*80 : (360, 840)

Std Deviation

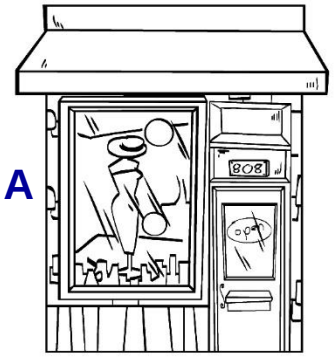
1) **Standard deviation used to provide a sense of variation in the data**

Std Deviation

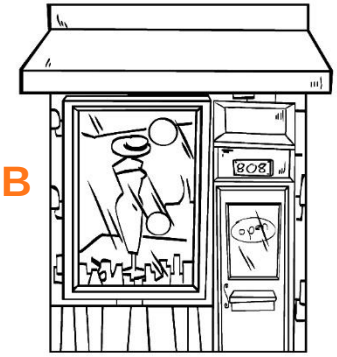
1) <u>Standard deviation used to provide a sense of variation in the data</u>

Example: Supposing we had data on customer spend by store.

STORE A



STORE B



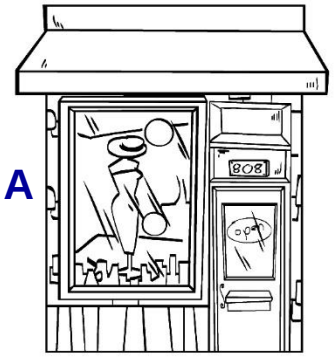
Std Deviation

1) <u>Standard deviation used to provide a sense of variation in the data</u>

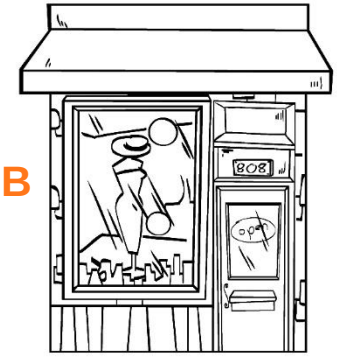
Example: Supposing we had data on customer spend by store.

Average spend per customer in Store A was $150, with a std deviation of $35, and the average spend per customer in Store B was $145, with a std deviation of $15.

STORE A



STORE B



Std Deviation

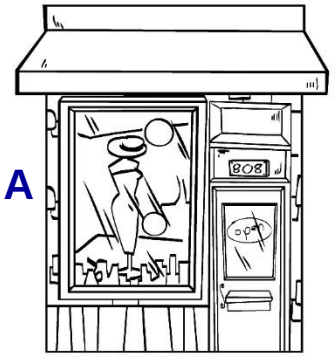
1) <u>Standard deviation used to provide a sense of variation in the data</u>

Example: Supposing we had data on customer spend by store.

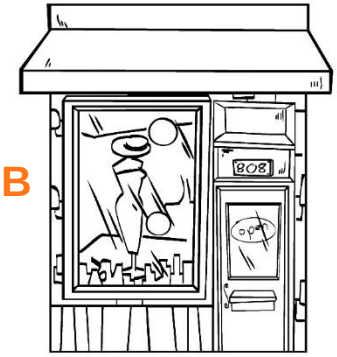
Average spend per customer in Store A was $150, with a std deviation of $35, and the average spend per customer in Store B was $145, with a std deviation of $15.

In which store are sales higher?

STORE A



STORE B



Std Deviation

2) <u>Standard deviation used as a measure or risk</u>

Example:

You are trying to pick stock for investing in the equity market.

- Stock A has an annual return of 15%, with a std deviation of 30%
- Stock B has an annual return of 12%, with a std deviation of 8%
- If you were risk averse, which would you choose?

TRY AND SOLVE IT!

Recap

Summary / Descriptive Statistics:

- Help us summarize information in a meaningful way
- Help us create a mental picture of data

Recap

Summary / Descriptive Statistics:

- Help us summarize information in a meaningful way
- Help us create a mental picture of data

Different summary statistics including:

- Measures of central tendency
- Measures of variation
- Measures of shape

DATA SCIENCE

STATISTICAL CONCEPTS

What does Statistics cover?

1. **Summary Statistics**
2. Inferential Statistics

Sample v/s Population

Probability Theory

Probability Distribution Concepts

Types of Distributions

Shape Measures

Shape Measures

- We have reviewed average measures and range measures. Another measure used in descriptive statistics is Shape

Shape Measures

- We have reviewed average measures and range measures. Another measure used in descriptive statistics is Shape

- One measure of shape is the degree of skewness

Shape Measures

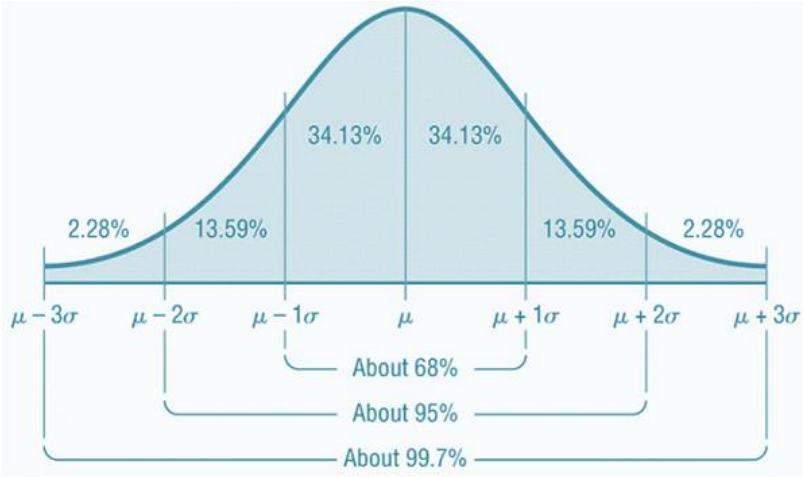
- We have reviewed average measures and range measures. Another measure used in descriptive statistics is Shape
- One measure of shape is the degree of skewness
- **Skewness is the absence of symmetry**

Shape Measures

- We have reviewed average measures and range measures. Another measure used in descriptive statistics is Shape
- One measure of shape is the degree of skewness
- **Skewness is the absence of symmetry**
- A symmetric shape is one where the left side of the data distribution is a mirror image of the right (side relative to mean)

Shape Measures

- We have reviewed average measures and range measures. Another measure used in descriptive statistics is Shape
- One measure of shape is the degree of skewness
- **Skewness is the absence of symmetry**
- A symmetric shape is one where the left side of the data distribution is a mirror image of the right (side relative to mean)



Shape Measures

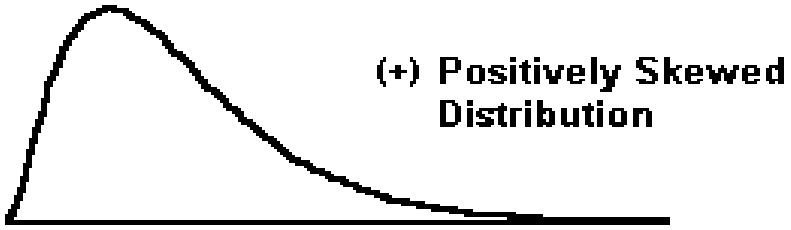
Skewness: Degree of Asymmetry

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$

Shape Measures

Skewness: Degree of Asymmetry

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$



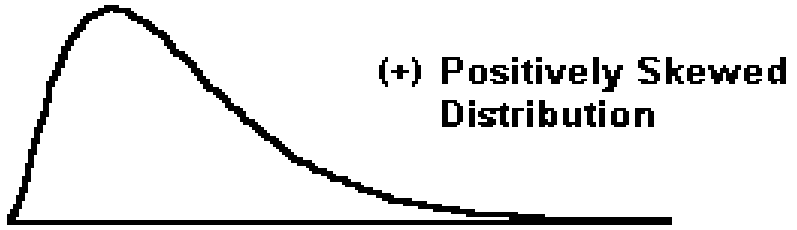
- Positive Skew: Long tail to the Right

Shape Measures

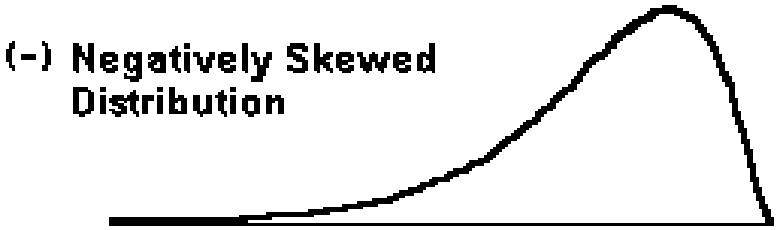
Skewness: Degree of Asymmetry

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$

- Positive Skew: Long tail to the Right



- Negative Skew: Long tail to the Left

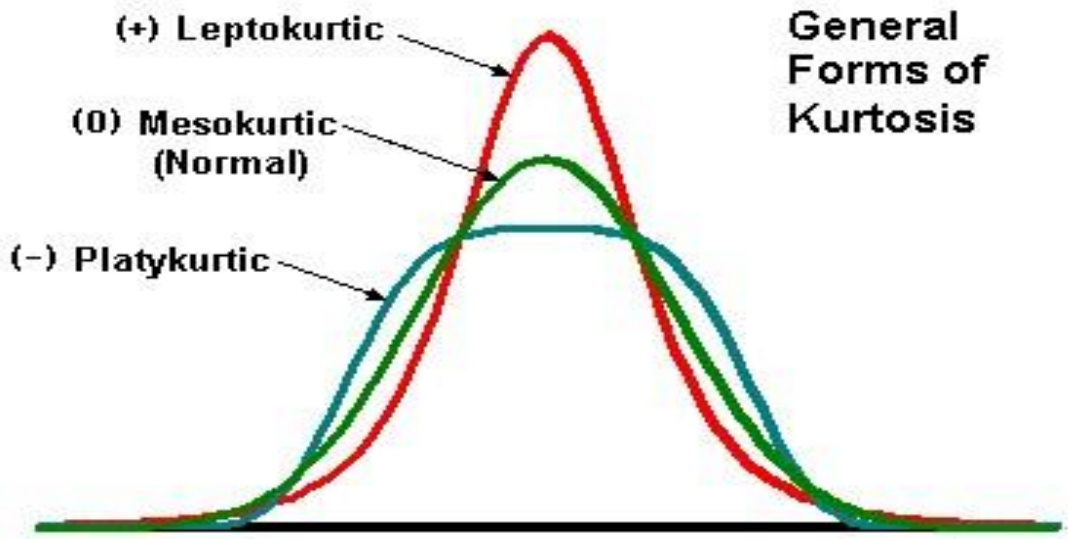


Shape Measures

Kurtosis: Sharpness of the peak of the distribution

Shape Measures

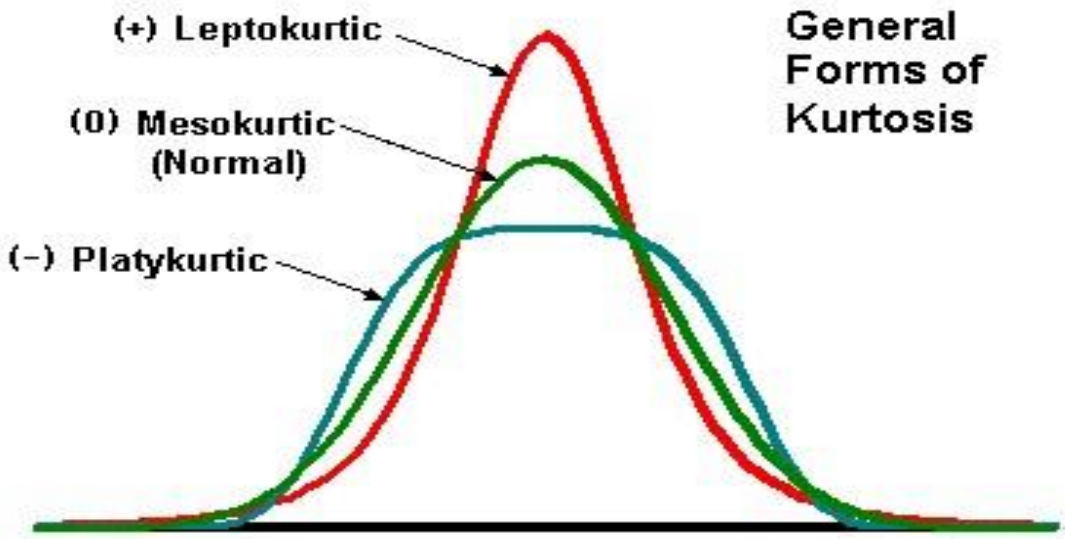
Kurtosis: Sharpness of the peak of the distribution



Shape Measures

Kurtosis: Sharpness of the peak of the distribution

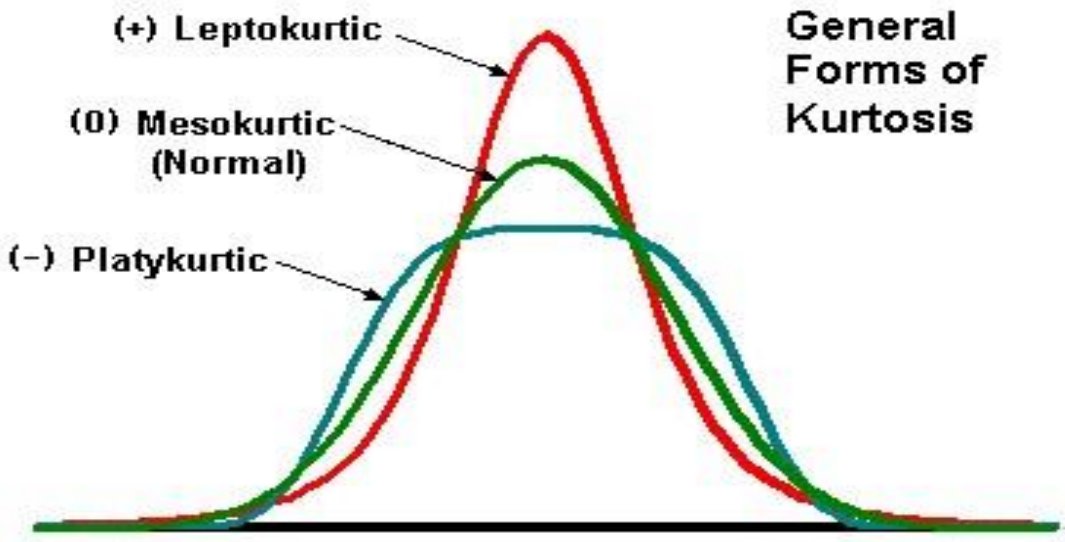
- A high kurtosis distribution has a sharp peak and fat tails



Shape Measures

Kurtosis: Sharpness of the peak of the distribution

- A high kurtosis distribution has a sharp peak and fat tails
- A low kurtosis distribution has a flat peak and thin tails

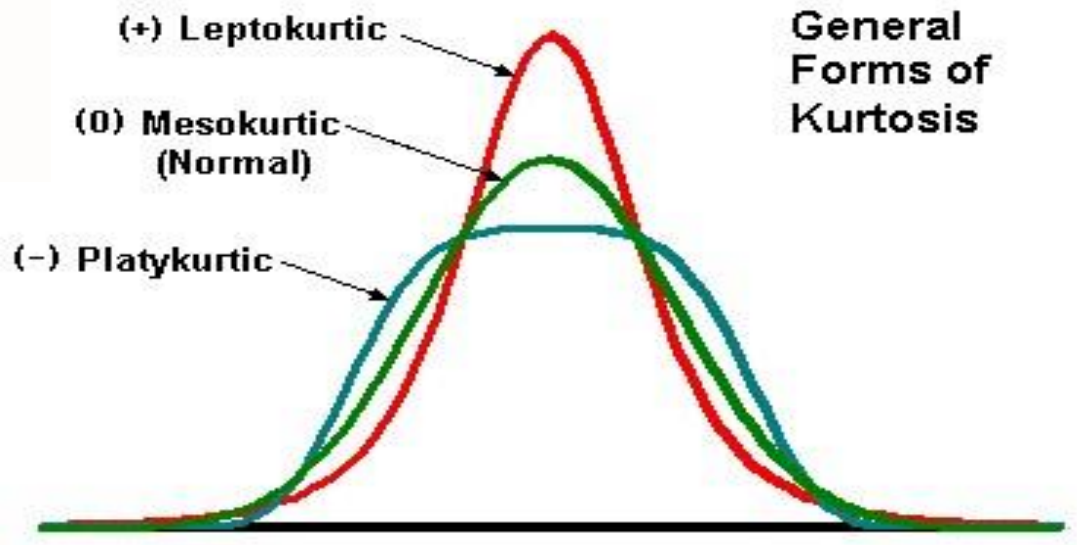


Shape Measures

Kurtosis: Sharpness of the peak of the distribution

- A high kurtosis distribution has a sharp peak and fat tails
- A low kurtosis distribution has a flat peak and thin tails

$$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4}$$



Descriptive Statistics

We have covered a variety of descriptive statistics:

Descriptive Statistics

We have covered a variety of descriptive statistics:

- Measures of central tendency – mean, median, mode

Descriptive Statistics

We have covered a variety of descriptive statistics:

- Measures of central tendency – mean, median, mode
- Measures of dispersion – range, variance, std deviation

Descriptive Statistics

We have covered a variety of descriptive statistics:

- Measures of central tendency – mean, median, mode
- Measures of dispersion – range, variance, std deviation
- Measures of shape – skewness, kurtosis

Descriptive Statistics

We have covered a variety of descriptive statistics:

- Measures of central tendency – mean, median, mode
- Measures of dispersion – range, variance, std deviation
- Measures of shape – skewness, kurtosis

These descriptive statistics are used to help describe data, especially when we are dealing with very large data sets.

Descriptive Statistics: Examples

Database of Car Prices

	A	B	C	D	E	F	G	H	I	J	K
1		MSRP	SUV	city	high	luggage	horse	Cyl	Disp	fuel	Non-SUV
2		30880	0	19	29	13.6	260	6	3.2	17.2	1
3		20465	0	24	32	14.6	140	4	2.2	14.1	1
4		13270	0	32	37	12.9	115	4	1.7	13.2	1
5		21635	0	20	29	14.6	175	6	3.4	14.1	1
6		12482	0	32	39	11.5	92	4	1.5	12.4	1
7		10480	0	34	41	13.6	108	4	1.5	11.9	1
8		31845	0	23	31	13.8	180	4	1.8	14.5	1
9		29745	0	19	27	9.5	184	6	2.5	16.6	1
10		15675	0	24	32	13.2	115	4	2.2	14.1	1
11		13330	0	25	33	11.8	130	4	2.0	12.8	1
12		39647	0	18	27	15.3	275	8	4.6	19.0	1
13		21170	0	20	29	16.7	175	6	3.1	17.5	1
14		23274	0	20	29	18.0	205	6	3.8	18.5	1
15		16433	0	21	38	12.4	140	4	2.2	14.1	1
16		13545	0	28	40	11.4	100	4	1.9	12.1	1
17		38150	0	21	29	13.1	236	5	2.3	18.0	1
18		21060	0	21	32	15.8	180	6	3.4	17.0	1
19		21015	0	21	28	16.3	142	4	2.4	16.3	1
20		21985	0	20	29	16.3	200	6	3.0	16.3	1
21		18705	0	20	29	10.9	190	6	3.8	15.5	1
22		21960	0	26	32	13.6	150	4	2.3	17.1	1
23		23835	0	19	29	16.0	200	6	3.8	17.5	1
24		19850	0	24	33	13.8	157	4	2.4	18.5	1
25		51100	0	19	28	13.3	405	8	5.7	18.5	1
26		83000	0	11	21	9.2	460	10	8.0	19.0	1
27		68665	0	18	26	4.6	320	6	3.6	16.9	1
28		33845	0	22	31	7.8	180	4	1.8	14.5	1

Sheet1 Sheet2 Sheet3

Recap

Recap

Why is it important to learn about summary statistics?

Recap

Why is it important to learn about summary statistics?

- Description of a large number of data points

Recap

Why is it important to learn about summary statistics?

- Description of a large number of data points
- Generate inferences from the summary statistics

Recap

Why is it important to learn about summary statistics?

- Description of a large number of data points
- Generate inferences from the summary statistics

Example:

You work for a credit card company and are required to understand drivers of default.

Recap

Why is it important to learn about summary statistics?

- Description of a large number of data points
- Generate inferences from the summary statistics

Example:

You work for a credit card company and are required to understand drivers of default.

You have access to:

Recap

Why is it important to learn about summary statistics?

- Description of a large number of data points
- Generate inferences from the summary statistics

Example:

You work for a credit card company and are required to understand drivers of default.

You have access to:

1. Billing data

Recap

Why is it important to learn about summary statistics?

- Description of a large number of data points
- Generate inferences from the summary statistics

Example:

You work for a credit card company and are required to understand drivers of default.

You have access to:

1. Billing data
2. Demographic data

Recap

You divide the data into customers who have:

Recap

You divide the data into customers who have:

1. A **great payment record**, and

Recap

You divide the data into customers who have:

1. A **great payment record**, and

2. Customers who have been **late with payments at least 3 times** in the last year

Recap

You divide the data into customers who have:

1. A **great payment record**, and

2. Customers who have been **late with payments at least 3 times** in the last year

Good	Not so good

Recap

You divide the data into customers who have:

1. A **great payment record**, and

2. Customers who have been **late with payments at least 3 times** in the last year

Good
Mean Income: $37K
Std Dev of Income: $5K

Recap

You divide the data into customers who have:

1. A **great payment record**, and

2. Customers who have been **late with payments at least 3 times** in the last year

Good	Not so good
Mean Income: $37K	Mean Income: $26K
Std Dev of Income: $5K	Std Dev of Income: $9K

Recap

What does it mean?

Mean income – capacity to pay back

- Lower in Not so good, but higher in Good

Good	Not so good
Mean Income: $37K	Mean Income: $26K
Std Dev of Income: $5K	Std Dev of Income: $9K

Recap

What does it mean?

Mean income – capacity to pay back

- Lower in Not so good, but higher in Good

Std Deviation – larger the std deviation, greater the variation in income

Good	Not so good
Mean Income: $37K	Mean Income: $26K
Std Dev of Income: $5K	Std Dev of Income: $9K

Recap

What does it mean?

Mean income – capacity to pay back

- Lower in Not so good, but higher in Good

Std Deviation – larger the std deviation, greater the variation in income

Variation in income is higher in Not so good, implying there is an overlap in income in Good and Not so good

Good	Not so good
Mean Income: $37K	Mean Income: $26K
Std Dev of Income: $5K	Std Dev of Income: $9K