Introduction to Statistics

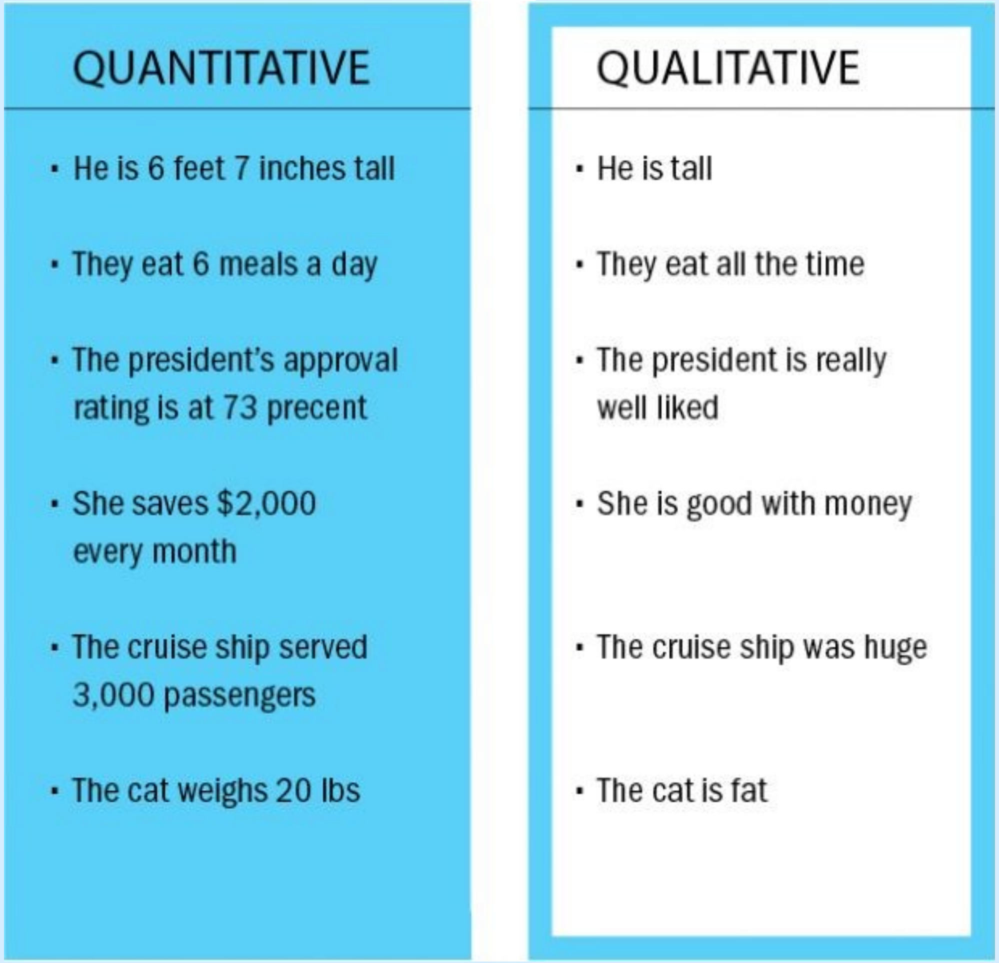
Agenda

- Data
- Statistics – What, when, why?
- Descriptive Statistics

Data

- Snapshot of the world
  - Represents a view of the world at any given point of time
- Different types:
  - Discrete and Continuous
  - Qualitative and Quantitative (Numeric vs Categorical
  - Scale – Nominal, Ordinal, Interval, Ratio

Data – Qualitative vs Quantitative

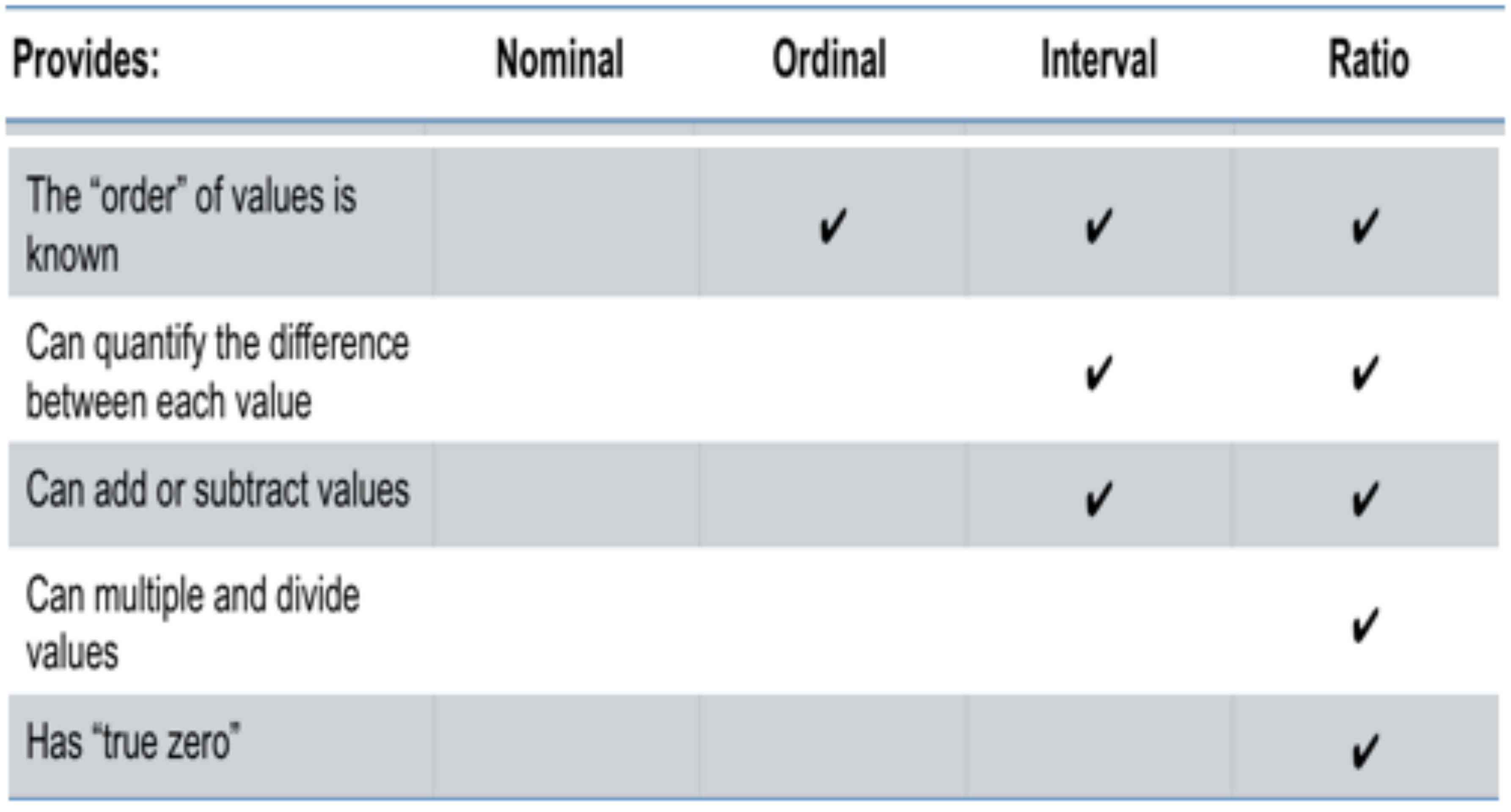


QUANTITATIVE	QUALITATIVE
▪ He is 6 feet 7 inches tall	▪ He is tall
▪ They eat 6 meals a day	▪ They eat all the time
▪ The president's approval rating is at 73 precent	▪ The president is really well liked
▪ She saves $2,000 every month	▪ She is good with money
▪ The cruise ship served 3,000 passengers	▪ The cruise ship was huge
▪ The cat weighs 20 lbs	▪ The cat is fat

Data - Scale

- Nominal
  - Categorical allocation. Categories don’t have any order
  - Gender, Sports type, nationality
- Ordinal
  - Categorical, but now with logical ordering
  - Income group, university ranking
- Interval
  - Specify difference but not scales
  - Zero point of scale is arbitrary (differences are meaningful)
  - Date
- Ratio
  - Comparison of values. Clearly defined zero value
  - Height, weight,

Data - Scale



Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✓	✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓

Why data?

- Presents the truth – fact rather than opinion
- Provides raw material
  - Answer questions in a factual manner
- We have opinions/theories
  - Data helps us validate/invalidate those
- Example: Do people make money in stock market?
  - When asked individually, answers would be yes, no, maybe
  - Better, let us gather trades data
  - Create a model and validate

Sources of data

- Each such question is an experiment
- Data could be collected in many ways
  - Experimental: Drug use data
    - Accurate and randomized. Cannot have experiments for everything
  - Observational (Complete): Census data
    - Expensive, not all data might be needed
  - Sampling: Only a representative subset of data
    - Inexpensive, ensure it is bias free

Example

- If the student knew his/her internal assessment marks and previous year CGPA can they get an idea of how they might perform in the finals?

- Suppose that a student had the following information on 50 students from the previous batch:
  - Marks in the final examination
  - Marks in 3 internal assessment tests held during the academic year
  - CGPA obtained in the previous year

- Simply using this data and few statistical tools we can answer various interesting questions:
  - Can a student predict what range his\her final examination score will lie in?
  - Is it correct to say someone that did well(or not so well) in internals will do well (or not so well) in the finals?
  - Does the previous year's CGPA, which doesn't depend on the present course, affect the final scores?

Marketing Research Example

- The scenario:
- There are 4 stores : 'OfficeStar', 'Paper & Co.', 'Office Equipment', 'Supermarket'
- There are some customers that have visited and made purchases from each of these stores
- The stores collect certain feed back from each of the customers. Each customer rates each store on a scale from 1 to 5 (1 being the lowest and 5 the highest) on the following attributes:
  - Large choice (wide variety)
  - Low prices
  - Service quality
  - Product quality
  - Convenience
  - Preference Score (overall satisfaction score)

- These stores are interested in answering the following questions:
  - What part of the variation in the ratings between stores is because of the customers and not the stores themselves?
  - Does a particular class of customers (age wise, gender wise, locality wise etc.) prefer a particular store?
  - Does a particular store serve a particular class of people more efficiently?
  - Statistics can help provide answers to the above questions with a reasonable level of accuracy.

General Observation

- The difficulty in providing straight-forward answers to all the above questions arises from the fact that there is variability
  - Different people, different ages, different weights etc
- The idea behind introducing the above few examples is to emphasize on the need for statistical investigation when a question needs to be answered or a hypothesis tested for accuracy in the presence of variation.

Statistics type

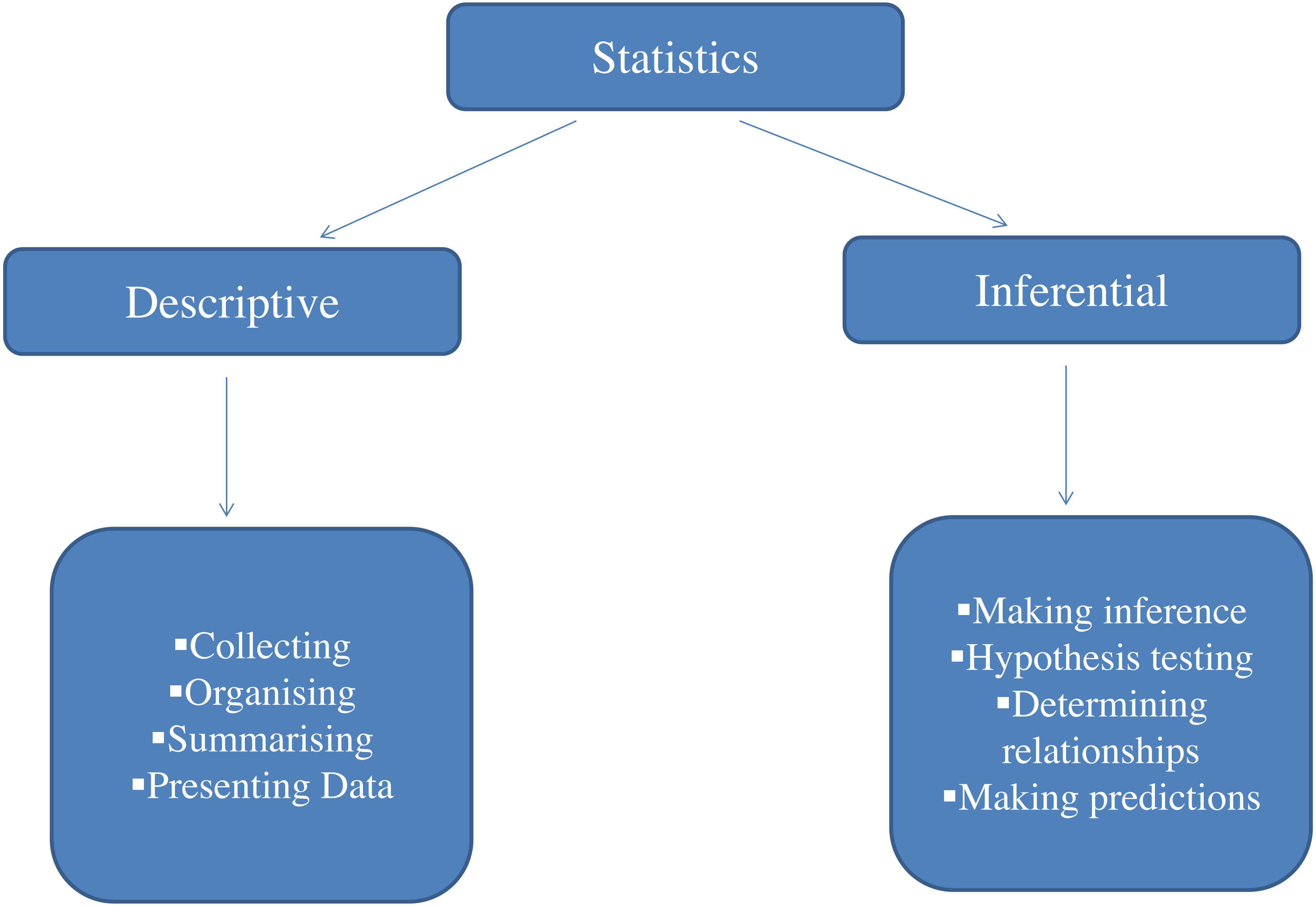
- Descriptive Statistics: Getting a better sense of data
  - Mean, Standard Deviation, Median, Quartiles, Distribution
- Inferential Statistics
  - Drawing conclusions about the population based on sample data
  - Properties of a single variable

Descriptive statistics

- Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that patterns might emerge from the data.
- Does not allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made.
- Simply a way to describe the data.

Inferential statistics

- Inferential statistics is concerned with making predictions or inferences about a population from observations and analysis of a sample.
- We can take the results of an analysis using a sample and can generalize it to the larger population that the sample represents.



Statistics

Descriptive

Inferential

- Collecting
- Organising
- Summarising
- Presenting Data

- Making inference
- Hypothesis testing
- Determining relationships
- Making predictions

Three types of analysis

- Univariate analysis
  - the examination of the distribution of cases on only one variable at a time (e.g., weight of college students)
- Bivariate analysis
  - the examination of two variables simultaneously (e.g., the relation between gender and weight of college students)
- Multivariate analysis
  - the examination of more than two variables simultaneously (e.g., the relationship between gender, race and weight of college students)

Purpose of diff. types of analysis

- Univariate analysis
  - Purpose: mainly description
- Bivariate analysis
  - Purpose: determining the empirical relationship between the two variables
- Multivariate analysis
  - Purpose: determining the empirical relationship among multiple variables

Data Analysis

- Defined steps for data analysis:
- Exploratory
  - Cleaning
  - Summarization
    - Centrality
    - Dispersion
    - Concentration
  - Visualization
- Inferential

Central Tendency

- Measure to summarize a large data set with a central value "average value"
- Makes comparison of datasets possible by comparing averages
- Most common measure is arithmetic mean or average

$$\overline{y} = (1/n)\sum_{i=1}^{n} y_i$$

Central Tendency

- Many more measures of central tendency
  - Harmonic Mean
  - Geometric Mean
  - Median
  - Mode
  - Quartiles

Different Measures

- Situation 1: Consider the following example of salary break up in a small firm visiting your campus for placement:
- Table 1

Employee	Salary (Monthly)
CEO (only 1)	3,00,000
Senior Analyst (10 of them)	70,000
Junior Analyst (20 of them)	50,000
Computer Scientist (2 of them)	35,000
Intern (2)	15,000

- Arithmetic Mean=

(3,00,000+ 10*70,000+ 20*50,000+2*35,000+2*15,000)/35= 60,000

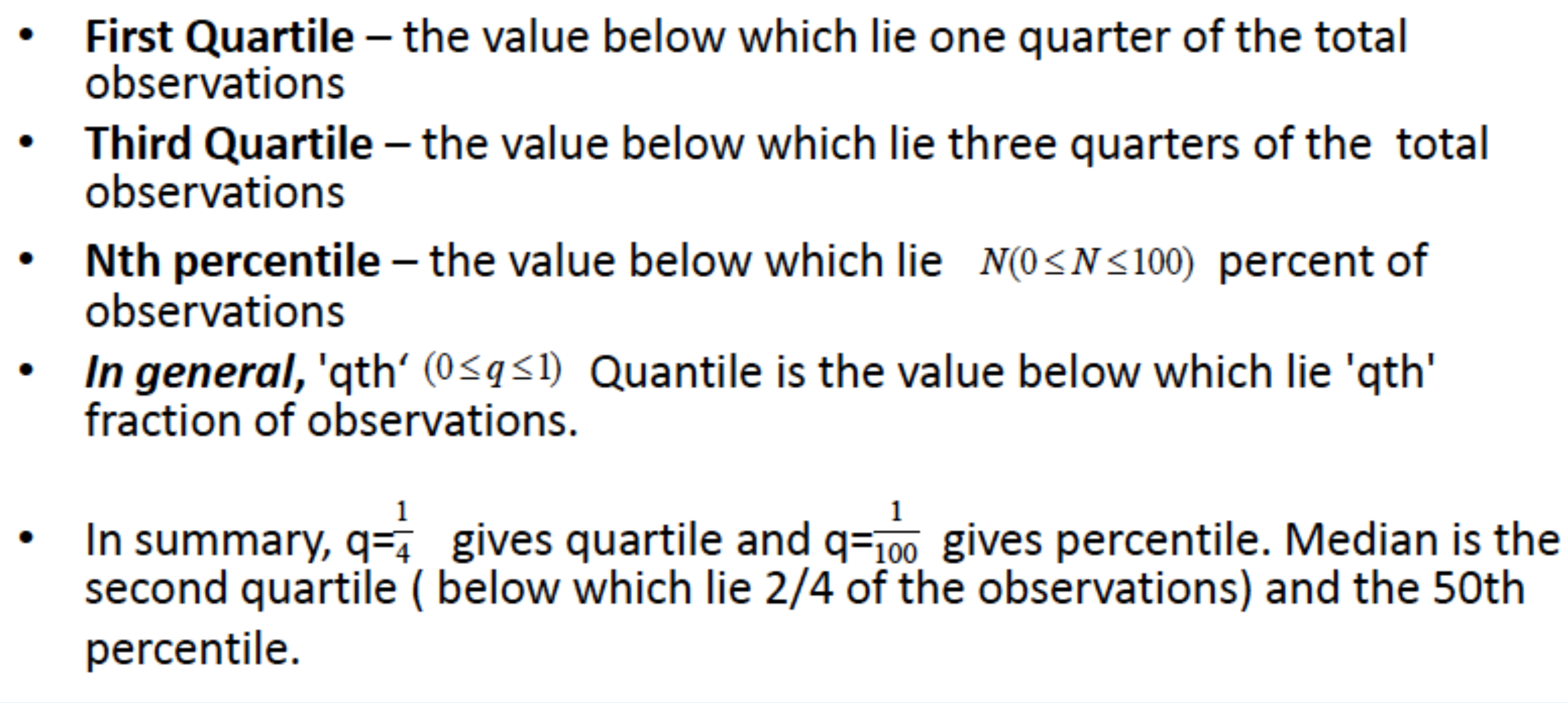
- However, more than half the employees, 24 out of 35, get salary less than 50,000!

- AM doesn't seem representative- The salary of the CEO pulls it up!

- Also, as a college graduate, you know that you are not going to be a CEO or an intern so you are hardly interested in the values of extreme observations. (Too high a value for CEO and too small a value for intern)

- Median comes handy in such situations
- Median is middlemost observation in dataset.
- What is median salary in previous example?
  - 50,000
- Less susceptible to extreme observations
- An example of positional measure. Many more positional measures are available
  - Quartiles
  - Percentiles

Quartile and Percentile

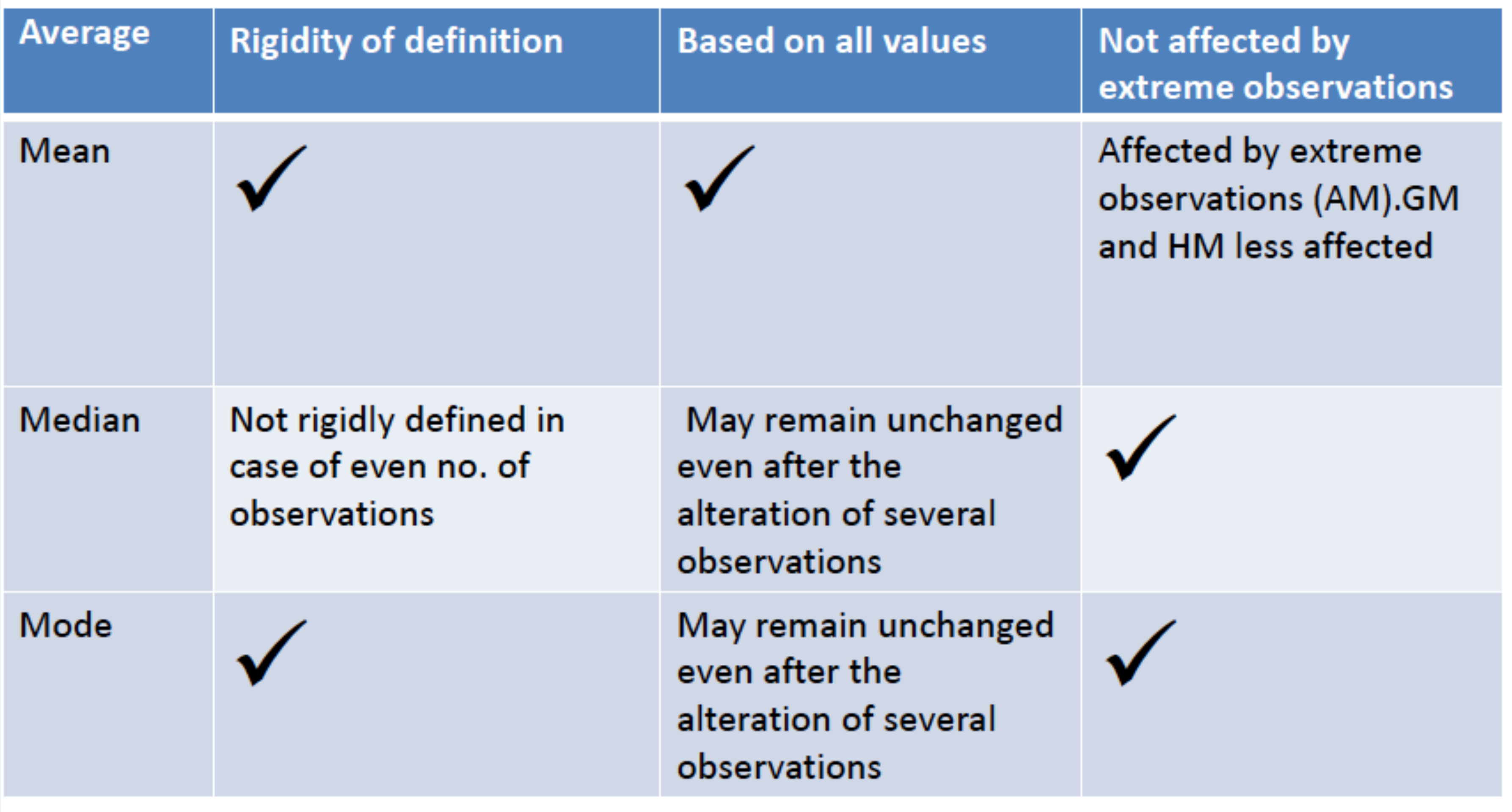


- **First Quartile** – the value below which lie one quarter of the total observations
- **Third Quartile** – the value below which lie three quarters of the total observations
- **Nth percentile** – the value below which lie $N(0 \leq N \leq 100)$ percent of observations
- ***In general,*** 'qth‘ $(0 \leq q \leq 1)$ Quantile is the value below which lie 'qth' fraction of observations.
- In summary, q=$\frac{1}{4}$ gives quartile and q=$\frac{1}{100}$ gives percentile. Median is the second quartile (below which lie 2/4 of the observations) and the 50th percentile.

Mode

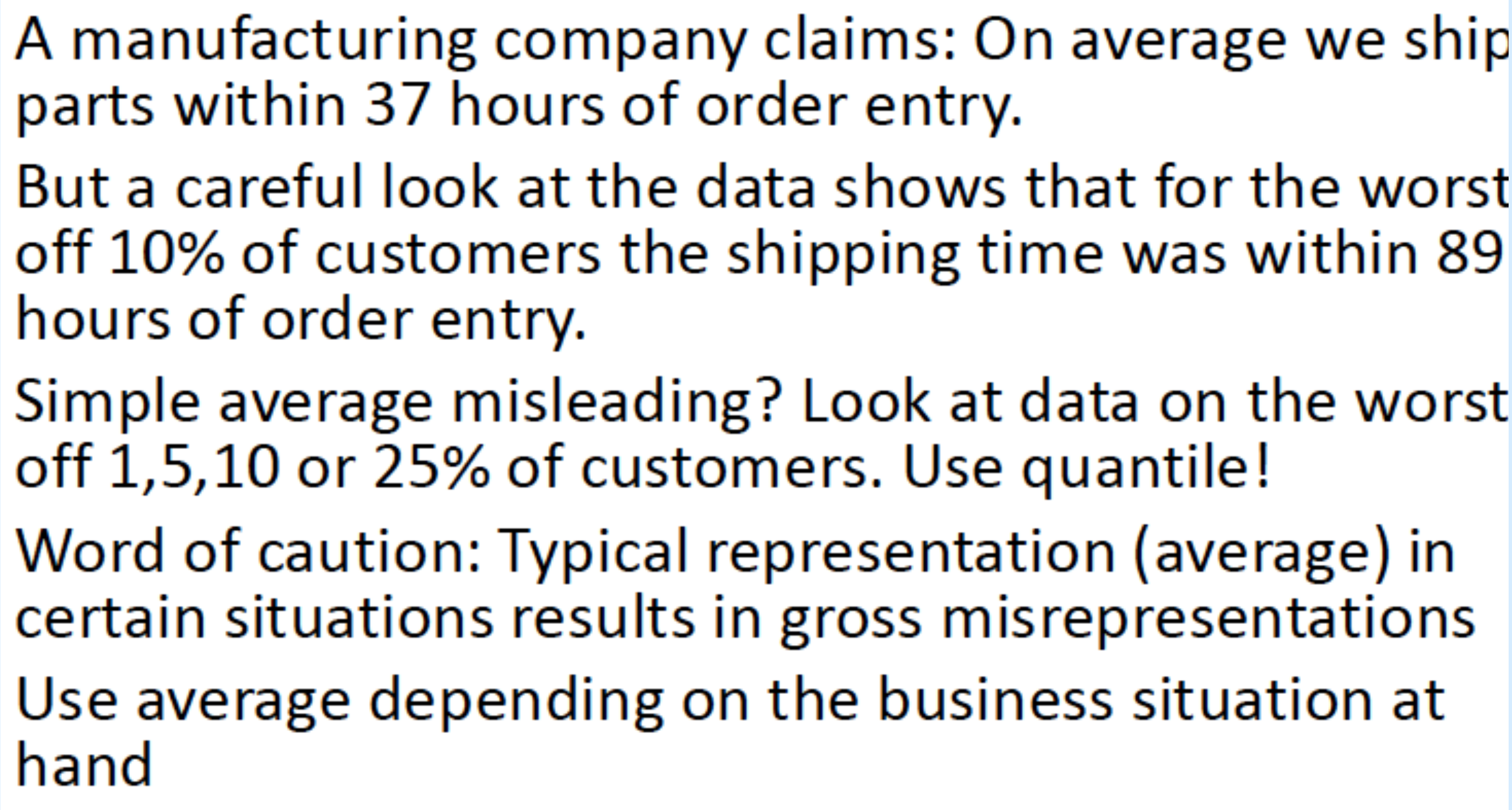
- Value with highest occurrence.
- Consider three colors of flowers: 15 with red color, 20 with white, 22 with yellow.
- Mode here would be yellow.

Different Measures



Average	Rigidity of definition	Based on all values	Not affected by extreme observations
Mean	✓	✓	Affected by extreme observations (AM).GM and HM less affected
Median	Not rigidly defined in case of even no. of observations	May remain unchanged even after the alteration of several observations	✓
Mode	✓	May remain unchanged even after the alteration of several observations	✓

Example



A manufacturing company claims: On average we ship parts within 37 hours of order entry.

But a careful look at the data shows that for the worst off 10% of customers the shipping time was within 89 hours of order entry.

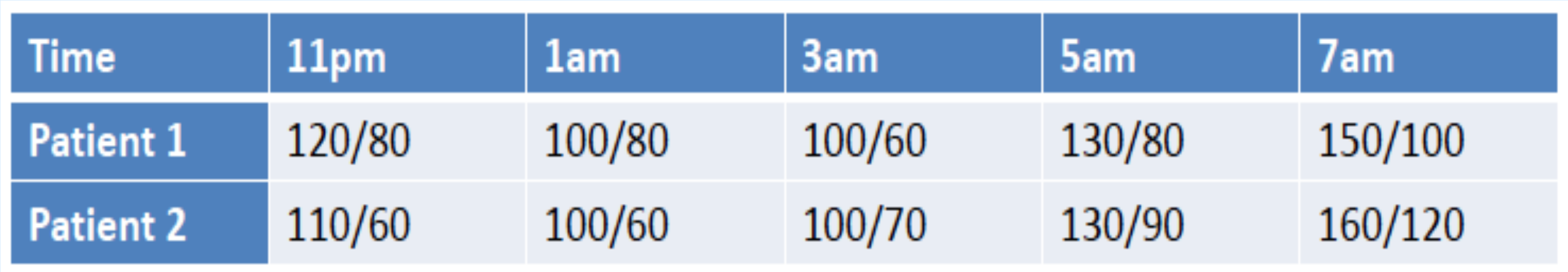
Simple average misleading? Look at data on the worst off 1,5,10 or 25% of customers. Use quantile!

Word of caution: Typical representation (average) in certain situations results in gross misrepresentations

Use average depending on the business situation at hand

Dispersion

- Two patients are admitted into the Intensive Care Unit of a hospital. The night before their operation, the doctor makes the last visit at 9pm and blood pressure for Patient 1 is 110/80 and for Patient 2 it is 120/70. Although they are normal, for precautionary reasons, the Doctor asks the nurse to check their blood pressure every 2 hours. At 7.30 the next morning, the nurse reports that the average blood pressure for both the patients was normal, 120/80. The chart of their actual blood pressures was:



Time	11pm	1am	3am	5am	7am
Patient 1	120/80	100/80	100/60	130/80	150/100
Patient 2	110/60	100/60	100/70	130/90	160/120

- In this case can we truly say that the two patients have similar blood pressure and advise same treatment for both?

Dispersion

- Measured along several dimensions
- Boundaries of dataset:
  - Max and min values
  - Range: max-min
  - Inter-quartile range
- Deviations of individual points from central measure
  - Mean absolute deviation
  - Variance
  - Standard deviation

Higher Moments

- Variance $= \frac{\sum(Data\ value - Mean)^2}{n}$ is the 'sum of squares of deviations from the mean divided by *n*' or the 'expected value of squared deviation of X from its mean'
- Expected values of higher powers of deviations from mean, give additional information about the distribution of data
- Expected value of any power of the deviations from mean of a variable X (say r^{th} power) is called the r^{th} **central moment** of that variable

$$r^{th}\ central\ moment = \mu_r = \frac{\sum(x-\bar{x})^r}{n} = E((x-\bar{x})^r)$$

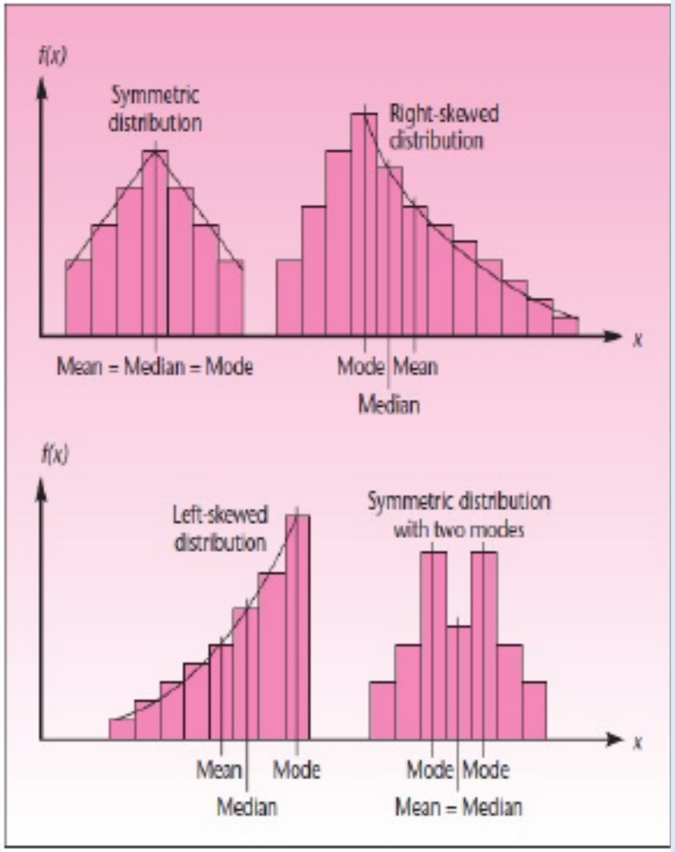
- Central moments depict the spread and shape of data
- Variance is 2nd central moment
- Measures using the 3rd and 4th central moments are useful to understand the shape of the distribution

Skewness

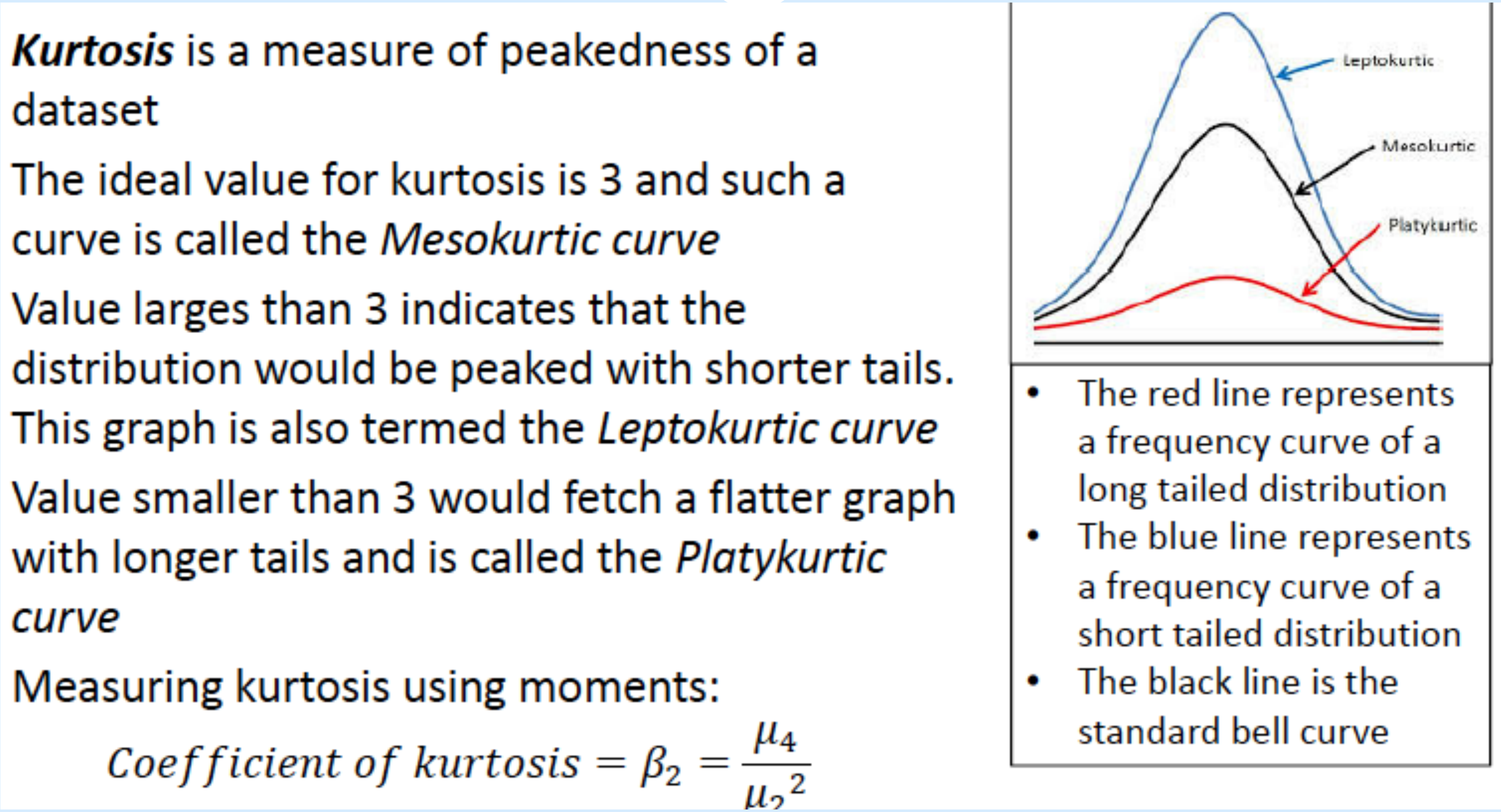
- ***Skewness*** is a measure of symmetry (or the lack of it) in a dataset
- A distribution is right-skewed or positively skewed if it stretches asymmetrically to the right
- It is left or negatively skewed if the asymmetric stretch is on the left
- Measuring skewness using moments:

$$Coefficient\ of\ skewness = \beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3}$$

- Important to note that if a distribution is perfectly symmetric, $\mu_3 = 0$
- The sign of the coefficient = the sign of μ_3
- A 'coefficient of skewness' value closer to zero, indicates a highly symmetric distribution



Kurtosis



Kurtosis is a measure of peakedness of a dataset

The ideal value for kurtosis is 3 and such a curve is called the *Mesokurtic curve*

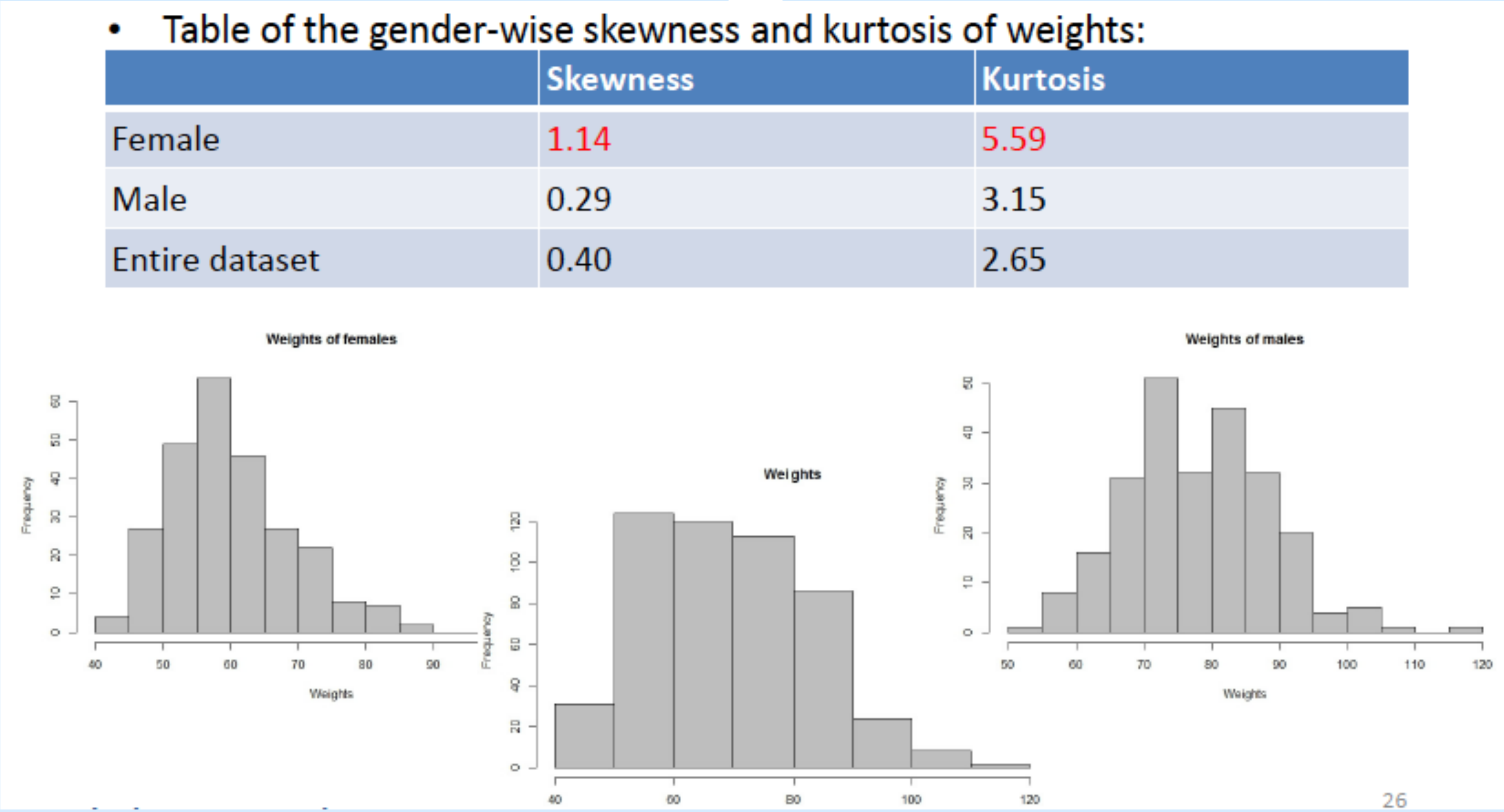
Value larges than 3 indicates that the distribution would be peaked with shorter tails. This graph is also termed the *Leptokurtic curve*

Value smaller than 3 would fetch a flatter graph with longer tails and is called the *Platykurtic curve*

Measuring kurtosis using moments:

$$Coefficient\ of\ kurtosis = \beta_2 = \frac{\mu_4}{{\mu_2}^2}$$

- The red line represents a frequency curve of a long tailed distribution
- The blue line represents a frequency curve of a short tailed distribution
- The black line is the standard bell curve



- Table of the gender-wise skewness and kurtosis of weights:

	Skewness	Kurtosis
Female	1.14	5.59
Male	0.29	3.15
Entire dataset	0.40	2.65

We see that skewness and kurtosis captures the numeric measure of the information presented in a histogram

We see that the histogram of ‘weights of females’ is highly stretched on the right, leading to a positive and high skewness measure of 1.14

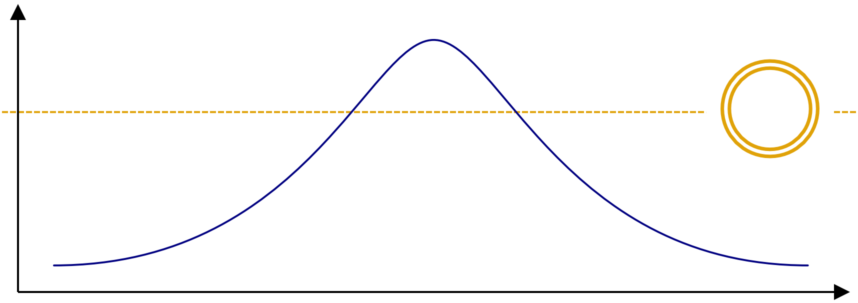
The stretch of histogram for weights of the entire dataset is moderate and much lesser than that for weights of females. This is reflected in the slightly lower skewness of 0.40

The weights of males are stretched almost equally on both sides of the centrality giving a skewness measure as close to zero as 0.29

Skewness and Kurtosis shed light on important characteristics such as symmetry and peakedness

Give additional information about distribution of data, than the measures of central tendency and measures of dispersion

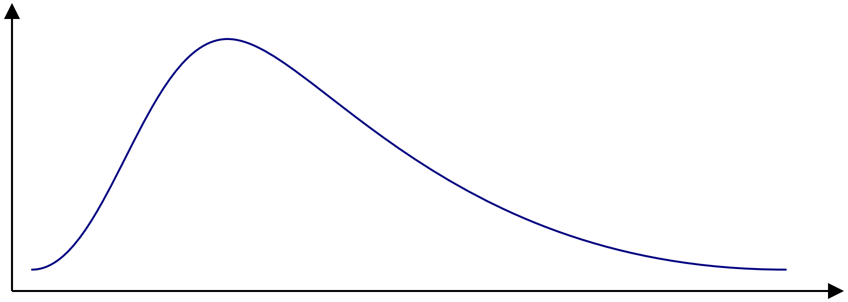
Shapes of distribution



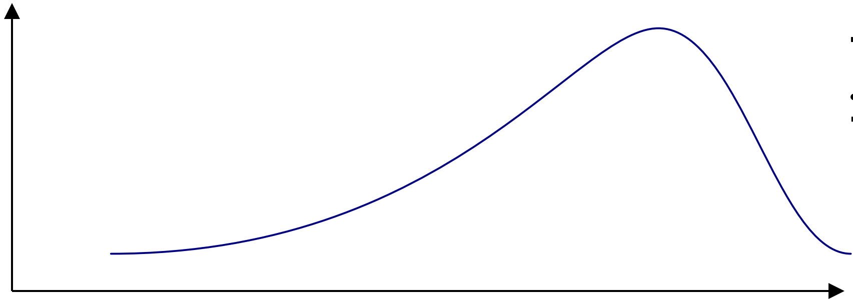
←**Normal distribution**: symmetrical Bell-shaped curve

symmetrical

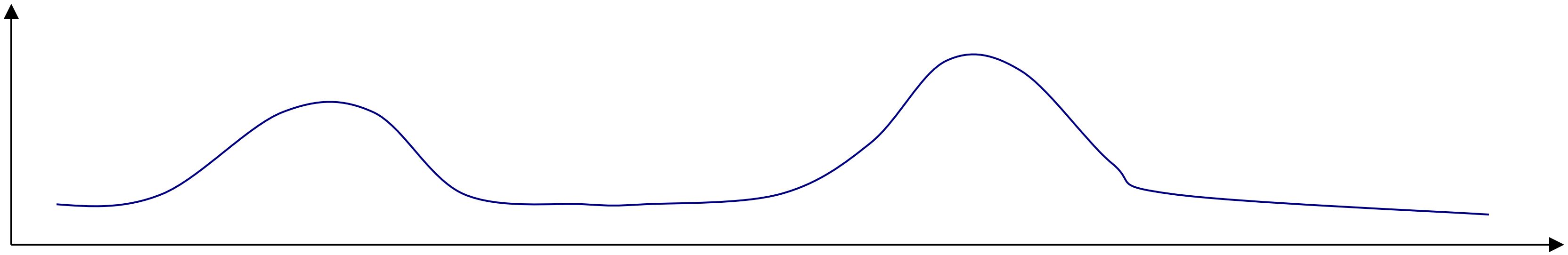
asymmetrical



Positively skewed:
tail on the right, cluster towards low end of the variable

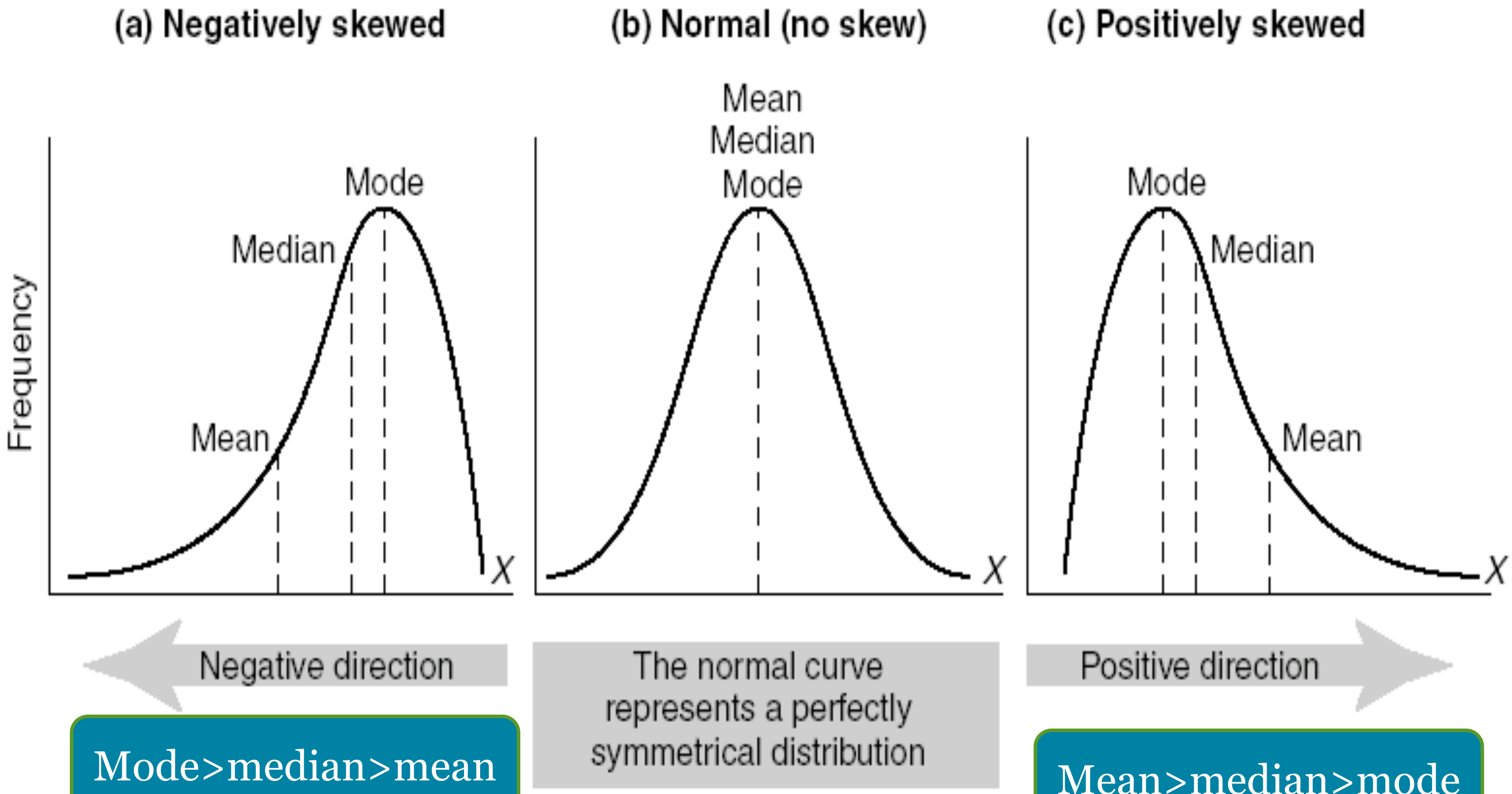


Negatively skewed:
tail on the left, cluster towards high-end of the variable



Bimodality: A double peak

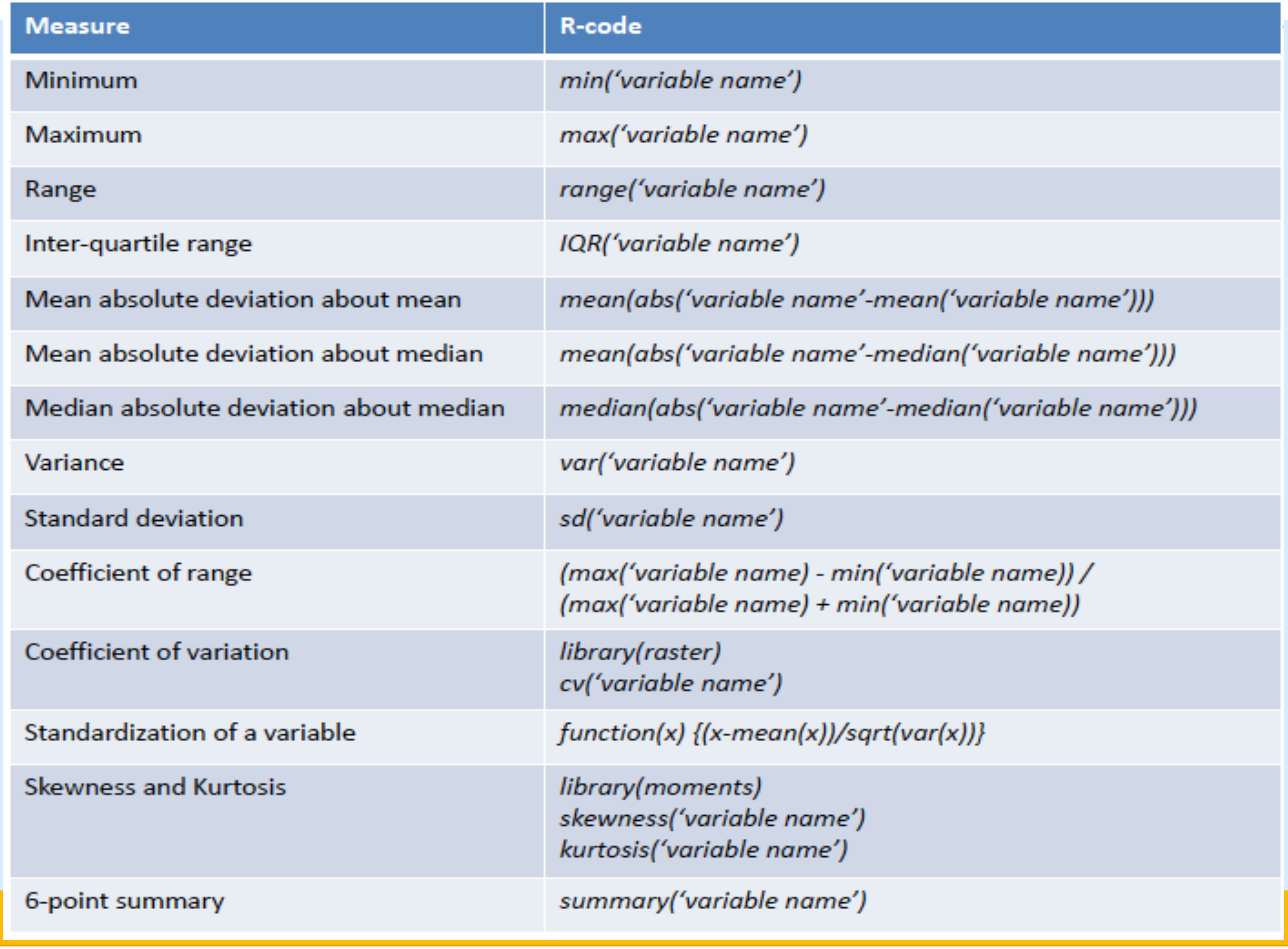
Position of mean median mode



Mode>median>mean

Mean>median>mode

R Code



Measure	R-code
Minimum	*min(‘variable name’)*
Maximum	*max(‘variable name’)*
Range	*range(‘variable name’)*
Inter-quartile range	*IQR(‘variable name’)*
Mean absolute deviation about mean	*mean(abs(‘variable name’-mean(‘variable name’)))*
Mean absolute deviation about median	*mean(abs(‘variable name’-median(‘variable name’)))*
Median absolute deviation about median	*median(abs(‘variable name’-median(‘variable name’)))*
Variance	*var(‘variable name’)*
Standard deviation	*sd(‘variable name’)*
Coefficient of range	*(max(‘variable name) - min(‘variable name)) / (max(‘variable name) + min(‘variable name))*

Coefficient of variation	*library(raster)*
	cv(‘variable name’)

Standardization of a variable	*function(x) {(x-mean(x))/sqrt(var(x))}*

Skewness and Kurtosis	*library(moments)*
	skewness(‘variable name’)
	kurtosis(‘variable name’)

6-point summary	*summary(‘variable name’)*

Frequency Table

- Table listing frequency counts for each value
- Provides quick info about data values
- Easily constructed and understood
- Can be used for both categorical and numerical data
- Eg. Cars dataset
- Let us explore it
  - What is this data about: Used cars in USA
  - What variables are given: 12.
  - How many observations: 804. Price, make, model, doors, cylinders…
  - R Code:table(variable name)

- For *Cars* data, let us take a look at various frequency tables:

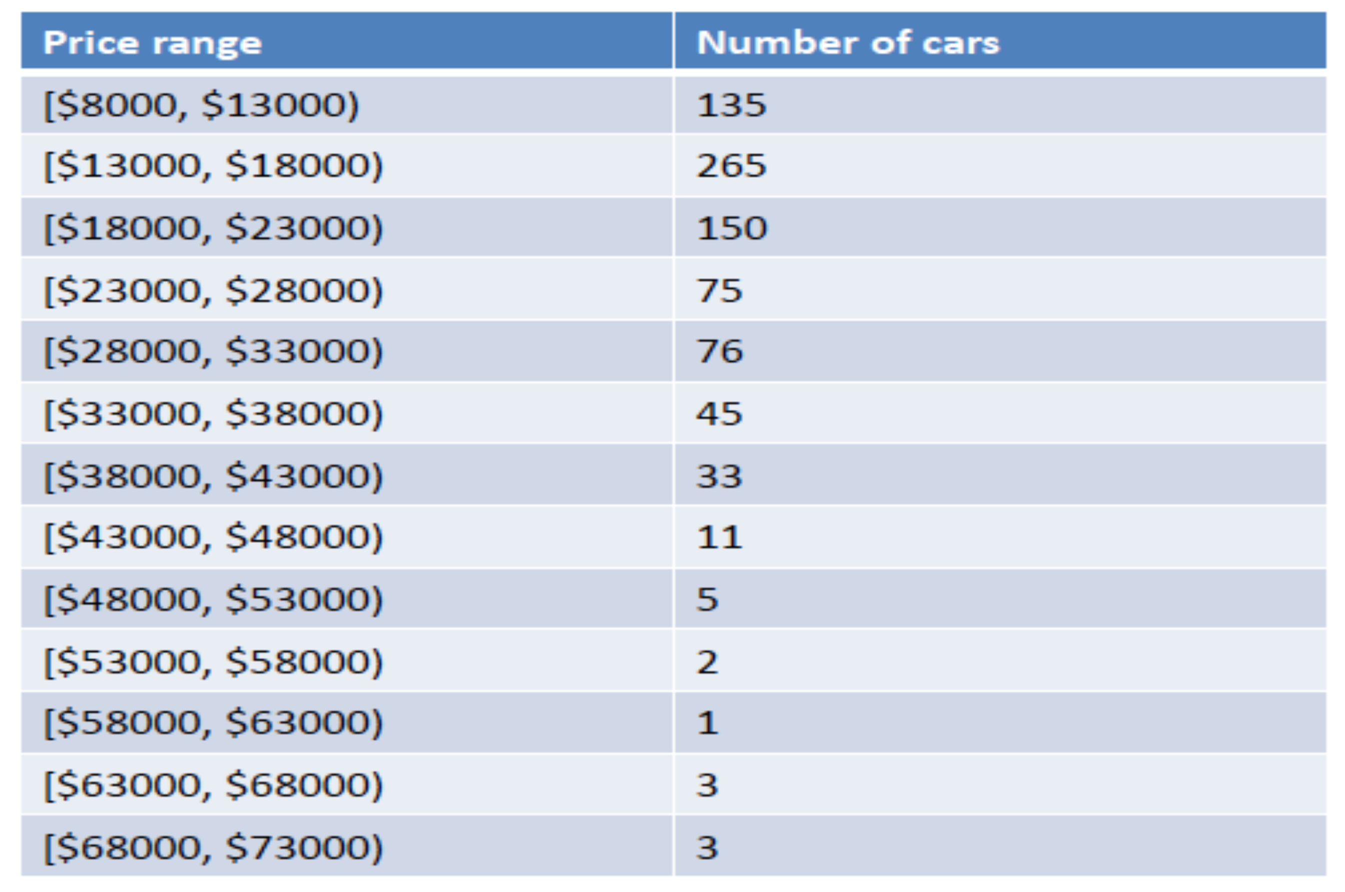
Car make	Frequency of each make
Buick	80
Cadillac	80
Chevrolet	320
Pontiac	150
SAAB	114
Saturn	60

No. of cylinders	Frequency of cars with corresponding no. of cylinders
4	394
6	310
8	100

Price	Frequency
8638.93	1
8769	1
8870.95	1
9041.91	1
9220.83	1
9482.22	1
9506.05	1
9563.79	1
9654.06	1
9665.85	1
9720.98	1
...	...
...	...
...	...

There are 798 unique prices!

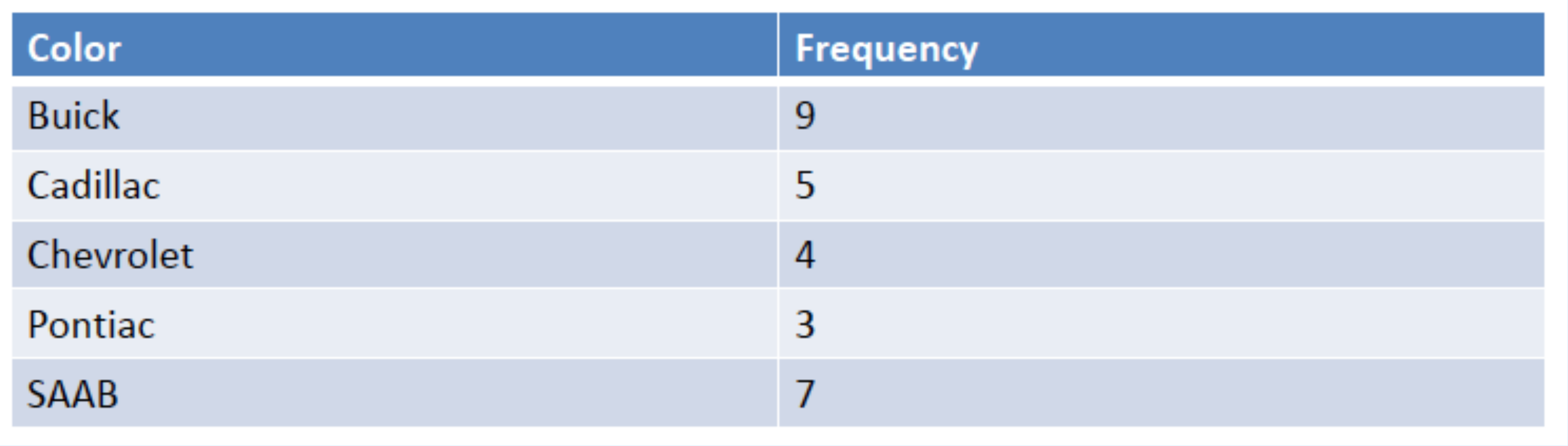
- What about prices?
  - We can visualize them by putting into bins



Price range	Number of cars
[$8000, $13000)	135
[$13000, $18000)	265
[$18000, $23000)	150
[$23000, $28000)	75
[$28000, $33000)	76
[$33000, $38000)	45
[$38000, $43000)	33
[$43000, $48000)	11
[$48000, $53000)	5
[$53000, $58000)	2
[$58000, $63000)	1
[$63000, $68000)	3
[$68000, $73000)	3

Example

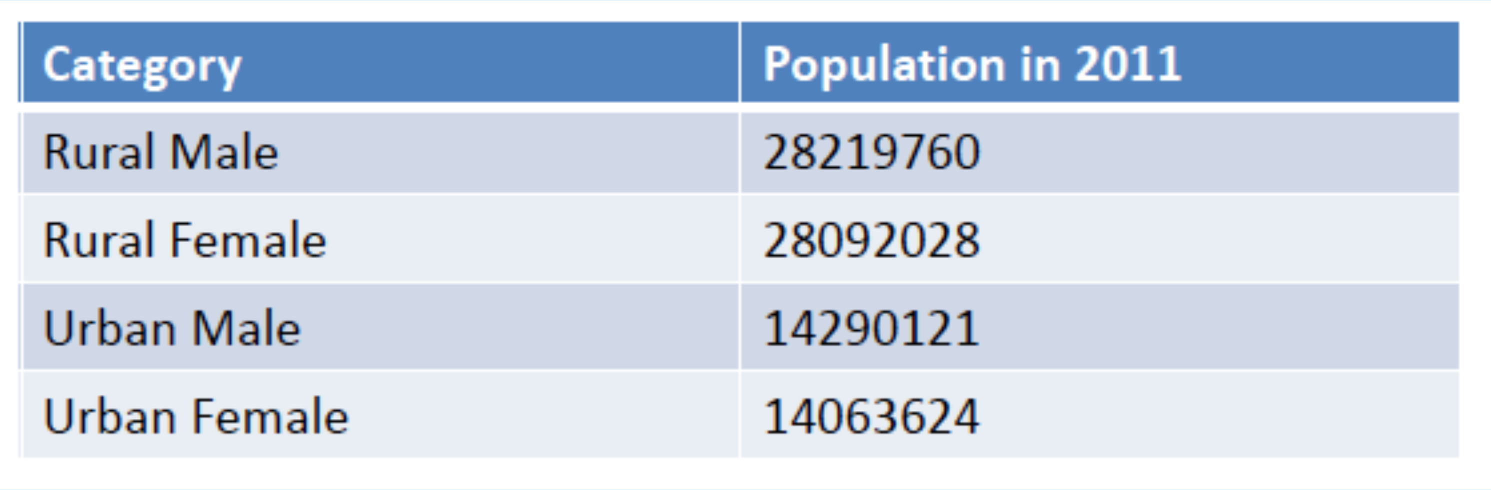
- A market survey firm is conducting a survey on the popularity of different makes of car in the USA. For this, it investigates various car showrooms and lists down the various car varieties in each showroom
- Suppose it obtains the following data on the makes of 28 cars in one such showroom :
  - Buick, Cadillac, Buick, Chevrolet, Buick, Buick, Buick, Pontiac, Cadillac, Chevrolet, SAAB, SAAB, SAAB, Cadillac, Chevrolet, Pontiac, Buick, Buick, Buick, SAAB, Cadillac, SAAB, Pontiac, SAAB, Chevrolet, Buick, SAAB, Cadillac.
- Relevant Questions:
  - Which is the most common car in the showroom?
  - Which is the least common car in the showroom?
- Frequency table can help answer them succintly



Color	Frequency
Buick	9
Cadillac	5
Chevrolet	4
Pontiac	3
SAAB	7

Graphs

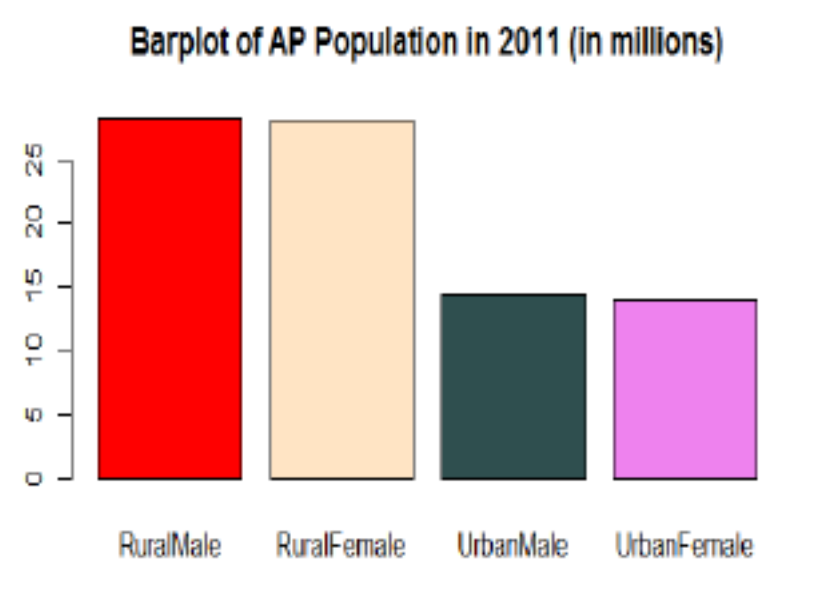
- In addition to frequency table, we can use graphs
  - Bar Chart
  - Pie Chart
- Example. Consider the following

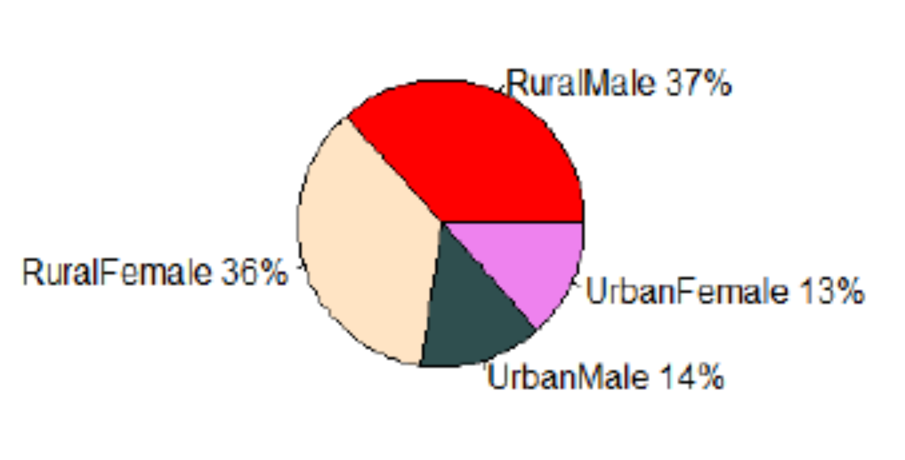


Category	Population in 2011
Rural Male	28219760
Rural Female	28092028
Urban Male	14290121
Urban Female	14063624

Bar & Pie Chart

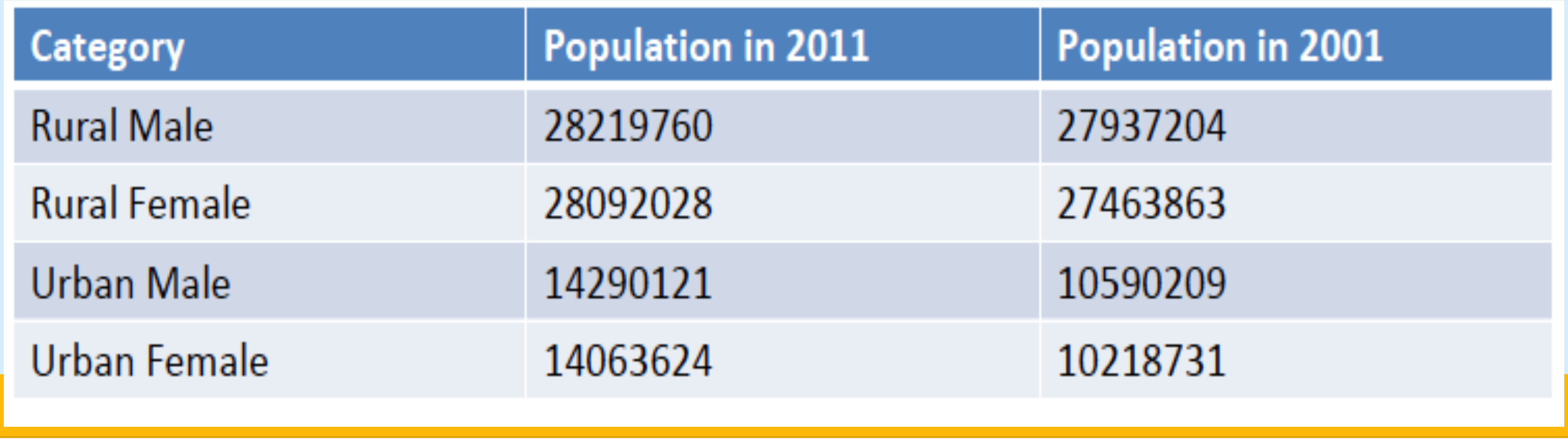
- Say, you want to know population composition in a state.
  - Bar chart allows you to do that
- How about relative share?
  - Pie chart helps here





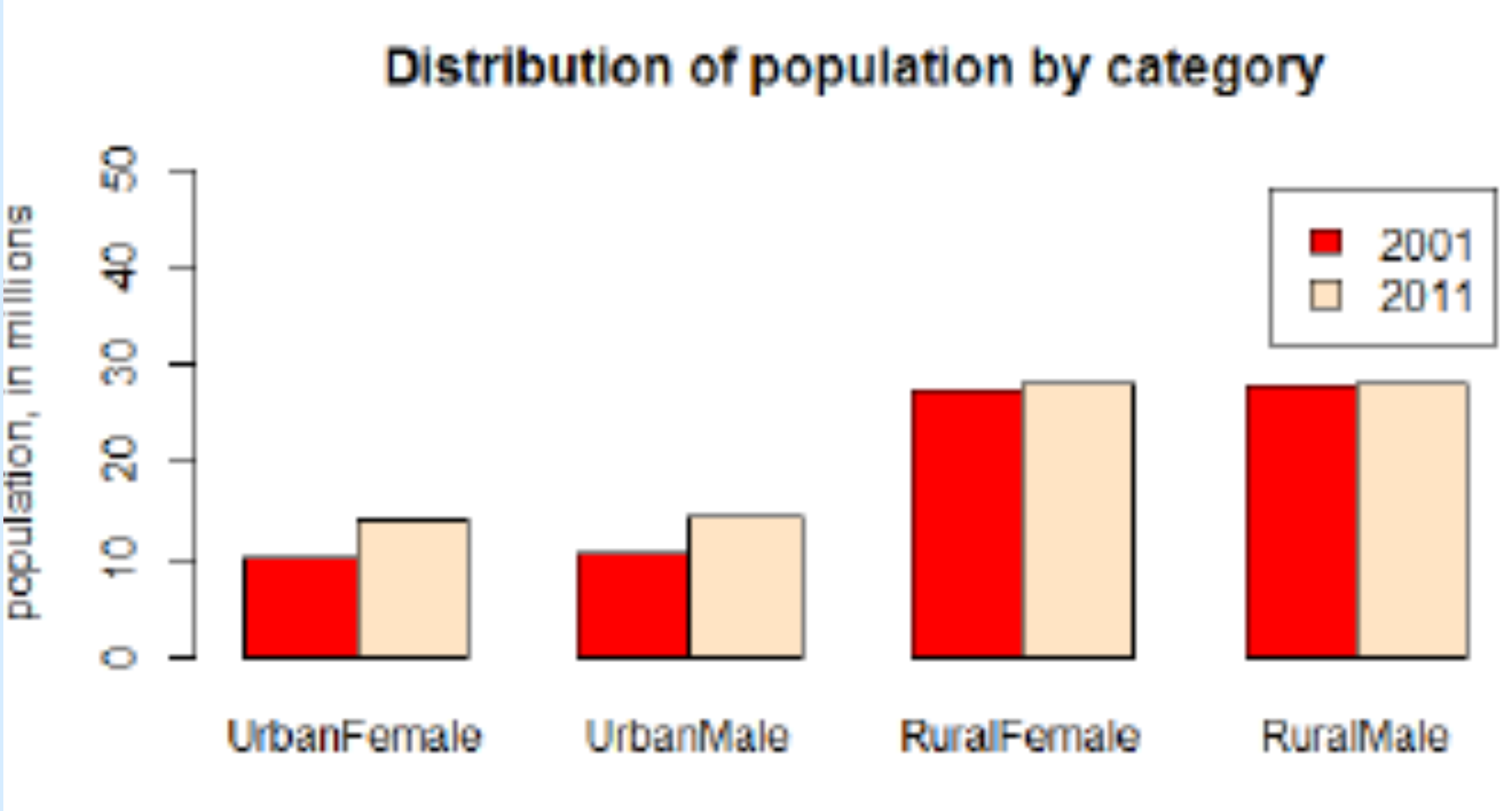
Multiple Bar Graph

- Suppose we want to look at time series properties
  - Population growth
- Bar and pie charts for population in 2001 and 2011
- They allow comparison within category or over years, but not both
- Multiple bar graph helps here



Category	Population in 2011	Population in 2001
Rural Male	28219760	27937204
Rural Female	28092028	27463863
Urban Male	14290121	10590209
Urban Female	14063624	10218731

Multiple Bar Graph



R Code

- #data
- APPopulation = cbind(c(28219760,28092028,14290121,14063624),c(27937204,27463863,10590209,10218731))
- rownames(APPopulation) = c("RuralMale","RuralFemale","UrbanMale","UrbanFemale")
- colnames(APPopulation) = c("2011","2001")

- #barplot
- colors=c("red", "bisque", "darkslategray", "violet")
- barplot(APPopulation[,"2011"]/1000000,col=colors)
- title(main="Barplot of AP Population in 2011 (in millions)")

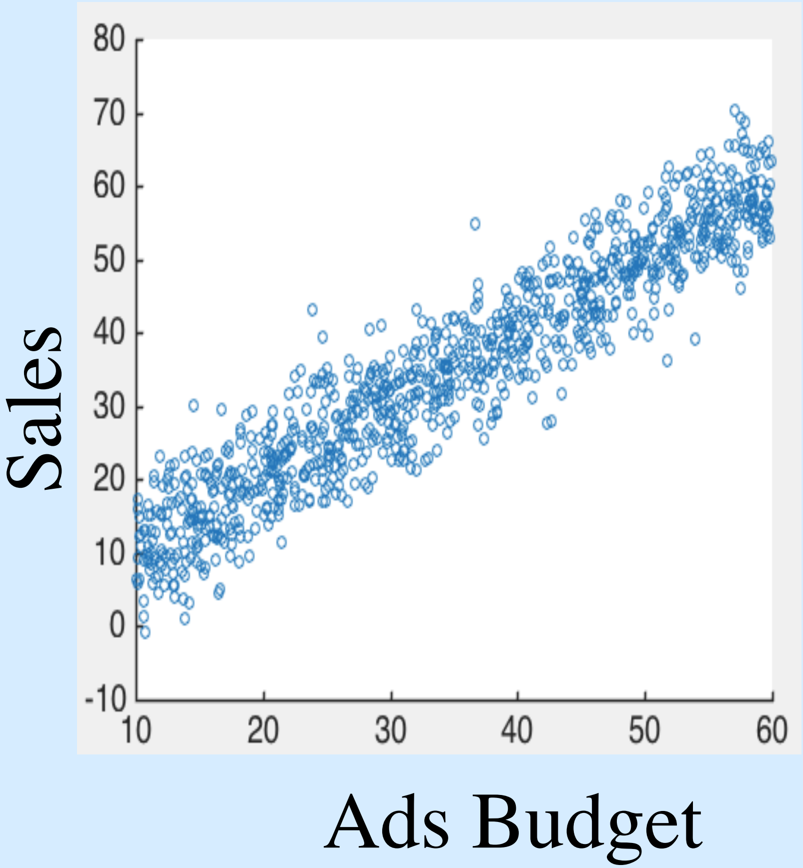
- # Multiple Bar Graph:
- A = matrix(c(10218731,10590209,27463863,27937204,14063624,14290121,28092028,28219760), nrow=2, ncol=4, byrow = TRUE)
- colors=c("red", "bisque")
- barplot(A/1000000,names.arg=rev(rownames(APPopulation)),legend.text=c(2001,2011),beside=TRUE,main="Distribution of population by category",xlab="Categories", ylab="population, in millions",ylim=c(0,80),col=colors)

R Code

- # Pie Chart
- colors=c("red", "bisque", "darkslategray", "violet")
- slices <- c(27937204,27463863,10590209,10218731)
- lbls <- c("RuralMale","RuralFemale","UrbanMale","UrbanFemale")
- pct <- round(slices/sum(slices)*100)
- lbls <- paste(lbls, pct) # add percents to labels
- lbls <- paste(lbls,"%",sep="") # ad % to labels
- pie(slices,labels = lbls, col=rainbow(length(lbls)), main="Pie Chart of APPopulation in 2011")

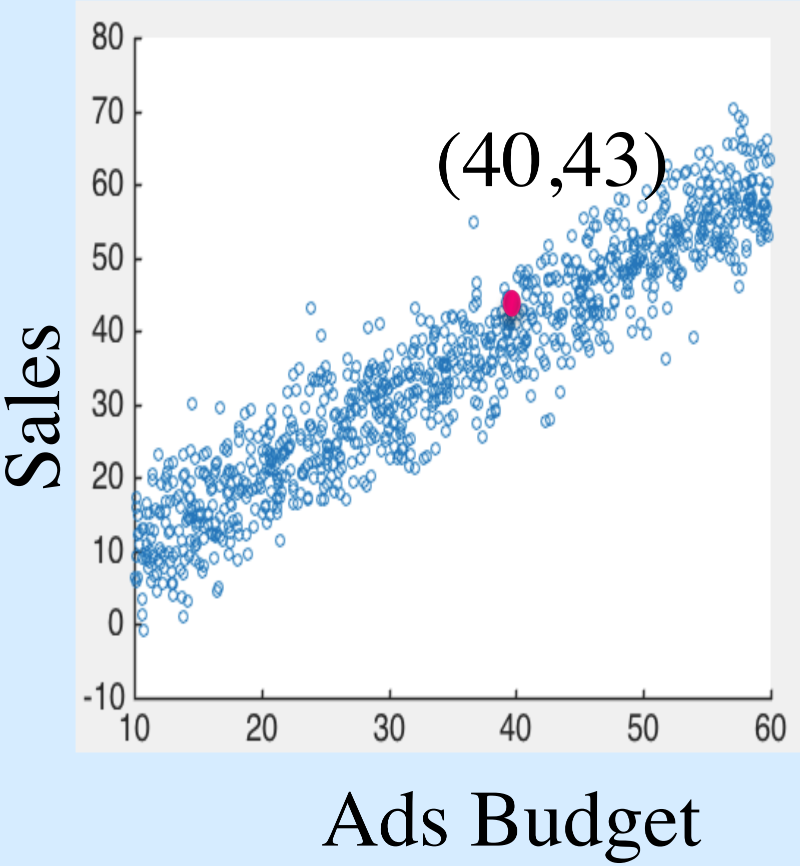
Scatter Plots

Ads Budget	Sales
40	43
15	18
27	24
35	38
10	8
17	14
:	:



Scatter Plots

Ads Budget	Sales
40	43
15	18
27	24
35	38
10	8
17	14
:	:



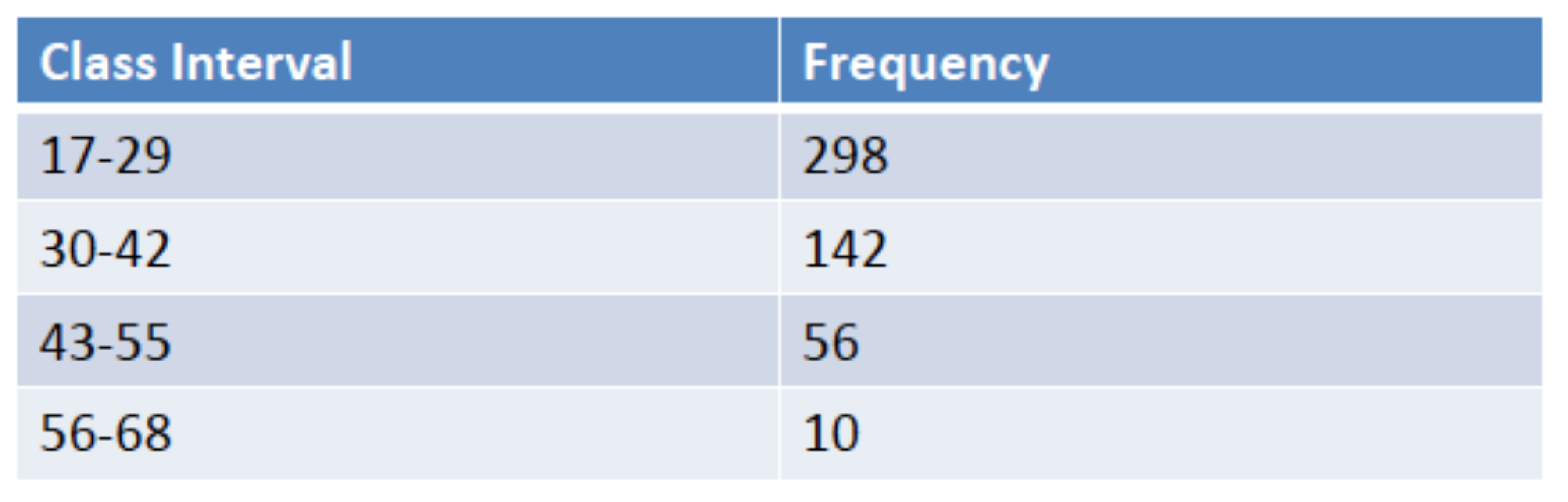
Example

- A private insurance firm interested in marketing it's insurance products in region A. To target precisely, needs to know age distribution.
- Questions-
  - In which age group does the highest number of people lie.
  - Needs to divide population into 4 different age groups, to sell 4 different products

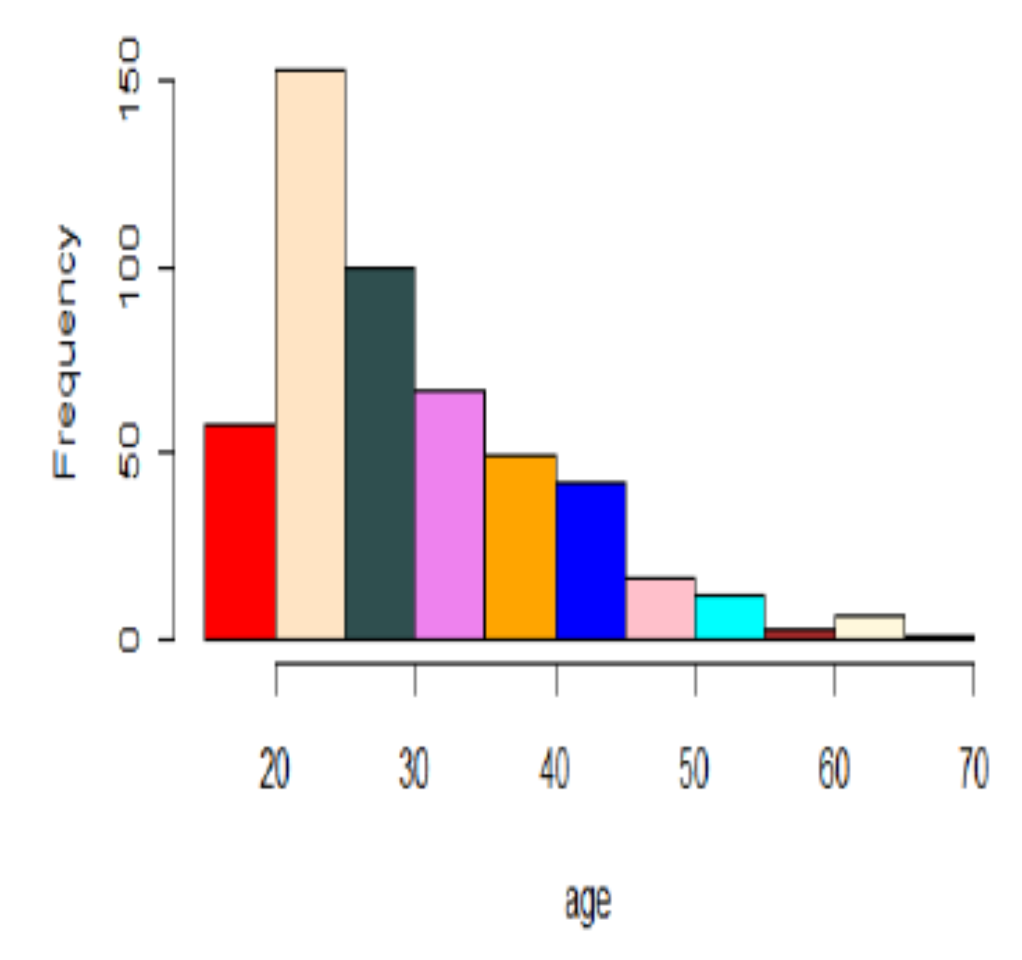
- Data
- 23,21,23,26,22,27,29,37,55,53,21,19,20,18,32,20,28,19,23,33,40,28,24,36,23,29,34,31,34,42,45,23,46,26,30,25,20,37,24,36,28,29,23,23,25,24,37,42,30,28,29,39,26,20,21,20,19,20,40,25,45,28,21,22,19,24,24,20,29,27,27,40,43,22,22,21,24,23,23,45,20,25,25,33,21,23,20,34,20,41,25,32,24,65,28,25,38,23,22,20,35,34,67,38,33,26,25,52,21,32,43,24,28,62,45,40,21,23,30,20,28,41,32,26,37,38,27,23,50,25,23,43,33,22,26,37,32,23,37,23,27,23,27,24,21,25,23,23,46,34,25,29,45,44,35,55,25,31,19,45,34,19,20,29,33,37,21,23,51,31,27,27,37,25,37,33,25,29,25,20,25,28,24,31,25,27,23,20,28,40,21,62,44,49,34,25,29,19,20,20,26,19,36,34,24,27,23,20,28,40,21,62,44,49,34,25,29,19,20,20,26,19,36,34,24,27,22,22,48,21,27,33,34,54,25,35,22,21,41,23,19,29,27,36,21,20,20,24,35,33,25,45,55,49,30,28,25,23,26,21,26,32,32,32,35,19,26,22,23,25,38,30,43,60,32,26,23,24,21,28,25,20,64,39,27,32,23,24,23,29,44,20,24,42,27,43,37,20,47,45,20,28,21,37,27,26,22,21,62,27,27,22,22,52,42,30,19,19,19,24,21,36,32,52,26,56,30,23,21,44,37,51,38,23,44,26,23,20,44,25,18,22,35,24,25,23,22,24,26,26,28,34,24,33,46,51,25,19,35,19,19,20,41,33,44,19,29,35,33,22,33,44,29,46,19,30,26,20,32,20,27,22,40,42,29,31,22,29,36,37,25,46,25,43,43,24,24,19,46,29,26,32,29,34,26,34,22,25,41,38,21,34,37,56,28,35,29,22,22,24,36,40,40,37,23,34,20,23,40,20,30,32,30,21,39,37,22,39,49,24,20,40,24,39,32,24,22,20,27,21,26,28,26,18,30,22,30,18,52,25,28,42,23,41,32,22,24,25,27,24,27,31,35,21,36,20,23,19,25,31,32,40,41,36,43,34,26,29,23,45,33,29,29,45,48,19,38,26,48,22,32,44,44,19,32,30

Derive Insights

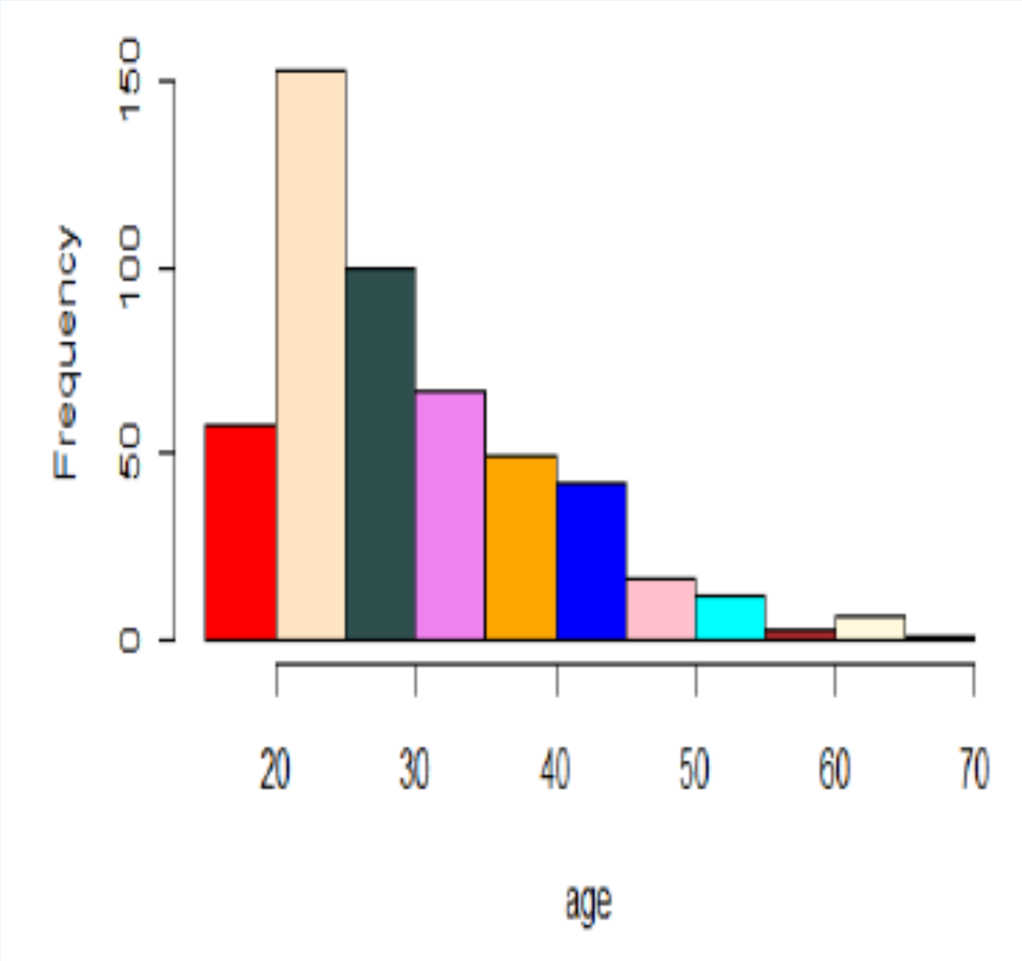
- What can you infer from the data? Practically nothing!
- First thing, let us create a frequency table
- Note: Age is, in theory, a continuous variable as it can assume any value.
- But here the variable is, age, in whole years, which is discrete.
- But 44 distinct values in your data!
- Hence frequency table with 44 rows and one frequency column
- So, should we look at individual ages or groups?
  - What does question asks? 4 categories for 4 products
- How to create groups?
  - Find max and min. Choose suitable class width= (max-min)/(desired no of classes)



Class Interval	Frequency
17-29	298
30-42	142
43-55	56
56-68	10

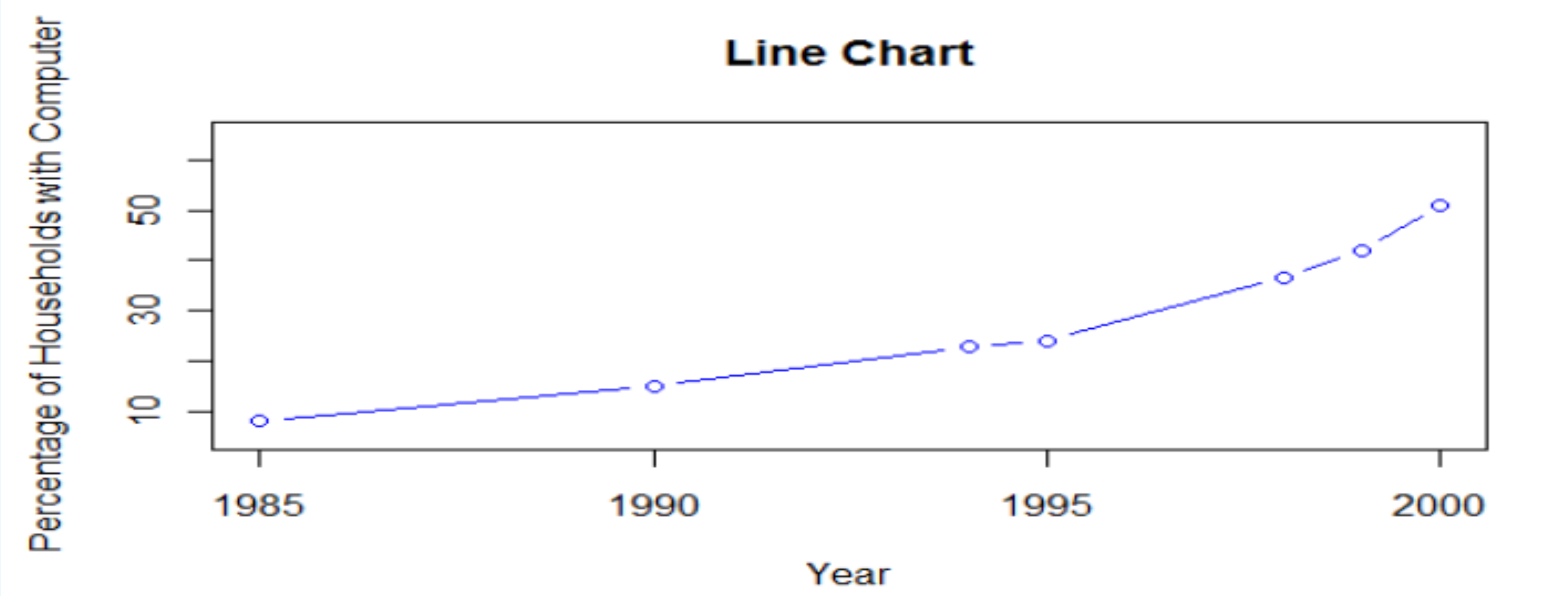

Frequency
0
50
100
150
20
30
40
50
60
70
age

- Tell us shape
  - Not symmetric
  - Right Skewed
- Variation?



Line Chart

- Examples before are cross sectional data
- For time-series data, we use line charts

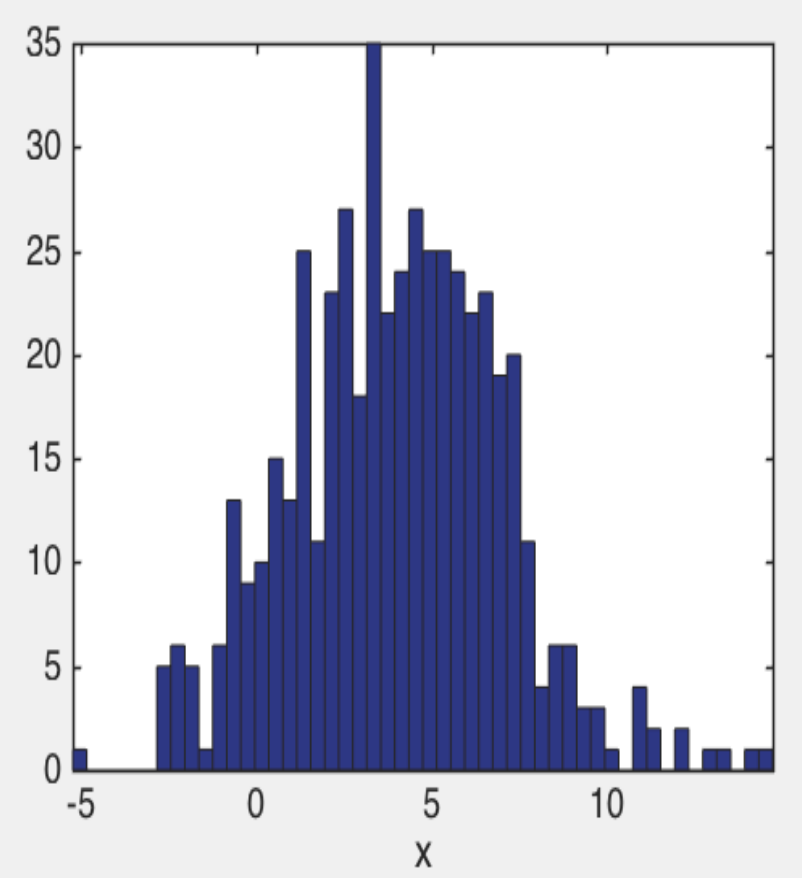


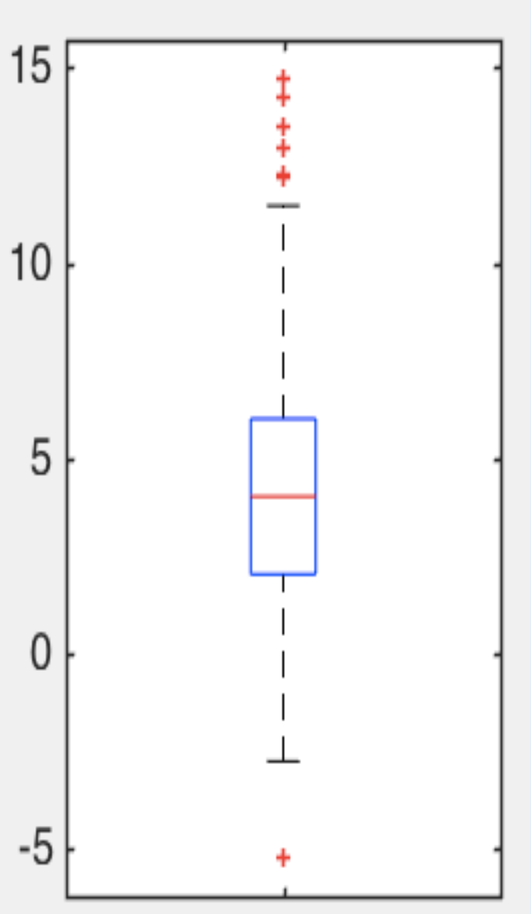
- Inferences?

Box Plot

- Visual tool for exploratory data analysis
- Visualize important summary statistics measures
  - Mean, Spread, Distribution, Symmetry, Skewness
- Useful for looking at tail observations

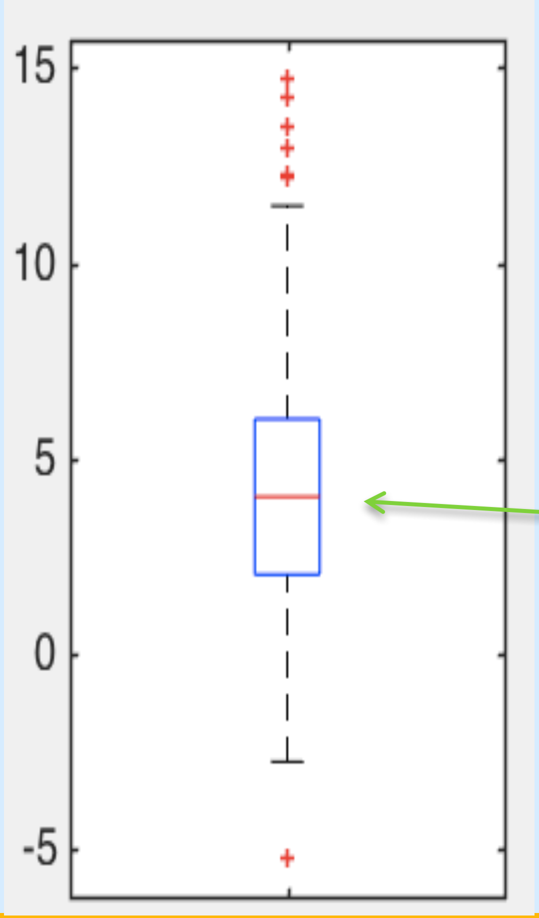
Box Plot





Box Plot

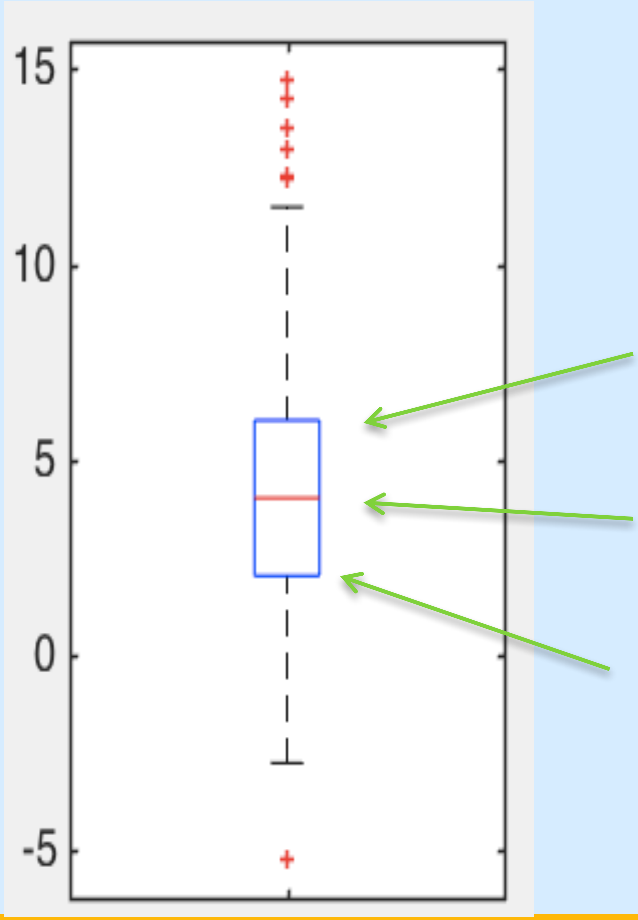
- Alternative to a histogram



Median, Q_2

Box Plot

- Alternative to a histogram



Third quartile, Q_3

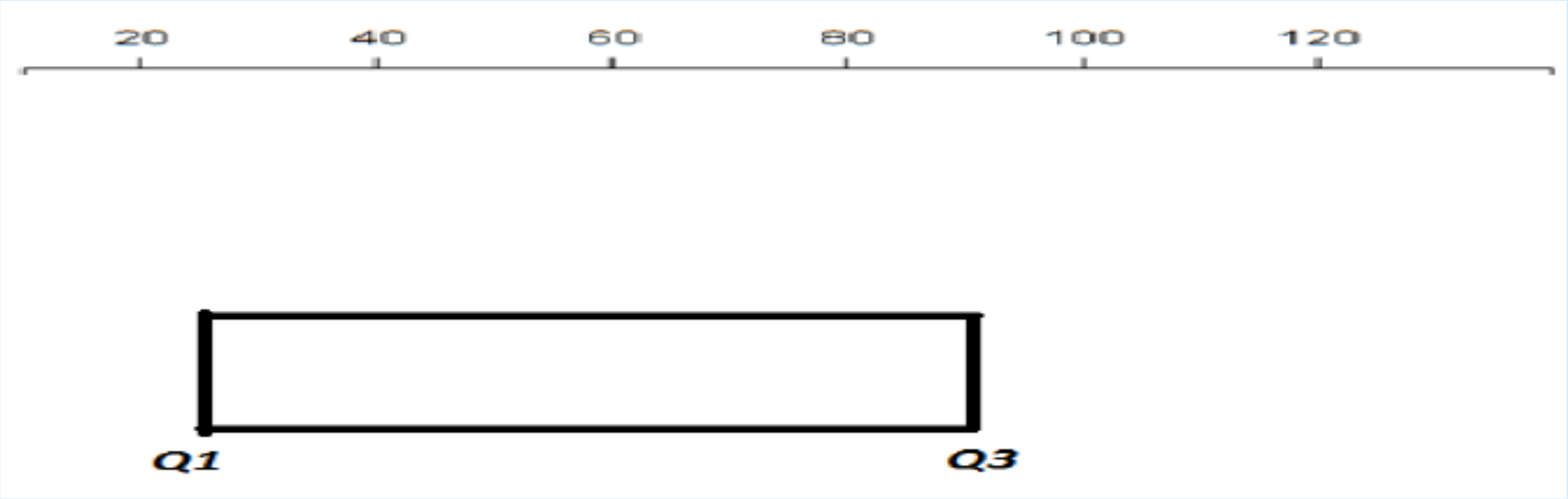
Median, Q_2

First quartile, Q_1

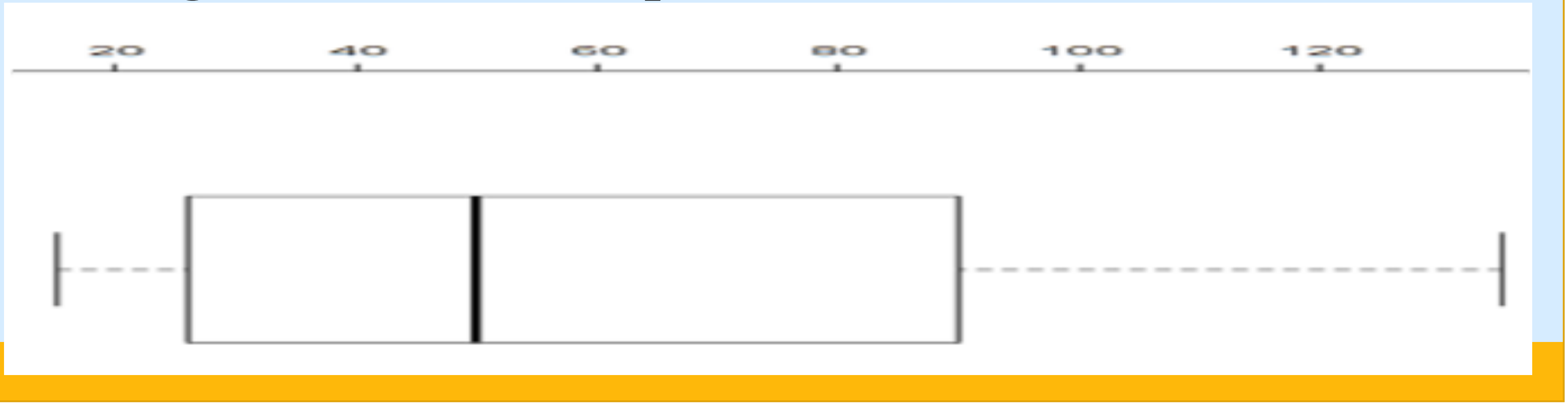
Box-Plot

- Data set 1. Suppose we have data on a batch (variable)
  - 90, 41, 22, 135, 15, 72, 50, 26, 105
- Step 1: Arrange the data in the increasing order:
  - 15, 22, 26, 41, 50, 72, 90, 105, 135
- Step 2: Get the Five-point Summary, consisting of (i) the Minimum, (ii) First quartile, (iii) Median, (iv) Third quartile and (v) the Maximum
- For the above data, the Five-point Summary is:
  - Minimum= 15
  - First Quartile= 26
  - Median= 50
  - Third Quartile= 90
  - Maximum= 135

- Step 3: Draw a box of length equal to **(Q3 – Q1)**. For now, we can choose the width as per convenience. The lower and upper hinges of the box represent the first and third quartiles.

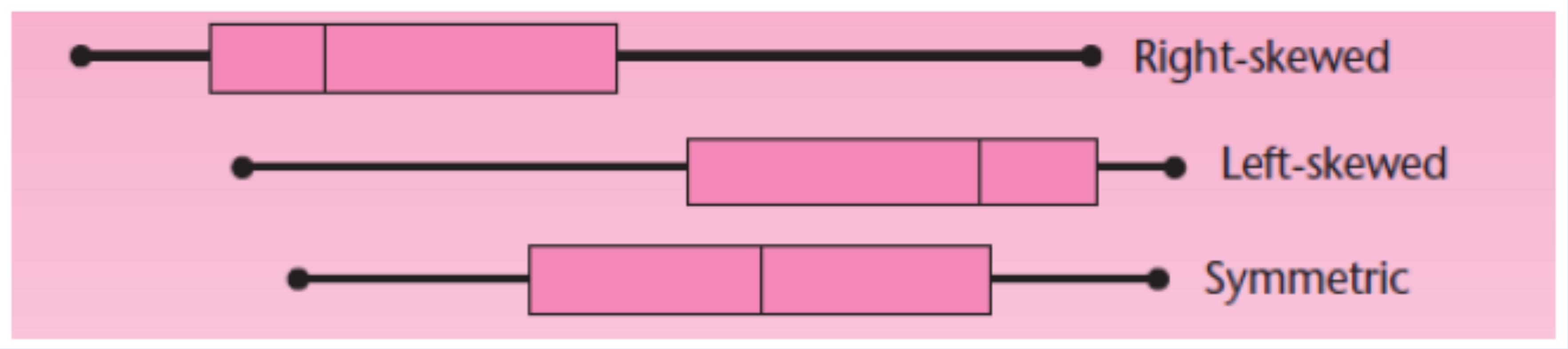


- Step 4: From the middle of the lower hinge draw a line (parallel to the lines corresponding to the length of the box) up to the minimum. Similarly draw a line from the middle of the upper hinge (parallel to the lines corresponding to the length of the box) up to the maximum. These lines are called the whiskers.
- Step 5: Draw a line at the median parallel to the hinges, dividing the box into two parts.

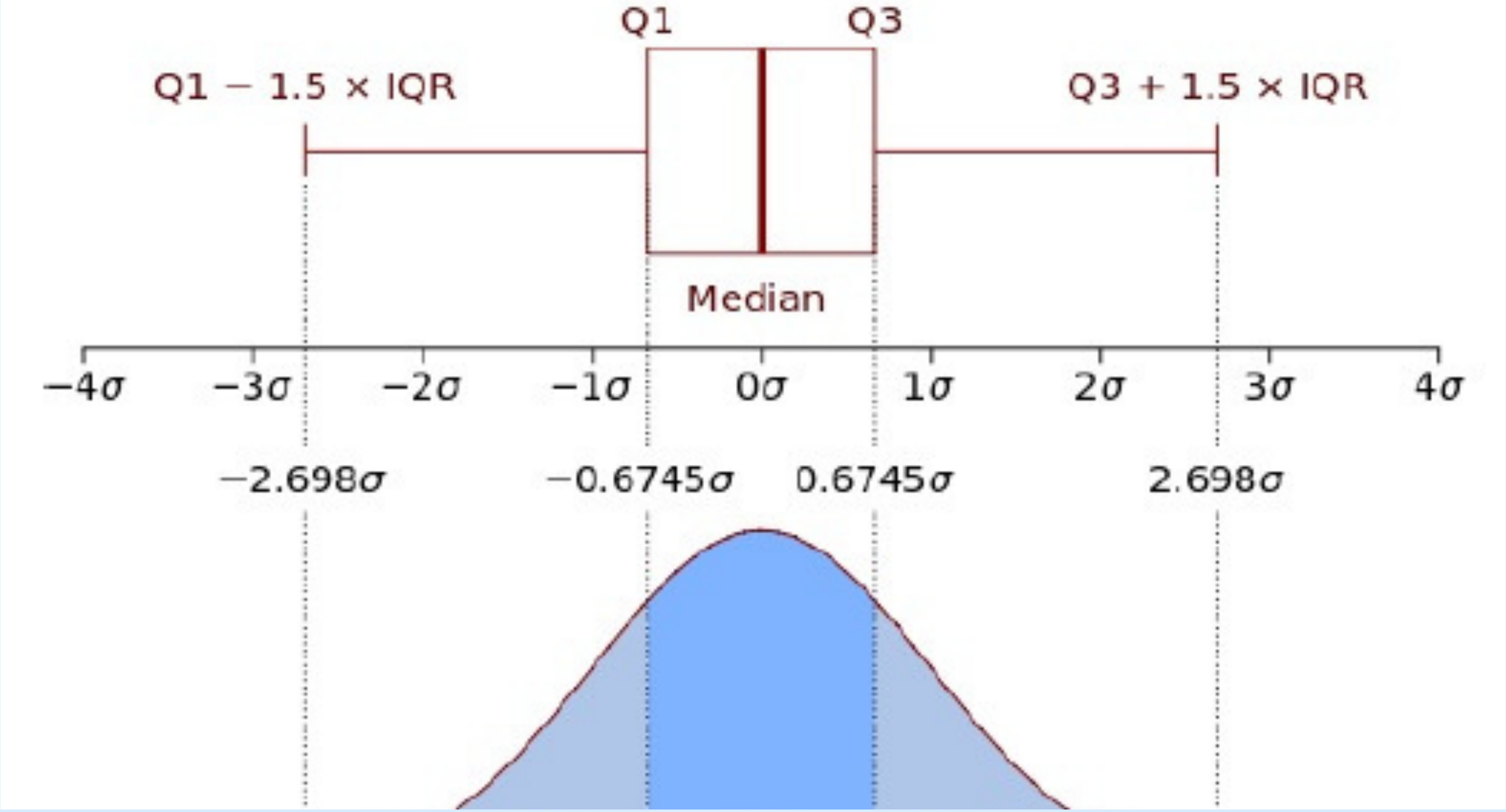


Box-Plot

- Info from a box plot
  - Spread, concentration of values
  - What else?



Bit more about Box-Plot

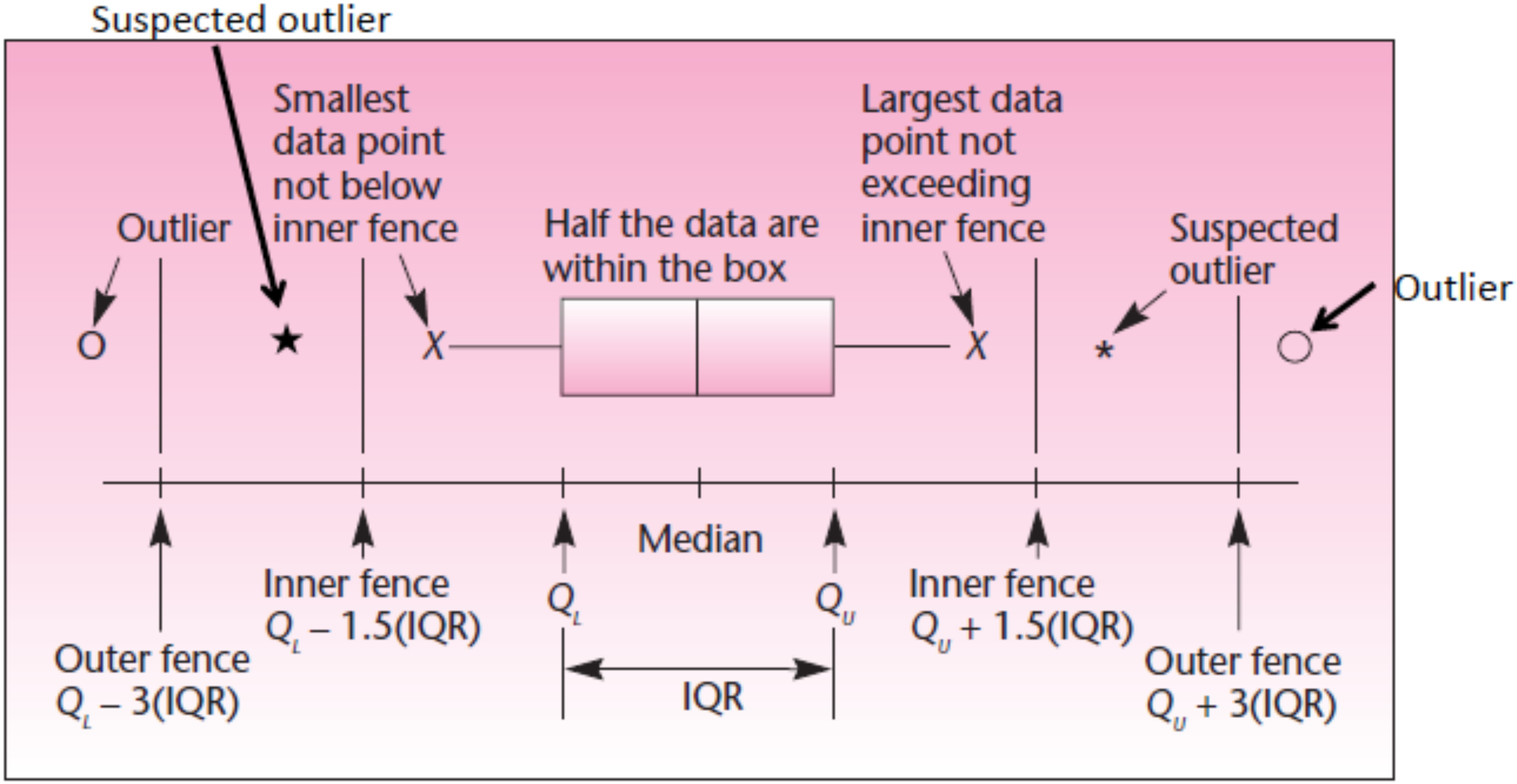


From the previous figure, we see that for a normally distributed data, 99.3% of the data lies in the interval

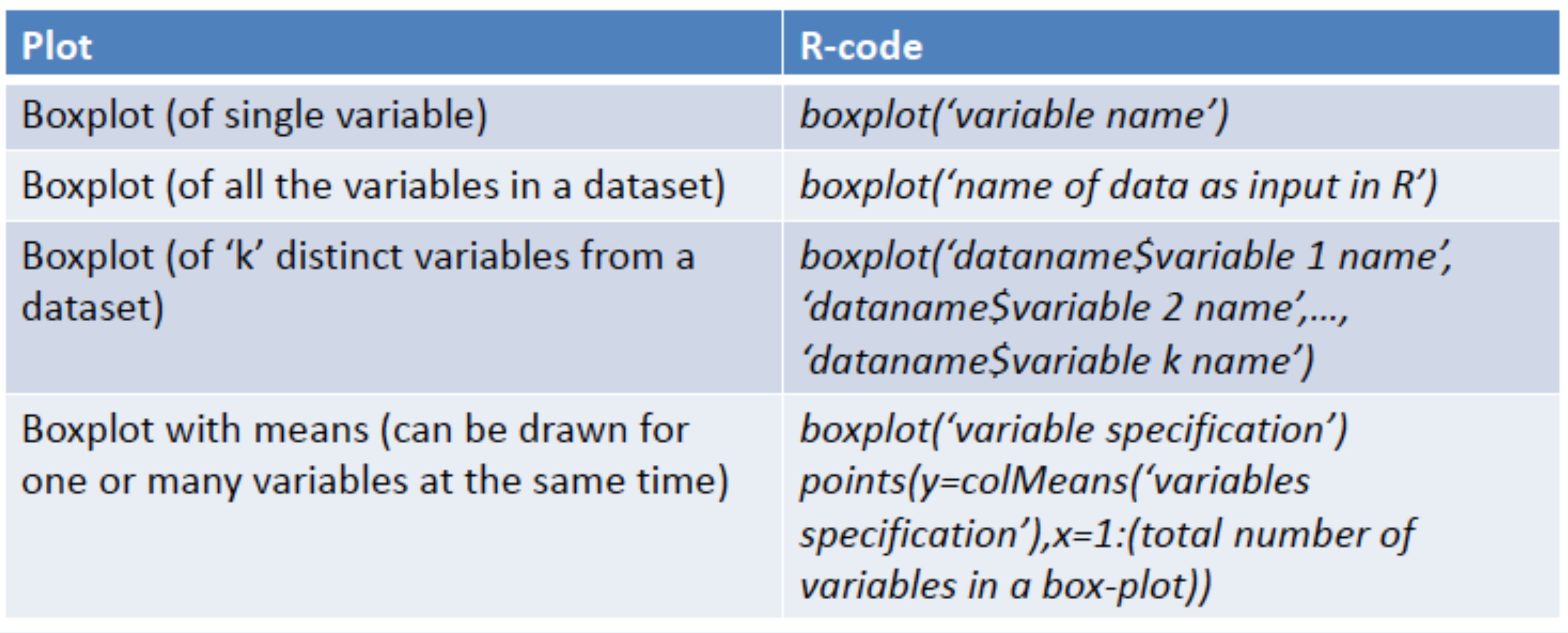
$$\boxed{(Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1))}$$

Also, only 3 out of a million or 0.003% observations are expected to be present outside the interval

$$\boxed{(Q_1 - 3(Q_3 - Q_1), Q_3 + 3(Q_3 - Q_1))}$$


Suspected outlier
Smallest data point not below inner fence
Outlier
Half the data are within the box
Largest data point not exceeding inner fence
Suspected outlier
Outlier
O
★
X
X
*
○
Median
Inner fence
$Q_L - 1.5(IQR)$
Q_L
Q_U
Inner fence
$Q_U + 1.5(IQR)$
Outer fence
$Q_L - 3(IQR)$
IQR
Outer fence
$Q_U + 3(IQR)$

R-Code



Plot	R-code
Boxplot (of single variable)	*boxplot('variable name')*
Boxplot (of all the variables in a dataset)	*boxplot('name of data as input in R')*
Boxplot (of 'k' distinct variables from a dataset)	*boxplot('dataname$variable 1 name', 'dataname$variable 2 name',…, 'dataname$variable k name')*
Boxplot with means (can be drawn for one or many variables at the same time)	*boxplot('variable specification') points(y=colMeans('variables specification'),x=1:(total number of variables in a box-plot))*