Descriptive Statistics

Introduction to Data Science

M. Sakthi Balan

Associate Professor
LNM Institute of Information Technology
Jaipur

Contents

LNMIIT

What is EDA?

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

- EDA was promoted by *John Tukey* to encourage statisticians to explore the data first
- Also to formulate hypotheses that could lead to new data collection and experiments.
- Tukey promoted the use of five number summary of numerical data – the two extremes (maximum and minimum), the median, and the quartiles.
  - defined for all distributions, unlike the mean and standard deviation
  - the quartiles and median are more robust to skewed or heavy-tailed distributions than traditional summaries

What is EDA?

Examine the variables one by one by getting a summary and try to get a visualization that may give a clear view of the distribution of the variable. This type of analysis is called as *Univariate data analysis*

- Useful method to check the quality and the quantity of the data
- Inconsistencies and unexpected values can be addressed by some appropriate means
- It gives the clear distribution of the data:
  - How dense the data is?
  - Any outliers?
  - Does the outliers make sense?
  - Any peculiar pattern in the data?

What is EDA?

Examine the variables one by one by getting a summary and try to get a visualization that may give a clear view of the distribution of the variable. This type of analysis is called as *Univariate data analysis*

- Useful method to check the quality and the quantity of the data
- Inconsistencies and unexpected values can be addressed by some appropriate means
- It gives the clear distribution of the data:
  - How dense the data is?
  - Any outliers?
  - Does the outliers make sense?
  - Any peculiar pattern in the data?

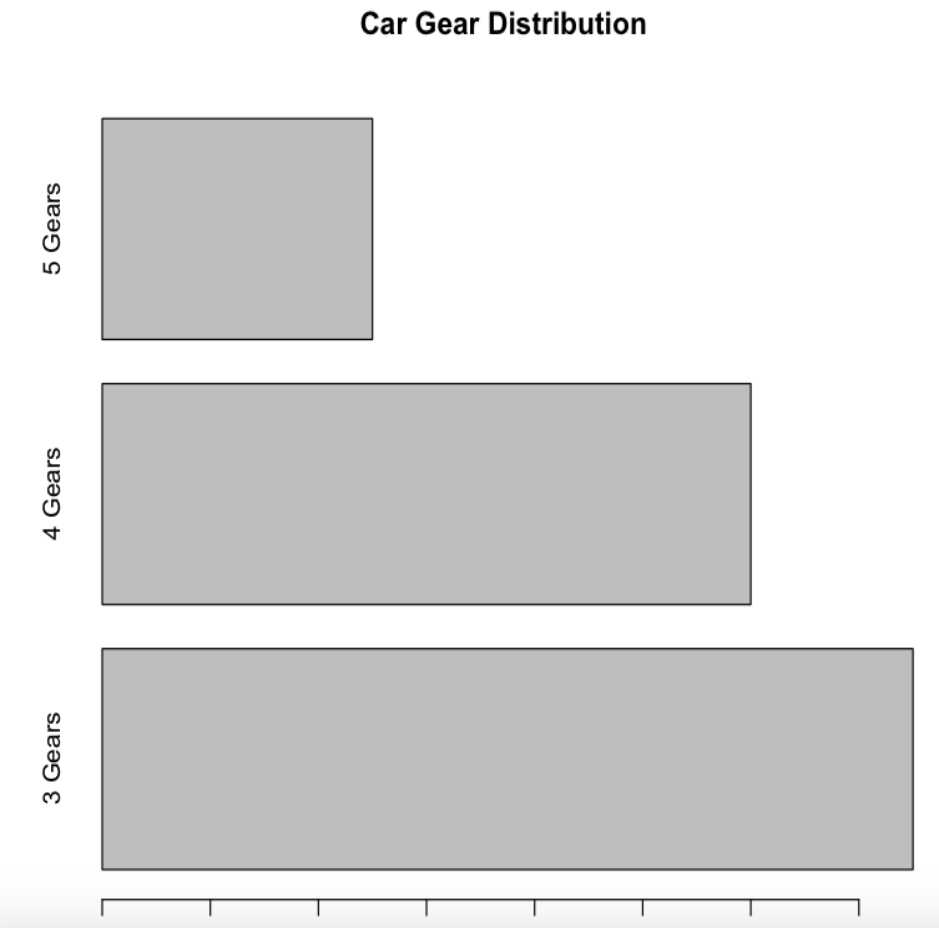
- Summarizing the variables, visualizing it and studying about the relationship between two variables (sometimes more) is known as exploratory data analysis
- How to summarize and visualize and study the relationship between variables depends on what kind of data that we are working with – categorical or quantitative

LNMIIT

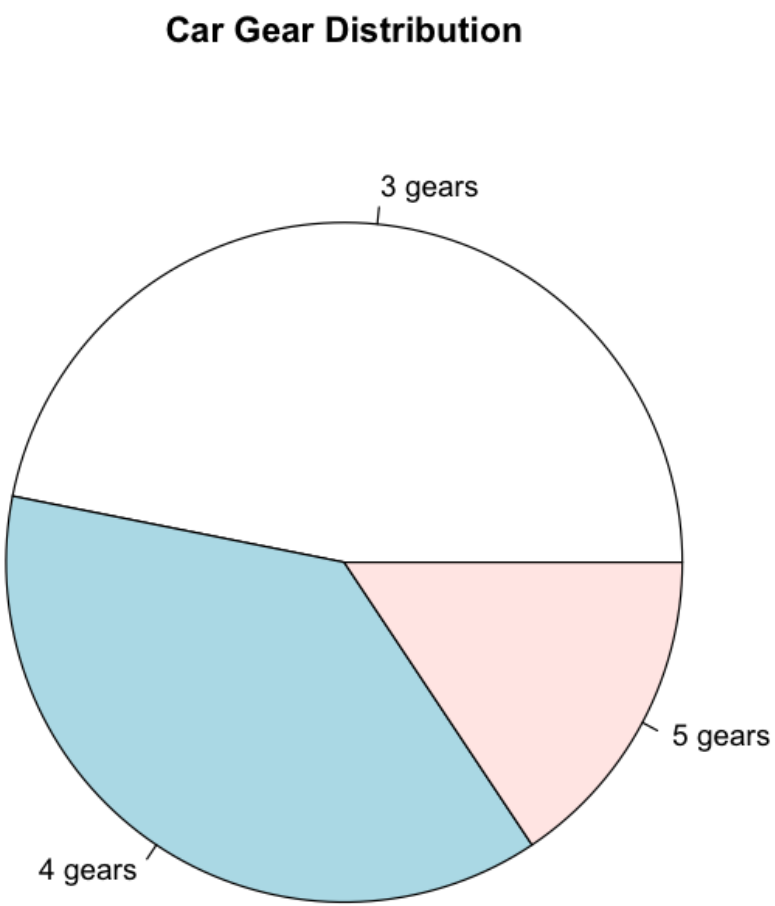
`mtcars` is a sample data set on cars that comes with R. We will use this for our visualization purpose.

```
> counts <-table(mtcars$gear)
> counts
 3  4  5
15 12  5
```

```
> barplot(counts, main="Car Gear Distribution",
names.arg=c("3 Gears", "4 Gears", "5 Gears")
,horiz=TRUE)
```

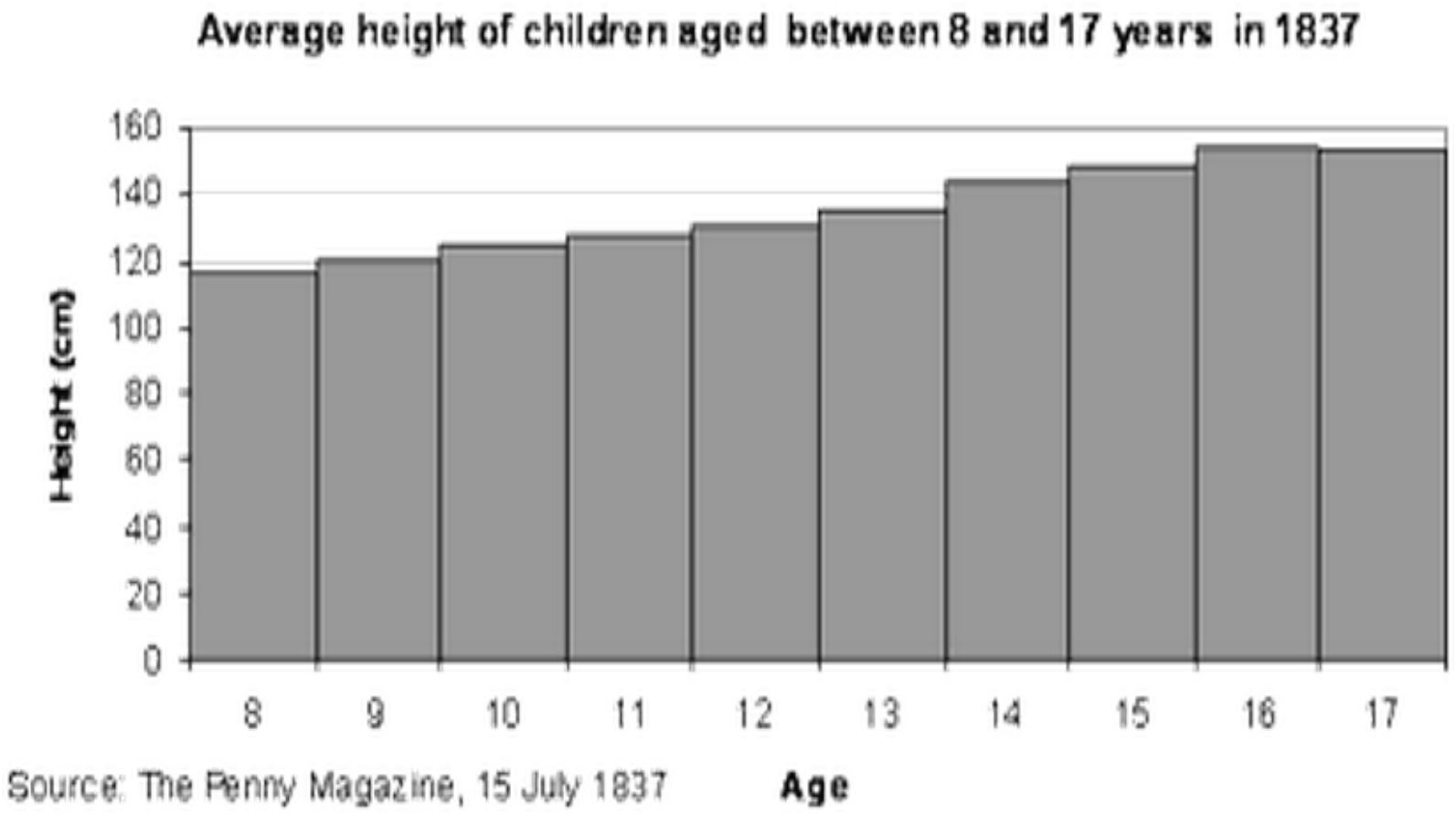


```
> pie(counts, main="Car Gear Distribution",
labels=c("3 gears","4 gears","5 gears"))
```



Histogram

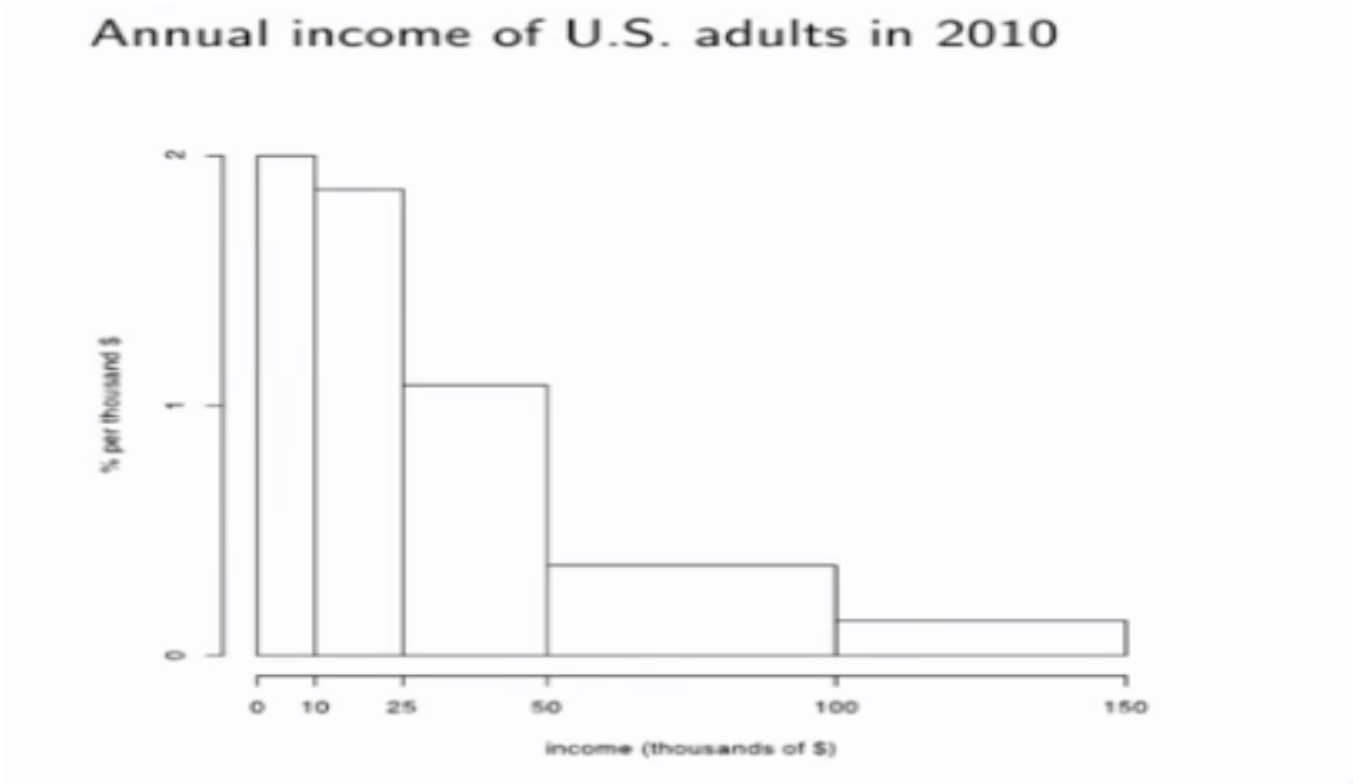
Histograms are special forms of Bar chart. Here the variable is a continuous one rather than a discrete categories



Histogram Vs Bar Chart

- Histogram is for quantitative data – the number of bars depends on the user or to the software
- Bar Chart is for categorical data – the number of bars depends on the number of categories

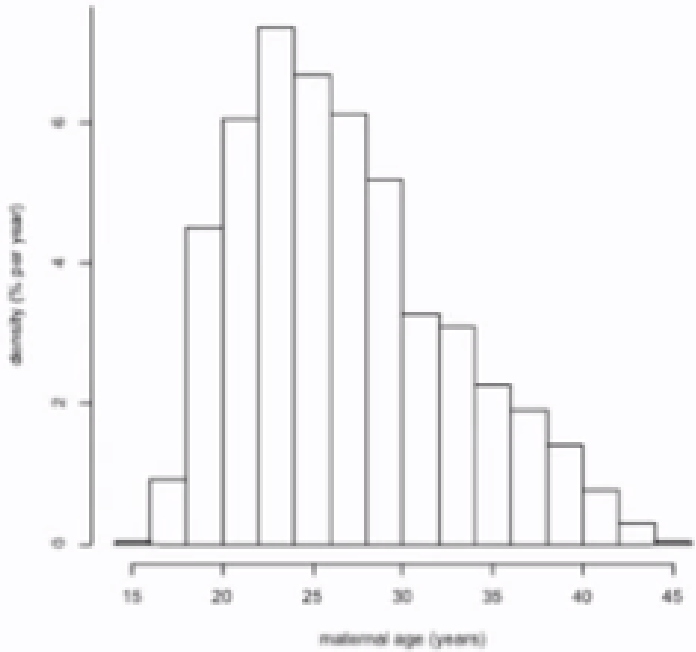
Histogram



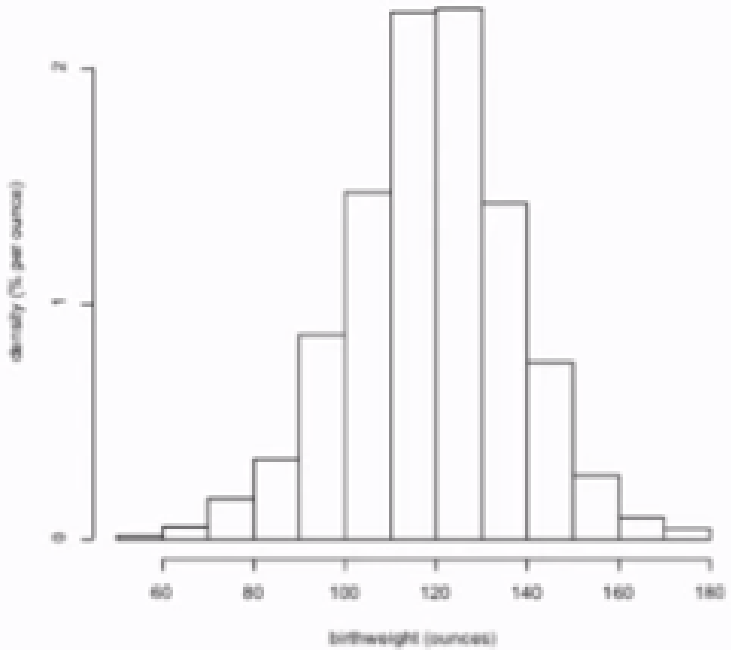
Picture from edx

Histogram

Ages of mothers who gave birth at a local hospital



Birthweights of their babies



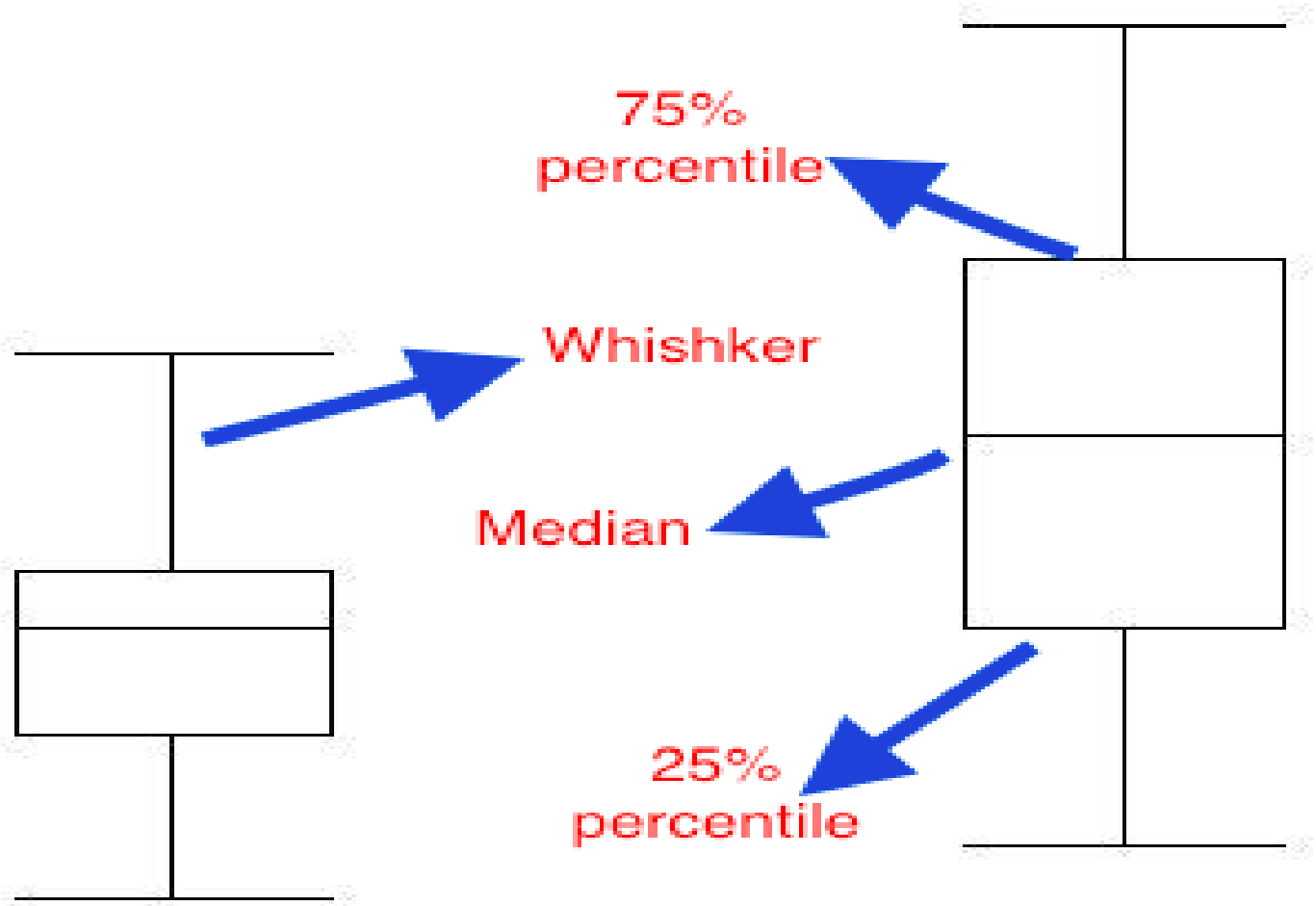
Pictures from edx

LNMIIT

Box Plots

- Box plot are a very nice summarization of data with five important measures:
  - First quartile
  - Median
  - Third quartile
  - Minimum
  - Maximum

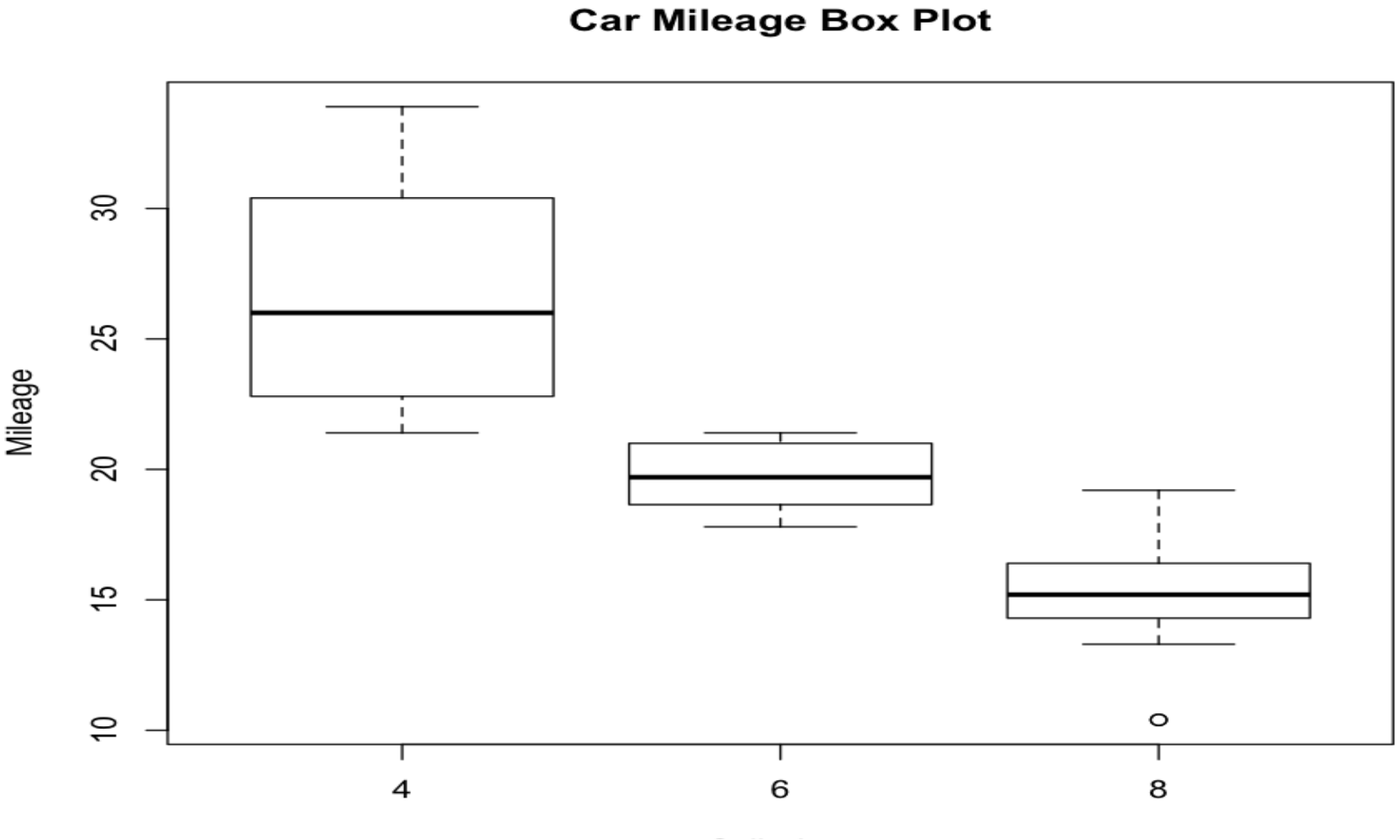
Box Plots



LNMIIT

Box Plots

```
boxplot(mpg ~ cyl,data=mtcars,xlab="Cylinders",
ylab="Mileage",main="Car Mileage Box Plot")
```



Stem and Leaf Plot

Race Running Times in Seconds

Stem	Leaves
12	2 6
13	0 2 5
14	1 2 4 6
15	2 3 7 8
16	1 2 4 6 8
17	5 7 8
18	1 3

Key: 14 | 2 = 14.2 seconds

Marks Scored:

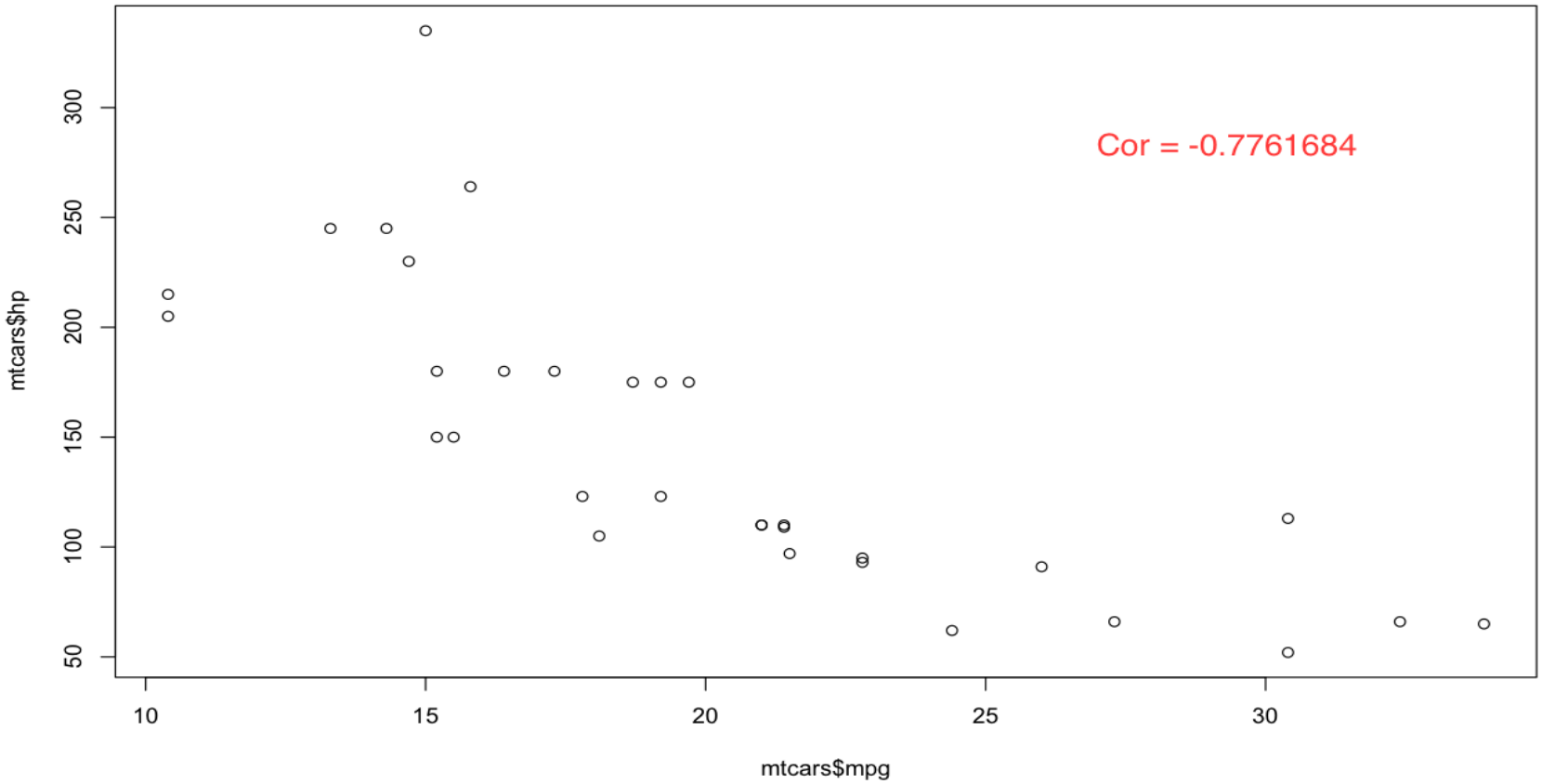
Stem and Leaf Plot

4	1				
5	2	7	8		
6	5	6			
7	0	5	8	8	8
8	0	0			
9	5				

LNMIIT

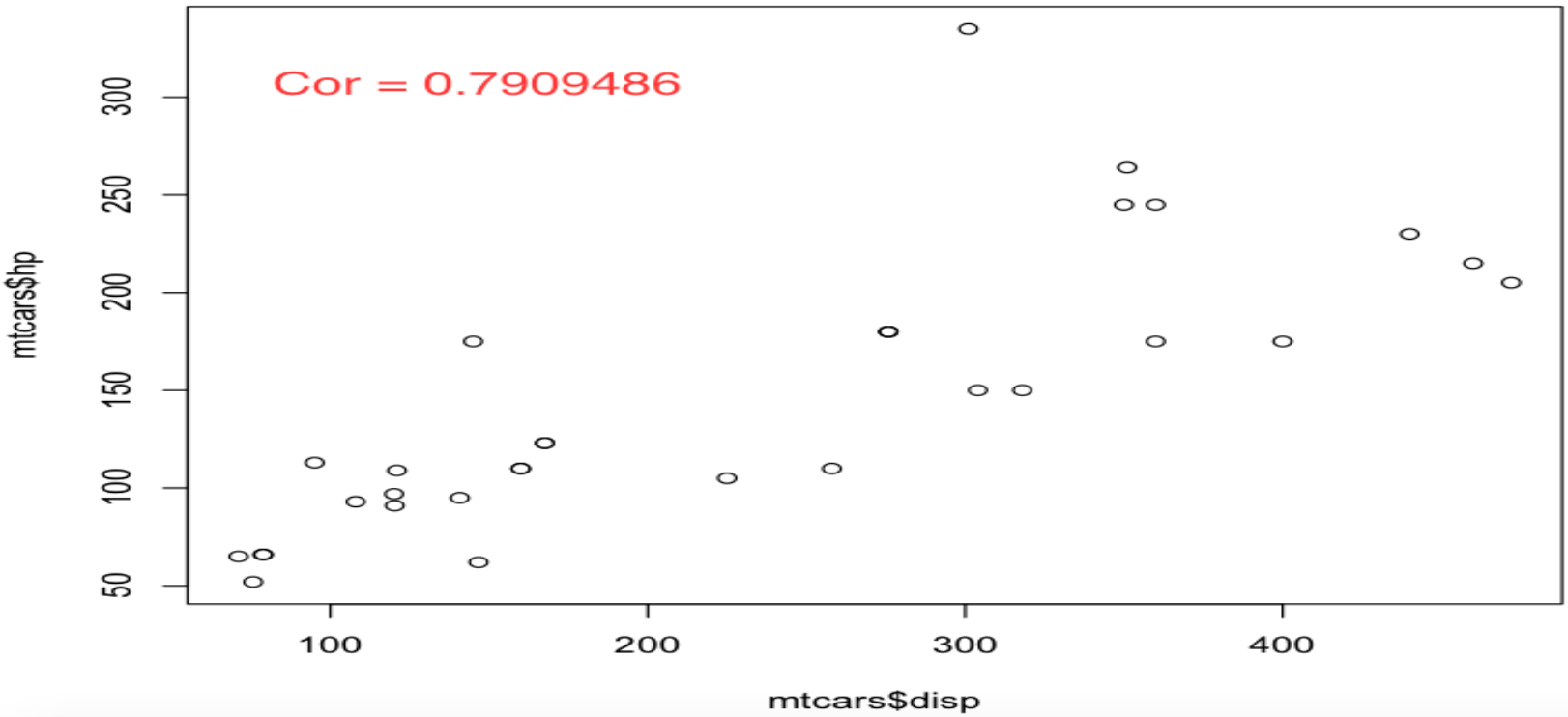
Scatter Diagram

```
> cor(mtcars$mpg,mtcars$hp)
[1] -0.7761684
> plot(mtcars$mpg,mtcars$hp)
```

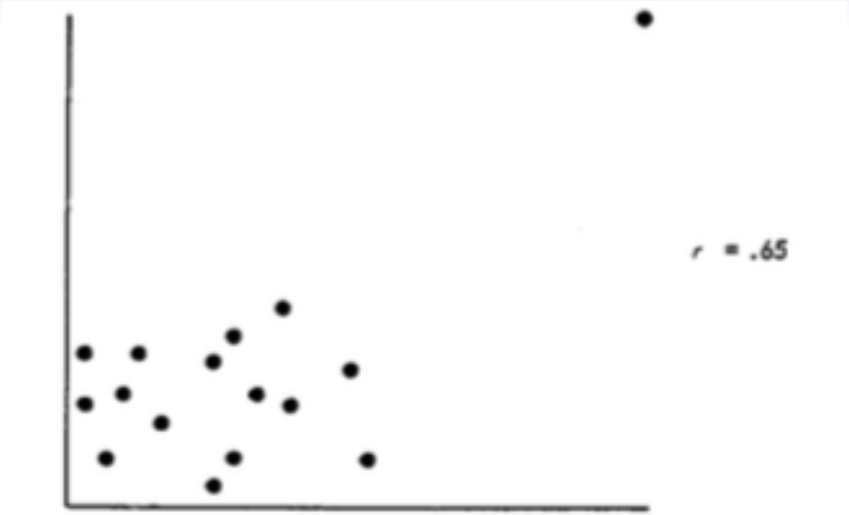


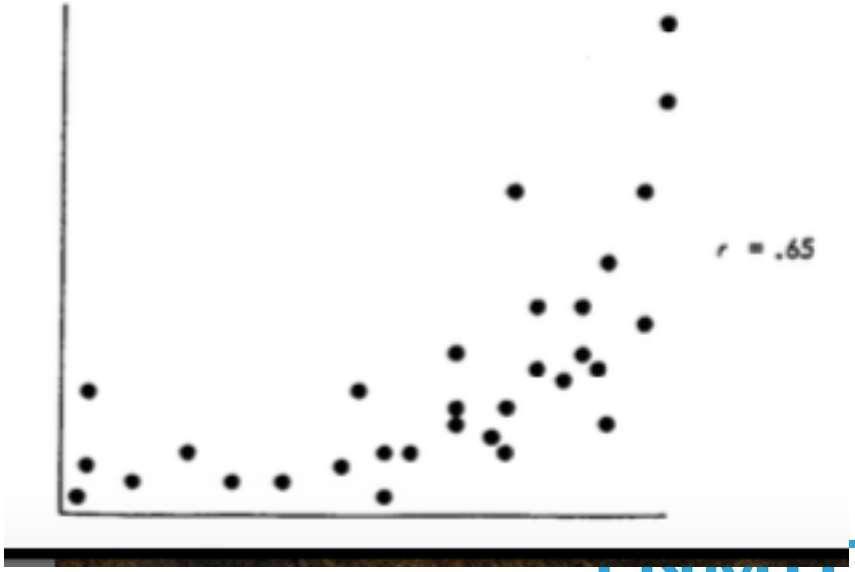
Scatter Diagram

```
> cor(mtcars$disp,mtcars$hp)
[1] 0.7909486
> plot(mtcars$disp,mtcars$hp)
```



Scatter Diagram





Scatter Diagram

- Simple to do and to visualize
- First step to understand the relationship
- This method cannot quantify the strength of correlation

Contingency Table

	j			
i	Democrat	Republican	Independent	Total
Women	68	56	32	156
Men	52	72	20	144
Total	120	128	52	300

Sex \ Handedness	Right handed	Left handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Basic Statistics

1. Mean
2. Median
3. Mode
4. Measures of location
5. Measures of spread

Basic Statistics

Mean	Average of the data – $\frac{\sum_{i=1}^{n} x_i}{n}$
Median	Value that divides the data into two halves
Mode	Most common value that is observed

Measure of Central Tendency

1. A single value to capture the whole picture of the data
2. The center of the distribution of the data
3. Need not be a data from the data set itself
4. Each measure has its own advantages and disadvantages – what we need to take depends on the situation and what kind of data are we dealing with

The Mode

1. The value that occurs the most number of times
2. Histogram – it denotes the value that has the highest bar
3. Not very commonly used like the other two measures but it has its own uses!
4. Best applied to nominal data
5. Applications?

The Mode

Advantages

1. Easy and quick to find out
2. Actual value in the data
3. Extreme scores does not affect it

Disadvantages

1. Sometimes not very informative
2. Can change from sample to sample

LNMIIT

The Median

1. Median is the value that divides the data into two halves where one half is less than the median value and the other half greater
2. Data should be sorted first to find the median
3. Also called as 50^{th} Percentile

The p^{th} percentile of a list of numbers is the smallest number that is at least as large as p% of the list

Examples

Suppose the list of number is $0, 2, 4, 7, 7, 12$ what is the 50^{th} and 25^{th} percentile?

- Lower Quartile – 25%
- Median – 50%
- Upper Quartile – 75%

NMIIT

The Median

1. Median is the value that divides the data into two halves where one half is less than the median value and the other half greater
2. Data should be sorted first to find the median
3. Also called as 50^{th} Percentile

The p^{th} percentile of a list of numbers is the smallest number that is at least as large as $p\%$ of the list

Examples

Suppose the list of number is $0, 2, 4, 7, 7, 12$ what is the 50^{th} and 25^{th} percentile?

- Lower Quartile – 25%
- Median – 50%
- Upper Quartile – 75%

NMIIT

The Median

1. Median is the value that divides the data into two halves where one half is less than the median value and the other half greater
2. Data should be sorted first to find the median
3. Also called as 50^{th} Percentile

The p^{th} percentile of a list of numbers is the smallest number that is at least as large as $p\%$ of the list

Examples

Suppose the list of number is $0, 2, 4, 7, 7, 12$ what is the 50^{th} and 25^{th} percentile?

- Lower Quartile – 25%
- Median – 50%
- Upper Quartile – 75%

The Median

Advantages

1. Relatively easier to find out
2. Actual value in the data (according to which method we follow)
3. Resistant to outliers

Disadvantages

1. Not very informative in some cases
2. Consider the data – $1, 2, 3, 3, 3, 3, 9, 15, 16$. Does the median give a good picture of the data?

LNMIIT

The Mean

1. Most commonly used measure
2. Mean value need not be from the set of given values
3. Also called average – center of gravity!!
4. Considers all values and gives a holistic picture of the set of values with a single measure

The mean is calculated as

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

Examples

Suppose the list of number is $0, 2, 4, 7, 7, 12$ what is the mean?
Suppose the list of number is $0, 2, 4, 7, 7, 100$ what is the mean?

The Mean

1. Most commonly used measure
2. Mean value need not be from the set of given values
3. Also called average – center of gravity!!
4. Considers all values and gives a holistic picture of the set of values with a single measure

The mean is calculated as

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

Examples

Suppose the list of number is $0, 2, 4, 7, 7, 12$ what is the mean?
Suppose the list of number is $0, 2, 4, 7, 7, 100$ what is the mean?

The Mean

1. Most commonly used measure
2. Mean value need not be from the set of given values
3. Also called average – center of gravity!!
4. Considers all values and gives a holistic picture of the set of values with a single measure

The mean is calculated as

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

Examples

Suppose the list of number is $0, 2, 4, 7, 7, 12$ what is the mean?
Suppose the list of number is $0, 2, 4, 7, 7, 100$ what is the mean?

Mean, Median and Mode

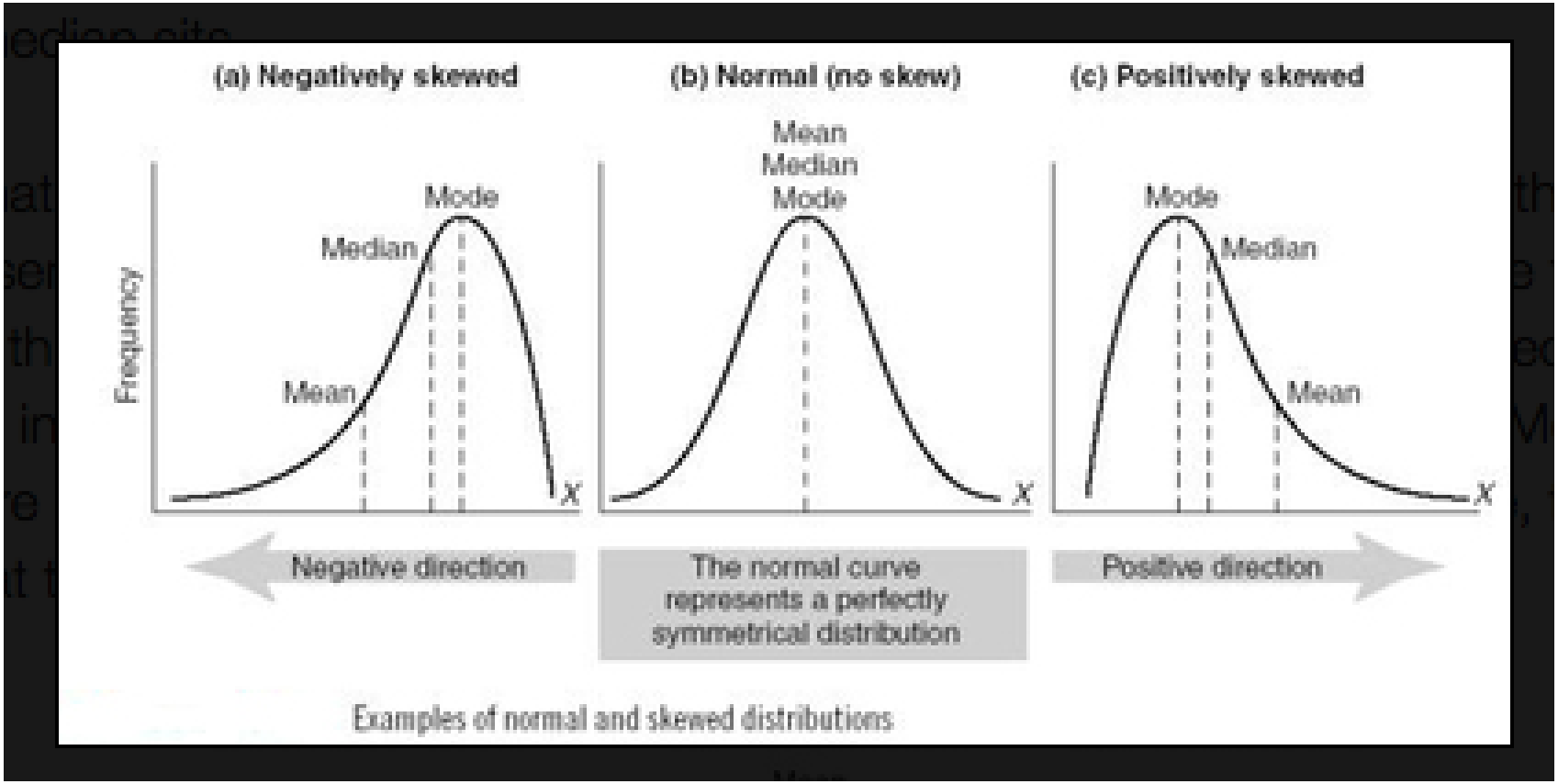
- What is (are) the measure(s) you will take to get the holisitc view of the performance of students in a exam?
- Suppose you write a reference letter for a student for an admission in an university for higher studies. What kind of measure(s) you will use?
- Suppose Bata company wants to cater to the majority of people in the world what kid of measures it will use?
- Suppose your company wants to revise your salary using the salary of the other companies (at the same level as you) as reference. What measure(s) it will use?
- Polling is done in India to select the popular party what measure we use here?

LNMIIT

Mean, Median and Mode

- Can mean, mode and median be same for a given data. Or it will be different for all the data?
- What are skewed distributions?
- Data is cross-sectional or longitudinal?

Mean, Median and Mode



Picture taken from Internet just for learning purpose

Cross-sectional versus Longitudinal

Age	20-30	30-40	40-50	50-60	60+
Ave-height	69.3	69.5	69.4	69.2	68.3

Is it that the people size increases and then decreases with age? **NO!**

- Data is cross-sectional – meaning, across different section of people
- For comparing the data the data has to be longitudinal – meaning, same set of people whose age is measures when they grow up.

LNMIIT

Cross-sectional versus Longitudinal

Age	20-30	30-40	40-50	50-60	60+
Ave-height	69.3	69.5	69.4	69.2	68.3

Is it that the people size increases and then decreases with age? **NO!**

- Data is cross-sectional – meaning, across different section of people
- For comparing the data the data has to be longitudinal – meaning, same set of people whose age is measures when they grow up.

Measures of Location

Question

Suppose the class average score in the exam is 30%. What portion of students scored more than 90%?

Markov's Inequality

If the list consists of only non-negative numbers then the proportion of entries that are at least as large as k times the average is at most $1/k$

Measures of Location

Question

Suppose the class average score in the exam is 30%. What portion of students scored more than 90%?

Markov's Inequality

If the list consists of only non-negative numbers then the proportion of entries that are at least as large as k times the average is at most $1/k$

LNMIIT

Measures of Location

Question

How much of the portion of people are of greater than half of the average?

Answer

It is $\frac{1}{1/2} \implies 200\%$ (Is that correct!?)

Markov's inequality is good but sometimes it gives a very loose bound. Reason is here we use only the mean.

Measures of Location

Question

How much of the portion of people are of greater than half of the average?

Answer

It is $\frac{1}{1/2} \implies$ 200% (Is that correct!?)

Markov's inequality is good but sometimes it gives a very loose bound. Reason is here we use only the mean.

LNMIIT

Measures of Location

Question

How much of the portion of people are of greater than half of the average?

Answer

It is $\frac{1}{1/2} \implies 200\%$ (Is that correct!?)

Markov's inequality is good but sometimes it gives a very loose bound. Reason is here we use only the mean.

LNMIIT

Measures of Spread

Standard Deviation

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

called as root mean square (rms) of deviation from average

Measures of Spread

Chebychev's Inequality

The proportion of entries that are k or more SDs away from the average is at most $1/k^2$

Question

Suppose the class average score in the exam is 30% and the SD is 10. What portion of students scored more than 90%?

Measures of Spread

Problem

Average time is 30 minutes to cross the Howrah Bridge during peak hours. The SD of the time is 15 minutes. What's the largest fraction of the time it could take more than 2 hours to cross the bridge?

Question

Question

Suppose a student gets a mark more than his class average. Does it guarantee that he is in the top-half of the class? Explain

Question

Suppose a student is in the top-half of the class. What can you say about the mean and mode?

Question

Question

Average age = 20, SD = 5. How many are >=80 according to Markov's and Chebechev's inequality?

Calculating Correlation Coefficient

1. Let X be the list containing x_i's and Y contains the list y_i's. Let $|X| = |Y| = n$.
2. Convert both the list to standard units.
3. Multiply corresponding pairs of standard units.
4. Correlation Coeefficient r is the average of the above products.

Formula

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \mu_x)}{\sigma_x} \frac{(y_i - \mu_y)}{\sigma_y}$$

LNMIIT

Calculating Correlation Coefficient

1. Association is not causation
2. r measures linear association
3. Even one outlier might affect r

Changing Units and Z-Score

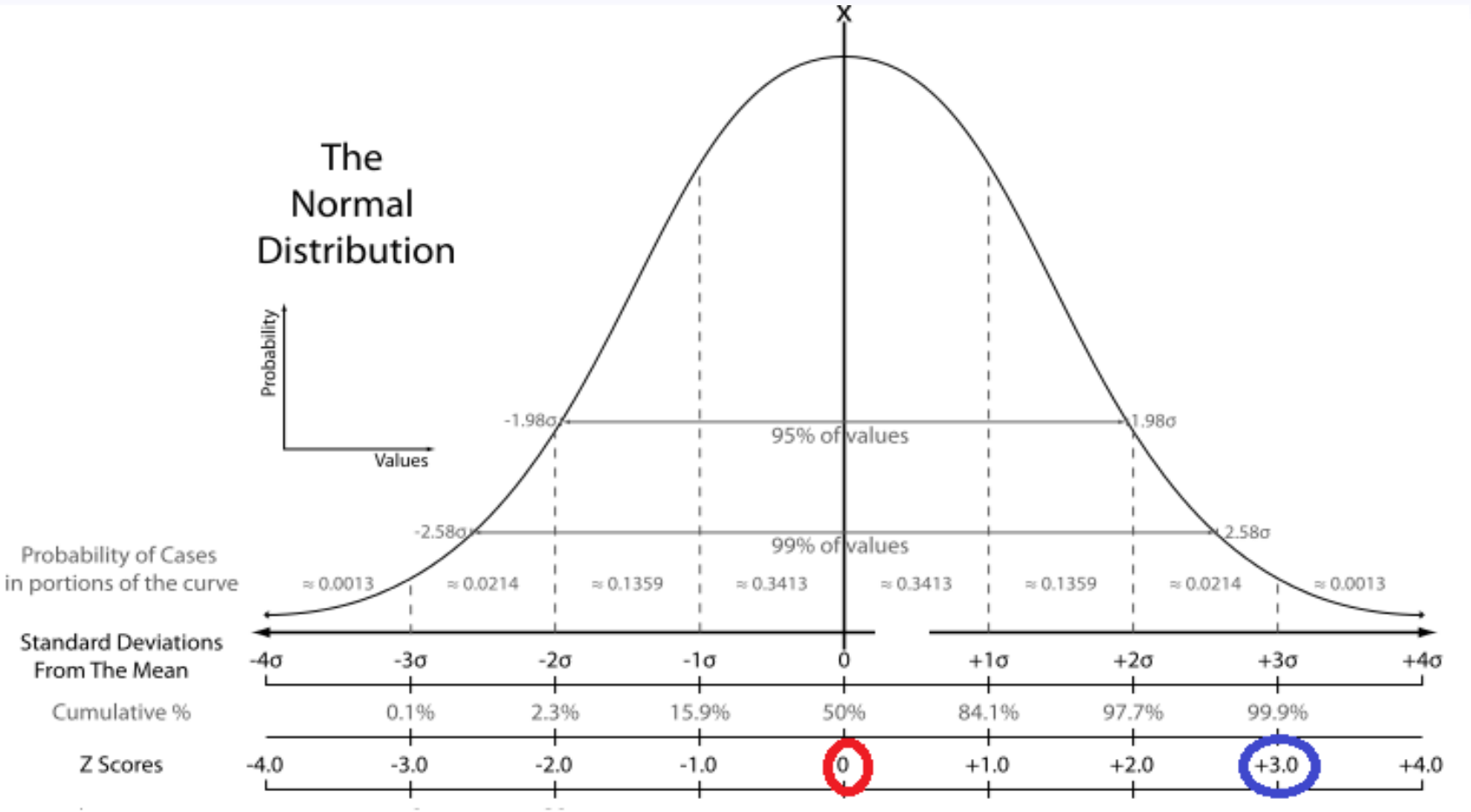
- You have a dataset S and its average m. Suppose you add a constant n to all the data items in S will there be any change in the average for the new data set S'? How about SD of the new data set?
- Same case as above but now you multiply by a constant n. What happens to the average and the SD?

Changing Units and z-Score

- Consider a distribution with mean μ and SD σ
- We define a new unit in terms of σ as follows:
  - Take the mean score as 0
  - Take the value z as $\frac{x-\mu}{\sigma}$
  - Now $z = 1$ if $x = \mu + \sigma$ and $z = -1$ if $x = \mu - \sigma$
  - Likewise for a constant $k > 0$,
    $z = k$ if $x = \mu + k\sigma$, and
    $z = -k$ if $x = \mu - k\sigma$

LNMIIT

Normal Distribution and z-score



Picture from StatisticsHowTo

LNMIIT

z-score

Exam 1

One student scored 28 where the mean was 21 and the SD was 5

Exam 2

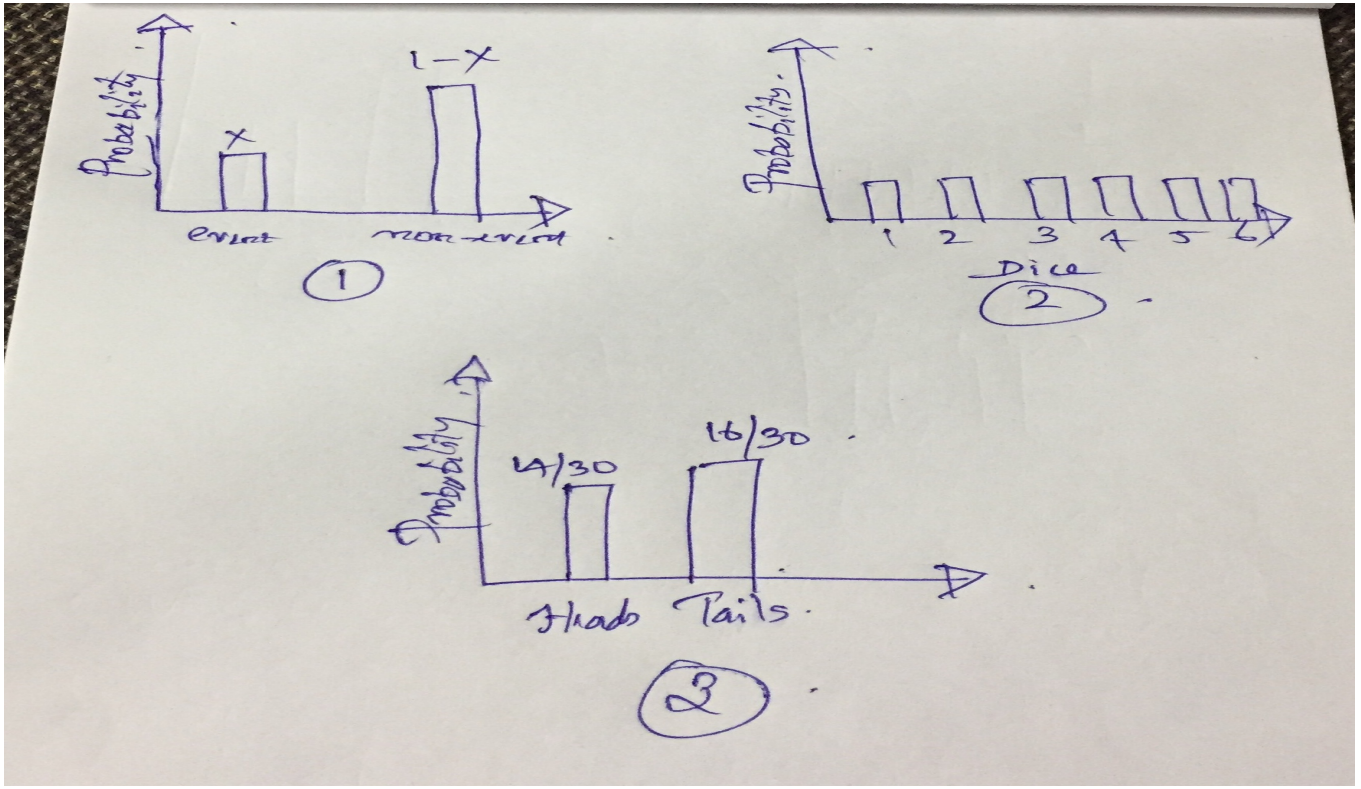
Another student scored 2100 where the mean was 1500 and SD was 325

Which is better?

Random Variables

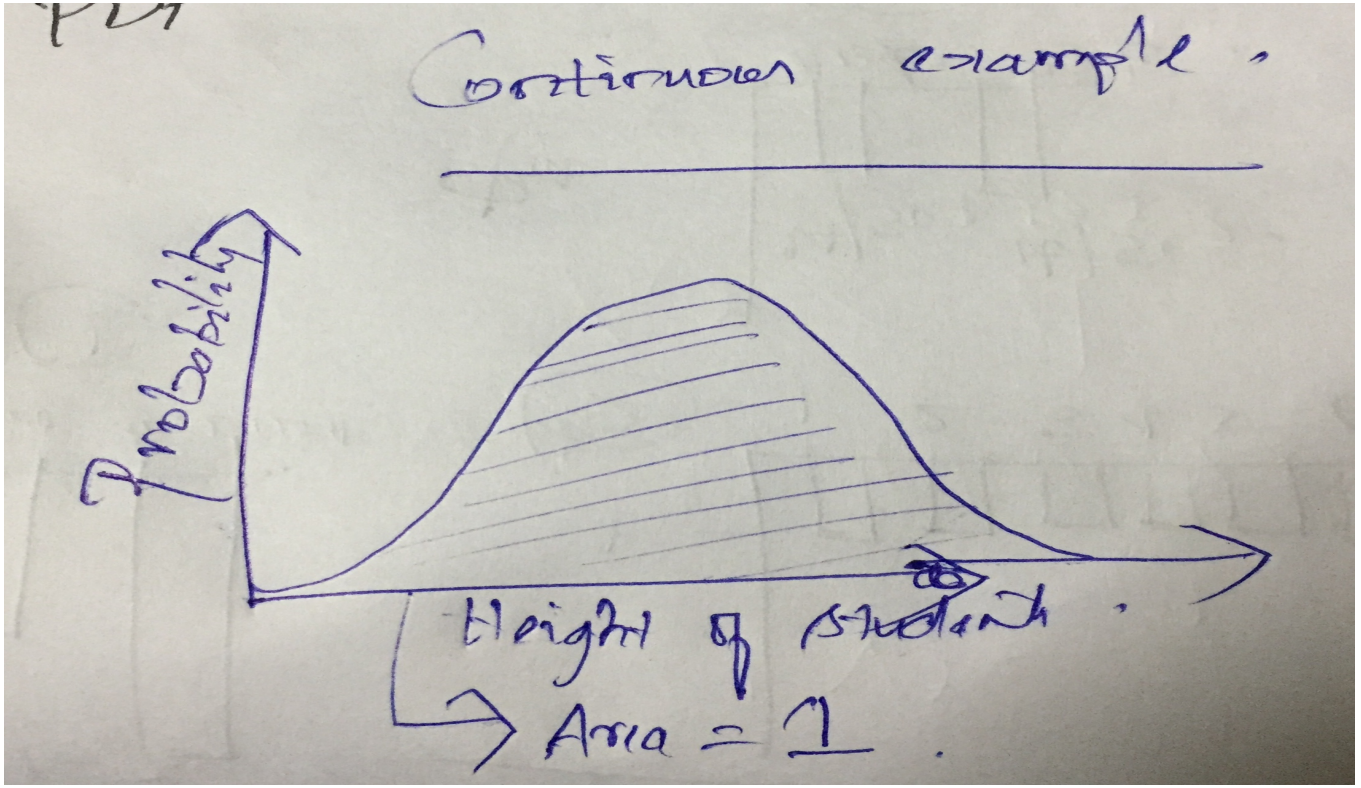
1. The quantity of interest that is determined by the result of an experiment is called as random variables.
2. Outcome of the experiment has a specific probability associated with it. We assign probability to all its possible values.
3. Random variables – discrete or continuous.

Random Variables

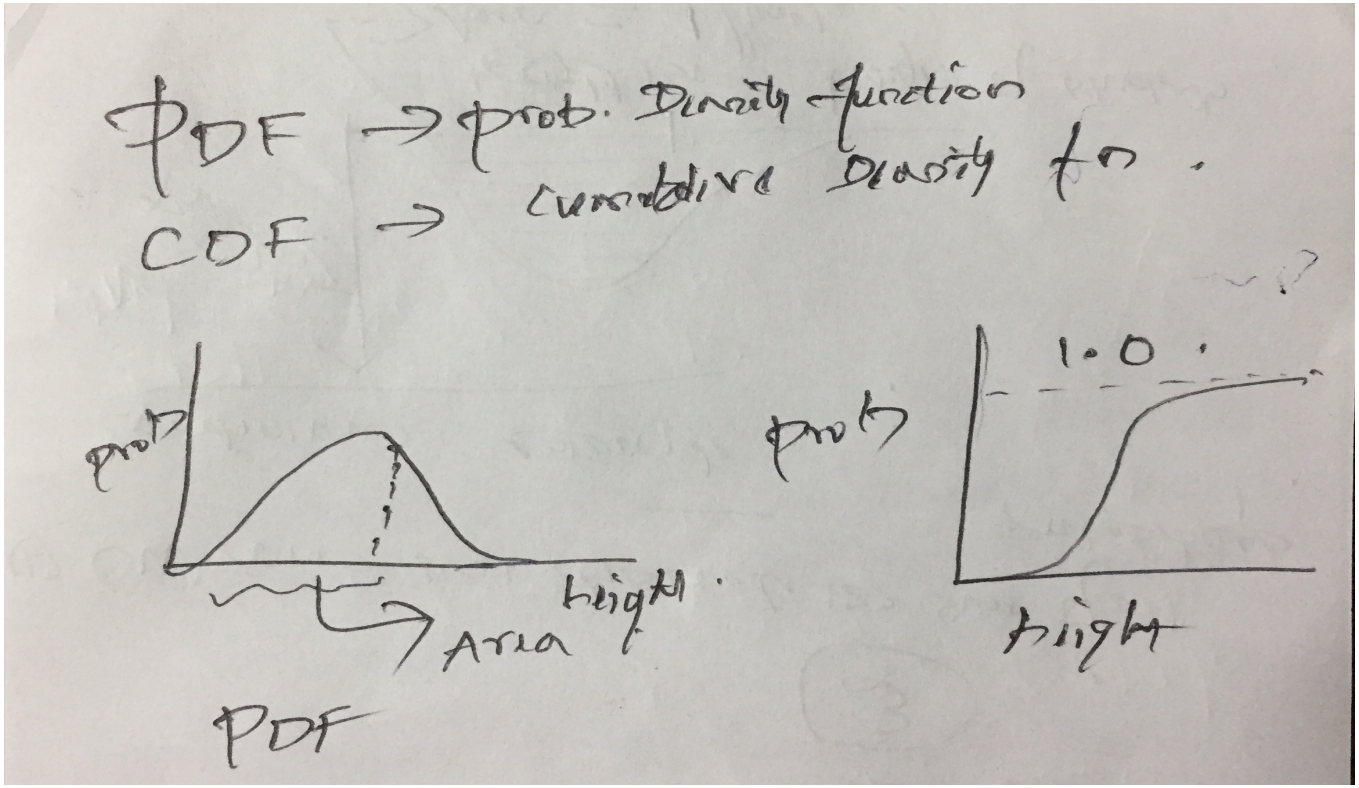


Picture from Intro to Data Analytics Course, NPTEL

Random Variables and Probability Distributions

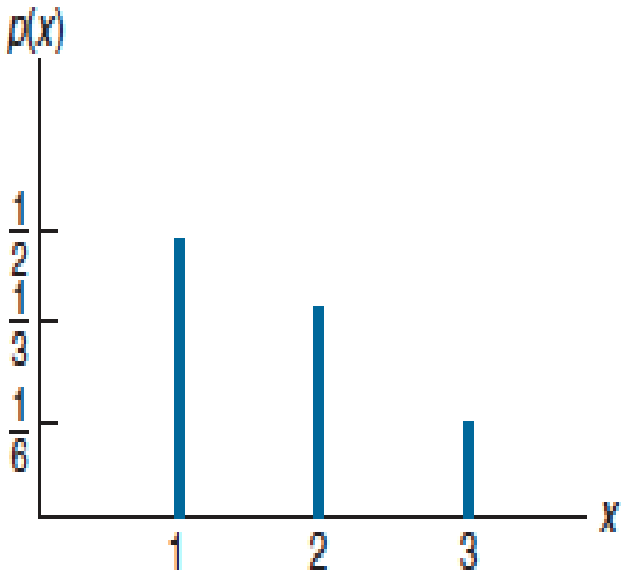


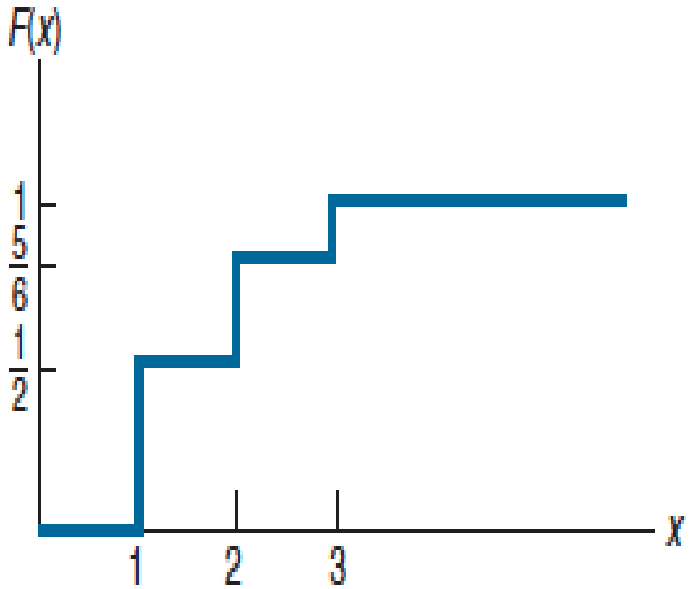
Random Variables and Probability Distributions



Random Variables and Probability Distributions

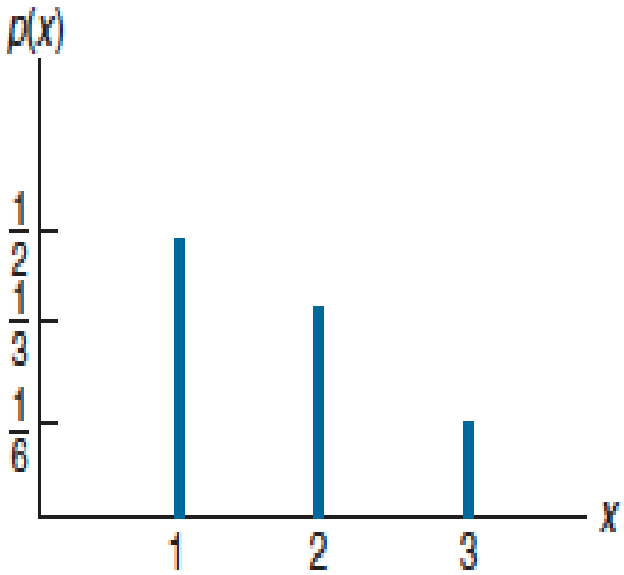
Consider the random variable X where X can take the values $1, 2$ or 3. Let $p(1) = 1/2$, $p(2) = 1/3$ and $p(3) = 1/6$. Draw the graph $p(x)$ (Probability Mass Function) and $F(x)$ (Cummulative Distribution Function).

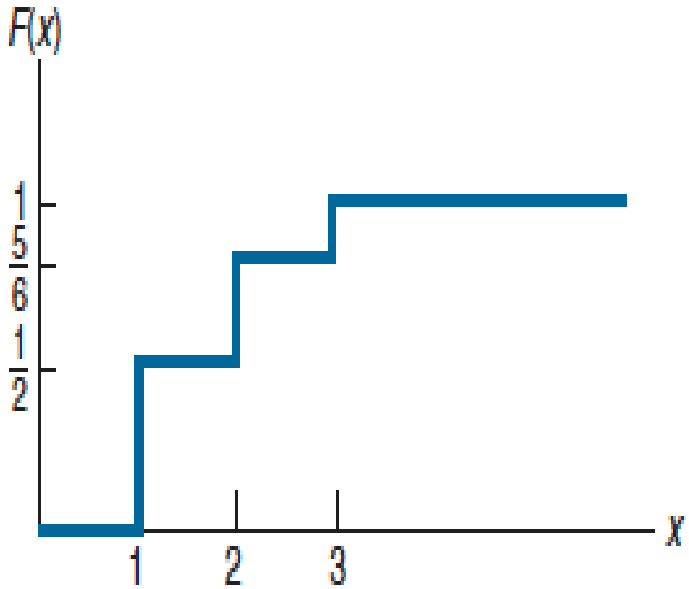




Random Variables and Probability Distributions

Consider the random variable X where X can take the values $1, 2$ or 3. Let $p(1) = 1/2$, $p(2) = 1/3$ and $p(3) = 1/6$. Draw the graph $p(x)$ (Probability Mass Function) and $F(x)$ (Cummulative Distribution Function).





Random Variables and Probability Distributions

Another way of representing CDF:

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{2} & 1 \leq a < 2 \\ \frac{5}{6} & 2 \leq a < 3 \\ 1 & 3 \leq a \end{cases}$$

LNMIIT

Random Variables and Probability Distributions

Continuous Random variable:

EXAMPLE 4.2b Suppose that X is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the value of C?

(b) Find $P\{X > 1\}$.

SOLUTION **(a)** Since f is a probability density function, we must have that $\int_{-\infty}^{\infty} f(x)\,dx = 1$, implying that

$$C \int_0^2 (4x - 2x^2)\,dx = 1$$

or

$$C\left[2x^2 - \frac{2x^3}{3}\right]\Big|_{x=0}^{x=2} = 1$$

or

$$C = \tfrac{3}{8}$$

Probability Distributions

- Uniform
- Bernoulli
- Binomial
- Poisson
- Geometric
- Exponential
- Normal Distribution

Uniform Distribution

- All cases equally likely
- Throwing a dice and finding the value it shows (Discrete)
- Randomly choosing a person and finding his age between 30 to 40 (Continuous)
- Note down the seconds from time when a kid makes noise (Continuous)
- Not easy to find in real life scenario

Formulas

- $f(x) = 1/(b-a)$ for $a \leq x \leq b$ (Probability Mass Function)
- $F(x) = (x-a)/(b-a)$
- Mean = $(b+a)/2$
- Variance = $\frac{1}{12}(b-a)^2$

Bernoulli Distribution

- A discrete distribution having two possible values 0 (success) or 1 (failure)
- Probability of the values are given as $f(1) = p$ and $f(0) = 1 - p$
- If X represents the number of successes that occur in the n trials, then X is said to be a binomial random variable with parameters (n, p)
- Actually if we do this trial morethan one time then this distribution becomes Binomial distribution

LNMIIT

Binomial Distribution

- Discrete distribution
- Example: How many times you will get 5 heads when you toss a coin 10 times?
- Example: Real life scenario – what is the probability of having 3 defective products when you manufacture 100 products?
- Probability Mass Function (PMF): $\binom{n}{k} p^k (1-p)^{n-k}$
- CDF: It is just the sum of PMF
- Ideal of small values of n
- Mean: np and Variance: $np(1-p)$

LNMIIT

Poisson Distribution

- Discrete distribution over a certain period of times or space
- Example: Number of people arriving over the next hour, number of requests got by a processor
- Number of people can be any number!
- PMF: $\frac{\lambda^k}{k!} e^{-\lambda}$
- Mean: λ and Variance: λ

LNMIIT

Geometric Distribution

- Informally this is a counterpart of Binomial – number of attempts before a desired event
- Example: How many times you need to toss a coin before you get the first head
- PMF: $(1-p)^{k-1}p$
- CDF: $1-(1-p)^k$
- Mean: $\frac{1}{p}$ and Variance: $\frac{1-p}{p^2}$

LNMIIT

Exponential Distribution

- Inter-arrival times of Poisson Distribution – counterpart of Poisson
- Continuous distribution – How long should I wait – time which is continuous
- PDF: $\lambda e^{-\lambda x}$
- CDF: $1 - e^{-\lambda x}$
- Mean: $\frac{1}{\lambda}$
- Variance: $\frac{1}{\lambda^2}$

LNMIIT

Formulas

- PDF to CDF: $F(x) = \int_{-\infty}^{x} f(x)dx$
- CDF to PDF: $f(x) = \frac{d}{dx}F(x)$
- Mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Mean: $E[x] = \sum_{i=1}^{\infty} p_i x_i$
- Mean: $E[x] = \int_{-\infty}^{\infty} xf(x)dx$
- SD: $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
- SD: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x - \mu)^2 p_i}$
- SD: $\sqrt{\int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2}$

LNMIIT

Normal Distribution

- Bell shaped curve
- Example: Height, Weight, Marks distributions
- PDF: $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Different Scenarios

1. Annual Income of people?
2. Maternal age of mothers?
3. Birthweight of babies?
4. Distribution of people getting monthly income upto 1,00,00,000
5. Distribution of people getting monthly income upto 90,000

PDF can be written in terms of standard unit z: $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{z^2}{2}}$

Normal Distribution

- Bell shaped curve
- Example: Height, Weight, Marks distributions
- PDF: $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Different Scenarios

1. Annual Income of people?
2. Maternal age of mothers?
3. Birthweight of babies?
4. Distribution of people getting monthly income upto 1,00,00,000
5. Distribution of people getting monthly income upto 90,000

PDF can be written in terms of standard unit z: $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{z^2}{2}}$

Normal Distribution

- $68\% - 95\% - 99.7\%$ rule (use z-table)
- Distribution of heights with mean at 67in and SD 3in. What percentage are between 63 and 67?
- What is the 40th percentile of heights? (see the z-table and do this)

LNMIIT

Normal Distribution

- Montly rents in a neighbourhood:
  Mean: $900
  SD: $600
  Can this be normal?
- In the above case, a good bound helps than finding approximations. Use Chebechev's inequality!
- Good bound is better than a very weak approximation!

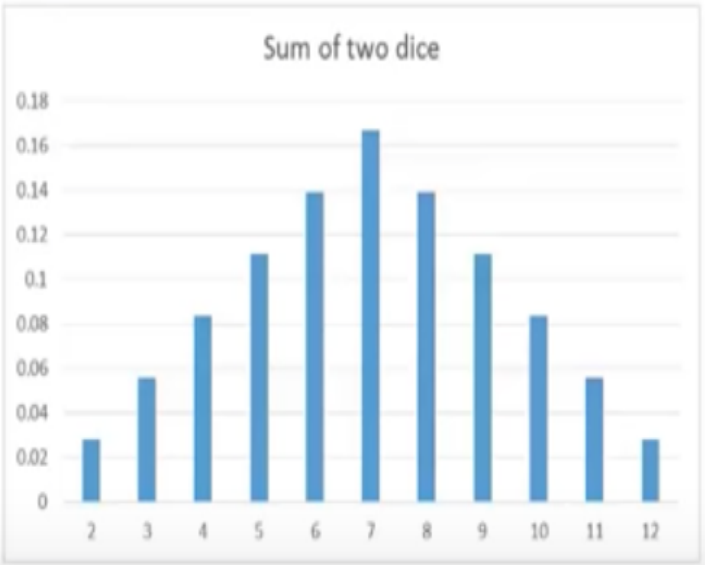
Normal Distribution

- Binomial Approximation with large values of n can be approximated by a Normal distribution.
- Use Mean as np and Variance as $np(1-p)$ and draw the normal curve. This will be a good approximation.
- Main objective is to avoid the intense computations needed when n is very large.
- Getting an approximated value using lean computations is better than getting exact value with intense computations in many scenarios.

LNMIIT

Central Limit Theorem

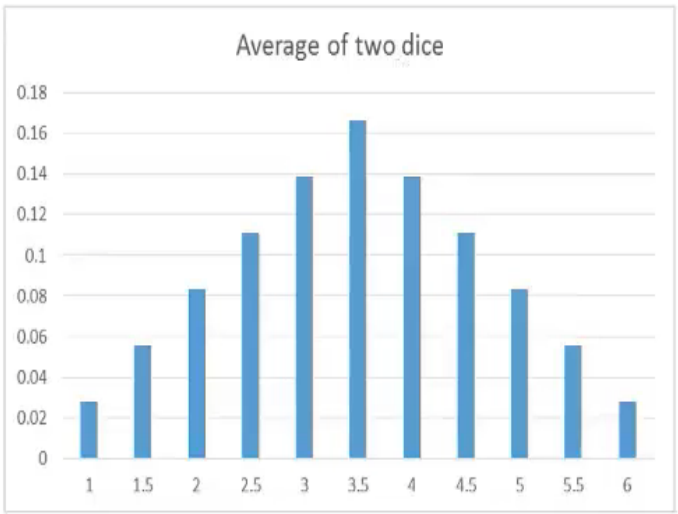
If we aggregate a sufficiently large number of independent random variables then it will lead to a random variable that is approximately normal!

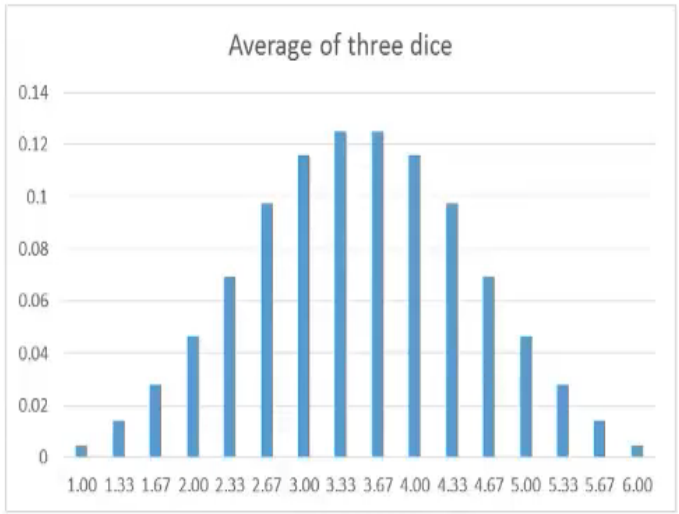


LNMIIT

Central Limit Theorem

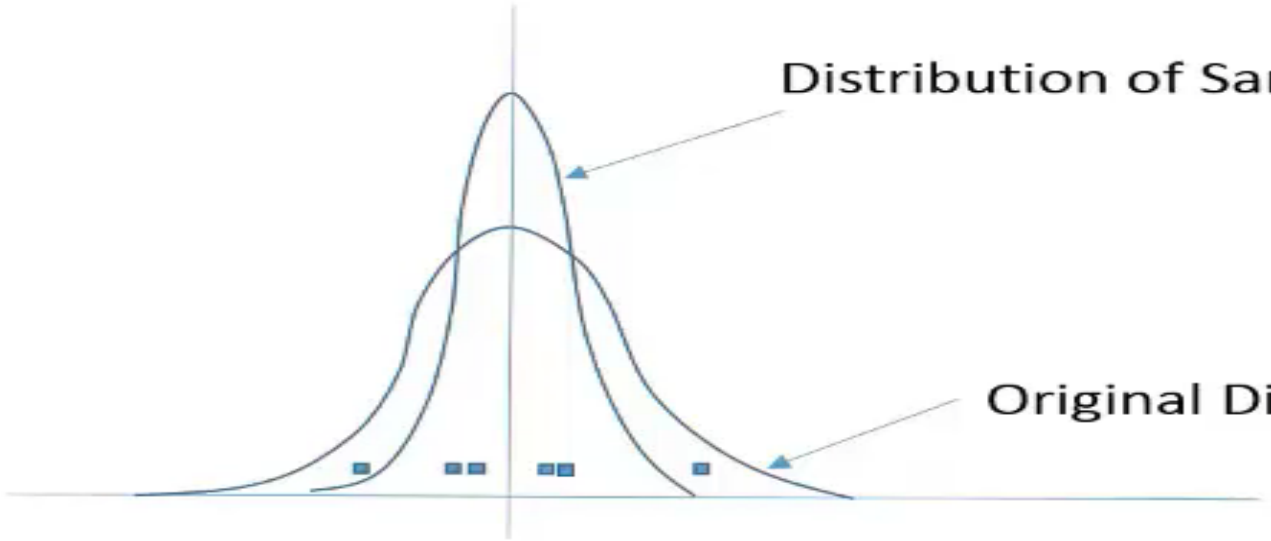
If we aggregate a sufficiently large number of independent random variables then it will lead to a random variable that is approximately normal!

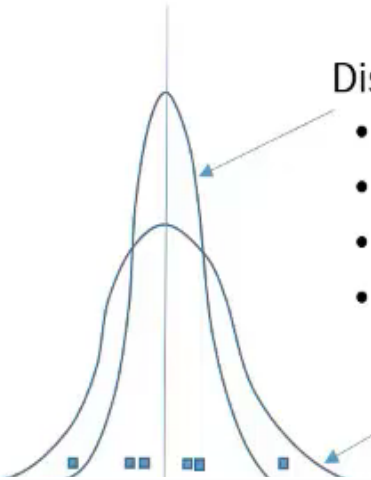




LNMIIT

Sampling





Distribution of Sample means

- What is its shape?
- What is its mean?
- What is its standard deviation?
- Can there be a distribution for sample standard deviations?

Original Distribution