CS 201: Probability and Statistics for Computer Science

Introduction to Statistics

- Statistics is the art of learning from data
- It is concerned with collection of data, its subsequent description and its analysis
- This often leads to the drawing of conclusion

Data Collection

- Prior Availability -
  - govt collects and releases data – such as yearly precipitation totals, earthquake occurrences, unemployment rate, GDP, rate of inflation etc.
  - other organizations also release data such as WHO, World Bank
- Experimental Generation –
  - When the data is not available – it may be generated by designing an appropriate experiment
  - The experiment design depends on the use that the generated data will be put to

Caution in Designing Experiment

- While designing experiment, care should be taken such that none of the results is more likely to come than the other due to the experimental design
- This is essential for drawing valid conclusion from the data
- The accepted way of doing this is to choose values ‘at random’
- At the end of such an experiment, the data is described for example, the mean, standard deviation etc. may be reported

Descriptive v/s Inferential Statistics

- In the previous slide, we discussed that the data generated through experiment is described using various methods
- In fact, any data that we work on is described by using different methods
- We shall learn more about this later
- This part of Statistics, that is concerned with description and summarization of data is known as Descriptive Statistics.

Inferential Statistics and Probability Models

- After we describe and summarize our data the next possible thing to do is to draw conclusions from it
- This part of Statistics that is concerned with drawing conclusions from the data is known as Inferential Statistics
- To be able to draw a conclusion from the data, we must take into account the possibility of chance
- We usually make assumptions regarding the chances of obtaining different values of data. The totality of these assumptions is referred to as the Probability Model of the data.

Populations and Samples

- In statistics, we are interested in obtaining information about a total collection of elements, we call this *population*
- Often, the population is too large for us to examine each of its members
- In such cases, we choose a sub-group from the population and examine each of its elements
- The subgroup is called a *sample* of the population
- If the sample is to be informative about the total population, it must be in some sense, representative of that population
- This is achieved by choosing the members of the sample in a totally random fashion without any prior considerations of the elements that will be chosen

Examples

A university plans on conducting a survey of its recent graduates to determine information on their yearly salaries. It randomly selected 200 recent graduates and sent them questionnaires dealing with their present jobs. Of these 200, however, only 86 were returned. Suppose that the average of the yearly salaries reported was $75,000.

(a) Would the university be correct in thinking that $75,000 was a good approximation to the average salary level of all of its graduates? Explain the reasoning behind your answer.

(b) If your answer to part (a) is no, can you think of any set of conditions relating to the group that returned questionnaires for which it would be a good approximation?

Examples

An article reported that a survey of clothing worn by pedestrians killed at night in traffic accidents revealed that about 80 percent of the victims were wearing dark colored clothing and 20 percent were wearing light-colored clothing. The conclusion drawn in the article was that it is safer to wear light-colored clothing at night.

(a) Is this conclusion justified? Explain.

(b) If your answer to part (a) is no, what other information would be needed before a final conclusion could be drawn?

Course Overview

- **Introduction:** Data Collection and Descriptive Statistics, Inferential Statistics and probability Models, Population and Samples.
- **Descriptive Statistics:** Describing Datasets, Single Point Summarization, Paired Datasets.
- **Probability:** Sample Space and Events, Axioms of Probability, Conditional Probability.
- **Random Variables and Expectations:** Random variables, Jointly Distributed Random variables, Expectation, Variance, Co-variance, Probability Distributions. Parameter Estimation-Maximum Likelihood Estimates; Regression Analysis; Applications, Markov Process, Poisson Process.

Other Course related Information

- Introduction to Probability and Statistics for Engineers and Scientists, Sheldon M. Ross
- A first course in Probability, Ninth Edition, Sheldon M. Ross
- Supplementary material – as documents, videos and web links will be provided from time to time through the google classroom

Descriptive Statistics

Describing Datasets, Single Point Summarization, Paired Datasets

Describing Data Sets

- The data generated by an experiment should be presented to a user in such a manner that they are able to quickly obtain a feel of the data
- It should also be clear and concise
- The most common ways of doing so is to use tables and graphs for data representation
- They help in revealing important features such as the range, the degree of concentration and the symmetry of data

Frequency Tables

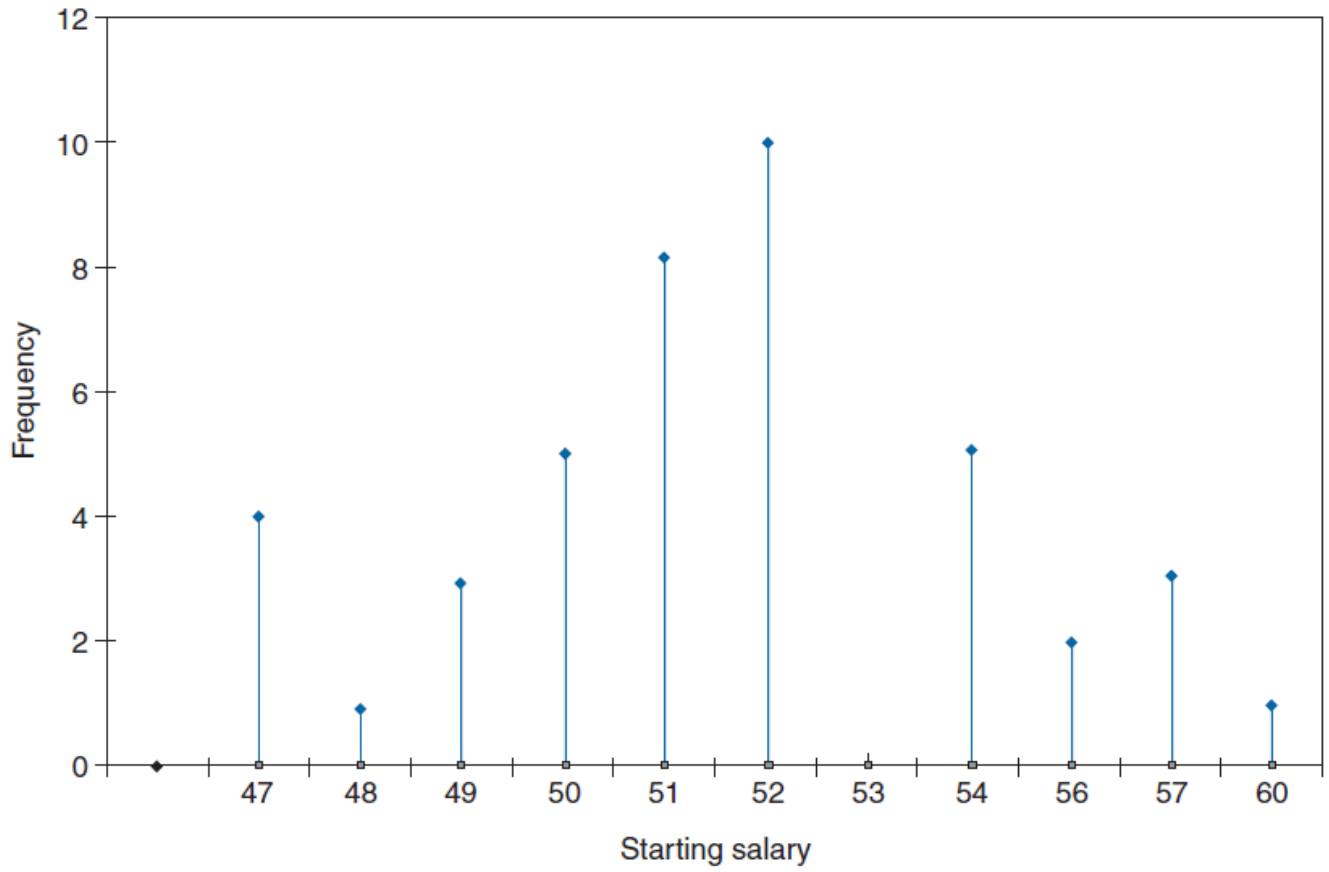
- These are useful presenting data set with relatively small number of distinct values
- Lowest value
- Highest value
- Most common value

TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

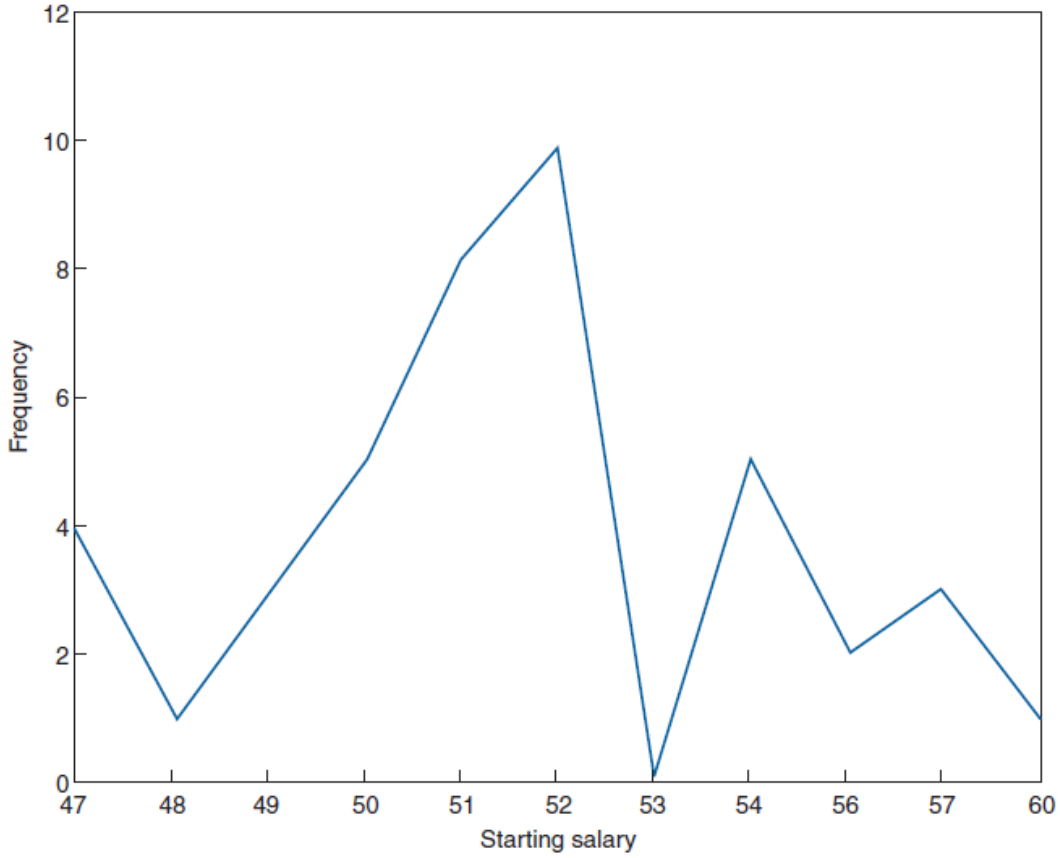
Line Graph

- Data from a frequency table can be represented using a line graph
- The horizontal axis contains the data values
- Vertical axis contains the frequencies
- When instead we use thick lines for plotting, the graph is known as *Bar Graph*



Frequency Polygon

- In frequency polygon, the horizontal and vertical axes are same as in case of line graph or bar graph
- Instead of using lines or bars, the data values are plotted (as points) and the plotted points are then joined together using straight lines



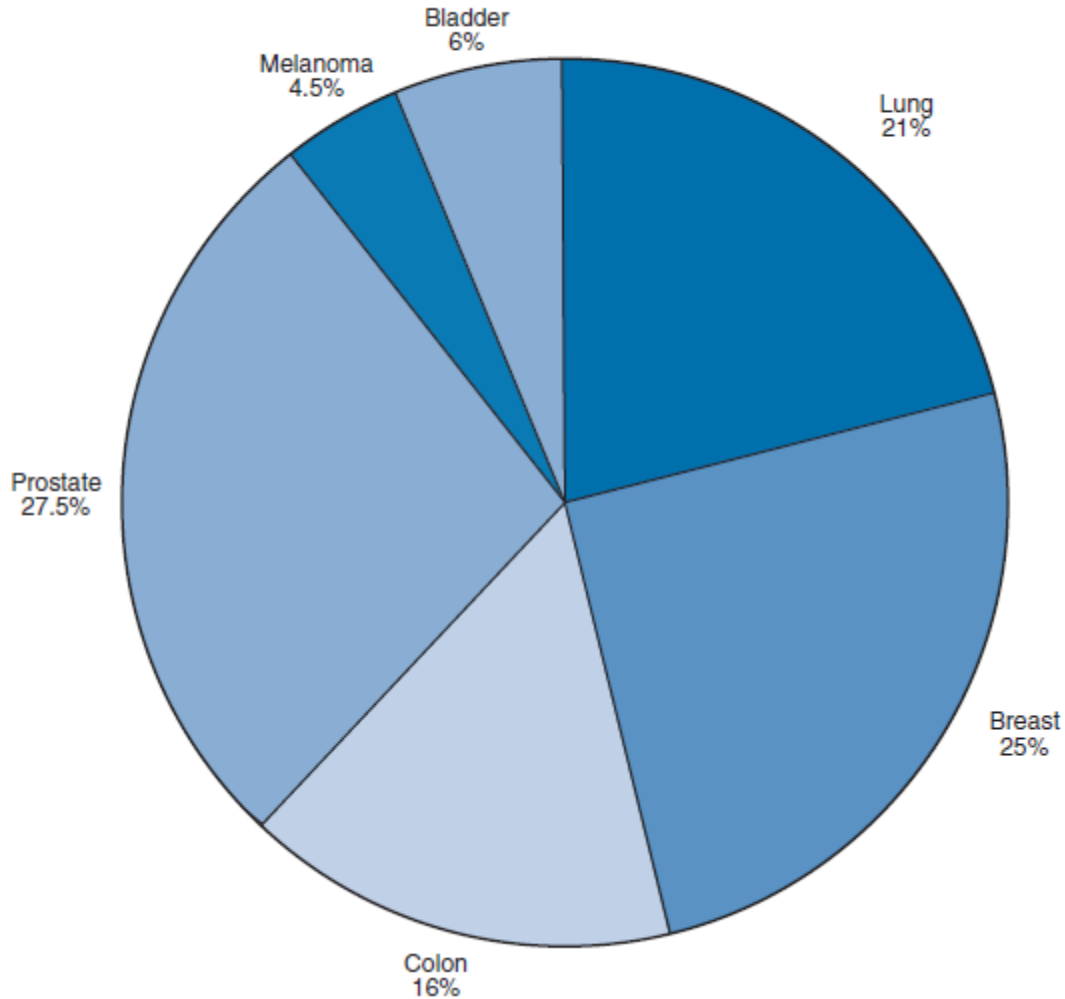
Relative Frequency Table

- Suppose data set consists of n values and f is the frequency of a particular data value
- Then, relative frequency is defined as: f/n
- It is portion of data that have that value
- Relative frequencies can also be used for plotting in the same manner as before
- Line graph, bar graph or relative frequency polygon
- The vertical axis in all such cases will represent relative frequency in place of frequency

Pie Chart

- Used to indicate relative frequencies when the data are not numerical
- For each distinct nature data, a slice is created on a circle
- The area of the slice depends on the relative frequency of the data item
- The area for a particular value of data item is decided by multiplying the relative frequency value with the area of the circle

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06


Bladder
6%
Melanoma
4.5%
Lung
21%
Prostate
27.5%
Breast
25%
Colon
16%

Grouped Data

- When the size of the dataset increases, frequencies of data values also becomes very large and becomes difficult to utilize
- In such cases, the most common approach is to divide the data values into different groups or *class intervals*
- We then plot the data values according to the class interval they fall in
- How to chose class-interval?
  - choosing too few classes might cause losing too much information about the data
  - choosing too many classes will cause frequencies of classes to become smaller and thus making it difficult to find a pattern in the data
- If there is no prior idea about what could be the appropriate distribution of classes, several different number of class intervals may be tried to identify which of them is most revealing of the data

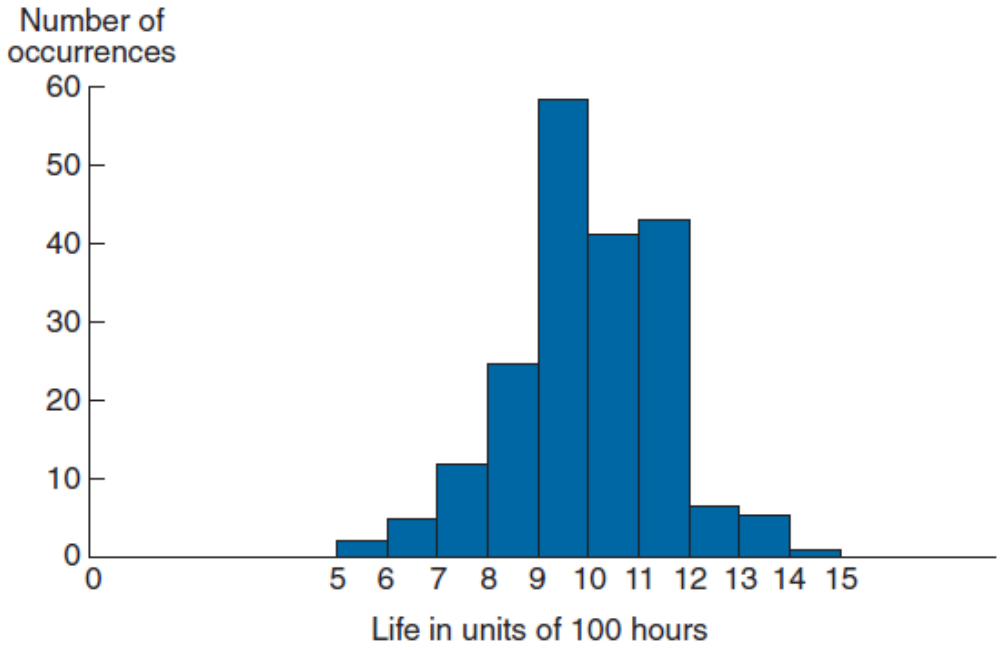
- The class intervals are called *class boundaries*
- Usually, the left-end inclusion convention is followed for grouping data values
- So, a class of values 20-30 will contain all data values that are **greater than or equal to 20** but **less than 30**
- Table shows lifetimes (in hours) of 200 incandescent lamps

TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

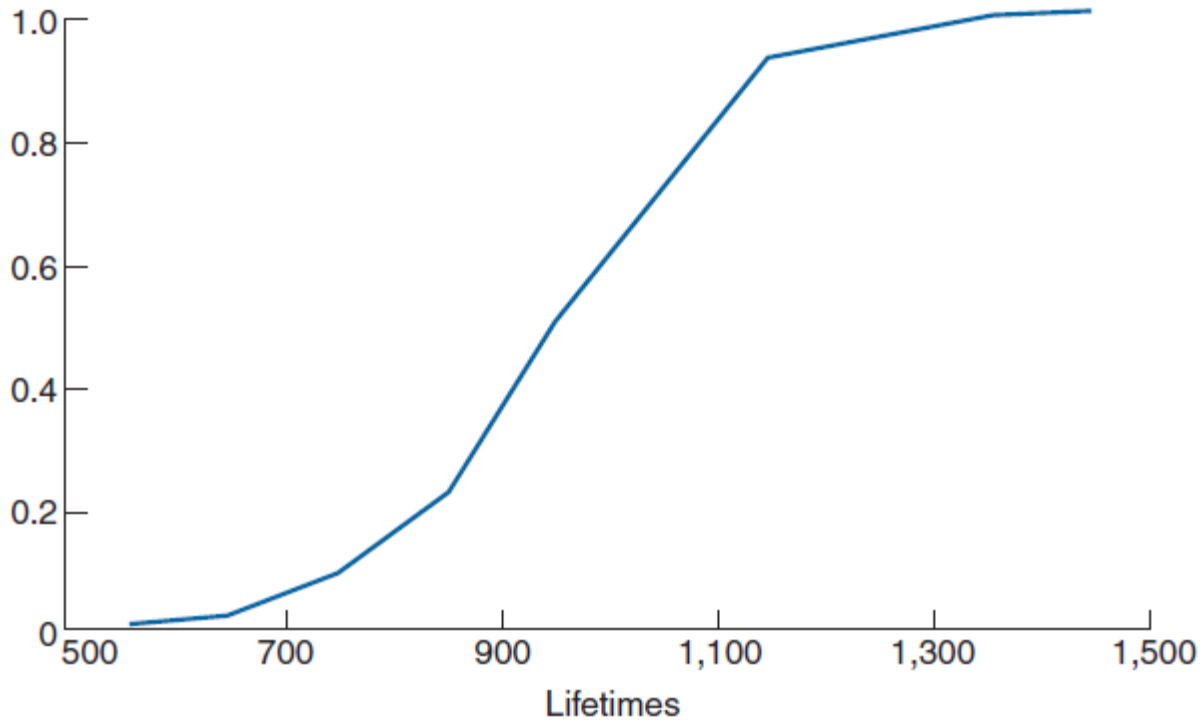
Histogram

- It is a bar graph plot of the class data
- The bars in histogram are placed adjacent to each other
- The horizontal axis contains the class values
- The vertical axis can represent either class frequency or relative class frequency
- Accordingly, we call the histogram as frequency histogram or relative frequency histogram



Cumulative Frequency Plot

- In this plot, the horizontal axis represents data values
- The vertical axis represents the frequency or relative frequency of the data which is less than or equal to it
- What you see in right is a cumulative relative frequency plot
- A cumulative frequency plot is also known as *ogive*



Stem and Leaf Plot

- This type of plot is an excellent way of organizing small to moderate-sized data
- Obtained by first dividing each data value into two parts – stem and leaf
- For example, if data consists of only two digit number, the tens digit may be regarded as the stem and the ones digit as leaf
- Thus, a value of 54 may be represented as:

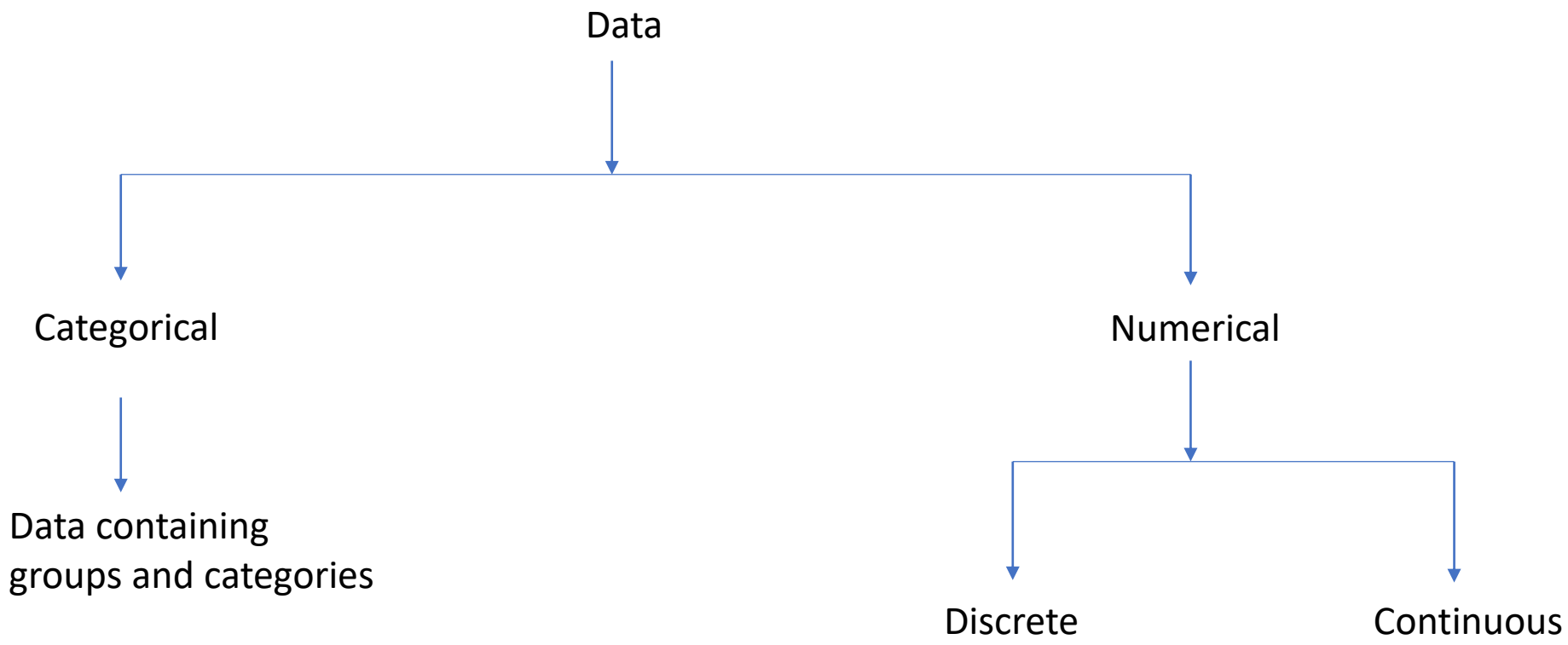
Stem	Leaf
5	4

This can be used for representing data for temperature, precipitation etc.

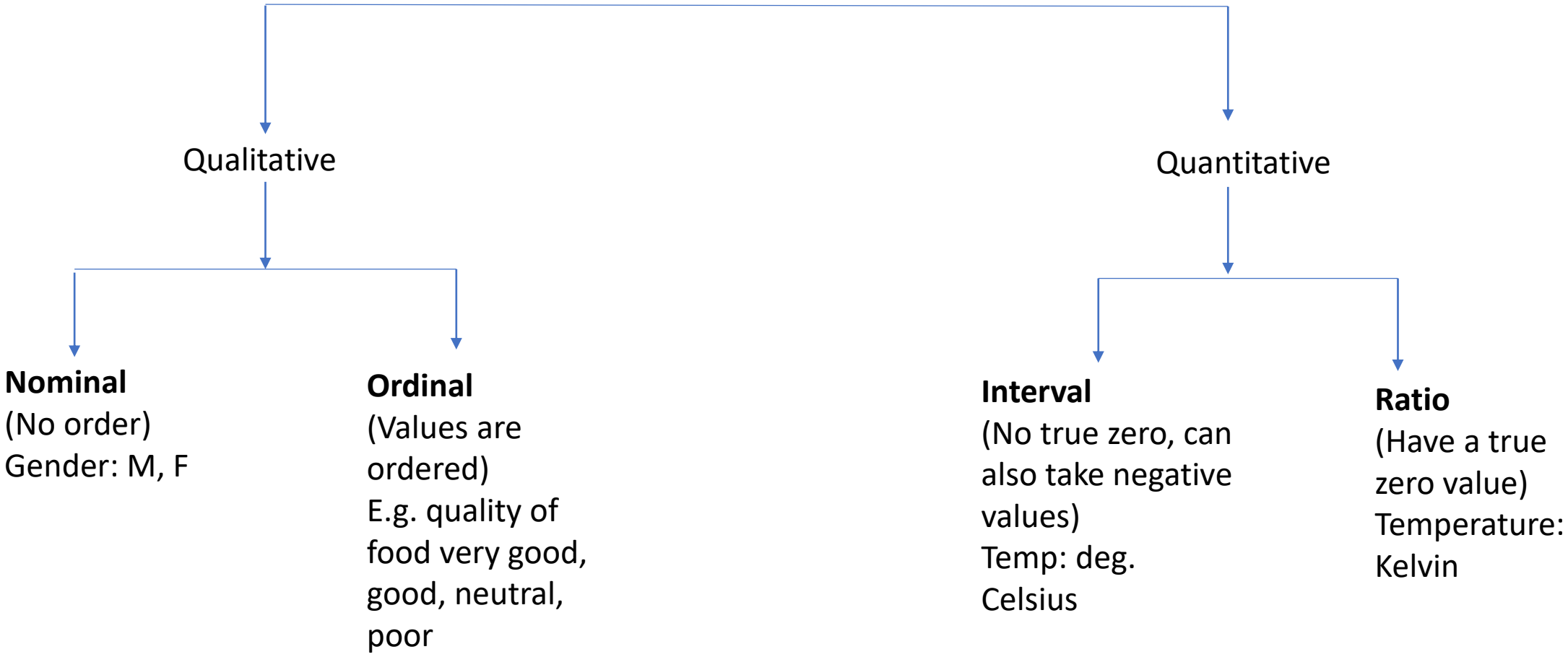
E.g., the annual average daily minimum temperatures of a city may be represented as follows using stem and leaf plot.

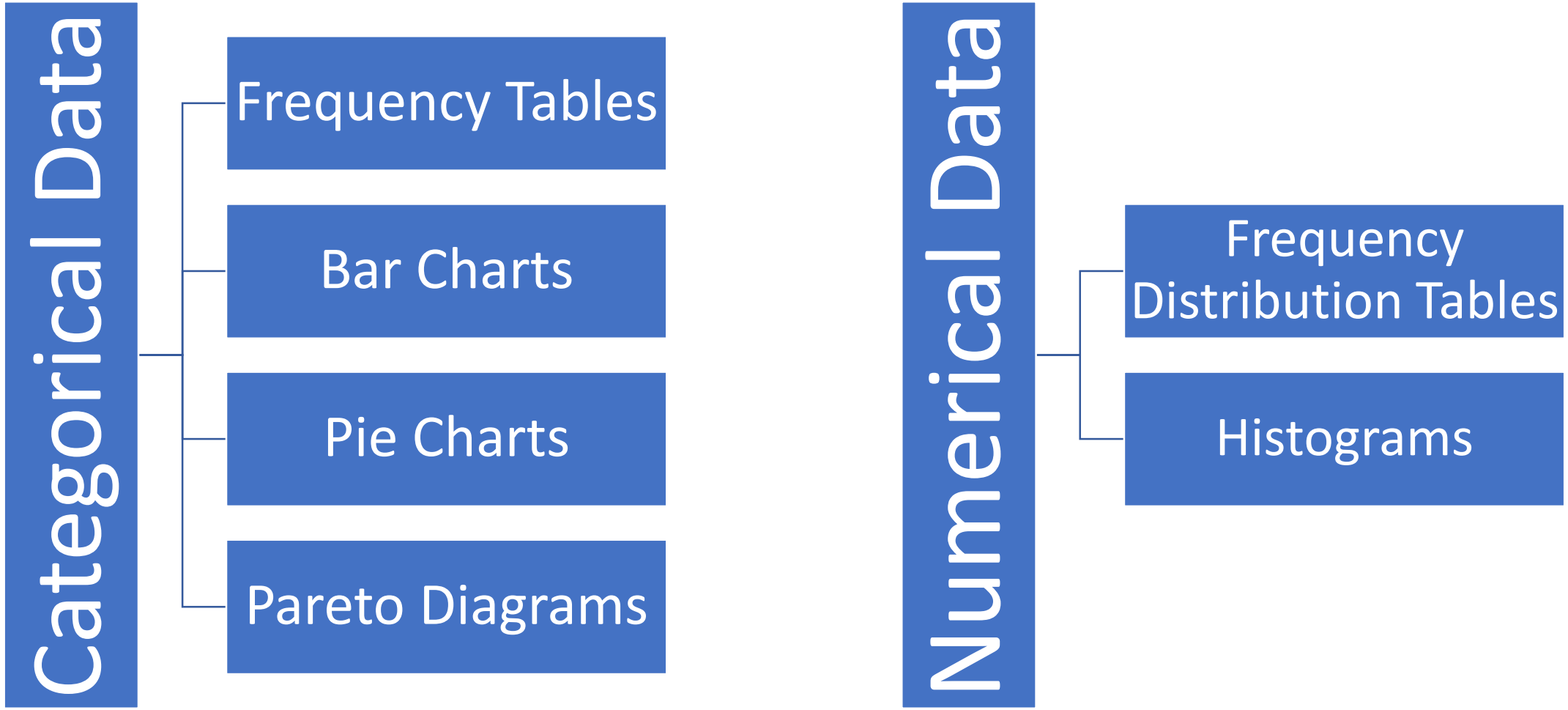
7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2	9.0, 9.8

Types of Data



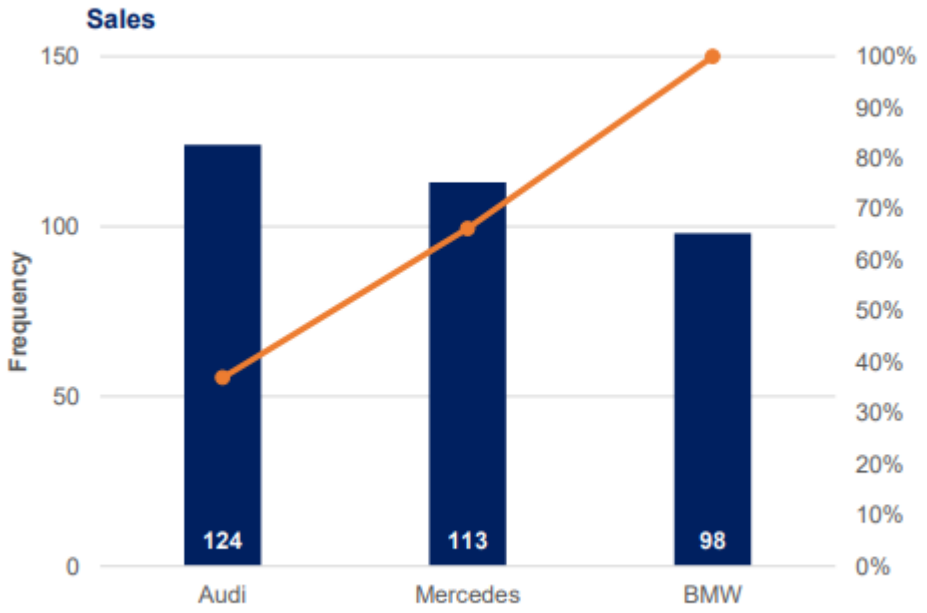
Measurement Scales (Levels)




Categorical Data
Frequency Tables
Bar Charts
Pie Charts
Pareto Diagrams
Numerical Data
Frequency Distribution Tables
Histograms

Pareto Diagrams

- It is a special type of bar chart where categories are shown in descending order of frequency and a separate curve is drawn over it that represents the cumulative frequency



Calculation of Interval Width

- Interval width is calculated as follows:

$$\text{Interval width} = \frac{\text{Highest value} - \text{Smallest value}}{\text{Number of intervals desired}}$$

- Either absolute value or rounded up value may be taken as per need

Describing relationship between variables

- When we want to compare two different variables we use two different types of plots viz.,

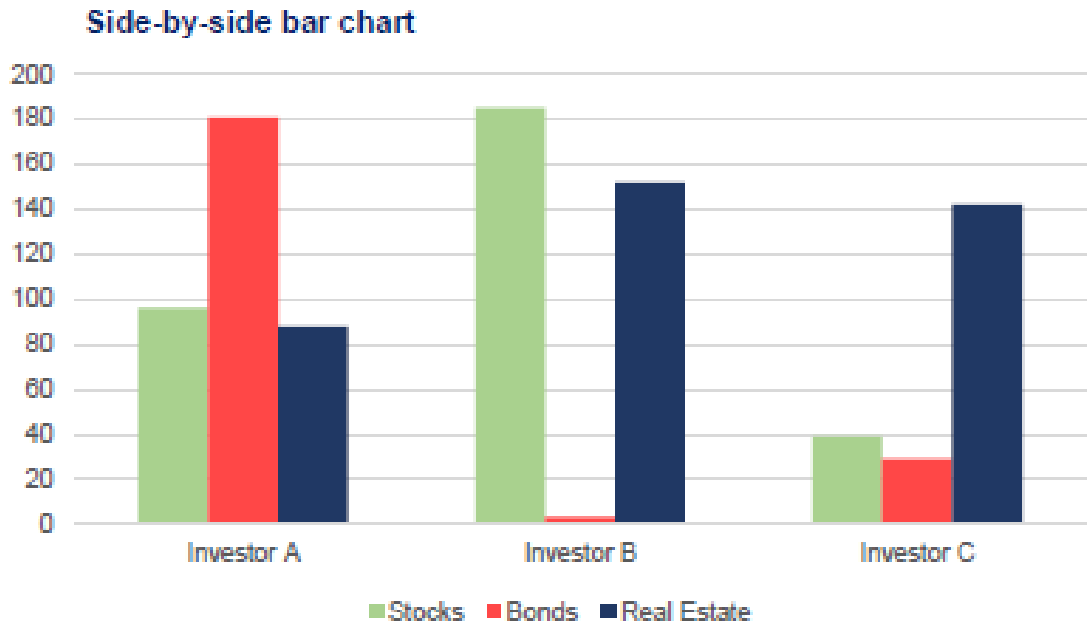
- Cross tables (contingency tables)
- Scatter Plots

Cross Tables & Scatter Plots

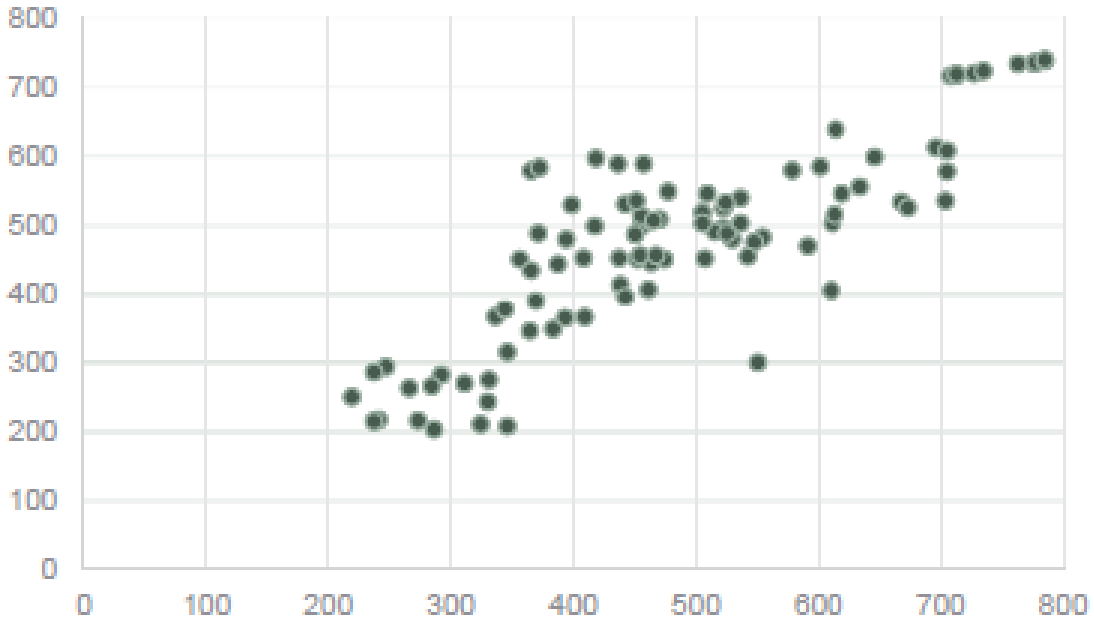
- Cross tables are used to represent two categorical variables.
- One set of categories is represented on the x-axis
- The other set is represented on the y-axis
- The tables may also be constructed using relative frequencies
- These are usually represented using side-by-side bar chart
- For representing two numerical variables, scatter plot is used

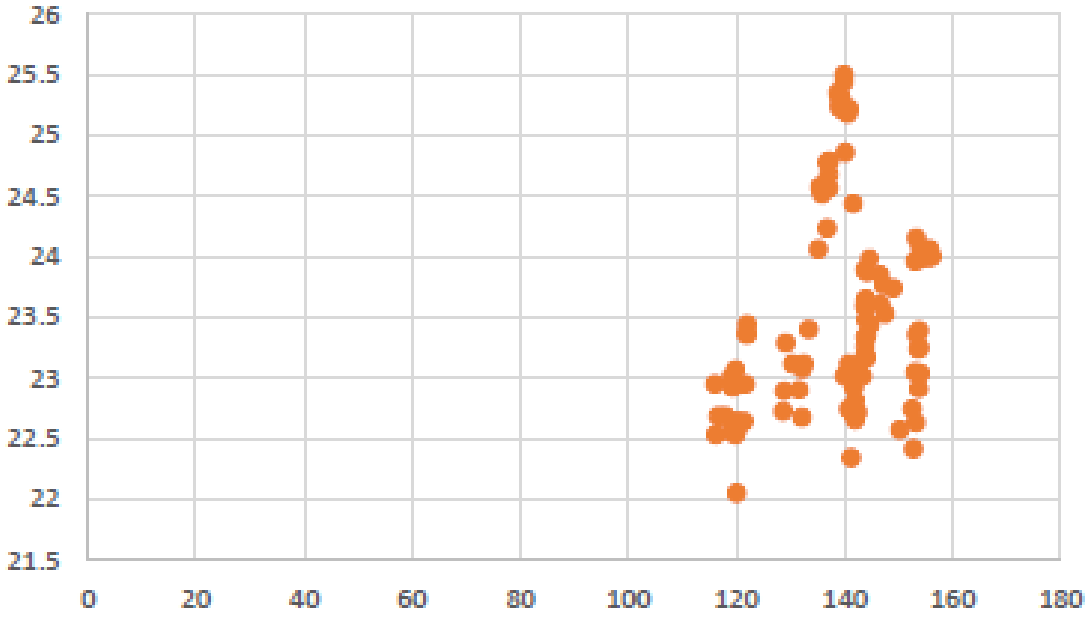
Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
Total	365	340	210	915

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	0.10	0.20	0.04	0.35
Bonds	0.20	0.00	0.03	0.23
Real Estate	0.10	0.17	0.16	0.42
Total	0.40	0.37	0.23	1.00



Scatter Plot





Descriptive Statistics

Data Summarization

Data Summarization

- In order to obtain a feel of very large amount of data, it is important to be able to summarize it in some manner
- There are several techniques that can be used for providing summary of data sets
- The summarization techniques are formally called as *statistics*.
- A statistic by definition is a numerical quantity whose value is determined by the data

Single Point Summarization

- The first kind of summarization statistics that we are going to learn try to describe the center of the data values
- This process is also formally known as measurement of *central tendency*
- There are three types of such statistics and are:
  - Sample Mean
  - Sample Median and
  - Sample Mode

Sample Mean

- Suppose, we have n numerical data values x1, x2… xn.
- The sample mean is defined as the arithmetic average of these data values

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

A simpler way of calculating mean

- Suppose, the data points can be represented using the following relation for some constants a and b

$$y_i = ax_i + b \qquad i = 1,2 \ldots n$$

Then, the sample mean $\bar{y}$ can be represented as:

$$\bar{y} = \sum_{i=1}^{n} \frac{(ax_i + b)}{n}$$

$$= \sum_{i=1}^{n} \frac{ax_i}{n} + \sum_{i=1}^{n} \frac{b}{n} \quad = a\bar{x} + b$$

The winning scores in the U.S. Masters golf tournament in the years for 1982 to 1991 were as follows:

284, 280, 277, 282, 279, 285, 281, 283, 278, 277

Find the sample mean of these scores.

Step 1: Choose any value as constant (b) for subtraction, say 284

Step 2: Subtract 284 from each of the data values

0, -4, -7, -2, -5, 1, -3, -1, -6, -7

Step 3: Calculate the mean of new data values obtained:

-34/10 = -3.4

Step 4: $\bar{x} = \bar{y} + 284 \quad = -3.4 + 284 = 280.6$

Mean using Frequency Table

- When data is presented in the form of frequency table, we can consider mean to be a weighted average of distinct data values
- The weights are the relative frequency values in this case

$$\bar{x} = \sum_{i=1}^{k} v_i f_i / n$$

$$= \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \frac{f_3}{n} v_3 \ldots \frac{f_k}{n} v_k$$

Sample Median

- Sample median can be thought to give middle value of the data items
- It gives the middle element obtained by arranging the data values according to non-decreasing order
- When the exact middle of data is not possible to define (in case of even number of data points), median is calculated by taking the average of n/2th and (n/2+1)th elements

Mean v/s Median

- Both mean and median are useful statistics for summarizing data
- Depending upon the purpose of collecting data one or the other may be useful for describing it
- Mean is calculated using all the data values present in the sample and is therefore affected by extreme values
- Median on the other hand selects the middle value of the data points (at most two middle elements), it is therefore not affected by extreme values if present in the data

Example

- You are given two scenarios, identify in which of them sample mean will a better statistic for summarization and in which case sample median will be better.

a. A state government has a flat rate income tax and is trying to estimate its total revenue from the tax

b. The state govt. is thinking about constructing middle income houses, and wants to determine the proportion of its population able to afford the houses

Sample Mode

- Sample mode is another statistic used to indicate the central tendency of the data values
- It is defined to be the data value that occurs with highest frequency
- It is possible that no single value occurs with highest frequency in the data, in such a case, all the values with highest frequency are reported as *modal values*

Exercise

- The following frequency table is given:

Value	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

calculate the sample mean, sample median and sample mode for this data.

Variance and Standard Deviation

- Apart from describing the central tendency of data, we may be interested in knowing the variability of data
- The statistic used to describe variability of data is known as sample variance or standard deviation
- For describing the variability in data values, we must choose some point about which variability may be described, usually this selected point is the mean of the data
- When the variability of data is described about the mean, the statistic is called sample variance

Definition

- For a given data set x1, x2…xn, the sample variance s^2 is defined as:

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$$

- The division is performed with (n-1) instead of n because the expected value of sample variance and population variance should be same. We shall discuss more about it later.

Example

- Find the sample variance of the following two data sets:

  A: 3, 4, 6, 7, 10 and B: -20, 5, 15, 24

Sample mean A: 6

Sample mean B: 6

Sample variance A: $\frac{((3-6)^2+(4-6)^2+(6-6)^2+(7-6)^2+(10-6)^2)}{4} = 7.5$

Sample variance B: $\frac{((-20-6)^2+(5-6)^2+(15-6)^2+(24-6)^2)}{3} \approx 360.67$

Interpretation

- Thus, two data sets with same mean can actually be pretty different from each other in terms of variability of data points
- You can now understand why variance is an important statistic for describing data
- In the following sections, we shall discuss some important results related to variance

Identity 1

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

$$= \sum_{i=1}^{n} {x_i}^2 - \sum_{i=1}^{n} 2\, x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2$$

$$= \sum_{i=1}^{n} {x_i}^2 - 2n\, \bar{x}^2 + n\bar{x}^2$$

$$= \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$$

Simplifying the calculation of variance

Suppose $y_i = ax_i + b \qquad i = 1,2 \ldots n$

Then, the sample variance can be described as below:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = a^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Proof: $\bar{y} = a\bar{x} + b$

$$y_i - \bar{y} = ax_i + b - (a\bar{x} + b)$$

$$= a(x_i - \bar{x})$$

Properties of Variance

Thus, if ${s_y}^2$ and ${s_x}^2$ are the respective sample variances of x and y then,

$$ {s_y}^2 = a^2 {s_x}^2 $$

- Adding a constant to each data item does not change its variance
- However, multiplying a constant value to each data item changes its variance

Example

The following data give the worldwide number of fatal airline accidents of commercially scheduled air transports in the years from 1985 to 1993.

Year	1985	1986	1987	1988	1989	1990	1991	1992	1993
Accidents	22	22	26	28	27	25	30	29	24

Find the sample variance of the number of accidents in these years

Step 1: Subtract (say 22) from each data value

0,0,4,6,5,3,8,7,2

Step 2: $\bar{y} = \frac{(0+0+4+6+5+3+8+7+2)}{9} = 35/9$

Step 3: Calculate $\sum_{i=1}^{n} {y_i}^2 = (4^2 + 6^2 + 5^2 + 3^2 + 8^2 + 7^2 + 2^2) = 203$

Step 3: Calculate standard deviation for y values

$$= \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2$$

$$= 203 - 9(35/9)^2 \approx 8.361$$

Standard Deviation

- Sample Standard deviation is defined as the positive square root of the sample variance
- Unlike sample variance, the sample standard deviation has a unit
- The unit of measurement of standard deviation is same as units of measurement of data
- The sample standard deviation ‘s’ is denoted as:

$$s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \Big/ n - 1}$$

Percentiles

- For some value of $0 <= p <= 1$, the 100p percentile of data values can be defined as that value such that 100p percent of the data values are less than or equal to it

- Thus, for $p = 0.8$, the 80^{th} percentile will be the value such that 80% of the data values are less than or equal to it.

Definition

- The sample 100p percentile of the data values is the value such that 100p data values are less than or equal to it
- 100(1-p) values are greater than or equal to it.
- If two data values satisfy this condition then the sample 100p percentile is the arithmetic average of these two data values

Percentile Calculation

- Suppose the sample consists of n data values
- To calculate the 100p percentile, the data values are first arranged in increasing order
- Next, we need to find out a data values such that np values are greater than or equal to it and n(1-p) values are smaller than or equal to this
- It turns out, that if np is an integer then the (np)th and (np +1)th data values in the sorted order satisfy the previous condition, we take the average of these two values and report it as the quartile
- On the other hand, if np is not an integer then, it can be easily checked that the next higher data value satisfies the previous condition

Example

22 22 26 28 27 25 30 29 24 13

For the above data calculate the sample 80 percentile.

Here, n = 10
Ascending order sequence of data: 13, 22, 22, 24, 25, 26, 27, 28, 29, 30

80^{th} percentile position = 0.8 * 10 = 8
Thus, 80^{th} percentile will be 8^{th} data value in ascending order sequence of data
= (28+29)/2

Quartiles

- Quartiles are specific percentile values
- The sample 25 percentile is called the first quartile
- The sample 50 percentile is called the second quartile (this gives median value of data)
- The sample 75 percentile is called the third quartile
- Of course, the fourth quartile defined in this manner will be the highest data value (and hence not very useful)
- The quartile break up of the data thus contains roughly 25% data values in first quartile, 25% between second and third quartile and roughly 25% greater than third quartile

Example

- The following data give noise levels measured at 36 different times in Manhattan.
- 82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85, 69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65
- Determine the quartiles.
- The stem and leaf plot of data values is given below:

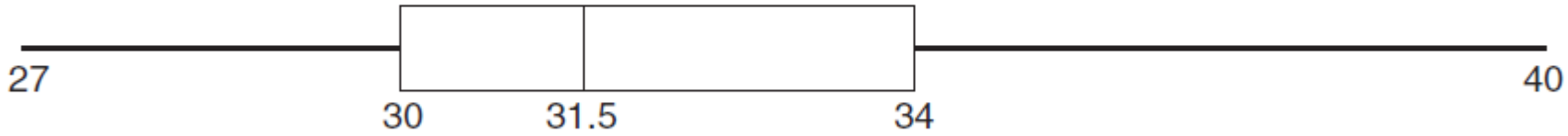
6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

Solution

- First Quartile: p = 0.25, n = 36
- Position = np = 0.25 x 36 = 9
- This implies that the first quartile is average of 9^{th} and 10^{th} data values = (75+77)/2 = 76
- Second Quartile: p = 0.5
- Position = np = 18
- Second quartile value = (89+90)/2 = 89.5
- Third quartile: p = ¾
- Position = 27
- Third quartile value = (102 + 107)/2 = 104.5

Box Plots

- A box plot is often used to plot some of the summarizing statistics of the data
- In a box plot, a straight line is drawn stretching from the lowest to the highest data value, thus, the length of this straight line depicts the range of data values
- This line is imposed by a box which starts at the first quartile and stretches till the third quartile of data
- The second quartile lies inside the box and is indicated with a vertical line
- In the box plot representation, the length of the straight line gives the range of the data and the length of the box itself gives what we call as the *interquartile range* and it gives the difference between the first and the third quartile values



An example of box plot

Exercise: Represent the previous noise example data using box plot.
Also determine the interquartile range for the noise data.

Paired Data Sets and Correlation

- Sometimes, we are interested in finding out the relationship between pairs of data values
- Specifically, we might be interested in knowing how the values of two data pairs change with each other
- For example, we may be interested in knowing the relationship between number of hours studied and performance in examination of a set of students
- There are plenty of real life situations where we are interested in finding out relationship between pairs of data values

Plotting paired datasets

- A useful way of portraying paired data values is to plot them on a two-dimensional scatter plot
- In the scatter plot, the x and y-axes each represent one of the data values
- In the scatter plot, we might roughly observe one of the following:
  - Larger x-values tend to pair with larger y-values
  - Larger x-values tend to pair with smaller y-values
  - None of the above is observed so the x and y values are distributed randomly on the plane

The Correlation Coefficient

- Suppose, we define a metric like $(x_i - \bar{x})$ for x values and similarly, $(y_i - \bar{y})$ for y-values
- Now, if the condition no. 1 holds true then, for larger x_i values $(x_i - \bar{x})$ will also be larger and for larger y_i values $(y_i - \bar{y})$ will also be larger
- Similarly, if x_i is small then, $(x_i - \bar{x})$ will also be small (may be negative) and similar will hold true for $(y_i - \bar{y})$
- Thus, in both the cases, the product of the two quantities i.e.,
- $(x_i - \bar{x})\ (y_i - \bar{y})$ will be a large positive quantity
- This is because the signs of the two numbers will tend to be same either both positive or both negative

- Thus, the statistic defined by paired sum of deviations from mean for the two data sets can give the trend of the relation between the two variables.
- We defined this statistic mathematically as below:

$$\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

- Thus, if larger x-values tend to pair with larger y-values then the above summation is also a large value
- Similarly, if larger x-values tend to pair with smaller y-values then the summation is a small value
- The above result is also followed from a mathematical result known as Hardy's Lemma
- We can standardize the above equation by dividing it with (n-1) and by the standard deviation of both the samples to arrive at our statistic called as the sample *correlation coefficient*.

Definition

- Let $s_x \; and \; s_y$ denote the standard deviation of x and y data values respectively, the sample correlation coefficient ‘r’ is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

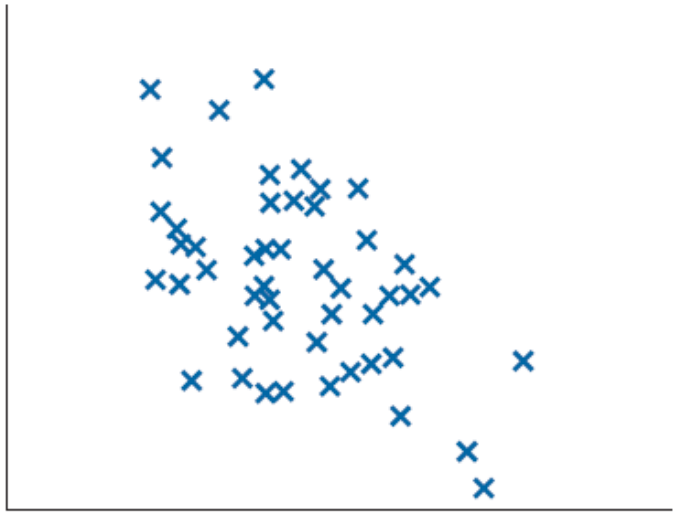
Here, r > 0 indicates positive correlation and r < 0 indicates negative correlation

Properties of r

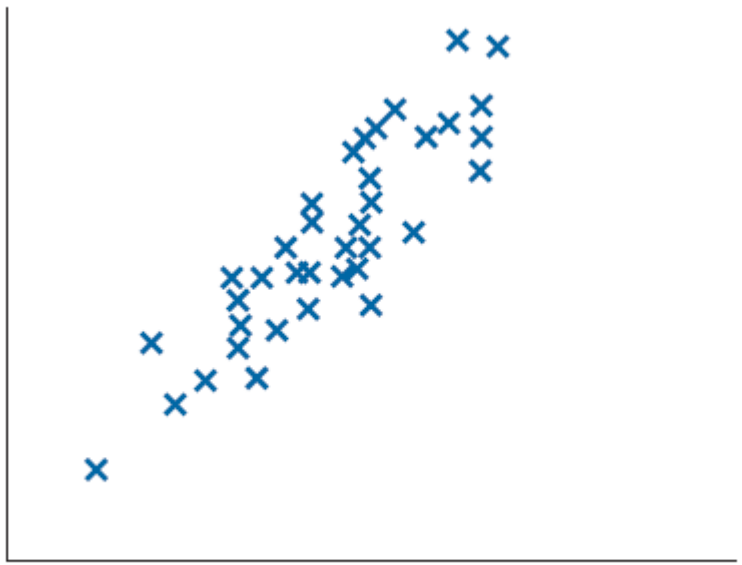
- -1 <= r <= 1
- If for constants a and b, where b > 0
  - $y_i = ax_i + b$ then, r = 1 (x and y are positively correlated)
- If for constants a and b, where b < 0
  - $y_i = ax_i + b$ then, r = -1 (x and y are negatively correlated)
- If r is the sample correlation coefficient for data points xi and yi, i = 1,2…n then, r is also the sample coefficient for
  - $a + bx_i \quad and \quad c + dy_i \quad i = 1, 2, \ldots n$

  provided that b and d are either both positive or both negative

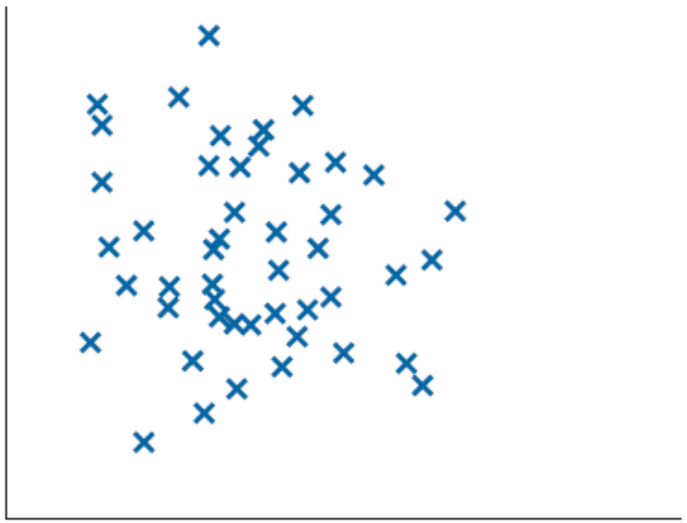
- The absolute value of correlation coefficient r gives the strength of relation between the data pair
- A value of $|r| = 1$ indicates a perfect linear relation between x and y
- A value of $|r| \cong 0.8$ implies a strong correlation although the relation between x and y cannot be represented using a straight line
- A value of $|r| \cong 0.3$ implies a weak correlation
- In general, a value of $|r| > 0.5$ is considered to be a strong correlation
- The sign of r indicates the direction of correlation
- For positive correlation, the sign of r is positive
- And in case of negative correlation between pairs of data values, the sign of r is negative



$r = -.50$



$r = +.90$



$r = 0$

Association v/s Causation

- It is important to note that correlation does not necessarily represent a causative relation between the two variables under study
- This means that a strong positive correlation between two variables does not directly imply that one is the cause of the other
- For making such claims, we must gather more information about the data
- As an example, consider the results of a study in which a strong negative correlation was observed between an individual's years of school and his resting pulse rate
- In this case, it is obvious that we cannot conclude that the lower resting pulse rate was caused by additional years of school attended. Can you think of some factors that might have caused this?

Proof of Property 1

$$\sum \left(\frac{(x_i-\bar{x})}{s_x} - \frac{(y_i-\bar{y})}{s_y}\right)^2 \geq 0$$

$$\sum \frac{(x_i-\bar{x})^2}{{s_x}^2} + \sum \frac{(y_i-\bar{y})^2}{{s_y}^2} - 2\sum \frac{(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \geq 0$$

$$n-1+n-1-2(n-1)r \geq 0$$
$$2n-2-2nr+2r \geq 0$$
$$2n(1-r)-2(1-r) \geq 0$$
$$2(1-r)(n-1) \geq 0$$
$$r \leq 1$$

- Proof for r >= -1

$$\sum \left(\frac{(x_i-\bar{x})}{s_x}+\frac{(y_i-\bar{y})}{s_y}\right)^2 \geq 0$$

$$\sum \frac{(x_i-\bar{x})^2}{{s_x}^2}+\sum \frac{(y_i-\bar{y})^2}{{s_y}^2}+2\sum \frac{(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \geq 0$$

$$n-1+n-1+2(n-1)r \geq 0$$

$$2n-2+2nr-2r \geq 0$$

$$2n(1+r)-2(1+r) \geq 0$$

$$2(1+r)(n-1) \geq 0$$

$$r \geq -1$$

Exercise

- Prove that r = 1 is only possible when there is linear relationship between $x_i \; and \; y_i$

Hint: Note that previous equation can only become equal to 0 if r = 1

Thus, for r = 1 we have $\frac{(x_i - \bar{x})}{s_x} = \frac{(y_i - \bar{y})}{s_y}$

Chebyshev's Inequality

- Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set. Assuming that s>0, Chebyshev's inequality states that for any value of k>= 1, greater than 100(1-1/k^2) percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$.

- For example, let k = 3/2 then, 100x(5/9) = 55.55 i.e., greater than 55.55% of data will lie within $\bar{x} - 1.5s$ to $\bar{x} + 1.5s$.
- Similarly, for k = 2, >75% of data lie within $\bar{x} - 2s$ to $\bar{x} + 2s$
- For k = 3, > 88.9% of data lie within 3 sample standard deviations of $\bar{x}$
- The Chebyshev's inequality can be sharpened when the size of the dataset is known

Example

TABLE 2.7 *Top 10 Selling Cars for 1999*

	1999	
1.	Toyota Camry	448,162
2.	Honda Accord	404,192
3.	Ford Taurus	368,327
4.	Honda Civic	318,308
5.	Chevy Cavalier	272,122
6.	Ford Escort	260,486
7.	Toyota Corolla	249,128
8.	Pontiac Grand Am	234,936
9.	Chevy Malibu	218,540
10.	Saturn S series	207,977

For the present data, $\bar{x}$=298,217.8 and s = 124,542.9

According to Chebyshev's inequality 100(5/9) i.e., greater than 55.55% of data lies within $\bar{x} - 1.5s$ to $\bar{x} + 1.5s$ i.e., (173,674.9, 422,760.67)

However, from the table we can see that in actuality 90% of the data falls within this limit.

Chebyshev's Inequality

Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set containing x1, x2,…xn values, where s>0. Let

$$S_k = \{i, 1 \leq i \leq n: |x_i - \bar{x}| < ks\}$$

And let N(S_k) be the number of elements in the set S_k. Then, for any $k \geq 1$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

Proof:

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

$$= \sum_{i \in S_k}(x_i - \bar{x})^2 + \sum_{i \notin S_k}(x_i - \bar{x})^2$$

$$\geq \sum_{i \notin S_k}(x_i - \bar{x})^2$$

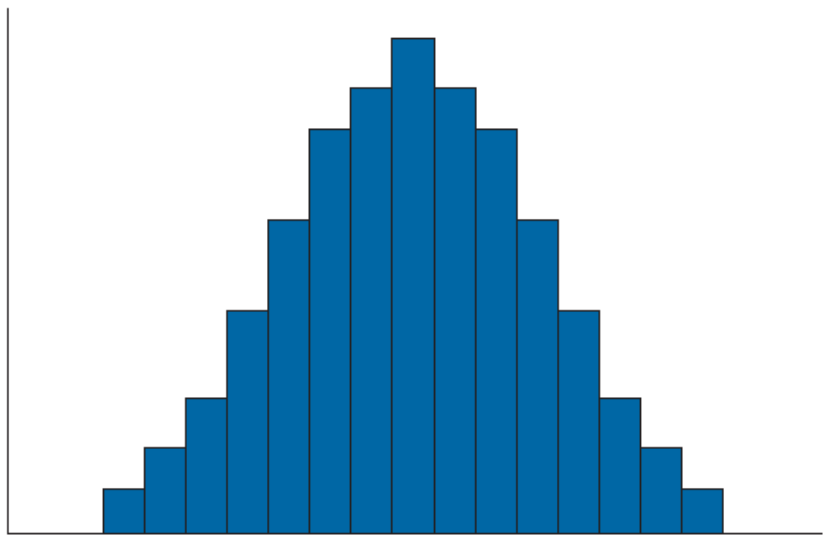
$$\geq \sum_{i \notin S_k} k^2 s^2$$

$$= k^2 s^2 (n - N(S_k))$$

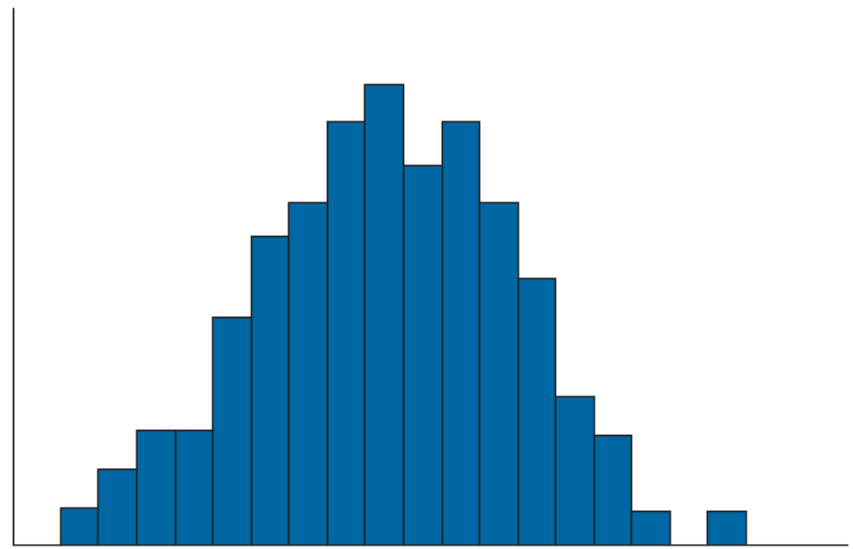
$$\frac{n-1}{nk^2} \geq 1 - \frac{N(S_k)}{n}$$

Normal Data Sets

- Many of the large data sets observed in real-life have histograms that are similar in shape
- These histograms reach their peaks near the sample median and decrease on both sides thus forming a bell-shape
- Such data sets are called normal data sets and their histograms are known as normal histograms
- For normal datasets, the sample mean and sample median are both equal



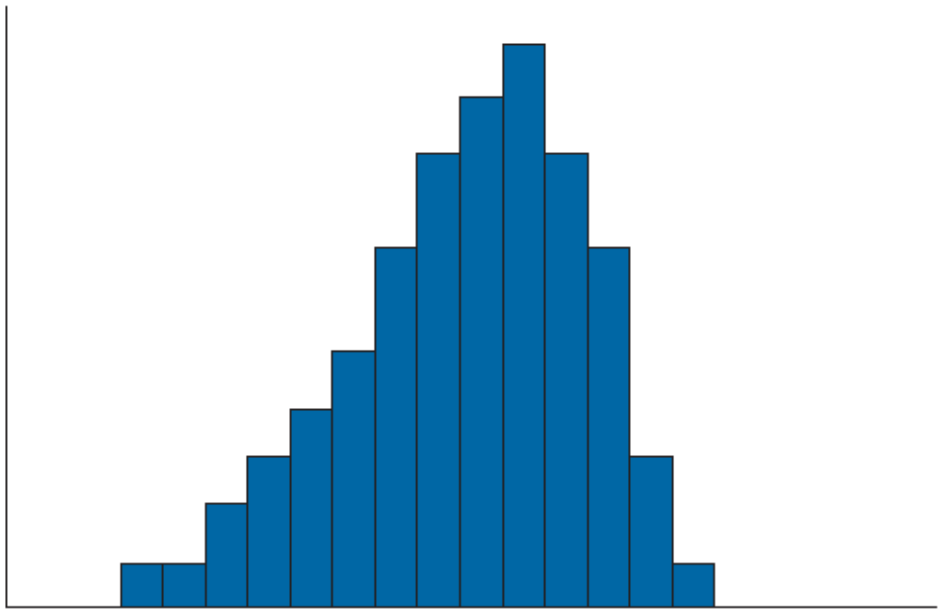
Histogram of a normal data set.



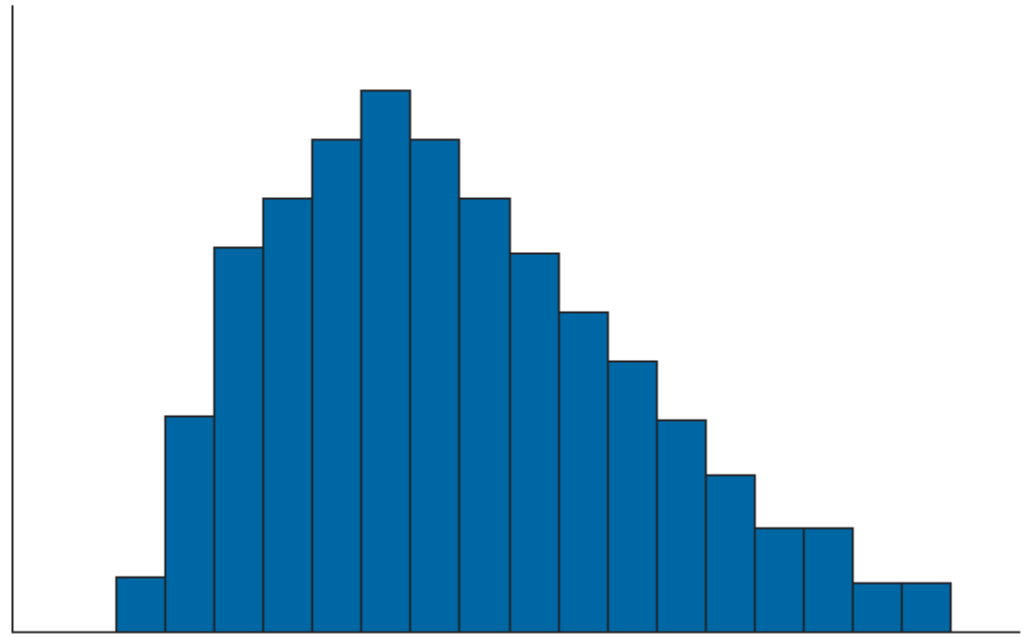
Histogram of an approximately normal data set.

Approximately Normal Datasets and Skewness

- If the histogram of a data set is close to normal (but not exactly bell shaped or normal) we call that dataset approximately normal.
- It follows from the definition of normal histograms that for approximately normal datasets, the sample mean and sample median are approximately equal
- We also come across data sets for which the histograms deviate too much from the bell shape. Such data sets are called *skewed*.
- *Skewed* data sets are not approximately normal about their medians
- Such dataset is called right skewed if it has a long tail to the right
- And it is called left skewed if it has a long tail to the left



Histogram of a data set skewed to the left.



Histogram of a data set skewed to the right.

Probability

Sample Space and Events, Axioms of Probability, Conditional Probability

Introduction

- There exist many real life situations where we do not know in advance about the happening of an event
- For example, in tossing of a coin we don’t know before tossing whether a Head or a Tail will come
- However, in most such cases the possible outcomes are known beforehand
- Probability is concerned with the chances of occurrence of a possible outcome out of the set of all possible outcomes
- There are generally two ways in which we can think of probability:
  - Frequency Interpretation
  - Subjective Interpretation

Frequency Interpretation

- Here, it is thought that the probability ‘p’ of occurrence of an outcome E, implies that if an experiment is designed to check the occurrence of E; this experiment is repeated many times and number of times E occurs is recorded then the proportion of times E occurs when experiments is repeated for very large number of times is ‘p’.
- In this interpretation, the probability of a given outcome is a property of the outcome itself and not of the person doing the experiment
- This interpretation is most prevalent among scientists

Subjective Interpretation

- In the subjective interpretation, the probability of an outcome is thought of as the belief of the person quoting the probability.
- Thus, the probability value now does not depend on the outcome but on the belief of the person and has no meaning outside of expressing one's degree of belief
- This interpretation of probability is often favoured by Philosophers and some economic decision makers
- The discussions we will have here are independent of any particular interpretations of probability

Sample Space and Events

- Let us consider an experiment whose outcome is not predictable in advance however, the set of all possible outcomes of the experiment is known.
- This set of all possible outcomes of an experiment is known as the *Sample Space* of the experiment and is denoted by S.
- Any subset E of the sample space is known as an *Event*.
- Thus, E is a set consisting of possible outcomes of a sample space S
- If the outcome of an experiment is contained in E then, we say that E has occurred.

Examples: Sample Space

- In the coin tossing experiment, the sample space consists of two values head and tail
  - i.e., S = {H, T}
- In the experiment of rolling of a die, the sample space consists of 6 values 1,2..6
  - i.e., S = {1, 2, 3, 4, 5, 6}
- If the experiment consists of running of a race among 5 horses having post positions given as 1, 2, 3, 4, 5 then,
  - S = {all possible orderings of (1, 2, 3, 4, 5)}

Examples: Event

- In the coin tossing experiment, ‘a head comes in the outcome’ is an event and can be denoted as:
  - E = {H}
- In the experiment of rolling of a die, ‘an even number appears’ is an event and can be denoted as:
  - E = {2, 4, 6}
- In the horse racing experiment, ‘Horse 2 wins the race’ is an event and can be denoted as:
  - E = {all outcomes of S starting with 2}

Properties of Events

- *Union*: For any two events E and F, we defined a new event $E \cup F$ as the union of E and F. The union consists of all outcomes of S that are either in E or in F or in both E and F.
- *Example:* In the rolling of a die experiment, Let E be the event that an even number appears and F be the event that a prime number appears, then $E \cup F$ represents the event that either a prime number or an even number appears. In this case,

$$E \cup F = \{1,2,3,4,5,6\}$$

- *Intersection*: For any two events E and F, a new event EF is defined as the intersection of E and F. EF consists of all the outcomes that are both in E and F. Thus, EF is said to occur only when both E and F occur
- *Example:* In the horse racing experiment, let E be the event denoting Horse 2 wins the race and F be the event that Horse 5 comes second. Then, $E \cap F$ *or EF* is the event that Horse 2 wins the race and Horse 5 comes second. Thus, EF consists of all permutations of numbers 1 through 6 which begin with 2 and are followed by 5.

- *Mutually Exclusive:* When the intersection of two events E and F does not contain any outcomes, we call the events E and F mutually exclusive. Thus, for mutually exclusive events $E \cap F = \emptyset$
- *Complement:* We define the complement of an event E as another event

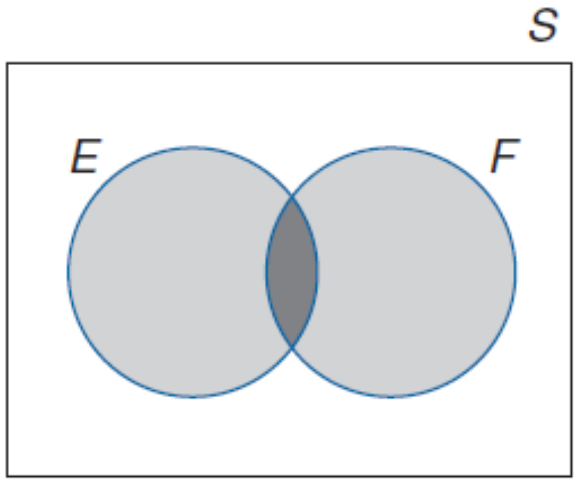
E^c consisting of all outcomes in the sample space S that are not in E. Thus, E^c occurs if and only if E does not occur or $E \cap E^c = \emptyset$

- Example: In rolling of a die experiment where we had defined an event E 'an even number appears', the event E^c consists of all the outcomes of the sample space S where an odd number appears.

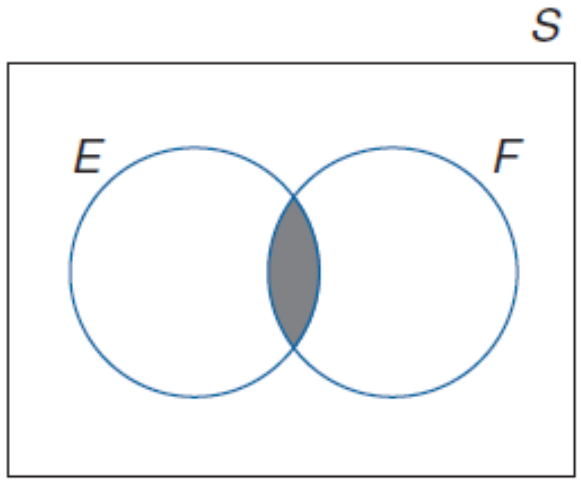
  i.e. E^c = {1,3,5}
- *Subset*: For any two events E and F, we say that E is contained in F if, all the outcomes of E are in F. E is thus a subset of F and is denoted by $E \subset F$
- If $E \subset F$ and $F \subset E$ then, E = F. i.e., E and F are identical events
- The definition of union and intersection can be extended to any number of events

Venn Diagrams

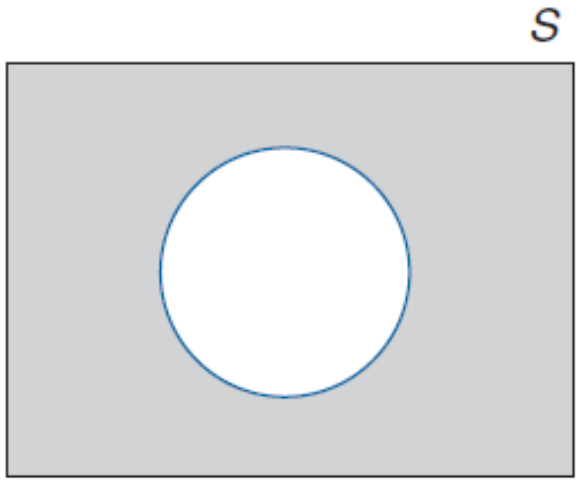
- Venn diagram is a technique used for representation of logical relations between events.
- Here, the Sample space S is depicted as a large rectangle and any event is represented as a circle within that rectangle
- Remember that sample space consists of all possible outcomes and is translated to all points within the rectangle
- An event is a subset of sample space and is therefore, shown by a circle within that rectangle. All the points within the circle depict the outcomes of the experiment that form the event
- Events of interest in Venn diagram are indicated by shading the appropriate portion of the diagram (rectangle)



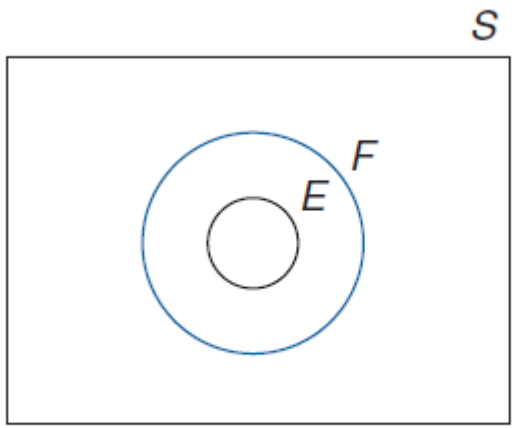
(a) Shaded region: $E \cup F$



(b) Shaded region: EF



(c) Shaded region: E^c



$E \subset F$

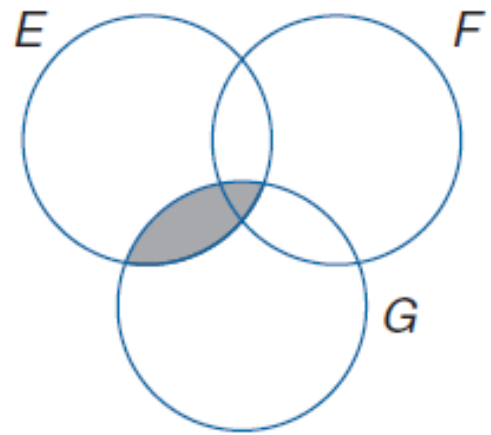
Laws obeyed by Operations on Events

- Commutative Law: $E \cup F = F \cup E \quad and \quad E \cap F = F \cap E$
- Associative Law: $(E \cup F) \cup G = E \cup (F \cup G) \quad and \quad (EF)G = E(FG)$
- Distributive Law: $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$

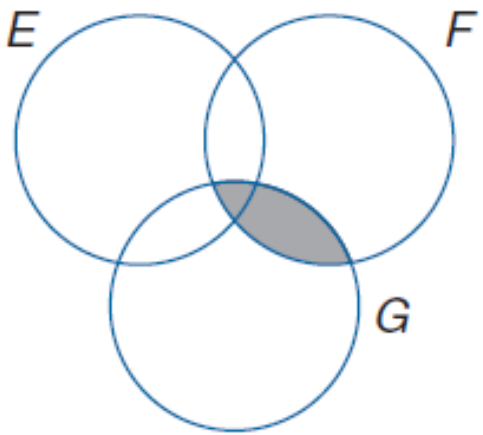
$$EF \cup G = (E \cup G) \cap (F \cup G)$$

- DeMorgan's Law: $(E \cup F)^c = E^c \cap F^c$

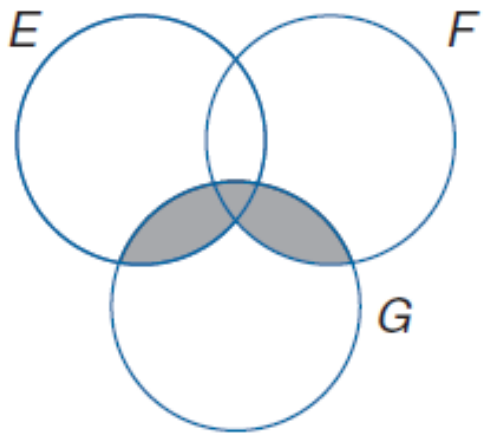
$$(E \cap F)^c = E^c \cup F^c$$



(a) Shaded region: EG



(b) Shaded region: FG



(c) Shaded region: $(E \cup F)G$

$(E \cup F)G = EG \cup FG$

Axioms of Probability

- When we speak of probability of an event E, we in our mind think of it as a constant value obtained by repeating the same experiment under same conditions for very large number of times and counting the proportion of times a particular event occurs or the outcome of the experiment is contained in E
- From a purely mathematical viewpoint, we suppose that the probability of an event E for an experiment with sample space S (denoted by P(E)) follows three axioms.

- Axiom 1: $0 \leq P(E) \leq 1$
- Axiom 2: $P(S) = 1$
- Axiom 3: For any sequence of mutually exclusive events $E_1, E_2, E_3 \ldots E_n$

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$$

n = 1, 2, .. ∞

- When all the above three axioms hold, we call P(E) as the probability of the event E
- Check that the relative frequency interpretation of probability is still satisfied

Proposition 1

- $P(E^c) = 1 - P(E)$

Proof:

Remember, $E \; and \; E^c$ are mutually exclusive and $E \cup E^c = S$

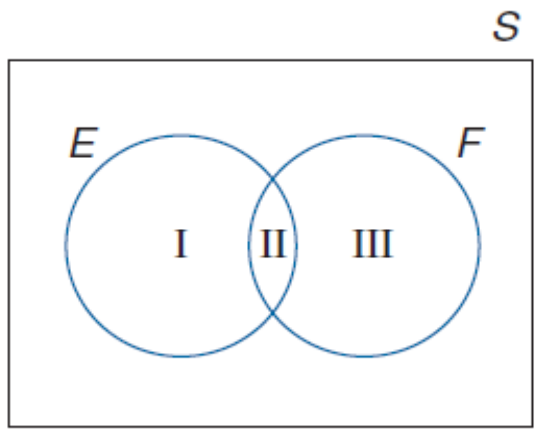
Thus, $\mathrm{P}(S) = P(E \cup E^c) = P(E) + P(E^c)$

Note that $P(S) = 1$

Therefore, $P(E^c) = 1 - P(E)$

Proposition 2

- $P(E \cup F) = P(E) + P(F) - P(EF)$



Proof: From the Venn diagram, we have the following:

$P(E \cup F)$ = P(I) +P(II) + P(III)

P(E) = P(I) + P(II)

P(F) = P(II) + P(III)

Thus, $P(E \cup F) = P(E) + P(F) - P(II)$

Example 1

- A total of 28 percent of males smoke cigarettes, 7 percent smoke cigars, and 5 percent smoke both cigars and cigarettes. What percentage of males smoke neither cigars nor cigarettes?

Odds of an Event

- The odds of an event E are defined as follows:

$$\frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}$$

- The odds of an event A tells us how likely it is that A occurs than that it does not occur.
- Example: Let P(A) = ¾ then, odds of A = (3/4)/(1/4) = 3
- Thus, it is 3 times more likely that A occurs than that it does not occur

The case of Equally Likely Outcomes

- Many experiments that we perform have equally likely outcome
- This implies that all the outcomes in the sample space are equally likely to occur
- Thus, if the sample space consists of N points denoted as

  S = {1, 2, 3… n},
- we have, P(1) = P(2)… = P(n) = *p* (say)
- Therefore, P(S) = 1 = P(1) + P(2) + … + P(n) = Np

$$\Rightarrow P(i) = 1/N$$

Calculating Probability of an Event

- Using the above relation, we can describe the probability of any event E as:

$$P(E) = \frac{Number\ of\ Points\ in\ E}{N}$$

- Thus, for experiments with equally likely outcomes, the probability of an event E is proportional to the number of points in the sample space that are contained in E
- In all such cases, we can compute the probabilities by simply *counting the number of ways a particular event can occur*

Basic Principle of Counting

- Suppose two experiments are performed. Let experiment 1 can result in m possible outcomes and for each outcome of experiment 1, experiment 2 can result in n possible outcomes, then there are total of mn possible outcomes of the two experiments
- Above method can be used to calculate the probability of events in various scenarios
- The basic principle of counting can be generalized to any number of r experiments
- This is same as counting the permutations of r numbers i.e., r!
- Here 0! = 1! = 1

Example 2

- Mr. Jones has 10 books that he is going to put on his bookshelf. Of these, 4 are mathematics books, 3 are chemistry books, 2 are history books, and 1 is a language book. Jones wants to arrange his books so that all the books dealing with the same subject are together on the shelf. How many different arrangements are possible?

$$4!\,(4!\,3!\,2!\,1!) =$$

Example 3

- A class in probability theory consists of 6 men and 4 women. An exam is given and the students are ranked according to their performance. Assuming that no two students obtain the same score, **(a)** how many different rankings are possible? **(b)** If all rankings are considered equally likely, what is the probability that women receive the top 4 scores?
- a) 10!
- b) 4!6!/10!

Counting where ordering is not important

- Suppose, we want to know the number of ways in which r items may be selected from a set of n items
- If in making such selections, the order of selection is not important then we need to divide the count from r

i.e., $$P(E) = \frac{n.(n-1).(n-2)\ldots(n-r+1)}{r!}$$

$$= \frac{n!}{(n-r)!r!}$$

- The above expression is called as combination and denoted as $\binom{n}{r}$

Example 4

- A committee of size 5 is to be selected from a group of 6 men and 9 women. If the selection is made randomly, what is the probability that the committee consists of 3 men and 2 women?

$$6_{C_3}.9_{C_2} = 5.4.9.4 = 720$$

P(E) = 720/(15_{C_5}) = 240/1001

Example 5

- From a set of n items a random sample of size k is to be selected. What is the probability a given item will be among the k selected?

$$P(E) = (1.{}^{n-1}_{k-1}C)/{}^{n}_{k}C$$

Example 6

- If n people are present in a room, what is the probability that no two of them celebrate their birthday on the same day of the year?

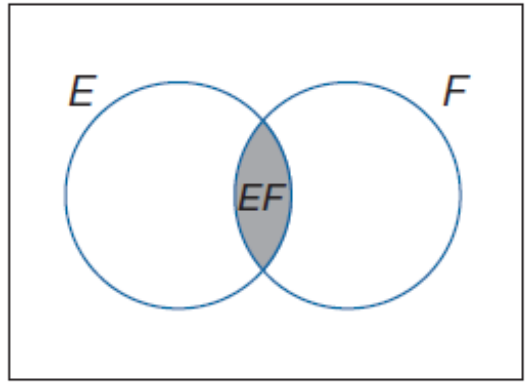
$$\frac{365.364.363\ldots(365-n+1)}{(365)^n}$$

Conditional Probability

- By conditional probability we mean that the probability of occurrence of an event given that some other event has happened
- The advantage of calculating conditional probabilities is two fold:
  - We are interested in finding the probability of occurrence of an event in the light of some additional information
  - Sometimes the easiest way of calculating the probability of an event is to condition it on some other event

Definition

- Suppose E and F are two events and we are interested in finding out the probability of occurrence of E given that F has occurred.
- Clearly, E given F consists of all and only those outcomes which are common in both E and F
- Since F has already occurred, we can think of the new sample space as consisting of all outcomes in F
- P(E|F) = (n1/N)/(n/N) => P(EF)/P(F)
- It is defined only when P(F) > 0
- Above equation can also be interpreted in the form of long-run relative frequency



Example 7

- A bin contains 5 defective (that immediately fail when put in use), 10 partially defective (that fail after a couple of hours of use), and 25 acceptable transistors. A transistor is chosen at random from the bin and put into use. If it does not immediately fail, what is the probability it is acceptable?
- $P(A|NF) = P(ANF)/P(NF) = (25/40)/(35/40)$

$P(A|NF) = 25/35 = 5/7$

Example 8

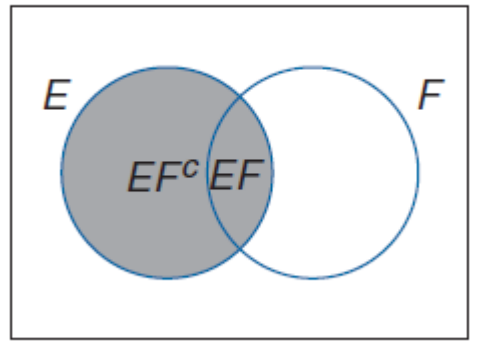
- Ms. Perez figures that there is a 30 percent chance that her company will set up a branch office in Phoenix. If it does, she is 60 percent certain that she will be made manager of this new operation. What is the probability that Perez will be a Phoenix branch office manager?

- $P(M) = P(M|B)P(B) = 0.6 \times 0.3 = 0.18$ (18% chance)

Bayes' Formula

- Let E and F be two events then, irrespective of what events E and F are, we can always write E as:

$$E = EF \cup EF^c$$



- Clearly, $EF \; and \; EF^c$ are mutually exclusive therefore,

$$P(E) = P(EF) + P(EF^c)$$

$$= P(E|F)P(F) + P(E|F^c)P(F^c)$$

$$= P(E|F)P(F) + P(E|F^c)[1-P(F)]$$

Interpretation

- The previous equation gives a very important result
- The probability of an event E can be calculated as a weighted average conditioned on probability of occurrence and non-occurrence of another event F
- Each conditional probability is weighted as much as the probability the event it is conditioned on has of occurring
- Thus, this formula allows us to determine the probability of an event by first conditioning it on whether or not some other event has occurred. This is useful because many times it is difficult to calculate the probability of an event directly, however it can be easily calculated if we know whether or not some second event has happened.

Example 9

- An insurance company believes that people can be divided into two classes — those that are accident prone and those that are not. Their statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability .4, whereas this probability decreases to .2 for a non-accident-prone person. If we assume that 30 percent of the population is accident prone, what is the probability that a new policy holder will have an accident within a year of purchasing a policy?

A: a new policy holder will meet an accident within 1 year

AP: person is accident prone

$P(AP) = 0.3$

$P(A) = P(A|AP).P(AP) + P(A|NAP).P(NAP)$

$=0.4x0.3 + 0.2x0.7 = 0.12 + 0.14 = 0.26$

Example 10

- Reconsider Example 9 and suppose that a new policy holder has an accident within a year of purchasing his policy. What is the probability that he is accident prone?

$P(AP|A) = P(AP \cap A)/P(A) = P(A|AP)P(AP)/P(A) = 0.4 \times 0.3/0.26 =$

Example 11

- A laboratory blood test is 99 percent effective in detecting a certain disease when it is, in fact, present. However, the test also yields a “false positive” result for 1 percent of the healthy persons tested. (That is, if a healthy person is tested, then, with probability .01, the test result will imply he or she has the disease.) If .5 percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive?

D: Person is having the disease

Pt: Test result is positive

$P(D|Pt) = P(Pt \cap D)|P(Pt) = P(Pt|D)P(D)/P(Pt) = 0.99x0.005/P(Pt) = 0.332$

$P(Pt) = P(Pt|D).P(D) + P(Pt|ND).P(ND) = 0.99x0.005 + 0.01x0.995$

Example 12

- At a certain stage of a criminal investigation, the inspector in charge is 60 percent convinced of the guilt of a certain suspect. Suppose now that a new piece of evidence that shows that the criminal has a certain characteristic (such as left-handedness, baldness, brown hair, etc.) is uncovered. If 20 percent of the population possesses this characteristic, how certain of the guilt of the suspect should the inspector now be if it turns out that the suspect is among this group?

G: the person is guilty

C: a person is having the characteristic

$P(G|C) = P(GC)/P(C) = P(C|G)P(G)/P(C) = 1x0.6/P(C) =$

$P(C) = P(C|G)P(G) + P(C|NG)P(NG) = 1x0.6 + 0.2x0.4 = 0.6 + 0.08 = 0.68$

Example 13

- Let us now suppose that the new evidence is subject to different possible interpretations, and in fact only shows that it is 90 percent likely that the criminal possesses this certain characteristic. In this case, how likely would it be that the suspect is guilty (assuming, as before, that he has this characteristic)?

$$P(G|C) = P(GC)/P(C) = P(C|G)P(G)/P(C) = 0.9x0.6/P(C) =$$

$$P(C) = P(C|G)P(G) + P(C|NG)P(NG) =0.9x0.6 + 0.2x0.4 =$$

Generalization of Bayes' Formula

- The Bayes' formula given on slide 34 can be generalized as follows:
- Suppose $F_1, F_2, \ldots F_n$ are mutually exclusive events that

$$\bigcup_{i=1}^{n} F_i = S$$

In other words, exactly one of the events $F_1, F_2, \ldots F_n$ must occur. Also

$$E = \bigcup_{i=1}^{n} E \cap F_i$$

$$P(E) = \sum_{i=1}^{n} P(EF_i)$$

$$= \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Thus, for given events $F_1, F_2, \ldots F_n$ of which one and only one must occur, we can compute P(E) by first conditioning on which one of the F_i has occurred.

- Suppose now that E has occurred and we are interested in determining which one of the F_j also occurred

$$P\left(F_j\middle|E\right) = \frac{P(EF_j)}{P(E)}$$

$$= \frac{P\left(E\middle|F_j\right)P(F_j)}{\sum_{j=1}^{n} P\left(E\middle|F_j\right)P(F_j)}$$

This equation is known as the Bayes formula after the English philosopher Thomas Bayes.

Example 14

- A plane is missing and it is presumed that it was equally likely to have gone down in any of three possible regions. Let $1-\alpha i$ denote the probability the plane will be found upon a search of the ith region when the plane is, in fact, in that region, i =1, 2, 3. (The constants αi are called overlook probabilities because they represent the probability of overlooking the plane; they are generally attributable to the geographical and environmental conditions of the regions.) What is the conditional probability that the plane is in the ith region, given that a search of region 1 is unsuccessful, i = 1, 2, 3?

Solution

- Ri = search in region i is unsuccessful
- Ei = plane is in region i
- P(R1|E1) = $\alpha 1$, P(R2|E2) = $\alpha 2$, P(R3|E3) = $\alpha 3$

- P(E1|R1) = P(E1∩R1)/P(R1) = P(R1|E1)P(E1)/P(R1) = $\alpha 1(\frac{1}{3})$ /P(R1) = $\alpha 1/\ \alpha 1$ + 2
- P(R1) = P(R1|E1)P(E1)+ P(R1|E2)P(E2) + P(R1|E3)P(E3) = $\alpha 1(\frac{1}{3})$ + $1.\left(\frac{1}{3}\right) + 1.\left(\frac{1}{3}\right)$

- P(E2|R1) = P(E2∩R1)/P(R1) = P(E2∩R1)/P(R1) = P(R1|E2)P(E2)/P(R1)

- P(E3|R1) =

Independent Events

- In general the conditional probability $P(E|F)$ is not equal to the unconditional probability of E i.e. $P(E)$ or $P(E) \neq P(E|F)$ generally.
- Thus, knowing that F has occurred usually changes the chances of occurrence of E leading to change in the probability value
- However, in some special cases $P(E|F) = P(E)$ i.e., occurrence of E is unaffected by occurrence of F.
- In such cases, we say that the event E is independent of F
- Thus, event E is independent of F if knowledge of occurrence of F does not change the probability that E occurs

Definition

- Thus, for an event E independent of F we have,

$$P(E|F) = P(E)$$

- Since

$$P(E|F) = P(EF)/P(F)$$

- We have,

$$P(EF) = P(E)P(F)$$

- Since, the above equation is symmetric we conclude that if E is independent of F then F is also independent of E
- Thus, two events are independent if $P(EF) = P(E)P(F)$. If two events are not independent then they are called dependent

Independence of E and F^c

- If E and F are independent, then E and F^c are also independent

Proof:

$$E = EF \cup EF^c$$

As $EF\ and\ EF^c$ are mutually exclusive, we have:

$$P(E) = P(EF) + P(EF^c)$$

$$\mathrm{P(E)} = \ P(E)P(F) + P(EF^c)$$

$$P(EF^c) = P(E)\big(1 - P(F)\big)$$

$$P(EF^c) = P(E)P(F^c)$$

- Thus, if E is independent of F then occurrence of E is unchanged by the information as to whether or not F has occurred
- If E is independent of F and is also independent of G. Then, E is not necessarily independent of FG.
- **Example:** Two fair dice are thrown. Let E7 denote the event that the sum of the dice is 7. Let F denote the event that the first die equals 4 and let T be the event that the second die equals 3. What is the probability of occurrence of E7?

Definition

- Three events E, F and G are said to be independent if

$$P(EFG) = P(E)P(F)P(G)$$

$$P(EF) = P(E)P(F)$$

$$P(EG) = P(E)P(G)$$

$$P(FG) = P(F)P(G)$$

- Note that, if E, F and G are independent, then E will be independent of any event formed from F and G as well
- Example: E is independent of F U G

$$P(E(FUG)) = P(EF \cup EG)$$
$$= P(EF) + P(EG) - P(EFG)$$
$$= P(E)P(F) + P(E)P(G) - P(E)P(F)P(G)$$
$$= P(E)(P(F) + P(G) - P(F)P(G))$$
$$= P(E)P(F \cup G)$$

- Above definition can be extended to any number of events i.e. events E1, E2… En are independent if for every subset of r events (r <= n) we have,

$$P(E1', E2', \ldots Er') = P(E1')P(E2')\ldots P(Er')$$

Example 15

- A system composed of n separate components is said to be a parallel system if it functions when at least one of the components functions. For such a system, if component i, independent of other components, functions with probability pi , i = 1, . . . , n, what is the probability the system functions?

$$1 - \prod_{i=1}^{n} (1 - pi)$$

Random Variables

Random variables, Jointly Distributed Random variables, Expectation, Variance, Co-variance

Random Variables

- When random experiments are performed, the quantities of interest can often be represented as some function of the outcome of experiment and we are interested in finding the values of these functions rather than the outcome itself
- For example, when tossing a coin we might be interested in knowing the number of times Head came or the sum of values obtained when rolling two dice
- Such quantities can be represented by defining a real-valued function on the sample-space of the experiment
- We call such a function a *Random Variable*

Random Variable and Probability

- Since each outcome can be associated with some real-value of the random variable
- We can talk about the probability of the random variable taking a particular value in a similar way as we talk about the probability of occurrence of a particular event
- Thus, each possible value of the random variable can be associated with a probability
- We denote it as $P(X= x_i)$
- Here, X is the random variable which is taking a particular value xi

Example 1

- Suppose an experiment is performed where 3 fair coins are tossed and the number of Heads obtained is recorded. Suppose X denotes the random variable number of Heads obtained then the probabilities can be represented as follows:
- P(X = 0) =P(TTT) = $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$
- P(X=1) = P{(HTT),(THT),(TTH)} = $\frac{1}{2}\frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{3}{8}$
- P(X=2) = P{(HHT),(HTH),(THH)} = $\frac{3}{8}$
- P(X=3) = P(HHH) = $\frac{1}{8}$

Total Probability

- We have defined the random variable on the outcomes of the experiment
- This implies that the random variable must take some value when the experiment is performed
- Therefore, the probabilities of all possible values of the random variable must sum to 1

$$1 = P(S) = P\left(\bigcup_{i=1}^{3} X = i\right) = \sum_{i=1}^{3} P(X = i)$$

Example 2

- A life insurance agent has 2 elderly clients, each of whom has a life insurance policy that pays Rs. 1,00,000 upon death. Let Y be the event that the younger one dies in the following year, and let O be the event that the older one dies in the following year. Assume that Y and O are independent, with respective probabilities P(Y) = 0.05 and P(O) = 0.1. If X denotes the total amount of money (in units of 1,00,000) that will be paid out this year to any of these client's beneficiaries, then X is a random variable that takes on the one of the possible values 0, 1, 2 and the respective probabilities can be given as:

  P(X=0) = $P(Y^c O^c) = (0.95 \times 0.9) =$

  P(X=1)= $P(Y^c O) + P(Y O^c) = 0.95 \times 0.1 + 0.05 \times 0.9 =$

  P(X=2)= $P(YO) =$

Example 3

- Suppose that an individual purchases two electronic components each of which may be either defective or acceptable. In addition, suppose that the four possible results — (d, d), (d, a), (a, d), (a, a) — have respective probabilities .09, .21, .21, .49. If we let X denote the number of acceptable components obtained in the purchase, then X is a random variable taking on one of the values 0, 1, 2. The respective probabilities would be:

  P(X=0) = 0.09

  P(X=1) = 0.21 + .021 = 0.42

  P(X=2) = 0.49

Question: How do we find probabilities for random variable at least one acceptable component?

P(I=1) = 0.42 + 0.49 =

P(I=0) = 0.09

Types of Random Variables

- Random variables are categorized based on the possible values that can be taken be them. There are two types of random variables:
  - Discrete random variable
  - Continuous random variable
- When the random variable can take at most countable number of possible values, it is called as Discrete random variable
- Continuous random variable can take all possible values. For example, the random variable denoting lifetime of a bulb

Cumulative Distribution Function

- Also known as simply *distribution function* is defined as a function F on a random variable X for any real number x by

$$F(x) = P\{X \leq x\}$$

- In other words, F(x) is the probability that the random variable X takes on a value less than or equal to x
- X∼F is used as a notation to denote that F is the distribution function of X

Calculating Probabilities using Distribution Function

- Distribution function can be used to calculate all the probabilities related to X
- For example, the probability that X lies between two values a and b i.e. $P\{a < X \leq b\}$ can be calculated as follows:

$$P\{X \leq b\} = P(X \leq a) + P\{a < X \leq b\} \qquad \text{from axiom 3}$$

$$F(b) = F(a) + P(a < X \leq b)$$

$$P(a < X \leq b) = F(b) - F(a)$$

Example

Suppose the random variable X has distribution function

$$F(x) = \begin{cases} 0 & x <= 0 \\ 1-\exp(-x^2) & x>0 \end{cases}$$

What is the probability that x exceeds 1?

Solution

- $P(X > 1) = 1 - P(X \leq 1)$

$= 1 - F(1)$

$= 1 - (1 - e^{-1})$

$= e^{-1}$

Discrete Random Variables

- A discrete random variable takes sequence of values or countable values
- The probability of the discrete random variable X taking any of the possible values is given by probability mass function p(a) of X
- We denote this as follows:

  p(a) = P{X=a}

- Suppose, X must assume one of the values x1, x2… . Then,

  p(xi) > 0 i = 1, 2…

  p(x) = 0 for all other values of x

- Since X must take one of the values from x1, x2… we have,

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

Example

- Consider a random variable X taking values 1, 2 or 3. Suppose p(1) = 1/2, p(2) = 1/3. Then, we can calculate the value of p(3) as

  p(3) = 1-{p(1) + p(2)} = 1/6

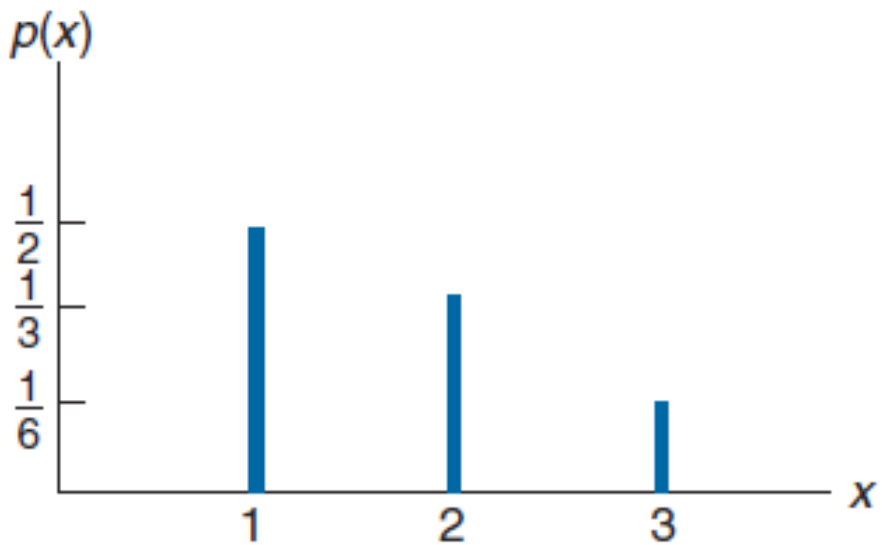
- The cumulative distribution function for a discrete random variable can be expressed as:

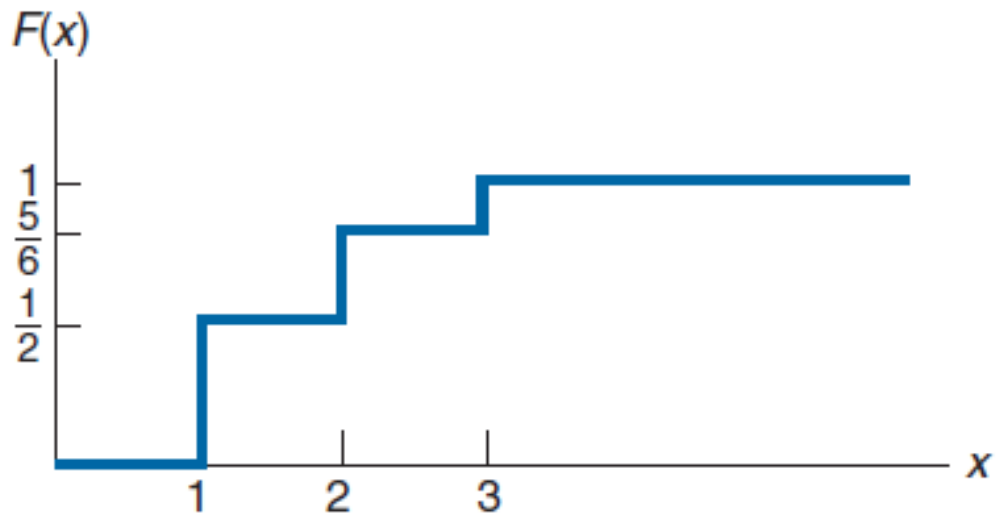
$$F(a) = \sum_{x \leq a} p(a)$$

- Clearly, for discrete values x1, x2… the cumulative distribution function will be a step function

- The cumulative distribution function for the discrete random variable X shown in previous example is given by:

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


p(x)
1/2
1/3
1/6
1
2
3
x


F(x)
1
5/6
1/2
1
2
3
x

Continuous Random Variable

- Many times the set of possible values for a random variable is not a sequence but an interval, such random variables are known as *continuous* random variable
- X is said to be a continuous random variable if there exists a non-negative function f(x) defined over $x\epsilon(-\infty,\infty)$ such that for any set B of real numbers

$$\mathrm{P}\{x\epsilon B\} = \int_B f(x)dx$$

- The function f(x) is called as the probability density function of the random variable X

- This implies that the probability that X will belong to the interval B can be obtained by integrating the probability density function over the set B
- Again, since X must assume some value, i.e.,

$$P\{X \epsilon (-\infty, \infty)\} = 1$$

Therefore,

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

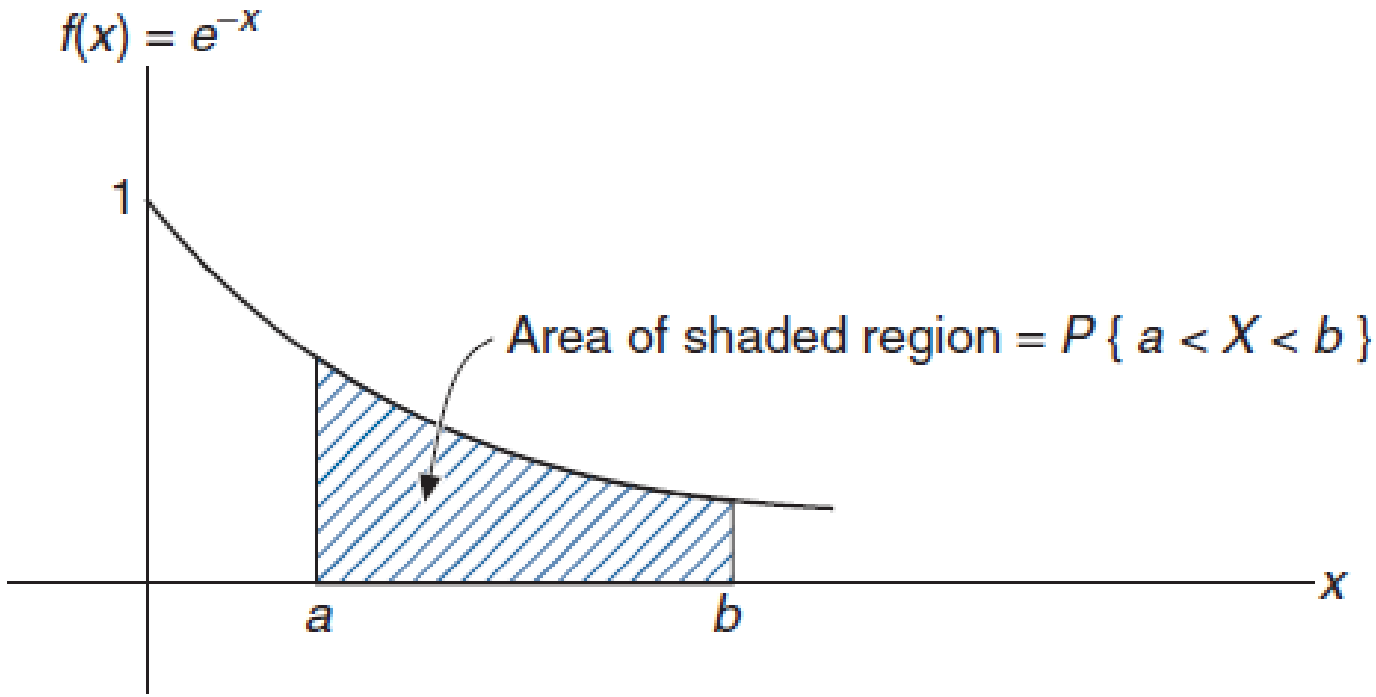
- Similar to probability mass function of discrete random variable, all probability statements about continuous random variable can be answered in terms of probability density function f(x)
- E.g. $P(a \leq X \leq b) = \int_a^b f(x)dx$

Value of Density Function at a point

- In the previous example, we calculated the probability of random variable X lying between the interval [a,b]
- Suppose, we set a = b. That is, we are interested in finding the probability of continuous random variable X taking a particular value ‘a’
- From the previous equation we get,

$$P(X = a) = \int_a^a f(x)dx = 0$$

- This implies that the probability of a continuous random variable taking a particular value is **zero**.



The probability density function $f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$.

- A more intuitive explanation of the probability density function can be obtained by representing the density function as follows:

$$P\left\{a - \frac{\varepsilon}{2} \leq X \leq a + \frac{\varepsilon}{2}\right\} = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)dx \approx \varepsilon f(a)$$

where, ε is very small

- Density function at some point a denoted as f(a) is the measure of how likely it is that the random variable will be near a

Distribution Function for Continuous Random Variable

- The cumulative distribution function for a continuous random variable is defined in a similar fashion as for the discrete case

$$F(a) = P\{X\epsilon(-\infty, a)\} = \int_{-\infty}^{a} f(x)dx$$

Further, differentiating both sides w.r.t. a gives:

$$\frac{d}{da}F(a) = f(a)$$

- Thus, density is the derivative of distribution function

Example

- Suppose that X is a continuous random variable whose probability density function is given by:

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

1. Find the value of C.
2. Find P(X>1)

$$\int_{-\infty}^{\infty} C\,(4x\ - 2x^2)dx = 1$$

$$= \int_0^2 C(4x - 2x^2)dx = 1 \quad \rightarrow \text{evaluate to obtain C}$$

$$P(X > 1) = 1 - P(X \leq 1) = 1 - \int_{-\infty}^{1} C(4x - 2x^2)dx$$

- $P(EF) = P(F)\,P(E|F) = P(E)\,P(F|E)$

Jointly Distributed Random Variables

- Many times we are interested in the joint distribution of random variables and not only in their individual probability distributions
- E.g. number of cigarette smoked daily and the age of contracting cancer
- The joint cumulative probability distribution of two random variables X and Y can be expressed as:

$$F(X,Y) = P\{X \leq x, Y \leq y\}$$

Joint Probability Mass Function

- Suppose X and Y are two random variables, the joint probability mass function X = x1, x2... and Y = y1, y2… is expressed as:

$$p(x_i, y_j) = P\{X = x_i\ , \mathrm{Y} = y_j\}$$

- The above function can be used to obtain the probability mass functions of individual random variables X and Y
- Suppose, we want to find the value of probability mass function for X then, we reason that since Y must take some value yj therefore, the probability X<=xi can be obtained by summing up the values over all j's for mutually exclusive events $P\{X = x_i\ , \mathrm{Y} = y_j\}$

$$\{X = x_i\} = \bigcup_j \{X = x_i\ , \mathrm{Y} = y_j\}$$

$$P\{X = x_i\} = \mathrm{P}\left(\bigcup_j \{X = x_i\ , \mathrm{Y} = y_j\}\right)$$

$$= \sum_j p(x_i, y_j)$$

- Similarly, P{Y = yj} can be obtained as:

$$\sum_i p(x_i, y_j)$$

- Thus, knowing joint pmf we can always find out the pmf of individual random variables X and Y
- But, the reverse is not always true
- That is, knowing the individual pmf of two random variables cannot be used to obtain the joint pmf of the two variables

Example

Suppose that 3 batteries are randomly chosen from a group of 3 new, 4 used but still working, and 5 defective batteries. If we let X and Y denote, respectively, the number of new and used but still working batteries that are chosen, then find the joint probability mass function of X and Y , p(i, j) = P{X = i, Y = j}.

$$p(i,j) = \frac{\binom{3}{i}\binom{4}{j}\binom{5}{3-i-j}}{\binom{12}{3}}$$

$$p(0,0) = \frac{\binom{3}{0}\binom{4}{0}\binom{5}{3}}{\binom{12}{3}} = 10/220$$

$P\{X = i, Y = j\}$

i \ j	0	1	2	3	Row Sum $= P\{X = i\}$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
Column Sums $= P\{Y = j\}$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

Marginal of X

Marginal of Y

Example

- Suppose that 15 percent of the families in a certain community have no children, 20 percent have 1, 35 percent have 2, and 30 percent have 3 children; suppose further that each child is equally likely (and independently) to be a boy or a girl. If a family is chosen at random from this community, then, find the joint probability mass function of the two random variables B, the number of boys in the family and G, the number of girls in the family.
- Also find out the probability of having at least one girl.

- P(child = 0)=0.15, P(child = 1)=0.2, P(child = 2)=.35, P(child = 3)=.3
- P(B=0, G=0) = 0.15
- P(B=0, G=1) = P(child=1)P(G=1) = 0.2x0.5 = 0.1
- P(B=0, G=2) =
- P(B=0, G=3)
- P(B=1, G=0)
- P(B=1, G=1) = P(child=2)P(B=1)P(G=1) = 0.35x0.5 = 0.175 (two favourable cases (b,g) and (g,b) out of four (b,b), (g,g), (b,g), (g,b))

$P\{B = i, G = j\}$

i \ j	0	1	2	3	Row Sum $= P\{B = i\}$
0	.15	.10	.0875	.0375	.3750
1	.10	.175	.1125	0	.3875
2	.0875	.1125	0	0	.2000
3	.0375	0	0	0	.0375
Column Sum = $P\{G = j\}$	.3750	.3875	.2000	.0375	

Joint Probability Density Function

- Two random variables X and Y are said to be jointly continuous if there exists a function f(x,y) defined for every real x and y such that for every set C of pairs of real numbers

$$P\{(X,Y)\epsilon C\} = \iint_{(x,y)\epsilon C} f(x,y)dxdy$$

- The function f(x,y) is called the joint probability density function of X and Y

- If for a set of real numbers A and B we define C = {(x,y): xϵA and yϵB} then from the previous equation we have,

$$P\{(X,Y)\epsilon C\} = \int_B \int_A f(x,y)dxdy$$

Further,

$$F(a,b) = P\{X\epsilon(-\infty,a], Y\epsilon(-\infty,b]\}$$

$$= \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy$$

Probability Density Functions of X and Y

- If the two variables X and Y are jointly continuous then they are also individually continuous
- The probability density functions of the variables X and Y can be obtained as:

$$P\{X \epsilon A\} = \int_A \int_{-\infty}^{\infty} f(x,y) dy dx$$

$$\int_A f_X(x) dx$$

- Here, $f_X(x) = \int_{-\infty}^{\infty} f(x,y) dy$ is the probability density function of X
- Similarly, $f_Y(y) = \int_{-\infty}^{\infty} f(x,y) dx$ is the probability density function of Y

Example

- The joint density function of X and Y is given by

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & otherwise \end{cases}$$

- Calculate
  - P{X>1, Y<1}
  - P{X<Y}
  - P{X<a}

Solution

$$\int_1^{\infty}\int_0^{1} 2e^{-x}e^{-2y}\,dydx$$

$$\int_0^{\infty}\int_0^{y} 2e^{-x}e^{-2y}\,dxdy$$

$$\int_0^{a}\int_0^{\infty} 2e^{-x}e^{-2y}\,dydx$$

Independent Random Variables

- The random variables X and Y are said to be independent if for any two sets of real numbers A and B

$$P\{X\epsilon A, Y\epsilon B\} = P\{X\epsilon A\}P\{Y\epsilon B\}$$

- In other words, X and Y are independent if, for all A and B, the events $E_A = \{X\epsilon A\}\ and\ F_B = \{Y\epsilon B\}$ are independent

Joint Distribution Function

- Previous equation also implies

$$P\{X \leq a, Y \leq b\} = P\{X \leq a\}P\{Y \leq b\}$$

- Thus, in terms of joint distribution function of X and Y, continuous random variables X and Y are independent if

$$F(a, b) = F_X(a)F_Y(b)$$

- If X and Y are discrete random variables then, the condition of independence can be expressed as

$$p(x, y) = p_X(x)p_Y(y) \qquad \text{for all x,y}$$

- Here, p_X and p_Y are probability mass functions of X and Y

Example

- Suppose that X and Y are independent random variables having the common density function

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & otherwise \end{cases}$$

Find the density function of the random variable X/Y.

Solution

$$F_{X/Y} = P\{X/Y \leq a\}$$

$$= \int_{\frac{x}{y} \leq a} \int f(x,y)dxdy$$

$$= \int_{\frac{x}{y} \leq a} \int e^{-x}e^{-y}dxdy$$

$$\int_0^{\infty} \int_0^{ay} e^{-x}e^{-y}dxdy \;=\; 1 - \frac{1}{a+1}$$

- The joint probability density function can be obtained by differentiating the above equation with respect to a

$$f_{\frac{X}{Y}}(a) = \frac{d}{da}\left(1 - \frac{1}{a+1}\right) = \frac{1}{(a+1)^2}, \qquad 0 < a < \infty$$

Distribution over n random variables

- The joint probability distribution over n random variables are defined in the same manner as for n = 2
- The joint cumulative probability distribution function F(a1,a2…an) of the n random variables X1, X2,…, Xn is defined by

$$F(a1, a2, \dots, an) = P\{X1 \leq a1, X2 \leq a2, \dots, Xn \leq an\}$$

- If these random variables are discrete, their joint probability mass function is defined as

$$p(x1, x2, \dots, xn) = P\{X1 = x1, X2 = x2, \dots, Xn = xn\}$$

- If these random variables are continuous, their joint probability density function for any set C in n-space is defined as:

$$P(X1, X2, \dots, Xn) = \int_{(x1,x2,\dots,xn)\epsilon C} \int \dots \int f(x1, x2, \dots, xn) dx1 dx2 \dots . dxn$$

Independence

- The condition of independence can be defined from the above equation as follows
- N random variables X1, X2,…, Xn are said to be independent if for all sets of real numbers A1, A2, …, An,

$$P\{X1\epsilon A1, X2\ \epsilon A2, \dots, Xn\epsilon An\} = \prod_{i=1}^{n} P\{Xi\epsilon Ai\}$$

- The above condition is equivalent to:

$$P\{X1 \leq a1, X2 \leq a2, \dots, Xn \leq an\} = \prod_{i=1}^{n} P\{Xi \leq ai\}$$

- An infinite collection of random variables is independent if every finite sub collection of them is independent

Example

- Suppose that the successive daily changes of the price of a given stock are assumed to be independent and identically distributed random variables with probability mass function given by

$$P\{\text{daily change is } i\} = \begin{cases} -3 & \text{with probability .05} \\ -2 & \text{with probability .10} \\ -1 & \text{with probability .20} \\ 0 & \text{with probability .30} \\ 1 & \text{with probability .20} \\ 2 & \text{with probability .10} \\ 3 & \text{with probability .05} \end{cases}$$

If Xi denote the the change on the ith day, what is the probability that the stock's prices will increase successively by 1, 2 and 0 points in the next three days?

Solution

- $P\{X1 = 0, X2 = 1, X3 = 2\} = P\{X1 = 0\}P\{X2 = 1\}P\{X3 = 2\}$

$= (0.2)(0.1)(0.3) = 0.006$

Conditional Distributions

- For two events E and F, the conditional probability is given as

$$P(E|F) = \frac{P(EF)}{P(F)}$$

- We can use the same analogy to find the conditional distribution of two random variables
- If X and Y are discrete random variables, the conditional probability mass function of X given that Y=y, by

$$p_{X|Y}(x|y) = P\{X = x, Y = y\}$$

$$= \frac{P\{X = x, Y = y\}}{P\{Y = y\}}$$

$$= \frac{p(x,y)}{p_Y(y)} \qquad p_Y(y) > 0$$

Example

- If we know, in the previous example, that the family chosen has one girl, compute the conditional probability mass function of the number of boys in the family.
- Note that P(G=1) = 0.3875

$$P\{B=0|G=1\} = \frac{P\{B=0,G=1\}}{P\{G=1\}} = \frac{.10}{.3875} = 8/31$$

$$P\{B=1|G=1\} = \frac{P\{B=1,G=1\}}{P\{G=1\}} = \frac{.175}{.3875} = 14/31$$

$$P\{B=2|G=1\} = \frac{P\{B=2,G=1\}}{P\{G=1\}} = \frac{.1125}{.3875} = 9/31$$

$$P\{B=3|G=1\} = \frac{P\{B=3,G=1\}}{P\{G=1\}} = 0$$

Conditional Distribution of Continuous Random Variables

- If two continuous random variables X and Y have a joint probability density function denoted by f(x,y), then the conditional probability density function of X, given that Y = y, is defined for all values of y such that $f_Y(y) > 0$ as:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

Example

- The joint density of X and Y is given by

$$f(x,y) = \begin{cases} \frac{12}{5}x(2-x-y) & 0 < x < 1, 0 < y < 1 \\ 0 & otherwise \end{cases}$$

- Compute the conditional density of X, given that Y=y, where 0<y<1.

Solution

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

$$= \frac{f(x,y)}{\int_{-\infty}^{\infty} \frac{12}{5} x(2-x-y)dx}$$

$$= \frac{\frac{12}{5} x(2-x-y)}{\int_0^1 \frac{12}{5} x(2-x-y)dx}$$

$$= \frac{x(2-x-y)}{\frac{2}{3} - y/2}$$

Expectation

Definition

- One of the most important concept in probability theory
- For a discrete random variable X, the expectation of X or the expected value of X, denoted by E[X] is defined as:

$$\boldsymbol{E[X] = \sum_i x_i P\{X = x_i\}}$$

- Thus, expected value of X is the weighted average of the possible values that X can take on
- The weights are the probability of assuming the particular value X = xi

Example

- Suppose X is the random variable with pmf as below:
- p(0) = p(1) = ½

- Then, the expected value of X E[X] will be:

$$E[X] = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2}$$

- Thus, the expected value is simply the mean

Example

- Now, suppose the pmf of X is as below:
- p(0) = 1/3 and p(1) = 2/3

- Then, the expected value of X E[X] will be:

$$E[X] = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}$$

Interpretation

- Motivation for definition of expectation can be drawn from the frequency interpretation of probability
- In general terms, we can think of E[X] as representing the average value that the random variable will take when the experiment is repeated very large number of times

Example

- Find E[X] where X is the outcome when we roll a fair die.

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + \cdots 6 \times \frac{1}{6} = \frac{7}{2}$$

- Note that, 7/2 is not the value that the random variable can take when the experiment is performed
- Thus, it is not appropriate to assume that “expectation” of X represents the “expected value that X can take”
- It is better to understand E[X] as the average value of X over large runs of the experiment

Indicator Variable

- A random variable I is said to be an indicator random variable for an event A if

$$I = \begin{cases} 1 & if\ A\ occurs \\ 0 & if\ A\ does\ not\ occur \end{cases}$$

- In this case,

$$E[I] \quad = 0 \times \{1 - P(A)\} + 1 \times P(A) \ = P(A)$$

- That is, expectation of an indicator random variable is just the probability of occurrence of its associated event

Expectation of Continuous Random Variable

- The expectation of a continuous random variable can be defined in a similar manner
- Suppose X is a continuous random variable with pdf f
- For small value dx we have

$$P\{x < X < x + dx\} \approx f(x)dx$$

- It follows, that the expectation i.e. the weighted average of X over all possible values of X is nothing but integral of $xf(x)dx$ over all values of x. Therefore,

$$\boldsymbol{E[X] = \int_{-\infty}^{\infty} xf(x)dx}$$

Example

- Suppose that you are expecting a message at some time past 5 P.M. From experience you know that X, the number of hours after 5 P.M. until the message arrives, is a random variable with the following probability density function:

$$f(x) = \begin{cases} \frac{1}{1.5} & if\ 0 < x < 1.5 \\ 0 & otherwise \end{cases}$$

Solution

$$E[X] = \int_0^{1.5} \frac{x}{1.5}\, dx = 0.75$$

- Thus, on average the waiting time before the message is received will be three-fourth of an hour

Information carried by probability statement

- Entropy of a random variable can be roughly understood as the amount of information conveyed by the statement X = x
- Clearly the more unlikely that X =x, the more information is carried by the statement X = x
- E.g. if X represents the sum of values obtained by rolling two fair dice, then more information is conveyed by X = 12 than by X = 7,

$$P(X = 12) = 1/36 \quad \text{and} \quad P(X = 7) = 1/6$$

- We denote by *I(p)*, the amount of information contained in the message that an event, whose probability is p, has occurred

Entropy of a random variable

- We define I(p) as follows:

  $I(p) = -c \log_2 p,$ in general we keep c = 1

- Suppose, a random variable X takes values x1, x2…, xn with respective probabilities p1, p2,…, pn.
- The information conveyed by the message X = xi will be $-\log_2 p_i$
- Therefore, the expected amount of information conveyed when X is transferred is:

$$\boldsymbol{H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i}$$

- The quantity H(X) is used in information theory and denotes the entropy of the random variable X

Properties of Expectation

- We shall now look at the expected values of:
    1. A function of a random variable
    2. Sums of random variables

Expected value of a function of a RV

- Proposition:

a. If X is a discrete RV with pmf p(x), then for any real-valued function g,

$$\boldsymbol{E}[(\boldsymbol{g}(\boldsymbol{X})] = \sum_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x})\boldsymbol{p}(\boldsymbol{x})$$

b. If X is a continuous random variable with probability density function f(x), then for any real-valued function g,

$$\boldsymbol{E}[\boldsymbol{g}(\boldsymbol{X})] = \int_{-\infty}^{\infty} \boldsymbol{g}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})\boldsymbol{dx}$$

Example

- Suppose X has the following probability mass function

  p(0) = .2, p(1) = .5, p(2) = .3

Calculate E[X^2].

$$E[X^2] = 0^2 \times p(0) + 1^2 \times p(1) + 2^2 \times p(2)$$

$$= 0^2 \times 0.2 + 1^2 \times 0.5 + 2^2 \times 0.3$$

$$= 1.7$$

Example

- The time, in hours, it takes to locate and repair an electrical breakdown in a certain factory is a random variable— call it X — whose density function is given by

$$f_X = \begin{cases} 1 & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

- If the cost involved in a breakdown of duration x is x3, what is the expected cost of such a breakdown?

Solution

- We are required to identify the expected value of X^3

$$E[X^3] = \int_0^1 x^3 f(x) dx = \int_0^1 x^3 dx$$

$$= \frac{x^4}{4}\Big|_0^1 = \frac{1}{4}$$

Corollary

- If a and b constants, then

$$\boldsymbol{E[aX + b] = aE[X] + b}$$

Proof:

When X is discrete:

$$E[aX + b] = \sum_x (ax + b)p(x)$$

$$a\sum_x xp(x) + b\sum_x p(x)$$

$$= aE[X] + b$$

When X is continuous:

$$E[aX + b] = \int_{-\infty}^{\infty} (ax + b) f(x) dx$$

$$= a \int_{-\infty}^{\infty} x f(x) dx + b \int_{-\infty}^{\infty} f(x) dx$$

$$= aE[X] + b$$

Few Important Points

- In the above equation, if we put a = 0 then, E[b] = b
- That is, expectation of a constant is the constant itself
- Similarly, if we put b = 0 then, E[aX] = aE[X]
- That is, expectation of a constant times a random variable is equal to the constant times the expectation of the random variable
- The expected value of a random variable X, E[X] is also known as the mean or the first moment of X
- E[X^n] for n>= 1, denotes the nth moment and is defined as below:

$$\boldsymbol{E[X^n] = \begin{cases} \sum_{x} x^n p(x) & \textit{if X is discrete} \\ \int_{-\infty}^{\infty} x^n f(x)dx & \textit{if X is continuous} \end{cases}}$$

Expected Value of Sums of RVs

- If X and Y are two random variables and g is a function of X and Y, then

$$\boldsymbol{E[g(X,Y)] = \sum_{x}\sum_{y} g(x,y)p(x,y) \qquad for\ discrete\ X\ and\ Y}$$

$$\boldsymbol{= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy \qquad for\ continuous\ X\ and\ Y}$$

- Suppose, g(X,Y) = X + Y, then, in the continuous case

$$E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x,y)dydx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f(x,y)dxdy$$

$$= E[X] + E[Y]$$

- Similar result can be proven for discrete X and Y

Generalized Equation

- The previous equation, can be generalized for any n number of random variables as follows:
- For n random variables X1, X2,…, Xn, the expected value of the sum of the random variables is given by

$$\boldsymbol{E[X1 + X2 + \ \ldots + Xn] = E[X1] + E[X2] + \ \ldots + E[Xn]}$$

Example

- A construction firm has recently sent in bids for 3 jobs worth (in profits) 10, 20, and 40 (thousand) dollars. If its probabilities of winning the jobs are respectively .2, .8, and .3, what is the firm's expected total profit?

- Let Xi denote the random variables denoting the profit of the firm from job i
- Then, the profit of the firm is another random variable Y = X1 + X2 + X3
- We want to find the expected value of Y
- E[Y] = E[X1 + X2 + X3] = E[X1] + E[X2] + E[X3]
- E[X1] = 10x0.2 + 0x0.8
- E[X2] = 20x0.8 + 0x0.2
- E[X3] = 40x0.3 + 0x0.7

Variance

- Just like we summarize a data set by giving its arithmetic average or mean, a random variable can be summarized by giving its expected value E[X]
- The expectation of X is nothing but the average value of X
- The expectation of X however, does not give any indication about the variability of the values assumed by X
- Thus, two different random variables with very different variability in their possible values, can have same expected value
- For this reason, another quantity called variance of a random variable is defined

Definition

- Let X be a random variable with mean μ, then the variance of X, denoted by Var(X) is defined as

$$Var(X) = E[(X - \mu)^2]$$

From the above equation,

$$E[(X - \mu)^2] = E[X^2 + \mu^2 - 2\mu X]$$

$$= E[X^2] + E[\mu^2] - E[2\mu X]$$

$$= E[X^2] + E[\mu^2] - 2\mu E[X]$$

$$= E[X^2] + \mu^2 - 2\mu^2$$

$$that\ is, \qquad \boldsymbol{Var(X) = E\left[X^2\right] - \mu^2 = E\left[X^2\right] - (E[X])^2}$$

Example

- Compute Var(X) when X represents the outcome when we roll a fair die.

$$Var(X) = E[X^2] - (E[X])^2$$

X ={1,…,6}

P(X=1) =….=P(X=6) =1/6

Variance of Indicator Random Variable

- We have (for some event A)

$$I = \begin{cases} 1 & if\ A\ occurs \\ 0 & if\ A\ does\ not\ occur \end{cases}$$

$$E[I^2] = \ 0^2 \times \{1 - P(A)\} + \ 1^2 \times P(A) = P(A)$$

$$\begin{aligned} Var(I) &= E[I^2] - (E[I])^2 \\ &= P(A) - P(A)^2 \\ &= P(A)\{1 - P(A)\} \end{aligned}$$

Identity

$$\boldsymbol{Var(aX + b) = a^2 Var(X)}$$

Proof:

$$Var(aX + b) = E[(aX + b)^2] - (E[aX + b])^2$$

$$E[a^2X^2 + b^2 + 2abX] - (aE[X] + b)^2$$

$$a^2E[X^2] + b^2 + 2abE[X] - (a^2(E[X])^2 + b^2 + 2abE[X])$$

$$a^2(E[X^2] - (E[X])^2)$$

- By setting a = 0 in the above equation we get,

$$Var(b) = 0$$

- Thus, variance of a constant is zero
- By setting a = 1 we get,

$$Var(X + b) = Var(X)$$

- Thus, variance of a constant plus a random variable is same as the variance of the random variable
- Standard Deviation: The quantity $\sqrt{Var(X)}$ is defined as the standard deviation of the random variable X.

Covariance

- The covariance of two random variables X and Y, is defined as:

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

Here, μ_x and μ_y are the means of X and Y respectively

$$Cov(X,Y) = E[XY - \mu_x Y - \mu_y X + \mu_x \mu_y]$$

$$= E[XY] - \mu_x E[Y] - \mu_y E[X] + \mu_x \mu_y$$

$$= E[XY] - \mu_x \mu_y$$

$$\boldsymbol{Cov(X,Y) = E[XY] - E[X]E[Y]}$$

From the above eqn, we have

- Cov(X,Y) = Cov(Y,X)
- Cov(X,X) = Var(X)
- Cov(aX,Y) = aCov(X,Y)

Lemma

- Covariance has additive property

$$\boldsymbol{Cov}(\boldsymbol{X}+\boldsymbol{Z},\boldsymbol{Y}) = \boldsymbol{Cov}(\boldsymbol{X},\boldsymbol{Y}) + \boldsymbol{Cov}(\boldsymbol{Z},\boldsymbol{Y})$$

$$= E[(X+Z)Y] - E[X+Z]E[Y]$$

$$= E[XY+ZY] - \mathrm{E[X]E[Y]} - \mathrm{E[Z]E[Y]}$$

$$= E[XY] - E[X]E[Y] + E[ZY] - E[Z]E[Y]$$

$$= Cov(X,Y) + Cov(Z,Y)$$

Generalization

$$Cov\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} Cov(X_i, Y)$$

Above can be used to derive the following

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

Proof:

$$Cov\left(\sum_{i=1}^{n} X_i , \sum_{j=1}^{m} Y_j\right)$$

$$= \sum_{i=1}^{n} Cov(X_i, \sum_{j=1}^{m} Y_j)$$

$$= \sum_{i=1}^{n} Cov\left(\sum_{i=1}^{m} Y_j, X_i\right)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

Corollary

$$\boldsymbol{Var}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i}\right) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{Var}(\boldsymbol{X_i}) + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\substack{\boldsymbol{j=1} \\ \boldsymbol{i \neq j}}}^{\boldsymbol{n}} \boldsymbol{Cov}(\boldsymbol{X_i}, \boldsymbol{X_j})$$

- For n = 2,

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Covariance of Independent Random Variables

- If X and Y are two independent random variables, then

$$Cov(X, Y) = 0$$

And therefore, for X1, X2,…Xn

$$\boldsymbol{Var}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i}\right) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{Var}(\boldsymbol{X_i})$$

Proof

- We need to prove that E[X,Y] = E[X]E[Y]

$$E[XY] = \sum_i \sum_j x_i y_j P\{X = x_i, Y = y_j\}$$

$$\sum_i \sum_j x_i P\{X = x_i\} y_j P\{Y = y_j\}$$

$$\sum_i x_i P\{X = x_i\} \sum_j y_j P\{Y = y_j\}$$

$$E[X]E[Y]$$

Example

- Compute the variance of the sum obtained when 10 independent rolls of a fair die are made

$$Var\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} Var(X_i)$$

$$= 10\{Var(X_i)\}$$

Correlation

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

For x=1 and y=1

$$Cov(X,Y) = P\{X=1, Y=1\} - P\{X=1\}P\{Y=1\}$$

Now if Cov(X,Y) > 0

$$\Rightarrow P\{X=1, Y=1\} > P\{X=1\}P\{Y=1\}$$

$$\Rightarrow \frac{P\{X=1,Y=1\}}{P\{X=1\}} > P\{Y=1\}$$

$$\Rightarrow P\{Y=1|X=1\} > P\{Y=1\}$$

Correlation

- Thus, a positive covariance implies that the value of Y tends to increase when X is increased
- In order to standardize this and make the value dimensionless, we divide the covariance by the standard deviations of X and Y
- This quantity is known as Correlation

$$\boldsymbol{Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}}$$

- This quantity lies between -1 and 1

Moment Generating Function

- The moment generating function ϕ(t) of the random variable X is defined for all values of t by

$$\boldsymbol{\phi(t) = E[e^{tX}] = \begin{cases} \sum_{x} e^{tx} p(x) & \textit{if X is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx & \textit{if X is continuous} \end{cases}}$$

- ϕ(t) is called the moment generating function of X because all the moments of X can be generated by successively differentiating ϕ(t)

$$\phi'(t) = \frac{d}{dt} E[e^{tX}]$$

$$\phi'(t) = E[\frac{d}{dt} e^{tX}]$$
$$\phi'(t) = E[Xe^{tX}]$$
$$\phi'(0) = E[X]$$

$$\phi''(t) = \frac{d}{dt}\phi'(t) = E[X^2 e^{tX}]$$

$$\phi''(0) = E[X^2]$$

- similarly, nth moment can be derived as

$$\phi^n(0) = E[X^n], \qquad n \geq 1$$

Independent Random Variables

- If X and Y are two independent random variables then the moment generating of the sum of X and Y is the product of individual moment generating functions

$$\phi_{X+Y}(t) = E[e^{t(X+Y)}]$$

$$= E[e^{tX}e^{tY}]$$

$$= E[e^{tX}]E[e^{tY}]$$

$$= \phi_X(t)\phi_Y(t)$$

- **The moment generating function uniquely determines the distribution. That is, the each distribution has a unique moment generating function**

Distributions

Motivation

- We shall look at certain types of random variables that occur over and over again
- The probability distributions of such random variables are given special names
- Here, we shall look at some such random variables namely
- Bernoulli Random Variable
- Normal Random Variable
- Poisson Random Variable

Bernoulli Random Variable

- This random variable is associated with the experiments where only two outcomes are present
- We call these as “success” and “failure”
- Suppose the probability of “success” in the experiment is p then, the pmf of the random variable can be written as:

$$P\{X = 1\} = p$$

$$P\{X = 0\} = 1 - p$$

- Such a random variable is also known as Bernoulli Random Variable after Swiss Mathematician James Bernoulli
- The expected value of the Bernoulli random variable is the probability that it takes value 1 i.e. p

Binomial Random Variable

- Suppose, X represents the number of successes in the n independent trails of this experiment each having constant probability of success denoted by ‘p’ then, X is said to be a binomial random variable with parameters (n,p).

- The probability mass function of this random variable X can be given as:

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \qquad i = 0,1 \ldots n$$

Example

- It is known that disks produced by a certain company will be defective with probability .01 independently of each other. The company sells the disks in packages of 10 and offers a money-back guarantee that at most 1 of the 10 disks is defective. What proportion of packages is returned? If someone buys three packages, what is the probability that exactly one of them will be returned?

Solution

$$P\{X > 1\} = 1 - P\{X \leq 1\}$$

$$1 - \{P(X = 0) + P(X = 1)\}$$

- $1 - \left\{\binom{10}{0}(0.99)^{10} + \binom{10}{1}(0.01)(0.99)^{9}\right\} \approx 0.005$
- $\binom{3}{1}(0.005)(0.995)^{2} = 0.015$

Mean and Variance

The n independent trials can be represented using n Bernoulli random variables as below

$$X_i = \begin{cases} 1 & if\ ith\ trial\ is\ a\ success \\ 0 & otherwise \end{cases}$$

In the above case,

$$E[X_i] = p$$
$$Var(X_i) = E\left[{X_i}^2\right] - p^2$$
$$= p(1-p)$$

- The binomial random variable can be expressed as a sum of the n Bernoulli random variables, that is,

$$X = \sum_{i=1}^{n} X_i$$

Therefore,

$$E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = np$$

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p)$$

Bernoulli Distribution Function

$$P\{X \leq i\} = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}$$

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

$$P\{X = k+1\} = \binom{n}{k+1} p^{k+1} (1-p)^{n-k-1}$$

Taking ratio of the above two equations we get,

$$P\{X = k+1\} = \frac{p}{1-p} \frac{n-k}{k+1} P\{X = k\}$$

The above eqn can be used to calculate the distribution function

Poisson Random Variable

- A random variable X taking values 0,1,2… is said to be a Poisson random variable with parameter $\lambda, \lambda > 0$, if the probability mass function of X is of the form:

$$P\{X = i\} = \frac{e^{-\lambda}\lambda^{i}}{i!}, \qquad i = 0,1, \ldots$$

Probability Distribution Function, Mean and Variance

$$\sum_{i=1}^{\infty} p(i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1$$

$$\phi(t) = E[e^{tX}]$$

$$= \sum_{i=0}^{\infty} e^{ti} e^{-\lambda} \frac{\lambda^i}{i!}$$

$$e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!}$$

$$e^{-\lambda} e^{\lambda e^t} = \exp\{\lambda(e^t - 1)\}$$

- After differentiation we get,

$$\phi'(t) = \lambda e^t \exp\{\lambda(e^t - 1)\}$$

$$\phi''(t) = (\lambda e^t)^2 \exp\{\lambda(e^t - 1)\} + \lambda e^t \exp\{\lambda(e^t - 1)\}$$

- Putting t = 0, we get

$$\phi'(0) = E[X] = \lambda$$

$$Var(X) = \phi''(0) - (E[X])^2 = \lambda$$

Poisson Distribution Function

- If X is a random variable with mean λ, then

$$\frac{P\{X = i+1\}}{P\{X = i\}} = \frac{\left(e^{-\lambda}\lambda^{i+1}\right)/(i+1)!}{e^{-\lambda}\lambda^{i}/i!} = \frac{\lambda}{i+1}$$

Approximation of Binomial Random Variable

- Poisson random variable is useful because it may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small
- The parameter of Poisson random variable in this case will be

$$\lambda = np$$

Example

- Suppose that the average number of accidents occurring weekly on a particular stretch of a highway equals 3. Calculate the probability that there is at least one accident this week.
- X = number of accidents in a week
- P{X>= 1} = 1-P{X=0} = $1 - \frac{e^{-\lambda}\lambda^0}{0!} = 1 - e^{-3} =$

Example

- Suppose the probability that an item produced by a certain machine will be defective is .1. Find the probability that a sample of 10 items will contain at most one defective item. Assume that the quality of successive items is independent.
- P{X<=1} = P{X=0} + P{X = 1} = $\binom{10}{0}(0.9)^{10}+\binom{10}{1}(0.1)^{1}(0.9)^{9}$ =
- P{X<=1} = P{X=0} + P{X=1} = $e^{-\lambda} + \lambda e^{-\lambda} = e^{-\lambda}(1+\lambda) = 2e^{-1} =$

Normal Random Variables

- A random variable is said to be normally distributed with parameters μ and σ^2, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty$$

- Above forms a bell-shaped curve that is symmetric about the mean μ

$E[X] = \mu$ and attains its maximum value ($\frac{1}{\sigma\sqrt{2\mu}} \approx \frac{0.399}{\sigma}$) also at $x = \mu$

Why is Normal Distribution important?

- It was first used to approximate probabilities associated with binomial random variable when the parameter n is very large
- Later, this result was extended and it was found that, many random phenomena obey, at least approximately, a normal probability distribution
- Examples, height of a person, error made in measurement, velocity of molecule in any direction etc.

Mean and Variance

- The mean and variance of the normal random variable can be calculated with the help of moment generating function
- We get the values as:

$$E[X] = \phi'(0) = \mu$$

$$E[X^2] = \phi''(0) = \sigma^2 + \mu^2$$

Therefore,

$$Mean = E[X] = \mu$$

$$Variance = E[X^2] - (E[X])^2 = \sigma^2$$

Standard Normal Random Variable

- If X is a normal random variable with mean μ and variance σ^2 and if we define another random variable $\mathrm{Y} = \alpha X + \beta$ then, the following hold:

1. Y is also a normal random variable
2. Y has a mean $\alpha\mu + \beta$ and has variance $\alpha^2\sigma^2$

- Above result can be used to define a special type of normal random variable known as the standard normal variable Z, which has a mean 0 and variance 1, in the following manner:

$$Z = \frac{X - \mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma}$$

- $\alpha = 1/\sigma$
- $\beta = -\mu/\sigma$
- $\Rightarrow \alpha\mu + \beta = 0,$
- $that\ is\ mean\ of\ standard\ normal\ random\ variable\ is\ zero$
- $\alpha^2\sigma^2 = \frac{1}{\sigma^2} \times \sigma^2 = 1$
- $that\ is\ variance\ of\ the\ standard\ normal\ random\ variable\ is\ 1$

Distribution Function

- The distribution function of the standard or unit normal distribution is given as:

$$\phi(x) = \frac{1}{\sqrt{2\sigma}} \int_{-\infty}^{\infty} e^{-y^2/2} dy, \qquad -\infty < x < \infty$$

- The conversion to standard normal variable allows us to write the probability values for X in terms of Z
- For example, if we want to find the probability P{X<b} then it can be calculated by noting that if X<b holds then $\frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}$ also holds
- Therefore, the probability can be expressed as:

$$P\{X < b\} = P\left\{\frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right\} = P\left\{Z < \frac{b-\mu}{\sigma}\right\} = \phi\left(\frac{b-\mu}{\sigma}\right)$$

- Similiarly, P{a<X<b} can be written as:

$$P\{a < X < b\} = P\left\{\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right\} = \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)$$

- Thus, if we know values of $\phi(x)$ for different values of x then, probabilities for any normal random variable can be calculated
- The values of $\phi(x)$ calculated to some pre-defined accuracy are provided in most cases and we can use the for calculating probabilities for normal random variable X
- Further, the probability $\phi(-x)$ can be calculated by using the symmetry of the normal distribution along the mean
- We have,

$$\phi(-x) = P\{Z < -x\}$$

$$= P\{Z > x\} \qquad by\ symmetry$$

$$= 1 - \phi(x)$$

Example

- If X is a normal random variable with mean μ = 3 and variance $\sigma 2$ = 16, find
- P{X<11}
- P{X>-1}
- P{2<X<7}

$$P\{X < 11\} = P\left\{\frac{X-3}{4} < \frac{11-3}{4}\right\} = P\{Z < 2\}$$

$$= \phi(2) = 0.9772$$

- P{X>-1} = P{X<1}

$$P\{X < 1\} = P\left\{\frac{X-3}{4} < \frac{1-3}{4}\right\} = P\{Z < -0.5\} = P\{Z > 0.5\}$$
$$= 1 - P\{Z < 0.5\}$$

P{2<X<7}

$$= P\left\{\frac{2-3}{4} < \frac{X-3}{4} < \frac{7-3}{4}\right\} = P\{-0.25 < Z < 1\}$$

$$= P\{Z < 1\} - 1 + P\{Z < 0.25\}$$

Example

- The power W dissipated in a resistor is proportional to the square of the voltage V. That is,

$$W = rV^2$$

- where, r is a constant. If r = 3, and V can be assumed (to a very good approximation) to be a normal random variable with mean 6 and standard deviation 1, find
- E[W]
- P{W > 120}

Solution

$$E[W] = E[3V^2]$$
$$= 3E[V^2]$$
$$= 3(Var[V] + E^2[V])$$
$$= 3(1 + 36)$$

$$P\{W > 120\} = P\{3V^2 > 120\}$$
$$= P\{V > \sqrt{40}\}$$
$$= P\{V - 6 > \sqrt{40} - 6\}$$
$$= P\{Z > 0.3246\}$$
$$= 1 - \phi(0.3246)$$
$$= 0.3727$$

Sum of Normal Random Variables

- The sum of independent normal random variables is a random variable
- Its mean and variance are:

$$\mu = \sum_{i=1}^{n} \mu_i \qquad and \quad \sigma = \sum_{i=1}^{n} {\sigma_i}^2$$

Chi-Square Distribution

- If Z1, Z2,…Zn are n independent standard normal random variables, then X, defined by

$$X = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_n}^2$$

is said to have a chi-square distribution with n degrees of freedom and is denoted as

$$X \sim {\chi_n}^2$$

Distribution of Sampling Statistics

- In order to draw conclusion about the population from samples, we assume certain kind of relations to hold between the population and sample
- One such assumption is that there exists an underlying probability distribution of the population such that the measurable values from the population can be thought to be independent random variables having this distribution
- If the sample data is chosen from this distribution randomly, then we can assume that the samples are also independent random variables

Definition

- If X1, X2, … Xn are independent random variables having a common distribution F, then we say that they constitute a *sample* from the distribution F
- If the form of the underlying distribution is known and we are only interested in estimating its parameters then the inference process is known as Parametric inference
- If neither the parameter nor the form of the distribution is known, we call the inference non-parametric

Statistic

- A statistic is a random variable whose value is determined by the sample data
- We are interested in finding probability distributions of certain statistics
- Here, we shall discuss mean and variance

Sample Mean

- Suppose, we are obtaining data regarding some numerical quantity from the population such as height, age, annual income etc.
- Then, the values obtained corresponding to any element of the population may be regarded as a value of a random variable with mean μ and variance σ^2 where μ and σ^2 are population mean and population variance
- Let X1, X2, … Xn be a sample of values from this population, then the sample mean will be defined as

$$\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$$

Mean and Variance

- Since, the sample mean as defined is also a random variable we can calculate its mean and variance

$$E[\bar{X}] = E\left[\frac{X1 + X2 + \cdots + Xn}{n}\right] = \frac{E[X1 + X2 + \cdots + Xn]}{n} = \frac{1}{n}(n\mu) = \mu$$

$$Var(\bar{X}) = \mathrm{Var}\left(\frac{X1+X2+\cdots+Xn}{n}\right) = \frac{1}{n^2}[\mathrm{Var(X1)} + \mathrm{Var(X2)} + \cdots + \mathrm{Var(Xn)}]$$

(because of independence)

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- Thus, the variability of sample decreases as the size of sample increases

The Central Limit Theorem

- Let X1, X2,… Xn be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for large n, the distribution of

$$X = X1 + X2 + \cdots + Xn$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$

- This implies that the quantity

$$\frac{(X1 + X2 + \cdots + Xn - n\mu)}{\sigma/\sqrt{n}}$$

Is approximately a standard normal random variable. Thus for large values of n

$$P\left\{\frac{(X1 + X2 + \cdots + Xn - n\mu)}{\sigma/\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

Example

- An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

$$P\{X > 8.3 \times 10^6\}$$

$$= P\left\{\frac{X - 25000 \times 320}{540/\sqrt{25000}} > \frac{8.3 \times 10^6 - 25000 \times 320}{540/\sqrt{25000}}\right\}$$

$$= P\{Z > 3.51\} \approx 0.00023$$

Binomial Approximation

- We now have two possible approximations for binomial random variable –
- Poisson approximation holds when n is large and p is small
- Normal approximation holds when np(1-p) is large
- In general, it holds good for np(1-p)>= 10

Approximate Distribution of Sample Mean

- As a constant times a normal random variable is also a normal random variable, we can use the central limit theorem to state that the *sample mean is a normal random variable*.
- Thus,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Has an approximate standard normal distribution

Example

- The weights of a population of workers have mean 167 and standard deviation 27. If a sample of 36 workers is chosen, approximate the probability that the sample mean of their weights lies between 163 and 170.

$$P\{163 < X < 170\}$$

$$\mu = 167, \sigma = 27$$

$$\sigma/\sqrt{n} = \frac{27}{\sqrt{36}} = 4.5$$

How large a sample is needed?

- Practically, no matter how non normal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal
- In general, the normal approximation is valid for even smaller datasets

Sample Variance

- We already know that the sample variance is denoted by statistic S^2

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- It can be shown that E[S^2] = σ^2

Parameter Estimation

- Let X1, X2, …, Xn be a random sample from a distribution F_θ that is specified up to a vector of unknown parameters θ.
- The central problem is to make inferences about the unknown parameters
- One such method is the Maximum likelihood estimate for unknown parameters that gives point estimates
- We shall also look at an example of interval estimate

Maximum Likelihood Estimators

- Any statistic used to estimate the value of an unknown parameter θ is said to be an estimator of θ.
- Estimate is the observed value of the estimator
- Let X1, X2,…Xn, whose joint distribution is available except for an unknown parameter θ
- f(x1, x2,…xn) = $f_{X1}(x1)f_{X2}(x2) \dots f_{Xn}(xn)$
- $f(x_1, x_2, \dots x_n|\theta)$ represents the likelihood that values $x_1, x_2, \dots x_n$ will be observed when θ is the true value of the parameter
- MLE $\hat{\theta}$ is the value of θ that maximizes $f(x_1, x_2, \dots x_n|\theta)$ where $x_1, x_2, \dots x_n$ are the observed values
- $f(x_1, x_2, \dots x_n|\theta)$ is called the likelihood function

MLE estimator of Bernoulli Parameter

- Let us consider n independent trials of experiment each with a probability p of success

$$P\{X = x\} = p^x(1-p)^{1-x}, \qquad \text{x} = 0,1$$

$$\begin{aligned} f(x_1, \ldots, x_n|p) &= P\{X_1 = x_1, \ldots, X_n = x_n|p\} \\ &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\Sigma_1^n x_i}(1-p)^{n-\Sigma_1^n x_i}, \quad x_i = 0, 1, \quad i = 1, \ldots, n \end{aligned}$$

$$\log f(x_1, \ldots, x_n|p) = \sum_1^n x_i \log p + \left(n - \sum_1^n x_i\right) \log(1-p)$$

Differentiation yields

$$\frac{d}{dp} \log f(x_1, \ldots, x_n|p) = \frac{\sum_1^n x_i}{p} - \frac{\left(n - \sum_1^n x_i\right)}{1-p}$$

$$\frac{\sum_{1}^{n} x_i}{\hat{p}} = \frac{n - \sum_{1}^{n} x_i}{1 - \hat{p}}$$

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

MLE estimator of a Poisson Parameter

- Suppose X1, . . . , Xn are independent Poisson random variables each having mean λ

$$f(x_1, \ldots, x_n|\lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$
$$= \frac{e^{-n\lambda}\lambda^{\Sigma_1^n x_i}}{x_1! \ldots x_n!}$$

$$\log f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_1^n x_i \log \lambda - \log c$$

$$\frac{d}{d\lambda} \log f(x_1, \ldots, x_n|\lambda) = -n + \frac{\sum_1^n x_i}{\lambda}$$

$$\hat{\lambda} = \frac{\sum_1^n x_i}{n}$$

Example

- The number of traffic accidents in Berkeley, California, in 10 randomly chosen non rainy days in 1998 is as follows:

  4,0,6,5,2,1,2,0,4,3

- Use these data to estimate the proportion of non rainy days that had 2 or fewer accidents that year.
- Soln: First estimate the mean and then use it to calculate the probability P{X<=2}

MLE of Normal Population

$$\hat{\mu} = \sum_{i=1}^{n} x_i / n$$

$$\hat{\sigma} = \left[\sum_{i=1}^{n} (x_i - \hat{\mu})^2 / n \right]^{1/2}$$

- The MLE estimators of $\mu \; and \; \sigma$ are given by

$$\bar{X} \;\; and \left[\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}\right]^{1/2}$$

- It should be noted that the MLE estimator of standard deviation σ differs from the sample standard deviation S

$$S = \left[\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}\right]^{1/2}$$

- However, for n of reasonable size, these two estimators of σ are approximately equal

Interval Estimates

For the population σ is known while the mean is estimated as $\bar{X}$

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma}$$

$$P\left\{-1.96 < \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma} < 1.96\right\} = .95$$